Databases for gene expression

George M. Church

Levels of mRNA in cells are often used as an approximation of protein activity levels. The latest report on the mRNA front by López-Nieto and Nigam[1] exploits statistical trends in protein-coding regions of DNA sequence databases to help find cell-type specific transcripts. Now that several genomes have been sequenced, spanning all major kingdoms of life[2], we realize that, relative to genome sequencing, sequencing mRNAs as cDNAs is not a "shortcut" to obtaining a complete list of genes. This is because, in genomic DNA, each gene is equally abundant, whereas, when surveying all cell types, some mRNAs are found at high levels in each cell while other nonabundant mRNAs may be down near the transcriptional noise level. For vast number of cell types, environment influences and potential transcriptional leakiness[3] could mean that each basepair of any genome may be detectably transcribed (at least one transcript per cell in at least some cell). Nevertheless, as each genome sequence is completed, gene discovery efforts will shift to gene function discovery, and so at least to mRNA quantification.

RNA quantification methods include cDNA array hybridizations, in situ hybridizations, counting of short expressed sequence tags (EST, SAGE), differential display (DD), and high-specificity PCR[1,4]. The last two methods yield somewhat predictable amplification product sizes for given primer pair sets. Conventional DD is anchored (primed) at the 3′ end of the RNA, with one of three oligo-dT (A,C,G) primers, and toward the 5′ end by a specific but arbitrary primer (or a specific but arbitrary restriction enzyme sequence[5]). High-specificity PCR requires knowledge of the sequence and possibly 100,000 primer pairs to cover all human coding regions.

In both cases, about 400 different product sizes can be resolved by DNA gel electrophoresis, so in theory, a minimum of 250 arbitrary primers (or pools of specific primer pairs) and gel lanes would be required per cell type measured. López-Nieto and Nigam explore the zone between these two methods with what could be called codon-optimized differential display (CODD). By computationally simulating the fraction of genes detected as a function of both primers chosen, they predict that the best set of 870 "arbitrary" primer pairs (the product of 30 forward and 29 reverse primers) should be sufficient to detect 73% of expressed genes, compared to the 10% expected of a truly arbitrary set of 870. The advantages of CODD over high-specificity PCR are that the recommended 60 primers are far fewer than previously predicted 200,000 and that one need not know all of the gene sequences in advance. Of course, as we approach the gene discovery endgame, we will want to know how to get the 27% of genes missed by this method down to less than 1 gene in 100,000 missed.

As we approach the gene discovery end-game, we will want to know how to get the 27% of genes missed by this method down to less than 1 gene in 100,000 missed.

As with many other methods, the shortest and lowest abundance coding regions will be the most easily missed. One source of background that will limit the hunt for the lowest RNA abundances detectable is mismatched priming. Mispriming effects are just beginning to be addressed[1,6,7] in tests using degenerate primers for large gene families such as G-protein coupled receptor genes. Another limiting factor is the dynamic range of the methods; a "pure" cell type could cover five orders of magnitude in mRNA abundance, and the number of cell types could number in the billions. Occasionally expression levels are experimentally lumped into compartments, such as "brain," containing numerous cell types. Such studies may be superseded by higher specificity definitions in terms of constellations of gene product levels, cell lineages, cellular and subcellular locations[8–10].

In this context, we are reminded that, unlike DNA sequences, gene expression data lack a widespread systematic database that goes beyond sharing and browsing to embrace pattern searching and integrated modeling. This will require standardized data fields including anatomical 3D stereotactic coordinates, quantification of colocalization of gene products and histochemicals, estimated detection limits to define "nonexpressing" cells, environmental conditions relevant to expression, physiological or pathological state, cell genotype, and, of course, accession numbers for stable references. Journals and research communities have come to expect speedy submission of sequences and 3D coordinates to public electronic databanks. These banks have, in turn, dramatically catalyzed improved computational tools and applications. With López-Nieto and Nigam and others showing us interesting uses for sequence databank withdrawals and new ways to garner gene expression capital, where shall we make our next expression level deposits?

1. Lopez-Nieto, C.E. and Nigam, S. 1996. *Nature Biotechnology* **14**:857-861.
2. Williams, N. 1996. *Science* **272**:481.
3. Sarkar, G. and Sommer, S.S. 1989. *Science* **244**:331-334.
4. Nowak R. 1995. *Science* **270**:368-371.
5. Kato, K. 1995. *Nucleic Acids Res.* **23**:3685-3690.
6. Yoshikawa ,T., Xing, G.Q., and Detera-Wadleigh, S.D. 1995. *Biochimica et Biophysica Acta* **1264**:63-71.
7. Libert, F. et al. 1989 *Science* **244**:569-572.
8. Ringwald, M. et al. 1994. *Science* **265**:2033-2034.
9. http://www.informatics.jax.org/doc/gxdgen.html The Gene Expression Information Resource Project.
10. http://eatworms.swmed.edu/VLhome.html The *Caenorhabditis elegans* page in the WWW Virtual library.

George M. Church is a Howard Hughes Medical Institute Investigator in the department of genetics, Harvard Medical School, Boston, MA 02115 (church@rascal.med.harvard.edu).

Silent genes and everlasting fruits and vegetables?

Don Grierson

Death and taxes are the only certainties in life. Although biotechnology cannot change either, plant gene research over the past eight years has shown that the onset of deterioration and senescence can be greatly delayed by inactivating specific genes. This can improve the quality, storage life, and general appeal of fruit and vegetables. By reducing spoilage and waste, there are also benefits for the food supply chain. This may be particularly important in developing countries, where 50% or more of fresh produce may never reach the consumer. Until now, this research has been pioneered in tomato. In this issue of *Nature Biotechnology*, Ayub et al.[1] demonstrate for the first time that similar improvements can be obtained in melon.

Don Grierson is head of the plant science section and the BBSRC Research Group in Plant Regulation, Sutton Bonnington campus, University of Nottingham, Loughborough, UK.