Mining functional information from cereal genomes – the utility of expressed sequence tags

N. Sreenivasulu[†], P. B. Kavi Kishor[#], R. K. Varshney[†] and L. Altschmied[†,*]

[†]Institute of Plant Genetics and Crop Plant Research, Corrensstr. 3, 06466 Gatersleben, Germany
[#]Department of Genetics, Osmania University, Hyderabad 500 007, India

The vast number of expressed sequence tags (ESTs) generated from cereals with large genome size is an important complement method to the whole genome sequencing projects. As technology-driven revolution sweeps through, we are submerged in an avalanche of new information about genes and their function. To accomplish functional roles to the available ESTs, proper annotation of EST data is crucial, which could be achieved through the employment of tools of bioinformatics generated in the recent past. In this review, the critical steps involved in EST-based gene discovery and the employment of available web-based bioinformatic tools for annotation of ESTs is discussed. The current status of application of EST clones in the development of molecular markers in cereal species and utilizing ESTs as a resource for the construction of arrays is summarized as well. We also focus on large-scale gene expression data analysis methods and the challenges for computational biologists to extract functional information from such large-scale gene expression data.

SINGLE-pass sequencing of randomly chosen cDNA clones is currently the most efficient method for the discovery of many genes from cereals with large genomes. Management and analysis of the enormous amount of low-quality sequence data require great care and powerful computational methods for annotation. On the basis of annotated expressed sequence tags, novel molecular markers can be developed and methods for global expression analysis established. These functional genomic approaches hold great promise for the future and have just begun to unravel their power for the investigation of complex metabolic and regulatory networks, which determine the development of plants and their response to the environment.

Expressed sequence tags – an introductory glimpse

To meet the future challenges presented by an ever-increasing population on earth, genes with the potential to improve various steps in food production and processing need to be identified to use them for the genetic modification of crop species. Prime targets are the genes of rice, wheat, maize, barley and sorghum, which belong to the ten most important crop species worldwide. They are all members of the grass family and their genomes contain large syntenic segments of conserved gene order[1]. Rice has the smallest genome among these five species, with a size of only 430 Mbp. Consequently, the complete sequence of the rice genome as well as loci of interest could be identified via genomic sequencing, with an acceptable investment. The genomes of the other species are considerably larger: sorghum 800 Mbp, maize 2500 Mbp, barley 5500 Mbp and wheat 16,000 Mbp, which presently precludes this approach. As an alternative to a genomic sequencing programme, partial, single-pass sequencing of more or less randomly chosen cDNA clones from libraries at all stages of plant growth and life cycle allows fast and affordable gene identification at a large scale[2,3]. This so-called expressed sequence tag (EST) approach targets the sequencing efforts to the most important part of the genome, namely the transcribed genes. Large EST programmes for grasses and other crop species are currently under way in many research laboratories worldwide, which leads to a steadily increasing number of entries in the EST databases (Table 1). At present (2001–2002) the EST database (dbEST; http://www.ncbi.nlm.nih.gov/dbEST) contains 504,466 ESTs from monocotyledonous plants, of which 119,158 are reported from maize, 107,278 from sorghum, 101,709 from rice, 95,487 from barley and 73,395 from wheat.

EST-based gene discovery – its merits and inherent limitations

Gene discovery via ESTs is comprised of three steps which include (i) the construction of cDNA libraries and single-pass sequencing of (randomly) selected clones, (ii) EST quality check – the removal of vector and low-quality sequences, (iii) the alignment of ESTs to identify the number of represented genes, and (iv) the annotation of these genes or the partial sequences which are available thereof.

[*]For correspondence. (e-mail: lothar@ipk-gatersleben.de)

Table 1. Expressed sequence tags of major cereals in dbEST

Species	EST	cDNA library	Low quality	≤ 100 b/≥ 800 b	*E. coli*
Oryza sativa	101,709	27	8889	140/1464	289
Triticum aestivum	73,395	38	4068	82/1793	198
Zea mays	119,158	31	3850	186/1352	16
Hordeum vulgare	95,487	31	5043	637/24,916	178
Sorghum bicolor	84,712	10	132	349/18	132
S. propinquum	21,387	2	41	10/-	31
S. halepense	1179	1	–	–/-	10

For the major cereals the number of entries in dbEST (January 2002) and the number of cDNA libraries from which more than 500 ESTs were derived are listed. Critical quality parameters include the number of ESTs containing low-quality segments (≥ 3 ambiguities/25 bases), short (≤ 100 bases) and overly long ESTs (≥ 800 bases) as well as contaminations, e.g. *E. coli* sequences (> 100 bases with ≥ 95% identity).

