Published online 23 February 2021

NAR Genomics and Bioinformatics, 2021, Vol. 3, No. 1 **1**
https://doi.org/10.1093/nargab/lqab008

TMSNP: a web server to predict pathogenesis of missense mutations in the transmembrane region of membrane proteins

Adrián Garcia-Recio [1,2,†], José Carlos Gómez-Tamayo [3,†], Iker Reina [1], Mercedes Campillo [1], Arnau Cordomí [1,4,*] and Mireia Olivella [2,4,*]

[1]Laboratori de Medicina Computacional, Facultat de Medicina, Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain, [2]Bioinformatics and Medical Statistics Group, Facultat de Ciències i Tecnologia, UVIC-UCC, 08500 Vic, Barcelona, Spain, [3]Pharmacoinformatics Group, Research Program on Biomedical Informatics (IMIM/UPF), 08003 Barcelona, Spain and [4]Bioinformatics Department, ESCI-UPF, 08003 Barcelona, Spain

Received May 18, 2020; Revised December 16, 2020; Editorial Decision January 25, 2021; Accepted January 27, 2021

ABSTRACT

The massive amount of data generated from genome sequencing brings tons of newly identified mutations, whose pathogenic/non-pathogenic effects need to be evaluated. This has given rise to several mutation predictor tools that, in general, do not consider the specificities of the various protein groups. We aimed to develop a predictor tool dedicated to membrane proteins, under the premise that their specific structural features and environment would give different responses to mutations compared to globular proteins. For this purpose, we created TMSNP, a database that currently contains information from 2624 pathogenic and 196 705 non-pathogenic reported mutations located in the transmembrane region of membrane proteins. By computing various conservation parameters on these mutations in combination with annotations, we trained a machine-learning model able to classify mutations as pathogenic or not. TMSNP (freely available at http://lmc.uab.es/tmsnp/) improves considerably the prediction power of commonly used mutation predictors trained with globular proteins.

INTRODUCTION

Whole genome and exome sequencing have revealed that Mendelian rare disease-causing missense mutations are more frequent than previously thought and collectively affect millions of patients worldwide (1). Thus, there is an urgent need to understand the relation between genotype and phenotype in order to identify disease-causing genetic variants within candidate variants. For this purpose, variant prioritization tools are widely used to predict the effect of mutations. These are mostly based on evolutionary conservation and expected impact on structure and function using evolutionary conservation parameters and physicochemistry properties of amino acids from sequence data [SIFT (2), Provean (3), MutationTaster (4)], while some tools such as Polyphen-2 (5) also incorporate features related to structural data [see (6) for a review].

Membrane proteins represent 25% of all human proteins (7) and perform essential roles in cellular functions (8). Consequently, they are the target for 50% of drugs in the market (9). Moreover, 90% of membrane proteins present disease-associated missense mutations that may affect protein folding, stability and/or function (10). Some of them have been related to various diseases, including cardiopathies, neurological diseases, cystic fibrosis and cancer (11,12). In fact, mutations in membrane proteins are more likely to cause diseases than in globular proteins (13). Membrane proteins differ from globular proteins in terms of amino acid composition, distribution, inter-residue interactions and structure (14,15). The main differences are in the transmembrane (TM) region of the proteins because of the different surface environments, that is lipid exposed versus water exposed. Current variant prioritization tools, which are mainly based on data from globular proteins, present low reliability for predicting the pathogenicity of mutations in membrane proteins (13). Thus, there is a need for computational tools specific for membrane proteins to understand its relation between sequence and structure (16). This is especially important given the scarce number of membrane protein structures compared to globular proteins due to experimental limitations (17). Mutation prediction tools and databases specific for membrane proteins are starting to emerge, such

*To whom correspondence should be addressed to Mireia Olivella. Tel: +34 932954710; Email: mireia.olivella@uvic.cat
Correspondence may also be addressed to Arnau Cordomí. Tel: +34 935813812; Email: arnau.cordomi@uab.cat
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

as Mut-HTP (18), Pred-MutHTP (10) or BorodaTM (19), which are all based on evolutionary conservation parameters, the former including structure descriptors and the latter focusing on regions with known structure.

