ELSEVIER
Volume 258
Issue 2
21 May 2009
ISSN 0022-5193
Journal of Theoretical Biology

Journal of Theoretical Biology 258 (2009) 289–293

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Predicting DNA- and RNA-binding proteins from sequences with kernel methods

Xiaojian Shao [a], Yingjie Tian [b], Lingyun Wu [c], Yong Wang [c], Ling Jing [a], Naiyang Deng [a,*]

[a] *College of Science, China Agricultural University, Beijing 100083, China*
[b] *Research Center on Fictitious Economy and Data Science, Chinese Academy of Sciences, Beijing 100190, China*
[c] *Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China*

ARTICLE INFO

Article history:
Received 23 August 2008
Received in revised form
8 December 2008
Accepted 19 January 2009
Available online 6 February 2009

Keywords:
Bioinformatics
Protein function
Conjoint triad
Nucleic-acid-binding protein
Support vector machine

ABSTRACT

In this paper, support vector machines (SVMs) are applied to predict the nucleic-acid-binding proteins. We constructed two classifiers to differentiate DNA/RNA-binding proteins from non-nucleic-acid-binding proteins by using a conjoint triad feature which extract information directly from amino acids sequence of protein. Both self-consistency and jackknife tests show promising results on the protein datasets in which the sequences identity is less than 25%. In the self-consistency test, the predictive accuracy is 90.37% for DNA-binding proteins and 89.70% for RNA-binding proteins. In the jackknife test, the predictive accuracies are 78.93% and 76.75%, respectively. Comparison results show that our method is very competitive by outperforming other previously published sequence-based prediction methods.

1. Introduction

DNA/RNA-binding proteins are functional proteins in a cell. DNA-binding proteins play key roles in many biological processes ranging from DNA packaging, replication to gene expression control (Luscombe and Thornton, 2002). While RNA-binding proteins interact with various RNAs at different stages of protein synthesis to control the process of protein synthesis (Siomi and Dreyfuss, 1997). How to differentiate DNA-, RNA-binding proteins from other proteins is a very important research topic in the proteomics fields.

Many lines of evidences have indicated that computational approaches can provide useful information for both basic research and drug discovery in a timely manner, such as structural bioinformatics (Chou, 2004), molecular docking (Chou et al., 2003; Li et al., 2007; Wang et al., 2008), pharmacophore modeling (Sirois et al., 2004), QSAR (Du et al., 2005, 2008; Prado-Prado et al., 2008), protein subcellular location prediction (Chou and Shen, 2007c, 2008a), identification of membrane proteins and their types (Chou and Shen, 2007b), identification of enzymes and their functional classes (Shen and Chou, 2007b), identification of proteases and their types (Chou and Shen, 2008b), protein cleavage site prediction (Chou, 1996; Shen and Chou, 2008), and signal peptide prediction (Chou and Shen, 2007a; Shen and Chou, 2007a). The present study is devoted to develop a computational approach by constructing support vector machines (SVMs) to predict the DNA- and RNA-binding proteins using sequence information.

SVMs have been widely used for the prediction of DNA-, RNA-binding proteins. The previous work use different representations of the protein to study these problems, such as encoding using the sequence information and/or the structural information (e.g. Bhardwaj et al., 2005; Cai and Lin, 2003; Han et al., 2004; Yu et al., 2006, etc.). More recently, there are some work on DNA-binding proteins recognition based on evolutionary profile (Kumar et al., 2007) and combination of ChIP–chip/expression and genomic sequences (Zhou and Liu, 2008). Although many studies on DNA-, RNA-binding proteins recognition, there is still a great need in improving the accuracy of the classifiers.

In this paper, we propose a new method based on SVMs for predicting DNA-binding and RNA-binding proteins. Two SVM classifiers are trained to identify the two types of nucleic-acid-binding proteins (which we call DNA-binding SVM, RNA-binding SVM, respectively). Importantly, a conjoint triad feature is applied to describe amino acids sequences, which has been successfully used for predicting protein–protein interactions (PPIs) (Shen et al., 2007). Our new approach is tested on the protein datasets in which the sequence identity is less than 25%. The results demonstrate the effectiveness and efficiency of our approaches. Besides these two classifiers, we also construct a classifier to differentiate DNA-binding proteins from RNA-binding proteins.