cDNA library generation

The production of ESTs starts with the construction of cDNA libraries. Within a certain tissue of defined developmental and physiological status, only a specific fraction of all genes of an organism is expressed and the abundance of mRNAs for different genes varies widely. This makes it less likely to identify low expressed genes and leads to redundant sequencing of the ones that are highly expressed. In addition to the construction of several cDNA libraries to cover a wider spectrum of expressed genes, various strategies have been applied to circumvent or minimize redundant sequencing. cDNA libraries can be normalized either during their synthesis by subtractive hybridization or a related approach[4], or afterwards by techniques such as oligonucleotide fingerprinting[5]. The exclusion of already sequenced cDNAs in the database or even complete libraries representing a high degree of redundancy, provides another valid alternative to minimize the costs of uncovering new genes. Table 1 provides an overview of the total number of ESTs obtained from cereals and the number of relevant cDNA libraries employed in the respective sequencing programmes. Despite these efforts, it was shown for species with completely sequenced genomes that the number of genes represented by ESTs is significantly smaller than the number of predicted genes. For instance, more than 113,000 ESTs from *Arabidopsis* represent less than 16,200 of the 25,556 genes predicted in the genome.

Quality of ESTs

After isolation of cDNA clones, plasmid preparation and single-pass sequencing, several quality issues have to be addressed. Vector and low-quality sequences need to be removed from the raw sequence data, as well as bacterial sequences or other contaminations. No generally accepted standards exist for these procedures, so that the quality of submitted sequences does depend on the submitting laboratory.

Many ESTs can be identified which contain low-quality sequences. Table 1 lists the number of database entries which contain segments with more than three ambiguities in 25 bases. These ESTs might represent only part of the problem, because base-calling software such as Phred will not assign ambiguous bases, but rather use a quality score, which is rarely provided in sequence databases. Wrong bases as well as small insertions and deletions (indels) go undetected in single-pass sequences. Especially indels occur frequently at short homopolymer stretches at greater read length. In these regions, the base-calling software has to determine the number of bases from the width of a single merged peak, which leads to a significant reduction in its reliability. Hence, sequences should be trimmed at a certain read length. This has not been done for many database entries, as can be seen by the large number of ESTs with a length of more than 800 bases (Table 1).

The removal of bacterial contaminations is also not a routine procedure, because ESTs which represent various pieces of the *E. coli* genome can be identified (Table 1). Not easy to recognize and therefore more serious are contaminations from other eukaryotic organisms. They result from the use of non-sterile plant tissue or material deliberately infected with plant pathogens. Furthermore, handling errors or lane tracking problems in gel-based sequence analysis cause wrong assignments of clones and sequences. Such errors cannot be recognized in databases, but will become apparent when the cDNA clones have to be used, e.g. for the construction of cDNA arrays (see below).

EST clustering/gene content

The assembly of gene sequences or parts thereof from a collection of ESTs to determine the number of represented genes is a non-trivial task. The above-mentioned problems with sequence quality and possible sequence errors together with the high number of gene families with closely related members in plant genomes[6], present huge challenges. Special program packages such as the Phred/Phrap/Consed system (http://www.phrap.org/), UniGene[7], Genexpres Index[8], TIGR_ASSEMBLER[9], STACK_PACK[10,11], CAP3[12], PCP/CAP4 (www.paracel.com/ products), HarvESTer (http://mips.gsf.de/proj/gabi/news/bioinformatics.html) and others have been and continue to be developed for the assembly of large EST collections. The result of the assembly process can be divided in so-called singletons, sequences which do not assemble with any other sequence, and groups of assembled sequences which might be called clusters, contigs, tentative consensus, tentative genes, unique genes (unigenes), etc.

Usually the sum of singletons and assemblies is larger than the true number of represented genes, for several reasons. Sequences may not be assembled even though they belong to the same gene. Long mRNAs for example, may yield various truncated cDNA clones, resulting in sequences which do not overlap. Existing overlaps may not be assembled as a result of low sequence quality, which prevents recognition or acceptance of the overlap by the used algorithm. Furthermore, clones that do not correspond to genes of the species of interest (see above) will contribute to the number of singletons and assemblies. The grouping of ESTs which do not represent the same gene will also occur, but seems to be more rare. A typical example is a family of closely related genes for which coding sequences from the 5′-ends of cDNA clones may be assembled as a result of sequence conservation. 3′-end sequences of the same clones would distinguish the family members, because they consist mainly of less conserved 3′-UTRs. However, 3′-sequences are often not available in EST projects. In addition, chimeric clones may link ESTs encoding unrelated genes. In large collections this presents a problem, which is difficult to resolve. A complete genomic sequence would reveal that the genes in question are from different loci, but it is not available for cereals with large genomes.

Several institutions provide pre-calculated assemblies of ESTs, sometimes including completely sequenced cDNA clones and genomic sequences to improve the results. Prominent examples are the gene indexes at the Institute for Genomic Research, Rockville, USA (TIGR; http://www.tigr.org). Table 2 provides an overview of gene indices of the species which are in focus in this article. Even though certain quality issues of ESTs are addressed by TIGR, one should keep in mind that the number of unique sequences should not be interpreted as the number of genes identified in a certain species.