With the aim of contributing to emerging mutation predictor servers for the TM region of membrane proteins, here we present TMSNP (accessible at http://lmc.uab.es/tmsnp/), a database of TM missense mutations (pathogenic and non-pathogenic) and a predictor server trained using evolutionary conservation parameters.

MATERIALS AND METHODS

Database of pathogenic and non-pathogenic mutations in transmembrane proteins

Our selected set of membrane proteins consisted of all human membrane proteins tagged as reviewed in the UniProt database (20,21). For each protein, we retrieved all disease-causing/pathogenic mutations associated with Mendelian disorders as reported in ClinVar (22) and SwissVar (23). We only kept in mutations occurring in the TM helices because these are the regions that mostly differ from globular proteins. The ranges of TM segments were taken from the UniProt database (20,21). We also retrieved non-pathogenic missense mutations and their population allele frequency from GnomAD (24) and ClinVar (22). The database (accessible at http://lmc.uab.es/tmsnp/tmsnpdb) resulted in 196 705 non-pathogenic, 2624 pathogenic and 437 likely pathogenic mutations in the TM region of membrane proteins.

TMSNP predictor

Filtering, homology reduction and dataset balancing. To ensure that mutations used in the machine-learning models were linked to protein function and/or structure alteration, we discarded all mutations in proteins for which no single pathogenic disease-causing mutations have been reported, that is those likely involved in complex diseases or recessive inheritance and tagged as 'non-pathogenic proteins' (25). Thus, the obtained pathogenic and non-pathogenic missense mutations were used to classify human TM proteins as 'pathogenic proteins' (358 proteins), when at least one disease-causing pathogenic mutation has ever been reported for this protein and as 'non-pathogenic proteins' (2420 proteins), elsewhere. We next performed homology reduction by discarding mutations for proteins belonging to the same Pfam family (26) that resulted in the same amino acid change in the same aligned position. The dataset after homology reduction contained 2704 pathogenic and likely pathogenic mutations and 19 292 non-pathogenic mutations. Data were subsequently subsampled to obtain a balanced dataset (50% pathogenic and 50% non-pathogenic mutations), by selecting the non-pathogenic mutations with the highest population allele frequency according to GnomAD (24). The final dataset used for training in the machine-learning model presented 5408 missense mutations, which implies a reduction of non-pathogenic mutations by 1/6.

Feature extraction. Multiple sequence alignments for the different families of the proteins in our dataset were taken from the Pfam database (26). For each missense mutation in the balanced dataset, we computed four variables related to evolutionary conservation and the likelihood that an amino acid change is tolerated in a position: (i) frequency of the wild-type amino acid in the Pfam alignment, (ii) frequency of the mutated amino acid, (iii) substitution matrix score (as a measure of similar physicochemical properties between wild-type and mutated amino acid) and (iv) entropy of the position (as a measure of sequence variability or information content) (27). For the substitution matrix score, we used the PHAT 75/73 matrix, which is specific for membrane proteins (28). The entropy of the position i is maximal ($= 1$) if all 20 amino acids at the position i present equal frequencies and is minimal ($= 0$) if only 1 amino acid has been observed at this position. Four additional variables: type of the reference and the mutated amino acids, Pfam and UniProt accession codes were included through one-hot-encoding of the qualitative variables. In the final dataset, each missense mutation had information encoded in eight variables contributing to a total of 569 features (20 features for reference and 20 for the mutated amino acids, 358 features for the UniProt accession codes and 167 for the Pfam accession codes).

