Amino Acids (2014) 46:2025–2035
DOI 10.1007/s00726-014-1760-9

ORIGINAL ARTICLE

Combination use of protein–protein interaction network topological features improves the predictive scores of deleterious non-synonymous single-nucleotide polymorphisms

Yiming Wu · Runyu Jing · Lin Jiang · Yanping Jiang · Qifan Kuang · Ling Ye · Lijun Yang · Yizhou Li · Menglong Li

Received: 18 September 2013 / Accepted: 3 May 2014 / Published online: 22 May 2014

Abstract Single-nucleotide polymorphisms (SNPs) are the most frequent form of genetic variations. Non-synonymous SNPs (nsSNPs) occurring in coding region result in single amino acid substitutions that associate with human hereditary diseases. Plenty of approaches were designed for distinguishing deleterious from neutral nsSNPs based on sequence level information. Novel in this work, combinations of protein–protein interaction (PPI) network topological features were introduced in predicting disease-related nsSNPs. Based on a dataset that was compiled from Swiss-Prot, a random forest model was constructed with an average accuracy value of 80.43 % and an MCC value of 0.60 in a rigorous tenfold crossvalidation test. For an independent dataset, our model achieved an accuracy of 88.05 % and an MCC of 0.67. Compared with previous studies, our approach presented superior prediction ability. Results showed that the incorporated PPI network topological features outperform conventional features. Our further analysis indicated that disease-related proteins are topologically different from other proteins. This study suggested that nsSNPs may share some topological information of proteins and the change of topological attributes could provide clues in illustrating functional shift due to nsSNPs.

Keywords Non-synonymous single-nucleotide polymorphisms · Disease · Protein–protein interaction network · Topological features · Prediction

Introduction

Single-nucleotide polymorphisms (SNPs) are very common genetic variations in a single base of DNA. They are tightly associated with human evolution. Among them, non-synonymous single-nucleotide polymorphisms (nsSNPs) can cause substitutions of amino acids in proteins. Such substitutions have potential to affect protein structures and functions, which would be more likely related to human inherited diseases (Stenson et al. 2003; Gibbs et al. 2003; Ramensky et al. 2002). In recent years, new genomic techniques have been developed for large-scale identification of human SNPs. The dbSNP database (http://www.ncbi.nlm.nih.gov/projects/SNP/) had an explosive growth as well as other databases such as the Human Gene Mutation Database (HGMD, http://www.hgmd.cf.ac.uk/ac/index.php) and the Human Genome Variation database of Genotype–Phenotype (HGVbase2GP, http://www.hgvbaseg2p.org/-index) (Calabrese et al. 2009).

Rapid accumulation of SNPs data presented the challenge of sorting out disease-related nsSNPs from functional neutral ones. However, confirming disease-associated nsSNPs one by one through wet laboratory experiments is labor-consuming and time-costing. It would, therefore, be desirable to develop an efficient way to detect them. Machine learning methods were widely used to achieve the goal in the last decade, such as k-nearest neighbors (Huang et al. 2010a), artificial neural networks (Ferrer-Costa et al. 2004), decision trees (Hu and Yan 2008; Dobson et al.

Electronic supplementary material The online version of this article (doi:10.1007/s00726-014-1760-9) contains supplementary material, which is available to authorized users.

Y. Wu · R. Jing · L. Jiang · Y. Jiang · Q. Kuang · L. Ye · L. Yang · Y. Li (✉) · M. Li (✉)
College of Chemistry, Sichuan University, Chengdu 610064, People's Republic of China
e-mail: liyizhou_415@163.com

M. Li
e-mail: liml@scu.edu.cn

2006), random forest (Li et al. 2011; Bao and Cui 2005), and support vector machines (SVMs) (Ye et al. 2007; Calabrese et al. 2009; Capriotti et al. 2006, 2007; Tian et al. 2007). The performance proved these computational approaches efficient and reliable. Several works explored new ways to classify nsSNPs and the corresponding web server was established. An early predictor was sorting tolerant from intolerant (SIFT) (Ng and Henikoff 2003), which was developed by Ng and Henikoff based on sequence homology. Another popular one named PolyPhen (Ramensky et al. 2002), which was upgraded to PolyPhen2, demonstrated that the selection pressure against deleterious SNPs depends on the molecular function of the proteins (Li et al. 2011). PANTHER PSEC (Thomas et al. 2003) based on Hidden Markov Model families can give a likelihood score for a mutation to measure its impact on protein function. These scores could also be treated as features for model input (Wang et al. 2011; Bao and Cui 2005).

At present, most approaches exploited information from sequence profile and three-dimensional structures of proteins. However, limited structural data restricted their practice. The evolutionary information encoded in the sequence profile was more widely used and then sequence-derived information had been becoming basic features in these prediction tasks over the years. Meanwhile, for sake of improving prediction performance, great efforts were made to develop powerful features from multiple pipelines. Recently, few researchers turned to concentrate on investigating the origin of a mutation site and its impact on gene products (Capriotti et al. 2007). Huang and Capriotti demonstrated that protein level information is helpful in detecting deleterious nsSNPs (Calabrese et al. 2009; Huang et al. 2010a). They adopted KEGG scores to reflect the importance of proteins in a pathway and gathered GO information to illustrate biological functions of proteins, respectively.

In this study, we adopted another kind of protein level information, protein–protein interaction (PPI) network topological features. They have been widely used in bioinformatics for predicting disease-related genes and drug targets. Gandhi et al. (2006) analyzed human interaction map and found that genes ascertained from the OMIM database that was associated with a human disease preferentially interacted with other disease-causing genes to those without any known disease association. Xu and Li (2006) demonstrated that heritable disease genes shared some topological features in the PPIs network, whereas the non-disease genes did not. By supposing that disease-associated genes may receive more evolutionary pressure which could be reflected by its corresponding nsSNPs, we made a combination use of topological features and sequence profile-derived features to predict disease-related nsSNPs. Several conventionally used features and the microenvironment were additionally taken into account. Finally, we compiled a dataset with 57 features. Through a strict protein level tenfold crossvalidation and independent test, satisfactory results were achieved by our method. By comparison on the same dataset, our method outperformed other implementations. Moreover, as was found in our further analysis, disease-related mutations tended to occur in proteins with high degree scores, while the polymorphism mutations in proteins whose neighbors tended to contact with each other and thus were more likely to cluster and form a subnetwork.