Employing bioinformatics tools for annotation of ESTs

In addition to the number of genes represented by ESTs, it is important to collect information about their (potential) function and to associate this information with the respective clones. This process, called annotation, will help identify promising targets for further research and to interpret results of downstream applications which employ these clones, respectively their sequences, e.g. global expression analysis. The annotation process has to face the same difficulties as the annotation of unknown genes in genomic sequences (except splice site prediction), but is further complicated by the partial information and the high, yet undefined error content of ESTs. To minimize these problems, consensus sequences of aligned ESTs should be used whenever available, because they contain more information of increased reliability with respect to individual ESTs. Figure 1 shows a basic scheme about how an annotation process might be structured.

The first question which needs to be addressed is whether the EST is identical with or similar to a known gene. It can be approached by comparing sequence with appropriate databases using Blast or FASTA programs. Comparisons at the nucleotide level will identify closely related database entries, whereas comparisons at the amino acid level, after translation of the EST in all (meaningful) reading frames, can be used to uncover less related genes. The public availability of databases and of Blast[13] and FASTA[14] programs as well as the low price

Table 2. Summary of gene indexes provided by TIGR

Species	Rice		Wheat		Maize		Barley		Sorghum (*S. bicolor*)	
	29 May 2001		12 March 2001		12 July 2001		13 November 2001		29 June 2001	
Version	TC	Single	TC	Single	TC	Single	TC	Single	TC	Single
EST	52,097	22,115	41,123	16,475	82,214	12,605	49,876	40,114	55,256	12,209
Et	3719	3205	437	308	1652	314	782	190	110	61
Unique sequence	8551	25,320	6814	16,738	12,205	12,919	8556	40,304	9065	12,270
Unique total	33,871		23,552		25,124		48,860		21,335	

Listed are the main properties of the gene indexes of rice, wheat, maize, barley and sorghum. The number of ESTs and ETs (completely sequenced cDNA clones or known genes) included in assemblies (TC, tentative consensus) or those that remain as singletons (single) are shown as well as the sequences resulting from the assembly process.

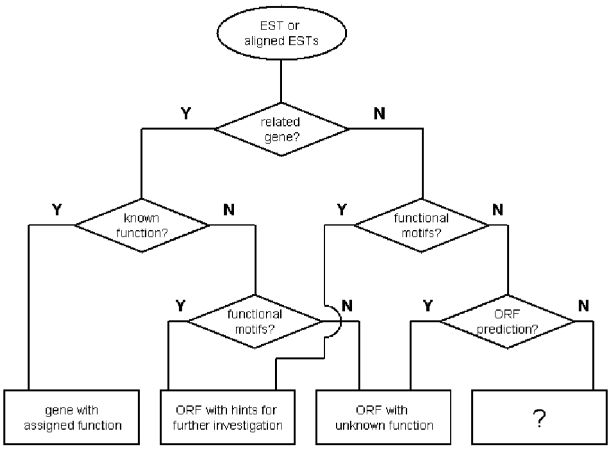


Figure 1. Annotation process for ESTs. A decision tree is depicted which will lead to four different categories of ESTs with respect to the knowledge that is available about an encoded gene. In contrast to the 'yes' and 'no' decisions depicted here, real answers are of probabilistic nature and will depend on the definition of threshold values, e.g. scores of BLAST results. Annotation is further complicated by the rapid increase of information stored in databases, which requires the constant revision of all decisions in this tree. The question mark, in most of the cases, stands for 5′ or 3′ untranslated sequences or parts of incompletely spliced mRNAs (introns).

of high computing power make it feasible to run many thousand comparisons at low costs within a moderate time. Yet, the incomplete sequence information with respect to the cDNA clone itself, and with respect to the gene content of the genome, usually precludes a precise answer. The main reasons are that minor sequence differences may distinguish members of gene families, but could also result from sequencing errors. Furthermore, the known part of a sequence might show similarity, but differences might be hidden in the unknown part. As a consequence, the answer is associated with a certain probability that is difficult, if not impossible, to quantify. This situation is even more complicated by the fact that the possible answers evolve quickly, because the content of databases used for comparisons increases rapidly.

When the question of function is approached, the difficulties in finding an answer do increase further. Usually the description and references contained in a database entry related to an EST provide a quick access to the relevant information, but several problems are associated with this approach. Mainly, as a result of genomic sequencing, many hypothetical genes will be encountered for which no functions are known. The description of a database entry might be outdated or, even worse, it may propagate annotation errors. To obtain a higher level of confidence, specialized databases which are curated and also provide more detailed information can be used for sequence comparisons, e.g. SwissProt[15], TRANSFAC for transcription factors[16], BRENDA for enzymes[17].

In case no related genes can be identified for an EST, or if the related gene does not provide information with respect to function, attempts can be made to identify functional motifs, which may guide further investigations. The identification of protein patterns from the PROSITE database[18], Pfam[19] and other databases, the prediction of targeting signals and transmembrane helices as well as the prediction of open reading frames provide several opportunities.

