AGRICULTURAL BIOLOGY,
2017, V. 52, № 1, pp. 63-74
(SEL'SKOKHOZYAISTVENNAYA BIOLOGIYA)

ISSN 2412-0324 (English ed. Online)
ISSN 0131-6397 (Russian ed. Print)
ISSN 2313-4836 (Russian ed. Online)

UDC 633.491:004.65:[631.5+632.9

doi: 10.15389/agrobiology.2017.1.63rus
doi: 10.15389/agrobiology.2017.1.63eng

DEVELOPMENT OF METHODS FOR AUTOMATIC EXTRACTION OF KNOWLEDGE FROM TEXTS OF SCIENTIFIC PUBLICATIONS FOR THE CREATION OF A KNOWLEDGE BASE SOLANUM TUBEROSUM

O.V. SAIK[1], P.S. DEMENKOV[1, 2], T.V. IVANISENKO[1], N.A. KOLCHANOV[1], V.A. IVANISENKO[1]

[1] *Federal Research Center Institute of Cytology and Genetics SB RAS,* Federal Agency of Scientific Organizations, 10, prosp. Akademika Lavrent'eva, Novosibirsk, 630090 Russia, e-mail saik@bionet.nsc.ru, demps@bio-net.nsc.ru, itv@bionet.nsc.ru, kol@bionet.nsc.ru, salix@bionet.nsc.ru;
[2] *Novosibirsk State University,* 2, ul. Pirogova, Novosibirsk, 630090 Russia
ORCID:
Demenkov P.S. orcid.org 0000-0001-9433-8341
Ivanisenko T.V. orcid.org 0000-0002-0005-9155
Kolchanov N.A. orcid.org 0000-0001-6800-8787
Ivanisenko V.A. orcid.org 0000-0002-1859-4631

The authors declare no conflict of interests

Acknowledgements:
Supported by Budget Project of Institute of Cytology and Genetics SB RAS for Potato Program
Received November 30, 2016

Abstract

Currently there are hundreds of scientific journals that publish research results in various fields of plant biology and agrobiology. Hundreds of thousands of international patents contain a variety of information on agricultural biotechnology. The number of articles and patents is increasing over time in an exponential progression. For example, there are more than 1.5 million publications devoted to the study of *Solanum tuberosum* that is one of the most important crops in the world. Analysis of such huge number of experimental facts presented in text sources (scientific publications and patents), requires the use of automated methods for knowledge extraction (text-mining). Intelligent automatic text analysis techniques are already widely used in biology and medicine to extract information about the properties and functions of molecular genetic objects. Unlike search engines such as Google, Yandex and others, that search documents by keywords, such text-mining methods are aimed at the automatic extraction of knowledge presented in the documents, knowledge integration and formalization according to the defined ontology. Among the known systems for intelligent knowledge extraction from scientific publications STRING, LMMA, ConReg, GeneMania and others can be listed. For the first time in Russia, we have previously developed a system, named ANDSystem, for automatic intelligent knowledge extraction in biomedicine. ANDSystem contains more than 10 million facts about molecular-genetic interactions extracted from more than 25 million scientific publications. For knowledge extraction in ANDSystem, specially developed semantic and linguistic rules are used for recognition of interactions between biological objects such as, proteins, genes, metabolites, drugs, microRNA, biological processes, diseases and others in natural language texts. However, the problem of development of methods for automatic knowledge extraction from the texts in plant biology, agrobiology and agrobiotechnology remains still unsolved and has a high relevance. The aim of this work was to adapt the methods of automatic knowledge extraction, presented in ANDSystem, to the field of crop production and to create on this basis a SOLANUM TUBEROSUM knowledge base, containing information on genetics, markers, breeding and selection of potatoes, its pathogens and pests, storage and processing technologies and others. The knowledge base ontology contains dictionaries, corresponding to more than 20 types of objects, including molecular genetic objects (proteins, genes, metabolites, microRNA, biological processes, biomarkers, etc.), potato varieties and their phenotypic traits, diseases and pests of potato, biotic and abiotic environmental factors, biotechnologies of cultivation, processing and storage of potato, and others. Also, the ontology contains more than 25 types of interactions that describe various relationships between the above listed objects, including molecular interactions, regulatory events and associative links. More than 5 thousand semantic templates were created to extract information about the interactions. The accuracy and recall of knowledge extraction by the developed method were assessed with the expert manual analysis of the text corpus and reached more than 65 % and 70 %, respectively. The full-scale version of the knowledge base SOLANUM TUBEROSUM will be created on the basis of the developed approaches.

Keywords: *Solanum tuberosum,* ANDSystem, text-mining, database, automatic knowledge extraction from texts

Currently, investigation of molecular genetic systems becomes top priority in genomic, proteomic, metabonomic and transcriptomic studies in different fields of biology, including plant growing [1-4]. New approaches take on special significance to the investigation of the genotype-phenotype relationship. Reconstruction and analysis of gene networks replace traditional approaches based on the search for individual genes responsible for the formation of phenotypic plant characteristics, including complex economically valuable features, such as resistance to diseases and pests, tolerance to abiotic factors, and yield [5-8]. The most important source of information on molecular genetic interactions occurring at the intracellular, intercellular and organismic levels of plant organization are databases that summarize the exploratory results, scientific publications and patents. The number of publications increases exponentially each year. Even a simple search query by the keyword ‘potato’ in the Web of Science and Google Patents, yields information on more than 60,000 papers and 900,000 patents. Many of them contain data on molecular genetic interactions. The prompt accumulation of new knowledge presented in scientific publications and databases is significantly associated with the development of experimental high-performance 'Omic' technology (genomic, transcriptomic, proteomic and metabolic). The use of high-performance sequencing technology allowed reading the genome of the potato.

