Child Health Research Institute
University of Nebraska Medical Center
Children's NEBRASKA

B-Cell Acute Lymphoblastic Leukemia Subtype Identification with An Ensemble Random Projection-Based Machine Learning Model

Lusheng Li[1], Hanyu Xiao[1], and Shibiao Wan[1]
[1]Department of Genetics, Cell Biology and Anatomy, University of Nebraska Medical Center, Omaha, NE 68198

According to the American Cancer Society, the estimated number of new cancer cases for children (ages 0 to 14 years) in the US in 2023 is 9,910 and the most diagnosed cancer is leukemia, of which the majority (75%) is acute lymphoblastic leukemia (ALL). As the most common pediatric malignancy, B-cell ALL (B-ALL) has multiple distinct subtypes characterized by somatic and germline genetic alterations like chromosomal alteration, transcription factor rearrangement or kinase inhibition. The treatment of B-ALL patients should be personalized based on specific subtypes, as different subtypes of B-ALL may respond differently to various treatments. Identification of B-ALL subtypes can facilitate risk stratification and enable tailored therapeutic approaches. Existing methods for B-ALL subtyping primarily depend on immunophenotypic, cytogenetic and genomic analyses, which would be costly, complicated, and laborious in clinical practice applications. To overcome these challenges, we present RanBALL (an Ensemble Random Projection-Based Model for Identifying B-Cell Acute Lymphoblastic Leukemia Subtypes), an accurate and cost-effective model for B-ALL subtype identification based on transcriptomic profiling only. The RanBALL utilizes random project (RP) techniques to construct an ensemble of SVM classifiers. Specifically, the transcriptomic profiling features were projected onto low-dimensional spaces by random projection matrices whose elements conform to a distribution characterized by zero mean and unit variance. To ensure reliable and robust performance, we selected 20 subspace dimensions ranging from 600 to 2500, with intervals of 100. The transformed low dimensional data matrix was used for training an ensemble of multi-class support vector machine (SVM) classifiers, each corresponding to one of the RP matrices of various dimensions. The predicted probabilistic scores of each B-ALL subtype were integrated for determining the final decision. Results based on 10 times 10-fold cross validation tests demonstrated that the proposed model achieved an accuracy of 93.7%, indicating promising prediction capabilities of RanBALL for B-ALL subtyping. Furthermore, the 30% held-out tests suggested that the model's robustness and consistency to maintain high confidence levels for accurate predictions. The high accuracies of RanBALL suggested that our model can effectively capture underlying patterns of transcriptomic profiling for accurate B-ALL subtype identification. To extend the impact of RanBALL, we have established a free and publicly available python package for RanBALL available at https://github.com/wan-mlab/RanBALL. We believe RanBALL will facilitate the discovery of B-ALL subtype-specific marker genes and therapeutic targets, and eventually have consequential positive impacts on downstream risk stratification and tailored treatment design.