[BIOINFO 2024]

PONYTA: Prioritization of phenotype-related genes from mouse KO events using PU learning on a biological network

Jun Hyeong Kim[1,†], Bonil Koo[2,3,†], and Sun Kim[1,2,3,4,5,*]

[1] *Bio-MAX/N-Bio, Seoul National University*
[2] *Interdisciplinary Program in Bioinformatics, Seoul National University*
[3] *AIGENDRUG Co., Ltd*
[4] *Department of Computer Science and Engineering, Seoul National University*
[5] *Interdisciplinary Program in Artificial Intelligence, Seoul National University*
[†] *These authors contributed equally to the work*
[*] *Corresponding author: sunkim.bioinfo@snu.ac.kr*

Transcriptome data from gene knock-out (KO) experiments in mice provide crucial insights into the intricate interactions between genotype and phenotype. Differentially expressed gene (DEG) analysis and network propagation (NP) are well-established methods for analyzing transcriptome data. To determine genes related to phenotype changes from a KO experiment, we need to choose a cutoff value for the corresponding criterion based on the specific method. Using a rigorous cutoff value for DEG analysis and NP is likely to select mostly positive genes related to the phenotype, but many will be rejected as false negatives. On the other hand, using a loose cutoff value for either method is prone to include a number of genes that are not phenotype-related, which are false positives. Thus, the research problem at hand is how to deal with the trade-off between false negatives and false positives.

We propose a novel framework called PONYTA for gene prioritization via positive-unlabeled (PU) learning on biological networks. Beginning with the selection of true phenotype-related genes using a rigorous cutoff value for DEG analysis and NP, we address the issue of handling false negatives by rescuing them through PU learning. Evaluations on transcriptome data from multiple studies show that our approach has superior gene prioritization ability compared to benchmark models. Therefore, PONYTA effectively prioritizes genes related to phenotypes derived from gene KO events and guides in vitro and in vivo gene KO experiments for increased efficiency.