* Corresponding author. Tel.: +86 10 62736511; fax: +86 01 62736777.
E-mail address: dengnaiyang@vip.163.com (N. Deng).

0022-5193/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2009.01.024

2. Materials and methods

2.1. Data sets

We select the DNA-binding and RNA-binding proteins from Swiss-Prot database (version 52.0) (Boeckmann et al., 2003) by using the keywords "DNA-binding" and "RNA-binding", respectively. We follow the similar procedure by Cai and Lin (2003) to get the "contrast" data set by using a list of keywords suspicious of implying DNA/RNA-binding function. Then the proteins in the "contrast" data set are removed from Swiss-Prot database to get the "unlabel"data set.

As indicated in the paper (Yu et al., 2006), some proteins contain irregular amino acid characters and there exist redundancy among sequences in those data sets. These data are removed. In order to compare our method with that of Yu et al., we select proteins mainly from the materials which are supplied by Yu et al. except for some proteins which are excluded from the new released Swiss-Prot database (version 52.0).

Finally, we label 1090 DNA-binding proteins and 358 RNA-binding proteins as "positive" data for DNA-binding and RNA-binding prediction task, respectively, and 16 932 non-binding proteins as "unlabel" data. In order to maintain a balance between positive and negative training data in SVM training procedure, we create the "negative" dataset by randomly selecting a subset which has the equal size to the "positive" dataset from the "unlabel" dataset. As for distinguishing DNA-binding proteins from RNA-binding proteins, the 1090 DNA-binding proteins and 358 RNA-binding proteins are used as positive dataset and negative dataset, respectively, to construct the classifier. All the lists of the non-redundant DNA-binding, RNA-binding proteins, and the "unlabel" dataset used, in this paper, are available in supplementary materials.

2.2. Feature vector

To represent a protein, we use a novel descriptor named conjoint triad, which considers the properties of one amino acid and its vicinal amino acids and regards any three continuous amino acids as an unit. This descriptor has been successful applied in PPI prediction (Shen et al., 2007). First, the 20 amino acids are clustered into seven classes according to their dipoles and volumes of the side chains. The classes are as follows: $\{A, G, V\}$, $\{I, L, F, P\}$, $\{Y, M, T, S\}$, $\{H, N, Q, W\}$, $\{R, K\}$, $\{D, E\}$, $\{C\}$. Then, the conjoint triads are obtained from the clustered elements. Thus, a protein is described by a feature vector with $7 \times 7 \times 7 = 343$ dimensions (Shen et al., 2007), where each element of the vector has the value of the frequency of the corresponding triad.

2.3. Support vector machines

Given a training dataset $\mathscr{X} = \{(x_i, y_i) : x_i \in R^n, y_i \in \{+1, -1\}\}_{i=1}^m$, where each x_i is labeled by y_i. Linear SVM finds a boundary that separates two different classes of feature vectors with a maximum margin (Vapnik, 1995; Schölkopf et al., 2004). It leads to the following convex quadratic programming:

$$\min_{w,b} \quad \frac{1}{2} w^T w + C \sum_{i=1}^{m} \xi_i \tag{1}$$

$$\text{s.t.} \quad y_i(w^T x_i + b) \geqslant 1 - \xi_i \tag{2}$$

$$\xi_i \geqslant 0, \quad i = 1, 2, \ldots, m \tag{3}$$

A nonlinear SVM projects feature vectors into a high dimensional feature space by using a kernel function such as a Gaussian kernel function $K(x_i, x_j) = exp\{-\|x_i - x_j\|^2/2\sigma^2\}$. The linear SVM procedure is then applied to the feature vectors in this feature space.

