predatoR: an R package for network-based mutation impact prediction

Berk Gurdamar¹,* , Osman Ugur Sezerman²

¹Department of Biostatistics and Bioinformatics, Institute of Health Sciences, Acibadem Mehmet Ali Aydinlar University, Istanbul, Turkey
²Department of Biostatistics and Medical Informatics, School of Medicine, Acibadem Mehmet Ali Aydinlar University, Istanbul, Turkey

*To whom correspondence should be addressed.

Abstract

Motivation: Classification of a mutation is important for variant prioritization and diagnostics. However, it is still a challenging task that many mutations are classified as variant of unknown significance. Therefore, in silico tools are required for classifying variants with unknown significance. Over the past decades, several computational methods have been developed but they usually have limited accuracy and high false-positive rates. To address these needs, we developed a new machine learning-based method for calculating the impact of a mutation by converting protein structures to networks and using network properties of the mutated site.

Results: Here, we propose a novel machine learning-based method, predatoR, for mutation impact prediction. The model was trained using both VariBench and ClinVar datasets and benchmarked against currently available methods using the Missense3D datasets. predatoR outperformed 32 different mutation impact prediction methods with an AUROC value of 0.941.

Availability: predatoR tool is available as an open-source R package at GitHub (https://github.com/berkgurdamar/predatoR).

Contact: berkgurdamar@gmail.com

1 Introduction

Single nucleotide variations in the coding regions of the genome can alter the amino acid sequence of a protein and may result in change in the protein structure and function. Such variations are called missense variants, or single amino acid variants (SAVs), which are the most studied form of variations (Yates et al., 2014). SAVs may affect the protein stability and cause alteration in the active site, protein-protein interaction sites and thus affect the function of the protein (Tan et al., 2021).

Developments in the next-generation sequencing technologies have enabled identification of several pathogenic mutations and establishment of disease-causing mutation databases such as ClinVar (Landrum et al., 2014), OMIM (Hamosh et al., 2005) and HGMD (Stenson et al., 2014). These databases facilitated prompt diagnosis of patients if their variants are present in the database. However, majority of the variations are still classified as variant of unknown significance (VUS) due to low frequency of such variants and lack of information about the association of the variant with a certain phenotype. Especially in rare diseases, most of the variants in a patient are not present in the databases. Therefore, there is an imminent need for more accurate impact prediction methods, which will provide a cost-efficient and time-efficient solution in contrast to experimental methods (Chen et al., 2020).
In the last decade, several computational methods and in-silico tools have been developed for mutation impact prediction. Most mutation impact prediction tools exploit the sequence, structure, and thermodynamic information and use several machine learning-based methods for prediction. However, accuracies of the methods remain limited with high false-positive rates (Ancien et al., 2018).

Different approaches have been made, such as GERP++ (Davydov et al., 2010) and SIFT (Ng and Henikoff, 2003) which make predictions using multiple sequence alignment-based conservation measuring. On the other hand, methods such as CADD (Rentzsch et al., 2019) uses a machine learning model with multiple genomic features. Moreover, Rhapsody (Ponzoni et al., 2020) uses structural information along with sequence and thermodynamic-based information. Another approach is the meta-predictors whose main approach is aggregating predictions of other predictors. BayesDel (Feng, 2017), REVEL (Ioannidis et al., 2016), and FATHMM-XF (Rogers et al., 2018) are such examples of meta-predictors. Thus far, network formalization have not been used for mutation impact prediction. ProSNEx (Aydinca et al., 2019) is the only method uses network formalization from protein structures, but the tool provides structure exploration not impact prediction.

Considering the downfalls of the aforementioned in-silico methods, we developed predatoR, a new machine learning-based method for mutation impact prediction. predatoR was implemented in R (R Core Team, 2021), and uses a network formalization approach that accepts protein structures as networks and uses 24 unique features for calculating the impact of a mutation. By putting all the factors that influence the impact of a mutation, we demonstrate that predatoR outperforms 32 different impact prediction methods currently in widespread use.

2 Materials and Methods
2.1 Overview
Main aim of this study was to build a new machine learning-based method for mutation impact prediction by converting protein structures into networks. For calculating network properties, PDB structures were converted into networks by calculating distances between atoms and assigning edges according to interatomic interaction distance cutoffs. Twenty-four unique features were taken into consideration. Machine learning models were trained using VariBench (Sasidharan Nair and Vihinen, 2013) and ClinVar (Landrum et al., 2014) datasets. Seven different machine learning algorithms were employed and the performance of the models were tested using the Missense3D (Ittisoponpisan et al., 2019) dataset. Performance of the novel method was compared with currently available 32 different mutation impact prediction methods by using the predictions on the Missense3D dataset. Details of this study are described in the following sections.

