RESEARCH ARTICLE

ENTPRISE-X: Predicting disease-associated frameshift and nonsense mutations

Hongyi Zhou, Mu Gao, Jeffrey Skolnick*

Center for the Study of Systems Biology, School of Biological Sciences, Georgia Institute of Technology, Atlanta, Georgia, United States of America

* skolnick@gatech.edu

OPEN ACCESS

Citation: Zhou H, Gao M, Skolnick J (2018) ENTPRISE-X: Predicting disease-associated frameshift and nonsense mutations. PLoS ONE 13 (5): e0196849. https://doi.org/10.1371/journal.pone.0196849

Editor: Alexandre G. de Brevern, UMR-S1134, INSERM, Université Paris Diderot, INTS, FRANCE

Received: November 14, 2017

Accepted: April 20, 2018

Published: May 3, 2018

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by Grant No. 1R35GM-118039 of the Division of General Medical Sciences of the National Institutes of Health to JS. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

To exploit the plethora of information provided by Next Generation Sequencing, the identification of the genetic mutations responsible for disease in general or cancer in particular, among the thousands of neutral germline or somatic variations is a crucial task. Genome-wide association studies for the detection of disease-associated genes or cancer drivers can only identify common variations or driver genes in a cohort of patients. Thus, they cannot discover unique disease-associated mutations or cancer driver genes on a personal basis. Moreover, even when there are such common variations, their significance is unknown. Here, we extend the machine learning based approach ENTPRISE developed for predicting the disease association of missense mutations to frameshift and nonsense mutations. The new approach, ENTPRISE-X, is shown to outperform the state-of-the-art methods VEST-indel and DDIG-in for predicting the disease association of germline frameshift mutations in terms of balanced measure Matthew's correlation coefficient, MCC, with a MCC of 0.586 for ENTPRISE-X, versus 0.412 by VEST-indel and 0.321 by DDIG-in, respectively. Large scale testing on the ExAC dataset shows ENTPRISE-X has a much lower fraction of 16% of variations classified as disease causing, as compared to VEST-indel's 26% and DDIG-in's 65% of predictions as being disease-associated. A web server for ENTPRISE-X is freely available for academic users at http://cssb2.biology.gatech.edu/entprise-x.

Introduction

Identifying the causative genetic variations of a disease or cancer is a key step towards diagnosis and cure. A typical patient has thousands of genetic variations involving thousands of genes. It is thus a nontrivial task to pinpoint the single variation or group of variations responsible for the given disease. Traditionally, genome wide association studies (GWAS) are employed for detecting common variations among a cohort of same disease patients [1, 2]. Due to the heterogeneity of human disease [3], GWAS can only cover a small fraction of patients. Thus, for the majority of patients, new approaches for personalized detection of disease-associated variations are needed. Since disease is a complex system phenomenon that is often caused by the synergetic effects of many factors including genetic, environmental, and life style effects, one can only infer that a specific mutation is associated with a disease, which means it could contribute to the onset of a disease. Even so, it might not be the only cause of the disease. A mutation, in particular frameshift and nonsense ones, could result in a loss or gain of function of the protein. If the function of the normal protein is essential for life, the mutated protein with functional consequences could result in disease subject to extrinsic factors. Because of their importance, many methods have been developed to predict disease-associated mutations [4–12]. Some can also help annotate somatic mutations in cancer [13]. Approaches specific for cancer driver mutations have also been developed [14–17]. More generally, our recently developed machine learning based method ENTPRISE outperforms previous methods in terms of the ability to distinguish neutral from disease-associated missense mutations in the same protein, false positive rate, and performance in cancer driver prediction [18].

Although ENTPRISE has outstanding performance for both Mendelian and cancer driver mutations and requires only single patient exome information, like many other methods, it can only handle missense amino acid substitutions. In practice, based on the v76 version of the COSMIC database [19], while 64.4% of the mutations are missense amino acid substitutions, 21.7% are silent, 8.4% are nonsense and frameshift mutations, 1.4% are in-frame indels (insertions/deletions), and the remaining 4.1% are uncharacterized. Furthermore, since both nonsense and frameshift mutations disrupt protein structure more severely than missense mutations and in-frame indels, they are more likely to have pathogenic consequences. In that regard, there are a number of methods dedicated to assessing the disease association of frameshift and/or nonsense mutations [17, 20–22]. However, as in the case of missense variations, these approaches often have quite high false positive rates, which limit their practical application when applied to entire exomes.

In this work, we extend the ENTPRISE approach [18] that successfully classifies the disease association of missense mutations to include nonsense and frameshift mutations. Since both nonsense and frameshift mutations disrupt protein structure at a given location within a protein's sequence, we will treat them identically. ENTPRISE uses a set of features derived from the protein's sequence and predicted tertiary structure in a boosted tree regression method [23] to learn a model from the training dataset. Here, we shall use similar features derived from predicted protein structures and include additional features related to the global functionality of the protein. The extended approach which is called ENTPRISE-X will be benchmarked against the best current state-of-the-art DDIG-in [20] and VEST-indel [21] methods for frameshift and nonsense mutations. DDIG-in is a SVM-based machine-learning method [24] that uses both DNA and protein sequence information as well as predicted accessible surface area and can address both frameshift and nonsense variations. VEST-indel deals with frameshift and in-frame variations and uses Random Forest classifiers [25] to train a model using 24 features including a "PubMed" feature for disease relevance. In addition to comparing to these established approaches, we also perform a large scale test on the ExAC dataset of 60,706 exomes from unrelated individuals [3]. While ExAC is not necessarily completely neutral (i.e. all variations are not disease associated), the number of variations classified by a method to be disease-associated should be strongly correlated with the false positive rate of the method. The method developed in this work is useful for personalized Mendelian disease and protein target identification and is freely available for academic users at http://cssb2.biology.gatech.edu/entprise-x.

Materials and method

Datasets

Training set. Pathogenic data (ClinVar) were downloaded from the NCBI site ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/tab_delimited/. Two neutral data sets are used. One is EPS6500 from http://evs.gs.washington.edu/EVS/ [26], and the other is from the 1000 Genomes Project phase 3 data [27]. The mutations employed for training exclude mutations in proteins having sequence identity $\geq$ 35% to any proteins in the test data set described below.

