Plant Physiol. Biochem. 39 (2001) 181–193
© 2001 Éditions scientifiques et médicales Elsevier SAS. All rights reserved
S0981942801012426/REV

Review

Genome annotation

Sébastien Aubourg, Pierre Rouzé*

Laboratoire associé de l'Institut national de la recherche agronomique (France), Department of Plant Genetics, Ghent University, Flanders Interuniversity Institute of Biotechnology (VIB), K.L. Ledeganckstraat 35, B-9000 Gent, Belgium

Received 18 September 2000; accepted 30 January 2001

Abstract – Today, the public international sequence databases contain more than nine billion nucleotides and the flow of new sequences is increasing dramatically. For scientists, the challenge is to exploit this huge amount of sequences. To extract biological knowledge from anonymous genomic sequences is the main objective of genome annotation. To meet the expectations of scientists, allowing them to use genomic knowledge for further experimentation as quickly as possible, the extensive use of computer tools is needed to minimize the slow and costly human interventions. This is the reason why annotation is often synonymous with prediction. The annotation work is divided into two steps: structural annotation, which consists mainly of localizing gene elements; and functional annotation, which aims at assigning a biochemical function to the deduced gene products. The different tools and strategies used to convert sequences to useful data will be discussed in detail with their advantages and bottlenecks. By focusing on plant genomes and especially on the *Arabidopsis thaliana* genome, the general results and their different display will be presented. The international annotation effort allows us to have an interesting overview of the *Arabidopsis* genome organization: general gene features and functions, classification into multigene families, importance of duplication events and chromosome structure. Furthermore, the limits and errors of these annotations are highlighted in order to use the sequence databases at their best and to consider some novel approaches to deepen the understanding of the regulation and the biological function of the genes. © 2001 Éditions scientifiques et médicales Elsevier SAS

annotation / *Arabidopsis* / bioinformatics / database / gene prediction / genomics

AGI, *Arabidopsis* Genome Initiative / BAC, bacterial artificial chromosome / COG, cluster of orthologous genes / EST, expressed sequence tag / IRGSP, International Rice Genome Sequencing Project / MIPS, Munich Information Centre for Protein Sequences / TIGR, The Institute for Genomic Research / YAC, yeast artificial chromosome

1. INTRODUCTION

1.1. Some genomes and sequences

In the last few years, terms such as genomics, high-throughput, databases and bioinformatics have invaded research in biology. This revolution is linked to important progresses in methodology and technology that allow the massive production of sequences. Indeed, with capillary sequencers, robots and the whole genome shotgun method, genomic sequencing has now become routine. Academic and private institutions understood that the repercussions of this genomic revolution will be, or already are, huge in term of knowledge and economy. Nowadays, twenty-seven bacterial, six archeal and three eukaryotic genomes (yeast, worm and fly) have been completely sequenced. A first draft of the human genome has been produced far in advance than initially scheduled and more than 150 sequencing projects are running [47]. For a number of organisms with agronomic, medical or fundamental interests, the complete sequence of the chromosomes will be known. Therefore, the change in thinking of scientists is prominent. Increasingly, the starting point of research projects is based on genomic sequence(s) found in a public database [68]. Systematic sequencing projects allow a direct access to all genes, without functional or practical bias towards the more expressed ones. Genes can now be studied in a large context and considered like elements of an ensemble of homologous genes or implicated in the same physiological function. But building the link between raw DNA sequence and gene function is far from being an easy task [15].

**Correspondence and reprints:* fax +32 9 264 53 49.
E-mail address: pierre.rouze@gengenp.rug.ac.be (P. Rouzé).

1.2. Annotation: definition and justification

Classically, annotation of sequences is done by authors by using information linked to each experiment. The problem we face is not so much the tremendous increase in sequence database size than paradigm shift of genomics. Indeed, in contrast to sequences obtained through individual gene cloning experiments, high-throughput genome sequences are blunt (without any associated biological information).

The increasing number of Mb available in the international public databases (GenBank, EMBL and DDBJ) constitutes a complex problem for biologists. Without transcript sequences, a genomic sequence cannot be directly used and 'Where are the genes?' is the first of a long list of questions. Furthermore, the public data release policies force the sequencing centres to deposit as quickly as possible the raw sequences in the databases. To meet the scientific community's expectations together with the massive sequence production, the first step is to extract rapidly a maximum of biological information from the sequences to establish a basis for further functional studies. This information has then to be associated with the sequences in order to label the genes, allowing biologists to find sequences of interest. This is the goal of genome annotation. In return, we should also be able to update annotations and to apply to genes the results of experimental analyses performed later on. This notion of integration is essential from a genomics point of view. Finally, based on nucleic and protein sequences, annotation is a process whereby the biological information is extracted, collected and displayed in a clear form amenable to query.

Bench work cannot follow the release pace of thousands of new potential genes and experimentally documented genes represent a minor fraction of most genomes. In order to compensate for this transitional lack of secure data, prediction tools are extensively used to analyse the sequences and extract putative information. By merging statistics, computer sciences and biological knowledge, bioinformaticians have developed many prediction tools. The analysis of the anonymous sequences combining different prediction programs and the results obtained by bioanalysts are the starting points of the annotation process. For this reason, the annotation work is mostly a predictive work and the result has to be considered as such. Whole genome annotation should not be taken as definitive and proven, but rather as indications to help biologists in the 'sequence jungle' and to drive future experimental approaches [60]. This point is often forgotten when annotations are used.

When in front of a large genomic sequence, the first problem is to localize all the genes on both strands and, more precisely, the different structural elements of these genes. This step is called structural annotation, to clearly distinguish it from the following one, the functional annotation, which tries to find signs of function from the deduced protein sequences. As described below, deep and detailed annotation implies numerous complementary analyses and checking. Unfortunately, because of the cost in time and money of this human expertise, genome annotation is generally restricted to the prediction of coding exons, to deduce the protein sequence of potential genes and to label it with the function of the closest homologue. We will compare and discuss the fast high-throughput annotation used in the systematic sequencing programs and the possibilities of a deeper, but slower, annotation with two objectives: to highlight the dangerous traps when automatic annotations are blindly used and to present a few novel approaches and applications in genome annotation.

1.3. In the plant kingdom

Even if, for obvious reasons, the human and mouse genomes are the main targets of sequencing centres and experts in bioinformatics, the other branches of the tree of life are not left aside. With a genome of only 130 Mb, detailed genetic and physical maps, low-repeat sequences and numerous practical advantages for genetic experiments in greenhouses, the mustard *Arabidopsis thaliana* is the perfect model for plant genomics. Initiated in 1996, the systematic sequencing of the five chromosomes is realized by an international consortium (*table I*), named AGI for *Arabidopsis* Genome Initiative [9]. By the time of publication of this review, the euchromatin part of the *Arabidopsis* genome will be fully sequenced, including chromosomes 2 and 4 that were published earlier [49, 56]. The chloroplastic and mitochondrial genomes of *Arabidopsis* have also been entirely sequenced [54, 71]. The annotations of the *Arabidopsis* genes will be used as the main source of examples. The knowledge extracted from the *Arabidopsis* sequences could be generalized to members of the Brassicaceae family but also to numerous dicot plants. For monocots, the *Oryza sativa* genome will be the privileged source of data, because it possesses the smallest genome among crops [37]. For this reason, the International Rice Genome Sequencing Project (IRGSP) is working on the sequencing and annotation of BACs from rice (*table I*). Recent work by Monsanto, which has produced a working draft version of nearly the whole genome as well as an

Table I. Description of the international sequencing projects for *Arabidopsis* and rice genomes.

