

# Gene expression microarrays and the integration of biological knowledge

Michiel O. Noordewier and Patrick V. Warren

Large-scale parallel measurement of whole-genome RNA expression is now possible with high-density arrays of cDNA or oligonucleotides. Using this technology efficiently will require the integration of other sources of biological information, such as gene identity, biomedical literature and biochemical pathway for a given gene. Such integration is essential to understand the cellular program of gene expression and the molecular physiology of an organism. Advances in microarray technology, and the expected rapid rise in microarray data will lead to new insight into fundamental biological problems such as the prediction of gene function from expression profiles and the identification of potential drug targets from biologically active compounds.

The genomes of many microbial species, including model organisms and pathogens, have now been sequenced completely, and those of several higher organisms have also been determined (or will soon be available), including yeast<sup>1</sup>, *Drosophila*<sup>2</sup>, mouse and human<sup>3,4</sup>. Although the ability to determine the sequence of all possible genes within such organisms is satisfying, the daunting task of predicting the biochemical function and cellular role of each gene product remains. An even more ambitious goal is to understand the myriad of interactions within the genome and its products as a whole. Only the complexity of the first model genome need be considered to realize the scope of the problem. The DNA sequence of bacteriophage  $\lambda$  is two orders of magnitude smaller than that of *Escherichia coli* and five orders of magnitude smaller than that of human, yet more than 30 years of classical genetic and molecular techniques have resulted in an entire branch of scientific literature to describe the physiology of this not-quite-organism<sup>5</sup>.

Characterizing the expression of a gene is a logical step towards understanding its biological role. Measuring the first step in gene expression (the formation of mRNA) is particularly appealing because it can be accomplished in a general way by hybridization with DNA of complementary sequence (cDNA). Such methods have been reviewed recently<sup>6</sup> but, in brief, measurements can be made in large-scale parallel fashion by printing arrays of cDNAs or oligonucleotides at high density onto glass slides. The availability of complete genomic sequences and of parallel gene expression analysis enables novel

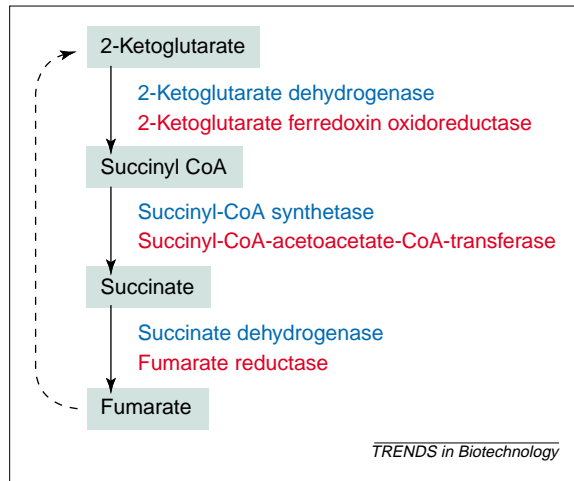
biological inquiry. Of particular interest are the characterization of the temporal order of gene expression within a cell, the determination of the cellular location of gene products, prediction of the function of resulting proteins and the effect of perturbations of the cellular environment on the program of gene expression by changing environmental conditions or administering drugs. It should be clear that such questions cannot be answered solely by inspection of spots on a glass slide but must also incorporate other sources of related biological knowledge. This includes, but is not limited to, databases of metabolic and regulatory pathways, transcriptional regulation, regulatory pathways, enzymes and/or reactions, transport mechanisms, and techniques such as metabolic reconstruction and pathway simulation algorithms.

The collection and collation of primary data are crucial to any microarray expression experiment. The challenges associated with experimental microarray methods should not be underestimated and include such hurdles as the lack of an objective scale of gene expression levels. These issues are beyond the scope of this article and readers are referred to recent reviews on methodology, data management and lab automation<sup>7</sup>, as well as mathematical tools for data analysis<sup>8</sup>. In this article, we focus on the integration of experimental expression data with other sources of knowledge, with the aim of understanding the roles of the studied genes.

