Minireview

The *Arabidopsis thaliana* cDNA sequencing projects

Michel Delseny*, Richard Cooke, Monique Raynal, Françoise Grellet

University of Perpignan, Laboratoire de Physiologie et Biologie Moléculaire des Plantes, UMR 5545 CNRS, Avenue de Villeneuve 66860, Perpignan cédex, France

Received 14 January 1997

Abstract Nearly 30 000 *Arabidopsis thaliana* EST (Expressed Sequence Tags) have been produced by a French-American consortium. Despite redundancy, these sequences tag about half of the expected *Arabidopsis* genes. Approximately 40% of the non-redundant EST can be assigned a putative function by simple homology search. This programme allowed the identification of a large number of genes which would have been very difficult to isolate by other classical techniques. It considerably stimulated many areas of plant biology by the rapid discovery a large number of genes, by revealing multigene families and by allowing the analysis of differential expression of the different members. Finally this programme facilitated construction of physical maps of the chromosomes and opened the way for complete sequencing of the *Arabidopsis* genome and comparative mapping of the major plant crops.

Key words: Expressed Sequence Tag; cDNA; Gene function; Genome sequencing; Crucifer

1. Introduction

Arabidopsis thaliana, a small crucifer weed, was chosen by a number of plant biologists a few years ago as a model to analyse plant genome structure and gene expression as well as plant development. The major advantages of this plant are its small genome, short life cycle, genetics and relative easiness of transformation [1]. With a genome size of 120 Mbp, *Arabidopsis* compares well with animal models such as *Drosophila* or *Caenorhabditis* and is one of the smallest plant genomes. More than 1000 mutants have been characterized and half of them have been mapped to one of its five chromosomes [2], making it a suitable model to dissect metabolism, development and transduction pathways using molecular genetics. Due to specific difficulties, as well as a much smaller scientific community, gene isolation and characterization in higher plants has been lagging behind animals. Thus, at the end of 1991, no more than 500 different plant protein sequences were represented in databases. In order to increase rapidly our knowledge of this genome and, hopefully, that of other crop plant genomes, two major strategies have been developed. The first consisted in partially sequencing a large number of cDNA in order to make a catalogue of *Arabidopsis* expressed genes (EST, or Expressed Sequence Tags) as initially described for human genes [3]. The second strategy consisted in sequencing the complete genome. Although the two strategies are complementary and part of a world-wide integrated project [2], this minireview will be restricted to the EST projects and their impact on plant science.

2. EST strategy and production

Most *Arabidopsis* EST sequencing has been carried out by two consortia: one in France, started by the end of 1991 with national support and continued after 1993 with European Union funding as part of the ESSA (European Scientists Sequencing *Arabidopsis*) programme, and one in the US initiated a few months later at Michigan State University. The initial French project was to obtain the largest number of EST, to identify as many of them as possible by homology search, and to map them on the genetic and physical maps. In contrast to their American colleagues, who prepared a single library from a mixture of mRNA from different tissues, the French scientists worked with specialized libraries corresponding to specific tissues, organs or developmental stages. This strategy initially helped to avoid overlaps and redundancy between groups as well as maintaining a strong biological interest for the participants and it allows one to compare the different libraries. A computer network was organised so as to automatically identify redundancies and submit sequence information to the EMBL database. All sequences were first edited and analyzed by homology search in each individual laboratory. After validation of the sequence and annotation, non-redundant EST were then transferred to the EMBL database from where they were automatically integrated into dbEST [4]. Another difference between the two projects was that American EST were sequenced only from the 5′ end and redundancy was not eliminated at the submission stage whereas in the French project non-redundant clones were sequenced from both ends.

Altogether, the two programmes have produced close to 30 000 EST, placing *Arabidopsis* genome in fourth position after human and mouse EST, immediately behind *Caenorhabditis elegans* and ahead of rice [5]. The French groups produced approximately 12 000 EST, but only 6500, which were not redundant with previously submitted EST, were deposited in dbEST. Most of the corresponding clones are available from the Ohio State University *Arabidopsis* Biological Resource Centre (ABRC) in Columbus, Ohio.

It is difficult to determine precisely how many different genes these EST represent because EST sequences are not 100% reliable because the collection is a mixture of 5′ and 3′ ends and because cDNA are not always full length. An additional problem is that many cDNA correspond to mem-

*Corresponding author. Fax: (33) 4-68-66-84-99.
E-mail: delseny @univ-perp.fr

This paper was presented at the 24th FEBS Meeting in Barcelona.

