Bioinformatics, 36(20), 2020, 4977–4983
doi: 10.1093/bioinformatics/btaa618
Advance Access Publication Date: 5 August 2020
Original Paper

Genome analysis

IDRMutPred: predicting disease-associated germline nonsynonymous single nucleotide variants (nsSNVs) in intrinsically disordered regions

Jing-Bo Zhou[1], Yao Xiong [1], Ke An[1], Zhi-Qiang Ye [1,2,*] and Yun-Dong Wu[1,2,3,*]

[1]Lab of Computational Chemistry and Drug Design, State Key Laboratory of Chemical Oncogenomics, Peking University Shenzhen Graduate School, Shenzhen 518055, China, [2]Shenzhen Bay Laboratory, Shenzhen 518055, China and [3]College of Chemistry and Molecular Engineering, Peking University, Beijing 100871, China

*To whom correspondence should be addressed.
Associate Editor: Pier Luigi Martelli

Received on October 18, 2019; revised on June 28, 2020; editorial decision on June 29, 2020; accepted on July 1, 2020

Abstract

Motivation: Despite of the lack of folded structure, intrinsically disordered regions (IDRs) of proteins play versatile roles in various biological processes, and many nonsynonymous single nucleotide variants (nsSNVs) in IDRs are associated with human diseases. The continuous accumulation of nsSNVs resulted from the wide application of NGS has driven the development of disease-association prediction methods for decades. However, their performance on nsSNVs in IDRs remains inferior, possibly due to the domination of nsSNVs from structured regions in training data. Therefore, it is highly demanding to build a disease-association predictor specifically for nsSNVs in IDRs with better performance.

Results: We present IDRMutPred, a machine learning-based tool specifically for predicting disease-associated germline nsSNVs in IDRs. Based on 17 selected optimal features that are extracted from sequence alignments, protein annotations, hydrophobicity indices and disorder scores, IDRMutPred was trained using three ensemble learning algorithms on the training dataset containing only IDR nsSNVs. The evaluation on the two testing datasets shows that all the three prediction models outperform 17 other popular general predictors significantly, achieving the ACC between 0.856 and 0.868 and MCC between 0.713 and 0.737. IDRMutPred will prioritize disease-associated IDR germline nsSNVs more reliably than general predictors.

Availability and implementation: The software is freely available at http://www.wdspdb.com/IDRMutPred.

Contact: yezq@pku.org.cn or ydwu@pku.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The paradigm of 'sequence-structure-function' states that the protein sequence will fold into well-defined domain structure, and the folded structure will then fulfill specific function (Edsall, 1995). This well-established dogma has been dominant in structural biology and has led to great successes in revealing the structural basis of numerous fundamental biological processes, including oxygen transport of hemoglobin (Marengo-Rowe, 2006), protein translation of ribosome machine (Schmeing and Ramakrishnan, 2009), enzymatic catalysis and so on. It has also played a vital part in understanding the disease mechanisms of genetic variants (Stefl *et al.*, 2013), and in structure-based drug discovery (Anderson, 2003). On the other hand, a large fraction of protein segments, enriched with hydrophilic and charged residues, lack stable structures (Babu *et al.*, 2012; Lise and Jones, 2004; Romero *et al.*, 2001; Uversky *et al.*, 2000). These segments are named intrinsically disordered regions (IDRs), and a large percentage of human proteins (35–44% due to different predictors and statistic methods) contain IDRs longer than 30 residues (Pentony *et al.*, 2010; Van Der Lee *et al.*, 2014; Ward *et al.*, 2004). Previously regarded as 'useless', IDRs have been confirmed to have versatile regulatory functions by acting as entropic chains, effectors, scavengers, assemblers, chaperones and display sites (Tompa, 2002, 2005; Van Der Lee *et al.*, 2014). Malfunctioning of IDRs is associated with a broad spectrum of human diseases, such as cancers (Iakoucheva *et al.*, 2002; Uversky *et al.*, 2014), cardiovascular diseases (Cheng *et al.*, 2006), neurodegenerative diseases (Raychaudhuri *et al.*, 2009) and type 2 diabetes (Uversky *et al.*, 2008), and computational strategies for rational drug discovery that targets IDR are also emerging (Ruan *et al.*, 2019). Overall, it was estimated that a substantial portion (~20%) of pathogenic nonsynonymous single nucleotide variants (nsSNVs) are located in IDRs

(Vacic *et al.*, 2012). For example, variant R306C in MECP2 leads to Rett syndrome (Vacic and Iakoucheva, 2012), a host of variants in BRCA1 promote the development of breast cancer (Mark *et al.*, 2005) and the dileucine motif gain by variants in GLUT1 causes the GLUT1 deficiency syndrome (Meyer *et al.*, 2018). Hence, studies on disease variants located at IDRs are as crucial as those in structured domains.

Recent years have witnessed the unprecedented advances in next-generation sequencing techniques, and their wide applications in disease research such as whole exome sequencing have generated a huge amount of variant data (Goodwin *et al.*, 2016). Since the identified variants contain both disease-associated and non-disease (neutral) ones, the major challenge is to discriminate them in a high throughput manner (Cooper and Shendure, 2011). While the experimental approaches are labor intensive and time consuming, a multitude of computational predictors have been developed, whose results can largely narrow down the variants pool for further experimental validation (Niroula and Vihinen, 2016).

