Predicting Functional Impact of Single Amino Acid Polymorphisms by Integrating Sequence and Structural Features

Mingjun Wang[1], Hong-Bin Shen[2], Tatsuya Akutsu[3], Jiangning Song[1,4,]*
[1]State Engineering Laboratory for Industrial Enzymes, Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin 300308, China
[2]Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong University, Shanghai 200240, China
[3]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan
[4]Department of Biochemistry and Molecular Biology, Faculty of Medicine, Monash University, Melbourne, VIC 3800, Australia

*Corresponding author. E-mail: song_jn@tib.cas.cn

Abstract- **Single amino acid polymorphisms (SAPs) are the most abundant form of known genetic variations associated with human diseases. It is of great interest to study the sequence-structure-function relationship underlying SAPs. In this work, we collected the human variant data from three databases and divided them into three categories, i.e. cancer somatic mutations (CSM), Mendelian disease-related variant (SVD) and neutral polymorphisms (SVP). We built support vector machine (SVM) classifiers to predict these three classes of SAPs, using the optimal features selected by a random forest algorithm. Consequently, 280 sequence-derived and structural features were initially extracted from the curated datasets from which 18 optimal candidate features were further selected by random forest. Furthermore, we performed a stepwise feature selection to select characteristic sequence and structural features that are important for predicting each SAPs class. As a result, our predictors achieved a prediction accuracy (ACC) of 84.97, 96.93, 86.98 and 88.24%, for the three classes, CSM, SVD and SVP, respectively. Performance comparison with other previously developed tools such as SIFT, SNAP and Polyphen2 indicates that our method provides a favorable performance with higher Sensitivity scores and Matthew's correlation coefficients (MCC). These results indicate that the prediction performance of SAPs classifiers can be effectively improved by feature selection. Moreover, division of SAPs into three respective categories and construction of accurate SVM-based classifiers for each class provides a practically useful way for investigating the difference between Mendelian disease-related variants and cancer somatic mutations.**

Keywords: single amino acid polymorphisms (SAPs); non-synonymous SNPs; support vector machine; random forest; feature selection

I. INTRODUCTION

Single nucleotide polymorphisms (SNPs) are the most abundant form of genetic variations, accounting for approximately 90% of DNA polymorphisms in humans [1, 2]. It is estimated that on average there is a SNP for every 300 base-pairs. SNPs in coding and regulatory regions may play a direct role in diseases or differing phenotypes [3]. Among them, the single amino acid polymorphisms (SAPs, also referred to as non-synonymous SNPs or nsSNPs) [3] are of special interest, as they lead to the change of amino acid types in the resulting protein products.

As many as 200,000 SAPs are estimated to be present in the human genome [4] and roughly 24,000-60,000 in an individual [5, 6]. This implies that there are 1-2 mutants per protein product. However, most of these mutants do not change the function of proteins. Therefore, discriminating the neutral and non-neutral mutants is urgently needed to understand the genotype/phenotype correlations and find the cure for diseases. Analyses of protein structure and function have suggested that single amino acid substitutions are responsible for certain disease types [7-9]. For example, it has been reported that about 60% of Mendelian disease is caused by amino acid substitutions [9, 10]. As the consequence of large-scale efforts, e.g. the HapMap project (http://www.hapmap.org) and the whole genome association studies [11], experimental SAPs data are accumulating rapidly in public databases including dbSNP, Swissprot variants and COSMIC (Catalogue of Somatic Mutation in Cancer) databases.

Previous analyses [3, 5, 12-15] generally divided the amino acid mutations into two classes, i.e. neutral and non-neutral mutants [9]. However, the non-neutral mutants can affect the function of proteins with varying levels of severity of phenotypic effects [5]. Hence, we further divided the non-neutral mutants into two classes: i) Mendelian disease-related variants (SVD) and ii) cancer somatic mutations (CSM) in addition to the neutral polymorphisms (SVP). Gong and Blundell have performed a similar analysis recently [9]. They analyzed disease-related variants and cancer somatic mutations. Compared to the conventional binary classification of SAPs, the three-class division has a more practical significance in discriminating different functional effects of SAPs and is able to shed light on the nature of the sequence-structure-function relationships of human SAPs.

In the past few decades, a variety of bioinformatic methods have been developed to predict possible disease association or functional effect of a given variant [3, 14-21]. A consensus of these methods is that they employ sequence or/and structural features and use them as input to train classifiers with various algorithms. They were based on statistical rules, decision trees, support vector machines (SVMs), neural networks, random forests and Bayesian networks and were applied to annotate mutants data. With the increasing availability of SAPs data, however, computational methods that are capable of predicting the functional effects of SAPs with better accuracy are consistently urgently needed.

In this study, we describe a new approach to classify SAPs into three different categories (CSM, SVD and SVP) and predict possible disease associations of SAPs, using SVMs augmented with efficient feature selection by random forest. We benchmarked this approach based on the rigorous 5-fold cross-validation tests and compared the prediction performance with other published tools. As a result, 18 optimal candidate features were selected from an initial set of 280 sequence and structural features. The SVM classifiers trained with the 18 selected optimal features achieved an accuracy of 84.97, 96.93, 86.98 and 88.24%, for the three-class, individual CSM, SVD and SVP class predictions, respectively.

II. METHODS

A. *Datasets*

We followed the same procedures as described in [9] to retrieve and compile a high-quality structural dataset of human variants. In particular, the SVD annotations were extracted from the UniProt [22] human sequence variations (release 57.5). CSM was taken from the COSMIC database (Catalogue of Somatic Mutation in Cancer, version: 48) [23] from which we chose the mutations that led to amino acid changes. SVP was taken from the Ensembl human variation database [24] (version 59_37d). In this study, we only extracted and used the verified SNPs to construct the structural dataset.

B. *Sequence and structural features*

Sequence feature: We extracted four different types of sequence-based features that proved useful in improving prediction performance. They include: (1) position-specific scoring matrics (PSSMs) generated by PSI-BLAST [25]; (2) predicted secondary structure by PSIPRED [26]; (3) predicted solvent accessibility by SCRATCH [27]; (4) predicted native disorder by DISOPRED [28]. Combination of these sequence-derived features has been shown to improve prediction performance in our recent work and that of others [29-32].

