BMC Bioinformatics

Research article

Searching for interpretable rules for disease mutations: a simulated annealing bump hunting strategy

Rui Jiang[†], Hua Yang[†], Fengzhu Sun and Ting Chen*

Address: Molecular and Computational Biology, University of Southern California. MCB201, 1050 Childs way, Los Angeles, CA 90089–2910, USA

Email: Rui Jiang - ruijiang@usc.edu; Hua Yang - huayang@usc.edu; Fengzhu Sun - fsun@usc.edu; Ting Chen* - tingchen@usc.edu

* Corresponding author   †Equal contributors

Published: 19 September 2006

BMC Bioinformatics 2006, **7**:417   doi:10.1186/1471-2105-7-417

Received: 29 June 2006
Accepted: 19 September 2006

This article is available from: http://www.biomedcentral.com/1471-2105/7/417

Abstract

Background: Understanding how amino acid substitutions affect protein functions is critical for the study of proteins and their implications in diseases. Although methods have been developed for predicting potential effects of amino acid substitutions using sequence, three-dimensional structural, and evolutionary properties of proteins, the applications are limited by the complication of the features and the availability of protein structural information. Another limitation is that the prediction results are hard to be interpreted with physicochemical principles and biological knowledge.

Results: To overcome these limitations, we proposed a novel feature set using physicochemical properties of amino acids, evolutionary profiles of proteins, and protein sequence information. We applied the support vector machine and the random forest with the feature set to experimental amino acid substitutions occurring in the *E. coli* lac repressor and the bacteriophage T4 lysozyme, as well as to annotated amino acid substitutions occurring in a wide range of human proteins. The results showed that the proposed feature set was superior to the existing ones. To explore physicochemical principles behind amino acid substitutions, we designed a simulated annealing bump hunting strategy to automatically extract interpretable rules for amino acid substitutions. We applied the strategy to annotated human amino acid substitutions and successfully extracted several rules which were either consistent with current biological knowledge or providing new insights for the understanding of amino acid substitutions. When applied to unclassified data, these rules could cover a large portion of samples, and most of the covered samples showed good agreement with predictions made by either the support vector machine or the random forest.

Conclusion: The prediction methods using the proposed feature set can achieve larger AUC (the area under the ROC curve), smaller BER (the balanced error rate), and larger MCC (the Matthews' correlation coefficient) than those using the published feature sets, suggesting that our feature set is superior to the existing ones. The rules extracted by the simulated annealing bump hunting strategy have comparable coverage and accuracy but much better interpretability as those extracted by the patient rule induction method (PRIM), revealing that the strategy is more effective in inducing interpretable rules.

Background

Variants in single bases of DNA sequences yield single nucleotide polymorphisms (SNPs), among which non-synonymous single nucleotide polymorphisms (nsSNPs) occurring in protein coding regions lead to amino acid substitutions in protein products, potentially affect protein functions, and are closely related to human inherited diseases. Hence, predicting potential effects of non-synonymous single nucleotide polymorphisms and their resulting amino acid substitutions on protein functions is of central importance in modern pathological and pharmaceutical studies [1]. Recently, increasing amounts of amino acid substitutions occurring in human proteins have been detected and collected in various databases such as the *Swiss-Prot* database [2], the *Human Gene Mutation Database* (HGMD) [3], and the *Online Mendelian Inheritance in Man*$^+$ (OMIM) database [4]. Stand alone data sets such as the unbiased laboratory mutagenesis data derived from experiments on the *E. coli* lac repressor [5,6] and the bacteriophage T4 lysozyme [7] are also available. With these data sources, the prediction is typically based on a set of features derived from the sequence and structural properties, as well as the phylogenetic information of the proteins containing the substitutions. For instance, Chasman and Adams derived sequence and structure-based features from a structural model and the phylogenetic information [8]. Sunyaev et al. analyzed amino acid substitutions on the basis of protein three-dimensional structure and multiple alignments of homologous sequences [9,10]. Ferrer-Costa et al. characterized disease-associated substitutions in terms of substitution matrix, secondary structure, accessibility, free energies of transfer from water to octanol, etc. [11,12]. Saunders and Baker created mutation models by means of a variety of structural and evolutionary features [13]. Krishnan and Westhead used the physicochemical classes of residues, sequence conservation score, secondary structure, solvent accessibility, and buried charge, etc. [14]. Ng and Henikoff utilized the sequence conservation and the BLOSUM amino acid substitution matrices [15]. With a set of features ready, the prediction is conventionally performed by making use of either the standard machine learning methods such as the decision tree [16], the support vector machines [17,18], the random forest [19], the statistical and classification models [8,9,13], or certain specifically designed methods such as the SIFT (Sorting Intolerant From Tolerant amino acid substitutions) [15].

No matter what kind of method is used, the quality of the features plays an important role in predicting the potential effects of given amino acid substitutions. To construct these features, amino acid substitutions were mapped to protein 3D structures [8-10,13]; evolutionary properties were measured from statistical models [8,9]; secondary structure and accessibility were computed from various prediction programs [11,13]; database annotations were also included [12]. However, the availability of protein or homologous proteins' structures limits the scope of the applications of these methods. In addition, most of these prediction methods are complicated and the prediction results are difficult to interpret, because a large number of complicated features are used and many of them rely heavily on other computational models. Although in some methods simple features were used with some specifically designed statistical models [15], the prediction accuracy is not as high as those methods using combined multiple features [1,8-14].

A good feature set should contain as few features as possible, while each feature should have clear physicochemical meaning and is easy to be interpreted in biological terms. To achieve these objectives, we derived a novel feature set (including a continuous form and a discrete form) based on three physicochemical properties (*molecular weight, pI value*, and *hydrophobicity scale*) of amino acids, three relative frequencies of occurrences of amino acids in the secondary structures (*helices, strands*, and *turns*) of proteins with known secondary structural information, and the evolutionary profile of the protein containing the substitution. We compared the quality of the proposed feature set with other more complicated ones by applying the decision tree [16], the support vector machine [17,18], and the random forest [19] to the experimental amino acid substitution data of the *E. coli* lac repressor [5,6] and the bacteriophage T4 lysozyme [7], as well as to a wide range of human amino acid substitutions collected in the Swiss-Prot database. The results showed that our simple yet interpretable feature set was superior to other published ones [15,14,20] in terms of the area under the receiver operating characteristic (ROC) curve (AUC), the balanced error rate (BER), and the Matthews' correlation coefficient (MCC).

Although existing machine learning methods could make predictions, they acted like "black boxes" in that they were not capable of capturing physicochemical principles behind the predictions. In many circumstances, however, these hidden principles were of more importance since they could reveal how amino acid substitutions affected protein functions and why some amino acid substitutions would result in diseases. In order to explore these principles and associate amino acid substitutions with biological knowledge, we would use rule induction methods to automatically search rules for amino acid substitutions. These rules should be (1) *interpretable*, consisting of a small set of simple features; (2) *high-quality*, with very few exceptions; and (3) *general*, capable of explaining a significant number of substitutions.

In this paper, we considered rules as sub-regions (boxes) in the feature space composed of amino acid substitutions. More specifically, the boxes were defined in terms of the feature intervals. A previous method for finding boxes in the feature space was the patient rule induction method (PRIM) [21], which searched for optimal boxes using a steepest-ascent approach and was intuitively referred to as a "bump hunting" method. When applied to our problem, the PRIM had drawbacks in that the imbalance between the numbers of data samples in different categories was not considered, and some redundant features in the boxes should be manually removed and the quality of the resulting boxes significantly decreased. To overcome these shortcomings, we incorporated a new criterion called the discrimination power to take the imbalance between the numbers of data samples in different categories into consideration, and developed a novel simulated annealing bump hunting strategy which made use of the simulated annealing method to automatically discard redundant features while extracting high quality rules. We validated this strategy using heterogenous experimental amino acid substitutions occurring in both the *E. coli* lac repressor [5,6] and the bacteriophage T4 lysozyme [7], and showed that our approach could extract rules with comparable converge and accuracy but much better interpretability as those extracted by the original PRIM method. We then applied our strategy to annotated human amino acid substitutions collected in the Swiss-Prot database and successfully identified several rules which could be interpreted using physicochemical terms and were consistent with the current biological knowledge. We further applied the induced intolerant rules to unclassified human amino acid substitution data, and the results showed that these rules could cover a large portion of data samples and most of the covered samples showed good agreement with predictions made by either the support vector machine or the random forest. Beyond the highly confident predictions, these rules more importantly revealed the physicochemical principles behind the covered amino acid substitutions and explained why these substitutions would result in diseases.

Results

Data sources

A large number of amino acid substitutions occurring in human proteins have been collected in the Swiss-Prot protein database [2]. In version 50.0 (released on May-30-2006), the Swiss-Prot database contained 25,994 amino acid substitution entries in 4,324 human proteins, with each substitution being annotated as "Disease", "Polymorphism", or "Unclassified". For a clear and concise presentation, we would refer to amino acid substitutions with the annotation "Disease" as intolerant ones and those with the annotation "Polymorphism" as tolerant ones. In this paper, we studied human proteins having at least 20 homologous proteins in the Pfam database [22] (version 20.0, released in May-2006), and focused on the substitutions occurring in known protein domains. In total, we collected 9, 610 intolerant substitutions, 4, 556 tolerant substitutions, and 1,487 unclassified ones in 2, 579 human proteins.

In order to validate the proposed feature set and the simulated annealing bump hunting strategy, two sets of experimental amino acid substitution data for the *E. coli* lac repressor [5,6] and the bacteriophage T4 lysozyme [7] were used. In these data sets, the effects of amino acid substitutions on the function of the corresponding protein were rated and classified to four categories. In the case of the lac repressor, the four categories were "+" (no effect), "+-" (slight effect), "-+" (larger effect), and "-" (complete absence). In the case of the T4 lysozyme, the four categories were "++" (no effect), "+" (slight effect), "+/-" (larger effect), and "-" (complete absence). Following the definition used by Chasman and Adams [8], as well as by Krishnan and Westhead [14], substitutions falling into the "no effect" category were treated as tolerant ones, and substitutions in the other categories were regarded as intolerant ones. In total, for the *E. coli* lac repressor, we collected 1,187 intolerant substitutions and 1,760 tolerant ones. For the T4 lysozyme, we collected 494 intolerant substitutions and 1,048 tolerant ones.

