A PROGRAM METHOD OF CONSTRUCTING ONTOLOGY OF PHENOTYPIC ABNORMALITIES FOR *ARABIDOPSIS THALIANA*

***Ponomaryov D.*[*1], *Omelianchuk N.*[2], *Mironova V.*[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: development, phenotype, computer analysis, algorithms, ontology

SUMMARY

Motivation: Modern researches prove that studying gene networks and development of organisms at a higher level of abstraction allows for a better understanding of mechanisms of developmental processes and their interactions. We develop a classification of phenotypic abnormalities of Arabidopsis to distinguish and prove the existence of specific functional modules in the plant, to identify key points of abnormal development and to find parallel regulatory pathways leading to abnormalities.

Results: We have developed an algorithm that takes a description of phenotypic abnormality caused by a mutation as an input and generates a graph of relations of this abnormality to other abnormalities of mutant and transgenic phenotypes, basing on data from AGNS database. As *ontology* is usually viewed as a set of terms and relations between them, we call this graph an ontology of phenotypic abnormalities of Arabidopsis.

Availability: Phenotypic data used by the algorithm are available at http://wwwmgs2.bionet.nsc.ru/agns/

INTRODUCTION

AGNS database (Omelianchuk *et al*., 2006) has a AGNS_PD module that contains information about phenotypes of Arabidopsis in the form of statements: $PD(Phenotype\_ID, Anatomy\_Element, Stage, Abnormality)$, where *Phenotype_ID* is a name for allele or transgene, *Anatomy_Element* is a name for organ, tissue or cell, *Stage* is a name for developmental stage of the *Anatomy_Element* and *Abnormality* is a name for phenotypic abnormality. For instance, the fact $PD(CLV1-1, Floral\_Meristem, FDS3, Enlarged)$ states that the floral meristem at *FDS3* developmental stage is enlarged in mutant plants homozygous for the clv*1-1* allele. Facts of this sort have been extracted from publicly available papers describing separate experimental results on mutant and transgenic plants of *Arabidopsis thaliana* (Omelianchuk, 2006). In our work, we found necessary to summarize these data in order to identify the same abnormalities that are described differently in different research groups and to build a general classification of phenotypic abnormalities of Arabidopsis. By this we aim to distinguish and prove the existence of specific functional modules in the plant, to identify key points of abnormal development and to find parallel regulatory

pathways leading to abnormalities (Ponomaryov, Omelianchuk, 2006). This in turn could allow for a better understanding of mechanisms of developmental processes and their interactions and would be a step to reconstructing the underlying gene networks. Modern researches (Gunsalus *et al*., 2005; Roy, Morris, 2005) show that studying this problem from a higher level of abstraction can benefit in lots of cases, where analysis of too detailed data could not finally lead to sound models.

METHODS AND ALGORITHMS

Let us consider four sets A, E, S and T, where

1. A is a set of phenotypes (alleles and transgenes) of Arabidopsis.
2. $E-$ is a partially ordered set of anatomical elements (cells, tissues, organs and structural elements, such as specially distinguished layers or zones; the whole plant is also considered as anatomical element), with the order $\rhd$, defined in the following way: (for all $e_1, e_2 \in E$)($e_1 \rhd e_2$, if and only if e_2 develops from e_1).
3. S – is a partially ordered set of developmental stages of anatomical elements with the order $>$ defined as follows: (for all $s_1, s_2 \in S$) ($s_2 > s_1$, if and only if s_1 is before s_2 in time).
4. T is a set of types of phenotypic abnormalities (*abnormal position, abnormal shape, delayed development, increased number*, etc.).

Define a relation $R \subseteq E \times S$ with the property (for all $e \in E$ and $s \in S$) ($(e,s) \in R$, if and only if s is a developmental stage of e). Introduce also a relation $\succ \subseteq E \times S \times E \times S$ as follows: (for all $e_1, e_2 \in E$ and $s_1, s_2 \in S$) ($(e_1, s_1, e_2, s_2) \in \succ$, if and only if e_1 at the developmental stage s_1 exists in e_2, when e_2 undergoes stage s_2). Note that under several additional assumptions, the structure $< E \cup S, \rhd, >, \succ >$ can be considered as a model for the logical theory described in (Ponomaryov, Omelianchuk 2006).

Following the level of abstraction, at which abnormalities are presented in the AGNS_PD module, we define a phenotypic abnormality as a 4-tuple: $N =< G, e, s, t >$, where $G \subseteq A$, $e \in E$, $s \in S$, $t \in T$. AGNS_PD consists precisely of a collection of such 4-tuples with assigned textual names. These names represent short characterizations of abnormalities that have been extracted from papers describing separate experimental results on mutant and transgenic plants of Arabidopsis. It follows immediately from our definition of abnormality that two different names denote the same abnormality in the AGNS_PD, if they correspond to the same 4-tuples. Note that this is not the only rule to identify the same abnormalities.

