Enabling Personalized Medicine: A Novel Deep Learning Tool for Classifying Genetic Mutations Using Text from Clinical Evidence

Ping, Jason (School: Bergen County Academies)

The understanding of genetic mutations and their effects is the foundation of personalized medicine. Currently, this interpretation is time-consuming, costly, and susceptible to bias, involving the manual reviewing of thousands of scientific texts on individual mutations. To address these issues, a deep-learning natural language processing tool was developed to automatically classify genetic variants and their effects. Opensource data on genetic variants and related clinical literature was utilized to engineer features that represent the relationship between variations and their impacts. Text-mining algorithms such as term frequency-inverse document frequency, coupled with high dimensional vector representations, were performed on the text corpus to embed the relationships between terms. Additionally, physicochemical properties of the substituted amino acids and their respective Grantham scores were used to map the severity of the changes and the amino acid evolutionary distances. The machine-learning system is the concatenation of a Multi-Layer Perceptron and a bidirectional Long Short-Term Memory Network that incorporates dimensionality reduction to capture principle features and mitigate noise. After training, the predictor achieved a high accuracy of 92.3% and an F1-score of 85.5. The tool was validated based on feature prioritization and previously annotated mutations through cross-validation. The deep learning predictor was then applied to currently unclassified genetic variations and identified 13 as novel oncogenic mutations. Ultimately, this study not only helps solve one of precision medicine's primary limitations, but also presents industry viability, significantly streamlining the research process and potentially leading to the development of new therapies.

Awards Won:
Intel ISEF Best of Category Award of $5,000
First Award of $3,000