Predicting Clinical Relevance of Missense Mutations with Machine Learning Models and Protein Structural Information

The human genome contains approximately 76 million potential missense variants, some of which are likely benign, while others are potentially pathogenic. Variants that lead to protein truncation, known as protein-truncating variants, are often classified as pathogenic because they are frequently associated with the loss of protein function. In contrast, missense variants—the majority of clinically relevant variants—exhibit highly variable effects on protein structure and function, with only a small proportion being pathogenic.

Advances in next-generation sequencing (NGS) have enabled the identification of millions of missense variants in the human genome. Many of these variants are cataloged in public aggregation databases. Analyzing these variants is clinically important and can improve our understanding of genetic diseases. Recently, efforts have focused on characterizing missense variants experimentally using high-throughput techniques, such as deep mutational scanning. However, these approaches are labor-intensive and limited in scope, leaving a large population of missense variants with uncertain significance.

In recent studies, several state-of-the-art variant effect predictors (VEPs) have been developed to address the challenges of interpreting missense variants. Examples of these tools include REVEL, SNPs&GO, and FATHMM which employ computational methods to predict the clinical relevance of missense variants. These VEPs are built on machine learning models such as Random Forest, Support Vector Machines, Gradient Boosting, and Neural Networks. These models are trained on diverse features, including variant frequency, evolutionary conservation, and the physicochemical properties of amino acids. To further enhance predictive performance, some VEPs integrate outputs from multiple predictors.

Despite their strengths, VEPs have limitations. Most precompute scores for a limited set of variants in disease-associated genes that were available at the time of their development. These scores are often accessible through online platforms, allowing researchers to query the pathogenicity of specific variants. However, not all disease-associated genes and their missense variants are curated or included in these precomputed datasets. Consequently, researchers often face challenges when analyzing uncharacterized variants by these VEPs. This challenge is particularly pronounced with the widespread adoption of next-generation sequencing (NGS), where the whole genome of a patient can be sequenced to identify missense variants associated with a particular disease condition. In such cases, the success of a query depends on whether the score of such variants has already been precomputed.

This study aims to develop a high-performing VEP that incorporates structural information derived from predicted protein 3D structures. Unlike existing VEPs which rely on precomputed scores for predictions, the proposed VEP will dynamically predict the pathogenicity of missense variants based on their specific features. This functionality will provide researchers with greater flexibility to analyze uncharacterized missense variants. Different machine learning models will be trained using high-quality variant data labeled with clinical significance. Additionally, data from deep mutational scanning (DMS) experiments will be used as an independent benchmark to evaluate and validate the VEP's performance. The model with the highest predictive performance will be implemented and made publicly available.

References:

Zhao, H., Du, H., Zhao, S., Chen, Z., Li, Y., Xu, K., ... & Wu, N. (2024). SIGMA leverages protein structural information to predict the pathogenicity of missense variants. Cell Reports Methods, 4(1).

Fowler, D.M., and Fields, S. (2014). Deep mutational scanning: A new style of protein science. Nat. Methods 11, 801–807.

Majithia, A.R., Tsuda, B., Agostini, M., Gnanapradeepan, K., Rice, R., Peloso, G., Patel, K.A., Zhang, X., Broekema, M.F., Patterson, N., et al. (2016). Prospective functional classification of all possible missense variants in PPARG. Nat. Genet. 48, 1570–1575.

Cheng, J., Randall, A., & Baldi, P. (2006). Prediction of protein stability changes for single-site mutations using support vector machines. Proteins: Structure, Function, and Bioinformatics, 62(4), 1125-1132