

Using Chou's pseudo amino acid composition based on approximate entropy and an ensemble of AdaBoost classifiers to predict protein subnuclear location

Xiaoying Jiang · Rong Wei · Yanjun Zhao ·
Tongliang Zhang

Received: 19 November 2007 / Accepted: 17 January 2008 / Published online: 7 February 2008
© Springer-Verlag 2008

Abstract The knowledge of subnuclear localization in eukaryotic cells is essential for understanding the life function of nucleus. Developing prediction methods and tools for proteins subnuclear localization become important research fields in protein science for special characteristics in cell nuclear. In this study, a novel approach has been proposed to predict protein subnuclear localization. Sample of protein is represented by Pseudo Amino Acid (PseAA) composition based on approximate entropy (ApEn) concept, which reflects the complexity of time series. A novel ensemble classifier is designed incorporating three AdaBoost classifiers. The base classifier algorithms in three AdaBoost are decision stumps, fuzzy K nearest neighbors classifier, and radial basis-support vector machines, respectively. Different PseAA compositions are used as input data of different AdaBoost classifier in ensemble. Genetic algorithm is used to optimize the dimension and weight factor of PseAA composition. Two datasets often used in published works are used to validate the performance of the proposed approach. The obtained results of Jackknife cross-validation test are higher and more balance

than them of other methods on same datasets. The promising results indicate that the proposed approach is effective and practical. It might become a useful tool in protein subnuclear localization. The software in Matlab and supplementary materials are available freely by contacting the corresponding author.

Keywords Protein subnuclear localization · Pseudo amino acid composition · Approximate entropy · Ensemble classifier · AdaBoost algorithm

Introduction

In eukaryotic cells, the nucleus is the largest, most prominent structure organelle. It organizes the assembly of genes and the life processes of cell, such as directing cellular reproduction and controlling cellular differentiation during the development of the organism, etc. Compartmentalization of cell nucleus is closely related to several nuclear processes, and has potential influence of cancer-related alternations on gene express (Fraser and Bickmore 2007; Schneider and Grosschedl 2007). Mis-localized nuclear proteins can lead to human genetic disease and cancer (Sutherland et al. 2001; Zaidi et al. 2007). Thus, the knowledge of proteins subnuclear localization is essential for understanding cell life processes and genomic regulation. Prediction of protein subnuclear location is thus an important topic in bioinformatics. Although the protein subnuclear localization can be determined by experiments ways, it is time-consuming and costly. The gap between the number of protein sequences and the number of identified proteins is rapidly increasing. It is highly desired to develop computational methods for fast identifying the proteins' subnuclear localization in cell nucleus.

X. Jiang
School of Chemistry and Chemical Engineering,
Henan Institute of Science and Technology, Xinxiang,
Henan 453003, People's Republic of China

R. Wei · Y. Zhao
College of Science, Hebei Polytechnic University,
Tangshan, Hebei 063009, People's Republic of China

T. Zhang (✉)
College of Information Sciences and Technology,
Donghua University, Shanghai 201620,
People's Republic of China
e-mail: tl.zhang@mail.dhu.edu.cn

Various approaches for protein subcellular localization prediction have been developed (Cai et al. 2002; Cai and Chou 2003; Chou 2001; Chou and Cai 2002, 2004; Chou and Shen 2006a, b, 2007a, 2008; Gao et al. 2005a, b; Park and Kanehisa 2003; Shen and Chou 2006, 2007a, b, c, d; Xiao et al. 2005, 2006; Zhang et al. 2006b, c, d, 2007; Chen and Li 2007a, b; Zhou and Doctor 2003) [see Chou (2000), Feng (2002) and Chou and Shen (2007b) for a comprehensive review in this area] since the algorithm proposed by Nakashima and Nishikawa (1994). Many recent prediction algorithms have been built as web servers freely available for scientist. However, the prediction algorithms for proteins subnuclear localization are far less than those for subcellular localization. Yet, only few studies have been carried out. Shen and Chou (2005) developed the first algorithm for prediction of nine classes subnuclear compartments localization, of which protein sequence is represented by Pseudo Amino Acid (PseAA) composition (34-D) and Optimized Evidence-Theoretic KNN (OET-KNN) is used as prediction engine. Lei and Dai (2005) employ support vector machine (SVM) for prediction of six classes subnuclear localization. Huang et al. (2007) developed algorithm named ProLoc using SVM with automatic selection from physicochemical composition features. Encouraged by the concept of PseAA discrete model introduced by Chou (2001), three predicting approaches based on PseAA composition have been proposed (Mundra et al. 2007; Li and Li 2008; Shen and Chou 2007a, b, c, d, e). Meanwhile, many Web-servers for predicting subcellular localization of proteins in various organisms have been established. Recently, two protocols with a step-by-step guide were published (Chou and Shen 2008; Emanuelsson et al. 2007; Shen and Chou 2007d) to help experimental scientists how to use some important Web-servers to predict the results they need. Three reasons are mainly responsible for limited study in this field (Lei and Dai 2005; Mundra et al. 2007): (1) proteins within the cell nucleus face no apparent physical barrier like a membrane; (2) the nucleus is far more compact and complicated in comparison with other compartments in a cell; and (3) protein complexes within the cell nucleus are not static.

Compared with the conventional amino acid composition (AAC), the PseAA as originally introduced by Chou can incorporate much more information of a protein sequence so as to remarkably enhance the power of using a discrete model to predict various attributes of a protein. Based on the concept of PseAA composition, a series of follow-up studies have been made to predict protein subcellular localization and other protein's attributes (Chen et al. 2006a, b; Chen and Li 2007a, b; Diao et al. 2008; Ding et al. 2007; Du and Li 2006; Fang et al. 2008; Gao et al. 2005b; Kurgan et al. 2007; Li and Li 2008; Lin and Li 2007a, b; Mondal et al. 2006; Mundra et al. 2007; Pu et al.

2007; Shi et al. 2007; Xiao and Chou 2007; Xiao et al. 2006; Zhang et al. 2006a, b, d, 2007, 2008; Zhang and Ding 2007; Zhou et al. 2007a, b). The promising results obtained from the approaches based on PseAA composition indicate that the PseAA discrete model can represent protein sequence in different subnuclear compartments effectively. In this study, we propose a prediction system for prediction subnuclear localization based on ensemble classifier, where sample of protein is represented with PseAA characterized by approximate entropy (ApEn). The ApEn is a non-negative number that denotes the complexity of time series (Pincus 1991; Richman and Moorman 2000). When the amino acids along a protein chain are replaced by a series of numbers, the protein sequence can be imaged as a short time series. Various studies based on ensemble classifier have been executed in protein attributes (Chou and Shen 2006a, b, 2008; Shen and Chou 2006; Shen and Chou 2007a, b, c, d, e; Nanni and Lumini 2007; Kedarisetti et al. 2006). According to the concept of Chou's PseAA discrete model, the weight factor is essential for PseAA composition. Note that λ is an uncertain parameter. Shen and Chou (2006) used an ensemble approach by fusing PseAA component with different λ and it has been successfully used to enhance the prediction quality in a number of relevant areas (see, e.g., Chou and Shen 2007a; Chou and Shen 2007b). The ensemble classifier is the ensemble of three AdaBoost classifiers which is one of Boosting ensemble methods that have the ability of generating a strong classifier from a weak method (Freund and Schapire 1997) and it has been used in predicting protein structural classes (Niu et al. 2006).

