Meeting report

Microarray gene expression database: progress towards an international repository of gene expression data

Paul Kellam

Address: Wohl Virion Centre, Department of Immunology and Molecular Pathology, Windeyer Institute, University College London, London W1T 4JF, UK. E-mail: p.kellam@ucl.ac.uk

Published: 2 May 2001

Genome ***Biology*** 2001, **2(5)**:reports4011.1–4011.3

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2001/2/5/reports/4011

© BioMed Central Ltd (Print ISSN 1465-6906; Online ISSN 1465-6914)

A report on the third Microarray Gene Expression Database group meeting (MGED3), Stanford University, Palo Alto, California, USA, 29-31 March, 2001.

In November 1999, the European Bioinformatics Institute (EBI; Hinxton, UK) organized the first Microarray Gene Expression Database group meeting (MGED) at Hinxton. At that meeting the idea was established that the world-wide microarray community desired a public database that would allow the submission, sharing and analysis of microarray data. It was recognized that this was a huge undertaking and that it required the participation of the whole microarray community, from small groups to large labs, in order to realize this vision. One and a half years later, MGED3 hosted by Stanford University (USA) revealed just how close we are to having such a resource. MGED initially divided the work associated with its aims into five working groups, namely: experimental description and data representation standards; microarray data XML exchange format; ontologies for sample description; normalization and quality control; and future user group (now disbanded). Three pre-meeting workshops were held on microarray data annotation, ontologies and data formats. These provided opportunities for members of the working groups to contribute to the future direction and content of their groups. The MGED3 meeting was organized into reports on the advances of the working groups and also included expert seminars on different aspects of array work, so as to provide a cross-disciplinary educational element.

The main meeting began with a keynote address by Pat Brown (Stanford University Medical School, USA), who presented his earlier work on understanding the complexities of transcriptional control at the level of the promoter and mRNA turnover rates. Brown questioned what the best 'filter' for array analysis is. According to Brown one clear requirement is whether the results make biological sense. This was qualified, however, when Brown referred to the often overly simple 'cartoonish' view of biological networks that exists at present. This was only too apparent in Brown's studies of yeast polyphosphate biosynthesis: he commented on the misnomer of 'non-essential genes' in a given phenotype and that "parachutes are only essential if you are falling out of an aircraft". As illustrated by Brown, in microarray studies and other functional genomics methods, context is everything, because an observed phenotype is specific for the conditions under study. Brown also reiterated his commitment to open and free science publishing (akin to *Genome Biology*'s philosophy): at present scientists produce, review, pay to publish and pay to read their work for the profits of journals and this is often paid for with public funds.

Transcriptome analysis of different regions of the brains from mice that have different behavioral phenotypes was presented by David Lockhart (Salk Institute, La Jolla, USA). The talk showed the need for the use of replicate samples in array experiments and for careful analysis of the array data before it is possible to gain meaningful biological insights; in this case a program called 'Data Triage' was used for analysis. Michael Radmacher (National Cancer Institute, Bethesda, USA) illustrated the processes and statistical methods needed to achieve class prediction from array data. Class prediction involves identifying significant groups of genes that allow the accurate prediction of a phenotype (or class), for example which genes are predictive of a particular cancer type compared to another related but distinct cancer type.

A number of available microarray databases were presented at the meeting including Gene Expression Omnibus (GEO) [http://www.ncbi.nlm.nih.gov/geo/] (Alex Lash, National Center for Biotechnology Information (NCBI), Bethesda, USA), ArrayExpress [http://www.ebi.ac.uk/arrayexpress/]

(Ugis Sarkans, EBI), GeneX [http://www.ncgr.org/research/genex/] and the DNA Data Bank of Japan [http://www.ddbj.nig.ac.jp/] (Yoshio Tateno, National Institute of Genetics, Mishima, Japan), but currently only GEO is able to act as a public repository where array data can be submitted and an accession number is issued.

Margaret Gardiner-Garden (Entigen Corporation, Sydney, Australia) documented in her poster presentation a comparison of many of the available array databases that provides a good starting point for newcomers to the field who have array results and wish to deposit them in a database. Natalia Novoradovskaya (Stratagene, La Jolla, USA) presented a poster on the use of universal human reference RNA for two-channel microarray experiments. She described the progress by Stratagene towards producing and quality-controlling a pool of nine RNA preparations from different cell lines that can be used as a standard reference for microarrays. If such a common RNA standard is embraced by array labs, this will go some way towards facilitating array comparisons in the future. A full list of poster and presentation abstracts can be found at the third MGED meeting website [http://www.dnachip.org/mged3].

