GeneFarm, structural and functional annotation of Arabidopsis gene and protein families by a network of experts

Sebastien Aubourg, Veronique Brunaud, Clémence Bruyère, J. Mark Cock, Richard Cooke, Annick Cottet, Arnaud Couloux, Patrice Dehais, Gilbert Deléage, Aymeric Duclert, et al.

Nucleic Acids Research, 2005, Vol. 33, Database issue ***D641–D646***
doi:10.1093/nar/gki115

GeneFarm, structural and functional annotation of *Arabidopsis* gene and protein families by a network of experts

Sébastien Aubourg[1,*], Véronique Brunaud[1], Clémence Bruyère[2], Mark Cock[3], Richard Cooke[4], Annick Cottet[6], Arnaud Couloux[2], Patrice Déhais[7], Gilbert Deléage[8], Aymeric Duclert[9], Manuel Echeverria[5], Aimée Eschbach[10], Denis Falconet[6], Ghislain Filippi[1], Christine Gaspin[11], Christophe Geourjon[8], Jean-Michel Grienenberger[10], Guy Houlné[10], Elisabeth Jamet[10], Frédéric Lechauve[3], Olivier Leleu[7], Philippe Leroy[12], Régis Mache[6], Christian Meyer[13], Hafed Nedjari[12], Ioan Negrutiu[14], Valérie Orsini[10], Eric Peyretaillade[12], Cyril Pommier[1], Jeroen Raes[7], Jean-Loup Risler[15], Stéphane Rivière[14], Stéphane Rombauts[7], Pierre Rouzé[7], Michel Schneider[16], Philippe Schwob[6], Ian Small[1], Ghislain Soumayet-Kampetenga[10], Darko Stankovski[2], Claire Toffano[2], Michael Tognolli[16], Michel Caboche[1] and Alain Lecharny[1,2]

[1]Unité de Recherche en Génomique Végétale (INRA/CNRS/UEVE) 2 Rue Gaston Crémieux, CP 5708, 91057 Evry Cedex, France, [2]Institut de Biotechnologie des Plantes (CNRS/UPS) Bâtiment 630, Université Paris-Sud, 91405 Orsay Cedex, France, [3]Station Biologique de Roscoff (CNRS/UPMC) Place Georges Tessier, BP 74, 29682 Roscoff Cedex, France, [4]Laboratoire de Physiologie et de Biologie Moléculaire des Plantes (CNRS/UP) and [5]Laboratoire Génome et Développement des Plantes (CNRS/UP), 52 Avenue de Villeneuve, Université de Perpignan, 68860 Perpignan Cedex, France, [6]Laboratoire Plastes et Différenciation Cellulaire (CNRS/UJF) Université J. Fourrier, BP 53, 38041 Grenoble Cedex 09, France, [7]Plant Systems Biology (INRA/VIB/GU) K.L. Ledeganckstraat 35, 9000 Ghent, Belgium, [8]Institut de Biologie et Chimie des Protéines (PBIL/CNRS) 7 Passage du Vercors, 69367 Lyon Cedex 7, France, [9]Unité de Recherche en Génomique-Info (INRA/GENOPLANTE) 523 Place des Terrasses, 91000 Evry, France, [10]Institut de Biologie Moléculaire des Plantes (CNRS) 12 Rue du Général Zimmer, 67084 Strasbourg Cedex, France, [11]Equipe Statistique et Informatique (INRA) Chemin de Borde Rouge, Auzeville BP 27, 31326 Castanet-Tolosan, France, [12]UMR Amélioration and Santé des Plantes (INRA/UBP) 234 Avenue du Brézet, Domaine de Crouelle, 63039 Clermont-Ferrand Cedex 2, France, [13]Laboratoire Nutrition Azotée des Plantes (INRA) Route de Saint-Cyr, 78026 Versailles Cedex, France, [14]Laboratoire de Reproduction et Développement des Plantes (INRA/CNRS/ENS/UL) 46 Allée d'Italie, Université Lyon I, 69364 Lyon Cedex 07, France, [15]Laboratoire Génome et Informatique (CNRS) 523 Place des Terrasses, 91034 Evry, France and [16]Swiss Institute of Bioinformatics (Swiss-Prot) CMU, 1 Michel Servet, CH-1211 Genève 4, Switzerland

