

# Functional genomics

DNA sequence data provided by genome projects have spawned the new field of functional genomics. This approach will yield exciting insights into the pathways to which specific genes belong and will provide clues to their roles in health and disease.

It is estimated that there are between 70 000 and 100 000 genes in the mammalian genome<sup>1</sup>. To turn this genetic 'blueprint' into a functioning organism, each of these genes must be expressed in specific temporal and spatial contexts. In the past, the responses of cells or organisms have been studied on a small scale or in a limited context; for example, one gene or pathway at a time. The goal of the new **functional genomics** field<sup>2</sup> is to bring new technologies to bear on studying gene expression on a large scale and/or in a high-throughput manner. This approach will not yield the detailed understanding of biological processes that biochemistry, cell biology, physiology and pharmacology have traditionally provided, but the hope is that it will give biologists a new, comprehensive and holistic understanding of complex systems and narrow the gap between sequence and function.

It is only possible to think about genome-wide expression studies because of the large volume of genomic and cDNA (**EST**) sequence data (from many organisms) that has accumulated over the past few years. For example, the complete sequences of nearly two dozen prokaryotic genomes and the unicellular eukaryote yeast are available (see Entrez Genomes Division at <http://www.ncbi.nlm.nih.gov/Entrez/Genome/org.html>). The first metazoan genome to be sequenced, that of *Caenorhabditis elegans*, should be finished by the end of 1998. The sequence of ~5% of the human genome has now been completed (G. Schuler, pers. commun.) – the remainder should be available by 2005 – and the complete sequence of the mouse genome should be finished by 2008 (Ref. 3). These sequences typically lack any functional information apart from that inferred by sequence homology studies. Clearly, biologists, both now and well into the 21st century, will require new technologies to study and understand the functions implicit in all of these sequence data.

Much effort is currently being directed towards developing such technologies, starting with methods to study mRNA profiles by large-scale, high-throughput serial and parallel methods (see the URLs box). Researchers have begun by developing technologies involving the measurement of mRNA levels, because

## Michael J. Brownstein

Section on Genetics, National Institute of Mental Health/National Human Genome Research Institute, NIH, Bethesda, MD 20892, USA.

[mike@codon.nih.gov](mailto:mike@codon.nih.gov)

## Jeffrey M. Trent

Cancer Genetics Branch, National Human Genome Research Institute, NIH, Bethesda, MD 20892, USA.

[jtrent@nhgri.nih.gov](mailto:jtrent@nhgri.nih.gov)

## Mark S. Boguski

National Center for Biotechnology Information, National Library of Medicine, NIH, Bethesda, MD 20894, USA.

[boguski@ncbi.nlm.nih.gov](mailto:boguski@ncbi.nlm.nih.gov)

nucleic acids can be simply and powerfully manipulated. It is implicitly assumed, however, that **proteomics**<sup>4</sup> will ultimately provide more detailed and precise information on gene expression effects. The development of these technologies and resources (e.g. arrays of antibodies specific for all human proteins) is just beginning (see below). Nevertheless, there is much to be learned by performing **mRNA expression profiling**, and there are some very effective technologies that are already available for this purpose.

## Gene expression arrays

These experimental approaches require the availability of large collections of cDNAs (EST clone inserts) immobilized on glass<sup>5</sup> (termed microarrays) or sets of synthetic oligonucleotides immobilized on silica wafers or chips<sup>6,7</sup> (termed probe arrays). Both types of array are the conceptual descendants of target nucleic acids immobilized on filters or membranes and detected with complementary radioactive probes<sup>8,9</sup>. Such filter-based systems are now available from several commercial sources (see the URLs box) and might still provide a cost-effective alternative to newer technologies.

Regardless of the array chosen, experiments are performed in much the same way (Fig. 1). RNA has to be extracted from the cells or tissue to be studied, and 'tagged' cDNA or cRNA has to be made from mRNA in the extract. At that point, the product (tagged with radioactive or fluorescent nucleotide analogs) is hybridized to the array, which is then washed to remove unhybridized or weakly hybridized material. Subsequently, the array is optically scanned or 'read', and the resulting data are analyzed. Numerous challenges remain for those who want to use and improve existing methods. These fall into three areas: biochemistry, instrumentation and informatics. We shall consider each of these in turn.

