nature genetics
volume 23 no. 3 november 1999

Connecting the dots

Carina Dennis

Mark Boguski scans the horizon.

Carina Dennis

Stuart Kim susses out *C. elegans*.

As expressed by Francis Collins (director of the National Human Genome Research Institute (NHGRI)) in *The Chipping Forecast*[1], it is too early to predict the ultimate impact of microarrays on our understanding of biology, and genetics in particular. Data presented at a recent *Nature Genetics* conference* indicate that estimates of impact and rate of progress should not be conservative. Two previous microarray meetings[2] focused more heavily on technological aspects. Whereas aspects of the technology received some airtime at this year's meeting, the substantive nature of presentations that focused on biological questions indicates that the field, such as it is, is advancing quickly. Abstracts of oral and poster presentations are freely available (http://genetics.nature.com/microarray).

Comprehensive analysis in yeast. The most extensive array analyses have been carried out in *Saccharomyces cerevisiae*, where approximately 250 sequence-specific regulators interact with at least 100 components of the transcription apparatus to effect the expression of over 6000 genes[3]. The hope is that microarray analysis will permit a detailed understanding of the transcriptional pathways that underlie cellular metabolism. From the data to hand, it is clear that gene expression patterns are clustered; groups of genes are expressed in similar patterns throughout the cell cycle and under a variety of environmental conditions. The hypothesis that these genes are co-regulated at the transcriptional level is corroborated by a growing number of studies.

Joseph DeRisi (University of California, San Francisco) described his approach to understanding the coordination of the 'proteasome' gene cluster, whose members encode all known proteasome subunits, through seeking sequences conserved between promoters. He discovered a conserved non-degenerate sequence common to all members of the cluster, and, by way of a modified one-hybrid screen, identified RPM4 (a putative transcription factor) as a candidate proteasome regulator. Whereas yeast lacking RPM4 are viable, gratifyingly, deletion mutants are more sensitive to proteasome inhibitors. Comparing expression profiles of stressed-out wild-type and *rpm4* mutants confirmed that *RPM4* affects the synthesis of all bona-fide proteasome components, in addition to genes not previously implicated in protein degradation. DeRisi plans to use similar strategies to identify the regulatory elements and the respective transcriptional regulators of genes in other clusters—with the aim of obtaining a comprehensive view of transcriptional patterns (for an updated version of advice and protocols see http://cmgm.stanford.edu/pbrown/mguide/).

Richard Young (Massachusetts Institute of Technology (MIT)) uses a similar approach to enable 'circuit discovery' (ref. 3; see also http://web.wi.mit.edu/young/

**The Microarray Meeting—Technology, Application and Analysis, Scottsdale, Arizona, September 22–25, 1999.*

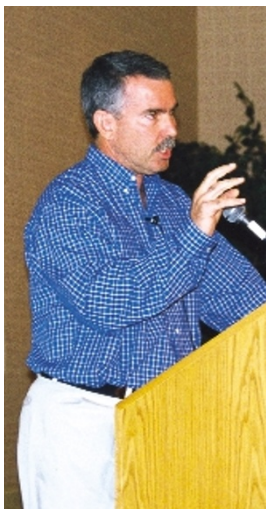

Michael Ronemus

Eric Lander classes cancers.


Carina Dennis


Michael Ronemus

Patrick Brown and **Michael Eisen**: "More data are good."


Carina Dennis

Jeffrey Trent: mining melanoma

expression/); he described time-course experiments designed to identify signal-transduction events downstream of a specified environmental change. Using mutant yeast strains in which particular genes were deleted or rendered temperature-sensitive, he verified the roles of individual signalling molecules and transcription factors. Young also described forays into the world of the nucleosome, which is thought to repress transcription through modifying chromatin structure. Surprisingly, he discovered that histone depletion causes derepression of only approximately 15% of the genes assayed, and, moreover, represses a significant fraction (around 10%). The expression of the remaining genes seems unaffected by nucleosome density and positioning. Reassuringly, Young also mentioned that he finds a high degree of consistency between his results (obtained using Affymetrix oligonucleotide arrays) and those obtained by others with cDNA arrays.

Arraying the worm. Multicellular organisms provide an additional layer of complexity to the challenge of designing experimental systems and framing questions. *Caenorhabditis elegans* is the only multicellular animal with a fully sequenced genome, and Stuart Kim (Stanford University) hopes that array analysis will help to annotate the function of many of its 19,099 genes. As with yeast, a wealth of well-characterized mutants allows comparison between normal and 'perturbed' systems, and material is not limiting—a million worms, typically used to prepare target material, are relatively easy to collect. Attempting a molecular description of germ-cell development, Kim compared hermaphrodite wild-type worms with three mutant strains and, interrogating 12,000 genes, discovered distinct sets whose mRNAs are enriched in sperm and oocytes. Approximately 15% of the sperm-enriched class encode kinases and phosphatases, suggesting that phosphorylation has a substantive role in sperm development. Oocytes specifically express genes known to be involved in inductive signalling pathways, consistent with maternal proteins governing early axis formation. Taking advantage of the ease of performing functional depletion experiments in the worm, the role of individual genes can now be tested.

