Journal of Molecular Biology Research; Vol. 4, No. 1; 2014
ISSN 1925-430X E-ISSN 1925-4318
Published by Canadian Center of Science and Education

Combining Network Topological Characteristics With Sequence and Structure Based Features for Predicting Protein Stability Changes Upon Single Amino Acid Mutation

Lijun Yang[1], Qifan Kuang[1], Yanping Jiang[1], Ling Ye[1], Yiming Wu[1], Menglong Li[1] & Yizhou Li[1]

[1] College of Chemistry, Sichuan University, Chengdu, China

Correspondence: Yizhou Li, College of Chemistry, Sichuan University, Chengdu 610064, China. E-mail: liyizhou_415@163.com

Received: March 6, 2014 Accepted: March 31, 2014 Online Published: April 14, 2014
doi:10.5539/jmbr.v4n1p20 URL: http://dx.doi.org/10.5539/jmbr.v4n1p20

Abstract

It has been shown that the stability of protein structure could be significantly changed by single amino acid substitution. Accurate prediction of protein stability changes caused by single amino acid substitutions is valuable for understanding the relationship between protein structures and functions as well as designing new proteins. Currently, various computational methods have been developed to study the effect of single amino acid mutation on protein stability. In this study, by combining network topological characteristics extracted from Protein Structure Network (PSN) with other physicochemical features obtained from protein sequence or structure, a Support Vector Machine (SVM) model was developed to distinguish the stabilizing mutants from the destabilizing mutants. 20-fold cross-validation was implemented for performance evaluation. An accuracy of 0.88 and a Matthews Correlation Coefficient (MCC) of 0.71 were obtained for the dataset with 1925 variants.

Our method is superior to the existing machine learning approaches evaluated under the same datasets. It suggests that such a combining strategy should be valuable in predicting protein stability changes upon amino acid mutation. In our study, the topological parameters are informative for prediction upon substitutions. Moreover, it is indicated that the Protein Structure Network (PSN) could be effectively used for representing the three-dimensional structure of protein and such network parameters are associated with the changes of protein function and structure.

Keywords: amino acid mutation, protein stability prediction, Protein Structure Network (PSN), Support Vector Machine (SVM), topological features

1. Introduction

Single amino acid substitution shows great impact on protein. Such substitutions can cause a series of changes of proteins, including the physicochemical properties, structures as well as functions and etc. (Shirley, Stanssens, Hahn, & Pace, 1992), which may lead to protein destabilization and even diseases. Variations are mostly owing to non-synonymous single nucleotide polymorphisms (nsSNPs). Each person may have 24,000-40,000 nsSNPs, and there are a total of more than 67,000 common nsSNPs in the human population (Cargill et al., 1999). These nsSNPs may result in amino acid changes in proteins. The majority of nsSNPs are neutral for protein function, while the remains may affect protein function and lead to diseases. It is estimated that about 25% of nsSNPs may be deleterious to protein function (Yue & Moult, 2006). Moreover, the majority of disease-associated mutations are caused by protein destabilization (Wang & Moult, 2001). Therefore, if a mutation causes the protein destabilization, it is more likely to cause disease. So, accurate prediction of how single amino acid mutation affects the stability of proteins is invaluable for understanding the protein structure, function and designing new proteins. However, with the increasing number of mutant data, it is not enough to determine the effect of each mutation on protein through experiment alone for its time and labor-consuming. Thus, more and more machine learning based methods have been developed owing to its quick and effective prediction performance.

The selections of feature vectors and algorithms are really important in machine learning methods. The machine learning methods can be designed to predict either the sign of the stability free energy change ($\Delta\Delta G$, classification) upon mutation (stabilizing if $\Delta\Delta G > 0$, destabilizing if $\Delta\Delta G < 0$) or the actual value of the stability free energy change ($\Delta\Delta G$, regression). Until now, many machine learning methods have been applied in the prediction of protein stability changes upon single amino acid mutation. These methods took advantage of sequence or structural information of proteins for classification or regression or for both. However, for most biological applications, accurate prediction of the sign of ΔΔG (classification) is more relevant than estimating the actual value of ΔΔG (regression) (Capriotti, Fariselli, & Casadio, 2004). In this research, the classification models were constructed to predict the sign of ΔΔG instead of predicting the actual value of ΔΔG.

Cheng et al. (Cheng, Randall, & Baldi, 2006) used Support Vector Machine (SVM) with sequence and structural information of proteins to predict protein stability changes upon single amino acid substitutions, and achieved 84% accuracy. Teng et al. (Teng, Srivastava, & Wang, 2010) developed a sequence feature-based model to predict protein stability changes upon amino acid substitutions and an overall accuracy of 84.59% was obtained. It was shown that the relevant sequence features can be used for the prediction of protein stability changes upon single amino acid substitutions. Folkman et al. (Folkman, Stantic, & Sattar, 2013) proposed a number of evolutionary features and predicted structural features acquired from protein sequence to predict the protein stability changes and achieved satisfying results. Yang et al. (Yang, Chen, Tan, Vihinen, & Shen, 2013) reported a structure-based method with the best accuracy of 87% which combined the contact energy with other physicochemical properties of amino acids. Based on the studies mentioned above, it is clear that either the protein sequence or structure implies useful information for predicting protein stability changes upon single amino acid substitutions.

From another perspective, a protein is truly an interacting network system because of the connections between nodes (amino acids). Both the location of residues and the interactions among them are important for the function and stability of protein. Bagler et al. and Greene et al. described the small-word property of protein structure network (Bagler & Sinha, 2005; Greene & Higman, 2003). Dokholyan et al. and Vendruscolo et al. demonstrated that a selected set of residues played an important role in protein folding from protein structure network (Dokholyan, Li, Ding, & Shakhnovich, 2002; Vendruscolo, Dokholyan, Paci, & Karplus, 2002). Also, another study showed that a set of hub residues in protein structure network primarily contributed to both the folding and stability of proteins (Brinda & Vishveshwara, 2005). Amitai et al. took advantage of such protein structure network for studying functional residues (Amitai et al., 2004). Li et al. predicted disease-associated substitution upon single site mutation by analyzing residue interactions in protein structure network (Li et al., 2011). Therefore, the protein structure network involves much potential information for biochemical studies.