Testing set. To compare ENTPRISE with VEST-indel and DDIG-in, we use the same frameshift testing set as employed in the VEST-indel [21]. It consists of 184 pathogenic mutations from the ClinVar database and 1,340 neutral ones from the SIFT-indel method constructed from inter-species multiple sequence alignment [22]. After mapping the protein sequences to structures in our predicted human exome protein structure database [18], we obtain 82 pathogenic and 1,025 neutral frameshift mutations. Since there are no independent data for comparing methods on nonsense mutations, we adopt a similar 10 fold cross-validation approach as was employed in the DDIG-in method [20] to benchmark ENTPRISE-X for nonsense mutations.

ExAC set. For large scale testing of our method in comparison to existing algorithms, we downloaded the ExAC dataset [28] that excludes the Cancer Genome Atlas (TCGA) data (http://cancergenome.nih.gov/). The goal of ExAC is to provide a global "reference set" for filtering out harmless genetic variants observed in patients with some disease. However, since the data only excluded childhood diseases, it could still contain variations causing disease in adults. The ExAC variations excluding the Cancer Genome Atlas (TCGA) data are considered to be mostly, but not completely, neutral [28]. A method with a larger false positive rate will result in a larger fraction of data as being classified as disease causing. Thus, to a rough first approximation, this large-scale test can be used to estimate the false positive rate of a method. After mapping proteins to our predicted protein structure database and excluding mutations in proteins having a sequence identity $\geq$ 35% to any protein in the training set, we obtained 56,917 putative neutral frameshift and 45,131 putative neutral nonsense mutations.

A brief summary of training and testing datasets on frameshift and nonsense used for ENTPRISE-X mutations are given in Table 1.

Table 1. Summary of variations in the ENTPRISE-X training and testing data sets.

Data Set				Usage
Training set				1. For training a model in future applications. 2. For feature reduction. 3. For large scale, ten-fold cross-validation test on nonsense mutations in comparison to DDIG-in, and on frameshift mutations in comparison to DDIG-in & SIFT-indel (see Table 3).
Frameshift		Nonsense		
Pathogenic	Neutral	Pathogenic	Neutral	
ClinVar: 6,513	ESP6500: 1,604	ClinVar: 5,023	ESP6500: 181	
	1000 GP: 366		1000 GP: 3,171	
Total numbers (sum of each column)				
6,513	1,970	5,023	32,51	
Independent testing sets (not used in training)				Usage
VEST-indel set				For test on frameshift variations in comparison to VEST-indel & DDIG-in methods (see Table 2).
Frameshift		Nonsense		
Pathogenic	Neutral	Pathogenic	Neutral	
ClinVar: 82	Inter-species: 1,025	—	—	
ExAC set				For large scale false positive rate test on frameshift & nonsense variations in comparison to the VEST-indel & DDIG-in methods
—	ExAC: 56,917	—	ExAC: 45,131	

https://doi.org/10.1371/journal.pone.0196849.t001

Features and machine learning

Owing to the success of the features in the ENTPRISE method [18] for missense mutations, we tested the same features for frameshift and nonsense mutations. These are derived from the protein's sequence and predicted protein three-dimensional (3D) structural information. However, since for nonsense mutations the mutated protein has no amino acids after the mutation, the features describing the amino acid types of the mutated protein in ENTPRISE are not needed for nonsense mutations, whereas the features describing the amino acid type before the mutation are kept. For frameshift mutations, since the major result is similar to nonsense mutations (dysfunction of the part after mutation point), we also neglect the amino acid type right after the mutation. A detailed description of these features is found in Ref. [18]. Here, we give a brief summary and describe the newly introduced features specific for frameshift and nonsense mutations:

(a). The reference amino acid type of the mutated protein position. This requires 20 variables to represent the 20 types of amino acids. For a given residue type, the specific variable is set to the value of 1 and the rest of the variables are set to 0. This feature reflects the fact that mutations of certain types of amino acids are more likely to cause disease than others [29]. The wild type amino acid as well as the following contact composition and entropy information encode the functional importance of the mutated position when the protein functions normally.

(b). 20 variables for the amino acid composition of the residues in contact with the mutated protein position. This reflects the protein's local structural environment at the mutated position and is defined as the amino acid composition of all residues whose C_α atom is within 12 Å from the mutated position's C_α atom assuming that the structure of the protein is in the same native conformation. The consideration of this feature is that if a mutation is at or close to a protein-protein interface or protein-ligand binding pocket, the mutation is more likely to be disease associated [29]. Since certain residues are usually conserved in interface and binding sites, the amino acid composition around the mutated position encodes information if the mutation is close to an interface or binding pocket. Since for the majority of human proteins, experimental structures are not available, this work uses predicted protein structures from TASSER$^{\mathrm{VMT}}$ [30] as in ENTPRISE [18]. The composition is calculated by

$$q(a) = N(a)/\Sigma N(b), \tag{1}$$

where $N(a)$ is the number of residues of type a contacting the mutated position, and $\Sigma N(b)$ sums over all 20 types of residues.

(c). 20 variables for the composition of the domain that contains the mutated protein position. The reason for considering this feature is that certain domains/proteins are more likely to be disease associated than others [7]. Proteins/domains are distinguishable by their amino acid composition. This feature encodes the characteristics of the domain at the mutated position. The domains of the mutated proteins are predicted by threading the sequence using the fast HHpred threading algorithm [31] against the SCOP domain database [32] in which domains are manually defined. The computation of domain composition is the same as Eq (1) with $N(x)$ being replaced by the number of residues of type x in the whole domain.

(d). Sequence entropy of the mutated protein position. This characterizes the evolutionary conservation at the mutated position. Our assumption is that evolutionarily conserved residues are either functionally and/or structurally important. The sequence entropy is given by

$$S_r = S - < S >, \tag{2}$$

where $S = -\sum_{l=1}^{20} f_l \ln(f_l)$ is the sequence entropy of the given position, f_l is the normalized amino acid frequency profile generated by sequence search method PSI-BLAST [33]. $<>$ stands for averaging over the entire protein sequence.

The above 61 features are the same as in the ENTPRISE method and describe the features of the mutated position. We next introduce five additional new features that describe partial or whole protein functionality. The newly introduced features specific for frameshift & nonsense mutations are:

(e). Fraction of the affected protein structure defined as $(N_r - k)/N_r$ where N_r is the total number of residues of the protein, k is the sequential position of the mutation from the N-terminus. This feature describes how much of the protein is affected by the mutation that disrupts the structure. Our assumption is that the greater the fraction of the protein that is affected, the more likely it is that the mutation is disease associated.

