Amino Acids (2008) 34: 103–109
DOI 10.1007/s00726-007-0568-2
Printed in The Netherlands

Amino Acids

Predicting DNA-binding proteins: approached from Chou's pseudo amino acid composition and other specific sequence features

Y. Fang, Y. Guo, Y. Feng, and **M. Li**

College of Chemistry, Sichuan University, Chengdu, China

Received March 25, 2007
Accepted May 23, 2007
Published online July 12, 2007;

Summary. DNA-binding proteins play a pivotal role in gene regulation. It is vitally important to develop an automated and efficient method for timely identification of novel DNA-binding proteins. In this study, we proposed a method based on alone the primary sequences of proteins to predict the DNA-binding proteins. DNA-binding proteins were encoded by autocross-covariance transform, pseudo-amino acid composition, dipeptide composition, respectively and also the different combinations of the three encoded methods; further, these feature matrices were applied to support vector machine classifiers to predict the DNA-binding proteins. All modules were trained and validated by the jackknife cross-validation test. Through comparing the performance of these substituted modules, the best result was obtained from pseudo-amino acid composition with the overall accuracy of 96.6% and the sensitivity of 90.7%. The results suggest that it can efficiently predict the novel DNA-binding proteins only using the primary sequences.

Keywords: DNA-binding proteins – Autocross-covariance transform – Pseudo-amino acid composition – Dipeptide composition – Support vector machine

1. Introduction

DNA-binding proteins (DNA-BPs) play a key role in the regulation of gene expression. It is estimated that in the human genome the total number of transcription factors alone can be as high as 3000 or about 10% of all protein-coding genes (Lander et al., 2001). With increasing availability of protein sequence data, there is an urgent need for computational tools that can rapidly and reliably identify DNA-BPs. Hence, there has been significant interest in developing computational methods for identification of amino acid residues that participate in protein-DNA interactions based on the integrated information of sequence, structure and evolution, and also the chemical or physical properties of amino acids. Jones et al. (2003) analyzed residue patches on the surface of DNA-BPs and used electrostatic potentials of residues to predict DNA-binding sites. They further applied this method to identify three specific classes of DNA-BPs, based on the presence of solvent accessible DNA-binding structure motifs (Shanahan et al., 2004). As for the related work, Tsuchiya et al. (2004) used a structure-based method to identify DNA-BPs based on electrostatic potentials and surface shape and Keil et al. (2004) trained a neural network classifier to identify patches that likely to be DNA-binding motifs based on physical and chemical properties of the patches. Neural network classifiers have also been used to identify DNA-BPs based on a combination of sequence neighbor and structure information (Ahmad et al., 2004). Recently Ahmad and Sarai have proposed a sequence-based method for predicting DNA-BPs. This method incorporated sequence alignment profiles into the input (Ahmad and Sarai, 2005). Kuznetsov et al. (2006) predicted DNA-BPs based on evolutionary and structural information of proteins and Bhardwaj et al. (2005) constructed a kernel-based machine learning protocol for predicting DNA-binding proteins based on electrostatic potentials and amino acid composition.

The aim of this paper is to develop a method that is independent of any DNA-BPs prediction both at training and predicting steps, but only the primary sequences. So a new investigation of autocross-covariance (ACC) transform, pseudo-amino acid composition and dipeptide composition with Support Vector Machine (SVM) was implemented to predict the DNA-BPs. First, the inquired primary sequence was transformed to numeric series by ACC transform, pseudo-amino acid composition and dipeptide composition technology respectively; then we in-

tegrated the numeric series of the sequences of proteins; finally, each of the numeric series was used as feature matrices to construct SVM modules. The results indicate that pseudo-amino acid composition substituted module by jackknife testing performs better than other modules, and it also suggest that this method can efficiently make predictions of DNA-BPs only using the primary sequences.

2. Materials and methods

2.1 Data sets

A positive data set of 118 DNA-BPs was obtained from a union of datasets used in previously reported studies (Jones et al., 2003; Stawiski et al., 2003; Ahmad et al., 2004; Bhardwaj et al., 2005). The negative dataset of 231 non-DNA-BPs was also adopted from a union of datasets used in earlier studies (Stawiski et al., 2003; Bhardwaj et al., 2005). These proteins have less than 35% sequence identity between each pairs. A complete list of all the PDB codes was list in Appendix A.

