J Membrane Biol (2016) 249:551–557
DOI 10.1007/s00232-016-9904-3

Classifying Multifunctional Enzymes by Incorporating Three Different Models into Chou's General Pseudo Amino Acid Composition

Hong-Liang Zou[1,2] · **Xuan Xiao**[2,3,4]

Received: 23 October 2015 / Accepted: 11 April 2016 / Published online: 25 April 2016

Abstract With the avalanche of the newly found protein sequences in the post-genomic epoch, there is an increasing trend for annotating a number of newly discovered enzyme sequences. Among the various proteins, enzyme was considered as the one of the largest kind of proteins. It takes part in most of the biochemical reactions and plays a key role in metabolic pathways. Multifunctional enzyme is enzyme that plays multiple physiological roles. Given a multifunctional enzyme sequence, how can we identify its class? Especially, how can we deal with the multi-classes problem since an enzyme may simultaneously belong to two or more functional classes? To address these problems, which are obviously very important both to basic research and drug development, a multi-label classifier was developed via three different prediction models with multi-label K-nearest algorithm. Experimental results obtained on a stringent benchmark dataset of enzymes by jackknife cross-validation test show that the predicting results were exciting, indicating that the current method could be an effective and promising high throughput method in the enzyme research. We hope it could play an important complementary role to the existing predictors in identifying the classes of enzymes.

Keywords Multifunctional enzyme · Prediction model · Multi-label K-nearest algorithm · Jackknife cross-validation test

Introduction

Enzyme plays a key role in catalyzing various biological reactions in the cell. Enzymes are considered as one of the most important biological catalysts in the metabolism of all organisms, they have attracted the attention of various investigators in the past decades. Identification and classification of enzymes are extremely beneficial in understanding their cellular functions and consequently in the design and development of drugs from a therapeutic perspective (Zou et al. 2013). According to the Enzyme Commission (EC) organizes, enzyme mainly divided into the following six classes: (1) oxidoreductases, (2) transferases, (3) hydrolases, (4) lyases, (5) isomerases, and (6) ligases.

Because of the class of enzyme keeps closely correlation with its functions, knowledge about the class of enzyme is constructive in understanding the mechanism of metabolism. Although the class of an enzyme may be determined by carrying out various biochemical experiments, it is both time-consuming and costly, so it is an urgent to developing an automated computed method for accurately and efficiently identifying the classes of the query enzyme.

In the past several decades, many efforts have been made in identifying the functional class of enzyme, such as Cai et al. (2004) using CTD (composition, translation, and

Electronic supplementary material The online version of this article (doi:10.1007/s00232-016-9904-3) contains supplementary material, which is available to authorized users.

✉ Hong-Liang Zou
hongliangzou@126.com

[1] Department of Mechanical and Electronic Information Engineering, Jiangxi University of Applied Science, Nanchang 330100, China

[2] Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 333046, China

[3] Information School, ZheJiang Textile & Fashion College, Ningbo 315211, China

[4] Gordon Life Science Institute, 53 South Cottage Road, Belmont, MA 02478, USA

distribution) with support vector machine (SVM) predicting the classes of enzymes; Zhou et al. (2007) using amphiphilic pseudo amino acid composition and support vector machine for prediction the classes of enzyme subfamily; Shen and Chou (2007) predicting enzyme functional classes and subclasses by a top-down method, and many others (Chou 2005; Chou and Cai 2004a; Chou and Elrod 2003; Chou and Cai 2004b; Khan et al. 2015; Shen and Chou 2007; Zhou et al. 2007).

At present, the phenomena of multi-label very widespread, and many multi-systems have been established (Chou and Shen 2007a, 2010b; Huang and Yuan 2013b, 2015; Mei 2012; Shen and Chou 2009; Wang et al. 2015; Zou and Xiao 2015). Although the above-mentioned methods have each of their advantages and did play a key role in stimulating the development in this area, they were established under the assumption that an enzyme only with one functional class. However, an enzyme may simultaneously belong to two or more classes. Enzymes with multiple classes are particularly interesting, because they may have some unique biological functions worthy of our special notice (Glory and Murphy 2007; Smith 2008). Particularly, when the enzyme with several different functions, that is multifunctional enzymes. Thus, the current existing prediction methods are not suitable to the situation. Therefore, it is urgent and meaningful to develop a predictor to deal with multifunctional enzymes sequences with single and multiple functional classes (Huang and Yuan 2013a).