The NCBI Genomes database (https://www.ncbi.nlm.nih.gov/genome) contains a version of the potato genome GCA_000226075.1 SolTub_3.0 [9, 10], which includes an annotation of 37,966 proteins. Another database, NCBI Gene (https://www.ncbi.nlm.nih.gov/gene), provides information on the sequences and functions of 33,037 potato genes.

To establish interactions between biological objects, experimental methods of direct analysis of protein-protein interactions (yeast two-hybrid systems), transcriptomic analysis (differential gene expression and co-expression) and others are often used. The GEO database (https://www.ncbi.nlm.nih.gov/gds) presents data from more than 1,300 experiments on the expression of potato genes obtained using transcriptomic technology. For example, Y. Ou et al. [11] in their paper presented in GEO (GSE43237) performed a full-genomic analysis of microRNA targets in tubers stored in the cold. The analysis identified 53 known and 59 new miRNAs, as well as 70 target genes potentially involved in the response to low-temperature storage. Another paper [12], also presented in the GEO database (GSE56333), investigated the effects of the potato Y-virus infection on potato resistance to the Colorado potato beetle larvae using high-performance sequencing.

Databases containing information on molecular genetic interactions obtained based on the analysis of factual databases and scientific publications are being rapidly developed worldwide. In particular, the PlantCyc database [13-16] contains information on molecular genetic networks for more than 22 plant species, including potatoes. The PotatoCyc database (the potato section in PlantCyc) contains information on 558 biological pathways, 5,790 enzymes, 3,122 reactions and 2,413 metabolites. However, this database was created based on the non-automated analysis of scientific publications, which guarantees the yield of high quality data, but inevitably leads to a delayed presentation of facts published in scientific articles.

The current number of publications and patents is the so-called ‘big data’ (extremely large amounts of data), the effective processing of which requires the use of automated text analysis (text mining). The technology of automatic extraction of knowledge from scientific publications is most rapidly developed in biomedicine [17-21]. Among the widely used systems for text analysis on the specified subjects, STRING [22-24], LMMA [25] and ConReg [26] can be distinguished. The STRING system includes descriptions of protein-protein interactions that are confirmed experimentally or predicted by various methods (including predictions of the distance of genes in the genome, phylogenetic profiles, or co-expression). The system STRING uses information extracted from databases, as well as obtained from publications using methods of automated text analysis. The LMMA system is designed for the reconstruction of biological networks based on the integration of literature data on molecular interactions and evidence on gene expression obtained in microchip experiments. It is based on estimates of the statistical significance of the concurrent occurrence of biological terms in texts from PubMed. ConReg is a plug-in for the Cytoscape system [27] that focuses on the study of genetic regulation in eukaryotic model organisms. Here the evidence on genetic regulation are taken from different databases and supplemented with information about the predicted binding sites for transcription factors, as well as information extracted using the automated analysis of PubMed texts.

We have previously developed the ANDSystem for the automated extraction of medical and biological knowledge from the PubMed texts using semantic template methods [28-30]. ANDSystem includes a linguistic analysis module that automatically extracts from an arbitrary text flow the factual information related to a specific subject (problem) domain according to a given ontology. The module of linguistic analysis consists of three main parts, such as a morphological analyzer, a problem-oriented ontology and a semantic analyzer. The morphological analyzer implements the following functions: descriptive text markup (recognition in the text the concepts included in the ontology, including terminological word combinations); lemmatization; and POS marking. A problem-oriented ontology forms a conceptual model of the problem domain. The semantic analyzer implements the functions of conceptual search in the text of the document and the user interface. The system operation is provided by two main dictionaries: the grammar dictionary supports lemmatization, POS markup and recognition of word combinations based on the linear context; the ontology supports semantic analysis, including elements of limited logic output. In addition, a defining dictionary is used (a word or a word combination as a concept), integrated into an ontology.

In the present work, the methods of the ANDSystem were adapted and adjusted for automatic extraction of the knowledge on genetics, markers, ‘omic’ resources, breeding, seed production, diagnostics of disease pathogens, protective means and potato storage technologies in order to create a knowledge base called SOLANUM TUBEROSUM. Setting up ANDSystem involved the creation of a subject domain ontology and semantic linguistic rules (templates) for analyzing natural language texts and extracting knowledge formalized according to a given ontology. Important components of the subject domain ontology are dictionaries of the objects, the information about the interactions between which is extracted from texts using templates. The developed ontology of the SOLANUM TUBEROSUM knowledge base contains dictionaries for more than 20 types of objects. Proteins, genes, metabolites, microRNAs, biological processes, biomarkers, etc. are considered as the molecular genetic objects. Separate dictionaries are potato varieties and their phenotypic signs, including potato diseases. A large section of ontology is devoted to potato pests, as well as to environmental biotic and abiotic factors. The ontology also contains dictionaries of cultivating agrobiotechnology, and biotechnology of potato processing and storage. The analysis of quality of the knowledge extraction using created templates demonstrated appropriate accuracy (65 %) and completeness (70 %).

Ontology-based model. We used the term ontology to mean the

$O = <C, R, F>$ set, where $C = C_t + C_o + C_{sp} + C_{ti} + C_i + C_r$ is the set of concepts of the subject domain (Fig. 1), represented by the following components: C_t is the set of object types, C_o is the set of molecular genetic objects, diseases, processes, cellular components, etc., C_{sp} is the set of organisms, C_{ti} is the set of types of interrelations between objects, C_i is the set of interrelations between objects, C_r is the set of object roles in the interrelations. $R = \{is\_a, role, present, exists\_in\} + R_1$ describes the set of relations between the concepts of a given domain and, in turn, consists of subsets describing the different types of relationships between objects, the admitted roles of objects in relationships of a particular type, the relation linking molecular genetic objects with the organisms in which they occur, etc. The third component is represented by the set $F = \{f\colon R_1 \text{ -> } C_i\}$, describing the interpretation function, which consists of the one-to-one mapping of the set of the R_1 relations onto the set of types of objects C_i.