Prediction of the experimental amino acid substitutions

We first show that the proposed feature set can outperform other published feature sets in the prediction of potential effects of experimental amino acid substitutions occurring in the *E. coli* lac repressor [5,6] and the T4 lysozyme [7]. We performed 10-fold cross-validation experiments using both the support vector machine and the random forest with the proposed feature set on the substitution samples, calculated the area under the ROC curve (AUC), the minimum balanced error rate (BER), and the maximum Matthews' correlation coefficient (MCC), and compared them with other published results. Detailed descriptions regarding the prediction methods and the definition of the criteria are presented in the method section.

The cross-validation results using our feature set (both the continuous form and the discrete form) are shown in Table 1. First, we can see from the table that for the experimental substitutions occurring in homogenous proteins, our feature set works well with both the support vector machine and the random forest. When working with the random forest, the discrete form of our feature set can produce an AUC of 0.944, a BER of 0.125, and a MCC of 0.741 for the experimental substitutions occurring in the *E. coli* lac repressor, suggesting that about 88% of the substitutions can be predicted accurately. When working with

Table 1: Results for predicting potential effects of experimental amino acid substitutions occurring in the *E. coli lac* repressor, the bacteriophage T4 lysozyme, and the mixed samples.

		Support vector machine			Random forest		
		AUC	BER	MCC	AUC	BER	MCC
Continuous	Lac repressor	0.912	0.152	0.694	0.939	0.143	0.723
	T4 lysozyme	0.897	0.177	0.614	0.907	0.167	0.640
	Mixed	0.889	0.185	0.612	0.921	0.158	0.678
Discrete	Lac repressor	0.905	0.170	0.654	**0.944**	**0.125**	**0.741**
	T4 lysozyme	0.878	0.199	0.588	**0.911**	**0.167**	**0.651**
	Mixed	0.887	0.187	0.622	**0.927**	**0.148**	**0.693**

the support vector machine, the results are slightly worse, but the continuous form of our feature set can still predict about 85% substitutions accurately. For experimental substitutions occurring in the T4 lysozyme, we obtain similar results. Second, the results show that our feature set can also work well for experimental substitutions occurring in heterogenous proteins. When applied to the mixed samples occurring in both the *E. coli* lac repressor and the T4 lysozyme, the random forest with the discrete form of our feature set can produce an AUC of 0.927, a BER of 0.148, and a MCC of 0.693, suggesting that about 85% of the substitutions can be predicted accurately. Thirdly, we notice that in our studies, the random forest works slightly better than the support vector machine with our feature set in terms of the AUC, the BER, and the MCC.

We compared the cross-validation results using our feature set with those obtained by the SIFT (Ng and Henikoff [15]) and another published feature set (Krishnan and Westhead [14]). As a sequence homology-based method, the SIFT can achieve BERs of 33% and 34% for experimental amino acid substitutions occurring in the *E. coli* lac repressor and the T4 lysozyme, respectively. By comparison, the continuous form of our feature set can achieve corresponding BERs of 14% and 17% when working with the random forest (15% and 18% when working with the support vector machine), respectively. These results suggest that our feature set can outperform the SIFT in the prediction of potential effects of experimental amino acid substitutions occurring in homogenous proteins. The published feature set by Krishnan and Westhead [14] uses 16 features, including 13 sequence based ones (the residue identities of the original and mutated residue, the physicochemical classes of these residues (hydrophobic, polar, charged, glycine), sequence conservation score at the mutated position, molecular mass shift on mutation, and hydrophobicity difference), and 3 structure based ones (secondary structure, solvent accessibility, and buried charge). When working with the support vector machine, this feature set can achieve BERs of 27%, 29%, and 28% for experimental amino acid substitutions occurring in the *E. coli* lac repressor, the T4 lysozyme, and the mixture of them, respectively, while the continuous form of our feature set can achieve corresponding BERs of 15%, 18%, and 19%, respectively. When working with the decision tree, the published feature set can achieve BERs of 16%, 20%, and 21% for experimental amino acid substitutions occurring in the *E. coli* lac repressor, the T4 lysozyme, and the mixture of them, respectively, while the continuous form of our feature set can achieve corresponding BERs of 16%, 18%, and 19%, respectively. These results suggest that our feature set can work as good as or outperform the published feature set [14] in the prediction of potential effects of experimental amino acid substitutions occurring in both homogenous and heterogenous proteins.

Prediction of the disease related amino acid substitutions

We performed 10-fold cross-validation experiments using both the support vector machine and the random forest with the proposed feature set on amino acid substitutions occurring in highly heterogenous human proteins and collected in the Swiss-Prot database, and compared the results with other published results (Bao and Cui [20]).

The published method [20] used a complicated feature set. For every substitution pair, they directly used two three-dimensional structural information predicted by the ENVIRONMENT program [23], one secondary structural information predicted by the STRIDE program [24], and one statistical score calculated by the SIFT program [15]. Their feature set also included another feature derived from the prediction results of these programs, and the wild-type amino acid identity. Altogether, they used six features. Five of them were three-dimensional structural or statistical ones, and needed to be calculated using other programs. Due to the limited availability of three-dimensional structural information, only a small fraction of available substitutions (3, 686 intolerant ones in 323 proteins and 532 tolerant ones in 305 proteins) in the Swiss-Prot database could be considered in their method. In

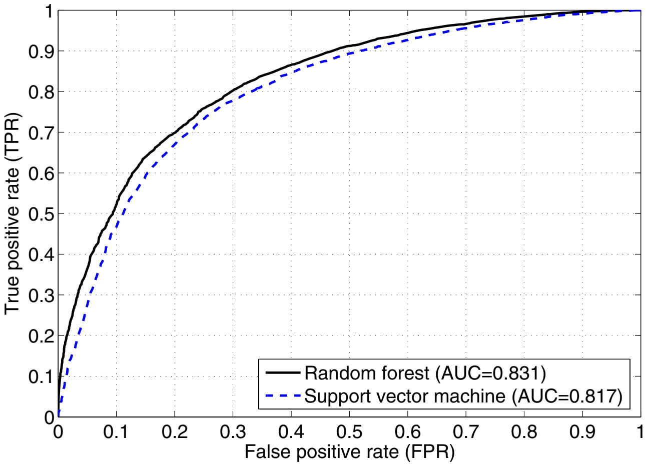


Figure 1
The ROC curves for predicting amino acid substitutions occurring in human proteins using the support vector machine and the random forest with the continuous form of the proposed feature set. For those using the discrete form, the curves (not shown) are similar.

contrast, our proposed feature set used only sequence information and evolutionary profiles, and did not depend on any other prediction programs. Consequently, we could predict more substitutions (9, 610 intolerant ones and 4, 556 tolerant ones) in a wider range of (2, 579) human proteins.

For comprehensive measures, Figure 1 shows the ROC curves for the support vector machine (AUC = 0.817) and the random forest (AUC = 0.831) using the proposed continuous form of our feature set. When compared with the SIFT and the method used by Bao and Cui (both presented in [20]), we can see clearly that both methods using our feature set produce better ROC curves (see Figure 1 and Fig.1 in [20]), indicating that the proposed feature set is superior to both the SIFT and the feature set presented in [20] in terms of comprehensive prediction power (the area under the ROC curve). For the discrete form, the AUC is 0.806 for the support vector machine and 0.817 for the random forest, and the ROC curves (not shown) are similar to those using the continuous form.

More specifically, we compared the two criteria for a certain single decision threshold, as shown in Table 2. When working with the support vector machine, the continuous form of our feature set leads the SIFT (results presented in [20]) by about 4% (26% vs. 30%) in BER and about 0.15 (0.46 vs. 0.31) in MCC. When working with the random forest, the continuous form of our feature set leads the SIFT by about 5% (25% vs. 30%) in BER and about 0.18 (0.49 vs. 0.31) in MCC. Similar results are obtained when comparing the discrete form of our feature set with the SIFT. These results suggest that our feature set can outperform the SIFT in the prediction of amino acid substitutions occurring in human proteins. When comparing our results with those obtained using the feature set proposed in [20], we can see from the table that for both prediction methods using the proposed feature set, the BERs are much smaller while the MCCs are much larger than the corresponding method using the feature set presented in [20], indicating that our feature set are much better than the published one.

Correlation and relative importance of the proposed features

For better understanding of the relationship between the proposed features, we calculated the pairwise Pearson's correlation coefficients between the proposed features (continuous form) based on the amino acid substitutions occurring in human proteins and presented the (upper triangle) correlation matrix in Figure 2. We divided the features to 7 groups according to their definitions in the method section, and named these groups at the top of the matrix. First, we can see from the matrix that the two evolutionary conservation scores (features 43 and 44) have very weak correlations with other 42 features. Second, for the original amino acid group (features 1 to 6), the window-sized group (features 13 to 18), and the column-weighted group (features 19 to 24), features derived from the same (physicochemical or relative frequency) properties (e.g., 1-13-19, 2-14-20, etc.) show medium positive correlations, as illustrated in region 1, 2, and 3 in the matrix. Third, the relative change features (features 25 to 42) show strong positive correlations with the substitution features derived from the same properties (e.g., 25–7, 26–8, etc.) and strong negative correlations with the orig-

Table 2: Results for predicting potential effects of annotated amino acid substitutions occurring in human proteins and collected in the Swiss-Prot database.

	Support vector machine			Random forest		
	AUC	BER	MCC	AUC	BER	MCC
Bao & Cui [20]	N/A	0.318	0.274	N/A	0.292	0.315
Continuous form	0.817	0.258	0.463	**0.831**	**0.245**	**0.491**
Discrete form	0.806	0.259	0.457	0.817	0.262	0.451

inal, the window-sized, or the column-weighted features (e.g., 25–1, 37–19, etc.), as illustrated in region 4, 5, and 6, respectively. Finally, as shown in region 7 in the matrix, the relative change features derived from the same properties show strong positive correlations (e.g., 25-31-37, 26-32-38, etc.). We also calculated the correlation matrix for the discrete form of the proposed features based on the amino acid substitutions occurring in human proteins and observed similar results (data not shown). These observations, though can be intuitively explained from the definitions and calculation schemes of the features (see the method section for details), provide us informative understanding and quantitative measurement of the relationship between the proposed features and can be used as evidences in the future feature selection procedure.

