Update on Genomics

Functional Genomics in Plants

David Bouchez and Herman Höfte*

Laboratoire de Biologie Cellulaire, Institut National de la Recherche Agronomique, Route de Saint-Cyr, 78026 Versailles cedex, France

The sequencing of the first genome of a higher plant, Arabidopsis, is progressing at a rapid pace. As we are writing these words, almost one-third of the estimated 100 Mb is available in public databases, and with appropriate funding the complete genome sequence is predicted to be finished by the year 2001. Sequencing programs for other plant genomes such as rice are planned, and with new generations of more efficient sequencing machines (Marshall and Pennisi, 1998), the sequence of several other plant genomes may become available in the coming decade. In the wake of these sequencing efforts, plant research enters an exciting period in which genome-wide approaches are becoming an integral part of plant biology, with potentially highly rewarding but as yet unpredictable biotechnological applications. This is reflected by the current frenzy with which new agricultural biotechnical companies are being founded and the rate at which existing companies are investing in the development of tools to exploit and further expand this wealth of information.

The term functional genomics can be referred to as the "development and application of global (genome-wide or system-wide) experimental approaches to assess gene function by making use of the information and reagents provided by structural genomics" (Hieter and Boguski, 1997). With these approaches the focus of the analysis is shifted from individual components to biological systems. It involves the use of high-throughput methods for the study of large numbers of genes (ideally the entire set) in parallel. Gene “function” can be considered from several points of view: it can mean biochemical function (e.g. protein kinase), cellular function (e.g. a role in a signal transduction pathway), developmental function (e.g. a role in pattern formation), or adaptive function (the contribution of the gene product to the fitness of the organism). Having identified a new sequence, the comparison with sequence databases is the simplest way to obtain (essentially biochemical) functional information.

Currently, about 50% of newly identified genes show sequence similarity to previously described genes. However, computerized analyses are generally not sufficient to define gene function with a high level of confidence, and experimental confirmation is needed in most cases. Indirect information on cellular or developmental function can be obtained from spatial and temporal expression patterns; for example, the presence of mRNA and/or protein in different cell types, during development, during pathogen infection, or in different environments. The subcellular localization and posttranslational modifications of proteins can be informative as well. Knocking out or overexpressing the gene permits the gene sequence to be linked to a phenotype from which a cellular role or a role in development may be deduced. Finally, the fitness of plants carrying mutations or natural variants for the gene can be compared in different environments with wild-type plants to study the adaptive function.

In this Update we provide a review of the ever-growing toolbox for the global study of gene function in plants, indicating the potential and the limitations of the different techniques. Where appropriate, we will also draw a parallel with the more advanced technologies in bacterial, yeast, and animal systems.

LARGE-SCALE SEQUENCING OF PLANT GENOMES

ESTs

A rapid way to establish an inventory of expressed genes is by determining partial sequences of cDNA called ESTs. In this approach, single-pass sequences of 300 to 500 bp are determined from one or both ends of randomly chosen cDNA clones. The sequences are sloppy and have a relatively high error rate, but they are sufficiently accurate to unambiguously identify the corresponding gene in most cases.

Thousands of sequences can thus be determined with a limited investment. EST information is present in public databases for a variety of species, including a number of plants (Höfte et al., 1993; Newman et al., 1994; Cooke et al., 1996; Yamamoto and Sasaki, 1997). A recent release (March 23, 1998) of the EST database of the National Center for Biotechnology Information (Table I) lists more than 39,000 Arabidopsis ESTs. Powerful programs for rapid sequence comparisons with databases are accessible via the Internet (Table I). The Institute for Genome Research (Table I) very conveniently has ordered overlapping ESTs into tentative contigs. It is difficult to know exactly how many different

* Corresponding author; e-mail hofte@versailles.inra.fr; fax 33–1–30–83–30–99.

1 This study was supported by grants from the BIOTECH program of the European Economic Community (nos. BIO4-CT95-0183 and BIO4-CT96-0689 to D.B.); from Groupement de Recherche du Genome (contract nos. 9147A and 5-95 to D.B. and no. 10-95 to H.H.); and from Action Incitative Programme Institut National de la Recherche Agronomique (no. P176 to D.B. and H.H.).

