DAMpred: Recognizing Disease-Associated nsSNPs through Bayes-Guided Neural-Network Model Built on Low-Resolution Structure Prediction of Proteins and Protein–Protein Interactions

Lijun Quan[1,2], Hongjie Wu[2,3], Qiang Lyu[1,4] and Yang Zhang[2,5,]

*1 - **School of Computer Science and Technology,** Soochow University, Suzhou, Jiangsu 215000, China*
*2 - **Department of Computational Medicine and Bioinformatics,** University of Michigan, Ann Arbor, MI 48109, USA*
*3 - **School of Electronic and Information Engineering,** Suzhou University of Science and Technology, Suzhou, Jiangsu 215000, China*
*4 - **Collaborative Innovation Center of Novel Software Technology and Industrialization,** Nanjing, Jiangsu 215000, China*
*5 - **Department of Biological Chemistry,** University of Michigan, Ann Arbor, MI 48109, USA*

***Correspondence to Qiang Lyu and Yang Zhang:**Q. Lyu is to be contacted at: School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu 215000, China; Y. Zhang is to be contacted at: Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA. qiang@suda.edu.cn, zhng@umich.edu*
https://doi.org/10.1016/j.jmb.2019.02.017
Edited by Michael Sternberg

Abstract

Nearly one-third of non-synonymous single-nucleotide polymorphism (nsSNPs) are deleterious to human health, but recognition of the disease-associated mutations remains a significant unsolved problem. We proposed a new algorithm, DAMpred, to identify disease-causing nsSNPs through the coupling of evolutionary profiles with structure predictions of proteins and protein–protein interactions. The pipeline was trained by a novel Bayes-guided artificial neural network algorithm that incorporates posterior probabilities of distinct feature classifiers with the network training process. DAMpred was tested on a large-scale data set involving 10,635 nsSNPs from 2154 ORFs in the human genome and recognized disease-associated nsSNPs with an accuracy 0.80 and a Matthews correlation coefficient of 0.601, which is 9.1% higher than the best of other state-of-the-art methods. In the blind test on the TP53 gene, DAMpred correctly recognized the mutations causative of Li–Fraumeni-like syndrome with a Matthews correlation coefficient that is 27% higher than the control methods. The study demonstrates an efficient avenue to quantitatively model the association of nsSNPs with human diseases from low-resolution protein structure prediction, which should find important usefulness in diagnosis and treatment of genetic diseases.

Introduction

Recent advances in the next generation of sequencing technologies have created high-volume mutation data for comparative genome analyses. By now, more than 6000 human diseases have been identified to be associated with non-synonymous single-nucleotide polymorphisms (nsSNPs). Recognition of the disease-associated genome mutations may help understand the mechanisms of the genetic disorders and improve the chance for early diagnosis and treatment of such diseases [1]. While considerable effort has been made along this line, it remains a significant unsolved problem to precisely recognize the disease-causing nsSNPs from dominant neutral mutations (NMs) [2].

Several methods have been developed for computational recognition of the disease-associated mutations (DMs), which can be generally categorized into two groups: statistical and machine-learning methods. The statistical methods aim to distinguish the deleterious from NMs by taking the advantage of the wealth of existing disease and mutation data sets. For example, MuSiC identifies the genes that have a significantly higher mutation rate than the background mutations, using a multidimensional statistical evaluation of the next-generation-derived cancer data sets [3]. However, estimate of the background mutation rate is often

challenging and the approach has difficulty for the identification of driver genes with low-frequency of recurrence. To address the issue, Oncodrive-fm [4] proposed to detect candidate cancer drivers that do not rely on the recurrence of mutations. It identifies genes under positive selection in tumor development by assessing their bias toward the accumulation of mutations with high functional impact across a cohort of tumor samples, so that many low-recurrent candidate cancer drivers can be successfully identified. E-Drive [5] is another statistical method that exploits the internal distribution of somatic missense mutations. Similarly, SIFT [6] examines the impact of mutations on protein functions by the degree of conservation of the amino acid residues in multiple sequence alignments, built on the assumption that mutations of highly conserved protein positions tend to be more deleterious. However, the relation between protein functions and disease association is complicated, in particular when the intricate interactions of the target protein with the environment molecules are involved, so the accuracy of mutation classifications that rely only on traditional statistics and evolutionary analyses are often limited [7–9].

The machine-learning methods [10–12] are designed to train predictive models on known positive (e.g., functional mutation, deleterious mutation, and DM) and negative (NMs) samples through the extraction of various physicochemical features. PolyPhen-2 [13] and SNAP2 [14] are two of such classical tools that train, respectively, naïve Bayes classifier and neural network models on both sequence and structural features. SNPMuSiC is one of most recently developed approaches that classify deleterious mutation and NM in terms of protein fold stabilities, with the latter trained on various statistical potentials and structural features by artificial neural network [15]. Despite the success, many of the machine-learning methods are built on experimental structures that are not available to most of the disease-associated proteins. Meanwhile, the training and benchmark data sets are often derived from a few common databases, such as UniProtKB/Swiss-Prot [16], OMIM [17], and HumVar [18], which are highly redundant and can result in illegitimately high-performance estimates [12,19].

In this work, we explore the possibility to combine sequence and physicochemical characteristics with three-dimensional structure information generated from the cutting-edge protein structure prediction [20,21] to improve the accuracy of DM recognitions. Meanwhile, the biological assembly structures are modeled through the dimeric threading neural net [22], which helps examine the impact of nsSNPs on the interactions of the target protein with the environment molecules. To overcome the sampling issues of traditional machine learning approaches, a new hybrid training method, Bayes-guided artificial neural network (BANN), is developed to incorporate the posterior probabilities of specific classifiers with the network training process for improving the training efficiency. The pipeline will be carefully benchmarked in multiple data sets, including both cross-validations and blind tests, compared to the current state-of-the-art approaches. The flowchart of the developed pipeline, named Disease-Associated Mutant Predictor (DAMpred), is depicted in Fig. 1, while the online server and standalone package are made freely available at http://zhanglab.ccmb.med.umich.edu/DAMpred/.

