Amino Acids (2008) 34:635–641
DOI 10.1007/s00726-007-0016-3

ORIGINAL ARTICLE

Combing ontologies and dipeptide composition for predicting DNA-binding proteins

Loris Nanni · Alessandra Lumini

Received: 18 November 2007 / Accepted: 6 December 2007 / Published online: 4 January 2008

Abstract Given a novel protein it is very important to know if it is a DNA-binding protein, because DNA-binding proteins participate in the fundamental role to regulate gene expression. In this work, we propose a parallel fusion between a classifier trained using the features extracted from the gene ontology database and a classifier trained using the dipeptide composition of the protein. As classifiers the support vector machine (SVM) and the 1-nearest neighbour are used. Matthews's correlation coefficient obtained by our fusion method is ≈ 0.97 when the jackknife cross-validation is used; this result outperforms the best performance obtained in the literature (0.924) using the same dataset where the SVM is trained using only the Chou's pseudo amino acid based features. In this work also the area under the ROC-curve (AUC) is reported and our results show that the fusion permits to obtain a very interesting 0.995 AUC. In particular we want to stress that our fusion obtains a 5% false negative with a 0% of false positive. Matthews's correlation coefficient obtained using the single best GO-number is only 0.7211 and hence it is not possible to use the gene ontology database as a simple lookup table. Finally, we test the complementarity of the two tested feature extraction methods using the Q-statistic. We obtain the very interesting result of 0.58, which means that the features extracted from the gene ontology database and the features extracted from the amino acid sequence are partially independent and that their parallel fusion should be studied more.

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

Keywords DNA-binding proteins · Gene ontology · Dipeptide composition · Chou's pseudo amino acid composition · Multi-classifier

Introduction

DNA-binding proteins participate in a fundamental role in the regulation of gene expression, since the number of proteins sequence data increase rapidly it is important to develop an automatic method to identify DNA-binding proteins (Lander et al. 2001).

Several methods are yet proposed in the literature, some of them are based on chemical or physical properties of amino acids (Jones et al. 2003; Shanahan et al. 2004; Keil et al. 2004; Tsuchiya et al. 2004). More recent methods are based on machine learning classifiers (e.g. neural networks). The authors of Ahmad et al. (2004) and Ahmad and Sarai (2005) propose a method based on neural network classifiers where the features are based on sequence neighbour, structure information and sequence alignment profiles. The authors of Kuznetsov et al. (2006) propose a method based on evolutionary and structural information of proteins. The authors of Bhardwaj et al. (2005) extract the features from the electrostatic potentials and from the amino acid composition of the protein.

In Fang et al. (2007), several methods based only on the primary sequences are compared, each feature vector is used to train a given support vector machine. The results reported in that paper indicate that Chou's pseudo-amino acid composition obtains the best results (Matthews's correlation coefficient = 0.924).

In the literature several methods are proposed to extract a feature vector from the primary sequence of a protein. The main part of these methods are proposed for the

sub-cellular location prediction (Shen and Chou 2006, 2007a, b, c, d).

Another interesting method to extract the features from a given protein it is to use the Gene Ontology (GO) Database (Chou and Shen 2007d). Each protein in the GO Database is described by several GO numbers, each GO number reflects the "biological reality" of a particular protein. In the GO Database there are several thousands of GO numbers, a protein can be represented by a vector where to the *i*th component is assigned 1 if to that protein is assigned the *i*th GO number. In Chou and Shen (2007a) this feature extraction method is coupled with an ensemble of nearest neighbour based classifiers. The ensemble of classifier was built by fusing many single classifiers where each classifier has a different value for the number K of considered neighbours.