2.2 Autocross-covariance transform

ACC transform, a simplified approach of the covariant discriminant algorithm (Chou and Maggiora, 1998; Liu and Chou, 1998; Zhou, 1998; Zhou and Assa-Munt, 2001), has been applied in several studies (Wold et al., 1993; Sjöström et al., 1995; Edman et al., 1999; Du and Li, 2006; Guo et al., 2006b). The sequences of DNA-BPs and non-DNA-BPs were translated into numerical arrays by representing each amino acid with three z-scales derived by Hellberg et al. (1987). The three descriptor scales are the principal components of 29 physicochemical properties of amino acids and represent hydrophobicity (z1), steric properties (z2) and electronic properties (z3) respectively. The details of 29 physicochemical properties of amino acids were list in Table 1. The ACC terms were calculated according to Eq. (1) with lags [-lg, lg]. The result is a new multivariate data matrix with dimensionality m (the number of sequences) times $(2 \times \lg + 1) \times P^2$ (variables).

$$\mathrm{ACC}_{x(j,k),\mathrm{lag}} = \sum_{i=1}^{N_x-|\mathrm{lag}|} \left(x_j(i+\mathrm{lag}) - \frac{1}{N_x}\sum_{i=1}^{N_x} x_j(i) \right) \left(x_k(i) - \frac{1}{N_x}\sum_{i=1}^{N_x} x_k(i) \right) \tag{1}$$

Here P is the number of descriptor scales and lg is the maximum lag (lag = [−lg, lg]); indices j and k are used for the scales ($j = 1, \ldots, P$ and $k = 1, \ldots, P$); N_x is the length of the xth sequence ($x = 1, \ldots, N_x$); indice i is the position of a given sequence of protein; $x_k(i)$ is the ith amino acid of a given protein coded by the kth scale. Here the descriptor scales are the three z-scales of 29 physicochemical properties of amino acids, so P equals to 3 and according to the results of Sjöström et al. (1995), lg equals to 25. So the sequences of variable lengths are transformed into the 459 $((2 \times 25 + 1) \times 3^2)$-length feature vectors in this way.

Table 1. The variables of 29 physicochemical properties of amino acids (Hellberg et al., 1987)

Variable no.	Property
1	molecular weight
2	pK_{COOH} (COOH on C_α)
3	p$K_{\mathrm{NH_2}}$ (NH_2 on C_α)
4	pI, pH at the isoelectric point
5	substituent van der waals volume
6	^{1}H NMR for C_α–H (cation)
7	^{1}H NMR for C_α–H (dipolar)
8	^{1}H NMR for C_α–H (anion)
9	^{13}C NMR for C=O
10	^{13}C NMR for C_α–H
11	^{13}C NMR for C=O in tetrapeptide
12	^{13}C NMR for C_α–H in tetrapeptide
13	R_f for 1-N-(4-nitrobenzofurazono)amino acids in ethyl acetate/pyridine/water
14	slope of plot $1/(R_f\text{-}1)$ vs. mol% H_2O in paper chromatography
15	dG of transfer of amino acids from organic solvent to water
16	hydration potential or free energy of transfer from vapor phase to water
17	R_f, salt chromatography
18	log P, partition coefficient for amino acids in octanol/water
19	log D, partition coefficient at pH 7.1 for acetylamide derivatives of amino acids in octanol water
20	$dG = RT \ln f$; f = fraction of buried/accessible amino acids in 22 proteins
21–29	HPLC retention times for nine combinations of three different pH and three eluent mixtures

2.3 Pseudo-amino acid composition

To approximately incorporate the sequence-order effects (Chou, 2000a), the concept of the pseudo-amino acid composition was proposed (Chou, 2000b, 2001, 2005a, b) and has been used via various approaches to enhance the prediction quality (Chou and Cai, 2003; Gao et al., 2005; Xiao et al., 2005a, b, 2006a, b, c; Chou and Shen, 2006d). Recently, a very powerful predictor based on pseudo-amino acid composition was developed to predict the protein–protein interaction (Chou and Cai, 2006). The sequences of DNA-BPs and non-DNA-BPs were translated into numerical order series by representing each amino acid with the first principal component (z1) of 29 physicochemical properties of amino acids which represented hydrophobicity. Now following the same procedure as described by Chou (2001, 2005b), a protein P can be expressed by a vector or a point in a $(20 + \lambda)$D space; that is

$$\mathbf{P} = (P_1, P_2, \ldots, P_{20}, P_{20+1}, P_{20+2}, \ldots, P_{20+\lambda})^{\mathrm{T}} \tag{2}$$

where T is the transpose operate, and

$$P_k = \begin{cases} \dfrac{f_i}{\sum_{i=1}^{20} f_i + w \sum_{j=1}^{\lambda} \tau_j} & 1 \le k \le 20 \\ \dfrac{w h_j}{\sum_{i=1}^{20} f_i + w \sum_{j=1}^{\lambda} \tau_j} & 20 + 1 \le k \le 20 + \lambda \end{cases} \tag{3}$$

where f_i is normalized occurrence frequency of the 20 amino acids in the protein $\mathbf{P}$, τ_j is the j-rank sequence coupling factor computed according to Chou's method (Chou, 2005a, b), and w is the weight factor for the sequence-order effect. Here we chose $w = 0.05$. As we can see in Eqs. (2) and (3), the first 20 components reflect the effect of amino acid composition, whereas the components from $20 + 1$ to $20 + \lambda$ reflect the effect of sequence order.