Significant progress has been made in developing tools to predict disease-associated nsSNVs (Riera *et al.*, 2014). Most of them are trained on variants from diverse protein families, and are thus general-purpose predictors. Among them, conservation and structural stability features have been used most widely (Riera *et al.*, 2014). However, the broadly distributed IDRs have neither ordered structures, nor are they as conserved as ordered regions (ORs) (Brown *et al.*, 2011), raising the concern that the disease-association prediction of variants in IDRs using general tools is supposed to be inferior. In fact, previous research observed that general predictors like SIFT encountered more misclassification on disease-associated nsSNVs in IDRs (Mort *et al.*, 2010). Several studies have reported that the amount, the distribution and other characteristics of IDR variants (Uversky *et al.*, 2008, 2014; Vacic *et al.*, 2012), but there is currently no disease-association predictors for them. Considering the indispensable roles played by IDRs in function regulation and pathogenesis, and that the widely used sequencing technologies are continuing to generate a huge amount of variants in IDRs, it is highly desirable to build a disease-association predictor specifically for IDR variants.

In this work, we have developed a machine learning-based disease-association predictor specifically for germline nsSNVs in IDRs (Fig. 1). First, a training and two testing datasets containing germline variants from IDRs was curated. Second, we extracted 175 features, and selected 17 optimal ones through feature selection. Third, three tree-based ensemble machine learning algorithms were adopted to train prediction models, and the comparison with other general prediction tools was conducted. Finally, a standalone package and its web server, namely IDRMutPred, was developed.

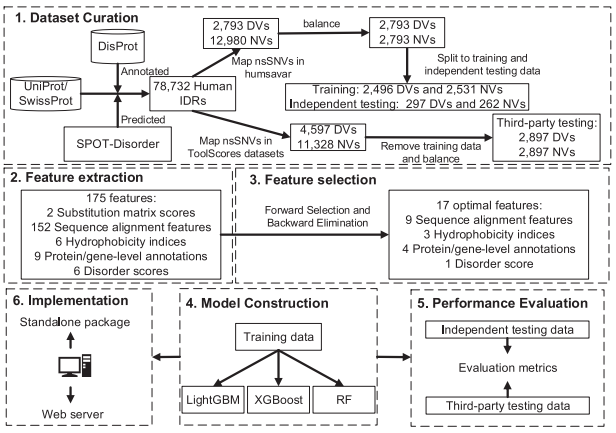

Fig. 1. The pipeline of building the IDR disease nsSNV predictor. DVs and NVs represent disease-associated and neutral nsSNVs, respectively

2 Materials and methods

The overall pipeline of our work is illustrated in Figure 1, and is described as follows.

2.1 Datasets curation

Human protein sequences were derived from UniProt/Swiss-Prot database (Release 2019_01 of January 16, 2019) (The UniProt Consortium, 2019), and the germline nsSNVs were parsed accordingly from the file 'humsavar.txt' from the UniProt FTP server. The nsSNVs labeled with 'Disease' and 'Polymorphism', representing variants reported to be implicated in disease or not, were regarded as positive and negative samples, respectively.

We regarded that a residue is located in IDR if it was experimentally annotated in DisProt (Release 7) database (Piovesan *et al.*, 2017) or could be predicted as disordered by SPOT-Disorder (Hanson *et al.*, 2016). By screening out the nsSNVs located outside IDRs, we obtained the IDR nsSNV dataset. Random sampling was utilized to build a balanced dataset. About one-tenth of the dataset was kept for independent testing, while the remaining served as the training dataset for feature selection and model training. Another third-party dataset for further evaluation was based on ToolScores datasets (Grimm *et al.*, 2015) from VariBench (Nair and Vihinen, 2013). After aggregating its five member datasets, we deleted the nsSNVs with conflicting class labels, kept only one occurrence for duplicates with consistent class labels, and retained the IDR nsSNVs. The third-party dataset was then constructed by removing those that have occurred in the training set and selecting equal number of positive and negative samples.

2.2 Feature extraction

A total of 175 features were calculated for each of the IDR nsSNVs (Supplementary Table S1), and were subjected to further feature selection. These features can be categorized into 5 groups, including 2 substitution matrix scores, 152 sequence alignment features, 6 amino acid hydrophobicity scores, 9 protein-/gene-level annotations and 6 disorder scores, all feature values in the training data were standardized (Supplementary Methods). Feature values in the independent testing and the third-party dataset were transformed accordingly with the parameters derived during the standardization of training data.

2.3 Feature selection

Feature selection is a necessary part in machine learning because the initial feature set often contain unnecessary, irrelevant and redundant features, which may slow down the training procedure or introduce over-fitting (Drotar *et al.*, 2015). In this work, we implemented a feature selection strategy by combining forward selection and backward elimination, which wrapped a machine learning algorithm as the backend engine for evaluating the goodness of the feature subsets (Supplementary Methods and Supplementary Fig. S1).

2.4 Prediction model training

We attempted three tree-based machine learning algorithms, including random forest (RF) (Breiman, 2001), extreme gradient boosting (XGBoost) (Chen and Guestrin, 2016) and light gradient boosting machine (LightGBM) (Ke *et al.*, 2017). RF is a classic machine learning framework by constructing a multitude of decision trees in parallel, while XGBoost and LightGBM are two more modern ones by constructing gradient boosting trees. Remarkably, LightGBM has faster training speed and lower memory usage (Ke *et al.*, 2017).

For each algorithm, we randomly tried a series of hyperparameter combinations (random search), and we chose the one with the best AUC in the G10FCV (described in Section 2.5). After obtaining the best hyperparameter combination, we trained the final prediction models using all the training data accordingly. The feature importance was outputted to compare the relative contributions between different features, and Mann–Whitney *U* test was performed when comparing each feature between disease and neutral nsSNVs. The python packages including scikit-learn v0.20.1

(Pedregosa *et al.*, 2011), xgboost v0.82 and lightgbm v2.2.1 were adopted in our work.