(f). Essentiality of mutated protein. If the protein is homologous (defined as having protein sequence identity $>$ 35%) to an essential protein in the database of essential genes [34], a feature with a value of one is set; otherwise, it is set to zero. The consideration of this feature is based on the idea that if the protein is essential, its functional change is more likely to be associated with disease.

(g). Pathogenicity of the affected part of the protein. This feature is derived from the ENTPRISE calculation on the whole protein sequence. It reflects the average effects of the disrupted structure if all the positions have a missense mutation in turn, one position a time. The assumption is that if the probability of a missense mutation being disease associated in the missing dysfunctional part of a protein is larger, a frameshift or nonsense mutation that disrupts this region will be more likely to be disease associated. For any position in the reference protein sequence, there are 19 possible missense mutations. The pathogenicity of a position is defined as the ratio of the number of mutations having an ENTPRISE score $\geq$ 0.45 divided by 19. (Note that an ENTPRISE score $\geq$ 0.45 is considered to be pathogenic). The pathogenicity of the affected part of the protein is the average pathogenicity of all positions after the nonsense/frameshift position in the sequence.

(h). Disease involvement of the protein. This feature provides the information that a protein is associated with disease when its function is altered. It is obvious a consideration for a mutation on the protein to be disease associated or not when it is mutated. If the protein is related to any disease as defined by the GeneCards database [35], this feature receives a value of one. In addition, we have collected gene-disease associations from other two sources [36, 37] that were used in our recent work [38] for predicting gene-disease association. If a protein is associated with any disease in this additional dataset, this feature receives another value of one. In all other cases, this feature is set to zero.

(i). Number of protein-protein interactions inferred from the HIPPIE protein-protein interaction database [39]. This feature reflects how many other proteins that it might affect if the protein's function changes because of the mutation. Its consideration for disease association is that a protein with more interactions will likely affect more biological processes when its function is altered, and thus this increases the possibility of a mutation being disease associated.

We shall employ the machine learning boosted tree regression method that has been employed in many other applications [23, 40]. It generates a sequence of decision trees; each grows on the basis of the residuals of all previous trees [23, 41]. Here, a decision tree regression is implemented with a maximal depth of eight. The scoring function is represented as a boosted decision tree [23]:

$$f(x) = \sum_{m=1}^{N_{tree}} \varepsilon T_m. \tag{3}$$

Where T_m is a decision tree, ε is the shrinkage factor or learning rate, and N_{tree} is the number of trees. Two objective function values are adopted: 1 for disease-associated mutations, and 0 for neutral mutations. The learning rate ε is set to 0.005 and the number of trees, N_{tree}, is set to 2000. These values are purely empirical and were not optimized. As mentioned above, we treat frameshift and nonsense mutations exactly the same. Since in the training sets, the total number of disease-associated mutations (11,536) is roughly twice the total number of neutral mutations (5,222), to train the model on balanced datasets, we shall train two models: one with a randomly selected half of the disease-associated mutations and all neutral ones, the other with the remaining half of the disease-associated mutations and all neutral ones. This way, both models are trained on balanced datasets. In prediction/testing, a prediction will be given by the average score from the two models. In case of a jackknife or cross-validation test, the same training/testing protocol was applied to the training/testing subsets.

Feature reduction

Overall, we have tested the 66 features described above. However, not all types of features are useful. We shall employ a feature reduction procedure to eliminate features that have negative impact on performance. To this end, we cluster the whole training proteins into 10 clusters with a 35% protein sequence identity cutoff and employ a leave one cluster out cross validation (LOOCV) test (also see ten-fold cross-validation subsection). Then, for each cluster, we use nine clusters for training and the one that is left out for testing. To eliminate features, for each type of feature, we remove it from training and perform a LOOCV test; if the feature's removal results in a better Matthew's Correlation Coefficient (MCC) with a default cutoff of 0.5, then the feature is discarded. As a result of this procedure, the 20 features of domain composition (type (c) above) were eliminated, leaving us with total of 46 final features including all five newly introduced features.

Results

Comparison to other methods

We next tested ENTPRISE-X using the 46 final features for frameshift mutations on the VEST-indel test set [21] and compared the results with the VEST-indel and DDIG-in methods [20, 21] which according to Ref [21] are the two most accurate methods for frameshift variations. The results are presented in Table 2. Comparison of different approaches is made on the consensus subset (all three methods have predictions) of the full test set. The full set has 82 pathogenic and 1025 neutral mutations, whereas the consensus has 70 pathogenic and 914 neutral mutations. The difference in numbers of cases is due to the fact that some methods like

Table 2. Performance on the VEST-indel test set for frameshift variations.

Method	MCC[b]	Sensitivity	Specificity	F-score[c]	False positive rate	False discovery rate
Evaluation on the consensus subset[a]						
ENTPRISE-X	0.626 (0.749)	0.943	0.916	0.620 (0.767)	8.4%	54%
VEST-indel	0.440 (0.585)	0.914	0.814	0.421 (0.615)	18.6%	73%
DDIG-in	0.321 (0.441)	0.943	0.663	0.297 (0.439)	33.7%	82%
Evaluation on the full test set						
ENTPRISE-X	0.586	0.878	0.912	0.590	8.8%	55%
Baseline[d]	0.323	0.988	0.621	0.294	37.9%	83%
Baseline[e]	0.224	0.598	0.775	0.271	22.5%	83%
ENTPRISE-X_1[f]	0.570	0.878	0.905	0.574	9.5%	57%
ENTPRISE-X_2[f]	0.555	0.854	0.905	0.562	9.5%	58%
ENTPRISE-X_10alt[f]	0.587±0.006	0.887±0.006	0.910±0.003	0.590±0.006	9.0%±0.3%	55.8%±0.7%
ENTPRISE-X-nolocal	0.481	0.707	0.914	0.509	8.6%	60%
ENTPRISE-X-nonew	0.099	0.793	0.390	0.168	61.0%	90%
ENTPRISE-X-noratio	0.513	0.890	0.871	0.509	12.9%	64%
ENTPRISE-X-noessential	0.574	0.890	0.903	0.575	9.7%	58%
ENTPRISE-X-nopathogen	0.543	0.866	0.896	0.546	10.4%	60%
ENTPRISE-X-nodisease	0.368	0.683	0.859	0.396	14.1%	72%
ENTPRISE-X-nointeract	0.586	0.890	0.909	0.588	9.1%	56%

[a] To be fair to all methods, only the consensus mutations of three methods are evaluated in comparison to the other methods.

[b] Matthew's Correlation Coefficient. The numbers in parentheses are the maximal possible values.

[c] 2(precision×recall)/(precision+recall), where precision = (true positive)/(true positive + false positive), recall = (true positive)/(true positive + false negative). Numbers in parentheses are the maximal possible values.

[d] When only the feature representing if the gene is disease-associated or not is used.

[e] When only the feature representing if the gene is essential or not is used.

[f] When using one of the 2 models trained on each half of the pathogenic data and training ENTPRISE-X for 10 different random partitions of the pathogenic part of the training set were used.

https://doi.org/10.1371/journal.pone.0196849.t002

DDIG-in cannot interpret all of the input data. Results for VEST-indel (http://www.cravat.us/CRAVAT) and DDIG-in (http://sparks-lab.org/ddig/) are obtained from their respective webservers. Table 2 clearly shows that ENTPRISE-X with a default cutoff score 0.5 has the best overall performance in terms of its Matthew's Correlation Coefficient and F-score defined as 2 (precision×recall)/(precision+recall), where precision = (true positive)/(true positive + false positive), recall = (true positive)/(true positive + false negative). It also has the lowest false positive rate defined as (the number of neutral variations classified as disease-associated)/(the total number of neutral variations) and is related to specificity by 1-specificity. DDIG-in has the worst MCC and F-score. It tends to over-predict disease association and thus has the largest false positive rate. VEST-indel performs in the middle for all measures. In Fig 1, we compare the Receiver Operating Characteristic (ROC) of ENTPRISE-X, VEST-indel and DDIG-in. Fig 1 shows that ENTPRISE-X performs significantly better than VEST-indel and DDIG-in, especially in the low false positive regime that is the most important for practical application. For example, at a false positive rate of 2%, ENTPRISE-X has a true positive rate of 66%, whereas VEST-indel only has a 33% true positive rate.

Since we have used the feature representing whether the given gene is disease-associated or not as learned from literature (similar to the "PubMed" feature in VEST-indel [21]), it is of

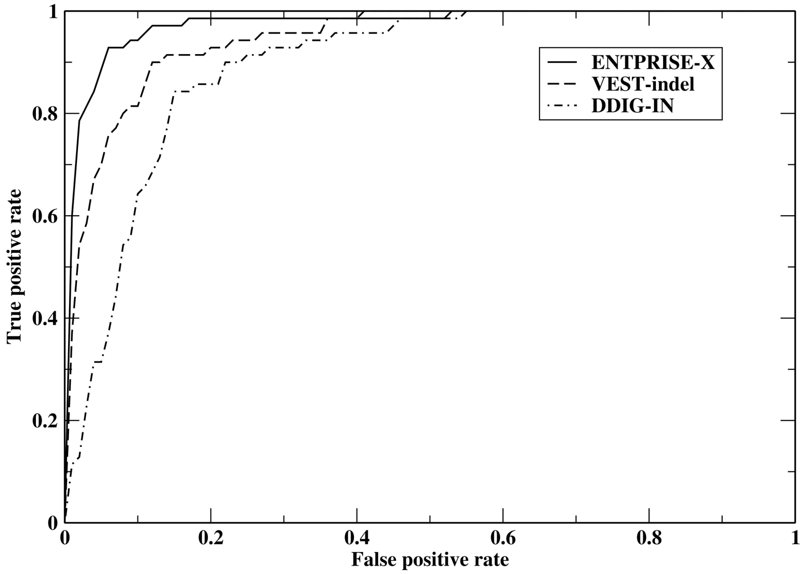

Fig 1. Receiver operating characteristic curves of ENTPRISE-X, VEST-indel and DDIG-in.

https://doi.org/10.1371/journal.pone.0196849.g001

interest to establish how well it performs when this is the only feature used for prediction. We call this scenario the *baseline* method, with the results also given in Table 2. Clearly, it has a too large false positive rate and the overall performance is much worse than ENTPRISE-X. It is also interesting to use the essentiality of protein as another baseline approach. It has much lower sensitivity and specificity than those of ENTPRISE-X. We tested another version of ENTPRISE-X that removes all the features encoding local information of the mutated position ((a)-(d)) used in ENTPRISE. This scenario is called ENTPRISE-X-nolocal and is shown in Table 2. It too has worse sensitivity and overall performance than full ENTPRISE-X. It is also important to examine the significance of the five new features by removing them from the 46 features and the resulting variant is called "ENTPRISE-X-nonew", its performance is much worse than that of ENTPRISE-X (see Table 2). This indicates that the newly introduced features provide the major contribution to performance of ENTPRISE-X. It is also interesting to see how different the individually trained models are from the default (average of two models) prediction. The results are shown in Table 2 with names ENTPRISE-X_1 and ENTPRISE-X_2. They differ slightly from each other and are slightly worse than the default prediction.

To see the importance of new features, we performed five separate tests by removing each of the five new features, one at a time:

ENTPRISE-X-noratio—removing the fraction of affected protein part;

ENTPRISE-X-noessential—removing the essentiality of protein;

ENTPRISE-X-nopathogen—removing the pathogenicity score predicted by ENTPRISE;

ENTPRISE-X-nodisease—removing the disease involvement of the protein;

ENTPRISE-X-nointeract—removing number of protein-protein interactions.

Table 2 shows that the order of the effect from strongest to weakest as ranked by their MCC & F-score are: (1) disease involvement of the protein; (2) fraction of the affected protein structure part; (3) pathogenicity of the affected part of the protein as predicted by ENTPRISE; (4) essentiality of mutated protein; and (5) number of protein-protein interactions. The effect of

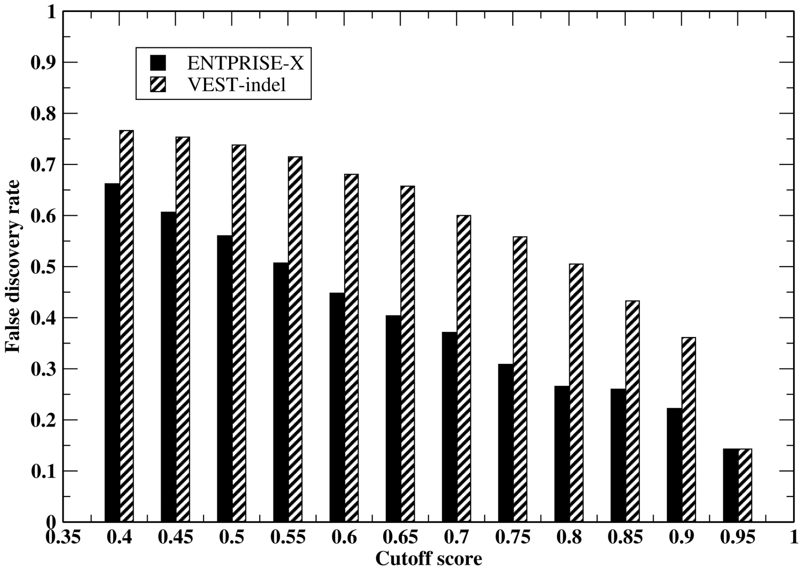

Fig 2. False discovery rate by the ENTPRISE-X and VEST-indel methods at various cutoffs on the VEST-indel test set.

https://doi.org/10.1371/journal.pone.0196849.g002

the number of protein-protein interactions has no visible effect on the MCC but has a small effect on the F-score and false positive rate.

In practice, due to the fact that neutral variations always dominate over true disease-associated variations, even a small false positive rate will result in large number of false positive predictions. We employ the "false discovery rate" (FDR) defined as (number of false positive predictions)/(total number of predictions). This rate depends not only on the false positive rate, but also on the relative abundance of true positives and true negatives in the evaluated data set. Fig 2 compares the false discovery rate by ENTPRISE-X and VEST-indel methods at various cutoffs on the VEST-indel test set. At the default cutoff score 0.5, both methods are dominated by false positive predictions with ENTPRISE-X having 56%, VEST-indel having 74% false predictions. Even at a cutoff of 0.9, ENTPRISE-X has 22% and VEST-indel has 36% false predictions for this particular data set. The high FDR for this set at the default cutoff value is mainly due to the overwhelmingly number of neutral variations (true positive/true neutral ~ 0.08). Even if all true positives are recalled, with 9% of neutral variations falsely classified as being disease-associated, the number of false predictions will be larger than that of correct positive predictions.

Ten-fold cross-validation

In order to benchmark ENTPRISE-X for nonsense as well as frameshift mutations, we performed a ten-fold cross-validation of ENTPRISE-X using the complete training dataset. Proteins in the whole training set are clustered into clusters with a 35% sequence identity cutoff. We then randomly partitioned these clusters into ten sets. To test frameshift and nonsense mutations in each set, we use the mutations in the remaining nine sets for training. The test is done in three scenarios: (a) evaluating the test set on frameshift and nonsense mutations together; (b) evaluating the test set only on frameshift mutations, and (c) evaluating the test set only on nonsense mutations. The results are compiled in Table 3 which shows that performance of ENTPRISE-X for frameshift and nonsense variations are very similar. Overall, the false positive rate of ENTPRISE-X in cross-validation is 18%, much more than that of the

Table 3. Ten-fold cross-validation of ENTPRISE-X on the whole training set[a].

Variation type	MCC	Sensitivity	Specificity	F-score	False positive rate	False discovery rate
			ENTPRISE-X			
Frameshift & nonsense	0.655	0.851	0.815	0.871	18.5%	10.8%
Frameshift	0.616	0.871	0.815	0.909	18.5%	4.9%
Nonsense	0.619	0.806	0.815	0.789	18.5%	22.8%
			DDIG-in			
Frameshift & nonsense	0.156	0.755	0.393	0.721	60.7%	31.0%
Frameshift	0.059	0.727	0.341	0.771	65.9%	17.8%
Nonsense	0.253	0.821	0.415	0.638	58.5%	47.9%
			SIFT-indel			
Frameshift	0.201	0.837	0.367	0.842	63.3%	15.3%

[a] To be fair to all methods, only the consensus mutations of the compared methods are evaluated.

https://doi.org/10.1371/journal.pone.0196849.t003

VEST-indel test set of 9.1%. This could be due to some contamination of true positives in the neutral set. The false discovery rates are relatively smaller than those on the VEST-indel test set due to the fact that the relative abundance of true positives is larger in the Ten-fold cross validation set than in the VEST-indel test set. We also compare our method for this set to the DDIG-in method (VEST-indel has used the same dataset for training, thus not compared) and SIFT-indel [22]. The SIFT-indel result was obtained from the server at http://sift.jcvi.org/www/SIFT_chr_coords_indels_submit.html. Clearly, due to its larger false positive rate, DDIG-in has a much larger false discovery rate. SIFT-indel is better than DDIG-in and worse than ENTPRISE-X in terms of its MCC, F-score and false positive rate.

Large scale test on the ExAC set

We next performed large scale testing of ENTPRISE-X, VEST-indel and DDIG-in on the ExAC set. The results are compiled in Table 4. Again, results by VEST-indel and DDIG-in are obtained from their respective webserver. Consistent with the test on the VEST-indel test set, VEST-indel classifies around 10% more disease causing mutations than ENTPRISE-X for frameshift mutations. ENTPRISE-X classifies around 16% of both frameshift and nonsense mutations as disease causing. The fractions of disease causing mutations classified by DDIG-in are around 65% for both frameshift and nonsense mutations and are far too high. These rates are well correlated with the false positive rates in Table 2. The 16% rate of ENTPRISE-X is very close to the false positive rate for ten-fold cross-validation; indicating that the ExAC set is similar to the neutral part of the training set in terms of the probability of being contaminated with

Table 4. Comparison of the percentage of disease causing variations in the ExAC set[a].

Method	Frameshift	Nonsense
	Evaluated #: 48123	Evaluated #: 40482
ENTPRISE-X	16.5%(0.73:6.0%)	15.7%(0.73:5.6%)
VEST-indel	26.2%(0.82:9.5%)	-
DDIG-in	64.4%(0.81:33.3%)	65.4%(0.81:35.1%)

[a] To be fair to all methods, only the consensus mutations of three methods are reported. Numbers in parenthesis are cutoff:false positive rate using the cutoff that maximizes the MCC in Table 2.

https://doi.org/10.1371/journal.pone.0196849.t004

some true positives because ENTPRISE-X has roughly the same ~16% false positive rate for both sets.

