ScienceWatch Home
Interviews

Featured Interviews
Author Commentaries
Institutional Interviews
Journal Interviews
Podcasts

Analyses

Featured Analyses
What's Hot In...
Special Topics

Data & Rankings

Sci-Bytes
Fast Breaking Papers
New Hot Papers
Emerging Research Fronts
Fast Moving Fronts
Research Front Maps
Current Classics
Top Topics
Rising Stars
New Entrants
Country Profiles

About Science Watch

Methodology
Archives
Contact Us
RSS Feeds

Inside This Month...

scienceWATCH.com
TRACKING TRENDS & PERFORMANCE IN BASIC RESEARCH

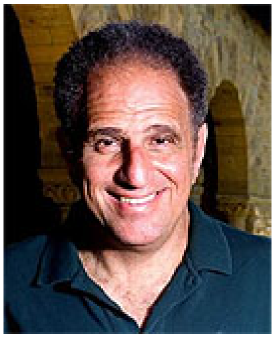
Interviews Analyses Data & Rankings

2008 : October 2008 - Author Commentaries : Stanford's Robert Tibshirani on Significance Analysis of Microarrays

AUTHOR COMMENTARIES - 2008

October 2008

PDF



"...microarray allows us to measure the expression of all the genes in the human genome in a cell with one assay."

Stanford's Robert Tibshirani on Significance Analysis of Microarrays

The *Science Watch*® Newsletter Interview

The introduction of a remarkable new technology into science can typically be divided into two phases. First, researchers begin to use it in ever-increasing numbers. Then they learn how to use it correctly. This was the case with DNA microarrays, introduced in 1995 with the promise of revolutionizing our understanding of gene functions. Rather than studying the expression of one or a few genes at once, researchers could study tens of thousands with their microarrays. "The breadth of array-based observations almost guarantees that surprising findings will be made," as one 2000 article in Nature *put it (405 [6788]: 827-36). But the critical question, of course, was determining what percentage of those surprising findings would turn out to be real.*

This was the question that lured Stanford University statistician and public-health scientist Robert Tibshirani into the fast-breaking field of microarray research. It also earned him a remarkable trifecta in *Thomson Reuter's* Essential Science Indicators*SM he's currently the third-most-cited author in mathematics, while ranked in the top 200 in clinical medicine and the top 250 in computer science. Tibshirani's top four papers of the last decade have collectively accumulated more than 7,500 citations. The leader, a 2001 article in* PNAS *titled "Significance analysis of microarrays applied to the ionizing radiation response," has garnered over 2,700 citations alone (see adjoining table).*

Tibshirani, 52, began his career studying math at the University of Waterloo in Ontario. He switched to statistics and computer science, he says, when he found that math got "too abstract, too quickly." He received his bachelor's degree in 1979 and followed it up with a master's in statistics from the University of Toronto a year later. In 1984, studying under Brad Efron, he obtained his Ph.D. in statistics from Stanford. Tibshirani then spent the next 14 years back at the University of Toronto, first as an assistant professor and then full professor, in statistics, preventive medicine, and public-health sciences. In 1998, Tibshirani returned to Stanford, where he now holds joint appointments in the Department of Health Research and Policy and the Department of Statistics.

From his office in Palo Alto, Tibshirani spoke to* Science Watch *correspondent Gary Taubes.

SW: What initially motivated you to take your expertise in statistics and apply it to public-health and biology applications?

Well, I always liked the applied aspect of statistics, especially medical applications and subjects related to biology and health. When I came to Stanford in 1998, I got interested in genomics. It's a hotbed for this kind of research.

SW: Your most highly cited papers, by far, involve significance analysis of microarrays (SAM). How did you get started with SAM and its role in microarray analysis?

I collaborated with Gil Chu, who's an oncologist here at Stanford. These were the early days of gene-expression microarrays. Gil and his student, Virginia Tusher, did an experiment and had a problem they needed help with. They had patients who would get radiation treatment for skin cancer, and they found that for some people the radiation treatment was actually worse than the disease itself. So they wanted to use microarrays to be able to predict ahead of time which patients would have these severe reactions. They tried to study it with microarrays.

From the statistical point of view, it was just like comparing two groups, except that with microarrays you're not looking at how one parameter changes between them—you're looking at how tens of thousands change. You're faced with the problem of how to compare thousands of parameters between different groups. Gil, Virginia, and I developed a statistical method that worked well for the context of their particular experiment. Then we realized that this was something people could use quite generally.