Generally, one can note that the computational annotation of ESTs is still in its infancy (Table 3). Software tools have to be improved significantly to meet the challenges provided by a rapidly increasing number of ESTs and to cope with their specific problems. Especially for cereals with large genomes, EST development will be important because complete genomic sequences will not be available in the near future.

Applications of EST clones and sequences

EST projects provide a wealth of sequence information and a large number of corresponding cDNA clones which can be utilized in various ways. Currently, the most interesting uses are large-scale transcript profiling and the development of molecular markers. In the foreseeable future, expression cloning and the production of protein arrays may be added to this list. Protein arrays produced from expressed cDNA clones would facilitate functional studies of proteins with respect to enzymatic activity and ligand binding, including the search for interacting partners and the isolation of antibodies. Cloning systems, which allow the high-throughput transfer of individual inserts or even whole libraries between different vectors for these purposes, became available during the last few years (GATEWAY cloning – http://www.tcd.ie/Genetics/staff/Gateway_Manual.pdf; CREATOR^{TM} Gene cloning and expression system – http://www.clontech.com/products/families/creator/index. shtml).

Development of molecular markers by EST approach

ESTs allow the efficient development of highly valuable molecular markers, because genes often represent single- or low-copy sequences. These are of great value in the cereal genomes, which consist of up to 80% of highly repetitive DNA. Hence classical, hybridization-based RFLP markers have been developed from ESTs and used extensively for the construction of high-density genetic linkage maps in rice[20], and maize[21], as well as for the construction of a physical map in rice[22]. Currently, a United States Department of Agriculture (USDA)-funded North American consortium and groups at the Institute of Plant Genetics and Crop Plant Research (IPK) in Gatersleben, Germany, funded by the German plant genome project (GABI) are localizing ESTs on BAC clones of barley and mapping them genetically. In addition, a large

Table 3. Web sites useful for EST annotation

Program	Purpose	URL
Blast	Sequence comparison	http://www.ncbi.nlm.nih.gov/BLAST/
Fasta	Sequence comparison	http://www.ebi.ac.uk/fasta33
SwissProt	Protein sequence comparison	http://www.expasy.org/sprot/
Pfam	Protein sequence comparison	http://www.sanger.ac.uk/Software/Pfam/
PROSITE	Protein pattern finding	http://www.expasy.ch/prosite/
TRANSFAC	Transcription factor detection	http://transfac.gbf.de/TRANSFAC/
BRENDA	Enzyme functional data collection	http://www.brenda.uni-koeln.de/
TMPRED	Transmembrane prediction	http://www.ch.embnet.org/software/TMPRED_form.html
TMHMM	Transmembrane helix prediction	http://www.cbs.dtu.dk/krogh/TMHMM/
FRAMED	GC content	http://www.toulouse.inra.fr/FrameD/cgi-bin/FD
GENEMARK	Prediction of ORF	http://genemark.biology.gatech.edu/GeneMark/
GENESCAN	Prediction of ORF	http://202.41.10.146/
BESTORF	Prediction of ORF	http://genomic.sanger.ac.uk/gf/gf.html

Tools which are useful for the annotation of ESTs are listed. Some of the publicly available tools are listed, which might be used to annotate translated ESTs with respect to functional motifs, but none of them has been designed or adjusted to handle ESTs specifically and to take care of associated problems.

project for physical mapping of ESTs in wheat using deletion stocks is in progress under a National Science Foundation (NSF) co-ordinated EST project at Kansas State University, USA.

Often EST-based RFLP markers allow comparative mapping across different grass species, because sequence conservation is high in the coding regions. The resulting anchor points in genetic and physical maps are especially important for grasses, because their genomes consist of large syntenic blocks with a highly conserved order of genes[1]. Hence, marker development and map-based cloning in one species will profit directly from data, which are available in any other species. Especially the upcoming genomic sequence of rice will greatly facilitate this comparative approach, because simple sequence comparisons can then be used to infer a map location of ESTs in one of the other cereal genomes. This approach has been demonstrated for barley[23].

