Mutformer-Deciphering the Language of Genetic Variants: Using a Transformer-Based Language Model To Identify Pathogenic Missense Mutations Associated With Human Inherited Diseases and Cancer

Jiang, Theodore (School: Palisades Charter High School)

Missense mutations are a form of genetic variant that result in substitutions of amino acids in protein sequences. These mutations account for approximately half of human inherited diseases and are often associated with cancers. However, the identification of such pathogenic mutations from non-pathogenic variants in the human genome remains challenging. Taking advantage of deep learning and the transformer architecture, I developed MutFormer, a transformer-based language model, to accurately predict pathogenicity of missense mutations. MutFormer mimics human language processing to understand the language of proteins. When training on reference protein sequences and mutated protein sequences resulting from common genetic variants (assumed to be benign), MutFormer utilizes an adaptive vocabulary where the “words” considered are patterns of amino acids learned. In pathogenicity prediction of missense mutations, the MutFormer architecture outperforms the current state-of-the-art transformer model by >10%. MutFormer also outperforms or matches current methods of pathogenicity prediction with or without, respectively, manually-annotated homology information or other additional data. With accurate and efficient identification of disease-associated mutations via a language model of proteins, MutFormer could assist in early diagnosis through genetic screening or the development of new cures for cancers.

Awards Won:

Fourth Award of $500

Fourth Award of $500