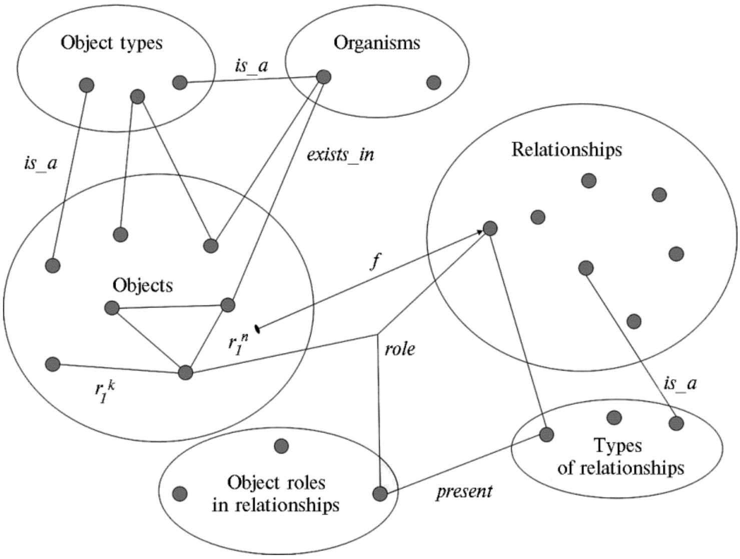


Fig. 1. A graphical presentation of the ontology of associative semantic networks, used to design the SOLANUM TUBEROSUM knowledge base.

The knowledge base structure. Using the developed ontological model of data representation, the SOLANUM TUBEROSUM knowledge base was designed. It includes a database containing molecular genetic information, information about technologies, diseases, environmental factors obtained as a result of the analysis of texts of scientific publications, patents and databases. In addition, the knowledge base contains methods that were used to extract knowledge from texts, and methods designed to analyze the molecular genetic networks available in the database. The MySQL 5.6 relational DBMS was used when developing the knowledge base. The database contains 18 tables describing the following sections: Plant, Potato Pathogens and Pests, Environment, Technology, Bioinformatics, Associative networks (see Fig. 2).

The Plant Section. The Plant section is intended to describe molecular genetic data. In the current version of the knowledge base, this section provides information on potatoes supplemented with information on seven model plants (*Solanum lycopersicum*, *Nicotiana tabacum*, *Arabidopsis thaliana*, *Oryza sativa* Indica Group, *Oryza sativa* Japonica Group, *Zea mays*, *Triticum aestivum*). Molecular genetic data include dictionaries of names and their synonyms for genes (> 140,000 terms), proteins (> 19,000 terms), metabolites (> 42,594 terms), microRNAs (> 10,000 terms), genetic biomarkers (> 20 terms) and biological processes (> 100,000 terms). Separate dictionaries represent potato varieties (206 varieties), selection qualities, economically valuable characteristics and consumer attributes (> 1,300 terms). Specialized dictionaries have been developed that describe more than 100 physiological (phenotypic) signs of potatoes and diseases.

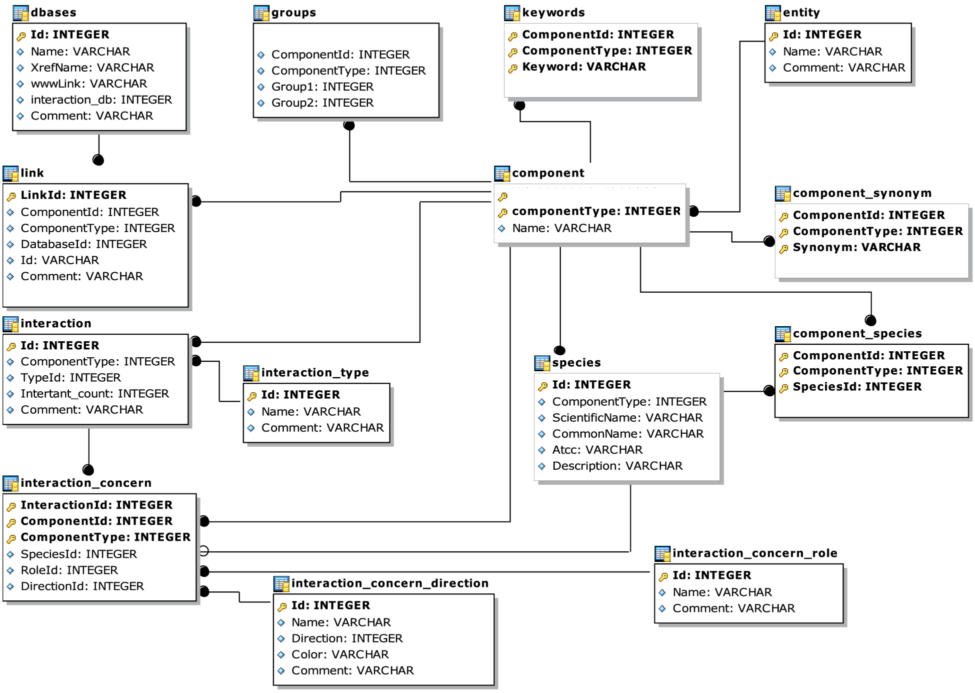


Fig. 2. The structure of the major tables of the relational database in the developed SOLANUM TUBEROSUM knowledge base.