The corresponding dual quadratic programming to (1)–(3) for nonlinear SVM is as follows:

$$\min_{\alpha} \quad \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j K(x_i, x_j) - \sum_{i=1}^{m} \alpha_i \tag{4}$$

$$\text{s.t.} \quad \sum_{i=1}^{m} y_i \alpha_i = 0, \quad 0 \leqslant \alpha_i \leqslant C, \ i = 1, \ldots, m \tag{5}$$

where C is a constant controlling the trade-off between maximizing the margin and minimizing the errors. After the determination of the solution α^* of (4)–(5) and b^*, we can get the decision function

$$f(x) = \text{sgn}\left\{ \sum_{i=1}^{m} \alpha_i^* y_i K(x_i, x) + b^* \right\} \tag{6}$$

where the function sgn is a sign function

$$\text{sgn}(x) = \begin{cases} 1 & \text{if } x > 0 \\ -1 & \text{otherwise} \end{cases}$$

In the present work, the Gaussian kernel function $K(x_i, x_j) = exp\{\|x_i - x_j\|^2/2\sigma^2\}$ is used. The SVM training algorithm is implemented in the LIBSVM (Chang and Lin, 2003), and the tunable parameters are selected by the grid method.

The performance of the SVMs has been evaluated by the following measures (Baldi et al., 2000): sensitivity $=$ TP/(TP $+$ FN), specificity $=$ TN/(TN $+$ FP), and the overall accuracy: accuracy $=$ (TP $+$ TN)/(TP $+$ TN $+$ FP $+$ FN). Here TP, TN, FP, FN are the number of true positives, true negatives, false positives, and false negatives, respectively. In addition, the area under the receiver operating characteristic curve (AUC), a common, unbiased measure of prediction accuracy, is also used to evaluate the performance of SVM.

3. Results

We compare our results with the recent work of Yu et al. As mentioned above, the differences between the training datasets are very small. For example, only 63 samples are excluded in our training data compared with that of Yu et al. (the total number of their training data is 1153) for DNA-binding SVM. In the following, our method with the triad features is denoted by triad-SVM while that of Yu et al. is denoted by Seq-SVM. Three methods are used to evaluate the performance. First, we use self-consistency test (see Bhardwaj et al., 2005; Yu et al., 2006, etc.) to show the performance of the new SVM, which will demonstrate how well SVM has turned into internal knowledge. Table 1 describes the result of our triad-SVM and Seq-SVM. It shows the new triad-SVMs perform better on both DNA-binding and RNA-binding SVMs.

Next the cross-validation test is conducted. 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 (2008a) and demonstrated by Eq. 50 of Chou and Shen (2007c), among the three cross-validation methods, the jackknife test is deemed the most objective that can always yield a unique result for a given benchmark dataset, and hence has been increasingly used by investigators to examine the accuracy of various predictors (see, e.g. Chen et al., 2008a, b; Chen and Li, 2007a, b; Du and Li, 2008; Jiang et al., 2008; Jin et al., 2008; Li and Li, 2008; Lin, 2008; Lin et al., 2008; Niu et al., 2008; Shi et al., 2008; Wu and Yan, 2008; Zhang and Fang, 2008; Zhou et al.,

Table 1
Comparison of triad-SVM with Seq-SVM on self-consistency test.

Protein	Sample size	Method	Accuracy (%)	Specificity (%)	Sensitivity (%)
DNA-binding	2180	triad-SVM	90.37	73.75	91.40
		Seq-SVM	74.37	66.78	81.96
RNA-binding	716	triad-SVM	89.70	90.24	89.55
		Seq-SVM	83.21	80.21	86.21

Table 2
Comparison of triad-SVM with Seq-SVM on jackknife test.