Features	Sequencing and annotation centres	Web sites	Chromosomes
Arabidopsis thaliana (Arabidopsis Genome Initiative (AGI))			
Dicot	ESSA & Génoscope-MIPS (Europe)	www.mips.biochem.mpg.de/proj/thal/	3+4+5
Brassicaceae	KAOS (Japan)	www.kazusa.or.jp/arabi	3+5
Columbia 0	TIGR (USA)	www.tigr.org/tdb/ath1/htmls/index.html	1+2+3
	CSH-WU-ABI (USA)	nucleus.cshl.org/protarab/ genome.wustl.edu/gsc/arab/arabidopsis.html	4+5
5 chromosomes	SPP (USA)	Sequence-www.stanford.edu/ara/SPP.html	1
130 Mbp		genome.salk.edu/	
111 745 ESTs		pgec-genome.ars.usda.gov/	
Oryza sativa (International Rice Genome Sequencing Project (IRGSP))			
Monocot	RGP (Japan)	rgp.dna.affrc.go.jp/Seqcollab.html	1+6+7+8
Poaceae	CUGI-CSH (USA)	nucleus.cshl.org/riceweb/	3+10
ssp. *japonicus*		stein.cshl.org/perl/ace/search/ricegenes	
Nipponbare GA3	TIGR (USA)	www.tigr.org/tdb/rice/	10
or spp. *indica*(*)	KRGRP (South Korea)	bioserver.myongji.ac.kr/ricemac.html	1
	NCGR (China)	www.ncgr.ac.cn/Ls/index.html	4(*)
12 chromosomes	ASPGC (Taiwan)	genome.sinica.edu.tw/	5
420 Mbp	GénoScope (France)	www.genoscope.cns.fr/	12
62 011 ESTs	NSTDA (Thailand)	www.cs.ait.ac.th/nstda/biotec/biotec.html	9
	PGIR (USA)	pgir.rutgers.edu/News.html	3+10
	JIC (UK)	www.jic.bbsrc.ac.uk/	2
	UW (USA)	www.gcow.wisc.edu/Rice/index.htm	11
	AGT (India)		8

excellent physical map of the twelve rice chromosomes, should help the IRGSP reach quickly its objectives [70].

Genome sequencing is not the only systematic approach in plant structural genomics. Numerous laboratories are sequencing extremities of plant cDNAs to have a direct access to the transcribed genome. These EST projects are running not only for *Arabidopsis* and rice, but also for various plants such as soybean, tomato, maize, sorghum, loblolly pine, barley, wheat, cotton and potato. For a few of them, physical maps are also under construction, using BAC ends sequencing, with genome sequencing in mind. The dbEST division of GenBank is the central repository for the tag sequences from all species [13]. The production of ESTs is complementary to genome sequencing and has to be taken into account during the annotation process.

2. STRATEGY AND TOOLS

For understanding and best use of annotations, it is essential to know how annotations are made. In the AGI project, each sequencing centre is also an annotation centre and analyses its own sequences. The only exception is Europe where the sequences produced by the two European consortiums are annotated by MIPS (*table I*). For the annotation process, the same general strategy is followed, but in the details, the prediction and analysis tools as well as the reference databases used by the annotation centres differ partially and, consequently, generate heterogeneous annotations.

2.1. Structural annotation

The prediction of the gene elements is a complex problem and its issue is primordial because of its consequences on all the following analyses. Eukaryotic genes with their mosaic structure are more difficult to find than prokaryotic ones which are simple open reading frames. The presence of introns complicates the problem, although the binding sites of the spliceosome may be used to predict the exact position of the exon borders. According to the prediction tools, the result of the prediction concerns the splice sites, the exons or the whole gene (gene modelling software) [69]. These prediction programs are based on two different approaches. The first one, called intrinsic, is based on the features of the genes and genomes [17, 34]. Therefore, a significant and representative set of only experimentally characterized genes is necessary to develop such efficient prediction programs. Furthermore, the origin of the training set has to be species-

Table II. The computational tools used by the five annotation centres of the *Arabidopsis* genome. * http://www.ibc.wustl.edu/bio_data/genefinder/. ** http://ftp.genome.washington.edu/RM/RepeatMasker.html.

Annotation tools		MIPS	KAOS	TIGR	CSH.WU	SPP
NetPlantGene	splice site prediction [39]	+		+	+	+
NetGene2	splice site prediction [79]		+			
SplicePredictor	splice site prediction [19]		+			
GeneMark	exon prediction [16]	+				
Grail	exon prediction [83]	+	+	+	+	+
Xpound	exon prediction [78]					+
Mzef	exon prediction [85]				+	+
XGrail	exon prediction [83]			+		
GeneFinder	gene modelling*	+			+	
GenScan	gene modelling [22]	+	+	+	+	+
Gene.Finder BCM	gene modelling [74]			+	+	+
AAT	gene modelling [41]					+
RepeatMasker	repeat seq. search**				+	+
tRNAscan-SE	tRNA gene prediction [51]		+	+	+	+
tRNA scanner	tRNA gene prediction [33]	+				
BLAST (prot., EST)	similarity search	+	+	+	+	+
Motifs, Blocks…	motifs search	+			+	+
FASTA, S&W…	full-length alignment	+		+	+	
Distribution of annotated sequences		39 %	18 %	20 %	10 %	13 %

specific because each genome has its own features and style [29]. The second approach to find genes and exons, named extrinsic [17], uses the similarities detected in homologous genes (in general at the protein level to increase the sensibility of the search) or, even better, identities with cognate transcript sequences. Only in this ideal last situation is the resulting gene structure asserted and not merely putative. Numerous prediction tools using one or several of these methods with a statistical support, have been developed and detailed in specialized reviews [23, 25]. In the framework of AGI, the annotation protocol often combines several programs to reinforce the efficiency of the predictions [69]. The different softwares used by each annotation centre are listed (first part of *table II*). The performances (sensitivity and specificity) of these softwares differ a lot [62] and, consequently, the final predictions are not identically reliable. The structural annotation is generally semi-automatic because human intervention is necessary to integrate the results of the different predictions, in the final gene structure. Although a fully integrated annotation platform is still lacking, some annotation centres use an interactive interface to visualize the output of the prediction tools and sequence similarity searches to help in this critical decision step. The 'Annotator' tool used by TIGR is a good example of this kind of user interface.

The structural annotation of the AGI does not only concern the genes encoding proteins but also the RNA genes, the different types of transposons and retroelements and the short repeat sequences for which specific databases are created. Similarities are widely detected to perform these annotations, but specific computational tools, such as tRNAscan-SE and RepeatMasker may also be employed (*table II*).