For different datasets, lambda (λ) usually has different optimal value (Chou, 2001). The maximum λ is chosen as 30, because the minimum length of the sequences of proteins is 35 and the λ should be less than it. The results of different λ to the performance were shown in the Fig. 1. It shows that the results are influenced greatly when λ is between 1 and 10 and hardly influenced when λ is between 10 and 30. So the λ can be chosen a number between 10 and 30. For the current study, the optimal value of λ was chosen as 20. Given a protein, the $(20 + 20) = 40$ pseudo-

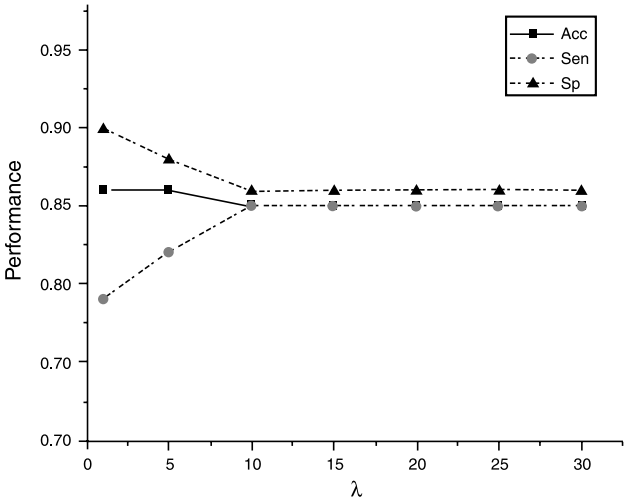

Fig. 1. The performance based on the pseudo-amino acid composition was influenced at different lambda. The abscissa represents the performance of overall accuracy, sensitivity and specificity and the ordinate represents the different lambda

amino acid components can be easily derived by following the procedures given by Chou and Cai (2006). A brief and clear description for how to use pseudo-amino acid composition has been given by Chou and Cai (2005). Because the pseudo amino acid composition discrete model has been widely used, recently a Web-server called PseAA was established at http://chou.med.harvard.edu/bioinf/PseAA/. Using the Web-server, one can easily generate the pseudo amino acid components for any given protein sequence.

2.4 Dipeptide composition

The dipeptide composition (Liu and Chou, 1998) has been successfully used to predict protein secondary structure contents and mitochondria proteins (Tan et al., 2006). The dipeptide composition used as input can provide global information on protein features in the form of fixed-length vector. It is calculated as follow for each protein.

$$\mathrm{Fdip(i)} = \frac{\text{total number of dip(i)}}{\text{total number of all dipeptides}} \tag{4}$$

where Fdip(i) is the fraction of dip(i) that the *i*th dipeptide out of 400 dipeptides.

Compared with native-amino acid composition (the fraction of each native amino acid in a protein), the advantage of dipeptide composition is that it incorporates some sequence-order information. With dipeptide composition coding scheme, each protein was represented as a fixed pattern length of 400 (20×20) elements.

2.5 Support vector machine

Support vector machine is a kind of learning machine based on statistical learning theory presented by Vapnik (1998). A brief and clear description for how to use SVM to do classification has been given by Chou and Cai (2002) and Cai et al. (2003). In this particular work, the DNA-BPs were defined as one class (labeled as $+1$) and the non- DNA-BPs were defined the other one (labeled as -1). The SVMs were implemented in MATLAB7.0. Radial basic function (RBF) was chosen as the kernel function and quadratic programming (QP) method was introduced to solve the optimization problem. All the parameters were kept constant except for C (regulatory parameter) and σ (the kernel width parameter). In the training process, C and σ were optimized.

