Received: 1 July 2019 | Revised: 5 November 2019 | Accepted: 20 November 2019

DOI: 10.1002/humu.23961

INFORMATICS

Human Mutation HGVS HUMAN GENOME VARIATION SOCIETY WILEY

Pred-MutHTP: Prediction of disease-causing and neutral mutations in human transmembrane proteins

A. Kulandaisamy[1]* | Jan Zaucha[2]* | Ramasamy Sakthivel[1] | Dmitrij Frishman[2,3] | M. Michael Gromiha[1,4]

[1]Department of Biotechnology, Bhupat and Jyoti Mehta School of BioSciences, Indian Institute of Technology Madras, Chennai, Tamilnadu, India

[2]Department of Bioinformatics, Wissenschaftszentrum Weihenstephan, Technische Universität München, Freising, Germany

[3]Department of Bioinformatics, Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russian Federation

[4]Advanced Computational Drug Discovery Unit, Tokyo Tech World Research Hub Initiative (WRHI), Institute of Innovative Research, Tokyo Institute of Technology, Yokohama, Japan

Correspondence
Dmitrij Frishman, Department of Bioinformatics, Technische Universität München, Wissenschaftszentrum Weihenstephan, Freising 85354, Germany.
Email: d.frishman@wzw.tum.de

M. Michael Gromiha, Department of Biotechnology, Bhupat and Jyoti Mehta School of BioSciences, Indian Institute of Technology Madras, Chennai 600 036, Tamilnadu, India.
Email: gromiha@iitm.ac.in

Funding information
Russian Science Foundation, Grant/Award Number: 16-44-02002; Department of Science and Technology, Ministry of Science and Technology, Grant/Award Number: INT/RUS/RSF/P-09

Abstract

Membrane proteins are unique in that segments thereof concurrently reside in vastly different physicochemical environments: the extracellular space, the lipid bilayer, and the cytoplasm. Accordingly, the effects of missense variants disrupting their sequence depend greatly on the characteristics of the environment of the protein segment affected as well as the function it performs. Because membrane proteins have many crucial roles (transport, signal transduction, cell adhesion, etc.), compromising their functionality often leads to diseases including cancers, diabetes mellitus or cystic fibrosis. Here, we report a suite of sequence-based computational methods "Pred-MutHTP" for discriminating between disease-causing and neutral alterations in their sequence. With a data set of 11,846 disease-causing and 9,533 neutral mutations, we obtained an accuracy of 74% and 78% with 10-fold group-wise cross-validation and test set, respectively. The features used in the models include evolutionary information, physiochemical properties, neighboring residue information, and specialized membrane protein attributes incorporating the number of transmembrane segments, substitution matrices specific to membrane proteins as well as residue distributions occurring in specific topological regions. Across 11 disease classes, the method achieved accuracies in the range of 75–85%. The model designed specifically for the transmembrane segments achieved an accuracy of 85% on the test set with a sensitivity and specificity of 86% and 83%, respectively. This renders our method the current state-of-the-art with regard to predicting the effects of variants in the transmembrane protein segments. Pred-MutHTP allows predicting the effect of any variant occurring in a membrane protein—available at https://www.iitm.ac.in/bioinfo/PredMutHTP/

KEYWORDS

disease-causing, machine learning, missense variant, mutation, neutral, transmembrane proteins

1 | INTRODUCTION

Diminishing costs of sequencing technologies have facilitated determining a large number of genetic variants including insertions, deletions and single nucleotide variants (SNVs). The sheer amounts of data generated allowed elucidating the relationship between

*A. Kulandaisamy and Jan Zaucha contributed equally to this study.

© 2019 Wiley Periodicals, Inc.

genotype and phenotype in genetic diseases (Bamshad et al., 2011; Buermans & Den Dunnen, 2014). Among all types of variants, SNVs are the most common within the human population (Collins, Guyer, & Chakravarti, 1997; Mooney, 2005). Nonsynonymous SNVs, known as missense variants, occur in the coding regions and alter the protein sequence; they have been extensively studied revealing their effects on protein stability, structure or function, and on the strength of interactions among members of protein complexes (Chaudhary, Naganathan, & Gromiha, 2016; Cui, Zhao, & Korkin, 2018; Dobson, 2003; Jemimah, Yugandhar, & Michael Gromiha, 2017; Sahni et al., 2015; Stefl, Nishi, Petukh, Panchenko, & Alexov, 2013; Stone & Sidow, 2005; Teng, Madej, Panchenko, & Alexov, 2009; Witham, Takano, Schwartz, & Alexov, 2011; Brender & Zhang, 2015). Specific missense variants have been implicated in several diseases including cancer, cystic fibrosis, and cardiomyopathy (Anoosha, Sakthivel, & Gromiha, 2016; Ganesan, Kulandaisamy, Priya, & Gromiha, 2019 and Kulandaisamy et al., 2018). However, an exhaustive experimental characterization of the effects of all possible variants is not achievable in the foreseeable future. Hence, a wide number of sequence- and structure-based *in silico* tools have been developed to predict the consequences of missense variants.

Computational approaches can be divided into three categories based on the type of features used in the prediction, namely sequence, structure, and meta. Although structure-based features can, in principle, provide greater accuracy, their limiting factor is the coverage of PDB, which, although constantly accelerating in growth (Burley et al., 2018), does not keep up with the rate of discovery of new sequences (The UniProt Consortium, 2017). Therefore, many commonly used prediction methods have been designed to rely on sequence data alone. Tools such as SIFT (Ng & Henikoff, 2003), PROVEAN (Choi & Chan, 2015), FATHMM (Shihab et al., 2013), PrimateAI (Sundaram et al., 2018), MutationTaster (Schwarz, Cooper, Schuelke, & Seelow, 2014), and MutationAssessor (Reva, Antipin, & Sander, 2011) arrive at their predictions based on evolutionary information supported by the physiochemical properties of the wild type and alternative residues (charge, volume, polarity and hydrophobicity, etc.). To leverage the full potential of the information acquired on proteins to date, several prediction methods such as SNPs3D (Yue, Melamud, & Moult, 2006), SNPeffect (Reumers, Maurer-Stroh, Schymkowitz, & Rousseau, 2006), and Polyphen-2 (Adzhubei, Jordan, & Sunyaev, 2013) have been developed to use both sequence and the structure-based features. The latter include solvent accessibility, energy terms, secondary structure, and intrinsic disorder annotations as well as residue contact networks. These methods have been developed utilizing machine learning algorithms such as naive Bayes, random forest, support vector machines, and neural networks (Mah, Low, & Lee, 2011; Tang & Thomas, 2016). Finally, metamethods, which combine the results from other tools to arrive at a consensus prediction, have been added on top: PredictSNP (Bendl et al., 2014), Meta-SNP (Capriotti, Altman, & Bromberg, 2013), and REVEL (Ioannidis et al., 2016).