Machine-learning models. We built three datasets from the 5408 missense mutations (see http://lmc.uab.es/tmsnp/datasets): (i) 8V dataset, containing all variables (569 features); (ii) 6V dataset, lacking UniProt and Pfam accession code variables, which are informative of the tendency of a protein or a protein family to pathogenesis, but still including one-hot encoded reference and mutated amino acids (60 features) and (iii) 4V dataset containing only conservation variables (wild-type and mutated frequencies, substitution matrix score and entropy). For each dataset, five different training (80%) and test (20%) sets were created by random sampling under certain restrictions for internal validation. For external-validation, in the 8V dataset and 4V datasets, mutations with the same UniProt code were equally split between training set and test set while for the 6V dataset Pfam accession codes were used to split mutations either in the training ser or in the validation set. Machine-learning models were built using Flame (https://github.com/phi-grib/flame; a Python modeling framework which wraps scikit-learn (http://scikit-learn.sourceforge.net)) or Keras (https://keras.io/). Various predictive models using different algorithm settings, applicability domain and dataset were built and internally validated using K-fold ($K = 5$) cross-validation. In specific we used Random Forest (RF), Gradient Boosting (XGBoost), Supporting Vector Machines (SVM) and a sequential neural network. The conformal prediction was used as an applicability domain technique (29) by testing our models at three different confidences: 95, 90 and 80%.

Web server

TMSNP web application tool was constructed using a Python backend (v.3.7) with the Flask framework (v.1.0.2). Both the application and the associated datasets used for

Table 1. Model statistics in cross-validation

Dataset	Algorithm	Confidence	Sensitivity	Specificity	MCC	Coverage	Accuracy
8V (569 features)	RF	0.95	0.92	0.88	0.80	0.42	0.90
	RF	0.90	0.89	0.83	0.72	0.62	0.86
	RF	0.80	0.82	0.77	0.58	0.88	0.79
	XGBOOST	0.95	0.96	0.93	0.89	0.39	0.94
	XGBOOST	0.90	0.92	0.88	0.80	0.58	0.90
	XGBOOST	0.80	0.85	0.81	0.66	0.86	0.83
	SVM	0.95	0.96	0.93	0.89	0.46	0.95
	SVM	0.90	0.92	0.90	0.82	0.63	0.91
	SVM	0.80	0.88	0.84	0.71	0.86	0.85
6V (44 features)	RF	0.95	0.91	0.71	0.63	0.13	0.81
	RF	0.90	0.87	0.72	0.60	0.29	0.79
	RF	0.80	0.79	0.67	0.46	0.58	0.72
	XGBOOST	0.95	0.88	0.73	0.60	0.08	0.79
	XGBOOST	0.90	0.81	0.74	0.55	0.25	0.77
	XGBOOST	0.80	0.76	0.69	0.45	0.57	0.72
	SVM	0.95	0.90	0.77	0.68	0.08	0.84
	SVM	0.90	0.81	0.74	0.55	0.25	0.77
	SVM	0.80	0.76	0.69	0.45	0.57	0.72
4V (4 features)	RF	0.95	0.86	0.59	0.46	0.12	0.72
	RF	0.90	0.77	0.64	0.42	0.29	0.71
	RF	0.80	0.71	0.63	0.34	0.57	0.67
	XGBOOST	0.95	0.87	0.72	0.59	0.09	0.79
	XGBOOST	0.90	0.82	0.72	0.54	0.21	0.76
	XGBOOST	0.80	0.73	0.67	0.40	0.48	0.70
	SVM	0.95	0.93	0.30	0.29	0.15	0.71
	SVM	0.90	0.90	0.41	0.37	0.34	0.69
	SVM	0.80	0.73	0.65	0.38	0.67	0.69

The table shows quality metrics (5-fold) for the machine-learning models created using 8V, 6V and 4V datasets with different conformal significance. MCC stands for Matthews correlation coefficient, which is a measure that combines sensitivity and specificity. Coverage stands for the percentage of samples inside the applicability domain.

training and testing the predictor were built automatically using Python/Bash scripts that collected the required data and stored it in a MySQL database (v.8.0.18), facilitating regular updates. All scripts could be found in a GitHub repository (https://github.com/adriangarciarecio/TMSNP).

