Protein Cell 2012, 3(1): 38–43
DOI 10.1007/s13238-011-1130-2

Protein & Cell

COMMUNICATION

SySAP: a system-level predictor of deleterious single amino acid polymorphisms

Tao Huang[1,2], Chuan Wang[1], Guoqing Zhang[1], Lu Xie[2] ✉, Yixue Li[1,2] ✉

[1] Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China
[2] Shanghai Center for Bioinformation Technology, Shanghai 200235, China
✉ Correspondence: xielu@scbit.org (L. Xie), yxli@sibs.ac.cn (Y. Li)
Received November 4, 2011 Accepted November 14, 2011

ABSTRACT

Single amino acid polymorphisms (SAPs), also known as non-synonymous single nucleotide polymorphisms (nsSNPs), are responsible for most of human genetic diseases. Discriminate the deleterious SAPs from neutral ones can help identify the disease genes and understand the mechanism of diseases. In this work, a method of deleterious SAP prediction at system level was established. Unlike most existing methods, our method not only considers the sequence and structure information, but also the network information. The integration of network information can improve the performance of deleterious SAP prediction. To make our method available to the public, we developed SySAP (a System-level predictor of deleterious Single Amino acid Polymorphisms), an easy-to-use and high accurate web server. SySAP is freely available at http://www.biosino.org/SySAP/and http://lifecenter.sgst.cn/SySAP/.

KEYWORDS deleterious single amino acid polymorphisms, predictor, web server

INTRODUCTION

With the rapid development of sequencing technology, more and more single nucleotide polymorphisms (SNPs) have been discovered (Sherry et al., 2001) and about 90% of the human genetic changes are caused by the SNPs (Burke et al., 2007). If SNPs occur in coding region and change the amino acid of encoded protein, they are called non-synonymous SNPs (nsSNPs), also known as single amino acid polymorphisms (SAPs). Since nsSNPs change the sequence of encoded protein, they may affect the function of protein and cause human genetic diseases (Stenson et al., 2003; Hamosh et al., 2005). Besides those disease-associated nsSNPs, there are also functionally neutral ones which do not cause any diseases. Maybe the neutral ones do not occur in functional important proteins or they do not cause severe structural change of the protein. Distinguishing disease-associated nsSNPs from neutral ones is important for the investigation of human genetic diseases. There are already a lot of tools to predict the deleterious SAPs, such as SIFT (Ng and Henikoff, 2003), PolyPhen (Ramensky et al., 2002), and sapred (Ye et al., 2007). Most of them are based on either the sequence features or structure features. It is difficult to improve the prediction performance anymore if only sequence and structure features are used. In fact, it is hard to believe that the effect of SAP can be accurately predicted only based on the sequence and structure features.

In our previous work, a method of deleterious SAP prediction at system level was established (Huang et al., 2010b). The rationale of our method is simple and easy-to-understand: If a SAP occurs in the protein with important functions and it can severely change the sequence and structure of the protein, it has a high possibility of causing disease (Huang et al., 2010b). Unlike most existing methods, our method not only considers the sequence and structure information, but also the network information which represents the importance of the protein. The integration of network information can improve the performance of deleterious SAP prediction. In fact, our method has higher accuracy than most other methods. To make our method public available, we developed SySAP (a System-level predictor of deleterious Single Amino acid Polymorphisms), an easy-to-use and highly accurate web server. SySAP is freely available at http://www.biosino.org/SySAP/ and http://lifecenter.sgst.cn/SySAP/.

RESULTS

We tested SySAP with Leave-One-Out Cross-Validation (LOOCV) based on data from UniProt Release 2010_12 and the prediction result is shown in Table 1. The sensitivity (Sn), specificity (Sp), accuracy (ACC) and Matthews's correlation coefficient (MCC) were 0.668, 0.907, 0.823 and 0.602, respectively.