Materials and methods

Dataset

Care et al. (2007) carefully examined several nsSNPs databases and suggested choosing Swiss-Prot data for this type of prediction. Following the advice, we compiled our nsSNPs data from Swiss-Prot protein knowledgebase. Two datasets (release 57.4 and release 57.13) used by Huang et al. (2010a) were chosen in our study. Release 57.4 (16 June 2009) was used for tenfold crossvalidation. NsSNPs data added in release 57.13 (19 January 2010) after release 57.4 were treated as the independent set. The final dataset was retrieved from the above two releases data by obeying the following rules: (1) the mutation site must be related to diseases or polymorphisms alternatively (unclassified nsSNPs were excluded), (2) the mutation corresponding proteins can map to the STRING database, (3) the position number of a mutation site should not beyond the sequential length of the corresponding protein. Mutation sites not matching the above conditions were excluded.

Finally, we attained a training dataset consisting of 20,252 disease-related nsSNPs (similar to SAPs or single amino acid substitutions) and 26,744 polymorphism nsSNPs. These disease-related and neutral nsSNPs were contained with 2,268 and 9,013 proteins, respectively. The independent set contained 872 disease-associated nsSNPs and 2,619 neutral polymorphisms, respectively.

Feature construction

Our task was to predict whether a given single amino acid substitution is a causation of human genetic diseases. We compiled 57 dimensional features (Fig. 1) to represent each of the mutation sites. Some of the features were typically used in previous investigations, especially the strong discriminators such as conservation scores, PSSM, hydrophobicity, disorder regions and so on. We were particularly interested in the performance of PPI topological features that were widely used in detecting disease-related genes and drug targets. The attributes were described as below.

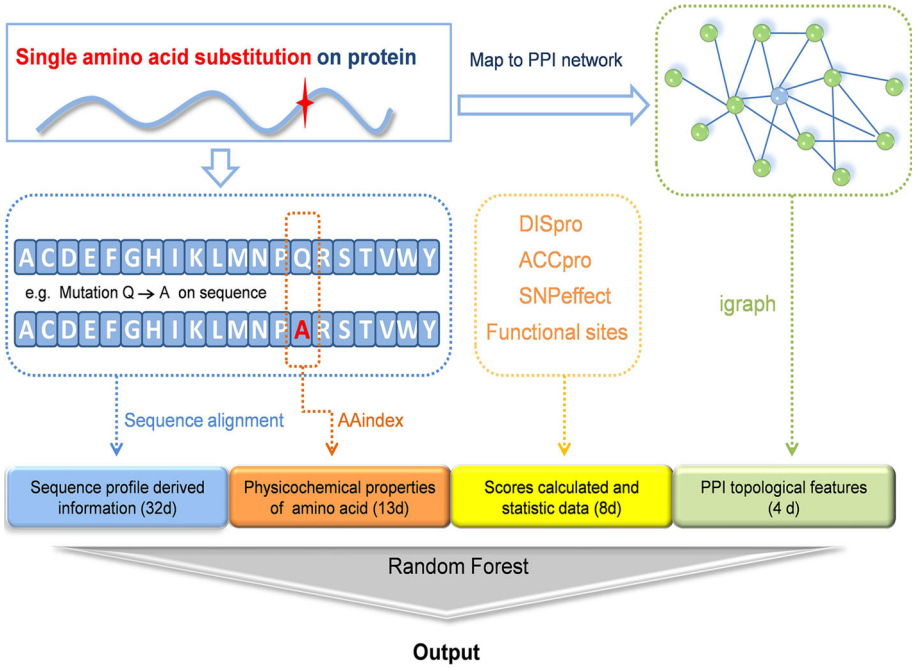

Fig. 1 Flow chart of our random forest-based method. All features were divided into four categories according to their sources. The sequence profile-derived information includes smooth window-derived 20 PSSM scores, five conservation scores, three mutation likelihood scores, average score of sliding smooth window, mutation frequency, PAM250 and HLA family. The physicochemical properties include four indexes for wild-type, variant and their difference ($4 \times 3 = 12$), Grantham. The attributes derived from tools and statistic data include disorder region, solvent accessibility, four SNPeffect scores, and two function site scores. The PPI topological features include four topological parameters

The PPI network topological parameters

STRING is a database that collects known and predicted protein interactions. The database currently covers 5,214,234 proteins from 1,133 organisms. The interactions include direct (physical) and indirect (functional) associations. All human PPI contained in STRING (version 9.0) (Szklarczyk et al. 2011) were compiled to construct the network. The R package igraph (Csardi and Nepusz 2006) was employed to calculate four well-established PPI network topological parameters: degree, transitivity, closeness and betweenness.

The degree is the most basic structural property of a vertex in a network, which means the number of a vertex's adjacent edges. A higher degree indicates that the node has more direct interactions with other proteins. It can be defined as:

$$\delta(i) = \sum_{j \in N} a_{i,j} \tag{1}$$

In which $a_{i,j}$ is a direct interaction between vertex i and vertex j, and N is the set of all vertices.

The transitivity measures the probability that the adjacent vertices of a vertex are connected. Transitivity is also called the clustering coefficient. It can be defined as

$$\mathrm{TR}(i) = 2e_i/\delta_i(\delta_i - 1), \tag{2}$$

where $\delta_i(\delta_i - 1)/2$ is the maximum possible number of edges between the neighbors of vertex i. The e_i is the number of existing edges.

Closeness centrality measures how many steps are required to access every other vertex from a given vertex. It can be calculated as

$$C(i) = N - 1/\sum_{i \neq j} d_{i,j}, \tag{3}$$

where $d_{i,j}$ is the shortest path between vertices i and j, N is the total number of vertices.

The betweenness is defined by the number of geodesics going through a vertex. It can be calculated as

$$B(i) = \sum_{j,k \in N, j \neq k} \frac{n_{j,k}(i)}{n_{j,k}}, \tag{4}$$

where $n_{j,k}(i)$ is the number of shortest paths connecting j and k passing through vertex i.

All the above topological parameters measure the importance of proteins in the PPI network. If an amino substitution, caused by nsSNP, happened on those important proteins and affects their functions, it could disturb the protein-related biological function system and was

considered as a risk. In present work, we used topological parameters of a protein to represent corresponding nsSNPs.