Overall, the studies mentioned above suggest that it is feasible for predicting protein stability changes based on sequence or structure information of proteins and that the protein structure network is a valuable method for utilizing the structure information of proteins. In this study, the protein structure network was introduced which regards a protein structure as a network and then four topological parameters were calculated for each residue in a protein. Our results indicate that only using topological features can achieve a satisfying accuracy in prediction with SVM (accuracy of 0.83 for S1925 dataset). Better results can be obtained when combining the topological features with other physicochemical properties. It suggests that the topological features we introduced are useful for predicting protein stability changes upon single amino acid substitutions. It further indicats that the network parameters of amino acids are associated with the changes of protein function and structure. In addition, the calculated environmental features are also informative for the prediction from the research and eight sequence neighboring residues are enough for providing environmental information of the variant.

2. Materials and Methods

2.1 Data Sets

In this work, two datasets were used for training and testing the models. The datasets S1925 and S2760 were originally extracted from ProTherm database (Bava, Gromiha, Uedaira, Kitajima, & Sarai, 2004) (http://gibk26.bio.kyutech.ac.jp/jouho-u/Protherm/protherm.html). The dataset S2760 was collected from (Yang et al., 2013) and S1925 dataset was originally compiled by (Masso & Vaisman, 2008).

S1925 contained 1925 single mutations for 55 proteins with 582 positive and 1343 negative cases. While in S2760, there were 2758 single mutations for 75 different proteins with 872 positive and 1886 negative cases, in which two mutations were removed due to the mismatch (renamed as SR2760). The two data sets both contained six attributes for each record including the PDB ID, the mutation, pH, Temperature (T), ΔΔG and the solvent accessibility (ASA of the variant residue).

2.2 Protein Structure Network

In this study, a network was constructed from a protein PDB structure where the nodes were residues and the interactions between residues were edges. If the distance between any two atoms from two residues was smaller than the sum of their Vander Waals radius plus 0.5Å (Greene et al., 2003), the two residues were considered to interact with each other. Based on the constructed protein structure network, then, the igraph package in R (Ihaka & Gentleman, 1996) was used to extract the topological parameters of each residue from the network.

2.3 Features

Thirty-four attributes were used to encode each data instance, including network topological features, environmental features and other physicochemical properties. We divided these features into the following classes.

2.3.1 Network Topological Features

In this study, four network topological features were calculated, including Degree, Clustering Coefficient, Betweenness and Closeness.

Degree is the number of edges that directly connect to node *i*. It reflects the number of the nearest neighbors for vertex *i*. It is calculated as

$$D\,(i) = \Sigma_{j\epsilon N}\, a_{i,j} \tag{1}$$

Where $a_{i,j}$ is the number of edges between nodes *i* and *j*, and N is the total number of vertices incident to vertex *i*.

The *Clustering Coefficient* reflects how well connected are the neighbors of node *i*. It can be calculated as

$$CC\,(i) = \frac{2e_i}{D_i(D_i - 1)} \tag{2}$$

Where e_i is the virtual number of edges among the neighborhoods of vertex *i*, and D_i is the degree of node *i*.

Betweenness reflects the probability of a vertex occurs on the shortest paths between other vertices. It is calculated as

$$B\,(i) = \sum_{j,k \in N, j \neq k} \frac{n_{j,k}(i)}{n_{j,k}} \tag{3}$$

Where $n_{j,k}$ is the number of shortest paths connecting vertices *j* and *k*. $n_{j,k}(i)$ is the number of shortest paths linking *j* and *k* which pass through vertex *i*.

Closeness describes the adjacent level of vertex *i* and all other vertices. It is calculated as

$$C\,(i) = \frac{N-1}{\sum_{i \neq j} d_{i,j}} \tag{4}$$

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

For more detailed information of the four parameters, please refer to the references (Newman, 2003; Watts & Strogatz, 1998)

2.3.2 Amino Acid Properties of the Variant

Eleven amino acid properties were introduced in this work, which were obtained from AAindex (Kawashima & Kanehisa, 2000) (http://www.genome.jp/aaindex/), Protscal (Gasteiger et al., 2005) (http://web.expasy.org/protscale/) and the reference (Collantes & Dunn, 1995). These properties consist of Bulkiness (Bu), Average area buried on transfer from standard state to folded protein (Aa), Conformational parameter for alpha helix (Al), Beta-sheet (Be) and Coil (Co), Polarity (P), Hydropathicity (H), Transmembrane tendency (Tt), Flexibility (F), Electronic charge index (ECI) and Isotropic surface area (ISA).

Bulkiness (Bu) (Zimmerman, Eliezer, & Simha, 1968), reflecting the role of individual residues in the entire protein configuration, may influence the local conformation of protein. Average area buried on transfer from standard state to folded protein (Aa) (Rose, Geselowitz, Lesser, Lee, & Zehfus, 1985) estimates a residue's mean area buried upon folding. Protein secondary structures can be divided into three types, including alpha-helix, beta-sheet and coil conformations. Each type of amino acid has a different preference to form one of the three secondary structures. So, the Conformational parameters for alpha helix (Al) (Chou & Fasman, 1978), beta-sheet (Be) (Chou et al., 1978), and coil (Co) (Deleage & Roux, 1987) were introduced. Polarity (P) (Grantham, 1974) reflects the intermolecular interactions between the residues of positive and negative charge. Hydropathicity（H）(Kyte & Doolittle, 1982) takes the hydrophilic and hydrophobic properties of each 20 amino acid side-chains into consideration. It is important for keeping the protein structure as it is critical for amino acid side chain packing and protein folding. Transmembrane tendency (Tt) is related to biological hydrophobicity, which was reported by Zhao and London (Zhao & London, 2006). Flexibility (Vihinen, Torkkila, & Riikonen, 1994) of protein structure is important for catalysis, binding, and allostery, which shows correlation to protein stability. Electronic charge index (ECI) (Collantes et al., 1995) is a measure of the charge density for amino acid. Isotropic surface area (ISA) (Collantes et al., 1995) approximates the hydrophobic property of the side chain substituent, which directly reflects the effect of the variant on structure.

The difference between the variant and the original in each amino acid property was calculated as the variant characterization and the prefix d was used to characterize the information. For example, the attribute dBu was defined as the difference between the variant and the original in amino acid Bulkiness (Bu). So, the remaining ten attributes of the variant were dAa, dAl, dBe, dCo, dP, dH, dTt, dF, dECI and dISA.

2.3.3 Environmental Features

Additionally, the environmental features were also considered by using four network topological features and eleven amino acid properties of each residue in protein. The prefix Env was used to characterize the environmental features. There were fifteen environmental features defined as Env_CC, Env_D, Env_B, Env_C, Env_Aa, Env_Al, Env_Be, Env_Co, Env_P, Env_H, Env_Tt, Env_F, Env_ECI and Env_ISA.

To calculate the environmental features, we took a subsequence of *w* consecutive residues from a protein sequence into consideration, where *w* was called the window size. The mutation site was located in the middle of the subsequence, and the other (*w*-1) neighboring residues provided the environment information for the substitution site. Then, the average value obtained from neighboring residues under each attribute was reported to be the corresponding environmental features. Here, took the attribute Env_P for instance.

The Polarity (P) for each of the 20 amino acid can be obtained from AAindex. When the window size was set to 7, 6 neighboring residues' polarity were averaged to represent the environmental features of the variant. The Env_P was calculated as

$$Env\_P\,(i) = \frac{\sum P_j}{w-1} \qquad (5)$$

Where P_j is the polarity of residue j and *w* is the window size.

In this study, we investigated nine kinds of window size to obtain different environmental features and analyzed their effect on classifier.

2.3.4 Evolutionary Feature

In a protein sequence, residues in different sites bear different evolutionary constraint. Specifically, functionally important sites tend to be more conserved. Here, we introduced an evolutionary feature, SIFT score(S).

SIFT (Ng & Henikoff, 2001) is a method for predicting whether an amino acid substitution affects the function of a protein and the prediction is based on the degree of conservation of amino acids in the alignment derived from PSI-BLAST. SIFT scores were scaled from 0 to 1. The amino acid substitutions are considered to be damaging if the scores below 0. 05. The SIFT Sequence tool was used to calculate the SIFT scores online (http://sift.jcvi.org/www/SIFT_seq_submit2.html). It was run on the Uniprot-TrEMBL database with sequences more than 90% identical to the query removed and the median conservation of sequences was set to 2.75.

2.3.5 Other Features

Additional three features were the solvent accessibility of the variant residue (ASA), the experimental pH and Temperature (T). All of the three attributes were extracted from ProTherm database.

2.4 Support Vector Machine

Support Vector Machine (SVM) has been successfully applied in a wide range of biological applications. It can learn from examples to assign labels to objects. Indeed, for understanding SVM classification, one only needs to comprehend the four basic concepts: (I) the separating hyperplane, (II) the maximum-margin hyperplane, (III) the soft margin and (IV) the kernel function (Noble, 2006). There are four basic kernels for SVM: linear, sigmoid, polynomial and radial basis function (RBF). In general, the RBF kernel is an optimal choice as demonstrated by Capriotti et al (Capriotti, Fariselli, Calabrese, & Casadio, 2005b). So, we chose the RBF kernel as SVM kernel function. The RBF kernel was calculated as

$$\mathrm{K}\,(\underset{x}{\rightarrow}, \underset{y}{\rightarrow}) = e^{-\gamma\left\|\vec{x}-\vec{y}\right\|^2}, \gamma>0 \qquad (6)$$

Where $\underset{x}{\rightarrow}$ and $\underset{y}{\rightarrow}$ are two vectors, and γ is a training parameter.

In the research, the SVMlight software package (http://svmlight.joachims.org/) was used to construct the classifiers. Various values of C and γ parameters were examined to optimize classifier performance. The best C and γ for each dataset were then used to construct the final classifier models.

2.5 Classifier Evaluation

The 20-fold cross-validation method was employed to evaluate the performance of the classifier under variant wise as many other studies did. In the case, all positive and negative instances were randomly divided into 20 folds without considering proteins. 19 folds were used to train a model, and then the remaining one fold was used to test the model. The process was repeated 20 times until each fold was used as the test set, then the evaluation value was obtained from the average result of the 20-fold cross validation. We computed the following performance measures:

$$\text{Accuracy (ACC)} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$$

$$\text{True positive rate (TPR)} = \frac{\text{TP}}{\text{TP+FN}}$$

$$\text{True negative rate (TNR)} = \frac{\text{TN}}{\text{TN+FP}}$$

$$\text{Positive predictive value (PPV)} = \frac{\text{TP}}{\text{TP+FP}}$$

$$\text{Negative predictive value (NPV)} = \frac{\text{TN}}{\text{TN+FN}}$$

$$\text{Matthews Correlation Coefficient (MCC)} = \frac{(\text{TP} \times \text{TN})-(\text{FP} \times \text{FN})}{\sqrt{(\text{TP+FN})(\text{TP+FP})(\text{TN+FN})(\text{TN+FP})}}$$

where TP, TN, FP, and FN are true positives, true negatives, false positives and false negatives, respectively. Matthews Correlation Coefficient (MCC) reflects the relationship between predictions and the real class labels.

In this paper, the area under the Receiver Operating Characteristic (ROC) curve (AUC) was also used as a measure of a classifier's performance (Bradley, 1997). Weak classifiers are with AUC values close to 0.5 and a perfect classifier has the maximum AUC value of 1.

3. Results and Discussion

3.1 Analysis of the Network Topological Features for Stabilization and Destabilization-Associated Variants

We constructed a classifier with four network topological features, including the Clustering Coefficient, Degree, Betweenness and Closeness of the variant. The results of S1925 and SR2760 dataset were listed in Table 1. For S1925, the obtained accuracy was 0.83 and MCC was 0.57. For SR2760 dataset, the accuracy of the model was about 0.81 and MCC was 0.54. Therefore, it is believed that the network topological features contained valuable information for predicting protein stability changes upon single amino acid mutation. Moreover, it can be demonstrated that the parameters reflecting the importance of nodes in protein structure network are related with the changes of protein function and structure.

Table 1. Predictive performance of classifiers constructed by topological features of the variants

Dataset	TPR	TNR	PPV	NPV	Accuracy	MCC	AUC
S1925	0.64	0.91	0.75	0.85	0.83	0.57	0.81
SR2760	0.62	0.90	0.74	0.83	0.81	0.54	0.80

Figure 1 shows that the two types of mutations differ in distributions of network topological features. The frequency was the ratio of mutations in the range of (x, x+d) to the total number of mutations. *Degree* describes the number of direct connections to a residue. Figure 1a shows that the destabilization-associated variants tend to have more neighbors than the stabilization-associated ones in the protein structure network. Residues with higher degree are more likely to be hub residues and therefore their mutations have more important influence on the protein structure and protein stability.

Similarly, in figure 1b, it can be observed that the destabilization-associated variants are with higher closeness compared with the stabilization-associated variants. It had been reported that residues associated with protein function have higher closeness than those neutral to protein function (Amitai et al., 2004). So, a centrally located residue in the protein structure network is more likely to associate with protein stability.