**Temporal and spatial programs of gene expression**  
The immediate result of a microarray expression experiment is a catalog of transcriptional activity for the set of genes represented on the grid<sup>9,10</sup>. Given that this set can potentially include the coding sequences of an entire genome, there must be facile links to external information sources that describe known structural and functional assignments for every possible gene, as well as cellular processes and pathways. The US National Center for Biotechnology Institute's Entrez databases and software<sup>11</sup> are such a source. These allow the retrieval of the scientific literature associated with each gene according to a variety of

Michiel O. Noordewier\*  
Patrick V. Warren  
Dept Bioinformatics,  
GlaxoSmithKline  
Pharmaceuticals, 1250  
South Collegeville Rd,  
UP 1345, PO Box 5089,  
Collegeville, PA 19426,  
USA.  
\*e-mail:  
Mick\_O\_Noordewier@  
sbphrd.com

Fig. 1. Non-homologous genes can encode proteins serving the same function. The citric acid cycles of *Helicobacter pylori* (in blue type) and *Saccharomyces cerevisiae* (in red type) demonstrate two different sets of catalytic enzymes to accomplish a common function.



criteria, including sequence, MeSH headings, keywords and content.

However, a challenge arises because databases of molecular sequence do not have a common lexicon for describing the roles that genes have in organisms. For example, although early sequence comparisons between eukaryotic genomes<sup>12</sup> have revealed a remarkably high degree of sequence and functional conservation, the underlying literature describing the molecular physiology of the organisms has evolved with different terms for congruent functions. The Gene Ontology Consortium<sup>13</sup> is addressing this problem by trying to produce a common and controlled vocabulary to describe the cellular roles of genes in any organism. More daunting still is the task of identifying relevant general biological knowledge from broad biomedical databases, such as subcellular location, associated diseases and drug interactions for specific genes. Although this field is still in an early state, recent work in automated text extraction has attempted to relate gene-related scientific text to the output of microarray experiments<sup>14-16</sup>.

After the establishment of a controlled vocabulary with well-defined semantics, significant work remains in the interpretation of the large volume of data produced by parallel expression experiments. Current databases of biological information associated with molecular sequence data have been structured to return information about individual genes, but more insight will be gained by asking questions about subsets of genes and gene products, based on their expression patterns. For example, genes of known function might be found to be related by temporal expression and suspected of belonging to a common metabolic or regulatory pathway. Recent efforts at integrating such information include the 'What Is There' (WIT) system<sup>17</sup>, which provides metabolic reconstructions. This approach to reconstruction is an attempt to identify the pathways and gene assignments of an organism on the basis of examination of the genomic sequence, using databases of metabolic

pathways such as the Metabolic Pathways Database (MPW)<sup>18</sup>. Another knowledge source for metabolic pathways and genomics is the Kyoto Encyclopedia of Genes and Genomes (KEGG)<sup>19</sup>, which tabulates functional information for gene catalogs of organisms. These functions include higher order cellular processes such as membrane transport, cell cycle and signal transduction. In addition, KEGG encodes information about positionally coupled genes, as well as components of cellular metabolism such as enzyme substrates and chemical cofactors.

Data sources that are expected to be relevant to extending these analyses include transcriptional regulation databases, such as the object-oriented Transcription Factors Database (ooTFD)<sup>20</sup>, RegulonDB (the database of transcriptional regulation and operon organization in *E. coli* K-12)<sup>21</sup>, the Eukaryotic Promoter Database (EPD)<sup>22</sup> and regulatory-pathway databases such as GeneNet<sup>23</sup> and the Database for Cell Signaling Networks (CSNDB)<sup>24</sup>. In addition to data structures engineered into these databases, methods have been continually refined to identify common motifs in DNA sequences adjacent to co-expressed genes. Current pattern recognition techniques for this purpose have been reviewed elsewhere<sup>25,26</sup>.

