BIOINFORMATICS ORIGINAL PAPER
Vol. 21 no. 10 2005, pages 2185–2190
doi:10.1093/bioinformatics/bti365

Genome analysis

Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information

Lei Bao and Yan Cui*

Department of Molecular Sciences, Center of Genomics and Bioinformatics, University of Tennessee Health Science Center, 858 Madison Avenue, Memphis, TN 38163, USA

Received on October 20, 2005; revised on February 17, 2005; accepted on February 28, 2005
Advance Access publication March 3, 2005

ABSTRACT

Motivation: There has been great expectation that the knowledge of an individual's genotype will provide a basis for assessing susceptibility to diseases and designing individualized therapy. Non-synonymous single nucleotide polymorphisms (nsSNPs) that lead to an amino acid change in the protein product are of particular interest because they account for nearly half of the known genetic variations related to human inherited diseases. To facilitate the identification of disease-associated nsSNPs from a large number of neutral nsSNPs, it is important to develop computational tools to predict the phenotypic effects of nsSNPs.

Results: We prepared a training set based on the variant phenotypic annotation of the Swiss-Prot database and focused our analysis on nsSNPs having homologous 3D structures. Structural environment parameters derived from the 3D homologous structure as well as evolutionary information derived from the multiple sequence alignment were used as predictors. Two machine learning methods, support vector machine and random forest, were trained and evaluated. We compared the performance of our method with that of the SIFT algorithm, which is one of the best predictive methods to date. An unbiased evaluation study shows that for nsSNPs with sufficient evolutionary information (with not $<$10 homologous sequences), the performance of our method is comparable with the SIFT algorithm, while for nsSNPs with insufficient evolutionary information ($<$10 homologous sequences), our method outperforms the SIFT algorithm significantly. These findings indicate that incorporating structural information is critical to achieving good prediction accuracy when sufficient evolutionary information is not available.

Availability: The codes and curated dataset are available at http://compbio.utmem.edu/snp/dataset/

Contact: ycui2@utmem.edu

Supplementary information: The curated dataset is available at http://compbio.utmem.edu/snp/dataset/

INTRODUCTION

In humans, ~90% of sequence variants are differences in single bases of DNA, called single nucleotide polymorphisms (SNPs) (Collins *et al*., 1998). Among them, non-synonymous SNPs (nsSNPs) that lead to an amino acid change in the protein product are most relevant to human inherited diseases (Stenson *et al*., 2003). Whereas a large number of nsSNPs may be functionally neutral, others may cause deleterious effects on protein functions and are hence disease associated. Given the vast number of nsSNPs discovered (Irizarry *et al*., 2000; Fredman *et al*., 2002), a major challenge is to predict which of them are potentially disease associated. Recent studies have discovered a variety of potential predictors discriminating disease-associated nsSNPs from neutral nsSNPs. Empirical rule-based and machine learning approaches were used to classify these two types of nsSNPs. Empirical rules discriminating disease-associated and neutral nsSNPs were derived based on structural information (Wang and Moult, 2001), evolutionary information (Ng and Henikoff, 2001) or both (Sunyaev *et al*., 2001). Other recent studies (Chasman and Adams, 2001; Saunders and Baker, 2002; Krishnan and Westhead, 2003) developed classification models automatically learned from the training data. Except for the work of Wang and Moult (2001), all the mentioned studies used some form of position-specific evolutionary information contained in the multiple sequence alignments. The prediction accuracy depends heavily on the existence of a sufficient number of homologous sequences. Saunders and Baker (2002) showed that the prediction accuracy decreased significantly when fewer than 5–10 homologous sequences are available. Incorporating structural information is crucial in such cases (Saunders and Baker, 2002). Here we developed classifiers combining structural and evolutionary information to discriminate disease-associated nsSNPs from neutral nsSNPs. We prepared a curated training dataset from the UniProt knowledgebase (Apweiler *et al*., 2004). This dataset consists of natural nsSNPs, in contrast to *in vitro* mutational data used in previous studies (Chasman and Adams, 2001; Krishnan and Westhead, 2003). The structural environments (Bowie *et al*., 1991) and substitution properties of nsSNPs were used as predictors. We applied two machine learning methods, support vector machine (SVM) (Vapnik, 1998) and random forest (RF) (Breiman, 2001). We showed that for nsSNPs with insufficient homologous sequences, our method outperformed the SIFT algorithm (Ng and Henikoff, 2003) on account of the incorporated structural information. In the cases where sufficient homologous sequences were available, the performance of our method was comparable with the SIFT algorithm.