RESULTS AND DISCUSSION

We initially constructed a database of missense mutations in human membrane proteins that exclusively focused on TM helices (accessible at http://lmc.uab.es/tmsnp/tmsnpdb; see 'Materials and Methods'). The database currently contains 2624 pathogenic, 437 likely pathogenic and 196 705 non-pathogenic mutations. We used a subset of this database (see 'Materials and Methods') to develop machine-learning models able to classify mutations as pathogenic or not. We assessed three different algorithms (RF, SVM and XGBoost) and three different datasets (8V, 6V and 4V). All combinations showed good performance in both internal (5-fold cross-validation; Table 1) and external validation (independent 20% test set; Table 2) with none of them clearly outperforming the other two. RF showed the best performance on the 4V dataset, while XGBoost and SVM were the best on 6V and 8V, respectively. Although we also tested a sequential neural network, we could not find advantages of using this method despite performance being close to the other algorithms used in this study. Models that use only four features reach a maximum accuracy of ∼70%, without the increase in the confidence of the model bringing additional improvement. The two additional features (type of reference and mutated residues) included in the 6V dataset increase the average performance up to ∼80% accuracy (at the maximum confidence) except for SVM which keeps at ∼70%. 8V dataset clearly shows the best performance both in internal cross validation and external validation. SVM provides the best models, which reach 94% accuracy with 46% coverage (95% confidence), or 85% with 86% coverage (80% confidence). XGBoost and RF models follow closely although with slightly worse performance (<5%).

In order to check for possible overfitting of the models using the largest (8V) dataset, we performed feature selection for the three different algorithms. Supplementary Table S1 compares the performance of the models with the original and the reduced features using *K*-best feature selection performed with $K = 60$ or 30 (number of variables reduced to 10% and ∼5%, respectively). The mean accuracy loss for RF, XGBoost and SVM was, respectively, 2%, -2% and 3% for $K = 60$, and 4%, 0% and 6% for $K = 30$. These small differences suggest lack of overfitting and also point out that SVM is less robust than RF or XGBoost algorithms. In order to assess the presence of bias due to dataset balancing (see 'Materials and Methods'), we generated an additional dataset containing the first 3000 non-pathogenic mutations following those used in training. This bias might lead to unrealistic predictions, possibly translated to an excess of false positives. Supplementary Table S2 shows prediction results using 8V models at 95%, 90% and 80% confidence. Minimum true negatives/false positives ratio is ∼4, being most of the predictions either negatives or out of the applicability domain and always sticking to the confidence restraints.

Table 2. Model statistics at external validation

Dataset	Algorithm	Confidence	Sensitivity	Specificity	MCC	Coverage	Accuracy
8V (569 features)	RF	0.95	0.90	0.86	0.76	0.38	0.88
	RF	0.90	0.86	0.82	0.68	0.58	0.84
	RF	0.80	0.81	0.75	0.56	0.86	0.78
	XGBOOST	0.95	0.90	0.88	0.78	0.30	0.89
	XGBOOST	0.90	0.85	0.84	0.70	0.48	0.85
	XGBOOST	0.80	0.79	0.78	0.57	0.77	0.78
	SVM	0.95	0.91	0.89	0.80	0.39	0.90
	SVM	0.90	0.86	0.85	0.71	0.55	0.85
	SVM	0.80	0.77	0.79	0.56	0.81	0.78
6V (44 features)	RF	0.95	0.86	0.66	0.54	0.78	0.74
	RF	0.90	0.81	0.69	0.51	0.76	0.73
	RF	0.80	0.76	0.66	0.42	0.71	0.70
	XGBOOST	0.95	0.87	0.69	0.58	0.09	0.80
	XGBOOST	0.90	0.82	0.70	0.53	0.24	0.77
	XGBOOST	0.80	0.77	0.68	0.45	0.54	0.73
	SVM	0.95	0.90	0.77	0.68	0.08	0.84
	SVM	0.90	0.81	0.74	0.55	0.25	0.77
	SVM	0.80	0.76	0.69	0.45	0.57	0.72
4V (4 features)	RF	0.95	0.85	0.60	0.47	0.13	0.75
	RF	0.90	0.79	0.66	0.46	0.29	0.73
	RF	0.80	0.71	0.67	0.37	0.56	0.69
	XGBOOST	0.95	0.79	0.65	0.45	0.09	0.73
	XGBOOST	0.90	0.76	0.67	0.43	0.21	0.72
	XGBOOST	0.80	0.69	0.66	0.35	0.47	0.68
	SVM	0.95	0.92	0.38	0.37	0.14	0.74
	SVM	0.90	0.84	0.55	0.41	0.35	0.73
	SVM	0.80	0.72	0.68	0.40	0.67	0.70

The table shows performance metrics in external validation (20% of the original dataset) for the machine-learning models created using 8V, 6V and 4V datasets with different conformal significance. MCC and coverage are described in Table 1.

