ClinPred: Prediction Tool to Identify Disease-Relevant Nonsynonymous Single-Nucleotide Variants

Most of the genetics diseases come from several types of mutation such as indels and frameshift. Our genes are the information bank of our body, so the mutation gene might cause unimaginable effects to our body. Moreover, the advancement of the high-throughput genome sequencing allows us to sequence DNA at the cheaper cost. The sequenced data has some regions that might differ from the others in the population which can cause the disease or not. Distinguishing between non-effect variant or cause-disease variant is an essential technology for personalised medicine. There are several pathogenicity predictors available but the performance is limited and needs to be improved. This project aims to build a pathogenicity prediction from DNA sequence by using machine learning algorithms that learned from the existing sample from several data sources. ClinPred shows the highest performance in several matrices such as sensitivity and specificity.

In addition, the researcher found that allele frequency plays a crucial role in predicting pathogenic variants which leads to significant performance improvement since it was not be considered from others predictors. The dataset was separated into 2 parts, the first is the training set. ClinVar database was retrieved and filtered by removing the variants that have conflict in interpretation to ensure the confidence of the data. The last part is testing set, the researcher retrieved data from several databases which are ClinVar, Mutageneix, DoCM, FORGE Canada and BRCA1 dataset from “A Database of Functional Classification of BRCA1 Variants based on Saturation Genome Editing”. All of the data were transformed to be features by using ANNOVAR in dbNSFP v.3a.

Figure 1: The comparison of performance among ClinPred and other competitors in both testing dataset. The result was calculated for 5-fold cross-validation.

To define models, they used tree-based machine learning algorithms which are Random Forest (cforest) and Gradient Boosting (xgboost) with balanced weight model scheme to handle the unbalanced dataset. The output of the model will be range from 0 to 1, representing the probability of being a pathogenic variant. To evaluate the model, they performed the 5-fold cross-validation method and computed seven evaluation metrics including sensitivity, specificity, accuracy, precision, F1 Score and Matthew correlation coefficient (MCC). From Figure 1, the performance comparison between ClinPred and other available predictor software result that ClinPred outperforms other competitors especially in sensitivity and specificity from both testing sets.