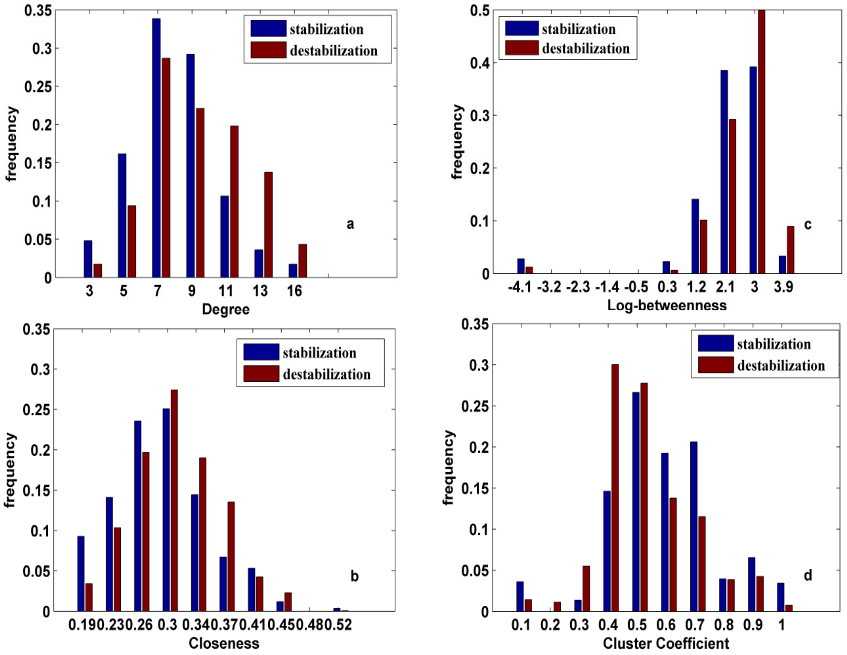


Figure 1. The frequency distributions of a) Degree; b) Closeness; c) Betweenness; d) Cluster Coefficient for stabilization and destabilization-associated mutants of S1925

For Betweenness and Clustering coefficient, the distributions of the two types of mutations are less distinct. However, as shown in figure 1c, higher frequencies were obtained for destabilization-related variants in the high-scoring region. Betweenness is an important centrality index of the interaction network. Residues with higher betweenness are more central residues, which make more important contribution for protein structure (Del Sol & O'Meara, 2005). Thus, the protein stability is more likely to be changed by mutations on the residues with higher betweenness. Figure 1d shows the distribution of *Clustering Coefficient*. As shown, higher frequencies were obtained for stabilization-associated variants in the high-scoring region. The clustering coefficient shows how well connected are the neighbors of a vertex in a network. Residues with high clustering coefficient reflect their compact environment. It can be found that the effect of substitution is related with the environment of the mutant site. More compact environment of the mutant site is of higher tolerance to substitution.

As mentioned above, the distributions of the four topological features between destabilization-associated and stabilization-associated mutations are different. The differences of the four parameters between the two types of substitutions are statistically significant (P value = 2.87E-19 for CC, P value = 5.86E-21 for D, P value = 4.43E-16 for B, P value = 1.15E-09 for C). Therefore, they can be identified as features with distinction power in discriminating destabilization-associated substitutions from other substitutions. The frequency distributions of the network topological features for SR2760 dataset can be obtained in Supplementary Figure 1.

3.2 The Optimal Choice of Window Size for Environmental Features Calculation

To calculate the environmental features, we took a subsequence of *w* consecutive residues into consideration, where *w* was the window size. The details about the calculation can be found in section 2.3.3. An optimal window size was chose for environmental features calculation by performing a conditional experiment. As the 11 amino acid properties and four topological parameters can be used to calculate the corresponding environmental features, we divided the environmental features into two types. One was the amino acid environmental features and the other was the topological environmental features. Then, two individual experiments were implemented. One took the 11 amino acid properties into consideration, and the other took the amino acid properties as well as topological parameters into consideration.

First, the amino acid properties were used for analyzing the optimal window size for environmental features calculation. For all the models, the amino acid properties of the variant were treated as baseline features. A model with the baseline features (i.e. no environmental features, *w* = 1) was constructed. Then, the classifiers with the baseline features and corresponding environmental features calculated from different window size were constructed for comparison. The results for S1925 dataset were listed in Table 2. As shown, the prediction was affected by environmental features. The model constructed without any environmental information (*w* = 1) gave a prediction accuracy of 0.75, but when the environmental information was considered (*w* > 1), the models' prediction strength was improved significantly with the accuracy of 83%-84%. The window sizes of 7 and 9 were both suitable for environmental features calculation,

Table 2. Effect of window sizes on classifiers constructed by amino acid properties for S1925 dataset

Window size	TPR	TNR	PPV	NPV	Accuracy	MCC	AUC
1	0.42	0.90	0.64	0.78	0.75	0.37	0.69
3	0.67	0.88	0.72	0.86	0.82	0.56	0.86
5	0.66	0.91	0.75	0.86	0.83	0.59	0.86
7	0.67	0.91	0.76	0.86	0.84	0.61	0.86
9	0.68	0.91	0.77	0.87	0.84	0.61	0.87
11	0.67	0.92	0.78	0.86	0.84	0.61	0.87
13	0.59	0.93	0.80	0.84	0.83	0.58	0.86
15	0.67	0.91	0.77	0.86	0.84	0.61	0.87
31	0.66	0.90	0.74	0.86	0.83	0.58	0.86

The effect of window sizes on SVM classifiers prediction strength was also illustrated by ROC curves. As shown in Figure 2, the ROC curves of the models for *w* = 7, 9 were better than that for *w* = 1. However, the classifier performance was not improved when using *w* = 31. It indicates that the environment of variants may also affect their influence on protein stability and .the window size is not the bigger the better.

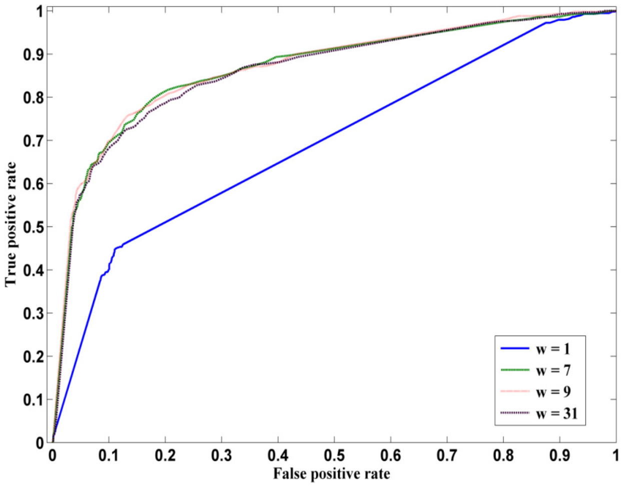


Figure 2. ROC curves to show the effect of window size on the performance of classifiers constructed with amino acid properties for S1925 dataset

To further study the best choice of window size, we combined the topological parameters and amino acid properties to detect their performance on classifiers using S1925 dataset. For these models, the baseline features were the amino acid properties and topological features of the variant (*w* = 1). A classifier with the baseline features (i.e. no environmental features, *w* = 1) was constructed. Then, the classifiers with the baseline features and corresponding environmental features calculated from different window size were constructed for comparison. Results were shown in Table 3. As shown, when no environmental features were considered (*w* = 1), the accuracy decreased, and other evaluation indexes also declined.