For a multi-label learning system, each sample in the training set may be associated with not limited to one label, and the mission is based on the model induced from multi-label training samples with known label sets to predict a label set for each unseen instance. In this study, a multi-label algorithm called ML-KNN, i.e., multi-label K-nearest neighbor, was adopted. ML-KNN was stemmed from the classical K-nearest neighbor (KNN) algorithm. Firstly, for every test sample, its K-nearest neighbors in the training set are identified. Then, according to statistical information gained from the label sets of those neighboring samples, i.e., the number of neighboring samples belonging to each possible class, maximum a posteriori principle is utilized to determine the label set for the test sample (Zhang and Zhou 2007).

To establish a powerful predictor, the following several procedures should be considered (Chou 2011): (1) construct or select a stringent benchmark dataset to train and test the predictor; (2) use a valid mathematical expression to formulate the sequences, which can truly reflect the intrinsic correlation with the target to be predicted; (3) introduce or develop a powerful algorithm (or engine) to operate the prediction; (4) properly perform cross-validation tests to objective examine the anticipated accuracy of the predictor.

Materials and Methods

Benchmark Dataset

All of the enzymes sequences were collected from the Enzyme nomenclature database at website http://enzyme.expasy.org/. To construct a high and updated benchmark dataset for developing a predictor to identifying the classes of enzymes, the following steps should be considered:

Step 1 Only those sequences with keyword "multifunctional enzyme" were collected.

Step 2 The sequences annotated with "fragment" should be removed.

Step 3 The sequences with length less than 50 amino acid residues were also removed, because these sequences may be belong to fragment.

Step 4 To reduce the influence of redundancy and homology bias, the program CD-HIT (Huang et al. 2010) was used to exclude these enzymes that had more than 80 % pairwise sequence identify to any other in a same subset.

Finally, we obtained 3095 different enzyme sequences. These sequences together form the benchmark dataset S which is used in the current study, and it covers 6 different classes and can be formulated as follows (Lin et al. 2013a):

$$S = S_1 \cup S_2 \cup S_3 \cup S_4 \cup S_5 \cup S_6 \quad (1)$$

where $\cup$ stands for the symbol for "union" in the set theory, while S_1 represents the subset of "oxidoreductases," S_2 for "transferases," S_3 for "hydrolases," and so on. The particular information about the dataset is listed in Table 1.

To establish an effective predictor for statistically predicting classes of multifunctional enzymes based on the sequence information, one of the most important steps is to formulate the sequences with an efficient mathematics expression that can truly reflect the correlation with the target to be identified (Chou 2011). To represent the protein sample, the following two models were often used:

Table 1 The benchmark dataset constructed in this study

Order	Class of enzyme	Number of enzyme
1	Oxidoreductases	800
2	Transferases	1931
3	Hydrolases	1351
4	Lyases	655
5	Isomerases	166
6	Ligases	139
Total number of virtual enzymes		5042
Total number of different enzymes		3095

Of the 3095 different enzymes sequences, 1302 belong to one class, 1647 to two classes, 138 to three classes, 8 to four classes—i.e., there are total 5042 enzyme sequences

sequential model and discrete model. In the sequential model, the sequence similarity search-based tools were used to conduct the prediction. However, the method would lose its function when an uncharacterized protein did not exist significant homology to attribute-known proteins. Thus, to address the problem, various discrete models were proposed.

Representation of Enzyme Sample

Among the discrete models, amino acid composition may be the simplest, which is short of AAC (Nakashima et al. 1986). According to the AAC-discrete model, protein sequence P can be formulated like (Chou and Zhang 1994; Lin et al. 2013a; Xiao et al. 2012):

$$P = [f_1 f_2 \ldots f_{20}]^{\mathrm{T}}, \tag{2}$$

where $f_i (i = 1, 2, \ldots, 20)$ represent the occurrence frequencies of the 20 native amino acids in protein P, while T is the symbol of transposing operator. Although AAC-discrete model has been employed for predicting a lot of protein attributes (Chou and Elrod 1999; Nakashima et al. 1986; Zhou 1998; Zhou and Assa-Munt 2001; Zhou and Doctor 2003), there exists a fatal disadvantage that if AAC model as the only one feature extract method was utilized to extract the information of the protein P, all of its sequence-order and sequence length information would be lost. Therefore, to avoid the situation arise, the pseudo amino acid composition (PseAAC) (Chou 2001) was put forward to replace the simple amino acid composition (AAC) to represent the sample of protein.