2.2 Data Collection
Neutral datasets (DS2, DS4, DS6, DS8, DS10, DS12, DS14, DS16, MMR and dbSNP), pathogenic datasets (DS3 and DS5) and training dataset of Rhapsody were downloaded from VariBench web server.
Duplicated and conflicting mutations (the same mutation labeled as both pathogenic and neutral) were removed. The combined VariBench dataset contained 6983 neutral and 23563 pathogenic mutations. For increasing the number of neutral mutations and preventing class imbalance, benign and likely benign variants from ClinVar were also collected. Variants in the ClinVar datasets were only mapped on genomic positions. For converting genomic positions to PDB-based positions, VarMap (Stephenson et al., 2019) was used. From the output, PDB position containing variants were filtered and combined with the VariBench dataset. Duplicated, conflicting and synonymous mutations were removed, and the final dataset contained 48005 mutations (24496 neutral and 23509 pathogenic variants).

Experimental datasets of Missense3D were selected for validation of machine learning models and comparison with available impact prediction methods. 2 datasets were downloaded from Missense3D web server. The combined dataset contains 10229 different mutations (4652 neutral, 5577 pathogenic variants). Only UniProt-based positions of the mutations were available in the dataset. For converting UniProt positions to PDB-based positions, the PDDBSWS (Martin, 2005) dataset was used. Variants with PDB-based positions were filtered and common variants between Missense3D dataset and merged ClinVar-VariBench dataset were removed. Final size of the Missense3D dataset decreased to 6554 mutations (3569 neutral, 2985 pathogenic variants).

### 2.3 Processing

For model training, we used 24 different features divided into 3 categories: Network-based, gene-based and amino acid-based. Network-based features were calculated by turning protein structures into networks from PDB structures. For this approach, PDB structures were downloaded by using the Bio3D (Grant et al., 2006) R package. Two different network building approaches were tested: Building network using only carbon alpha (Ca) atoms and all atoms in the structure. For each approach, networks were built using 2 different interatomic distance cutoffs: ≤ 7 angstrom (Å) and ≤ 5Å. According to the approach, edge lists were created by calculating distance between each atom. From the edge list, networks were created for each chain separately via igraph (Csardi and Nepusz) R package. These networks were used for calculating Eigen Centrality, Betweenness Centrality, Degree Centrality, PageRank Centrality and Shortest Path Centrality measures by using the igraph R package and node-specific Clique Score by using in-house scripting. For Shortest Path Centrality, sum of all shortest path lengths of a node were calculated and the sum of the shortest path lengths were used as a score of a node. For node-specific Clique Score, the number of edges between neighbors of a node was calculated for each node separately. For normalizing the scores, scores of network properties were turned into Z-scores for each chain separately for preventing bias between high and low number of atom containing structures. Z-scores of the Ca atoms were accepted as the scores of the positions.

For gene-based features, each mutation was annotated according to its corresponding gene identifier. However, not all the mutations in the datasets contained gene identifier information. For completing the missing gene identifiers, Ensembl BioMart (Smedley et al., 2009) web server was used for mapping PDB-chain IDs to associated gene identifiers. Nine different gene-based features were used. "pLoF Metrics by Gene" dataset was downloaded from gnomAD (Karczewski et al., 2020) web server. From this dataset Missense Z-scores, Synonymous Z-scores and pL1 scores were filtered. All human KEGG pathways and associated
genes were downloaded by using EnrichmentBrowser (Geistlinger et al., 2016) R package and number of pathways associated with each gene were calculated. Ensembl BioMart web server was used for collecting biological process domain related and experimental evidence-coded (EXP, IDA, IPI, IMP, IGI, IEP, HAD, HMP and HEP) GO terms and gene associated GO term numbers were calculated. Disease-Gene association dataset was downloaded from DisGeNET (Piñero et al., 2017) database and gene related disease numbers were computed. Moreover, Genic Intolerance (Petrovski et al., 2013) scores and gene essentiality scores from OGEE v3 (Gurumayum et al., 2021) were downloaded. The median TPM expression dataset of the GTex portal (Lonsdale et al., 2013) was collected and median gene expression values across 54 different tissue types were calculated. Each feature was annotated/calculated according to the associated gene name in the dataset. If there were multiple genes associated with the same PDB-chain ID, the gene having higher gnomAD scores across all 3 metrics was selected.

Nine different amino acid-based features were used. Reference amino acids and the mutant amino acids were used as 2 categorical features. Accessible Surface Area (ASA) and Hydrophobicity Scale values were collected from aaSEA (D.M, 2019) R package. ASA and Hydrophobicity Scale of reference amino acids, mutant amino acids and the difference between mutant and the reference amino acids were calculated. Lastly, BLOSUM62 scores between reference and the mutant amino acids were also used as another feature.