2.5 Cross-validation and performance evaluation

Tenfold cross-validation was utilized for finding the best hyperparameter combination. In the separation of data into 10 parts, we required that nsSNVs from the same protein would not be split into different data parts, i.e. the nsSNVs were split at the protein level. Technically, we implemented this process using the Python module GroupKFold in scikit-learn package (Pedregosa *et al.*, 2011), and it was referred to as grouped 10-fold cross-validation (G10FCV). Given a hyperparameter combination, the average performance obtained by G10FCV was adopted to measure the goodness of this hyperparameter combination. By trying various combinations of hyperparameters, one can select the optimal one.

We adopted four comprehensive performance metrics, including accuracy (ACC), Matthew's Correlation Coefficient (MCC), F1 score and area under the receiver operating characteristic curve (AUC). The detailed definitions are provided in Supplementary Methods.

Using the testing datasets, we compared the performance of our models with 17 popular general predictors, including SIFT (Ng and Henikoff, 2001), PolyPhen2 (Adzhubei *et al.*, 2010), PhD-SNP (Capriotti *et al.*, 2006), MutationAssessor (Reva *et al.*, 2011), FATHMM (Shihab *et al.*, 2013), PON-P2 (Niroula *et al.*, 2015), PROVEAN (Choi and Chan, 2015), PANTHER-PSEP (Tang and Thomas, 2016a), Eigen (Ionita-Laza *et al.*, 2016), REVEL (Ioannidis *et al.*, 2016), PMut2017 (Lopez-Ferrando *et al.*, 2017), MutPred2 (Pejaver *et al.*, 2017), CADD (Rentzsch *et al.*, 2019), LIST (Malhis *et al.*, 2019), MetaSVM (Dong *et al.*, 2015), MetaLR (Dong *et al.*, 2015) and M-CAP (Jagadeesh *et al.*, 2016). Among them, two versions of PolyPhen2 (pph2-HumDiv and pph2-HumVar) and FATHMM (weighted fathmm-W and unweighted fathmm-U) were both adopted in the comparison. The prediction results of PolyPhen2, MutationAssessor, PON-P2, PMut2017 and LIST were obtained using their web servers, and those of Eigen, REVEL, CADD, metaSVM/metaLR and M-CAP were downloaded from dbNSFP v3.5a (Liu *et al.*, 2016). The prediction results of all other predictors were obtained by running their standalone programs with default settings locally.

2.6 Implementation of the predictor

To ease the application of our models, we implemented a standalone package to automate the whole process. Based on Python 3.6.6, a set of modules were implemented to calculate all necessary features. By feeding them into the trained models, this package can make predictions in high throughput.

A web server was built to further facilitate the researchers without bioinformatics skills. The Django (v2.1.0) (https://www.djangoproject.com/) and Bootstrap library (v3.3.7) (https://getbootstrap.com/) were utilized to construct the web framework, MySQL (https://www.mysql.com/) database management system was adopted to temporarily store the prediction results and to manage the submitted job queue, and Apache httpd (https://httpd.apache.org) was chosen to provide the web services.

3 Results

3.1 Construction of the training and testing datasets

In total, 29 544 disease and 39 801 neutral nsSNVs located in 12 518 human proteins were derived from humsavar.txt provided by UniProt/Swiss-Prot (The UniProt Consortium, 2019), and 78 732 IDRs were obtained. The integration of nsSNV and IDR resulted in 2793 disease-associated and 12 980 neutral nsSNVs located in IDRs from 6337 proteins (Fig. 1).

A total of 2793 neutral nsSNVs were randomly sampled and were combined with the 2793 disease nsSNVs to build a balanced dataset. From it, 297 disease nsSNVs and 262 neutral nsSNVs (~1/10 of the balanced dataset) were randomly chosen for independent testing, while the remaining were kept for training (Table 1). We required that all nsSNVs from the same protein were either in the testing set or in the training set.

The ToolScores datasets from VariBench database contain 4597 disease-associated and 11 328 neutral nsSNVs in IDRs (Grimm *et al.*, 2015; Nair and Vihinen, 2013). After removing nsSNVs that have occurred in the training dataset, we randomly sampled 2897 neutral nsSNVs to combine with the 2897 disease-associated ones, serving as a third-party testing dataset for further evaluation (Table 1).

3.2 Optimal feature subset

Our feature selection strategy is involved with a huge number of training iterations on various feature combinations, so the computational cost is demanding. Due to its speediness, we adopted LightGBM (Ke *et al.*, 2017) as the backend engine of feature selection. The procedure of feature selection on all of the 175 candidates resulted in an optimal feature subset with 17 features, including 9 sequence alignment features, 4 protein-/gene-level annotations, 3 hydrophobicity features and 1 disorder feature (Table 2). The detailed feature definitions are provided in Supplementary Methods. In the alignment features, four of them are related to the frequencies of wild-type and mutant residues (#2, #3, #6, #7 in Table 2) in the forms of proportion or number of wild-type or mutant residues, or position weight matrix score; three directly describe the conservation of the nsSNV site in terms of relative entropy (#4, #8, #9); two represent the quality of the alignment (#1, #5). The selected gene-/protein-level features measure the variation tolerance (#12, #13), essentiality (#14) and recessive disease-association probability (#11) of the gene that bears the nsSNV. Other selected features measure the hydrophobicity of the microenvironment around the nsSNV site (#15, #17), the hydrophobicity difference between wild-type and mutant residues (#16), and the disorder level at the nsSNV site (#10).

3.3 Prediction model training and evaluation

Based on G10FCV on the training dataset with the optimal features, we determined the best hyperparameter combination from 1000 randomly generated ones using RF, XGBoost and LightGBM, respectively. The best hyperparameters and the performance metrics of cross-validation are listed in Supplementary Tables S2 and S3. Using these best hyperparameters accordingly, the model parameters of the three prediction models were trained on the whole training dataset. Notably, our cross-validation is based on splitting nsSNVs at the protein level, which has avoided the so-called type 2 circularity (Grimm *et al.*, 2015). This strategy will decrease the risk of overly fitting the prediction models to protein-/gene-level features.