At the same time, I was getting inundated with requests from people around Stanford to help them with their microarray experiments, and I realized that I needed a software package that I could give to people so I wouldn't be doing this same kind of analysis over and over—these thousands of T-tests, as they're called, for every microarray study. So that's what I did, with the help of my colleague Balasubramanian Narasimhan.

"There are many questionable analyses published these days," says Robert Tibshirani of Stanford University. *"As science gets very big and expensive, there's a lot of pressure to get positive results."*

SW: For those of us who aren't as up on our statistics as we should be, could you explain what a T-test is?

You have a parameter, such as the expression of a gene, and you've measured it in a bunch of normal samples and, say, a bunch of samples from tumors. You want to know if the parameter is higher or lower in the tumors. So you take the average of the normals and the average of the tumors, and you compare them. If the average is 2 in the normals and 3.5 in the tumors, then you might think, okay, this parameter is higher in tumors—maybe we should pay attention. But you don't know if that difference is very large or something that you're likely to see just by chance.

So a T-test takes the difference between two averages and compares it to how much the parameter actually varies within the normals or within the tumor group. If, within the normal group, the parameter varies, for example, from 2 to 2.1, and in the tumor group it varies from 3 to 3.1, then the difference of 2 to 3 between groups is probably important. But if the normals are varying from 2 to 10 and it's the same with the tumors, then the 2 to 3 difference is probably not that meaningful. That's what a T-test tells you. And this is the basis of what you're asking in gene-expression studies. You want to answer that question about a gene under particular circumstances.

SW: Why does this become problematic in microarray analyses?

Well, the microarray allows us to measure the expression of all the genes in the human genome in a cell with one assay. So, suppose I take 10 patients with breast cancer and 10 normals, and then take a sample of their breast tissue and measure the expression of all the genes in those breast cells. My data then consists of something like 20,000 genes, all the genes in the human genome, measured in each of 20 patients, and I want to know which of these 20,000 genes are more active, or significantly less active, in the cancer group versus the controls. I can do the T-test, but

Highly Cited Papers by Robert Tibshirani and Colleagues, Published Since 2000
(Ranked by total citations)

Rank	Papers	Cites
1	V.G. Tusher, R. Tibshirani, G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *PNAS*, 98 (9): 5116-21, 24 April 2001.	2,817
2	A.A. Alizadeh, *et al.*, "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature*, 403(6769): 503-11, 3 February 2000.	2,665
3	T. Sorlie, *et al.*, "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications," *PNAS*, 98(19): 11 September 2001.	1,298

4	J.D. Storey, R. Tibshirani, "Statistical significance for genomewide studies," *PNAS*, 100(16): 9440-5, 5 August 2003.	938
5	T. Sorlie, *et al.*, "Repeated observation of breast tumor subtypes in independent gene expression data sets," *PNAS*, 100(14): 8418-23, 8 July 2003.	653

SOURCE: Thomson Reuters *Web of Science*®

I'm doing it for each of 20,000 genes in the array, not just one, so now I have a multiple-comparison problem. I have to worry about false positives: even if all the genes are null—if every gene behaved in tumors just as it does in normals—we would still expect that some genes would be significantly higher, just by chance, and some genes would be lower in the tumors versus the normals. So how do I account for the fact that I'm looking at the most extreme values out of 20,000 possibilities?

SW: And this is what significance analysis for microarrays does?

Yes. SAM gives you a way of estimating the false-positive, or "false-discovery" rate. You put data for 20,000 genes into the algorithm and it tells you the 100 genes, let's say, showing the largest change in tumors versus normals. And it also tells you that, in the list of 100 genes, you can expect a false-discovery rate of, say, 10%, which means that out of 100 genes, probably 10 are false positives. The other 90 are true positives. That's useful information for the experimenter. It says that the experiment was well done; it didn't have much noise.

But if the false-discovery rate is, for example, 50%, then that means probably half of these genes—half of this list—are likely false positives. It says maybe you shouldn't spend a year in the lab studying these particular 100 genes—you should probably go back and do a new experiment, maybe get more tumor samples, improve the lab procedures to get less noise. So it tells you how informative your experiment is in terms of the false-discovery rate.

SW: What are the most significant factors affecting the false-discovery rate in microarray experiments?

Probably the biggest factor is sample size. The human genome, as I've mentioned, has maybe 20,000 to 30,000 genes in it. You're trying to study them—you want to know about all of those genes. But there could be hundreds, maybe thousands, involved in any specific disease process. The problem is that the number of human samples you get for any one disease is usually in the tens or hundreds. Now, it would be nice if you had samples from a million tumors. Then you could have a very low false-discovery rate. But that's expensive, so typically you get a few hundred, and there's biological variability in there as well. In fact, they're not all the same disease, but different flavors of the same disease. So that's the main problem: small sample sizes.