## Biochemistry

To facilitate interpretation of expression data, the source of the mRNA to be studied should be as homogeneous as possible. Cultured cells, altered by the insertion or deletion of a gene or by pharmacological or physiological manipulations, are ideal for expression studies because

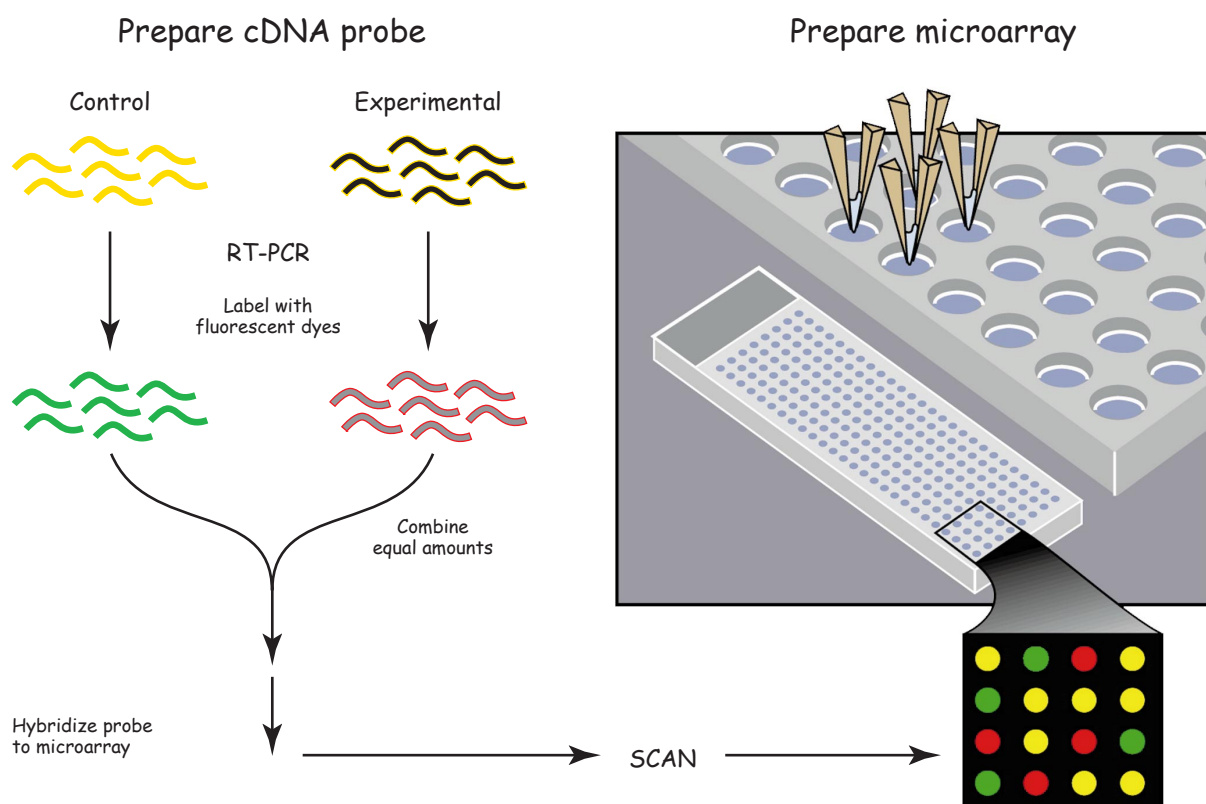


Fig. 1. Schematic overview of cDNA array fabrication and experimental procedures.

ample amounts of RNA can be prepared from them. Ideally, however, we would like to be able to study human or animal tissues in addition to cell lines, and most tissues are composed of numerous different cell types. Microdissection methods<sup>10</sup> permit the isolation of specific cells for analysis, but very little RNA can be prepared from the numbers (<1000) of such cells obtained. Consequently, methods are being devised that permit reverse transcription, linear amplification and labeling of tiny amounts of RNA; hopefully, these methods will also maintain the relative concentrations of the starting materials.

One of the most important, but rarely discussed, problems with the preparation of mRNA from various cells and tissues is the variability of specimen handling prior to RNA extraction and the effects that this might have on altering expression patterns. Even cultured cells present some difficulties in this context, but it is not hard to conceive how vexing the problem can be in human tissue samples obtained as a result of biopsy, surgical treatment or postmortem examination. Varying periods of anoxia and storage at room temperature, which are an almost unavoidable part of routine processing for pathology, will undoubtedly introduce complications into the interpretation of gene expression patterns. Hopefully, a set of reproducible artefacts will be identified that can be electronically subtracted in the course of data analysis to reveal the original and specific changes that result from the biological state or phenomenon under study.

The second challenge for the biochemist – to develop methods for proteomics<sup>4</sup> – is outside the scope

of this brief review. However, in the future, it would be interesting to look at all the proteins and peptides in a tissue extract and compare their identities, activities and post-translational modifications. In addition, one could envisage using protein arrays to probe for specific interactions with other proteins, nucleic acids, carbohydrates, lipids or even small molecule ligands.