Experimental description and data standards

Alvis Brazma (EBI) reported on the progress of the MIAME (Minimal Information About a Microarray Experiment) working group. The group has now collectively developed a draft specification (MIAME 1.0) of information that should be reported about any microarray experiment to ensure interpretability of the results. The principles of MIAME require that microarray data are revealed in sufficient detail and with sufficient annotation to be of use to third parties. The full draft of MIAME 1.0 is available from the Annotations Working Group website [http://www.mged.org/Annotations-wg/index.html].

Microarray data XML exchange format

Paul Spellman (University of California, Berkeley, USA) discussed MAML (MicroArray Markup Language), a data format used for information transfer; it is based on the widely used web language XML. The XML working group has been developing an XML DTD (document type definition) for communicating microarray data in a platform- and database-independent manner. XML is an emerging standard for the structuring of documents: in short, XML documents consist of elements that are textual data structure tags. The XML DTD describes the structure of elements (tags) of an XML document (see also Achard *et al.*: *Bioinformatics* 2001, **17:**115-125). MAML can incorporate all of the recommendations laid out in MIAME 1.0. The MAML DTD is due to be submitted to the OMG (an international standards organization), and the working group is currently negotiating a final data standard with other interested parties. The current draft MAML DTD is available at SourceForge [http://sourceforge.net/] and at the National Center for Biotechnology Information MAML Specification page [http://www.ncbi.nlm.nih.gov/geo/maml/].

Ontologies and array annotation

Chris Stoeckert (University of Pennsylvania, Philadelphia, USA) reported progress in the array annotation and ontologies working group. Stoeckert's report revealed how difficult and time-consuming it is to develop an ontology for reporting microarray-sample and experimental details. (An ontology is a structural vocabulary that represents a specification of a conceptualization design to allow the reuse of information across multiple applications and implementations; see also Karp PO: *Bioinformatics* 2000, **16:**269-285.) Provisional ontologies are available from the MGED3 Ontology Working Group web pages [http://www.cbil.upenn.edu/Ontology/MGED_ontology.html] and [http://www.cbil.upenn.edu/Ontology/MGED3OWG.htm], and Stoeckert welcomed the increased use of these ontologies by MEGD members, who seem enthusiastic about using the ontologies with their own array experiments. The power of ontologies for knowledge discovery was well illustrated by Michael Ashburner (EBI) of the Gene Ontology (GO) consortium. Their aim is to produce a dynamic controlled gene-description vocabulary that can be applied to all eukaryotes. GO has already been used by the model-organism genome databases FlyBase [http://flybase.bio.indiana.edu/], *Saccharomyces* Genome Database [http://genome-www.stanford.edu/Saccharomyces/], and the Mouse Genome Database [http://www.informatics.jax.org]. The use of GO terms in annotating arrayed genes in a future international array database will clearly provide a very powerful resource. Peter Karp (SRI International, California, USA) talked about ontologies for genetic networks using EcoCyc.

Data normalization and experimental design

Roger Bumgarner (University of Washington, Seattle, USA) and John Quackenbush (The Institute for Genomic Research, Rockville, USA) provided notable highlights of the meeting. Bumgarner presented his latest methods to normalize array data from two-color microarrays using a curve-fitting method to take into account the non-linear relationship, at low and high signal levels, of Cy3:Cy5 or Cy5:Cy3 ratios (that is, the ratios of changes of 'test' samples relative to 'reference' samples). These methods will soon be available in the form of software, but can also be implemented manually in Microsoft Excel. Quackenbush provided a wonderful example of the use of existing methods, such as hierarchical clustering, K-means clustering, self-organizing maps and principal-component analysis to analyze array data. This was all performed through a demonstration of a user-friendly, workstation-based program called TIGR Multi Experiment Viewer (TMEV), which is freely available from

the Institute for Genome Research Software pages [http://www.tigr.org/softlab/]. The beauty of this talk was the use of a 'synthetic' dataset that enabled the clear and easy visualization of what each different analysis method actually does to the data. As a research tool this looks superb. But perhaps more importantly, the use of the program and the synthetic dataset are just the place for people to start to explore and 'play' with array data. Only by becoming comfortable with the different analysis algorithms can a researcher proceed with confidence. For anyone involved in teaching microarray-analysis methods, Quackenbush's material is also the place to start.

At the close of MGED3 it was agreed that the blend of updates, tutorials and invited speakers provided an excellent basis for this sort of cross-disciplinary meeting. MGED4 will happen next year, probably on the East coast of the United States, and is certainly a must for people involved in microarray research.