Using the independent testing dataset, we directly compared the performance of our models with 14 of the 17 general-purpose predictors (Table 3 and Supplementary Fig. S2). The comparison shows that our models rank on the top tier for all of the four-performance metrics. In detail, our best ACC (0.868), MCC (0.737), F1 score (0.872) and AUC (0.934) have improved 3.3, 4.1, 1.2 and 0.5 percentage point, respectively, when compared to the best one in the other 13 predictors. The most significant improvement comes from MCC, a robust performance metric that balances positive and negative predictions.

We are curious about whether the homologous relationship between proteins from the testing set and those from the training set have conferred overly optimistic performance. Hence, we removed the testing data whose proteins are homologous to those in the training set using the cd-hit webserver (Huang *et al.*, 2020,b) with 30% as cutoff. The datasets and the performance comparison before and after removing homologs (Supplementary Tables S4 and S5) demonstrate that the performance of our predictors remain similar, indicating that these homologous sequences in the testing dataset do not lead to overly optimistic results.

The third-party dataset contains 2897 disease and 2897 neutral IDR nsSNVs, and the performance comparison on this dataset shows similar results (Table 3 and Supplementary Fig. S3). In detail,

Table 1. Summary of the training and testing datasets

Dataset	Number of disease nsSNVs	Number of neutral nsSNVs	Number of IDRs	Number of proteins
Training	2496	2531	2821	2390
Independent testing	297	262	313	262
Third-party testing	2897	2897	2914	2562

Table 2. The 17 selected optimal features

#	Feature name	Description
1	b9_eva_nal_w	Weighted number of sequences in the alignment based on BLAST against UniRef90 with E-value of 10E-45
2	b9_all_rwt	Proportion of wild-type residue at the nsSNV site in the alignment (UniRef90, E-value: default)
3	b9_eva_rmt_w	Weighted proportion of mutant residue at the nsSNV site in the alignment (UniRef90, E-value: 10E-45)
4	b9_eva_ree	Relative entropy based on the alignment (UniRef90, E-value: 10E-45)
5	b1_eva_naa_w	Weighted number of residues at the nsSNV site in the alignment (UniRef100, E-value: 10E-75)
6	b1_hum_nmt_w	Weighted number of mutant residues at the nsSNV site in the alignment (UniRef100 human, E-value: default)
7	b1_nhu_pwm_w	Weighted position weight matrix score based on the alignment (UniRef100 non-human, E-value: default)
8	b1_hum_ree	Relative entropy based on the alignment (UniRef100 human, E-value: default)
9	b1_nhu_ree	Relative entropy based on the alignment (UniRef100 non-human, E-value: default)
10	pos_spo	SPOT-Disorder score of the wild-type residue at the nsSNV site (Hanson *et al.*, 2016)
11	pro_Prec	Estimated probability that a gene is a recessive disease gene (MacArthur *et al.*, 2012)
12	pro_RVIS_ExAC	ExAC-based RVIS score (Petrovski *et al.*, 2013)
13	pro_GDI_Phred	Phred-scaled GDI score (Itan *et al.*, 2015)
14	pro_Essential_gene	Gene essentiality (Georgi *et al.*, 2013)
15	hww_9	Sum of Wimley–White hydropathy index of neighboring residues with a window of 9 (Wimley and White, 1996)
16	hwo_d	Difference of octanol–water free energy transfer index (Eisenberg and McLachlan, 1986)
17	hwo_3	Sum of octanol–water free energy transfer index of neighboring residues with a window of 3

Table 3 Performance comparison on the independent testing dataset and third-party testing dataset

Method[a]	Independent testing dataset				Third-party dataset			
	ACC	MCC	F1 score	AUC	ACC	MCC	F1 score	AUC
Predictors without protein-/gene-level features								
SIFT	0.793	0.584	0.803	0.863	0.660	0.321	0.642	0.723
pph2-HumDiv	0.773	0.542	0.791	0.854	0.637	0.279	0.611	0.693
pph2-HumVar	0.782	0.572	0.781	0.872	0.666	0.353	0.604	0.745
PhD-SNP	0.816	0.664	0.799	–[b]	0.673	0.412	0.552	–[b]
MutationAssessor	0.773	0.559	0.770	0.856	0.663	0.341	0.614	0.755
fathmm-U	0.753	0.518	0.745	0.831	0.615	0.252	0.514	0.665
PROVEAN	0.785	0.590	0.774	0.862	0.636	0.310	0.522	0.675
PANTHER-PSEP	0.805	0.598	0.845	0.858	0.527	0.068	0.496	0.649
Eigen	0.801	0.556	0.707	0.840	0.671	0.357	0.610	0.735
PMut2017	0.834	0.696	0.826	0.924	0.772	0.365	0.452	0.765
MutPred2	0.812	0.657	0.795	0.906	0.634	0.344	0.466	0.761
LIST	0.736	0.495	0.790	0.904	0.703	0.437	0.749	0.809
Predictors containing protein-/gene-level features								
fathmm-W	0.801	0.615	0.795	0.889	0.842	0.686	0.835	0.898
PON-P2	0.835	0.670	0.860	0.929	0.803	0.614	0.822	0.896
REVEL	0.825	0.637	0.701	0.915	0.744	0.555	0.662	0.908
CADD	0.725	0.448	0.672	0.787	0.675	0.352	0.658	0.729
RF-based model	0.857	0.716	0.861	0.927	0.858[c]	0.718[c]	0.853	0.926
XGBoost-based model	0.859	0.719	0.863	0.934[c]	0.856	0.713	0.854	0.929[c]
LightGBM-based model	0.868[c]	0.737[c]	0.872[c]	0.931	0.858[c]	0.718[c]	0.856[c]	0.929[c]

[a]Both PolyPhen2 and fathmm have two versions, so this table contains 16 lines for the 14 general-purpose predictors.
[b]No AUC was calculated for PhD-SNP due to lack of continuous prediction scores.
[c]The best value in each column is underlined.

the best ACC (0.858), MCC (0.718), F1 score (0.856) and AUC (0.929) of our models have improved 1.6, 3.2, 2.1 and 2.1 percentage points, respectively, when compared to the best one in the other 13 predictors. The improvement of MCC is also the most significant in this comparison.

