

Functional genomics: Learning to think about gene expression data

Roger Brent

Three recent studies of gene expression patterns in whole cells provide examples of the inferences one can make from this type of information. They also provide examples of the non-traditional types of reasoning we will need to use to make such inferences.

Address: The Molecular Sciences Institute, 2168 Shattuck Avenue, Berkeley, California 94704, USA.

Current Biology 1999, 9:R338–R341
<http://biomednet.com/elecref/09609822009R0338>

© Elsevier Science Ltd ISSN 0960-9822

Although molecular biology has deep roots and continuing success in the study of gene transcription, these advances have not been matched by an understanding of the transcripts that are actually expressed under different conditions in cells, tissues, and organisms. The development of methods to visualize gene expression by hybridization to arrays of DNAs carried on chips promised to help correct that ignorance, and, as shown by three recent studies [1–3], late last year these methods began delivering. Because they enable observation of the previously unobservable, these methods are sometimes compared to the telescope and microscope. Just as development of the telescope and microscope was followed by periods during which science was mostly done by observation rather than experiment, it is possible that the development of gene-expression monitoring and other functional genomic methods may presage a phase in which biology once again becomes more observational.

Two of the recent papers, by Chu *et al.* [1] and Holstege *et al.* [2], explored transcription of the entire complement of genes in the budding yeast, *Saccharomyces cerevisiae*. The third paper, by Iyer *et al.* [3], explored transcription of more than 8,000 genes in human fibroblasts. The three studies produced some insights into biology but also illustrate a current limitation, namely that making new technology work may be easier than using it to discover truth. The sheer abundance of data requires better methods for handling data and thinking about results than now exist. Moreover, the inferences one makes from these and other functional genomic methods will often come from kinds of reasoning that do not closely match the traditional picture of hypothesis-driven experimentation.

The ability to survey transcripts from whole genomes required progress in the methods for manufacturing miniaturised arrays, manipulating mRNAs, and visualizing data. Arrays of oligonucleotides corresponding to large numbers of genes on chips were first made by direct

photolithographic synthesis by Fodor and coworkers [4]. The generation of probes by faithful amplification, using the polymerase chain reaction (PCR), of a population of mRNAs to a population of cDNAs in the same relative abundance was first perfected by David Lockhart, Gene Brown, and their coworkers [5]. The potential of this approach sparked efforts in a number of groups, particularly those of Brown, Botstein, and Davis (see [6]), to make lower density, cheaper, non-patent-protected arrays using a robot-controlled pen to spot PCR-generated longer DNA fragments onto glass. The first clustering algorithms, to group together samples with similar expression patterns, were developed in a far-reaching study by Weinstein and his coworkers [7], who examined changes in about 100 markers, including gene expression, in 60 cancer cell lines in response to 60,000 drugs, and needed to make sense of their results.

One of the recent papers, by Chu *et al.* [1], examined the patterns of gene expression at different times after sporulation in yeast. The authors applied a number of methods to make inferences about the function of genes expressed during this process. The authors term one of the most powerful of these inference-making methods “guilt by association”: genes that are expressed together may function together. Here is an example of the type of inference derived by this method: since genes of the proteolytic anaphase-promoting complex, which normally functions in mitosis, are expressed midway through sporulation, it is possible that proteolysis by this complex might also be needed for the sporulating yeast to exit one or both of its meiotic divisions. The ability of the authors to make inferences based on coexpression was facilitated by their development of improved clustering algorithms that grouped profiles of similarly expressed genes together to facilitate their comparison by eye [8].