Table 2. Indices introduced to evaluate the DNA-binding protein based on support vector machine method

Index	Definition and formula
Acc	$(TP + TN)/(TP + TN + FP + FN)$
Sen	$TP/(TP + FN)$
Sp	$TP/(TP + FP)$
R	$\dfrac{2(TP/(TP+FN) - FP/(TN+FP))}{1 + \mathrm{abs}(TP/(TP+FN) - FP/(TN+FP))}$
MCC	$\dfrac{TP \cdot TN - FN \cdot FP}{\sqrt{(TP+FN)\cdot(TP+FP)\cdot(TN+FN)\cdot(TN+FP)}}$

TP (*true positive*) The number of observed positive samples, predicted positive samples
TN (*true negative*) The number of observed negative samples, predicted negative samples
FP (*false positive*) The number of observed negative samples, predicted positive samples
FN (*false negative*) The number of observed positive samples, predicted negative samples
Acc Overall accuracy; *Sen* sensitivity; *Sp* specificity; *R* reliability; *MCC* Matthews's correlation coefficient

2.6 Performance evaluation

The jackknife (leave-one-out) test has been considered as one of the most objective and rigorous test procedure in examining the power of a prediction method, as illustrated in a comprehensive review article (Chou and Zhang, 1995). It has been increasingly utilized by leading investigators to examine the quality of various prediction methods (see, e.g., Zhou, 1998; Du et al., 2003; Zhou and Doctor, 2003; Wang et al., 2004, 2006; Chou and Cai, 2005; Shen and Chou, 2005a, b, 2006, 2007a, b, c, d; Chou and Shen, 2006a, b, c, d, 2007a, b, c; Chen et al., 2006; Du and Li, 2006; Du et al., 2006; Gao and Wang, 2006; Guo et al., 2006a; Mondal et al., 2006; Shen et al., 2006, 2007; Xiao et al., 2006a, b; Zhang et al., 2006; Lin and Li, 2007; Liu et al., 2007a, b). In this paper, a jackknife procedure was carried out. All the protein sequences in the datasets were in turn singled out as a 'testing set' and all the remaining proteins as the 'training set'. Five parameters were employed to evaluate the performance of each module, including Acc, Sen, Sp, MCC and R. Details of these indices are listed in Table 2 (Liu et al., 2006).

3. Results and discussion

3.1 Prediction results

The performance of all modules developed in this study is shown in Table 3. The performance of all modules was evaluated by jackknife testing. The pseudo-amino acid composition based SVM module yielded 96.6% overall accuracy and 90.7% sensitivity. The performance of dipeptide composition based SVM module was satisfactory but gave with the relatively lower overall accuracy (85.4%) and sensitivity (68.6%) in comparison with the pseudo-amino acid composition based module. In the case of the ACC based module, the overall accuracy was nearly 25% lower than the pseudo-amino acid composition based module and nearly 10% lower than the dipeptide compo-

Table 3. The performance of the methods based on different substituted models in identifying DNA-binding proteins by jackknife testing

Approach	Acc	Sen	Sp	MCC	*R*
PseAA-based(A)	0.966	0.907	0.996	0.924	0.949
dp based(B)	0.854	0.686	0.939	0.664	0.769
ACC based(C)	0.756	0.280	1.00	0.452	0.438
NAA-based	0.799	0.483	0.961	0.536	0.615
Hybrid1(A + B)	0.897	0.881	0.905	0.774	0.880
Hybrid2(A + C)	0.756	0.280	1.00	0.452	0.438
Hybrid3(B + C)	0.756	0.280	1.00	0.452	0.438
Hybrid(A + B + C)	0.756	0.280	1.00	0.452	0.438

PseAA Pseudo-amino acid composition; *dp* dipeptide composition; *ACC* autocross-covariance transform; *NAA* native amino acid composition

sition based module. A module based on the native-amino acid composition was also constructed. The performance of this module was not good (Table 3) compared to the pseudo-amino acid composition based module and the dipeptide composition based module but a little better than the ACC based module. Thus, pseudo-amino acid composition, which provided information about amino acid composition as well as local order of amino acids, is a better feature for predicting DNA-binding proteins. This observation is consistent with the suggestion that DNA-binding residues are likely to be conserved (because of their function).

To further study the three encoding methods, hybrid modules on the basis of various features of proteins were constructed. The first hybrid module (hybrid1) was developed on the basis of pseudo-amino acid composition and dipeptide composition of proteins. The prediction overall accuracy and sensitivity of hybrid1 module was 89.7 and 88.1%, respectively, which was better than the dipeptide composition based module but worse than the pseudo-amino acid composition based module. The other three hybrid modules (hybrid2, hybrid3, hybrid) containing ACC substituted matrices were shown the same performance as the ACC based module alone through the jackknife testing. These three hybrid modules were respectively developed on basis of pseudo-amino acid composition and ACC (hybrid2), dipeptide composition and ACC (hybrid3), pseudo-amino acid composition, dipeptide composition and ACC (hybrid), as shown in Table 3. These hybrid approaches have no any improvements in identifying the DNA-binding proteins. The reason may be that the three descriptor scales of the principal components of 29 physicochemical properties of amino acids can not exactly represent the sequences of proteins and these feature vectors can not be concatenated in this simple way. Comparing the eight substituted modules, the best prediction performance was the module on the basis of the pseudo-amino acid composition. So in this study, the pseudo-amino acid composition module was applied for differentiating the DNA-binding proteins from the non-DNA-binding proteins.