The sequence-derived features

In sequence-derived information, evolutionary conservation score was considered as an important feature in this type of prediction (Ye et al. 2007; Li et al. 2011; Yang et al. 2012). A conserved amino acid in a particular position of the corresponding protein means that it hardly changes in a specific evolutionary time. The amino acid residue may closely relate with functional domains. Once it changes, it may alter the protein's function and structure significantly. A Position-Specific Iterative BLAST (PSI BLAST) (Altschul et al. 1997) was implemented against the Swiss-Prot sequence dataset with an E value cutoff of 1E−3 and two iterations. The output Position-Specific Scoring Matrix (PSSM) contains frequencies of 20 types of amino acids at each sequence position. The frequency weighted occurring probabilities of 20 types of residues at a specific sequence position in the evolutionary history. Then, the conservation score was calculated by following formula:

$$\mathrm{CS}_i = -\sum_{j=1}^{20} p_{i,j} \log_2 p_{i,j}, \tag{5}$$

where $p_{i,j}$ is the frequency of amino acid j at position i. A lower value indicates more conserved at a position and vice versa. Besides, the mutation frequency difference between a wild-type and variant was treated as a feature for having been proved as a powerful attribute (Ng and Henikoff 2001).

Microenvironment of mutation site

It is a common view that the microenvironment of the mutation site plays an important role in prediction (Ye et al. 2007; Calabrese et al. 2009; Huang et al. 2010a). Taking evolutionary microenvironment into account, we adopted sliding and smooth window method to get information from PSSM. The Nearest Neighbor Algorithm was adopted to search the best combination of sliding window size and smooth window size. The sliding window size ranged from 3 to 11 and the smooth window size ranged from 1 to 11. We employed a 'grid-search' strategy to select the pair with the best cross-validation accuracy. Finally, we attained sliding window size 11 and smooth window size 9 with accuracy of 65.5 %. Hence, we got an $11 \times 20 = 220$ dimensional vector to represent the evolutionary microenvironment. To reduce data dimensions, the mean value of 220 dimensional data was calculated to represent the sliding-smooth window. It is calculated by

$$\mathrm{Averslid} = \sum_{i=1}^{220} p_i / 220, \tag{6}$$

where p_i is the score in position i of the sliding window. The Averslid is one-dimensional.

Besides, the smooth window-derived PSSM scores were kept and the conservation microenvironment was the amino acid composition window of five residues centred on the mutation site.

Physicochemical characteristic

AAindex is a database of numerical indices representing various physicochemical and biochemical properties of amino acids and pairs of amino acids (ftp://ftp.genome.jp/pub/db/community/aaindex/) (Kawashima et al. 2008). We compiled five physicochemical properties from AAindex1 and AAindex2: B values, hydrophobicity, polarity, side chain volume and Grantham scores. They were widely used in previous studies of predicting deleterious missense mutation (Qin et al. 2012). Except Grantham scores, the rest indices scores were applied for representing wild-type and variant as well as their difference. Hence this section contains 13 attributes.

Near functional sites

Dobson et al. (2006) found that some terms (most are functional sites) listed in the Swiss-Prot features table are most discriminatory. The nsSNPs related to the feature table terms are predominantly associated with diseases. However, these terms of features table are not efficiently used as strong discriminators for very limited nsSNPs related to them. To overcome this limitation, Ye et al. (2007) defined some new attributes by calculating the distance between a SAP position and its closest functional site. It was realized that an amino acid substitution occurs near or exactly on a protein function site which may affect the protein function. In the wake of Ye's idea, we defined two attributes to represent a SAP near function sites: one was that whether a SAP near any function sites within five sequential distances, the other was that whether a SAP near any function sites within 5 Å spatial distances. Five function sites were chosen from Swiss-Prot feature table: ACT_SITE, BINDING, METAL, MOD_RES and CARBOHYD. SAPs near these function sites were considered having the potential to change protein function.

HLA family

Human leukocyte antigen (HLA) family is a group of proteins known as the HLA complex, which is the human

version of the major histocompatibility complex. The HLA complex helps the immune system to distinguish the body's own proteins from proteins made by foreign invaders such as viruses and bacteria. Thus, nsSNPs associated with HLA proteins naturally have some correlations with diseases. Some methods (Ye et al. 2007; Yang et al. 2012; Li et al. 2011) took that whether an nsSNP-related protein belongs to the HLA family as an attribute. We followed the procedure that ran BLAST against IMGT/HLA database (Robinson 2003; Robinson et al. 2009). The protein hits a sequence with e value <0.01 and identity over 70 % was treated as the HLA complex.

Features calculated by tools

There were various bioinformatics tools designed for reflecting properties of proteins. We selected several sequence-based tools to calculate scores which could help predicting deleterious nsSNPs.

Intrinsically disordered proteins do not form a fixed three-dimensional structure in their putatively native states, neither in their entirety nor in part (Sickmeier et al. 2007). Although lacking a fixed structure, the disorder region closely related to vital biological functions, typically involved in regulation, signaling and control. Amino acid substitutions in these regions putatively change their normal functions and, therefore, tend to be disease related. In this study, we used DISpro (Cheng et al. 2005b) to detect disorder regions and judged whether SAPs locate in the regions. DISpro can accurately predict disorder regions in a protein based on the sequence. The stand-alone version could be downloaded from SCRATCH (Cheng et al. 2005a), which supplies several tools for predicting tertiary structure and structural features.

The solvent accessibility of an amino acid residue in a protein measures the degree of the residue's exposure to the surrounding solvent in the protein structure. Solvent accessibility is a customary feature used in amino acid substitution prediction and has been confirmed as a strong discriminator (Ng and Henikoff 2006; Saunders and Baker 2002). It was summarized that disease nsSNPs tend to be buried and neutral nsSNPs tend to be exposed (Dobson et al. 2006). Here, we use ACCpro to detect whether nsSNPs are buried or exposed. ACCpro is a sequence-based tool for detecting whether an amino acid residue is buried or exposed at each sequence position. The same as DISpro, the stand-alone version was downloaded from SCRATCH.