The GO-based features are widely studied by the research group of Professor Chou, in whose works the GO-based features are combined ONLY in a serial way with the features extracted from the primary sequence of a protein. For a systematical description how to hybridize the ab-initio sequence-based approach (e.g. Chou's pseudo amino acid composition) and the higher-level GO (gene ontology) approach, the reader is referred to a recent review (Chou and Shen 2007d) and the references therein. Simply, if a protein has a hit in the GO database then the GO-based features are used else the pseudo amino acid composition is used. The pseudo amino acid composition was originally introduced by Chou to improve the prediction quality for protein subcellular localization and membrane protein type (Chou 2001), as well as for enzyme functional class (Chou 2005). It can be used to represent a protein sequence with a discrete model yet without completely losing its sequence-order information. Since the concept of Chou's pseudo amino acid composition was introduced, various pseudo amino acid composition approaches have been stimulated to deal with varieties of problems in proteins and protein-related systems (Aguero-Chapin et al. 2006; Caballero et al. 2007; Cai and Chou 2006; Chen et al. 2006a, b; Chen and Li 2007a, b; Chou and Shen 2006a, b, 2007a, b, c, e; Diao et al. 2007a; 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 et al. 2006a, b, 2007; Zhang and Ding 2007; Zhou et al. 2007). Because of its wide usage, recently a very flexible pseudo amino acid composition generator, called "PseAAC" (Shen and Chou 2007e), was established at the website http://chou.med.harvard.edu/bioinf/PseAA/, by which users can generate 63 different kinds of pseudo amino acid composition. In this work we show that the parallel fusion is very useful.

The good performance, with respect to that obtained by a stand-alone method, of the ensemble of classifiers (Chou 2000a, b) are well known, several examples are published in the bioinformatics literature. Several ensemble methods are applied on 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), and signal peptide prediction (Chou and Shen 2007e).

Particularly interesting are Nanni and Lumini (2006a, b), where multiple physicochemical properties of amino-acids are used to train a given classifier. The selection of the best physicochemical properties to be combined is performed by sequential forward floating selection (Nanni and Lumini 2006a).

In this paper, we deal with the DNA-binding protein prediction problem proposing a parallel fusion between a classifier trained using the features extracted from the GO database and a classifier trained using the dipeptide composition of the protein (Nanni and Lumini 2006a). Our results are very interesting; the fusion outperforms both the stand-alone methods and the best published results on the same dataset. We test two different classifiers: support vector machine (Cristianini and Shawe-Taylor 2000); 1-nearest neighbour (Duda et al. 2000).

The proposed fusion obtains an interesting result of ≈ 0.97 Matthews's correlation coefficient when the jack-knife cross-validation is used, while the best performance obtained in the literature is 0.924.

Finally, we test the complementarity of the two tested feature extraction methods by the Q-statistic (Kuncheva and Whitaker 2003). We obtain the very interesting result of 0.58, which means that the features extracted from the GO database and the features extracted from the amino acid sequence are partially independent.

Materials and methods

In this paper, we propose a system for dealing with the DNA-binding protein prediction problem (see Fig. 1) combining two classifiers: the first classifier is trained using the features extracted using the GO database; the latter classifier is trained using the well known 2-gram (i.e. the dipeptide composition).

Finally, these two classifiers are combined by sum rule (Kittler et al. 1998). The sum rule selects as final score the sum of the scores of the two classifiers.

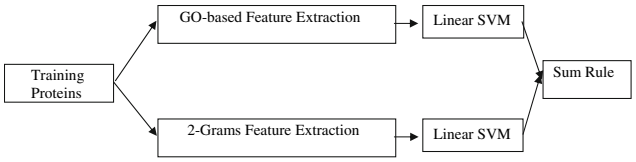

Fig. 1 Global schema of our system

As base classifier we have tested the support vector machine (the state-of-the-art of the machine learning approaches; Cristianini and Shawe-Taylor 2000) and the 1-nearest neighbour. Moreover, for each classifier we have tested also the random subspace version. The random subspace method (Ho 1998) modifies the training data set generating K new training sets ($K = 50$ in this paper), builds classifiers on these modified training sets, and then combines them into a final decision rule (sum rule in this paper). The new training sets contain only a subset (50% in this paper) of the all features.

We obtain the best results using as classifier the linear version of SVM[1] (LSVM).

2-Grams

Each 2-gram is a couple of values (v_i, c_i), where v_i is the feature and c_i is the counts of this feature in a protein sequence. The features v_i are all the possible combinations of 2 amino-acids (Nanni and Lumini 2006). An example is shown in Fig. 2. Since there are 20 amino-acids, we have 400 2-grams for each protein. The feature vector extracted from the 2-gram i is given by dividing c_i with the length of the protein. Notice that this normalization is very important since the proteins can have a very different length.

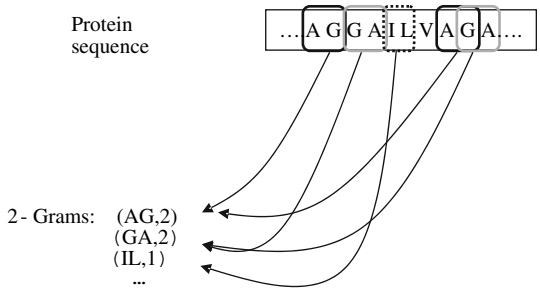

Fig. 2 An example of 2-grams (from Nanni and Lumini 2006)

Gene ontology

To extract the feature vector **P** from a protein using the Gene Ontology we need to search the protein in the GO database, then if the ith GO number is assigned to that protein to the ith component, of the feature vector that describes that protein, is assigned the value 1 (i.e. $\mathbf{P}_i = 1$) otherwise 0.

Thus, for each protein it is possible to extract a binary vector. Given a protein which biological activities are sulfotransferase activity (GO number 0008146) and transferase activity (GO number 0016740), the feature vector **P** that describes that protein has value 0 for each component except $\mathbf{P}_{8146}$ and $\mathbf{P}_{16740}$ that have value 1.

Notice that some GO-numbers do not exist and these features (that are 0 for all the proteins) are not considered in the feature extraction step (see Chou and Shen 2007d) for details).[2]

Results and discussion

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.[3] These proteins have less than 35% sequence identity between each pairs.

As performance indicator we use the Matthew's correlation coefficient (MCC) (Fang et al. 2007) and the area under the ROC-curve (Fawcett 2004) (AUC). The ROC-curve is a two-dimensional measure of classification performance that plots the probability of classifying correctly the positive examples against the rate of incorrectly classifying negative examples.

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

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.

We use the objective and rigorous leave-one-out cross validation method (Feng 2002) as testing protocol. In statistical prediction, the sub-sampling test and jackknife test are two cross-validation methods often used in literatures

[1] The OSU svm Matlab toolbox was used, the parameter C of Linear SVM is set to the default value 1.

[2] In this dataset we have 336 GO-numbers with some hits.

[3] 107 DNA-binding proteins and 196 Non-DNA-binding proteins have a hit in the GO database.

for examining the accuracy of a predictor (Chou and Zhang 1995). However, as demonstrated by Eq. (50) in a recent comprehensive review (Chou and Shen 2007d), the sub-sampling (e.g. fivefold cross-validation) test cannot avoid arbitrariness even for a very simple benchmark dataset. Accordingly, the jackknife test has been increasingly and widely adopted by investigators (Chen et al. 2006a, b, 2007; Chou and Shen 2006a; Diao et al. 2007b; 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 et al. 2006a, 2007; Zhang and Ding 2007; Zhou 1998; Zhou and Doctor 2003; Zhou et al. 2007) to test the power of various predictors.

In Tables 1, 2 we report the AUC and the MCC obtained by the stand-alone classifiers and by them the random subspace version.

In Table 3 we report the performance obtained combining the two stand-alone linear support vector machine and the two random subspace of linear support vector machine. In both the methods the first classifier is trained by the GO-based features, while the latter classifier is trained by the 2-gram features. Moreover, it is interesting to note that a random subspace of linear support vector machine based on the 2-gram features outperforms the support vector machine trained using the Chou's pseudo amino-acid composition.

Table 1 Performance obtained using the GO-based features

Onthology	MCC	AUC
Stand-alone		
LSVM	0.928	0.993
1-NN	0.892	0.990
Random subspace		
LSVM	0.935	0.991
1-NN	0.921	0.990

Table 2 Performance obtained using the 2-gram features

2-gram	MCC	AUC
Stand-alone		
LSVM	0.848	0.982
1-NN	0.657	0.989
Random subspace		
LSVM	0.942	0.989
1-NN	0.401	0.903

Table 3 Comparison among several methods

Methods	MCC	AUC
Fusion between the two stand-alone LSVM	0.971	0.995
Fusion between the two random subspace of LSVM	0.971	0.994

The results reported in Table 3 show that the parallel fusion is very useful, the fusion drastically outperforms the base classifiers.

In order to confirm the benefit of our method the DET curve has been also considered. The DET curve (Martin et al. 1997) is a two-dimensional measure of classification performance that plots the probability of false alarm (i.e. false positive) against the probability of mis-detection (i.e. false negative). In Fig. 3 the DET curves obtained by fusion between the two stand-alone linear support vector machine (FUS) and by the linear support vector machine trained using the 2-gram features (2G) are plotted.

As further experiment, we run the Wilcoxon Signed-Rank test (Demsar 2006) for comparing the results (the MCC is used as performance indicator) of FUS and the linear support vector machine trained using the Gene Ontology based features. The null hypothesis is that there is no difference between the accuracies of the two classifiers (Demsar 2006). We reject the null hypothesis (level of significance 0.05) and accept that the two classifiers have significant different accuracies. This result confirms that FUS outperforms the stand-alone approaches.

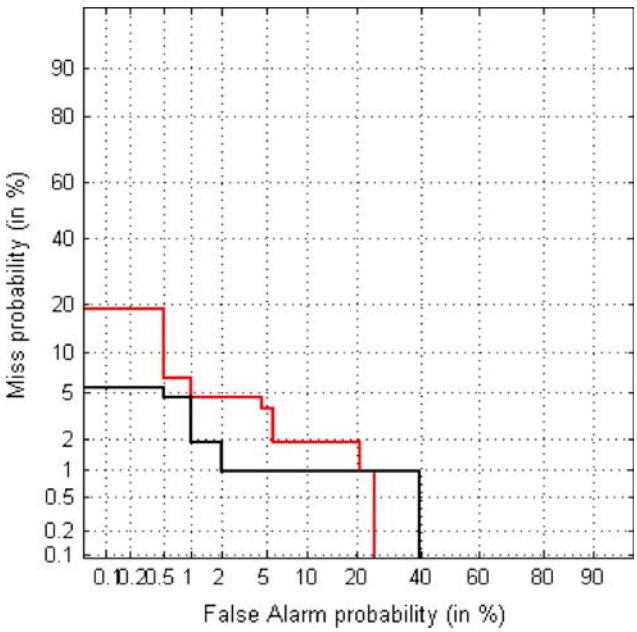

Fig. 3 The DET curves obtained by FUS (*black line*) and by 2G (*red line*)

Finally, we run the Q-statistic method to measure the complementarity between the two classifiers that belong to the method FUS. We obtain a value of 0.58; this result explains the good performance of our parallel fusion.

For two classifier D_i and D_k the Yule's Q-statistic (Kuncheva and Whitaker 2003) is

$$Q_{i,k} = \frac{ad - bc}{ad + bc}$$

where a is the probability of both classifiers being correct, d is the probability of both classifiers being incorrect, b is the probability that the first classifier is correct and second is incorrect, c is the probability that the second classifier is correct and the first is incorrect. For statistically independent classifiers, $Q_{i,k} = 0$. Q varies between -1 and 1. Classifiers that tend to classify the same patterns correctly will have $Q > 0$, and those which commit errors on different patterns, will have $Q < 0$.