3.2 Comparison with other prediction methods

The performance of the pseudo-amino acid composition module developed in this study was compared with existing methods that were also developed from the same dataset. The performance of the previously reported studies are Bhardwaj et al. (2005), with sensitivity of 80.6%, Jones et al. (2003) with sensitivity of 67.8% and Kuznetsov et al. (2006), with sensitivity of 79.2%. These previous approaches in the classification of DNA-BPs mainly based on the structure factors such as overall charge, electrostatic calculations etc. The results demonstrated that the performance of pseudo-amino acid composition module is superior to those previous studies.

4. Conclusions

In this work, we compared several different substituted modules in differentiating the DNA-BPs from non-DNA-BPs based on the primary sequences of proteins. The classifier of pseudo-amino acid composition with SVM offers the best performance for identifying DNA-BPs from other proteins. The module based on the pseudo-amino acid composition gives the overall accuracy of 96.6% and sensitivity of 90.7%. The good result indicates that this method may be helpful to further study the details of the specific interactions of the DNA-BPs on the base of pseudo-amino acid composition. We can draw a conclusion that it is reasonable and feasible to develop a successful method only using the primary sequences of proteins to predict the DNA-BPs, which is helpful for annotating the DNA-binding proteins in the absence of experiment data. Such methods can be a supplement to biochemical experiments and help to provide insight in finding the targets of proteins for drug discovery. Further works on samples collection for DNA-binding proteins, refined negative samples selection, and feature vector selection will further improve the performance of the machine learning methods for predicting the DNA-protein interaction sites.

Acknowledgement

This work was sustentated by Student Innovation Found of Sichuan University of the People's Republic of China (No. 2006L012). The authors would like to express their cordial thanks to the unknown reviewers for providing comments on the manuscript.

Appendix A

Complete list of proteins with less than 35% identity used for this study

PDB codes of protein-DNA complexes positive cases (DNA-binding)

1A1H	1CKT	1HDD	1REP	
1A31	1CMA	1HLO	1RVA	3MHT
1A36	1CRX	1HRY	1SKN	6CRO
1A3Q	1CRZ	1HWT	1SVC	1CW0
1A73	1D02	1IF1	1T7P	1A02
1AOI	1D66	1IGN	1TAU	1A74
1AU7	1DCT	1IHF	1TC3	1AAY
1AZP	1DFM	1J59	1TF3	1AZQ
1B3T	1DIZ	1LMB	1TRO	1J59
1BC8	1DMU	1MDY	1TRR	1AM9
1BDT	1DP7	1MHD	1TSR	1BDT
1BF4	1ECR	1MNM	1TUP	1MJO
1BF5	1EMH	1MSE	1UBD	1PAR
1BG1	1EON	1OCT	1VAS	1QPZ
1BHM	1EQZ	1PDN	1XBR	1SRS
1BL0	1EWN	1PER	1YTB	1VOL
1BNK	1FJL	1PNR	1YUI	1YRN
1BP7	1FOK	1PUE	1ZQF	1YSA
1BPY	1GCC	1PVI	2BOP	2CGP
1C0W	1GD2	1PYI	2DNJ	2GLI
1C9B	1GAT	1QPI	2DNJ	3CRO
1CDW	1GDT	1QPS	2HDC	2IRF
1CF7	1HCQ	1QRV	2HMI	1DDN
1CJG	1HCR	1QUM	3HTS	

PDB codes of non-protein-DNA complexes negative cases (non-DNA-binding)

