Deep Learning: Structural and Characteristic Classification of nsSNP Deleterity

Lee, Michael

Huffman, Raymond

Genetic evaluation can predict cancers and other genetic diseases before expression in the phenotype, allowing for early preventative treatment. Single nucleotide polymorphisms (SNP) are the most frequently occurring mutations in humans, but there is currently no reliable method for determining the deleterity of a given SNP. One solution is to leverage artificial intelligence by creating an algorithmic machine learning model. Based on Sim et al. (2012) and Collingridge et al. (2012), it was hypothesized that ortholog-only MSA creation would be more effective than current MSA-building methods used by other models. The algorithm queried from Ensembl and UniRef using BLAST and biomaRt queries and ran InParanoid to identify orthologs, aligning MSAs using the shell program Clustal Omega. Domain conservation was quantified using BLOSUM62, and physicochemical properties were queried from pre-computed files. These scores were input as features for the machine learning algorithm written using the Python package Keras. The model was trained and tested on public mutation databases HUMDIV and HUMVAR, with analysis done by plotting ROC and PRC graphs to analyze the area under the curve (AUC), precision and recall, kappa statistics, and accuracy. In a direct comparison, the study's algorithm was more effective than other algorithms, including PolyPhen2, on HUMVAR. The use of orthologous domain conservation allowed for better performance on a dataset with more variant nsSNPs, like HUMVAR. The best accuracy achieved was 86%, with a kappa statistic of 0.64 and an AUC of 90%. The study revealed that orthologous domain conservation, CpG change, and accessible surface area change are important indicators of deleterity, creating a model that is effective in evaluating variant mutations.

Awards Won: