Amino Acids (2012) 43:545–555
DOI 10.1007/s00726-011-1143-4

ORIGINAL ARTICLE

Predicting protein submitochondria locations by combining different descriptors into the general form of Chou's pseudo amino acid composition

Guo-Liang Fan · Qian-Zhong Li

Received: 28 June 2011 / Accepted: 27 October 2011 / Published online: 20 November 2011

Abstract Knowledge of the submitochondria location of protein is integral to understanding its function and a necessity in the proteomics era. In this work, a new submitochondria data set is constructed, and an approach for predicting protein submitochondria locations is proposed by combining the amino acid composition, dipeptide composition, reduced physicochemical properties, gene ontology, evolutionary information, and pseudo-average chemical shift. The overall prediction accuracy is 93.57% for the submitochondria location and 97.79% for the three membrane protein types in the mitochondria inner membrane using the algorithm of the increment of diversity combined with the support vector machine. The performance of the pseudo-average chemical shift is excellent. For contrast, the method is also used to predict submitochondria locations in the data set constructed by Du and Li; an accuracy of 94.95% is obtained by our method, which is better than that of other existing methods.

Keywords Submitochondria location · Increment of diversity · Average chemical shift · Support vector machine · Chou's pseudo amino acid

Introduction

The mitochondrion is a semiautonomous, self-reproducing organelle that occurs in the cytoplasm of all cells of most, but not all, eukaryotes (Scharfe et al. 2000; Cotter and Guda 2004). Each mitochondrion is surrounded by a double limiting membrane: the inner membrane and outer membrane. Inner and outer mitochondrial membranes enclose two spaces: the matrix and intermembrane space. Mitochondria are the sites of the reactions of oxidative phosphorylation, which result in the formation of ATP. Proteins located in mitochondria play important roles in the energy metabolism process. Proteins in different submitochondria play distinctive roles in biological processes like triggering programmed cell death (Gottlieb 2000) and ionic homeostasis (Jassem and Heaton 2004). Therefore, knowing their submitochondria locations can provide useful hints to understand the protein functions and assist with drug design for many diseases related to mitochondria defects, ranging from rare monogenic to common age-related disorders (Alberts et al. 2002). However, experimental approaches for identifying the protein submitochondria locations are costly and time consuming. Therefore, it is becoming crucial to develop a reliable automatic submitochondria localizer for identifying protein subcompartment locations.

Recently, some computational methods for predicting protein submitochondria locations have been proposed in the literature: SUBmito (Du and Li 2006), Gp-Loc (Nanni and Lumini 2008), and Predict_subMITO (Zeng et al. 2009). Both of these methods used the data set constructed by Du and Li (2006) with 317 proteins and considered three submitochondria locations: the mitochondria inner membrane, mitochondria outer membrane, and mitochondria matrix. SUBmito considered the sequence-order information, and used the amino acid composition (AAC), dipeptide composition (DC) (Bhasin and Raghava 2004), and pseudo-amino acid composition (PseAAC) of nine physicochemical properties to construct the feature vector. Moreover, the protein was segmented into two segments. The predictive accuracy was 85.5% for the inner

G.-L. Fan · Q.-Z. Li (✉)
Department of Physics, School of Physical Science
and Technology, Inner Mongolia University,
Hohhot 010021, China
e-mail: qzli@imu.edu.cn

membrane, 94.5% for the matrix, and 51.2% for the outer membrane using the jackknife test. Gp-Loc enhanced the prediction accuracy of the matrix and outer membrane using genetic programming extracting 15 "artificial" features as its PseAAC, but the accuracy of the inner membrane was 83.21%, lower than that of SUBmito. Predict_subMITO used the auto covariance (AC) approach to transform numerical vectors of eight physicochemical properties of amino acids into uniform matrices and then used Chou's PseAAC to construct the vector. The overall jackknife cross-validation predictive accuracy was 89.3%.