2.4 Model training/testing and comparison

After the processing step (Fig. 1A), the size of the dataset decreased to 34968 mutations (17469 neutral, 17499 pathogenic variants) due to missing information. The dataset was split into 90% training and 10% testing datasets with conserving the proportions of pathogenic and neutral variants. Impact prediction models were built using 7 different machine learning algorithms, XGBoost, Naive Bayes, Random Forest, Elastic Net, GLM, GBM and Adaboost (Fig. 1B). Parameters were optimized through 3 repeated 10-fold cross-validation and for each model, same training-testing datasets were used. For method comparison, the Missense3D dataset went through the same processing steps, as described above (Fig. 1C). The final size of the dataset was 4403 mutations (1803 neutral, 2600 pathogenic variants) after the processing step. For annotating the Missense3D dataset with available impact prediction methods, ANNOVAR (Wang et al., 2010) was used. However, ANNOVAR uses genomic positions as an input. For converting PDB-based positions to genomic positions, TransVar (Zhou et al., 2015) was used. Impact prediction scores of 31 different methods were annotated onto the dataset via ANNOVAR using the DBNSFP 4.2a database. We also collected the pathogenicity predictions from Rhapsody on the Missense3D dataset by using the Rhapsody web server.

3 Results

3.1 Performances of impact prediction models

Performance of the machine learning models were evaluated using ROC curves and AUROC curve values via testing dataset (Fig. 2). According to the AUROC curve values, for all atoms and Ca atoms approaches, 7Å-Adaboost model (AUROC: 0.9762) and 5Å-Adaboost model (AUROC: 0.9808) showed the best performance respectively. These two models were selected as final models and used for comparison with available impact prediction methods.

3.2 Comparison with other methods

Scores of 31 different impact prediction methods were annotated for the Missense3D dataset via ANNOVAR and predictions of Rhapsody were collected using their web server. The performance of all the prediction methods were visualized and evaluated using ROC curves and AUROC curve values (Fig. 3). According to the AUROC curve values, 7Å-all atoms approach used Adaboost model outperformed all other prediction tools with an AUROC curve value of 0.941. Feature importances of the final models represented in the Fig. 4.

3.3 The predatoR R package
The predatoR package takes an input with 5 mandatory and 1 optional argument, PDB ID, chain ID, position (PDB-based), reference and mutant amino acids with 3 letter codes, and gene identifier (optional). If a gene identifier is not provided in the input, associated gene identifier is found using as described before. Package contains 2 models for prediction, 7Å-all atoms approach used Adaboost model and 5Å-Cα atoms approach used Adaboost model. User can decide which model they are going to use. Each utility function of the package can be used separately if necessary. Also, predatoR can be used for exploratory purposes, such as networks can be built with using different cutoffs. In exploratory purposes, all 24 features annotated to the input dataset and returned as the output. However, predictions cannot be made using a cutoff other than 5Å and 7Å. The predatoR package can be installed from GitHub (https://github.com/berkgurdamar/predatoR).

Two different models from two different network formalization approaches were selected as final models and compared with 32 different mutation impact prediction methods with using Missense3D datasets. 7Å-all atom approach used model outperformed all 32 other methods with an AUROC curve value of 0.941. Even though 5Å-Cα atom approach used model could not outperform all the other methods, it is still performed well and showed better AUROC curve value compared with most of the methods. Therefore, it is included in the predatoR package and user can decide which model to use for prediction. In addition, exploratory analysis can be done by using predatoR. Main aim of exploratory analysis is that new models and new approaches can be made with using different interatomic interaction cutoffs or with combination of different features. We hope the community benefits from exploratory analysis.

For model training and testing, VariBench datasets were selected. However, after combining VariBench datasets, we found that most of the datasets contain pathogenic variants and number of neutral variants containing datasets are few or the datasets contain few number of variations. In addition, different datasets have conflicting labels (pathogenic and neutral) for the same mutations and those mutations had to be removed and it resulted in a decrease in the number of mutations. For preventing class imbalance and increase the number of neutral variants, benign and likely benign variants from ClinVar were used. Still, many mutations in the ClinVar dataset were filtered after the conversion from genomic coordinates to PDB positions due to the missing PDB annotations. With an update in the available databases and increase in the number of datasets, better models and methods can be built for mutation impact prediction and more advancements can be done in our field.

In summary, we developed a new machine learning-based method, predatoR, for mutation impact prediction by converting protein structures into networks and making prediction with using 24 different unique features. Apart from performing impact prediction, exploratory analysis also can be done. Networks can be built, and network properties can be calculated with desired thresholds and further analysis can be done. predatoR is presented as an R package and designed as user-friendly as possible.

