SEMANTICALLY RICH ONTOLOGY OF ANATOMICAL STRUCTURE AND DEVELOPMENT FOR *ARABIDOPSIS THALIANA* L.

Ponomaryov D.*[*1], *Omelianchuk N.*[2], *Kolchanov N.*[2], *Mjolsness E.*[3], *Meyerowitz E.*[4]

[1] Institute of Informatics Systems, Novosibirsk, 630090, Russia; [2] Institute of Cytology and Genetics, SB RAS, 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: ponomaryov@ngs.ru

Key words: plant, anatomy, morphology, development, phenotype, ontology

SUMMARY

Motivation: An ontology of a subject domain is a set of statements that are true for every possible situation in this subject domain. The aim of our work is a formal ontological description of anatomical structures in development of *Arabidopsis thaliana* wild-type plant, which serves as the target subject domain. The reason for developing this type of ontology in presence of TAIR and Plant Ontology was to support analysis and processing of expression and phenotypic data from AGNS database. For this task, it was necessary to capture the fact that the hierarchy of anatomical elements changes within the developmental process. This means that we should have been able to express not only development of one structure into another, but also changes in the number of contained entities, as well as containers for anatomy elements.

Results: We have developed an ontology that consists of a core structure for representing statements of the kind "*anatomy element X at stage Sx exists in anatomy element Y at stage Sy*" (where Sx and Sy are development stages of X and Y respectively) and is filled in with facts from Arabidopsis development studies, which have been extracted from publicly available articles. In general, the problem of describing anatomical structures in development (without taking spatial orientation into account) was reduced to ordering and inclusion of developmental stages.

Availability: The ontology was built as a wrapper around the AGNS database, which is available at http://wwwmgs2.bionet.nsc.ru/agns/

INTRODUCTION

AGNS (Arabidopsis GeneNet Supplementary database) is an Internet-available resource that provides access to description of the functions of the known Arabidopsis genes at various levels—the levels of mRNA, protein, cell, tissue, and ultimately at the levels of organs and the organism in both wild type and mutant backgrounds (Omelianchuk *et al.*, 2006). At present, the data are collected, assembled, and curated using the formats developed in two AGNS modules: expression, and phenotype databases. In addition, while annotating, it appeared necessary to built two controlled vocabularies around contents of annotated data and an ontology of development of the cell types, tissues, organs based on the data presented by authors of publications on gene expression patterns and mutant phenotypes and by the publications describing developmental processes. While controlled vocabularies were developed for paper curators, the aim in constructing ontology was to apply it in algorithms of analysis of expression and phenotypic data. The benefits of use of formal ontology in processing of experimental data are recognized in bioinformatics (Karp, 2000; Bard, 2004). In this paper, we explain the results of formalizing plant anatomy in development by example of navigating in the AGNS database. We believe this use case to be rather simple, yet very illustrative for our work. We show that by having a concrete practical task to be solved, the choice of formalization becomes well-founded and easier to evaluate, than in some abstract case.

METHODS AND ALGORITHMS

We formulate the task of navigating the database via ontology as follows:

- query the database in ontological terms (concepts);
- use relations between ontological terms to broaden/narrow data extraction from the database.

In practice, AGNS consists of two databases: ED (expression database) and PD (phenotype database) and the AGNS facts, which we denote by predicates with corresponding names *ED* and *PD*, have the following form:

$ED(Gene, Anatomy\_Element, Stage, Express\_Level, IsAbnormal)$,

$PD(Phenotype\_ID, Anatomy\_Element, Stage, Abnormality)$.

Here *Anatomy_Element* is an organ, tissue or cell, *Stage* is a developmental stage of *Anatomy_Element* and *Abnormality* is a textual name for the phenotypic abnormality. Our first step is to navigate in the expression database ED. A typical example of navigation in this case would be like this: if we are interested in expression of gene *AG* in some organ *X*, then obviously we would like to know its expression in sub-organs or tissues of *X*. Or we would like to restrict ourselves to only certain parts of *X* or certain development stages of *X*. This implies that the ontology should have the necessary relations between anatomy elements and developmental stages. It turns out that in order to formulate queries and manage extraction of data from the database, one needs to query the ontology itself. In our case, two typical queries, the answers which ontology should provide are:

(Q1) X is an anatomy element, find all elements Y, belonging to X

(Q2) S is a developmental stage, find all stages earlier/later, than S.

The main problem in representing plant anatomy in development may be formulated as follows: if *X*, *Y* are two anatomy elements, then *X* belongs to *Y* at developmental stage S_i does not necessary imply that *X* belongs to *Y* at another stage S_k. The number of organs in a plant may change within development and moreover, anatomy elements may have different direct containers (i.e. elements to which it directly belongs) at different stages. With respect to the task of navigation, this directly affects what anatomy elements are considered and what expression data are extracted for a queried plant/organ/tissue developmental stage. From the facts given above one can notice that the relation "belongs to" is temporal. Fortunately, Arabidopsis development is well studied, "what-where-when" is known, and stages are discrete. Eventually, the information we would like to be expressed in the ontology are statements of the kind:

"*Anatomy element X at stage S_X exists in anatomy element Y at stage S_Y*" (I)

We only need to represent this concept by a combination of suitable binary predicates.