The Potato Pathogens and Pests Section. This contains dictionaries of molecular genetic objects for 24 pathogens and pests of potatoes. Molecular genetic data, similar to the Plant section, are represented by genes (3,451 genes), proteins (476 proteins), metabolites and biological processes. Separate dictionaries describe the markers of resistance to plant protection products, as well as molecular targets for chemical plant protection products.

The Environment Section. This involves the dictionaries for two types of objects, such as biotic and abiotic environmental factors (> 100 and > 50 terms, respectively).

The Technology Section. It should be noted that, along with molecular-genetic objects and environmental factors, various technologies for selection, cultivation, protection and diagnostics of potato diseases, potato processing and storage are presented as independent objects in the developed knowledge base. More than 100 technologies are described in the current version of the knowledge base.

The Associative networks Section. An associative semantic network was used as an informational model of the subject domain, which was in the form of an oriented bipartite graph, the state points of which corresponded to the objects of the domain, and arcs defined relations between them. The relations of the following types are used to describe interactions between molecular genetic objects: 1st type, physical interactions, i.e. the formation of short-lived or permanent molecular complexes; 2nd type, chemical interactions (catalytic reactions and processes) of the substrate-enzyme-product type, in which proteins (enzymes) and low-molecular compounds (metabolites) are involved, including proteolytic cleavage reactions of one protein (substrate) by another protein (proteolytic enzyme), post-translational modifications of proteins (phosphorylation, glycosylation, etc.); 3rd type, regulatory interactions, including regulation of gene expression by transcription factors, regulation of protein activity or function by other proteins, regulation (or implementation) of transport of some proteins by other proteins, regulation of stability or degradation of some proteins by other proteins or metabolites (regulatory events will also be subdivided according to the effect one object exerts onto another one, i.e. the enhancement or weakening of the process); 4th type, co-expression (simultaneous expression of several genes), which was caused by shared regulatory mechanisms that activate expression under varying conditions in the cell; 5th type, associative connections (this category includes unclassified relations between molecular genetic objects, as well as links between molecular genetic objects and objects that match the concepts of breeding, phenomics and seed production, phytopathology, diagnostics, means of protection, agrobiotechnology of cultivation and biotechnology of potato processing and storage). The relations between the concepts of breeding, phenomics and seed production, diseases, diagnostic techniques and means of protection, technologies are based on various types of regulatory links (upregulation and downregulation), as well as links describing the involvement, application, associative links, etc.

The Bioinformatics Section. A specially developed section is closely associated with the knowledge base, which presents bioinformatic methods for analyzing experimental data on molecular mechanisms of functioning of the analyzed biological systems, gene prioritization, prediction of markers, planning experiments, etc. The analysis is performed based on the experimental findings entered by the user, as well as data on the network of molecular genetic interactions automatically extracted from the Associative Networks section.

Currently, bioinformatic analysis of experimental data is actively developing, related to the task of prioritizing the genes when identifying among them those the most important for the studied biological processes (including the response to biotic and abiotic environmental factors), phenotypic (physiological) features, diseases, etc. [31, 32].

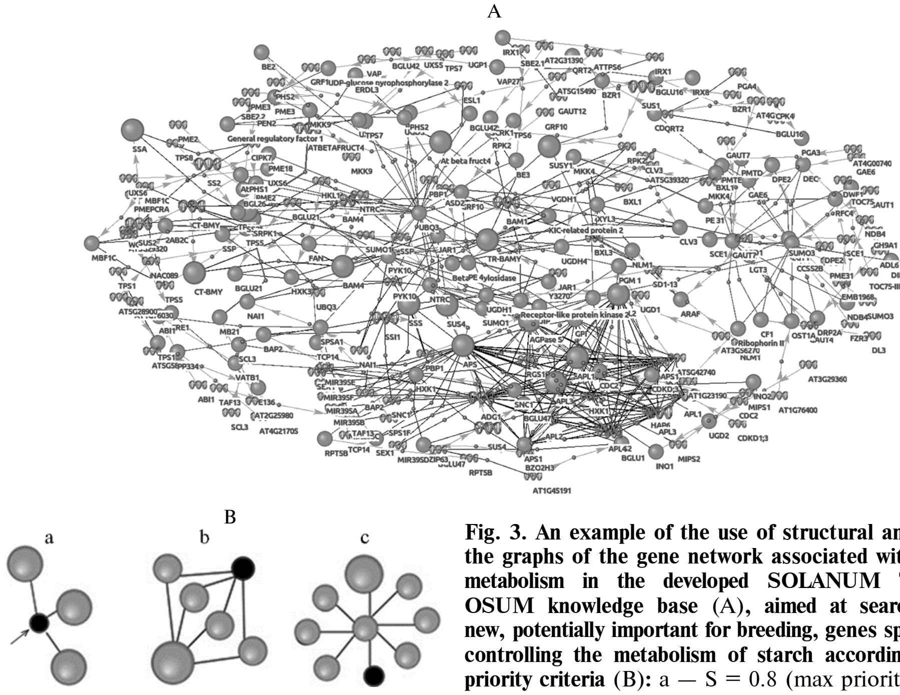


Fig. 3. An example of the use of structural analysis of the graphs of the gene network associated with starch metabolism in the developed SOLANUM TUBEROSUM knowledge base (A), **aimed at searching for new, potentially important for breeding, genes specifically controlling the metabolism of starch according to the priority criteria** (B): a — S = 0.8 (max priority), the

candidate gene (marked by a black circle) interacts directly with three known key participants in starch metabolism (reference genes, represented by larger circles); b — S = 0.0346, the candidate gene is linked to the reference gene through four proxy genes; c — S = 0.0115 (min priority), the candidate gene is linked to the reference gene through a hub. The identified candidate gene with the highest priority is potato starch branching enzyme 22.1 (marked with an arrow).