Materials and methods

Datasets

Two datasets often used in published works are adopted to validate the performance of the proposed approach. The one is the SNL9 (Shen and Chou 2005), which have 370 proteins localized in 9 subnuclear compartments: 10 Cajal body, 59 chromatin, 31 heterochromatin, 65 nuclear diffuse, 25 nuclear pore, 15 nuclear speckle, 115 nucleolus, 10 PcG body and 40 PML body. Another is SNL6 (Lei and Dai 2005) which contains 504 proteins localized in 6 subnuclear compartments: 38 PML body, 61 chromatin, 75 nuclear diffuse, 219 nucleolus, 56 nuclear speckle, and 55 nuclear lamina.

Representation of protein sequence

According to the concept of Chou's PseAA (Chou 2001) composition, a sample of protein sequence is a point in $(20 + \lambda)$ -D space.

$$X = [x_1, x_2, \dots, x_{20}, x_{21}, \dots, x_{20+\lambda}]^T \in \mathbb{R}^{(20+\lambda)} \quad (1)$$

$$x_i = \begin{cases} \frac{f_i}{\sum_{j=1}^{20} f_j + w \sum_{j=1}^{\lambda} p_j} & (1 \leq i \leq 20) \\ \frac{w \mu_i}{\sum_{j=1}^{20} f_j + w \sum_{j=1}^{\lambda} p_j} & (21 \leq i \leq 20 + \lambda) \end{cases} \quad (2)$$

where, the f_i ($1 \leq i \leq 20$) in Eq. (2) is the occurrence frequencies of 20 amino acids in sequence, i.e., the AAC which was often used as representation of protein sequence in early studies (Chou 1995; Zhang et al. 1995; Nakashima and Nishikawa 1994; Shen et al. 2005). p_i ($21 \leq i \leq 20 + \lambda$) is the additional factors that incorporate some sort of sequence order information. The parameter w is weight factors. Encouraged by our previous success in the design of PseAA composition with ApEn for prediction of protein structural classes (Zhang et al. 2008) and subcellular localization (Zhang et al. 2006c), the ApEn values of protein sequence still are adopted as additional factors in PseAA composition. The ApEn values of a sample of protein sample could be easily computed [see Eqs. (3)–(8) in Zhang et al. (2008)].

Ensemble classifier prediction systems

It is well known that in many situations combining the output of several classifiers leads to an improved classification results (Opitz and Maclin 1999; Alexandre et al. 2001). The proposed prediction system consists of three AdaBoost classifiers that are combined into an ensemble. The AdaBoost is one of ensemble method that has the ability of generating a strong classifier from a weak classifier (Freund and Schapire 1997). The architecture of ensemble is illustrated in Fig. 1, where the weak classifiers of three AdaBoost classifiers are decision trumps (Schapire and Singer 1999), Fuzzy K-nearest neighbor classifier (FKNN) (Keller et al. 1985; Huang and Li 2004; Shen et al. 2006), and radial basis-SVMs (Cristianini and Shawe-Taylor 2000), respectively. The AdaBoost algorithm is obtained from the classification toolbox in Matlab (Duda et al. 2001). The description of AdaBoost algorithm is illuminated in next section. The AdaBoost method usually applies to two-class problems. For the current case of multi-class problems, the “one-Vs-one” strategy is adopted.

AdaBoost algorithm

Given one or more classification methods, one of the most natural ways of obtaining more accurate classifiers is the use of ensembles (Rodríguez and Maudes 2007). The

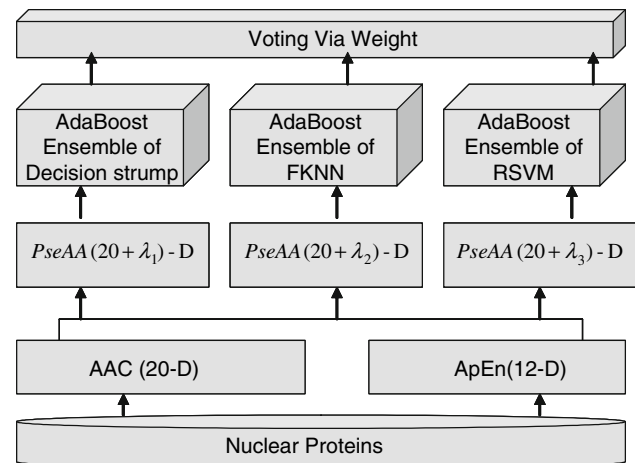


Fig. 1 The architectures of the ensemble of AdaBoost classifiers

ensemble method of Boosting is one of most successful methods. There are several variants, AdaBoost is the most well known.