The prediction accuracy of SySAP is higher than the most widely used deleterious SAP predictor—SIFT (Sorting Tolerant from Intolerant) (Ng and Henikoff, 2003). The reported accuracies of SIFT were often around 70% (Ng and Henikoff, 2002; Huang et al., 2010b; Li et al., 2011), much lower than us. This indicates that the integration of network information can improve the performance of deleterious SAP prediction. To our knowledge, SySAP is the first deleterious SAP prediction tool that includes both amino acid level features (PSSM conservation scores, disorder score, AAFactors, GRANTHAM score) and protein network level features (betweenness, KEGG enrichment scores). The network features can not only improve the prediction performance but also illustrate the functional association of SAP better.

Table 1 Confusion matrix of prediction result evaluated by Leave-One-Out Cross-Validation

Confusion matrix		Predicted	
		Disease SAP	Polymorphism
Actual	Disease SAP	13,157	6529
	Polymorphism	3377	32,822

The sensitivity (Sn), specificity (Sp), accuracy (ACC) and Matthews's correlation coefficient (MCC) were 0.668, 0.907, 0.823 and 0.602, respectively.

DISCUSSION

In our previous work (Huang et al., 2010b), we used mRMR (Maximum Relevance Minimum Redundancy) and IFS (Incremental Feature Selection) to reduce the number of features and then applied NNA (Nearest Neighbor Algorithm) to do the prediction. In SySAP, we used more sophisticated machine learning method LiblineaR (Fan et al., 2008) to predict the effect of query SAP with all features. LiblineaR (Fan et al., 2008) can make prediction very fast, even when the numbers of both samples and features are extremely large.

As an easy-to-use highly accurate tool for deleterious SAP identification, SySAP could be useful for medical geneticists and facilitate the post genome-wide association studies.

METHODS

According to a recent comprehensive review (Chou, 2011), to develop a useful predictor for biological systems, the following points were usually needed to consider: (1) benchmark dataset construction or selection, (2) mathematical formulation for biological sequence samples, (3) operating algorithm (or engine), (4) anticipated accuracy, and (5) web-server establishment. Below, we will elaborate these procedures one by one.

Dataset

Care et al. (2007) compared several widely used SAP datasets and thought the UniProt (Universal Protein Resource) dataset is the best training data for deleterious SAP prediction. In this study, SAP data from UniProt (http://www.uniprot.org/docs/humsavar, Release 2010_12) were downloaded to train and test the deleterious SAP prediction model. Each SAP in UniProt is annotated as either 'disease' (SAP with disease associated), 'polymorphism' (SAP with no known disease associated) or 'unclassified' (SAP which has too little information to be classified into former two classes). After excluding 'unclassified' SAPs and removing the redundancy, there were 36,199 unique polymorphism SAPs and 19,686 unique disease SAPs.

To avoid homology bias and remove the redundant sequences from the benchmark dataset, a cutoff threshold of 25% was imposed in (Chou and Shen, 2007; Chou, 2011; Chou et al., 2011; Wu et al., 2011) to exclude those proteins from the benchmark datasets that have 25% or greater sequence identity to any other in the same subset. However, in this study we did not use such a stringent criterion because even at the same site of the same protein, there could be several different single amino acid mutations with different effects.

Workflow of SySAP

Figure 1 shows the framework of SySAP server to predict whether the query SAP is deleterious. First, the query SAP is encoded into a 455-dimension vector which includes network, structure and sequence features. Then the query SAP is compared with the training vector with known effects, i.e. deleterious or not. Finally the LibliеaR (Fan et al., 2008) model makes its prediction: whether the query SAP is deleterious.

Feature space of SAP

To represent the SAP, we used 455 network, structure and sequence features. In the following, each kind of features will be briefly described.

The network features

As mentioned above, one major difference between our method and others is the integration of network information. We added two kinds of network information into our model.