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

- The Gene Ontology based feature extraction permit to obtain a reliable method;
- The random subspace permits to improve the performance of linear support vector machine when the 2-grams features are used; the random subspace of support vector machine obtains a Matthew's correlation coefficient of 0.942, when it is trained by the 2-gram features. In the literature it is well known that random subspace works well when the features are correlated (Ho 1998), our results confirm that the 2-grams features are correlated and that random subspace could be an interesting method to solve these drawbacks;
- All the tested fusions permit to outperform the method proposed in (Fang et al. 2007), where a Matthew's correlation coefficient of 0.924 is reported (on the same dataset), when the Chou's pseudo amino-acid composition is used;
- Our results confirm that Chou's pseudo amino-acid composition outperforms the 2-grams features, but we show that the performance of a classifier based on 2-grams features increases if an ensemble of classifiers is used.

Conclusions

We investigated the fusion of classifiers for predicting DNA-binding proteins. To enforce the diversity we combine two classifiers based on features extracted from the proteins. It is believed that classifiers based on different features offer complementary information about the patterns to be classified. In our test the Q-statistic between the two combined classifiers is only 0.58. Our method describes a given protein by the features extracted from the Gene Ontology database and from the amino-acid sequence.

We can draw the conclusion that it is reasonable to study the parallel fusion between the two tested feature extraction methods and we want to stress that our method can be successfully used in all the methods where the serial fusion is used (e.g. Chou and Shen 2007a).

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.

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

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 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

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

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 (Erratum: ibid., 2001, Vol.44, 60)

Chou KC (2005) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21:10–19
Chou KC, Cai YD (2004) Prediction of protein subcellular locations by GO-FunD-PseAA predictor, Biochem Biophys Res Commun 320:1236–1239
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, 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, Cambridge
Demsar J (2006) Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res 7:1–30
Diao Y, Li M, Feng Z, Yin J, Pan Y (2007a) 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 (2007b) Using pseudo amino acid composition to predict transmembrane regions in protein: cellular automata and Lempel-Ziv complexity. Amino Acids. 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 (2007) Predicting DNA-binding proteins: approached from Chou's pseudo amino acid composition and other specific sequence features. Amino Acids. doi:10.1007/s00726-007-0568-2
Fawcett T (2004) ROC graphs: notes and practical considerations for researchers. Technical report, HP Laboratories, Palo Alto
Feng ZP (2002) An overview on predicting the subcellular location of a protein. In Silico Biology 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
Ho TK (1998) The random subspace method for constructing decision forests. IEEE Trans Pattern Anal Mach Intell 8:832–844
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
Kuncheva LI, Whitaker CJ (2003) Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Mach Learn 51:181–207
Kurgan LA, Stach W, Ruan J (2007) Novel scales based on hydrophobicity indices for secondary protein structure. J Theor Biol 248:354–366
Kuznetsov I, 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. doi:10.1007/s00726-007-0545-9
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 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
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. Prot J 24:385–389
Martin A et al (1997) The DET curve in assessment of decision task performance. In: Proc. of EuroSpeech, pp 1895–1898
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
Niu B, Cai YD, Lu WC, Zheng GY, Chou KC (2006) Predicting protein structural class with AdaBoost learner. Protein Pept Lett 13:489–492
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. doi:10.10.1016/j.ab.2007.10.012
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. 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
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 (Erratum, ibid. 2005, 29:301) 28:395–402
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 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 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
Zhang TL, Ding YS (2007) Using pseudo amino acid composition and binary-tree support vector machines to predict protein structural classes. Amino Acids. doi:10.1007/s00726-007-0496-1
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