It is worth noting that the independent testing dataset and the third-party dataset have no overlap with our training set. However, some of the testing nsSNVs may be in the training set of other predictors, which may not properly estimate their performance. For example, PMut2017 and PON-P2 used humsavar (October 2016

release) and VariBench for training, respectively (Lopez-Ferrando *et al.*, 2017; Niroula *et al.*, 2015). Therefore, the above comparisons may have underestimated the improved magnitude of our models. When a testing dataset having no overlap with any of the training datasets of all compared predictors is available, the improvement would hopefully be larger.

Moreover, when comparing the performance of each predictor between the independent testing and the third-party dataset, our predictors are stable, while many others manifest large variance. All of these comparisons demonstrate that our IDR-specific models are robustly better in prioritizing pathogenic IDR nsSNVs from neutral ones than general predictors.

As for the other three of the 17 general-purpose predictors, i.e. MetaSVM, MetaLR and M-CAP, our models also showed superior performance on the two testing datasets (Supplementary Table S6), with the only exception that the F1 score of M-CAP is slightly better. Because MetaSVM and MetaLR directly used allele frequency (AF) as a feature, and M-CAP adopted the prediction scores of MetaSVM/MetaLR as their features, we also supplemented ExAC-based AF to our optimal feature subset to re-train our models for the fair comparison with them.

Another concern is how our models will perform on nsSNVs whose proteins contain both disease and neutral nsSNVs. To inspect this, from ToolScores datasets (Grimm *et al.*, 2015) we removed nsSNVs that have occurred in the training data, and then selected proteins that contain both disease and neutral variants. After this filtering, we obtained a dataset with 2475 disease-associated and 873 neutral IDR nsSNVs from 321 proteins, and conducted an additional evaluation (Supplementary Table S7). Although the ACC and MCC have decreased, they are still high (though not ideal) with MCC greater than 0.61 and ACC greater than 0.84. Moreover, the performance of our models remains on the top tier, with other predictors dropping more. These results also demonstrated that predicting the disease-association of nsSNVs whose proteins have both disease and neutral variants are more challenging.

3.4 Feature analysis

To investigate relative contribution of the features in the optimal feature subset, we plotted the feature importance of each feature for the three models (Fig. 2 and Supplementary Fig. S4). Although different models have no identical rank of relative feature importance, they provide some consensus insights: several alignment and gene-/protein-level features contribute more than others. The distributions of the standardized *Z*-scores of each feature in the disease and neutral group are shown in Supplementary Figure S5. We inspect several representative features in detail here.

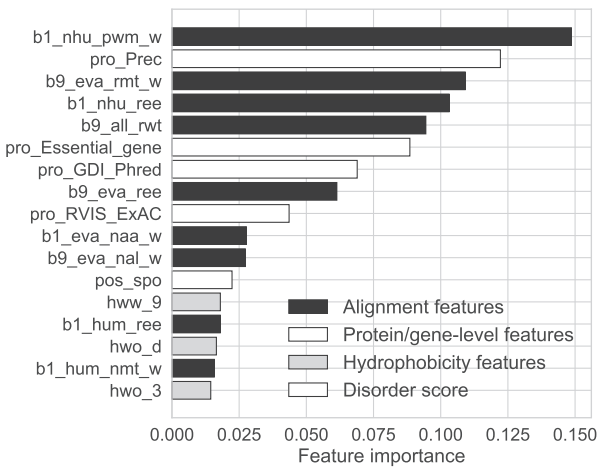

Fig. 2. The feature importance based on RF. The importance is defined as the average gain of splits that use the feature in RF

Conservation features are one group of the most distinguishable features for predicting disease-associated nsSNVs (Niroula and Vihinen, 2016). One may concern whether these features remain powerful in predicting disease nsSNVs in IDRs, as they are of lower sequence conservation (Brown *et al.*, 2011; Tang and Thomas, 2016b). Our results show that several conservation features were selected and ranked on top (Fig. 2 and Supplementary Fig. S4). To inspect them further, we compared one of these features (the relative entropy based on the alignment with homologous non-human proteins in UniRef100, feature #9 in Table 2) between variants in IDRs and in ORs. The conservation levels in IDRs are indeed lower than in ORs (the lower the relative entropy, the higher the conservation level according to our definition in Supplementary Methods), which is consistent with previous knowledge (Fig. 3A). Even though, the distributions of this feature are evidently distinguishable between disease and neutral nsSNVs, either in ORs or in IDRs. In OR nsSNVs, the feature medians of disease and neutral nsSNVs are 2.73 and 3.84, respectively (*P*-value: 0, Mann–Whitney *U* test); in IDR nsSNVs, the medians are 2.85 and 4.56, accordingly (*P*-value: 0, Mann–Whitney *U* test), showing even more evident separation in IDR nsSNVs (Fig. 3A). Hence, it is reasonable that the conservation or conservation-related features were selected.