#### Predicting gene function using microarrays

The databases described so far begin with genes of known function and try to determine biological activity using methods such as the presence of pathways or the co-regulation of expression. However, when novel sequences are encountered from the genomic data, a fundamental task is the assignment of function to those sequences. Although function is often inferred by comparing sequences, non-homologous genes can code for proteins that serve the same function, a phenomenon referred to as 'gene displacement' (Fig. 1). Conversely, in 'gene recruitment', genes with identical sequences can code for wholly different functions<sup>27-29</sup>. Gene recruitment can arise by, for example, post-translational modification to produce polymorphism of protein isoforms. As a result, sequence comparison alone is insufficient to infer function.

Using biological intuition alone, it might be assumed that coordinated repression, induction or overexpression of genes in response to external stimuli (e.g. ultraviolet light or stress response) indicates the cellular role of a gene. In an expression experiment, the similarities between the temporal expression pattern of genes are assumed to imply related function; a notion recently termed 'guilt by association'<sup>30</sup>. Indeed, yeast expression experiments suggest many speculative functional assignments for many of the previously uncharacterized genes in that genome<sup>9,10</sup>. This evidence, however, although enticing, is not

Fig. 2. Genes can be clustered to predict function on the basis of co-occurrence in evolution. Profiles of the occurrences of genes in genomes can be clustered to predict function on the basis of co-occurrence in evolution. Similarly, genes can be characterized as vectors of normalized expression levels or as a matrix of tissues and treatments.

		Organisms						
		<u>a</u>	<u>b</u>	<u>c</u>	<u>d</u>	<u>e</u>	<u>f</u>	...
Genes	x:	0	1	1	0	1	0	
	y:	1	1	0	0	1	1	
	z:	1	1	0	0	1	1	

TRENDS in Biotechnology

sufficient unambiguously to assign biochemical function or cellular role by observation of correlated gene expression alone. Such inference is hindered by the presence of constitutively expressed genes and the loose coupling between the levels of mRNA and protein. It is important to look for other methods to support functional assignments, including the following.

#### *Phylogenetic profiles*

Pellegrini and colleagues have created profiles of occurrences of genes in genomes that can be clustered to predict function on the basis of co-occurrence in evolution<sup>31</sup>. Proteins within a common metabolic pathway or structural complex are defined as being functionally linked. Functionally linked proteins are presumed to evolve in a coordinated fashion and should therefore display homologs in the same sets of organisms (Fig. 2). For example, flagella proteins should be observed only in bacteria that possess flagellae. The availability of complete genomic sequences for a substantial set of bacteria makes possible the inference of functionally linked proteins.

#### *'Rosetta Stone' proteins*

Often, protein domains exist separately in one organism but are seen fused in some other organism<sup>32,33</sup>. If two protein domains, A and B, are involved in a related function, fusion of the domains into a single peptide chain would increase the apparent affinity of A for B, by reducing the association free energy of A to B. This model for the evolution of protein-protein interactions is termed the 'Rosetta Stone' hypothesis.

#### *Operon clusters*

In many cases, the genes encoding proteins in the same pathway are also clustered along the genome. The PFMP database<sup>34</sup> uses chromosomal location, among other types of information, to construct models of cellular function and dynamics. The database incorporates evidence of structure, cellular location and functional neighbors in determining genetic function.

#### *Integration*

To demonstrate the power of integrating these inference methods, Marcotte and colleagues

combined expression patterns, phylogenetic profiles and domain fusion patterns for 20 fully sequenced genomes<sup>35</sup>. In doing so, they built functional relationships among the 6217 yeast proteins and suggested a general function for more than half of the 2557 previously uncharacterized proteins in yeast. In the same way that systematic approaches are required to identify the biological function of novel genes, methods are required to characterize the effect of biologically active compounds on cell metabolism. Hughes and colleagues<sup>36</sup> have constructed a reference database of expression profiles corresponding to 300 mutations and chemical treatments in *Saccharomyces cerevisiae*, and have shown that cellular pathways affected by such perturbations can be identified by pattern matching. The internally consistent expression data set so constructed allowed the identification of co-regulated sets of genes and led to more detailed analysis of transcriptional control for a variety of cellular processes. Examples of pathways identifiable by this method included ergosterol and cell wall synthesis, as well as mitochondrial respiration and translation.