#### Instrumentation

Currently, the most advanced expression systems are very expensive and, as a result, have only been available to large pharmaceutical companies or the laboratories in which the technologies are being developed. Even filter-based systems, the lower-cost alternative to microarrays or probe arrays, are expensive to produce or buy. To make all but the smallest filter array, one would need to construct a robot (see <http://cmgm.stanford.edu/pbrown/array.html>), order or amplify cDNAs, sequence them to be sure of their identity, and array them. This is impractical for investigators at most research centers. Even filter-based systems are relatively expensive and may contain a relatively small number of genes, and can be stripped and re-probed only a few times before they have to be discarded. The advantage of such an array is that it can be scanned with a phosphorimager (or even with X-ray film for qualitative results), obviating the need to buy an expensive fluorescence detector. Until the cost per experiment falls, the number of experiments that can be performed and the number of replicates per experiment will be limited. With few replicates, complete confidence in the results will only

be obtained by large (e.g. tenfold) changes in mRNA levels. Once this 'low-lying fruit' (i.e. the most obvious new phenomena) is plucked, biologists will need to pursue much more subtle and numerous 'signals' to make new discoveries. It is unclear whether current experimental and statistical/informatics tools are yet up to this task.

### Informatics

Large-scale, high-throughput experimental methods require information processing and analysis systems to match. Software and database systems to design arrays, track materials, collect and analyze, and interpret data from gene expression studies are still in their infancy<sup>11</sup>. Among other things, such systems have to catalog the expression behavior of thousands of genes in a single experiment and, subsequently, make comparisons across tissues, developmental and pathological states, or cellular perturbations.

Very large quantities of data have to be managed both before and after the experiment, because direct access is required to all sequences, annotations and physical DNA resources for the genes of the organism studied. Following hybridization and readout of relative expression levels observed in the sites on an array, the data must be stored and preserved so that it is available for image processing and statistical-biological analysis. The latter includes identifying the transcripts that show statistically significant<sup>12</sup> changes in absolute or relative quantity. Once this is done, several tasks need to be performed, the most obvious and straightforward of which is to provide information about the structures and functions of the gene products of interest. Interpreting this information is the responsibility of the investigator, who should potentially be able to interrogate the data sets in other ways. Biochemical pathways to which a particular transcript belongs could be identified or genes with which the transcript is thought to interact could be found. In a time-course experiment, sets of genes with similar temporal expression profiles could be sought. In the long run, the software could, indeed should, be made capable of pre-interpreting the data (using a biochemical knowledge base and a set of **heuristics**) and presenting the investigator with alternative hypotheses or explanations about its meaning. It is only in this way that experiments involving tens of thousands of genes, of which a considerable fraction shows changes, can be managed.

It is exciting to anticipate a time when data from thousands of gene expression experiments will be available for **meta-analysis**<sup>13</sup>, which has the potential to balance out artifacts from many individual studies, thus leading to more subtle findings and robust results. This will require that data adhere to some type of uniform structure and format that would ideally be independent of the particular expression technology used to generate it. The advantages and disadvantages of various publication modalities for these large electronic data sets have been discussed previously<sup>11</sup>.

### References

- 1 Fields, C. *et al.* (1994) *Nat. Genet.* 7, 345–346
- 2 Hieter, P. and Boguski, M. (1997) *Science* 278, 601–602
- 3 Collins, F. *et al.* *Science* (in press)
- 4 Kahn, P. (1995) *Science* 270, 369–370
- 5 Schena, M. *et al.* (1995) *Science* 270, 467–470
- 6 Fodor, S.P. *et al.* (1991) *Science* 251, 767–773
- 7 Fodor, S.P. *et al.* (1993) *Nature* 364, 555–556
- 8 Gress, T.M. *et al.* (1992) *Mamm. Genome* 3, 609–619
- 9 Lennon, G.G. and Lehrach, H. (1991) *Trends Genet.* 7, 314–317
- 10 Simone, N.L. *et al.* (1998) *Trends Genet.* 14, 272–276
- 11 Ermolaeva, O. *et al.* (1998) *Nat. Genet.* 20, 19–23
- 12 Chen, Y., Dougherty, E.R. and Bittner, M.L. (1997) *Biomed. Optics* 2, 364–374
- 13 Hedges, L.V. and Olkin, I. (1985) *Statistical Methods for Meta-analysis*, Academic Press

#### Information about various transcript profiling technologies

[http://www.ncbi.nlm.nih.gov/ncicgap/expression\\_tech\\_info.html](http://www.ncbi.nlm.nih.gov/ncicgap/expression_tech_info.html)

