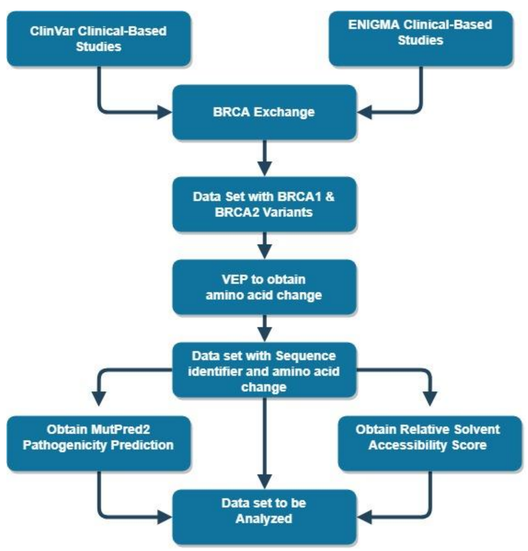
Testing the Performance of the MutPred2 Variant Pathogenicity Predictor on a set of BRCA1 and BRCA2 Variants and the Relevance of Solvent Accessibility in Predicting Pathogenicity

Diego Quezada-Munoz, Nicholas Lenz, Louis Gil-Acevedo, Melissa Cline
California State University Monterey Bay, University of California Los Angeles,
University of Puerto Rico, University of California Santa Cruz

Abstract

Mutations in the BRCA1 and BRCA2 genes are known to be associated with an increased risk of breast and ovarian cancer. There are variants of BRCA1 and BRCA2 genes that have been clinically assessed for pathogenicity, yet there are new and undiscovered variants that have not. These gene variants are classified as Variants of Uncertain Significance (VUSs) and individuals who have a variant in this class could possibly be in risk of developing cancer. In absence of clinical data, In Silico Prediction Algorithms are techniques that predict the probability of a variant being pathogenic by using parameters such as protein structure of the variant and amino acid substitutions. The purpose of this study is to test the performance of the In Silico prediction software known as MutPred2 and also determine if there is an advantage in using solvent accessibility as a predictor for variant pathogenicity. To address this question we applied the MutPred2 prediction software on a set of single-nucleotide change BRCA1 and BRCA2 variants that were added to the BRCA Exchange after the MutPred2 release date. Performance is based on how MutPred2 output compares to pathogenicity scores in the ENIGMA and ClinVar database which has data on clinical-based studies. We identified outliers that MutPred2 had difficulty in predicting pathogenicity and check to see if solvent accessibility can aid in predicting pathogenicity. If MutPred2 and solvent accessibility predictions compare well to ENIGMA and ClinVar classifications, then we have evidence that solvent accessibility is valuable in predicting pathogenicity.

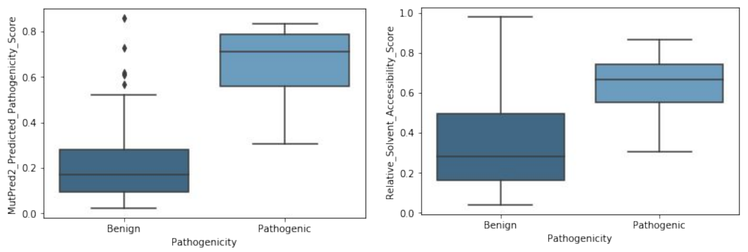
Data Processing

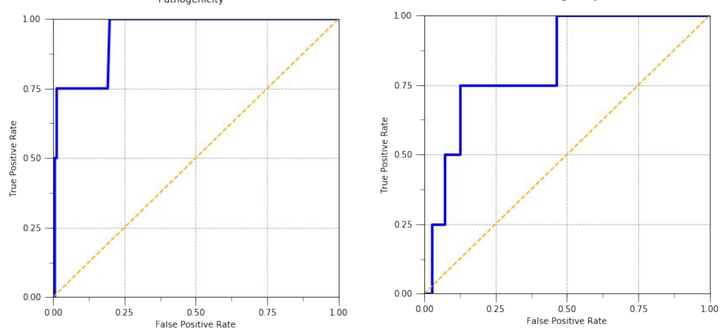


Methods

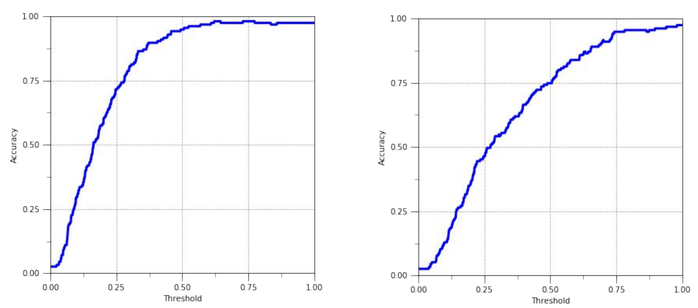
1. We downloaded the MutPred2 Gene Variant Pathogenicity Predictor and successfully reproduced the example output file using the example input file to ensure its concordance. Both files were provided by the developers and included in the software download.
2. We assembled a training set with 386 BRCA1 and BRCA2 gene variants. The training set was first analyzed by another variant pathogenicity predictor called VEP to obtain an amino acid change from each genomic coordinate.
3. We then assembled a data set that contained a protein sequence identifier and it's amino acid change for each variant.
4. The previous data set was used to create the FASTA format input file required for MutPred2 and was used to obtain the relative solvent accessibility at the amino acid change position for each variant.
5. We then assembled another data set that included the MutPred2 pathogenicity prediction and the relative solvent accessibility at the amino acid position. Both of these values are a score between 0 and 1. Clinical-based classifications by ClinVar and ENIGMA wee also included.