SYSTEMS AND METHODS

Dataset

Human nsSNPs were extracted via analysis of the VARIANT field in the corresponding Swiss-Prot entries (Apweiler *et al*., 2004). nsSNPs annotated as

*To whom correspondence should be addressed.

'Disease' are disease associated, and those annotated as 'Polymorphism' are neutral nsSNPs. Major histocompatibility complex proteins and membrane proteins were excluded. We focused our analysis on nsSNPs with experimentally determined structure or structural homologs. Each nsSNP variant was searched against the ASTRAL database (Chandonia *et al.*, 2004) using the BLASTP program (Altschul *et al.*, 1990) to find representative homologous 3D structures. Hits were retained if they met the following criteria:

(1) sequence identity to the query sequence was not $<30\%$, for the conservation of basic structural characteristics,

(2) the number of identical amino acids was not <20,

(3) gap content was $<15\%$ and

(4) the hit sequence had the same amino acid as the query sequence at the substitution site.

In case of multiple representative PDB entries, the one with highest sequence identity was chosen. These filters resulted in 532 neutral nsSNPs within 305 genes and 3686 disease-associated nsSNPs within 323 genes. To evaluate the discriminative power of our method on nsSNPs with insufficient evolutionary information, we split all the 4218 nsSNPs into two sets according to the number of homologous sequences. 4013 nsSNPs with not <10 homologous sequences were used as training samples (502 neutral and 3511 disease-associated nsSNPs), while the remaining 205 nsSNPs were used as independent test samples (30 neutral and 175 disease-associated nsSNPs). The datasets are available at http://compbio.utmem.edu/snp/dataset/.

SIFT score

The SIFT program (Ng and Henikoff, 2003) was used to calculate the SIFT score, a score measuring the tolerance of a substitution based on the mutability of the substitution position. SIFT used PSI-BLAST (Altschul *et al.*, 1997) to search against the EMBL non-redundant protein database (Apweiler *et al.*, 2004) for homologous sequences and construct a multiple sequence alignment. The multiple sequence alignment was converted into a position-specific scoring matrix. Each matrix entry P_{ij} is the probability of amino acid j occurring at position i. The P_{ij} was estimated as a weighted average of the observed frequencies at the position and the Dirichlet pseudocounts (Henikoff and Henikoff, 1996). To reduce multiple contributions from closely related members of a sequence family, the sequences were weighted (Henikoff and Henikoff, 1996). SIFT uses an empirical threshold: substitutions with normalized probabilities <0.05 are predicted as deleterious while others are predicted as tolerated.