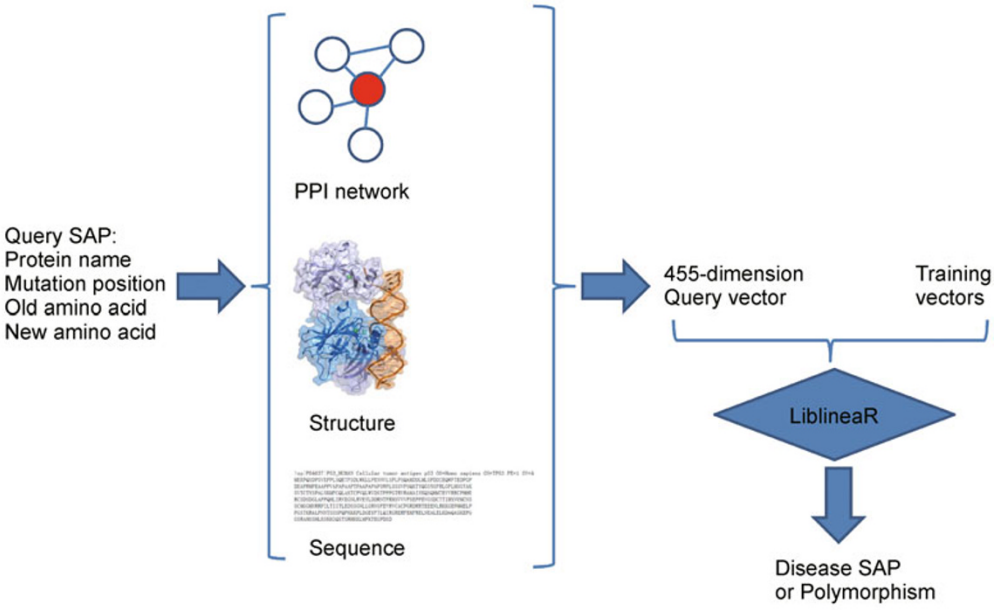

Figure 1. The framework of SySAP server to predict whether the query SAP is deleterious. First, the query SAP is encoded into a 455-dimension vector which includes network, structure and sequence features. Then the query SAP is compared with the training vector with known effects, i.e. deleterious or not. Finally the LibliеaR model makes its prediction: whether the query SAP is deleterious.

The first kind of network feature is betweenness (Freeman, 1979). It measures the information flow of the network. High betweenness indicates that there are multiple paths going through the node, while low betweenness means there are only few paths. In protein-protein interaction network, betweenness measures the ways in which signals pass through. We used R package tnet (http://opsahl.co.uk/tnet) to calculate the betweenness. The protein-protein interaction network was downloaded from STRING v8.3 (http://string-db.org/) (Jensen et al., 2009). The second kind of network feature is KEGG enrichment score. The function of one protein can be represented by its immediate neighbors on protein interaction network (Sharan et al., 2007). We defined the KEGG enrichment score of one protein by its neighbors on STRING network (Jensen et al., 2009). The value of KEGG enrichment score equals $-\log_{10}$ of the hypergeometric test p value of its neighbors. The larger the KEGG pathway enrichment score is, the more this KEGG pathway is overrepresented. Both the KEGG enrichment scores and betweenness were network level features. In total, there were 215 network features including 214 KEGG enrichment score features and one feature of betweenness.

The PSSM conservation scores

If the amino acid at certain site of a protein is evolutionarily conserved, it often means that this amino acid is in an important functional region of the protein and mutation of it could cause a significant structural and functional change of the protein. In this study, we calculated the Position Specific Scoring Matrix (PSSM) (Ahmad and Sarai, 2005) conservation score with Position Specific Iterative BLAST (PSI BLAST) Release 2.2.24 (Altschul et al., 1997) to quantify the conservation status of each amino acid site in the protein sequence. PSSM conservation score has been successfully used in the studies of post translational modifications (Niu et al., 2010; Cai et al., 2011) and effects of mutations (Huang et al., 2011b).

