Gene Embedding: A Novel Computational Hybrid Approach to Somatic Mutation-Based Primary Cancer Type Identification and Biomarker Discovery

Xu, Sidra (School: The Harker School)

Cancer is a leading cause of death worldwide. Because the organ and cell type that generated the tumor determine a patient's response to therapies, quick and accurate identification of the primary site is critical for guiding effective treatment. Yet there is no rapid and effective approach available to aid early screening. Somatic point mutation-based cancer typing has the potential of differentiating similar tumors and delivering accurate results, but researchers face limited accuracy due to the challenge of modeling complex gene interactions. To address this issue, a novel approach is implemented in this study: to harness the power of gene expression data and to include it in a somatic mutation-based cancer identification model through embeddings. An embedding is a mapping of complex categorical variables to vectors of continuous numbers with the capability of uncovering relationships between these variables that are otherwise regarded as independent entities. By introducing portable gene expression embeddings, the model can harvest information in both somatic mutation and gene expression without requiring the latter from patients, often not available in clinical settings. Results in this work show that when a gene expression embedding extracted from all cancer-related genes in TCGA databases is applied, the model provides a prediction accuracy of 76% on 12 tumor classes, an improvement of more than 15% compared with previous studies without embeddings. My research is the first to demonstrate the success of a hybrid genetic model. Furthermore, feature ranking analyses reveal a number of genetic markers specific to each cancer type, which could be studied for a better understanding of each cancer and utilized as therapeutic targets in the future.