Results

Data set construction and method evaluation

Proteomic database mapping

DAMpred starts with the construction of a set of derivative data sets (feature types, access ID, BioUnit, and resMAP, etc.) from two primary sequence and structure databases, UniProtKB [23] and PDB [24], with the mapping pipeline described in Text S1 and Fig. S1 in Supporting Information, SI. The mapping architecture of the derivative data sets is constructed using SQLite, which can be downloaded at http://zhanglab.ccmb.med.umich.edu/DAMpred/download/human.sqlite.

Construction of non-redundant data sets

A set of experimentally validated missense mutations are collected from the UniProt [19], HumDiv [13], and a data set derived from previous study [25]. These mutant variants are from the human genome. In case that >60 mutations occur on one gene, up to 60 variants were randomly selected in the gene, with one variant picked up at each residue position, to reduce redundancy and bias in training. This filtering process results in 10,634 mutations involved in 2154 proteins, containing 5355 DMs in 617 proteins and 5279 NMs in 1836 proteins (see Table S1); the neural mutations in the latter protein set do not change physicochemical property of the proteins. This data set named D10634 can be downloaded at http://zhanglab.ccmb.med.umich.edu/DAMpred/download/D10634.xlsx.

Here, we have limited the maximum number of mutations per gene for two considerations. First, as it was pointed out previously [12], conventional data sets are highly biased in that many proteins with disease-associated variations have no or very little neutral variation data. Likewise, most proteins containing neutral variations do not have disease-associated variations. Thus, limiting the number of mutant samples from a specific gene can help retain the balance of disease and neutral samples in the final data set. Second, one assumption of DAMpred is that the physicochemical features associated with neutral and deleterious mutations are conserved in protein evolution; this assumption has been used in the structure-

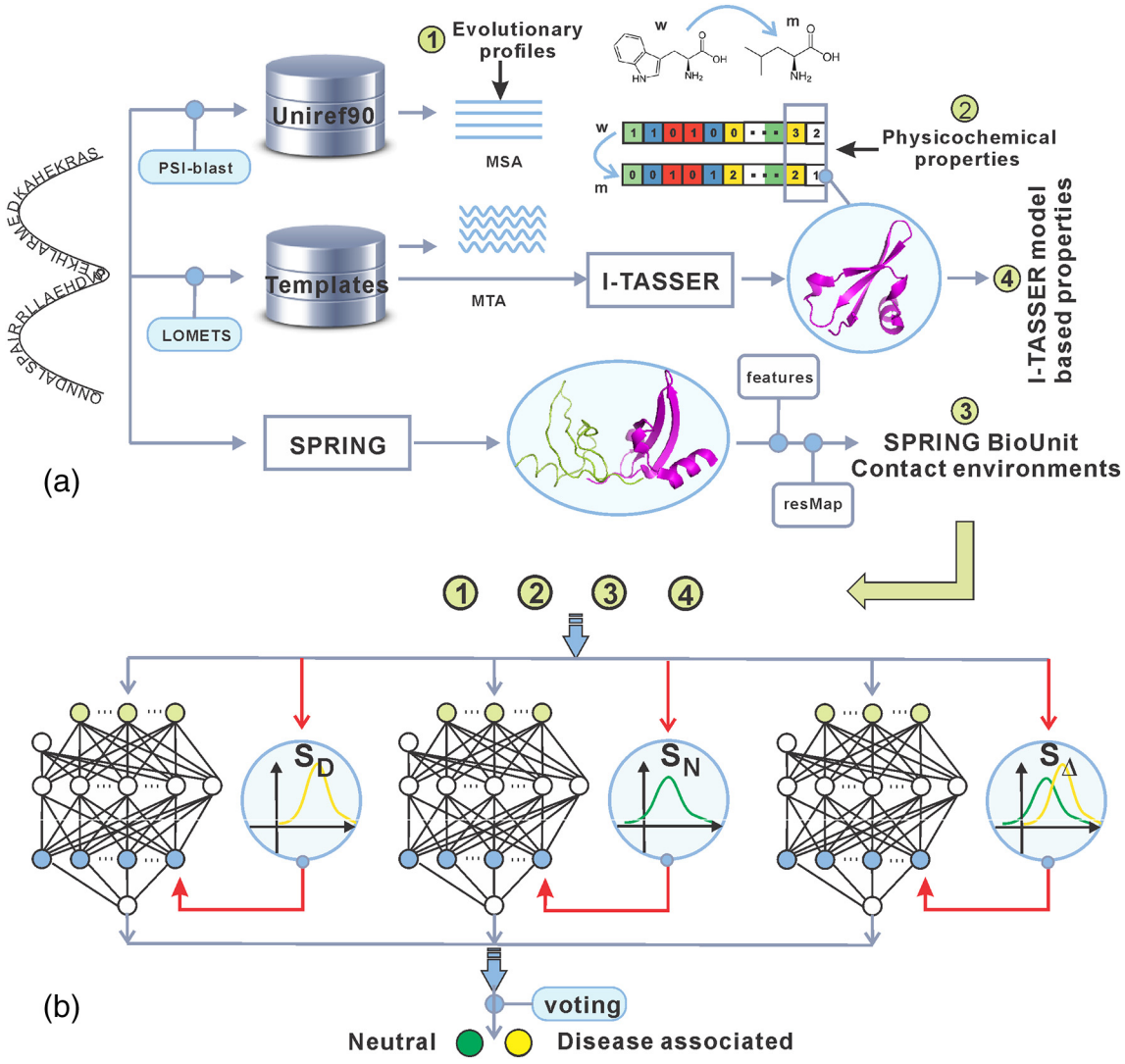

Fig. 1. Flowchart of DAMpred for DM prediction. The pipelines for (a) feature extraction and (b) BANN training.

based feature design and selection. Without appropriate filter to remove plenty of similar physicochemical and structural environment samples from the same proteins, it can result in biased model training and/or overestimated performance in DAMpred. When filtering the samples, we tried to select mutation samples from different positions along the sequence to reduce the potential loss of mutants involved in the important residues.