Given the input dataset $S = \{(x_1, y_1), \dots, (x_n, y_n)\}$ where $x_i \in \mathbb{R}^m$ is the i th vector in m -D (dimensional) space, and $y_i \in \{-1, +1\}$ is the binary label of x_i . AdaBoost calls a weak learning algorithm repeatedly in a series of time intervals $t = 1, 2, \dots, T$. In the iteration t , a weight $D_t(x_i)$ is associated to the training sample x_i . The method generates a base classifier h_t , taking into account the weights distribution. It is necessary to determine a real value α_t . It is the weight associated to h_t and it depends on the training error of that classifier. The AdaBoost algorithm is illustrated in Fig. 2.

Performance measurement

In statistical prediction, the following three cross-validation tests are often used to examine the power of a predictor: sub-sampling (e.g., fivefold, sevenfold, etc.), jackknife, and independent dataset tests. Sub-sampling test,

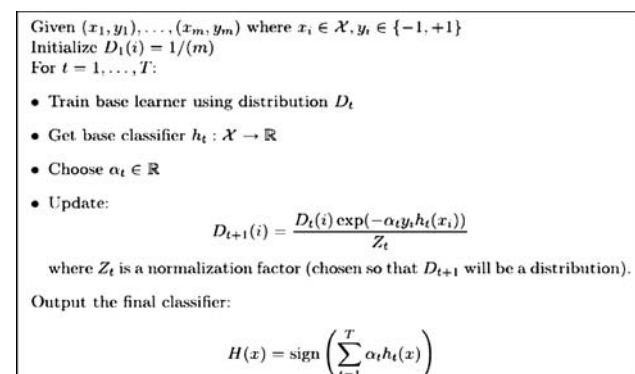


Fig. 2 AdaBoost algorithm, reproduced from Rodríguez and Maudes (2007) with permission

such as fivefold approach as often used in literatures, cannot avoid arbitrariness and yield a unique outcome even for a same benchmark dataset as illustrated by Eq. 50 of (Chou and Shen 2007a, b). Of these three, the jackknife test is thought the most rigorous and objective one [see (Chou and Zhang 1995) for a comprehensive review in this regard], and hence has been used by more and more investigators (see, e.g., Chen et al. 2007; Chou and Shen 2007b; Diao et al. 2007, 2008; Ding et al. 2007; Fang et al. 2008; Gao et al. 2005a, b; Guo et al. 2006; Li and Li 2008; Liu et al. 2007; Niu et al. 2006; Shen and Chou 2007a, b, c; 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. 2005, 2006; Zhang et al. 2006a, b, c, d, 2008; Zhang and Ding 2007; Zhou et al. 2007a) in examining the power of various prediction methods.

In statistic prediction study, it is convenient to introduce an accuracy matrix $[M_{ij}]$ of size $c \times c$ (c is the number of compartments to be predicted). The element M_{ij} of accuracy matrix is the number of proteins predicted to be in subnuclear compartment j , which are actually in the compartment i .