We then evaluated the relative importance of the proposed features using the scheme included in the random forest and presented the results in Figure 3. In the random forest, the raw importance of a feature is calculated by randomly permuting the values of the feature in the Out-Of-Bag (OOB) cases, calculating the difference of classification errors between the original and the permuted cases, and averaging this difference over all the trees in the forest [19]. To make the measurement of importance more understandable, a normalization procedure is further applied to the raw importance of each feature by dividing the raw importance with the maximum raw importance over all the features (assuming it to be a positive number). Consequently, the relative importance of every proposed feature (a real number which is less than or equal to 1.0) is obtained. For the continuous form of the proposed features (Figure 3A), we can see that the two evolutionary conservation scores (features 43 and 44)

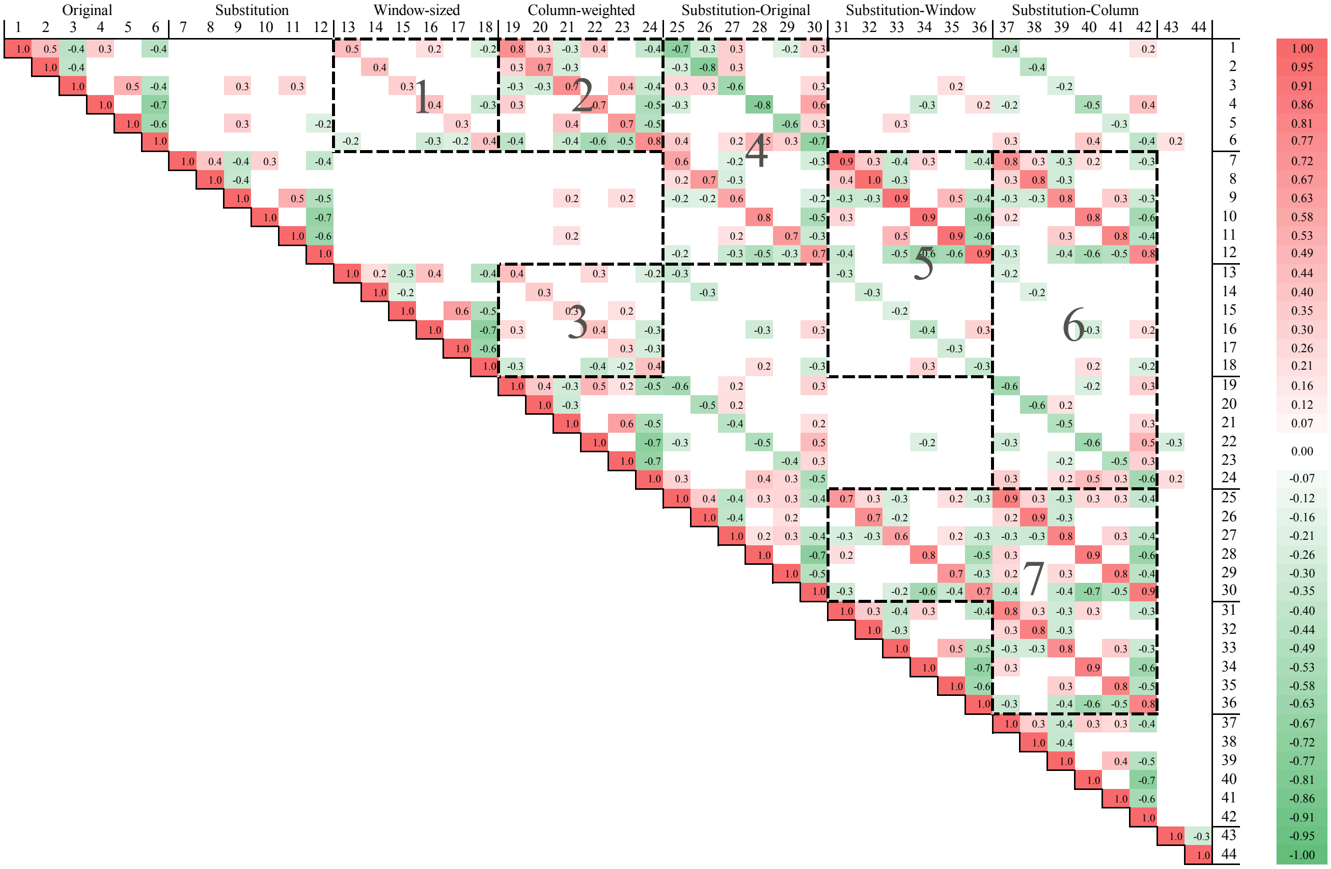


Figure 2
The Pearson correlation coefficient matrix (upper triangle) of the proposed features (continuous form). The Pearson correlation coefficients are calculated based on the amino acid substitutions occurring in human proteins and collected in the Swiss-Prot database. For a clear and concise presentation, correlation coefficients with absolute values less than 0.2 are ignored in the figure. For the discrete form of the proposed features, the correlation coefficient matrix (not shown) is similar.

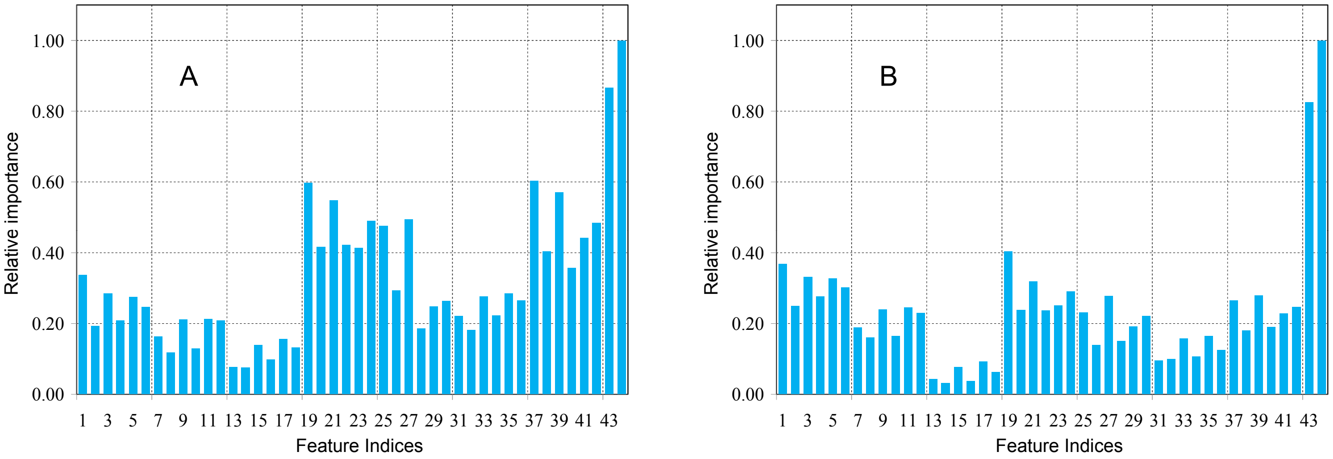


Figure 3

The relative importance of the proposed features. (A) the continuous form. (B) the discrete form. The raw importance for the features are calculated by the random forest [19]. The normalization is performed by dividing the raw importance with the maximum raw importance over all the features.

are of the most importance. The column-weighted group (features 19–24) and the substitution-column group (features 37–42) have similar importance and follow the evolutionary score group. For other groups of features, the order of importance is the substitution-original group (features 25–30) > the original group (features 1–6) > the substitution-window group (features 31–36) > the substitution group (features 7–12) > the window-sized group (features 13–18). For individual features, the first 10 most important features are ordered as $X_{44} > X_{43} > X_{37} > X_{19} > X_{39} > X_{21} > X_{27} > X_{24} > X_{42} > X_{25}$. On the one hand, all of the 10 features except for X_{27} in this order are calculated with the evolutionary conservation scores (see the method section for details), revealing the significant importance of the evolutionary information in the prediction of potential effects of amino acid substitutions. On the other hand, the frequent appearances of the features derived from the molecular weight (X_{19}, X_{25}, and X_{37}), the hydrophobicity scale (X_{21}, X_{27}, and X_{39}), and the relative frequency in turns (X_{24} and X_{42}) in this order suggest the importance of these properties in the identification of human disease related amino acid substitutions. For the discrete form of the proposed features (Figure 3B), the results show that the relative importance of the window-sized, the column-weighted, and the relative change features are not as good as their continuous forms, suggesting that the discretization procedure causes information loss for individual features. When looking at the order of the top 10 most informative features ($X_{44} > X_{43} > X_{19} > X_1 > X_3 > X_5 > X_{21} > X_6 > X_{24} > X_{39}$), we confirm the importance of the evolu-

Table 3: Results for validating the simulated annealing bump hunting strategy using the mixed experimental amino acid substitutions occurring in the *E. coli lac* repressor and the bacteriophage T4 lysozyme.

		Training set			Test set		
	Number of features	Box mean	Box size	Discrimination power	Box mean	Box size	Discrimination power
Intolerant	1	0.933	0.033	23.374	0.938	0.021	25.057
	2	0.934	0.020	23.794	0.917	0.015	18.383
	3	0.926	0.040	20.782	0.900	0.020	15.034
	4	0.953	0.021	33.947	0.955	0.015	35.042
	5	0.947	0.041	29.729	0.936	0.021	24.228
Tolerant	1	0.917	0.056	6.588	0.933	0.120	8.377
	2	0.955	0.052	12.734	0.944	0.048	10.168
	3	0.919	0.054	6.820	0.947	0.050	10.633
	4	0.983	0.060	35.248	0.927	0.064	7.613
	5	0.961	0.051	14.673	0.936	0.052	8.741