In addition to the D10634, three other data sets are constructed for independent testing and validation of the methods. First, the D2186 contains 2186 mutations from 233 proteins collected from the ENTPRISET-balance set [12], which have not been included in the D10634. Second, a small data set of D146 contains 146 mutations involved in 90 proteins, which is a subset of the D10634 but consists of mutations from proteins with a sequence identity $<30\%$ to the training sets of the control methods of this study; this is to give a fair comparison of DAMpred with other control methods when DAMpred is trained on a protein set non-homologous to D146. The third data set is collected from mutations on the p53 protein, a predominant tumor suppressor in human genome. As shown in Fig. S2A, the mutations on p53 have been classified as "non-neutral," "neutral," or "uncertain." The non-neutral consists of 618 SNP mutations with experimentally validated effects on the protein function, in which 67 are causative of Li–Fraumeni syndrome or Li–Fraumeni-like syndrome and therefore grouped in DMs. The 22 mutations have been experimentally validated as neutral. The rest are 870 mutations without definitive experimental validation and therefore categorized as "uncertain." Since p53 is a highly enriched with deleterious mutations, it represents a particularly difficult data set for recognizing the neutral mutants to generate balance performance. These three data sets are downloadable at

http://zhanglab.ccmb.med.umich.edu/DAMpred/download/benchmark.tar.bz2.

Modeling of monomer and complex protein structures

The accuracy of protein structure models is critical to structure-based modeling of DMs. Fig. S3A shows a histogram distribution of the TM score of the first models generated by I-TASSER [20,26] for the 2154 target proteins, where all homologous templates with a sequence identity >30% to the query or detectable by PSI-BLAST with *E*-value <0.05 have been excluded. It is shown that the majority of the proteins (86%) can be modeled with a correct fold (TM score > 0.5) [27], although no close homology template is used. Fig. S3B also lists the RMSD of the models to the native, where 94% of them have an RMSD of <5 Å. The average TM score and RMSD for the 2154 testing proteins are 0.74 and 2.9 Å, respectively.

Protein–protein complex structure are modeled by matching the query sequences through a non-redundant library of known BioUnit structures collected from the PDB, using the multimeric threading SPRING algorithm [22]. Among the 2154 test proteins, SPRING successfully constructed complex models for 2116 proteins, after excluding templates having a sequence identity >80% to the query proteins (with an average sequence identity 31%). Figs. S3C and S3D show that the majority of the complex models (80%/92%) have a correct fold and orientation (with TM score > 0.5 and RMSD <5 Å) compared to the experimental structures. Here, since there are much less multimeric templates than monomer ones in the PDB library, SPRING has taken a looser homologous cutoff (80%) than I-TASSER does (30% plus PSI-BLAST E-value), so that there are sufficient number of testing targets with complex BioUnit structures.

Protein-level cross-validation

The accuracy of machine-learning based methods can often be over-optimistically assessed due to the overlaps between training and test data sets; these include the case of mixing the training and testing mutants from the same protein, which has been shown to result in artificial correlation of test and training samples in cross-validation studies [28,29]. Here we used a rigorous "protein-level" cross-validation in our experiment, in which the training and testing samples are collected from different non-homologous proteins. To do this, we first cluster the 2154 proteins in D10634 by BLASTclust [30] using a sequence identity cutoff of 30%, which results in 1644 clusters, each with 1 to 43 members. Next, we construct ten mutation subsets, each with roughly similar size, by randomly taking mutations from different proteins but with a constraint that mutations from the same protein cluster must be in the same subset and mutations in different subsets are from different protein clusters. In the 10-fold cross-validation, nine subsets are chosen as training data set and the remaining one is used as test, where all subsets are rotated as a test only once. Table S1 lists the distribution of the mutations in the 10 subsets, where a detailed list of the 10 subsets can be downloaded at http://zhanglab.ccmb.med.umich.edu/DAMpred/download/tenFold.xlsx.

Assessment of different feature groups on deleterious mutant recognition

DAMpred collects multiple features to train the models. In Table S2, we list all the 70 individual features, together with their mean scores for the DM and NM (columns 4–5). In columns 6–7, we also list the Matthews correlation coefficient (MCC) between the score and the experimental data (disease or neutral), where a score cutoff is specified for each feature to define the positive or negative predictions. To quantitatively examine the difference, column 8 lists the *p* value of two-side Mann–Whitney test between DM and NM data sets, where the histogram distributions of the features between DM and NM are described in Figs. S4–S10 in SI. Next, we examine and highlight several important feature groups in more detail.

Evolutionary features

The evolution-based features show the highest correlation with the experimental data with an average MCC = 0.29 (Fig. S4). Accordingly, this group of features has the lowest *p* value between DM and NM, where all but one (JSD score for wild-type residue, JSD_w) have a *p* value below 10^{-12}. Among them, the position-specific independent count (PSIC) scores based on MSA from Uniref90 have the highest distinguishing power and their *p* values are all below 10^{-200} (or 0). Most of the wild-type amino acids in the DM data set have a higher $PSIC_w$ score ($PSIC_w$ = 1.57) than that in NM data set ($PSIC_w$ = 0.91). On the contrary, the mutant amino acids in the DM have a lower score (PSICm = −0.42) than that in the NM ($PSIC_m$ = 0.24). Accordingly, the dPSIC score is the most closely correlated to the experimental data among all individual features, which has an MCC = 0.54 when a cutoff of −1.1 is used. The same trend holds true for the tPSIC scores based on multiple threading alignments constructed by the monomeric threading LOMETS program [31]. Apparently, both PSIC and tPSIC features have shown that the mutations from a desirable amino acid (with a higher positive PSIC score) to an undesirable amino acid (with a lower negative PSIC score) tend to cause a disease.

The Pfam database is a large collection of protein domain families represented by profile-HMMs. Here, the Pfam score reports a posterior probability for each residue that aligns with a "match" or "insert" state in a

profile-HMM. We can find that the DMs are more likely to be found in the Pfam families than the neutral ones. For those mutations occurred at Pfam families, for example, the wild-type amino acids tend to have a lower profile score ($Pfam_w$) in the DM data set than those in the NM data set, while the mutant amino acids have a higher profile score ($Pfam_m$) in the DM data set than in the NM data set.