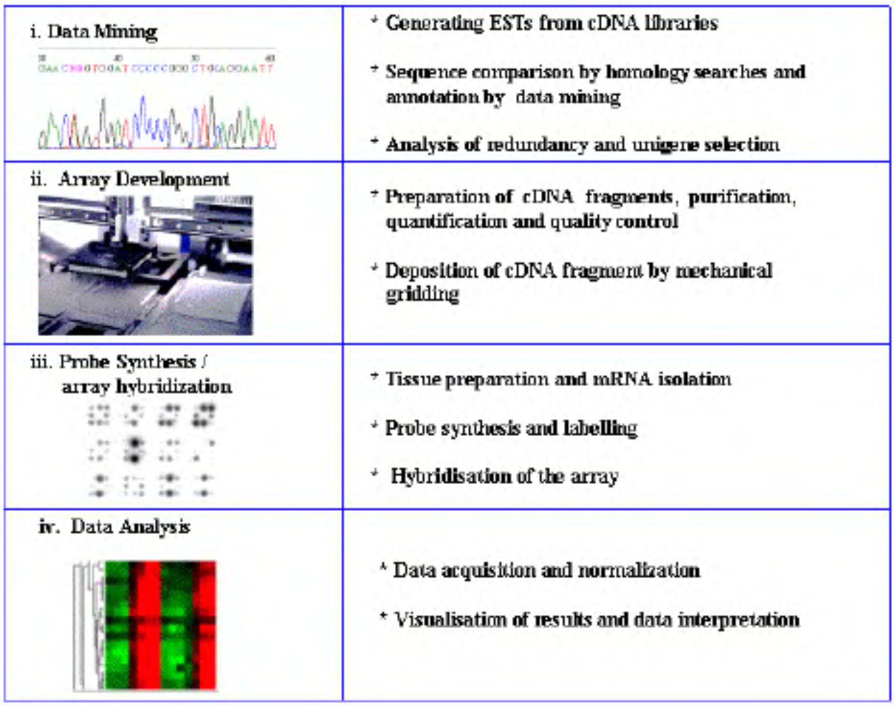
ESTs also allow a computational approach to the development of SSR (simple sequence repeat) and SNP (single nucleotide polymorphism) markers[24,25], for which previous development strategies have been expensive. Pattern-finding programs can be employed to identify SSRs in ESTs. The available sequence information allows the design of primer pairs, which can be used to screen cultivars of interest for length polymorphisms. For SNP development, two strategies have been employed. One strategy uses ESTs from the 3′-end of cDNA clones, which consists mainly of 3′-UTRs, to maximize the chance of finding sequence variations. Primer pairs can be derived from the EST sequence, and the amplification of corresponding regions from several genotypes followed by sequence comparison may reveal SNPs. Alternatively, one can use clusters of ESTs which contain sequences from different cultivars and identify potential SNPs computationally. An experimental verification of these potential SNPs is indispensable, because the sequence quality of ESTs cannot be guaranteed. Currently, the generation and mapping of such EST-derived SSRs and SNPs is in progress for several important cereal species such as wheat[24] and barley[26].

High-throughput transcript profiling

ESTs also provide the main resource for the construction of cDNA arrays in cereals. The DNA-chip technology has been developed[27,28] and is well established. The construction and use of such EST arrays for high-throughput transcript profiling can be divided in four general steps, which are depicted in Figure 2. These steps comprise (i) identification of a non-redundant set of cDNA clones, (ii) synthesis and deposition of hybridization targets on an appropriate surface, (iii) preparation of mRNA from the tissue of interest, labelling of the hybridization probe and the hybridization to the array, and finally (iv) data acquisition and evaluation.

Data mining: The development of a non-redundant unigene set from ESTs has been covered above. It serves the purpose to minimize the number of samples on a cDNA array mainly for technical reasons. However, a low degree of redundancy will provide data for quality control[29].

Array development: Several different approaches, which are summarized in Table 4, could be taken for the construction of a cDNA array. The least expensive approach is the PCR amplification of cDNA fragments using vector primers and their spotting on nylon membranes or chemically modified glass or plastic surfaces (for review specifically on plant cDNA arrays, see ref. 30). To serve the purpose, cDNA clones from the EST project have to be available. All handling errors with respect to the



i. Data Mining

- Generating ESTs from cDNA libraries
- Sequence comparison by homology searches and annotation by data mining
- Analysis of redundancy and unigene selection

ii. Array Development

- Preparation of cDNA fragments, purification, quantification and quality control
- Deposition of cDNA fragment by mechanical gridding

iii. Probe Synthesis / array hybridization

- Tissue preparation and mRNA isolation
- Probe synthesis and labelling
- Hybridisation of the array

iv. Data Analysis

- Data acquisition and normalization
- Visualisation of results and data interpretation

Figure 2. A diagrammatic representation of EST-array technique. Four major steps involved in EST-array production technology are: (i) Database mining; (ii) Array development; (iii) Probe synthesis/array hybridization and (iv) Data analysis. The sub-steps followed in every major step have been provided with star mark on the right.

Table 4. Design principle of arrays used for expression analysis

Target on array	Array surface	Target application	Features (cm^{-2})	Label
cDNA fragment	Nylon membrane	Spotting	100	^{33}P
cDNA fragment		Spotting	4000	Fluorescent dye
Oligonucleotide (50–80 mer)	Modified glass or plastic	Spotting	4000	Fluorescent dye
Oligonucleotide (25 mer)		On-chip synthesis	300,000	Fluorescent dye

Array designs used for expression analysis differ widely with respect to the hybridization targets, solid support, method of application of hybridization targets and their density, as well as the label which is used to detect hybridization intensities.

clones will be reflected on the array. To provide gene-specific hybridization targets for different members of certain gene families, 3′-end sequences could be used together with one gene-specific primer to amplify the 3′-UTR (ref. 31). The requirement of gene-specific primers will increase the set-up costs dramatically, if an array with a decent number of genes is constructed.