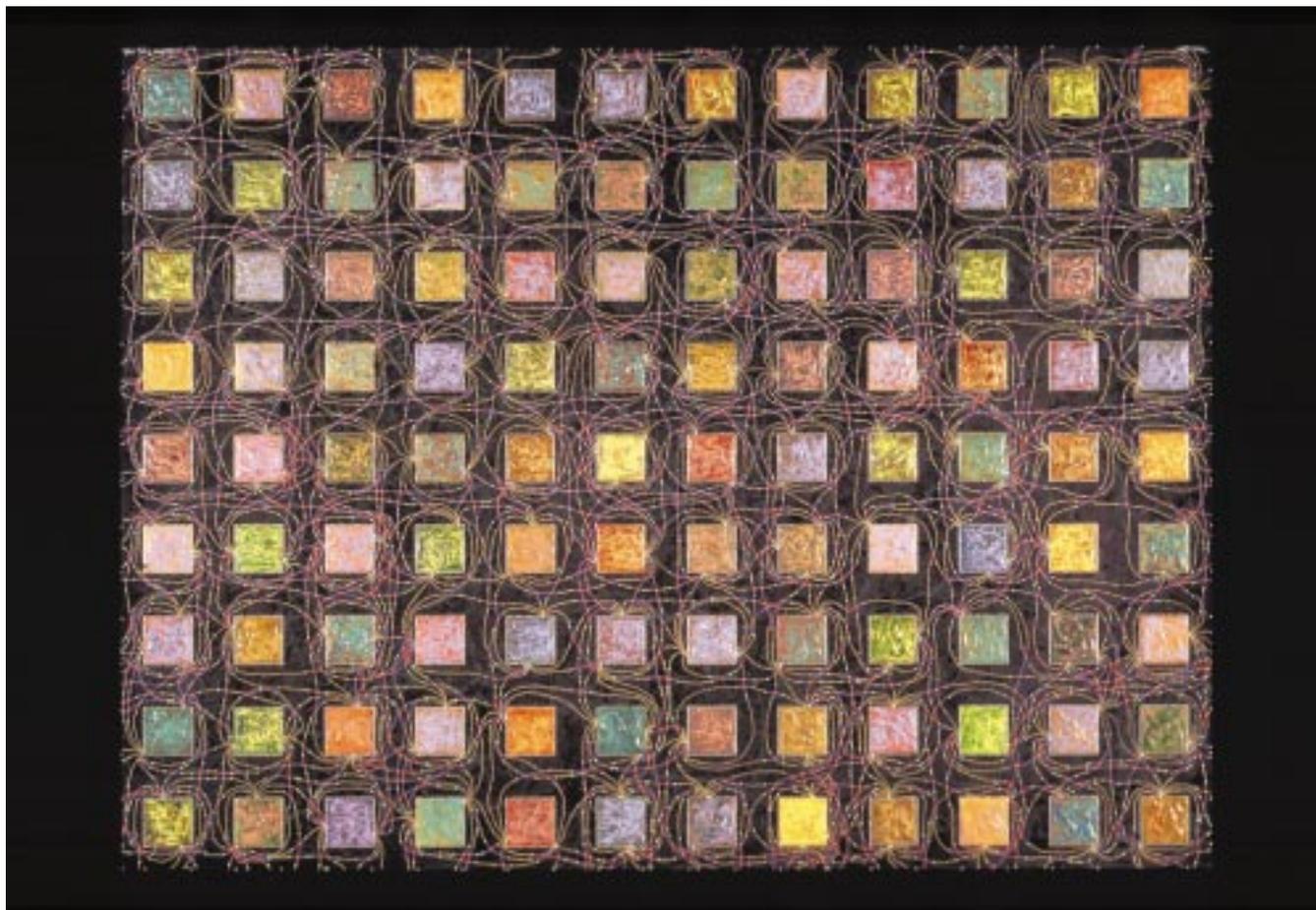
In another inference-generating tactic, the authors used coexpression as a criterion to give them comfort that regions of sequence similarity upstream of coregulated genes might identify conserved regulatory elements. Here, the authors’ combination of expression data with sequence data made weak sequence similarity more believable. Both of these reasoning tactics illustrate an important point: biologists can make progress by generating testable inferences, even if those inferences fall well short of ‘conclusions’ in the traditional sense. The systems being explored are adventitiously evolved, and the reasoning of their explorers does not have to be held to the same standard to which one holds university students in a logic class.