Statistical significance of ENTPRISE-X predictions

In order to estimate the false positive rate at an arbitrary prediction score, a smooth fitting function of the false positive rate distribution is needed. This is equivalent to finding the p-value (the probability of falsely classifying a mutation being disease-associated when it is a true neutral at a given cutoff score) given an arbitrary prediction score. To derive a p-value measuring the statistical significance of ENTPRISE-X predictions, we utilize a total of 1,025 neutral frameshift mutations from the VEST-indel test set. The distribution of ENTPRISE-X scores is shown in Fig 3. The fitting function rapidly approaches zero at a score ~ 1. Based on the shape of the distribution in Fig 3, we tried fitting the curve with Normal, Weibull, Extreme value, Log-logistic and Burr distributions. We found that the extreme value distribution is the best two-parameter fit with mean μ = 0.0706 and variation σ = 0.1331 using least root mean squared deviation criterion. The p-value of score x by ENTPRISE-X can be obtained with $p(x) = 1 - exp\left[-\exp\left(\frac{\mu - x}{\sigma}\right)\right]$. The default cutoff score 0.5 has a p-value of 0.039 that is smaller than the observed false positive rate 9% for the VEST-indel test set. This could be due to statistical fluctuations of the test data. Fig 4 shows the p-value against cutoffs. Similar fitting for VEST-indel score gives μ = 0.1371 and variation σ = 0.1583 that results in a p-value of 0.096 at the default cutoff 0.5.

Next, we apply the p-value that is equivalent to theoretical false positive rate at a given cutoff score to real exome data to estimate the false discovery rate of ENTPRISE-X prediction along with DDIG-in method. We obtained data from 13 individuals (private communication) and annotated their frameshift and nonsense variations using ENTPRISE-X. Table 5 summarizes the annotation results. The numbers of genes predicted to be disease-associated are around 30–70 when a default cutoff of 0.5 is used. Since the true neutral input variations are not known, here, we estimate the false predictions at given cutoff by pN_{total}, where p is the p-value

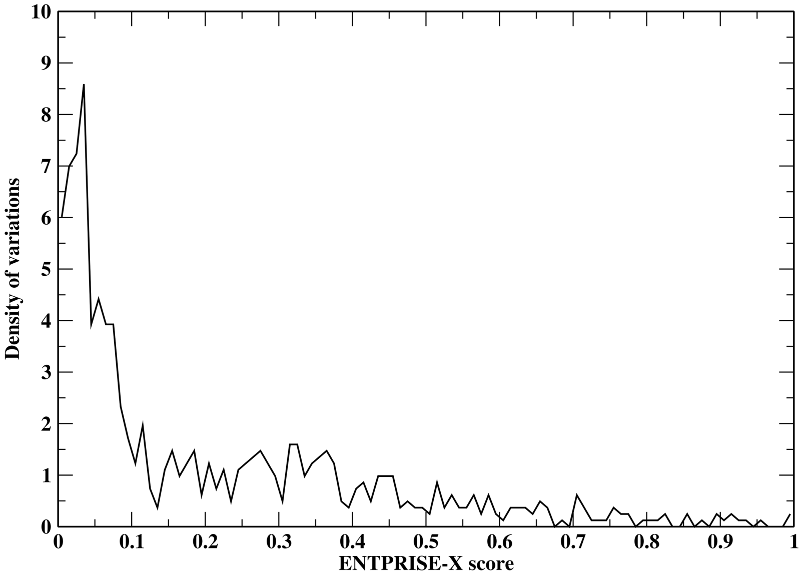

Fig 3. Distribution of ENTPRISE-X scores for the neutral variations in the VEST-indel test set. The area under the curve is normalized to one.

https://doi.org/10.1371/journal.pone.0196849.g003

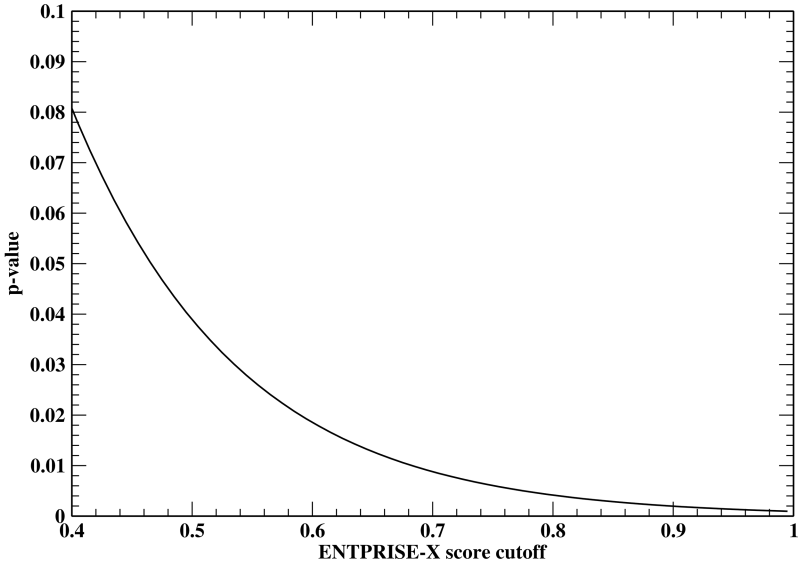

Fig 4. P-value derived from the VEST-indel test set by fitting to an extreme value distribution versus ENTPRISE-X score cutoffs.

https://doi.org/10.1371/journal.pone.0196849.g004

at the cutoff, N_{total} is the total number of genes evaluated as an approximation of the total number of neutral input genes. Then the false discovery rate can be estimated by pN_{total} / N_{pred}, where N_{pred} is number of predicted genes to be disease-associated. The average false discovery rate over the 13 individuals is around 25% with the default cutoff 0.5. The dependence of the average false discovery rate for 13 human exomes on the cutoff is shown in Fig 5. It has a minimum at a cutoff score of 0.92 with a 10% false discovery rate and then it increases with increased cutoff. This is due to the observation that N_{pred} drops faster than the p-value does. Obviously, the FDRs of DDIG-in is too high to be practically useful as its FDR ranges from 0.55 to 0.88.

