BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 12 2007, pages 1444–1450
doi:10.1093/bioinformatics/btm119

Genome analysis

Finding new structural and sequence attributes to predict possible disease association of single amino acid polymorphism (SAP)

Zhi-Qiang Ye[1], Shu-Qi Zhao[1], Ge Gao[1], Xiao-Qiao Liu[1], Robert E. Langlois[2], Hui Lu[2] and Liping Wei[1,*]

[1]Center for Bioinformatics, National Laboratory of Protein Engineering and Plant Genetic Engineering, College of Life Sciences, Peking University, Beijing 100871, P. R. China and [2]Bioinformatics Program, Department of Bioengineering, University of Illinois at Chicago, Chicago, IL 60607, USA

Received on December 21, 2006; revised on February 25, 2007; accepted on March 16, 2007

Advance Access publication March 24, 2007

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: The rapid accumulation of single amino acid polymorphisms (SAPs), also known as non-synonymous single nucleotide polymorphisms (nsSNPs), brings the opportunities and needs to understand and predict their disease association. Currently published attributes are limited, the detailed mechanisms governing the disease association of a SAP remain unclear and thus, further investigation of new attributes and improvement of the prediction are desired.

Results: A SAP dataset was compiled from the Swiss-Prot variant pages. We extracted and demonstrated the effectiveness of several new biologically informative attributes including the structural neighbor profiles that describe the SAP's microenvironment, nearby functional sites that measure the structure-based and sequence-based distances between the SAP site and its nearby functional sites, aggregation properties that measure the likelihood of protein aggregation and disordered regions that consider whether the SAP is located in structurally disordered regions. The new attributes provided insights into the mechanisms of the disease association of SAPs. We built a support vector machines (SVMs) classifier employing a carefully selected set of new and previously published attributes. Through a strict protein-level 5-fold cross-validation, we attained an overall accuracy of 82.61%, and an MCC of 0.60. Moreover, a web server was developed to provide a user-friendly interface for biologists.

Availability: The web server is available at http://sapred.cbi.pku.edu.cn/

Contact: sapred@mail.cbi.pku.edu.cn

Supplementary information: Supplementary data are available at http://sapred.cbi.pku.edu.cn/supp.do

1 INTRODUCTION

After the completion of the human genome project, increasing attention has focused on the identification of human genomic variations, especially single nucleotide polymorphisms (SNPs). It is estimated that the world population contains a total of ~10 million SNP sites, resulting in an average density of one variant per 300 bases (Gibbs *et al.*, 2003; Kruglyak and Nickerson, 2001; Reich *et al.*, 2003). SNPs in coding and regulatory regions may play a direct role in diseases or differing phenotypes (Gibbs *et al.*, 2003; Pastinen *et al.*, 2006). Among them, the single amino acid polymorphisms (SAPs, conventionally known as non-synonymous SNPs or nsSNPs), which cause amino acid substitutions in the protein product, are of major interest because they account for about 50% of the gene lesions known to be related to genetic diseases (Krawczak *et al.*, 2000). Through large-scale efforts such as the HapMap project (http://www.hapmap.org), The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov) and whole genome association studies (Liu *et al.*, 2006), available SAP data is accumulating rapidly in databases such as dbSNP (Sherry *et al.*, 2001), HGVbase (Fredman *et al.*, 2004), Swiss-Prot variant page (Yip *et al.*, 2004) and many allele-specific databases. However, because of the high-throughput nature of these efforts, many SAPs could not be experimentally characterized in terms of their possible disease association. Furthermore, the underlying mechanisms that explain why a SAP may be associated with disease and have deleterious functional effect are not yet fully understood.

In the past five years, several bioinformatics methods have been developed to use sequence and structural attributes to predict possible disease association or functional effect of a given SAP. A popular sequence-based method is SIFT (Ng and Henikoff, 2001) which predicts whether an amino acid substitution is deleterious or tolerated based on the evolutionary conservation of the SAP site from multiple sequence alignment. More recent methods incorporate both sequence and structural attributes and use a range of classifiers such as rule-based, decision trees, support vector machines (SVMs), neural networks, random forests and Bayesian networks to annotate SAPs (Bao and Cui, 2005; Cai *et al.*, 2004; Dobson *et al.*, 2006; Ferrer-Costa *et al.*, 2004; Karchin *et al.*, 2005; Krishnan and Westhead, 2003; Ramensky *et al.*, 2002;

*To whom correspondence should be addressed.

Sunyaev *et al*., 2001; Yue and Moult, 2006). Methods that employ machine learning generally obtained better performance.