Received August 12, 2004; Revised and Accepted October 15, 2004

ABSTRACT

Genomic projects heavily depend on genome annotations and are limited by the current deficiencies in the published predictions of gene structure and function. It follows that, improved annotation will allow better data mining of genomes, and more secure planning and design of experiments. The purpose of the GeneFarm project is to obtain homogeneous, reliable, documented and traceable annotations for *Arabidopsis* nuclear genes and gene products, and to enter them into an added-value database. This re-annotation project is being performed exhaustively on every member of each gene family. Performing a family-wide annotation makes the task easier and

*To whom correspondence should be addressed. Tel: +33 1 60 87 45 16; Fax: +33 1 60 87 45 49; Email: aubourg@evry.inra.fr

more efficient than a gene-by-gene approach since many features obtained for one gene can be extrapolated to some or all the other genes of a family. A complete annotation procedure based on the most efficient prediction tools available is being used by 16 partner laboratories, each contributing annotated families from its field of expertise. A database, named GeneFarm, and an associated user-friendly interface to query the annotations have been developed. More than 3000 genes distributed over 300 families have been annotated and are available at http://genoplante-info.infobiogen.fr/Genefarm/. Furthermore, collaboration with the Swiss Institute of Bioinformatics is underway to integrate the GeneFarm data into the protein knowledgebase Swiss-Prot.

INTRODUCTION

The GeneFarm project was launched in 2001 soon after the announcement of the near-completion of the *Arabidopsis thaliana* genome (1). The initial annotation released at the same time as the assembled sequence of the five chromosomes was largely a compilation of independent annotations from different members of the *Arabidopsis* Genome Initiative (AGI) consortium. The generally recognized drawback of this otherwise invaluable resource was that this annotation was often faulty and misleading. Important discrepancies have been identified, for example when these initial annotations were later compared to more time-consuming expert-driven annotations, especially in the definitions of intron–exon boundaries and in erroneous names for genes or gene products (2,3). Owing to the cost in time and money of human expertise, genome annotation has often been restricted to the prediction of coding exons and to the labelling of the deduced protein with the function of its closest homologue (4) resulting in the under-annotated databases where errors are often multiplied by a snowball effect (5). Finally, the source of a specific annotation feature, such as whether the annotation feature originates from the external documented information or from prediction software has rarely been stated (6).

During the last four years, the TIGR institute has made available five updated versions of the *Arabidopsis* chromosome sequences with associated structural and functional annotation (7). The structural semi-automatic annotation has been greatly improved by the development of new prediction software using the rapidly expanding transcript resources, mainly expressed sequence tags (ESTs) and full-length cDNAs (8–10). For functional predictions, TIGR has made an important effort to search known protein motifs and to classify the predicted genes according to the Gene Ontology method (11). Nevertheless, the computational part of the automatic gene annotation has been globally limited using heterogeneous intrinsic (sequence composition, signals, etc.) and extrinsic (cognate transcripts, similarities, etc.) data. The outcome of the automated annotation process is constrained by general rules defined to limit the number of false positive and false negative predictions. The biological complexity which includes many atypical situations in gene structure and organization along the chromosomes (alternative events, U12 splicing sites, pseudogenes, micro-exons, overlapping genes, etc.) cannot be described using satisfactory models and this constitutes a significant limitation to the annotation pipelines (7,12). An overview of the last release (TIGR R5.0) of the *Arabidopsis* chromosomes shows that the associated annotation is still not optimal and, considering its pivotal role as a reference plant resource and as a tool for genomic projects, an improved annotation would certainly be of wide general interest. It would allow better planning and design of future experiments such as high-throughput functional analysis of genes (13) and characterization of interaction networks.