Kim's experiments are feasible because the germ line makes up a substantial fraction of the worm's cells, and an even larger fraction of some of the mutants. It is difficult to detect changes in expression patterns if the proportion of target cells is small (and impossible if mRNA levels increase in some cells and decrease in others). As the task to dissect a million worms does not inspire, Kim and colleagues are working on strategies of obtaining tissue-specific gene expression patterns. They are also expanding their arrays to represent all of the worm's genes and are offering to probe mRNA samples of others. A database (http://cmgm.stanford.edu/~kimlab/wmdirectorybig.html) of *C. elegans* expression data allows researchers to compare their own results—or mine for information on their favourite genes, potentially gaining leads *in silico*.

Experiments on higher organisms (and their organs) are at a more preliminary state, but data presented by David Lockhart (Affymetrix), who compared gene expression in brains of wild-type and mutant mice, and Jonathan Pevsner (Johns Hopkins University), who seeks to identify pathways relevant to Rett syndrome and autism, indicated a low level of brain-to-brain variation between control samples and reproducible differences in samples from mouse mutants and human patients.

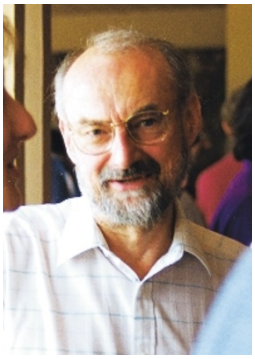
Human disease mechanisms. Experiments are sometimes inspired by a well-defined question, leading to a hypothesis such as: "triplet repeat expansion upstream of *DMPK* causes myotonic dystrophy by altering *DMPK* expression". Rolf Krahe (Ohio State University) used oligonucleotide arrays to test this hypothesis—the alternative being that expanded repeats bind cellular proteins critical to appropriate expression of various genes and that sequestration of these proteins results in global expression deregulation and loss of function of critical genes. Comparing muscle tissue from patients and controls, Krahe obtained expression profiles that support the latter view.


Carina Dennis

Francis Collins: keying up for the future


Michael Ronemus

David Lipman: it takes one to tango?


Carina Dennis

Edwin Southern navigates a coffee break.


Carina Dennis

John Quackenbush emphasizes the need to detect small changes in gene expression.

In contrast are experiments to which Thomas Shenk (Princeton University) jokingly referred as 'ignorance-driven' research. He used expression profiles of cells infected by human cytomegalovirus with a view to obtaining a better understanding of virus-host interactions[4]. Surprisingly, intact virus and viral particles inactivated by ultraviolet radiation induced identical transcriptional responses in infected cells, indicating that viral structural components are critical in terms of eliciting cellular response. Several host targets regulated by viral entry are likely to protect the virus from immune attack. Curiously, however, one cluster of induced genes encodes interferons and related molecules. This seems inconsistent with a viral strategy of avoiding immune response—until one realizes that some viral genes contain interferon-response elements and depend upon a transient surge of interferons to induce transcription.

High-throughput genotyping using single-nucleotide polymorphisms is another application of microarrays with extraordinary potential. Theoretical considerations (Leonid Kruglyak, Fred Hutchinson Cancer Research Center) complemented by comprehensive results on variation within specific genes (Aravinda Chakravarti, Case Western Reserve University; Robert Lipshutz, Affymetrix) will allow the design of studies surveying human variation and exploring its influence on disease susceptibility.

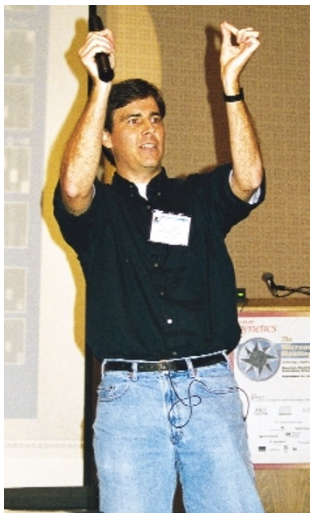
Cancer classification. Despite recent advances in cancer biology, diagnosis, prognosis and selection of appropriate treatment for a particular cancer still amount to a daunting task; cancers that are pathologically indistinguishable have disparate behaviours and react differently to treatment. There is hope that microarray-based approaches will aid genetic classification and diagnosis and provide insights into the molecular events underlying tumour development and progression.

Eric Lander (MIT) presented data that demonstrated class discovery and class prediction of acute myeloid leukaemia and acute lymphoid leukaemia[5]. These are morphologically indistinguishable but biologically distinct. Comparing their expression patterns using oligonucleotide arrays, he selected 50 'predictor' genes differentially expressed in the two cancers that direct correct classification of random samples. The method can also be used to discover classes whose significance can then be tested by selecting a set of predictor genes and determining predictive strength on a random sample. Along similar lines, Louis Staudt (National Cancer Institute) reported how the 'lymphochip', an array of cDNA probes enriched for their expression by lymphocytes, can reveal 'diseases within a disease'. Using results obtained with this array, he was able to subdivide B-cell malignancies and observed a significant difference in the survival rate of patients with different tumour sub-types.

