Meeting report

Making sense of microarrays

James N Siedow

Address: DCMB/Biology, Levine Science Research Center, Duke University, Durham, NC 27708-91000, USA. E-mail: jim.siedow@duke.edu

Published: 7 February 2001

Genome ***Biology*** 2001, **2(2)**:reports4003.1–4003.2

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2001/2/2/reports/4003

A report on the 'Critical Assessment of Microarray Data Analysis' (CAMDA 2000) meeting, Durham, North Carolina, USA, December 18-19, 2000.

Researchers gathered at Duke University on 18-19 December 2000 to review and critically evaluate methods for analyzing the large data sets generated from gene expression microarrays. As anyone who has perused the pages of *Genome Biology* can attest, the amount of information contained within a microarray data set seems overwhelming at first glance and no formal standards for their analysis have been established. The CAMDA 2000 meeting brought together scientists from such diverse fields as mathematics, statistics, computer science and the still somewhat ill-defined field known as bioinformatics. In addition, there were a number of biologists in attendance, including me. As I am a biochemist/molecular biologist, there were two features of this meeting that I had never encountered previously. First, as a way of normalizing the various analytical methods being compared, speakers and poster presenters alike were supposed to make use of one of two published sets of microarray data: those from Golub *et al.* (*Science* 1999, **286:**531-537) and Spellman *et al.* (*Mol Biol Cell* 1998, **9:**3273-3297). I gather this is not an uncommon practice among computational scientists, and makes a fair amount of sense. Second, the meeting participants selected, by secret ballot, the best presentation at the meeting - an interesting exercise.

Considering the two papers: Golub *et al.* analyzed a microarray data set of roughly 7,000 genes from a series of 38 leukemia patients with the dual goals of assigning leukemias to known classes (class prediction) and identifying new tumor classes (class discovery). They could distinguish between acute myeloid leukemia and acute lymphoblastic leukemia when they focused on the expression patterns of a subset of 50 genes. When their analysis was applied to a test set of microarrays from 34 additional patients, they were able to correctly predict the nature of the leukemia in all but four cases. Spellman *et al.* looked to create a catalog of all yeast genes whose transcript levels varied periodically (and systematically) during the course of the cell cycle. They further analyzed the roughly 800 genes identified as being cell-cycle-regulated to look for both known and new promoter elements that could be predictive of cell-cycle regulation.

In addition to the 15 talks addressing the two data sets, opening remarks from John Weinstein (National Cancer Institute, National Institutes of Health, Bethesda, USA) and closing remarks from Athel Cornish-Bowden (Centre National de la Recherche Scientifique, Marseilles, France) were presented and there were two keynote lectures. Weinstein did a particularly nice job of setting the context of the meeting by noting that he had never seen a field where the literature was so far ahead of the reality. He felt this was a result of the fact that not only are the analytical methodologies being applied to microarrays under active development but the technology itself is far from becoming stabilized, and it is still uncertain which technologies will survive at the end of the day. He saw this state of flux as being a healthy thing for the field and one that will continue to exist for the near term. Weinstein also attempted to categorize the field of bioinformatics, in which he currently sees two sub-fields, applied and developmental bioinformatics. The former is primarily the province of biologists and mathematical scientists, whereas the latter includes development of both algorithms and software and requires a synergy between computer and computational scientists. Cornish-Bowden's closing remarks were wide-ranging, focusing more on functional genomics and less on the analysis of microarrays *per se*. He also pointed out that even if microarray data indicate changes in the level of any single enzyme, these have only a limited ability to significantly alter metabolic flux rates, but that they can dramatically affect metabolite concentrations, and he suggested methodologies for addressing this issue.

The two keynote speakers each addressed broad issues related to microarray analysis. Mike West (Duke University, Durham, USA) used the analysis of expression arrays from a set of breast cancer patients to point out where future analytical improvements were needed, including better methods for selecting useful gene subsets for any given analysis and accounting for measurement errors. These were themes that remained in play throughout the meeting. Gavin Sherlock (Stanford University, USA) brought more of a biological perspective to his presentation and stressed the importance of the need for good database annotation if the vast amount of information contained in the ever-expanding number of microarrays is to be of broad use to biologists.

With respect to the 15 presentations on the two datasets, I came away with a sense that the Golub *et al.* data set may have been a bit too 'easy' to analyze. All but two of the speakers focused on this data set and the majority were able to do at least as well as Golub *et al.* in classifying the test set of leukemias (31/34 patients correctly classified) and many speakers correctly predicted all but one of the subtypes. This led some to question whether the one outlier might not be incorrectly classified clinically. Interestingly, the most accurate analyses were those that focused on a more limited subset of genes than the 50 used by Golub *et al.*, leading several speakers to also suggest that including too many genes in the subset being analyzed results in a reduced ability to discriminate between the two leukemia types. In fact, simply using the expression behavior of only one gene, that encoding zyxin, correctly classified 31 of the 34 patients. It seems clear that as microarrays are used increasingly for clinical diagnosis, learning how to identify which subset of genes to focus on for which diagnostic purpose will become increasingly important.

Although most of the talks were well presented, I will confess to often not having followed their statistical nuances. On the other hand, the panel discussions at the end of each half-day's set of talks were always quite interesting because the questions that arose addressed the many unresolved issues facing those currently developing approaches to microarray analysis. One point that came up repeatedly was the need for replication (for example, there was no replication in the data set of Golub *et al.*) to get a better sense of the variability associated with the data. This, in turn, led to the observation that in microarray studies, replication is needed for many reasons, including chip-to-chip variability, variability in RNA isolation, and tissue-to-tissue variability along with the related problem of tissue heterogeneity. The panel discussions also repeatedly emphasized the fact that the 'best' approach for analyzing any given data set will be related to the question being asked of that data set - a point that should keep many statisticians occupied for some time to come. For the record, the award for Best Presentation went to Chris Stoeckert (University of Pennsylvania, Philadelphia, USA) who developed a non-parametric approach he named PaGE (Pattern from Gene Expression [http://www.cbil.upenn.edu/PaGE]). Full details of the conference can be found on the CAMDA 2000 home page [http://bioinformatics.duke.edu/camda/].

Although this meeting represented the first of its kind, it is probably unfortunate that the Golub *et al.* study was so focused on the specific question of classifying the leukemias, because that drove most of the participants' analyses. Nevertheless, it was a good first effort, and I suspect that additional meetings having this format will appear regularly in the future as mathematical, computer and biological scientists continue to develop and use microarray technology.