The second paper [2] describes gene expression monitoring with chips made by the biotechnology company Affymetrix; it examines the contribution of different RNA polymerase II (polII) holoenzyme genes and transcription cofactors to gene transcription. This work contains a number of experiments that deliberately perturbed the system in small, genetically defined, ways. Here, the chief technique was to compare mRNAs expressed in two cultures of yeast, one of which lacked a particular transcription-associated protein. In the past few years, the number of such proteins has greatly proliferated, and this proliferation has been reflected in a burgeoning and confusing literature concerning the effect of changes in the amount of cofactor *x* on transcription of genes *a*, *b*, and *c*. In this context, the allure of whole-genome gene-expression monitoring was obvious, if only to obviate the need to remember the conclusions of numerous one-off studies. And the results were striking. The work identified an unexpected diversity in the needs of different genes for different transcription-associated proteins. The authors' work divided the polII-transcribed genes of *S. cerevisiae*

into overlapping sets whose transcription did or did not depend on specific cofactors, histone deacetylases, and the like. Furthermore, the work revealed that one set of genes — those induced during pheromone response — depended for its transcription on a particular cofactor, Srb5. Finally, the work revealed that another cofactor, Srb10, was needed to repress the transcription of genes that are turned on in nutrient-limiting conditions. These last two results suggest that at least these two transcription-associated proteins are far-downstream members of signal transduction pathways, and thus suggest another mechanism — control of the activity of individual transcription components — by which extracellular and intracellular events can affect gene expression.

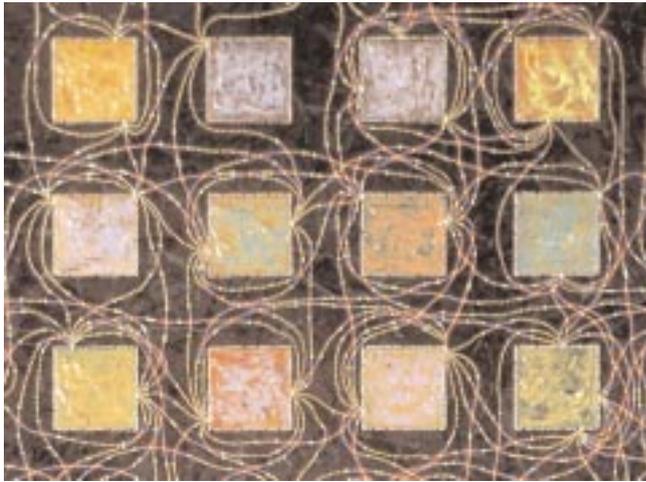
The third paper, by Iyer *et al.* [3], simply examined the expression of 8,613 known and unknown human genes in fibroblasts that had been starved of serum and then given fresh serum to induce them to proliferate. Here, the experiments showed not only the large numbers of genes known to be induced by the serum signal but also hitherto

Figure 1



A painting, *Ken's light*, by Andrea Way in 1993, bears a resemblance to the way gene expression data from microarrays are often represented.