Protein	Sample size	Method	Accuracy (%)	Specificity (%)	Sensitivity (%)
DNA-binding	2180	triad-SVM	78.93	66.74	84.86
		Seq-SVM	71.64	63.90	79.38
RNA-binding	716	triad-SVM	76.75	74.81	78.70
		Seq-SVM	77.51	74.59	80.42

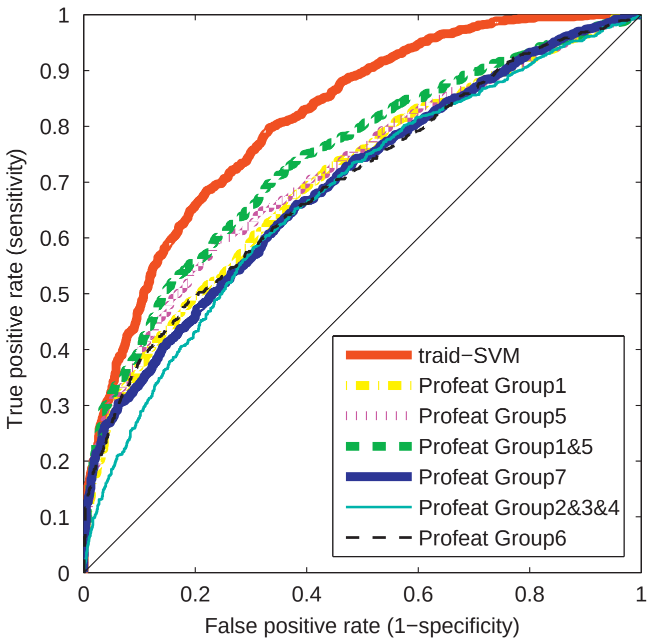

Fig. 1. ROC curves for cross validation tests on DNA-binding SVM with different feature encoding methods. Solid line (red): our triad-SVM, dash–dot line (yellow): profeat group1, dotted line (magenta): profeat group5, dashed line (green): profeat group1&5, solid line (blue): group7, solid line (cyan): group2&3&4 and dash line (black): group6. (For interpretation of the references to colors in this figure legend, the reader is referred to the web version of this article.)

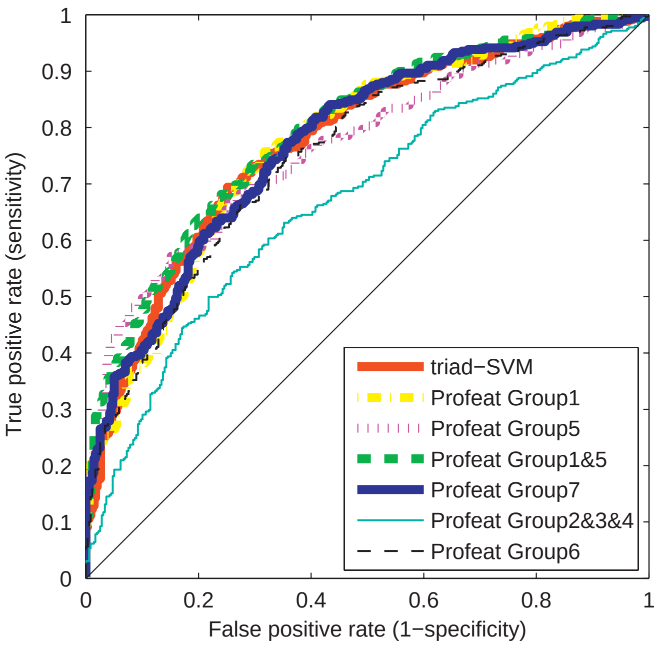

Fig. 2. ROC curves for cross validation tests on RNA-binding SVM with different feature encoding methods. Solid line (red): our triad-SVM, dash–dot line (yellow): profeat group1, dotted line (magenta): profeat group5, dashed line (green): profeat group1&5, solid line (blue): group7, solid line (cyan): group2&3&4 and dash line (black): group6. (For interpretation of the references to colors in this figure legend, the reader is referred to the web version of this article.)