Completion and correction of the existing semi-automatic gene prediction will require a more in-depth approach and, for this, the manual intervention of expert biologists is unavoidable (14,15). An expert-based approach is the solution that has been chosen for the construction of the Swiss-Prot library in which the information associated with specific sequences is generated and rigorously controlled by expert annotators (16). This task is time consuming and limits the quantity of proteins that can be processed. For instance, in July 2004, Swiss-Prot contained 2853 *Arabidopsis* entries as compared with the 10 times greater number of predicted genes in the *Arabidopsis* genome. The goal of GeneFarm is to actively participate in this manual annotation effort and to extend it at the gene/nucleic acid level. The GeneFarm project is based on a network of scientists working in different fields of research allowing an extensive and curated annotation of *Arabidopsis* nuclear genes. In order to optimize the added value of this human expertise, the annotation process focuses on gene families since many of the features and much of the information mined in the literature or predicted for one gene can often be extrapolated to some or all the homologous genes (17). Performing a gene family-based annotation makes the task easier and more efficient than a gene-by-gene approach. Indeed, due to their common origin, genes from the same family quite often share the same gene intron–exon structure. Furthermore, sequence comparisons of all the members of a protein family help to highlight conserved motifs responsible for shared biochemical function(s) and point to specific features characteristic of one or a subset of paralogous genes. The complete functional study of a given gene that belongs to a family of duplicated paralogs (as is frequently the case in plants) should take into consideration its evolutionary relationship with the other members of the family. Therefore, systematically characterizing gene families in *Arabidopsis* and identifying particular characteristics of each member is an essential step to define orthologous relationships with genes from other plant species.

THE GeneFarm PHILOSOPHY

The main motivating aims during the definition of the GeneFarm database were (i) to obtain a consistent annotation across the different annotators, (ii) to track the annotation sources and (iii) to use a common bioinformatics toolbox to reduce annotation heterogeneity to a minimum. Based on precise evaluations of both the automatic annotation

bottlenecks (18) and of the performances of prediction software (19), a minimum annotation protocol (i.e. mandatory steps) was defined. For example, at the gene structure level, the minimum protocol uses the Eugene (20) and GeneMark.hmm (21) programs, which were specially trained with *Arabidopsis* datasets and showed the best results compared with other programs, both at the exon and model gene levels. Other examples of mandatory steps, but this time at the protein level, are the Predotar program used to predict targeting peptides (22) and a combination of DSC (23), PHD (24) and SOPMA (25) for the prediction of secondary structures. Whatever the annotation steps and the software used, the results are always checked and compared by the biologist partners before being accepted. When available, experimental results coming from participant's laboratories or from publications are given precedence over the results of prediction software. In order to make the loading task easy, robust and traceable, two web submission interfaces were developed for the annotators, one for the gene and a second for the family descriptions. In the GeneFarm database, each piece of information is clearly justified either by experimental proof (unpublished data or bibliographic references), an accession number (motifs, structure, sequence, etc.) or reference to a prediction software. Each biologist partner is in charge of annotating several *Arabidopsis* gene families that are targets of their own research field. Often, results have been produced for another purpose, such as research into gene function, but have not been published in a form that is usable for the scientific community. The GeneFarm approach delivers an annotation of high quality with precise and detailed features and numerous links to the pertinent literature. Furthermore, the close examination by an expert annotator ensures that the best and most up-to-date nomenclature and ontology is used to name all the genes of the same family. In GeneFarm, the definition of a gene family is based on sequence similarities and on evidence for a common evolutionary origin (homology). The boundary between different families is not always easy to define and the expert annotators play an important role in defining this. Some of the GeneFarm partners are involved in methodological approaches, which provide additional aid for the identification of homologous genes. For example, the PHYTOPROT resource, in which all available plant proteins are clustered by an all-by-all systematic comparison (26), is being used as a starting point to define gene families, and a comparison of predicted secondary structures is also being exploited with the aim of detecting highly divergent homologous proteins (27).

THE CONTENT OF THE GeneFarm DATABASE

The GeneFarm database contains gene entries and family entries. The family entries contain the description of the families including common features shared by all of the homologous genes (signature, biochemical function, keywords, paper review, etc.). The gene entries contain the complete annotation of the genes including data specific to each gene. This information is organized into different sections: target plant and genomic sequence, gene name and synonym(s), references to all cognate transcripts, intron–exon structure(s), deduced protein(s), regulatory motifs in promoters, biochemical function, protein localization, motifs and domains, secondary structure, post-translational maturation sites, biological function(s), mutant phenotype(s), expression condition(s), cross-references with other databases and bibliographic references. Each gene entry is linked to its corresponding family entry.