In this article, we constructed the most up-to-date submitochondria data set, which has 1,105 proteins (denoted as M1105) derived from SWISS-PROT (Release 2010_12 of 30-Nov-2010) (Wu and Apweiler 2006), and then used the ID_SVM approach combined increment of diversity (ID) (Li and Lu 2001) with the support vector machine (SVM) (Chang and Lin 2011) by using many features to enhance the prediction performance for submitochondria locations. Some researchers have pointed out that proteins localized in the same subcellular location have similar amino acid compositions (AAC), which may reflect the physicochemical properties, because they are adapted to the micro environment (Andrade et al. 1998). However, the AAC lost the sequence order information; dipeptide composition, PseAAC, and other representational features were extracted to construct the substitution model of a protein for subcellular location (Cai and Chou 2000; Bhasin and Raghava 2004; Chou and Shen 2006a, b, 2010a, b, c; Chen and Li 2007a, b; Li and Li 2008b; Lee et al. 2008; Cai et al. 2010; Gu et al. 2010b; Wang and Geng 2011), and the features were used for submitochondria locations (Du and Li 2006; Nanni and Lumini 2008; Zeng et al. 2009). In this article, six representative features are used, including AAC, dipeptide composition (DC), evolutionary information (PSSM), gene ontology (GO) information, reduced physicochemical properties (Hn), and a novel constructed feature, pseudo-average chemical shift (PseACS). The DC and PseACS information was input to the ID, and then each feature was selected as an input to multiclass SVM. Here, the overall predictive accuracy was 93.57% for submitochondria locations and 97.79% for membrane protein types in our data set. In order to compare the prediction performance, we got an overall predictive accuracy of 94.95% for submitochondria locations of the data set (denoted as M317) constructed by Du and Li in jackknife tests.

According to a recent comprehensive review (Chou 2011), to establish a really useful statistical predictor for a protein system, we need to consider the following procedures: (1) construct or select a valid benchmark data set to train and test the predictor; (2) formulate the protein samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the attribute to be predicted; (3) introduce or develop a powerful algorithm to operate the prediction; (4) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor; (5) establish a user-friendly Web server for the predictor that is accessible to the public. Below, we describe how to deal with these steps.

Materials and methods

Data sets

We constructed the submitochondria data set derived from SWISS-PROT [Release 2010_12 of 30-Nov-2010 (Wu and Apweiler 2006)] by searching with '*KW*' containing 'mitochondrion,' and then the following steps were used to confine the quality data set. (1) The sequences that had any ambiguous annotation words such as '*probable*,' '*potential*,' '*possible*,' and '*by similarity*' were excluded. (2) The sequences containing ambiguous residues such as 'X,' 'B,' and 'Z' were removed. (3) The sequences that located more than one submitochondria location were also excluded. (4) The sequences annotated with '*fragment*' were excluded. (5) Sequences with a length less than 20 were dropped. (6) Proteins located at the inner membrane without a membrane protein type such as 'multi-pass membrane protein,' 'single-pass membrane protein,' or 'peripheral membrane protein' were also excluded. (7) To avoid homology bias and remove the redundant sequences from the benchmark data set, a cutoff threshold of 25% (Chou et al. 2011; Xiao et al. 2011a, b) was imposed to exclude those proteins from the benchmark data sets that have ≥25% sequence identity to any other in a same subset. However, in this study we did not use such a stringent criterion because the currently available data do not allow us to do so. Otherwise, the numbers of proteins for some subsets would have been too few to have statistical significance. We used the CD-HIT (Li et al. 2001) program to exclude the proteins with a sequence identity higher than 40%.

Finally, we obtained 1,164 sequences classified into four submitochondria locations, including 589 mitochondria inner membrane proteins, 59 mitochondria intermembrane space membrane proteins, 280 mitochondria matrix membrane proteins, and 236 mitochondria outer membrane proteins. Owing to the number of proteins located at the mitochondria intermembrane space is smaller than others, the data set used in this work contained three submitochondria locations. The final data set contained 1,105 sequences distributed in the three submitochondria locations listed in Table 1, denoted as M1105. For the 589 mitochondria inner membrane proteins, we divided them into three membrane protein types according to their

Table 1 The distribution of data set M1105

Label	Compartment	Membrane protein type	Sequence no.	
1	Inner membrane	Single-pass membrane protein	119	589
		Peripheral membrane protein	158	
		Multi-pass membrane protein	312	
2	Matrix membrane			280
3	Outer membrane			236
Total				1,105

annotation. The data sets are listed on our website (http://wlxy.imu.edu.cn/college/biostation/fuwu/mito/index.asp) and can be obtained from the author.