2007). So here we use jackknife test to evaluate the performance of the triad-SVM. Table 2 describes the performance of our new triad-SVM with Seq-SVM on both DNA-binding and RNA-binding proteins. In this test, we get better results on DNA-binding SVM while get comparable results on RNA-binding SVM. Furthermore, we compare our triad-encoding feature with another popular sequence-based encoding feature method (profeat) (Li et al., 2006). Profeat has been successfully used to encoding protein in many applications. It has seven groups of features and the group1, group5 and group7 are the most reported features to represent protein. Figs. 1 and 2 show receiver operating characteristic (ROC) curves for DNA-binding SVM and RNA-binding SVM according to different feature encoding methods(triad-SVM and profeat-SVM), respectively. One can see that in DNA-binding protein the triad-SVM performs better than all other profeat-based SVMs (that is, use profeat-group1, profeat-group5, both group1&5, profeat-group7, combination of profeat-group2&3&4, and group6, respectively), average areas under the curves are 0.796, 0.701, 0.721, 0.733, 0.696, 0.677 and 0.698 for triad-SVM, profeat-group1, group5, profeat-group1&5, profeat-group7, profeat-group2&3&4 and profeat-group6. And it gets the comparable results on RNA-binding SVMs, the AUCs are 0.776, 0.773, 0.767, 0.792,0.775, 0.684 and 0.754, respectively.

In addition, we use holdout test (Bhardwaj et al., 2005) to validate the performance of the newly constructed SVMs. The accuracy achieved in this test could be treated as the one from

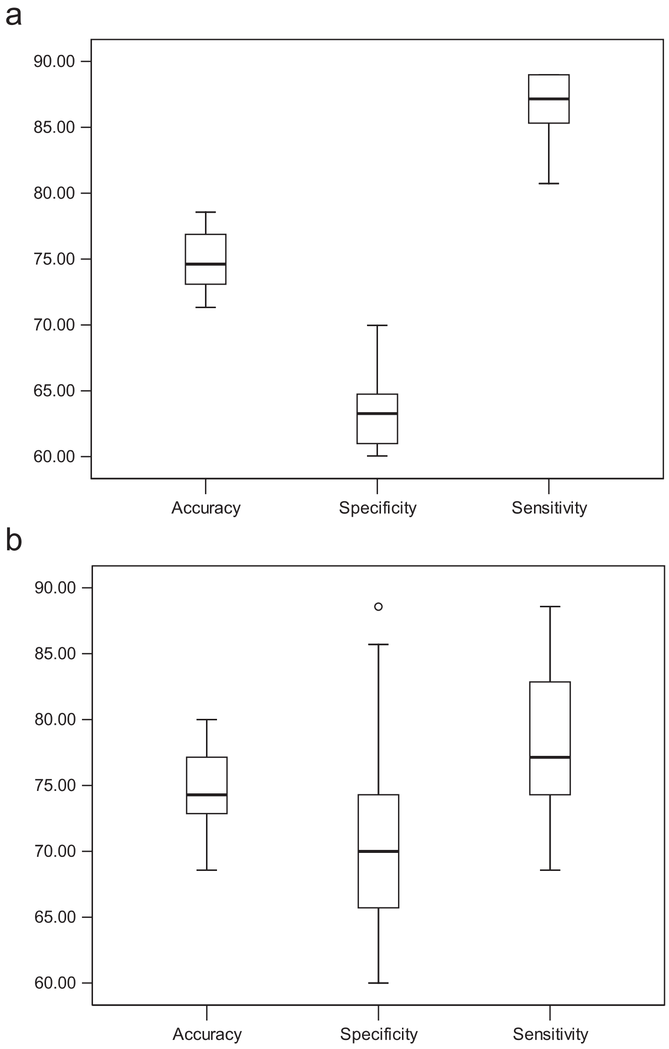

Fig. 3. Box plots for the performance of DNA/RNA-binding SVM for holdout test.

true blind prediction. However, holdout test has a very high variance depending on the division of the data into training and testing subset. Fig. 3 displays the variance on the three measures. The accuracies of DNA-binding SVM vary from 71.33% to 79.00%, with an average of 75.92%, and the average specificity of DNA-binding SVM is 64.52%, the average sensitivity is 86.31%. While the accuracies of RNA-binding SVM range between 68.57% and 80.00%, it gets an average of 71.63% and 77.96% by specificity and sensitivity measures.

