mICKEY: Memory-Efficient Deep Learning for Personalized Biomarker Discovery and Cancer Origin Prediction From DNA Methylation Data

Tussanapirom, Pakanan (School: Triam Udom Suksa)
Aewsrisakul, Kasidech (School: Triam Udom Suksa)

Cancer accounts for nearly 10 million deaths annually, and treatment relies on identifying tissue origin. Traditional diagnostics are time-consuming, with wait times up to 120 days from symptom onset to biopsy, potentially worsening disease progression. Early cancer detection using blood-based molecular markers, like CpG site methylation changes, offers a promising alternative. However, standard methylation profiling technologies are expensive and computationally intensive, while conventional statistical methods lack complexity and generalizability for modeling large-scale, non-linear biological data. Addressing these limitations, we collected 8,846 samples from 7,781 patients via the GDC Data Portal and preprocessed data to reduce bias and account for missing values. We employed a deep learning model with variational inference layers, dense layers, and attention layer to select CpG sites with the highest feature importance scores. Our model demonstrated over 0.95 sensitivity and specificity across 18 cancer types, in different stages, and demographic, using fewer than 100 CpG sites. We also achieved over 0.85 sensitivity and specificity in cell-free DNA prediction with domain adaptation techniques. A permutation test assessed the importance of each CpG site for specific cancer types, discovering known and novel biomarker genes. We developed a web application providing graphical and interpretative results for medical professionals, enabling informed decision-making in patient care. mICKEY offers a memory-efficient deep learning solution for personalized biomarker discovery and cancer origin prediction, facilitating improved early cancer detection and treatment strategies.

Awards Won: