Amino Acids (2009) 36:167–175
DOI 10.1007/s00726-008-0044-7

ORIGINAL ARTICLE

An ensemble of reduced alphabets with protein encoding based on grouped weight for predicting DNA-binding proteins

Loris Nanni · Alessandra Lumini

Received: 30 November 2007 / Accepted: 7 February 2008 / Published online: 21 February 2008

Abstract It is well known in the literature that an ensemble of classifiers obtains good performance with respect to that obtained by a stand-alone method. Hence, it is very important to develop ensemble methods well suited for bioinformatics data. In this work, we propose to combine the feature extraction method based on grouped weight with a set of amino-acid alphabets obtained by a Genetic Algorithm. The proposed method is applied for predicting DNA-binding proteins. As classifiers, the linear support vector machine and the radial basis function support vector machine are tested. As performance indicators, the accuracy and Matthews's correlation coefficient are reported. Matthews's correlation coefficient obtained by our ensemble method is ≈ 0.97 when the jackknife cross-validation is used. This result outperforms the performance obtained in the literature using the same dataset where the features are extracted directly from the amino-acid sequence.

Keywords Multi-classifier · Amino-acid alphabets · Support vector machine · DNA-binding proteins · Ensemble classifier

Introduction

In this paper, we deal with the DNA-binding proteins (Ahmad and Sarai 2005; Hwang et al. 2007; Kuznetsov et al. 2007; Lander et al. 2001; Ofran et al. 2007; Wang and Brown 2006; Yan et al. 2006) prediction problem proposing an ensemble (Nanni and Lumini 2006) of support vector machines (Cristianini and Shawe-Taylor 2000). Our results are very interesting; the fusion outperforms the best-published results on the same dataset when the features are extracted directly from the amino-acid sequence. We present a new multi-classifier construction based on the combination between the encoding method based on grouped weight (Zhang et al. 2006c) and a set of reduced alphabets obtained by a Genetic Algorithm (Nanni and Lumini 2008a).

Notice that the encoding method based on grouped weight obtain performance lower than that obtained by the pseudo amino acid composition, but when the encoding method based on grouped weight is combined with a set of reduced alphabets, the performance drastically increases.

In the literature, it is well known that the fusion among classifiers (Chou 2000) is a feasible method for improving the performance of a stand-alone method. Several ensemble methods are published in the bioinformatics literature: protein secondary structure prediction (Riis and Krogh, 1996); protein fold pattern prediction (Shen and Chou 2006); protein subcellular localization prediction (Shen and Chou 2006, 2007a, b, c, d); membrane protein type prediction (Chou and Shen 2007c); signal peptide prediction (Chou and Shen 2007e); peptide classification (Nanni and Lumini 2006a, b). Recently, based on the approach by fusing many individual classifiers, a powerful web-server package, called Cell-PLoc, was established (Chou and Shen 2008) for predicting the subcellular localization of proteins in various different organisms.

To extract features directly from the amino-acid sequence, the most known method is the pseudo amino acid composition. A pseudo amino acid composition generator (Shen and Chou e) was established at the website http://chou.med.harvard.edu/bioinf/PseAA/, originally introduced by Chou for protein subcellular localization (Chou 2001) and for enzyme functional class (Chou 2005). In the last year, a huge number of pseudo amino acid composition approaches have been proposed (Aguero-Chapin et al. 2006;

L. Nanni (✉) · A. Lumini
DEIS, IEIIT—CNR, Università di Bologna,
Viale Risorgimento 2, 40136 Bologna, Italy
e-mail: loris.nanni@unibo.it

Caballero et al. 2007; Cai and Chou 2006; Chen et al. 2006a; Chen et al. 2006b; Chen and Li 2007a, b; Chou and Shen 2006a, b, 2007a, b, c, e; Diao et al. 2007; Ding et al. 2007; Du and Li 2006; Fang et al. 2007; Gao et al. 2005; Gonzalez-Diaz et al. 2006, 2007a, b, c; Kurgan et al. 2007; Li and Li 2007; Lin and Li 2007a, b; Liu et al. 2005a, b; Mondal et al. 2006; Mundra et al. 2007; Pan et al. 2003; Pu et al. 2007; Shen and Chou 2005a, b, 2006, 2007a, b, c, d, f, g, h; Shen et al. 2006, 2007; Shi et al. 2007; Wang et al. 2004, 2006; Xiao and Chou 2007; Xiao et al. 2006a, b; Zhang and Ding 2007; Zhang et al. 2006a, b, 2007; Zhou et al. 2007).

The encoding method based on grouped weight (Zhang et al. 2006c) is based on the division of the 20 amino acid residues into four different classes; in the original paper, this division is performed considering the hydrophobicity and charged character, and then to extract a set of features from different subsequences of each protein. To improve the method, we apply the method for building reduced alphabets proposed in Nanni and Lumini (2008b) for peptide classification. A different support vector machine (Cristianini and Shawe-Taylor 2000) is trained on each feature set (each extracted from a different alphabet). Finally, this pool of classifiers is combined by the vote rule (Duda et al. 2000).

Notice that we use the same reduced alphabets proposed in Nanni and Lumini (2008b); these reduced alphabets are obtained to maximize the area under the receiver operating characteristic curve in two problems: HIV-protease, and recognition of T-cell epitopes. In these two datasets, the ensemble of classifiers obtains an average area under the receiver operating characteristic curve of 0.984.

Notice that the proteins of the problem of this paper are not considered for obtaining the reduced alphabets.

Several methods are proposed in the literature to deal with the DNA-binding proteins: Jones et al. (2003), Shanahan et al. (2004), Keil et al. (2004), and Tsuchiya et al. 2004) propose systems based on chemical or physical properties of amino acids; Ahmad and Sarai (2005), and Ahmad et al. (2004) propose systems based on neural network classifiers; Kuznetsov et al. (2006) propose a system based on structure information and sequence alignment profiles; Bhardwaj et al. (2005) propose a system based on electrostatic potentials and the amino acid composition of the protein; Fang et al. (2007) propose a system based on the Chou's pseudo-amino acid composition; and Nanni and Lumini (2008b) propose to combine the dipeptide composition of the protein with the features extracted from the gene ontology database (Chou and Shen 2007a, d).

Materials and methods

We have used exactly the same datasets used in Fang et al. (2007). This dataset contains 118 DNA-binding proteins and 231 non-DNA-binding proteins.[1] These proteins have less than 35% sequence identity between each pairs.

The proposed method combined 11 classifiers where each classifier is trained using the encoding method based on grouped weight combined with a different reduced alphabet. Finally, these classifiers are combined by vote rule (Kittler et al. 1998). Notice that our ensemble is built by 11 classifiers since we have used $K = 5$ (as in Nanni and Lumini 2008b) different alphabets for each value of the size $S = 4$ of the reduced alphabets and for a given value of N ($N = 1$ or $N = 2$). Moreover, the baseline implementation based on the alphabets of Zhang et al. (2006c) also belongs to the ensemble.

In this work, an alternative way for the construction of reduced alphabets is applied; it is based on a Genetic Algorithm for grouping amino-acids (Nanni and Lumini 2008b). The objective function of the Genetic Algorithm is the maximization of the area under the receiver operating characteristic curve (Fawcett 2004) for a given classification problem using the N-gram feature extraction. K ($K = 5$) different alphabets are created for each value of the size S of the reduced alphabets and for a given value of N. The ith reduced alphabet is built considering the previous reduced alphabets of the same size S and of the same value of N. Simply, for the calculation of the objective function of the ith iteration of GA, the scores obtained by the ith reduced alphabet are combined by the mean rule with the scores obtained by the previous $i-1$ reduced alphabets. The mean rule (Eq. 1) selects as final score $[score(\mathbf{s},c)]$ the mean score of a pool of K classifiers:

$$score(\mathbf{s}, c) = \frac{1}{K} \sum_{j=1...K} \mathrm{sim}_j(\mathbf{s}, c) \quad (1)$$

Where: $\mathrm{sim}_j(\mathbf{s},c)$ is the similarity of the pattern (protein) $\mathbf{s}$ to the class c, obtained by the jth classifier; K is the number of combined classifier.

In this work, we use the ten alphabets obtained with: $S = 4$, $N = 1$; $S = 4$, $N = 2$. The area under the receiver operating characteristic curve, as in Nanni and Lumini (2008b), is maximized for two problems of peptide classification: HIV-protease; and recognition of T-cell epitopes.

The reduce alphabets are:

C1 = 'GPSV'; C2 = 'ACDFNW'; C3 = 'EKR';
C4 = 'HILMQTY'
C1 = 'QRSW'; C2 = 'AFHIKLMPV'; C3 = 'DENTY';
C4 = 'CG'
C1 = 'MNRTW'; C2 = 'ACHILPV'; C3 = 'DEFQ';
C4 = 'GKSY'
C1 = [ ]; C2 = 'FHMNRSVWY'; C3 = 'ACEGIKLP';
C4 = 'DQT'

[1] 107 DNA-binding proteins and 196 non-DNA-binding proteins have a hit in the GO database.

C1 = 'CDES'; C2 = 'AGHQTY'; C3 = 'FKVW';
C4 = 'ILMNPR'
C1 = 'HKR'; C2 = 'AGIMNQVWY';
C3 = 'CDEFPST'; C4 = 'L'
C1 = 'CMPS'; C2 = 'ADEFGHKLTVY'; C3 = 'IQ';
C4 = 'NRW'
C1 = 'AIPRSV'; C2 = 'CGHLNQW'; C3 = 'DEY';
C4 = 'FKMT'
C1 = 'FGHL'; C2 = 'ADKNPW';
C3 = 'CEIMQRSTVY'; C4 = '[ ]'
C1 = 'AKMWY'; C2 = 'FHILRST'; C3 = 'CDNQ';
C4 = 'EGPV'

Now, we describe the method used in this paper for the encoding method based on grouped weight; the original paper divided the 20 amino acid residues into four different classes (Zhang et al. 2006c): C1 = 'GAVLIMPFW'; C2 = 'QNSTYC'; C3 = 'DE'; C4 = 'HKR'. Since the amino acid residues are divided into four different classes, we can partition the amino acid residues into the following disjoint groups: C1 + C2 versus C3 + C4, or C1 + C3 versus C2 + C4, and C1 + C4 versus C2 + C3.

Given a protein **p** we calculate three binary sequences (Eq. 2):

$$\begin{aligned} \mathbf{H1}_{\mathrm{p}}(j) &= 1 \text{ if } \mathbf{p}(j) \in \mathrm{C1} + \mathrm{C2} \text{ else } \mathbf{p}(j) = 0 \\ \mathbf{H2}_{\mathrm{p}}(j) &= 1 \text{ if } \mathbf{p}(j) \in \mathrm{C1} + \mathrm{C3} \text{ else } \mathbf{p}(j) = 0 \\ \mathbf{H3}_{\mathrm{p}}(j) &= 1 \text{ if } \mathbf{p}(j) \in \mathrm{C1} + \mathrm{C4} \text{ else } \mathbf{p}(j) = 0 \end{aligned} \quad (2)$$

We partition each binary sequence into L pieces of subsequence. In our implementation, the jth feature ($j = 1, \ldots, L$) is given by $\mathrm{sum}(\mathbf{H1}_{\mathrm{p}}(1{:}j \times S) = 1)/(j \times S)$, where S = length of the protein/L and the function sum(**x**) gives the number of 1 in the vector **x**. We create L features for **H1**, **H2**, **H3** and then we concatenate these three vectors.

For building the ensemble, we simply use the values of C1, C2, C3, and C4 obtained by the Genetic Algorithm.

Notice that to avoid any doubt in the implementation of the feature extraction method, we report the Matlab code in the "Appendix".

Results and discussion

As performance indicators, we use Matthew's correlation coefficient (MCC) (Fang et al. 2007) and the accuracy (see Eq. 3).

$$\mathrm{MCC}(i) = \frac{\mathrm{TP}(i) \times \mathrm{TN}(i) - \mathrm{FP}(i) \times \mathrm{FN}(i)}{\sqrt{(\mathrm{TP}(i) + \mathrm{FP}(i)) \times (\mathrm{TP}(i) + \mathrm{FN}(i)) \times (\mathrm{TN}(i) + \mathrm{FN}(i)) \times (\mathrm{TN}(i) + \mathrm{FP}(i))}} \quad (3)$$

where TP(i) is the number of correctly predicted DNA-binding proteins (true positives); TN(i), FP(i) and FN(i) are the numbers of true negatives, false positives, and false negatives, respectively.

As testing protocol, the rigorous jackknife test (Feng 2002, Chou and Zhang 1995) is used since it is demonstrated (Chou and Shen 2007d) that it is the best method for comparing different classifiers in a given dataset. For this reason, the jackknife test is widely adopted in the literature (Chen et al. 2006a, b, 2007; Chou and Shen 2006a; Diao et al. 2008; Ding et al. 2007; Du and Li 2006; Fang et al. 2007; Gao et al. 2005; Guo et al. 2006; Kedarisetti et al. 2006; Li and Li 2007; Lin and Li 2007a, b; Liu et al. 2007; Mondal et al. 2006; Niu et al. 2006; Shen and Chou 2007g; Shen et al. 2007; Shi et al. 2007; Sun and Huang 2006; Tan et al. 2007; Wang et al. 2005; Wen et al. 2006; Xiao and Chou 2007; Xiao et al. 2005a, b, 2006a; Zhang and Ding 2007; Zhang et al. 2006a, 2007; Zhou 1998; Zhou and Doctor 2003; Zhou et al. 2007).

We have used three different classifiers: linear support vector machine (LSVM); radial basis function support vector machine with C = 1 and Gamma = 1 (RSVM1); radial basis function support vector machine with C = 0.1 and Gamma = 1,000 (RSVM2). In Fig. 1, we report the performance of these three classifiers, using only the baseline encoding method based on grouped weight method, varying the value of L. In Fig. 1, we also report the performance of our ensemble (ENS) of radial basis function support vector machine[2] (with cost of the constrain violation = 0.1 and parameters of the radial based kernel = 1,000) classifiers. It is clear that the proposed method outperforms the stand-alone methods.

In Table 1, we report the performance of each stand-alone classifiers that build the ensemble (L = 10).

In Table 2, we report the MCC of other state-of-the-art methods, from Nanni and Lumini (2008a) and Fang et al. (2007), for predicting DNA-binding proteins.

From the analysis of the experimental results, the following observations may be made:

1. The encoding method based on grouped weight permits to obtain a reliable method.
2. The proposed ensemble permits to improve the performance of the baseline encoding method based on grouped weight; the proposed ensemble of support vector machines obtains a Matthew's correlation

[2] Implemented as in the OSU svm matlab toolbox

coefficient of 0.968 when it is trained by 11 different set of reduced alphabets.
3. The tested fusion permits to outperform several other state-of-the-art methods.
4. Our results show that Chou's pseudo amino-acid composition outperforms the encoding method based on grouped weight, but we show that the performance of the encoding method based on grouped weight increases if our method to build an ensemble of classifiers is used.

Conclusions

We investigated how to build an ensemble of classifiers for the encoding method based on grouped weight. For building the ensemble, we used the same reduced alphabets proposed in Nanni and Lumini (2008b); these reduced alphabets are obtained to maximize the area under the receiver operating characteristic curve in two problems: HIV-protease, and recognition of T-cell epitopes.

The proposed ensemble of 11 support vector machines obtains a Matthew's correlation coefficient of 0.968; moreover, the validity of the novel approach is proved by the performance improvements obtained with respect to other state-of-the-art methods in the tested problem.

As a future study, we want to combine methods trained using different feature extraction methods, e.g., we can combine an ensemble builds using the onthology-based features, an ensemble based on the 2-gram composition, an ensemble based on the Chou's pseudo amino-acid composition, and the ensemble proposed in this work.

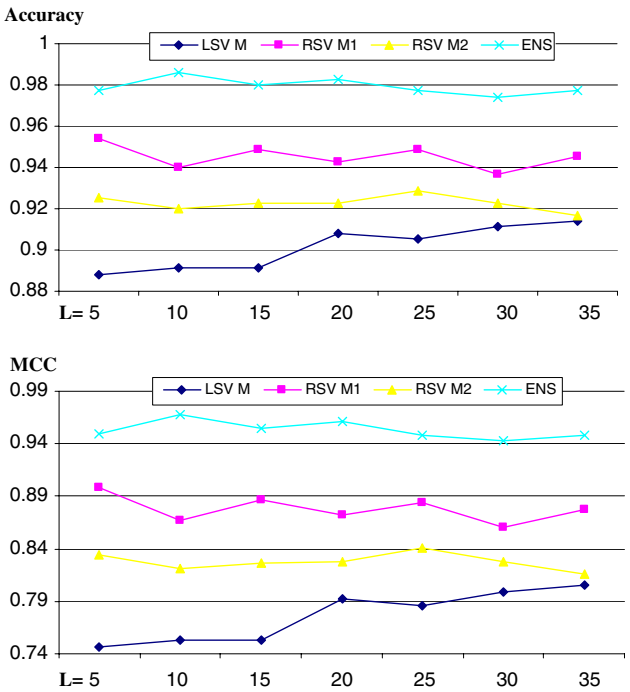

Fig. 1 Accuracy and MCC of three stand-alone methods and of the proposed ensemble

Table 1 Performance of the classifiers that build the ensemble

Method	Accuracy	MCC
Baseline[a]	0.939	0.8674
First reduced set	0.954	0.8972
Second reduced set	0.959	0.9101
Third reduced set	0.942	0.8733
Fourth reduced set	0.942	0.872
Fifth reduced set	0.868	0.7096
Sixth reduced set	0.937	0.8592
Seventh reduced set	0.948	0.886
Eighth reduced set	0.931	0.8512
Ninth reduced set	0.934	0.8568
Tenth reduced set	0.945	0.8793

[a] The set used in Zhang et al. (2006c)

Table 2 Performance, in the same dataset, of other state-of-the-art methods for predicting DNA-binding proteins

Features	Classifier	MCC
Onthology-based	Stand-alone LSVM	0.928
Onthology-based	Random subspace LSVM	0.935
2-gram	Stand-alone LSVM	0.848
2-gram	Random subspace	0.942
Chou's pseudo amino-acid composition		0.924
Proposed ensemble		0.968

Acknowledgments The author would like to thank Y. Fang for sharing the dataset.

Appendix: Matlab code

The following function implements the base feature extraction method as detailed in "Materials and methods":

```
function F=featuresEBGW(ALFABETO,P,L)
C1=ALFABETO{1};
C2=ALFABETO{2};
C3=ALFABETO{3};
C4=ALFABETO{4};

uno=strcat(C1,C2);
due=strcat(C1,C3);
tre=strcat(C1,C4);
t=1;

H=[];
for j=1:length(P)
   if sum(P(j)==uno)
      H(j)=1;
   else
      H(j)=0;
   end
end
S=length(P)/L;
for j=1:L
   F(t)=sum(H( 1:j*S )==1)/(j*S);
   t=t+1;
end

H=[];
for j=1:length(P)
   if sum(P(j)==due)

      H(j)=1;
   else
      H(j)=0;
   end
end
for j=1:L
   F(t)=sum(H( 1:j*S )==1)/(j*S);
   t=t+1;
end

H=[];
for j=1:length(P)
   if sum(P(j)==tre)
      H(j)=1;
   else
      H(j)=0;
   end
end
for j=1:L
   F(t)=sum(H( 1:j*S )==1)/(j*S);
   t=t+1;
end
```

The following rows implements the main of the system:

```
ALFABETO{1}='GAVLIMPFW';
ALFABETO{2}='QNSTYC';
ALFABETO{3}='DE';
ALFABETO{4}='HKR';

load c:\AlfabetiUniti.mat Alfabeto
%Alfabeto stores the reduced sets obtained in (Nanni and Lumini, 2008a)

for NK=1:11
        if NK==1
           ALF=ALFABETO;
        elseif NK>1 & NK<7
           ALF=Alfabeto{3}{NK-1};
        else
           ALF=Alfabeto{4}{NK-6};
        end


        F=[];
        for i=1:size(SEQ,2)
           F(i,:)=featuresEBGW(ALF,SEQ{i},L);
        end


        res=F;
        for op=1:size(SEQ,2)
           TR=res;yTR=y;TR(op,:)=[];yTR(op)=[];
           TE=res(op,:);
           massimo=max(TR)+0.00001;
           minimo=min(TR);
           training=[];
           testing=[];
           training=TR;
           testing=TE;
           for i=1:size(TR,2)
              training(1:size(TR,1),i)=double(TR(1:size(TR,1),i)-minimo(i))/(massimo(i)-minimo(i));
           end
           for i=1:size(TE,2)
              testing(1:size(TE,1),i)=double(TE(1:size(TE,1),i)-minimo(i))/(massimo(i)-minimo(i));
           end
           tra=[];

           TR=training;
           TE=testing;
           [AlphaY, SVs, Bias, Parameters, nSV, nLabel] = rbfSVC(double(TR)', double(yTR),0.1,1000);
           [ER, Decision(op), Ns, ConfMatrix, lab(op)]= SVMTest(double(TE)', double(y(op)), AlphaY, SVs, Bias, Parameters, nSV,
nLabel);
        end
        VALUE(1)=sum(lab==y);
        %calcoloM MCC
        for i=1:3
           VALUE(i+1)=CalcoloMCC(lab,y,i)
        end
        labP(tt,:)=lab;
        tt=tt+1
     end
end
% vote rule
Q=[];
for i=1:size(SEQ,2)
   for j=1:max(y)
      Q(i,j)=sum(labP(:,i)==j);
   end
end
[a,b]=max(Q');
```

References

Aguero-Chapin G, Gonzalez-Diaz H, Molina R, Varona-Santos J, Uriarte E, Gonzalez-Diaz Y (2006) Novel 2D maps and coupling numbers for protein sequences. The first QSAR study of polygalacturonases; isolation and prediction of a novel sequence from *Psidium guajava* L. FEBS Lett 580:723–730

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

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

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

Caballero J, Fernandez L, Garriga M, Abreu JI, Collina S, Fernandez M (2007) Proteometric study of ghrelin receptor function variations upon mutations using amino acid sequence autocorrelation vectors and genetic algorithm-based least square support vector machines. J Mol Graph Model 26:166–178

Cai YD, Chou KC (2006) Predicting membrane protein type by functional domain composition and pseudo amino acid composition. J Theor Biol 238:395–400
Chen YL, Li QZ (2007a) Prediction of apoptosis protein subcellular location using improved hybrid approach and pseudo amino acid composition. J Theor Biol 248:377–381
Chen YL, Li QZ (2007b) Prediction of the subcellular location of apoptosis proteins. J Theor Biol 245:775–783
Chen C, Tian YX, Zou XY, Cai PX, Mo JY (2006a) Using pseudo-amino acid composition and support vector machine to predict protein structural class. J Theor Biol 243:444–448
Chen C, Zhou X, Tian Y, Zou X, Cai P (2006b) Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network. Anal Biochem 357:116–121
Chen J, Liu H, Yang J, Chou KC (2007) Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. Amino Acids 33:423–428
Chou KC (2000) Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. Biochem Biophys Res Commun 278:477–483
Chou KC (2001) Prediction of protein cellular attributes using pseudo amino acid composition. Proteins 43:246–255 (Erratum: ibid, 2001, 44:60)
Chou KC (2005) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21:10–19
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 (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) MemType-2L: a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. Biochem Biophys Res Commun 360:339–345
Chou KC, Shen HB (2007d) Review: recent progresses in protein subcellular location prediction. Anal Biochem 370:1–16
Chou KC, Shen HB (2007e) Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides. Biochem Biophys Res Commun 357:633–640
Chou KC, Shen HB (2008) Cell-PLoc: a package of web-servers for predicting subcellular localization of proteins in various organisms. Nat Protoc 3:153–162
Chou KC, Zhang CT (1995) Review: prediction of protein structural classes. Crit Rev Biochem Mol Biol 30:275–349
Cristianini N, Shawe-Taylor J (2000) An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press, London
Diao Y, Li M, Feng Z, Yin J, Pan Y (2007) The community structure of human cellular signaling network. J Theor Biol 247:608–615
Diao Y, Ma D, Wen Z, Yin J, Xiang J, Li M (2008) Using pseudo amino acid composition to predict transmembrane regions in protein: cellular automata and Lempel–Ziv complexity. Amino Acids 34(1):111–117. doi:10.1007/s00726-007-0550-z
Ding YS, Zhang TL, Chou KC (2007) Prediction of protein structure classes with pseudo amino acid composition and fuzzy support vector machine network. Protein Pept Lett 14:811–815
Du P, Li Y (2006) Prediction of protein submitochondria locations by hybridizing pseudo-amino acid composition with various physicochemical features of segmented sequence. BMC Bioinformatics 7:518
Duda RO, Hart PE, Stork G (2000) Pattern classification, 2nd edn. Wiley, New York
Fang Y, Guo Y, Feng Y, Li M (2008) Predicting DNA-binding proteins: approached from Chou's pseudo amino acid composition and other specific sequence features. Amino Acids 34(1):103–109. doi:10.1007/s00726-007-0568-2
Fawcett T (2004) ROC graphs: notes and practical considerations for researchers. HP Laboratories, technical report, Palo Alto
Feng ZP (2002) An overview on predicting the subcellular location of a protein. In Silico Biol 2:291–303
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
Gonzalez-Diaz H, Perez-Bello A, Uriarte E, Gonzalez-Diaz Y (2006) QSAR study for mycobacterial promoters with low sequence homology. Bioorg Med Chem Lett 16:547–553
Gonzalez-Diaz H, Aguero-Chapin G, Varona J, Molina R, Delogu G, Santana L, Uriarte E, Podda G (2007a) 2D-RNA-coupling numbers: a new computational chemistry approach to link secondary structure topology with biological function. J Comput Chem 28:1049–1056
Gonzalez-Diaz H, Perez-Castillo Y, Podda G, Uriarte E (2007b) Computational chemistry comparison of stable/nonstable protein mutants classification models based on 3D and topological indices. J Comput Chem 28:1990–1995
Gonzalez-Diaz H, Vilar S, Santana L, Uriarte E (2007c) Medicinal chemistry and Bioinformatics—current trends in drugs discovery with networks topological indices. Curr Top Med Chem 10:1015–1029
Guo YZ, Li M, Lu M, Wen Z, Wang K, Li G, Wu J (2006) Classifying G protein-coupled receptors and nuclear receptors based on protein power spectrum from fast Fourier transform. Amino Acids 30:397–402
Hwang S, Gou Z, Kuznetsov IB (2007) DP-Bind: a web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins. Bioinformatics 23(5):634–636
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
Kedarisetti KD, Kurgan LA, Dick S (2006) Classifier ensembles for protein structural class prediction with varying homology. Biochem Biophys Res Commun 348:981–988
Keil M, Exner TE, Brickmann J (2004) Pattern recognition strategies for molecular surfaces: III. Binding site prediction with a neural network. J Comput Chem 25:779–789
Kittler J, Hatef M, Duin R, Matas J (1998) On combining classifiers. IEEE Trans Pattern Anal Mach Intell 3:226–239
Kurgan LA, Stach W, Ruan J (2007) Novel scales based on hydrophobicity indices for secondary protein structure. J Theor Biol 248:354–366
Kuznetsov IB, Gou Z, Li R, Hwang S (2006) Using evolutionary and structural information to predict DNA-binding sites on DNA-binding proteins. Proteins 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
Li FM, Li QZ (2007) Using pseudo amino acid composition to predict protein subnuclear location with improved hybrid approach. Amino Acids 34:119–125
Lin H, Li QZ (2007a) Predicting conotoxin superfamily and family by using pseudo amino acid composition and modified Mahalanobis discriminant. Biochem Biophys Res Commun 354:548–551
Lin H, Li QZ (2007b) Using pseudo amino acid composition to predict protein structural class: approached by incorporating 400 dipeptide components. J Comput Chem 28:1463–1466

Liu H, Wang M, Chou KC (2005a) Low-frequency Fourier spectrum for predicting membrane protein types. Biochem Biophys Res Commun 336:737–739

Liu H, Yang J, Wang M, Xue L, Chou KC (2005b) Using Fourier spectrum analysis and pseudo amino acid composition for prediction of membrane protein types. Protein J 24:385–389

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

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

Mundra P, Kumar M, Kumar KK, Jayaraman VK, Kulkarni BD (2007) Using pseudo amino acid composition to predict protein subnuclear localization: approached with PSSM. Pattern Recognit Lett 28:1610–1615

Nanni L, Lumini A (2006a) An ensemble of K-local hyperplane for predicting protein–protein interactions. Bioinformatics 22:1207–1210

Nanni L, Lumini A (2006b) MppS: an ensemble of support vector machine based on multiple physicochemical properties of amino-acids. Neurocomputing 69:1688–1690

Nanni L, Lumini A (2008a) A genetic approach for building different alphabets for peptide and protein classification. BMC Bioinformatics 9:45

Nanni L, Lumini A (2008b) Combing ontologies and dipeptide composition for predicting DNA-binding proteins. Amino Acids. doi:10.1007/s00726-007-0016-3

Niu B, Cai YD, Lu WC, Zheng GY, Chou KC (2006) Predicting protein structural class with AdaBoost learner. Protein Pept Lett 13:489–492

Ofran Y, Mysore V, Rost B (2007) Prediction of DNA-binding residues from sequence. Bioinformatics 23(13):347–353

Pan YX, Zhang ZZ, Guo ZM, Feng GY, Huang ZD, He L (2003) Application of pseudo amino acid composition for predicting protein subcellular location: stochastic signal processing approach. J Protein Chem 22:395–402

Pu X, Guo J, Leung H, Lin Y (2007) Prediction of membrane protein types from sequences and position-specific scoring matrices. J Theor Biol 247:259–265

Riis SK, Krogh A (1996) Improving prediction of protein secondary structure using neural networks and multiple sequence alignments. J Comput Biol 3:163–183

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 Commun 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) EzyPred: a top-down approach for predicting enzyme functional classes and subclasses. Biochem Biophys Res Commun 364:53–59

Shen HB, Chou KC (2007b) 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 (2007c) 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 (2007d) Nuc-PLoc: a new web-server for predicting protein subnuclear localization by fusing PseAA composition and PsePSSM. Protein Eng Des Sel 20:561–567

Shen HB, Chou KC (2007e) PseAAC: a flexible web-server for generating various kinds of protein pseudo amino acid composition. Anal Biochem 373:386–388

Shen HB, Chou KC (2007f) Signal-3L: a 3-layer approach for predicting signal peptide. Biochem Biophys Res Commun 363:297–303

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

Shen HB, Chou KC (2007h) 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 33:57–67

Shi JY, Zhang SW, Pan Q, Cheng Y-M, Xie J (2007) Prediction of protein subcellular localization by support vector machines using multi-scale energy and pseudo amino acid composition. Amino Acids 33:69–74

Sun XD, Huang RB (2006) Prediction of protein structural classes using support vector machines. Amino Acids 30:469–475

Tan F, Feng X, Fang Z, Li M, Guo Y, Jiang L (2007) Prediction of mitochondrial proteins based on genetic algorithm—partial least squares and support vector machine. Amino Acids 33:669–675

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

Yan C, Terribilini M, Wu F, Jernigan RL, Dobbs D, Honavar V (2006) Predicting DNA-binding sites of proteins from amino acid sequence. BMC Bioinformatics 7:262

Wang L, Brown SJ (2006) Prediction of DNA-binding residues from sequence features. J Bioinform Comput Biol 4(6):1141–1158

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 M, Yang J, Chou KC (2005) Using string kernel to predict signal peptide cleavage site based on subsite coupling model. Amino Acids 28:395–402 (Erratum: ibid, 2005, 29:301)

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

Wen Z, Li M, Li Y, Guo Y, Wang K (2006) Delaunay triangulation with partial least squares projection to latent structures: a model for G-protein coupled receptors classification and fast structure recognition. Amino Acids 32:277–283

Xiao X, Chou KC (2007) Digital coding of amino acids based on hydrophobic index. Protein Pept Lett 14:871–875

Xiao X, Shao S, Ding Y, Huang Z, Chen X, Chou KC (2005a) Using cellular automata to generate image representation for biological sequences. Amino Acids 28:29–35

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 SH, Ding YS, Huang ZD, Chou KC (2006a) Using cellular automata images and pseudo amino acid composition to predict protein subcellular 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

Zhang TL, Ding YS (2007) Using pseudo amino acid composition and binary-tree support vector machines to predict protein structural classes. Amino Acids 33(4):623–629. doi: 10.1007/s00726-007-0496-1

Zhang SW, Pan Q, Zhang HC, Shao ZC, Shi JY (2006a) Prediction protein homo-oligomer types by pseudo amino acid composition: approached with an improved feature extraction and naive Bayes feature fusion. Amino Acids 30:461–468

Zhang T, Ding Y, Chou KC (2006b) Prediction of protein subcellular location using hydrophobic patterns of amino acid sequence. Comput Biol Chem 30:367–371

Zhang Z-H, Wang Z-H, Zhang Z-R, Wang Y-X (2006c) A novel method for apoptosis protein subcellular localization prediction combining encoding based on grouped weight and support vector machine. FEBS Lett 580:6169–6174

Zhang SW, Zhang YL, Yang HF, Zhao CH, Pan Q (2007) Using the concept of Chou's pseudo amino acid composition to predict protein subcellular localization: an approach by incorporating evolutionary information and von Neumann entropies. Amino Acids. doi:10.1007/s00726-007-0010-9

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

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

Zhou XB, Chen C, Li ZC, Zou XY (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