Table 5. Summary of patient annotations using ENTPRISE-X[a].

patient	# of annotated variations	# of annotated genes	# of disease associated genes	false discovery rate
1	396/337	274/201	40/77	0.267/0.880
2	473/437	313/245	53/119	0.230/0.694
3	561/629	341/297	72/182	0.185/0.550
4	417/395	270/207	40/94	0.263/0.742
5	397/393	262/190	38/86	0.269/0.745
6	410/377	271/226	34/105	0.311/0.725
7	454/340	284/198	34/79	0.326/0.845
8	380/352	261/205	39/97	0.261/0.712
9	431/ 414	278/227	48/96	0.226/0.797
10	402/372	255/206	47/94	0.212/0.739
11	494/451	319/256	51/123	0.244/0.701
12	572/588	353/299	69/184	0.200/0.548
13	547/580	342/301	67/166	0.199/0.611

[a] In each cell, first number is from ENTPRISE-X, second number is from DDIG-in method.

https://doi.org/10.1371/journal.pone.0196849.t005

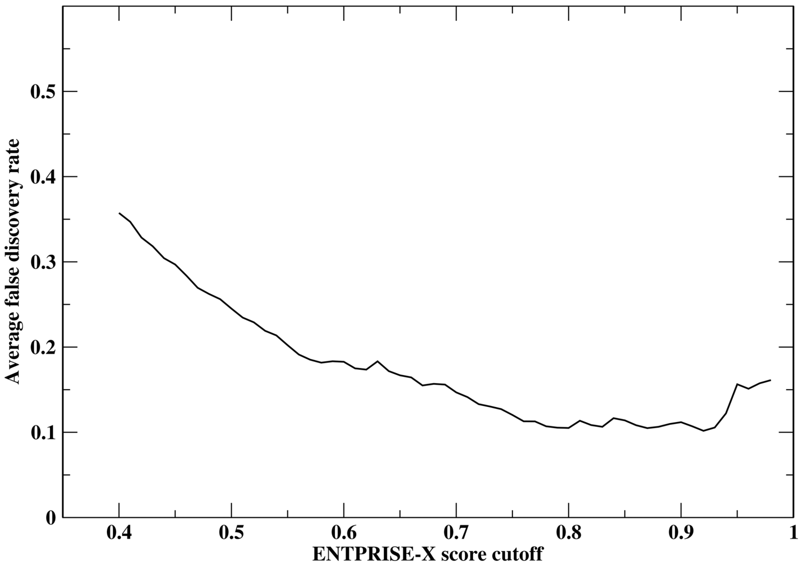

Fig 5. Dependence of average false discovery rate of ENTPRISE-X for 13 human exomes on score cutoffs.

https://doi.org/10.1371/journal.pone.0196849.g005

Discussion

In this work, we have extended our machine learning approach ENTPRISE for predicting disease association of missense mutations to address frameshift and nonsense mutations in ENTPRISE-X. We show that ENTPRISE-X's performance is superior to the state-of-the-art VEST-indel and DDIG-in methods. As with ENTPRISE, ENTPRISE-X has a much lower false positive rate than its peers while maintaining comparable sensitivity. In practice, for an entire exome, a large false positive rate will result in a method not useful because of the large number of mutations incorrectly identified as being disease-associated. The better performance of ENTPRISE-X demonstrates that selection of proper features and their integration yields better results. For example, the protein structure based local features help ENTPRISE-X to increase its sensitivity while maintaining specificity. As was shown for ENTPRISE [18], use of boosted tree regression gives superior performance compared to training using SVM based training. A natural extension of the current work is to place disease-causing genes in the content of a functional network to establish whether they would be useful disease-associated drug targets arising from missense mutations or if the protein is missing due to frameshift/nonsense mutations, what is the best downstream protein to target to treat the disease.

Supporting information

S1 File. All training and testing data sets.
(ZIP)

Acknowledgments

The authors thank Dr. Bartosz Ilkowski for managing the cluster on which this work was conducted.

Author Contributions

Conceptualization: Hongyi Zhou.

Data curation: Hongyi Zhou, Mu Gao.

Investigation: Hongyi Zhou.

Supervision: Jeffrey Skolnick.

Validation: Hongyi Zhou.

Writing – original draft: Hongyi Zhou.

Writing – review & editing: Mu Gao, Jeffrey Skolnick.

References

1. Reich D, Lander E. On the allelic spectrum of human disease. Trends Genet. 2001; 17(9):502–10. PMID: 11525833
2. Manolio T, Guttmacher AE, Manolio TA. Genomewide association studies and assessment of the risk of disease. N Engl J Med. 2010; 363(2):166–76. https://doi.org/10.1056/NEJMra0905980 PMID: 20647212
3. McClellan J, King M-C. Genetic Heterogeneity in Human Disease. Cell. 2010; 141(2):210–7. https://doi.org/10.1016/j.cell.2010.03.032 PMID: 20403315
4. Kumar P, Henikoff S, Ng P. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 2009; 4(7):1073–81. https://doi.org/10.1038/nprot.2009.86 PMID: 19561590
5. Adzhubei I, Schmidt S, Peshkin L, Ramensky V, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010; 7(4):248–9. https://doi.org/10.1038/nmeth0410-248 PMID: 20354512
6. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res. 2011; 39(17):e118. https://doi.org/10.1093/nar/gkr407 PMID: 21727090
7. Shihab H, Gough J, Cooper D, Stenson P, Barker G, Edwards K, et al. Predicting the Functional, Molecular and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models. Hum Mutat. 2013; 34:57–65. https://doi.org/10.1002/humu.22225 PMID: 23033316
8. Capriotti E, Calabrese R, Casadio R. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. Bioinformatics. 2006; 22:2729–34. https://doi.org/10.1093/bioinformatics/btl423 PMID: 16895930
9. Bromberg Y, Yachdav G, Rost B. SNAP predicts effect of mutations on protein function. Bioinformatics. 2008; 24:2397–98. https://doi.org/10.1093/bioinformatics/btn435 PMID: 18757876
10. Li B, Krishnan V, Mort M, Xin F, Kamati K, Cooper D, et al. Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics. 2009; 25:2744–50. https://doi.org/10.1093/bioinformatics/btp528 PMID: 19734154
11. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, et al. PANTHER: A Library of Protein Families and Subfamilies Indexed by Function. Genome Res. 2003; 13:2129–41. https://doi.org/10.1101/gr.772403 PMID: 12952881
12. Calabrese R, Capriotti E, Fariselli P, Martelli P, Casadio R. Functional annotations improve the predictive score of human disease-related mutations in proteins. Hum Mutat. 2009; 30:1237–44. https://doi.org/10.1002/humu.21047 PMID: 19514061
13. Gnad F, Baucom A, Mukhyala K, Manning G, Zhang Z. Assessment of computational methods for predicting the effects of missense mutations in human cancers. BMC Genomics. 2013; 14(Suppl 3):S7.
14. Hou JP, Ma J. DawnRank: discovering personalized driver genes in cancer. Genome Medicine. 2014; 6:56–71. https://doi.org/10.1186/s13073-014-0056-8 PMID: 25177370
15. Carter H, Chen S, Isik L, Tyekucheva S, Velculescu V, Kinzler K, et al. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. Cancer Res. 2009; 69(16):6660–7. https://doi.org/10.1158/0008-5472.CAN-09-1133 PMID: 19654296
16. Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC, et al. MuSiC: Identifying mutational significance in cancer genomes. Genome Res. 2012; 22(8):1589–98. https://doi.org/10.1101/gr.134635.111 PMID: 22759861
17. Kircher M, Witten D, Jain P, O'Roak B, Cooper G, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014; 46:310–5. https://doi.org/10.1038/ng.2892 PMID: 24487276