Based on the concept of PseAAC, a query protein P can be formulated by

$$P = [\phi_1 \phi_2 \phi_3 \cdots \phi_\tau]^{\mathrm{T}}, \tag{3}$$

where the subscript τ is a positive integer and its value rely on what information we want to extract from the protein sequence of P. Below, we would detailed introduce how to extract the information from protein sequence.

Chou's Pseudo Amino Acid Composition (PseAAC)

Ever since the concept of pseudo amino acid composition was introduced by Chou in 2001, it has widely used in bioinformatics and computational proteomics (Chen et al. 2009; Esmaeili et al. 2010; Li and Li 2008; Xiao et al. 2006). Because it has been widely and increasingly used, five open access softwares, called 'PseAAC' (Shen and Chou 2008), 'PseAAC-Builder' (Du et al. 2012), 'propy' (Cao et al. 2013), 'PseAAC-General' (Du et al. 2014) and 'Pse-in-one' (Liu et al. 2015b), were established: the 1st and 2nd ones are for generating various models of Chou's special PseAAC; the 3rd and 4th ones are for generating various Chou's general PseAAC; and the 5th one not only can generate varieties of PseAAC defined by users themselves but also can generate various feature vectors for DNA/RNA sequences. According to PseAAC, a protein sequence can be converted into a $20 + \lambda$ dimension vector, among the $20 + \lambda$ elements, the first 20 represent the amino acid composition of the 20 native amino acids, while the latter λ elements represent the sequence-order information. The sequence-order information can be indirectly represented by the following expression:

$$\delta_\eta = \frac{1}{L-\eta} \sum_{i=1}^{L-\eta} \Omega(R_i, R_{i+\eta}), \ (\eta = 1, 2, \ldots, \lambda \text{ and } \lambda < L), \tag{4}$$

where L represents the length of the sequence and the δ_η is the ηth correlation factor with which harbors the sequence-order information between all the η most contiguous residues. The correlation function $\Omega(R_i, R_j)$ can be defined as follows:

$$\Omega(R_i, R_j) = \frac{1}{3}\Big\{ \big[F(R_j) - F(R_i)\big]^2 + \big[G(R_j) - G(R_i)\big]^2 + \big[H(R_j) - H(R_i)\big]^2 \Big\} \tag{5}$$

where $F(R_i)$, $G(R_i)$, and $H(R_i)$ are the evaluated values of hydrophobicity, hydrophilicity, and mass, respectively. Before the three types of values were used, a standard conversion should be conducted using Eq. (4) of Huang and Yuan (2013a).

The numerical values of the three physical–chemical (PC) properties for each of the 20 native amino acids are listed in Table 2.

Thus, a protein sequence P with L amino acid residues can be formulated:

$$P = [x_1, x_2, \ldots, x_{20}, x_{20+1}, \ldots, x_{20+\lambda}]^{\mathrm{T}}, \lambda < L, \tag{6}$$

where

$$x_\varphi = \begin{cases} \dfrac{f_\varphi}{\sum_{i=1}^{20} f_i + w \sum_{\eta=1}^{\lambda} \delta_\eta}, & (1 \le \varphi \le 20) \\ \dfrac{w\delta_{\varphi-20}}{\sum_{i=1}^{20} f_i + w \sum_{\eta=1}^{\lambda} \delta_\eta}, & (20+1 \le \varphi \le 20+\lambda), \end{cases} \tag{7}$$

where w is the weight factor, $f_i (i = 1, 2, \ldots, 20)$ represent the normalized occurrence frequencies of the 20 native amino acids, and δ_η is the η-tier sequence-correlation factor, which can be computed by Eq. (4). According to Eqs. (4)–(7), we can see that the value of w and λ is very important to the prediction performance, by preliminary computation and analyse, we find that when $\lambda = 20$ and $w = 0.5$ the best results would be obtained.