Abbreviations: EST, expressed sequence tag; SAGE, serial analysis of gene expression.
genes are represented by these ESTs, because cDNA clones frequently are truncated at the 5' end, and sequencing of 5' ends of cDNA clones often yields nonoverlapping sequences corresponding to the same mRNA. However, the fact that 56% of the predicted genes exactly matched an EST within 1.9 Mb of contiguous genomic sequence (Bevan et al., 1998) suggests that more than one-half of the total set of Arabidopsis genes is already represented by an EST. It should be noted that several companies possess large private EST databases for various crop plants (e.g., maize and soybean), the access to which can be negotiated on a case-by-case basis.

While we await the complete genome sequence, EST databases have proven to be a tremendous resource for finding genes and for interspecies sequence comparison, and have provided markers for genetic and physical mapping and clones for expression analyses. The relative abundance of ESTs in libraries prepared from different organs and plants in different physiological conditions also provides preliminary information on expression patterns for the more abundant transcripts.

### Genomic Sequencing

EST sequencing programs in Arabidopsis and rice have been extremely successful in the discovery of new genes. However, rare transcripts and transcripts of genes that are induced under specific conditions (e.g. biotic and abiotic stresses) are not represented in EST databases. The only sure way to gain access to the entire set of genes is to determine the complete genomic sequence. This sequence also provides information on the global structure of the genome, including the relative order of genes on the chromosomes, which is extremely valuable for positional cloning strategies. At present the genomes of rice and Arabidopsis, both model plant species with small, information-rich genomes, are being sequenced on the basis of coordinated multinational efforts. For the Arabidopsis Genome Initiative, specialized sequencing laboratories from the United States, Europe, and Japan are continuously releasing sequence data in public databases (Table I) annotated with the results from database searches and gene-prediction algorithms (Rounsley et al., 1998).

The first results from the large-scale sequencing of a 1.9-Mb region (Bevan et al., 1998) confirmed that the Arabidopsis genome is extremely gene rich (one gene every 5 kb, on average) and poor in repeated elements such as retrotransposons. The total gene complement of Arabidopsis is estimated to be around 21,000 protein-coding genes. About one-half of the predicted genes on the 1.9-Mb segment can be assigned to a functional category based on similarity with known proteins.

A major problem with genomic sequences is how to distinguish coding regions from noncoding intergenic sequences and introns. Comparisons with EST and cDNA sequences and sequence similarity to known coding sequences can be used to assign intron positions for many genes. However, for the genes that do not match sequences in the databases, the coding sequences need to be predicted from the genomic sequence. For the human genome, programs such as GRAIL (Table I) have been developed to predict the beginnings and ends of genes and intron positions with high reliability (Uberbacher and Mural, 1991). In plants splicing signals are still poorly defined, and in Arabidopsis introns are often very small; the average size is around 200 bp and on average six or seven introns are found per gene. A gene-prediction algorithm based on neuronal networks was developed for plant sequences. The program, NetPlantGene, which is freely available (Table I), allows for a reliable prediction of introns (Hebsgaard et al., 1996; Tolstrup et al., 1997). A better comprehension of splicing mechanisms in plants is needed to improve gene-prediction algorithms in the future.

### FUNCTIONAL ANALYSIS OF PLANT GENES

#### Gene Expression

Knowing when and where a gene product (RNA and/or protein) is expressed can provide important clues to its biological function. The facility with which a high-throughput approach can be used for the study of gene expression depends largely on the level of regulation that is being addressed. Technologies have been developed for the parallel study of mRNA and protein levels in yeast, animals, and plants. Additionally, procedures for the systematic study of protein-protein interactions using two-hybrid...
screens (Fromont-Racine et al., 1997) are being developed for yeast and may be adopted for plant systems in the future.

**SAGE**

SAGE is a logical extension of EST sequencing. An inventory of transcripts is established based on very short cDNA sequence tags (9–11 bp) from the 3’ part of mRNA molecules, concatenated, and cloned before sequencing (Velculescu et al., 1995). These short tags have proven sufficiently long to unambiguously identify corresponding genes in databases (Velculescu et al., 1997). Expression patterns for the different genes are reflected by the relative abundance of individual tags.