SNPeffect (Reumers et al. 2005) is a database for phenotyping human SNPs. The database contains all known human protein variants from UniProt. Besides, the sever SNPeffect 4.0 (De Baets et al. 2012) integrates aggregation prediction (TANGO), amyloid prediction (WALTZ), chaperone-binding prediction (LIMBO) and protein stability analysis (FoldX) for structural phenotypic. Aggregation property was thought as a key factor of disease susceptibility (Reumers et al. 2009; Belli et al. 2011). Amyloid-like structures result from protein aggregation. Once the structure changed, the protein may be processed with an entirely different biological mechanism (Maurer-Stroh et al. 2010). The chaperone-binding sites on the protein are meaningful for interaction between proteins and molecular chaperones (Van Durme et al. 2009). All four attributes could be disturbed by single amino acid substitutions, thereby affecting the protein function, so we took them into consideration for training the model.

Random forest

The random forest classifies mutations into disease-related (desired output set to yes) and polymorphism ones (desired output set to no) in our study. Particularly, the random forest package in R was applied for training and prediction. The random forest is an outstanding ensemble classifier based on decision trees which was enormously used in classification and regression works (Breiman 2001). It is widely used owing to the following merits: (1) enhancing prediction accuracy, (2) robust against overfitting, (3) good tolerance of noisy data. In addition, it is so user-friendly that it only has two main parameters: the number of variables in the random subset at each node (mtry) and the number of trees in the forest (ntree). In our study, we used grid-search method to get the best combination of ntree and mtry. The combination was determined by tenfold cross-validation overall accuracy. However, the result was not so sensitive to the parameters as Liaw found (Liaw and Wiener 2002). We, therefore, set ntree as 100 and mtry as default value for a rapid calculation. Random forest packages also supplied a module to evaluate variable importance. Two calculated indexes, mean decrease accuracy (MDA) and mean decrease Gini (MDGI), were employed to measure the importance of features in our study. The mean decrease in accuracy a variable causes is determined during the out of bag error calculation phase. The more the accuracy of the random forest decreases due to the addition of a single variable, the more important the variable is deemed. The mean decrease in Gini coefficient is a measure of how each variable contributes to the homogeneity of the nodes and leaves in the resulting random forest. Each time a particular variable is used to split a node, the Gini coefficient for the child nodes is calculated and compared to that of the original node. The changes in Gini are summed for each variable and normalized at the end of the calculation. Variables that result in nodes with higher purity have a higher decrease in Gini coefficient (Breiman 2001; Nicodemus 2011). After training and

prediction, accuracy, sensitivity, specificity and Matthew's correlation coefficient (MCC) were adopted for evaluating the performance. They are defined as

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{FN} + \text{TN} + \text{FP}) \quad (7)$$

$$\text{Sensitivity} = \text{TP}/(\text{TP} + \text{FN}) \quad (8)$$

$$\text{Specificity} = \text{TN}/(\text{TN} + \text{FP}) \quad (9)$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FN})(\text{TN} + \text{FP})}}, \quad (10)$$

where TP is the number of true positive (positive sample correctly predicted as positive sample), TN is the number of true negative (negative sample correctly predicted as negative sample), FP is the number of false positive (negative sample incorrectly predicted as positive sample), and FN is the number false negative (positive sample incorrectly predicted as negative sample). Thus, the sensitivity represents the ratio of correctly predicted positive samples to all positive samples, the specificity represents the ratio of correctly predicted negative samples to all negative samples and the accuracy is the overall percentage of nsSNPs correctly predicted. MCC is an important parameter for evaluating the performance when the dataset is unbalanced.

Results and discussion

Test on independent set

Our independent set was compiled by executing the procedures as described in "Materials and methods", which included 872 diseased-related nsSNPs and 2,619 polymorphisms derived from 116 and 1,054 proteins, respectively. We also adopted a strict prediction condition for the independent set. The proteins related to independent set were excluded from the training data. Satisfactory results were obtained from independent set with an accuracy of 88.05 % and an MCC of 0.67.

Comparisons with previous studies

Here, Huang's method (Huang et al. 2010a) was taken for our comparison for using the same datasets and both methods used the protein level information. In Huang's work, 220 dimensional protein level features were employed named KEGG scores, which were defined to measure function of neighbors in KEGG pathways and proved to be the most contributive features in their method. Integrating with conventionally used features, a dataset was constructed with 472 attributes. After a Jackknife crossvalidation through Nearest Neighbor Algorithm, they attained a prediction accuracy of 83.27 % and prediction accuracy of 80.0 % on independent set.

To perform a fair comparison, the same predicting procedure (Nearest Neighbor Algorithm and Jackknife crossvalidation) was adopted on the same data set (release 57.4). As a result, our model attained an accuracy of 83.8 % and an MCC of 0.67. For the independent set, an accuracy of 83.5 % and an MCC of 0.55 were attained.

Since the protein level features indistinguishably assigned the same score to nsSNPs (both disease-related ones and neutral ones) on a same protein, it would make the prediction sensitive to data division, i.e. these nsSNPs got the same feature values, while some of them distributed in test set and the rest in training set, affected by unbalanced data of diseased ones and neutral ones in training set, the test data were easily classified as the majority side. Accordingly, most nsSNPs were correctly predicted after a crossvalidation procedure and this might lead to a bias result. To avoid such situation, a reasonable approach would be strictly classifying the dataset into test set and training set at the protein level.

We, therefore, presented our predictor tested and trained with a stringent crossvalidation procedure, i.e. nsSNPs in the same proteins were kept in the same fold. However, the predictive power of Nearest Neighbor Algorithm reduced obviously. The accuracy and MCC value decreased to 75.93 % and 0.51, respectively. Compared with Nearest Neighbor Algorithm, it was observed that SVM and Random Forest could get better performance. We turned to use Random Forest as the classifier for it is not susceptible to parameters (Krishnan and Westhead 2003; Liaw and Wiener 2002). After a protein level tenfold crossvalidation, our model attained an accuracy of 80.43 % and the MCC was 0.60. Our independent set was compiled by following Huang's procedure. For a fair comparison, by randomly eliminating definite number of samples, the number of independent set was reduced to reach the same as Huang's. The final independent set was also predicted in a stringent condition. We repeated this procedure 10 times with the worst accuracy of 85.5 % and an MCC of 0.65, which was obviously higher than Huang's accuracy of 80 %.