Currently, the GeneFarm database contains more than 3000 gene entries distributed among 300 gene families (Figure 1). The sizes of the families range from 2 paralogous genes (40% of cases, a consequence of the ancient duplication which affected almost the entire *Arabidopsis* genome) to 270 members (the cytochrome P450 family). An overview of the GeneFarm database shows that the annotation of the gene entries includes more than 35 000 cross-references with GeneBank/EMBL/DDBJ, 750 with Swiss-Prot, 6000 with motif databases, 2700 literature references, 3500 transcription proofs and detailed descriptions of more than 1700 expression conditions. The GeneFarm website contains a list of the annotator partners and their assigned families, lists of annotated gene and gene family entries and an interface to query the database. This interface allows access to genes and gene families using their names, their AGI or GeneFarm accession number (GF AC), keywords, expression conditions or sequence comparisons (with BLAST). The page results display the annotations with dynamic web links to the referenced databases. There are links between the genes and their corresponding family. In order to help the users to quickly have a general idea of the extent of the annotation, in term of details and experimental support, two scores (from 1 to 5) have been defined for structural and functional annotations. Figure 2 shows the distribution of the annotated genes as a function of these two scores and describes the scoring system in more detail.

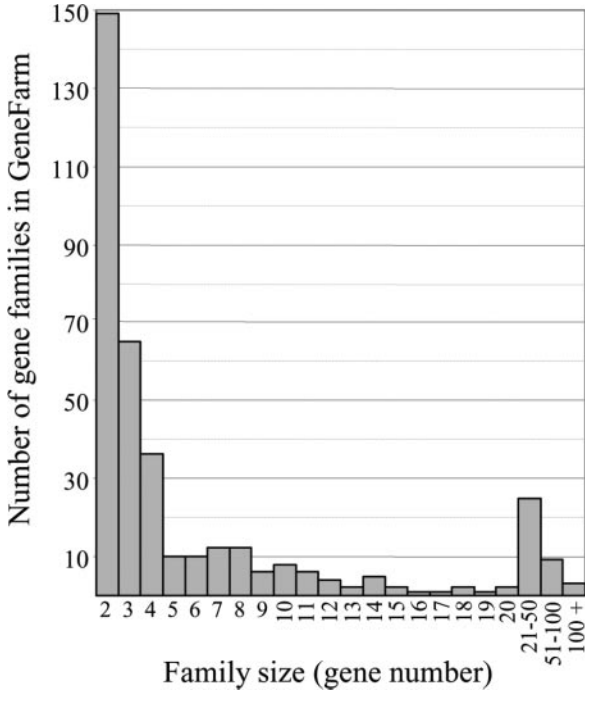

Figure 1. Distribution of the gene families in the GeneFarm database according to the number of annotated paralogs in the *Arabidopsis thaliana* genome.

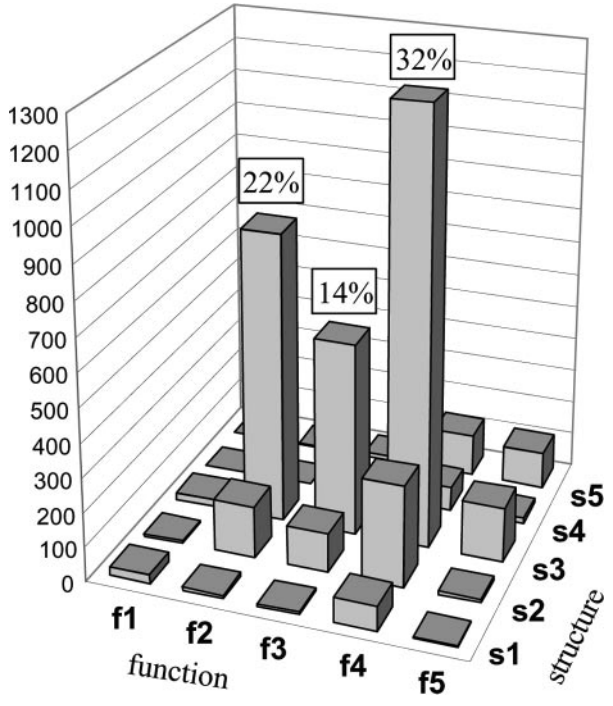

Figure 2. Distribution of the genes annotated in the GeneFarm database according to their scores at the structural and functional levels. The structural score depends on the origin of the annotated intron–exon structure: s1, prediction software only; s2, prediction software and similarities with homologous genes; s3, the gene structure is partially covered by a transcript (EST, RT–PCR product, etc.); s4, the whole CDS is covered by a transcript; and s5, a cognate full-length cDNA is available (TSS and UTR are known). The functional score: f1, unknown function (no information); f2, some predicted clues (motif, signal, etc.); f3, similarities with a known gene; f4, biochemical function proved; and f5, biological function experimentally shown.