#### *Bioinformatics, cheminformatics and pharmacology*

The integration of biological information and methods need not be limited to assignment of gene function. Scherf and colleagues<sup>37</sup> combined cDNA microarray gene expression experiments with drug sensitivity relationships in tumor cell lines to investigate the molecular pharmacology of cancer. cDNA arrays containing ~8000 genes were used to measure mRNA levels in 60 separate human cancer cell lines. These same cell lines were used to screen 70 000 chemical compounds for growth inhibition, as a measure of potential anticancer activity (Fig. 3). The growth-inhibiting activity of compounds against the panel of cell lines was correlated with the expression levels of proteins (targets) in the same cell lines, as measured by mRNA expression. Compound-target pairs with high positive or negative correlations are candidates for further investigation of gene-drug relationships. Although Scherf and colleagues are investigating molecular pharmacology, the experiments provide an unexpectedly powerful link with both cheminformatic methods and protein sequence analysis. The initial results match individual chemical compounds with single genes but it is further possible to cluster compounds according to chemical descriptors and sequences by means of phylogenetic relationships. In this way, it is at least theoretically possible to associate classes of compounds with families of proteins by means of their correlated respective activities and expression levels.

		Cell lines							
		A	B	C	D	E	F	G	H
Compounds	U	+	-	+	+	-	-	-	+
	V	-	+	-	-	+	-	+	-
	W	+	-	+	-	+	-	-	-
	X	-	-	-	-	+	+	-	-
	Y	-	-	-	+	+	-	-	+
	Z	-	+	+	+	-	+	-	+
		Cell lines							
		A	B	C	D	E	F	G	H
Genes	L	+	-	+	-	+	-	-	+
	M	-	+	-	-	+	+	+	-
	N	+	-	+	+	-	-	-	-
	O	-	+	+	-	+	+	-	-
	P	+	-	+	-	+	-	-	-
	Q	-	+	-	-	-	+	-	+

*TRENDS in Biotechnology*

Fig. 3. Microarrays of cDNAs can be used to correlate gene expression with growth inhibition of cancer cells. (a) Compounds were assayed for their ability to inhibit the growth of tumor cell lines (+ indicates growth, - indicates no growth). (b) The expression levels of various genes were measured in the same cell lines (+ indicates expression, - indicates no expression). Where a compound's pattern of growth inhibition correlates with the pattern of expression of a gene (outlined), an inference can be made that the gene is the cellular target of the compound. Compound-target pairs with high positive or negative correlations are candidates for further investigation of gene-drug relationships.

## Conclusions

Ultimately, there are no shortcuts to characterization of the molecular physiology of an organism. Any understanding of the function of a set of genes must incorporate all available knowledge of each step in the central dogma of gene expression (DNA to RNA to protein). Parallel expression microarray experiments provide an experimental approach to understanding the first of these steps, the pattern of mRNA formation. Combined with our recent ability to sequence entire genomes, we can for the first time contemplate investigating the entire program of gene expression within a cell.

Continued progress will rely on two areas of endeavor. The first of these is the attempt to gather large-scale experimental evidence about other parts of the gene expression process, such as individual rates of mRNA turnover, control of message translation, post-translational modification of peptides, protein stability and protein-protein interactions. The second is the effort to improve our ability to integrate information from high-throughput parallel expression experiments with other sources of knowledge about metabolism, gene regulation, transport and signal transduction. Despite these continuing challenges, even well-studied biological systems have benefited from the initial use of microarray expression tools.