SAGE patterns have been studied in human and in yeast (Polyak et al., 1997; Velculescu et al., 1997), but to our knowledge, the technique has not yet been applied to plants. A prerequisite for the identification of the tags is the availability of large sequence databases for the species under study. The technique is powerful but not very convenient for the comparison of many different samples and for the study of the rarer transcripts.

**Chip in for DNA Chips**

An alternative approach for monitoring mRNA levels is based on hybridization techniques. A “reverse northern” technique is used, whereby DNA fragments or oligonucleotides corresponding to different genes or cDNAs are immobilized on a solid support and hybridized to probes prepared from total mRNA pools extracted from cells, tissues, or whole organisms and converted to cDNA (Fig. 1A). The hybridization signal for each individual spot can be quantified automatically, and in principle reflects the relative abundance of the corresponding mRNA in the total mRNA pool. The value of this approach is its propensity for miniaturization, allowing huge numbers of gene fragments to be analyzed in parallel. For example, DNA fragments corresponding to the entire set of more than 6400 yeast open reading frames could be contained on a single 18×18-mm microscope slide (DeRisi et al., 1997).

Different systems have been developed depending on the source of target DNA and the nature of the solid support and the detection system. The simplest system,
which uses nylon filters in combination with radioactive probes, can be put to work with a minor investment. It requires a commercial, all-purpose gridding robot and a radioactivity detection system such as a Phosphorimager. Software for the detection and quantification of signals is commercially available. The system was tested out on some 800 partially sequenced cDNA clones in Arabidopsis, with reproducible results (Desprez et al., 1998). Filters with a density of up to 6144 cDNA on a 12-×-8-cm surface have been made (H. Höfte, unpublished data). The technique can be used to detect low-abundance mRNAs (down to 1:10,000 of the total mRNA population).

A more sophisticated system uses DNA microarrays printed at a high density on pretreated glass slides (Schena et al., 1995; DeRisi et al., 1997). This system allows the use of fluorescent probes, and hybridization signals are detected using an adapted confocal laser microscope (Fig. 1B). The use of two different fluorophores allows the simultaneous detection of hybridization signals of two probes in a single hybridization experiment, resulting in highly reproducible data. The system has a high sensitivity: mRNAs of an abundance down to 1:100,000 can be reliably detected, which in yeast corresponds to approximately 0.15 mRNA molecules per cell. The major drawback of this technique is its cost and the requirement for a specialized arraying robot and scanner. Also, the arrays cannot be reused, which further increases cost.

An alternative technology is referred to as DNA chips. It is based on a method to synthesize large amounts of different oligonucleotides in situ on a glass support using light-directed, solid-phase, combinatorial chemistry developed by Affymetrix (Santa Clara, CA). Oligonucleotides corresponding to the genes of interest are hybridized with fluorescent probes, and signals are detected using techniques similar to the ones used for the microarrays. A hallucinating experiment was described recently (Wodicka et al., 1997) in which 260,000 25-mers, corresponding to a nearly complete set of yeast genes, were synthesized in situ on four 1.28-cm2 grids. Each open reading frame predicted from the genomic sequence was represented by about 20 oligonucleotides to cancel out differences in hybridization behavior of different oligonucleotides and cross-hybridization with related sequences. In addition, the hybridization value for each oligonucleotide was corrected for the hybridization value for a second, negative control oligonucleotide with an identical sequence except for a single base difference. The method showed a very high sensitivity and reliably detected a transcript present at one copy in 10 to 20 cells. The company is planning to market a human chip with elements representing 40,000 genes for expression analysis. DNA chips for the complete set of Arabidopsis genes may become possible in the not-too-distant future, provided appropriate funding is available for manufacturing the photolithographic masks. For the moment, the technology remains inaccessible to plant scientists in public laboratories.

In the near future, genome-wide analysis of gene expression will become possible for plants. Several public laboratories are currently setting up systems for Arabidopsis (Desprez et al., 1998; S. Somerville, personal communication), and a few companies are already offering high-density filters or services including probe hybridization and data analysis with cDNA microarrays.

