

Enhancing biomedical knowledge discovery through hybrid text-mining and graph neural network approaches in ANDSystem

Ivanisenko T.V.^{1, 2*}, Demenkov P.S.^{1, 2, 3}, Kolchanov N.A.^{1, 2, 3}, Ivanisenko V.A.^{1, 2, 3}

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

² Kurchatov Genomic Center of the Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia

³ Novosibirsk State University, Novosibirsk, Russia

*itv@bionet.nsc.ru

Key words: text-mining; graph neural networks; large language models; ANDSystem; Associative Gene Networks

Motivation and Aim: The exponential growth of biomedical literature complicates the extraction and understanding of complex biological systems. Traditional text-mining methods, while essential, often yield fragmented knowledge that may overlook implicit connections within biological data. This work introduces a novel hybrid approach integrated into the new version of ANDSystem [1], combining classical text-mining techniques with advanced graph neural networks (GNNs) and fine-tuned large language models (LLMs) to enhance the extraction and interpretation of biomedical information.

Methods and Algorithms: The updated ANDSystem incorporates a hybrid methodology that synergizes traditional text-mining with modern machine learning technologies, specifically graph neural networks (GNNs) and large language models (LLMs), to enhance the reconstruction and interpretation of biomedical knowledge graphs. The pipeline includes the following steps:

Named Entity Recognition: Molecular-biological entities are identified in unstructured scientific texts using ANDSystem's ontology dictionaries.

Entity Verification: The system verifies the identified names against their corresponding entity types using the context from the article, with false recognitions filtered out by the fine-tuned PubMedBERT model.

Relation Extraction: Interactions between entities are automatically established using ANDSystem's rule sets and templates, which are enhanced by information extracted from external databases.

Computation of co-occurrences: The system calculates co-occurrence values for entity pairs using the ANDDigest module to gauge the likelihood of interactions.

Graph Neural Network Training: A graph neural network is trained to represent vertices of the knowledge graph vectorially using the GraphSAGE algorithm [2].

Link Prediction Model Training: A binary classification model is trained to predict interactions between vertex pairs based on their vector representations and co-occurrence values.

Graph-based Interaction Prediction: The trained classifier is applied to predict new interactions between graph vertices.

Contextual Interaction Prediction: A fine-tuned generative LLM is used to predict interactions based on contextual cues and to provide explanations for these predictions.

The system searches for literature mentioning the interacting pairs using ANDDigest [3] and applies the fine-tuned model to detect and verify information about the connections. *Results:* The application of the developed hybrid approach within ANDSystem demonstrated improved accuracy and comprehensiveness in identifying and predicting biologically relevant interactions. Our results indicate significant enhancements in the system's ability to not only identify known interactions but also predict novel ones based on contextual understanding derived from LLMs and GNNs.

Conclusion: The integrated approach of ANDSystem facilitates a more comprehensive understanding of complex biological networks. This methodology not only increases the reliability of the reconstructed networks but also provides a platform for detailed biological interpretation, thus supporting advanced research in biomedical sciences.

Funding: The study is supported by the budget project, FWNR-2022-0020.

References

1. 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(1):34. doi 10.1186/s12859-018-2567-6
2. Hamilton W., Ying Z., Leskovec J. Inductive representation learning on large graphs. In: 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA. 2017
3. Ivanisenko T.V., Demenkov P.S., Kolchanov N.A., Ivanisenko V.A. The New Version of the ANDDigest Tool with Improved AI-Based Short Names Recognition. *Int J Mol Sci*. 2022;23(23):14934. doi 10.3390/ijms232314934