The disorder score

Disordered regions of protein do not have fixed three dimensional structures, but they play important roles in signaling transduction and gene regulation. In this study, we used the disorder score, calculated by VSL2 (Peng et al., 2006), to quantify the disorder status of each amino acid site in the protein sequence.

The AAFactors

AAindex (http://www.genome.ad.jp/aaindex/) is a database of more than five hundred numerical indices representing different physicochemical and biochemical properties of amino acids (Kawashima et al., 1999). Based on the factor analysis, Atchley et al. (2005) summarized and transformed the AAindex attributes of amino acids into five

multidimensional patterns of attribute covariation that reflected polarity, secondary structure, molecular volume, codon diversity, and electrostatic charge. We called these five transformed scores "AAFactors" and used them to encode the amino acid in our research.

GRANTHAM score

GRANTHAM score measures the differences of physico-chemical properties between amino acids (Grantham, 1974). Using it, we defined the feature of GRANTHAM score for each SAP that reflected the physicochemical difference between the original amino acid and changed amino acid.

One SAP includes 10 amino acids to encode: the original and changed amino acids of the SAP, the upstream 4 amino acids of the SAP and the downstream 4 amino acids of the SAP. Hence, each SAP has 1 betweenness, 214 KEGG enrichment scores, $5 \times 10 = 50$ AAFactors, $20 \times 9 = 180$ PSSM conservation scores, 9 disorder scores and 1 GRANTHAM score. In total, there were 455 features. By utilizing the concept of pseudo amino acid composition (PseAAC) (Chou, 2001), a SAP **P** can be generally formulated as vector with 455 components; i.e.,

$$\mathbf{P}=[\psi_1 \psi_2 \cdots \psi_u \cdots \psi_{455}]^T \quad (1)$$

where ψ_1 represents the 1st feature of the SAP, ψ_2 the 2nd feature, and so forth.

Predictor construction and evaluation

In this study, we used LibliearR (Fan et al., 2008) to classify the query SAP into disease SAP or polymorphism. LibliearR is an R interface to LIBLINEAR, a C/C++ library for large linear classification (Fan et al., 2008). LIBLINEAR not only has good theoretical properties, but also shows promising performance in practice (Fan et al., 2008; Hsieh et al., 2008; Keerthi et al., 2008; Lin et al., 2008). L2-regularized L2-loss support vector classification model (Fan et al., 2008) in LibliearR was applied to construct the predictor.

In statistical prediction, the following three cross-validation methods are often used to examine a predictor for its effectiveness in practical application: independent dataset test, subsampling test, and jackknife test (Chou and Zhang, 1995). However, as elucidated in Chou and Shen (2008) and demonstrated by Eqs.28-32 of Chou (2011), among the three cross-validation methods, the jackknife test is deemed the least arbitrary (most objective) that can always yield a unique result for a given benchmark dataset, and hence has been increasingly used and widely recognized by investigators to examine the accuracy of various predictors (Georgiou et al., 2009; Zeng et al., 2009; Esmaeili et al., 2010; Mohabatkar, 2010; Qiu et al., 2010; Hu et al., 2011a, 2011b; Huang et al., 2011a, 2011b; Lin et al., 2011; Wang et al., 2011; Xiao et al., 2011). Accordingly, the jackknife test, also known as Leave-One-Out Cross-Validation (LOOCV) (Huang et al., 2008; Cai et al., 2010; Huang et al., 2009, 2010a, 2010b) was adopted here to examine the quality of the present predictor. During LOOCV, each sample in the data set is used as test sample in turn and predicted by the model trained by the other samples. The sensitivity (Sn), specificity (Sp), accuracy (ACC) and Matthews's correlation coefficient (MCC) (Baldi et al., 2000) were calculated to measure the prediction performance:

$$\begin{cases} S_n = \dfrac{TP}{TP+FN} \\ S_p = \dfrac{TN}{TN+FP} \\ ACC = \dfrac{TP+TN}{TP+TN+FP+FN} \\ MCC = \dfrac{TP \times TN - FP \times FN}{\sqrt{(TP+FN)\times(TN+FP)\times(TP+FP)\times(TN+FN)}} \end{cases} \quad (2)$$

where TP, TN, FP and FN stand for the number of true positive, true negative, false positive and false negative samples, respectively.