Contact features

DAMpred considers both intra- and inter-chain contact features derived from SPRING protein–protein interaction models. Fig. S5 shows the histogram distribution of various contact scores between DM and NM. From the intra-chain contact score (Intra), the DMs tend to appear in more crowded environment, as Intra scores have a higher value in the DM than in the NM data set. The differences between the two data sets are statistically highly significant, with a p value being 5.80E−245. In addition, the inter-chain contact score (Inter) also shows that the mutations adjacent to or in the proteins-proteins interface are more likely associated with disease than NMs.

TASSER model-based features

Fig. S6 presents the histogram distribution of the features derived from the I-TASSER-predicted structure models. The DMs have a higher probability in the surface concave regions than the NMs, as reflected by the cavity score calculated by ConCavity [32]. The average depths of DM and NM are 6.72 and 5.50 Å, respectively, which suggests that the DMs tend to occur at a deeper region of the protein structure. This is partially because the core regions are usually more tightly packed than the surface areas and mutations in the deep core regions have a higher chance to affect the function and stability of the protein and thus cause diseases. Consistent with the previous observation that most of the mutations decrease the folding stability of proteins [28], the DMs have a larger free-energy reduction (average $\Delta\Delta G$ = 1.62 kcal/mol) than the neutral ones (average $\Delta\Delta G$ = 0.52 kcal/mol). The difference is statistically significant with a p value 8.97E−91 in the Mann–Whitney test, where 2033 out of 4277 mutations with $\Delta\Delta G > 1.62$ kcal/mol are associated with disease and 3021 out of 4029 mutations with $\Delta\Delta G < 0.52$ kcal/mol are neutral. Simply using the energy cutoff with <0.52 kcal/mol for NMs and >1.62 kcal/mol for mutations associated with diseases, we can obtain recognitions with an appreciable MCC = 0.24 between $\Delta\Delta G$ score and the experimental data (disease or neutral).

Fig. S6 also shows correlation data for several other structure derivative terms, including structural profile score from EvoDesign [33], van der Waals, and side-chain packing from CIS-RR [34], which demonstrate again the tendency that the mutations associated with more drastic changes on protein structures are more likely to be disease-associated than neutral ones.

Overview of other individual features

The histogram distributions for all other features used in DAMpred are listed in Figs. S7–S10 in SI. Overall, the features derived from mutant amino acid are more sensitive than the wild-type amino acid, and the feature differences between wild-type and mutations is generally more sensitive than the individual wild-type or mutant features in distinguishing the DMs and the neutral ones. However, the p value is lower than 0.05 for almost all the feature types as shown in Table S2, suggesting that the designed feature functions have the potential to recognize the DM and NM.

Prediction of DMs on the D10634 data set

Comparison of BANN with other machine-learning methods

One of the important innovations of DAMpred is the employment of the BANN, in which the posterior probabilities of the features for classification are integrated into the neural network model. To examine the efficiency of BANN, we present its performance in Table S3 by comparing with other four models trained by the gradient boosting classifier [35], K-nearest neighbor classifier [36], support vector classifier [37], and artificial neural network. Here, all the trainings are implemented by the Scikit-learn toolkit [38], where the tests are performed in protein-level 10-fold cross-validation.

The data show that although there are some variations among the results by different methods, the BANN model consistently outperforms the models trained by other machine-learning methods. In particular, based on the MCC, which is the most important assessment parameter to balance precision and recall, the BANN model with 70 features is 15.7% higher than the second-best model by ANN (0.601 *versus* 0.580). The p value of the difference in the Mann–Whitney test is 3.99E−3, indicating that the difference is statistically significant; when trained on the top 20 features (see below), the difference between BANN and other methods becomes much more significant with p value <2.25E−10. Here, the only difference between BANN and ANN is that the former incorporates the posterior probabilities of the features into the network training, where the improvement of BANN over ANN suggests the efficiency of the inclusion of the Bayes classifier into the network training.

We also test the data using different feature sets. In the first set, we examine the power of individual features by calculating the p value of Mann–Whitney of their distributions in the disease-associated and neutral data sets, and then select the top 20 features that have the lowest p values. The second set

includes all 70 features. The results show that BANN outperforms the control training methods in both feature sets, demonstrating the robustness of the BANN training. Interestingly, the MCC of the models trained by top 20 features is only slightly worse than that the full-set models, indicating that the performance of the DAMpred mainly relies on the efficient features that have the significant distinguishing power. We also test the model trained on the worst 20 features with the highest *p* value, where the performance is much worse with a significantly lower MCC (0.317). Nevertheless, the higher MCC value achieved by the full-set feature model shows that the use of more features is still needed to achieve the best distinguishing power for the DAMpred model. In the following, we will report the results trained by the full-set of features unless specifically clarified.

Impact of protein stability and PPI features on DAMpred performances

Two core feature groups introduced in DAMpred is the protein-structure stability from I-TASSER models and the contact environments from SPRING PPI prediction. To get a quantitative assessment of the impact of these feature groups on mutation classifications, we re-trained two DAMpred models by dropping off the stability and PPI-associated features separately. The cross-validation tests show that dropping off each of the feature groups can significantly impact the performance of DAMpred with the MCC reduced from 0.601 to 0.593 (by dropping off stability) and 0.588 (by dropping off PPI), which correspond to a *p* value of 7.07E−08 and 9.77E−20, respectively. Ninety-five (or 1994) DMs, which were successfully identified by DAMpred with 70 features, are incorrectly classified into neutral group by the model without $\Delta\Delta G$ (or PPI) feature. These data suggest that while both features are needed to achieve the optimal classification performance, DAMpred seems more sensitive to the PPI than the stability features as the reduction of performance by the former is more significant.

Comparison of DAMpred with other mutation prediction methods

In Table 1 (top panel), we conduct a protein-level 10-fold cross-validation test of the DAMpred on the D10634 data set, in control with three state-of-the art mutation prediction methods, SIFT [6], SNAP2 [6,14], and PolyPhen2 [13], which are all installed in our local computer and run with the default setting. It is observed that DAMpred has an average MCC of 0.601, which is 9.1% higher than the second-best method from PolyPhen2 (0.551). The sensitivities of SIFT, SNAP2, and PolyPhen2 in positive and negative cases are severely unbalanced (0.873 for the positive and 0.561 for the negative on average), which indicates that these methods have incorrectly categorized too many NMs as DMs. In contrast, the DAMpred prediction has a more balanced prediction on the positive and negative samples with a sensitivity of 0.812 and 0.788, respectively. Accordingly, the relative sensitivity and specificity in the positive samples are also more balanced in DAMpred, which is one of the main reasons that DAMpred can have a higher global performance as judged by MCC than the control methods.

Impact of BANN on control methods

Both SNAP2 [14] and PolyPhen2 [13] models recognize the mutations based on traditional machine-learning methods, where SNAP2 uses the artificial neural network and PolyPhen2 uses the naive Bayes classifier. To further examine the impact of the BANN training to the performance of the mutation recognition, we extracted 20 features of the SNAP2 and PolyPhen2 programs with the lowest *p* value, and then retrain the models separately by BANN. The cross-validation results are summarized in Table 1, labeled as "SNAP2 + BANN" and "PolyPhen2 + BANN," respectively.

It is shown that the MCC of SNAP2 was improved by 0.134 using BANN (from 0.451 to 0.585), and the MCC of PolyPhen2 is improved by 0.029 using BANN (from 0.551 to 0.580). The *p* value of the changes in the Mann–Whitney test is 1.42E−84 and 3.32E−24, respectively, indicating that the improvement is statistically significant. The data suggest again that BANN is more efficient than the traditional artificial neural network and the naive Bayes classifier methods used by the original programs of SNAP2 and PolyPhen2. Although the DAMpred is only slightly (but statistically significantly) better than SNAP2 + BANN and PolyPhen2 + BANN, the difference of sensitivity of the DAMpred predictions (0.024) between positive and negative cases is less than that of SNAP2 + BANN (0.07) and PolyPhen2 + BANN (0.125), which suggests that the features exploited in DAMpred have probably a more balanced recognition ability for discriminating the DM and NM data sets.

Test of DAMpred on three independent data sets from D2186, D146, and p53 protein

In addition to the D10634, DAMpred was tested on three other independent data sets (Table 1). First, on the D2186, DAMpred generates prediction results with MCC/ACC (=0.503/0.752) lower than that in the cross-validation on the D10634 (0.601/0.800). However, the values are still significantly higher than the three control methods with a *p* value below 2.45E−45 in McNemer's test for all the comparisons.

Since both D10634 and D2186 may contain samples from the training data sets used by the control methods,

Table 1. Comparison of different methods for recognizing disease-causing mutation: DM and NM.

Methods	MCC	ACC	Positive		Negative		p[a]
			SEN	SPE	SEN	SPE	
D10634 data set in 10-fold cross-validation							
DAMpred	**0.601**	**0.800**	0.812	**0.796**	**0.788**	0.805	
SIFT	0.536	0.763	0.861	0.721	0.664	0.826	2.17E−54
SNAP2	0.451	0.732	0.871	0.715	0.551	0.767	1.50E−120
PolyPhen2	0.551	0.768	**0.887**	0.718	0.648	**0.850**	6.62E−80
SNAP2 + BANN[b]	0.585	0.792	0.827	0.776	0.757	0.811	4.41E−5
PolyPhen2 + BANN[c]	0.580	0.788	0.850	0.757	0.725	0.828	1.95E−19
D2186 data set							
DAMpred	**0.502**	**0.753**	0.742	0.725	0.762	0.777	
SIFT	0.442	0.706	0.843	0.635	0.589	0.816	2.45E−45
SNAP2	0.485	0.713	**0.917**	0.628	0.541	**0.885**	3.11E−86
PolyPhen2	0.406	0.683	0.856	0.613	0.534	0.810	2.912E−51
D146 data set							
DAMpred	**0.521**	**0.781**	0.767	**0.600**	**0.786**	0.890	
SIFT	0.394	0.674	0.814	0.493	0.609	0.875	2.10E−3
SNAP2	0.397	0.667	0.837	0.500	0.581	0.877	1.88E−3
PolyPhen2	0.461	0.699	**0.884**	0.494	0.621	**0.928**	3.126E−4
SNPMuSiC[d]	0.177	0.644	0.465	0.408	0.718	0.763	4.61E−3
TP53 data set							
DAMpred	**0.401**	**0.787**	0.881	**0.843**	**0.500**	0.579	
SIFT	0.279	0.719	0.791	0.828	**0.500**	0.440	8.32E−2
SNAP2	0.316	**0.787**	**0.970**	0.793	0.227	**0.714**	3.71E−4
PolyPhen2	0.295	0.742	0.836	0.824	0.455	0.476	5.67E−1
SNPMuSiC[d]	0.295	0.750	0.830	0.845	0.471	0.444	3.39E−2

The results on D10634 data set are by the protein-level 10-fold cross-validation, and those on D2186, D146, and TP53 data sets are from the models trained on non-homologous samples. Bold fonts highlight the best predictor in each category.

[a] p Value of MCC comparison in McNemar's test (for D10634) or Student's t test (for TP53) is calculated between DAMpred other predictors.

[b] SNAP2 + BANN: model re-trained by BANN using the top 20 features from SNAP2 selected with the lowest p value.

[c] PolyPhen2 + BANN: model trained by BANN using the top 20 features from PolyPhen2 selected with the lowest p value.

[d] SNPMuSiC result from the online server for the data sets with known experimental structures.

we tested the methods on a second set of D146, which contains 146 mutation samples from the D10634 but from the proteins that are non-homologous to the training sets of the control methods. Here, SNPMuSiC [15] was included in the test with data taken from the online server, where the experimental structures from the PDB were specified when submitting the on-line jobs. Because D146 is a subset of D10634 and the DAMpred model trained on the entire D10634 data set might be an over-fit for the test on the D146, we re-trained DAMpred specifically on a subset of samples from D10634, which are non-homologous to D146 with a sequence identity $<30\%$. Rows 15–19 of Table 1 show the comparison of the retrained DAMpred with other control methods on the D146 data set. The result shows again that DAMpred outperforms the control methods in both MCC and ACC. The p values between DAMpred and other methods are higher in D146 than other data sets due to the relatively smaller size of the D146, but they are still statistically significant with all being lower than 4.6E−3. These data suggest that the superiority of DAMpred is quite robust and not dependent on the data sets tested.