We also compared with two famous predictors in predicting deleterious amino acid substitutions, the SIFT and PolyPhen-2. Sorting tolerant from intolerant is an early predictor which uses sequence homology to predict whether an amino acid substitution would affect protein function. Since it was released and the corresponding web server was established, SIFT has become a standard tool for characterizing missense variation. PolyPhen is another predictor which measures the impact of single amino acid substitutions on protein structure and function with supplying possibilities. PolyPhen uses several powerful

Table 1 Comparison with SIFT and PolyPhen-2 on a same dataset

Method	Sensitivity (%)	Specificity (%)	Accuracy (%)	MCC
SIFT	75.85	75.08	75.41	0.504
PolyPhen-2	80.85	73.43	76.53	0.536
Present work (KNN)	68.68	81.42	75.93	0.506
Present work (RF)	72.35	86.54	80.43	0.598

features derived from functional annotations, structural parameters and sequence alignments. In recent year, the method updated to PolyPhen-2 (Adzhubei et al. 2010), which was superior to PolyPhen. Due to the updated database, our data were of <0.1 % difference from those of SIFT and PolyPhen-2, which would place quite limited effect on our comparison. Three methods were implemented on the same dataset (release 57.4), and the results are shown in Table 1. Our method got higher specificity, accuracy and MCC values. The sensitivity was lower than the other methods, because our training data were unbalanced that negative samples larger than positive ones. It could be deduced that the overall performance of our method was better than SIFT and PolyPhen-2.

Potential bias in predictions

The major issue of these protein level information using predictions is that homology of proteins may lead to an overestimation of the classifier. To evaluate the effect derived from homology on our method, protein cluster procedures were implemented before cross-validation.

Specifically, all 9,623 proteins were divided into 7,634 clusters through CD-HIT (Huang et al. 2010b) with a cut-off value 0.4. And then, a tenfold cross-validation was performed in which the mutations of the same cluster were kept in one subset. The results (accuracy 80.23 %, sensitivity 72.50 %, specificity 86.09 %, MCC 0.592) were almost the same as that of original protein level cross-validation.

Similarly, to avoid the overestimation, homolog proteins to those in the independent dataset were removed from the training dataset. The comparing results to SIFT and PolyPhen-2 on the independent dataset are listed in Table 2. Furthermore, considering mutations belonging to hybrid proteins (proteins containing both disease-related and neutral mutations) were difficult to be precisely predicted, we supplemented a competition among three methods based on these mutations. Under a very harsh condition that homology proteins were simultaneously removed from main training and independent dataset, our method obtained competitive results (as shown in Table S1).

Table 2 Comparison with SIFT and PolyPhen-2 on independent dataset

Method	Sensitivity (%)	Specificity (%)	Accuracy (%)	MCC
SIFT	75.69	77.47	77.03	0.480
PolyPhen-2	80.16	77.41	78.10	0.518
Present work	69.61	94.11	87.99	0.668

Table 3 Performance comparison by eliminating attributes group from model

Eliminated attributes group	ACC (%)	MCC
Topological parameters	76.13	0.50
Evolutionary conservation and sequence-derived information	77.28	0.53
Physicochemical characteristics in AAindex	79.88	0.59
Scores calculated by tools	79.76	0.58
None	80.43	0.60

Feature group performance

For detecting the performance of PPI topological parameters, all features were grouped into four categories for sorting from the same sources or having similar properties (Fig. 1): the first set contained four topological parameters; the second group was composed of a 32 dimensional vector of evolutionary conservation, sequence profile and sequence-derived information; the third group consisted of a 13 dimensional data containing physicochemical characteristics; the rest were calculated scores by bioinformatics tools. Each category was eliminated from all categories in turn and the rest three categories were adopted for a tenfold crossvalidation. A stringent cross-validation was also performed here to prevent an overestimation of the predictor. When eliminating different group of features, predicting accuracy and MCC started toppling with various degrees in value. As shown in Table 3, it was observed that the prediction accuracy decreased by 4.3 % when topological features were removed, which was close to that of SIFT and PolyPhen-2. The addition of these attributes significantly improved results in prediction, which demonstrated that PPI topological features could provide significant information in discovering nsSNPs associated with human hereditary diseases.

Feature evaluation

Our dataset consisted of 57 different features and each of them made a different contribution to our model. We had

