Meeting report

The miracle of microarray data analysis

Yuk Fai Leung, Dennis Shun Chiu Lam and Chi Pui Pang

Address: Department of Ophthalmology and Visual Sciences, The Chinese University of Hong Kong, 147K Argyle Street, Kowloon, Hong Kong.

Correspondence: Chi Pui Pang. E-mail: cppang@cuhk.edu.hk

Published: 29 August 2001

Genome ***Biology*** 2001, **2(9)**:reports4021.1–4021.2

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2001/2/9/reports/4021

A report on the tenth Annual Bioinformatics and Genome Research meeting of the Cambridge Healthtech Institute's Beyond Genome 2001 series, San Francisco, USA, 17-19 June 2001.

Life scientists are facing an unprecedented data-analysis challenge. The large datasets from genomic or proteomic studies are not easily analyzed by traditional one-by-one genetics. As a result, we have to bring together expertise from statistics, mathematics, bioinformatics, and computer science to handle such tasks. This conference successfully facilitated interaction between experts from various fields, sparked off many new ideas, and addressed challenges from several areas, such as microarray data mining, gene prediction, and pathway reconstruction.

Several speakers emphasized the importance of getting better tests of statistical significance for the patterns of differentially expressed genes discovered using microarrays. Thomas D. Wu (Genentech Inc., San Francisco, USA) started the conference by reviewing various types of supervised and unsupervised learning methods for analyzing microarray data. Tom Downey (Partek Inc., St. Charles, USA) pointed out that commonly used visual-analysis methods are limited because their interpretation is subjective and the statistical significance of the results is not measurable. As a result, such methods are difficult to automate. Downey discussed several statistical tests, such as Student's *t* test and analysis of variance (ANOVA) that can be used to identify significantly different gene expression in different experimental conditions. He recommended applying statistical methods to microarray data in order to automate data analysis and make it objective. Kenneth Hess (University of Texas, Houston, USA) explored the role of replication in experiments. Using the fact that the variation in intensity ratio between replicates is inversely proportional to the magnitude of the ratio, he identified the differentially expressed genes from his microarray experiments as those with intensity-ratio variation different from the expected variation. He too suggested using *t*-statistics to identify the genes that are most reliably differentially expressed. Stanley N. Cohen (Stanford University, USA) has developed a rule-based system for microarray data analysis, which allows users to define a series of rules prior to the analysis, from their knowledge of experimental conditions, gene expression patterns and functions, and correlations between the expression of different genes. The rules are applied uniformly to the dataset, avoiding any subjective interpretation of the unsupervised gene groupings. This system greatly helps the user to understand the reasons for the assignment of genes to clusters.

David Haussler (University of California Santa Cruz, USA) reviewed the progress of the assembly of human-genome working draft by the public and private projects and the controversial question of gene-number estimation. The real number of genes will never be certain until they are all experimentally verified. Daniel Shoemaker (Rosetta Inpharmatics Inc., Kirkland, USA) described a feasible method for experimental annotation of the human genome using ink-jet oligonucleotide arrays, for both exon-only and tiling arrays (the latter consist of overlapping sequences that cover an entire genomic region). In the exon array, 15,511 different 60-nucleotide probes, derived from 8,183 predicted exons from chromosome 22q, were used to screen for expression of exons. Expression of 567 genes was verified by this approach. The tiling array contains 60-nucleotide probes that overlap in 10-nucleotide steps. The hybridization process not only verified the expression of the exons but also efficiently confirmed the splice-site boundaries. David Kulp (Affymetrix Inc., Santa Clara, USA) further stressed the importance of genome sequence analysis for designing the

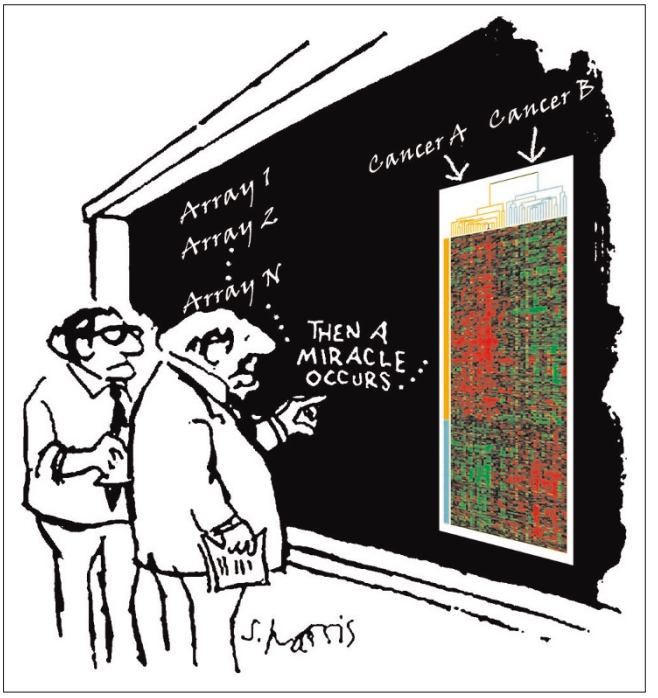

Figure 1
"I think you should be more explicit here in step two." Modified with permission from a cartoon by Sidney Harris and from an image provided by Patrick Brown.

best oligonucleotides for manufacturing DNA chips capable of differentiating between similar members of a gene family.

A better-annotated human genome sequence is also a valuable tool for the reconstruction of molecular pathways. Edgar Wingender (BIOBASE Biological Databases GmbH, Braunschweig, Germany) and colleagues have focused their efforts on developing the TRANSFAC and TRANSPATH databases as an integrated knowledge base for allocation of genes and gene products into regulatory networks. Understanding of upstream regulatory elements and transcription factors, when combined with experimental data from microarray experiments, has provided a powerful method for reconstructing cellular pathways (see, for example, Wolfsberg *et al.*, *Genome Res* 1999, **9:**775-792).

Life scientists often feel that it is miraculous to be able to analyze and make sense of the tremendous amount of data generated from microarray or other genomic studies (see Figure 1). Although most appreciate that such data mining can provide a better global picture of biological systems and clues for future research directions, some still try to avoid involvement in such an important task and leave the statisticians, mathematicians or computer scientists to deal with it. Should we still wait for more miracles to come, or should we work closely with experts from other fields to formulate better and more biologically relevant analyses?