The *Arabidopsis* genome has been the first plant genome to be systematically sequenced and annotated. Therefore, the *Arabidopsis* project is a model for the annotation of the rice genome. As for AGI, IRGSP partners (*table I*) annotate themselves the genomic sequences that they have produced. However, the IRGSP is confronted with the lack of efficient prediction tools that are specifically created for rice, because the number of experimentally identified genes in rice is very low compared to that of *Arabidopsis*. In addition to running sequence similarity searches, the annotation centres predict exons with GenScan, GeneMark.hmm and GeneFinder [70]. The Japanese Rice Genome Research Program (RGP) has adapted GenScan for rice and a version of GeneMark.hmm is also available for the analysis of rice sequences [52], but the low size of the training sets may limit their efficiency. TIGR even uses a maize-adapted GenScan version. The rice EST and cDNA sequencing should soon allow the development of more efficient prediction programs. Meanwhile, scientists who use this information must be aware that the structural annotation of the rice genome is approximate. This problem is even more pronounced for other plant genomes.

2.2. Functional annotation

At the present time, the functional genome annotation is based on the idea that some sequence similarities detected between two proteins mean that they are homologues, i.e. that they come from the same ancestor and share the same biochemical function. Therefore, for each predicted gene, the protein is deduced from the coding region and is compared through BlastP with the protein databases. If the similarities detected are considered relevant, the name (function) of the putative homologue protein is associated with the prediction. This minimum approach allows, in the best cases, the biochemical function of the gene product to be suggested. The high-throughput annotation realized by the annotation centres is too basic and quick to extract reliable information on the biological function. Some annotators confirm and complete the Blast results by full-length alignments between the query protein and the closest homologue detected, and by looking for motifs and family signatures (*table II*). This way appears to be the best to attribute one or several biochemical functions to a predicted protein [40].

The names attributed to the predicted genes/proteins depend on the results of the homologous sequence research. Four categories of genes have been defined, but the associated nomenclature is not very homogeneous (see below). The tendency is nevertheless the following: when a predicted gene product is 100 % identical to an already characterized protein, it receives the same name, whereas sequences with stringent similarity to known proteins are called 'putative' proteins of the same name. The sequences for which only similarities to ESTs are detected are named 'unknown' proteins. Finally, genes without similar sequences and, hence, only deduced from intrinsic prediction programs are labelled 'hypothetical'.

2.3. Annotation display

Although the policy varies between the different AGI annotators, in general, during the sequencing work, the partial sequences are available in the High-Throughput Genome (HTG) database. Once the sequence is complete, it may be annotated before or after its deposition in the public databases. In this last case, the annotations are available in an updated version of the sequence and are presented in the feature section of the sequence entries. The deduced products of the predicted genes are available in the TrEMBL and GenPep databases, which are the protein versions (translated according to the annotations) of EMBL and GenBank, respectively. In contrast to other AGI groups, the annotations produced by the Kazusa institute (KAOS) are not deposited in the sequence entries of the databases. The results of the annotation process are only available on their web site and also regularly published in the 'DNA Research' journal (see [72] for example).

The annotations associated with sequences and accessible by accession numbers are very brief and mainly limited to the co-ordinates of the coding exons and the putative function of the proteins. In order to obtain more detailed information and to know which different data contribute to the final annotation, the users have to consult the web site of the concerned annotation centre (indicated in the header of the sequence entry). The URLs of the different annotators of the AGI are indicated in *table I*.

In general, the web sites provide an access to interactive physical maps of the sequenced and annotated regions or chromosomes. These maps indicate the positions of the contigs of YAC, BAC or P1 clones. For each sequence, a graphical representation of the annotation results is available, presenting the positions of the predicted genes with their intron-exon structure. Other elements, such as repeat sequences, markers and transposons are sometimes included. A few web sites also propose the graphical outputs of the prediction software and similarity researches against protein, genomic sequences and EST databases. These illustrations are a practical way of giving immediately a good idea of the prediction validity. In addition to these graphical displays, some tables or windows provide functional information and propose hypertext links to the concerned gene, deduced protein, cognate transcripts and homologous sequences. Search engines have been developed to find easily a specific annotation by using a gene name, a function or an accession number. The web site built by the University of Pennsylvania and the 'Genome viewer' interface of MIPS are particularly pleasant and efficient to search, evaluate and use the *Arabidopsis* genome annotations. Sometimes, the web sites provide additional analyses and annotations absent from the sequence entries. For instance, MIPS performs a systematic prediction of the sub-cellular localization of the deduced proteins using the TargetP program [31].

3. WHAT DOES ANNOTATION TELL US ABOUT PLANT GENOMES?

Until now, AGI partners have annotated about 80 % of the *Arabidopsis* genome. This global annotation,

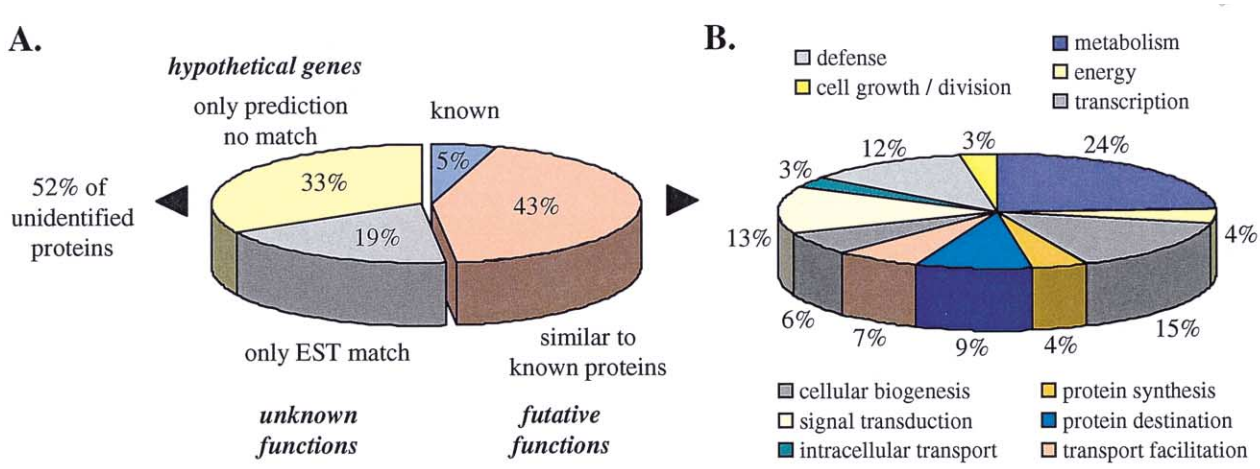

Figure 1. General results of the annotation process in *Arabidopsis*. **A**, Proportion of the different classes of annotated genes. This distribution is based on the type of similar or identical sequences detected during the annotation process. **B**, Classification of the annotated genes according to their biochemical function. The predicted genes with a function characterized experimentally or deduced from sequence comparisons with known proteins (48 % of the total) have been classified according to the functional catalogue defined by the MIPS institute [59]. This result was obtained from the analysis of chromosome 4 [56].

especially of chromosomes 2 and 4, combined with more detailed analyses of shorter genomic fragments and with a few other specific approaches allow us to have a general overview of the *Arabidopsis* genes in terms of structure, function and organization.

