

SmartCrop cognitive platform: modules for automatic knowledge extraction from scientific publications, patents and factual databases for reconstruction and analysis of gene networks of stress response in rice and wheat

Ivanisenko V.A.^{1, 2*}, Demenkov P.S.^{1, 2}, Ivanisenko T.V.^{1, 2}, Antropova E.A.¹, Volyanskaya A.R.^{1, 2}, Zubairova U.S.^{1, 2}, Makarova A.A.^{1, 2}, Orlov Y.L.^{1, 3, 4}, Chen M.⁵

¹ Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia

² Novosibirsk State University, Novosibirsk, Russia

³ I.M. Sechenov First Moscow State Medical University of the Russian Ministry of Health (Sechenov University), Moscow, Russia

⁴ Agrarian and Technological Institute, Peoples' Friendship University of Russia, Moscow, Russia

⁵ College of Life Sciences, First Affiliated Hospital of Medical School, Zhejiang University, Hangzhou, China

* salix@bionet.nsc.ru

Key words: cognitive platform; gene networks; stress response; rice; wheat

Motivation and Aim: Rice (*Oryza sativa* L.) and wheat (*Triticum aestivum* L.) have high food, technical, and feed value and are among the most important agricultural crops. Almost half of the world's population uses rice as a staple food. This product is most widespread in Asia. Wheat is the staple food for 35 % of the world's population and is the most important agricultural crop for Russia. Such widespread use of rice and wheat as food products determines their central role in ensuring global food security. However, extreme environmental impacts, including those caused by adverse climate change, plant diseases and insect pests lead to significant crop losses. The study of molecular genetic mechanisms of plant resistance to adverse biotic and abiotic factors (high or low temperature, drought, salinity, soil pollution with metals, diseases, the action of pathogens and pests, etc.) requires analysis of large multi-omics experimental data and reconstruction of gene networks, including complex signaling, regulatory, transport and metabolic pathways.

To create the cognitive software and information platform “Smart Crop”, providing a full cycle of knowledge engineering in the field of plant biology necessary for solving problems of studying molecular genetic mechanisms of the genotype-phenotype-environment relationship for agriculturally valuable rice and wheat crops, the efforts of Chinese and Russian groups were combined. The Chinese group, led by Prof. Ming Chen, possesses advanced data mining information technologies for analyzing plant omics data [1–5]. The Russian team from the ICG SB RAS is a leader in the field of automatic knowledge extraction from scientific publication texts and reconstruction of gene networks [6, 7].

The platform is focused on solving such substantive tasks as interpretation of the results of omics data (establishing a connection between gene sets and biological processes, phenotypic traits, etc.); reconstruction of gene networks describing the relationship between molecular genetic objects and objects corresponding to the concepts of breeding, phenomics and seed production, phytopathology, diagnostics, protection

means, etc.; identification of regulatory and signaling pathways of plant response to specific growth conditions, biotic and abiotic stresses (high or low temperature, drought, salinity, soil contamination with metals, response to fertilizers, the effect of hormones, elicitors, etc.); prediction of candidate genes for genotyping; search for markers for marker-assisted selection; targets for directed impact of substances (including external factors) on plants to solve problems of early/uniform germination, better vegetative growth, efficient absorption of nutrients, etc.

The aim of this work is to present the results on the development of the “Smart Crop” platform module, which performs automatic analysis of texts of scientific publications, patents and factual databases and reconstruction on this basis of gene networks of rice and wheat response to stress.

Methods and Algorithms: Automatic analysis of scientific publications and patent texts was carried out using an integrated approach that includes an ontological description of the subject area and a set of methods for extracting knowledge from texts based on semantic-linguistic patterns, large linguistic neural network models, and graph neural networks. The development of semantic-linguistic patterns was carried out by adapting the patterns of the ANDSystem software and information system developed by us earlier [6]. The knowledge extracted from the texts was represented as a knowledge graph (associative gene networks) describing the relationships between the entities of the ontology. The technology was configured for the subject areas of breeding, phenomics and seed production, phytopathology, diagnostics, protection tools, etc. of wheat and rice.

Results: A knowledge extraction unit of the “Smart Crop” cognitive system has been developed, which includes modules for domain-specific ontology, information extraction from scientific texts and databases, and gene network reconstruction. Dictionaries for various entities related to rice and wheat have been created within the ontology. The initial filling of the “Smart Crop” knowledge base was performed using automatic analysis of scientific publications from 2023, utilizing semantic-linguistic patterns and information from factual databases. The analysis was conducted on a subset of articles (33,000 for rice and 29,000 for wheat) to identify errors and refine dictionaries and templates. The templates extracted over 200,000 interactions for wheat and 45,000 for rice, with “Association” and “Regulation” interactions being the most prevalent. Additionally, around 220,000 interactions were extracted from factual databases for both crops. The extracted interactions demonstrated good quality in terms of completeness and accuracy.

Conclusion: In the future, it is planned to integrate the knowledge extraction unit developed by the Russian group and the multi-omics data analysis unit developed by the Chinese group within the framework of the “Smart Crop” cognitive platform, which will provide full-featured capabilities of the platform for solving problems of knowledge engineering in the field of plant biology necessary for studying the molecular genetic mechanisms of the relationship between the genotype and phenotype of agriculturally valuable rice and wheat crops, taking into account environmental factors.

Funding: The work of IVA, DPS, ITV, AEA, VAR, ZUS and OYL was supported by the Russian-Chinese grant from the Russian Science Foundation No. 23-44-00030. The work of MCh was supported by National Natural Science Foundation of China (32261133526).

References

1. Chen D., Chen M., Altman T., Klukas C. Bridging Genomics and Phenomics. In: Approaches in Integrative Bioinformatics. 2014;299-333

2. Chen M., Liu L., Mei Q., Zhang Z. Rikenet: Genome-scale multi-level network reconstruction of rice. *J Comp Sci Syst Biol.* 2013;6:4
3. Liu L., Jiang L., Chen M. The Organelle-focused Proteomes and Interactomes in Rice. *Curr Protein Peptide Sci.* 2014;15(6):583-590
4. Liu L., Mei Q., Yu Z., Sun T., Zhang Z., Chen M. An Integrative Bioinformatics Framework for Genome-scale Multiple Level Network Reconstruction of Rice. *J Integr Bioinf.* 2013;10(2):223
5. Rahaman M.M., Chen D., Gillani Z., Klukas C., Chen M. Advanced phenotyping and phenotype data analysis for the study of plant growth and development. *Front Plant Sci.* 2015;6:619
6. 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
7. Ivanisenko V.A., Demenkov P.S., Ivanisenko T.V., Mishchenko E.L., Saik O.V. A new version of the ANDSystem tool for automatic extraction of knowledge from scientific publications with expanded functionality for reconstruction of associative gene networks by considering tissue-specific gene expression. *BMC Bioinformatics.* 2019;20(Suppl. 1):H.34