Predictors

The predictors we used are listed in Table 1. The first three predictors in Table 1 represent the structural environment of a substitution site. The structural environment of each nsSNP was annotated by the ENVIRONMENT program developed by Bowie *et al.* (1991). The program combined three structural parameters (area buried, fraction polar and secondary structure) to define the structural environment of a site. Briefly, the buried area of a residue was determined by placing imaginary solvent spheres around each atom and calculating the difference between the side-chain area covered by solvent-accessible sample points in a protein site and in a Gly-X-Gly tripeptide. The fraction polar of a residue was calculated as the fraction of the number of sample points covered by polar atoms (or exposed to solvent) to the number of total sample points. By setting empirical cutoffs for these two structural parameters, Bowie *et al.* (1991) defined six environment classes: B1, B2, B3, P1, P2 and E (Figure 4 of Bowie *et al.*, 1991). Combining the six environment classes with three-state (helix, sheet and coil) secondary structures gave a total of 18 environment classes. The STRIDE program (Frishman and Argos, 1995) was used to assign the secondary structures. In essence, each position in a 3D structure could be assigned to 1 of the 18 environment classes. It is of importance that different environment classes had different amino acid preferences, as was measured by 3D–1D compatibility scores (Figure 5 of Bowie *et al.*, 1991). To assess the differences between the wild-type (original) and mutated amino acids, we derived a structural environment-specific grouping of the 20 amino acids (Table 2). The grouping of the 20 amino acids was based both on their physicochemical properties and compatibility with the structural environment (Table S1). If the wild-type and mutated amino acids fell into the same group, the indicator of 'change of amino acid group' got a value of 0; otherwise, the indicator got a value of 1. The SIFT score was calculated by the SIFT program as described above.

Table 1. The predictors used

Predictor	Value	Information type	Description
Buried area[a]	Continuous	Structural environment	Indicator of solvent accessibility
Fraction polar[a]	Continuous	Structural environment	Indicator of the environmental polarity
Secondary structure	Categorical	Structural environment	Three-state secondary structure: α-helix, β-sheet and coil
Wild-type amino acid	Categorical	Amino acid identity	The identity of the wild-type residue
Amino acid group change	Categorical	Substitution change	Whether the wild-type and mutated amino acid belong to the same group
SIFT score[b]	Continuous	Substitution change	Whether the substitution is tolerated in a multiple sequence alignment

[a]Predictors developed by Bowie *et al.* (1991).
[b]Predictor developed by Ng and Henikoff (2001).

Table 2. Structural environment-specific grouping of amino acids

Environment[a]	Amino acid grouping
B1H	G VLIM FYW CSTANQDEKRH P
B1S	G VLIM FYW CSTANQDEKRH P
B1C	G VLIM FYW CSTANQDEKRH P
B2H	G VLIM FYW STANDEKR HQC P
B2S	G VLIM FYW CSTANQDEKRH P
B2C	G VLIM FYW CSTANQDEKR H P
B3H	G VLIM FYW KRHQE NDCSTA P
B3S	G VLIM FYW KRHQ NDECSTA P
B3C	G VLIM FYW KRHQN DECSTA P
P1H	G VLI FYW KRHQNDEM C STA P
P1S	G VLI FYW KRHQNDEM C STA P
P1C	G VLI FYW KRHQNDEM C STAP
P2H	G VLI FYW KRHQNDE CSTAM P
P2S	G VLI FYW KRHQNDEST CAM P
P2C	G VLI FYW KRHQNDEP CSTAM
EH	G VLI FYW KRHQNDECSTAM P
ES	G VLI FYW KRHQNDECSTAM P
EC	G VLI FYW KRHQNDECSTAMP

[a]Eighteen structural environments. B, buried; P, partially buried; E, exposed; H, α-helix, S, β-sheet; C, coil. The environmental polarity is denoted by a number following the code, i.e. B2 is more polar than B1 and so on. Please refer to Figure 4 of Bowie *et al.* (1991) for details.

Evaluation of classification accuracy

Classification accuracy was evaluated using a 10-fold cross-validation. The data were randomly split into 10 equal parts. One was used for testing and the others for training. The procedure was repeated 10 times so that each sample was used exactly once for testing. The results of five independent 10-fold cross-validation experiments were averaged to get a fair evaluation. Since the dataset contains many more disease-associated nsSNPs (positives) than neutral nsSNPs (negatives), we used Matthew's correlation coefficient (MCC) (Matthews, 1985) to evaluate the performance,