Three indexes are applied to evaluate the prediction accuracy, i.e., sensitivity (S_n), specificity (S_p), and Matthew's correlation coefficients (CC).

$$S_n = \frac{M_{ii}}{\sum_{j=1}^c M_{ij}} \quad (3)$$

$$S_p = \frac{M_{ii}}{\sum_{j=1}^c M_{ji}} \quad (4)$$

$$CC = \frac{M_{ii} \left(\sum_{k \neq i}^c \sum_{j \neq i}^c M_{jk} \right) - \left(\sum_{j \neq i}^c M_{ij} \right) \times \left(\sum_{j \neq i}^c M_{ji} \right)}{\left[\left(M_{ii} + \sum_{j \neq i}^c M_{ij} \right) \left(M_{ii} + \sum_{j \neq i}^c M_{ji} \right) \left(\sum_{k \neq i}^c \sum_{j \neq i}^c M_{jk} + \sum_{j \neq i}^c M_{ji} \right) \left(\sum_{k \neq i}^c \sum_{j \neq i}^c M_{jk} + \sum_{j \neq i}^c M_{ij} \right) \right]^{1/2}} \quad (5)$$

$$A_c = \left(\sum_{i=1}^c M_{ii} \right) / \left(\sum_{i=1}^c \sum_{j=1}^c M_{ij} \right) \quad (6)$$

S_n represents the accuracy, and S_p represents the reliability in procedure of prediction. The CC is a single parameter characterizing the matching extent between the observed and predicted subnuclear compartments.

Results and discussion

According to the concept of Chou's PseAA (Chou 2001), the weight factor and the dimension of addition feature are

essential parameter. In this study, genetic algorithm (GA) toolbox in Matlab is used to optimize the weight factors and lambda (λ) in Eq. (2). The ranges of parameters selection: weight factor $w \in [0, 1]$ and the dimension of ApEn $\lambda \in [1, 12]$. The overall accuracy value of jackknife test with given classifier is used as the result of fitness function in GA, where the give classifiers is three AdaBoost classifiers, respectively. The crossover rate $P_c = 0.9$ and mutation rate $P_m = 0.2$. When the weight factors (w) and the dimension of ApEn (λ) are determined using by the three AdaBoost classifier, respectively, the input PseAA compositions of each base classifier within ensemble are also determined. The optimized parameters of three basic classifiers are listed in Table 1.

Three Adaboost classifiers are trained with the optimized parameters, dimension of ApEn (λ), and weight factors (w) in Eq. 2, respectively. The results of three Adaboost classifiers are fusing through weight voting, of which the voting weight is the success rate of each basic classifier. The fusion scheme has been used in several protein function prediction studies with ensemble classifier (Shen and Chou 2006, 2007e). The final results are listed in Table 2, where for facilitating comparison, the results by other methods on the same dataset are also listed. The overall accuracy (A_c) is 83.2%, distinctly higher than the methods of SVM (Cai et al. 2002), OET-KNN (Shen and Chou 2005), and PSSM (Mundra et al. 2007). The accuracies of each subnuclear compartment are higher or same as them of the method of PSSM (Mundra et al. 2007).

In order to validate the performance of the approach further, the dataset SNL 6 constructed by Lei and Dai

(2005) is also used, which is composed by 504 protein sequences classified in 6 subnuclear compartments. Same as the optimization process using by the dataset of SNL9, the optimized results of three base classifiers trained by the dataset of SNL6 are listed in Table 3. When the λ and w are determined, the PseAA composition as input data are also determined. The results of jackknife cross-validation test are deposited in Table 4. In order to compare with other methods, the results of the prediction methods on the same dataset are also listed in Table 4. The overall accuracy (A_c) of proposed approach is 73.2%, higher than the methods of Lei-SVM (Lei and Dai 2005) and ESVM (Huang et al. 2007). Comparing the accuracy (S_n) of subnuclear

Table 1 Optimization results of input data of three basic classifiers in AdaBoost with the dataset SNL9

Base classifier in AdaBoost	Parameters	λ	w	Jackknife accuracy (%)
Decision stumps	–	2	0.59	77.3
FKNN	$K = 3; P = 1.05$	9	0.79	78.4
RSVM	$C = 100; \gamma = 0.05$	10	0.16	80.6