At the chromosome level, the gene density is as expected very high, since the mean gene size is approximately 4.5 kb [80]. Nevertheless, the gene density becomes very low in the pericentromeric regions. In contrast, these regions are very rich in repeat sequences and transposons [57]. The approximately 700 rDNA genes are clustered at the extremities of the short arms of chromosomes 2 and 4. Surprisingly, a recent insertion of 75 % of the mitochondrial genome has been found in chromosome 2 and is the result of a large organellar-nuclear DNA transfer event [49]. On average, an *Arabidopsis* gene contains five exons, but this number is probably somewhat underestimated because the predictions concern exclusively the coding regions. Until now, the exons localized in the 5'- and 3'-untranslated regions (UTR) cannot be predicted, and a cognate full-length transcript sequence is necessary to characterize them.

The functional annotation of more than 16 000 genes gives a good idea of the proportion of the different classes of predicted genes. Between 90 and 95 % of the genes revealed by the systematic sequencing are novel genes [10], but a putative function can be attributed by sequence similarities to about 43 % of them (*figure 1A*). The distribution of these genes among the different functional categories defined by the MIPS institute [59] shows that metabolism and transcriptional machinery regroup 40 % of the known genes (*figure 1B*). The assignation of a biochemical function fails for more than half of the predicted genes. These unidentified genes are estimated at 13 000 in the whole *Arabidopsis* genome and will be the main targets of functional genomics in the coming years. Independently of their function, approximately 45 % of predicted genes, obviously the more expressed ones, are tagged by at least one EST or cDNA [80]. This proportion has probably increased with the recent release of numerous additional ESTs in public databases [1].

The TIGR Gene Indices project aims to exhaustively assemble the cognate ESTs and transcripts and to provide, if possible, links to the predicted genes [64]. This powerful approach is in progress for several organisms including *Arabidopsis*, rice, tomato, maize, potato and soybean. The research of perfect EST matches with genomic sequences and their clustering are very fruitful for the annotation of genomic fragments [3] but also of gene families [28, 32].

The importance of the multigene families in the *Arabidopsis* genome is one of the most relevant data revealed by the annotations. Indeed, the systematic searches of similar sequences show that the groups of paralogous genes are more common than unique genes [68]. The size of multigene families varies from a few members to more than 200 paralogous genes [2]. The

presence of these families is directly linked to the dynamic of the genome. Clearly, the genesis of such numerous and large multigene families implicates a high number of duplication events. Duplication is a major motor for the *Arabidopsis* genome evolution [27]. This fact is confirmed by a recent systematic analysis illustrating that chromosome rearrangements, such as duplications and translocations, have occurred extensively, suggesting that *Arabidopsis* could be a degenerated tetraploid [12]. Duplication mechanisms can implicate both short and large genomic regions and even parts of chromosomes [77] and can generate complex interlocked repeat sequences [76] and large gene clusters. Indeed, the annotations show that nearly 20 % of the predicted *Arabidopsis* genes occur as clusters from two to twelve tandemly organized related genes [10].

In spite of the huge differences in genome size, the total number of genes seems quite similar in *Arabidopsis* and cereals. The main source of variation between the genomes is the size of the intergenic regions and, to a lesser extent, the size of introns. The first IRGSP annotations reveal a gene density of one gene every 12 kb in the rice genome. In maize and wheat, the gene distribution on the chromosomes is different: the annotation of a few genomic fragments in maize and barley discloses that genes are clustered in short islands spread in a sea of several repeated elements, mainly transposons [61, 82]. Furthermore, fractionation of DNA from Gramineae followed by gene localization experiments has shown that the genes are specifically localized in DNA fragments (the 'gene space') representing only approximately 20 % of the genome [5]. An interesting strategy for future sequencing and annotation of large cereal genomes would be to focus on these coding sequence-enriched DNA regions [65].

4. WEAKNESSES AND BOTTLENECKS

Large-scale annotation work is a relatively novel task and, by definition, a difficult and risky process. Essentially based on prediction tools, computer technology and biological knowledge always in evolution, the annotation of a genomic sequence is never perfect and always inevitably incomplete. Classical errors and limits inherent to the annotations will be discussed here.

The multinational and, therefore, fragmented organization of genome annotation allows the scientists to follow the daily huge sequence production, but sets the problem of the heterogeneous character of the results. Each AGI annotation centre has its own protocol (*table II*) and provides annotations that are difficult to compare and to exploit by automatic routines. The IRGSP consortium profits from the *Arabidopsis* experience and has defined an almost common protocol.

All annotated genes do not have the same reliability. The validity level of each prediction is rarely specified in the feature section of the sequences, making it necessary to consult the annotation centre's web site. The ideal case is when the different prediction softwares are in agreement with each other and with the detected conserved regions (*figure 2A*). In this too rare situation, the final prediction is very reliable. More frequently, the similarities found are too low and small to influence a choice between different predictions (*figure 2B*) and the validity of the annotation is extremely difficult to estimate. An extreme situation in which the gene modelling programs disagree and the sequence is poorly conserved is illustrated in *figure 2C*. The rebuilt gene has little chance of being true. As shown by the graphical interfaces of the web sites, the data at the origin of annotations can be very different in terms of quality and quantity, and the resulting genes have to be considered with caution.

Classical errors in the structural annotation, as a result of prediction failure, are gene splitting (two genes are predicted instead of one) and gene merging (the opposite situation). Even with exon prediction tools becoming more and more efficient, it is still difficult to predict whether the exons are internal or external. Indeed, the gene extremities are not easily predicted for many reasons. The nucleotide content of introns and intergenic regions is very similar and, at least in *Arabidopsis*, very long introns (up to 4 kb) and very short or even rarely no intergenic regions (overlapping genes) can be found. Consequently, gene modelling programs have difficulties in making the distinction. Furthermore, very few experimental data on plant promoter sequences and translation initiation sites are available to help the 5'-prediction. In the 3'-extremity, the canonical polyadenylation signal is rarely found in plant genes where differential polyadenylation may occur [38] and, hence, cannot be used to predict final exons as is the case for mammals and yeast. Sometimes, only the detection of similarities with one or two different proteins allows the discrimination between one or two genes (*figure 2B*). For this reason, gene merging is more frequent in genomic regions that contain gene clusters.