PSSM profiles: To generate the PSSMs, PSI-BLAST was run to search against the NCBI nr database with three iterations. Then the alignment profile and the obtained PSSMs were retained.

Solvent accessibilities: We used the NACCESS program [33] to calculate the absolute and relative solvent accessibilities of all atoms, total side chain, main chain, non-polar side chain and all-polar side chain, respectively.

Neighboring functional sites: If a mutation position is neighboring or close to the functional sites of a protein, it is more likely to be deleterious or disease associated. The annotations regarding the functional sites can be found in the "FT" line in UniProt database [22]. In our work, we extracted eight different types of UniProt functional features: ACT_SITE, BINDING, CA_BIND, DISULFID, DNA_BIND, LIPID, MENTAL, NP_BIND and MOD_RES. In addition, two other different types of neighboring functional sites were also taken into account. They included sequence neighbors and spatial neighbors, which were defined using the sequence and structural distances, respectively [3].

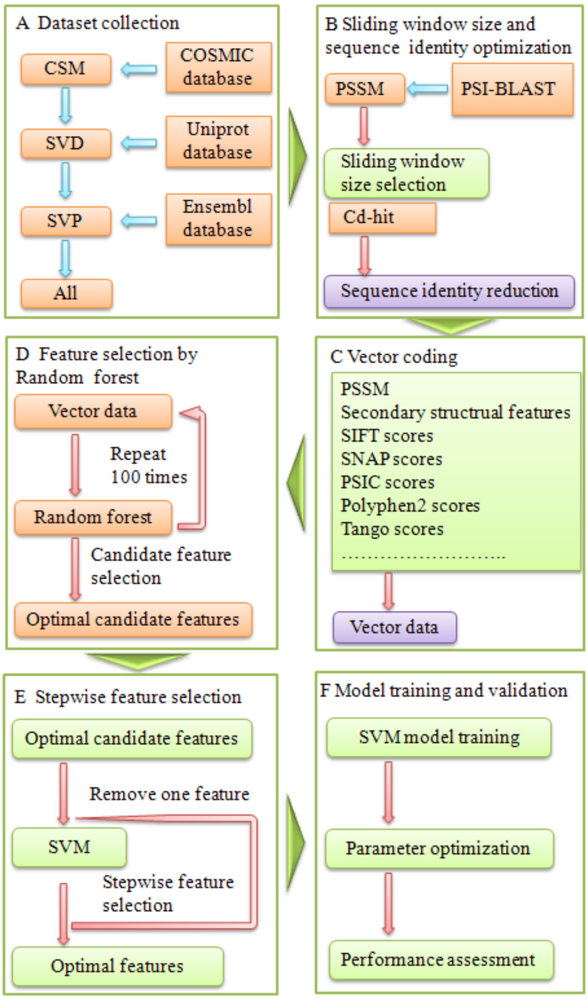
A Dataset collection
CSM
COSMIC database
SVD
Uniprot database
SVP
Ensembl database
All
B Sliding window size and sequence identity optimization
PSSM
PSI-BLAST
Sliding window size selection
Cd-hit
Sequence identity reduction
C Vector coding
PSSM
Secondary structrual features
SIFT scores
SNAP scores
PSIC scores
Polyphen2 scores
Tango scores
Vector data
D Feature selection by Random forest
Vector data
Repeat 100 times
Random forest
Candidate feature selection
Optimal candidate features
E Stepwise feature selection
Optimal candidate features
Remove one feature
SVM
Stepwise feature selection
Optimal features
F Model training and validation
SVM model training
Parameter optimization
Performance assessment

Figure 1. Schematic representation of the data collection, homology reduction, stepwise feature selection, parameter optimization, SVM model training and validation processes.

Aggregation properties: We used TANGO [34] to calculate the residue β-aggregation properties at mutation sites [3]. Particularly, we investigated whether inclusion of this feature could result in a performance improvement of deleterious mutation prediction.

Secondary structure features: We used DSSP [35] to extract the secondary structure annotations, including the number of H-bonds and disulfide bonds, solvent-accessible surface area, dihedral angle, Cα atom coordinates, protein backbone torsion angles (PHI and PSI angles) and so on.

Scores calculated by other softwares: 1) PSIC score: it represents how likely it is for a particular amino acid to occupy a specific position in protein sequence, calculated by PSIC [36]; 2) SIFT score: SIFT is a program that uses sequence homology to predict whether a substitution affects protein function [37]. For each mutation, five scores were calculated by SIFT and all of them were included in our feature sets; 3) SNAP score: SNAP is a method that predicts the functional effects of single amino acid substitutions [5, 38]. It calculated two scores and both were selected as features; 4) Polyphen2 score: Polyphen2 is a tool for predicting damaging effects of missense mutations [39] and it calculated four scores, all of which were used as features.

C. Selection of local sliding window size and reduction of sequence homology

We used PSSMs as input to the SVM models in order to select the optimal local sliding window size of *L*. 11 different window sizes were examined and compared, i.e. *L*=1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21.

To perform sequence homology reduction and avoid the potential overfitting problem, we used Cd-hit [40] to cluster protein sequences in our datasets, with varying sequence identity (SI) levels of 40, 50, 60, 70, 80, 90 and 100%, respectively.

D. Random forest feature selection

Random forest (RF) is an ensemble classifier based on decision trees [12, 41], which can be used for solving classification and regression tasks.

The RF package in R has been widely employed in bioinformatics, such as protein-RNA binding sites [42] and protein-protein interaction prediction [43]. More recently, RF has been successfully applied to perform feature selection in combination with SVM, for example, in the prediction of siRNA potency [44] and domain linker [45].

We used the mean decrease Gini index (MDGI) to select the more informative features, as calculated by the RF package in R [41]. MDGI represents the importance of individual vector element for correctly classifying SAPs. The mean MDGI Z-Score of each vector element is defined as:

$$MDGI \cdot Z - Score = (x_i - \overline{x}) / \sigma$$

where x_i is the mean MDGI of the *i*-th feature and σ is the standard deviation (SD), respectively. The vector element with MDGI Z-Score larger than 1.0 was selected as an optimal feature candidate (OFC). The feature selection based on RF was applied to the 3-class classification, but not to the binary classifications of CSM, SVD and SVP.