PII S0014-5793(97)00075-6

bers of multigene families which are not easily distinguishable. A first estimation consists in assembling overlapping EST into contigs: the 30 000 EST can be organized into about 15 000 contigs [6] but this is certainly an overestimate of gene number and a more realistic speculation is that 10 000 distinct genes have been tagged by at least one EST. This figure has to be compared with the gene number which is now estimated to be around 20 000 and with the 50 or 60 genes which have been isolated using T-DNA or transposon tagging or map-based cloning [7]. Thus, just by random sequencing of cDNA, roughly half of the genes have been tagged; they probably correspond to the most abundantly expressed genes. Sequencing work is almost completed in both consortia because redundancy is now becoming too high for the programmes to be cost effective. Tagging genes with rare transcripts cannot be achieved using the present strategies and will require further developments.

EST are usually ca. 300 bp long and represent only part of a gene. However, about 150 full-length cDNA have now been completely sequenced, starting from an EST. In addition, if a gene is abundantly expressed, overlapping EST can be assembled in order to reconstruct a full-length cDNA *in silico*.

3. What genes have been identified by EST sequencing?

A preliminary identification of a gene relies on the presence of an already identified homologous sequence in databases. Each sequence was translated in all six possible reading phases. Using the BLAST and FASTA algorithms [8,9], they were aligned with the non-redundant protein sequence database. In this way, more than 40% of the non-redundant EST were found to have significant homology, at least at the amino acid sequence level, with an already known gene.

Lists of putative functions for EST can be found in the papers reporting major progress in the programmes [10–12]. Such an analysis simply gives a hint of what the function might be and, in many cases, extensive biochemical and biological work is necessary to unambiguously identify a gene and its function. This task is clearly impossible for the sequencing labs, therefore, relevant clones were distributed to colleagues who have an interest in a specific gene.

Similarity searches must be regularly repeated because information in databases is increasing daily at an exponential rate. The percentage of putatively identified proteins should progressively increase when more full-length cDNA or genes are analyzed and when more functions are described in plants and other organisms. Very often when a new gene is isolated by classical methods, identical EST with no previously assigned function can be found in the databases: this situation was met, for instance, for most of the fatty acid desaturases. Many new genes, which would have been very difficult to isolate using any other strategy, or which were not even suspected to occur in plants, have been tagged in a very short time. Examples are many EST corresponding to well-known proteins of the translation and transcription machineries, of the cell death and protein degradation programmes, to enzymes in major metabolic pathways, and to many proteins involved in various transduction pathways [12,13]. Such results point to the large number of genes which are shared by most organisms. However, nearly 60% of the non-redundant EST have no match with anything which might indicate that a large number of genes are plant specific. Alternatively the sequenced region might correspond to a variable region in a conserved gene.

A major surprise was to discover that many genes belonged to multigene families: for example, we found that there were at least five distinct thioredoxin h genes which are differentially expressed whereas there are three genes in yeast and a single one in human [14]. Similarly we reconstructed 106 different cDNA coding for 50 distinct types of ribosomal proteins. The majority are coded by two, three and even four genes which differ essentially in their non-coding regions [15].

4. Impact of the *Arabidopsis* EST programmes

One of the biggest impacts of the *Arabidopsis* EST programmes is on plant biology studies. A crude measure of this impact is the thousands of clone requests which were received by the ABRC. Availability of the EST sequences accelerated the isolation of many new genes for which a partial protein sequence was available. It also stimulated and facilitated the obtention of recombinant proteins and the analysis of differential expression of members of multigene families by allowing one to prepare gene-specific probes [14,16–19]. Having available a large number of EST allows analysis of overall gene expression. cDNA clones can be gridded on high-density filters and hybridized with various complex probes corresponding to different development stages, different stimuli or different genetic backgrounds [20]. Alternatively oligonucleotides derived from the EST can be spotted on DNA chips [21]. When several cDNA libraries have been used, the frequency of occurrence of the various clones can also be analysed and compared. Another type of physiological study consists in attempting to classify cDNA and genes according to their control by a given regulatory gene. This strategy was developed for a number of genes expressed during seed maturation to determine whether their expression was dependent upon regulatory gene *ABI 3* [22].

EST programmes have an important impact on mapping and sequencing the *Arabidopsis* genome. From the very beginning, mapping the EST on the chromosomes was a major goal of the French consortium because it was anticipated that it would facilitate positional cloning. It was essentially carried out on YAC clones using PCR-pooling strategies or colony hybridization. As a result, about 600 EST have been now located on YACs, in France. Meanwhile several hundred additional EST have been positioned by two other groups [7]. Therefore within the last 2 years approximately 1500 different EST have been positioned on the physical map. This mapping effort has considerably accelerated the organisation of contigs on the various chromosomes [23,24]. Only part of this information has been released, but it should be publicly available as soon as the physical map of the *Arabidopsis* genome is completed.