Nevertheless, it has recently become apparent that the existing methods are severely limited in their reliability for predicting the effects of variants occurring in membrane proteins (Kulandaisamy, Priya, Sakthivel, Frishman, & Gromiha, 2019). A potential reason for this is that the physiochemical requirements for embedding such polypeptides within the phospholipid bilayer make them significant outliers in the landscape of the entire cellular proteome, which serves as the training set for the universal prediction tools (Almén, Nordström, Fredriksson, & Schiöth, 2009; Gromiha & Ou, 2014). A loss of function of such proteins affected by deleterious mutations leads to various diseases including cystic fibrosis (Cheng et al., 1990) and different types of cancer. While membrane proteins amount to roughly a quarter of the entire proteome, about 50–60% of them are used as drug targets (Overington, Al-Lazikani, & Hopkins, 2006), demonstrating their importance in disease and affirming the need for developing specialized tools for characterizing their variation. Recently, we have developed a method, BorodaTM, for discriminating the effect mutations on membrane proteins by utilizing several structure-based parameters (Popov, Bizin, Gromiha, Kulandaisamy, & Frishman, 2019). However, there is no sequence-based method available for distinguishing between disease-causing and neutral variants specifically in membrane proteins.

In this work, we have developed a sequence-based predictor named Pred-MutHTP, which discriminates between disease-causing and neutral mutations in membrane proteins. Different features such as evolutionary information, contact potentials, substitution matrices and parameters specific to membrane proteins were employed in the prediction. Our method showed an average accuracy of 78% on the training data set in 10-fold group-wise cross-validation. We have also developed topology-specific models (cytosol-, membrane-, and extracellular-specific) and obtained an average accuracy of 80% on the respective test datasets. The comparison of the performance of the present method with existing methods showed an improvement of 4–11% in balanced accuracy. The method allows identifying harmful variants relevant to different diseases and we hope that it will facilitate the development of mutation-specific drugs.

2 | MATERIALS AND METHODS

2.1 | Data collection and curation

We have collected information on missense mutations from the MutHTP database (Kulandaisamy et al., 2018). This integrated database provides various sequence and structural features, topology, number of transmembranes (TM) segments, disease class, and source of the annotation for each mutation. MutHTP includes data from Humsavar (http://www.uniprot.org/docs/Humsavar), SwissVar (Mottaz, David, Veuthey, & Yip, 2010), 1000 Genomes (Genomes Project Consortium, 1000, 2015), COSMIC (Forbes et al., 2014), and ClinVar (Landrum et al., 2014) databases. To ensure the reliability of the annotations used, we have considered only the mutations that are present in at least two databases.

2.2 | Nonredundant data set

To remove the redundancy from the data set, first, we have clustered the proteins based on sequence identity (cut-off of >40%) using CD-HIT (Fu, Niu, Zhu, Wu, & Li, 2012). The representative proteins from each cluster along with their mutations were included in the data set. Further, the mutations in the nonrepresentative sequences in each cluster were scrutinized by checking the conservation of mutation positions; if a mutation position is conserved, we included only one mutation at that position. Finally, for the mutations at the conserved position, we considered the type of amino acid mutation (A→T, G→S, etc.); we have excluded the mutations occurring at the conserved position and featuring the same substitutions. The final data set comprises 11,846 disease-causing and 9,533 neutral mutations, which have occurred in 1,014 and 2,958 proteins, respectively (Supp. Table S1). The complete data set is available at https://www.iitm.ac.in/bioinfo/PredMutHTP/datasets.php.

2.3 | Feature extraction

2.3.1 | Physicochemical properties

We have collected 685 physical, chemical, energetic, and conformational parameters from the following resources: (a) 566 physicochemical properties of 20 amino acids in AAindex1 database (Kawashima & Kanehisa, 2000), (b) 49 physical, chemical, energetic and conformational parameters from literature (Gromiha, 2005) and (c) 70 parameters from the ProtScale server and literature, which includes physicochemical properties as well as the transmembrane, helical and sheet propensity values of 20 amino acids (Gasteiger et al., 2005; Morita et al., 2011; Simm, Einloft, Mirus, & Schleiff, 2016; Xiao & Shen, 2015). Properties with missing data or those deemed irrelevant to this study were excluded. Further, if any two properties had a Pearson correlation of >0.85, we retained only one. This procedure yielded a set of 253 properties, which were normalized into the range between 0 and 1. In each case, the change in the value of a property between the wild type and mutant residue is computed as

$$\Delta P_{\text{mutation}} = P_{\text{mutant}} - P_{\text{wild-type}} \tag{1}$$

where, $P_{\text{wild-type}}$ and P_{mutant} are the property values of wild-type and mutant residues, respectively, and $\Delta P_{\text{mutation}}$ is the change in property due to the mutation.

2.3.2 | Substitution matrices and contact potentials

A set of 94 substitution matrices and 47 pair-wise contact potential matrices of amino acids were collected from AAindex2 and AAindex3 databases (Kawashima & Kanehisa, 2000), respectively. Substitution matrices were directly assigned to each type of mutation. For contact potentials, the difference in an amino acid's contact potential due to a mutation is obtained by subtracting the contact potential value of the N/C-terminal neighbors of the mutated position to the wild-type residue from the values corresponding to the N/C-terminal neighbors to the mutant residue.

2.3.3 | Evolutionary information

The evolutionary information reveals the functional importance of each particular residue within the protein sequence. We computed 18 different types of normalized conservation scores using the standalone AACon tool (Manning, Jefferson, & Barton, 2008; Valdar, 2002). Additionally, three properties were extracted from the position-specific scoring matrix (PSSM) including the value of wild-type residue in mutation site ($\text{PSSM}_{\text{wild-type}}$), the probability of mutant residue located in the mutation position ($\text{PSSM}_{\text{mutant}}$) and the difference between them ($\text{PSSM}_{\text{mutant}} - \text{PSSM}_{\text{wild-type}}$). The PSSM was constructed with three iterations of searching (PSI-BLAST; Altschul et al., 1997) against the UniRef90 (Suzek et al., 2014) database sequences with an *E*-value cutoff of 0.001.

2.3.4 | Neighboring residue-based features

The information on neighboring residues of the mutant residue is included using the following equation:

$$\Delta P_{\text{local}} = P_i(\text{mutant}) - \frac{\sum_{n=\text{i}-j}^{\text{i}+j} P_{\text{n}}(\text{wild-type})}{2j + 1} \tag{2}$$

where $j = (1, 2, 3)$ for window lengths of 3, 5, and 7, respectively. $2j + 1$ is the total window length and "i" is the position of the mutation.