1A8E	1ATG	1BX7	1DRW	1HOE	1MAZ	1PDO	1RZL
1A8P	1AUK	1BYB	1DUN	1HPM	1MBA	1PEA	1SBP
1A8Y	1AV4	1C52	1DXY	1HTP	1MLA	1PGS	1SEK
1A53	1AXN	1CA1	1EAF	1HXN	1MOQ	1PHD	1SFP
1AAC	1AYL	1CEC	1ECY	1HYP	1MPP	1PHK	1SKF
1ABE	1B0B	1CEM	1EDG	1IAE	1MRP	1PHP	1SMD
1AC5	1B6A	1CFB	1ESC	1IDO	1MSK	1PHR	1SRA
1AEW	1B51	1CHD	1EZM	1IFC	1MUP	1PHT	1SUR
1AH7	1BA3	1CIY	1FCE	1INP	1NAR	1PLC	1SVB
1AHO	1BB9	1CLC	1FDS	1IOV	1NDH	1PMI	1SVY
1AIR	1BD8	1CNV	1FIT	1JDW	1NEU	1PNE	1SYM
1AJJ	1BDB	1COT	1FKJ	1JER	1NFP	1POA	1TCA
1AL3	1BDO	1CPO	1FMK	1KLO	1NG1	1POC	1TDE
1ALH	1BEA	1CPQ	1FNC	1KOE	1NIF	1POT	1TEN
1ALU	1BEO	1CPT	1FRB	1KPF	1NKR	1PPN	1TFE
1ALY	1BFD	1CSN	1FUA	1KTE	1NLS	1PRN	1THV
1AMF	1BG6	1CTJ	1FUS	1KUH	1NNC	1PTF	1TML
1AMK	1BGC	1CTT	1G3P	1LAM	1NOX	1PUC	1TMY
1AMM	1BHE	1CV8	1GAI	1LBU	1NPK	1RA9	1TN3
1AMP	1BHP	1CVL	1GCA	1LCL	1NSJ	1RB9	1TON
1AMX	1BJ7	1CYO	1GEN	1LED	1OBR	1RCB	1TRY
1AOA	1BK0	1CZJ	1GKY	1LFO	1OPR	1REC	1TUL
1AOL	1BOB	1DDT	1GND	1LID	1OPS	1RFS	1UCH
1AQB	1BPI	1DFX	1GOF	1LIT	1OPY	1RH4	1UOK
1ARB	1BQK	1DHN	1GPR	1LKI	1OSA	1RHS	1USH
1ARU	1BR9	1DHR	1GSA	1LST	1OYC	1RIE	1UTG
1ASH	1BS9	1DIN	1HCZ	1LTM	1PBE	1RKD	1VLS
1ASS	1BTN	1DOI	1HFC	1MAI	1PBV	1RMG	1WL9
1AT0	1BV1	1DPE	1HKA	1MAT	1PDA	1RSY	

References

Ahmad S, Gromiha MM, Sarai A (2004) Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. Bioinformatics 20: 477–486

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

Bhardwaj N, Langlois RE, Zhao G, Lu H (2005) kernel-based machine learning protocol for predicting DNA-binding proteins. Nucleic Acids Res 33: 6486–6493

Cai YD, Zhou GP, Chou KC (2003) Support vector machines for predicting membrane protein types by using functional domain composition. Biophys J 84: 3257–3263

Chen C, Tian YX, Zou XY, Cai PX, Mo JY (2006) Using pseudo-amino acid composition and support vector machine to predict protein structural class. J Theor Biol 243: 444–448

Chou KC (2000a) Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. Biochem Biophys Res Commun 278: 477–483

Chou KC (2000b) Review: prediction of protein structural classes and subcellular locations. Curr Protein Pept Sci 1: 171–208

Chou KC (2001) Prediction of protein cellular attributes using pseudo amino acid composition. Proteins 43: 246–255

Chou KC (2005a) Review: progress in protein structural class prediction and its impact to bioinformatics and proteomics. Curr Protein Pept Sci 6: 423–436

Chou KC (2005b) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21: 10–19

Chou KC, Cai YD (2002) Using functional domain composition and support vector machines for prediction of protein subcellular location. J Biol Chem 277: 45765–45769

Chou KC, Cai YD (2003) Prediction and classification of protein subcellular location: sequence-order effect and pseudo amino acid composition. J Cell Biochem 90: 1250–1260

Chou KC, Cai YD (2005) Prediction of membrane protein types by incorporating amphipathic effects. J Chem Inf Model 45: 407–413

Chou KC, Cai YD (2006) Predicting protein–protein interactions from sequences in a hybridization space. J Proteome Res 5: 316–322

Chou KC, Maggiora GM (1998) Domain structural class prediction. Protein Eng 11: 523–538

Chou KC, Shen HB (2006a) Hum-PLoc: a novel ensemble classifier for predicting human protein subcellular localization. Biochem Biophys Res Commun 347: 150–157

Chou KC, Shen HB (2006b) Large-scale predictions of Gram-negative bacterial protein subcellular locations. J Proteome Res 5: 3420–3428