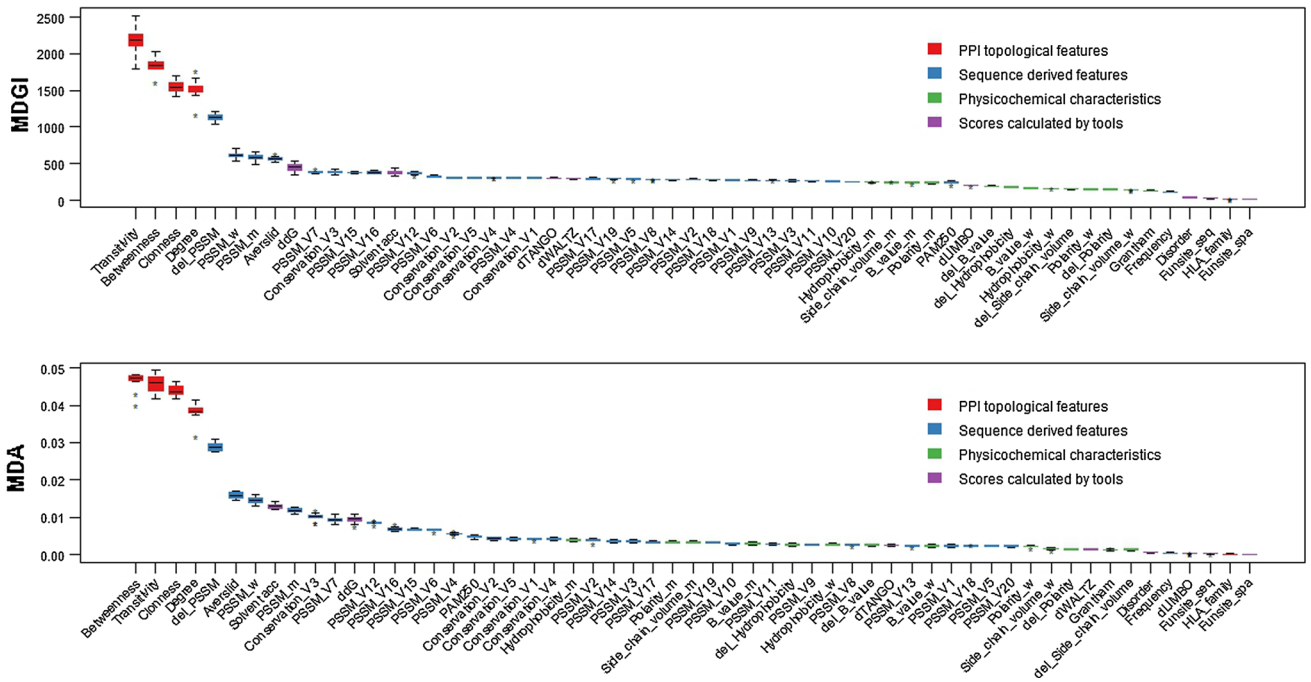

Fig. 2 Feature evaluation results by random forest. The MDGI and MDA scores of features were measured 10 times during the tenfold crossvalidation procedure in R. Features were ranked by their average scores in a descending order. The features were *marked with colors* according to their categories (color figure online)

roughly evaluated the importance of topological features by measuring accuracy decrease when removing them from dataset and found they were critical for predicting. Actually, the random forest package in R provided modules to evaluate the features in an unbiased way. Mean decrease accuracy and mean decrease Gini index were adopted to weigh the role by which variables played in the model. We used box plots to reflect the scores of MDA and MDGI for they were calculated during every iteration in tenfold crossvalidation (Fig. 2). It could be observed that the MDA and MDGI scores of topological parameters were top ranked.

The human PPI network topological features were employed to discover disease genes (Nibbe et al. 2011). In a similar way, we used them to predict disease-related nsSNPs which occur in coding regions of genes. It was supposed that human diseases which share a common biochemical mechanism could be caused by mutations in one of the several genes that interact at the protein level in a pathway or subnetwork, which means that genes may share features of their products in a pathway or subnetwork. Grandhi et al. (2006) supported this hypothesis and found that genes ascertained from the OMIM database that was associated with a human disease preferentially interacted with other disease-causing genes over those without known disease association and demonstrated the existence of disease subnetwork. These preferentially interacting

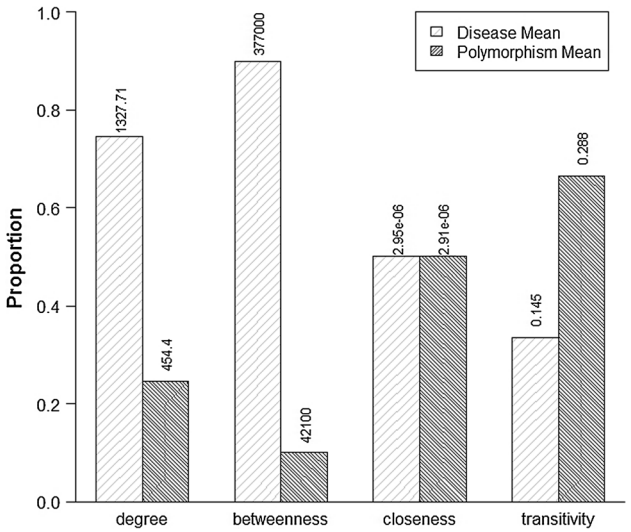

Fig. 3 Compare the mean values of topological features between disease samples and polymorphism ones. The topological features mean value of disease samples and polymorphism ones were calculated, respectively, and they are *listed on the bars*. To every attribute, the sum of disease mean value and polymorphism mean value was set to one and the heights of the bars reflect their ratios, respectively

gene products were also related at the protein level that leaded to form protein subnetworks. As for different type of subnetworks, they contained different numbers of nodes

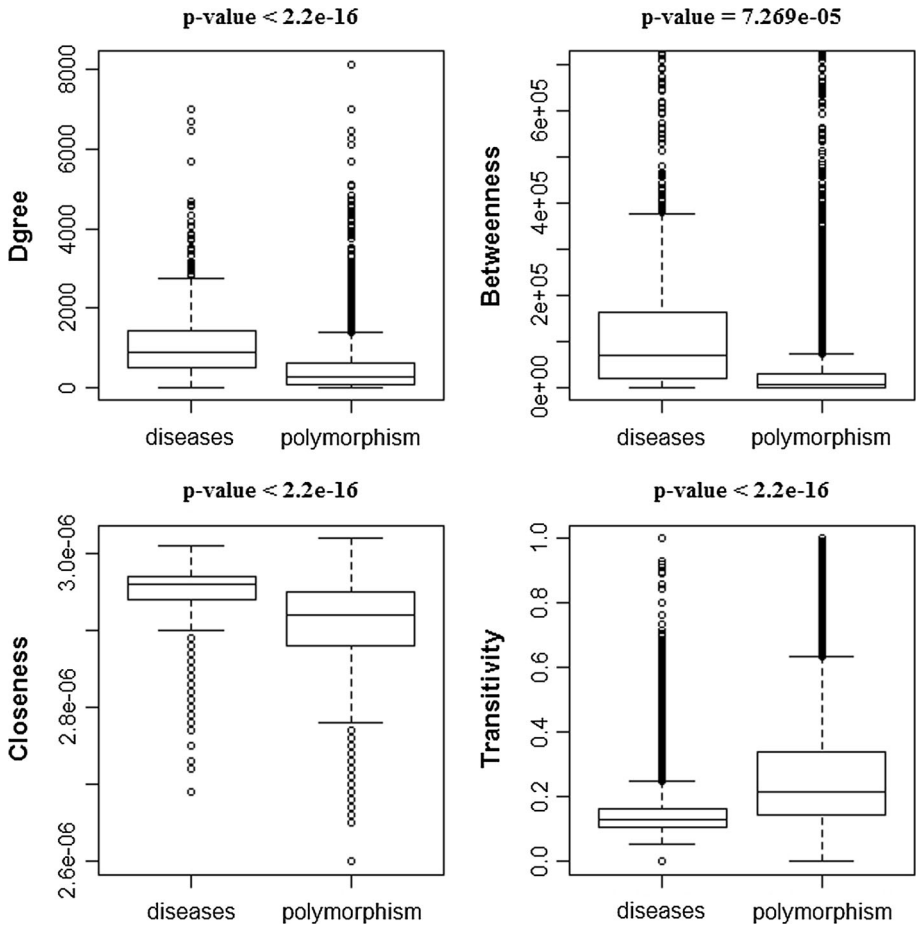

Fig. 4 The statistic analysis of topological features. Four topological features were analyzed and the *box plots* reflect the data distribution of disease samples and polymorphism ones. Compared with polymorphism data, the distributions of disease data are more disperse in degree and betweenness, but more centralized in closeness and transitivity. In general, the distribution of disease-related data and that of polymorphism data are different

and edges. It meaned network topological parameters of a protein were similar with those among the same subnetwork but different from proteins in other groups. This property could make proteins be classified as those with similar topological attributes, so do corresponding nsSNPs of proteins. We supposed that this may be the reason why PPI topological parameters enhanced the predicting result.