## References

- Goffeau, A. *et al.* (1996) Life with 6000 genes. *Science* 274, 563-567
- Adams, M.D. *et al.* (2000) The genome sequence of *Drosophila melanogaster*. *Science* 287, 2185-2195
- International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860-921
- Venter, J.C. *et al.* (2001) The sequence of the human genome. *Science* 291, 1304-1351
- Hendrix, R.W. *et al.* (1983) *Lambda II*. Cold Spring Harbor Laboratory
- Harrington, C.A. *et al.* (2000) Monitoring gene expression using DNA microarrays. *Curr. Opin. Microbiol.* 3, 285-291
- Bassett, D.E. *et al.* (1999) Gene expression informatics - it's all in your mine. *Nat. Genet.* 21 (Suppl.), 51-55
- Bittner, M. *et al.* (1999) Data analysis and integration: of steps and arrows. *Nat. Genet.* 22, 213-215
- DeRisi, J.L. *et al.* (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278, 680-686
- Cho, R.J. *et al.* (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell* 2, 65-73
- McEntyre, J. (1998) Linking up with Entrez. *Trends Genet.* 14, 39-40
- Schervitz, S.A. *et al.* (1999) Using the *Saccharomyces* Genome Database (SGD) for analysis of protein similarities and structure. *Nucleic Acids Res.* 27, 74-78
- Ashburner, M. *et al.* (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.* 25, 25-29
- Shatkay, H. *et al.* (2000) Genes, themes and microarrays. *ISMB 2000*, 317-328
- Blaschke, C. *et al.* (1999) Automatic extraction of biological information from scientific text: protein-protein interactions. *ISMB 1999*, 60-66
- Craven, M. *et al.* (1999) Constructing biological knowledge bases by extracting information from text sources. *ISMB 1999*, 77-86
- Dsouza, M. *et al.* (1997) Searching for patterns in genomic data. *Trends Genet.* 13, 497-498
- Selkov, E. (2000) MPW: the metabolic pathways database. *Nucleic Acids Res.* 26, 43-45
- Kanehisa, M. and Goto, S. (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 29-34
- Ghosh, D. (2000) Object-oriented transcription factors database (ooTFD). *Nucleic Acids Res.* 28, 308-310
- Salgado, H. *et al.* (2001) RegulonDB: transcriptional regulation and operon organization in *E. coli* K-12. *Nucleic Acids Res.* 29, 72-74
- Perier, R.C. *et al.* (2000) The eukaryotic promoter database (EPD). *Nucleic Acids Res.* 28, 302-303
- Kolpakov, F.A. *et al.* (1998) GeneNet: a gene network database and its automated visualization. *Bioinformatics* 14, 529-537
- Takai-Igarashi, T. *et al.* (1998) A database for cell signaling networks. *J. Comput. Biol.* 5, 747-754
- Burset, M. and Guigo, R. (1996) Evaluation of gene structure prediction programs. *Genomics* 34, 353-367
- Reese, M.G. *et al.* (2000) Genome annotation assessment in *Drosophila melanogaster*. *Genome Res.* 10, 483-501
- Sun, Y.J. *et al.* (1999) The crystal structure of a multifunctional protein: phosphoglucose isomerase/autocrine motility factor/neuroleukin. *Proc. Natl. Acad. Sci. U. S. A.* 96, 5412-5417
- Kojima, H. *et al.* (1997) The dual functions of *Tetrahymena* citrate synthase are due to the polymorphism of its isoforms. *J. Biochem.* 122, 998-1003
- Segovia, L. *et al.* (1997) Two roles for mu-crystallin: a lens structural protein in diurnal marsupials and a possible enzyme in mammalian retinas. *Mol. Vis.* 9, 3-9
- Chu, S. *et al.* (1998) The transcriptional program of sporulation in budding yeast. *Science* 282, 699-705
- Pellegrini, M. *et al.* (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. U. S. A.* 96, 4285-4288
- Marcotte, E.M. *et al.* (1999) Detecting protein function and protein-protein interactions from genome sequences. *Science* 285, 751-753
- Enright, A.J. *et al.* (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature* 402, 86-90
- van Helden, J. *et al.* (2000) Representing and analysing molecular and cellular function using the computer. *Biol. Chem.* 381, 921-935
- Marcotte, E.M. *et al.* (1999) A combined algorithm for genome-wide prediction of protein function. *Nature* 402, 83-86
- Hughes, T.R. *et al.* (2000) Functional discovery via a compendium of expression profiles. *Cell* 102, 109-126
- Scherf, U. *et al.* (2000) A gene expression database for the molecular pharmacology of cancer. *Nat. Genet.* 24, 236-244