Figure 2



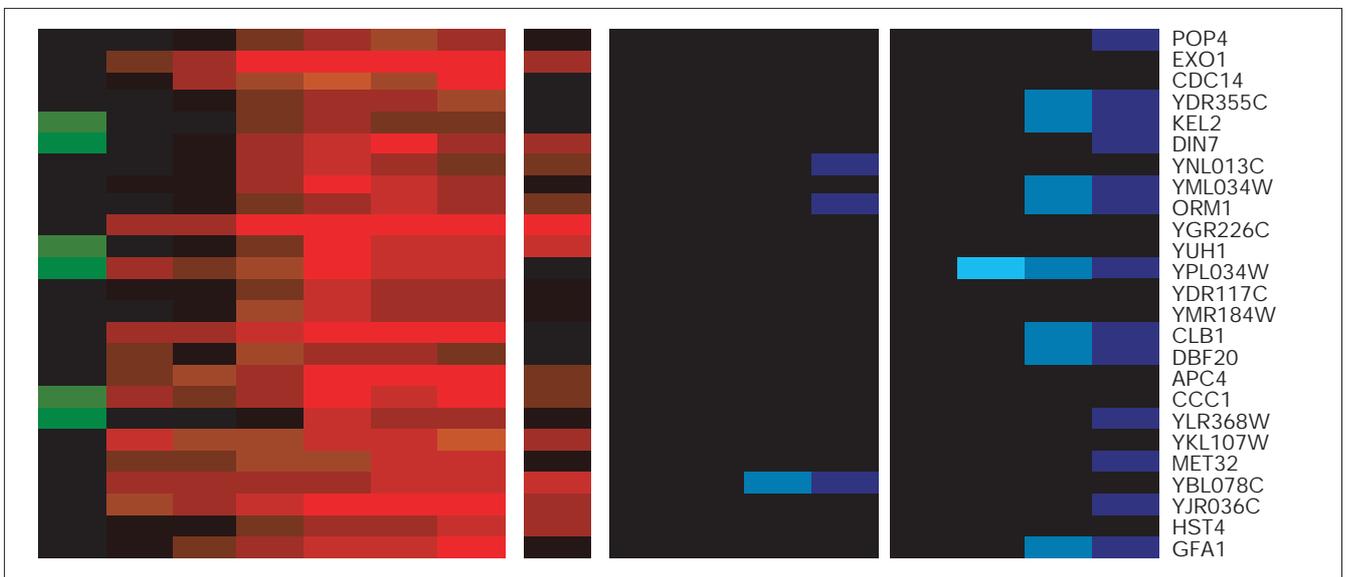
A detail of the Andrea Way painting shown in Figure 1 reveals apparent connections among network elements. It thus symbolizes a level of functional understanding we would like to be able to use genomic data to reach.

unexpected changes in the expression of other genes, such as protein phosphatases, that might be involved in modulating or adjusting the cellular response to serum. But perhaps the most striking conclusion of these experiments

comes from the guilt-by-association technique: the response to serum also induced large numbers of genes that function in wound healing. Wound healing is, after all, one of the main things fibroblasts do, and, in retrospect, it makes sense that expression of such genes might normally be observed at times when fibroblasts might be proliferating to invade a wound site. One can only wonder whether there are other biologically plausible processes handled by the 200 genes of unknown function that the authors also showed to be induced.

These early gene expression studies illustrate a number of issues that are likely to become important in the next few years. First, their findings come from an observational, or at least ‘not so hypothesis-driven’ approach. While this constitutes something of a departure from the mainstream biology that has developed during the past 30 years, it has sound roots. After all, biology started with individuals looking, seeing, describing, and making simple inferences from what they saw. Now this process must be speeded up. We can expect progress in using clustering methods to facilitate inference production, since these methods can be used to take expression data together with other kinds of quantitative information — for example, susceptibility to inhibition by different chemical inhibitors. Similarly (although it is a much more difficult problem), we can expect progress in learning to layer different types of weak

Figure 3



A figure based on the data from [1] shows cluster representation of genes (listed on the right) whose transcription is induced immediately after the onset of yeast sporulation. In such representations [7,8], the value of a measurement is represented by its color, measurements deemed similar by some criterion are presented together and units of time are represented by successive blocks along the horizontal. These

displays are now being used to facilitate, for example, inspection of the transcription patterns of genes whose transcription is induced under similar circumstances. Although these displays afford a step up from the raw data, they do not represent the information at a high enough level of abstraction to easily extract much meaning from it.

data on top of one another to create stronger inferences. Such inferences will not always be the end point: we will often need to test them, and one consequence of these recent studies has been to focus the attention of many people on the need to devise new, high-throughput, experimental ways of testing specific ideas.