Chou KC, Shen HB (2006c) Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers. J Proteome Res 5: 1888–1897

Chou KC, Shen HB (2006d) Predicting protein subcellular location by fusing multiple classifiers. J Cell Biochem 99: 517–527

Chou KC, Shen HB (2007a) Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. J Proteome Res 6: 1728–1734

Chou KC, Shen HB (2007b) Large-scale plant protein subcellular location prediction. J Cell Biochem 100: 665–678

Chou KC, Shen HB (2007c) Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides. Biochem Biophys Res Comm 357: 633–640

Chou KC, Zhang CT (1995) Review: prediction of protein structural classes. Crit Rev Biochem Mol Biol 30: 275–349

Du P, Li Y (2006) Prediction of protein submitochondria locations by hybridizing pseudo-amino acid composition with various physi-

cochemical features of segmented sequence. BMC Bioinformatics 7: 518

Du QS, Jiang ZQ, He WZ, Li DP, Chou KC (2006) Amino acid principal component analysis (AAPCA) and its applications in protein structural class prediction. J Biomol Struct Dyn 23: 635–640

Du QS, Wei DQ, Chou KC (2003) Correlation of amino acids in proteins. Peptides 24: 1863–1869

Edman M, Jarhede T, Sjöström M, Wieslander A (1999) Different sequence patterns in signal peptides from mycoplasmas, other gram-positive bacteria, and *Escherichia coli*: a multivariate data analysis. Proteins 35: 195–205

Gao QB, Wang ZZ (2006) Classification of G-protein coupled receptors at four levels. Protein Eng Des Sel 19: 511–516

Gao Y, Shao SH, Xiao X, Ding YS, Huang YS, Huang ZD, Chou KC (2005) Using pseudo amino acid composition to predict protein subcellular location: approached with Lyapunov index, Bessel function, and Chebyshev filter. Amino Acids 28: 373–376

Guo J, Lin Y, Liu X (2006a) GNBSL: a new integrative system to predict the subcellular location for Gram-negative bacteria proteins. Proteomics 6: 5099–5105

Guo Y, Li M, Lu M, Wen Z, Huang Z (2006b) Predicting GPCR-G-protein coupling specificity based on autocross-covariance transform. Proteins Struct Func Bioinformatics 65: 55–60

Hellberg S, Sjostrom M, Skagerberg B, Wold S (1987) Peptide quantitative structure-activity relationships, a multivariate approach. J Med Chem 30: 1126–1135

Jones S, Shanahan HP, Berman HM, Thornton JM (2003) Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins. Nucleic Acids Res 31: 7189–7198

Keil M, Exner TE, Brickmann J (2004) Pattern recognition strategies for molecular surfaces: III. Binding site prediction with a nural network. J Comput Chem 25: 779–789

Kuznetsov I, Gou Z, Li R, Hwang S (2006) Using evolutionary and structural information to predict DNA-binding sites on DNA-binding proteins. Proteins Struc Funct Bioinformatics 64: 19–27

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860–921

Lin H, Li QZ (2007) Predicting conotoxin superfamily and family by using pseudo amino acid composition and modified Mahalanobis discriminant. Biochem Biophys Res Commun 354: 548–551

Liu DQ, Liu H, Shen HB, Yang J, Chou KC (2007a) Predicting secretory protein signal sequence cleavage sites by fusing the marks of global alignments. Amino Acids 32: 493–496

Liu H, Yang J, Liu DQ, Shen HB, Chou KC (2007b) Using a new alignment kernel function to identify secretory proteins. Protein Pept Lett 14: 203–208

Liu LX, Li ML, Tan FY, Lu MC, Wang KL, Guo YZ, Wen ZN, Jiang L (2006) Local sequence information-based support vector machine to classify voltage-gated potassium channels. Acta Biochim Biophys Sin 38: 363–371

Liu W, Chou KC (1998) Singular points of protein beta-sheets. Protein Sci 7: 2324–2330

Mondal S, Bhavna R, Mohan Babu R, Ramakumar S (2006) Pseudo amino acid composition and multi-class support vector machines approach for conotoxin superfamily classification. J Theor Biol 243: 252–260

Shanahan HP, Garcia MA, Jones S, Thornton JM (2004) Identifying DNA-binding proteins using structural motifs and the electrostatic potential. Nucleic Acids Res 32: 4732–4741

Shen HB, Chou KC (2005a) Predicting protein subnuclear location with optimized evidence-theoretic K-nearest classifier and pseudo amino acid composition. Biochem Biophys Res Comm 337: 752–756