Besides the task of name disambiguation, we also distinguish six relations between abnormalities that we aim to extract by analyzing AGNS data. Here we only list their names and give a brief informal explanation.

We have developed an algorithm that takes a representation of abnormality in the form of a tuple $N =< G, e, s, t >$ as an input and outputs a graph with vertices referring to abnormalities, which are connected by edges that correspond to the relations above. In other words, vertices in the resulting graph correspond to the tuples from the AGNS_PD and are labeled with textual descriptions of abnormalities, while edges are labeled with the names of the relations. In particular, those vertices that have no incoming edges labeled with "Consequence_of" are considered to be candidates for initial points of abnormality development and are checked against AGNS gene expression data by the algorithm. To resolve ambiguous cases when AGNS data is insufficient, the algorithm

uses additional information about known gene functions, as well as information about known gene interactions, which is available in a separate database.

Table 1. Relations between phenotypic abnormalities – informal definitions

Name of relation between abnormalities	Informal definition by an example
Blocked_by	One abnormality is blocked by another one, if it is not present, when the other abnormality is observed.
Consequence_of	An abnormality is a consequence of another one, if it is the result of development of this abnormality within time.
Specialization_of	An abnormality, to which another one is a specialization, is a stronger abnormality.
Inverse_to	Two abnormalities are inverse to each other, if they are presented by opposite phenotypic changes.
Composite_of	An abnormality is a composite of several others, if it is caused by all of them together.
Alternative_to	One abnormality occurs in another percentage of cases, in comparison to that of the second one.

For developing the algorithm it was necessary to give an explicit formal definition to all of the listed relations. As a result, we defined a set of rules, which are the necessary conditions defining these relations. These rules were produced by analysis of all possible cases of structural differences between two arbitrary abnormalities in the form: $N_1 =< G_1, e_1, s_1, t_1 >$, $N_2 =< G_2, e_2, s_2, t_2 >$. Clearly, there is a restricted number of cases how these abnormalities (defined as tuples) can differ from each other. They originate from considering all possible set-theoretic relations between the sets G_1, G_2, as well as relations $\rhd, >, \succ$ pair-wise between the elements e_i, s_i, t_i, $i = 1,2$. Due to paper size limitations, we list below only some of the rules used in our algorithm.

Table 2. Some of the rules for inferring relations between abnormalities

Premise	Conclusion
$G_1 = G_2$ and $s_1 < s_2$ and $e_1 = e_2$ and $t_1 = t_2$	Abnormality $N_2 =< G_2, e_2, s_2, t_2 >$ is a consequence of abnormality $N_1 =< G_1, e_1, s_1, t_1 >$.
$G_1 = G_2$ and $s_1 = s_2$ and $e_1 = e_2$ and $t_1 \neq t_2$	N_1 and N_2 are alternative abnormalities to each other.
$G_1 = G_2 = G_3$ and $s_1 = s_2 = s_3$ and $(e_1, s_1, e_3, s_3) \in \succ$ and $(e_2, s_2, e_3, s_3) \in \succ$	N_3 is a composite of N_1 and N_2.

IMPLEMENTATION AND RESULTS

We have introduced an axiomatic semantics for the relations considered above by a system of axioms restricting their interpretation. On one hand, it has allowed for constructing a declarative language for description of abnormalities and relations between them. On the other hand, we have used these axioms to define post-conditions for implementation of our algorithm. By this we have proved a total correctness of the implemented algorithm using the Floyd's method of program proving (the preconditions have been taken according to the formal definition of abnormality in our work).

DISCUSSION

Presently, we are evaluating the developed algorithm on different datasets from the AGNS_PD module of the AGNS database. In particular, we use phenotypic information regarding only certain organs or certain periods of development. The aim is to estimate the adequacy of the output results of the algorithm by checking, whether the inferred relations are correct and also whether theoretically predicted modules are present. As adequacy of the resulting graph potentially depends on the amount of data available for processing, such testing process also helps to discover points of potentially incomplete information in the AGNS database.

ACKNOWLEDGEMENTS

This work was supported by 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

Gunsalus K.C., Ge H. *et al*. (2005) Predictive models of molecular machines involved in caenorhabditis elegans early embryogenesis. *Nature*, **436**, 861–865.

Omelianchuk N.A. *et al*. (2006) AGNS – A Database on Expression of Arabidopsis Genes. Bioinformatics of Genome Regulation and Structure II, Springer Verlag, 433–442.

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

Roy P.J., Morris Q. (2005) Network news: functional modules revealed during early embryogenesis in *C. elegans*. *Developmental Cell*, **9**, 307–315.