Web server implementation

To facilitate the disease associated SAPs prediction, we implemented an automated pipeline of our method and developed a web server interface. Tomcat/Apache served as a J2EE container for JSP. The prediction model was implemented with LiblineaR (http://cran.r-project.org/package = LiblineaR)—a package of R programming language (http://www.r-project.org/). The web server runs on Linux system, and works with both the Microsoft Internet Explorer and Mozilla Firefox browsers.

For the users who have a large number of SAPs to predict or want to set up the mirror website of SySAP, the code of SySAP can be downloaded and easily run in command line. They only need to install R programming environment and R package LiblineaR first. All the data needed for prediction were pre-computed and can be directly downloaded from our website to speed up the computation efficiency.

To use SySAP, only the basic information that defines the SAP was needed: the protein name, mutation position, old amino acid and new amino acid at the mutation position. The protein name should be UniProt accession number of human protein, such as P04637. The mutation position should be the position where the SAP occurs and should not exceed the length of protein sequence. The old amino acid is the original amino acid at the mutation position. The new amino acid is the changed amino acid after mutation. Both the old and new amino acids should be one of the 20 standard amino acids.

Based on the input information, SySAP can predict the query whether SAP is disease SAP or polymorphism. If the protein name is not UniProt accession number or the mutation

position exceeds the length of protein sequence, or the old and new amino acid is not one of the 20 standard amino acids, the corresponding error information will be printed. So the user can correct their improper input and try again.

ACKNOWLEDGEMENTS

This work was supported by the National Basic Research Program of China (973 Program) (Grant Nos. 2011CB910204, 2010CB529206, and 2010CB912702), Research Program of the Chinese Academy of Sciences (KSCX2-EW-R-04, KSCX2-YW-R-190, 2011KIP204), National Natural Science Foundation of China (Grant Nos. 30900272 and 31070752), National Key Technology R&D Program in the 11th Five Year Plan of China (No. 2008BAI64B01), National High-Tech R&D Program of China (863 Program) (Grant No. 2009AA02Z304) and National Scientific-Basic Special Fund (No. 2009FY120100).

REFERENCES

Ahmad, S., and Sarai, A. (2005). PSSM-based prediction of DNA binding sites in proteins. BMC Bioinformatics 6, 33.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25, 3389–3402.

Atchley, W.R., Zhao, J., Fernandes, A.D., and Drüke, T. (2005). Solving the protein sequence metric problem. Proc Natl Acad Sci U S A 102, 6395–6400.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C.A., and Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 16, 412–424.

Burke, D.F., Worth, C.L., Priego, E.M., Cheng, T., Smink, L.J., Todd, J. A., and Blundell, T.L. (2007). Genome bioinformatic analysis of nonsynonymous SNPs. BMC Bioinformatics 8, 301.

Cai, Y., Huang, T., Hu, L., Shi, X., Xie, L., and Li, Y.Prediction of lysine ubiquitination with mRMR feature selection and analysis. Amino Acids. 2011 Jan 26. [Epub ahead of print].

Cai, Y.D., Huang, T., Feng, K.Y., Hu, L., and Xie, L. (2010). A unified 35-gene signature for both subtype classification and survival prediction in diffuse large B-cell lymphomas. PLoS One 5, e12726.

Care, M.A., Needham, C.J., Bulpitt, A.J., and Westhead, D.R. (2007). Deleterious SNP prediction: be mindful of your training data! Bioinformatics 23, 664–672.