From a theoretical point of view, we consider formal ontology as a set of sentences in First Order Logic language. But for practical purposes we restrict ourselves to a decidable fragment of FOL and to formulas of a special kind, which will be reflected by the choice of the OWL-DL language for implementation. In a formal ontology we distinguish a signature that is a set of predicate, functional, constant symbols and axioms that restrict possible interpretations of these symbols. Throughout this paper, we will simultaneously use names for binary predicates and the term "relation" to denote the same things.

The core elements in the ontology are the predicates listed below and axioms that will be introduced further in this section.

$\mathit{Anatomy\_Element}^1$
$\mathit{Development\_Stage}^1$
$\mathit{Has\_Development\_Stage}^2$ *(Anatomy_Element xDevelopment_Stage)* (II)
Before^2 *(Development_Stage xDevelopment_Stage)*
$\mathit{Occurs\_In}^2$ *(Development_Stage xDevelopment_Stage)*

We assume that binary predicates are defined on Cartesian products of those sets, that are defined by unary ones and are mentioned informally in the parenthesis. Let us denote the statement (I) by predicate $\mathit{ExistsIn}(X,S_X,Y,S_Y)$, i.e. the predicate is true, whenever statement (I) holds for X, S_X,Y, S_Y. We define this predicate by the following combination of binary predicates that were introduced above:

$$\mathit{ExistsIn}\,(X,S_X,Y,S_Y) \leftrightarrow$$
$$\mathit{Has\_Developmen\ t\_Stage}\,(X,S_X)\ \&\ \mathit{Has\_Developmen\ t\_Stage}\,(Y,S_Y)\ \&\ \mathit{Occurs\ \_In}(S_X,S_Y)$$

An example for Arabidopsis development is the case, when Y is Leaf, S_Y is LDS4 (the fourth leaf development stage), X is midrib and S_X is MDS1 (the first midrib development stage). One should notice that there are no direct relations between anatomy elements. Instead, the relation *Occurs_In* serves for inclusion of anatomy elements into each other. The ontology also includes axioms restricting interpretation of the predicates introduced above. We do not mention them here due to paper size limitations. Core concepts defined by unary predicates in (II) together with relations defined by the mentioned binary predicates and these axioms make up a formal ontology, which we consider as an *ontology for anatomical structure and development of plants.*

The facts that help to evaluate the proposed ontology originate from the experience of its instantiation with real data from Arabidopsis development studies. Textual information about Arabidopsis anatomy and development was extracted from 100+ publicly available articles and entered into the ontology in the form defined by formal constructs from the previous section. During this process we have encountered two main problems regarding incompleteness of information, namely:

1. stages were not defined for all anatomy elements we would like to include in the ontology;
2. development of lots of anatomy elements was described in the manner "*X has the following properties, when Y is at stage S*" (e.g. "*LDS2 leaf shows meristematic divisions throughout the mesophyll*"), but these descriptions were given only for some stages of Y.

This has lead to the need of creation of "artificial" development stages for anatomy elements. Even though such stages may not be distinguished in reality, we need to have them in the ontology to provide a proper unfolding of organ anatomy structure. On the other hand, they also allowed for a more detailed description of developmental processes of Arabidopsis in comparison with the well-known of TAIR ontology.

IMPLEMENTATION AND RESULTS

In the implementation we were guided by two objectives: to have a widely supported expressive ontology language with full reasoning capabilities and to use an ontology editor with rich import-export and visualization functions. This resulted in the choice of the *OWL* language and *Protégé* ontology editor (version 3.1, http://protege.stanford.edu/). The problem of representing rule-like axioms was not considered as a significant one, because the number of such axioms in our ontology is small. Thus, it is possible to rewrite them manually into the language of the needed inference engine. In our case we used the *Algernon* engine for *Protégé.* A part of the axioms was implemented as constraints, other axioms – as forward or backward chaining rules. The notion of transitivity is present in

OWL as the *TransitiveProperty* construct, but as far as *OWL* is not yet fully supported by existing reasoners, we decided to implement transitivity axioms both in *OWL* and in the inference engine language.

DISCUSSION

We introduced a formal framework for representing plant anatomy in development by the example of an ontology for *Arabidopsis thaliana* (L.) Heynh. The proposed formalization was developed with orientation to the task of navigating in a gene expression database for the plant. The main problem of representing plant anatomy in development was reduced to ordering and inclusion of developmental stages. The built ontology has a small core, yet with strong expressive capabilities. It presents a richer language for describing anatomical structures in development in comparison to the known controlled vocabularies. This is the reason we call this ontology *semantically rich*. It became possible to define *Develops_From* relation for anatomy elements not as a separate one, but on top of core structures of the ontology. Information about Arabidopsis development taken from 100+ publicly available articles was presented in a formal way. Each stage in the ontology was provided with textual description with a reference to the source article. The implemented ontology is a deductive database consisting of facts about plant development and inference rules.

ACKNOWLEDGEMENTS

This work was supported in part by RF Ministry of Education grant “The development of the higher education scientific potential” (project No. 8328), Russian Foundation for Basic Research (grants No. 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

Bard J.B., Rhee S.Y. (2004) Ontologies in biology: design, applications and future challenges. *Nat. Rev. Genet.*, **5**, 213–222.

Karp P.D. (2000) An ontology for biological function based on molecular interactions. *Bioinformatics*, **16**, 269–285.

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.