However, to better predict the possible disease association of SAPs, existing methods still need to be improved in several aspects. First, more biologically informative structural and sequence attributes need to be investigated to further understand the underlying mechanism of how a SAP may be associated with a disease. Second, several studies used imbalanced datasets which impeded the performance of their classifiers. Third, by using more biologically informative attributes and a better dataset, the overall accuracy of the prediction can be improved.

2 METHODS

We constructed a relatively balanced dataset from the Swiss-Prot variant pages (Yip *et al*., 2004). We then investigated the most complete set of structural and sequence attributes to date and identified a number of biologically informative new attributes that could explain why a SAP may be associated with disease. Finally, we incorporated these attributes into an SVM-based machine-learning classifier. Two previous studies (Bao and Cui, 2005; Dobson *et al*., 2006) used datasets derived from Swiss-Prot variant pages and thus could be compared with our method. At the end, we evaluated our method by protein-level cross-validation.

2.1 Data collection

We collected SAPs from the Variant pages of the Swiss-Prot knowledgebase version 49 (Yip *et al*., 2004). Swiss-Prot classifies variants into three categories, 'Disease', 'Polymorphism' and 'Unclassified': a variant is classified as 'Disease' if it is proven experimentally to be directly causing the disease (similar to 'deleterious'); a variant is classified as 'Polymorphism' if no disease association has been reported (similar to 'neutral' or 'tolerated'); and a variant is 'Unclassified' if its disease association is unclear (http://www.expasy.org/tools/variant_man.html#status). We focused only on 'Disease' and 'Polymorphism'.

Previous work calculated structural attributes of a SAP by mapping from the homologous protein structure template. This limited not only the types of attributes that could be calculated but also the accuracy of the attribute values. We chose to use the structure models in the ModSNP database instead (Yip *et al*., 2004) which contains models for both the wild-type and the variant proteins built by homology modeling software. Swiss-Prot variant web pages were downloaded from EXPASY website and subsequently parsed with an in-house Perl script. URLs to the corresponding ModSNP PDB files, if available, were extracted from the Variant pages. Following these URLs, the ModSNP PDB files were downloaded as well. For SAPs with more than one pair of homology models, only the first pair was used. Further quality control removed problematic PDB files with incorrect residue substitution, erroneous position of variation, non-printable characters or lack of coordinates. Finally, we obtained a dataset containing 3438 SAPs located in 522 proteins, consisting of 2249 'Disease' and 1189 'Polymorphism' SAPs, with a balanced ratio of ~2 to 1.

2.2 Calculation of structural and sequence attributes

We investigated a large set of structural and sequence attributes including both commonly used ones such as residue frequency and solvent accessibility and new ones that are novel to this kind of study. We were particularly interested in new attributes that are biologically informative. Most attributes were calculated from both the wild-type and the variant protein sequence and structure, from which the difference between the wild-type and variant and the absolute value of the difference are also calculated. We describe the most interesting attributes below. Detailed descriptions of other attributes are available in the Supplementary Material.

2.2.1 Residue frequency and conservation The difference between the observed frequencies of the wild-type and variant residue in homologous protein sequences have been shown to be a powerful attribute (Ng and Henikoff, 2001). The higher the difference, the more likely the SAP is deleterious or associated with disease. We used the same definition of residue frequencies as the SIFT program (Ng and Henikoff, 2001), searching for homolog in the Swiss-Prot 49 database. A slight variation from SIFT is that we calculated the difference (and the absolute value of the difference) between the wild-type and variant residue frequencies as opposed to the ratio.

We defined a second attribute to measure the level of sequence conservation among homologous proteins of the SAP position and its neighboring residue positions. The conservation score of a sequence position is defined as the information content of the amino acid frequency distribution at this position in a multiple sequence alignment (Schneider *et al*., 1986),

$$\text{Conservation} = -\sum_{i=1}^{20} p_i \log_2 p_i,$$

where p_i represents the frequency of residue type i. The conservation score ranges from 0 to 4.32; the lower the value, the more conservative the position is. We calculated the conservation score for the SAP position as well as the three positions to its left and three positions to its right. Previous studies only considered the SAP position itself (Bao and Cui, 2005; Dobson *et al*., 2006).