4. Discussion

Our work of predicting DNA-binding proteins is motivated by the successful results of designing the triad descriptor for predicting protein-protein interactions. Encouragingly, we get better prediction rate (78.93%) for discriminating DNA-binding proteins from non-DNA-binding proteins. It shows the efficiency of combining the new feature representation and the kernel-based machine learning method (SVM). Although the RNA-binding SVM gets similar accuracy with that of Yu et al., our protein representation is much simpler. In the future, we will use other features in the construction of the descriptors to further improve the predictive accuracy, such as second structural information and moment information (Ahmad and Sarai, 2004) etc.

Table 3
Results of discriminating DNA-binding proteins from RNA-binding proteins.

Test method	Accuracy (%)	Specificity (%)	Sensitivity (%)
Self-consistency	92.03	82.97	94.27
Jackknife	80.25	72.71	82.73
Holdout	78.39	70.08	81.12

RNA- and DNA-binding proteins share many similar characteristics (Bhardwaj et al., 2005). However, we note DNA- and RNA-binding have different functional mechanisms. One evidence is that we get different performance by our method on DNA- and RNA-binding datasets. DNA-binding prediction has better accuracy than RNA-binding prediction. Possible reason is that the triad feature captures the local binding site pattern in DNA protein binding. Furthermore, we also investigate how well our current descriptor can discriminate these two protein classes. It is exciting that the classifier also gets high accuracy. Table 3 shows the result of the prediction on DNA-binding and RNA-binding proteins. It shows that the features abstracted from sequences by using of the conjoint triad are suitable to grasp the important information of the difference of DNA-binding and RNA-binding proteins.

We implement the experiments on the computer with 1.6 GHz CPU. The CPU time consumed for training DNA-, RNA-binding SVMs, and the DNA-RNA binding SVM are 9.84, 1.15 and 3.70 seconds, respectively. And for the jackknife testing, they are 6.09 hour, 11.58 minutes, and 1.5 hour, respectively.

Acknowledgments

The authors would like to thank the editor Janet Stein and the anonymous reviewers for valuable suggestions. This work is supported by the key project of the National Natural Science Foundation of China (no. 10631070) and the National Natural Science Foundation of China (no. 10601064).

Appendix A. Supplementary data

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jtbi.2009.01.024.

References

Ahmad, S., Sarai, A., 2004. Moment-based prediction of DNA-binding proteins. J. Mol. Biol. 341, 65–71.

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

Bhardwaj, N., Langlois, R.E., Zhao, G.J., Lu, H., 2005. Kernel-based machine learning protocol for predicting DNA-binding proteins. Nucleic Acids Res. 33 (20), 6486–6493.

Boeckmann, B., Bairoch, A., Apweiler, R., et al., 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. 31 (1), 365–370.

Cai, Y.D., Lin, S.L., 2003. Support vector machines for predicting rRNA-, RNA-, and DNA-binding proteins from amino acid sequence. Biochim. et Biophys. Acta 1648, 127–133.

Chang, C.C., Lin, C.J., 2003. LIBSVM: a library for support vector machines. Available from: ⟨http://www.csie.ntu.edu.tw/cjlin/papers/libsvm.pdf⟩.

Chen, C., Chen, L.X., Zou, X.Y., Cai, P.X., 2008a. Predicting protein structural class based on multi-features fusion. J. Theor. Biol. 253, 388–392.

Chen, K., Kurgan, L.A., Ruan, J., 2008b. Prediction of protein structural class using novel evolutionary collocation-based sequence representation. J. Comput. Chem. 29, 1596–1604.

Chen, Y.L., Li, Q.Z., 2007a. Prediction of apoptosis protein subcellular location using improved hybrid approach and pseudo amino acid composition. J. Theor. Biol. 248, 377–381.

Chen, Y.L., Li, Q.Z., 2007b. Prediction of the subcellular location of apoptosis proteins. J. Theor. Biol. 245, 775–783.