The third independent test is on the mutations from the p53 protein, one of the most important tumor suppressors that regulates cell growth pathways through DNA binding. Table 1 (bottom panel) shows a summary of the predictions, where only deleterious and neutral mutant samples were considered. DAMpred achieves an MCC of 0.401, compared to 0.279, 0.316, 0.295, and 0.295 by SIFT, SNAP2, PolyPHen2, and SNPMuSiC, respectively. Since P53 protein is highly enriched in with DMs, it is very difficult to correctly recognize the neutral variants, which is part of the reason that the overall performance on p53 was considerably worse than that on other data sets.

Nevertheless, the obvious higher MCC value achieved by DAMpred suggests that the pipeline generated a relatively more balanced classification than the control methods. In Fig. S2B, we present the structure model by I-TASSER for the isoform P04637-1 of p53 protein, where the majority of the DMs are located in the core domain, which accommodates the sequence-specific DNA binding activities. These data help further explain the functional annotation by DAMpred. For example, mutations of N235S and R290L in the disease-associated group are correctly identified only by DAMpred and SIFT. The feature table in the derivative DAMpred databases marks N235 and R290 as BINDING and SIGNAL, where their directly

intra contacts (15 and 10, respectively) are all functional. Furthermore, V97I in the neutral group is correctly recognized only by DAMpred, where V97 is labeled as nonfunctional site. On further view of all 70 features of the deleterious (R290L) and neutral V97I) mutations, the numbers of features with a higher probability in the deleterious data set (compared with that in neutral data set) are 35 and 18, respectively; these data also highlight the importance of the integration of the inherent probability of the different features.

Unfortunately, no method has correctly recognized P309S as disease-associated. As shown in Fig. S2B, P309 is located at the loop region and there are no directly intra or inter contacts. The number of features with a higher probability in disease-associated data set is only 15, where dominant higher scores in the neutral data set have led DAMpred to incorrectly assign the mutation as neutral. Thus, there is still considerable room for further improvement. The overall prediction results by the four predictors can be downloaded at https://zhanglab.ccmb.med.umich.edu/DAMpred/download/TP53.xlsx.

Conclusion

We have developed a new pipeline, DAMpred, for recognizing the DMs in the human genome. A major uniqueness of the pipeline is the employment of the features extracted from low-resolution structure prediction of both monomer and BioUnit structure of the target protein, in addition to other resources of pharmacophore and evolutionary profiles. Second, a novel machine-learning method (BANN) is introduced to integrate the posterior probabilities of Bayesian classifiers with the neural network training to improve the efficiency of neural network training.

The pipeline was tested on four benchmarking data sets, D10634, D2186, D146, and the p53 protein, which demonstrated advantage in sensitive disease mutation recognition compared to four state-of-the-art methods built on evolution and machine learning, respectively. Detailed analysis shows that the performance gap between DAMpred and the control methods can be partly reduced by applying the BANN training algorithm to the control programs, further demonstrating the advantage of the BANN that mainly stems from the non-linear combination of the classifier instead of the linear combination as taken by the traditional artificial network training. Even with the same training algorithm, DAMpred still shows a better performance than the control methods, suggesting the advantage of feature selections of DAMpred by combining multiple sources of features, especially those from structure prediction and BioUnit structures.

While the results of DAMpred are promising, several of its limitations should be acknowledged. First, there is a modest correlation observed between structural modeling accuracy and DAMpred performance. Although the majority of the structure models (86%) by the cutting-edge tools such as I-TASSER could have a correct fold with a TM score of >0.5, the local structure error, especially those involved in the functional regions, could compromise the accuracy of DAMpred. Second, only part of targets (20% in our data set) could not have BioUnit structure reliably constructed by multimeric threading SPRING due to the lack of homologous templates of complex structures even with a relatively loose homology filter. Therefore, the BioUnit contact information cannot be fully implemented in a considerable portion of sequences. Apparently, with the continuous progress of on the computational structure modeling, as witnessed by both CASP and CAPRI examples [39–41], the DAMpred pipeline should benefit from the improvement of the modeling accuracy from both monomer and complex structure predictions in the future.

Methods

DAMpred consists of two general steps of feature collection and Bayes-guided artificial neural network training, where a flowchart is depicted in Fig. 1.

Multi-source feature collections

As a machine-learning based approach, the design and selection of the training features play a key role in the disease mutation prediction in DAMpred. Many common feature properties of proteins, including residue physicochemical properties and conservation profile in evolution, have been shown to have strong correlations with the mutation classification in previous studies [3,6,15,29]. One of the major motivations in the DAMpred development is to examine the impact of structural characteristics of proteins, in particular those from low-resolution homologous structure predictions, on the functional and deleterious classification of SNP mutations. In this regard, we designed multiple levels of residue contact and van der Waals interaction features, built on the tertiary and quaternary complex structure prediction by I-TASSER [20,26] and SPRING [22], which are combined with a set of classical features on pharmacophore and evolutional profiles that are extended from the previous studies. Overall, DAMpred collects 70 individual features, the details of which are listed in Table S2 with the feature extraction process depicted in Fig. 1a and Text S2. These features are categorized into four groups based on their properties.

Physicochemical properties

The physicochemical property features in DAMpred include pharmcophore of the target residues [29] and

the mutation-induced environmental pharmacophore changes, which are described in Fig. S11. In addition, we consider the common physicochemical properties, including the volume and weight from the wild-type and mutant residues (see Text S2).

Evolutionary profiles

Evolution is a major driven force for protein structure and function determination, where sequence profiles from multiple sequence alignments contain information on how the protein families evolve. To identify evolution relations between sequences, which are often distantly homologous, three sequence profiles are collected in DAMpred by PSI-BLAST [42], LOMETS [31], and Pfam [43], separately (see Eqs. S2–S3 in SI).

