SIGMA LEVERAGES PROTEIN STRUCTURAL INFORMATION TO PREDICT PATHOGENICITY OF MISSENSE VARIANTS

Summary

Recent studies have seen the development of numerous in-silico prediction tools as an alternative approach to laboratory experimental studies for interpreting the pathogenicity of missense variants in the human population. These tools utilize various methodologies, including evolutionary conservation, sequence alignment, and physicochemical attributes of mutated residues, and apply machine learning models such as Support Vector Machines, Random Forests, Gradient Boosting, and Neural Networks in their predictions.

Assessing the pathogenicity of missense variants is key to interpreting genetic data. A promising strategy involves evaluating variant effects in the context of protein structure. However, the scarcity of known 3D protein structures has hindered the exploitation of structural information, as protein structure is critically important for ensuring proper molecular function. The advent of AlphaFold2 has begun to address this challenge by enabling extensive, accurate predictions of 3D structures. In this study, the researchers introduced the Structure-Informed Genetic Missense Mutation Assessor (SIGMA), an in-silico predictive tool that leverages protein structural information from AlphaFold2 predictions to evaluate the effects of missense variants in the context of predicted protein structures.

Training and testing datasets

A total of 27,165 benign and 22,957 pathogenic missense variants from 3454 proteins were obtained from the ClinVar and gnomAD databases. Additionally, 27,928 variants of six proteins (BRCA1, P53, MSH2, PTEN, VKORC1, and HRAS), systematically evaluated by Deep Mutational Scanning (DMS) experiments, were included to independently assess the performance of the Variant Effect Predictors (VEPs). Wild-type protein structure predictions in PDB format for the 3454 proteins considered in the study were retrieved from the AlphaFold Protein Structure Database. Variants were mapped to their predicted protein 3D structures, and those that could not be mapped due to inconsistent isoforms were excluded.

For each variant, 57 features were derived from their 3D protein structures. These features were classified into three categories:

- Protein-level features: These characterize the general properties of the wild-type predicted protein structures. The Stability command of the FoldX program was used to estimate these features.
- Residue-level features: These characterize the structural context of the missense variants. The DSSP (Dictionary of Secondary Structure of Proteins) program was used to extract these features.
- Mutation-level features: These characterize the impact of missense variants on protein stability after mutation. The PositionScan command of the FoldX program was used to estimate these features.

Model development and evaluation

The dataset was split into training (80%) and testing (20%) subsets. The additional dataset from DMS studies, which are independent of the labeled dataset, was used for independent performance assessments of the tool. For the training dataset:

- Dichotomous variables (e.g., disulfide bond cleavage and secondary structures) were normalized using the one-hot encoding technique.
- Continuous variables (e.g., RSA and ΔΔG) were normalized using Z-score normalization. Normalization parameters derived from the training set were applied to the test datasets.

After preprocessing, a Gradient Boosting Machine (GBM) model was trained to classify pathogenic versus benign variants. Hyperparameters (e.g., learning rate, maximum tree depth, and sample rate per tree) were tuned using a Cartesian grid search procedure with a 5-fold cross-validation strategy. Various models were built using different hyperparameters, and out-of-fold predictions were used to estimate model generalization performance. The Area Under the Receiver Operating Characteristic (ROC) Curve (AUC) was used to evaluate the model's ability to discriminate between benign and pathogenic variants. The GBM model with the highest AUC on the validation set was selected as the final predictor.

To determine each feature's contribution to the predictor's discriminative ability, feature importance scores were retrieved from the GBM model. These scores were calculated as the average improvement in the performance metric (squared error) for all trees, then rescaled to a fixed range (0–1). The optimal threshold for binary classification was determined by maximizing the Youden index. Metrics such as AUC, accuracy, sensitivity, specificity, positive predictive value, and negative predictive value were used to evaluate the model's performance.

Combining SIGMA with other VEPs

To create a more comprehensive predictor for variant pathogenicity, SIGMA was combined with four high-performing VEPs: DEOGEN2, EVE, PROVEAN, and MutPred. The combined predictor, SIGMA+, was constructed using elastic-net-penalized logistic regression with a 5-fold cross-validation strategy based on the same training set. To evaluate SIGMA's performance, SIGMA scores were compared with the predicted scores of 28 other in-silico tools. These VEPs were categorized into:

- Individual predictors (n = 17): Tools that do not rely on other VEPs.
- Meta-predictors (n = 11): Tools that integrate the outputs of other VEPs as input features.

The performance of SIGMA, SIGMA+, and the other VEPs was evaluated on the independent DMS dataset. Spearman's correlation coefficient was computed between functional scores from the DMS dataset and prediction scores from VEPs for each of the six proteins. The overall performance of a VEP was defined as the mean of the correlation coefficients across all six proteins. In comparison with existing predictors across labeled variant datasets and experimental datasets, SIGMA demonstrated superior performance in predicting missense variant pathogenicity (AUC = 0.933).

Notably, the relative solvent accessibility of the mutated residue contributed significantly to SIGMA's predictive ability. Combining SIGMA with other top-tier predictors to create SIGMA+ further enhanced prediction performance (AUC = 0.966). To facilitate the application of SIGMA, pre-computed scores for over 48 million possible missense variants across 3,454 disease-associated genes were made available through an interactive online platform (https://www.sigma-pred.org/). Overall, SIGMA leverages protein structure information to provide an accurate, structure-based approach to evaluating missense variant pathogenicity.

Statistical analysis

Associations between structural features and variant pathogenicity were assessed based on feature types:

- Dichotomous features: Contingency tables were built, and the chi-squared test was used to determine associations.
- Continuous features: Logistic regression was used to assess associations, with the strength of association quantified using odds ratios (ORs) and 95% confidence intervals (CIs).

All statistical analyses and data visualizations were conducted using R software (version 3.6.3) and the following packages: h2o, caret, pROC, forestplot, ggpubr, ggsci, viridis, and cutpointr.