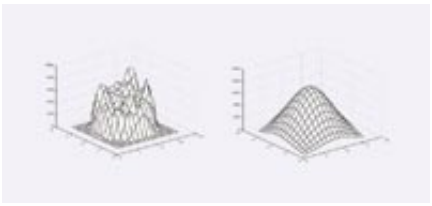
Genomics & Proteomics

An Old Insight Leads to New Tools

Woes in genomics and proteomics research include the lack of robust results researchers are willing to trust for further work. The ideas of an eighteenth-century English vicar are driving new and sophisticated statistical approaches directed at those woes

By Mark Greener
Mark Greener is a freelance writer based in Cambridge, UK.

Thomas Bayes seems an unlikely person to help transform the statistical rigor of experiments in the vanguard of microarray analyses and proteomics. For one thing, he died almost 250 years ago in Tunbridge Wells, an English town that remains sleepy even today. And rather than being a scientist, Bayes was a nonconformist minister. Yet today's microarray and proteomic researchers increasingly use the statistical approach bearing his name and the related maximum entropy model. They are finding it helps resolve some of the worries that their results are not robust enough to withstand further scrutiny.



click the image to enlarge

The first step in the spot quality control method developed by Tampere University of Technology is to extract the features using Gaussian fitting. The figure on the left is the original spot distribution. The figure on the right is the resulting two-dimensional Gaussian fit, which allows users to determine several features that facilitate quality control. (Source: Signal Processing Algorithm Group, Tampere University of Technology)

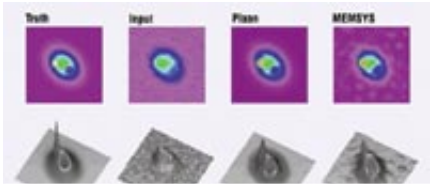
Certainly, a plethora of problems face microarray researchers. Microarray papers are often based on replicates in single figures. Missing values can influence clustering, classification, and network design. Quality control may be unsystematic, whereas expression levels often prove highly sensitive to experimental conditions. A widely used statistical approach can miss subtle but potentially important findings, so researchers increasingly apply sophisticated statistical techniques that solve these fundamental problems and allow results to withstand the rigors of peer review.

"Microarray research suffers from a flood of results [on which] nobody is really willing to follow up, because they sense the data might not be robust," says Paul Pavlidis, PhD, assistant professor of biomedical informatics at Columbia University, New York. For example, removing even a single sample from a small microarray data set can yield entirely different, often unreliable, results.

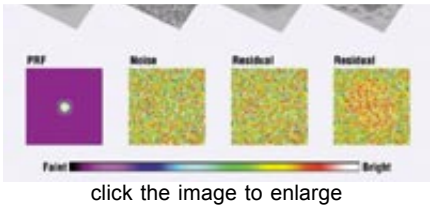
How many replicates?
In a recent study, for example, Pavlidis and collaborators used random sampling of 16 published gene expression microarray experiments. They found that having fewer than five replicates rarely produces stable results. On the other hand, running more than 10 to 15 replicates usually yields little further improvement in stability.

"Some researchers prefer to treat microarrays as an initial coarse screening method. They then go back to the lab to weed out false positives, and, if they are fortunate, confirm a fraction of their findings. This [approach] will tend to miss subtle but potentially important findings, while successfully identifying the 'most robust' effects," Pavlidis says. In this case, he shoots for a higher level of specificity and sensitivity by doing more replicates. While this approach incurs more expense at the outset, he says, the eventual gains in biological insight as well as in reduced effort and frustration outweigh the initial investment.

In many such cases, researchers use analysis of variance (ANOVA) to produce "plausible" results. But the small number of replicates means that the results can fail to reach statistical significance. Essentially, ANOVA estimates the difference between the means of two or more sets of results and accounts for the measurements' variability. Thus, a large difference between means might be unreliable if the variance is very large. Similarly, small differences may be statistically significant if the variance is small. The researcher generates a *p*- value, which is an estimate of the probability that the difference would arise by chance. "ANOVA is simple, fast, and generally accepted as a good method," Pavlidis says. "Other methods tend to be more experimental, not as well standardized, or at least, more difficult to use than ANOVA." On the other hand, more powerful statistical analysis might find a difference that ANOVA missed.



The vicar's contribution
Thomas Bayes (1702-1761) established the theoretical foundation for the method to infer an event's probability on the basis of the frequency at which it occurred in the past. After Bayes' death, his friend Richard Price sent the paper to the Royal Society in London, and it was published in the Society's Philosophical Transactions in 1764. This laid the foundation of modern Bayesian statistics, which allows researchers to include incomplete or partial data. They can update it as more information becomes available.



click the image to enlarge

The Pixon method begins with a "maximum likelihood" technique that produces too much information but provides a good fit to the data. The system then asks what information can be taken away and still fit. This is iterated until the fit is good everywhere and no more information can be removed. (Source: Pixon)

In genomics and proteomics, Bayesian methods help solve, for example, the problem posed by small sample sizes when using ANOVA and help quantify gene expression. "In DNA microarray analysis we don't ultimately care what the technical error variances are," says Jeffrey Townsend, PhD, a research fellow at the Miller Institute for Basic Research In Science, University of California, Berkeley. "But we would like to acquire the best estimates of gene expression levels, integrating across all realistic error variances in proportion to their 'reasonableness.' Bayesian methods offer the best method to achieve an estimate with error bars." He adds that the next step is to analyze a broad set of Bayesian approaches to determine the best model for DNA microarray analysis.

Furthermore, users can develop a Bayesian model based on their understanding of the interactions between components in a biological system. Soumya Raychaudhuri, MD, PhD, of Stanford University, Stanford, Calif., notes that in genetic networks, each variable is a gene, the expression of which is influenced by the expression of other genes. Building such networks requires so-called training data. During training, researchers run the Bayesian model using well-characterized examples before analyzing test data.

Bayesian analysis can be misleading if the defined model is not appropriate, Raychaudhuri says. "However, Bayesian analysis offers the user great insight into the data. Not only can it make predictions, but it is easy to see which key points influence the prediction." Indeed, a growing number of genomics researchers routinely employ Bayesian methods to solve problems that can arise during analysis of microarray data. Here is a look at several examples of Bayesian analysis put to work.

Case 1: Bayesian analysis hones sequence analysis
In a typical application, Bayesian analysis helps predict secondary structures from the primary protein sequence. In this case, the wealth of data allows researchers to train the Bayesian method very effectively. "For instance, in the case of predicting the secondary structure of a protein, it is well understood which other amino acid positions might be most influential," Raychaudhuri says. There is also general consensus regarding the appropriate statistical models for sequence analysis.

Case 2: Bayesian analysis compares gene expression
Townsend used microarray and Bayesian methods to compare gene expression levels among four natural isolates of the wine yeast *Saccharomyces cerevisiae* , from Montalcino, Italy. He showed variations in amino acid metabolism, sulfur assimilation and processing as well as protein degradation. In many cases, the difference in gene expression was less than two-fold. It remains an open question, though, to what extent differential gene expression allows the organism to adapt to new or changing environments.

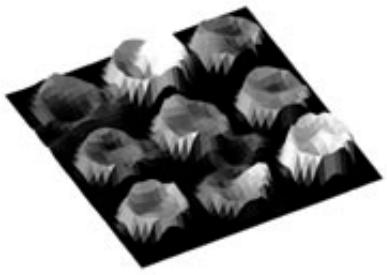
Case 3: Bayesian analysis tackles outliers
Raphael Gottardo, a PhD candidate in statistics at the University of Washington, Seattle, uses Bayesian methods to detect and place less weight on outliers in gene expression analysis. Outliers can occur because of scratches or dust on the slide, imperfections in the glass or array production. "If an experiment only has, let's say, four replicates, it is almost impossible to know if a data point is an outlier or the gene is just more variable," he says. In a paper submitted for publication and available on his Web site, Gottardo developed a Bayesian model that uses all the genes to detect outliers.