Further, amino acids were grouped into six categories based on physicochemical properties including aliphatic (G, A, L, I, and V), aromatic (F, Y, and W), sulfur-containing (M and C), polar (N, Q, S, T, and P), negatively (D and E), and positively charged (R, H, and K) residues. Additionally, we used the category distributions (the number of residues in each category surrounding the mutation site in window lengths of 3–21) as additional features. On top of that, we considered the residues surrounding the site of the mutation at window lengths of 1–13 (1–6 residues towards each terminus). Finally, the wild type and mutant residue, as well as the mutation position and atomic frequencies in all residues, were also used.

2.3.5 | Membrane protein-based features

We have extracted several membrane protein-specific features including the number of transmembrane segments, the topological location of the mutation site and fractions of residues corresponding to the signal-peptide, cytoplasmic, membrane, and extracellular

TABLE 1 Performance of classification models on a different type of datasets

Data set	Number of features	Data type/validation	SN	SP	ACC	BAC	MCC	AUC
Whole data	20	10-fold-group-wise	76.32	72.46	74.62	74.39	0.48	0.82
		Test	78.14	78.60	78.41	78.31	0.5	0.86
Inside	15	10-fold	72.32	73.54	72.93	72.92	0.454	0.79
		Test	75.44	76.13	75.64	75.75	0.47	0.81
Membrane	15	10-fold-group-wise	81.38	74.81	79.33	78.09	0.54	0.84
		Test	86.64	83.80	85.42	85.21	0.70	0.91
Outside	19	10-fold-group-wise	73.35	74.84	74.50	74.09	0.44	0.81
		Test	78.71	74.90	77.24	76.83	0.53	0.84

Abbreviations: ACC, accuracy; AUC, area under the curve; BAC, balanced accuracy; MCC, Matthews correlation coefficient; SN, sensitivity; SP, specificity.
Note: For details, refer Section 2.

regions. The topology information was retrieved from CCTOP (Dobson, Reményi, & Tusnády, 2015) and TOPCONS (Bernsel, Viklund, Hennerdal, & Elofsson, 2009) servers. In addition, we used the substitution matrices for whole human membrane proteins as well as additional ones specific to topological regions including the cytosol, TM, and extracellular space from our previous study (Kulandaisamy et al., 2019).

2.3.6 | Feature selection and classification

For feature selection and classification, we have utilized the Waikato Environment for Knowledge Analysis (WEKA) platform (Hall et al., 2009). In WEKA, features were selected using the following two methods: (a) CfsSubsetEval evaluator with the genetic and BestFirst search and (b) Consistency subset evaluator with the genetic and BestFirst search. Next, the classification was carried out by all the available methods in WEKA and based on the performance in the cross-validation, we have selected the voting algorithm for classifying mutations into disease-causing or neutral. For the whole data model, the voting algorithm utilizes the classification (disease/neutral) obtained by MultilayerPerceptron, RandomForest, logistic regression and classification via regression (based on REPTree) algorithms independently and gives the average result for discrimination of disease-causing and neutral mutations (Supp. Table S1). All algorithms were trained using the default values of all parameters in WEKA (Supp. Table S2).

2.3.7 | Model evaluation and validation

The classification models were assessed by 10-fold group-wise cross-validation in which mutations were split according to protein sequence identity clusters and divided into 10 groups. We have considered each cluster as a single group. This ensures that the proteins used for the training are evolutionarily independent from the proteins used in the validation set at each fold of the cross-validation procedure. We have evaluated performance measures including sensitivity, specificity, accuracy, balanced accuracy and Matthew's correlation coefficient (MCC) to assess the reliability of each classification model; the equations for all of the aforementioned measures are given in the supplementary information.

The disease-causing mutations are considered as the positive class and neutral mutations as the negative class. In addition, we plotted the true positive rate against the false positive rate (receiver operating characteristic curve), which is used to estimate the trade-off between sensitivity and specificity at different thresholds. Finally, the model was tested using a holdout test data set (20%), which is evolutionarily independent of the training set.

3 | RESULTS AND DISCUSSION

3.1 | Discrimination of human membrane protein mutations

We have computed 19 different types of features from various categories including evolutionary information, physiochemical properties, neighboring residue information, and membrane protein-specific parameters (see below) and utilized WEKA for feature selection and developing computational models (whole data) to discriminate between disease-causing and neutral mutations in membrane proteins. In this model, we have considered all variants in human transmembrane proteins together, irrespective of their topological locations. The features selected in the final models include conservation scores at mutation sites, PSSM values of the mutant residue, difference between the PSSM value of mutant and wild-type residues, the Grantham score, hydrophobicity, polarity, flexibility, Chou-Fasman sheet propensity, number of polar residues around the mutation site, first residue type at the mutation site towards the N-terminal, membrane protein-specific features including the specialized substitution matrix (SLIM) for transmembrane proteins, number of transmembrane (TM) segments, topological class of the mutation site, and the fraction of residues in each topological region. The combination of the above features could discriminate between disease-causing and neutral mutations in human membrane proteins with a sensitivity, specificity, accuracy, and AUC of 76.3%, 72.4%, 74.6%, and 0.82, respectively, on 10-fold group-wise cross-validation (Table 1 and Figure 1).

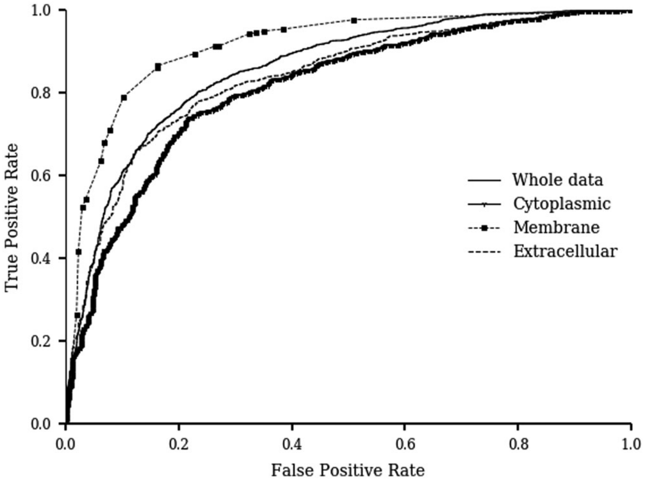

FIGURE 1 Receiver operating characteristic curve for the classification models for whole data, cytoplasmic, membrane, and extracellular mutants

3.1.1 | Performance of Pred-MutHTP on a different type of datasets

Since mutations are not evenly distributed over individual or groups of proteins, we grouped the mutations based on function, number of transmembrane segments of a protein, number of mutations linked with a particular protein, and different disease classes for a condition-specific evaluation of the method.