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

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. MCC has been widely used as an evaluation criterion of machine learning performance in bioinformatics studies (Bhasin and Raghava, 2004; Chen *et al*., 2004). When there is an obvious disparity in the number of positive samples and negative samples, MCC is usually a better evaluation criterion of performance than the overall accuracy (TP + TN), because in the extreme case when all the samples are assigned to the majority class (when all nsSNPs are labeled as disease-associated), overall accuracy may still be high. An alternative solution to this problem is the balanced error rate (BER),

$$\mathrm{BER} = \left(0.5 \times \frac{\mathrm{FN}}{\mathrm{TP}+\mathrm{FN}}\right) + \left(0.5 \times \frac{\mathrm{FP}}{\mathrm{TN}+\mathrm{FP}}\right)$$

This measure assumes equal weights of the prediction errors for positive and negative samples. Finally, receiver operating characteristic (ROC) curves (Zhou *et al*., 2002) were used to compare the performances graphically. Samples are first ranked according to their decision function values. By varying the decision cutoffs, ROC curve plots true positive rate [TP/(TP + FN)] against false positive rate [FP/(TN + FP)]. A good classifier is characteristic for its ROC curve climbing rapidly toward upper left hand corner of the graph. SVM and SIFT algorithm both output decision function value for each sample. For RF, we use the fraction of votes for the positive class of each sample as its decision function value.

Support vector machine

Support vector machine (SVM) (Vapnik, 1998) is a classifier seeking an optimal hyperplane to separate two classes of samples. SVM uses kernel functions to map original data to a feature space of higher dimensions and locate an optimal separating hyperplane there. We used SVM-light, an implementation of the SVM algorithm by Joachims (1999). The performance of SVM is mainly controlled by the kernel function and the regularization parameter C. The kernel function determines the sample distribution in the feature space. Regularization parameter C is used to trade between training errors and larger hyperplane margins. A larger C value assigns a higher penalty to the training errors. Polynomial kernels functions with powers of 1, 2 or 3 and radial basis kernels (g = 0.01, 0.1, 1.0, 5.0 and 10.0) were tested in combination with different C values (0.1, 1.0, 5.0, 10.0 and 50.0) to tune for good performance.

Random forest

Random forest (RF) is a classifier consisting of an ensemble of tree-structured classifiers (Breiman, 2001). RF takes advantage of two powerful machine learning techniques: bagging (Breiman, 1996) and random feature selection. In bagging, each tree is trained on a bootstrap sample of the training data, and predictions are made by majority vote of the trees. When using bootstrap samples of the training data, about one-third of the cases are left out, which is called out-of-bag (OOB) data. OOB data can be used to get an unbiased estimate of the classification error during the training process. The details in growing (training) of an individual tree can be found in Breiman *et al.* (1984). RF is a further development of bagging. Instead of using all features, RF randomly selects a subset of features to split at each node when growing a tree. Breiman (2001) deduced an upper bound on the generalization error and concluded that RF does not suffer from the overfitting problem. Several recent studies demonstrated the better performance of RF over other machine learning approaches (Wu *et al*., 2003; Gunther *et al*., 2003; Svetnik *et al*., 2003). We used the R language implementation of RF (Svetnik *et al*., 2003). The number of trees to grow was set to 1000. RF uses a parameter mtry to specify the number of random features to be searched at each tree node. We used cross-validation to determine the best mtry value.