Shen HB, Chou KC (2005b) Using optimized evidence – theoretic K-nearest neighbor classifier and pseudo amino acid composition to predict membrane protein types. Biochem Biophys Res Commun 334: 288–292

Shen HB, Chou KC (2006) Ensemble classifier for protein fold pattern recognition. Bioinformatics 22: 1717–1722

Shen HB, Chou KC (2007a) Gpos-PLoc: an ensemble classifier for predicting subcellular localization of Gram-positive bacterial proteins. Protein Eng Des Sel 20: 39–46

Shen HB, Chou KC (2007b) Hum-mPLoc: an ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites. Biochem Biophys Res Commun 355: 1006–1011

Shen HB, Chou KC (2007c) Using ensemble classifier to identify membrane protein types. Amino Acids 32: 483–488

Shen HB, Chou KC (2007d) Virus-PLoc: a fusion classifier for predicting the subcellular localization of viral proteins within host and virus-infected cells. Biopolymers 85: 233–240

Shen HB, Yang J, Chou KC (2006) Fuzzy KNN for predicting membrane protein types from pseudo amino acid composition. J Theor Biol 240: 9–13

Shen HB, Yang J, Chou KC (2007) Euk-PLoc: an ensemble classifier for large-scale eukaryotic protein subcellular location prediction. Amino Acids (doi: 10.1007/s00726-006-0478-8)

Sjöström M, Rännar S, Wieslander Å (1995) Polypeptide sequence property relationships in *Escherichia coli* based on auto cross covariances. Chemometr Intell Lab Syst 29: 295–305

Stawiski EW, Gregoret LM, Gutfreund YM (2003) annotating nucleic acid-binding function based on protein structure. J Mol Biol 326: 1065–1079

Tan F, Feng X, Fang Z, Li M, Guo Y, Jiang L (2006) Prediction of mitochondrial proteins based on genetic algorithm – partial least squares and support vector machine. Amino Acids (published online Oct 15, 2006, doi: 10.1007/s00726-006-0465-0)

Tsuchiya Y, Kinoshita K, Nakamura H (2004) Structure-based prediction of DNA-binding sites on proteins using the empirical preference of electrostatic potential and the shape of molecular surfaces. Proteins 55: 885–894

Vapnik VN (1998) Statistical learning theory. J Wiley, New York

Wang M, Yang J, Liu GP, Xu ZJ, Chou KC (2004) Weighted-support vector machines for predicting membrane protein types based on pseudo amino acid composition. Protein Eng Des Sel 17: 509–516

Wang SQ, Yang J, Chou KC (2006) Using stacked generalization to predict membrane protein types based on pseudo amino acid composition. J Theor Biol 242: 941–946

Wold S, Jonsson J, Sjöström M, Sandberg M, Rännar S (1993) DNA and peptide sequences and chemical processes multivariately modelled by principal component analysis and partial least-squares projections to latent structures. Anal Chim Acta 277: 239–253

Xiao X, Shao S, Ding Y, Huang Z, Chen X, Chou KC (2005a) An application of gene comparative image for predicting the effect on replication ratio by HBV virus gene missense mutation. J Theor Biol 235: 555–565

Xiao X, Shao S, Ding Y, Huang Z, Huang Y, Chou KC (2005b) Using complexity measure factor to predict protein subcellular location. Amino Acids 28: 57–61

Xiao X, Shao S, Ding Y, Huang Z, Chou KC (2006a) Using cellular automata images and pseudo amino acid composition to predict protein sub-cellular location. Amino Acids 30: 49–54

Xiao X, Shao SH, Huang ZD, Chou KC (2006b) Using pseudo amino acid composition to predict protein structural classes: approached with complexity measure factor. J Comput Chem 27: 478–482

Xiao X, Shao SH, Chou KC (2006c) A probability cellular automaton model for hepatitis B viral infections. Biochem Biophys Res Commun 342: 605–610

Zhang ZH, Wang ZH, Zhang ZR, Wang YX (2006) A novel method for apoptosis protein subcellular localization prediction combining encoding based on grouped weight and support vector machine. FEBS Lett 580: 6169–6174

Zhou GP (1998) An intriguing controversy over protein structural class prediction. J Protein Chem 17: 729–738

Zhou GP, Assa-Munt N (2001) Some insights into protein structural class prediction. Proteins 44: 57–59

Zhou GP, Doctor K (2003) Subcellular location prediction of apoptosis proteins. Proteins 50: 44–48

Authors' address: Menglong Li, College of Chemistry, Sichuan University, Chengdu 610064, P.R. China,
Fax: +86-28-85412356, E-mail: liml@scu.edu.cn