Functions: Based on functions, the data set was classified into five groups including enzymes, receptors, transporters, GPCR family proteins, and others. For enzymes and transporters, we have obtained a balanced accuracy of 81% and 83%, respectively. Among these five groups, GPCR proteins are widely used as drug targets, and our method could correctly predict 85% of their mutations. Similarly, we have achieved a balanced accuracy of 84% in receptors and 86% in the others class (Figure 2). Altogether, this indicates that our method consistently provides reliable predictions regardless of the protein functional class.

Number of mutations annotated in proteins: The data set was grouped into seven classes based on the number of mutations annotated within a particular protein. For proteins having 1–10 mutations we have obtained a balanced accuracy of 83%; in proteins with 31 to 50 and 51 to 100 mutations, our method could discriminate between the disease-causing and neutral mutations with a balanced accuracy of 87% and 85%, respectively. The accuracy was up to 81% for the proteins with more than 100 mutations (Figure 2).

Number of transmembrane segments: Mutations were divided into 17 groups based on the number of transmembrane segments (TM) present in a particular protein. Here, our method showed accuracies in the range of 82–94%. Specifically, the accuracy is 82–86% for proteins with 1, 2, 3, 7, 9, 15–20, 23–29 TM segments and >86% of accuracy for 4, 5, 6, 8, 10, 11, 13, 14, and 22 TM segments (Figure 3).

Disease classes: We classified the mutations into 11 groups based on their associations with a specific disease type using the information available in the MutHTP database. For the cardiovascular, nervous system, and urinary system diseases, we have obtained a balanced accuracy of >80%. For rest of the disease classes, the balanced accuracy ranges from 74% to 79% (Figure 4).

3.2 | Discrimination of mutations in different topological regions

Based on the nature of the environment, the membrane protein mutations are grouped into three regions including the cytoplasmic (inside the cell or organelle), membrane and extracellular space (outside of the cell cell or organelle), and we developed prediction models for each case separately.

For the cytoplasmic regions, we used 4,416 and 2,958 disease-causing and neutral mutations, respectively. A set of 15 features, which include the PSSM profile, contact potentials, change in compressibility, hydrophobicity, conservation score, number of transmembrane segments and fraction of residues in the inside and outside regions of the cell allowed discriminating between the disease-causing and neutral mutations with a sensitivity, specificity and accuracy of 75%, 76%, and 76%, respectively, for test set

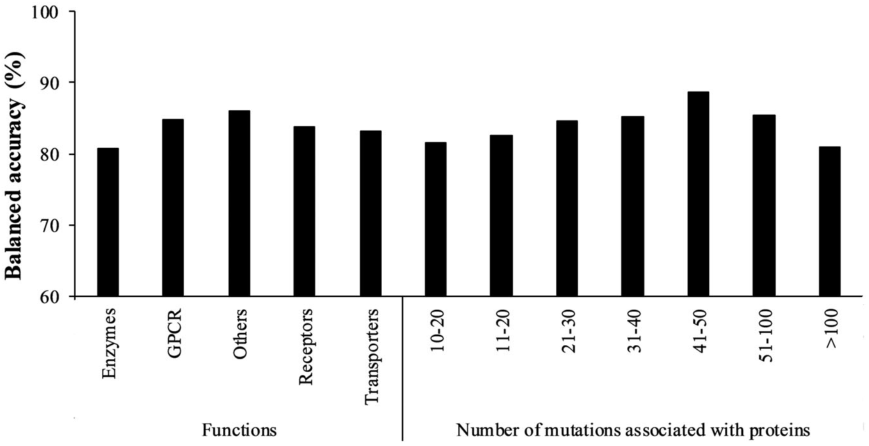

FIGURE 2 Performance of Pred-MutHTP on different functions and number of mutations associated with a particular protein

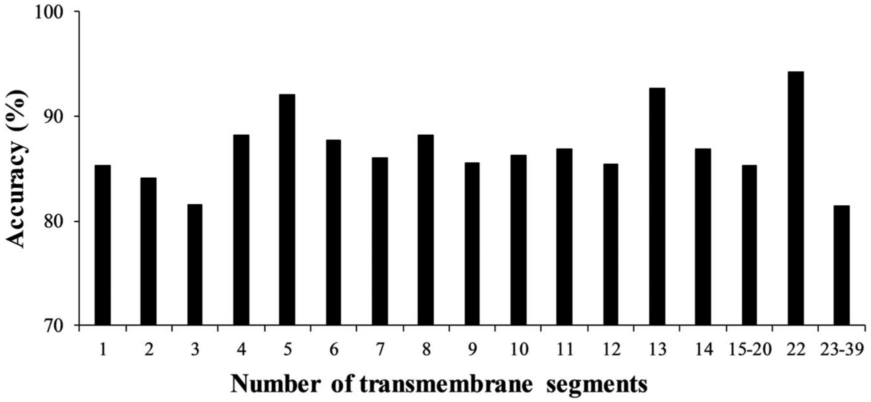

FIGURE 3 Performance of Pred-MutHTP on number of transmembrane segments

(Table 1). On the other hand, for the extracellular region, our method achieved accuracy and AUC of 74% and 0.8, respectively, on 10-fold group-wise cross-validation using a combination of 19 features (Figure 1 and Supp. Table S3).

For the TM region, we used 2,421 disease-causing and 1,285 neutral mutations. Our method utilized the following features: transmembrane tendency, molecular weight, isoelectric point, different substitution matrices, PSSM value of the mutant residue, difference between the wild-type and mutant PSSM value, conservation score, number of TM segments, fraction of residues in different topological regions (Supp. Table S3) and obtained an accuracy of 79% in the 10-fold group-wise cross-validation (Table 1).