Feature vectors

To develop a powerful predictor for a protein system, one of the keys is to formulate the protein samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the attribute to be predicted (Chou 2011). To realize this, the concept of pseudo amino acid composition (PseAAC) was proposed (Chou 2001) to replace simple amino acid composition (AAC) for representing the protein sample. For a brief introduction to Chou's PseAAC, visit the Wikipedia Web page at http://en.wikipedia.org/wiki/Pseudo_amino_acid_composition. Ever since the concept of PseAAC was introduced, it has been widely used to study various problems in proteins and protein-related systems [see, e.g., (Chen et al. 2009; Ding et al. 2009; Esmaeili et al. 2010; Georgiou et al. 2009; Gu et al. 2010a; Jiang et al. 2008a, b; Li and Li 2008a; Lin 2008; Lin et al. 2008; Mohabatkar 2010; Mohabatkar et al. 2011; Qiu et al. 2010; Yu et al. 2010; Zeng et al. 2009; Zhang et al. 2008; Zhou et al. 2007)]. For various different modes of PseAAC, see Chou (2009). According to a recent comprehensive review (Chou 2011), the general form of Chou's pseudo amino acid composition (PseAAC) can be formulated as [see Eq. 6 of (Chou 2011)]:

$$P = [\psi_1, \psi_2, \ldots \psi_{\mathrm{u}} \cdots \psi_{\Omega}]^T \tag{1}$$

where T is a transpose operator, while the subscript Ω is an integer, and its value as well as the components ψ_1, ψ_2, ... will depend on how to extract the desired information from the amino acid sequence of P. Here, we used a combination of the amino acid composition, dipeptide composition, reduced physicochemical property, gene ontology, evolutionary information, and pseudo-average chemical shift to represent the protein samples.

Amino acid composition

The amino acid composition may represent the average physicochemical properties of the molecule. Therefore, we considered the amino acid composition. The sequence was divided into three segments. The absolute occurrence frequencies of 20 amino acids from each segment were calculated. Then these vectors from each segment were merged together. Thus, the feature vector of AAC can be expressed by 20 × 3 = 60D coordinates.

$$V_{i1} = \frac{1}{L_i}\left[n_{i1}, n_{12}, \ldots n_{ij}, \ldots, n_{i20}\right] \quad (i = 1, 2, 3;\ j = 1, 2, 3, \ldots 20) \tag{2}$$

where L_i is the length of the ith segmentation; n_{ij} is the jth residue occurrence frequencies in the ith segment.

Dipeptide composition

We used the DC of two consecutive residues to express the sequence order information. Like for AAC, the protein sequence was also divided into three segments; then, the feature vectors of DCs extracted from each segment were inputted into ID; finally, the dimension of DC was $3 \times 3 = 9$D, denoted by V_{i2} ($i = 1, 2, 3$). The performance of the ID algorithm was good, so that it reduced the dimension from 1200 to 9D and improved the accuracy from 79.5 to 84.2% for M317.

Reduced physicochemical properties

The amino acid composition of the sequence was correlated with the average physicochemical properties of the molecular surface. Therefore, we used 6 characters to represent the 20 amino acids according to the following physicochemical properties: strongly hydrophilic or polar (R, D, E, N, Q, K, H), strongly hydrophobic (L, I, V, A, M, F), weakly hydrophilic or weakly hydrophobic (S, T, Y, W), proline (P), glycine (G), and cysteine (C) (Chen and Li 2007a; Li and Li 2008a, b). The protein sequence was divided into eight segments; then, the composition of six characters for each segment was chosen, denoted by H_n ($n = 1, 2, \ldots, 8$). The dimension was $6 \times 8 = 48$D.