To this end, the methods of the well-known GUILD program package (http://sbi.imim.es/web/index.php/research/software/guildsoftware) are integrated into the SOLANUM TUBEROSUM system, based on the analysis of the structure of the gene networks graph [33]. Figure 3 shows an example of the prioritization of genes that specifically control the metabolism of starch, which may be of interest as candidates for breeding. In the present case, the criterion for prioritization was the number of links between candidate genes with reference genes.

Another class of bioinformatic methods, implemented in SOLANUM TUBEROSUM, is based on estimates of the enrichment of biological processes by genes identified experimentally (for example, in transcriptomic analysis). Such methods are widely used in known computer systems intended for interpreting experimental data, e.g. DAVID [34], PANTHER [35, 36], GORILLA [37, 38], etc.

Extraction of knowledge using semantic-linguistic templates. In the ANDSystem, texts are recognized by a linguistic analysis module, which is fed into the input by a textual stream of factual information related to a specific subject (problem) domain. A problem-oriented ontology implemented in SOLANUM TUBEROSUM forms a conceptual model of knowledge. The module of linguistic analysis uses morphological and semantic analyzers. The morphological analyzer performs descriptor markup of the text (recognizing in the text the concepts included in the ontology, including terminological phrases), lemmatization (bringing the word to a normal form), and the POS markup. The semantic analyzer carries out a conceptual search for knowledge in the text, processed by the morphological analyzer, using semantic linguistic templates. A grammar dictionary is used for lemmatization and POS markup. The functional diagram of the system for extracting knowledge about the interactions between ontology objects in the SOLANUM TUBEROSUM knowledge base is given in Figure 4.

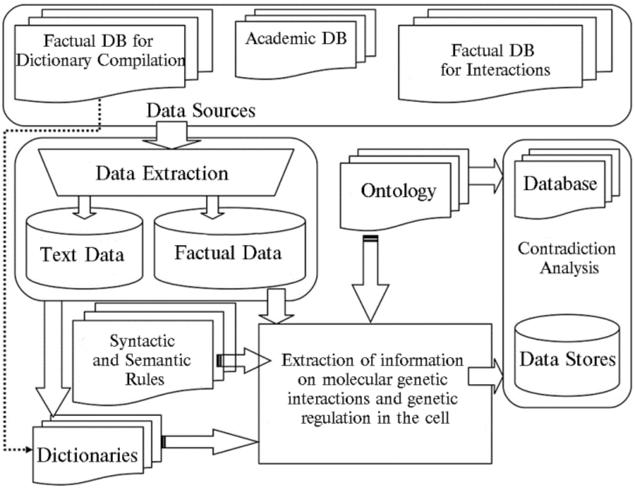
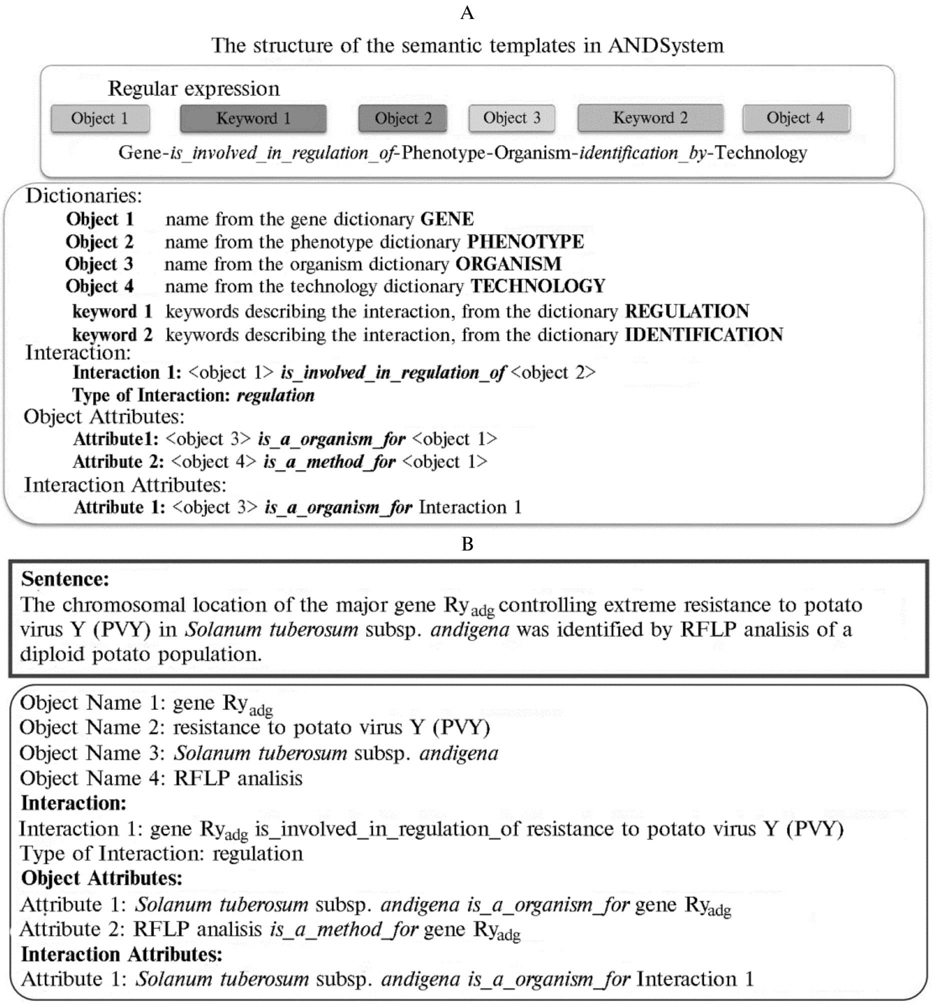

Factual DB for Dictionary Compilation
Academic DB
Factual DB for Interactions
Data Sources
Data Extraction
Text Data
Factual Data
Ontology
Database
Contradiction Analysis
Syntactic and Semantic Rules
Extraction of information on molecular genetic interactions and genetic regulation in the cell
Data Stores
Dictionaries

Fig. 4. The functional diagram of the system for extracting knowledge about the interactions between ontology objects in the SOLANUM TUBEROSUM knowledge base.