4 Discussion

In this study, we propose a new machine learning-based method for analyzing protein structures as networks and making predictions on the impact of a mutation with using 24 different unique features. Creating a network model from a protein structure is a novel and promising approach for mutation impact prediction. One of the advantages of this approach is that it uses structural information as network properties and takes less time, needs less computational power than thermodynamic simulations. Two different network formalization approaches (all atoms and Cα atoms) were tested for network building. For setting edges between atoms, two different interatomic distance cutoffs (5Å and 7Å) were tested. According to the results, models built with 7Å cutoff showed better AUROC curve values in most of the models on test dataset. Moreover, final model that uses all atoms approach showed better performance than the Cα atoms on the Missense3D dataset. This shows that, considering more interactions from the protein structure have positive effect on the prediction performance. Despite the better performance of all atom approach, Cα approach can build the networks and make predictions faster due to the fewer number of atoms used in the networks.

Fig. 4. Feature importance plots of the 10 most important feature of the final models. (A) 7Å-all atoms approach used Adaboost model. (B) 5Å-Cα atoms approach used Adaboost model.

Acknowledgements

Funding

European Joint Programme on Rare Diseases (EJP RD) provides B.G. with a monthly scholarship.

Conflict of Interest: none declared.

References

Ancien,F. et al. (2018) Prediction and interpretation of deleterious coding variants in terms of protein structural stability. Sci Rep, 8, 4480.

Aydmkal,R.M. et al. (2019) ProSNEx: a web-based application for exploration and analysis of protein structures using network formalism. Nucleic Acids Research, 47, W471–W476.
Chen, Y. et al. (2020) PremPS: Predicting the impact of missense mutations on protein stability. *PLoS Computational Biology, 16*, e1008543.

Csardi, G. and Nepusz, T. The igraph software package for complex network research. 9.

Davydov, E. V. et al. (2010) Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. *PLoS Computational Biology, 6*, e1001025.

D.M.R. S.S.R. (2019) aaSEA: Amino Acid Substitution Effect Analyser.

Feng, B.-J. (2017) PERCH: a unified framework for disease gene prioritization. *Hum Mutat, 38*, 243–251.

Geistlinger, L. et al. (2016) Bioconductor’s EnrichmentBrowser: seamless navigation through combined results of set- & network-based enrichment analysis. *BMC Bioinformatics, 17*, 45.

Grant, B. J. et al. (2006) Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics, 22*, 2695–2696.

Gurumayum, S. et al. (2021) OGEE v3: Online GEne Essentiality database with increased coverage of organisms and human cell lines. *Nucleic Acids Research, 49*, D998–D1003.

Hamosh, A. et al. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res, 33*, D514–D517.

Ioannidis, N. M. et al. (2016) REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am J Hum Genet, 99*, 877–885.

Ittisoponpisan, S. et al. (2019) Can Predicted Protein 3D Structures Provide Reliable Insights into whether Missense Variants Are Disease Associated? *J Mol Biol, 431*, 2197–2212.

Karczewski, K. J. et al. (2020) The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature, 581*, 434–443.

Landrum, M. J. et al. (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res, 42*, D980–D985.

Lonsdale, J. et al. (2013) The Genotype-Tissue Expression (GTEx) project. *Nat Genet, 45*, 580–585.

Martin, A. C. R. (2005) Mapping PDB chains to UniProtKB entries. *Bioinformatics, 21*, 4297–4301.

Ng, P. C. and Henikoff, S. (2003) SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res, 31*, 3812–3814.

Petrovski, S. et al. (2013) Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes. *PLOS Genetics, 9*, e1003709.

Piñero, J. et al. (2017) DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Research, 45*, D833–D839.

Ponzoni, L. et al. (2020) Rhapsody: predicting the pathogenicity of human missense variants. *Bioinformatics, 36*, 3084–3092.

R Core Team (2021) R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria.

Rentzsch, P. et al. (2019) CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research, 47*, D886–D894.

Rogers, M. F. et al. (2018) FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. *Bioinformatics, 34*, 511–513.

Sasidharan Nair, P. and Vihinen, M. (2013) VarBench: a benchmark database for variations. *Hum Mutat, 34*, 42–49.

Smedley, D. et al. (2009) BioMart – biological queries made easy. *BMC Genomics, 10*, 22.

Stenson, P. D. et al. (2014) The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum Genet, 133*, 1–9.

Stephenson, J. D. et al. (2019) VarMap: a web tool for mapping genomic coordinates to protein sequence and structure and retrieving protein structural annotations. *Bioinformatics, 35*, 4854–4856.

Tan, K. P. et al. (2021) Packpred: Predicting the Functional Effect of Missense Mutations. *Front Mol Biosci, 8*, 646288.

Wang, K. et al. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res, 38*, e164.

Yates, C. M. et al. (2014) SuSPect: Enhanced Prediction of Single Amino Acid Variant (SAV) Phenotype Using Network Features. *J Mol Biol, 426*, 2692–2701.

Zhou, W. et al. (2015) TransVar: a multilevel variant annotator for precision genomics. *Nat Methods, 12*, 1002–1003.