Another outcome of the EST programmes is facilitation of gene identification in the genomic sequencing programme. Plant genes have different nucleotide compositions, different sizes of exons and introns from those of animals and most of the gene prediction programmes developed for other organisms are inappropriate for *Arabidopsis* [25]. At the beginning of systematic genomic sequencing gene identification relied essentially on use of EST and isolation of cognate cDNA, which in turn could be used to train existing programmes. Now that 2.5 Mbp of genomic sequence have been determined

by the ESSA programme [7], about 40% of the predicted genes match an EST. This observation strongly supports the conclusion that nearly half of the transcribed genes have been tagged by the EST programmes.

EST sequences can be compared with other sequences from an evolutionary point of view. They have been searched for sequences coding for the Ancient Conserved Regions (ACR) [26] which are common to a number of proteins conserved in different phyla: more than two-thirds of the known ACR can be recognized in the *Arabidopsis* EST collection. Another comparison has been made with the rice EST, directly at the nucleotide level and it was demonstrated that, despite different GC content and codon usage, more than 300 non-redundant *Arabidopsis* EST matched a rice one with a BLAST X score greater than 400, which predicts cross-hybridization in reasonably stringent conditions [12,27].

5. Perspectives and conclusions

The random and partial sequencing of *Arabidopsis thaliana* cDNA has been a particularly rewarding programme, producing close to 30 000 sequences and tagging virtually half of the genes of this plant. Although EST sequencing, per se, is almost completely stopped there is still an enormous amount of data to analyse and exploit and several important tasks remain to be achieved.

The first is certainly to reconstruct as many full-length cDNA as possible from overlapping EST, and identify the members of multigene families.

The second consists in the identification of the functions of the genes tagged with an EST, particularly those which do not resemble anything known. Unfortunately homologous recombination is not efficient in plants to disrupt genes but alternative strategies consisting in saturating the genome with DNA insertions are being developed [28,29]. *Arabidopsis* should also largely benefit from gene function search programmes developed in other organisms such as yeast. Anyway it is clear that extensive biochemical work will be necessary to identify definitively all the gene functions.

A third task is to identify genes which have not yet been tagged by an EST and which may represent another half of the genes. They correspond essentially to the low expressed genes which are often important regulatory genes. However, it might be more cost effective to sequence the complete genome directly, a task which is expected to be completed within the next 7 or 8 years [1,7].

Finally, another extremely important task is to transfer as much of our knowledge of *Arabidopsis* genome as possible to economically important crop genomes. EST can not only be used to fish out the homologous genes in other plants using a variety of strategies, but also, at least a few percent can be used directly in comparative mapping strategies, to facilitate mapping of the main crop genomes and provide a framework of identified markers common to many plant species [27].

As a conclusion, the *Arabidopsis* EST programme is a strong foundation for many biological studies in the future and for complete genome sequencing. With the comparative Rice Genome Programme [30] it demonstrated the efficiency and feasibility of this technology in plants and proved that it was possible to tag a significant fraction of the expressed genes, at a reasonable cost and within a short time. As a consequence similar projects were induced in a variety of crop species. Last, but not the least, it profoundly changed the way of thinking and work habitudes of many plant biologists by forcing them to use computer network and database resources and to make use of the remarkable conservation of many genes in all living organisms to solve specific plant biology problems.

Acknowledgements: The authors wish to acknowledge stimulating discussions and communication of unpublished information from many colleagues in their laboratories and in the *Arabidopsis* scientific community. Work in their laboratories on *Arabidopsis* sequencing was supported by CNRS (URA 565 and GDR 1003), by GREG and by European Union Grants (Bio 2CT-930075).