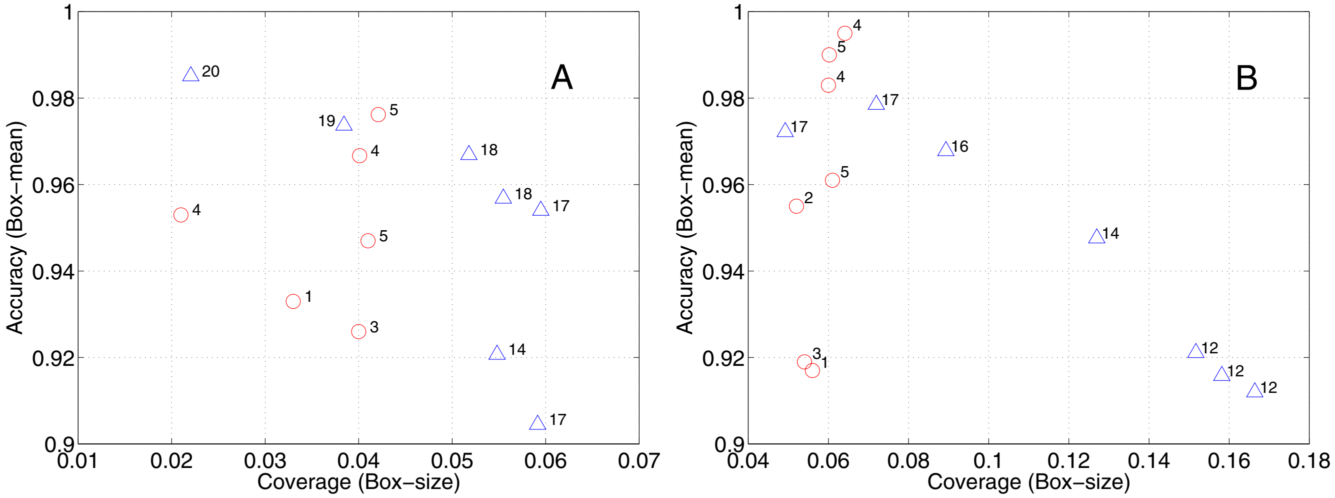


Figure 4
Comparison of the rules (boxes) extracted by the simulated annealing bump hunting strategy and the original bump hunting method (the SuperGEM software). (A) intolerant rules. (B) tolerant rules. Red circles are rules extracted by the proposed simulated annealing bump hunting strategy and green triangles are those extracted by the SuperGEM. The x-axes denote the coverage of the extracted rules, and the y-axes denote the accuracy of the extracted rules. The numbers beside the circles and the triangles denote the number of features included in the corresponding rules.

tionary information (X_{43} and X_{44}), the molecular weight (X_1 and X_{19}), the hydrophobicity scale (X_3, X_{21}, and X_{39}), and the relative frequency in turns (X_6 and X_{24}).

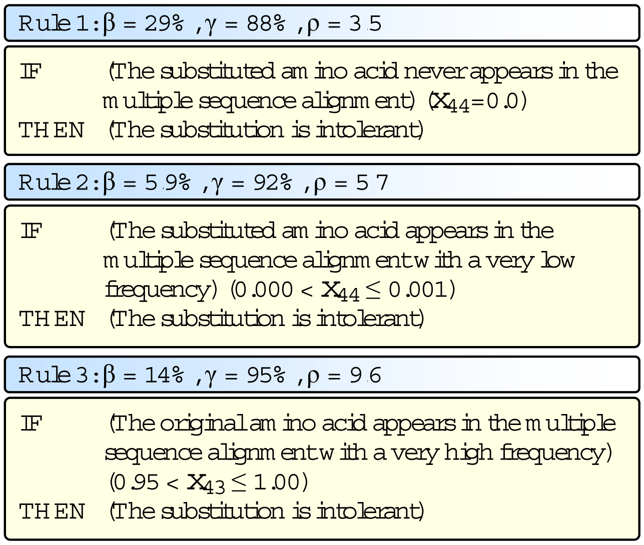
Validation of the simulated annealing bump hunting strategy

A merit property of the discrete form of our feature set is that every feature has strong physico-chemical meaning, which enables us to induce interpretable rules to explain the biological principles behind amino acid substitutions. We first validated the proposed simulated annealing bump hunting strategy using the heterogenous experimental amino acid substitution data. We randomly divided the mixed substitution samples occurring in the *E. coli* lac repressor and the T4 lysozyme into a training set (containing 2/3 of the data) and a test set (containing the rest 1/3 of the data), applied the simulated annealing bump hunting strategy to the training set, and evaluated the resulting rules on the test set. As an example, Table 3 lists ten rules (five intolerant ones and five tolerant ones, respectively) extracted by the simulated annealing bump hunting strategy. From the table, we can see that the extracted rules have comparable coverage (box-size), accuracy (box-mean), and discrimination power for the training and test set, suggesting that our strategy is capable of extracting general rules. For example, for intolerant rules, the simulated annealing bump hunting strategy extracted a 1-feature rule with a coverage of 0.033 and an accuracy of 0.933 from the training set, while the same rule have a coverage of 0.021 and an accuracy of 0.938 when evaluated using the test set.

We also made a comparison between the simulated annealing bump hunting strategy and the original patient rule induction method (PRIM), which was implemented in the SuperGEM software [21]. Some candidate rules (7 intolerant ones and 7 tolerant ones) are shown in Figure 4. From the figure, we can see that the rules extracted by the simulated annealing bump hunting strategy have comparable coverage and accuracy but much better interpretability (less number of features) as the rules extracted by the original PRIM method. For example, for intolerant rules (Figure 4A), the simulated annealing bump hunting strategy extracted a 5-feature rule with a coverage of 0.042 and an accuracy of 0.98, while the original PRIM method extracted a 19 feature rule with comparable coverage and accuracy. Similarly, for tolerant rules (Figure 4B), the simulated annealing bump hunting strategy extracted a 4-feature rule with a coverage of 0.06 and an accuracy of 0.98, while the original PRIM method extracted a 17 feature rule with comparable coverage and accuracy.

Amino acid substitution rules

We applied the simulated annealing bump hunting strategy with the discrete form of our feature set to the human amino acid substitution data and extracted several rules which were consistent with current biological knowledge. As examples, Figure 5 presents a group of three intolerant



Rule 1: β = 29% , γ = 88% , ρ = 3.5

IF (The substituted amino acid never appears in the multiple sequence alignment) (X_{44}=0.0)
THEN (The substitution is intolerant)

Rule 2: β = 5.9% , γ = 92% , ρ = 5.7

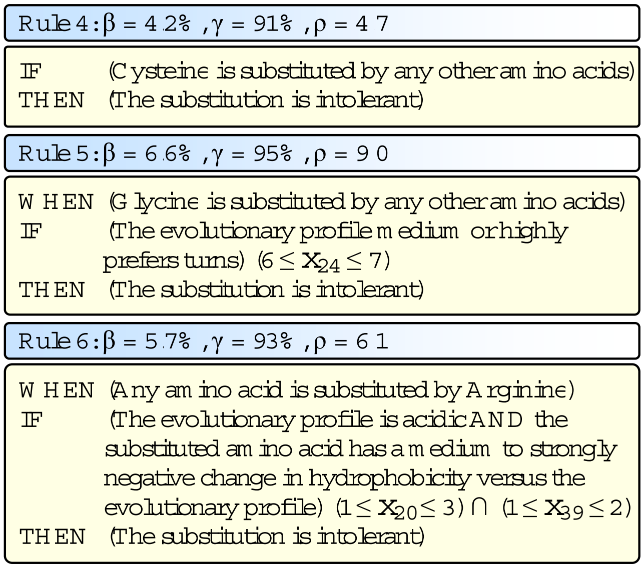
IF (The substituted amino acid appears in the multiple sequence alignment with a very low frequency) ($0.000 < X_{44} \le 0.001$)
THEN (The substitution is intolerant)

Rule 3: β = 14% , γ = 95% , ρ = 9.6

IF (The original amino acid appears in the multiple sequence alignment with a very high frequency) ($0.95 < X_{43} \le 1.00$)
THEN (The substitution is intolerant)

Figure 5
Three general intolerant rules. In the figure, β, γ, and ρ for each rule are the coverage, the accuracy, and the discrimination power of the corresponding rule, respectively.

rules which uses conservation scores and provides us general understanding regarding the intolerant substitutions. Detailed descriptions regarding the notations and definitions of the features are presented in the method section.



Rule 4: β = 4.2% , γ = 91% , ρ = 4.7

IF (Cysteine is substituted by any other amino acids)
THEN (The substitution is intolerant)

Rule 5: β = 6.6% , γ = 95% , ρ = 9.0

WHEN (Glycine is substituted by any other amino acids)
IF (The evolutionary profile medium or highly prefers turns) ($6 \le X_{24} \le 7$)
THEN (The substitution is intolerant)

Rule 6: β = 5.7% , γ = 93% , ρ = 6.1

WHEN (Any amino acid is substituted by Arginine)
IF (The evolutionary profile is acidic AND the substituted amino acid has a medium to strongly negative change in hydrophobicity versus the evolutionary profile) ($1 \le X_{20} \le 3$) $\cap$ ($1 \le X_{39} \le 2$)
THEN (The substitution is intolerant)

Figure 6
Three intolerant rules for individual amino acids. In the figure, β, γ, and ρ for each rule are the coverage, the accuracy, and the discrimination power of the corresponding rule, respectively.

Rule 1 in Figure 5 says that for a substitution pair, if the substituted amino acid never appears in the column (corresponding to the substitution position) of the Pfam multiple sequence alignment, the substitution is very likely to be intolerant. This rule uses a single feature (X_{44}) and covers 29% (4,137 out of 14,166) data samples with an accuracy of 88% and a discrimination power of 3.5.

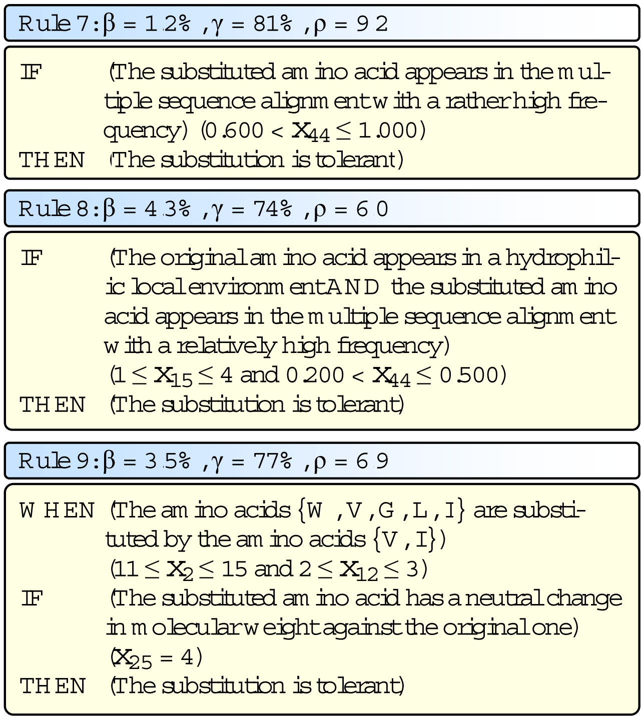
Rule 2 in Figure 5 says that for a substitution pair, if the substituted amino acid rarely (e.g., with a very low frequency ≤ 0.1%) appears in the corresponding column of the Pfam multiple sequence alignment, the substitution is very likely to be intolerant. This rule uses a single feature (X_{44}), covering 5.9% (836) data samples with an accuracy of 92% and a discrimination power of 5.7.