Using two sets of publicly available gene expression data, Gottardo found that, compared to raw log ratios, the Bayesian approach reduced between-replicate variability by 64% and 83%. Compared to other methods (including one using ANOVA), the Bayesian approach reduced between-replicate variation by more than 55% compared to the best alternative method.

Case 4: Bayesian analysis and microarray quality control
Even if only 1 per cent of microarray spots are of poor quality, researchers could face hundreds of unreliable ratios. However, researchers have not yet reached a consensus about the characteristics that comprise good quality spots or experiments. Partly, this reflects the dearth of studies on technical issues, such as quality control. To a certain extent, this reflects differences between labs' ad hoc QA (quality assurance) rules. And then there are differences due to the scanners and image analysis software used. Against this background, a Bayesian network developed by the Biological Data Analysis Group at the Institute of Signal Processing, Tampere University of Technology, Tampere, Finland, offers one of the first steps towards creating a systematic way to assess quality of a spot and the whole microarray slide.

One relatively simple quality assessment uses predefined cut-offs for spot size and intensities. For example, a spot could be considered unreliable if its area is under 50 pixels. Alternatively, a spot could be considered unreliable if both Cy3 and Cy5 intensities are below 100 fluorescent units. These cut-offs depend on normalizing the results against the microarray's controls, known as "within-slide normalization," to allow for differences in the experimental conditions. "Quality control with predefined cut-offs leads to an intriguing problem. The data should be filtered for quality before within-slide normalization. But the data should be within-slide normalized before quality filtering," says group leader Sampsa Hautaniemi, PhD.

Hautaniemi's approach resolves this paradox using predefined features, such as signal-to-noise ratio, spot morphology, and background intensity. The Bayesian network uses these features to classify each spot into a quality category: bad, medium, or good. "The Bayesian network strategy forces researchers to respond explicitly to several assumptions that are usually hidden behind statistics, such as the need for a training set. Quality control based on Bayesian networks allows researchers to incorporate new discrete and continuous features, which may



A three dimensional close-up of nine spots on a microarray consisting of 13,824 spots. x- and y-axis denote coordinates on a microarray slide and z-axis represents intensity. Some

not be easy using strategies based on non-Bayesian statistics," Hautaniemi says.

His group is currently assessing if ratios of low-intensity spots deviate more than ratios measured from higher intensity spots. They are applying the method to frequently used data sets, such as Spellman's cell cycle and trying to apply quality control based on Bayesian networks to emerging technologies, such as protein microarrays.

Maximum entropy
The maximum entropy method (MEM), a variation on the Bayesian theme, seeks to extract as much information from a measurement as is justified by the signal-to-noise ratio. Entropy refers to lack of order, so MEMs seek the minimum structure (order) that remains consistent with the data and the prior knowledge.

"The maximum entropy method is a Bayesian signal-processing technique that seeks the 'best' representation of a limited, noisy data set in terms of a usually linearly related quantity that facilitates interpretation of the measurement," says Peter Steinbach, PhD, chief of the Center for Molecular Modeling in the Center for Information Technology, at the National Institutes of Health, Bethesda, Md. "The goal is to transform only the meaningful information embodied in the measurement to the more convenient representation while leaving the statistical noise behind."

This statistical noise means, however, that several representations could be consistent with a data set. So the MEM chooses the one possessing the minimum structure. "Maximum entropy methods assume that the best models are those with the highest entropy that are still consistent with the training data," says Raychaudhuri. "Maximum entropy analysis avoids over committing to a sparse data set."

background and morphological problems are visible: several spots are connected to each other (bleeding). Also, the intensity distribution of many spots is vulcano-like rather than uniform, which may cause ratios to be unreliable. To reduce the impact of these problems to results, quality for each spot on a microarray slide should be assessed. (Source: Signal Processing Algorithm Group, Tampere University of Technology)

MEMs depend on a default representation that embodies the researcher's prior expectations for the measurement. For example, the researcher might expect that the representation is as uniform or as smooth as possible. Modifying this default representation allows researchers to examine a range of interpretations consistent with the measurement. Two typical cases exemplify the principal in practice.