Chou, K.C. (2001). Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins 43, 246–255.

Chou, K.C. (2011). Some remarks on protein attribute prediction and pseudo amino acid composition. J Theor Biol 273, 236–247.

Chou, K.C., and Shen, H.B. (2007). Recent progress in protein subcellular location prediction. Anal Biochem 370, 1–16.

Chou, K.C., and Shen, H.B. (2008). Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms. Nat Protoc 3, 153–162.

Chou, K.C., Wu, Z.C., and Xiao, X. (2011). iLoc-Euk: a multi-label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins. PLoS One 6, e18258.

Chou, K.C., and Zhang, C.T. (1995). Prediction of protein structural classes. Crit Rev Biochem Mol Biol 30, 275–349.

Esmaeili, M., Mohabatkar, H., and Mohsenzadeh, S. (2010). Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses. J Theor Biol 263, 203–209.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). LIBLINEAR: A library for large linear classification. J Mach Learn Res 9, 1871–1874.

Freeman, L.C. (1979). Centrality in social networks: Conceptual clarification. Soc Networks 1, 215–239.

Georgiou, D.N., Karakasidis, T.E., Nieto, J.J., and Torres, A. (2009). Use of fuzzy clustering technique and matrices to classify amino acids and its impact to Chou's pseudo amino acid composition. J Theor Biol 257, 17–26.

Grantham, R. (1974). Amino acid difference formula to help explain protein evolution. Science 185, 862–864.

Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A., and McKusick, V.A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 33, D514–D517.

Hsieh, C.-J., Chang, K.-W., Lin, C.-J., Keerthi, S.S., and Sundararajan, S. (2008). A dual coordinate descent method for large-scale linear SVM. In: Proceedings of the 25th international conference on Machine learning. Helsinki, Finland: ACM, 408–415.

Hu, L., Huang, T., Shi, X., Lu, W.C., Cai, Y.D., and Chou, K.C. (2011a). Predicting functions of proteins in mouse based on weighted protein-protein interaction network and protein hybrid properties. PLoS One 6, e14556.

Hu, L.L., Huang, T., Cai, Y.D., and Chou, K.C. (2011b). Prediction of body fluids where proteins are secreted into based on protein interaction network. PLoS One 6, e22989.

Huang, T., Chen, L., Cai, Y.D., and Chou, K.C. (2011a). Classification and analysis of regulatory pathways using graph property, biochemical and physicochemical property, and functional property. PLoS One 6, e25297.

Huang, T., Cui, W., Hu, L., Feng, K., Li, Y.X., and Cai, Y.D. (2009). Prediction of pharmacological and xenobiotic responses to drugs based on time course gene expression profiles. PLoS One 4, e8126.

Huang, T., Niu, S., Xu, Z., Huang, Y., Kong, X., Cai, Y.D., and Chou, K. C. (2011b). Predicting transcriptional activity of multiple site p53 mutants based on hybrid properties. PLoS One 6, e22940.

Huang, T., Shi, X.H., Wang, P., He, Z., Feng, K.Y., Hu, L., Kong, X., Li, Y.X., Cai, Y.D., and Chou, K.C. (2010a). Analysis and prediction of the metabolic stability of proteins based on their sequential features, subcellular locations and interaction networks. PLoS One 5, e10972.

Huang, T., Tu, K., Shyr, Y., Wei, C.C., Xie, L., and Li, Y.X. (2008). The prediction of interferon treatment effects based on time series microarray gene expression profiles. J Transl Med 6, 44.

Huang, T., Wang, P., Ye, Z.Q., Xu, H., He, Z., Feng, K.Y., Hu, L., Cui, W., Wang, K., Dong, X., *et al.* (2010b). Prediction of deleterious non-synonymous SNPs based on protein interaction network and hybrid properties. PLoS One 5, e11900.

