AGNS (ARABIDOPSIS GeneNet SUPPLEMENTARY DATABASE), RELEASE 3.0

***Omelianchuk N.A.*[*1], *Mironova V.V.*[1], *Poplavsky A.S.*[1], *Pavlov K.S.*[1], *Savinskaya S.A.*[2], *Podkolodny N.L.*[1], *Mjolsness E.D.*[3], *Meyerowitz E.M.*[4], *Kolchanov N.A.*[1]**

[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia; [2] Novosibirsk State University, Novosibirsk, 630090, Russia; [3] Institute for Genomics and Bioinformatics; University of California, Irvine CA 92607, USA; [4] Division of Biology, California Institute of Technology, Pasadena, CA 91125, USA

[*] Corresponding author: e-mail: nadya@bionet.nsc.ru

Key words: *Arabidopsis*, gene expression, phenotype, database

SUMMARY

Motivation: Arabidopsis GeneNet Supplementary DataBase (AGNS) provides an integrated view of genetic data for *Arabidopsis thaliana* (Omelyanchuk *et al.*, 2006). AGNS contains the description of the functions of the known Arabidopsis genes at the levels of mRNA, protein, cell, tissue, organs and the organism in both wild type and mutant backgrounds. AGNS annotates published papers on gene expression and function and by using a special format and interface integrates, systematizes, and classifies this heterogeneous, disparate, and scattered information.

Results: The AGNS structure, formats and content have now been dramatically revised in order to ensure the efficient use of the records, rapid growth of the content and integration with other databases. The Arabidopsis ontology used in AGNS has been brought up to international standard. A new section, SD, which collects data on mutation localization in gene nucleotide sequences and provides descriptions to transgenic constructs, has been added. A new web interface with data submission and database navigation capabilities has been designed and implemented. Major updates have been made to all the AGNS sections.

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

INTRODUCTION

To study development in plant, access is required to the various and scattered information in publications found as descriptions of gene expression patterns and morphological abnormalities in different mutant and transgenic lines, to compare these data with those for the wild type. Analysis of this body of data and its reduction to a single format would considerably facilitate analysis of the information. The different parts of this information started to be annotated, which is indicative of growing interest to the systematization of such data. TAIR (http://www.arabidopsis.org/) and Geneinvestigator (https://www.genevestigator.ethz.ch/) keep collecting information on gene expression in Arabidopsis obtained from large-scale experiments, AREX systematizes evidence-based (microarrays, in situ, promoter::reporter constructs, etc.) data on Arabidopsis gene expression in the root (http://www.arexdb.org/database.jsp). WatDB (http://www.watdb.nl/) and SeedGenes (http://www.seedgenes.org) systematize data on phenotypic aberrations in mutants. However, a complete understanding of the function of a gene can be reached only when the data about localization of mutations in nucleotide sequences and

information on expression patterns of genes in wild type, mutant and transgenic plants will be associated with detailed description of phenotypes of the wild type, mutant and transgenic plants (including double mutants and transgenic plants in the mutant background). AGNS was developed as a tool, which allow to trace the whole history from change in the nucleotide sequence of a particular gene, through further changes in the expression of that or associated genes, to phenotypic changes. In two previous releases the structure and format of three sections (Expression, Phenotype and Reference Databases) and two controlled vocabularies have been developed (Fig. 1). Expression Database describes gene expression in wild type, mutant, and transgenic plants. Phenotype Database contains information on phenotypic abnormalities of particular organs at particular stage in mutant and transgenic plants. Reference Database includes references to the papers and description of plant growth conditions with an indication of the ecotypes used as controls in the experiments. Detailed controlled vocabularies on growth stages and morphology were developed around the annotated data.

The rationale for AGNS release 3.0 was the need for improvements to the available data presentation format and user interface. The improved format serves to collect the results from different types of experiments on gene expression; to describe and systematize data on phenotypic abnormalities in single mutants, double mutants or combinations of mutant genes with transgenes; to provide cross database querying on Arabidopsis. The user interface in its current version let AGNS be updated easily; besides, it features new queries, which are required for data mining, creation of new hypothesis and modeling of gene expression regulation and developmental processes in Arabidopsis.

