Acta Biochimica Polonica

Vol. 45 No. 4/1998

929–934

QUARTERLY

This paper is dedicated to Professor Maciej Wiewiórowski, scientific mentor of one of us (J.A.R.), on his 80th birthday

New experimental and computational approaches to the analysis of gene expression

J. Antoni Rafalski✉, Mike Hanafey, Guo-Hua Miao, Ada Ching, Jian-Ming Lee, Maureen Dolan and Scott Tingey

DuPont Agricultural Biotechnology – Genomics. Delaware Technology Park, Newark, U.S.A.

Key words: DNA chip, DNA array, gene expression, EST, expressed sequence tag, soybean, embryo, elongation factor 1, EF1

Public and private EST (Expressed Sequence Tag) programs provide access to a large number of ESTs from a number of plant species, including Arabidopsis, corn, soybean, rice, wheat. In addition to the homology of each EST to genes in GenBank, information about homology to all other ESTs in the data base can be obtained. To estimate expression levels of genes represented in the DuPont EST data base we count the number of times each gene has been seen in different cDNA libraries, from different tissues, developmental stages or induction conditions. This quantitation of message levels is quite accurate for highly expressed messages and, unlike conventional Northern blots, allows comparison of expression levels between different genes. Lists of most highly expresses genes in different libraries can be compiled. Also, if EST data is available for cDNA libraries derived from different developmental stages, gene expression profiles across development can be assembled. We present an example of such a profile for soybean seed development.

Gene expression data obtained from Electronic Northern analysis can be confirmed and extended beyond the realm of highly expressed genes by using high density DNA arrays. The ESTs identified as interesting can be arrayed on nylon or glass and probed with total labeled cDNA first strand from the tissue of interest. Two-color fluorescent labeling allows accurate mRNA ratio measurements. We are currently using the DNA array technology to study chemical induction of gene expression and the biosynthesis of oil, carbohydrate and protein in developing seeds.

✉J. Antoni Rafalski, DuPont Agricultural Biotechnology – Genomics, Delaware Technology Park, Suite 200, 1 Innovation Way, P.O. Box 6104, Newark, DE 19714-6104, U.S.A., tel.: (302) 631 2612; fax (302) 631 2607; e-mail: j-antoni.rafalski@usa.dupont.com

Abbreviations: EF1, elongation factor 1; EST, expressed sequence tag; cDNA, complementary DNA; PCR, polymerase chain reaction; RT-PCR, reverse transcription PCR.

In the future, high density DNA arrays containing the whole gene complement of an organism will become available for the analysis of genome-wide patterns of gene expression.

In the recent years, genomics, a new subdiscipline of biology was created. Genomics deals with the study of the genome, the whole complement of the organism's DNA, all of its genes. The hallmark of genomics is the use of parallel, instead of more usual, serial approaches to the study of genes. Instead of studying genes one by one, a large number of genes, in some cases all of the genes of the organism, are studied simultaneously.

Expressed sequence tag (EST) sequencing, initially conceptualized by Sydney Brenner, and first realized on a large scale by C. Venter (Adams *et al.*, 1991), is the high throughput, parallel approach to gene discovery. This approach was made possible by improvements in fluorescent DNA sequencing technology and data processing and by the large number of genes already sequenced and deposited in genome data bases (Fig. 1). Newly sequences ESTs could frequently be assigned a function based on homology to a known protein or a gene from another species.