These results demonstrate lack of sampling bias. Interestingly, the RF model at 95% confidence correctly predicts 658 non-pathogenic mutations with only 6 false positives out of 3000 non-pathogenic mutations.

Random Forest 8V was selected as the final model to be implemented in the web application. Although the performance of RF was not the best, it demonstrated to be the more robust algorithm at different conditions. While XGBoost performed on average better than RF, it provides more importance to protein classification features rather than sequence conservation (amino acid frequencies, substitution matrix score and entropy), questioning its ability to generalize the predictions (Supplementary Table S3). On the other hand, SVM was discarded because it was less robust towards feature reduction and also because the implementation of the SVM algorithm using radial basis function kernel did not allow to inspect feature importances.

Feature importance analysis (Supplementary Table S4) shows that for the original RF model as well as for the 60 and 30 best feature-reduced models, conservation features contribute the most clearly driving the predictive power of the algorithms (30%, 60% and 75% of the total contribution, respectively). Pfam PF00520 (ion channel family) and UniProt P35498 (sodium channel protein) accession codes follow in contribution (∼2% each), probably indicating high sensitivity of this family of receptors to become pathogenic upon a mutation. Mutation to proline and to arginine (P_m and R_m features) appear as the next features in importance (∼1% each). This is compatible with the known distorting effects of these amino acids when present at TM helices (13). Accordingly, the loss of performance for the 6V or 4V dataset might be related to not using UniProt and Pfam codes as features, as they are related to different vulnerability of proteins and/or protein families to amino acid change in their transmembrane region.

TMSNP returns the unambiguous class prediction at the highest confidence possible. Predictions with a confidence below 0.75 are considered outside the domain of applicability. Table 3 shows the comparison between TMSNP models (8V dataset) generated at three levels of significance together with the results of SIFT, Polyphen-2 and Pred-MutHTP. Pred-MutHTP is also specific for membrane proteins, whereas both SIFT and Polyphen-2 are not. As reflected by the equilibrated sensitivity/specificity and the higher accuracy at the different confidence levels, TMSNP not only provides a more balanced model but also performs very well when tested against a full non-pathogenic dataset of 3000 mutations. Importantly, TMPSNP brings higher specificity compared to the other methods. The robustness of the algorithm relies on the quality of the extracted conservation features, which are the most important according to the feature contribution analysis of the model. Noteworthy, TMSNP is using conformal prediction as an applicability domain and uncertainty framework, providing predictions under confidence restraints. Our models demonstrated to be good at all confidence limits, which translates into the corresponding accuracy (i.e. a confidence of 0.8 sets to ∼0.2 the maximum error rate). The higher specificity and accuracy of TMSNP compared to SIFT and PolyPhen-2, which also rely on similar conservation parameters, might be related to using a dataset for training specific for membrane proteins, as the used evolutionary conservation parame-

Table 3. Sensitivity, specificity, Matthews correlation coefficient (MCC) and coverage of TMSNP model (8V dataset) and comparison to Pre-MutHTP, SIFT and Poyphen-2

Predictor types	Method	Sensitivity	Specificity	MCC	Coverage	Accuracy
Specific for membrane proteins	**TMSNP** (0.95 confidence)	0.90	0.86	0.76	0.38	0.88
	TMSNP (0.90 confidence)	0.86	0.82	0.68	0.58	0.84
	TMSNP (0.8 confidence)	0.81	0.75	0.56	0.86	0.78
	Pre-MutHTP (0.95 confidence)	0.96	0.54	0.56	0.76	0.64
	Pre-MutHTP (0.90 confidence)	0.96	0.53	0.55	0.76	0.67
	Pre-MutHTP (0.80 confidence)	0.96	0.53	0.56	0.76	0.71
Non-specific for membrane proteins	**Polyhen-2**	0.93	0.35	0.35	1	0.64
	SIFT	0.88	0.52	0.42	1	0.70

Data are shown at various levels of significance in external validation. MCC and coverage are described in Table 1.