E. Stepwise feature selection

In addition, we performed a stepwise feature selection by training and evaluating four different SVM classifiers based on the 5-fold cross-validation tests. We divided our dataset into five subsets- in each validation step, four subsets were used to learn and train a model, while the rest one was used to validate the model. This procedure was repeated five times such that every subset was used in the training and was validated in the testing.

The stepwise feature selection works by training an original SVM with an initial feature set (OFC) for the first round. Then in the next round, one feature will be removed from the initial feature set once a time. If the accuracy of the resulting SVM classifier achieved a better accuracy, such feature would be removed. This stepwise feature selection process was repeated until the accuracy no longer increased. Through this process, more important and informative features can be identified. The detailed procedures are depicted in Figure 1.

F. SVM classifiers

Support vector machine (SVM) is a sophisticated supervised machine learning technique based on statistical learning theory. SVM has been widely used in bioinformatics, such as protein-protein interaction prediction and domain linker prediction [45]. For the implementation of SVM in this study, we used the LIBSVM package [46]. We selected the radial basis function (RBF) as the kernel function, and employed ‘grid-search’ to optimize the SVM parameters, i.e. γ of the RBF kernel and the regularization parameter C based on 5-fold cross-validation tests. C and γ were set within the range of 2^{-8}-2^{8}. We used the ‘one-against-one’ (pair-wise) method to train the multi-class (three-class) SVM classifiers.

We built four SVM classifiers to predict the three classes and each individual class of SAPs. The four classifiers are thus termed as SVM^{3class}, SVM^{CSM}, SVM^{SVD} and SVM^{SVP}, respectively. Here, SVM^{3class} represents a 3-class SVM classifier for predicting CSM, SVD and SVP. SVM^{CSM} denotes a binary SVM classifier, where the CSM data were trained as positives and the remaining two classes of SVD and SVP were merged as negatives. SVM^{SVD} and SVM^{SVP} have similar meanings as SVM^{CSM}, for which SVD and SVP were in turn used as positives.

G. Performance Evaluation

We used Sensitivity (SN), Specificity (SP), Accuracy (ACC) and the Matthew’s correlation coefficient (MCC) to evaluate predictive performance of our method.

The Sensitivity (SN) is defined as:

$$SN = TP/(TP + FN)$$

The Specificity (SP) is defined as:

$$SP = TN/(TP + FP)$$

The overall Accuracy (ACC) is defined as:

$$ACC = (TP + TN)/(TP + TN + FP + FN)$$

The Matthew’s correlation coefficient (MCC) [43] is defined as:

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

where *TP* is the number of true positives, *TN* is the number of true negatives, *FP* is the number of false positives and *FN* is the number of false negatives, respectively.

III. RESULTS

A. Compilation of datasets

We followed the same procedures in previous work to collect variants data [9], i.e. from the following resources: 1) COSMIC database [23]; 2) UniProt human variants [48, 49] and 3) Ensembl human variation database [24] (See Materials and Methods for more details). The residue positions of variants from the source data, namely, the residue positions in protein sequences in UniProt, were mapped onto the corresponding locations in three-dimensional PDB structures if available [50, 51]. TABLE IV in the Supplementary Material shows the statistics of the collected variants from the resources data, the collected variants from the resources data, with mapping onto the UniProt sequence and PDB structure levels, respectively. Finally, we removed the overlapping variants data in CSM, SVD and SVP datasets.

B. Sliding window size selection and sequence homology reduction

We used the PSSM feature as input to train the SVM classifiers in order to determine the optimal sliding window size of *L*, by using the training sets with varying sequence identity (SI) levels, as clustered by Cd-hit [40]. Figure 2 shows the change of accuracies in relation to the SI levels and the sliding window size *L*. As can be seen, ACC increased by 1.5% with the SI decreasing from 100 to 40%. It is worth noting that ACC is not the only performance measure of the SVM classifiers in this study. In particular, for a highly unbalanced dataset, higher ACC does not always mean that the prediction performance of a predictor is satisfactory. Thus, in order to comprehensively evaluate the performance of classifiers, we also used other measures such as Sensitivity, Specificity and MCC. Secondly, the ACC does not necessarily correspond to the SI. In other words, lower SI level does not necessarily mean lower ACC would be achieved. In view of this, we think it is reasonable that the ACC increased by 1.5% with the SI decreasing from 100 to 40%. The window size of *L*=3 and SI=40% led to the overall highest ACC of 85%. In the following analysis, we then fixed the local window size at *L*=3 and used the training dataset clustered at the SI level of 40% to build the SVM classifiers and evaluate prediction performance.

C. Optimal feature selection by random forest

The optimal features were selected in two steps. In the first step, RF was used to evaluate the importance of a total of 280 features using the mean MDGI Z-Score. Finally, eighteen features with the mean MDGI Z-Score >1.0 were selected as the optimal feature candidates (OFCs) (Figure 3). The feature with the highest mean MDGI Z-Score is the structural distance between the mutation position and the DNA binding site, with Z-score of 9.393. Interestingly, the sequence distance between the mutation position and the DNA binding site also has a high Z-Score of 7.107.

We compared the MDGI Z-scores of 18 selected features in the CSM, SVD and SVP datasets and performed the ANOVA analysis [52, 53]. This analysis provides a statistical test of whether or not the means of several sources are equal and thus is useful for comparing the means of more than two samples. The results are shown in TABLE I. Most selected features are significantly different in different types of mutant datasets, with the *P*-value <<0.05. The only exception is that the SIFT_3 feature has a *P*-value of 0.0606, which is slightly larger than 0.05.

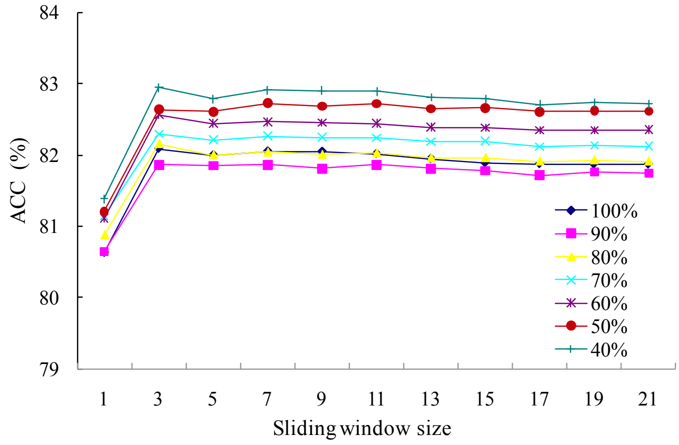

Figure 2. Correspondence between the overall Accuracy (ACC) and sequence identity (SI) levels, based on different sliding window sizes (*L*).