Currently, over 50% of randomly sequenced ESTs can be functionally identified by comparison with the GenBank sequences. Interspecific gene comparisons are very effective in identifying genes even across wide evolutionary distances. However, this still leaves us with around 40% of all sequenced ESTs without a functional identification. Therefore, functional genomics, that is identification of gene function using parallel approaches is becoming increasingly important. These include gene disruption (for example by transposon insertion), gene activation, gene expression studies at the mRNA and protein level, study of protein–protein interactions (yeast two-hybrid system), creation of transgenic organisms, for the purposes of

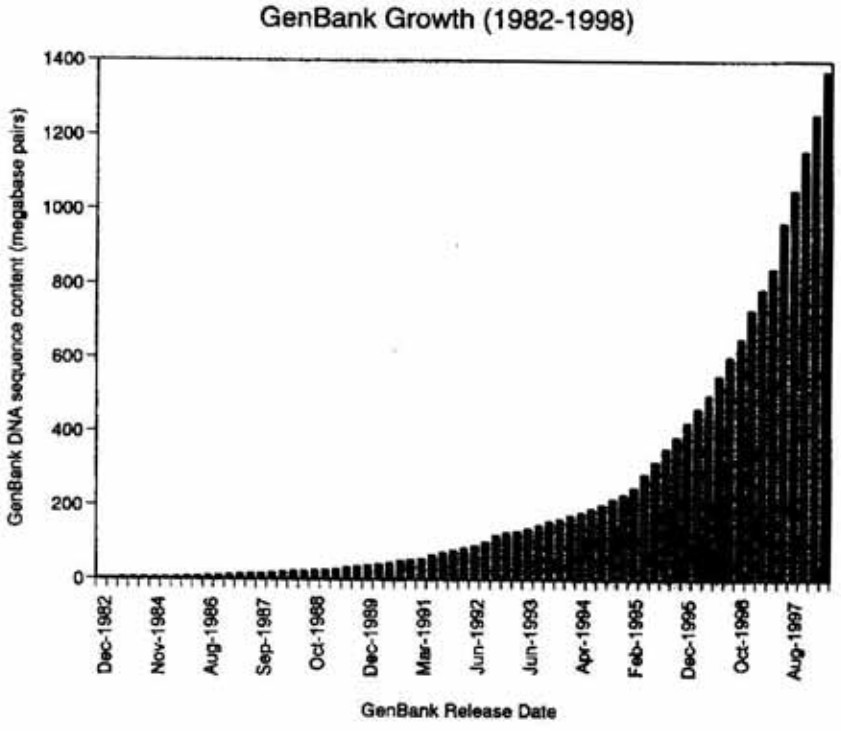

Figure 1. Growth of GenBank.

The amount of DNA sequence information in GenBank (in millions of base pairs) is plotted as a function of time. Re-plotted from the information made available by GenBank (ftp://ncbi.nlm.nih.gov/genbank/gbrel.txt).

overexpression or suppression of expression (sense or antisense), analysis of fitness of individuals carrying gene disruptions in populations. Here we will discuss two approaches to the analysis of gene expression: one experimental and one computational.

RESULTS AND DISCUSSION

Electronic Northerns: gene expression information from EST sequence data

Data bases of expressed sequence tags (ESTs) may be used to obtain gene expression data from the abundance of different messages in unbiased cDNA libraries. The idea is simple: if a message is abundant, it is likely to occur frequently among randomly sequenced EST. A rare message will occur only rarely among the EST. If the EST collection is sufficiently large, it is possible to collect enough data to produce a reasonably accurate estimate of mRNA abundance, at least for the more highly expressed genes.

As a part of DuPont's plant expressed sequence tag project, directional cDNA libraries from six different stages of soybean embryo were produced. Single lane DNA sequence information (about 450 bp) is available for each clone. All sequences have been compared to the GenBank collection of sequences using BlastX algorithm. Also, all clones were compared to each other using FastA algorithm. This analysis, coupled with custom developed software (M.K. Hanafey, in preparation) allowed us to evaluate the abundance of each EST type in each of the cDNA libraries, by counting how many times each type of cDNA occurs in our collection and also to calculate confidence intervals on the percentage mRNA abundance estimates.

Table 1 shows the example of such mRNA abundance data for early developing soybean embryo. It is worth noting that we can rather accurately estimate the concentration of abundant messages (for example elongation factor 1). In case of less abundant messages (cdc2 kinase), the relative error of the estimate increases dramatically. In order to decrease the error it would be necessary to sequence many more ESTs, which is impractical because of costs. Therefore this methodology can only be successfully applied to relatively abundant messages (0.1–0.2% total mRNA).