EXAMPLES OF ADDED VALUE

One of the strong points concerning GeneFarm is that annotators are members of a coordinated project with regular work meetings. Therefore, the work is not redundant and is of controlled quality. We have tried to estimate the gain in annotation quality of the expertised annotation compared to the semi-automatic annotation. It is evident that the gain should be higher for the functional annotation as compared to the structural one. Nevertheless, the former cannot be quantified and therefore we only present results of a systematic comparison of the GeneFarm and the TIGR CDS structures. Structural differences have been observed for 751 genes out of the 3501 that have been re-annotated (21%) within the framework of GeneFarm. Differences are more frequently observed for genes that do not have cognate cDNA or EST sequences. Indeed 254 out of 870 (29%) genes without transcript support differ in their CDS structure between the TIGR and GeneFarm resources.

A concrete example of the contribution of the GeneFarm effort is the collective ongoing annotation of the PPR family (PentratricoPeptide Repeat proteins). This huge family of 442 proteins is characterized by a complex arrangement of short motifs (28,29) deciphered using two different bioinformatics approaches, the MEME/MAST and the HMMER packages. In the TIGR annotation, most of the PPR genes are tagged by the motif PF01535 from the PFAM database (30). The structural annotation of this family is particularly poorly done by automatic procedures. Even the unique motif PF01535 does not cover all the repeats defined by the GeneFarm experts. Examples of the corrections proposed in GeneFarm for regions containing misleadingly annotated PPR genes are illustrated in Figure 3A and B. GeneFarm contains the complete re-annotation of a subgroup of 89 genes of the PPR family, named PCMP-H, and will soon contain data for the whole PPR family and thus provide a unified annotation reflecting as accurately as possible the complex structural organization of these proteins.

Incorrect structural annotations often lead to erroneous functional labelling of genes. For example, the gene *PCMP-H16* (GF AC 3179) is annotated as being homologous to the yeast SEC14 cytosolic factor in the TIGR annotation due to the fusion of two genes to create a single predicted gene *AT5G04780* (Figure 3B). This type of error is easily detected by expert analysis.

More surprisingly, even in the case of well-known families with relatively high-sequence conservation, erroneous gene predictions can be found that are contradicted by cognate transcripts, by the conserved positions of introns between paralogues and by the presence of Pfam motifs, as illustrated by the cytochrome P450 *AT4G20240* (Figure 3C).

Since molecular data are sometimes lacking, the gene models in the GeneFarm database should still be considered as predictions in many cases. However, we believe that owing to the manual comparisons performed between the different prediction approaches, including those carried out using TIGR and MIPS, together with the extensive analysis of the families by the annotators, the gene models have a high probability of corresponding to the real gene structure.

CONCLUSION

The GeneFarm network has carried out checked, curated, justified, homogeneous and deep annotation of more than 3000 nuclear genes distributed among 300 complete gene families in *Arabidopsis*. This resource is organized in a relational database and available on the GeneFarm website at http://genoplante-info.infobiogen.fr/Genefarm/. All the annotations corresponding to the protein sequences are also available in the UniProt knowledgebase (31). One of the partners of the project, the Swiss Institute of Bioinformatics (SIB), acts in synergy with GeneFarm annotators in order to improve annotations and to provide the scientific community with high-quality protein data via Swiss-Prot entries. To benefit from this dual expertise, a special DR (Database cross-Reference) line has been added to Swiss-Prot entries to point out to the corresponding GeneFarm entries. Reciprocally, each GeneFarm entry is cross-referenced to the relevant Swiss-Prot entry. Furthermore, the FLAGdb++ database (32) provides a graphical visualization of the GeneFarm gene structures in the context of the TIGR annotation. The GeneFarm project aims to provide, during the year 2005, a complete and detailed biological description of about 5500 *Arabidopsis* nuclear genes and more than 450 gene families. GeneFarm participates to the demanding collection of expert annotations also performed using TAIR (33) and AtGDB (34). In the long