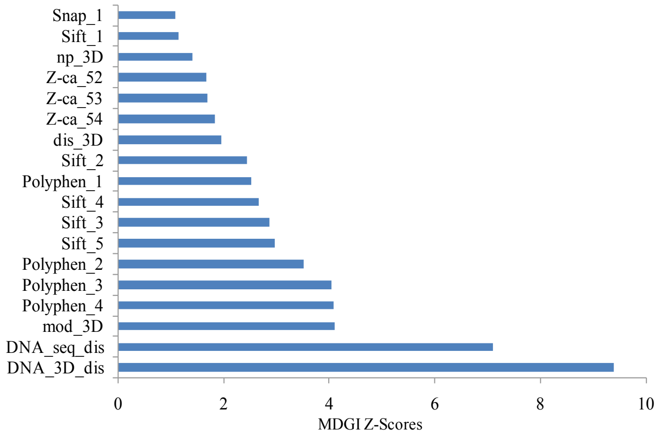

Figure 3. Ranking of the optimal feature candidates based on the MDGI Z-Scores that indicate their importance to performance improvement.

.

TABLE I. THE AVERAGE VALUES AND STANDARD DEVIATIONS FOR THE CSM, SVD AND SVP DATASETS.

Name	Annotation	MDGI Z-Score	CSM	SVD	SVP	*P*-value
			Average value±SD			
DNA_3D_dis	3D distance between variant mutation position and DNA_BIND site	9.39	20.58±6.05	26.04±8.93	59.91±26.33	2.20E-16
DNA_seq_dis	Sequence distance between variant mutation position and DNA_BIND site	7.11	100.18±51.80	170.57±104.15	255.65±120.64	2.20E-16
mod_3D	3D distance between variant and MOD site	4.11	20.425±14.08	40.58±24.49	47.041±30.07	2.20E-16
Polyphen_4	Polyphen2 true positive rate	4.07	0.45±0.37	0.38±0.37	0.79± 0.31	2.20E-16
Polyphen_3	Polyphen2 false positive rate	4.04	0.096±0.19	0.07±0.14	0.40± 0.40	2.20E-16
Polyphen_2	Polyphen2 probability	3.52	0.81±0.34	0.85±0.31	0.37±0.43	2.20E-16
Sift_5	Sequences at Position	2.97	171.71±64.47	173.44±136.40	125.05±123.26	2.20E-16

Sift_3	Median Information	2.86	2.89±0.23	2.88±0.29	2.86±0.41	0.061
Sift_4	Sequences at Position	2.67	165.98±64.54	170.10±136.41	119.54±122.76	2.20E-16
Polyphen_1	Polyphen2 prediction	2.52	/	/	/	/
Sift_2	SIFT Score	2.44	0.10±0.21	0.054±0.15	0.29±0.33	2.20E-16
dis_3D	3D distance between variant mutation position and origin of coordinates	1.95	32.06±25.60	54.33±32.65	55.78±31.78	2.20E-16
Z-ca_54	The Z coordinate of Cα in the residue after variant mutation residue	1.69	-3.41±26.09	20.36±35.86	20.84±33.45	2.20E-16
Z-ca_53	The Z coordinate of Cα in the mutation residue of variant	1.83	-3.46±26.13	20.37±35.91	21.16±33.61	2.20E-16
Z-ca_52	The Z coordinate of Cα in the residue before variant mutation residue	1.67	-3.30±25.87	20.20±35.99	21.15±33.71	2.20E-16
np_3D	3D distance between variant and NP_BIND site	1.42	67.56±47.92	49.31±33.44	46.14±33.94	7.66E-4
Sift_1	SIFT prediction	1.15	/	/	/	/
Snap_1	SNAP prediction	1.08	/	/	/	/