Even higher set-up costs have to be covered, if long oligonucleotides (50–80 mers) are synthesized and spotted instead of cDNA fragments. The advantages of this approach are that oligonucleotides can be designed to distinguish members of gene families, that only a cDNA sequence and not the clones needs to be available, and that handling errors with respect to the clones will not affect the array.

The third approach is the on-chip synthesis of short oligonucleotides (25 mers), which is offered by Affymetrix (http://www.affymetrix.com/). Again set-up costs are high; furthermore, the array design is rather static with respect to the gene content, because a new design would require a completely new set-up. Therefore, construction of these types of arrays is thought to be useful, if a genomic sequence is available to identify most of the genes or parts thereof, with a high degree of reliability. Recently, the construction of an array of that type, containing approximately 20,000 genes of rice, has been reported (San Diego January 2002; http://www.intl-pag.org/10/abstracts/).

Except for Affymetrix arrays, the oligonucleotides or cDNA fragments need to be transferred and permanently attached to the array surface. Usually this is accomplished by solid or slit pins which pick up the samples from microtiter plate wells and transfer them to the target locations on the array. Spot distances in the order of 100 to 400 μm, up to several thousand spots per array, and transferred volumes in the picolitre-range require high precision and high speed moving devices which perform this task in an environment with precisely controlled temperature and humidity. For the permanent bonding of oligonucleotides or cDNA fragments to the glass surface of microarrays, several different chemical modifications are currently in use which have a common property that they form covalent bonds with primary amines. Oligonucleotides need to be modified accordingly, whereas longer DNA fragments will bind to these surfaces without further modification.

Probe synthesis/array hybridization: This step of the cDNA array analysis involves the isolation of mRNA, probe synthesis and labelling as well as the hybridization to the array. To synthesize a labelled hybridization probe various protocols are available[32], which are too numerous to be covered here. Generally, ^{33}P-labelled nucleotides are employed when membrane-based arrays (macroarrays) are hybridized, because incorporation rates are high and sensitive phosphorimagers can be used for signal detection. Radioactive labels cannot be used for any kind of microarray, because the spatial resolution of the phosphoimager is not sufficient to separate signals of neighbouring spots. Usually, fluorescent dyes are incorporated either directly using dye modified nucleotides (CyDye™ fluorescent dyes: Amersham/Pharmacia – http://www.amershambiosciences.com/product/publication/lsn/lsn4/lsn4-17.html) or indirectly via aminoallyl-modified dUTP (Clontech – http://www.clontech.com/archive/JAN02UPD/pdf/PowerScript.pdf). Alternative strategies employ, for example, the incorporation of biotinylated nucleotides and labelling with phycoerythrin-conjugated streptavidin after the hybridization has been performed (Affymetrix). Hybridizations are carried out under the most stringent conditions possible to prevent cross-hybridization.

Data analysis: After hybridization, signals are detected using specialized scanners for microarrays and phosphoimagers for macroarrays. The resulting images are processed with a software for automatic spot detection to derive a list of signal intensities for all features on the array. The raw data have to be processed to gain biological knowledge. Important steps include (a) critical assessment of data reliability and normalization to allow the comparison of many experiments, and (b) categorizing of gene expression profiles and their biological interpretation. For these purposes several software packages are available commercially and in the public domain. An overview, not necessarily complete, is given in Table 5.

(a) Depending on the type of experiment, various procedures can be employed to normalize raw data for comparison with a series of other experiments. These procedures range from mathematical methods, which assume that the intensity distribution of signals does not change between experiments, to the use of reference signals, which are derived from housekeeping genes or foreign mRNAs included in probe synthesis. The choice of a method will often influence the experimental design and has to be made before an array is constructed. Our experience with macroarray experiments and Northern blot controls for many differentially expressed genes led to the conclusion that mathematical methods are sufficiently accurate[33,34]. Equally important is a careful evaluation of signal and array quality. Often, the initial data set will be reduced to a much smaller one of differentially expressed genes. Within this selected data set, experimental artefacts, which lead to large differences in signal intensity, will specifically accumulate and cause misleading interpretations. In addition, the biological variability will significantly influence the data and it is good practice to repeat each experiment with hybridization probes from independently harvested tissue samples. It seems difficult or even impossible to control all environmental variables to such an extent that no significant variation in gene expression can be observed in such repeats.

(b) As a consequence of the large number of data points obtained from just a few moderately-sized experiments, evaluation of the data has to be supported by computational methods. To categorize expression profiles, several methods from multivariate statistics can be employed, such as hierachical clustering[35], *K*-mean clustering[36], principal component analysis, self-organizing maps[37] and others. If they are used on a carefully controlled, reliable data set, they will yield similar, but not identical results.