Table 3. Effect of window sizes on classifiers constructed with amino acid properties and topological parameters for S1925 dataset

Window size	TPR	TNR	PPV	NPV	Accuracy	MCC	AUC
1	0.69	0.87	0.71	0.86	0.82	0.57	0.82
3	0.64	0.94	0.82	0.86	0.85	0.63	0.87
5	0.65	0.94	0.84	0.86	0.85	0.64	0.88
7	0.73	0.92	0.80	0.88	0.86	0.66	0.89
9	0.74	0.92	0.80	0.89	0.86	0.67	0.89
11	0.64	0.94	0.83	0.86	0.85	0.64	0.88
13	0.64	0.94	0.83	0.86	0.85	0.63	0.87
15	0.64	0.94	0.83	0.86	0.85	0.63	0.87
31	0.64	0.94	0.82	0.86	0.85	0.63	0.87

From Table 2 and Table 3, it can be found that the use of w = 9 was marginally better than the use of other window sizes. As shown, the results of w = 7 and w = 9 were similar with each other. But the AUC of w = 9 was slightly better than w = 7 in Table 2. In addition, the difference between TPR and TNR of w = 9 was marginally smaller than that of w = 7 which represented that the classifier constructed with w = 9 was more balance in predicting positive and negative cases. Though the excellence is small, we also choose w = 9 as a better window size in order to improve the predictive strength in details. It is clear that the environmental features made contribution to the classifier. Moreover, it demonstrates that eight sequence neighbors are enough for providing environment information of the variant and that such environment may also have influence on protein structure and function.

3.3 The Classifier Performance Assessment Using the Dataset SR2760 and S1925

34 features were finally used to construct the model. Besides the network topological features and amino acid properties of the variants and their corresponding environmental features, the features ASA, SIFT score(S), T and pH were also included. The performance of the model was evaluated by 20-fold cross-validation with SVM. With C = 16, $\gamma = 0.0625$, the best result (accuracy of 0.86 and MCC of 0.68) was obtained when the window size was set to 9 (w = 9) for SR2760 dataset. We also compared the result with other methods, like M47 (Yang et al., 2013), FoldX (Schymkowitz et al., 2005), I-Mutant 2.0 (Capriotti, Fariselli, & Casadio, 2005a) and MUpro (Cheng et al., 2006). As the M47 model was also compared with the other three methods, we just got the comparison results from literature (Yang et al., 2013). The comparison was listed in Table 4. It reveals that our method outperforms in the comparison.

Table 4. Prediction performance for the SR2760 dataset

Method	TPR	TNR	PPV	NPV	Accuracy	MCC
Our method	0.75	0.91	0.80	0.89	0.86	0.68
M47	0.66	0.94	0.84	0.86	0.85	0.65
FoldX	0.64	0.72	0.52	0.81	0.70	0.35
I-Mutant 2.0	0.64	0.93	0.81	0.85	0.84	0.61
MUpro	0.65	0.92	0.79	0.85	0.84	0.61

Table 5. Prediction performance for the S1925 dataset

Method	TPR	TNR	PPV	NPV	Accuracy	MCC
Our method	0.76	0.93	0.83	0.90	0.88	0.71
M47	0.76	0.92	0.80	0.9	0.87	0.68
AUTO-MUTE	0.70	0.9	0.75	0.87	0.84	0.61
FoldX	0.55	0.69	0.43	0.78	0.66	0.22
I-Mutant 2.0	0.56	0.91	0.73	0.83	0.80	0.51
Mupro	0.68	0.92	0.79	0.87	0.85	0.63

To further compare our method with other methods, we used the S1925 dataset to test the model. With C = 4, $\gamma = 0.0625$ and window size of 9 ($w = 9$), the accuracy of our model was 0.88 for S1925 dataset with 0.71 MCC. The result was compared with other five methods, M47, AUTO-MUTE (Masso et al., 2008), FoldX, I-Mutant 2.0 and Mupro. Our approach obtained the best prediction accuracy (Table 5).

3.4 Feature Importance Measures

The analysis of feature importance was addressed by adopting the feature estimation module of random Forest package in R (Breiman, 2001). Permutation accuracy importance measure was adopted to evaluate the importance of each of the 34 features used in this research. According to the principle of permutation importance, the higher score a feature gets, the more important it will be.

The evaluation process was repeated 100 times. The importance score of each parameter was shown as box plots in Figure 3. In a box plot, the middle bar represents the mean value of scores. The top three features were dBu, dP and ASA. It had been reported that Bu, P and ASA were relevant for predicting protein stability changes on single site mutation by (Teng et al., 2010; Yang et al., 2013). The attributes Betweenness (B) and Clustering Coefficient (CC) ranked in top 10. Moreover, the scores between topological features and the top rank parameters were not in huge difference. It can be believed that the topological parameters are comparable with previously reported valuable attributes. In addition, the environmental features we calculated were not in high score as they were used for measuring individually. When combining the environmental features together, they make valuable contribution for predicting as demonstrated in section 3.2.