What can be done with these techniques? First, they can be used to describe in a systematic way mRNA levels in different tissues in the course of development or in different environmental conditions. Such information can be stored in gene-expression databases that may be consulted just like sequence databases. Early versions of such databases exist for yeast and Arabidopsis (Table I). Second, global expression studies can be used to classify genes based on their spatial and kinetic expression patterns. For example, during the course of the diauxic shift of yeast cultures (from anaerobic to aerobic metabolism), classes of genes could be distinguished based on their induction or repression kinetics (DeRisi et al., 1997). New regulatory elements may be identified by comparing the regulatory sequences of genes of the same class. Third, the kinetics of changes in gene expression, combined with expression profiles of mutants for known regulatory genes, should allow the study of expression networks. For example, in yeast, changes in the expression of genes with known metabolic function were indicative of the metabolic reprogramming taking place during the diauxic shift. Along these same lines, expression profiles will be helpful in characterizing the defects in developmental mutants. Software tools to facilitate such comparisons need to be developed.

Other techniques such as systematic, whole-mount in situ hybridization studies on various organs are required to obtain more precise information on cell-type-specific expression. Alternatively, as discussed below, “enhancer” or “gene-trap” lines can also provide information on cell-type-specific gene expression.

What about Proteins?

Information on mRNA levels is not sufficient to obtain a complete picture of the way gene expression is regulated within the cell. Protein expression data are more informative, but are much more difficult to obtain in a parallel fashion. The term “proteome” was coined to refer to the total protein complement expressed by the genome. Two-dimensional PAGE is for the moment the method of choice to study the abundance and posttranslational modifications of several hundred proteins in parallel (Humphrey-Smith and Blackstock, 1997). Recently, the resolution and reproducibility of the separation system was greatly improved, and algorithms for automatic spot quantification have been developed. Furthermore, protein spots on gels are now more easily identified with improved methods for N-terminal and internal microsequencing and, more importantly, with the use of MS techniques (Jungblut and Thiede, 1997). The latter are based on the ability to determine very accurately the molecular mass of fragments released upon protease treatment of minute amounts of protein excised from gels. Comparison with theoretical molecular mass predicted from sequences in protein databases allows for the attribution of a sequence to the protein fragments, provided that a complete, organism-specific protein database is available.
Large two-dimensional-gel databases exist for *Escherichia coli*, yeast, and human. In plants more limited databases exist for maize, rice, and Arabidopsis. A database for plasma membrane proteins in Arabidopsis is under development (Table I), which not only provides information on abundance and posttranslational modification, but also on the intracellular localization of a subset of proteins. This approach can be extended to other organelles, and even multisubunit protein complexes, provided that efficient separation procedures are available.

Among the various applications, the proteome approach has been used in yeast to study gene function through the generation of knockout or overexpression mutants and for the analysis of changes in protein profiles on two-dimensional gels. In maize comparison of proteomes of lines nearly isogenic for the gene *opaque2* allowed the identification of new targets for the encoded transcription factor (Damerval and Le Guilloux, 1998). In Arabidopsis two-dimensional-gel profiles were used to characterize developmental mutants and allowed the hypothesis of the overproduction of cytokinins in one of the mutants, which was confirmed subsequently (Santoni et al., 1994).

Despite its improvements, the technique remains limited, allowing for the monitoring of only a few thousand abundant proteins. Other techniques, possibly based on MS of complex mixtures of protein fragments, may be developed in the future for the study of total protein complements.

**Forward and Reverse Genetics in Plants**

Genetics relies on the study of variants, either found in natural populations or induced by mutagenesis. The analysis of the inheritance of this variation in mapping populations allows us to determine the number of genetic factors responsible for the observed variation and their relative position on the chromosomes. This variation can be discrete, as is generally the case for mutations, or continuous, as with most complex traits (especially those of agronomic importance, such as yield and plant height), which are mostly controlled by large numbers of genes.