2.2.2 Solvent accessibilities Solvent accessibility has been shown to be the second most powerful attribute in predicting the disease association or functional effect of SAPs (Dobson *et al*., 2006; Saunders and Baker, 2002). We investigated different definitions of solvent accessibilities. For each wild-type and variant residue, we used NACCESS (Hubbard and Thornton, 1993) 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. Furthermore, the differences and absolute value of the differences between wild-type and variant were also calculated. We calculated C_β density as another measurement of solvent accessibility, defined as a count of the number of C_β atoms within 10 Å of the C_β on the wild-type and variant residue (Saunders and Baker, 2002). For glycine, an equivalent pseudo-atom is used. Calculation of C_β density does not require a full-atom model, thus allowing a wider application when only coarse structural models can be built.

2.2.3 Structural neighbor profiles We hypothesized that the 3D microenvironment around a SAP site may have a large influence on whether a SAP may have disease association. To study this, we defined a new attribute, the structural neighbor profile, to comprise a 20-D vector of the counts of different types of residues found in the 3D vicinity of a site: a count was obtained for each of the 20 residue types; a residue was considered as a 'Neighbor' if one or more of its heavy atoms fall within a specific radius around the C_α atom of the residue at the center. We calculated the structural neighbor profiles for both the wild-type and variant residue positions. The difference and the absolute value of the difference between the wild-type and variant were also calculated. We experimented with different radii ranging from 5 to 15 Å, at 1 Å interval.

2.2.4 Nearby functional sites Intuitively, if a SAP position is close to a protein functional site, then the SAP is more likely to be

associated with disease. The 'FT' fields in the Swiss-Prot database list residues known to be involved in functional sites. Several earlier studies mapped SAP positions to functional sites, but found that because too few SAPs could be mapped, the predictive power was limited (Dobson *et al.*, 2006; Ramensky *et al.*, 2002). To overcome this limitation, we defined a new group of attributes that, instead of considering only the SAPs that lie exactly on functional sites, calculate the distance between a SAP position and closest functional site. We considered ACT_SITE (active site), BINDING (binding site), METAL (metal ion binding site), MOD_RES (posttranslationally modified residue), DISULFID (disulfide bond) and TRANSMEM (transmembrane region). 'Distance' is calculated both along the sequence (rendered by the number of residues in-between) and in 3D structure (rendered by the spatial distance between two residues' C_β atoms; for glycine, C_α atom is used), represented by separate attributes.

2.2.5 Structure model energy A SAP may affect the stability of a protein, which in turn may affect its function. We investigated a new attribute that measure the stability of the structure models of both the wild-type and variant proteins. The attribute is calculated using the default energy evaluation function in MODELLER (Sali and Blundell, 1993). If the variant protein became unstable because of the SAP, its optimized structure model would have a higher energy score than the wild-type protein presumably.

2.2.6 Hydrogen bond The number of hydrogen bonds around a SAP site has been used in previous studies (Ramensky *et al.*, 2002; Wang and Moult, 2001). We calculated the number of hydrogen bonds using HBPLUS (McDonald and Thornton, 1994). We further calculated the difference (and the absolute value of the difference) in terms of the number of hydrogen bonds around the wild-type versus variant residue.

2.2.7 Disulfide bond SAP that damages a disulfide bond will likely destabilize a protein structure and affect its function. We utilized the additional output of HBPLUS program (McDonald and Thornton, 1994) to parse the number of disulfide bonds between a SAP site and its nearby residues in the wild-type and variant protein structure models, respectively. The difference and the absolute value of the difference were also calculated.

2.2.8 Disordered region Disordered regions of a protein lack a fixed tertiary structure, and are partially or fully unfolded. Contrary to earlier belief that they were 'useless', they have been found to be involved in many important functions, such as DNA recognition, modulation of specificity and affinity of protein binding (Dunker and Obradovic, 2001). To investigate disordered regions as a factor that may affect the disease association of a SAP, we defined a new attribute to indicate whether a SAP is located in disordered region or not. The collection of disordered regions was obtained from DisProt database version 3.3 (Vucetic *et al.*, 2005).

2.2.9 Aggregation properties Many diseases have been associated with excessive aggregation of proteins (Chiti *et al.*, 2002). We investigated whether a SAP may change the β-aggregation properties of a protein, and whether the change may be associated with disease. For the wild-type and the variant proteins, we defined the new attributes of residue β-aggregation properties at the SAP site as well as the fragment β-aggregation properties, calculated using TANGO (Fernandez-Escamilla *et al.*, 2004).

2.2.10 HLA family Histocompatibility leukocyte antigen (HLA) is a large family of proteins whose variations are used by the immune system to distinguish non-self from self-molecules (Fleisher 2006). Thus, SAPs in HLA may hold special characteristics and were investigated in terms of their disease association. We defined a new attribute to indicate whether the protein in which the SAP located belongs to the HLA family. The protein was BLASTed against IMGT/HLA database release 2.14 (Robinson *et al.*, 2003), and was considered as HLA if it could hit a sequence satisfying both the e-value less than 1.0 and the sequence identity over 80%.