The contact environments in SPRING biological assembly

DAMpred considers four types of contact-environment features deduced from the complex structural models built by SPRING [22]; these include the number of intramolecular contacts (Intra), the number of intramolecular contacts involving functional residues (FunIntra), the number of intermolecular contacts (Inter), and the number of intermolecular contacts involving functional residues (FunInter). In addition, DAMpred considers an enlarged contact environment by counting the residues that are in contact with the contacting partners of the mutant residues, called indirect contacts (see Fig. S12).

I-TASSER modeling and structure-based feature extraction

I-TASSER [20,26] was used to construct three-dimensional models for both wild-type and mutant sequences, where two groups of structure-based features, on protein surface and physics-based energy terms, are extracted from the I-TASSER models (Text S2).

Bayes-guided artificial neural network (BANN)

DAMpred models are trained by hybrid learning approach combing the artificial neural network with an extended form of naïve Bayesian classifier. Here, we consider two classes of mutations, i.e. C_D for the DM and C_N for the NM, respectively. Given a class variable C_k ($k = D$ or N) and n specific features $F = (f_1, f_2, \cdots, f_n)$, the posterior probability $P(C_k|F)$ of the feature F associated with the class C_k can be written, by the naïve Bayes classifier model, as:

$$P(C_k|F) \propto P(C_k) \prod_{i=1}^{n} P(f_i|C_k) \tag{1}$$

or

$$\begin{aligned} \log P(C_k|F) &\propto \log\left(P(C_k)\prod_{i=1}^{n} P(f_i|C_k)\right) \\ &= \sum_{i=1}^{n} \log P(f_i|C_k) + \log P(C_k) \end{aligned} \tag{2}$$

However, since the naïve assumption that the naïve Bayes classifier model is based on, that is, the considered features are independent from each other, is not always true, Eq. (2) can be written in a more general form for specific mutation classes (D and N):

$$\begin{cases} S_D = \log P(C_D|F) = \sum_{i=1}^{n} \alpha_i(F) \log P(f_i|C_D) + \alpha_{n+1}(F) \log P(C_D) \\ S_N = \log P(C_N|F) = \sum_{i=1}^{n} \alpha_i(F) \log P(f_i|C_N) + \alpha_{n+1}(F) \log P(C_N) \end{cases} \tag{3}$$

and

$$\begin{aligned} S_\Delta = S_D - S_N = &\sum_{i=1}^{n} \alpha_i(F)[\log P(f_i|C_D) - \log P(f_i|C_N)] \\ &+ \alpha_{n+1}(F)[\log P(C_D) - \log P(C_N)] \end{aligned} \tag{4}$$

This form of linear combination of the logarithm of prior and likelihood probabilities makes it possible to integrate the determination of the $n + 1$ weight parameters, $\alpha_i(F)$, with neural network (see Text S3). Here, we use the artificial neural network (ANN) for training the three sets of weights for S_D, S_N, and S_Δ, separately. The structure of network is depicted in Fig. S13. Once the $\alpha_i(F)$ values are determined from the network training, the probability of the mutation states and the difference can be estimated by S_D, S_N, and S_Δ in Eqs. (3) and (4), where a higher value of S_Δ indicates the higher possibility of the mutation to be disease-associated. Generally, if $S_D > S_D^0$, $S_N < S_N^0$ or $S_\Delta > S_\Delta^0$, the mutation will be classified as a DM. Here, we set S_D^0, S_N^0, and S_Δ^0 all equal to 0.5, where the final decision is made by voting from the three scores. The entire process of learning can be viewed in Fig. 1b, with detailed derivation of the BANN described in Text S3.

Acknowledgments

The work was supported in part by the National Institute of Allergy and Infectious Diseases (AI134678), National Institute of General Medical Sciences (GM083107, GM116960), National Science Foundation (DBI1564756), and National Natural Science Foundation of China (31801108, 61772357).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmb.2019.02.017.

Received 14 November 2018;
Received in revised form 9 February 2019;
Accepted 11 February 2019
Available online 21 February 2019

Keywords:
non-synonymous single nucleotide polymorphisms;
protein structure prediction;
Bayes-guided artificial neural network algorithm;
p53 protein;
protein–protein interaction

Abbreviations used:
DAMpred, disease-associated mutation prediction; BANN, Bayes-guided artificial neural network algorithm; DM, disease-associated mutation; NM, neutral mutation.

