[BIOINFO 2024]

Deep Learning-Based Prediction of Human Gene-Pathway Interactions using Gene and Pathway Literature information

Gyujin Son[1], Byeongkwon Choi[2], and Mina Rho[1, 2, 3, *]

[1] *Department of Artificial Intelligence, Hanyang University*
[2] *Department of Computer Science, Hanyang University*
[3] *Department of Biomedical Informatics, Hanyang University*
** Corresponding author: minarho@hanyang.ac.kr*

Gene-pathway interaction research is essential for understanding the mechanisms of diseases, designing new drugs, and developing personalized therapies by exploring the interactions between genes and biological pathways. Understanding these interactions also plays a crucial role in deciphering gene expression and regulatory mechanisms, contributing particularly to the analysis of complex polygenic diseases. The Kyoto Encyclopedia of Genes and Genomes Database(KEGG DB) provides vital information for understanding various biological phenomena by integrating data on biological pathways, genes, and metabolites. In particular, data on human genes are indispensable for this type of research, and the KEGG DB is widely used in most pathway studies. However, the KEGG DB has a limitation in that it does not cover all existing human genes. To overcome this limitation, this study proposes a deep-learning-based graph model to predict gene-pathway interactions. Instead of relying on traditional experimental methods, we extracted embedding values of literature information on genes and pathways using a pretrained language model such as Bidirectional Encoder Representations from Transformers and used them as learning features. This approach offers the advantage of reducing time and cost compared to conventional methods, enabling more efficient research. The model was evaluated using human genes from the KEGG DB, and it achieved impressive performance with a precision of 0.84, recall of 0.84, and F1 score of 0.83 across various pathway categories. Furthermore, this study suggests that it can assist in building high-quality candidates for predicting interactions between new genes and existing pathways, verified through experimental validation. These results are expected to enhance the efficiency of gene-based research across various fields and expand the potential for clinical applications.