2.3 SVM classifiers, attributes evaluation and selection, and performance evaluation

With a proper mapping furnished by a kernel function, support vector machines (SVM) classifiers separate transformed data with a hyper plane in a high-dimensional space. SVMs have been widely used in supervised classification problems in bioinformatics, such as in DNA-binding protein prediction (Bhardwaj *et al.*, 2005), membrane-binding protein prediction (Bhardwaj *et al.*, 2006), and protein fold recognition (Langlois *et al.*, 2006). We adopted the LIBSVM package to evaluate the attributes and build the final classifier, using the radial basis function (RBF) as the kernel function (Chang and Lin, 2001). We employed a 'grid-search' to select the proper values of the parameter γ of RBF and the penalty parameter (C) of the soft-margin SVM. C was set to $2^{-5}, 2^{-3}, \ldots 2^{15}$, and γ to $2^{-15}, 2^{-13}, \ldots 2^{3}$. We tried all the combinations of C and γ and selected the pair with the best cross-validation accuracy.

We evaluated the entire set of attributes in terms of their association with disease and selected a final subset with good predictive power. For 'single' attributes, mutual information was used for initial screening, followed by further evaluation based on cross-validation accuracies from the SVMs. Numerical attributes were converted into categorical ones using five equal frequency bins for calculating the mutual information. For 'group' attributes such as the 20 components of structural neighbor profiles, we used SVMs' cross-validation accuracies directly. An additional consideration is that the final subset should cover at least one attribute from each attribute type, representing broadly the involved biological significance.

As for the 5-fold cross-validation, we ensured that the dataset was split at the protein level in addition to the stratified partition. That is, the SAP data from the same protein was not permitted to reside in different groups. This 'double' stratification ensures a more rigorous evaluation. The detailed partition of the dataset into five groups is listed in Table S1 in the Supplementary Material.

The overall accuracy (ACC) is defined as

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}},$$

where TP is the number of true positives (here we took disease associated as positive), TN is the number of true negatives, FP is the number of false positives and FN is the number of false negatives. For the sake of comparison with previous studies, we also calculated the Matthew's correlation coefficient (MCC) (Matthew, 1985):

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

ACC is an intuitive measurement of performance on a balanced dataset whereas MCC is more realistic than ACC on an unbalanced dataset (Bao and Cui, 2005; Dobson *et al.*, 2006). The higher the MCC and ACC, the better the performance. In addition to ACC and MCC, we also calculated the recall and precision, which are defined as TP/(TP + FN) and TP/(TP + FP), respectively.

2.4 Web server implementation

To facilitate the prediction of disease association of newly identified SAPs, we implemented an automated pipeline of our method and developed a web server interface. Tomcat/Apache served as a J2EE

container for Java Servlet and JSP. A MySQL database was set up to store the user account information, uploaded data and prediction results. The web server runs on a Linux box with 2 AMD Opteron CPUs and 4 GB memory, and works with both the Microsoft Internet Explorer and Mozilla Firefox browsers.

Considering that a large amount of SAPs are located on proteins whose structures could not be reliably modeled by homology modeling, we implemented an option that allows users to make predictions using only sequence-based attributes. This option increases the applicability of the web server.

3 RESULTS

3.1 Structural and sequence attributes with predictive power

Previous studies have found sequence evolutionary information such as SAP residue frequencies and conservation to be the most powerful type of attributes (Bao and Cui, 2005; Dobson *et al.*, 2006). In our study, we confirmed that residue frequencies provided the best discrimination. Combination of wild-type and variant residue frequencies and their difference achieved an ACC of 77.5% and an MCC of 0.489 (Table 1). Conservation scores at the SAP position and six nearby positions achieved an ACC of 74.8%. The biological insight underlying the predictive power of this type of attribute is clear: a conserved residue position is more intolerant to amino acid substitutions and a SAP will thus be more likely associated with disease.

Previous studies found solvent accessibilities to be the type of attributes with the second most predictive power. We evaluated a full spectrum of definitions of solvent accessibilities produced by NACCESS (http://wolf.bms.umist.ac.uk/naccess/) and found that the 'total-side absolute' and 'total-side relative' solvent accessibilities had higher mutual information with SAP status than other definitions (Table S2 in the Supplementary Material). Combination of these two attributes achieved an ACC of 68.0% and MCC of 0.232 (Table 1). C_β density, the calculation of which did not require full atomic model, achieved an accuracy of 67.0% and MCC of 0.190 (Table 1). Thus the predictive power of solvent accessibilities is reasonable and consistent with previous studies (Dobson *et al.*, 2006).