Case 5: MEMs helps analyze protein kinetics
MEMs can aid analysis of a protein's kinetic motion, the link between structure and function. Time-resolved spectroscopy can assess conformational changes and protein folding to account for "transitions among intermediate states and any other processes that contribute to the measured signal." To achieve this, researchers need to determine the number, amplitude and rate of each kinetic process. Steinbach and colleagues [P. J. Steinbach, R. Ionescu, C. R. Matthews, Biophys J., vol. 82, pp. 2244-2255 (2002)] found that applying the MEM revealed six kinetic processes during the folding of dihydrofolate reductaseone more than previously reported.

Case 6: MEMs in complex classification tasks
For example, Raychaudhuri says, researchers classifying pathological specimens may need to consider the expression levels for several thousands of genes. Similarly, to assign gene ontology codes, they must consider the presence or lack of a particular word. In both cases, Raychaudhuri says, the researcher considers thousands of data points, with rarely more than about a few hundred training examples.

Beyond maximum entropy
The MEM has limitations, however. The calculations used in MEMs are intrinsically iterative. In other words, the computation is repeated numerous times, so MEMs can be computationally intensive and too slow for some high-throughput applications. Moreover, MEMs impose an average value for the information content in each location of the image. But the information content varies across the image. As a result, in some locations a MEM removes real picture information if there is more than this average. At others, it creates artificial, false information to make up the difference.

A method developed by Pixon LLC, Setauket, N.Y., avoids this artefact. According to Richard Puetter, PhD, chief technology officer with Pixon, the Pixon method uses a "maximum likelihood" technique that produces too much information, but provides a good fit to the data. The system then asks what information can be taken away and still fit. The program iteratively asks: "Where do I need to introduce more information because I have a bad fit, and where can I take information away and still maintain a good fit." This is iterated until the fit is good everywhere and no more information can be removed. As a result, the model is no more complicated than needed. As Pixon's method doesn't spend time calculating things that aren't real, Puetter claims it is faster than MEMs.

MEMs are often used with a uniform default representation. They can also use a "blurred" version of an intermediate result from MEMs. In this case, all parts of the image are blurred in the same way. In this case, however, the MEM can tend to over-smooth sharp features and under-smooth slowly varying features. In a recent paper, Steinbach proposes a method that counters this tendency and helps assess whether ripples are real or artifacts. "This approach is subjective, but can help determine the range of interpretations that are consistent with the data," Steinbach says.

As microarray technology matures, sophisticated statistical approaches should become as familiar as ANOVA or the Student t-test. But the pace of research regularly throws up novel problems. "This is what I like about the area," Gottardo says. "New problems arise everyday." One thing seems certain, however. The approach pioneered by an English vicar almost 250 years ago is set to remain an important part of the answer.

Organizations mentioned in this article:

A brief introduction to Bayes
Thomas Bayes, original essay
Peter Steinbach, PhD, Chief of the Center for Molecular Modeling in the Center for Information Technology, NIH
Steinbach: a brief introduction to the maximum entropy method
Lab of Jeffrey Townsend, Miller Institute for Basic Research In Science, University of California,Berkeley

Raphael Gottardo, University of Washington, home page
Research of Raphael Gottardo
Hautaniemi's research, Biological Data Analysis Group at the Institute of Signal Processing, Tampere University of Technology, Finland
SpotQuality
Institute of Signal Processing
Paul Pavlidis, PhD, Columbia University
Pixon LLC
A collection of Bayesian songs, hosted by Brad Carlin (includes: 'Biostat Division Blues,' 'These are Bayes,' and 'Bayesian Believer')