Table 2 Results of Jackknife test by different algorithms on SNL9

Compartments	SVM ^a	OET-KNN ^b	PSSM ^c	This paper		
				S_n (%)	S_p (%)	CC
Cajal body	NA	NA	20.0	90.0	100.0	0.948
Chromatin	NA	NA	66.7	86.4	92.7	0.882
Heterochromatin	NA	NA	74.2	83.9	78.8	0.799
Nuclear diffuse	NA	NA	65.0	76.9	78.1	0.720
Nuclear pore	NA	NA	58.3	56.0	77.8	0.642
Nuclear speckle	NA	NA	50.0	40.0	75.0	0.537
Nucleolus	NA	NA	93.9	93.9	85.7	0.848
PcG body	NA	NA	30.0	60.0	85.7	0.71
PML body	NA	NA	60.0	95.0	76.0	0.830
A_c (%)	33.8	64.32	71.23	83.2		

NA not available

^a From the work of Cai et al. (2002)^b From the work of Shen and Chou (2005)^c From the work of (Mundra et al. 2007)**Table 3** Optimization results of input data of three basic classifiers in AdaBoost with the dataset SNL6

Base classifier in AdaBoost	Parameters	λ	w	Jackknife accuracy (%)
Decision stumps	–	1	0.73	67.5
FKNN	$K = 3; P = 1.05$	12	0.53	70.8
RSVM	$C = 100; \gamma = 0.05$	11	0.24	69.2

compartments in proposed approach with them in other methods, the accuracies of PML body, chromatin, nuclear speckles, and nuclear lamina, the four compartments with less proteins number, are higher than them of other methods. We can see that the results of proposed approach are more balance than that of the methods of Lei-SVM (Lei and Dai 2005) and ESVM (Huang et al. 2006).

The results of Jackknife cross-validation test with two datasets indicate that the proposed approach is effective and practical. The PseAA composition based on ApEn indeed reflects the core feature of proteins in different subnuclear compartments. Comparing the results of base classifier with that of ensemble, performance of ensemble is stronger than that of base classifier. The ensemble of

Table 4 Results of Jackknife test by different algorithms on the dataset of SNL6

Compartments	Lei-SVM (Lei and Dai 2005)	ESVM (Huang et al. 2006)	This paper		
			S_n (%)	S_p (%)	CC
Chromatin	21.3	21.3	100.0	100.0	1.00
Nuclear lamina	36.4	36.4	65.5	75.0	0.667
Nuclear speckles	33.9	26.8	62.5	58.3	0.552
Nucleolus	83.1	90.3	88.6	83.6	0.748
Nuclear diffuse	28.0	42.7	38.7	41.4	0.300
PML body	10.5	18.4	36.8	42.4	0.350
A_c (%)	51.4	56.4	73.2		

three AdaBoost classifiers might become a useful tool in prediction of protein subnuclear localization.

Conclusions

A novel approach for protein subnuclear localization is proposed. Sample of protein sequence is represented by PseAA based on ApEn. Ensemble classifier is used as prediction engine. The ensemble classifier is combined with three AdaBoost classifiers in which base classification algorithms are decision stumps, FKNN, and RSVM, respectively. The input data of AdaBoost classifier in ensemble is PseAA composition, of which dimension (λ) and weight factor (w) are optimized by GA. Two datasets often used in various methods of this area are used to validate the performance of the novel approach. Promising results obtained by jackknife cross-validation test indicate that the proposed approach is effective and practical, and might become a useful tool for prediction protein subnuclear localization.

Acknowledgments The authors wish to thank the two anonymous reviewers, whose constructive comments are very helpful for improving the presentation of this study.