D. *Stepwise feature selection and SVM parameter optimization*

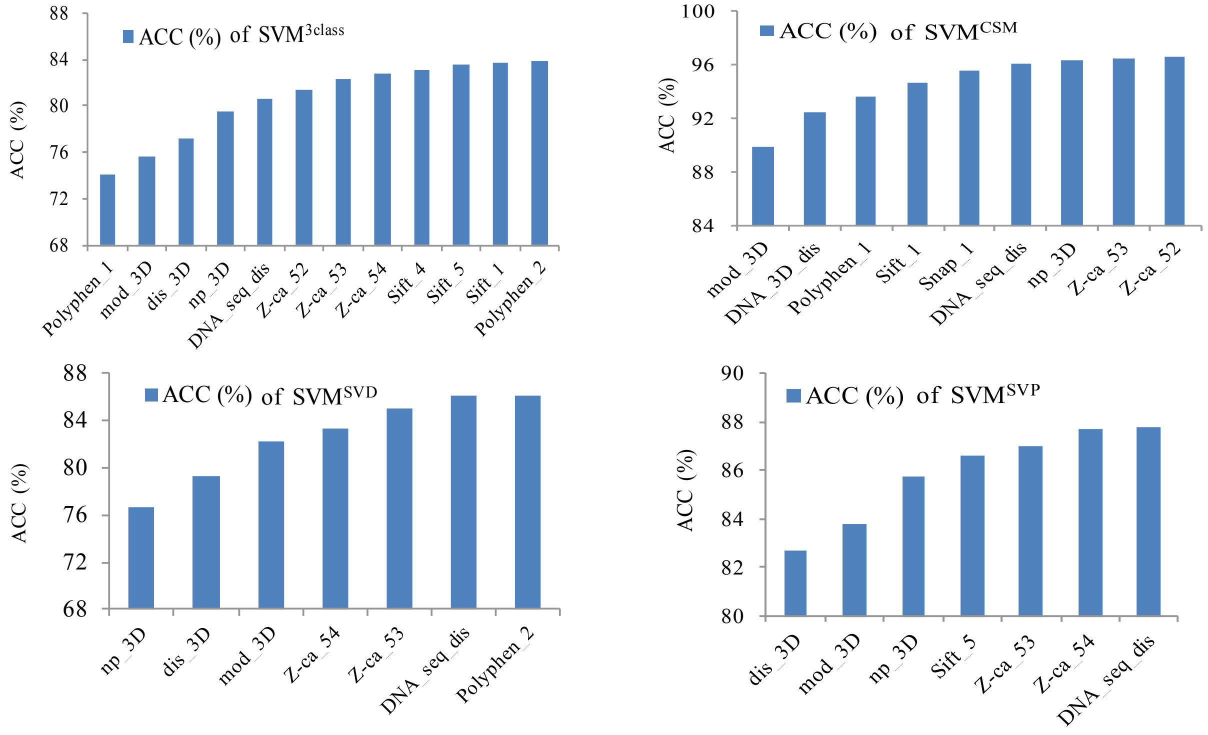

Figure 4. Performance improvement of the SVM classifiers during the stepwise feature selection.

The second step is a stepwise feature selection. If the removal of a feature leads to a higher prediction accuracy, that feature will be removed from the feature set. By iteratively examining and removing the redundant and less informative features in the initial feature set, it is expected that the prediction performance can be enhanced during this process. Figure 4 showed the resulting performance based on stepwisely selected features for four SVM classifiers. We can see that the SVM^{3class} classifier achieved the highest accuracy of 83.78%. It was built after 12 rounds of stepwise feature selection and as a result, 12 features were removed. In the case of SVM^{CSM}, it achieved an accuracy of 96.65%, with 9 features removed. In the case of SVM^{SVD}, it removed 7 features and attained the highest accuracy of 86.16%. SVM^{SVP} also removed a few less informative features and achieved a prediction accuracy of 87.77%.

We further calculated the prediction performance of the SVM classifiers trained with feature subgroups with various levels of Z-Scores and compared with the SVM classifier trained based on stepwise feature selection. The results indicate that the classifier based on stepwise feature group selection achieved the overall best ACC than any other feature subgroups, with Z-Scores larger than 1, 2 and 3, respectively (Supplementary Table V). Depending on the SAPs class, different feature groups have different influence on the prediction performance. For example, the SVM^{3class} classifiers based on the feature groups OFC-2 and OFC-3 achieved almost the same ACCs, while the SVM^{SVP}

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

classifier based on OFC-2 attained better accuracy than that based on OFC-3. These results indicate that feature selection by Z-Scores only may overlook important complementary features and inclusion of the seemingly ‘useless’ features may become useful for improving the prediction performance.

We optimized the parameters (C, γ) of SVM using the ‘grid.py’ module in LIBSVM. TABLE II presents the prediction performance of four SVM classifiers with the optimized parameters (C, γ). They changed after each stepwise feature selection, with the accuracy consistently improved. The ACC increased from 83.78 to 84.97%, from 96.65 to 96.93%, from 86.16 to 86.98% and from 87.77 to 88.24%, for the SVM^{3class}, SVM^{CSM}, SVM^{SVD} and SVM^{SVP} classifiers, respectively.

TABLE II. THE IMPROVEMENT OF ACCURACY BY PARAMETER OPTIMIZATION

	ACC (%)	Original C	Original γ	ACC (%)	Final C	Final γ
SVM^{3class}	83.78	1.0	0.015625	84.97	1.0	2.0
SVM^{CSM}	96.65	2.0	0.0625	96.93	1.0	0.015625
SVM^{SVD}	86.16	8.0	0.015625	86.98	2.0	0.0039063
SVM^{SVP}	87.77	1.0	0.0078125	88.24	2.0	0.0039063

E. Performance improvement by optimal feature selection and parameter optimization

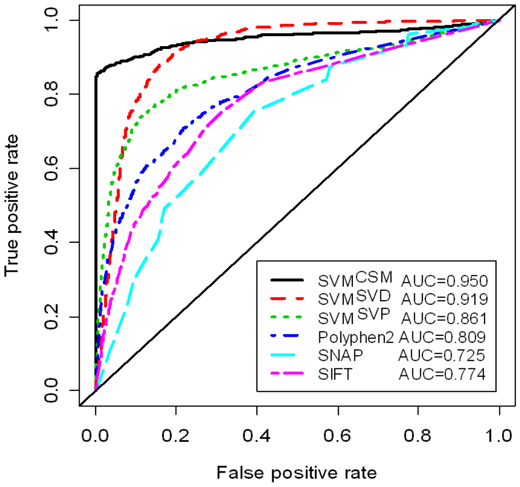

Figure 5. The ROC curves of four SVM classifiers based on 5-fold cross-validation tests of 5,109 variants, trained with the selected optimal features

To further evaluate the predictive performance, we plotted the receiver operating characteristic (ROC) curves using the “ROCR” package [54], as shown in Figure 5. The area under the ROC curve (AUC) is a measure of the overall quality of the prediction, incorporating both the Sensitivity and Specificity measures. The uppermost curve with the largest AUC indicates the best prediction model. From Figure 5, we can see that the classifier SVM^{CSM} has the best prediction performance with the AUC of 0.9504. The SVM^{SVD} classifier has an AUC value of 0.9190, while the SVM^{SVP} has an AUC of 0.8612. As the SVM^{3class} classifier is a three-class model, it is not applicable to generate the ROC curve.

F. Comparison with other prediction tools

We next compared the predictive performance of our SVM predictor with other previous prediction tools. SIFT is a program that uses sequence homology to predict whether a substitution affects protein function, and its output has five values [37, 55, 56]. The first value is “tolerant” or “deleterious”, which is similar to our SVM^{SVP} predictor as both are binary classifiers. In our SVM^{SVP} classifier, the positive class is the SVP subset, while the negative class is the CSM and SVD subsets. The performance comparison is shown in TABLE III. We can see that SVM^{SVP} achieved the Sensitivity, Specificity, Accuracy and MCC scores of 76.60%, 90.19%, 88.24% and 0.593, respectively, while SIFT achieved the Sensitivity, Specificity, Accuracy and MCC of 40.92%, 90.04%, 75.25% and 0.360, respectively. This suggests that our classifier provides a better performance than SIFT, with the accuracy significantly increased from 75.25 to 88.24%. However, the Specificity (90.19%) of our method is close to SIFT (90.04%). The Specificity measures the proportion of negatives that are correctly identified. Thus, our method and SIFT have the comparable capabilities to predict the negatives. Sensitivity (also referred to as Recall in the information retrieval field) measures the proportion of actual positives that are correctly identified as such. Hence, our method is more accurate than SIFT, with the Sensitivity improved from 40.92 to 76.60%. SNAP is another tool that predicts the functional effects of single amino acid substitutions. It outputs the value of “neutral” or “non-neutral”. To compare our method with SNAP, we calculated the SN, SP, ACC and MCC based on 5-fold cross-validation tests and listed the results in TABLE III. It can be seen that our SVM^{SVP} classifier achieved a higher prediction performance than SNAP in terms of the SN, SP, ACC and MCC measures. Similar to SIFT, SNAP also has a higher specificity of 86.33%, in contrast to the Sensitivity of 50.82%.