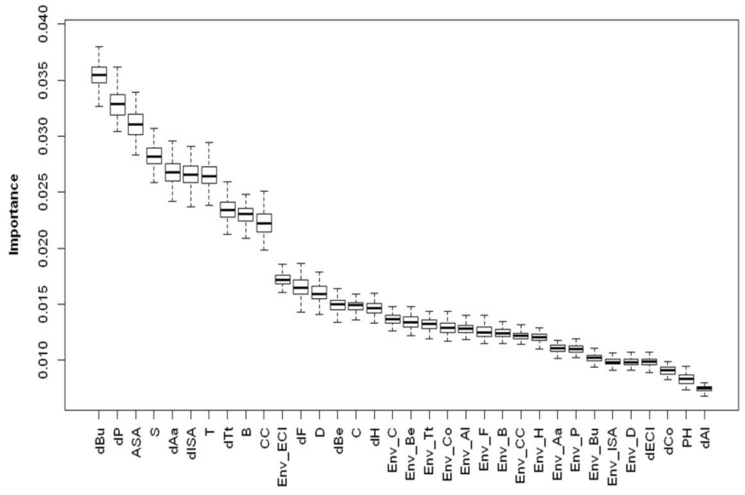


Figure 3. Importance score of each feature evaluated by the random Forest algorithm in the R package

3.5 The Analysis of the Mutant Rules

As SR2760 dataset almost covered the S1925 dataset, so we just used the SR2760 dataset for analyzing the mutant rules. Owing to the dataset contained more than one experimental ddG values for a variant under the same experimental conditions, the redundant data were removed to ensure that a unique ddG value for a variant under the same experimental conditions can be acquired. After a rigorous selection process, the final dataset contained 2218 mutations. We renamed the dataset as NSR2760 dataset. Based on NSR2760 dataset, we got the frequency of stabilizing and destabilizing mutations. The results were shown in Supplementary Table 1.

As shown in Supplementary Table 1, several rules can be observed. If the deleted residue is L while the introduced residue is A, the stability change will be negative (frequency = 88.6%). The mutants from A to S (frequency = 90.3%), Y to A (frequency = 100%), V to G (frequency = 90.5%), I to F (frequency = 100%), I to M (frequency = 93.8%) and so on also cause the protein destabilization. It may be believed that different mutations have their specific effect on protein stability changes. We hope such statistics will give valuable information for the protein engineering.

3.6 The Model Performance Assessment Using the New Cut Datasets

As the two datasets both contained more than one experimental ddG values for a variant under the same experimental conditions, in order to ensure that our method was not significantly affected by the redundant data, we made a selection for the datasets. After a rigorous selection process, the final dataset contained 1752 mutations with 531 positive and 1221 negative cases for S1925 dataset. We renamed the new dataset as NS1925 dataset. SR2760 which was renamed as NSR2760 dataset contained 2218 mutations with 696 positive and 1522 negative cases. The new datasets just involved a unique ddG value for a variant under the same experimental conditions.

We used the new datasets with 34 features to construct the classifier. The window size was set to 9 ($w = 9$). The model achieved accuracy of 0.86 with 0.67 MCC for NS1925 dataset. For NSR2760 dataset, the accuracy was 0.84 with MCC of 0.63. The results were shown in Table 6. For both datasets, the accuracy declined about 2% compared with the results obtained from original datasets, which may be caused by the change of dataset. However, it also achieved satisfying prediction performance. It could be believed that our method was effective for predicting protein stability changes upon single amino acid mutation.

Table 6. Results for NS1925 and NSR2760

Dataset	TPR	TNR	PPV	NPV	Accuracy	MCC	AUC
NS1925	0.74	0.91	0.79	0.89	0.86	0.67	0.90
NSR2760	0.73	0.89	0.75	0.88	0.84	0.63	0.88

4. Conclusions

In this study, the topological characteristics extracted from the protein structure network were introduced in predicting protein stability changes upon single-site mutations. These parameters exhibit satisfying predictive strength. It further indicates that the network parameters reflecting the importance of nodes in protein structure network are related with their importance in protein function and structure. Based on our study, eight sequence neighbors centered on the mutant site were enough for providing the mutant site's environment information in protein sequence. In addition, from the analysis of the mutant rules, some substitutions were observed to have the preference for protein stability changes. Such as the mutants of Y to A, I to F and etc are more likely to cause protein destabilization. Our result could be anticipated valuable for designing new proteins.

Acknowledgements

We would like to thank the anonymous reviewers for their patient review and constructive suggestions. This study was supported by the National Natural Science Foundation of China (21175095), Doctoral Fund of Ministry of Education of China (20120181110051).

References

Amitai, G., Shemesh, A., Sitbon, E., Shklar, M., Netanely, D., Venger, I., & Pietrokovski, S. (2004). Network analysis of protein structures identifies functional residues. *Journal of Molecular Biology, 344*(4), 1135-1146. http://dx.doi.org/ 10.1016/j.jmb.2004.10.055

Bagler, G., & Sinha, S. (2005). Network properties of protein structures. *Physica a-Statistical Mechanics and Its Applications, 346*(1-2), 27-33. http://dx.doi.org/ 10.1016/j.physa.2004.08.046

Bava, K. A., Gromiha, M. M., Uedaira, H., Kitajima, K., & Sarai, A. (2004). ProTherm, version 4.0: thermodynamic database for proteins and mutants. *Nucleic Acids Research, 32*, D120-D121.

http://dx.doi.org/ 10.1093/nar/gkh082

Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition, 30*(7), 1145-1159. http://dx.doi.org/ 10.1016/s0031-3203(96)00142-2

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32. http://dx.doi.org/ 10.1023/a:1010933404324

Brinda, K. V., & Vishveshwara, S. (2005). A network representation of protein structures: Implications for protein stability. *Biophysical Journal, 89*(6), 4159-4170. http://dx.doi.org/ 10.1529/biophysj.105.064485

Capriotti, E., Fariselli, P., & Casadio, R. (2004). A neural-network-based method for predicting protein stability changes upon single point mutations. *Bioinformatics (Oxford, England), 20 Suppl 1*, i63-68. http://dx.doi.org/ 10.1093/bioinformatics/bth928

Capriotti, E., Fariselli, P., & Casadio, R. (2005a). I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Research, 33*, W306-W310. http://dx.doi.org/ 10.1093/nar/gki375

Capriotti, E., Fariselli, P., Calabrese, R., & Casadio, R. (2005b). Predicting protein stability changes from sequences using support vector machines. *Bioinformatics (Oxford, England), 21 Suppl 2*, ii54-58. http://dx.doi.org/ 10.1093/bioinformatics/bti1109

Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., . . . Lander, E. S. (1999). Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genetics, 22*(3), 231-238.