SW: Are you concerned about the quality of the statistical analyses done in this kind of high-throughput genomics and proteomics?

Yes. There are many questionable analyses published these days. As science gets very big and expensive, there's a lot of pressure to get positive results. In some significant proportion of papers published, the analyses I see are exaggerated or just plain wrong.

For example, I had an experience a couple of years ago with a paper that appeared in *Science*. The authors had sequenced something like 30 colon cancers and 30 breast cancers, looking for single nucleotide polymorphisms, or SNPs, present in the tumors. They claimed to have found something like 140 new SNPs. I did a reanalysis of this with graduate student Holger Hoefling, along with Gaddy Getz, Todd Golub, and Eric Lander of the Broad Institute, and we found that their analysis was wrong. They made some very fundamental errors in estimation of false-discovery rates. Two other groups reanalyzed the paper and concluded the same thing we did. [Note: for discussion, view here] In this kind of statistics, you can get any answer you want just by perturbing your analysis in some way.

SW: How do you get around this problem? In your opinion, what in particular has to be done to improve the quality of analyses?

Well, one big problem is that people don't publish the scripts of their analyses. So if you want to know what people did in an analysis, you have to spend a lot of time reconstructing it from their description in the paper. A typical paper will say, "Here's what we did: we took this data, filtered it in this way, applied so-and-so program," and so forth. You'd think this would be enough to reconstruct it, but you'd be surprised.

And the point is, particularly with high-dimensional data, just a small change in the analysis can cause large change in results. So the details make a real difference. For instance, which version of the software program did they use? It sounds easy to take the data from their paper and actually replicate the answers that they got, but it's not!

SW: But papers are limited in length, so what's an author to do?

I think authors should at least be required to make the script of their analyses available on a supplemental web site. It's like a lab book, which says, "In step three, I put in this chemical, etc." If I see two papers, and one publishes a script and the others don't, even if I don't understand the script, the fact that the authors were willing to put it forward is a sign of good faith; maybe I can trust that paper. If there's no script published, I would wonder if the authors are trying to hide something. Did they have to do the analysis 400 different ways before they got the answer they wanted? So, publishing scripts would be a real step forward. It would help to clean up a lot of the shoddy stuff that's out there, because people wouldn't be able to hide behind their analyses anymore.

Another major issue is the file-drawer problem. A study is well done, but it doesn't find anything, so it's harder to publish, and it gets filed in a drawer somewhere and never sees the light of day. You can imagine 100 studies that look for chemical causes of cancer, and 99 are null. The one that gets published is the one that finds something. I've talked to my friend Patrick Brown of the Howard Hughes Medical Institute about this. [Note: see *Science Watch*, 13(3): 3-4, May/June 2002.] He's one of the founders of the Public Library of Science, PLoS, which has a series of journals, all open-access. *PLoS ONE* is the express journal. I suggest that they need a *PLoS ZERO* for null studies, those that find nothing. That way if someone reports that hot dogs cause breast cancer, you can immediately do a search and find out if 500 other studies looked at the association and found nothing. That's important information. You can't know the truth without it. It's critical to good science. It needs to be published somewhere.

SW: Your papers have been extremely influential in a trio of fields. Is there any one that gives you the most satisfaction?

Well, here's a different take on that question: my son Ryan is now a first-year graduate student in statistics at Stanford. We're the first parent-child combo ever in the department. Obviously this makes me both happy and proud, particularly because I never pushed him into the field; he did some summer internships in bio labs analyzing data, and fell in love with statistics on his own. The other thing is I've had some wonderful collaborators, especially Trevor Hastie, Jerome Friedman, and Brad Efron. The first two are coauthors of my book, *The Elements of Statistical Learning*. This book has sold very well, across many fields, and is probably the single scientific accomplishment that I'm most proud of.

Keywords: Robert Tibshirani, Rob Tibshirani, Stanford University, microarrays, significance analysis of microarrays, SAM, SNPs, T-test, genomics, genome-wide analysis.

PDF

back to top

2008 : October 2008 - Author Commentaries : Stanford's Robert Tibshirani on Significance Analysis of Microarrays

Scientific Home | About Scientific | Site Search | Site Map

Copyright Notices | Terms of Use | Privacy Statement