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

Sébastien Aubourg1,*, Véronique Brunaud1, Clémence Bruyère2, Mark Cock3, Richard Cooke4, Annick Cottet4, Arnaud Couloux2, Patrice Déhais7, Gilbert Deléage8, Aymeric Ducler9, Manuel Eccherria5, Aimée Eschbach10, Denis Falconet6, Ghislain Filippi1, Christine Gaspin11, Christophe Geourjon6, Jean-Michel Grienenberger10, Guy Houlné10, Elisabeth Jamet10, Frédéric Lechauve3, Olivier Leleu7, Philippe Leroy12, Régis Mache6, Christian Meyer13, Hafed Nedjari12, Ioan Negruțiu14, Valérie Orsini10, Eric Peyretaillade12, Cyril Pommier1, Jeroen Raes7, Jean-Loup Risler15, Stephane Rivière14, Stéphane Rombauts7, Pierre Rouzé7, Michel Schneider16, Philippe Schwob6, Ian Small1, Ghislain Soumaya-Kametenga10, Darko Stankovski2, Claire Toffano2, Michael Tognolli16, Michel Caboche1 and Alain Lecharny1,2

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

The online version of this article has been published under an open access model. Users are entitled to use, reproduce, disseminate, or display the open access version of this article for non-commercial purposes provided that: the original authorship is properly and fully attributed; the Journal and Oxford University Press are attributed as the original place of publication with the correct citation details given; if an article is subsequently reproduced or disseminated not in its entirety but only in part or as a derivative work this must be clearly indicated. For commercial re-use permissions, please contact journals.permissions@oupjournals.org.

© 2005, the authors

Nucleic Acids Research, 2005, Vol. 33, Database issue © Oxford University Press 2005; all rights reserved
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.](image-url)
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
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