Rule 3 in Figure 5 says that for a substitution pair, if the original amino acid very abundantly (with a very high frequency ≥ 0.95) appears in the corresponding column of the Pfam multiple sequence alignment, the substitution is very likely to be intolerant. This rule uses a single feature (X_{43}), covering 14% (2, 014) data samples with an accuracy of 95% and a discrimination power of 9.6.

These rules can be understood as follows. In the Pfam multiple sequence alignments, homologous proteins are aligned according to their functional units (protein domains). Hence, amino acids appearing in a certain column of an alignment would be those that are crucial in maintaining the protein function. On the contrary, amino acids rarely appearing in a certain column of an alignment would very likely be irrelevant to the protein function. Therefore, in Rule 1 and Rule 2, when an amino acid is substituted by another amino acid which never or rarely appears in the multiple sequence alignment, the function of the protein could hardly be maintained. In Rule 3, the very abundant appearance of the original amino acid in the multiple sequence alignment indicates that the amino acid is crucial in keeping the protein function. Therefore, when the amino acid is substituted, the protein would very likely be malfunction.

The second group of three rules uses physicochemical features and provides us specific understanding regarding the intolerant substitutions for individual amino acids, as illustrated in Figure 6.

Rule 4 in Figure 6 says that if a Cysteine is substituted, no matter what kind of amino acids it is substituted to, the substitution is very likely to be intolerant. This rule uses a single feature (X_1), covering 4.2% (596) data samples with an accuracy of 91% and a discrimination power of 4.7. The Cysteine is the only amino acid capable of forming disulfide bonds, and the disulfide bridges between Cysteines within a polypeptide support the protein's sec-



Rule 7: $\beta = 1.2\%$, $\gamma = 81\%$, $\rho = 9.2$

IF (The substituted amino acid appears in the multiple sequence alignment with a rather high frequency) ($0.600 < X_{44} \le 1.000$)
THEN (The substitution is tolerant)

Rule 8: $\beta = 4.3\%$, $\gamma = 74\%$, $\rho = 6.0$

IF (The original amino acid appears in a hydrophilic local environment AND the substituted amino acid appears in the multiple sequence alignment with a relatively high frequency) ($1 \le X_{15} \le 4$ and $0.200 < X_{44} \le 0.500$)
THEN (The substitution is tolerant)

Rule 9: $\beta = 3.5\%$, $\gamma = 77\%$, $\rho = 6.9$

WHEN (The amino acids {W, V, G, L, I} are substituted by the amino acids {V, I}) ($11 \le X_2 \le 15$ and $2 \le X_{12} \le 3$)
IF (The substituted amino acid has a neutral change in molecular weight against the original one) ($X_{25} = 4$)
THEN (The substitution is tolerant)

Figure 7
Three general tolerant rules. In the figure, β, γ, and ρ for each rule are the coverage, the accuracy, and the discrimination power of the corresponding rule, respectively.

ondary structure. Therefore, when a Cysteines is substituted, the structure would be destroyed, and the protein would lose its function.

Rule 5 in Figure 6 says that when a Glycine is substituted, if the evolutionary profile medium or highly prefers turns (x_{24} = 6,7), the substitution is very likely to be intolerant. This rule uses 2 features (X_1 and X_{24}), covering 6.6% (940) data samples with an accuracy of 95% and a discrimination power of 9.0. This rule can be understood from two aspects. First, the Glycine is the smallest amino acid. Therefore, when a Glycine is substituted by any other (bigger) amino acids, there might not be enough space to hold that amino acid, and thus the secondary structure of the polypeptide would be destroyed. As a result, the protein would lose its function. Second, the Glycine is one of the amino acids most prefer turns (only second to Proline). Hence, when the turn structure is important to the protein function (evolutionary profile medium or highly prefers turns) and a Glycine is substituted, the function of the protein would very likely change.

Rule 6 in Figure 6 says that when an amino acid is substituted by an Arginine, if the evolutionary profile is acidic (x_{20} = 1,2,3), and the substituted amino acid (the Arginine) has a medium to strongly negative change in hydrophobicity versus the evolutionary profile (x_{39} = 1,2), the substitution is very likely to be intolerant. This rule uses 3 features (X_9, X_{20}, and X_{39}), covering 5.7% (805) data samples with an accuracy of 93% and a discrimination power of 6.1. This rule can be understood from the following aspects. First, the Arginine is the most alkalic (with the highest pI value) and most hydrophilic amino acid. Second, an acidic evolutionary profile indicates that amino acids with small pI values are crucial to the protein's function. Thirdly, the substituted amino acid (the Arginine) having medium to strongly negative change in hydrophobicity scale suggests that hydrophilic amino acids are the majority in the homologous proteins (in other words, hydrophilic amino acids are crucial to the protein's function). Therefore, when an Arginine replaces the original amino acid, the above second and third conditions are violated, and thus the function of the protein would be destroyed.

The third group of three rules uses both the conservation scores and the physicochemical features, and provides us specific understanding regarding the tolerant substitutions, as illustrated in Figure 7.

Rule 7 in Figure 7 says that if the substituted amino acid appears in the multiple sequence alignment with a rather high frequency ($0.600 < X_{44} \le 1.000$), the substitution is very likely to be tolerant. This rule uses a single feature (X_{44}), covering 1.2% (171) data samples with an accuracy of 81% and a discrimination power of 9.2. This rule can be thought of as the opposite of the previous Rule 1 and Rule 2. Amino acids appearing in a certain column of a Pfam multiple sequence alignment would be those that are crucial in maintaining the protein function. Therefore, when an amino acid is substituted by another amino acid which appears in the multiple sequence alignment with a rather high frequency, the function of the protein could possibly be maintained, and the substitution is likely to be tolerant.

Rule 8 in Figure 7 says that if the original amino acid appears in a hydrophilic local environment ($1 \le X_{15} \le 4$) and the substituted amino acid appears in the multiple sequence alignment with a relatively high frequency ($0.200 < X_{44} \le 0.500$), the substitution is very likely to be tolerant. This rule uses 2 features (X_{15} and X_{44}), covering 4.3% (610) data samples with an accuracy of 74% and a discrimination power of 6.0. The understanding of this rule is similar to the previous Rule 7. Amino acids appearing in a certain column of a Pfam multiple sequence alignment would be those that relate to the maintenance of the protein function. Hence, when an amino acid is substituted by another amino acid which appears in the multiple sequence alignment with a relatively high frequency,

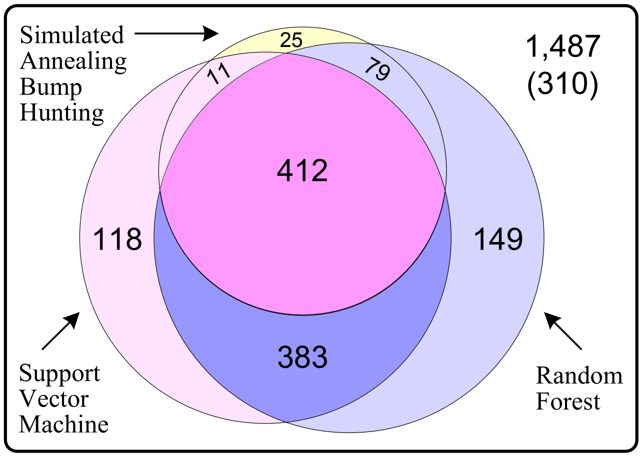


Figure 8
Prediction results for unclassified amino acid substitutions occurring in human proteins and collected in the Swiss-Prot database.

the function of the protein could possibly be maintained, and the substitution is likely to be tolerant.

Rule 9 in Figure 7 says that when one of the amino acids in the set of {W, V, G, L, I} is substituted by one of the amino acids the set of {V, I} ($11 \le X_2 \le 15$ and $2 \le X_{12} \le 3$), if the substituted amino acid has a neutral change in molecular weight against the original one ($X_{25} = 4$), the substitution is likely to be tolerant. This rule uses 3 features (X_2, X_{12}, and X_{25}), covering 3.5% (493) data samples with an accuracy of 77% and a discrimination power of 6.9. The principle behind this rule is that when amino acids are substituted by other amino acids having similar physicochemical properties, the structure of the protein is likely to be maintained, and thus the function of the protein is likely to be kept.

Prediction of the unclassified amino acid substitutions

We further applied the support vector machine and the random forest with the discrete form of our feature set to predict potential effects of unclassified amino acid substitutions in human proteins. We first used the 10-fold cross-validation experiments to determine the decision threshold for each method so that the balanced error rate (BER) could be minimized in the experiments, and then applied each method with the corresponding decision threshold on the unclassified data to make predictions. The results are shown in Figure 8.

Within the 1,487 unclassified amino acid substitutions, the support vector machine predicted 924 (412 + 383+118+11) as intolerant and 563 (310+149 + 79 + 25) as tolerant, while the random forest predicted 1023 (412 + 383 + 149 + 79) as intolerant and 464 (310 + 118 + 11 + 25) as tolerant. 795 (412 + 383) substitutions were predicted as intolerant and 335 (310 + 25) were predicted as tolerant by both methods. These overlapping predictions were therefore with high confidence.

We also applied the six intolerant rules induced by the simulated annealing bump hunting strategy in the previous section to the unclassified data. In total, the six intolerant rules covered 527 (412 + 79 + 11 + 25) data samples, and 412 out of them were also predicted as intolerant by both the support vector machine and the random forest. Besides, 90 samples covered by these rules were also predicted as intolerant by one of the prediction methods. Only 25 samples were not predicted as intolerant by either method. These statistics suggested that the induced interpretable rules were general (covering a significant proportion of data samples), and were of very high quality (with very few exceptions). More importantly, beyond the highly confident predictions, these rules also revealed the physicochemical principles behind the covered amino acid substitutions and explained why these substitutions would be intolerant.