References

[1] Somerville, C.R. and Meyerowitz, E.M. (1994) in: Arabidopsis (Somerville, C.R. and Meyerowitz, E.M. eds.) Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp. 1–6.
[2] Meinke, D., Caboche, M., Dennis, E., Flavell, R., Goodman, H., Jurgens, G., Last, R., Martinez-Zapater, J., Mulligan, B., Okada, K. and Van Montagu, M. (1995) The Multinational Coordinated *Arabidopsis thaliana* Genome Research Project. Progress Report: Year Five, 59 pp., National Science Foundation Publication 96–43, Arlington, VA (also available via the Internet through the NSF home page: http://www.nsf.gov).
[3] Adams, M.D., Kelley, J.M., Gocayne, J.D. et al. (1991) Science 252, 1651–1656.
[4] Boguski, M.S., Lowe, T.M.J. and Tolstoshev, S.H. (1993) Nature Genet. 4, 332–333 (http://www.ncbi.nlm.nih.gov/dbEST/index.html).
[5] dbEST Release 11.96.
[6] Rounsley, S., Glodek, A., Sutton, G., Adams, M.D., Somerville, C.R., Venter, J.C. and Kerlavage, A.R. (1996) Plant Physiol. 112, 1177–1183.
[7] Somerville, S. and Somerville, C.R. (1996) Plant Cell 8, 1917–1933.
[8] Pearson, W.R. and Lipman, D.J. (1988) Proc. Natl. Acad. Sci. USA 895, 2444–2448.
[9] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) J. Mol. Biol. 215, 403–410.
[10] Höfte, H., Desprez, T., Anselem, J. et al. (1993) Plant J. 4, 1051–1061.
[11] Newman, T., De Bruijn, F.J., Green, P., Keegstra, K., Keude, H., McIntosh, L., Ohlrogge, J., Raikhel, N., Somerville, S., Thomashow, M., Retzel, E. and Somerville, C.R. (1994) Plant Physiol. 106, 1241–1255.
[12] Cooke, R., Raynal, M., Laudié, M., Grellet, F., Delseny, M. et al. (1996) Plant J. 9, 101–124.
[13] Gallois, P., Makishima, T., Hecht, V., Desprez, B., Laudié, M., Nishimoto, T. and Cooke, R. (1997) Plant J. (in press).
[14] Rivera-Madrid, R., Mestres, D., Marinho, P., Jacquot, J.P., Decottignies, P., Miginiac-Maslow, M. and Meyer, Y. (1995) Proc. Natl. Acad. Sci. USA, 92, 5620–5624.
[15] Cooke, R., Raynal, M., Laudié, M. and Delseny, M. (1997) Plant J. (in press).
[16] Genschik, P., Durr, A. and Fleck, J. (1994) Mol. Gen. Genet. 244, 548–556.
[17] Phillips, A.L., Ward, D.A., Uknes, S., Appleford, N.E.J., Lange, T., Huttly, A.K., Gaskin, P., Graebe, J.E. and Hedden, P. (1995) Plant Physiol. 108, 1049–1057.
[18] Herzog, M., Dorne, A.M. and Grellet, F. (1995) Plant Mol. Biol. 27, 746–752.
[19] Xu, W., Campbell, P., Vargheese, A.K. and Braam, J. (1196) Plant J. 9, 879–889.
[20] Hervé, C., Perret, E., Tremousaygues, D. and Lescure, B. (1996) Plant Physiol. Biochem. 34, 425–430.
[21] Schena, M., Shalon, D., Davis, R.W. and Brown, P.O. (1996) Science 270, 467–470.
[22] Parcy, F., Valon, C., Raynal, M., Gaubier-Commella, P., Delseny, M. and Giraudat, J. (1994) Plant Cell 6, 1567–1582.
[23] Schmidt, R., West, J., Love, K., Lencham, Z., Lister, C., Thompson, H., Bouchez, D. and Dean, C. (1995) Science 270, 480–483.

[24] Zachgo, E.A., Wang, M.L., Dewdney, J., Bouchez, D., Camilleri, C., Belmonte, S., Huang, L., Dolan, M., Goodman, H. (1996) Genome Res. 6, 19–25.
[25] Hebsgoard, S.M., Korning, P.G., Tolstrup, N., Engelbrecht, J., Rouzé, P. and Brunal, S. (1996) Nucl. Acids Res. 24, 3439–3452.
[26] Green, P., Lipman, D., Hittier, L., Waterston, R., States D. and Claverie, J.M. (1193) Science 259, 1711–1716.
[27] Delseny, M., Raynal, M., Laudié, M., Varoquaux, F., Comella, P., Wu, H.J., Cooke, R. and Grellet, F. (1996) in: Unifying Plant Genomes (Heslop-Harrison, J.S. ed.), SEB Symposium No. 50. The Company of Biologists Ltd., Cambridge, UK, pp. 5–9.
[28] McKinney, E.C., Aali, N., Trant, A., Feldmann, K.A., Belotosky, D.A., McDowell, J.M. and Meagher R.B. (1995) Plant J. 8, 613–622.
[29] Krysan, P.J., Young, J.C., Tax, F. and Sussman, M.R. (1996) Proc. Natl. Acad. Sci. USA 93, 8145–8150.
[30] Sasaki, T., Song, J., Koga-Ban, Y. et al. (1994) Plant J. 6, 615–624.