RESULTS

Selected predictors and biological implications

Structural and functional constraints are believed to be the underlying mechanisms that determine the phenotypic effect of an nsSNP. We derived several predictors from the literature and our own studies; the predictors used in this work are listed in Table 1. Such constraints are related to the properties of the substitution site, the identity of the wild-type amino acid and the differences between the wild-type and the mutated amino acid. First, the structural environment class definition, originally introduced by Bowie *et al*. (1991) in 1D representation of protein structure in fold-recognition studies, is a good proxy for structural constraints on the substitution site. Here, we extended their application to the problem of predicting phenotypic effect of nsSNPs. Bowie *et al*. (1991) used combinations of three structural parameters (buried area, fraction polar and secondary structure) to define 18 structural environments. Buried area reflects the solvent accessibility constraint and it is known that disease-associated nsSNPs tend to occur at buried sites (Sunyaev *et al*., 2000). Fraction polar is an indicator of environmental polarity and reflects the hydrogen bond constraint (Bowie *et al*., 1991). Disease-associated and neutral nsSNPs also have a slightly different secondary structure propensity, with the former tend to occur at β-sheet sites (Sunyaev *et al*., 2000). Second, the identity of wild-type amino acid was used as a predictor. Third, two parameters were used to describe the substitution changes. The SIFT score (Ng and Henikoff, 2001) measures the tolerance for a substitution in a multiple sequence alignment and hence incorporates evolutionary information. The indicator of change in the amino acid group has been first proposed by us. We took both physicochemical properties and compatibility with the structural environment into consideration. Different structural environments have different groupings of amino acids (Tables 2 and S1). A substitution leading to a great change in amino acid physicochemical property and/or compatibility with the structural environment tends to be disease associated, while a substitution leading to a minor change tends to be neutral.

Performance of SVM and RF

For the 4013 training samples with sufficient evolutionary information (each had no $<$10 homologous sequences), we used cross-validation experiments to evaluate the performance of our method and to compare the results with that of the SIFT algorithm. For the 205 independent test samples with insufficient evolutionary information (each had $<$10 homologous sequences), the classifiers trained by the training samples made a prediction on each test sample. It was straightforward to compare the prediction accuracy between different methods. Various parameters were tested for SVM and RF classifiers. RF has a built-in measurement of the performance: the OOB prediction error (Breiman, 2001). Hence, cross-validation was not necessary. But for the purpose of comparison with SVM, we still performed cross-validation to determine the best RF parameter mtry. In fact, the OOB error was very similar to the classification error

Table 3. Prediction accuracies on the training set and the independent test set

Method	Training set[a]				Test set[b]			
	FPR (%)	FNR (%)	BER (%)	MCC	FPR (%)	FNR (%)	BER (%)	MCC
SVM	42.2	21.4	31.8	0.274	30.0	31.4	30.7	0.282
RF	37.8	20.6	**29.2**	**0.315**	30.0	24.0	**27.0**	**0.352**
SIFT	40.4	19.7	30.1	0.305	33.3	38.3	35.8	0.203

SVM, support vector machine; RF, random forest; SIFT, SIFT algorithm; FPR, False positive rate; FNR, False negative rate; BER, Balanced error rate; MCC, Matthew's correlation coefficient.

[a]The training set includes 4013 samples, each of which has no <10 homologous sequences. Shown are results from cross-validation experiments.

[b]The independent test set includes 205 samples, each of which has <10 homologous sequences. Highlighted are those with the best performances.

of cross-validation. Best performance was found using radial basis kernel with $g = 0.1$ and $C = 10$ among the tested SVM classifiers. For RF classifiers, best performance was found by setting mtry $= 2$. The cross-validation results of the selected SVM and RF are listed in Table 3 along with the prediction accuracy of the SIFT algorithm. Figure 1 plots the corresponding ROC curves. The result shows that RF outperforms SVM. A possible reason is that the last two predictors in Table 1 are partially correlated, and SVM has difficulty in dealing with correlated predictors. In contrast, correlated predictors are tractable to RF, because RF uses random feature selection technique. Table 3 and Figure 1A also show that for nsSNPs with sufficient evolutionary information (not <10 homologous sequences), our method is comparable with the SIFT algorithm. The BER and the MCC of our method are slightly better than the SIFT algorithm. These findings indicate that, for nsSNPs with sufficient evolutionary information, adding structural information only improves the prediction accuracy slightly. However, for the 205 independent test samples with insufficient evolutionary information, Table 3 and Figure 1B show that the improvement is significant. Therefore, for nsSNPs with insufficient evolutionary information, making use of structural information is critical for predicting the phenotypic effects of the nsSNPs.