Polyphen2 is a tool for predicting damaging effects of missense mutations. It divides the variants into three categories- “benign”, “probably damaging” and “possibly damaging”. In this study, we defined the “probably damaging” and “possibly damaging” as the negatives. The resulting SN, SP, ACC and MCC of Polyphen2 are listed in the row Polyphen2^{2class} in Table III. We can see that the SN, SP, ACC and MCC of Polyphen2^{2class} are higher than those of SIFT and SNAP, but are lower than those of our SVM^{SVP} classifier. In addition, if we did not combine the “probably damaging” and “possibly damaging” and instead defined the “probably damaging” as SVD-class, and “possibly damaging” as CSM-class, then Polyphen2 would become a three-class predictor. We then calculated the ACC of Polyphen2^{3class}, which was 61.74% and was much lower than our SVM^{3class} predictor (Table III). In conclusion, our method provides a favorable performance compared with the other three tools.

In this study, the overfitting issue is potentially alleviated by the following three strategies we adopted: 1) The training set used in this study was mapped to the PDB structures with sequence identity of 40%. The sequence and structural features were extracted from this non-redundant structural dataset; 2) We used a stepwise feature selection to select the optimal candidate features. The selected features this way constitute a reliable and robust feature subset; 3) We performed 100 iterations of Random Forest algorithm to calculate the Z-scores of each feature, and finally selected 18 features with Z-scores larger than 0. Therefore, based on the above strategies, we suggest that the overfitting issue is effectively alleviated and relatively less severe.

TABLE III. PERFORMANCE COMPARISON OF OUR SVM^{3CLASS}, SVM^{CSM}, SVM^{SVD}, SVM^{SVP} CLASSIFIERS WITH SIFT, SNAP, POLYPHEN2^{2CLASS} AND POLYPHEN2^{3CLASS}. THE PREDICTION PERFORMANCE WAS EVLUATED USING 5-FOLD CROSS-VALIDATION TESTS.

Method	SN (%)	SP (%)	ACC (%)	MCC
SVM^{3class}	-[a]	-	84.97	-
SVM^{CSM}	98.44	96.63	96.93	0.898
SVM^{SVD}	86.65	87.64	86.98	0.723
SVM^{SVP}	76.60	90.19	88.24	0.593
SIFT	40.92	90.04	75.25	0.360
SNAP	50.82	86.33	80.81	0.340
polyphen2^{2class}	55.89	89.84	83.00	0.464
polyphen2^{3class}	-	-	61.74	-

[a] "-" denotes that the prediction result at this specificity level is not available by this tool.

IV. DISCUSSION

There are a number of bioinformatic approaches developed to predict the functional impact of SAPs, classified as either 'deleterious' or 'neutral' in the training/validation stages. These include empirical rules and machine learning techniques such as decision tress, support vector machines, neural networks, etc. All machine learning methods require a dataset of SAPs data for model training and error rate estimation [2]. However, a critical question to address is how to select the appropriate training data. In a recent work by Care *et al.*, the authors showed that differences in training datasets derived by different ways can give rise to trained classifiers with varying error rates, thereby making some of them less ideal for SAPs prediction [2]. We appreciated this enlightening work and have carefully curated and checked the consistency of collected structural mutants by mapping them onto UniProt and carefully removing the unreliable data.

Different from the majority of previous works, in this study, we divided the human SAPs mutants into three detailed classes rather than two conventional classes (being deleterious or neutral). This provides an intuitively and conceptually better way to characterize the differences between cancer somatic and Mendelian disease-related variants. By doing so, some important and critical features between different SAPs classes can be better extracted and identified (as listed in TABLE I). For example, we found that the distance between the variant and some function sites is an important descriptor for predicting CSM, SVD and SVP, which is significantly different between different mutant types by the ANOVA statistical test.

Another important feature, DNA_3D_dis, describes the 3D distance between the variant mutation position and DNA binding sites. Inclusion of this feature is crucial for our SVM classifiers. This is understandable as the closer to the DNA binding site, the larger its chance to affect the DNA binding site, and consequently the more likely mutations at such position will change the DNA expression, thus making it an important descriptor. Similarly, other distance descriptors are important for the prediction of mutational effects, possibly due to the similar reason.

Z-ca is the coordinate of Cα atom of an amino acid residue. It can be seen that the average Z-ca of the CSM class is about -3.4, which is remarkably different from that of SVD and SVP (average of 20~21, *P*-value of 2.20E-16). The reason is that the majority of the variants in the CSM category (786 out of 964) are actually from a single protein, i.e. P04637 (PDB_ID: 3D06_A).

Compared with the 18 selected optimal features, secondary structure features are not included in the final feature set. This suggests that secondary structure features are less important in contrast to other optimal features that were finally selected.

To compare our method with SIFT, SNAP, Polyphen2, we applied these tools to the same datasets and calculated their prediction performances in this study. We then compared these tools with our method. SIFT is based on the principles of protein evolution and most of its features are sequence-derived. It was previously reported that SIFT could correctly predict that 69% of the substitutions associated with a certain type of disease [55]. In our study, the ACC of SIFT was 75.25% and the MCC was 0.360.