References

- Alexandre LA, Campilho AC, Kamel M (2001) On combining classifiers using sum and product rules. Pattern Recognit Lett 22:1283–1289

- Cai YD, Chou KC (2003) Nearest neighbour algorithm for predicting protein subcellular location by combining functional domain composition and pseudo-amino acid composition. *Biochem Biophys Res Commun* 305:407–411
- Cai YD, Liu XJ, Xu XB, Chou KC (2002) Support vector machines for prediction of protein subcellular location by incorporating quasi sequence-order effect. *J Cell Biochem* 84:343–348
- Chen YL, Li QZ (2007a) Prediction of the subcellular location of apoptosis proteins. *J Theor Biol* 245:775–783
- Chen YL, Li QZ (2007b) Prediction of apoptosis protein subcellular location using improved hybrid approach and pseudo-amino acid composition. *J Theor Biol* 248(2):377–318
- 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
- Cristianini N, Shawe-Taylor J (2000) An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press, London
- Chou KC (1995) A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. *Proteins Struct Funct Genet* 21:319–344
- Chou KC (2000) 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 Struct Funct Genet* 43(3):246–255
- 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 (2004) Prediction of protein subcellular locations by GO-FunD-PseAA predictor. *Biochem Biophys Res Commun* 320:1236–1239
- Chou KC, Shen HB (2006a) Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers. *J Proteome Res* 5:1888–1897
- Chou KC, Shen HB (2006b) 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) Review: recent progresses in protein subcellular location prediction. *Anal Biochem* 370:1–16
- 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
- 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
- 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 R, Hart P, Stork DG (2001) Pattern classification, 2nd edn. Wiley, New York
- Emanuelsson O, Brunak S, von Heijne G, Nielsen H (2007) Locating proteins in the cell using TargetP, SignalP and related tools. *Nat Protoc* 2:953–971
- 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
- Feng ZP (2002) An overview on predicting the subcellular location of a protein. *In Silico Biol* 2:291–303
- Fraser P, Bickmore W (2007) Nuclear organization of the genome and the potential for gene regulation. *Nature* 447(7143):413–417
- Freund Y, Schapire R (1997) A decision-theoretic generalization of online learning and an application to boosting. *J Comput Syst Sci* 55:119–139
- Gao QB, Wang ZZ, Yan C, Du YH (2005a) Prediction of protein subcellular location using a combined feature of sequence. *FEBS Lett* 579:3444–3448
- Gao Y, Shao SH, Xiao X, Ding YS, Huang YS, Huang ZD, Chou KC (2005b) 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 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
- Huang Y, Li Y (2004) Prediction of protein subcellular locations using fuzzy k-NN method. *Bioinformatics* 20(1):21–28
- Huang WL, Tung CW, Huang HL, Hwang SF, Ho SY (2007) ProLoc: prediction of protein subnuclear localization using SVM with automatic selection from physicochemical composition features. *BioSystems* 90(2):573–581
- Kedarisetti KD, Kurgan LA, Dick S (2006) Classifier ensembles for protein structural class prediction with varying homology. *Biochem Biophys Res Commun* 348:981–988
- Keller JM, Gray MR, Givens JA (1985) A fuzzy k-nearest neighbours algorithm. *IEEE Trans Syst Man Cybern* 15:580–585
- Kurgan LA, Stach W, Ruan J (2007) Novel scales based on hydrophobicity indices for secondary protein structure. *J Theor Biol* 248:354–366
- Lei ZD, Dai Y (2005) An SVM-based system for predicting protein subnuclear localizations. *BMC Bioinformatics* 6:291–298
- Li FM, Li QZ (2008) Using pseudo amino acid composition to predict protein subnuclear location with improved hybrid approach. *Amino Acids* 34(1):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 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
- Nakashima H, Nishikawa K (1994) Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. *J Mol Biol* 238(1):54–61