Biological interpretation of expression data

Finally, expression data are expected to yield insights into regulatory processes during plant development and stimulus response. To reach that goal, it is necessary to compare the pre-processed array data with known models of metabolic and regulatory networks as depicted in KEGG[38] (http://www.genome.ad.jp/kegg/metabolism.html), the Boehringer biochemical pathway database[39] (http://www.expasy.ch/cgi-bin/search-biochem-index) or the general literature, and to confirm or reject specific

Table 5. Analytical tools with application to gene expression

Organization	Primary function	URL
Academic software		
Array Viewer	Multi-experiment viewer	http://www.tigr.org/softlab/
Image/J	Image processing	http://rsb.info.nih.gov/ij/
Spot finder	Spot detection	http://www.tigr.org/softlab/
Scan Alyze	Spot detection	http://rana.lbl.gov/EisenSoftware.htm
Cluster	Data filtering/clustering	http://rana.lbl.gov/EisenSoftware.htm
Tree View	Cluster visualization	http://rana.lbl.gov/EisenSoftware.htm
Xcluster	Clustering, visualization	http://genome-www.stanford.edu/~sherlock/cluster.html
J-Express	Clustering, visualization	http://www.ii.uib.no/~bjarted/jexpress/
Genesis	Clustering, visualization	http://genome.tugraz.at
Amanda	Clustering, visualization	http://xialab.hku.hk/software
Data explorer	Data flow visual program	http://www.opendx.org/
The R language	Comprehensive statistical analysis, clustering, etc.	http://cran.us.r-project.org/
Cyber T	'*t*'-test variants for gene expression data sets	http://genomics.biochem.uci.edu/genex/cybert/
Commercial software		
Array-Pro	Spot detection	http://www.mediacy.com/arraypro.htm
Array Vision	Image visualization, spot detection	http://imaging.brocku.ca/products/Arrayvision.htm
Array Explorer	Clustering, visualization	http://www.spotfire.net/
Expressionist	Clustering, visualization	http://www.genedata.com/products/expressionist/
Gene Maths	Clustering, visualization	http://www.applied-maths.com/ge/ge.htm
Gene Sight	Clustering, visualization	http://www.biodiscovery.com/products/genesight/genesight.html
Gene Spring	Clustering, visualization and normalization	http://www.sigenetics.com/cgi/SiG.cgi/index.smf
JMA Viewer	calls KEGG, BLAST,	http://sequence.aecom.yu.edu:8000/jmaviewer/
Partek	Clustering, visualization, 3D gene expression data	http://www.partek.com/

Worldwide web addresses of software for array data analysis both from public domain as well as from private sectors.

hypotheses. Many successful examples have been provided already, for example, the analysis of seed development[40,41] or phytochrome A signalling[42] in *Arabidopsis*, and the analysis of salt stress in rice[43].

Most of this interpretation process is a manual task, which requires the simultaneous integration of many different information resources. Software tools to support this complicated process are still in their infancy. Implementation of powerful interactive simulation environments for metabolic and regulatory networks, such as Metabolika[44], with integrated access to the information about related genes, proteins and metabolites as well as the actual expression data will be a next important step. Until such tools are available, the development of new hypotheses from the data of expression analysis will continue to depend on human ingenuity.

Conclusions

In cereals, the EST data set provides the primary access to genes; these are the basis for molecular marker identification and gene expression analysis. Only recently, the first results of expression studies in cereals have been made public, which use unique sets of genes derived from annotated ESTs. Tissue-specific expression patterns have been investigated for seed development[33] and germination[34] in barley, and for the dissection of the response to salt-stress in rice[43]. Currently many laboratories are actively engaged in setting up various techniques and the bioinformatics support, so that many more studies are expected to appear within a short time. One of the main challenges we foresee in the near future will be to couple expression data with metabolic and regulatory network models for better interpretation, and access to the large amount of information that will be available soon in many cereals.

1. Bennetzen, J. L. and Freeling, M., *Genome Res.*, 1997, **7**, 301–306.
2. Adams, M. D. *et al.*, *Nature*, 1992, **355**, 632–634.
3. Rounsley, S. D., Glodek, A., Sutton, G., Adams, M. D., Somerville, C. R., Venter, J. C. and Kerlavage, A. R., *Plant Physiol.*, 1996, **112**, 1177–1183.
4. Kohchi, T., Fujisige, K. and Ohyama, K., *Plant J.*, 1995, **8**, 771–776.
5. Guerasimova, A. *et al.*, *Biotechniques*, 2001, **31**, 490–495.
6. Schoof, H. *et al.*, *Nucleic Acids Res.*, 2002, **30**, 91–93.