3.3 | Relevance of selected features for predicting the effects of mutations

We have developed a classification model for discriminating between disease-causing and neutral mutations in human membrane proteins and specific models for different topological regions using numerous evolutionary data, physiochemical properties, contact potentials, neighboring residues, and membrane protein-based features (Supp. Table S3). In all the models, evolutionary features including the PSSM profile, and conservation scores are the predominant sources of information used for arriving at the prediction; these features represent the paradigm of predicting the effects of mutations in proteins (Bromberg & Rost, 2009; Choi & Chan, 2015; Ng & Henikoff, 2003). We found that physiochemical, energetic, and conformational properties such as polarity, molecular weight, hydrophobicity, flexibility, total nonbonded energy, and transmembrane helix forming propensity also influence the mutation's tendency towards being disease-causing or neutral. These features have been previously reported to play an important role in determining the changes in protein stability and functional effects upon mutations (Folkman, Stantic, Sattar, & Zhou, 2016; Xiao & Shen, 2015). Contact potentials and neighboring residues information, which have also been used in our prediction, are important for capturing the short-range interactions in protein folding and stability (Cserzö & Simon, 1989) and have already been reported to be important for predicting the effect of somatic mutations in cancers (Anoosha, Huang, Sakthivel, Karunagaran, & Gromiha, 2015). Apart from these commonly used features, the membrane protein-specific features including the number of TM segments, topology, and the fraction of residues in different topological regions provide additional information for arriving at reliable results specific to the characteristics of membrane proteins. We also point out that the key to the performance of our method is the carefully selected training set, which provides an unbiased (nonredundant) set of features specific to the membrane protein class that has never before been used to train a predictor of the effects of mutations.

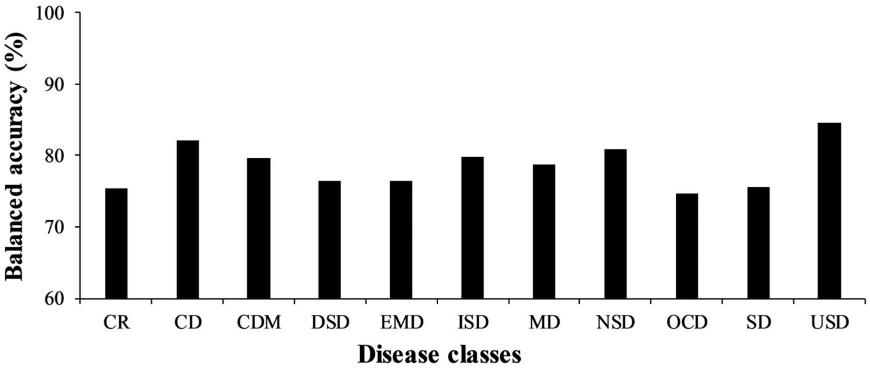

FIGURE 4 Performance of Pred-MutHTP on different disease classes. CD, cardiovascular diseases; CDM, congenital disorders of metabolism; CR, cancers; DSD, digestive system diseases; EMD, endocrine and metabolic diseases; ISD, immune system diseases; MD, musculoskeletal diseases; NSD, nervous system diseases; OCD, other congenital disorders; SD, skin diseases; USD, urinary system diseases

The feature importance of the model was determined by removing one feature at a time and evaluating the sensitivity, specificity, accuracy, balanced accuracy, and Matthew's correlation coefficient (MCC) change associated with the loss of that feature. The results are presented in Supp. Table S4 and show a maximum of 1–2% deviation with respect to the full model. These results suggest that no single feature is uniquely responsible for the major part of the prediction accuracy.

3.4 | Comparison of our method with existing generic methods

The model performance of our method was compared with 11 existing methods including FATHMM-weighted, FATHMM-Unweighted, FATHMM-MKL (Shihab et al., 2013), FATHMM-XF (Rogers et al., 2017), Polyphen2-Humvar, Polyphen2-Humdiv (Adzhubei et al., 2013), PROVEAN (Choi & Chan, 2015), SIFT (Ng & Henikoff, 2003), PrimateAI (Sundaram et al., 2018), MutationTaster (Schwarz et al., 2014), and MutationAssessor (Reva et al., 2011). For FATHMM-weighted, FATHMM-Unweighted, Polyphen2-Humvar, Polyphen2-Humdiv, PROVEAN, SIFT, and MutationAssessor methods, we retrieved the prediction results directly from each method using the UniProt id along with the mutation information and for FATHMM-MKL, FATHMM-XF, PrimateAI, MutationTaster the precomputed predictions were extracted from the dbNSFP v4.0 database (Liu, Wu, Li, & Boerwinkle, 2016).

To compare the performance of our method (Pred-MutHTP) with the above methods, we have identified the variants, which are not used in the training of the respective methods. Further, the proteins are clustered based on sequence identity and the performance of FATHMM-weighted, FATHMM-Unweighted, Polyphen2-Humvar, Polyphen2-Humdiv methods are measured with respect to a set of nonredundant evolutionarily independent protein mutations. For the other methods, our test set was used for comparing the performance. We observed that the performance of our prediction model is about 4–11% higher than that of the pre-existing methods published in the literature (Figure 5). Further, we divided the data set based on functions, number of mutations associated with a particular protein, number of TM segments, different disease classes, and compared the performance with other existing methods (Supp. Tables S5, S6, and S7). The average accuracy of our new method was 2–40% higher in the sub-groups related to specific functions, number of TM segments and number of mutations associated with each protein.

3.5 | Web server development

We have developed four classification models to effectively discriminate between the effects of missense variants in membrane proteins accounting for the topological location of the mutation site. Pred-MutHTP is freely available and can be accessed at https://www.iitm.ac.in/bioinfo/PredMutHTP/. The user can predict the effect of any mutation in a membrane protein by inputting the protein's UniProt id along with the information specifying the mutation (position and alternative amino acid residue). The Pred-MutHTP webserver automatically extracts the topology information from precomputed topology data and predicts the effect of a mutation in the membrane and other soluble regions using respective classification models. The precomputed topology information is retrieved from CCTOP (Dobson et al., 2015) and TOPCONS (Bernsel et al., 2009) servers. Pred-MutHTP provides the output in a table format, which includes a query, topology information, classification of the mutation (disease-associated or neutral), and the confidence score of each prediction. Moreover, users can download the results of their queries.

Further, we have precomputed the predictions of all possible amino acid variants in membrane protein sequences; the prediction results are available at https://www.iitm.ac.in/bioinfo/PredMutHTP/pred_db_search.php. Users can search for any specific protein or download the entire data set.