- Nanni L, Lumini A (2007) Ensemblator: an ensemble of classifiers for reliable classification of biological data. *Pattern Recognit Lett* 28:622–630
- Niu B, Cai YD, Lu WC, Li GZ, Chou KC (2006) Predicting protein structural class with AdaBoost learner. *Protein Pept Lett* 13:489–492
- Opitz D, Maclin R (1999) Popular ensemble methods: an empirical study. *J Artif Intell Res* 11:169–198
- Park KJ, Kanehisa M (2003) Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. *Bioinformatics* 19(13):1656–1663
- Pincus SM (1991) Approximate entropy as a measure of system complexity. *Proc Natl Acad Sci USA* 88:2297–301
- 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
- Richman JS, Moorman JR (2000) Physiological time-series analysis using approximate entropy and sample entropy. *Am J Physiol Heart Circ Physiol* 278:H2039–H2049
- Rodríguez JJ, Maudes J (2007) Boosting recombined weak classifiers. *Pattern Recognit Lett*. doi:[10.1016/j.patrec.2007.06.019](https://doi.org/10.1016/j.patrec.2007.06.019)
- Schapire RE, Singer Y (1999) Improved boosting algorithms using confidence-rated predictions. *Mach Learn* 37(3):297–336
- Schneider R, Grosschedl R (2007) Dynamics and interplay of nuclear architecture, genome organization, and gene expression. *Genes Dev* 21(23):3027–3043
- Shen HB, Chou KC (2005) 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 (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) 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, 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(11):561–567
- Shen HB, Chou KC (2007e) Using ensemble classifier to identify membrane protein types. *Amino Acids* 32(4):483–488
- Shen HB, Yang J, Liu XJ, Chou KC (2005) Using supervised fuzzy clustering to predict protein structural classes. *Biochem Biophys Res Commun* 334:577–581
- 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
- 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
- Sutherland HE, Mumford GK, Newton K, Ford LV, Farrall R, Dellaire G, Ca'ceres JF, Bickmore WA (2001) Large-scale identification of mammalian proteins localized to nuclear sub-compartments. *Hum Mol Genet* 10:1995–2011
- 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
- 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)
- 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, Huang Y, Chou KC (2005) Using complexity measure factor to predict protein subcellular location. *Amino Acids* 28:57–61
- Xiao X, Shao SH, Ding YS, Huang ZD, Chou KC (2006) Using cellular automata images and pseudo amino acid composition to predict protein subcellular location. *Amino Acids* 30:49–54
- Zaidi SK, Young DW, Javed A, Pratap J, Montecino M, van Wijnen A, Lian JB, Stein JL, Stein GS (2007) Nuclear microenvironments in biological control and cancer. *Nat Rev Cancer* 7(6):454–463
- Zhang TL, Ding YS (2007) Using pseudo amino acid composition and binary-tree support vector machines to predict protein structural classes. *Amino Acids* 33:623–629
- Zhang CT, Chou KC, Maggiora GM (1995) Predicting protein structural classes from amino acid composition: application of fuzzy clustering. *Protein Eng* 8:425–435
- 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 TL, Ding Y, Chou KC (2006b) Prediction of protein subcellular location using hydrophobic patterns of amino acid sequence. *Comput Biol Chem* 30:367–371
- Zhang ZH, Wang ZH, Zhang ZR, Wang YX (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 TL, Ding YS, Shao SH (2006d) Protein subcellular location prediction based on pseudo amino acid composition and Immune genetic algorithm. *ICIC 2006, LNBI 4115*, 534–542
- 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](https://doi.org/10.1007/s00726-007-0010-9)
- Zhang TL, Ding YS, Chou KC (2008) Prediction protein structural classes with pseudo amino acid composition: approximate entropy and hydrophobicity pattern. *J Theor Biol* 250:186–193
- Zhou GP, Doctor K (2003) Subcellular location prediction of apoptosis proteins. *Proteins Struct Funct Genet* 50:44–48
- Zhou XB, Chen C, Li ZC, Zou XY (2007a) Improved prediction of subcellular location for apoptosis proteins by the dual-layer support vector machine. *Amino Acids*. doi:[10.1007/s00726-007-0608-y](https://doi.org/10.1007/s00726-007-0608-y)
- Zhou XB, Chen C, Li ZC, Zou XY (2007b) Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. *J Theor Biol* 248:546–551