7. Boguski, M. S. and Schuler, G. D., *Nature Genet.*, 1995, **10**, 369–371.
8. Houlgatte, R., Mariage-Samson, R., Duprat, S., Tessier, A., Bentolilal, S., Lamy, B. and Aufray, C., *Genome Res.*, 1995, **5**, 272–304.
9. Sutton, G., White, O., Adams, D. and Kerlvage, A., *Genome Sci. Technol.*, 1995, **1**, 9–18.
10. Miller, R. T., Christoffels, A. G., Gopalakrishnan, C., Burke, J., Ptitsyn, A. A., Broveak, T. R. and Hide, W. A., *Genome Res.*, 1999, **9**, 1143–1155.
11. Christoffels, A., Miller, R. and Hide, W., *Am. J. Hum. Genet.*, 1999, **65**, 477.
12. Huang, X. Q. and Madan, A., *Genome Res.*, 1999, **9**, 868–877.
13. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J., *Nucleic Acids Res.*, 1997, **25**, 3389–3402.
14. Stoesser, G. *et al.*, *ibid*, 2002, **30**, 21–26.
15. Bairoch, A. and Apweiler, R., *ibid*, 2000, **28**, 45–48.
16. Miyamoto, Y., *Trends Glycosci. Glycotechnol.*, 2000, **12**, 351–360.
17. Schomburg, I., Chang, A. and Schomburg, D., *Nucleic Acids Res.*, 2002, **30**, 47–49.
18. Falquet, L., Pagni, M., Bucher, P., Hulo, N., Sigrist, C. J. A., Hofmann, K. and Bairoch, A., *ibid*, 2002, **30**, 235–238.
19. Bateman, A. *et al.*, *ibid*, 2002, **30**, 276–280.
20. Harushima, Y. *et al.*, *Genetics*, 1998, **148**, 479–494.
21. Davis, G. L. *et al.*, *ibid*, 1999, **152**, 1137–1172.
22. Kurata, N., Umehara, Y., Tanoue, H. and Sasaki, T., *Plant Mol. Biol.*, 1997, **35**, 101–113.
23. Smilde, W. D., Halukova, J., Sasaki, T. and Graner, A., *Genome*, 2001, **44**, 361–367.
24. Eujayl, I., Sorrells, M., Baum, M., Wolters, P. and Powell, W., *Euphytica*, 2001, **119**, 39–43.
25. Cho, Y. G. *et al.*, *Theor. Appl. Genet.*, 2000, **100**, 713–722.
26. Kota, R., Varshney, R. K., Thiel, T., Dehmer, K. J. and Graner, A., *Hereditas*, 2001, **135**, 145–151.
27. Schena, M., Shalon, D., Davis, R. W. and Brown, P. O., *Science*, 1995, **270**, 467–470.
28. Nguyen, C., Rocha, D., Granjeaud, S., Baldit, M., Bernard, K., Naquet, P. and Jordan, B. R., *Genomics*, 1995, **29**, 207–216.
29. Herwig, R., Aanstad, P., Clark, M. and Lehrach, H., *Nucleic Acids Res.*, 2001, **29**, 1–9.
30. Richmond, T. and Somerville, S., *Curr. Opin. Plant Biol.*, 2000, **3**, 108–116.
31. Yazaki, J. *et al.*, *DNA Res.*, 2000, **7**, 367–370.
32. Gupta, P. K., Roy, J. K. and Prasad, M., *Curr. Sci.*, 1999, **77**, 875–884.
33. Sreenivasulu, N., Altschmied, L., Panitz, R., Hähnel, U., Michalek, W., Weschke, W. and Wobus, U., *Mol. Genet. Genomics*, 2002, **266**, 758–767.
34. Potokina, E., Sreenivasulu, N., Altschmied, L., Michalek, W. and Graner, A., *Funct. Integr. Genomics*, 2002, **2**, 28–39.
35. Eisen, M. B., Spellman, P. T., Brown, P. O. and Botstein, D., *Proc. Natl. Acad. Sci. USA*, 1998, **95**, 14863–14868.
36. Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J. and Church, G. M., *Nature Genet.*, 1999, **22**, 281–285.
37. Tamayo, P. *et al.*, *Proc. Natl. Acad. Sci. USA*, 1999, **96**, 2907–2912.
38. Goto, S., Bono, H., Ogata, H., Fujibuchi, T., Nishioka, T., Sato, K. and Kanehisa, M., Proceedings of the Pacific Symposium on Biocomputing, World Scientific, Singapore, 1997, pp. 175–186.
39. Michal, G. (ed.), *Biochemical Pathways*, Boehringer Mannheim, Germany, 1993.
40. White, J. A. *et al.*, *Plant Physiol.*, 2000, **124**, 1582–1594.
41. Ohlrogge, J. and Benning, C., *Curr. Opin. Plant Biol.*, 2000, **3**, 224–228.
42. Tepperman, J. M., Zhu, T., Chang, H. S., Wang, X. and Quail, P. H., *Proc. Natl. Acad. Sci. USA*, 2000, **98**, 9437–9442.
43. Kawasaki, S. *et al.*, *Plant Cell*, 2001, **13**, 889–905.
44. Hofestädt, R. and Scholz, U., *Biosystems*, 1998, **47**, 91–102.

ACKNOWLEDGEMENT. We thank Prof. Ulrich Wobus and Dr Winfriede Weschke for helpful comments on the manuscript.

Received 8 February 2002; accepted 19 August 2002