Gene ontology

Molecular function was correlated to the subcellular location, and the gene ontology (GO) (Ashburner et al. 2000) was one of the databases that describes the molecular function. Chou and Cai had used GO to predict the subcellular locations (Chou and Cai 2004, 2005; Fyshe et al. 2008; Huang et al. 2008; Chou and Shen 2010a, b, c).

According to the (GO) consortium, the GO database was established based on three criteria: (1) biological process, (2) molecular function, and (3) cellular component. Since the cellular component refers to the place in the cell, we only used the GO of biological process and molecular function.

We could get ‘GO_terms_and_ids’ from GO (http://www.geneontology.org/GO.Downloads.files.shtml), then map the GO numbers, which were used in the submitochondria data set in a 2103D vector orderly. For example, the GO numbers were: GO: 0000001, GO: 0000002, GO: 0000003, GO: 0000010, and GO: 0000023…. GO: 0080010 will map to $a_1, a_2, a_3, a_4, a_5 \ldots a_{2,103}$ separately.

$$p = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_i \\ \vdots \\ a_{2,103} \end{bmatrix} \tag{3}$$

$$\text{where, } a_i = \begin{cases} 1 & \text{hit GO number} \\ 0, & \text{otherwise} \end{cases}. \tag{4}$$

Then we statistically analyzed each coordinate of the vector and found that the sum of several coordinates was only one or two. This denoted that, for certain GOs, only a few proteins have them; then these GOs were eliminated, and the dimension of the feature vector decreased to 410D, denoted by G_{410}.

Evolutionary information

To use the evolution information, the position-specific scoring matrix (PSSM) (Schaffer et al. 2001) was generated by using the PSI-Blast program (Schaffer et al. 2001) to search the SWISS-PROT database (released on 14 May 2011) through three iterations with 0.001 as the *E*-value cutoff for multiple sequence alignment against the protein sequence *P*; then we used the standardization procedure to normalization.

$$V_{i\to j} = \frac{V^0_{i\to j} - \bar{V}^0_i}{\mathrm{SD}(V^0_i)} \quad (i = 1, 2, \ldots, L;\ j = 1, 2, \ldots, 20) \tag{5}$$

where $V^0_{i\to j}$ is the score directly obtained by PSI-Blast, $\bar{V}^0_i$ is the mean of $V^0_{i\to j}$ over 20 amino acids, $\mathrm{SD}(V^0_i)$ is the standard deviation of $V^0_{i\to j}$, and L is the length of the protein sequence. Then the PSSM becomes:

$$P_{\mathrm{PSSM}} = \begin{bmatrix} V_{1\to 1} & V_{1\to 2} & & V_{1\to 20} \\ V_{2\to 1} & V_{2\to 2} & & V_{2\to 20} \\ & & & \\ V_{i\to 1} & V_{i\to 2} & & V_{i\to 20} \\ & & & \\ V_{L\to 1} & V_{L\to 2} & \ldots & V_{L\to 20} \end{bmatrix} \tag{6}$$

In order to use the sequence order information, we adapted the concept of pseudo amino acid composition (Chou 2001) and obtained the PsePSSM by the following equations:

$$P^{\lambda}_{\mathrm{PsePSSM}} = \left[\theta^{\lambda}_1, \theta^{\lambda}_2, \ldots, \theta^{\lambda}_i, \ldots \theta^{\lambda}_{20}\right] \tag{7}$$

$$\theta^{\lambda}_i = \frac{1}{L-\lambda} \sum_{j=1}^{L-\lambda} \left[V_{j\to i} - V_{(j+\lambda)\to i}\right]^2 \quad (i = 1, 2 \ldots 20;\ \lambda < L) \tag{8}$$

where θ^{λ}_i is the correlation factor of amino acid type *i*, whose contiguous distance is λ along the protein sequence. Especially for $\lambda = 0$, θ^0_i becomes the average score of the amino acid residues in the protein *P*, which is changed to amino acid type *i* during the evolution process. The λ factor reflects the rank of correlation and is a non-negative integer; there is a best number for a certain problem. We calculated $\lambda = 0$–8 from Fig. 1, and we can see that the accuracy was best for the M317 and M1105 data set when $\lambda = 3$. Then the PsePSSM would be expressed as:

$$P_{\mathrm{PsePSSM}} = \left[\theta^0_1, \theta^0_2, \ldots, \theta^0_{20}, \theta^1_1, \theta^1_2, \ldots, \theta^1_{20}, \theta^2_1, \theta^2_2, \ldots, \theta^2_{20}, \theta^3_1, \theta^3_2, \ldots, \theta^3_{20}\right] \tag{9}$$

It becomes a 20 × 4 = 80D vector.