Certain protein-/gene-level annotations can prioritize disease genes and provide information for further prioritization of variants (Itan *et al.*, 2015). Our work has selected four protein-/gene-level features, and all the three trained models ranked them within the top 10 (Fig. 2 and Supplementary Fig. S4). This observation hints that incorporation of gene-/protein-level features is beneficial in training predictors for disease nsSNVs. One of them is the score estimating the probability that a gene is a recessive disease gene (feature #11 in Table 2). This score has been widely used to discriminate Loss-of-Function tolerant genes (with low score) from recessive disease genes (with high score) (MacArthur *et al.*, 2012). In our dataset, the feature values in the disease group are significantly higher than those in the neutral group (Fig. 3B), indicating that disease-related nsSNVs tend to come from recessive disease genes. The medians of disease and neutral nsSNVs are 0.417 and 0.121 in IDRs, respectively (*P*-value: 1.01E-59, Mann–Whitney *U* test), while in OR these values are 0.333 and 0.152, respectively (*P*-value: 1.70E-70, Mann–Whitney *U* test). Similar to the conservation, the separation of this feature between disease and neutral nsSNVs in IDRs is much larger than that in ORs, illustrating that it may have

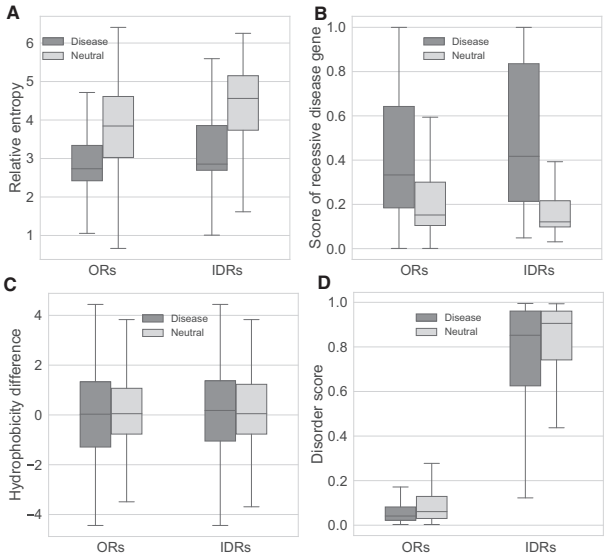

Fig. 3. The boxplots of four features in IDR and OR nsSNVs. (**A**) The relative entropy at the nsSNV site. (**B**) The score estimating the probability that a gene is a recessive disease gene. (**C**) The hydrophobicity difference between the mutant and wild-type residue at the nsSNV site. (**D**) The SPOT-Disorder score of the wild-type residue at the nsSNV site

more potential in predicting disease nsSNVSs in IDRs than in ORs. As protein-/gene-level features cannot differentiate nsSNVs from the same protein, it is necessary to conduct cross-validation at the protein level in optimizing the hyperparameters of the prediction models (G10FCV in this work), and to perform additional evaluations on the dataset derived from proteins containing both disease and neutral variants (Supplementary Table S7).

Hydrophobicity and disorder-related features have been demonstrated informative in developing general predictors (Adzhubei *et al.*, 2010; Huang *et al.*, 2010a; Ye *et al.*, 2007). In this work, several selected features measure the sum of hydrophobicity propensity of a short peptide segment centered with the nsSNV site, i.e. hydrophobicity microenvironment, or the hydrophobicity differences between substituted residues. The disorder score of the wild-type residue at the nsSNV site has also been selected. Although their feature importance ranks are relatively low when compared to other types of features (Fig. 2 and Supplementary Fig. S4), their distributions between disease and neutral nsSNVs are still informative (Fig. 3C and D). For the hydrophobicity difference, the medians of disease and neutral nsSNVs are 0.18 and 0.05 in IDRs, respectively (*P*-value: 0.001, Mann–Whitney *U* test); In ORs, these two values are 0.03 and 0.05 (*P*-value: 3.08E-7, Mann–Whitney *U* test). The larger median of disease nsSNVs in IDRs may indicate that a larger portion of disease nsSNV have been substituted from hydrophilic to hydrophobic residues. Larger separation between disease and neutral nsSNVs in IDRs than in ORs can be observed as well.

The disorder score and the DisProt annotation have been adopted to separate the IDRs from ORs in dataset curation, so the disorder scores of IDR nsSNVs are much larger than OR nsSNVs (Fig. 3D). Moreover, the medians of disease and neutral nsSNVs are 0.852 and 0.906 in IDRs, respectively (*P*-value: 6.84E-14, Mann–Whitney *U* test), while these values are 0.041 and 0.061 in ORs (*P*-value: 0, Mann–Whitney *U* test). The difference of medians between disease and neutral nsSNVs in IDRs is larger than that in ORs, which may also indicate that disorder scores have more potential to separate disease and neutral nsSNVs in IDRs than in ORs.

3.5 The standalone package and web server

Although the performance of the three models in our work is similar, we choose the LightGBM-based model as the default since it is much faster. The RF-based and XGBoost-based models are also provided as the options. The standalone package and the web server of our method, namely IDRMutPred, are freely available at http://www.wdspdb.com/IDRMutPred. Anaconda was utilized to install all necessary packages and to share the running environment (https://www.anaconda.com/), so the users can install and configure a local copy of IDRMutPred smoothly, which will be convenient for high throughput runs. The versions of the related Python packages are listed in Supplementary Table S8. In addition, a Docker image of the standalone IDRMutPred is also freely available at the website.

IDRMutPred requires that the user should provide a protein sequence and a list of amino acid substitutions. The output contains the prediction score in the range between 0 and 1, and the binary classification is based on the default cutoff of 0.5. The users can further prioritize the disease-associated nsSNVs by ranking the scores.