Prediction softwares evolve very rapidly but it may take some time to recognize which is the most effi-

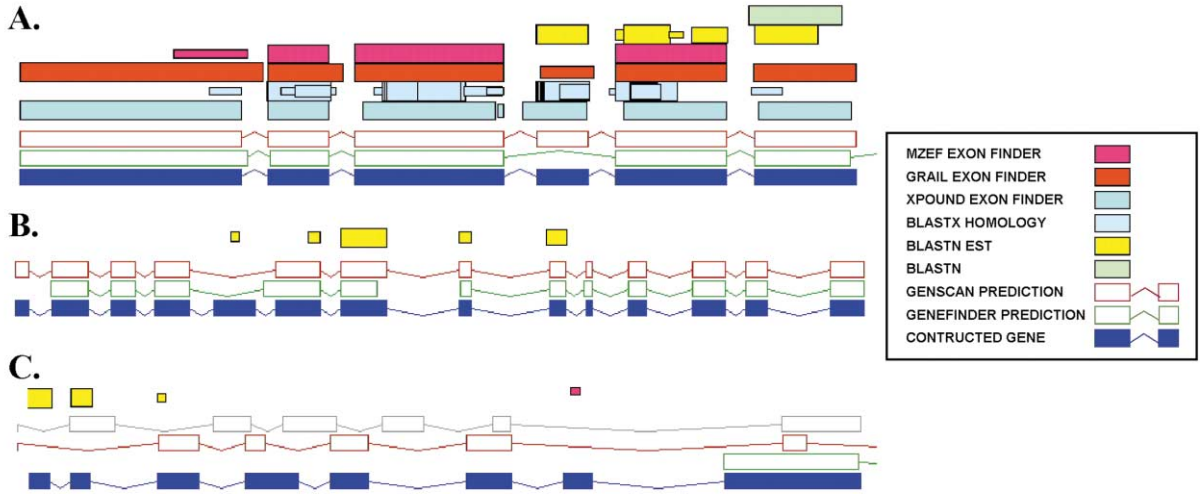

Figure 2. Three examples of structural prediction to illustrate the heterogeneous reliability in the annotations. The results given by the different prediction tools (exon finders and gene modellers) and of the search for similarities within several databases (proteins, ESTs and genomic sequences) are represented by coloured boxes. The height of the boxes is proportional to the prediction or match scores. These examples were found in the web site of the Plant Gene Expression Center (http://pgec-genome.pv.usda.gov) of the SPP consortium.

cient. For example, GeneMark.hmm [52] has been evaluated as the best gene modelling program for *Arabidopsis* but is surprisingly not used by the AGI annotation centres. To give an idea of the prediction efficiency, this program can find 80 % of the actual exons although the modelling of perfect genes is only approximately 40 % [62].

An obvious but important limit is that predictions are based on previously known data and, consequently, the rare and novel gene features (U12 splice sites, overlapping genes, atypical translation start, alternative splicing, etc.) cannot be found in the annotations. The systematic sequencing of full-length cDNA will probably be the unique long-term solution for structural annotations [46, 68].

In functional annotations, the major problem occurs when an erroneous function is attributed to a predicted gene and spread by recurring references in numerous other annotations. The automatic interpretation of a Blast result is very dangerous because it increases the background noise of high-scoring but biologically irrelevant matches. In several cases, the similarities detected by local alignments have to be more deeply analysed to avoid giving importance to a non-significant match [2]. The additional use of full-length alignments between the predicted protein and the best hit(s) detected by Blast reduces the error rate but is not systematically done (*table II*). Because the databases TrEMBL and GenPep are used to find similarities with each novel predicted protein, a false functional annotation can be used as a reference and so be propagated afterwards. This kind of problem is very frequent in the case of multidomain proteins, to which wrong functions attributed by crossed references are spread even when the similarities detected by Blast are significant [73]. By a snowball effect, annotations and databases that exploit them (for instance Pfam and Prodom) are really polluted. Transitively assigning function to a series of closely related sequences appears to be a risky issue [14].

To estimate the consequences of these different difficulties, the automatic annotation of a 400-kb region from *Arabidopsis* chromosome 4 has been compared to a manual annotation carried out by an expert in sequence analysis. The two annotation methods were totally in agreement (at structural and functional levels) for only 23 genes out of 106 [77]. These annotation errors are not specific to *Arabidopsis* and have also been reported in other genomes [20, 26, 48, 73]. At the present time, the fully automatic annotation methods are not satisfactory [36] and the integrated annotation platforms [4, 58] managed by bioanalysts and regularly actualized seem to be the solution.

The nomenclature of genes, proteins and their function is also a source of ambiguities. There is a clear lack of controlled vocabulary both in the literature and the databases. This problem is linked to sequence redundancy in the databases, which can contain sev-

eral times the same genes under different names. The resulting loss in time for the search and the annotation is very serious. Furthermore, the multi-origin of the annotations amplifies the diversity of the nomenclature. For example, the American annotators name as 'putative' or '-like' a function deduced from similarities, whereas the Japanese centre and MIPS use 'potential' and 'similar', respectively. For the latter, 'putative' applies to unknown proteins tagged by EST. The annotations of bacterial genomes suffer from the same kind of problems [48]. The ephemeral feature of some annotations, especially the number of EST matches and the closest homologue, has to be kept in mind when they are used and imposes a regular update [81].

The last important general problem is the lack of accuracy in the annotation sources. It is not always easy to know where the reality ends and where the prediction begins. Predicted and actual genes are considered identical for statistical analyses or definition of motifs and signature of families. The heterogeneity of the annotations can then introduce a significant bias in such studies. Once more, the consultation of the annotator's web sites is necessary, but is not optimized for automatic works.

Biologists have to consider all these weaknesses to optimize their searches and to be capable of exploiting fruitfully genome annotations.

5. FUTURE WAYS

Annotations may be enriched by a variety of emerging experimental and in silico complementary approaches, some of which are listed in *figure 3*. The discovery of a large number of genes allows multiple analysis with a true statistical background and finding novel criteria for gene classification. These novel points of view permit a better understanding of the selection pressures driving the gene evolution and contribute, indirectly, to the improvement of the prediction tools. For example, *Arabidopsis* genes have been classified according to their codon usage to improve gene prediction [55] but also to predict the subcellular localization of their product [24]. In the same way, the spatial structure of proteins will be increasingly used in functional annotation. Indeed, conventional sequence comparisons appear to be limiting for the detection of homologous proteins and the study of 3-D structures allows the increase of the sensitivity [21]. The protein structure being much more highly conserved than the primary sequence, extraction of data on biochemical mechanisms and functions might be greatly facilitated by the analysis of 3-D structures [44].

The building of the *Arabidopsis* Genome Annotation Database (AGAD) by TIGR should solve the problem of the heterogeneous annotation process in

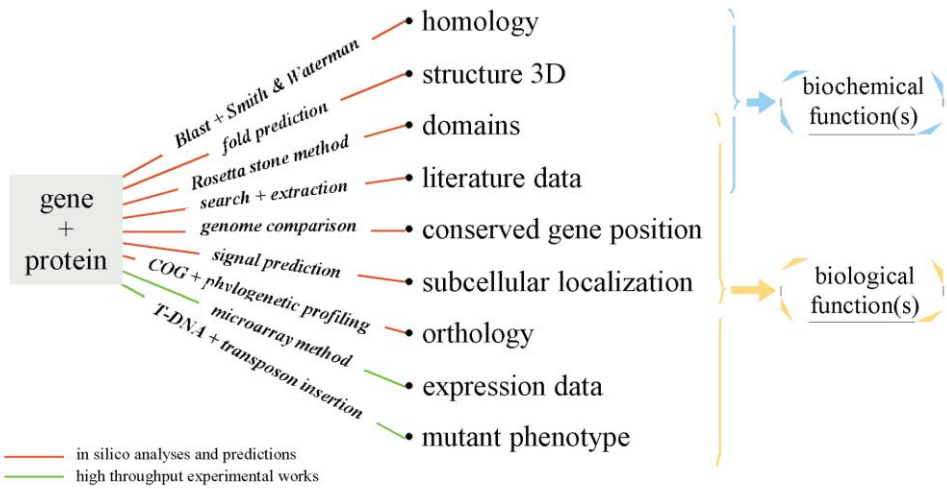

Figure 3. Complementary approaches for a complete functional annotation. The integration of high-throughput in silico and experimental approaches generates a pool of information and hypotheses on biochemical and biological functions, which, when considered all together, can lead to more deeply focused experiments.