By analyzing four topological features, we found that the differences do exist between disease-causing nsSNPs and polymorphism ones (Fig. 3). The average degree and betweenness of disease-related samples are noticeably higher than neutral samples, indicating that deleterious nsSNPs corresponding proteins occupy key positions in the biological mechanism. These proteins interact with a number of proteins and are involved in more pathways. Dysfunctional mutation happened on such protein could place more effects on the biological process. The closeness values are equal, indicating that proteins have similar abilities to connect with other proteins. Polymorphism nsSNPs' higher transitivity scores mean that the neighbors of corresponding proteins have strong ability of interconnection. This can be deduced to some relevance with the formation of protein function modules or sub networks. It can be reasonably inferred that proteins in modules implementing biological functions should be robust and steady. A similar opinion of Ramensky et al. (2002) is that the selection against these variants is likely to depend on the molecular function of proteins rather than on the type of structure or cellular localization. To make sure the mean value was not meaningless, we carried a statistical analysis of topological features on positive and negative samples, respectively. As shown in Fig. 4, it was observed that the four topological features were of different distribution for positives and negatives, respectively. Furthermore, all proteins were divided into three categories: (1) diseased proteins: proteins with only disease mutations, (2) neutral proteins: protein with only polymorphisms, (3) hybrid proteins: protein with both disease and polymorphic variants. We examined their network properties and the results are included in Figure S1. It is obvious that topological features of D.proteins are quite different from those of N.proteins. Unsurprisingly, similar topological features were observed between D.proteins and H.proteins. It indicates that the fragility of protein function to a mutation could be determined by its molecular specificity as well as its role in the whole biological network. The high degree

score for D.protein indicated that the corresponding genes are functionally significant and highly conserved; this result was consistent with the conclusion of a recent research (Khurana et al. 2013).

It is worth noting that more topological features did not mean a better result. We have tried using one to eight topological features in the model and found that different combination of topological features attained various results. Besides, single topological feature could not improve the predicting result significantly. With regard to the other attributes, the features derived from evolutionary conservation information still performed well in our study and occupied the second most important place of feature groups. Another frequently used feature, solvent accessibility, also represented its discriminative attribute in our model, which was congruent with previous studies. The Averslid that compiled from sliding window and smooth window outperformed most of PSSM scores.

As shown in Fig. 2, we found that some powerful features in previous studies were not so contributive in our method. Such as HLA family and near function sites, we suppose their low counts make them not so outstanding in large sample predictions. In detail, among our totally 46,996 mutations, only 3,459 mutation close to function sites and 1,505 mutations related to HLA family proteins. Similarly, two factors lead to the low performance of disorder region-associated feature: (1) the tool DISpro cannot correctly predict all disorder regions on proteins. (2) Not every protein contains disorder regions, but in our study, every protein was predicted with disorder regions. The false-positive rates restricted its discriminative ability in our method.

Conclusions

In this study, we found that the PPI network parameters, another kind of protein level features, could significantly increase predictive scores when added in a machine learning model and was proved to be superior to traditionally used powerful features in performance. Our method simply compiled 57 features without adopting protein structure files and outperformed in comparison with Huang, SIFT and PolyPhen-2. In particular, we got a higher MCC value on a large unbalanced data set. Furthermore, this work demonstrated that genes might share their products' information on protein level as Grandhi supposed. Our results may help to get a deeper understanding of relations between genes and their products at the protein level or in pathways and design a better predictor in finding disease-causing nsSNPs.

Acknowledgments We would like to thank Zhiqiang Ye for providing us data for comparison. We thank the anonymous reviewers for their patient review and constructive suggestions. This work was supported by the National Natural Science Foundation of China (21175095), Doctoral Fund of Ministry of Education of China (20120181110051).

Conflict of interest We declare that we have no conflict of interest.

References

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR (2010) A method and server for predicting damaging missense mutations. Nat Methods 7:248–249

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Bao L, Cui Y (2005) Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information. Bioinformatics 21:2185–2190

Belli M, Ramazzotti M, Chiti F (2011) Prediction of amyloid aggregation in vivo. EMBO Rep 12:657–663

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

Calabrese R, Capriotti E, Fariselli P, Martelli PL, Casadio R (2009) Functional annotations improve the predictive score of human disease-related mutations in proteins. Hum Mutat 30:1237–1244

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

Capriotti E, Arbiza L, Casadio R, Dopazo J, Dopazo H, Marti-Renom MA (2007) Use of estimated evolutionary strength at the codon level improves the prediction of disease-related protein mutations in humans. Hum Mutat 29:198–204

Care MA, Needham CJ, Bulpitt AJ, Westhead DR (2007) Deleterious SNP prediction: be mindful of your training data! Bioinformatics 23:664–672

Cheng J, Randall AZ, Sweredoski MJ, Baldi P (2005a) SCRATCH: a protein structure and structural feature prediction server. Nucleic Acids Res 33:W72–W76

Cheng J, Sweredoski MJ, Baldi P (2005b) Accurate prediction of protein disordered regions by mining protein structure data. Data Min Knowl Discov 11:213–222

Csardi G, Nepusz T (2006) The igraph software package for complex network research. Inter J Complex Syst 1695:38

De Baets G, Van Durme J, Reumers J, Maurer-Stroh S, Vanhee P, Dopazo J, Schymkowitz J, Rousseau F (2012) SNPeffect 4.0: online prediction of molecular and structural effects of protein-coding variants. Nucleic Acids Res 40:D935–D939