Chou, K.C., 1996. Review: prediction of HIV protease cleavage sites in proteins. Anal. Biochem. 233, 1–14.

Chou, K.C., 2004. Review: structural bioinformatics and its impact to biomedical science. Curr. Med. Chem. 11, 2105–2134.

Chou, K.C., Shen, H.B., 2007a. Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides. Biochem. Biophys. Res. Comm. 357, 633–640.

Chou, K.C., Shen, H.B., 2007b. MemType-2L: a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. Biochem. Biophys. Res. Comm. 360, 339–345.

Chou, K.C., Shen, H.B., 2007c. Review: recent progresses in protein subcellular location prediction. Anal. Biochem. 370, 1–16.

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

Chou, K.C., Shen, H.B., 2008b. ProtIdent: a web server for identifying proteases and their types by fusing functional domain and sequential evolution information. Biochem. Biophys. Res. Comm. 376, 321–325 doi:10.1016/j.bbrc.2008.08.125.

Chou, K.C., Zhang, C.T., 1995. Review: prediction of protein structural classes. Crit. Rev. Biochem. Mol. Biol. 30, 275–349.

Chou, K.C., Wei, D.Q., Zhong, W.Z., 2003. Binding mechanism of coronavirus main proteinase with ligands and its implication to drug design against SARS. (Erratum: Chou, K.C., Wei, D.Q., Zhong, W.Z., 2003. Binding mechanism of coronavirus main proteinase with ligands and its implication to drug design against SARS, vol. 310, 675). Biochem. Biophys. Res. Comm. 308, 148–151.

Du, P., Li, Y., 2008. Prediction of C-to-U RNA editing sites in plant mitochondria using both biochemical and evolutionary information. J. Theor. Biol. 253, 579–589.

Du, Q.S., Huang, R.B., Wei, Y.T., Du, L.Q., Chou, K.C., 2008. Multiple field three dimensional quantitative structure-activity relationship (MF-3D-QSAR). J. Comput. Chem. 29, 211–219.

Du, Q.S., Mezey, P.G., Chou, K.C., 2005. Heuristic molecular lipophilicity potential (HMLP): a 2D-QSAR study to LADH of molecular family pyrazole and derivatives. J. Comput. Chem. 26, 461–470.

Han, L.Y., Cai, C.Z., Lo, S.L., Chung, M.C.M., Chen, Y.Z., 2004. Prediction of RNA-binding proteins from primary sequence by a support vector machine approach. RNA 10, 355–368.

Jiang, X., Wei, R., Zhang, T.L., Gu, Q., 2008. Using the concept of Chou's pseudo amino acid composition to predict apoptosis proteins subcellular location: an approach by approximate entropy. Protein Pept. Lett. 15, 392–396.

Jin, Y., Niu, B., Feng, K.Y., Lu, W.C., Cai, Y.D., Li, G.Z., 2008. Predicting subcellular localization with AdaBoost learner. Protein Pept. Lett. 15, 286–289.

Kumar, M., Gromiha, M.M., Raghava, G.P.S., 2007. Identification of DNA-binding proteins using support vector machines and evolutionary profiles. BMC Bioinformatics 8, 463 doi:10.1186/1471-2105-8-463.

Li, F.M., Li, Q.Z., 2008. Predicting protein subcellular location using pseudo amino acid composition and improved hybrid approach. Protein Pept. Lett. 15, 612–616.

Li, Y., Wei, D.Q., Gao, W.N., Gao, H., Liu, B.N., Huang, C.J., Xu, W.R., Liu, D.K., Chen, H.F., Chou, K.C., 2007. Computational approach to drug design for oxazolidinones as antibacterial agents. Med. Chem. 3, 576–582.

Li, Z.R., Lin, H.H., Han, L.Y., Jiang, L., Chen, X., Chen, Y.Z., 2006. PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. Nucleic Acids Res. 34 (web server issue), W32–7.

Lin, H., 2008. The modified Mahalanobis discriminant for predicting outer membrane proteins by using pseudo amino acid composition. J. Theor. Biol. 252, 350–356.