18. Zhou H, Gao M, Skolnick J. ENTPRISE: An algorithm for predicting human diease-associated amino acid substitutions from sequence entropy and predicted protein structures. PLOS one. 2016; 11(3): e0150965. https://doi.org/10.1371/journal.pone.0150965 PMID: 26982818

19. Bamford S, Dawson E, Forbes S, Clements J, Pettett R, Dogan A, et al. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. British Journal of Cancer. 2004; 91:355–8. https://doi.org/10.1038/sj.bjc.6601894 PMID: 15188009

20. Folkman L, Yang Y, Li Z, Stantic B, Sattar A, Mort M, et al. DDIG-in: detecting disease-causing genetic variations due to frameshifting indels and nonsense mutations employing sequence and structural properties at nucleotide and protein levels. Bioinformatics. 2015; 31(10):1599–606. https://doi.org/10.1093/bioinformatics/btu862 PMID: 25573915

21. Douville C, Masica D, Stenson P, Cooper D, Gygax D, Kim R, et al. Assessing the Pathogenicity of Insertion and Deletion Variants with the Variant Effect Scoring Tool (VEST-Indel). Hum Mutat. 2016; 37 (1):28–35. https://doi.org/10.1002/humu.22911 PMID: 26442818

22. Hu J, Ng PC. SIFT Indel: Predictions for the Functional Effects of Amino Acid Insertions/Deletions in Proteins. PLOS ONE. 2013; 8(10):e77940. https://doi.org/10.1371/journal.pone.0077940 PMID: 24194902

23. Roe BP, Yang H-J, Zhu J. Boosted Decision Trees, A Powerful Event Classifier. Statistical Problems in Particle Physics, Astrophysics and Cosmology 12006. p. 139.

24. Cortes C, Vapnik V. Support-vector networks. Machine Learning. 1995; 20(3).

25. Ogutu J, Piepho H, Schiulz-Streeck T. A comparison of random forests, boosting and support vector machines for genomic selection. BMC Proceedings. 2011; 5(Suppl 3):S11.

26. Tennessen J, Bigham A, O'Connor T, Fu W, Kenny E, Gravel S, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science. 2012; 337:64–9. https://doi.org/10.1126/science.1219240 PMID: 22604720

27. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated map of structural variation in 2,504 human genomes. Nature. 2015; 526:75–81. https://doi.org/10.1038/nature15394 PMID: 26432246

28. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016; 536:285–91. https://doi.org/10.1038/nature19057 PMID: 27535533

29. Gao M, Zhou H, Skolnick J. Insights into disease-associated mutations in the human proteome through protein structural analysis. Structure. 2015; 23(7):1362–69. https://doi.org/10.1016/j.str.2015.03.028 PMID: 26027735

30. Zhou H, Skolnick J. Template-based protein structure modeling using TASSER^{VMT}. Proteins. 2011; 80(2):352–61. https://doi.org/10.1002/prot.23183 PMID: 22105797

31. Hildebrand A, Remmert M, Biegert A, Söding J. Fast and accurate automatic structure prediction with HHpred. Proteins. 2009; 77(Suppl 9):128–32.

32. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol. 1995; 247:536–40. PMID: 7723011

33. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl Acid Res. 1997; 25 3389–402.

34. Zhang R, Lin Y. DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes. Nucleic Acids Research. 2009; 37:D455–D8. https://doi.org/10.1093/nar/gkn858 PMID: 18974178

35. Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D. GeneCards: integrating information about genes, proteins and diseases. Trends in Genetics. 1997; 13:163. PMID: 9097728

36. Zhang Y, De S, Garner JR, Smith K, Wang SA, Becker KG. Systematic analysis, comparison, and integration of disease based human genetic association data and mouse genetic phenotypic information. BMC Medical Genomics. 2010; 3:1. https://doi.org/10.1186/1755-8794-3-1 PMID: 20092628

37. Menche J, Sharma A, Kitsak M, Ghiassian SD, Vidal M, Loscalzo J, et al. Uncovering disease-disease relationships through the incomplete interactome. Science. 2015; 347:841.

38. Zhou H, Skolnick J. A knowledge-based approach for predicting gene-disease associations. Bioinformatics. 2016; 32(18):2831–8. https://doi.org/10.1093/bioinformatics/btw358 PMID: 27283949

39. Schaefer MH, Fontaine J-F, Vinayagam A, Porras P, Wanker EE, Andrade-Navarro MA. HIPPIE: Integrating Protein Interaction Networks with Experiment Based Quality Scores. Plos One. 2012; 7(2): e31826. https://doi.org/10.1371/journal.pone.0031826 PMID: 22348130

40. Friedman JH, Meulman JJ. Multiple additive regression trees with application in epidemiology. Statistics in Medicine. 2003; 22:1365–83. https://doi.org/10.1002/sim.1501 PMID: 12704603

41. Friedman JH. Greedy function approximation: a gradient boosting machine. The Annals of Statistics. 2001; 29:1189–232.