However, our study identified two new types of attributes that showed higher predictive power than solvent accessibilities. Structural neighbor profiles were found nearly as powerful as residue frequencies (Table 1). We tested a range of radii varying from 5 to 15 Å. The predictive power, in terms of ACC and MCC, of each value of the radius is displayed in Figure 1A and B. Three observations were made from the figures. First, the structural neighbor profiles of the wild-type and variant had high predictive power. It indicated that certain microenvironments tolerate SAPs whereas others are more sensitive to SAPs which may lead to disease. This observation coincides with the previously reported importance of microenvironments (Wei and Altman, 2003). Second, the difference between the structural neighbor profiles of the wild-type and variant had a much lower predictive power, indicating that it was the microenvironment itself, not the change of it caused by the SAP, that is biologically informative in terms of disease association. Third, the predictive power of the wild-type and variant structural neighbor profiles peaked when the radius was set to 13 Å. We used this empirical value in our final attribute set. This group of novel attributes proved to be almost as powerful as sequence evolutionary information, contributing to the increased accuracy of disease association prediction and demonstrating the importance of microenvironments on the functional effect of SAPs.

We found a second new group of attributes, nearby functional sites, which showed higher predictive power than solvent accessibilities. SAPs in the vicinity (both spatial and sequential) of ACT_SITE, BINDING, METAL, MOD_RES and DISULFID had a higher ratio of 'Disease' to 'Polymorphism' than the distant ones. The most

Table 1. The predictive power of certain attribute(s) subsets in terms of ACC and MCC

Attribute groups	ACC (%)	MCC
Residue frequencies	77.5	0.489
13 Å structural neighbor profile	76.4	0.467
Conservation scores	74.8	0.425
Nearby functional sites	69.3	0.246
Solvent accessibilities	68.0	0.232
C_β density	67.0	0.190
Final attribute set	82.6	0.604

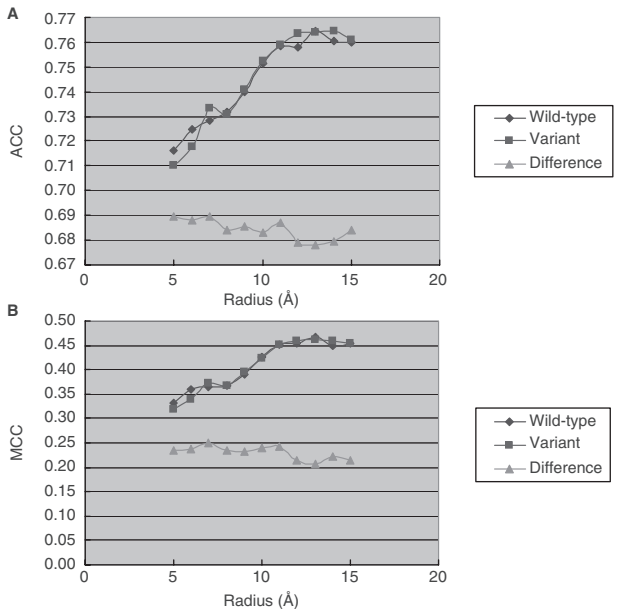

Fig. 1. (**A**) The predictive power (in terms of ACC) of structural neighbor profiles defined by different radii. (**B**) The predictive power (in terms of MCC) of structural neighbor profiles defined by different radii. The structural neighbor profiles of variant and wild-type protein, as well as their difference were evaluated. The ACCs and MCCs resulted from 5-fold cross-validation by varying the radius of the structural neighbor profile as evaluated by LIBSVM. Both (A) and (B) indicated that the structural neighbor profile defined by the radius of 13 Å was the best.

evident is the ACT_SITE: the closer the SAP was located to residues of ACT_SITE, the higher the ratio of 'Disease' to 'Polymorphism' (Fig. 2). SAPs near functional sites were more prone to alter the nature of the functional sites than the distant SAPs and thus contribute to disease. For TRANSMEM (transmembrane regions), only 52 SAPs were found in regions annotated as TRANSMEM, among which 49 were associated with disease. Although the total number was small, the mechanism was clear: transmembrane regions consist of conserved hydrophobic helices; SAPs in these regions are more likely to affect the localization and function of the protein. The nearby functional sites attributes yielded an ACC of about 69% (Table 1), exceeding the accuracies of solvent accessibilities, the previously reported second most powerful attribute.