Table 2 The numerical values of the three physical–chemical properties

AA	*PC*		
	Hydrophilicity	Hydrophobicity	Mass
A	−0.5	0.62	15.0
C	−1.0	0.29	47.0
D	3.0	−0.90	59.0
E	3.0	−0.74	73.0
F	−2.5	1.19	91.0
G	0.0	0.48	1.0
H	−0.5	−0.40	82.0
I	−1.8	1.38	57.0
K	3.0	−1.50	73.0
L	−1.8	1.06	57.0
M	−1.3	0.64	75.0
N	0.2	−0.78	58.0
P	0.0	0.12	42.0
Q	0.2	−0.85	72.0
R	3.0	−2.53	101.0
T	−0.4	−0.05	45.0
V	−1.5	1.08	43.0
W	−3.4	0.81	130.0
Y	−2.3	0.26	107.0

Split Amino Acid Composition (SAAC)

In the SAAC model, a protein sequence will be divided into three parts, and the amino acid composition of each part would be calculated separately. In view of this, the SAAC model was adopted in this study, and the enzyme sequence was divided into the following three segments: C termini, N termini, and the middle part. In these three parts, C termini and N termini contain 25 amino acid residues, respectively, and the others are included in the middle part. Thus, using a SAAC-based method, a sequence can be represented by a 60-dimension vector.

Grey Model (GM)

Grey model was first used in bioinformatics by Lin et al. (2011). It has been provided it is a useful tool in this area. Therefore, we also adopted it in this study to extract feature from sequence. According to grey model GM(2,1) (Lin et al. 2011), a sequence can be formulated as

$$P = [\phi_1 \phi_2 \cdots \phi_{20} \phi_{21} \phi_{22} \phi_{23}]^T, \tag{8}$$

where $\phi_i (i = 1, 2, \ldots, 20)$ are the occurrence frequencies of the 20 different types of amino acids in the protein concerned, while $\phi_j (j = 21, 22, 23)$ represent the absolute value of three coefficients. For the detailed description about the grey model, please refer to Lin et al. (2011).

Prediction Engine

In this study, the following multi-label algorithm was adopted to perform the prediction: multi-label K-nearest neighbor (ML-KNN). A detailed description about how the classifier works is clearly described in Zhang and Zhou (2007). The predictor established in this study can be used to predict the functional classes of both singleplex and multiplex multifunctional enzymes.

Results and Discussion

It is worthy point out that for a multi-label learning system like the current, which is different from the classical single-label learning system, hence those existed metrics used to evaluate the quality of a predictor on a single-label system will failed work when faced a multi-label problem like this. The metrics will be much more complicated for a multi-label learning system. Now, let us describe the metrics used in multi-label system in the following.

For a multi-label learning system contain N protein sequences, which belong to M functional classes, L is the label set that contain all of possible functional classes concerned. Thus, the $i-$th sequence P_i and its corresponding functional class can be expressed by

$$\{P_i, L_i\} \quad (i = 1, 2, \ldots, N), \tag{9}$$

where L_i is the subset that included all class label(s) for the ith protein. Obviously, we have

$$L_1 \cup L_2 \cup \cdots \cup L_N \subseteq L = \{l_1, l_2, \cdots l_M\}, \tag{10}$$

where $l_i (i = 1, 2, \ldots, M)$ corresponding to the label for the ith functional class. In this study, $N = 3095$ and $M = 6$. Assume L_i^* as the predicted label(s) for ith sample. Thus, the following five metrics can be used to measure the prediction quality of the multi-label system:

$$\begin{cases} \text{Absolute-false} = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\left(\dfrac{\|L_i \cup L_i^*\| - \|L_i \cap L_i^*\|}{M}\right) \\ \text{Accuracy} = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\left(\dfrac{\|L_i \cap L_i^*\|}{\|L_i \cup L_i^*\|}\right) \\ \text{Precision} = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\left(\dfrac{\|L_i \cap L_i^*\|}{\|L_i^*\|}\right) \\ \text{Recall} = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\left(\dfrac{\|L_i \cap L_i^*\|}{\|L_i\|}\right) \\ \text{Absolute true} = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\Delta(L_i, L_i^*) \end{cases}, \tag{11}$$

where N is the number of different multifunctional enzymes, M is the total number of classes, and here $N =$

3095 and $M = 6$. The symbols $\cup$ and $\cap$ represent "union" set theory and intersection, respectively. $\| \|$ represents the operator acting on the set therein to count the number of its elements, and

$$\begin{cases} \Delta(L_i, L_i^*) = 1, & \text{if all the labels in } L_i \text{ are identified to those in } L_i^* \\ \Delta(L_i, L_i^*) = 0, & \text{otherwise} \end{cases} \tag{12}$$

Among the five evaluation measures, the lower the absolute false is, the better the prediction quality will be. However, for the other four metrics, the situation is just opposite, i.e., the higher their rates are, the better the prediction quality will be.