The initial data are external sources of information, including three groups of factual databases used for compiling dictionaries, extracting knowledge on molecular genetic objects and extracting knowledge about molecular interactions in the cell and gene networks, as well as a database of bibliographic data (for extracting knowledge using semantic and linguistic templates about interactions between ontology objects of the SOLANUM TUBEROSUM knowledge base).

Semantic and linguistic templates are structured records with information about types of objects, dictionaries, rules for text analysis or regular expressions, and a meta description of interaction semantics. The template structure includes the following main groups of fields: Regular expression, Dictionaries, Interactions, Object Attributes, Interaction Attributes. A regular expression defines the arrangement of object names and special linking words that indicate the specified type of interactions between specific objects in the analyzed sentence. The structure of a regular expression is a sequence of identifiers for object dictionaries and dictionaries of linking words. The symbol "–" is used as a separator character between the identifiers of dictionaries. A regular expression can also specify the admissible number of words that are not object names, which can be placed between object names in the sentence. In addition, a regular expression can contain a negation. We developed a total of about 5,000 such semantic and linguistic templates to be used in ANDSystem for extracting knowledge from the texts of scientific publications.



Gene-*is_involved_in_regulation_of*-Phenotype-Organism-*identification_by*-Technology

Dictionaries:

Object 1 name from the gene dictionary **GENE**
Object 2 name from the phenotype dictionary **PHENOTYPE**
Object 3 name from the organism dictionary **ORGANISM**
Object 4 name from the technology dictionary **TECHNOLOGY**
keyword 1 keywords describing the interaction, from the dictionary **REGULATION**
keyword 2 keywords describing the interaction, from the dictionary **IDENTIFICATION**

Interaction:

Interaction 1: <object 1> ***is_involved_in_regulation_of*** <object 2>
Type of Interaction: ***regulation***

Object Attributes:

Attribute1: <object 3> ***is_a_organism_for*** <object 1>
Attribute 2: <object 4> ***is_a_method_for*** <object 1>

Interaction Attributes:

Attribute 1: <object 3> ***is_a_organism_for*** Interaction 1

Sentence:
The chromosomal location of the major gene Ry_{adg} controlling extreme resistance to potato virus Y (PVY) in *Solanum tuberosum* subsp. *andigena* was identified by RFLP analisis of a diploid potato population.

Object Name 1: gene Ry_{adg}
Object Name 2: resistance to potato virus Y (PVY)
Object Name 3: *Solanum tuberosum* subsp. *andigena*
Object Name 4: RFLP analisis

Interaction:
Interaction 1: gene Ry_{adg} is_involved_in_regulation_of resistance to potato virus Y (PVY)
Type of Interaction: regulation

Object Attributes:
Attribute 1: *Solanum tuberosum* subsp. *andigena is_a_organism_for* gene Ry_{adg}
Attribute 2: RFLP analisis *is_a_method_for* gene Ry_{adg}

Interaction Attributes:
Attribute 1: *Solanum tuberosum* subsp. *andigena is_a_organism_for* Interaction 1

Fig. 5. Examples of the structure of the semantic and linguistic template used in ANDSystem (A), **and computer output of the results of its tryout** (B) **when extracting information about the interactions of objects in the sentence from the paper by J.H. Hämäläinen et al. [39] in the SOLANUM TUBEROSUM knowledge base.**

Let us consider as an example a template for extracting interactions between the genes and phenotypes of the organism (Fig. 5, A), adapted for the SOLANUM TUBEROSUM knowledge base. In this template, the GENE, PHENOTYPE, ORGANISM and TECHNOLOGY dictionaries are used as objects, while the “regulation” and “identification” dictionaries are used as linking words. It follows from the regular expression that object 1 (a gene from the GENE dictionary) is involved in the regulation of object 2 (a phenotype from the FENOTYPE dictionary). As can be seen, the objects and the interactions between these objects have their own attributes. In this example, objects 3 and 4 are, respectively, the organism and technology for object 1. At the same time, object 3 is an interaction attribute for object 1, indicating the organism in which it is performed.

In fact, the template contains all information about the types of objects and the types of their interactions without specifying the names of specific objects. Tryout of a template results in identifying from the texts the specific object names that match a specified regular expression. When applying the considered template to the sentence “The chromosomal location of the major gene Ry_{adg} controlling extreme resistance to potato virus Y (PVY) in *Solanum tuberosum* subsp. *andigena* was identified by RFLP analysis of a diploid potato population” [39] (see Fig. 5, B), the response output is as follows. In the case under consideration, the Ry_{adg} gene corresponds to object 1, the phenotype of resistance to potato virus Y (PVY) to object 2, the organism *Solanum tuberosum* subsp. *andigena* to object 4, and the technology of RFLP analysis to object 5.