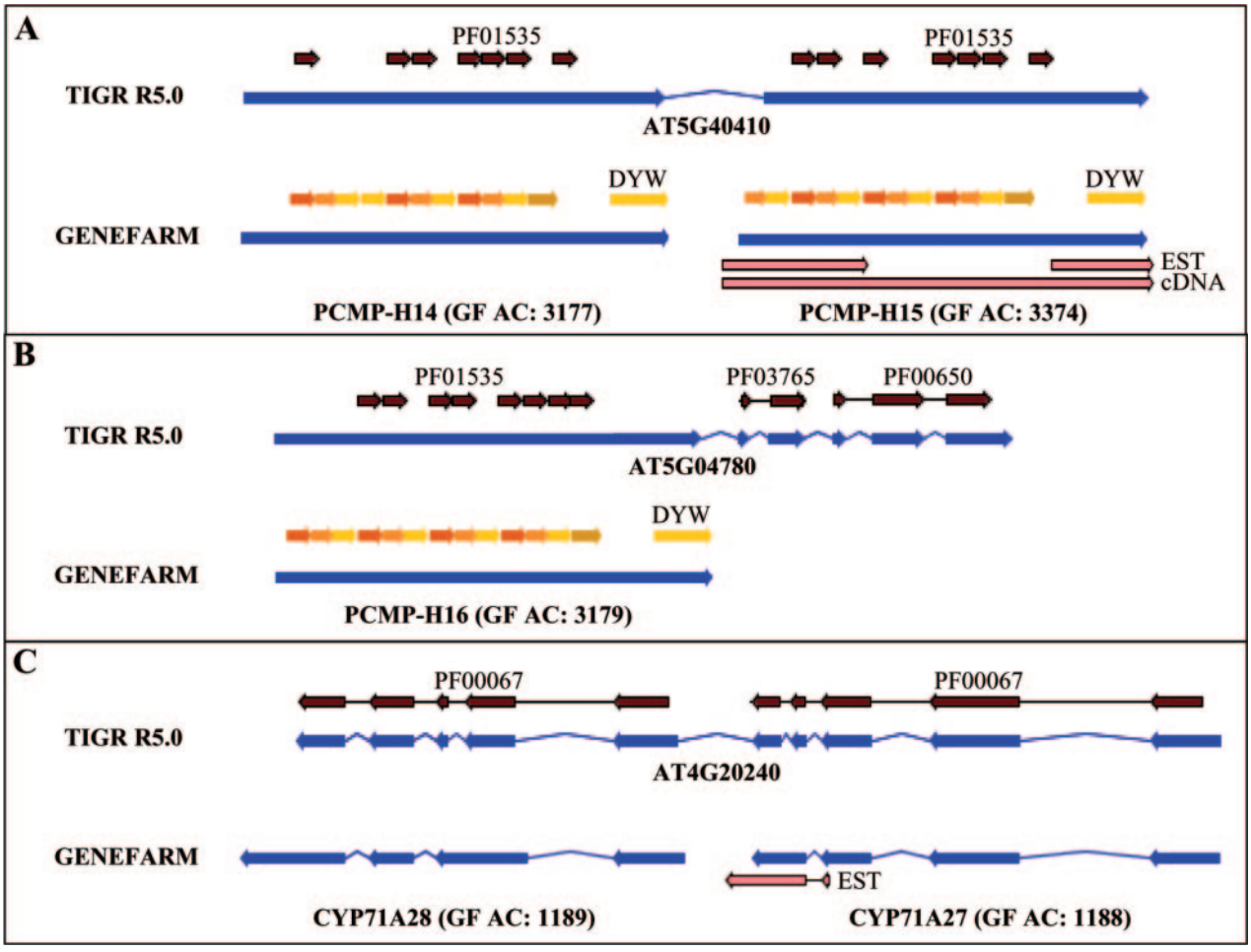

Figure 3. Examples of corrections to TIGR annotations proposed by GeneFarm. (**A**) Fusion of two PPR genes revealed by a detailed definition of the repeat motifs (4 different matrixes have been defined by GeneFarm annotators to exhaustively tag all the repeat motifs of the PPR family), presence of C-terminal DYW motifs and cognate transcripts. (**B**) The consequence of this fusion of a PPR gene with a downstream gene is the attribution of a function on the basis of the presence of PFAM motifs PF03765 and PF00650. GeneFarm suggests two genes instead of one based on the presence of a C-terminal DYW motif in the first gene. The second gene has not been re-annotated in the framework of GeneFarm. (**C**) Gene fusion and erroneous exon boundaries. The GeneFarm corrections are supported by the fact that the gene model is shared by other members of the CYP sub-group, a cognate EST and better scores with the Pfam motif PF00067. Blue arrows and lines: CDS exons and introns, respectively. Brown arrows: PFAM motifs mapped to exons. Pink arrows: transcript sequences. Other arrows: different types of PPR repeats.

term, it will be important to enlarge the GeneFarm effort to other plant species and, thus, to provide a database for curated orthologous relationships across the plant kingdom.

ACKNOWLEDGEMENTS

The authors are grateful to Vincent Thareau for his help in the comparison between the GeneFarm and TIGR annotations. The GeneFarm database has been filed at the 'Agence pour la Protection des Programmes' under the key number IDDN.FR.001.070013.001.S.C.2004.000.10300. This work is supported by GENOPLANTE grants under the projects Bi1999087 and Bi2001071.

REFERENCES

1. Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.
2. Bork,P. and Koonin,E.V. (1998) Predicting functions from protein sequences, where are the bottlenecks? *Nature Genet*., **18**, 313–318.
3. Terryn,N., Heijnen,L., De Keyser,A., Van Asseldonck,M., De Clercq,R., Verbakel,H., Gielen,J., Zabeau,M., Villarroel,R., Jesse,T. *et al*. (1999) Evidence for an ancient chromosomal duplication in *Arabidopsis thaliana* by sequencing and analysing a 400-kb contig at the APETALA2 locus on chromosome 4. *FEBS Lett*., **445**, 237–245.
4. Smith,T.F. and Zhang,X. (1997) The challenges of genome sequence annotation or 'the devil is in the details'. *Nat. Biotechnol*., **15**, 1222–1223.
5. Gilks,W.R., Audit,B., De Angelis,D., Tsoka,S. and Ouzounis,C.A. (2002) Modelling the percolation of annotation errors in a database of protein sequences. *Bioinformatics*, **18**, 1641–1649.
6. Aubourg,S. and Rouzé,P. (2001) Genome annotation. *Plant Physiol. Biochem*., **39**, 181–193.
7. Wortman,J.R., Haas,B.J., Hannick,L.I., Smith,R.K., Maiti,R., Ronning,C.M., Chan,A.P., Yu,C., Ayele,M., Whitelaw,C.A. *et al*. (2003) Annotation of the *Arabidopsis* genome. *Plant Physiol*., **132**, 461–468.
8. Haas,B.J., Delcher,A.L., Mount,S.M., Wortman,J.R., SmithR.K.,Jr, Hannick,L.I., Maiti,R., Ronning,C.M., Rusch,D.B., Town,C.D. *et al*. (2003) Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res*., **31**, 5654–5666.
9. Zhu,W., Schlueter,S.D. and Brendel,V. (2003) Refined annotation of the *Arabidopsis* genome by expressed sequence tag mapping. *Plant Physiol*., **132**, 469–484.
10. Castelli,V., Aury,J.M., Jaillon,O., Wincker,P., Clepet,C., Menard,M., Cruaud,C., Quetier,F., Scarpelli,C., Schachter,V. *et al*. (2004) Whole