We also found several other new types of attributes that have good predictive power. Three new attributes have particularly interesting biological significance. First, among 122 SAPs located in disordered regions, 114 (93%) are associated with disease. Thus SAPs occurring in disordered regions are highly likely to affect the functions of the proteins and be associated with disease. This highlighted the functional importance of disordered regions. Second, among 194 SAPs altering the fragment β-aggregation properties, 169 (87%) are associated with disease. Third, among 435 SAPs from 75 HLA proteins, all except one held the status of 'Polymorphism'. This follows from the nature of HLA that variation between HLAs from different individuals marks immune individuality, and is usually not related to disease.

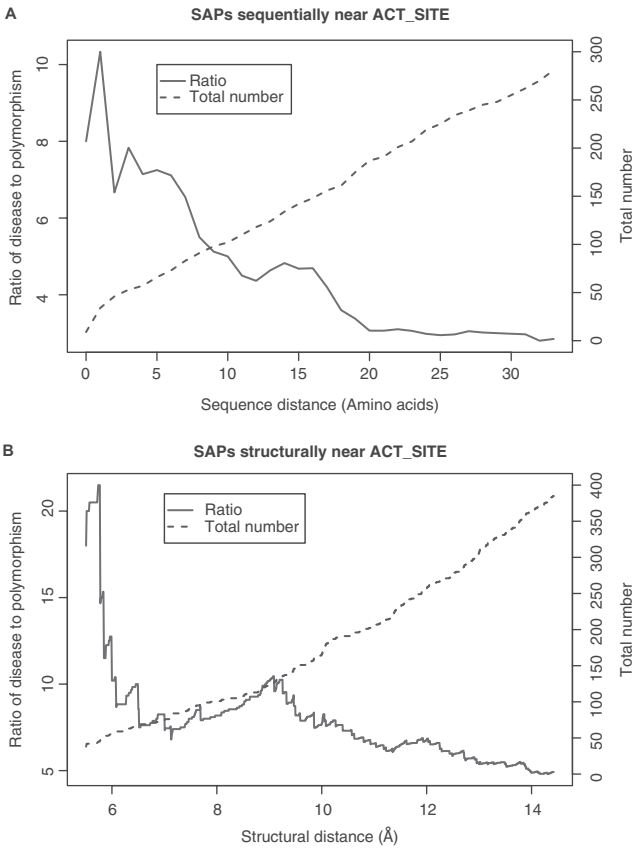

Fig. 2. (**A**) The correlation of both sequential SAP proximity and distribution with the ACT_SITE to the likelihood of disease association. (**B**) The correlation of both structural SAP proximity and distribution with the ACT_SITE to the likelihood of disease association. As the distance (sequence-based and structure-based) between the SAP and ACT_SITE widens, the total number of SAPs within this distance increases, while the ratio of disease to polymorphism in these SAPs decreases. Although the tendency in (B) is not as evident as that in (A), the global trend remains.

3.2 Other attributes

Other attributes also showed useful predictive power, although in most cases lower than that of the above-mentioned attributes. The final set of 60 attributes are summarized in the group style in Table S3, and the 1R rank generated by WEKA (Witten and Frank, 2005) of each single attribute is listed in Table S4 (Supplementary Material). For example, the difference between the structural model energies of the wild-type and variant proved to be of some, but limited predictive power according to the 1R rank (Table S4), a finding coincident with previous observation that the energy in force field does not correlate well with the quality of a structural model in some circumstances (Novotny *et al.*, 1984). Several other attributes were weakly correlated with the disease association partly because they occur with lower frequency and thus have low coverage in the dataset, such as SAPs with altered putative disulfide bonds. However, due to their biological significance we retained them in the final set.

3.3 Final model and web server

Using this final set we achieved an ACC of 82.61% and MCC of 0.604 in 5-fold cross-validation when C was set to 2 and γ to 0.0078125 (Table 1). The recall and precision were 93.86 and 82.11%, respectively. The increase in MCC of our method compared with SIFT on the same dataset was 0.124. Bao and Cui reported an MCC of 0.315 on similar dataset and an increase in MCC of 0.01 over SIFT on their same dataset (Bao *et al.*, 2005). Dobson *et al.* reported that the MCC of using all attributes on their 100% balanced dataset was 0.49, while using PSIC (a conservation score) alone achieved an MCC of about 0.425, which indicated an improvement of 0.065 (Fig. 1 in Dobson *et al.*, 2006). Thus, compared with previous methods based on similar datasets, our method achieved higher MCC as well as larger improvement over using conservation scores alone on the same dataset.