Thus, an initial version of the knowledge base has been created for storing information on genetics, breeding, seed production, diagnostics of disease pathogens, protective means and technologies of potato storage, and to do this, an appropriate ontology has been developed (which includes dictionaries of concepts on genetics, breeding, phenomics and seed production, agrobiotechnology of cultivation and biotechnologies for potato processing and storage, diseases, pests, diagnostic methods and means of protection, environmental factors, etc.). The ANDSystem methods were adjusted for extraction of knowledge from the texts of scientific publications, patent and factual databases in the subject domain defined by the created ontology, and previously developed user interfaces were adapted. Using this system, it is planned to conduct a large-scale automatic analysis of the texts of scientific publications and patent databases. It is also expected to significantly expand the volumes of the knowledge base dictionaries by extracting new object names during analysis.

REFERENCES

1. Fiehn O. Metabolomics — the link between genotypes and phenotypes. *Plant Mol. Biol.*, 2002, 48: 155-171 (doi: 10.1023/A:1013713905833).
2. Kristensen T.N., Pedersen K.S., Vermeulen C.J., Loeschcke V. Research on inbreeding in the «omic» era. *Trends Ecol. Evol.*, 2010, 25(1): 44-52 (doi: 10.1016/j.tree.2009.06.014).
3. Weckwerth W. Green systems biology — from single genomes, proteomes and metabolomes to ecosystems research and biotechnology. *J. Proteomics*, 2011, 75(1): 284-305 (doi: 10.1016/j.jprot.2011.07.010).
4. Kumar A., Pathak R.K., Gupta S.M., Gaur V.S., Pandey D. Systems biology for smart crops and agricultural innovation: filling the gaps between genotype and phenotype for complex traits linked with robust agricultural productivity and sustainability. *OMICS: A Journal of Integrative Biology*, 2015, 19(10): 581-601 (doi: 10.1089/omi.2015.0106).
5. Lachowiec J., Queitsch C., Kliebenstein D.J. Molecular mechanisms governing differential robustness of development and environmental responses in plants. *Ann. Bot.*, 2016, 117(5): 795-809 (doi: 10.1093/aob/mcv151).
6. Lee T., Kim H., Lee I. Network-assisted crop systems genetics: network inference and integrative analysis. *Curr. Opin. Plant Biol.*, 2015, 24: 61-70 (doi: 10.1016/j.pbi.2015.02.001).
7. Hammer G., Cooper M., Tardieu F., Welch S., Walsh B., van Eeuwijk F.,

Chapman S., Podlich D. Models for navigating biological complexity in breeding improved crop plants. *Trends Plant Sci.*, 2006, 11(12): 587-593 (doi: 10.1016/j.tplants.2006.10.006).

8. Vanhaeren H., Inzé D., Gonzalez N. Plant growth beyond limits. *Trends Plant Sci.*, 2016, 21(2): 102-109 (doi: 10.1016/j.tplants.2015.11.012).
9. Potato Genome Sequencing Consortium. Genome sequence and analysis of the tuber crop potato. *Nature*, 2011, 475(7355): 189-195 (doi: 10.1038/nature10158).
10. Rensink W.A., Iobst S., Hart A., Stegalkina S., Liu J., Buell C.R. Gene expression profiling of potato responses to cold, heat, and salt stress. *Funct. Integr. Genomics*, 2005, 5(4): 201-207 (doi: 10.1007/s10142-005-0141-6).
11. Ou Y., Liu X., Xie C., Zhang H., Lin Y., Li M., Song B., Liu J. Genome-wide Identification of microRNAs and their targets in cold-stored potato tubers by deep sequencing and degradome analysis. *Plant Mol. Biol. Rep.*, 2015, 33(3): 584-597 (doi: 10.1007/s11105-014-0771-8).
12. Petek M., Rotter A., Kogovšek P., Baebler Š., Mithöfer A., Gruden K. Potato virus Y infection hinders potato defence response and renders plants more vulnerable to Colorado potato beetle attack. *Mol. Ecol.*, 2014, 23(21): 5378-5391 (doi: 10.1111/mec.12932).
13. Chae L., Kim T., Nilo-Poyanco R., Rhee S.Y. Genomic signatures of specialized metabolism in plants. *Science*, 2014, 344(6183): 510-513 (doi: 10.1126/science.1252076).
14. Dreher K. Putting the plant metabolic network pathway databases to work: going offline to gain new capabilities. In: *Plant metabolism: methods and protocols. Ser. Methods in Molecular Biology*. G. Sriram (ed.). Springer Science+Business Media, NY, 2014, V. 1083: 151-171 (doi: 10.1007/978-1-62703-661-0_10).
15. Chae L., Lee I., Shin J., Rhee S.Y. Towards understanding how molecular networks evolve in plants. *Curr. Opin. Plant Biol.*, 2012, 15(2): 177-184 (doi: 10.1016/j.pbi.2012.01.006).
16. Zhang P., Dreher K., Karthikeyan A., Chi A., Pujar A., Caspi R., Karp P., Kirkup V., Latendresse M., Lee C., Mueller L.A. Creation of a genome-wide metabolic pathway database for *Populus trichocarpa* using a new approach for reconstruction and curation of metabolic pathways for plants. *Plant Physiol.*, 2010, 153(4): 1479-1491 (doi: 10.1104/pp.110.157396).
17. Gonzalez G.H., Tahsin T., Goodale B.C., Greene A.C., Greene C.S. Recent advances and emerging applications in text and data mining for biomedical discovery. *Brief. Bioinform.*, 2016, 17(1): 33-42 (doi: 10.1093/bib/bbv087).
18. Wu H.Y., Chiang C.W., Li L. Text mining for drug—drug interaction. In: *Biomedical Literature Mining. Ser. Methods in molecular biology*. V.D. Kumar, H.J. Tipney (eds.). Springer Science+Business Media, NY, 2014, V. 1159: 47-75 (doi: 10.1007/978-1-4939-0709-0_4).
19. Piedra D., Ferrer A., Gea J. Text mining and medicine: usefulness in respiratory diseases. *Archivos de Bronconeumología* (Engl. Ed.), 2014, 50(3): 113-119 (doi: 10.1016/j.arbr.2014.02.008).
20. Fluck J., Hofmann-Apitius M. Text mining for systems biology. *Drug Discov. Today*, 2014, 19(2): 140-144 (doi: 10.1016/j.drudis.2013.09.012).
21. Krallinger M., Erhardt R.A., Valencia A. Text-mining approaches in molecular biology and biomedicine. *Drug Discov. Today*, 2005, 10(6): 439-445 (doi: 10.1016/S1359-6446(05)03376-3).
22. Szklarczyk D., Franceschini A., Wyder S., Forslund K., Heller D., Huerta-Cepas J., Simonovic M., Roth A., Santos A., Tsafou K.P., Kuhn M. STRING v10: protein—protein interaction networks, integrated over the tree of life. *Nucl. Acids Res.*, 2014, 28: gku1003 (doi: 10.1093/nar/gku1003).
23. Von Mering C., Huynen M., Jaeggi D., Schmidt S., Bork P., Snel B. STRING: a database of predicted functional associations between proteins. *Nucl. Acids Res.*, 2003, 31(1): 258-261 (doi: 10.1093/nar/gkg034).
24. Snel B., Lehmann G., Bork P., Huynen M.A. STRING: a web-server to retrieve and display the repeatedly occurring neighborhood of a gene. *Nucl. Acids Res.*, 2000, 28(18): 3442-3444 (doi: 10.1093/nar/28.18.3442).
25. Li S., Wu L., Zhang Z. Constructing biological networks through combined literature mining and microarray analysis: a LMMA approach. *Bioinformatics*, 2006, 22(17): 2143-2150 (doi: 10.1093/bioinformatics/btl363).
26. Pesch R., Böck M., Zimmer R. ConReg: Analysis and visualization of conserved regulatory networks in eukaryotes (In: German Conference on Bioinformatics, 2012). *Dagstuhl research Online Publication Server*, 2012, 26: 69-81 (doi: 10.4230/OASIcs.GCB.2012.69).
27. Shannon P., Markiel A., Ozier O., Baliga N.S., Wang J.T., Ramage D., Amin N., Schwikowski B., Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, 2003, 13: 2498-2504 (doi: 10.1101/gr.1239303).
28. Demenkov P.S., Ivanisenko T.V., Kolchanov N.A., Ivanisenko V.A. ANDVisio: a new tool for graphic visualization and analysis of literature mined associative gene networks in the ANDSystem. *In Silico Biology*, 2012, 11(3, 4): 149-161 (doi: 10.3233/ISB-2012-0449).
29. Ivanisenko V.A., Saik O.V., Ivanisenko N.V., Tiys E.S., Ivanisenko T.V.,