Arabidopsis. Its objective is to analyse and re-annotate all the genomic sequences with their own routine and to periodically update the annotations when new tools become available. The results should be a complete and uniform annotation of the *Arabidopsis* genome, presented in a free access database. Concerning the problems of nomenclature, important efforts are in progress. Indeed, the Gene Ontology consortium (http://www.geneontology.org) and the Mendel database (http://genome-www.stanford.edu/Mendel/) try to produce a reference vocabulary for the gene names and functions and a common basis for genome annotation. The *Arabidopsis* Information Resource (TAIR, http://www.arabidopsis.org/) is a central site for this model plant. The web site centralizes all the links towards *Arabidopsis* databases, research laboratories and annotation centres, and also displays regularly updated genetic and physical maps.

Even if the annotations concerning the gene structure and the biochemical function of the proteins are far from perfect and should be largely improved, extraction of information on the biological function of the genes becomes more and more a priority [11]. Up to now, in silico studies were not very efficient to address the question 'in which cellular process is this gene involved?'. The access to several full genomes in different phylum opens interesting methods for functional genomics. Sequence comparison between different organisms is one of the keys for the research of orthologous genes through the definition of COG [75]. In closely related genomes, the conservation of gene positions in the chromosomes can also be considered to confirm the suspected orthologues [42]. Nevertheless, the orthology quest is pertinent for a limited number of genes. A lot of biological functions are not shared by all the compared organisms. For instance, more than 60 % of the annotated *Arabidopsis* genes seem to be plant specific [49].

The whole genome comparison and the research of synteny blocks are also more and more used in order to understand the organization and the evolution of the genomes [8]. For close organisms, the characterization of large conserved regions should provide data not only on the gene position and structure [7], but also on structural and functional elements, such as matrix attachment regions or regulation sequences. Still in the framework of comparative genomics, non-homology approaches such as the phylogenetic profiling [63], the 'Rosetta Stone' method [53] and an analysis supported by gene location [43], have been developed to characterize the function of proteins. With these methods, the gene is always studied in a large context: metabolic or signalling pathways, proteins bearing the same domains and genome environment. Now, the annotations concern not only individual genes but also gene networks in a biological context [35].

The general tendency for the annotation is the development of specialized databases to complement the basic annotation made by the sequencing centres. These specific databases are often built by experts in a method (COG research, motif definition, EST or gene clustering, etc.), an organism (TAIR, Flybase, YDB, etc.), or a multigene family (kinesins, protein kinases, etc.). The experience with multigene families, in particular concerning gene structure, conserved domains, degree of variability, and especially biochemical and biological functions, allows the detailed annotation of the different paralogous genes and proteins, making it possible to highlight distinctive features inside the family.

Even if the bioinformatical approaches may give strong hints on individual gene function, 'wet' biology is absolutely required to discover and finally prove the function of the predicted genes, especially at the level of the cell and organism [60]. High-throughput strategies are extensively used to discover the physiological role of plant genes [18]. Numerous mutant libraries are generated by T-DNA or transposon insertion in order to discover the knock-out phenotypes, and technologies for whole genome expression are producing a huge flow of data [50]. Because there is a tight connection between the expression profile and the biological function of a gene, the integration of the micro-array results into the annotation is primordial. On the other hand, the availability of consistent annotation databases and effective tools for data mining are necessary to fully exploit the various transcriptome analysis technologies [6]. The logical process is exactly the same for the data coming from proteomics research [84]. For these reasons, the development of computational methods for the clustering of genes according to their expression pattern and for the automatic extraction of information from the literature is a hot issue in bioinformatics [66]. In a complementary manner, the accumulation of knowledge on the promoter sequences and the construction of the related databases [67] and research tools should improve the annotation process increasingly.

6. CONCLUSION

The large majority of the structural information contained in the genomic sequences is revealed by

annotation. In *Arabidopsis*, genes completely forgotten by the structural annotation process seems to be exceptional. On the other hand, perfectly predicted genes are too rarely found. The time between the sequencing step and the annotation display is rather short and an important effort is done by annotation centres to propose web sites and relatively easy-to-use graphical interfaces. At the functional level, the problems are more important and over- and under-prediction errors generate a significant loss of information. However, an important work is ongoing to reduce their occurrence and to obtain not only the biochemical but also the biological function.

Until now, the high-throughput annotation, realized progressively and linearly on the genomic sequences, has extracted basic information to quickly satisfy the scientific community. Nevertheless, as explained by Smith and Zhang [73], our impatience for new data degrades the annotations and their longer-term utility. Future annotation should be performed in detail but with a more general view, considering the regulation of the genes, their synergy, and the interactions of their products. The complete annotation of the genes will be, more and more, the result of the integration of clues extracted from numerous approaches [15]. Therefore, the description of genes and genomes passes by the interconnection between specialized databases [30]. The visualization of the highly complex gene and protein networks and their regulation with the environment or development is a major challenge for biologists and computer scientists [45].

Acknowledgments. We would like to thank Raquel Tavares, Jeroen Raes and Martine De Cock for helpful comments and critical reading of the manuscript.

REFERENCES

[1] Asamizu E., Nakamura Y., Sato S., Tabata S., A large scale analysis of cDNA in *Arabidopsis thaliana*: generation of 12,028 non-redundant expressed sequence tags from normalized and size-selected cDNA libraries, DNA Res. 7 (2000) 175–180.

[2] Aubourg S., Boudet N., Kreis M., Lecharny A., In *Arabidopsis thaliana*, 1% of the genome codes for a novel protein family unique to plants, Plant Mol. Biol. 4 (2000) 603–613.

[3] Aubourg S., Takvorian A., Chéron A., Kreis M., Lecharny A., Structure, organization and putative function of the genes identified within a 23.9 kb fragment from *Arabidopsis thaliana* chromosome IV, Gene 199 (1997) 241–253.

[4] Bailey L.C., Fisher S., Schug J., Crabtree J., Gibson M., Overton G.C., GAIA: framework annotation of genomic sequence, Genome Res. 8 (1998) 234–250.

[5] Barakat A., Carels N., Bernardi G., The distribution of genes in the genomes of *Gramineae*, Proc. Natl. Acad. Sci. USA 94 (1997) 6857–6861.

[6] Bassett D.E., Eisen M.B., Boguski M.S., Gene expression informatics – it's all in your mine, Nature Genet. 21 (1999) 51–55.

[7] Batzoglou S., Pachter L., Mesirov J.P., Berger B., Lander E.S., Human and mouse gene structure: comparative analysis and application to exon prediction, Genome Res. 10 (2000) 950–958.