genome sequence comparisons and 'full-length' cDNA sequences: a combined approach to evaluate and improve *Arabidopsis* genome annotation. *Genome Res*., **14**, 406–413.
11. Camon,E., Magrane,M., Barrell,D., Lee,V., Dimmer,E., Maslen,J., Binns,D., Harte,N., Lopez,R. and Apweiler,R. (2004) The Gene Ontology Annotation (GOA): sharing knowledge in UniProt with Gene Ontology. *Nucleic Acids Res*., **32**, D262–D266.
12. Volfovsky,N., Haas,B.J. and Salzberg,S.L. (2003) Computational discovery of internal micro-exons. *Genome Res*., **13**, 1216–1221.
13. Hilson,P., Small,I. and Kuiper,M.T. (2003) European consortia building integrated resources for *Arabidopsis* functional genomics. *Curr. Opin. Plant Biol*., **6**, 426–429.
14. Hubbard,T. and Birney,E. (2000) Open annotation offers democratic solution to genome sequencing. *Nature*, **403**, 825.
15. Rhee,S.Y. (2004) *Carpe diem*: retooling the 'publish or perish' model into the 'share and survive' model. *Plant Physiol*., **134**, 543–547.
16. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.-C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al*. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*., **31**, 365–370.
17. Abascal,F. and Valencia,A. (2003) Automatic annotation of protein function based on family identification. *Proteins*, **53**, 683–692.
18. Mathé,C., Sagot,M.F., Schiex,T. and Rouzé,P. (2002) Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res*., **30**, 4103–4117.
19. Pavy,N., Rombauts,S., Déhais,P., Mathé,C., Ramana,D.V., Leroy,P. and Rouzé,P. (1999) Evaluation of gene prediction software using a genomic data set: application to *Arabidopsis thaliana* sequences. *Bioinformatics*, **15**, 887–899.
20. Foissac,S., Bardou,P., Moisan,A., Cros,M.-J. and Schiex,T. (2003) Eugène′hom: a generic similarity-based gene finder using multiple homologous sequences. *Nucleic Acids Res*., **31**, 3742–3745.
21. Lukashin,A.V. and Borodovsky,M. (1998) GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res*., **26**, 1107–1115.
22. Small,I., Peeters,N., Legeai,F. and Lurin,C. (2004) Predotar: a tool for rapidly screening proteomes for N-terminal targeting sequences. *Proteomics*, **4**, 1581–1590.
23. King,R.D. and Sternberg,M.J. (1996) Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. *Protein Sci*., **5**, 2298–2310.
24. Rost,B. and Sander,C. (1993) Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol*., **232**, 584–599.
25. Geourjon,C. and Deléage,G. (1995) SOPMA: significant improvement in protein secondary structure prediction by consensus prediction from multiple alignments. *Comput. Appl. Biosci*., **11**, 681–684.
26. Mohseni-Zadeh,S., Louis,A., Brezellec,P. and Risler,J.-L. (2004) PHYTOPROT: a database of clusters of plant proteins. *Nucleic Acids Res*., **32**, D351–D353.
27. Geourjon,C., Combet,C., Blanchet,C. and Deléage,G. (2001) Identification of related proteins with weak sequence identity by using secondary structure information. *Protein Sci*., **10**, 788–797.
28. Aubourg,S., Boudet,N., Kreis,M. and Lecharny,A. (2000) In *Arabidopsis thaliana*, 1% of the genome codes for a novel protein family unique to plants. *Plant Mol. Biol*., **42**, 603–613.
29. Lurin,C., Andres,C., Aubourg,S., Bellaoui,M., Bitton,F., Bruyere,C., Caboche,M., Debast,C., Gualberto,J., Hoffmann,B. *et al*. (2004) Genome-wide analysis of *Arabidopsis* pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis. *Plant Cell*, **16**, 2089–2103.
30. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L. *et al*. (2004) The Pfam protein families database. *Nucleic Acids Res*., **32**, D138–D141.
31. Apweiler,R., Bairoch,A., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M.*et al*. (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res*., **32**, D115–D119.
32. Samson,F., Brunaud,V., Duchêne,S., De Oliveira,Y., Caboche,M., Lecharny,A. and Aubourg,S. (2004) FLAGdb++: a database for the functional analysis of the *Arabidopsis* genome. *Nucleic Acids Res*., **32**, D347–D350.
33. Rhee,S.Y., Beavis,W., Berardini,T.Z., Chen,G., Dixon,D., Doyle,A., Garcia-Hernandez,M., Huala,E., Lander,G., Montoya,M. *et al*. (2003) The *Arabidopsis* Information Resource (TAIR): a model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. *Nucleic Acids Res*., **31**, 224–228.
34. Dong,Q., Schlueter,S.D. and Brendel,V. (2004) PlantGDB, plant genome database and analysis tools. *Nucleic Acids Res*., **32**, D354–D359.