ONTOGONY OF ARABIDOPSIS GENE NET SUPPLEMENTARY DATABASE (AGNS), CROSS DATABASE REFERENCES TO TAIR ONTOLOGY

Mironova V.V.*, Poplavsky A.S., Ponomaryov D.K., Omelianchuk N.A.
1 Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia; 2 Institute of Informatics System, SB RAS, Novosibirsk, 630090, Russia
* Corresponding author: e-mail: kviki@bionet.nsc.ru

Key words: ontology, plant development, controlled vocabulary, Arabidopsis

SUMMARY

Motivation: Inter-database queries on the same subject domain require a coordinated vocabulary of terms. The existing plant ontology (PO) has been made mainly for description of large-scale experiments and contains gaps, which should be filled to provide terms for description of gene expression patterns revealed by in situ hybridization and phenotype abnormalities in mutant and transgenic plants.

Results: We have created an Arabidopsis PO based on controlled vocabularies of AGNS (Arabidopsis Gene Net Supplementary) database, which contains a detailed information on anatomical elements and development stages from annotated papers on gene expression and phenotype abnormalities in mutant and transgenic plants. The comparison of ANGS ontology with PO allowed us to develop a common ontology providing both inter-database queries based on general standards and an opportunity to describe the whole scale of transcriptomic and phenomic data in the course of Arabidopsis development.

Availability: http://wwwmgs2.bionet.nsc.ru/agns

INTRODUCTION

To study development in Arabidopsis one needs to have an access to various and scattered information in publications and to different databases storing results of large-scale experiments. Description of this growing body of data using a common format would considerably facilitate analysis of the information and support inter-database queries. The Plant Ontology Consortium (POC) founded in 2002, has developed ontology for describing anatomical elements and developmental stages in Arabidopsis. This ontology consists of two controlled vocabularies: a controlled vocabulary describing morphological and anatomical structures representing organ, tissue and cell types and a controlled vocabulary on growth and developmental stages (Jaiswal et al., 2005). The main purpose of these vocabularies is to facilitate the cross-database querying and to foster consistent use of these vocabularies in annotation of tissue and/or growth stage specific expression of genes, proteins and phenotypes. Relationships between terms of such PO within controlled vocabularies and between them are represented as Directed Acyclic Graphs (DAGs.) A DAG is similar to a hierarchical structure but is superior because terms (representing concepts) within a DAG structure have the ability to have one or more parents. For example, in the dictionary of anatomy and morphology of plants, the following initial
classes of terms are defined: “cell”, “tissue”, “sporophyte”, etc. In this case, the term “root epidermis” corresponds to both parental classes: “tissue” and “sporophyte”. There are three types of relationships between PO classes: “is a”, “part of” and “develops from”. At present, the PO of Arabidopsis contains information on structure and stages of germplasm development and information on organs and stages of Arabidopsis development, which is sufficient for description of microarray experiments and gene annotation.

Our aims comprise accumulating data on gene expression from various experiments and information on phenotypic anomalies of Arabidopsis. These tasks are addressed in the AGNS database (Arabidopsis GeneNet Supplementary DataBase) (Omelyanchuk et al., 2005). Controlled vocabularies of AGNS contain detailed information on stages of development of separate organs and the whole plant, anatomy and morphology of the plant, as detailed as it could be found in numerous publications on Arabidopsis development. We have constructed the Arabidopsis PO on the basis of these controlled vocabularies. Besides the function of data systematization, this PO can be used for studying the plant development. We compare our data in PO AGNS with PO TAIR and discuss the results of this work and application of PO AGNS in this paper.

METHODS AND ALGORITHMS

There are two ways of representation of PO Arabidopsis in the AGNS: the navigation system of AGNS and the Protégé ontology editor 3.1. In the AGNS database one can see informational content of controlled vocabularies on separate pages, which are marked by corresponding tabs: Development (the vocabulary for stages of organs development) and Morphology (the vocabulary of organs morphology). On the same pages a search in vocabularies for detailed description of the particular stage of development and morphology of a selected organ is provided (Fig. 1a).

Figure 1. Representation of PO Arabidopsis in navigation system of AGNS by the example of hierarchy and description of early embryonic stages in the controlled vocabulary on development stages (a). Representation of PO Arabidopsis in Protégé system by the example of Whorl-4 development during flowering (b).