If EST-sequenced DNA libraries from several developmental stages are available, one could also determine developmental mRNA abundance profiles. Figure 2 shows the developmental profile of expression of elongation factor 1. As can be seen, there is a strong burst in the level of EF1 mRNA level early in development.

Using approach described above, developmental gene expression profiles of most moderately to highly expressed genes my be quantitated, to help understand the function of the corresponding gene products in the biological processes. Expression levels of ESTs of un-

Table 1. Example of the quantitation of gene expression levels in stage 1 of soybean embryo development, based on the sequencing of 3826 cDNA clones.

Several mRNA species representing different abundance levels were arbitrarily selected.

Gene identification	Number of occurrences	Total mRNA (%)	90% Confidence interval
EF1alpha	114	2.98	2.52–3.43
Annexin p33	59	1.54	1.21–1.86
40S ribosomal protein S8	29	0.76	0.52–0.98
40S ribosomal protein D17	26	0.68	0.46–0.89
40S ribosomal protein S4	20	0.52	0.33–0.71
Cdc2 kinase	5	0.13	0.03–0.22

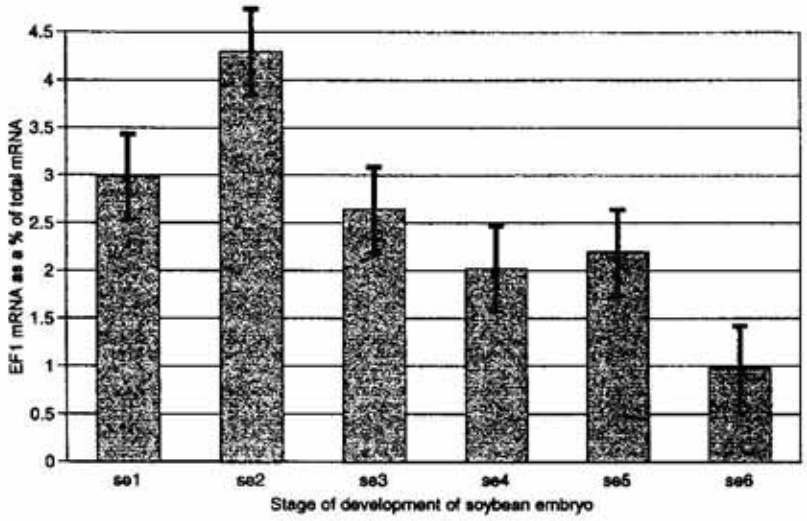

Figure 2. Expression level of elongation factor 1 mRNA, expressed as a percentage of total mRNA in six stages of soybean embryo development, se1 (early) to se6 (late), calculated on the basis of EF1 cDNA abundance in randomly sampled cDNA libraries (see text).

Error bars show 90% confidence intervals (M. Hanafey, unpublished data).

known function may help in their functional identification.

DNA arrays

Gene expression analysis could provide important information about gene function. For example, induction of the gene pathogens and wounding may indicate function in the defense processes. Gene expression studies could be done at the protein level, for example by performing two-dimensional gel electrophoresis analysis, followed by mass spectroscopy. More commonly, gene expression is performed at the messenger RNA level.

Northern blotting is the traditional method for the analysis of steady state RNA levels. Other methodologies, for example RT-PCR, could provide more quantitative results, but, like Northern blots, are usually limited in the total number of genes that could be simultaneously analyzed. Reverse dot blot provides an opportunity to analyze a large number of DNA samples simultaneously (Gress *et al.*, 1992; Nguyen *et al.*, 1995; Pietu *et al.*, 1996). DNA samples are deposited as small spots on a substrate, usually a Nylon membrane. The membrane is probed with labeled (^{32}P or ^{33}P is commonly used) first strand cDNA, obtained by reverse transcription of the poly(A)$^+$ RNA sample from tissue of interest. After hybridization and removal of unbound probe the radioactivity associated with individual spots of DNA is quantitated by autoradiography, or, more accurately, phosphoimaging. The intensity of the signal associated with each spot corresponds to the expression level of the gene represented by the spot. This method appears to give good results, and allows analysis of a large number of genes, if robotic devices for the preparation of densely spotted membranes are available.