Cheng, J. L., Randall, A., & Baldi, P. (2006). Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins-Structure Function and Bioinformatics, 62*(4), 1125-1132. http://dx.doi.org/ 10.1002/prot.20810

Chou, P. Y. & Fasman, G. D. (1978). *Adv. Enzym, 47*, 45-148

Collantes, E. R., & Dunn, W. J. (1995). Amino acid side chain descriptors for quantitative structure-activity relationship studies of peptide analogues. *Journal of Medicinal Chemistry, 38*(14), 2705-2713. http://dx.doi.org/ 10.1021/jm00014a022

Del Sol, A., & O'Meara, P. (2005). Small-world network approach to identify key residues in protein-protein interaction. *Proteins, 58*(3), 672-682. http://dx.doi.org/ 10.1002/prot.20348

Deleage, G., & Roux, B. (1987). An algorithm for protein secondary structure prediction based on class prediction. *Protein engineering, 1*(4), 289-294. http://dx.doi.org/ 10.1093/protein/1.4.289

Dokholyan, N. V., Li, L., Ding, F., & Shakhnovich, E. I. (2002). Topological determinants of protein folding. *Proceedings of the National Academy of Sciences of the United States of America, 99*(13), 8637-8641. http://dx.doi.org/ 10.1073/pnas.122076099

Folkman, Lukas, Stantic, Bela, & Sattar, Abdul. (2013). Sequence-only evolutionary and predicted structural features for the prediction of stability changes in protein mutants. *Bmc Bioinformatics, 14*. http://dx.doi.org/ 10.1186/1471-2105-14-s2-s6

Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R., Appel, R. D., & Bairoch, A. (2005). In J. M. Walker (Ed.), *The Proteomics Protocols Hadbook* (pp. 571-607). Humana Press.

Grantham, R. (1974). Amino acid difference formula to help explain protein evolution. *Science (New York, N.Y.), 185*(4154), 862-864. http://dx.doi.org/ 10.1126/science.185.4154.862

Greene, L. H., & Higman, V. A. (2003). Uncovering network systems within protein structures. *Journal of Molecular Biology, 334*(4), 781-791. http://dx.doi.org/ 10.1016/j.jmb.2003.08.061

Ihaka, R., & Gentleman, R. (1996). R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics, 5*(3), 299-314.

Kawashima, S., & Kanehisa, M. (2000). AAindex: Amino acid index database. *Nucleic Acids Research, 28*(1), 374-374. http://dx.doi.org/ 10.1093/nar/28.1.374

Kyte, J., & Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *Journal of molecular biology, 157*(1), 105-132. http://dx.doi.org/ 10.1016/0022-2836(82)90515-0

Li, Y. Z., Wen, Z. N., Xiao, J. M., Yin, H., Yu, L. Z., Yang, L., & Li, M. L. (2011). Predicting disease-associated substitution of a single amino acid by analyzing residue interactions. *Bmc Bioinformatics, 12*.

http://dx.doi.org/ 10.1186/1471-2105-12-14

Masso, M., & Vaisman, I. I. (2008). Accurate prediction of stability changes in protein mutants by combining machine learning with structure based computational mutagenesis. *Bioinformatics, 24*(18), 2002-2009. http://dx.doi.org/ 10.1093/bioinformatics/btn353

Newman, M. E. J. (2003). The structure and function of complex networks. *Siam Review, 45*(2), 167-256. http://dx.doi.org/ 10.1137/s003614450342480

Ng, P. C., & Henikoff, S. (2001). Predicting deleterious amino acid substitutions. *Genome Research, 11*(5), 863-874. http://dx.doi.org/ 10.1101/gr.176601

Noble, W. S. (2006). What is a support vector machine? *Nature Biotechnology, 24*(12), 1565-1567. http://dx.doi.org/ 10.1038/nbt1206-1565

Rose, G. D., Geselowitz, A. R., Lesser, G. J., Lee, R. H., & Zehfus, M. H. (1985). Hydrophobicity of amino acid residues in globular proteins. *Science (New York, N.Y.), 229*(4716), 834-838. http://dx.doi.org/ 10.1126/science.4023714

Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., & Serrano, L. (2005). The FoldX web server: an online force field. *Nucleic Acids Research, 33*, W382-W388. http://dx.doi.org/ 10.1093/nar/gki387

Shirley, B. A., Stanssens, P., Hahn, U., & Pace, C. N. (1992). Contribution of hydrogen bonding to the conformational stability of ribonuclease T1. *Biochemistry, 31*(3), 725-732. http://dx.doi.org/ 10.1021/bi00118a013

Teng, S. L., Srivastava, A. K., & Wang, L. J. (2010). Sequence feature-based prediction of protein stability changes upon amino acid substitutions. *Bmc Genomics, 11*. http://dx.doi.org/ 10.1186/1471-2164-11-s2-s5

Vendruscolo, M., Dokholyan, N. V., Paci, E., & Karplus, M. (2002). Small-world view of the amino acids that play a key role in protein folding. *Physical Review E, 65*(6). http://dx.doi.org/ 10.1103/PhysRevE.65.061910

Vihinen, M., Torkkila, E., & Riikonen, P. (1994). Accuracy of protein flexibility predictions. *Proteins, 19*(2), 141-149. http://dx.doi.org/10.1002/prot.340190207

Wang, Z., & Moult, J. (2001). SNPs, protein structure, and disease. *Human Mutation, 17*(4), 263-270. http://dx.doi.org/ 10.1002/humu.22

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature, 393*(6684), 440-442. http://dx.doi.org/ 10.1038/30918

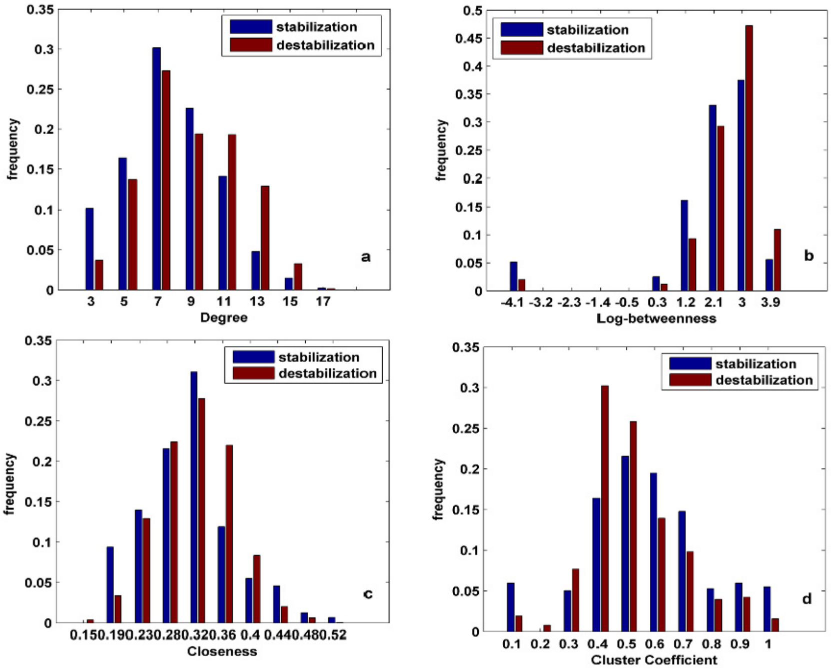
Yang, Y., Chen, B., Tan, G., Vihinen, M., & Shen, B. R. (2013). Structure-based prediction of the effects of a missense variant on protein stability. *Amino Acids, 44*(3), 847-855. http://dx.doi.org/ 10.1007/s00726-012-1407-7