Predictive power of the individual predictors

RF has a built-in measurement for the importance of individual predictor called mean decrease accuracy. It is calculated by randomly permuting the values of an individual predictor (predictor j) in the OOB cases. For each tree, the number of votes for the correct class in the predictor-j-permuted OOB data was subtracted from the number of votes for the correct class in the untouched OOB data, and the remainders were averaged over all trees in the forest. The resulting 'mean decrease accuracy' is a measure of predictor importance with respect to its contribution to the prediction accuracy. Table 4 shows the importance of individual predictors. SIFT score was the best among all the predictors. This is expected when sufficient evolutionary information exists, because the SIFT algorithm uses the information that the tolerance of a substitution has been naturally sampled during the evolution. The discriminating power of buried area and β-sheet was consistent with previous observations (Sunyaev *et al.*, 2000). Interestingly, the discriminating power of the wild-type amino acid was obvious for some amino acids like glycine, cysteine and the charged amino acids. This indicated that wild-type amino acid was differently distributed over the 18 structural environments between disease-associated and neutral nsSNPs. For example, in the EC structural environment, we found that the wild-type amino acid of disease-associated nsSNPs was much more likely to be glycine than that of neutral nsSNPs (64% versus 27%). Hydrophobic wild-type amino acids, in contrast, had the least discriminating power.

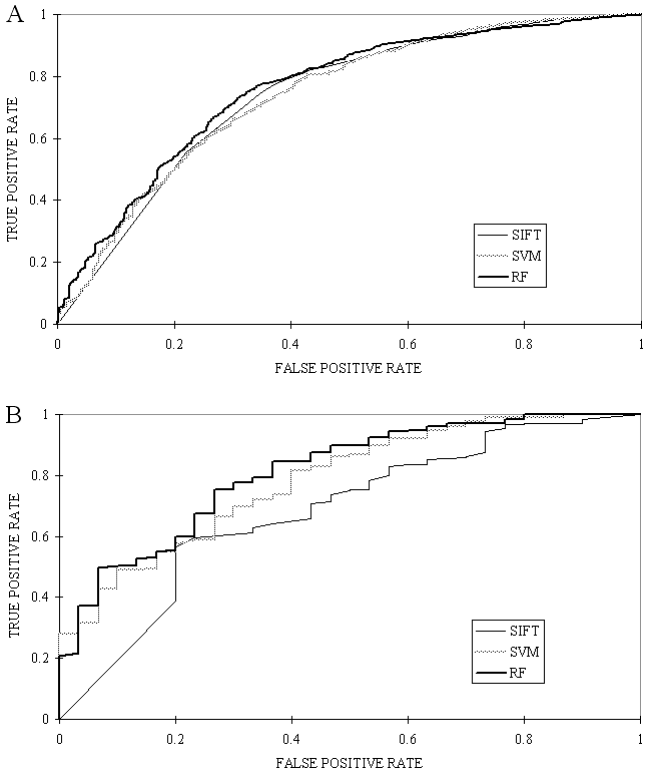

Fig. 1. ROC curves of RF, SVM and SIFT algorithm. (**A**) ROC curves on the training set using cross-validation. The training set includes 4013 samples, each of which has not <10 homologous sequences. (**B**) ROC curves on the test set. The test set includes 205 samples, each of which has <10 homologous sequences.