In statistical prediction, it would be meaningless to simple say a success rate of a predictor without specifying what method and benchmark dataset were used to test its accuracy (Wu et al. 2012). As is well known, the following three methods often used to evaluate the performance of a predictor: independent test, n-fold cross-validation test (sub-sampling test), and jackknife test (leave-one-out cross-validation test), respectively. Among these three methods, the jackknife test was considered the most objective method because it always yields a unique result for a given benchmark dataset, and hence, it has been widely recognized and increasingly used by various researchers to examine the power of the predictors (Chou and Shen 2007b; Hayat et al. 2012; Lin et al. 2013b; Wu et al. 2012; Xiao et al. 2013).

However, even though the jackknife test was used as the cross-validation method, a same predictor may still generate obviously different results when tested by different benchmark datasets. This is because the more stringent of a benchmark dataset in excluding homologous and high similarity sequences, the more difficult for a predictor to achieve a high overall success rate (Chou and Shen 2010a).

Listed in Table 3 are the results obtained based on the aforementioned benchmark dataset S by the jackknife test. From Table 3, we can see that for such a multiplex benchmark dataset, the absolute-true rate is high, while the absolute-false rate is much lower, indicating the method is quite a promising method for identifying the functional classes of multifunctional enzymes.

It is instructive to point out that, for such a multi-label learning system, only say the absolute-true rate for each individual multifunctional enzyme functional classes is meaningless and misleading. Therefore, instead of the absolute-true success rate for each of individual functional classes, the results about the absolute-true success rate for multifunctional enzymes with different numbers of functional classes (or labels) are listed in Tables 4, 5, 6. Furthermore, in order to facilitate comparison, the

Table 3 The results obtained by different models with jackknife test

Evaluate metrics	Methods		
	CPseAAC	SAAC	GM
Absolute-false	0.0941	0.0447	0.0711
Accuracy	0.7881	0.9057	0.8513
Precision	0.8200	0.9164	0.8705
Recall	0.8193	0.9219	0.8763
Absolute-true	0.7267	0.8801	0.8090

Table 4 A comparison of the absolute-true success rates by CPseAAC for the multifunctional enzymes with different numbers of functional classes

Number of functional classes or labels	Number of multifunctional enzymes	Absolute-true success rate		
		CPseAAC (%)	Completely random guess (%)	Weighted random guess (%)
1	1302	999/1302 = 76.73	2.78	7.01
2	1647	1129/1647 = 68.55	1.11	3.55
3	138	115/138 = 83.33	0.83	0.22
4	8	6/8 = 75.00	1.11	0.015

Table 5 A comparison of the absolute-true success rates by SAAC for the multifunctional enzymes with different numbers of functional classes

Number of functional classes or labels	Number of multifunctional enzymes	Absolute-true success rate		
		SAAC (%)	Completely random guess (%)	Weighted random guess (%)
1	1302	1146/1302 = 88.02	2.78	7.01
2	1647	1440/1647 = 87.43	1.11	3.55
3	138	131/138 = 94.93	0.83	0.22
4	8	7/8 = 87.50	1.11	0.015

Table 6 A comparison of the absolute-true success rates by GM for the multifunctional enzymes with different numbers of functional classes

Number of functional classes or labels	Number of multifunctional enzymes	Absolute-true success rate		
		GM (%)	Completely random guess (%)	Weighted random guess (%)
1	1302	1061/1302 = 81.49	2.78	7.01
2	1647	1319/1647 = 80.09	1.11	3.55
3	138	117/138 = 84.78	0.83	0.22
4	8	7/8 = 87.50	1.11	0.015

corresponding success rates by the completely random guess and weighted random guess are also provided in Tables 4, 5, 6. The detailed information about the completely random guess and weighted random guess can be found in Lin et al. (2013b) and Xiao et al. (2013).

From Tables 4, 5, 6, we can see that (1) though the enzyme with multiple functional classes, its absolute true is still high, even the overall success rate by the worst solution in each dataset is overwhelmingly higher than the completely randomized rate and weighted randomized rate; (2) although the number of enzymes which belongs to four functional classes is few, the result is still promising, indicating that the method is powerful.

