ipath: an interpretable approach for prediction of gain-of-function and loss-of-function variants

Halima Bensmail
Qatar Computing Research Institute

Aim: Uncovering the genetic basis of human disease requires determining the pathogenicity and functional impact (i.e., gain-of-function [GOF] or loss-of-function [LOF]) of a variant and results in a gene product with increased (gain-of-function; GOF) or diminished (loss-of-function; LOF) activity.

The long term is to create a ML/AI-based computational solution to provide a framework for the prediction of pathogenicity and functional impact of variants.
The project has two levels:

1- A challenging part which is related to curate and collect data from several database before using a ML algorithm. We predict we will have 9619 pathogenic GOF/LOF and 138,026 neural variants curated data.
2- Use an optimized gradient boosting machine (XGBoost) approach on the data with a total of 138 variant-level, 262 protein-level, and 103 genome-level characteristics (503 features).
3- Use feature prioritization approach such as SHAP scores machine and other statistical analyses to identify discriminative features.

Data: We will use variants from the Genome Aggregation Database (gnomAD release 2.0.2) that are considered to be likely neutral variants as our reference dataset to compare with GOF and LOF variants [1-2].

Features: we will use (1) variant level features, (2) gene-level features, and (3) protein-level features

If successful, this will be a very important work toward improving our understanding of how variants affect gene/protein function and may ultimately guide future treatment options.

Deliverables:

a- Build two groups of data, cluster 1 that has 9619 pathogenic GOF/LOF and 138,026 with neutral variants.
b- Build features from 138 variant levels, 262 protein level and 103 genome level
c- Run the recently published code produced by Fang Ge et al. (2023). VPatho: a deep learning-based two-stage approach for accurate prediction of gain-of-function and loss-of-function variants. *Briefings in Bioinformatics*, Volume 24, Issue 1, https://doi.org/10.1093/bib/bbac535

Required Skills:

Knowledge in programming with any language specifically Python and R if possible.

Internship Batch: Batch 1 from May 7 to June 29

Mentors

Halima Bensmail email: hbensmail@hbku.edu.qa

References:

[1] Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alfoldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581, 434–443.

[2] Ge, F., Li, C., Iqbal, I., Arif, M., Li, F., Thafar, M.A., Yan, Z., Worachartcheewan, A., Xu, Xiaofeng., Song, J et al. (2023). VPatho: a deep learning-based two-stage approach for accurate prediction of gain-of-function and loss-of-function variants. *Briefings in Bioinformatics*, Volume 24, Issue 1, January 2023, bbac535, https://doi.org/10.1093/bib/bbac535