References

[1] J.F. Baranoski, M.Y.S. Kalani, C.J. Przybylowski, J.M. Zabramski, Corrigendum: cerebral cavernous malformations: review of the genetic and protein–protein interactions resulting in disease pathogenesis, Front. Surg. 4 (2017) 31.
[2] C.M. Yates, M.J.E. Sternberg, Proteins and domains vary in their tolerance of non-synonymous single nucleotide polymorphisms (nsSNPs), J. Mol. Biol. 425 (2013) 1274–1286.
[3] N.D. Dees, Q. Zhang, C. Kandoth, M.C. Wendl, W. Schierding, D.C. Koboldt, et al., MuSiC: identifying mutational significance in cancer genomes, Genome Res. 22 (2012) 1589–1598.
[4] A. Gonzalez-Perez, N. Lopez-Bigas, Functional impact bias reveals cancer drivers, Nucleic Acids Res. 40 (2012) e169.
[5] E. Porta-Pardo, A. Godzik, E-Driver: a novel method to identify protein regions driving cancer, Bioinformatics. 30 (2014) 3109–3114.
[6] P. Kumar, S. Henikoff, P.C. Ng, Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm, Nat. Protoc. 4 (2009) 1073–1081.
[7] B. Reva, Y. Antipin, C. Sander, Predicting the functional impact of protein mutations: application to cancer genomics, Nucleic Acids Res. 39 (2011) E118–U85.
[8] Y. Choi, G.E. Sims, S. Murphy, J.R. Miller, A.P. Chan, Predicting the functional effect of amino acid substitutions and Indels, PLoS One 7 (2012).
[9] P. Katsonis, O. Lichtarge, A formal perturbation equation between genotype and phenotype determines the evolutionary action of protein-coding variations on fitness, Genome Res. 24 (2014) 2050–2058.
[10] E. Capriotti, R.B. Altman, Y. Bromberg, Collective judgment predicts disease-associated single nucleotide variants, BMC Genomics 14 (2013).
[11] Y. Itan, L. Shang, B. Boisson, M.J. Ciancanelli, J.G. Markle, R. Martinez-Barricarte, et al., The mutation significance cutoff: gene-level thresholds for variant predictions, Nat. Methods 13 (2016) 109–110.
[12] H.Y. Zhou, M. Gao, J. Skolnick, ENTPRISE: an algorithm for predicting human disease-associated amino acid substitutions from sequence entropy and predicted protein structures, PLoS One 11 (2016).
[13] I.A. Adzhubei, S. Schmidt, L. Peshkin, V.E. Ramensky, A. Gerasimova, P. Bork, et al., A method and server for predicting damaging missense mutations, Nat. Methods 7 (2010) 248–249.
[14] M. Hecht, Y. Bromberg, B. Rost, Better prediction of functional effects for sequence variants, BMC Genomics 16 (Suppl. 8) (2015) S1.
[15] F. Ancien, F. Pucci, M. Godfroid, M. Rooman, Prediction and interpretation of deleterious coding variants in terms of protein structural stability, Sci. Rep. 8 (2018) 4480.
[16] E. Boutet, D. Lieberherr, M. Tognolli, M. Schneider, P. Bansal, A.J. Bridge, et al., UniProtKB/Swiss-Prot, the Manually Annotated Section of the UniProt KnowledgeBase: How to Use the Entry View. Plant Bioinformatics: Methods and Protocols, 2nd edition1374, 2016 23–54.
[17] V.A. McKusick, Mendelian inheritance in man and its online version, OMIM, Am. J. Hum. Genet. 80 (2007) 588–604.
[18] E. Capriotti, R. Calabrese, R. Casadio, Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information, Bioinformatics. 22 (2006) 2729–2734.
[19] J. Bendl, J. Stourac, O. Salanda, A. Pavelka, E.D. Wieben, J. Zendulka, et al., PredictSNP: robust and accurate consensus classifier for prediction of disease-related mutations, PLoS Comput. Biol. e1003440 (2014) 10.
[20] J. Yang, R. Yan, A. Roy, D. Xu, J. Poisson, Y. Zhang, The I-TASSER suite: protein structure and function prediction, Nat. Methods 12 (2015) 7–8.
[21] C. Zhang, S.M. Mortuza, B. He, Y. Wang, Y. Zhang, Template-based and free modeling of I-TASSER and QUARK pipelines using predicted contact maps in CASP12, Proteins. 86 (Suppl. 1) (2018) 136–151.
[22] A. Guerler, B. Govindarajoo, Y. Zhang, Mapping monomeric threading to protein–protein structure prediction, J. Chem. Inf. Model. 53 (2013) 717–725.
[23] UniProt C, The Universal Protein Resource (UniProt) in 2010, Nucleic Acids Res. 38 (2010) D142–D148.
[24] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, et al., The Protein Data Bank, Nucleic Acids Res. 28 (2000) 235–242.
[25] M. Gao, H.Y. Zhou, J. Skolnick, Insights into disease-associated mutations in the human proteome through protein structural analysis, Structure. 23 (2015) 1362–1369.
[26] A. Roy, A. Kucukural, Y. Zhang, I-TASSER: a unified platform for automated protein structure and function prediction, Nat. Protoc. 5 (2010) 725–738.
[27] J. Xu, Y. Zhang, How significant is a protein structure similarity with TM-score = 0.5? Bioinformatics. 26 (2010) 889–895.
[28] L. Quan, Q. Lv, Y. Zhang, STRUM: structure-based prediction of protein stability changes upon single-point mutation, Bioinformatics. 32 (2016) 2936–2946.
[29] D.E. Pires, D.B. Ascher, T.L. Blundell, mCSM: predicting the effects of mutations in proteins using graph-based signatures, Bioinformatics. 30 (2014) 335–342.
[30] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.
[31] S. Wu, Y. Zhang, LOMETS: a local meta-threading-server for protein structure prediction, Nucleic Acids Res. 35 (2007) 3375–3382.

[32] J.A. Capra, R.A. Laskowski, J.M. Thornton, M. Singh, T.A. Funkhouser, Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure, PLoS Comput. Biol. 5 (2009), e1000585.
[33] P. Mitra, D. Shultis, Y. Zhang, EvoDesign: de novo protein design based on structural and evolutionary profiles, Nucleic Acids Res. 41 (2013) W273–W280.
[34] Y. Cao, L. Song, Z. Miao, Y. Hu, L. Tian, T. Jiang, Improved side-chain modeling by coupling clash-detection guided iterative search with rotamer relaxation, Bioinformatics. 27 (2011) 785–790.
[35] J.H. Friedman, Greedy Function Approximation: A Gradient Boosting Machine, 2001 1189–1232.
[36] T.M. Cover, P.E. Hart, Nearest Neighbor Pattern Classification, 13, 1967 21–27.
[37] C. Cortes, V. Vapnik, Support-Vector Networks, 20, 1995 273–297.
[38] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al., Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[39] J. Moult, K. Fidelis, A. Kryshtafovych, T. Schwede, A. Tramontano, Critical assessment of methods of protein structure prediction: progress and new directions in round XI, Proteins. 84 (Suppl. 1) (2016) 4–14.
[40] M.F. Lensink, S.J. Wodak, Docking, scoring, and affinity prediction in CAPRI, Proteins. 81 (2013) 2082–2095.
[41] A. Szilagyi, Y. Zhang, Template-based structure prediction of protein–protein interactions, Curr. Opin. Struct. Biol. 24 (2014) 10–23.
[42] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, et al., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402.
[43] R.D. Finn, P. Coggill, R.Y. Eberhardt, S.R. Eddy, J. Mistry, A. L. Mitchell, et al., The Pfam protein families database: towards a more sustainable future, Nucleic Acids Res. 44 (2016) D279–D285.