Conclusions

Prediction of the functional classes of multifunctional enzyme is a challenging and meaningful problem, particularly when the system concerned contains both singleplex and multiplex enzymes. In this paper, three different models were proposed to deal with multifunctional enzyme with single or multiple functional classes. The current approach represents a new strategy to handle the multi-label biological problems and hence may become a useful tool in the area of bioinformatics and proteomics (Wang and Li 2012).

As demonstrated in a series of recent publications (Chen et al. 2012; Ding et al. 2014; Jia et al. 2015; Liu et al. 2015a; Qiu et al. 2014; Xiao et al. 2015; Xu et al. 2013, 2014), user-friendly and publicly accessible web servers represent the further direction for developing practically more useful models, simulated methods, or predictors, and we shall make efforts in our future work to provide a web server for the method presented in this study.

References

Cai C, Han L, Ji Z, Chen Y (2004) Enzyme family classification by support vector machines. Proteins 55:66–76

Cao D-S, Xu Q-S, Liang Y-Z (2013) Propy: a tool to generate various modes of Chou's PseAAC. Bioinformatics 29:960–962

Chen C, Chen L, Zou X, Cai P (2009) Prediction of protein secondary structure content by using the concept of Chou's pseudo amino acid composition and support vector machine. Protein Pept Lett 16:27–31

Chen W, Lin H, Feng P-M, Ding C, Zuo Y-C, Chou K-C (2012) iNuc-PhysChem: a sequence-based predictor for identifying nucleosomes via physicochemical properties. PLoS ONE 7:e47843

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

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

Chou K-C (2011) Some remarks on protein attribute prediction and pseudo amino acid composition. J Theor Biol 273:236–247

Chou K-C, Cai Y-D (2004a) Using GO-PseAA predictor to predict enzyme sub-class. Biochem Biophys Res Commun 325:506–509

Chou KC, Cai YD (2004b) Predicting enzyme family class in a hybridization space. Protein Sci 13:2857–2863

Chou K-C, Elrod DW (1999) Protein subcellular location prediction. Protein Eng 12:107–118

Chou K-C, Elrod DW (2003) Prediction of enzyme family classes. J Proteome Res 2:183–190

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

Chou K-C, Shen H-B (2010a) Cell-PLoc 2.0: an improved package of web-servers for predicting subcellular localization of proteins in various organisms. Nat Sci 2:1090–1103

Chou K-C, Shen H-B (2010b) A new method for predicting the subcellular localization of eukaryotic proteins with both single and multiple sites: Euk-mPLoc 2.0. PLoS ONE 5:e9931

Chou K-C, Zhang C-T (1994) Predicting protein folding types by distance functions that make allowances for amino acid interactions. J Biol Chem 269:22014–22020

Ding H, Deng E-Z, Yuan L-F, Liu L, Lin H, Chen W, Chou K-C (2014) iCTX-Type: a sequence-based predictor for identifying the types of conotoxins in targeting ion channels. BioMed Res Int. doi:10.1155/2014/286419

Du P, Wang X, Xu C, Gao Y (2012) PseAAC-Builder: a cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions. Anal Biochem 425:117–119

Du P, Gu S, Jiao Y (2014) PseAAC-General: fast building various modes of general form of Chou's pseudo-amino acid composition for large-scale protein datasets. Int J Mol Sci 15:3495–3506

Esmaeili M, Mohabatkar H, Mohsenzadeh S (2010) Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses. J Theor Biol 263:203–209

Glory E, Murphy RF (2007) Automated subcellular location determination and high-throughput microscopy. Dev Cell 12:7–16

Hayat M, Khan A, Yeasin M (2012) Prediction of membrane proteins using split amino acid and ensemble classification. Amino Acids 42:2447–2460

Huang C, Yuan J-Q (2013a) A multilabel model based on Chou's pseudo–amino acid composition for identifying membrane proteins with both single and multiple functional types. J Membr Biol 246:327–334

Huang C, Yuan J-Q (2013b) Predicting protein subchloroplast locations with both single and multiple sites via three different modes of Chou's pseudo amino acid compositions. J Theor Biol 335:205–212

Huang C, Yuan J-Q (2015) Simultaneously identify three different attributes of proteins by fusing their three different modes of Chou's pseudo amino acid compositions. Protein Pept Lett 22:547–556

Huang Y, Niu B, Gao Y, Fu L, Li W (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics 26:680–682

Jia J, Liu Z, Xiao X, Liu B, Chou K-C (2015) iPPI-Esml: an ensemble classifier for identifying the interactions of proteins by incorporating their physicochemical properties and wavelet transforms into PseAAC. J Theor Biol 377:47–56