Discussion and conclusions

Most contemporary studies aiming at predicting potential effects of amino acid substitutions made use of complicated and not widely available properties of amino acids and proteins. To overcome these limitations, we proposed a feature set based on three physicochemical properties of amino acids, three relative frequencies of amino acids in the secondary structures of proteins with known secondary structure information, and two evolutionary conservation scores. We applied three machine learning methods (the decision tree, the support vector machine, and the random forest) with our feature set to experimental amino acid substitutions occurring in the *E. coli* lac repressor and the bacteriophage T4 lysozyme, and showed that the methods using our feature set could achieve preferred prediction results in terms of the area under the ROC curve, the balanced error rate, and the Matthews' correlation coefficient. We further applied the support vector machine and the random forest with our feature set to a large number of amino acid substitutions occurring in highly heterogenous human proteins, and showed that our feature set could be applied to a much wider range of human proteins and the prediction methods using our feature set were superior to those using the existing more complicated feature sets.

Although existing methods could produce reasonable predictions, they were not capable of capturing physicochemical principles behind the predictions. In many situations, however, these hidden principles were of more importance because they could uncover how amino acid substitutions affect protein functions and why some substitutions would result in diseases. In order to explore these principles, we used a novel designed rule induction method called the simulated annealing bump hunting strategy to automatically extract interpretable rules for amino acid substitutions. The induced rules were either consistent with current biological knowledge or providing new insights for the understanding of the physicochemical principles behind amino acid substitutions.

One limitation of our feature set is that we currently use the Pfam multiple sequence alignment to extract evolutionary information for the query protein sequence. As a result, we are limited to deal with amino acid substitutions occurring in known protein domains. This limitation can be overcome by using some other multiple sequence alignment method such as the PSI-BLAST and ClustalW instead of the Pfam. Another limitation of our feature set is that the number of features is large, and some of them are highly correlated. Although good results have been achieved, integrating feature selection mechanisms in prediction methods could further improve the prediction performance. This demand is especially urgent when using the support vector machine as the prediction method. A third limitation is regarding how to perform fair and comprehensive comparisons between feature sets and prediction methods proposed in different literatures, especially when the training and test samples are of different sizes and from different data sources. Although this direction is not the focus in this paper, it would be of great importance and necessity in developing a general benchmark system using unified statistical criteria in our future work.

As for the simulated annealing bump hunting strategy, there exist two free parameters (λ and β_0). Although free parameters incorporate more flexibility into the method, they make the computational burden heavier (in order to tune these parameters). How to design an automated mechanism to guide the determination of these free parameters remains an ongoing study. Also, although the nine presented rules could be well explained, there exist some other rules which are not easy to be interpreted by current biological knowledge, especially when the rules contain many features. How to simplify our feature set to make the rules more interpretable forms another research focus.

Despite the limitations, we showed that our results were reasonably good. When using our feature set with the support vector machine and the random forest, we obtained better ROC curves and smaller (balanced) prediction error rates in the cross-validation experiments. When applied to unclassified data, the six induced intolerant rules could cover a large portion of data samples, and most of the covered substitutions were also predicted as intolerant by either the support vector machine or the random forest. More importantly, beyond the highly confident predictions, these rules could also reveal the physicochemical principles behind the covered samples and explain why these substitutions would cause diseases.

Methods

The proposed feature set

We propose a set of 44 features which are general enough for most known proteins and are easy to be obtained by simple calculations. Our feature set has a continuous form, in which all the 44 features have continuous values, and a discrete form, in which 42 features have ordered categorical values and the other 2 have continuous values. The features are derived based on 3 physicochemical properties (*molecular weight*, *pI value*, and *hydrophobicity scale*) of amino acids, 3 relative frequencies for the occurrences of amino acids in the secondary structures (*helices*, *strands*, and *turns*) of proteins with known secondary structural information, and two evolutionary conservation scores. The unit of molecular weight is Dalton. The *p*I (Isoelectric Point) is the *p*H value at which a molecule carries no net electrical charge. The hydrophobicity scale of Kyte and Doolitle is derived from the physicochemical properties of amino acid side chains [25]. The three relative frequencies are calculated by counting the occurrences of amino acids in the corresponding secondary structure of proteins with known secondary structural information. All these six properties can either be obtained from the literature [25,26], or be calculated using only the sequential information of proteins [27].

The continuous form

For a given amino acid substitution pair (Org → Sub) in a certain query protein, the above 6 properties are calculated for the original (Org) and the substituted (Sub) amino acid, as well as in a window-sized situation which includes the neighbors of the original amino acids in the query protein sequence, and in a column-weighted circumstance in which the query protein sequence is aligned with its homologous proteins. The calculations of the properties for the original and the substituted amino acids are straightforward. The window-sized properties (with window size W) are calculated as the average of the corresponding properties for the original amino acid and its W - 1 neighbors in the query protein sequence. According to the known relationship between sequences and secondary structures of proteins (i.e., α helices are defined by repeated hydrogen bonds with a period of 4 amino acids, and have 3.6 amino acids per turn [26]), in this paper, we set the window size $W = 9$ so that the sequence information of the amino acids at the substitution positions and the α helices next to the substitution residues can be included. The column-weighted properties are calculated as follows. For the query protein, its homologous proteins

Table 4: Details of the proposed features.

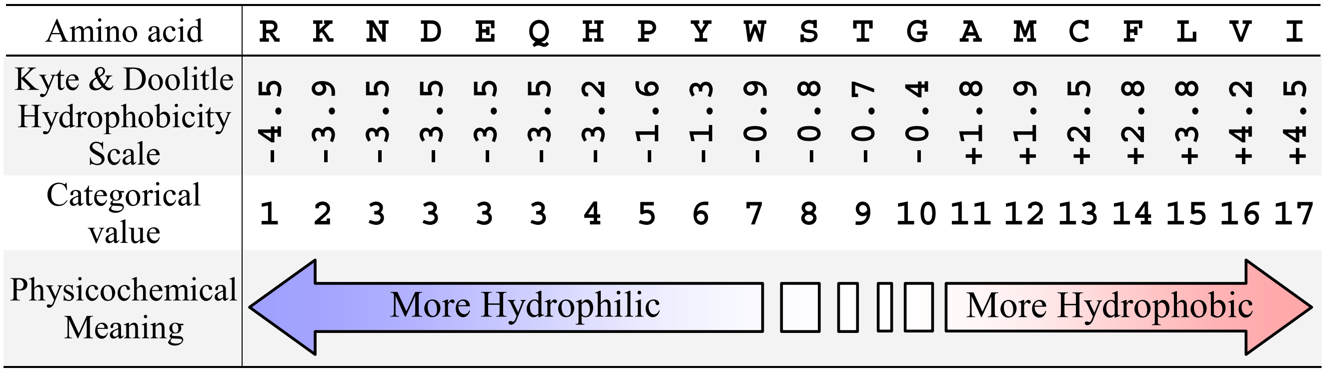
	Physicochemical			Relative frequency in			Conservation
	Molecular weight	pI value	Hydrophobicity	Helices	Strands	Turns	Frequency in MSA
Original	X_1	X_2	X_3	X_4	X_5	X_6	X_{43}
Substitution	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{44}
Window-sized	X_{13}	X_{14}	X_{15}	X_{16}	X_{17}	X_{18}	
Column-weighted	X_{19}	X_{20}	X_{21}	X_{22}	X_{23}	X_{24}	
Substitution-Original	X_{25}	X_{26}	X_{27}	X_{28}	X_{29}	X_{30}	
Substitution-Window	X_{31}	X_{32}	X_{33}	X_{34}	X_{35}	X_{36}	
Substitution-Column	X_{37}	X_{38}	X_{39}	X_{40}	X_{41}	X_{42}	

Each of the 6 amino acid properties is calculated in 7 situations, forming $X_1 \sim X_{42}$, while the conservation scores for the original and the substituted amino acids become X_{43} and X_{44}, respectively.

are extracted from the Pfam database [22]. Supposing that the substitution occurs at a position corresponds to a certain column of the alignment, the column-weighted properties are then calculated as the weighted average of the corresponding properties for all the 20 kinds of amino acids, where the weight of a certain kind of amino acid is the frequency of its occurrence in the corresponding column of the alignment.

In addition, for each substitution pair, three combinations of the above four situations are considered. First, each of the 6 properties of the original amino acid is subtracted from the corresponding property of the substituted amino acid, forming 6 features measuring the relative change of the substituted amino acid versus the original amino acid. Second, each of the 6 window-sized properties is subtracted from the corresponding property of the substituted amino acid, forming 6 features measuring the relative change of the substituted amino acid versus the local environment of the substitution position in the query protein. Thirdly, each of the 6 column-weighted properties is subtracted from the corresponding property of the substituted amino acid, forming 6 features measuring the relative change of the substituted amino acid versus the evolutionary profile of the substitution position in the homologous proteins.

Besides the above physicochemical and relative frequency features, our feature set also include two evolutionary conservation scores for the original and the substituted amino acids. The conservation scores are defined as the frequencies of occurrences of the amino acids (original or substituted) in the corresponding column of the Pfam multiple sequence alignment.



Amino acid	R	K	N	D	E	Q	H	P	Y	W	S	T	G	A	M	C	F	L	V	I
Kyte & Doolitle Hydrophobicity Scale	-4.5	-3.9	-3.5	-3.5	-3.5	-3.5	-3.2	-1.6	-1.3	-0.9	-0.8	-0.7	-0.4	+1.8	+1.9	+2.5	+2.8	+3.8	+4.2	+4.5
Categorical value	1	2	3	3	3	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Physicochemical Meaning	More Hydrophilic													More Hydrophobic						

Figure 9
An illustration of ordered categorical values for hydrophobicity scales for the original or substituted amino acids.

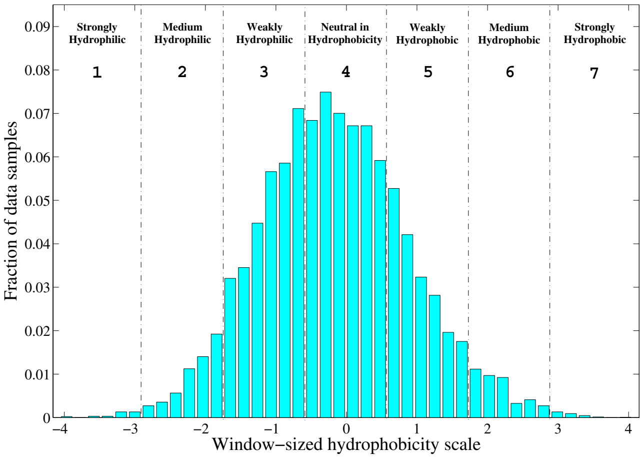


Figure 10
An illustration of ordered categorical values for the window-sized hydrophobicity.