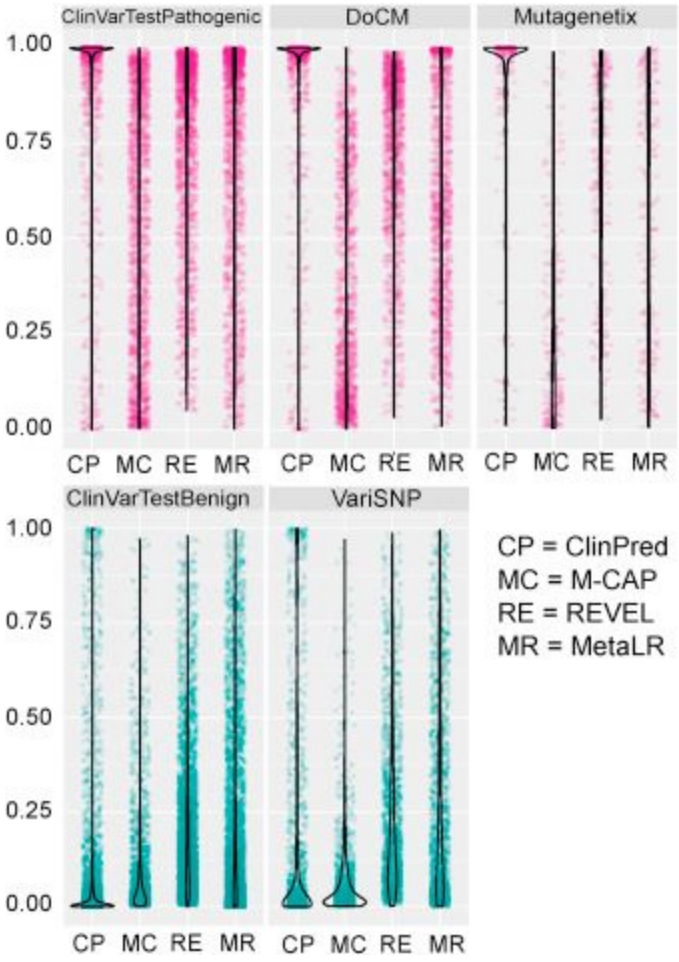

Figure 2: Comparison of Raw Scores of ClinPred, M-CAP, REVEL and MetaLR

In Figure 2, it shows the distribution of predicted score ranging from 0 to 1, benign and pathogenic respectively. by using ClinPred, M-CAP, REVEL and MetaLR datasets. Pink represents pathogenic datasets, and green represents benign dataset. The result shows that ClinVar predicts to be correct in most of the cases since the distribution in pathogenic dataset mostly packed in nearly one while the distribution in benign datasets mostly packed in nearly

zero. So, we can conclude that ClinPred well differentiates between benign and pathogenic variants compared to other methods.

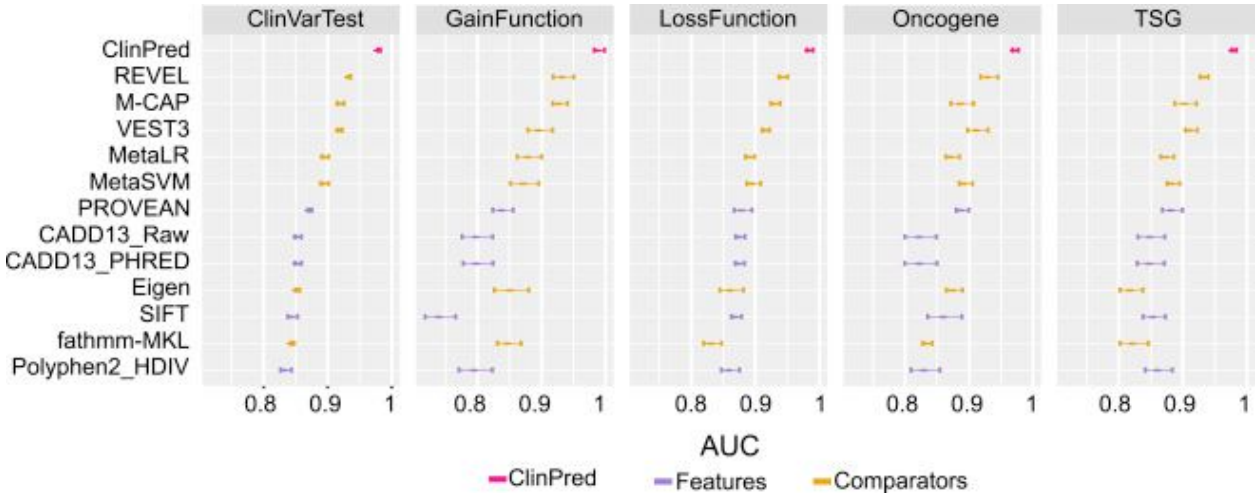

Figure 3: Comparison of AUC of ClinPred and other methods across distinct dataset based on pathogenic mechanisms.

The experiment changes in Figure 3 by categorised the variants into 5 datasets based on the effect of variant including ClinVarTest as a control set, Gain of Function, Loss of Function, Oncogene and Tumour Suppressor Gene (TSG). The position of the line for each method shows the value of AUC which ClinPred got highest AUC in all of the datasets, so ClinPred robust to several types of the mutation effects.

There are lots of pathogenic prediction method available but the performance is limited to distinguishing between pathogenic and background variants. ClinPred improved the performance over other methods by using machine learning since it showed the highest sensitivity, improved specificity and the best performance in evaluation metrics. The concern for this project which might lead to performance improvement which is a dataset. The model relies on the training set. If the dataset accuracy, the model should work better. Form the progress so far, this research will be beneficial for the future precision medicine.

Work Cited

- Alirezaie, N., Kernohan, K. D., Hartley, T., Majewski, J., & Hocking, T. D. (2018). ClinPred: Prediction Tool to Identify Disease-Relevant Nonsynonymous Single-Nucleotide Variants. *The American Journal of Human Genetics, 103*(4), 474-483. doi:10.1016/j.ajhg.2018.08.005