4 Discussion

The better performance of IDRMutPred may roughly stem from several aspects. First, general predictors are 'one-size-fits-all' models based on training data from heterogeneous protein families or protein segments (Riera *et al.*, 2014; Vacic and Iakoucheva, 2012), while IDRMutPred has been trained on pure IDR nsSNVs, which are more homogeneous. It is reasonable to train better specific predictors on homogeneous datasets due to lower noise. Developing specific predictors has been practiced in several studies like KinMutRF, wKinMut-2, iFish and others (Fechter and Porollo, 2014; Izarzugaza *et al.*, 2012; Pons *et al.*, 2016; Vazquez *et al.*, 2016; Wang and Wei, 2016).

Second, homogeneous training data help to highlight informative features accordingly (Torkamani and Schork, 2007). Intuitively, several features in our work have shown more evident contrast between disease and neutral nsSNVs in IDRs than in ORs, e.g. the relative entropy (Fig. 3A). If we combine the IDR with OR nsSNVs, the contrast between disease and neutral nsSNVs will be smaller, i.e. less informative. These features will contribute more in IDR nsSNVs predictors than in general ones, and will be supposed to result in the better performance of IDRMutPred.

In summary, our work presents the first IDR nsSNV-specific predictor, IDRMutPred, which will hopefully serve as a valuable tool in the research community that focuses on the study of nsSNVs, especially those located in IDRs.

Acknowledgements

The authors would like to thank Dr. Xin-Hao Zhang and Fan Jiang for helpful suggestions. They also gratefully acknowledge the support of the National Supercomputer Center in Guangzhou (NSCC-GZ) for computing resources.

Funding

This work was supported by the National Natural Science Foundation of China [31471243, 21933004, 30800641]; the Shenzhen Science and Technology Innovation Commission [JCYJ20170818085409785, JCYJ20170412150507046]; the Program for Guangdong Introducing Innovative and Entrepreneurial Talents; and the Shenzhen Municipal Health Commission [SZSM201809085].

Conflict of Interest: none declared.

References

Adzhubei,I.A. *et al.* (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.

Anderson,A.C. (2003) The process of structure-based drug design. *Chem. Biol.*, **10**, 787–797.

Babu,M.M. *et al.* (2012) Versatility from protein disorder. *Science*, **337**, 1460–1461.

Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.

Brown,C.J. *et al.* (2011) Evolution and disorder. *Curr. Opin. Struct. Biol.*, **21**, 441–446.

Capriotti,E. *et al.* (2006) Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics*, **22**, 2729–2734.

Chen,T. and Guestrin,C. (2016) XGBoost: a scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA. ACM, pp. 785–794.

Cheng,Y. *et al.* (2006) Abundance of intrinsic disorder in protein associated with cardiovascular disease. *Biochemistry*, **45**, 10448–10460.

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.

Cooper,G.M. and Shendure,J. (2011) Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat. Rev. Genet.*, **12**, 628–640.

Dong,C. *et al.* (2015) Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum. Mol. Genet.*, **24**, 2125–2137.

Drotar,P. *et al.* (2015) An experimental comparison of feature selection methods on two-class biomedical datasets. *Comput. Biol. Med.*, **66**, 1–10.

Edsall,J.T. (1995) Hsien Wu and the first theory of protein denaturation (1931). *Adv. Protein Chem.*, **46**, 1–5.

Eisenberg,D. and McLachlan,A.D. (1986) Solvation energy in protein folding and binding. *Nature*, **319**, 199–203.

Fechter,K. and Porollo,A. (2014) MutaCYP: classification of missense mutations in human cytochromes P450. *BMC Med. Genomics*, **7**, 47.

Georgi,B. *et al.* (2013) From mouse to human: evolutionary genomics analysis of human orthologs of essential genes. *PLoS Genet.*, **9**, e1003484.

Goodwin,S. *et al.* (2016) Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.*, **17**, 333–351.

Grimm,D.G. *et al.* (2015) The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Hum. Mutat.*, **36**, 513–523.

Hanson,J. *et al.* (2016) Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks. *Bioinformatics*, **33**, 685–692.

Huang,T. *et al.* (2010a) Prediction of deleterious non-synonymous SNPs based on protein interaction network and hybrid properties. *PLoS One*, **5**, e11900.

Huang,Y. *et al.* (2010b) CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, **26**, 680–682.

Iakoucheva,L.M. *et al.* (2002) Intrinsic disorder in cell-signaling and cancer-associated proteins. *J. Mol. Biol.*, **323**, 573–584.

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.

Ionita-Laza,I. *et al.* (2016) A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat. Genet.*, **48**, 214–220.

Itan,Y. *et al.* (2015) The human gene damage index as a gene-level approach to prioritizing exome variants. *Proc. Natl. Acad. Sci. USA*, **112**, 13615–13620.

Izarzugaza,J.M. *et al.* (2012) Prioritization of pathogenic mutations in the protein kinase superfamily. *BMC Genomics*, **13**, S3.

Jagadeesh,K.A. *et al.* (2016) M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat. Genet.*, **48**, 1581–1586.

Ke,G. *et al.* (2017) LightGBM: a highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.*, **30**, 30.

Lise,S. and Jones,D. (2004) Sequence patterns associated with disordered regions in proteins. *Proteins Struct. Funct. Bioinf.*, **58**, 144–150.

Liu,X. *et al.* (2016) dbNSFP v3. 0: a one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Hum. Mutat.*, **37**, 235–241.

Lopez-Ferrando,V. *et al.* (2017) PMut: a web-based tool for the annotation of pathological variants on proteins, 2017 update. *Nucleic Acids Res.*, **45**, W222–W228.

MacArthur,D.G. *et al.* (2012) A systematic survey of loss-of-function variants in human protein-coding genes. *Science*, **335**, 823–828.

Malhis,N. *et al.* (2019) Improved measures for evolutionary conservation that exploit taxonomy distances. *Nat. Commun.*, **10**, 1556.