With the above properties being calculated, we propose the continuous form of the feature set, including 42 physicochemical or relative frequency properties (each of the 6 amino acid properties being calculated in 7 different situations) and 2 conservation scores (for the original and the substituted amino acids). As a summary, Table 4 shows this feature set, with features labeled by X_i for i = 1,..., 44.

The discrete form

In order to make the features interpretable in physicochemical terms, we further discretize the physicochemical and relative frequency properties ($X_1 \sim X_{42}$ in Table 4). For each of the properties corresponding to the original or the substituted amino acids ($X_1 \sim X_{12}$), we first order the possible values (corresponding to the 20 amino acids) from the smallest to the largest, and then use the ranks as the categorical values for the property. By doing this, each categorical value corresponds to one or more amino acids, while the categorical values for a certain property have intrinsic order and clear physicochemical meaning. For example, Figure 9 illustrates the ordered categorical values for the hydrophobicity scale (X_3 or X_9). 20 amino acids are sorted according to their hydrophobicity scale, from the most hydrophilic (R) to the most hydrophobic (I). The ranks of the sorting results are then used as the categorical values for the amino acids. Since amino acids N, D, E, and Q have identical hydrophobicity in the Kyte and Doolitle scale, they are assigned the same categorical value (3). The physicochemical meaning of the ordered categorical value is straight forward: the smaller the value, the more hydrophilic the amino acid, and vice versa.

For each of the other properties ($X_{13} \sim X_{42}$), we first group all the possible values to 7 bins with each bin having equal interval, and then use the indices of the bins as the categorical values of the property. By doing this, each categorical value corresponds to several substitution pairs, while the categorical values for a certain property have intrinsic order and clear physicochemical meaning. As a demonstration, Figure 10 illustrates how the window-sized hydrophobicity scale is discretized. First, all the window-sized hydrophobicity scale values (X_{15}) in the data set are collected, and the minimum value (-4.0) and maximum value (+4.0) are determined. And then, 6 threshold values ({-2.86, -1.71, -0.57, 0.57, 1.71, 2.86}) are calculated so that the interval ([-4.0, +4.0]) can be cut into 7 equal bins. Finally, the indices of the bins are used as the categorical values. The physicochemical meaning of the ordered categorical value is straight forward: the smallest value corresponds to strongly hydrophilic local environment, while the largest value corresponds to strongly hydrophobic local environment.

With each feature having meaningful ordered categorical values, we propose the discrete form of our feature set, containing 42 physicochemical or relative frequency properties (each having ordered categorical values) and the 2 conservation scores. We would use the same method as shown in Table 4 to label these features, with $X_1 \sim X_{42}$ having ordered categorical values.

Prediction methods and evaluation criteria

When comparing the proposed feature set with other published ones using experimental substitution data, we use the decision tree, the support vector machine, and the random forest to predict the potential effects of amino acid substitutions. Recent studies regarding the random forest [19] have shown that prediction results can be significantly improved by growing a set of decision trees and letting them to vote. Hence, we adopt the support vector machine (SVM) [18] and the random forest (RF) [19] with the proposed feature set to predict the potential effects of human disease related substitutions. For the support vector machine, two crucial parameters are commonly referred to as C and g. We use a grid search, as included in the *libsvm* software package [18] to determine these parameters. For the random forest, two important parameters are in general referred to as mtry (the number of randomly selected features at each node of the internal decision trees) and jbt (the number of decision trees in the forest). We use jbt = 1000, and try different mtry (from 1 to 10) to select the one which can give us the best prediction performance.

The performance of each prediction method is evaluated using 10-fold cross-validation experiments, and the results of 10 independent experiments are averaged to get a fair evaluation. We use three criteria to evaluate the performance of a prediction method. The first criterion is the area under the receiver operating characteristic (ROC) curve (AUC), which provides us comprehensive understanding for the prediction power of a given method. The other two criteria are the balanced error rate (BER) and the Matthews' correlation coefficient (MCC) [28]. They take the imbalance of intolerant samples and tolerant samples into consideration and provide us more detailed understanding for the prediction power under certain decision threshold.

Given the 10-fold cross-validation results and a certain decision threshold, we can calculate the numbers of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) under the threshold. The balanced error rate (BER) and the Matthews' correlation coefficient (MCC) [28] under the decision threshold are then defined as

$$\mathrm{BER} = \frac{1}{2}\left(\frac{\mathrm{FP}}{\mathrm{TN}+\mathrm{FP}} + \frac{\mathrm{FN}}{\mathrm{TP}+\mathrm{FN}} \right),$$

and

$$\mathrm{MCC} = \frac{\mathrm{TP}\times\mathrm{TN} - \mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TN}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TP}+\mathrm{FP})}}.$$

In general, the small the BER and the large the MCC, the better the prediction method.

Rule induction for amino acid substitutions

With the meaningful features available, we can use rule induction methods to automatically extract interpretable rules for amino acid substitutions. A rule has a productive format

IF (*condition*)

THEN (*prediction*).

The condition part should include only a small number of features so that the rule can be easily interpreted, while the prediction part gives an assertion about the potential effects (tolerant or intolerant) of amino acid substitutions which satisfy the condition.

For a given amino acid substitution data sample, let $\mathbf{x} = (x_1,..., x_D)^T$ be the vector of all the features, where $D = 44$ is the total number of features. Let y be the indicator of the substitution type. In the case that we target to extract rules for intolerant substitutions, $y = 1$ if a substitution is intolerant and 0, otherwise. In the case that we aim at extracting rules for tolerant substitutions, y has the opposite meaning. Each substitution can be thought of as an observation of the output (y) produced by a certain unknown function, given the inputs ($\mathbf{x}$), and observations with similar outputs and similar inputs (or a subset of the inputs) define a rule. The similarity of the inputs can be specifically described by a "box" (sub-region) in the feature space, and defined by a set of feature intervals. The coverage of a rule can be represented by the size of the corresponding box (box-size), and the quality of a rule can be described by the average value of the output y for data samples inside the box (box-mean). Let N be the total number of data samples. The rule induction process is then mathematically formulated as:

Given repeated observations $\{(y_k, \mathbf{x}_k)\}_{k=1}^{N}$ composed of the substitutions (the outputs y_k), along with simultaneous values of the features (the inputs $\mathbf{x}_k$), search in the feature space optimal boxes such that the box-means are as large as possible while the box-sizes are not very small.

This problem can be addressed using the patient rule induction method (PRIM) [21], which is also referred to as a "bump hunting" method. Each rule is described using a "box" $\mathcal{B}$ in the feature space, defined as

$$\mathcal{B} = \bigcap_{d=1}^{D} \mathcal{B}_d,$$

where interval $\mathcal{B}_d = [b_{d-}, b_{d+}]$ is the boundary for the d-th dimension of the box. The location of the k-th data point in the d-th dimension can be described by an indicator

$$\delta_k^d = \begin{cases} 1, & b_{d-} \le x_k^d \le b_{d+}; \\ 0, & \text{otherwise.} \end{cases}$$

For the k-th data sample, another indicator is further introduced as

$$\delta_k = \prod_{d=1}^{D} \delta_k^d,$$

to describe whether the k-th data point locates inside the box ($\delta_k = 1$) or not ($\delta_k = 0$). The size of a box $\mathcal{B}$ is quantified by the (normalized) number of data points falling into the box as

$$\beta_{\mathcal{B}} = \frac{1}{N}\sum_{k=1}^{N} \delta_k.$$

The average value of the output y for data points locating inside the box $\mathcal{B}$ is referred to as the box-mean and calculated by

$$\gamma_{\mathcal{B}} = \frac{1}{N\beta_{\mathcal{B}}}\sum_{k=1}^{N}\delta_k\gamma_k.$$

These definitions make both the box-size and the box-mean taking values in the interval of [0, 1].

The PRIM then intends to search in the box space a box $\mathcal{B}$ which has maximum box-mean $\gamma_{\mathcal{B}}$, with the constraint $\beta_{\mathcal{B}} > \beta_0$ (β_0 is a predefined threshold). This is treated by a "top-down peeling" algorithm and a "bottom-up pasting" algorithm. The top-down peeling algorithm starts from the whole search space (the initial box) and repeatedly tries to maximize the box-mean by removing some bad data points ($\gamma = 0$) from the box. Since each peeling is performed without knowledge of later peels, it is possible that the final box can be refined by readjusting some of its boundaries. Hence, the bottom-up pasting algorithm repeatedly tries to put some good data points ($\gamma = 1$) back by growing the box. Smaller boxes often results in larger box-mean, the PRIM thus seeks for a reasonable tradeoff between the box size and the box mean. The tradeoff is typically done manually by looking at a box-size – box-mean trajectory plot. The final box represents the extracted rule. Considering that some redundant features may exist, a tradeoff between the complexity and goodness of the rule can be further considered by trying to remove some features from the rule. This is done after the final box is obtained by looking at how box-mean changes while removing some boundaries from the box.

Simulated annealing bump hunting strategy

The PRIM can be thought of as a steepest-ascent searching method in the box space. The final box is a (local) optimum without guarantee to be the global optimum. Also, the top-down peeling removes the data points permanently in iterations. Although some of the good data points can be put back by the bottom-up pasting, the repair to the box seems to be very limited. Moreover, it is doubtable that the process of removing some features from the rule after the final box is obtained could be an effective way to generate an optimum rule. These considerations motivate us to explore an automated feature selection methodology which can discard redundant features while extracting rules. The basic idea is to use the simulated annealing strategy instead of the steepest-ascent searching, while incorporating the automated feature selection process in the strategy.