DISCUSSION

Discovering relationships between genotypes and phenotypes is the central task of genetic studies. The links between genotype and phenotype of nsSNPs have received plenty of research attention because of their prevalence in genomes and close associations to inherited diseases. With more and more genotype and phenotype data available and with increasing knowledge of the properties of nsSNPs, it is now practical to predict the phenotype of an nsSNP (i.e. whether an nsSNP is disease-associated or neutral) from the genotype *in silico*. The SIFT server (Ng and Henikoff, 2003) and the PolyPhen server (Ramensky *et al.*, 2002) are the two representatives for this purpose. Instead of learning from data, they determine parameters manually based on the knowledge of a human expert. Several other studies have exploited machine learning approaches to classify disease-associated and neutral nsSNPs (Chasman and Adams, 2001; Saunders and Baker, 2002; Krishnan and Westhead, 2003). Our study is different from the others in that we used natural nsSNPs

Table 4. Mean decrease accuracy of predictors

Predictor	Mean decrease accuracy
SIFT score	0.62
Amino acid group change	0.43
Buried area	0.31
Wild-type amino acid: Gly	0.24
Fraction polar	0.18
Secondary structure: β-sheet	0.17
Wild type amino acid: Arg	0.17
Wild type amino acid: Cys	0.17
Wild type amino acid: Lys	0.16
Wild type amino acid: Asp	0.15
Wild type amino acid: Asn	0.11
Wild type amino acid: His	0.11
Wild type amino acid: Met	0.09
Wild type amino acid: Glu	0.06
Wild type amino acid: Leu	0.04
Wild type amino acid: Gln	0.03
Wild type amino acid: Ala	0.03
Wild type amino acid: Ser	0.02
Secondary structure: α-helix	0.00
Wild type amino acid: Thr	−0.02
Wild type amino acid: Ile	−0.04
Wild type amino acid: Tyr	−0.05
Wild type amino acid: Pro	−0.05
Wild type amino acid: Trp	−0.08
Wild type amino acid: Val	−0.08
Secondary structure: Coil	−0.08
Wild type amino acid: Phe	−0.10

rather than *in vitro* mutational data as the training set. Saunders and Baker (2002) also tested their method in natural nsSNPs, but their set contained a rather small number of samples. *In vitro* mutational data includes only two proteins and might introduce some bias. Previous studies showed that cross-validation accuracy of natural nsSNP data is generally lower than that of *in vitro* mutational data (Saunders and Baker, 2002), demonstrating that a fair evaluation of performance should use a natural nsSNP dataset.

Good prediction accuracy usually depends on two factors: informative predictors and superior machine learning approach. We introduced several novel informative predictors in combination with some predictors from the literature to achieve better discriminating power. We found that the structural parameters representing environments of nsSNPs as well as the environment-specific grouping of wild-type and mutated amino acids have considerable discriminating powers. Furthermore, two state-of-the-art machine learning methods—RF and SVM, were used to combine the discriminating powers of individual predictors in approximately optimal ways. RF was found to outperform SVM. A possible reason is that RF is superior to SVM in dealing with correlated predictors. The comparison of our method with the frequently used SIFT algorithm revealed that, for nsSNPs with insufficient evolutionary information, incorporating structural information remarkably increased the prediction accuracy.

Our method required 3D structures (or homologous structures) of the nsSNP variants, which limits its application when only sequence information is available. However, it is expected that the structural genomics project (Berman and Westbrook, 2004) will rapidly increase the number of experimentally derived protein structures. Furthermore, genome-wide protein 3D modeling projects (Schwede *et al*., 2003) and the progress in protein structure prediction (Hardin *et al*., 2002) will also increase the applicability of our method.

ACKNOWLEDGEMENTS

We thank Drs James Bowie, Roland Luethy and David Eisenberg for providing the computer program for calculating the structural environments. We thank Drs Pauline Ng and Steven Henikoff for providing access to the SIFT program. We thank Drs Leo Breiman, Andy Liaw and Matthew Wiener for providing access to the Random Forest package and helpful discussions. We thank Dr Thorsten Joachims for providing access to the SVM-light software. We also thank the two anonymous reviewers for their very helpful comments. This work was partly supported by a PhRMA Foundation grant to YC.

REFERENCES

Altschul,S.F. *et al*. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.

Apweiler,R. *et al*. (2004) UniProt: the universal protein knowledgebase. *Nucleic Acids Res*., **32**, D115–D119.