Mutational approaches have been extremely successful in recent years for the study of the genetic and molecular bases for any trait in plant biology. Classic chemical/physical mutagenesis procedures allow us to reach saturation relatively easily (meaning a high probability of recovering a mutation in every gene in the genome). Access to the mutated genes is obtained using positional cloning strategies (i.e. cloning the gene based on its position on the genetic map). This strategy is facilitated in model species such as Arabidopsis, for which dense genetic maps with many visible and molecular genetic markers exist, and for which an almost complete physical map consisting of a collection of overlapping DNA fragments cloned in yeast artificial chromosomes or bacterial artificial chromosomes, has been constructed (Schmidt et al., 1995, 1997; Zachgo et al., 1996; Camilleri et al., 1998). The total genomic sequence will provide the ultimate physical map. With all of these tools, especially with the genomic sequence in hand, positional cloning strategies will be greatly accelerated, the limiting factors being the time and effort required for constituting the mapping population and the fine mapping of the mutant locus. New mapping strategies based on “mapping chips” will accelerate this aspect as well.

**Insertional Mutagenesis (Transposons/T-DNA)**

The use of insertional mutagenesis in principle provides a more rapid way to clone a mutated gene. DNA elements that are able to insert at random within chromosomes, such as transposons (Martienssen, 1998) or the T-DNA of *Agrobacterium tumefaciens* (Azpiroz-Leehan and Feldmann, 1997), can be used as mutagens to create loss-of-function mutations in plants. Because the sequence of the inserted element is known, the gene in which it is inserted can be easily recovered using various cloning or PCR-based strategies (Fig. 2).
The key parameters here, as in any mutagenesis strategy, are the level of saturation (i.e. the probability of having at least one insertion in any gene, which depends on the number of independent insertions in the population), the randomness of insertion of the element, and the number of insertions per line. For the Arabidopsis genome, we estimate that a collection of around 120,000 independent inserts is needed to provide a high level of saturation (95% chance of hitting any gene at least once).

Various populations of mutagenized plants, either with heterologous transposons (mainly maize transposons such as Ac/Ds, En/Spm, or Mu) or the T-DNA of A. tumefaciens, have been produced in several plant species (Arabidopsis, petunia, snapdragon, tomato, maize, and rice). The level of saturation reached depends strongly on the size and organization of the genome.

In addition, the insertional mutagen can be engineered with reporter cassettes (enhancer or gene traps) that will report on the expression of the chromosomal gene at the site of insertion (Martienssen, 1998). For example, reporter cassettes containing a minimal promoter (enhancer trap) close to the end of the insertion element can be cis-activated when inserted close to a transcriptional enhancer that will drive the expression of the reporter gene. Mutagenized populations can be screened for lines expressing the reporter gene in specific cell types or in specific environmental conditions. Genes with interesting expression patterns and their promoters can be isolated from such lines. Such lines can also be used to mark certain cell types for developmental studies (Sundaresan et al., 1995). Activation tagging, which is based on the use of an insertion element carrying a strong enhancer or promoter directing transcription into the region flanking the insertion, enables the isolation of gain-of-function mutations in which ectopic activation of a flanking gene promotes a mutant phenotype (Wilson et al., 1996). Such mutations are generally dominant or semidominant.

The main advantage of transposon-based approaches is their relative facility in generating large populations of insertions and their ability to use the propensity of many transposable elements to transpose to linked sites, which makes it possible to remobilize the element for insertion in the vicinity of the starting insertion site. On the other hand, it is more difficult to achieve saturation with T-DNA insertional mutagenesis, but it results in fewer insertions (1–2 loci per line); insertions are stable, easy to maintain, and, as far as we know, do not show strong insertional biases (Azpiroz-Leehan and Feldmann, 1997). In Arabidopsis, highly efficient in planta transformation techniques have been instrumental in generating large populations of T-DNA insertion lines while minimizing the effect of somaclonal variation linked to in vitro culture and regeneration (Bechtold et al., 1993; Azpiroz-Leehan and Feldmann, 1997).

**Reverse Genetics: From Genes to Functions**

A direct way to obtain information on the function of a gene identified by sequencing is to create a loss-of-function mutation and study the phenotype of the resulting mutant. In many model organisms (e.g. microbes and mice), homologous recombination can be used efficiently to target mutations into specific genes by replacing the wild-type gene with a mutated allele. In plants controlling homologous recombination has proven extremely difficult because of the prevalence of illegitimate recombination events (Puchta and Hohn, 1996). Although some success has been reported (Miao and Lam, 1995; Kempin et al., 1997), gene replacement by recombination is not considered feasible on a large scale, and other strategies have to be implemented to allow the functional analysis of large numbers of genes.