SNAP and Polyphen2 used many features including sequence, structural, function site feature and many others. SNAP used neural networks, while Polyphen2 used the Naïve Bayes approach. As a comparison, our method used the random forest algorithm to select optimal feature candidates and applied SVM to train the prediction models based on the optimally selected features and SVM parameters. As a result, our method provides a favourable performance in comparison with the former three tools. However, the contribution or relative importance of each selected features to performance improvement of a SVM classifier is different, depending on a particular SAPs category (Figure 4). For instance, the scores of SIFT, SNAP and Polyphen2 are critical for SVM^{SVP}, only the sift_5 feature was removed from the final optimal features (Figure 4). Although sift_1, snap_1, and polyphen_1 were all included in the final optimal feature set in all three SAPs classes, they are particularly more important for improving the prediction accuracy of SVM^{SVP} and SVM^{SVD}, but are less important for SVM^{CSM}. The reason might be that although our datasets were divided into three classes, SIFT, SNAP and Polyphen2 did not specifically train a third additional predictor for the CSM class, i.e. the cancer somatic mutations. In a sense, the training dataset of SVM^{SVP} is more similar to that of SIFT, SNAP and Polyphen2, as they were specifically developed to predict the SVP class. Therefore, features such as sift_1, snap_1, and polyphen_1 are more important for SVM^{SVP}. Additionally, SNPs3D is another similar tool for predicting SAPs [57]. Nevertheless, as its coverage to our datasets is very low (909/5109), we did not include it in our comparison.

V. CONCLUSION

In this work, we have collected and selected 5,109 human variants from three public databases and categorized them into three classes, i.e. CSM, SVD and SVP. Important sequence and structural features in our datasets were extracted and selected using a random forest algorithm. Moreover, we have stepwisely selected the optimal features and four built predictors SVM^{3class}, SVM^{CSM}, SVM^{SVD} and SVM^{SVP} achieved the ACC of 84.97, 96.93, 86.98 and 88.24%, respectively. The ACC were further improved after SVM parameter optimization. To validate our approach, we compared it with three other tools SIFT, SNAP and Polyphen2 and showed that our method provides a favorable performance than these three methods in terms of SN, SP, ACC and MCC measures. We expect that our approach offers useful insights in predicting the functional impact of different types of SAPs with more available 3D structure data.

ACKOWLEDGMENTS

This work was supported by grants from the National Health and Medical Research Council of Australia (NHMRC) and the Hundred Talents Program of the Chinese Academy of Sciences (CAS). This work was also supported by the Innovation Program of Shanghai Municipal Education Commission (10ZZ17), the Knowledge Innovative Program of CAS (KSCX2-EW-G-8) and Tianjin Municipal Science & Technology Commission (10ZCKFSY05600). JS is an NHMRC Peter Doherty Fellow and the Recipient of the Hundred Talents Program of the CAS.

REFERENCES

[1] Collins, F.S., L.D. Brooks, and A. Chakravarti, A DNA Polymorphism Discovery Resource for Research on Human Genetic Variation. Genome Research, 1998. 8(12): p. 1229-1231.

[2] Care, M.A., et al., Deleterious SNP prediction: be mindful of your training data! Bioinformatics, 2007. 23(6): p. 664-672.

[3] Ye, Z.-Q., et al., Finding new structural and sequence attributes to predict possible disease association of single amino acid polymorphism (SAP). Bioinformatics, 2007. 23(12): p. 1444-1450.

[4] Halushka, M.K., et al., Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. Nat Genet, 1999. 22(3): p. 239-247.

[5] Bromberg, Y., G. Yachdav, and B. Rost, SNAP predicts effect of mutations on protein function. Bioinformatics, 2008. 24(20): p. 2397-2398.

[6] Cargill, M.e.a., Characterization of single-nucleotide polymorphisms in coding regions of human genes Nat. Genet, 1999. 22: p. 231–238.

[7] Sunyaev, S., V. Ramensky, and P. Bork, Towards a structural basis of human non-synonymous single nucleotide polymorphisms. Trends in genetics, 2000. 16(5): p. 198-200.

[8] Wang, Z. and J. Moult, SNPs, protein structure, and disease. Human Mutation, 2001. 17(4): p. 263-270.

[9] Gong, S. and T.L. Blundell, Structural and Functional Restraints on the Occurrence of Single Amino Acid Variations in Human Proteins. Plos One, 2010. 5(2): e9186.

[10] Botstein, D. and N. Risch, Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat Genet, 2003.

[11] Liu, Q.-R., et al., Addiction molecular genetics: 639,401 SNP whole genome association identifies many "cell adhesion" genes. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, 2006. 141B(8): p. 918-925.

[12] Li, Y.Z., et al., Predicting disease-associated substitution of a single amino acid by analyzing residue interactions. BMC Bioinformatics, 2011. 12.

[13] Huang, T., et al., Prediction of Deleterious Non-Synonymous SNPs Based on Protein Interaction Network and Hybrid Properties. Plos One, 2010. 5(7): e11900.

[14] Dobson, R., et al., Predicting deleterious nsSNPs: an analysis of sequence and structural attributes. BMC Bioinformatics, 2006. 7(1): p. 217.

[15] Bao, L. and Y. Cui, Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information. Bioinformatics, 2005. 21(10): p. 2185-2190.

[16] Cai, Z., et al., Bayesian approach to discovering pathogenic SNPs in conserved protein domains. Human Mutation, 2004. 24(2): p. 178-184.

[17] Ferrer-Costa, C., M. Orozco, and X. de la Cruz, Sequence-based prediction of pathological mutations. Proteins: Structure, Function, and Bioinformatics, 2004. 57(4): p. 811-819.

[18] Karchin, R., et al., LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. Bioinformatics, 2005. 21(12): p. 2814-2820.

[19] Krishnan, V.G. and D.R. Westhead, A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function. Bioinformatics, 2003. 19(17): p. 2199-2209.

[20] Ramensky, V., P. Bork, and S. Sunyaev, Human non‐synonymous SNPs: server and survey. Nucleic Acids Research, 2002. 30(17): p. 3894-3900.

[21] Yue, P. and J. Moult, Identification and Analysis of Deleterious Human SNPs. Journal of Molecular Biology, 2006. 356(5): p. 1263-1274.

[22] Bairoch, A., et al., The Universal Protein Resource (UniProt). Nucleic Acids Res, 2005. 33: p. D154 - 159.

[23] Forbes, S.A., et al., The Catalogue of Somatic Mutations in Cancer (COSMIC). Curr Protoc Hum Genet, 2008. Chapter 10: p. Unit 10.11.