Bhasin,M. and Raghava,G.P. (2004) ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. *Nucleic Acids Res*., **32**, W414–W419.

Berman,H.M. and Westbrook,J.D. (2004) The impact of structural genomics on the protein data bank. *Am. J. Pharmacogenomics*, **4**, 247–252.

Bowie,J.U. *et al*. (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, **253**, 164–170.

Breiman,L. (1996) Bagging predictors. *Mach. Learning*, **24**, 123–140.

Breiman,L. (2001) Random forest. *Technical Report*, Stat. Dept. UCB.

Breiman,L., Friedman,J.H., Olshen,R.A. and Stone,C. (1984) *Classification and Regression Trees*, Chapman and Hall, NY.

Chandonia,J.M. *et al*. (2004) The ASTRAL compendium in 2004. *Nucleic Acids Res*., **32**, D189–D192.

Chasman,D. and Adams,R.M. (2001) Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *J. Mol. Biol*., **307**, 683–706.

Chen,Y.C. *et al*. (2004) Prediction of the bonding states of cysteines using the support vector machines based on multiple feature vectors and cysteine state sequences. *Proteins*, **55**, 1036–1042.

Collins,F.S. *et al*. (1998) A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res*., **8**, 1229–1231.

Fredman,D. *et al*. (2002) HGVbase: a human sequence variation database emphasizing data quality and a broad spectrum of data sources. *Nucleic Acids Res*., **30**, 387–391.

Frishman,D. and Argos,P. (1995). Knowledge-based protein secondary structure assignment. *Proteins*, **23**, 566–579.

Gunther,E.C. *et al*. (2003) Prediction of clinical drug efficacy by classification of drug-induced genomic expression profiles *in vitro*. *Proc. Natl Acad. Sci. USA*, **100**, 9608–9613.

Hardin,C. *et al*. (2002) *Ab initio* protein structure prediction. *Curr. Opin. Struct. Biol*., **12**, 176–181.

Henikoff,J.G. and Henikoff,S. (1996) Using substitution probabilities to improve position-specific scoring matrices. *Comput. Appl. Biosci*., **12**, 135–143.

Irizarry,K. *et al*. (2000) Comprehensive EST analysis of single nucleotide polymorphism across coding regions of the human genome. *Nat. Genet*, **26**, 233–236.

Joachims,T. (1999) Making large-scale SVM learning practical. In Schölkopf,B., Burges,C. and Smola,A. (eds), *Advances in Kernel Methods—Support Vector Learning*. MIT Press, Cambridge, MA.

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

Matthews,BW. (1985) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.

Ng,P.C. and Henikoff,S. (2001) Predicting deleterious amino acid substitutions. *Genome Res*., **11**, 863–874,

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

Ramensky,V. *et al*. (2002) Human non-synonymous SNPs: server and survey. *Nucleic Acids Res*., **30**, 3894–3900.

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

Schwede,T. *et al*. (2003) SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Res*., **31**, 3381–3385.

Stenson,P.D. *et al*. (2003) Human gene mutation database (HGMD): 2003 update. *Hum. Mutat*., **21**, 577–581.

Sunyaev,S. *et al*. (2000) Towards a structural basis of human non-synonymous single nucleotide polymorphisms. *Trends Genet*., **16**, 198–200.

Sunyaev,S. *et al*. (2001) Prediction of deleterious human alleles. *Hum. Mol. Genet*., **10**, 591–597.

Svetnik,V. *et al*. (2003) Random forest: a classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci*., **43**, 1947–1958.

Vapnik,V. (1998) *Statistical Learning Theory*. Wiley, NY.

Wang,Z. and Moult,J. (2001) SNPs, protein structure, and disease. *Hum. Mutat*., **17**, 263–270.

Wu,B. *et al*. (2003) Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics*, **19**, 1636–1643.

Zhou,X.H., Obuchowski,N. and Obuchowski,D. (2002) *Statistical Methods in Diagnostic Medicine*. Wiley and Sons, NY.