DSSV 2023
Data Science, Statistics & Visualisation
ECDA 2023
European Conference on Data Analysis

Evaluation of network-guided random forest for disease gene discovery

Jianchang Hu[1] and Silke Szymczak[1]

[1]Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein,
Campus Lübeck, Lübeck, Schleswig-Holstein, Germany

Identification of biomarkers associated with complex diseases can improve patient risk prediction and foster understanding of underlying molecular pathomechanisms. Gene network information is believed to be beneficial for disease module and pathway identification. We investigate the performance of a network-guided random forest (RF) where the network information is summarized into a sampling probability of predictor variables which is further used in the construction of the RF. The identification of important genes is based on standard variable importance measures from RF. In our simulation study, we generate synthetic RNA sequencing (RNA-Seq) data along with the underlying network structure using the R package `SeqNet` [1]. Our results suggest that network-guided RF does not provide better disease prediction than the standard RF. In terms of disease gene discovery, when disease genes are randomly distributed within the network, network information only deteriorates the gene selection, but if they form disease module(s), network-guided RF identifies causal genes more accurately. We also find that when disease status is independent from expression of genes in the given network, spurious gene selection results can occur when using network information, especially on hub genes. Two balanced microarray and RNA-Seq breast cancer datasets from The Cancer Genome Atlas (TCGA) with 283 and 284 patients, respectively, along with protein-protein interaction network information from the STRING database [2] are investigated for classification of progesterone receptor (PR) status. Both datasets include 193 PR-positive patients. Standard and network-guided RFs can both detect the core genes including *PGR* and *ESR1* on both datasets. In addition, network-guided RFs can identify further genes from PGR-related pathways, which leads to a more connected module of identified genes. This demonstrates the potential gains in disease module and pathway identification by utilizing network information for complex diseases.

[1] T. Grimes and S. Datta, "SeqNet: an R package for generating gene-gene networks and simulating RNA-seq data," *Journal of Statistical Software*, vol. 98, no. 12, 2021.

[2] D. Szklarczyk, R. Kirsch, M. Koutrouli, K. Nastou, F. Mehryary, R. Hachilif, A. L. Gable, T. Fang, N. T. Doncheva, S. Pyysalo, *et al.*, "The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest," *Nucleic Acids Research*, vol. 51, no. D1, pp. D638–D646, 2023.