Lin, H., Ding, H., Feng-Biao Guo, F.B., Zhang, A.Y., Huang, J., 2008. Predicting subcellular localization of mycobacterial proteins by using pseudo amino acid composition. Protein Pept. Lett. 15, 739–744.

Luscombe, N.M., Thornton, J.M., 2002. Protein-DNA interactions: amino acid conservation and the effects of mutations on binding specificity. J. Mol. Biol. 320 (5), 991–1009.

Niu, B., Jin, Y.H., Feng, K.Y., Liu, L., Lu, W.C., Cai, Y.D., Li, G.Z., 2008. Predicting membrane protein types with bagging learner. Protein Pept. Lett. 15, 590–594.

Prado-Prado, F.J., Gonzalez-Diaz, H., de la Vega, O.M., Ubeira, F.M., Chou, K.C., 2008. Unified QSAR approach to antimicrobials. Part 3: first multi-tasking QSAR model for Input-Coded prediction, structural back-projection, and complex networks clustering of antiprotozoal compounds. Bioorg. Med. Chem. 16, 5871–5880.

Schölkopf, B., Tsuda, K., Vert, J.P., 2004. Kernel Methods in Computational Biology. MIT Press, Cambridge, MA.

Shen, H.B., Chou, K.C., 2007a. Signal-3L: a 3-layer approach for predicting signal peptide. Biochem. Biophys. Res. Comm. 363, 297–303.

Shen, H.B., Chou, K.C., 2007b. EzyPred: a top-down approach for predicting enzyme functional classes and subclasses. Biochem. Biophys. Res. Comm. 364, 53–59.

Shen, H.B., Chou, K.C., 2008. HIVcleave: a web-server for predicting HIV protease cleavage sites in proteins. Anal. Biochem. 375, 388–390.

Shen, J.W., Zhang, J., Luo, X.M., Zhu, W.L., Yu, K.Q., Chen, K.X., Li, Y.X., Jiang, H.L., 2007. Predicting protein–protein interactions based only on sequences information. Proc. Nat. Acad. Sci. USA 104 (11), 4337–4341.

Shi, M.G., Huang, D.S., Li, X.L., 2008. A protein interaction network analysis for yeast integral membrane protein. Protein Pept. Lett. 15, 692–699.

Siomi, H., Dreyfuss, G., 1997. RNA-binding proteins as regulators of gene expression. Curr. Opinion Genet. Dev. 7 (3), 345–353.

Sirois, S., Wei, D.Q., Du, Q.S., Chou, K.C., 2004. Virtual screening for SARS-CoV protease based on KZ7088 pharmacophore points. J. Chem. Inf. Comput. Sci. 44, 1111–1122.

Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer, New York.

Wang, J.F., Wei, D.Q., Chen, C., Li, Y., Chou, K.C., 2008. Molecular modeling of two CYP2C19 SNPs and its implications for personalized drug design. Protein Pept. Lett. 15, 27–32.

Wu, G., Yan, S., 2008. Prediction of mutations in H3N2 hemagglutinins of Influenza A virus from North America based on different datasets. Protein Pept. Lett. 15, 144–152.

Yu, X.J., Cao, J.P., Cai, Y.D., Shi, T.L., Li, Y.X., 2006. Predicting rRNA-, RNA-, and DNA-binding proteins from primary structure with support vector machines. J. Theor. Biol. 240, 175–184 doi:10.1016/j.jtbi.2005.09.018.

Zhang, G.Y., Fang, B.S., 2008. Predicting the cofactors of oxidoreductases based on amino acid composition distribution and Chou's amphiphilic pseudo amino acid composition. J. Theor. Biol. 253, 310–315.

Zhou, Q., Liu, J.S., 2008. Extracting sequence features to predict protein–DNA interactions: a comparative study. Nucleic Acid Res. 36 (12), 4137–4148.

Zhou, X.B., Chen, C., Li, Z.C., Zou, X.Y., 2007. Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. J. Theor. Biol. 248, 546–551.