ters differ between globular and membrane proteins. When compared to the previously reported membrane-specific predictor Pred-MutHTP, the higher specificity and accuracy of TMSNP might be related to the better curated non-pathogenic TM variants as the result of (i) only considering the highest allele frequencies in GnomAD and (ii) discarding non-pathogenic mutations in proteins for which no single causative disease mutation has been identified that might be affecting the structure and function of the protein without being related to pathogenesis (recessive inheritance and complex diseases). Compared to MutHTP database, TMSNP (i) contains mutations in the TM segments of membrane proteins but discards mutations in the extracellular and intracellular regions of membrane proteins, as these regions and domains of the proteins are exposed to an environment similar to globular proteins; (ii) is based on a bigger dataset of non-pathogenic variants as MutHTP does not include variants from GnomAD database (24); and (iii) does not include somatic mutations.

CONCLUSIONS

TMSNP is a regularly updated web server that presents two main functionalities: on the one hand, it brings a searchable database of reported pathogenic and non-pathogenic mutations in TM segments of membrane proteins; on the other hand, it provides a mutation prediction tool able to predict pathogenicity for previously non-reported TM missense mutations. The predictive model developed specifically for membrane proteins allows to improve the prediction power compared to unspecific mutation predictor servers.

DATA AVAILABILITY

TMSNP is available at http://lmc.uab.es/tmsnp/.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

FUNDING

Spanish Ministerio de Ciencia, Innovación y Universidades [SAF2015-74627-JIN and SAF2016-77830-R], ISCIII-Subdirección General de Evaluación [PI19/00348] that may include European Regional Development Fund (FEDER) funds. Funding for open access charge: this study was supported by project PI19/00348, financed by the ISCIII-Subdirección General de Evaluación and FEDER.

Conflict of interest statement. None declared.

REFERENCES

1. Chong,J.X., Buckingham,K.J., Jhangiani,S.N., Boehm,C., Sobreira,N., Smith,J.D., Harrell,T.M., McMillin,M.J., Wiszniewski,W., Gambin,T. *et al.* (2015) The genetic basis of Mendelian phenotypes: discoveries, challenges, and opportunities. *Am. J. Hum. Genet.*, **97**, 199–215.
2. Sim,N.-L., Kumar,P., Hu,J., Henikoff,S., Schneider,G. and Ng,P.C. (2012) SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.*, **40**, W452–W457.
3. Choi,Y. and Chan,A.P. (2015) PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics*, **31**, 2745–2747.
4. Schwarz,J.M., Cooper,D.N., Schuelke,M. and Seelow,D. (2014) MutationTaster2: mutation prediction for the deep-sequencing age. *Nat. Methods*, **11**, 361–362.
5. Adzhubei,I., Jordan,D.M. and Sunyaev,S.R. (2013) Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.*, **7**, Chapter 7:Unit7.20.
6. Niroula,A. and Vihinen,M. (2016) Variation interpretation predictors: principles, types, performance, and choice. *Hum. Mutat.*, **37**, 579–597.
7. Dobson,L., Langó,T., Reményi,I. and Tusnády,G.E. (2015) Expediting topology data gathering for the TOPDB database. *Nucleic Acids Res.*, **43**, D283–D289.
8. Gromiha,M.M. and Ou,Y.-Y. (2014) Bioinformatics approaches for functional annotation of membrane proteins. *Brief. Bioinform.*, **15**, 155–168.
9. Overington,J.P., Al-Lazikani,B. and Hopkins,A.L. (2006) How many drug targets are there? *Nat. Rev. Drug Discov.*, **5**, 993–996.
10. Kulandaisamy,A., Priya,S.B., Sakthivel,R., Frishman,D. and Gromiha,M.M. (2019) Statistical analysis of disease-causing and neutral mutations in human membrane proteins. *Proteins*, **87**, 452–466.
11. Kulandaisamy,A., Zaucha,J., Sakthivel,R., Frishman,D. and Michael Gromiha,M. (2020) Pred-MutHTP: prediction of disease-causing and neutral mutations in human transmembrane proteins. *Hum. Mutat.*, **41**, 581–590.
12. Hauser,A.S., Chavali,S., Masuho,I., Jahn,L.J., Martemyanov,K.A., Gloriam,D.E. and Babu,M.M. (2018) Pharmacogenomics of GPCR drug targets. *Cell*, **172**, 41–54.
13. Zaucha,J., Heinzinger,M., Kulandaisamy,A., Kataka,E., Salvádor,Ó.L., Popov,P., Rost,B., Gromiha,M.M., Zhorov,B.S. and Frishman,D. (2020) Mutations in transmembrane proteins: diseases, evolutionary insights, prediction and comparison with globular proteins. *Brief. Bioinform.*, bbaa132.
14. Olivella,M., Gonzalez,A., Pardo,L. and Deupi,X. (2013) Relation between sequence and structure in membrane proteins. *Bioinformatics*, **29**, 1589–1592.