In the Protégé system we have constructed an alternative version of PO using the OWL language (Ponomaryov et al., 2006). The proposed ontology (see Fig. 1b) describes plant anatomy in development and is filled in with data from Arabidopsis development studies that have been summarized by development of the AGNS controlled vocabulary.
IMPLEMENTATION AND RESULTS

The PO of the AGNS database consists of two hierarchically structured controlled vocabularies for stages of development and for anatomy and morphology of Arabidopsis. These vocabularies have been constructed around the contents of annotated data based on both the data provided by the authors of the annotated papers, and from specialized publications on plant developmental stages and morphology of the wild type. The most frequently used names of the stages and organs are highlighted and their synonyms are given, also description of stage and organ includes quantitative data and accompanied by detailed comments. The vocabularies are being supplemented with new research data as they become available. During construction of the controlled vocabularies we were consistent with the hierarchy requirement: organs and stages of development are organized with enclosure from the smaller to the greater, where it is possible. A characteristic feature of PO AGNS is that the controlled vocabulary for anatomy and morphology is presented in two parts: (1) as a structured terminological vocabulary like in TAIR, (2) as a complete vocabulary for descriptions of anatomy and morphology of Arabidopsis. In the terminological vocabulary anatomical elements are presented in sections “organs”, “tissues”, and “domains” with a brief description. This vocabulary is used in AGNS input system, and we have carried out its comparison and linking with PO TAIR. The complete controlled vocabulary is filled automatically with the entered combinations from sections of AGNS terminological vocabulary. The description of an anatomical element in this dictionary can be generalized as “organ, organ, …., domain, …. domain, tissue”, where description of organs is represented in the hierarchical form, and the name parts “domain” and “tissue” can be skipped or reiterated in various combinations. In the complete controlled vocabulary the information on description of an anatomical element, quantitative characteristics and other useful data with reference to publication are entered. This vocabulary is used for navigation within the database and for further analysis of the data. It may be also useful as a reference book for educational and research purposes.

In the controlled vocabulary for stages of development we did not separate “plant growth stages” or “body part developmental stages” in PO AGNS. Both types are joined and “plant growth stages” are higher in the hierarchy. Such structure divides whole ontogenesis into four stages: an embryogenesis, a seedling, a transition to flowering, and a reproductive phase. A term “vegetative growth” is used for description of development of a vegetative part during the last three stages.

We have carried out a comparison of terms in the controlled vocabularies of PO Arabidopsis AGNS with terms from PO Arabidopsis TAIR. Terms concerning the description of the same plant structures and identical stages of development in PO AGNS have been replaced by corresponding terms from PO TAIR. New terms that have not described in AGNS have been added.

In the controlled vocabulary for stages of development about 50 % of the terms have been compared. For the rest of the terms in PO TAIR no analogues have been found. These are the terms for development stages of some organs, such as embryo, leaf primordium, midvein, endosperm, and others. During large-scale experiments such details are usually omitted, but they are necessary for studying gene expression using in situ hybridization and for description of phenotypic anomalies.

The comparison of controlled vocabularies for anatomy and morphology has been carried out for the terminological dictionary at first. We have found out a lot of differences in terms of POs. This has occurred because there is no detailed description of anatomical structure for many organs in TAIR, which are selected by researchers during analysis of plant development. For example, there is no description of domains in the embryo, no description of some organs of flowers and leaves, etc. On the other hand, in TAIR PO description for germplasm, tissue culture and root is much better developed.
The comparison and in some cases replacement of terms has been performed for AGNS PO to be according to international nomenclature. Therefore, AGNS and TAIR vocabularies are partly overlapping in common terms, and comprehensively describe different parts of the plant organism. Their integration may become an important breakthrough in ontology development.

DISCUSSION

Earlier the construction of PO has been an additional task during description of large-scale experiments on gene expression. This approach has resulted in a too general character of such information. Recently a tendency towards description and systematization of other data type has appeared. These types of data include phenotypic abnormalities of plants in mutants, insertion and transgene lines, patterns of gene expression in normal and mutant types, obtained not only by large-scale experiments. A high accuracy of determining the plant stage and organ is typical for these data (up to number of cells expressing the selected gene). Systematization of such data requires to construct a more comprehensive PO, which would allow to describe not only general data, but all necessary details as well. The proposed PO Arabidopsis in AGNS meets these demands: it is used for description of gene expression obtained in various experiments, from large-scale down to in situ hybridizations, and for description of phenotypic anomalies in mutant and transgenic lines. Furthermore, the Protégé-version of the AGNS ontology can be a useful tool for studying the development, since it will allow to analyze phenotypic abnormalities and to reveal the causes and time of their appearances. We believe that the developed controlled vocabularies and ontology will be a useful resource for the entire plant science community.

ACKNOWLEDGEMENTS

This work was supported in part by INNOVATION PROJECT of Federal Agency of Science and innovation IT-CP.5/001 “Development of software for computer modeling and design in postgenomic system biology (system biology in silico)”, Russian Foundation for Basic Research (grants Nos 05-07-98012 and 03-04-48506), Russian Academy of Sciences (grant No. 10.4), Siberian Branch of Russian Academy of Sciences (Integration Project No. 119) and the US National Science Foundation (FIBR EF-0330786 Development Modeling and Bioinformatics).

REFERENCES