Recently, two novel approaches to the production of reverse Northern data have been described. Schena *et al.* (1995) used a robotic device to deposit up to several thousand samples of DNA on a glass microscope slide coated with poly-L-lysine. The DNA was then probed with a mixture of two probes labeled with different fluorophores and derived from different poly(A)$^+$ RNA samples of interest. By measuring the ratio of fluorescence at two different wavelengths for each of the DNA spots the relative expression levels can be quantitated. The advantage of this approach is that the ratio of gene expression in two samples is directly recorded, increasing the reliability of the assay. Relatively high densities of DNA deposition on glass can be obtained (up to 10000 or more samples per 25 mm × 75 mm glass slide), and sensitivity is reported to be very good (Derisi *et al.*, 1996; Schena *et al.*, 1996; Shalon *et al.*, 1996).

Affymetrix is a company known for their development of DNA "chip" – technology to synthesize oligonucleotides directly on a solid surface using photomask technology originally developed to produce semiconductor devices (Fodor, 1991; Chee *et al.*, 1996). They have recently described application of high density DNA "chips" to gene expression analysis (Lockhart *et al.*, 1996). Current versions of the "chips" contain up to 400000 DNA sequences and could be used for the analysis of expression levels of up to 20000 genes. This brings closer the possibility of analyzing the expression of all genes of an organism on a single "chip". This technology is still very expensive, but the prices are expected to go down, and "off the shelf" chips for more commonly studied organisms are expected to become available at reasonable prices. Genetic differences between individuals could cause some difficulty in applying this technology to the more polymorphic species.

Other approaches to the construction of DNA arrays that may be useful for expression analysis have also been described (Yershov *et al.*, 1996).

We are using DNA arrays to improve our understanding of corn seed development. Selected cDNAs, corresponding to genes of interest to us, in particular those of the enzymes involved in the biosynthesis of fatty acids, carbohydrates and amino acids in the developing corn seed, are PCR amplified and deposited on the slide using prototype equipment produced by Molecular Dynamics company. Expression levels of these genes in different developmental stages and different genotypes of corn are quantitated and correlations between the gene expression levels of individual genes and traits of interest are sought.

Application of the DNA array technologies is expected to produce huge amounts of gene expression information for the whole genomes of many organisms. Development of data bases and analytical tools able to handle such information volumes and identify biologically relevant gene expression patterns is a high priority.

CONCLUSIONS

Expressed sequence tag (EST) sequencing programs, in addition to the identification of a large number of novel genes, allow one to estimate the abundance of highly and moderately expressed mRNAs by simply counting the number of occurrences a cDNA of each kind in the set of randomly sampled and sequenced clones from each cDNA library. The assumption here is that the cDNA library adequately represent the abundance of individual mRNA molecules in the poly(A)$^+$ RNA sample.

The advantage of this "electronic northern" approach to mRNA quantitation is that it provides absolute, rather than relative, mRNA level information (in terms of percentage of total mRNA content), and therefore allows direct comparisons between steady state levels of different mRNAs. This is generally impossible using Northern blot hybridization, because the efficiency of labeling of different mRNA molecules is different.

The disadvantage of the electronic Northern approach lies in the fact that it relies on the availability of highly redundant DNA sequence data, which is expensive to produce, and in that accurate estimates of mRNA abundance could only be made for moderately to highly expressed genes.

DNA array technology is expected to develop rapidly and produce massive amount of gene expression data, allowing more quantitative description of gene expression patterns during development, biotic and abiotic stress induction, leading ultimately to a quantitative description of the regulatory and enzymatic networks involved.