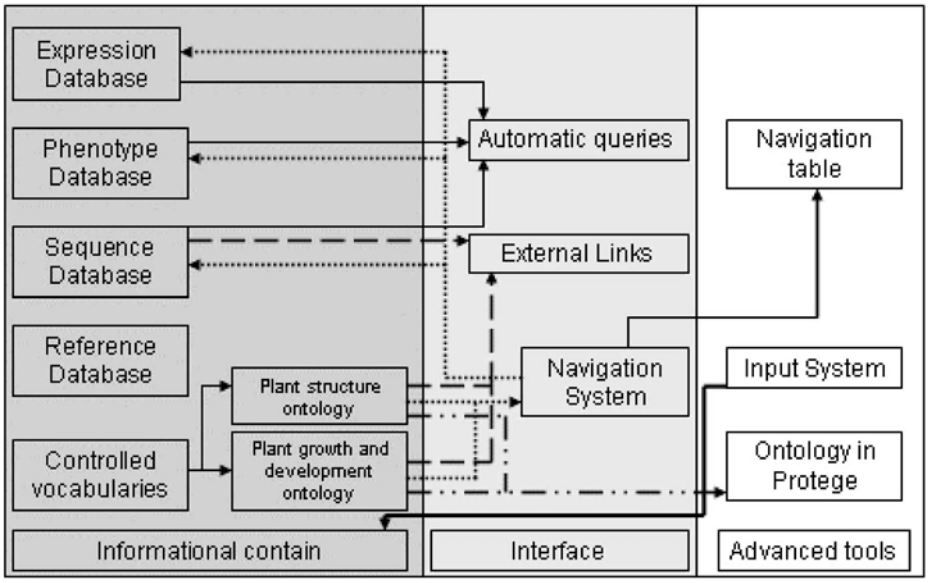

Figure 1. Structure and internal links in AGNS release 3.0.

METHODS AND ALGORITHMS

Under the AGNS project, a toolkit was developed for the modeling of new types of data, data mining and presenting the complex types of data. The software developed to serve the purpose is a set of CASE (Computer Aided Software Engineering) tools, which provide language and system support as follows:

1. a library for data format conversion between Java objects and XML documents stored in the database: resolves references to entries by their system ID, creates proxy-objects; supports separate indexation for objects in different classes.
2. a JSF (Java Server Faces)-based library for specialized GUI components.
3. a library for data management and the creation of skeleton WEB applications.

The system features the following main capabilities:

1. data storage and indexation
2. optimized advanced querying (WEB)
3. tools for creating and editing entries (WEB)
4. various navigation options (WEB)
5. an option to use custom data types for the benefit of external users (databases not linked originally to the system can, if so desired, be integrated into the system).

For efficient data storage, the system in its basic release uses Berkeley DB XML, an embedded XML database, with W3C XQuery as the basic query language.

The extensibility of the system was demonstrated on reflexive and model-based Java and XML technologies. This approach provides the maximum versatility of the system's main components and keeps it applicable to a semantically broad spectrum of tasks at the expense of a minor reduction in throughput. This approach (MDA, Model Driven Architecture) is believed to be absolutely adequate to most bioinformatics disciplines.

RESULTS AND DISCUSSION

The main problem concerned with extending and upgrading AGNS so far has been the tediousness of the annotation process. To address the problem, a software toolkit was developed (see Methods and Algorithms), which allowed us to create a specialized, "smart" data submission system for AGNS and an easy-to-use navigation system. The new interface makes AGNS upgrade rapid and more accurate.

Ontology is one of the backbones to AGNS. The AGNS ontology consists of two controlled vocabularies on plant developmental stages, anatomy and morphology. These vocabularies are used in the database format for annotation of data. The structure of the AGNS ontology is different from those of other databases (Jaiswal *et al.*, 2005): what makes it unique is that any plant organ or developmental stage can be described to the highest detail and, as a result, it is strongly extendable.