Dobson RJ, Munroe PB, Caulfield MJ, Saqi MA (2006) Predicting deleterious nsSNPs: an analysis of sequence and structural attributes. BMC Bioinform 7:217

Ferrer-Costa C, Orozco M, De La Cruz X (2004) Sequence-based prediction of pathological mutations. Proteins Struct Funct Bioinform 57:811–819

Gandhi T, Zhong J, Mathivanan S, Karthick L, Chandrika K, Mohan SS, Sharma S, Pinkert S, Nagaraju S, Periaswamy B (2006) Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. Nat Genet 38:285–293

Gibbs RA, Belmont JW, Hardenbol P, Willis TD, Yu F, Yang H, Chang LY, Huang W, Liu B, Shen Y (2003) The international HapMap project. Nature 426:789–796

Hu J, Yan C (2008) Identification of deleterious non-synonymous single nucleotide polymorphisms using sequence-derived information. BMC Bioinform 9:297

Huang T, Wang P, Ye ZQ, Xu H, He Z, Feng KY, Hu L, Cui W, Wang K, Dong X, Xie L, Kong X, Cai YD, Li Y (2010a) Prediction of deleterious non-synonymous SNPs based on protein interaction network and hybrid properties. PLoS One 5:e11900

Huang Y, Niu B, Gao Y, Fu L, Li W (2010b) CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics 26:680–682

Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M (2008) AAindex: amino acid index database, progress report 2008. Nucleic Acids Res 36:D202–D205

Khurana E, Fu Y, Chen J, Gerstein M (2013) Interpretation of genomic variants using a unified biological network approach. PLoS Comput Biol 9:e1002886

Krishnan VG, Westhead DR (2003) A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function. Bioinformatics 19:2199–2209

Li Y, Wen Z, Xiao J, Yin H, Yu L, Yang L, Li M (2011) Predicting disease-associated substitution of a single amino acid by analyzing residue interactions. BMC Bioinform 12:14

Liaw A, Wiener M (2002) Classification and regression by random forest. R News 2:18–22

Maurer-Stroh S, Debulpaep M, Kuemmerer N, de la Paz ML, Martins IC, Reumers J, Morris KL, Copland A, Serpell L, Serrano L (2010) Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. Nat Methods 7:237–242

Ng PC, Henikoff S (2001) Predicting deleterious amino acid substitutions. Genome Res 11:863–874

Ng PC, Henikoff S (2003) SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res 31:3812–3814

Ng PC, Henikoff S (2006) Predicting the effects of amino acid substitutions on protein function. Annu Rev Genomics Hum Genet 7:61–80

Nibbe RK, Chowdhury SA, Koyutürk M, Ewing R, Chance MR (2011) Protein–protein interaction networks and subnetworks in the biology of disease. Wiley Interdiscipl Rev Syst Biol Med 3:357–367

Nicodemus KK (2011) Letter to the editor: on the stability and ranking of predictors from random forest variable importance measures. Brief Bioinform 12:369–373

Qin W, Li Y, Li J, Yu L, Wu D, Jing R, Pu X, Guo Y, Li M (2012) Predicting deleterious non-synonymous single nucleotide polymorphisms in signal peptides based on hybrid sequence attributes. Comput Biol Chem 36:31–35

Ramensky V, Bork P, Sunyaev S (2002) Human non-synonymous SNPs: server and survey. Nucleic Acids Res 30:3894–3900

Reumers J, Schymkowitz J, Ferkinghoff-Borg J, Stricher F, Serrano L, Rousseau F (2005) SNPeffect: a database mapping molecular phenotypic effects of human non-synonymous coding SNPs. Nucleic Acids Res 33:D527–D532

Reumers J, Schymkowitz J, Rousseau F (2009) Using structural bioinformatics to investigate the impact of non synonymous SNPs and disease mutations: scope and limitations. BMC Bioinform 10:S9

Robinson J (2003) IMGT/HLA and IMGT/MHC: sequence databases for the study of the major histocompatibility complex. Nucleic Acids Res 31:311–314

Robinson J, Waller MJ, Fail SC, McWilliam H, Lopez R, Parham P, Marsh SG (2009) The IMGT/HLA database. Nucleic Acids Res 37:D1013–D1017

Saunders CT, Baker D (2002) Evaluation of structural and evolutionary contributions to deleterious mutation prediction. J Mol Biol 322:891–901

Sickmeier M, Hamilton JA, LeGall T, Vacic V, Cortese MS, Tantos A, Szabo B, Tompa P, Chen J, Uversky VN, Obradovic Z, Dunker AK (2007) DisProt: the database of disordered proteins. Nucleic Acids Res 35:D786–D793

Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NST, Abeysinghe S, Krawczak M, Cooper DN (2003) Human gene mutation database (HGMD®): 2003 update. Hum Mutat 21:577–581

Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P, Jensen LJ, von Mering C (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res 39:D561–D568

Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A (2003) PANTHER: a library of protein families and subfamilies indexed by function. Genome Res 13:2129–2141

Tian J, Wu N, Guo X, Guo J, Zhang J, Fan Y (2007) Predicting the phenotypic effects of non-synonymous single nucleotide polymorphisms based on support vector machines. BMC Bioinform 8:450

Van Durme J, Maurer-Stroh S, Gallardo R, Wilkinson H, Rousseau F, Schymkowitz J (2009) Accurate prediction of DnaK-peptide binding via homology modelling and experimental data. PLoS Comput Biol 5:e1000475

Wang M, Shen HB, Akutsu T, Song J (2011) Predicting functional impact of single amino acid polymorphisms by integrating sequence and structural features. In: 2011 IEEE international conference on systems biology (ISB), pp 18–26

Xu J, Li Y (2006) Discovering disease-genes by topological features in human protein–protein interaction network. Bioinformatics 22:2800–2805

Yang J, Li YY, Li YX, Ye ZQ (2012) Partition dataset according to amino acid type improves the prediction of deleterious non-synonymous SNPs. Biochem Biophys Res Commun 419:99–103

Ye ZQ, Zhao SQ, Gao G, Liu XQ, Langlois RE, Lu H, Wei L (2007) Finding new structural and sequence attributes to predict possible disease association of single amino acid polymorphism (SAP). Bioinformatics 23:1444–1450