As complete genome sequences of many organisms become available, attention will turn from structural genomics to functional ge-

nomics. DNA arrays containing all the genes of an organism will be used for parallel studies of gene expression. This information, in turn, will allow reconstruction of complex regulatory networks involved in regulating development, response to the environment and metabolic activity. Another type of DNA array, containing a representation of allelic diversity at many genetic loci distributed throughout the genome will allow rapid mapping of genes associated with traits of interest, and diagnosis of genetic mutations.

It is worthwhile to remember that these tools, while extremely useful, are only aids in understanding biological phenomena. With the new and better toolbox we return to studying biology in all of its phenotypic and biochemical diversity.

The authors would like to thank their colleagues at DuPont for fruitful discussions. We would also like to thank Barbara Mazur for creating a scientifically stimulating and collegial environment for research in plant biotechnology at DuPont.

REFERENCES

Adams, M.D., *et al.* (1991) Complementary DNA sequencing: Expressed sequence tags and the human genome project. *Science* **252**, 1651–1656.

Chee, M., Yang, R., Hubbell, E., Berno, A., Huang, X.C., Stern, D., Winkler, J., Lockhart, D.J., Morris, M.S. & Fodor, S.P.A. (1996) Accessing genetic information with high-density DNA arrays. *Science* **274**, 610–614.

Derisi, J., Penland, L., Brown, P.O., Bittner, M.L., Meltzer, P.S., Ray, M., Chen, Y.D., Su, Y.A. & Trent, J.M. (1996) Use of a cDNA microarray to analyze gene-expression patterns in human cancer. *Nature Genetics* **14**, 457–460.

Fodor, S.P.A. (1991) Light-directed, spatially addressable parallel chemical synthesis. *Abstracts of papers of the American Chemical Society* **202**, 90.

Gress, T.M., Hoheisel, J.D., Lennon, G.G., Zehetner, G. & Lehrach, H. (1992) Hybridization fingerprinting of high-density cDNA-library arrays with cDNA pools derived from whole tissues. *Mammalian Genome* **3**, 609–619.

Lockhart, D.J., Dong, H., Byrne, M.C. Follettie, M.T., Gallo, M.V., Chee, M.S., Mittman, M., Wang, C., Kobayashi, M., Horton, H. & Brown, E.L. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology* **14**, 1675–1680.

Nguyen, C., Rocha, D., Granjeaud, S., Baldit, M., Bernard, K., Naquet, P. & Jordan, B.R. (1995) Differential gene expression in the murine thymus assayed by quantitative hybridization of arrayed cDNA clones. *Genomics* **29**, 207–216.

Pietu, G., Alibert, O., Guichard, V., Lamy, B., Bois, F., Leroy, E., Mariage-Samson, R., Houlgatte, R., Soularue, P. & Auffray, C. (1996) Novel gene transcripts preferentially expressed in human muscles revealed by quantitative hybridization of a high density cDNA array. *Genome Res.* **6**, 492–503.

Schena, M., Shalon, D., Davis, R.W. & Brown, P.O. (1995) Quantitative monitoring of gene-expression patterns with a complementary-DNA microarray. *Science* **270**, 467–470.

Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P.O. & Davis, R.W. (1996) Parallel human genome analysis – microarray-based expression monitoring of 1000 genes. *Proc. Natl. Acad. Sci. U.S.A.* **93**, 10614–10619.

Shalon, D., Smith, S.J. & Brown, P.O. (1996) A DNA microarray system for analyzing complex DNA samples using 2-color fluorescent-probe hybridization. *Genome Res.* **6**, 639–645.

Yershov, G., Barsky, V., Belgovskiy, A., Kirillov, E., Kreindlin, E., Ivanov, I., Parinov, S., Guschin, D., Drobishev, A., Dubiley, S. & Mirzabekov, A. (1996) DNA analysis and diagnostics on oligonucleotide microchips. *Proc. Natl. Acad. Sci. U.S.A.* **93**, 4913–1918.