15. Mayol,E., Campillo,M., Cordomí,A. and Olivella,M. (2019) Inter-residue interactions in alpha-helical transmembrane proteins. *Bioinformatics*, **35**, 2578–2584.
16. Almeida,J.G., Preto,A.J., Koukos,P.I., Bonvin,A.M.J.J. and Moreira,I.S. (2017) Membrane proteins structures: a review on computational modeling tools. *Biochim. Biophys. Acta Biomembr.*, **1859**, 2021–2039.
17. Burley,S.K., Berman,H.M., Christie,C., Duarte,J.M., Feng,Z., Westbrook,J., Young,J. and Zardecki,C. (2018) RCSB Protein Data Bank: sustaining a living digital data resource that enables breakthroughs in scientific research and biomedical education. *Protein Sci.*, **27**, 316–330.
18. Kulandaisamy,A., Binny Priya,S., Sakthivel,R., Tarnovskaya,S., Bizin,I., Hönigschmid,P., Frishman,D. and Gromiha,M.M. (2018) MutHTP: mutations in human transmembrane proteins. *Bioinformatics*, **34**, 2325–2326.
19. Popov,P., Bizin,I., Gromiha,M., A,K. and Frishman,D. (2019) Prediction of disease-associated mutations in the transmembrane regions of proteins with known 3D structure. *PLoS One*, **14**, e0219452.
20. McGarvey,P.B., Nightingale,A., Luo,J., Huang,H., Martin,M.J., Wu,C. and UniProt Consortium (2019) UniProt genomic mapping for deciphering functional effects of missense variants. *Hum. Mutat.*, **40**, 694–705.
21. Consortium, T.U.The UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, **47**, D506–D515.
22. Landrum,M.J., Lee,J.M., Riley,G.R., Jang,W., Rubinstein,W.S., Church,D.M. and Maglott,D.R. (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.*, **42**, D980–D985.
23. Mottaz,A., David,F.P.A., Veuthey,A.-L. and Yip,Y.L. (2010) Easy retrieval of single amino-acid polymorphisms and phenotype information using SwissVar. *Bioinformatics*, **26**, 851–852.
24. Karczewski,K.J., Francioli,L.C., Tiao,G., Cummings,B.B., Alföldi,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.
25. Eilbeck,K., Quinlan,A. and Yandell,M. (2017) Settling the score: variant prioritization and Mendelian disease. *Nat. Rev. Genet.*, **18**, 599–612.
26. El-Gebali,S., Mistry,J., Bateman,A., Eddy,S.R., Luciani,A., Potter,S.C., Qureshi,M., Richardson,L.J., Salazar,G.A., Smart,A. *et al.* (2019) The Pfam protein families database in 2019. *Nucleic Acids Res.*, **47**, D427–D432.
27. Pei,J. and Grishin,N.V. (2001) AL2CO: calculation of positional conservation in a protein sequence alignment. *Bioinformatics*, **17**, 700–712.
28. Ng,P.C., Henikoff,J.G. and Henikoff,S. (2000) PHAT: a transmembrane-specific substitution matrix. *Bioinformatics*, **16**, 760–766.
29. Norinder,U., Carlsson,L., Boyer,S. and Eklund,M. (2014) Introducing conformal prediction in predictive modeling. A transparent and flexible alternative to applicability domain determination. *J. Chem. Inf. Model.*, **54**, 1596–1603.