[8] Bennetzen J.L., Comparative sequence analysis of plant nuclear genomes: microcolinearity and its many exceptions, Plant Cell 12 (2000) 1021–1029.

[9] Bevan M., Objective: the complete sequence of a plant genome, Plant Cell 9 (1997) 476–478.

[10] Bevan M., Murphy G., The small, the large and the wild: the value of comparison in plant genomics, Trends Genet. 15 (1999) 211–214.

[11] Bevan M., Bancroft I., Mewes H.W., Martienssen R., McCombie R., Clearing a path through the jungle: progress in *Arabidopsis* genomics, BioEssays 21 (1999) 110–120.

[12] Blanc G., Barakat A., Guyot R., Cooke R., Delseny M., Extensive duplication and reshuffling in the *Arabidopsis* genome, Plant Cell 12 (2000) 1093–1102.

[13] Boguski M.S., Lowe T.M.J., Tolstoshev C.M., dbEST-database for 'expressed sequence tags', Nature Genet. 4 (1993) 332–333.

[14] Bork P., Koonin E.V., Predicting functions from protein sequences – Where are the bottlenecks?, Nature Genet. 18 (1998) 313–318.

[15] Bork P., Dandekar T., Diaz-Lazcoz Y., Eisenhaber F., Huynen M., Yuan Y., Predicting function: from genes to genomes and back, J. Mol. Biol. 283 (1998) 707–725.

[16] Borodovsky M., McIninch J., GENMARK: parallel gene recognition for both DNA strands, Comput. Chem. 17 (1993) 123–133.

[17] Borodovsky M., Rudd K.E., Koonin E.V., Intrinsic and extrinsic approaches for detecting genes in a bacterial genome, Nucleic Acids Res. 22 (1994) 4756–4767.

[18] Bouchez D., Höfte H., Functional genomics in plants, Plant Physiol. 118 (1998) 725–732.

[19] Brendel V., Kleffe J., Carle Urioste J.C., Walbot V., Prediction of splice sites in plant pre-mRNA from sequence properties, J. Mol. Biol. 276 (1998) 85–104.

[20] Brenner S.E., Errors in genome annotation, Trends Genet. 15 (1999) 132–133.

[21] Brenner S.E., Chothia C., Hubbard T.J., Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships, Proc. Natl. Acad. Sci. USA 95 (1998) 6073–6078.

[22] Burge C., Karlin S., Prediction of complete gene structures in human genomic DNA, J. Mol. Biol. 268 (1997) 78–94.

[23] Burge C., Karlin S., Finding the genes in genomic DNA, Curr. Opin. Struct. Biol. 8 (1998) 346–354.
[24] Chiapello H., Ollivier E., Landes-Devauchelle C., Nitschke P., Risler J.L., Codon usage as a tool to predict the cellular location of eukaryotic ribosomal proteins and aminoacyl-tRNA synthetases, Nucleic Acids Res. 27 (1999) 2848–2851.
[25] Claverie J.M., Computational methods for the identification of genes in vertebrate genomes sequences, Hum. Mol. Genet. 6 (1997) 1735–1744.
[26] Claverie J.M., Do we need a huge new center to annotate the human genome?, Nature 403 (2000) 12.
[27] Clegg M.T., Cummings M.P., Durbin M.L., The evolution of plant nuclear genes, Proc. Natl. Acad. Sci. USA 94 (1997) 7791–7798.
[28] Cooke R., Raynal M., Laudié M., Delseny M., Identification of members of gene families in *Arabidopsis thaliana* by contig construction from partial cDNA sequences: 106 genes encoding 50 cytoplasmic ribosomal proteins, Plant J. 11 (1997) 1127–1140.
[29] Danchin A., The Delphic boat or what the genomic texts tell us, Bioinformatics 14 (1998) 383.
[30] Dicks J., Plant genome databases: from references to inferences tools, Briefings Bioinform. 1 (2000) 138–150.
[31] Emanuelsson O., Nielsen H., Brunak S., von Heijne G., Predicting subcellular localization of proteins based on their N-terminal amino acid sequence, J. Mol. Biol. 300 (2000) 1005–1016.
[32] Epple P., Apel K., Bohlmann H., ESTs reveal a multigene family for plant defensins in *Arabidopsis thaliana*, FEBS Lett. 400 (1997) 168–172.
[33] Fichant G.A., Burks C., Identifying potential tRNA genes in genomic DNA sequences, J. Mol. Biol. 220 (1991) 659–671.
[34] Fickett J.W., The gene identification problem: an overview for developer, Comput. Chem. 20 (1996) 103–118.
[35] Galperin M.Y., Brenner S.E., Using metabolic pathway databases for functional annotation, Trends Genet. 14 (1998) 332–333.
[36] Galperin M.Y., Koonin E.V., Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement, and operon disruption, In silico Biol. 1 (1998) 7.
[37] Goff S.A., Rice as a model for cereal genomics, Curr. Opin. Plant Biol. 2 (1999) 86–89.
[38] Graber J.H., Cantor C.R., Mohr S.C., Smith T.F., In silico detection of control signals: mRNA 3'-end-processing sequences in diverse species, Proc. Natl. Acad. Sci. USA 24 (1999) 14055–14060.
[39] Hebsgaard S.M., Korning P.G., Tolstrup N., Engelbrecht J., Rouzé P., Brunak S., Splice site prediction in *Arabidopsis thaliana* pre mRNA by combining local and global sequence information, Nucleic Acids Res. 24 (1996) 3439–3452.
[40] Henikoff S., Greene E.A., Pietrokovski S., Bork P., Attwood T.K., Hood L., Gene families: the taxonomy of protein paralogs and chimeras, Science 278 (1997) 609–614.
[41] Huang X., Adams M.D., Zhou H., Kerlavage A.R., A tool for analyzing and annotating genomic sequences, Genomics 46 (1997) 37–45.
[42] Huynen M.A., Bork P., Measuring genome evolution, Proc. Natl. Acad. Sci. USA 95 (1998) 5849–5856.
[43] Huynen M.A., Snel B., Lathe W., Bork P., Exploitation of gene context, Curr. Opin. Struct. Biol. 10 (2000) 366–370.
[44] Jones D.T., Protein structure prediction in the postgenomic area, Curr. Opin. Struct. Biol. 10 (2000) 371–379.
[45] Kaiser J., From genome to functional genomics, Science 288 (2000) 1715.
[46] Kato A., Suzuki M., Kuwahara A., Ooe H., Higano-Inaba K., Komeda Y., Isolation and analysis of cDNA within a 300 kb *Arabidopsis thaliana* genomic region located around the 100 map unit of chromosome 1, Gene 239 (1999) 309–316.
[47] Kyrpides N.C., Genomes OnLine Database (GOLD): a monitor of complete and ongoing genome projects world wide, Bioinformatics 15 (1999) 773–774.
[48] Kyrpides N.C., Ouzounis C.A., Whole-genome sequence annotation: 'Going wrong with confidence', Mol. Microbiol. 32 (1999) 886–887.
[49] Lin X., Kaul S., Rounsley S., Shea T.P., Benito M.I., Town C.D., et al., Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*, Nature 402 (1999) 761–768.
[50] Lockhart D.J., Winzeler E.A., Genomics, gene expression and DNA arrays, Nature 405 (2000) 827–836.
[51] Lowe T.M., Eddy S.R., tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence, Nucleic Acids Res. 25 (1997) 955–964.
[52] Lukashin A.V., Borodovsky M., GeneMark.hmm: New solutions for gene finding, Nucleic Acids Res. 26 (1998) 1107–1115.
[53] Marcotte E.N., Pellegrini M., Thompson M.J., Yeates T.O., Eisenberg D., A combined algorithm for genome-wide prediction of protein function, Nature 402 (1999) 83–86.
[54] Marienfeld J., Unseld M., Brennicke A., The mitochondrial genome of *Arabidopsis* is composed of both native and immigrant information, Trends Plant Sci. 12 (1999) 495–502.
[55] Mathé C., Déhais P., Pavy N., Rombauts S., Van Montagu M., Rouzé P., Gene prediction and gene classes in *Arabidopsis thaliana*, J. Biotechnol. 78 (2000) 293–299.
[56] Mayer K., Schüller C., Wambutt R., Murphy G., Volckaert G., Pohl T., et al., Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*, Nature 402 (1999) 769–777.
[57] McCombie W.R., de la Bastide M., Hebermann K., Parnell L., Dedhia N., Gnoj L., et al., The complete sequence of a heterochromatic island from a higher eukaryote, Cell 100 (2000) 377–386.
[58] Médigue C., Rechenmann F., Danchin A., Viari A., Imagene: an integrated computer environment for sequence annotation and analysis, Bioinformatics 15 (1999) 2–15.
[59] Mewes H.W., Albermann K., Bähr M., Frishman D.,