[24] Chen, Y., et al., Ensembl variation resources. BMC Genomics, 2010. 11(1): p. 293.

[25] Altschul, S.F., et al., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research, 1997. 25(17): p. 3389-3402.

[26] Jones, D.T., Protein secondary structure prediction based on position-specific scoring matrices. Journal of Molecular Biology, 1999. 292(2): p. 195-202.

[27] Cheng, J., et al., SCRATCH: a protein structure and structural feature prediction server. Nucleic Acids Research. 33(suppl 2): p. W72-W76.

[28] Ward, J.J., et al., Prediction and Functional Analysis of Native Disorder in Proteins from the Three Kingdoms of Life. Journal of Molecular Biology, 2004. 337(3): p. 635-645.

[29] Song, J., et al., HSEpred: predict half-sphere exposure from protein sequences. Bioinformatics, 2008. 24(13): p. 1489-1497.

[30] Song, J., et al., Cascleave: towards more accurate prediction of caspase substrate cleavage sites. Bioinformatics, 2010. 26(6): p. 752-760.

[31] Schlessinger, A., et al., Improved Disorder Prediction by Combination of Orthogonal Approaches. Plos One, 2009. 4(2): p. e4433.

[32] Song, J., et al., Predicting disulfide connectivity from protein sequence using multiple sequence feature vectors and secondary structure. Bioinformatics, 2007. 23(23): p. 3147-3154.

[33] Hubbard, S.J.a.T., J.M., 'NACCESS', computer program. Department Biochemistry and Molecular Biology, University College, London., 1993.

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

[34] Fernandez-Escamilla, A.-M., et al., Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. Nat Biotech, 2004. 22(10): p. 1302-1306.

[35] Kabsch, W. and C. Sander, Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers, 1983. 22(12): p. 2577-2637.

[36] Sunyaev, S.R., et al., PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. Protein Engineering, 1999. 12(5): p. 387-394.

[37] Ng, P.C. and S. Henikoff, Predicting deleterious amino acid substitutions. Genome Res, 2001. 11(5): p. 863-74.

[38] Bromberg, Y. and B. Rost, SNAP: predict effect of non-synonymous polymorphisms on function. Nucleic Acids Research, 2007. 35(11): p. 3823-3835.

[39] Adzhubei, I.A., et al., A method and server for predicting damaging missense mutations. Nat Meth, 2010. 7(4): p. 248-249.

[40] Li, W. and A. Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics, 2006. 22(13): p. 1658-1659.

[41] Liaw, A. and M. Wiener, Classification and Regression by randomForest. R news, 2002. 2: p. 18-22.

[42] Liu, Z.-P., et al., Prediction of protein–RNA binding sites by a random forest method with combined features. Bioinformatics, 2010. 26(13): p. 1616-1622.

[43] Chen, X.-W. and M. Liu, Prediction of protein–protein interactions using random decision forest framework. Bioinformatics, 2005. 21(24): p. 4394-4400.

[44] Wang, L., C. Huang, and J. Yang, Predicting siRNA potency with random forests and support vector machines. BMC Genomics, 2010. 11(Suppl 3): p. S2.

[45] Ebina, T., H. Toh, and Y. Kuroda, DROP: an SVM domain linker predictor trained with optimal features selected by random forest. Bioinformatics, 2011. 27(4): p. 487-494.

[46] Chang, C.-C. and C.-J. Lin, LIBSVM:a library for support vector machines. Software availabe at http://www.csie.ntu.edu.tw/~cjlin/libsvm. 2001.

[47] Matthews, B.W., Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA) - Protein Structure, 1975. 405(2): p. 442-451.

[48] The UniProt Consortium, The Universal Protein Resource (UniProt) 2009. Nucl. Acids Res., 2009. 37(suppl_1): p. D169-174.

[49] Yip, Y.L., et al., Annotating single amino acid polymorphisms in the UniProt/Swiss-Prot knowledgebase. Human Mutation, 2008. 29(3): p. 361-366.

[50] Berman, H.M., et al., The Protein Data Bank. Nucl. Acids Res., 2000. 28(1): p. 235-242.

[51] Gong, S. and T.L. Blundell, Discarding functional residues from the substitution table improves predictions of active sites within three-dimensional structures. PLoS Comput Biol, 2008. 4(10): p. e1000179.

[52] Bailey, R.A., Design of Comparative Experiments. 2008, Cambridge: Cambridge University Press.

[53] http://en.wikipedia.org/wiki/Analysis_of_variance.

[54] Sing, T., et al., ROCR: visualizing classifier performance in R. Bioinformatics, 2005. 21(20): p. 3940-3941.

[55] Ng, P.C. and S. Henikoff, Accounting for human polymorphisms predicted to affect protein function. 2002. 12(3): p. 436-46.

[56] Ng, P.C. and S. Henikoff, SIFT: predicting amino acid changes that affect protein function. Nucl. Acids Res., 2003. 31(13): p. 3812-3814.

[57] Yue, P., E. Melamud, and J. Moult, SNPs3D: Candidate gene and SNP selection for association studies. BMC Bioinformatics, 2006. 7(1): p. 166.

Supplementary Material

TABLE IV. Three types of mutation variants and their statistics

Source	Type	Abbreviation	Number of distinct variants from the sources	Mappped to UniProt	Mapped to PDB	Further refined	40% Sequence identity
UniProt	Disease	SVD	19270	19270	4495	3677	3153
Ensembl	Verified SNPs	SVP	43906	25425	1734	1242	992
COSMIC	Cancer mutations	CSM	11306	3382	1455	1029	964

TABLE V. Comparasion of ACC (%) of different feature groups

Feature group	Abbreviation	ACC (%)			
		SVM^{3class}	SVM^{CSM}	SVM^{SVD}	SVM^{SVP}
MDGI Z-Score>1.0	OFC-1	72.20	88.54	75.08	81.41
MDGI Z-Score >2.0	OFC-2	80.76	92.46	82.62	85.91
MDGI Z-Score >3.0	OFC-3	80.74	95.69	81.89	80.58
Optimal features by stepwise feature selection	OFC-s	83.78	96.65	86.16	87.77

978-1-4577-1666-9/11/$26.00 ©2011 IEEE