3.6 | Applications of Pred-MutHTP

In the age of cheap next-generation sequencing technologies, the characterization of the impacts of protein-coding variation in the

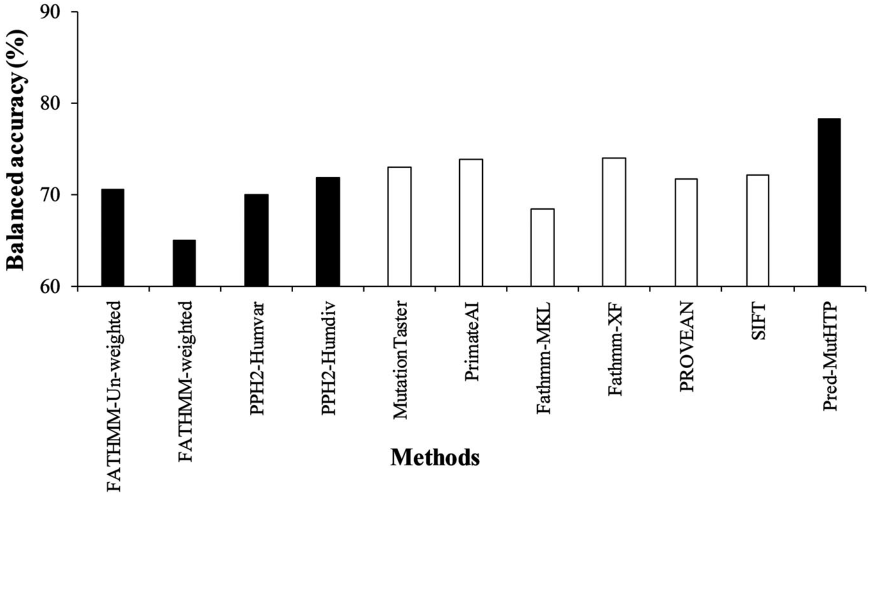

FIGURE 5 Comparison of the present work with other existing methods. Filled bars represent the accuracy achieved on the set of evolutionarily independent protein variants in the data set, which were not used in the model training; empty bars represent the accuracy obtained for the test set used in the present study

human population is of paramount importance for the advancement of medicine. In this context, the presented method Pred-MutHTP has the following potential applications: First of all, the developed method relies only on the proteins primary structure for predicting the effects of mutations. This is crucially important because less than 1% of membrane proteins have structural coverage within the Protein Data Bank (PDB; Scott, Kummer, Tremmel, & Plückthun, 2013). This is the first method designed especially for membrane proteins, which utilizes numerous membrane protein-specific features. The separate classification models specialized for each topological region provide further specificity towards the physicochemical properties characteristic to each environment. The developed tool will help reliably evaluating the effects of experimentally unannotated missense mutations occurring in membrane proteins and potentially aid the design of personalized medicine strategies.

4 | CONCLUSIONS

We have constructed a representative data set for membrane protein mutations extracted from the MutHTP database and developed a sequence-based predictor for discriminating between the disease-causing and neutral mutations in human membrane proteins using an ensemble of machine learning techniques. For the discrimination, we have utilized sequence-based features such as different physiochemical properties of the mutation, the number of neighboring residues at different window lengths, evolutionary information, contact potentials, as well as features specific to membrane proteins. On the whole data set, we achieved an average accuracy of 74.6% in 10-fold group-wise cross-validation and 78.4% on the test set. Further, we have grouped the mutations based on different topological regions of membrane proteins and obtained an average accuracy of 79% in 10-fold group-wise cross-validation. The performance of our prediction method is better than that of other existing methods in the literature, rendering it the current state-of-the-art for predicting the effects of mutations in membrane proteins.

ACKNOWLEDGMENTS

The authors thank the Bioinformatics facility, Department of Biotechnology and Indian Institute of Technology Madras for computational facilities. The work is partially funded by the Department of Science and Technology, Government of India (INT/RUS/RSF/P-09) to M. M. Gromiha and of the Russian Science Foundation for a research grant (16–44-02002) to D. Frishman. A. Kulandaisamy was supported by the Ministry of Human Resource and Development, India.

DATA AVAILABILITY STATEMENT

The complete data set used in this work is available at https://www.iitm.ac.in/bioinfo/PredMutHTP/datasets.php.

ORCID

Jan Zaucha http://orcid.org/0000-0003-3289-4590
Dmitrij Frishman http://orcid.org/0000-0002-9006-4707
M. Michael Gromiha http://orcid.org/0000-0002-1776-4096

REFERENCES

Adzhubei, I., Jordan, D. M., & Sunyaev, S. R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. *Current Protocols in Human Genetics*, *76*(1), 7–20. https://doi.org/10.1002/0471142905.hg0720s76

Almén, M. S., Nordström, K. J., Fredriksson, R., & Schiöth, H. B. (2009). Mapping the human membrane proteome: A majority of the human membrane proteins can be classified according to function and evolutionary origin. *BMC Biology*, *7*(1), 50. https://doi.org/10.1186/1741-7007-7-50

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–3402. https://doi.org/10.1093/nar/25.17.3389

Anoosha, P., Sakthivel, R., & Gromiha, M. M. (2016). Exploring preferred amino acid mutations in cancer genes: Applications to identify potential drug targets. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, *1862*(2), 155–165. https://doi.org/10.1016/j.bbadis.2015.11.006

Anoosha, P., Huang, L. T., Sakthivel, R., Karunagaran, D., & Gromiha, M. M. (2015). Discrimination of driver and passenger mutations in epidermal growth factor receptor in cancer. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, *780*, 24–34. https://doi.org/10.1016/j.mrfmmm.2015.07.005

Bamshad, M. J., Ng, S. B., Bigham, A. W., Tabor, H. K., Emond, M. J., Nickerson, D. A., & Shendure, J. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. *Nature Reviews Genetics*, *12*(11), 745–755. https://doi.org/10.1038/nrg3031

Bendl, J., Stourac, J., Salanda, O., Pavelka, A., Wieben, E. D., Zendulka, J., … Damborsky, J. (2014). PredictSNP: Robust and accurate consensus classifier for prediction of disease-related mutations. *PLoS Computational Biology*, *10*(1), e1003440.https://doi.org/10.1371/journal.pcbi.1003440

Bernsel, A., Viklund, H., Hennerdal, A., & Elofsson, A. (2009). TOPCONS: Consensus prediction of membrane protein topology. *Nucleic Acids Research*, *37*(suppl_2), W465–W468. https://doi.org/10.1093/nar/gkp363

Brender, J. R., & Zhang, Y. (2015). Predicting the effect of mutations on protein-protein binding interactions through structure-based interface profiles. *PLOS Computational Biology*, *11*(10), https://doi.org/10.1371/journal.pcbi.1004494. e1004494.