Demenkov P.S., Kolchanov N.A. ANDSystem: an Associative Network Discovery System for automated literature mining in the field of biology. *BMC Syst. Biol.*, 2015, 9(Suppl. 2): S2 (doi: 10.1186/1752-0509-9-S2-S2).

30. Saik O.V., Ivanisenko T.V., Demenkov P.S., Ivanisenko V.A. Interactome of the hepatitis C virus: literature mining with ANDSystem. *Virus Res.*, 2016, 218: 40-48 (doi: 10.1016/j.virusres.2015.12.003).
31. Yu B. Role of in silico tools in gene discovery. *Mol. Biotechnol.*, 2009, 41(3): 296-306 (doi: 10.1007/s12033-008-9134-8).
32. Li J., Lin X., Teng Y., Qi S., Xiao D., Zhang J., Kang Y. A Comprehensive evaluation of disease phenotype networks for gene prioritization. *PloS ONE*, 2016, 11(7): e0159457 (doi: 10.1371/journal.pone.0159457).
33. Guney E., Oliva B. Exploiting protein—protein interaction networks for genome-wide disease-gene prioritization. *PloS ONE*, 2012, 7(9): e43557 (doi: 10.1371/journal.pone.0043557).
34. Huang D.W., Sherman B.T., Lempicki R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, 2008, 4(1): 44-57 (doi: 10.1038/nprot.2008.211).
35. Thomas P.D., Kejariwal A., Guo N., Mi H., Campbell M.J., Muruganujan A., Lazareva-Ulitsky B. Applications for protein sequence—function evolution data: mRNA/protein expression analysis and coding SNP scoring tools. *Nucl. Acids Res.*, 2006, 34(Suppl 2): W645-W650 (doi: 10.1093/nar/gkl229).
36. Mi H., Poudel S., Muruganujan A., Casagrande J.T., Thomas P.D. PANTHER version 10: expanded protein families and functions, and analysis tools. *Nucl. Acids Res.*, 2015, 44(D1): D336-D342 (doi: 10.1093/nar/gkv1194).
37. Eden E., Lipson D., Yogev S., Yakhini Z. Discovering motifs in ranked lists of DNA sequences. *PLoS Comput. Biol.*, 2007, 3(3): e39 (doi: 10.1371/journal.pcbi.0030039).
38. Eden E., Navon R., Steinfeld I., Lipson D., Yakhini Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, 2009, 10: 48 (doi: 10.1186/1471-2105-10-48).
39. Hämäläinen J.H., Watanabe K.N., Valkonen J.P.T., Arihara A., Plaisted R.L., Pehu E., Miller L., Slack S.A. Mapping and marker-assisted selection for a gene for extreme resistance to potato virus Y. *Theor. Appl. Genet.*, 1997, 94(2): 192-197.