#### Plans and protocols for cDNA array technology

<http://cmgm.stanford.edu/pbrown/array.html>

#### A system for data management and analysis of gene expression arrays

<http://www.nhgri.nih.gov/DIR/LCG/15K/HTML/>

#### Examples of commercially available filter arrays<sup>a</sup>

##### GeneFilters™ (Research Genetics)

<http://www.resgen.com>

##### Gene Discovery Arrays (Genome Systems)

<http://www.genomesystems.com>

##### Atlas™ Arrays (CLONTECH)

<http://www.clontech.com>

<sup>a</sup> Please note that this list is not meant to be comprehensive nor does it imply endorsement of these products by the authors or the US Government. A more comprehensive list of technologies and suppliers can be found at [http://www.ncbi.nlm.nih.gov/ncicgap/expression\\_tech\\_info.html](http://www.ncbi.nlm.nih.gov/ncicgap/expression_tech_info.html).



URLS...

### Trends in Genetics

#### NEW – Genetic Nomenclature Guide

This special supplement summarizes the most important rules and guidelines for the genetic nomenclature of 19 model organisms used by geneticists and developmental biologists.

The *TIG Genetic Nomenclature Guide* is an essential 'quick-reference' resource, and includes many updates to the first edition published in 1995.

Copies of the *Guide* are available from November 1998

**Contact:** Thelma Reid ([t.reid@elsevier.co.uk](mailto:t.reid@elsevier.co.uk))

Elsevier Trends Journals, 68 Hills Road,  
Cambridge, UK CB2 1LA.

Tel: +44 1223 311114 • Fax: +44 1223 321410

# TRENDS GUIDE TO BIOINFORMATICS

Biological data, and DNA sequence data in particular, are accumulating at a phenomenal rate. By around 2005, it is likely that the DNA sequence of the complete human genome will have been determined. Although this achievement might seem an end in itself, in reality it is only the beginning. In order to exploit the wealth of DNA sequence and other biological data, a new science has arisen that fuses biology with mathematics and computer science – 'bioinformatics'.

To find the genes within the genomic sequence is a massive task in itself. Once apparent, otherwise uncharacterized coding regions must be assigned a function. Thereafter, the interactions between genes and gene products must be understood at all levels, not merely in the context of the pathways within and between cells but also in terms of the evolution of gene families within and between species. These questions can all be addressed using bioinformatics.

Bioinformatics touches all of biology, and straightforward access to data via the Internet means that a wealth of information is available, literally at our fingertips. However, the newcomer to bioinformatics might be discouraged because of the initially daunting computational and mathematical content. Others might simply be confused by the language and terminology of bioinformatics. To help overcome these barriers, the *Trends Guide to Bioinformatics* examines the background to this novel and rapidly evolving scientific discipline. A series of tutorials, written by expert authors, clearly explains the concepts and provides practical examples of how the immense store of data can be exploited. Technical terms are highlighted in each article and defined in the glossary on pp. 32–33. Topics covered range from retrieving and aligning sequences, to predicting structure and function of gene products. Whether you are interested in molecular structure or taxonomy of organisms, the *Trends Guide to Bioinformatics* is an essential tool.

- 1 Bioinformatics – a new era  
*Mark Boguski*
- 3 Text-based database searching  
*Fran Lewitter*
- 7 Fundamentals of database searching  
*Stephen Altschul*
- 9 Practical database searching  
*Steven Brenner*
- 12 Computational genefinding  
*David Haussler*
- 15 Multiple-alignment & -sequence searches  
*Sean Eddy*
- 18 Protein classification & functional assignment  
*Kay Hofmann*
- 22 Phylogenetic analysis & comparative genomics  
*James Lake and Jonathan Moore*
- 24 Databases of biological information  
*Minoru Kanehisa*
- 27 Functional genomics  
*Michael Brownstein, Jeffrey Trent and Mark Boguski*
- 30 The future of bioinformatics  
*Janet Thornton*
- 32 Glossary

## COPYRIGHT INFORMATION



©1998 Elsevier Science. All rights reserved. This supplement and the individual contributions contained in it are protected under copyright by Elsevier Science. See the box in the accompanying *Trends* journal for further terms and conditions that apply to the copyright. Except as outlined in the terms and conditions, no part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without prior written permission of the publisher.

## Guest Editors

Steven Brenner, Stanford University, CA, USA and Fran Lewitter, Whitehead Institute for Biomedical Research, Cambridge, MA, USA

## Editors

Mark Patterson and Michaela Handel

## Editorial Staff

Caroline Ash, Meran Owen, Rob Brines and John Pettigrew

## Editorial Administrator

Helen Steele

## Production Designer

Naomi Wright

## Publisher

Peter Desmond