Bromberg, Y., & Rost, B. (2009). Correlating protein function and stability through the analysis of single amino acid substitutions. *BMC Bioinformatics*, *10*(8), S8. https://doi.org/10.1186/1471-2105-10-S8-S8

Buermans, H. P. J., & Den Dunnen, J. T. (2014). Next generation sequencing technology: Advances and applications. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, *1842*(10), 1932–1941. https://doi.org/10.1016/j.bbadis.2014.06.015

Burley, S. K., Berman, H. M., Christie, C., Duarte, J. M., Feng, Z., Westbrook, J., … Zardecki, C. (2018). RCSB Protein Data Bank: Sustaining a living digital data resource that enables breakthroughs in scientific research and biomedical education. *Protein Science*, *27*(1), 316–330. https://doi.org/10.1002/pro.3331

Capriotti, E., Altman, R. B., & Bromberg, Y. (2013). Collective judgment predicts disease-associated single nucleotide variants. *BMC Genomics*, *14*(3), S2. https://doi.org/10.1186/1471-2164-14-S3-S2

Chaudhary, P., Naganathan, A. N., & Gromiha, M. M. (2016). Prediction of change in protein unfolding rates upon point mutations in two state proteins. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, *1864*(9), 1104–1109. https://doi.org/10.1016/j.bbapap.2016.06.001

Cheng, S. H., Gregory, R. J., Marshall, J., Paul, S., Souza, D. W., White, G. A., … Smith, A. E. (1990). Defective intracellular transport and processing of CFTR is the molecular basis of most cystic fibrosis. *Cell*, *63*(4), 827–834. https://doi.org/10.1016/0092-8674(90)90148-8

Choi, Y., & Chan, A. P. (2015). PROVEAN web server: A tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics*, *31*(16), 2745–2747. https://doi.org/10.1093/bioinformatics/btv195

Collins, F. S., Guyer, M. S., & Chakravarti, A. (1997). Variations on a theme: Cataloging human DNA sequence variation. *Science*, *278*(5343), 1580–1581. https://doi.org/10.1126/science.278.5343.1580

Cserzö, M., & Simon, I. (1989). Regularities in the primary structure of proteins. *International Journal of Peptide and Protein Research*, *34*(3), 184–195. https://doi.org/10.1111/j.1399-3011.1989.tb00229.x

Cui, H., Zhao, N., & Korkin, D. (2018). Multilayer view of pathogenic SNVs in human interactome through in silico edgetic profiling. *Journal of Molecular Biology*, *430*(18), 2974–2992. https://doi.org/10.1016/j.jmb.2018.07.012

Dobson, C. M. (2003). Protein folding and misfolding. *Nature*, *426*(6968), 884–890. https://doi.org/10.1038/nature02261

Dobson, L., Reményi, I., & Tusnády, G. E. (2015). CCTOP: A consensus constrained TOPology prediction web server. *Nucleic Acids Research*, *43*(W1), W408–W412. https://doi.org/10.1093/nar/gkv451

Folkman, L., Stantic, B., Sattar, A., & Zhou, Y. (2016). EASE-MM: Sequence-based prediction of mutation-induced stability changes with feature-based multiple models. *Journal of Molecular Biology*, *428*(6), 1394–1405. https://doi.org/10.1016/j.jmb.2016.01.012

Forbes, S. A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., … Kok, C. Y. (2014). COSMIC: Exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Research*, *43*(D1), D805–D811. https://doi.org/10.1093/nar/gku1075

Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics*, *28*(23), 3150–3152. https://doi.org/10.1093/bioinformatics/bts565

Ganesan, K., Kulandaisamy, A., Priya, S. B., & Gromiha, M. M. (2019). HuVarBase: A human variant database with comprehensive information at gene and protein levels. *PLOS One*, *14*(1), https://doi.org/10.1371/journal.pone.0210475. e0210475.

Gasteiger, E., Hoogland, C., Gattiker, A., Wilkins, M. R., Appel, R. D., & Bairoch, A. (2005). Protein identification and analysis tools on the ExPASy server, *In the proteomics protocols handbook* (pp. 571–607). Totowa, NJ: Humana Press. https://doi.org/10.1385/1-59259-890-0:571

Gromiha, M. M. (2005). A statistical model for predicting protein folding rates from amino acid sequence with structural class information. *Journal of Chemical Information and Modeling*, *45*(2), 494–501. https://doi.org/10.1021/ci049757q

Gromiha, M. M., & Ou, Y. Y. (2014). Bioinformatics approaches for functional annotation of membrane proteins. *Briefings in Bioinformatics*, *15*(2), 155–168. https://doi.org/10.1093/bib/bbt015

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. *ACM SIGKDD explorations newsletter*, *11*(1), 10–18.

Ioannidis, N. M., Rothstein, J. H., Pejaver, V., Middha, S., McDonnell, S. K., Baheti, S., … Cannon-Albright, L. A. (2016). REVEL: An ensemble method for predicting the pathogenicity of rare missense variants. *The American Journal of Human Genetics*, *99*(4), 877–885. https://doi.org/10.1016/j.ajhg.2016.08.016

Jemimah, S., Yugandhar, K., & Michael Gromiha, M. (2017). PROXiMATE: A database of mutant protein–protein complex thermodynamics and kinetics. *Bioinformatics*, *33*(17), 2787–2788. https://doi.org/10.1093/bioinformatics/btx312

Kawashima, S., & Kanehisa, M. (2000). AAindex: Amino acid index database. *Nucleic Acids Research*, *28*(1), 374–374. https://doi.org/10.1093/nar/28.1.374

Kulandaisamy, A., Priya, S. B., Sakthivel, R., Frishman, D., & Gromiha, M. M. (2019). Statistical analysis of disease-causing and neutral mutations in human membrane proteins. *Proteins: Structure, Function, and Bioinformatics*, *87*(6), 452–466. https://doi.org/10.1002/prot.25667

Kulandaisamy, A., Binny Priya, S., Sakthivel, R., Tarnovskaya, S., Bizin, I., Hönigschmid, P., … Gromiha, M. M. (2018). MutHTP: Mutations in human transmembrane proteins. *Bioinformatics*, *34*(13), 2325–2326. https://doi.org/10.1093/bioinformatics/bty054

Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., & Maglott, D. R. (2014). ClinVar: Public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research*, *42*(D1), D980–D985. https://doi.org/10.1093/nar/gkt1113

Liu, X., Wu, C., Li, C., & Boerwinkle, E. (2016). dbNSFP v3. 0: A one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Human Mutation*, *37*(3), 235–241. https://doi.org/10.1002/humu.22932

Mah, J. T., Low, E. S., & Lee, E. (2011). In silico SNP analysis and bioinformatics tools: A review of the state of the art to aid drug discovery. *Drug Discovery Today*, *16*(17-18), 800–809. https://doi.org/10.1016/j.drudis.2011.07.005

Manning, J. R., Jefferson, E. R., & Barton, G. J. (2008). The contrasting properties of conservation and correlated phylogeny in protein functional residue prediction. *BMC Bioinformatics*, *9*(1), 51. https://doi.org/10.1186/1471-2105-9-51