Yue, P., & Moult, J. (2006). Identification and analysis of deleterious human SNPs. *Journal of Molecular Biology, 356*(5), 1263-1274. http://dx.doi.org/ 10.1016/j.jmb.2005.12.025

Zhao, Gang, & London, Erwin. (2006). An amino acid "transmembrane tendency" scale that approaches the theoretical limit to accuracy for prediction of transmembrane helices: Relationship to biological hydrophobicity. *Protein Science, 15*(8), 1987-2001. http://dx.doi.org/ 10.1110/ps.062286306

Zimmerman, J. M., Eliezer, N., & Simha, R. (1968). The characterization of amino acid sequences in proteins by statistical methods. *Journal of theoretical biology, 21*(2), 170-201. http://dx.doi.org/10.1016/0022-5193(68)90069-6

Supplementary Materials



Supplementary Figure 1. The frequency distributions of a) Degree; b) Betweenness; c) Closeness; d) Cluster Coefficient for stabilization and destabilization-associated mutants of SR2760

Supplementary Table 1. Frequency of occurrence of stabilizing mutants obtained from NSR2760 dataset

	A	R	N	D	C	Q	E	G	H	I
A	-	0(2)	1(3)	2(5)	1(3)	0(4)	1(5)	6(19)	0(4)	3(5)
R	3(15)	-	0(0)	0(0)	2(3)	1(6)	3(7)	2(4)	1(7)	0(0)
N	8(27)	0(0)	-	12(21)	0(0)	0(1)	3(3)	1(5)	1(4)	7(7)
D	15(37)	4(4)	19(44)	-	1(2)	3(5)	10(13)	6(12)	9(12)	3(4)
C	2(8)	0(0)	0(0)	0(0)	-	0(0)	0(0)	0(1)	0(0)	0(1)
Q	7(18)	1(1)	2(3)	0(0)	1(1)	-	3(6)	2(8)	0(1)	1(1)
E	9(30)	2(3)	1(7)	1(7)	2(4)	9(25)	-	3(6)	3(4)	3(3)
G	17(44)	2(6)	0(2)	0(5)	0(1)	1(3)	2(2)	-	1(4)	0(0)
H	3(9)	1(4)	0(3)	1(4)	1(1)	3(6)	1(2)	4(8)	-	0(0)
I	1(51)	0(0)	2(2)	0(2)	0(2)	0(0)	0(5)	1(10)	0(1)	-
L	**8(70)**	2(4)	0(1)	0(3)	0(3)	0(1)	0(2)	0(7)	1(2)	1(10)
K	11(29)	5(9)	6(7)	0(2)	0(0)	1(8)	6(19)	3(10)	0(5)	1(2)
M	3(15)	1(1)	0(0)	3(3)	1(1)	0(0)	3(3)	0(5)	0(0)	3(10)
F	2(24)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(2)	0(1)	0(2)
P	3(25)	0(1)	0(1)	0(0)	0(0)	0(0)	0(0)	0(9)	0(0)	0(0)
S	12(35)	2(2)	2(4)	5(8)	0(2)	1(2)	3(3)	2(6)	1(5)	2(2)
T	3(27)	3(6)	3(11)	3(9)	1(5)	1(4)	4(14)	0(12)	1(5)	7(16)
W	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(2)	0(0)
Y	**0(11)**	0(1)	0(2)	1(2)	2(3)	1(2)	0(0)	0(6)	1(1)	0(0)
V	12(86)	1(4)	1(6)	2(4)	0(9)	0(1)	1(5)	**2(21)**	0(7)	17(35)

	L	K	M	F	P	S	T	W	Y	V
A	2(8)	1(6)	2(6)	1(3)	13(24)	**3(31)**	4(11)	0(2)	0(3)	11(18)
R	0(1)	1(8)	1(4)	0(0)	0(0)	0(3)	0(0)	0(0)	0(0)	0(0)
N	1(1)	2(5)	3(3)	1(1)	0(0)	3(7)	3(4)	0(0)	0(0)	2(2)
D	3(4)	6(11)	1(2)	3(5)	0(5)	7(10)	3(4)	1(2)	2(2)	2(3)
C	0(1)	0(0)	0(1)	0(0)	0(0)	2(11)	4(5)	0(0)	0(0)	0(5)
Q	4(5)	5(5)	0(1)	0(0)	0(4)	0(2)	0(1)	0(0)	0(0)	0(1)
E	5(7)	12(33)	3(4)	2(4)	3(4)	3(6)	1(5)	1(3)	2(3)	7(8)
G	0(1)	0(1)	0(0)	0(1)	1(3)	1(9)	0(1)	0(1)	0(0)	4(8)
H	5(5)	0(1)	0(0)	0(0)	7(7)	1(2)	0(2)	0(1)	7(8)	0(0)
I	11(22)	0(0)	**1(16)**	**0(11)**	0(3)	0(3)	2(16)	0(2)	0(3)	**8(58)**
L	-	0(1)	2(13)	2(9)	1(2)	0(1)	0(6)	1(1)	1(1)	**1(16)**
K	0(0)	-	3(5)	2(2)	2(3)	0(1)	0(0)	0(1)	1(1)	3(3)
M	4(23)	3(8)	-	2(3)	0(0)	0(0)	0(2)	0(0)	0(1)	1(8)
F	3(10)	0(1)	0(4)	-	0(0)	0(0)	0(1)	5(6)	5(8)	0(5)
P	3(4)	0(0)	0(0)	0(0)	-	1(5)	0(0)	0(1)	0(1)	0(1)
S	1(2)	2(2)	1(1)	4(4)	1(2)	-	4(7)	1(1)	1(2)	3(5)
T	1(6)	1(1)	1(1)	3(4)	0(3)	8(19)	-	1(1)	3(3)	11(27)
W	0(4)	0(0)	0(0)	0(15)	0(0)	0(0)	0(0)	-	0(7)	0(0)
Y	1(2)	1(1)	0(0)	14(35)	0(1)	0(3)	0(0)	2(5)	-	1(1)
V	11(21)	0(4)	4(13)	**0(10)**	2(4)	0(6)	**0(33)**	0(0)	0(8)	-

(1) Mutations are from columns of residue to rows of residue.

(2) The number of the stabilizing mutants is given outside the parentheses with the total number of the corresponding mutants in parentheses.

Copyrights