Gleissner A., Hani J., Heumann K., Kleine K., Maierl A., Oliver S.G., Pfeiffer F., Zollner A., Overview of the yeast genome, Nature 387 (1997) 7–84.
[60] Miklos G.L.G., Rubin G.M., The role of the genome project in determining gene function: insights from model organisms, Cell 86 (1996) 521–529.
[61] Panstruga R., Büschges R., Piffanelli P., Schulze-Lefert P., A contiguous 60 kb genomic stretch from barley reveals molecular evidence for gene islands in a monocot genome, Nucleic Acids Res. 26 (1998) 1056–1062.
[62] Pavy N., Rombauts S., Déhais P., Mathé C., Ramana D.V., Leroy P., Rouzé P., Evaluation of gene prediction software using a genomic data set: application to *Arabidopsis thaliana* sequences, Bioinformatics 15 (1999) 887–899.
[63] Pellegrine M., Marcotte E.M., Thompson M.J., Eisenberg D., Yeates T.O., Assigning protein functions by comparative genome analysis: protein phylogenetic profiles, Proc. Natl. Acad. Sci. USA 96 (1999) 4285–4288.
[64] Quackenbush J., Liang F., Holt I., Pertea G., Upton J., The TIGR Gene Indices: reconstruction and representation of expressed gene sequences, Nucleic Acids Res. 28 (2000) 141–145.
[65] Rabinowicz P.D., Schutz K., Dedhia N., Yordan C., Parnell L.D., Stein L., McCombie W.R., Martienssen R.A., Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome, Nature Genet. 23 (1999) 305–308.
[66] Renner A., Aszodi A., High-throughput functional annotation of novel genes products using document clustering, Pac. Symp. Biocomput. 12 (2000) 54–68.
[67] Rombauts S., Déhais P., Van Montagu M., Rouzé P., PlantCARE, a *cis*-acting regulatory element database, Nucleic Acids Res. 27 (1999) 295–296.
[68] Rounsley S., Lin X., Ketchum K.A., Large-scale sequencing of plant genomes, Curr. Opin. Plant Biol. 1 (1998) 136–141.
[69] Rouzé P., Pavy N., Rombauts S., Genome annotation: which tools do we have for it?, Curr. Opin. Plant Biol. 2 (1999) 90–95.
[70] Sasaki T., Burr B., International Rice Genome Sequencing Project: the effort to completely sequence the rice genome, Curr. Opin. Plant Biol. 3 (2000) 138–141.
[71] Sato S., Nakamura Y., Kaneko T., Asamizu E., Tabata S., Complete structure of the chloroplast genome of *Arabidopsis thaliana*, DNA Res. 5 (1999) 283–290.
[72] Sato S., Nakamura Y., Kaneko T., Katoh T., Asamizu E., Tabata S., Structural analysis of *Arabidopsis thaliana* chromosome 3. Sequence features of the region of 4,504,864 bp covered by sixty P1 and TAC clones, DNA Res. 7 (2000) 131–135.
[73] Smith T.F., Zhang X., The challenges of genome sequence annotation or 'The devil is in the details', Nature Biotech. 15 (1997) 1222–1223.
[74] Solovyev V., Salamov A., in: Gaasterland T., Karp P., Karplus K., Ouzounis C., Sander C., Valencia A. (Eds.), The Fifth International Conference on Intelligent Systems for Molecular Biology, Halkidiki, Greece, 1997, pp. 294–302.
[75] Tatusov R.L., Koonin E., Lipman D.J., A genomic perspective on protein families, Science 278 (1997) 631–637.
[76] Tavares R., Aubourg S., Lecharny A., Kreis M., Organization and structural evolution of four multigene families in *Arabidopsis thaliana*: AtLCAD, AtLGT, AtMYST and AtHD-GL2, Plant Mol. Biol. 5 (2000) 703–717.
[77] Terryn N., Heijnen L., De Keyser A., Van Asseldonck M., De Clercq R., Verbakel H., et al., Evidence for an ancient chromosomal duplication in *Arabidopsis thaliana* by sequencing and analyzing a 400-kb contig at the *APETALA2* locus on chromosome 4, FEBS Lett. 445 (1999) 237–245.
[78] Thomas A., Skolnick M.H., A probabilistic model for detecting coding regions in DNA sequences, IMA J. Math. Appl. Med. Biol. 11 (1994) 149–160.
[79] Tolstrup N., Rouzé P., Brunak S., A branch point consensus from *Arabidopsis* found by non circular analysis allows for better prediction of acceptor sites, Nucleic Acids Res. 25 (1997) 3159–3163.
[80] Wambutt R., Murphy G., Volckaert G., Pohl T., Düsterhöft A., Stiekema W., et al., Progress in *Arabidopsis* genome sequencing and functional genomics, J. Biotechnol. 78 (2000) 281–292.
[81] Wheelan S.J., Boguski M.S., Late-night thoughts on the sequence annotation problem, Genome Res. 8 (1998) 168–169.
[82] White S., Doebley J., Of genes and genomes and the origin of maize, Trends Genet. 14 (1998) 327–332.
[83] Xu Y.X., Uberbacher E.C., Automated gene identification in large-scale genomic sequences, J. Comput. Biol. 4 (1997) 325–338.
[84] Yates J.R., Mass spectrometry and the age of the proteome, J. Mass Spectrom. 33 (1998) 1–19.
[85] Zhang M.Q., Identification of protein coding regions in the human genome by quadratic discriminant analysis, Proc. Natl. Acad. Sci. USA 94 (1997) 565–568.