Khan ZU, Hayat M, Khan MA (2015) Discrimination of acidic and alkaline enzyme using Chou's pseudo amino acid composition in conjunction with probabilistic neural network model. J Theor Biol 365:197–203

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

Lin W-Z, Fang J-A, Xiao X, Chou K-C (2011) iDNA-Prot: identification of DNA binding proteins using random forest with grey model. PLoS ONE 6:e24756

Lin W-Z, Fang J-A, Xiao X, Chou K-C (2013a) iLoc-Animal: a multi-label learning classifier for predicting subcellular localization of animal proteins. Mol BioSyst 9(4):634–644

Lin W-Z, Fang J-A, Xiao X, Chou K-C (2013b) iLoc-Animal: a multi-label learning classifier for predicting subcellular localization of animal proteins. Mol BioSyst 9:634–644

Liu B, Fang L, Long R, Lan X, Chou K-C (2015a) iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition. Bioinformatics. doi:10.1093/bioinformatics/btv604

Liu B, Liu F, Wang X, Chen J, Fang L, Chou K-C (2015b) Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. Nucleic Acids Res 43:W65–W71

Mei S (2012) Multi-kernel transfer learning based on Chou's PseAAC formulation for protein submitochondria localization. J Theor Biol 293:121–130

Nakashima H, Nishikawa K, Tatsuo O (1986) The folding type of a protein is relevant to the amino acid composition. J Biochem 99:153–162

Qiu W-R, Xiao X, Chou K-C (2014) iRSpot-TNCPseAAC: identify recombination spots with trinucleotide composition and pseudo amino acid components. Int J Mol Sci 15:1746–1766

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

Shen H-B, Chou K-C (2008) PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition. Anal Biochem 373:386–388

Shen H-B, Chou K-C (2009) Gpos-mPLoc: a top-down approach to improve the quality of predicting subcellular localization of Gram-positive bacterial proteins. Protein Pept Lett 16: 1478–1484

Smith C (2008) Subcellular targeting of proteins and drugs. URL http://www.biocompare.com/Articles/TechnologySpotlight/976/Subcellular-Target-ing-Of-Proteins-An

Wang X, Li G-Z (2012) A multi-label predictor for identifying the subcellular locations of singleplex and multiplex eukaryotic proteins. PLoS ONE 7:e36317

Wang X, Zhang W, Zhang Q, Li G-Z (2015) MultiP-SChlo: multi-label protein subchloroplast localization prediction with Chou's pseudo amino acid composition and a novel multi-label classifier. Bioinformatics 31:2639–2645

Wu Z-C, Xiao X, Chou K-C (2012) iLoc-Gpos: a multi-layer classifier for predicting the subcellular localization of singleplex and multiplex gram-positive bacterial proteins. Protein Pept Lett 19:4–14

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

Xiao X, Wang P, Chou K-C (2012) inr-physchem: a sequence-based predictor for identifying nuclear receptors and their subfamilies via physical-chemical property matrix. PLoS ONE 7:e30869

Xiao X, Wang P, Lin W-Z, Jia J-H, Chou K-C (2013) iAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types. Anal Biochem 436:168–177

Xiao X, Min J-L, Lin W-Z, Liu Z, Cheng X, Chou K-C (2015) iDrug-Target: predicting the interactions between drug compounds and target proteins in cellular networking via benchmark dataset optimization approach. J Biomol Struct Dyn 33:2221–2233

Xu Y, Ding J, Wu L-Y, Chou K-C (2013) iSNO-PseAAC: predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition. PLoS ONE 8:e55844

Xu Y, Wen X, Wen L-S, Wu L-Y, Deng N-Y, Chou K-C (2014) iNitro-Tyr: Prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition. PLoS ONE 9:e105018

Zhang M-L, Zhou Z-H (2007) ML-KNN: a lazy learning approach to multi-label learning. Pattern Recogn 40:2038–2048

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

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

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

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

Zou H-L, Xiao X (2015) Predicting the functional types of singleplex and multiplex eukaryotic membrane proteins via different models of Chou's pseudo amino acid compositions. J Membr Biol. doi:10.1007/s00232-015-9830-9

Zou Q, Li X, Jiang Y, Zhao Y, Wang G (2013) BinMemPredict: a web server and software for predicting membrane protein types. Curr Proteomics 10:2–9