If we selected only the top 10 attributes according to the 1R rank, the accuracy would be slightly better, with ACC of 82.84% and MCC of 0.610. However, we still chose to use 60 attributes in the final model, favoring a representative attribute subset that covered many aspects of intuitive biological significance.

A web server interface of our method, named SAPRED, is freely available at http://sapred.cbi.pku.edu.cn/. SAPRED server requires as input a FASTA-format protein sequence, a mutation in the form of A#B, where A and B represent the single-letter code of amino acid and # represents the position of the substitution, and two PDB-format files describing the

structures of the wild-type and variant protein. The output contains the predicted result and the prediction confidence, as well as the values of all attributes to help elucidate the putative biological insights. A registered user has an additional benefit of having a private 'User Space' to manage his/her input data and results files on the SAPRED server, and can monitor all his/her job history online.

For proteins with no structural information available, SAPRED offers an alternative method called SAPRED_SEQ which requires only a FASTA sequence and a mutation as input, and makes the prediction based on sequence-derived attributes only. The ACC and MCC of SAPRED_SEQ are 81.5% and 0.577, respectively. It is still higher than that of previous methods based on similar datasets, which is largely due to contribution of the powerful new sequence-based attributes such as sequence-based nearby functional sites, aggregation properties and HLA families.

4 DISCUSSION

The new attributes studied in this work may further the understanding of the biological mechanism underlying the functional effect and disease association of SAPs. At the same time, they also contribute to the increase in accuracies of predicting the disease association of SAPs. In particular, the predictive power of structural neighbor profile is almost as high as that of residue frequencies, highlighting the importance of the microenvironment around a SAP. In addition, the predictive power of nearby functional sites is higher than solvent accessibilities, the second most powerful type of attributes in previous studies. By considering residues both at and near functional sites in terms of both sequence and structure, our method significantly enlarged the coverage and overcame the limitations in previous work that used only the functional site residues themselves (Dobson *et al*., 2006; Ramensky *et al*., 2002). The other new attributes we reported, such as disordered regions and aggregation properties, also provided direct biological insights into the study of SAPs and contributed to the overall accuracy of prediction.

Using structural models of both the wild-type and variant proteins allowed us to explore attributes that were not possible otherwise as well as to define them more precisely. In contrast, previous studies calculated the structural attributes of the homologous protein template, and then mapped them back to the protein that contained the SAPs (Bao and Cui, 2005; Dobson *et al*., 2006).

Previous studies based on Swiss-Prot variant page used unbalanced datasets (Bao and Cui, 2005; Dobson *et al*., 2006). The number of 'Disease' associated SAPs was seven and five times as many as that of 'Polymorphism' ones in the former and the latter, respectively. This imbalance may have restricted the discriminating power of their SVMs, random forests and decision tree classifiers. The latter study used under-sampling to create a balanced dataset and found that it could increase the prediction accuracy significantly. Here we generated a more balanced dataset with the ratio of 'Disease' to 'Polymorphism' of 2 to 1. This is important for accurate attribute selection and classifier training. Furthermore, we used a dataset partition strategy on the protein level for cross-validation. Such a strategy avoided the over-fitting of SAPs' disease association status to the proteins which the SAPs belonged to. We have provided a more rigorous evaluation than the partition on SAP level.

It is well known that grid-search of parameters in LIBSVM is important for better performance, as described in detail in the user guide of LIBSVM (http://www.csie.ntu.edu.tw/~cjlinpapers/guide/guide.pdf). Two contour plots of performance over $\log_2 C$ and $\log_2 \gamma$ are available in the Supplementary Material (http://sapred.cbi.pku.edu.cn/routine.do).

The homology models in ModSNP have relatively high quality. In future research, we will investigate the effect of using medium- to low-quality homology models generated directly from homology modeling software. We will also investigate the effect of using predicted functional sites instead of the ones annotated in Swiss-Prot.

ACKNOWLEDGEMENTS

We would like to thank Dr Qing-Rong Liu, Min Zhao and Chuanyun Li for helpful suggestions. We thank the two anonymous reviewers for their constructive comments. This work is supported by the Hi-Tech Research and Development Program of China (863 Program, No. 2006AA02Z314, 2006AA02Z334), National Keystone Basic Research Program of China (No. 2006CB910404) and China Ministry of Education 111 Project. H.L. is partially supported by NIH grant P01 AI060915. R.E.L. is supported by NIH training grant 5T32HL007692: Cellular Signaling in Cardiovascular System (PI, John Solaro).