The presence of a feature in a rule can be described as the presence of a boundary in a box and represented by an indicator ξ_d, where $\xi_d = 1$ if the d-th feature is included in the box and 0, otherwise. The indicator δ_k with ξ_d included then becomes $\delta_k = \prod_{d=1}^{D}\left(\delta_k^d\right)^{\xi_d}$. The formulas for the box size $\beta_{\mathcal{B}}$ and box-mean $\gamma_{\mathcal{B}}$ remain unchanged. Considering that the proportion of data samples from different categories may be very different, we further introduce a normalized quantity

$$\rho_{\mathcal{B}} = \frac{\sum_{k=1}^{N}\delta_k\gamma_k / \sum_{k=1}^{N}\gamma_k}{\sum_{k=1}^{N}\delta_k(1-\gamma_k) / \sum_{k=1}^{N}(1-\gamma_k)} = \frac{1}{\alpha}\times\frac{\gamma_{\mathcal{B}}}{1-\gamma_{\mathcal{B}}}$$

to measure the discrimination power of a box $\mathcal{B}$, where $\alpha = \sum_{k=1}^{N}\gamma_k / \sum_{k=1}^{N}(1-\gamma_k)$ is the ratio of the positive data samples against the negative data samples. We would take the possible imbalance between the data samples into consideration and maximize the discrimination power $\rho_{\mathcal{B}}$ for rule induction. Nevertheless, α is a constant with fixed number of data samples, maximizing the discrimination power $\rho_{\mathcal{B}}$ is therefore equivalent to maximizing the box mean $\gamma_{\mathcal{B}}$, and vice versa.

For boxes with comparable box-sizes and box-means, we prefer boxes have fewer features. This is achieved by rewarding boxes with less features using the quantity of $\tau_{\mathcal{B}} = \exp(-\lambda\sum_{d=1}^{D}\xi_d)$ where λ is a hyper-parameter. λ = 0 means that we do not take the number of features into consideration, while positive λ values give preference to less number of features. In this paper, we in general set λ = 1.0.

The simulated annealing strategy then intends to maximize the box-mean $\gamma_{\mathcal{B}}$ using as simple box as possible with the constraint that the box-size $\beta_{\mathcal{B}} > \beta_0$. We write this maximization problem as

$$\begin{array}{ll} \max & \gamma_{\mathcal{B}} + \tau_{\mathcal{B}}, \\ \text{s.t.} & \beta_{\mathcal{B}} > \beta_0. \end{array}$$

where β_0 is a predefined threshold (minimum size of a box, e.g. $\beta_0 = 0.05$). Define the energy function as $E = 1 - (\gamma_{\mathcal{B}} + \tau_{\mathcal{B}})/2$. The simulated annealing strategy repeatedly generates new boxes using meta-operations described below and seeking for energy decreasing. Let $\Delta E = E^{\text{new}} - E^{\text{old}}$. If a tentative new box can decrease the energy

($\Delta E < 0$), it is accepted; otherwise ($\Delta E \geq 0$), it is accepted with probability $\pi = \exp(-\kappa\Delta E/T)$, where κ is a normalization constant (e.g., $\kappa = 1.0$) and T is a pseudo-temperature (with initial value 1.0). Three meta-operations are used to generate a new box from the current one.

1. **Left side peeling/pasting**. Select a d-th dimension at random, then update the left bound ary. For continuous values, let $b_{d-} \leftarrow b_{d-} + N(0,1) \times (b_{d+} - b_{d-})$, where $N(0,1)$ is a real number sampled from a Normal distribution with mean 0 and standard derivation 1. For ordered categorical values, let $b_{d-} \leftarrow b_{d-} \pm 1$.

2. **Right side peeling/pasting**. Select a d-th dimension at random, then update the right boundary. For continuous values, let $b_{d+} \leftarrow b_{d+} + N(0,1) \times (b_{d+} - b_{d-})$. For ordered categorical values, let $b_{d+} \leftarrow b_{d+} \pm 1$.

3. **Feature including/excluding**. Select a d-th dimension at random, then update the d-th boundary by adding it to the box ($\xi_d \leftarrow 1$) or removing it from the box ($\xi_d \leftarrow 0$).

The simulated annealing bump hunting strategy can then be described as follows.

1. **Initialization**. Generate a box containing all the data samples.

2. **Random walk**. Execute one of the meta-operations at random on the current box, calculate

$$\Delta E = \frac{1}{2}(\gamma_{\mathscr{B}}^{\text{old}} + \tau_{\mathscr{B}}^{\text{old}} - \gamma_{\mathscr{B}}^{\text{new}} - \tau_{\mathscr{B}}^{\text{new}}).$$

3. **Acceptance**. If $\Delta E < 0$, accept the random walk; otherwise, accept the walk with probability

$\Pr(\text{accept}) = \exp(-k\Delta E/T)$.

4. **Temperature decay**. Decrease T by power law: $T \leftarrow T \times \Delta T$, where ΔT is a positive real number close to 1.0 (e.g., $\Delta T = 0.9999$).

5. Repeat 2 ~ 4 until convergence.

Authors' contributions

RJ designed the simulated annealing bump hunting strategy, performed the rule induction part and prepared the manuscript. HY designed the feature set and performed the prediction part. FS participated in the research design and helped to prepare the manuscript. TC initialized, designed and directed the research. All authors read and approved the final manuscript.

Acknowledgements

The authors appreciate Dr. Pauline Ng for providing experimental amino acid substitution data, and Dr. Carles Ferrer-Costa for communication regarding human amino acid substitution data. The authors thank two anonymous referees for suggested revisions. This work was partly supported by National Institutes of Health P50 HG 002790, National Institutes of Health R01 LM008991-01, National Institutes of Health/National Science Foundation Joint Mathematical Biology Initiative DMS-0241102, and Alfred Sloan Research Fellowship.

References

1. Terp BN, Cooper DN, Christensen IT, Jorgensen FS, Bross P, Gregersen N, Krawczak M: **Assessing the relative importance of the biophysical properties of amino acid substitutions associated with human genetic disease.** *Human Mutation* 2002, **20(2):**98-109.
2. Bairoch A, Apweiler R, Wu C, Barker W, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin M, Natale D, O'Donovan C, Redaschi N, Yeh L: **The Universal Protein Resource (UniProt).** *Nucleic Acids Research* 2005, **33:**D154-159.
3. Krawczak M, Ball EV, Fenton I, Stenson PD, Abeysinghe S, Thomas N, Cooper DN: **Human gene mutation database: a biomedical information and research resource.** *Human Mutation* 2000, **15:**45-51.
4. McKusick VA: *Mendelian Inheritance in Man. A Catalog of Human Genes and Genetic Disorders* 12th edition. Baltimore: Johns Hopkins University Press; 1998.
5. Markiewicz P, Kleina LG, Cruz C, Ehret S, Miller JH: **Genetic studies of the *lac* repressor XIV: Analysis of 4000 altered *Escherichia coli lac* repressors reveals essential and non-essential residues, as well as "spacers" which do not require a specific sequence.** *Journal of Molecular Biology* 1994, **240(5):**421-433.
6. Suckow YJ, Markiewicz P, Kleina LG, Miller J, Kisters-Woike B, Muller-Hill B: **Genetic studies of the *lac* repressor XV: 4000 single amino acid substitutions and analysis of the resulting phenotypes on the basis of the protein structure.** *Journal of Molecular Biology* 1996, **261(4):**509-523.
7. Renell D, Bouvier SE, Hardy LW, Poteete AR: **Systematic mutation of bacteriophage T4 lysozyme.** *Journal of Molecular Biology* 1991, **222:**67-88.
8. Chasman D, Adams RM: **Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation.** *Journal of Molecular Biology* 2001, **307(2):**683-706.
9. Sunyaev S, Ramensky V, Koch I, Lathe W III, Kondrashov AS, Bork P: **Prediction of deleterious human alleles.** *Human Molecular Genetics* 2001, **10(6):**591-597.
10. Ramensky V, Bork P, Sunyaev S: **Human non-synonymous SNPs:server and survey.** *Nucleic Acids Research* 2002, **30(17):**3894-3900.
11. Ferrer-Costa C, Orozco M, de la Cruz X: **Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties.** *Journal of Molecular Biology* 2002, **315(4):**771-786.
12. Ferrer-Costa C, Orozco M, de la Cruz X: **Sequence-based prediction of pathological mutations.** *Proteins: Structure, Function, and Bioinformatics* 2004, **57(4):**811-819.
13. Saunders CT, Barker D: **Evaluation of structural and evolutionary contributions to deleterious mutation prediction.** *Journal of Molecular Biology* 2002, **322(4):**891-901.
14. Krishnan VG, Westhead DR: **A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function.** *Bioinformatics* 2003, **19(17):**2199-2209.
15. Ng PC, Henikoff S: **Predicting deleterious amino acid substitutions.** *Genome Research* 2001, **11(5):**863-874.
16. Mitchell TM: *Machine Learning* U.S.A.: McGraw-Hill; 1997.
17. Vapnik NV: *Statistical Learning Theory* New York: Wiley-Interscience; 1998.
18. Fan RE, Chen PH, Lin CJ: **Working set selection using the second order information for training SVM.** *Journal of Machine Learning Research* 2005, **6:**1889-1918 [http://www.csie.ntu.edu.tw/~cjlin/libsvm].
19. Breiman L: **Random Forests.** *Machine Learning* 2001, **45:**5-32.

20. Bao L, Cui Y: **Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information.** *Bioinformatics* 2005, **21(10):**2185-2190.
21. Friedman JH, Fisher NI: **Bump hunting in high-dimensional data.** *Statistics and Computing* 1999, **9(2**123-143 [http://www-stat.stanford.edu/~jhf/SuperGEM.html].
22. Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer EL, Bateman A: **Pfam: clans, web tools and services.** *Nucleic Acids Research* 2006:D247-D251.
23. Bowie JU, Luthy R, Eisenberg D: **A method to identify protein sequences that fold into a known three-dimensional structure.** *Science* 1991, **253:**164-170.
24. Frishman D, Argos P: **Knowledge-based protein secondary structure assignment.** *Proteins* 1995, **23(4):**566-579.
25. Kyte J, Doolittle RF: **A simple method for displaying the hydropathic character of a protein.** *Journal of Molecular Biology* 1982, **157:**105-132.
26. Berg JM, Tymoczko JL, Stryer L: *Biochemistry* Fifth edition. W. H. Freeman and Company; 2002.
27. Durbin R, Eddy S, Krogh A, Mitchison G: *Biological sequence analysis: probabilistic models of proteins and nucleic acids* Cambridge University press; 1998.
28. Matthews BW: **Comparison of the predicted and observed secondary structure of T4 phage lysozyme.** *Biochim Biophys Acta* 1975, **405(2):**442-451.