Second, these studies, particularly those of Holstege *et al.* [2] and Iyer *et al.* [3], show that, to generate inferences, loose reasoning can work. The guilt-by-association technique for guessing gene function, although perhaps not acceptable in law courts or logic classes, has found wide use by other human institutions, from intelligence agencies to currency markets, and may also serve biology. Similarly, I believe we can consider rehabilitating another fallacy – *post hoc, ergo propter hoc*, or ‘after this, therefore because of this’. This fallacy should be useful in finding regulators: expression of a putative regulator followed by activation or repression of a number of genes may mean that the regulator activates them or represses them. What other ‘logical errors’ might prove useful?

Third, the recent studies show that current data-handling and visualization methods are inadequate. Existing clustering data-display methods are not in themselves sufficient to get all the information that must be present in the data. To address the data-handling issues, many biologists have turned for inspiration to standard texts in the field, only to find themselves betrayed by the cheerful promise lurking in phrases such as “the visual display of quantitative information” [9]; alas, this is a field for which there are no general-purpose solutions, only stylistic guidelines. It thus seems likely that improvements in ways to display and analyze gene expression and other functional genomic information will arise from *ad hoc* contributions of individual research groups. These contributions need not come from scientists: for biological data, where precise quantitation is almost never important at the beginning of a story, it may be that progress on data analysis will come from graphic artists or advertising agents as well as statisticians (see Figures 1–3). One ray of hope here is that (like sequence data) all expression data look alike, and insights into their display and analysis made once should stay valid. But such insights may not come quickly: the progress of publicly supported efforts to advance visualization and analysis of new functional genomic data types — never fast — has been further slowed by the exodus of many of the best computational and statistical biologists to industry.

Even though these are early days, however, the three recent studies [1–3] clearly reflect the first fruits of powerful, and possibly ultimately transformative, technologies. With luck, insights made from these methods may be as useful as those made from telescopes turned on the skies. They may also foreshadow the rise of a more exploratory school of biology, as opposed to the ‘physics-envy’ school,

in which single variables are changed and predictions proven or disproven. For the moment, read the papers for their glimpses of new planets and nebulae, but follow further developments in anticipation of the deeper insights that motivate us as biologists.

References

1. Chu S, DeRisi J, Eisen M, Muhlolland J, Botstein D, Brown PO, Herskowitz I: **The transcriptional program of sporulation in budding yeast.** *Science* 1998, **282**:699-705.
2. Holstege FCP, Jennings EG, Wyrick JJ, Lee TI, Hengartner CJ, Green MR, Golub TR, Lander ES, Young RA: **Dissecting the regulatory circuitry of a eukaryotic genome.** *Cell* 1998, **95**:717-728.
3. Iyer VR, Eisen MB, Ross DT, Schuler G, Moore T, Lee JCF, Trent JM, Staudt LM, Hudson J Jr, Boguski MS, *et al.*: **The transcriptional program in the response of human fibroblasts to serum.** *Science* 1999, **283**:83-87.
4. Fodor S, Read JL, Pirrung MC, Styer L, Lu AT, Solas D: **Light-directed, spatially addressable parallel chemical synthesis.** *Science* 1991, **251**:767-773.
5. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittman M, Wang C, Kobayashi M, Horton H, Brown EL: **Expression monitoring by hybridization to high-density oligonucleotide arrays.** *Nat Biotech* 1996, **14**:1675-1680.
6. DeRisi J, Iyer VR, Brown PO: **Exploring the metabolic and genetic control of gene expression on a genomic scale.** *Science* 1997, **278**:680-684.
7. Weinstein JN, Myers TG, O'Connor PM, Friend SH, Fornace AJ Jr, Kohn KW, Fojo T, Bates SE, Rubinstein LV, Anderson NL: **An information-intensive approach to the molecular pharmacology of cancer.** *Science* 1996, **275**:5298-5302.
8. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci USA* 1998, **95**:14863-14868.
9. Tufte ER: *The Visual Display of Quantitative Information.* New York: Graphics Press; 1992.