Conflict of Interest: none declared.

REFERENCES

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

Bao,L. *et al*. (2005) nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms. *Nucleic Acids Res*., **33**, W480–W482.

Bhardwaj,N. *et al*. (2005) Kernel-based machine learning protocol for predicting DNA-binding proteins. *Nucleic Acids Res*., **33**, 6486–6493.

Bhardwaj,N. *et al*. (2006) Structural bioinformatics prediction of membrane-binding proteins. *J. Mol. Biol*., **359**, 486–495.

Cai,Z.H. *et al*. (2004) Bayesian approach to discovering pathogenic SNPs in conserved protein domains. *Hum. Mutat*., **24**, 178–184.

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

Chiti,F. *et al*. (2002) Studies of the aggregation of mutant proteins in vitro provide insights into the genetics of amyloid diseases. *Proc. Natl Acad. Sci. USA*, **99** (Suppl. 4), 16419–16426.

Dobson,R.J. *et al*. (2006) Predicting deleterious nsSNPs: an analysis of sequence and structural attributes. *BMC Bioinformatics*, **7**, 217.

Dunker,A.K. and Obradovic,Z. (2001) The protein trinity — linking function and disorder. *Nat. Biotechnol*., **19**, 805–806.

Fernandez-Escamilla,A.M. *et al*. (2004) Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat. Biotechnol*., **22**, 1302–1306.

Ferrer-Costa,C. *et al*. (2004) Sequence-based prediction of pathological mutations. *Proteins*, **57**, 811–819.

Fleisher,T.A. (2006) Back to basics: primary immune deficiencies: windows into the immune system. *Pediatr. Rev*., **27**, 363–372.

Fredman,D. *et al*. (2004) HGVbase: a curated resource describing human DNA variation and phenotype relationships. *Nucleic Acids Res*., **32**, D516–D519.

Gibbs,R.A. *et al*. (2003) The International HapMap Project. *Nature*, **426**, 789–796.

Hubbard,S.J. and Thornton,J.M. (1993) 'NACCESS'. In: *Computer Program*. Department of Biochemistry and Molecular Biology, University College London.

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

Krawczak,M. *et al*. (2000) Human gene mutation database-a biomedical information and research resource. *Hum. Mutat*., **15**, 45–51.

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.

Kruglyak,L. and Nickerson,D.A. (2001) Variation is the spice of life. *Nat. Genet*., **27**, 234–236.

Langlois,R.E. *et al*. (2006) Improved protein fold assignment using support vector machines. *International Journal of Bioinformatics Research and Applications*, **1**, 319–335.

Liu,Q.R. *et al*. (2006) Addiction molecular genetics: 639,401 SNP whole genome association identifies many "cell adhesion" genes. *Am. J. Med. Genet. B. Neuropsychiatr. Genet*., **141**, 918–925.

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

McDonald,I.K. and Thornton,J.M. (1994) Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol*., **238**, 777–793.

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

Novotny,J. *et al*. (1984) An analysis of incorrectly folded protein models. Implications for structure predictions. *J. Mol. Biol*., **177**, 787–818.

Pastinen,T. *et al*. (2006) Influence of human genome polymorphism on gene expression. *Hum. Mol. Genet*., **15**, R9–R16.

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

Reich,D.E. *et al*. (2003) Quality and completeness of SNP databases. *Nat. Genet*., **33**, 457–458.

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

Sali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol*., **234**, 779–815.

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

Schneider,T.D. *et al*. (1986) Information content of binding sites on nucleotide sequences. *J. Mol. Biol*., **188**, 415–431.

Sherry,S.T. *et al*. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*., **29**, 308–311.

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

Vucetic,S. *et al*. (2005) DisProt: a database of protein disorder. *Bioinformatics*, **21**, 137–140.

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

Wei,L. and Altman,R.B. (2003) Recognizing complex, asymmetric functional sites in protein structures using a Bayesian scoring function. *J. Bioinform. Comput. Biol*., **1**, 119–138.

Witten,I.H. and Frank,E. (2005) *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco.

Yip,Y.L. *et al*. (2004) The Swiss-Prot variant page and the ModSNP database: a resource for sequence and structure information on human protein variants. *Hum. Mutat*., **23**, 464–470.

Yue,P. and Moult,J. (2006) Identification and analysis of deleterious human SNPs. *J. Mol. Biol*., **356**, 1263–1274.