Jensen, L.J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M., *et al.* (2009). STRING 8—a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res 37, D412–D416.

Kawashima, S., Ogata, H., and Kanehisa, M. (1999). AAindex: amino acid index database. Nucleic Acids Res 27, 368–369.

Keerthi, S.S., Sundararajan, S., Chang, K.-W., Hsieh, C.-J., and Lin, C.-J. (2008). A sequential dual method for large scale multi-class linear svms. In: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. Las Vegas, Nevada, USA: ACM, 408–416.

Li, S., Xi, L., Li, J., Wang, C., Lei, B., Shen, Y., Liu, H., Yao, X., and Li, B. (2011). In silico prediction of deleterious single amino acid polymorphisms from amino acid sequence. J Comput Chem 32, 1211–1216.

Lin, C.-J., Weng, R.C., and Keerthi, S.S. (2008). Trust region newton method for logistic regression. J Mach Learn Res 9, 627–650.

Lin, W.Z., Fang, J.A., Xiao, X., and Chou, K.C. (2011). iDNA-Prot: identification of DNA binding proteins using random forest with grey model. PLoS One 6, e24756.

Mohabatkar, H. (2010). Prediction of cyclin proteins using Chou's pseudo amino acid composition. Protein Pept Lett 17, 1207–1214.

Ng, P.C., and Henikoff, S. (2002). Accounting for human polymorphisms predicted to affect protein function. Genome Res 12, 436–446.

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

Niu, S., Huang, T., Feng, K., Cai, Y., and Li, Y. (2010). Prediction of tyrosine sulfation with mRMR feature selection and analysis. J Proteome Res 9, 6490–6497.

Peng, K., Radivojac, P., Vucetic, S., Dunker, A.K., and Obradovic, Z. (2006). Length-dependent prediction of protein intrinsic disorder. BMC Bioinformatics 7, 208.

Qiu, J.D., Huang, J.H., Shi, S.P., and Liang, R.P. (2010). Using the concept of Chou's pseudo amino acid composition to predict enzyme family classes: an approach with support vector machine based on discrete wavelet transform. Protein Pept Lett 17, 715–722.

Ramensky, V., Bork, P., and Sunyaev, S. (2002). Human non-synonymous SNPs: server and survey. Nucleic Acids Res 30, 3894–3900.

Sharan, R., Ulitsky, I., and Shamir, R. (2007). Network-based prediction of protein function. Mol Syst Biol 3, 88.

Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29, 308–311.

Stenson, P.D., Ball, E.V., Mort, M., Phillips, A.D., Shiel, J.A., Thomas, N.S., Abeysinghe, S., Krawczak, M., and Cooper, D.N. (2003). Human Gene Mutation Database (HGMD): 2003 update. Hum Mutat 21, 577–581.

Wang, P., Xiao, X., and Chou, K.C. (2011). NR-2L: a two-level predictor for identifying nuclear receptor subfamilies based on sequence-derived features. PLoS One 6, e23505.

Wu, Z.C., Xiao, X., and Chou, K.C. (2011). iLoc-Plant: a multi-label classifier for predicting the subcellular localization of plant proteins with both single and multiple sites. Mol Biosyst 7, 3287–3297.

Xiao, X., Wu, Z.C., and Chou, K.C. (2011). A multi-label classifier for predicting the subcellular localization of gram-negative bacterial proteins with both single and multiple sites. PLoS One 6, e20592.

Ye, Z.Q., Zhao, S.Q., Gao, G., Liu, X.Q., Langlois, R.E., Lu, H., and Wei, L. (2007). Finding new structural and sequence attributes to predict possible disease association of single amino acid polymorphism (SAP). Bioinformatics 23, 1444–1450.

Zeng, Y.H., Guo, Y.Z., Xiao, R.Q., Yang, L., Yu, L.Z., and Li, M.L. (2009). Using the augmented Chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach. J Theor Biol 259, 366–372.

