[BIOINFO 2024]

Optimization of Preprocessing Strategy for Developing a Machine Learning-based Disease Diagnosis Model Using Transcriptomics

Hye Jung Min[1], Ki Wook Lee[1], Hyun Woo Park[1], Ji Min Seo[1], Na Young Kwon[1], Ji Min Seo[1], Heeje cho[1], Minho Choi[1], Balachandran Manavalan[1], Young-Jun Jeon[1]*

[1]*Department of Integrative Biotechnology, Sungkyunkwan University, Suwon 16419, Gyeonggi-do, Korea*

**Corresponding author: jeon2020@skku.edu*

Current research in the fields of medicine and biology predominantly focuses on utilizing artificial intelligence (AI) for the analysis of patient-specific images, such as CT and MRI scans. However, the utility of the AI has been limited in NGS data, in particular, transcriptome. RNA sequencing is susceptible to biological and technical errors during data generation. The approaches to error correction vary significantly, and the selection of a specific method can yield different outcomes, potentially influencing subsequent analyses, suggesting that the clear guidelines using transcriptome data presents challenges in achieving optimal results. In this study, we aim to identify the optimal preprocessing strategies for the development of AI models utilizing transcriptomic data. To this end we employed RNA-sequencing dataset from 25 different cohorts including lung adenocarcinoma, colorectal cancer, diabetes, and other disease with over 6,000 patient data sets. Subsequently, 18 different combinatorial analyses using 6 different normalization methods (Raw, CPMTMM, RLE, UQ, RPKM, TPM) and 3 scaling methods (Raw, Minmax, Z) were systemically applied and the effects of each preprocessing approach were assessed through conventional bioinformatics analysis. Using the over 20,000 transcripts were preprocessed from each combination of normalization and scaling techniques, we developed the disease diagnosis model leveraging 15 machine learning & deep learning algorithms and evaluated the performances to determine the most effective preprocessing strategy. Ultimately, optimal preprocessing combinations for each algorithm were identified for developing AI models to develop the best performing model using transcriptome data. Our findings support the further application of AI for disease diagnosis by proposing optimized guidelines to develop the best model using transcriptomics.