Patrick Brown (Stanford University) described efforts to develop a cancer taxonomy based on expression profile. Comparing the expression patterns of over 50 primary breast cancers, several normal breast tissue samples and cell lines representing different cell types of the breast, he too found that pathologically indistinguishable tumours have highly specific expression signatures[6]. A set of informative genes was subsequently used to analyse and compare a mix of over 60 individual samples. Strikingly, samples from a tumour before and after treatment were found to be much more similar to one another than to any of the other samples. Similarly, the profiles of primary tumours and lymph node metastases from the same patient bore a closer relationship than with those obtained from different patients. These results suggest that biological heterogeneity between tumours is matched by measurable molecular diversity. The challenge will be to relate specific molecular variation to specific variation in tumour phenotype.

Cancer biology. Having observed that the expression patterns of melanoma cell lines and primary tumours bear a close similarity, Jeffrey Trent (NHGRI) went on to determine that a subset of melanomas are characterized by low-level expression of genes encoding proteins that mediate cell migration and invasion. Preliminary

Joseph De Risi talks transcription.



David Lockhart: "There is nothing wrong with a fishing expedition as long as you are trying to catch fish."

migration assays confirmed that tumour cells obtained from the relevant patients are less mobile than those with higher levels of 'mobility' gene expression. Consistent with these results were observations by Paul Meltzer (also of the NHGRI, who studied melanoma of the eye (known as uveal melanoma), a cancer that depends on its own vasculature for progression and metastasis. Comparison of profiles of cell lines derived from highly invasive and less aggressive tumours revealed higher expression levels of a set of genes (many of which coincide with those identified by Trent) encoding extracellular matrix proteins, intermediate filaments and proteolytic enzymes, as well as vasculogenic growth factors and their receptors. The more aggressive tumours were seen to form acellular, vessel-like structures—presumably composed of extracellular matrix—which may confer a survival advantage to the tumour (and disadvantage to the patient).

Algorithms for pattern recognition. As emphasized by a number of participants, array experiments are not immune to the need for an estimation of reproducibility and error rate. Many expression clusters in yeast, for example, are not characterized by 'dramatic' changes in expression. Rather, the 'tightness' of the cluster reflects a large data set. The detection of small differences in expression will depend on extensive repetition, allowing one to distinguish between random noise and non-random signal. To make sense of array data, a number of algorithms are in use and under development[7]. Classical statistical methods can be applied, but the size of data sets makes more computational methods, such as hierarchical clustering, more alluring. Other approaches involve self-organizing maps and neural networks. None of these can transform data acquired from ill-designed or ill-executed experiments, and none can reveal all of the information inherent in a given data set. Bioinformatics experts and computer scientists are testing modifications of current methods and exploring new ones. At present, however, there is no obvious algorithm of choice, and different methods may be required for the detection of different underlying patterns.

On the spot. With a reasonably powerful computer, one can cluster gene expression profiles within a few hours. The challenge is to interpret the results—a task made easier if the genes within the cluster are characterized. Appropriate annotation, as discussed by Terry Gaasterland (Rockefeller University), is of issue, as is the way in which information is made accessible. David Lipman (National Center for Biotechnology Information) reported on the NCBI's attempts to link public databases and announced the intention to launch a database for gene expression data next spring. The gene expression omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/) is the second initiative along these lines (the first was announced earlier this year by the European Bioinformatics Institute[8]). Clearly, coordination between these two groups—and careful attention to design—is in the best interests of the community. Even with improved databases, a gene-by-gene exploration of annotations and literature is tedious. Mark Boguski (NCBI) discussed two strategies to automate processing of expression data to prepare summaries of significant or recurrent themes. One involves the generation of cluster-specific annotation summaries, which are then combined into an 'executive summary'. A second approach is to retrieve documents relating to genes that demonstrate a change in expression, and then cluster the publications. Currently, such document clustering depends on informative features in titles and abstracts, and yields variable results.

The eventual success of this or any other attempt to transform information from large data sets into knowledge depends on a well-designed scientific information space in which smart computer programs can harvest relevant information. The potential benefit for our understanding of biological processes—the highlights of the meeting are just the tip of the iceberg—is challenging producers, publishers and users of large biological data sets to collaborate in establishing such a space. More data are better—as long as we can devise ways to share and make sense of them.

1. The Chipping Forecast. *Nature Genet.* **21**, 1–60 (1999).
2. Editorial. *Nature Genet.* **18**, 195–196 (1998).
3. Holstege, F.C.P. *et al. Cell* **95**, 717–728 (1998).
4. Zhu, H. *et al. Proc. Natl Acad. Sci. USA* **95**, 14470–14475 (1998).
5. Golub, T.R. *et al. Science* **286**, 531–537 (1999).
6. Perou, C. *et al. Proc. Natl Acad. Sci. USA* **96**, 9212–9217 (1999).
7. Bittner, M. *et al. Nature Genet.* **22**, 213–215 (1999).
8. Editorial. *Nature Genet.* **22**, 211–212 (1999).