An alternative, highly efficient procedure for obtaining mutants in genes identified in sequencing programs takes advantage of the availability of large collections of plants mutagenized by an insertion element (T-DNA or transposon). This procedure, based on methods originally developed in Drosophila melanogaster (Ballinger and Benzer, 1989) and Caenorhabditis elegans (Zwaal et al., 1993), makes use of the specificity and sensitivity of the PCR reaction to screen for insertions within regions of interest in a large population of mutagenized plant lines (McKinney et al., 1995; Krysan et al., 1996; Azpiroz-Leehan and Feldmann, 1997; Martienssen, 1998). Using oligonucleotide primers from the insertional element and from the gene of interest, it is possible to detect an insertion event within the gene, even in complex DNA samples (Figs. 2 and 3). The sensitivity of the PCR is so high that it is possible to detect such an event in large pools (up to a few thousand) of mutagenized plants (Fig. 3).

The first published examples of targeted mutations in Arabidopsis (Giilliland et al., 1998; Hirsch et al., 1998) show that careful analysis of mutants is often required to detect the deleterious effects of a given mutation. In our laboratory more than 50 mutated lines were obtained from PCR screening of a T-DNA insertion population (Bechtold et al., 1993), and most of the lines did not show strong visible phenotypes when grown in standard conditions (D. Bouchez, unpublished data). To reveal phenotypic differences, it may be necessary to identify the correct environmental conditions for the expression of the phenotype. An example is the potassium-channel mutant AKT1, for which a reduction in growth is visible only at low potassium concentrations (Hirsch et al., 1998). Some mutations may have very subtle effects that are not detectable in standard experimental conditions, and multigenerational population studies have to be devised to reveal phenotypic differences between mutant and wild type (Asmussen et al., 1998). Finally, it becomes clear from large-scale sequencing programs that eukaryotic genomes are quite redundant, with many genes duplicated in families (Bevan et al., 1998). The functional analysis of such gene families may often require the construction of lines carrying multiple mutations in different family members.

A further extension of this strategy involves the use of an EST-like approach to systematically sequence the flanking regions of insertions. Using PCR-based techniques, it is possible to isolate and sequence a large number of insertion sites from pooled or individual lines. The insertion sites can be mapped by comparing the sequences with the genomic sequence. Databases of such flanking sequences will be established and can be searched for hits in any...
particular gene. Seeds of the lines carrying the sequenced inserts will become available from the stock centers. Several public and private laboratories are involved in such programs, and these tools should become available in the near future. It is possible, however, that the growing interest from private companies will interfere with the public release of these resources.

CONCLUSIONS AND PERSPECTIVES

Together with the completion of the first sequence of a plant genome, a magnificent tool kit is becoming available that will dramatically improve the quality of research in all disciplines of plant biology. This tool kit comprises large collections of knockout mutants in model organisms for systematic gene function searches and methods for the study of the regulation of entire sets of genes at the mRNA or protein level. A challenge in the next decade will be to build integrated databases combining information on such things as sequence, map position, mRNA and protein expression, mutant phenotypes, metabolism, and allelic variation. Also, intelligent software tools will be required for the efficient mining of this wealth of information and to construct intelligible models describing the complex molecular interactions that constitute the regulatory networks. EST sequencing and mapping or genomic sequencing of a few strategic crop species may suffice to facilitate the transfer of information from model species to the majority of crop plants, to be used for crop improvement in the future.

NOTE ADDED IN PROOF

A new server for gene expression data for Arabidopsis can be consulted at the following address: http://www.monsanto.com/Arabidopsis. An interesting application of DNA chips, in which...
gene expression was monitored during the cell cycle of yeast, was reported in Cho RJ, et al. (1998).

ACKNOWLEDGMENTS

Shauna Somerville is thanked for providing information on cDNA arrays, Dominique Devienne for communicating unpublished results, and Patrick Brown for providing figures for Example 1.

Received July 22, 1998; accepted July 29, 1998.
Copyright Clearance Center: 0032-0889/98/118/0725/08.

LITERATURE CITED


Copyright © 1998 American Society of Plant Biologists. All rights reserved.