Marengo-Rowe,A.J. (2006) Structure-function relations of human hemoglobins. *Proc. (Bayl. Univ. Med. Cent.)*, **19**, 239–245.

Mark,W.-Y. *et al.* (2005) Characterization of segments from the central region of BRCA1: an intrinsically disordered scaffold for multiple protein–protein and protein–DNA interactions? *J. Mol. Biol.*, **345**, 275–287.

Meyer,K. *et al.* (2018) Mutations in disordered regions can cause disease by creating dileucine motifs. *Cell*, **175**, 239–253.

Mort,M. *et al.* (2010) In silico functional profiling of human disease-associated and polymorphic amino acid substitutions. *Hum. Mutat.*, **31**, 335–346.

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

Ng,P.C. and Henikoff,S. (2001) Predicting deleterious amino acid substitutions. *Genome Res.*, **11**, 863–874.

Niroula,A. *et al.* (2015) PON-P2: prediction method for fast and reliable identification of harmful variants. *PLoS One*, **10**, e0117380.

Niroula,A. and Vihinen,M. (2016) Variation interpretation predictors: principles, types, performance, and choice. *Hum. Mutat.*, **37**, 579–597.

Pedregosa,F. *et al.* (2011) Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.

Pejaver,V. *et al.* (2017) MutPred2: inferring the molecular and phenotypic impact of amino acid variants. 10.1101/134981.

Pentony,M.M. *et al.* (2010) Computational resources for the prediction and analysis of native disorder in proteins. *Methods Mol. Biol. (Clifton, N.J.)*, **604**, 369–393.

Petrovski,S. *et al.* (2013) Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet.*, **9**, e1003709.

Piovesan,D. *et al.* (2017) DisProt 7.0: a major update of the database of disordered proteins. *Nucleic Acids Res.*, **45**, D219–D227.

Pons,T. *et al.* (2016) KinMutRF: a random forest classifier of sequence variants in the human protein kinase superfamily. *BMC Genomics*, **17**, 396.

Raychaudhuri,S. *et al.* (2009) The role of intrinsically unstructured proteins in neurodegenerative diseases. *PLoS One*, **4**, e5566.

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

Reva,B. *et al.* (2011) Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.*, **39**, e118.

Riera,C. *et al.* (2014) Prediction of pathological mutations in proteins: the challenge of integrating sequence conservation and structure stability principles. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, **4**, 249–268.

Romero,P. *et al.* (2001) Sequence complexity of disordered protein. *Proteins Struct. Funct. Bioinf.*, **42**, 38–48.

Ruan,H. *et al.* (2019) Targeting intrinsically disordered proteins at the edge of chaos. *Drug Disc. Today*, **24**, 217–227.

Schmeing,T.M. and Ramakrishnan,V. (2009) What recent ribosome structures have revealed about the mechanism of translation. *Nature*, **461**, 1234–1242.

Shihab,H.A. *et al.* (2013) Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum. Mutat.*, **34**, 57–65.

Stefl,S. *et al.* (2013) Molecular mechanisms of disease-causing missense mutations. *J. Mol. Biol.*, **425**, 3919–3936.

Tang,H. and Thomas,P.D. (2016a) PANTHER-PSEP: predicting disease-causing genetic variants using position-specific evolutionary preservation. *Bioinformatics*, **32**, 2230–2232.

Tang,H. and Thomas,P.D. (2016b) Tools for predicting the functional impact of nonsynonymous genetic variation. *Genetics*, **203**, 635–647.

The UniProt Consortium. (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, **47**, D506–D515.

Tompa,P. (2002) Intrinsically unstructured proteins. *Trends Biochem. Sci.*, **27**, 527–533.

Tompa,P. (2005) The interplay between structure and function in intrinsically unstructured proteins. *FEBS Lett.*, **579**, 3346–3354.

Torkamani,A. and Schork,N.J. (2007) Accurate prediction of deleterious protein kinase polymorphisms. *Bioinformatics*, **23**, 2918–2925.

Uversky,V.N. *et al.* (2014) Pathological unfoldomics of uncontrolled chaos: intrinsically disordered proteins and human diseases. *Chem. Rev.*, **114**, 6844–6879.

Uversky,V.N. *et al.* (2000) Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins Struct. Funct. Genet.*, **41**, 415–427.

Uversky,V.N. *et al.* (2008) Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annu. Rev. Biophys.*, **37**, 215–246.

Vacic,V. and Iakoucheva,L.M. (2012) Disease mutations in disordered regions—exception to the rule? *Mol. Biosyst.*, **8**, 27–32.

Vacic,V. *et al.* (2012) Disease-associated mutations disrupt functionally important regions of intrinsic protein disorder. *PLoS Comput. Biol.*, **8**, e1002709.

Van Der Lee,R. *et al.* (2014) Classification of intrinsically disordered regions and proteins. *Chem. Rev.*, **114**, 6589–6631.

Vazquez,M. *et al.* (2016) wKinMut-2: identification and interpretation of pathogenic variants in human protein kinases. *Hum. Mutat.*, **37**, 36–42.

Wang,M. and Wei,L. (2016) iFish: predicting the pathogenicity of human nonsynonymous variants using gene-specific/family-specific attributes and classifiers. *Sci. Rep.*, **6**, 31321.

Ward,J.J. *et al.* (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.*, **337**, 635–645.

Wimley,W.C. and White,S.H. (1996) Experimentally determined hydrophobicity scale for proteins at membrane interfaces. *Nat. Struct. Biol.*, **3**, 842–848.

Ye,Z.Q. *et al.* (2007) Finding new structural and sequence attributes to predict possible disease association of single amino acid polymorphism (SAP). *Bioinformatics*, **23**, 1444–1450.