Predicted Pathogenicity | **Relative Solvent Accessibility**





AUC = .945 | **AUC = .828**



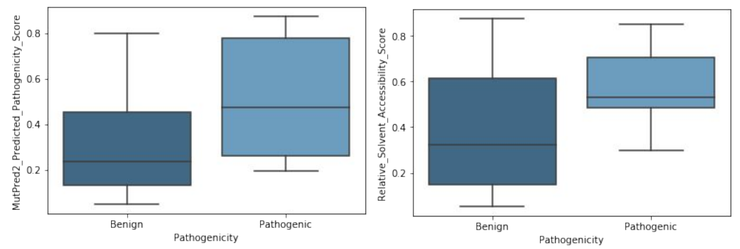
Max Threshold: .981
Median Threshold: .740

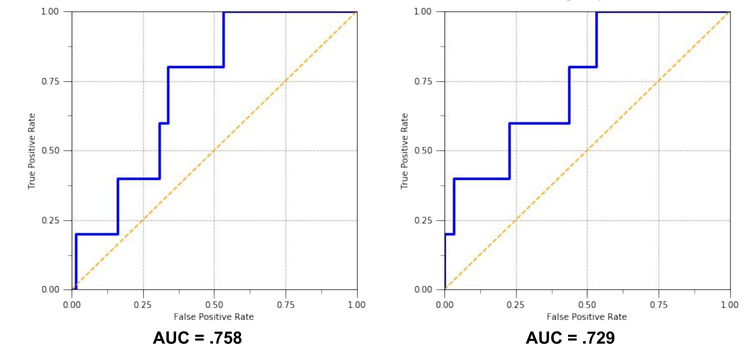
Max Threshold: .974
Median Threshold: .99

6. We selected the median threshold in both instances to use in the test set.

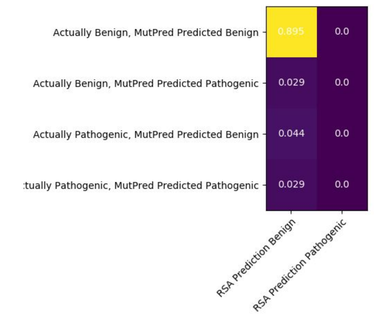
Results

Predicted Pathogenicity | **Relative Solvent Accessibility**





AUC = .758 | **AUC = .729**



Conclusion

At the threshold of .74 for the MutPred2 variant pathogenicity predictions and a threshold of .99 for the Relative Solvent Accessibility Pathogenicity Predictions we see that these thresholds yield AUCs less than those found in the training set. We also found that .029 of variants in the test set were incorrectly predicted pathogenic by MutPred2, were correctly predicted benign by using Relative Solvent Accessibility. Yet, 0 of the variants were incorrectly predicted benign, but were correctly predicted pathogenic by Relative Solvent Accessibility.

Potential Future Work

In the future we hope to construct a larger data set with an increase in number of variants. We would like to do an outlier analysis on the variants MutPred2 predicted. We would consider other data consortiums and other *In Silico* Prediction Software.

Acknowledgements

We thank Judith Canner, Zia Isola, and Steven Kim for their support. We also thank Hongde Hu, Raylene Potter, and Carla Gonzalez for their recommendation. We thank Carla Fresquez and Corin White for their support in research. We thank Jacob Pheil, Zack Fischmann, Amy Coffin, Vikas Pejaver, Kimberly Runyan, Tracey Ramirez, Nina Thoeni, the BD2K group, and the BRCA Exchange team.

References

1. Pejaver V, Urresti J, Lugo-Martinez J, Pagel KA, Lin GN, Nam H, Mort M, Cooper DN, Sebat J, Iakoucheva LM, Mooney SD, Radivojac P. MutPred2: inferring the molecular and phenotypic impact of amino acid variants. bioRxiv 134981; doi: https://doi.org/10.1101/134981.
2. Advanced Software Engineering Ltd. (2017) K-12 Public School Enrollment: State of California {2007-2017}. Retrieved from The Variant Call Format and VCFtools, Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert Handsaker, Gerton Lunter, Gabor Marth, Stephen T. Sherry, Gilean McVean, Richard Durbin and 1000 Genomes Project Analysis Group, Bioinformatics, 2011

Project GitHub Repository: https://github.com/BRCAChallenge/Summer2018
Personal GitHub Repositories: https://github.com/quez5481

R25MD010391-03