Mooney, S. (2005). Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis. *Briefings in Bioinformatics*, *6*(1), 44–56. https://doi.org/10.1093/bib/6.1.44

Morita, M., Katta, A. M., Ahmad, S., Mori, T., Sugita, Y., & Mizuguchi, K. (2011). Lipid recognition propensities of amino acids in membrane proteins from atomic resolution data. *BMC Biophysics*, *4*(1), 21. https://doi.org/10.1186/2046-1682-4-21

Mottaz, A., David, F. P., Veuthey, A. L., & Yip, Y. L. (2010). Easy retrieval of single amino-acid polymorphisms and phenotype information using SwissVar. *Bioinformatics*, *26*(6), 851–852. https://doi.org/10.1093/bioinformatics/btq028

Ng, P. C., & Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Research*, *31*(13), 3812–3814. https://doi.org/10.1093/nar/gkg509

Overington, J. P., Al-Lazikani, B., & Hopkins, A. L. (2006). How many drug targets are there? *Nature Reviews Drug Discovery*, *5*(12), 993–996. https://doi.org/10.1038/nrd2199

Popov, P., Bizin, I., Gromiha, M., Kulandaisamy, A., & Frishman, D. (2019). Prediction of disease-associated mutations in the transmembrane regions of proteins with known 3D structure. *PLOS One*, *14*(7), e0219452. https://doi.org/10.1371/journal.pone.0219452

Reumers, J., Maurer-Stroh, S., Schymkowitz, J., & Rousseau, F. (2006). SNPeffect v2. 0: A new step in investigating the molecular phenotypic effects of human non-synonymous SNPs. *Bioinformatics*, *22*(17), 2183–2185. https://doi.org/10.1093/bioinformatics/btl348

Reva, B., Antipin, Y., & Sander, C. (2011). Predicting the functional impact of protein mutations: Application to cancer genomics. *Nucleic Acids Research*, *39*(17), e118–e118. https://doi.org/10.1093/nar/gkr407

Rogers, M. F., Shihab, H. A., Mort, M., Cooper, D. N., Gaunt, T. R., & Campbell, C. (2017). FATHMM-XF: Accurate prediction of pathogenic point mutations via extended features. *Bioinformatics*, *34*(3), 511–513. https://doi.org/10.1093/bioinformatics/btx536

Sahni, N., Yi, S., Taipale, M., Bass, J. I. F., Coulombe-Huntington, J., Yang, F., … Kovács, I. A. (2015). Widespread macromolecular interaction perturbations in human genetic disorders. *Cell*, *161*(3), 647–660. https://doi.org/10.1016/j.cell.2015.04.013

Schwarz, J. M., Cooper, D. N., Schuelke, M., & Seelow, D. (2014). MutationTaster2: Mutation prediction for the deep-sequencing age. *Nature Methods*, *11*(4), 361–362. https://doi.org/10.1038/nmeth.2890

Scott, D. J., Kummer, L., Tremmel, D., & Plückthun, A. (2013). Stabilizing membrane proteins through protein engineering. *Current Opinion in Chemical Biology*, *17*(3), 427–435. https://doi.org/10.1016/j.cbpa.2013.04.002

Shihab, H. A., Gough, J., Cooper, D. N., Stenson, P. D., Barker, G. L., Edwards, K. J., … Gaunt, T. R. (2013). Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Human Mutation*, *34*(1), 57–65. https://doi.org/10.1002/humu.22225

Simm, S., Einloft, J., Mirus, O., & Schleiff, E. (2016). 50 years of amino acid hydrophobicity scales: Revisiting the capacity for peptide classification. *Biological Research*, *49*(1), 31. https://doi.org/10.1186/s40659-016-0092-5

1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Stefl, S., Nishi, H., Petukh, M., Panchenko, A. R., & Alexov, E. (2013). Molecular mechanisms of disease-causing missense mutations. *Journal of Molecular Biology*, *425*(21), 3919–3936. https://doi.org/10.1016/j.jmb.2013.07.014

Stone, E. A., & Sidow, A. (2005). Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome Research*, *15*(7), 978–986. https://doi.org/10.1101/gr.3804205

Sundaram, L., Gao, H., Padigepati, S. R., McRae, J. F., Li, Y., Kosmicki, J. A., … Xu, J. (2018). Predicting the clinical impact of human mutation with deep neural networks. *Nature Genetics*, *50*(8), 1161–1170. https://doi.org/10.1038/s41588-018-0167-z

Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., & Wu, C. H., UniProt Consortium. (2014). UniRef clusters: A comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, *31*(6), 926–932. https://doi.org/10.1093/bioinformatics/btu739

Tang, H., & Thomas, P. D. (2016). Tools for predicting the functional impact of nonsynonymous genetic variation. *Genetics*, *203*(2), 635–647. https://doi.org/10.1534/genetics.116.190033

Teng, S., Madej, T., Panchenko, A., & Alexov, E. (2009). Modeling effects of human single nucleotide polymorphisms on protein-protein interactions. *Biophysical Journal*, *96*(6), 2178–2188. https://doi.org/10.1016/j.bpj.2008.12.3904

The UniProt Consortium. (2017). UniProt: The universal protein knowledgebase. *Nucleic Acids Research*, *45*(D1), D158–D169. https://doi.org/10.1093/nar/gkw1099

Valdar, W. S. (2002). Scoring residue conservation. *Proteins: Structure, Function, and Bioinformatics*, *48*(2), 227–241. https://doi.org/10.1002/prot.10146

Witham, S., Takano, K., Schwartz, C., & Alexov, E. (2011). A missense mutation in CLIC2 associated with intellectual disability is predicted by in silico modeling to affect protein stability and dynamics. *Proteins: Structure, Function, and Bioinformatics*, *79*(8), 2444–2454. https://doi.org/10.1002/prot.23065

Xiao, F., & Shen, H. B. (2015). Prediction enhancement of residue real-value relative accessible surface area in transmembrane helical proteins by solving the output preference problem of machine learning-based predictors. *Journal of Chemical Information and Modeling*, *55*(11), 2464–2474. https://doi.org/10.1021/acs.jcim.5b00246

Yue, P., Melamud, E., & Moult, J. (2006). SNPs3D: Candidate gene and SNP selection for association studies. *BMC Bioinformatics*, *7*(1), 166. https://doi.org/10.1186/1471-2105-7-166

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Kulandaisamy A, Zaucha J, Sakthivel R, Frishman D, Michael Gromiha M. Pred-MutHTP: Prediction of disease-causing and neutral mutations in human transmembrane proteins. *Human Mutation*. 2020;41:581–590. https://doi.org/10.1002/humu.23961