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 Duclet9, Manuel Echeverria5, Aimée Eschbach10, Denis Falconet6, Ghislain Filippi1, Christine Gaspin11, Christophe Geourjon8, Jean-Michel Grienenberger10, Guy Houlné10, Elisabeth Jamet10, Frédéric Lechauve3, Olivier Leleu7, Philippe Leroy12, Régis Mache6, Christian Meyer13, Hafed Nedjari12, Ioan Negrutiu14, Valérie Orsini10, Eric Peyretaillade12, Cyril Pommier1, Jeroen Raes7, Jean-Loup Risler15, Stéphane Rombauts7, Pierre Rouzé7, Michel Schneider16, Philippe Schwob6, Ian Small1, Ghislain Soumayet-Kampetenga10, 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 Tessier, 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 (PBI/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, Auzelles 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...
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-complete of the Arabidopsis thaliana genome (1). The initial annotation released at the same time as the assembled sequence of 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 its annotation was of often faulty and misleading. Important discrepancies have been identified, for example when a set of initial annotations were later compared to the more time-consuming expert-driven annotations, especially in the definition 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 a gene annotation feature originates from the external documentation 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 its rapidly expanding transcript resources, mainly expressed sequence tags (ESTs) and full-length cDNAs (8–10). For functional prediction, 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 annotation (9) involves the following steps: (i) to obtain a consistent annotation across different members of the Arabidopsis chromosomes shows that the associated error is still no 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. This would allow better planning and design of future experiments such as high-throughput functional analysis of genes (12) and characterize annotation of interest it works.

Complete and correct identification of existing semi-automatic gene prediction will require a more in-depth approach and, for this, the manual insertion of expert biological annotation 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 10 times greater number of predicted genes in the Arabidopsis genome. The goal of GeneFarm is to actively participate in this process and, for instance, in the manual annotation effort and to extend it to the gene/nucleic acid level. The GeneFarm project is based on the 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 the 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 one or all the homologous genes (17).

Performing a gene family-based annotation makes it easier and more efficient than a gene-by-gene approach. Indeed, due to its common origin, genes from the same family are more or less share the same gene in the intron-exon structure. Furthermore, sequence comparisons of all the members of a protein family help highlight conserved motifs responsible for shared biochemical functions and point to specific features characteristic of one or a subset of paralogous genes. The complete functional study of a given gene belongs to a family of duplicated paralogs (as is frequently the case in plant species) should take in consideration its evolution and ionary 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 or logical 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 different Arabidopsis chromosomes, (ii) to track the annotation history, and (iii) to use a common bioinformatics toolbox to reduce annotation errors to a minimum. Based on precise evaluations of both the automatic annotation which includes many typical 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 the construction of a significant annotation of the Arabidopsis pipelines (7,12). An overview of this last release (TIGR R5.0) of the Arabidopsis chromosomes shows that the associated error 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. This would allow better planning and design of future experiments such as high-throughput functional analysis of genes (13) and characterize annotation of interest it works.

Complete and correct identification of existing semi-automatic gene prediction will require a more in-depth approach and, for this, the manual insertion of expert biological annotation 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 10 times greater number of predicted genes in the Arabidopsis genome. The goal of GeneFarm is to actively participate in this process and, for instance, in the manual annotation effort and to extend it to the gene/nucleic acid level. The GeneFarm project is based on the 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 the 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 one or all the homologous genes (17).

Performing a gene family-based annotation makes it easier and more efficient than a gene-by-gene approach. Indeed, due to its common origin, genes from the same family are more or less share the same gene in the intron-exon structure. Furthermore, sequence comparisons of all the members of a protein family help highlight conserved motifs responsible for shared biochemical functions and point to specific features characteristic of one or a subset of paralogous genes. The complete functional study of a given gene belongs to a family of duplicated paralogs (as is frequently the case in plant species) should take in consideration its evolution and ionary 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 or logical relationships with genes from other plant species.
by bottle necks (18) and of the performances of predic ion software (19), a minimum annotation protocol (i.e. mandatory steps) was defined. For example, at the gene structure level, he minimum protocol uses Eugene (20) and GeneMark.hmm (21) programs, which were specially trained with Arabidopsis datasets and showed the best results compared with other programs, both in exon and model gene levels. Of her examples of mandatory steps, for his time a he protein level, are he Predotar program used o predict arg peptides (22) and a combination of DSC (23), PHD (24) and SOPMA (25) he for he predic ion of secondary stures. Whatever he anno ion s and he soft ware used, he result s are always checked and compared by he biologists partners before being accepted. When available, experimental result s coming from par leipan’s laboratories or from publications are given precedence over he result s of predic ion soft ware. In order o make he loading ask easy, robus and raceable, wo web submission in erases were developed for he anno ions, one for he gene and a second for he family descriptions.

In he GeneFarm database, each piece of information is clearly justified either by experimental proof (unpublished data or bibliographic references), an accession number (nucleotides, sequence, etc.) or reference o a predicedion software. Each biologist partner is charged of annoating several Arabidopsis gene families ha are arg e s of heir own research field. Of en, results have been produced for and other purpose, such as research on gene function, but have not been published in a form ha is usable for he scientific community. The GeneFarm approach delivers an anno ion of high qua ity with precise and detailed features and numerous links o he per in m erase. Fur hermore, he close examination by an expert annoator ensures ha he bes est and most up-to-date nomenclature and ontology is used o name all he genes of he same family. In GeneFarm, he 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 he experts annoators play an important role in defining his. Some of he GeneFarm partners are involved in methodological approaches, which provide additional aid for he identification of homologous genes. For example, he PHYTOPROT resource, in which all available plant protein data 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 explored with he 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 he description of he families including common features shared by all of he homologous genes (signature, biochemical function, keywords, paper review, etc.). The gene entries contain he complete annoation of he genes including he specific s of each gene. This information is organized in o different sections: gene name and synonyms, references o all cognate transcripts, intron–exon sture(s), deduced protein(s), regulary motifs in
The GeneFarm project aims to provide, during the year 2005, a complete and detailed 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 this project is the Swiss Institute of Bioinformatics (SIB), which is in synergy with GeneFarm annotators in order to improve annotation and 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 annotators the corresponding GeneFarm entries. Reciprocally, each GeneFarm entry is cross-referenced to the relevant Swiss-Prot entry and vice versa.

**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, he work is not redundant and is of controlled quality. We have tried to emphasize the expression of annotation quality by 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 the GeneFarm project. Differences are more frequently observed for the TIGR and the GeneFarm CDS structures. Structural differences are 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/.

**CONCLUSION**

The GeneFarm project has carried out a detailed, 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 this project is the Swiss Institute of Bioinformatics (SIB), which is in synergy with GeneFarm annotators in order to improve annotation and 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 GeneFarm database (32) provides a graphical visualization of the GeneFarm gene structure 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 par ticipates in the demanding collection of expert annotations also performed using TAIR (33) and AtGDB (34).
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