Now AGNS ontology has been brought up to international nomenclature, which will enable cross database querying with other databases on arabidopsis.

A new section, Sequence DataBase, which contains references to the nucleotide (genomic, mRNA and protein-encoding) sequences of the genes for morphogenesis and the indications of where the mutations which disrupt morphogenesis are located within these sequences; descriptions to the transgenic constructs built using these genes and transgenic modules included in transgenic constructs with the intact coding sequence of the gene (for example, another promoter or untranslated mRNA regions). Depending on what data type, one of three data formats is provided in Sequence DataBase: gene description, mutant description and transgene description. An external link to TAIR through the AGI gene code improves AGNS cross database querying capabilities.

In order to enhance the efficiency of the use of data contained in AGNS, the formats were revised in other AGNS sections, too. The strategy for describing double mutants was changed in Phenotype DataBase: if the genes interact epistatically or additively, this information is entered, in a particular format, as comments to the main gene; if the phenotype of the double mutant is not identical to those of single mutations, this piece of information is annotated separately and the double mutant is necessarily indicated in the "allele/transgene" field.

Today the most thoroughly examined in AGNS data are the phenotypes of mutant and transgenic plants and the expression of the genes in wild type, mutant and transgenic plants for the genes expressed in the apical shoot meristems and involved in miRNA biogenesis. These data were used for the reconstruction of the gene network for the functioning of the apical shoot meristem of Arabidopsis (Mironova, Omelyanchuk, 2004), for the creation of the ontology of Arabidopsis development and developmental abnormalities in the Protégé system (Ponomaryov, 2006), for the development of cell automaton models (Akberdin *et al.*, 2006) and a mathematical model of meristem functioning (Nikolaev *et al.*, 2006).

Thus, the current release has altogether with a two times higher content (Table 1), the improved format, database structure, database model, web interface, data submission system and navigation system.

Table 1. AGNS content

Releases	Expre-ssion patterns	Genes	Pheno-type abnor-malities	Mutant alleles or transgenes	Anato-mical elements	Develop-mental stages	Papers annotated
Release 2	514	44	192	526	243	169	197
Release 3	1154	103	229	759	417	230	259

The progress made allows is, in the nearest future, to rapidly increase the volume of the AGNS content, to crosslink AGNS with other databases on the Arabidopsis genome, to extensively use AGNS references for the development of computer-based systems, which analyze AGNS data, and, following these analyses, to create models for plant organ development.

ACKNOWLEDGEMENTS

This work was supported in part by Russian Federal Agency of Science and innovation (IT-CP.5/001), Russian Foundation for Basic Research (grant No. 05-07-98012), Russian Academy of Sciences (grant No. 10.4 and Program “Origin and evolution of biosphere”), Siberian Branch of Russian Academy of Sciences (Integration Project No. 115) and the US National Science Foundation (FIBR EF-0330786 Development Modeling and Bioinformatics).

REFERENCES

Akberdin I.R. *et al*. (2006) A cellular automation to model the development of shoot apical meristem of *Arabidopsis thaliana*. *This issue*.

Jaiswal P. *et al*. (2005) Plant Ontology (PO): a controlled vocabulary of plant structures and growth stages. *Comparative and Functional Genomics*, **6**, 388–397.

Mironova V.V., Omelianchuk N.A. (2004) Gene network of the Arabidopsis developing shoot meristem and its description in the GeneNet computer system. *Proceedings of the Fourth International Conference on Bioinformatics of Genome Regulation and Structure (BGRS’2004*). ICG, Novosibirsk, **2**, 101–104.

Nikolaev S.V. *et al* (2006) A one dimensional model for the regulation of the size of the renewable zone in biological tissue. *This issue.*

Omelyanchuk N. *et al.* (2006) AGNS-a database on expression of Arabidopsis genes. In Kolchanov N., Hofestaedt R., Milanesi L., (eds), *Bioinformatics of Genome Regulation and Structure II*. Springer Science+Business Media, Inc. 433–442.

Ponomaryov D.K. *et al.* (2006) Semantically rich ontology of anatomical structure and development for *Arabidopsis thaliana*. *This issue*.