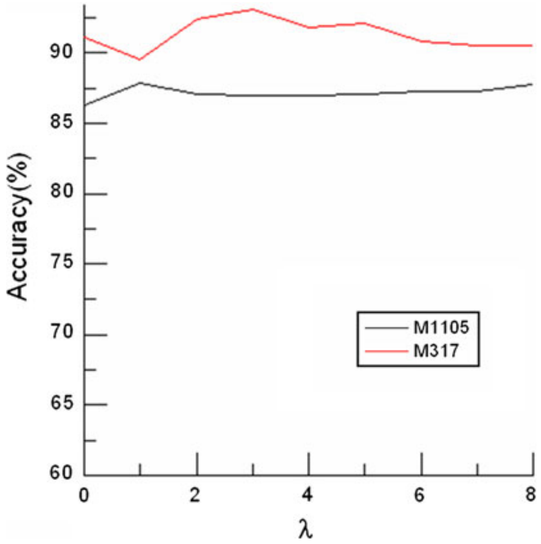


Fig. 1 The predictive accuracy varying with the λ for the PsePSSM descriptor

Pseudo-average chemical shift

Protons are sensitive to their chemical environment—an electron moving near them produces its own magnetic field, which changes the external field experienced by the proton. Protons in different chemical environments experience slightly different magnetic fields and absorb at different frequencies.

The resonance frequencies of the different protons are expressed as chemical shifts relative to a standard.

Chemical shifts, among the most important parameters, are measured by NMR spectroscopy. They are sensitive to local environments and can be used as indicators of local conformations. As an important example, the chemical shifts of protein backbone atoms are known to correlate strongly with the backbone dihedral angles or secondary structure types (Spera and Bax 1991; Wishart et al. 1991; Luginbuhl et al. 1995).

Several works have pointed out that the averaged chemical shift (ACS) of a particular nucleus in the protein backbone correlates well to its secondary structure (Sibley et al. 2003; Mielke and Krishnan 2003; Zhao et al. 2010), and the protein functions are determined by its structure.

Chemical shift values corresponding to the protein backbone atoms ^{15}N, ^{13}C$_\alpha$, ^{1}H$_\alpha$, and ^{1}H$_N$ were obtained from BMRB (http://www.bmrb.wisc.edu) (Seavey et al. 1991). By searching the BMRB QueryGrid Interface, the star files of proteins with 50 or more amino acid residues and matched with PDB (Berman et al. 2000) entries were considered and downloaded (see http://www.bmrb.wisc.edu/search/query_grid/query_1_2.html). In order to avoid redundancy and homology, we used the CD-HIT program to exclude the proteins with sequence identity higher than 40%. After the above steps, 1,552 proteins were selected. From the star file, we extracted chemical shift values of ^{15}N, ^{13}C$_\alpha$, ^{1}H$_\alpha$, and ^{1}H$_N$, four types of protein backbone atoms for every amino acid residue of protein *P*.

The averaged chemical shift of a protein backbone atom '*i*' for amino acid residue '*j*' with a second structure type '*k*' is defined as:

$$\mathrm{ACS}_i^k(j) = \frac{1}{N}\sum \omega_i^k(j) \qquad (10)$$

where $i = {}^{15}\mathrm{N}, {}^{13}\mathrm{C}_\alpha, {}^{1}\mathrm{H}_\alpha$, or $^{1}\mathrm{H}_N$, and *j* is the 20 native amino acid type; $k = $ H, E, and C, which express the three types of second structure. *N* is the total number of amino acid residues '*j*,' which has a secondary structure of '*k*' in 1,552 proteins. $\omega_i^k(j)$ is the chemical shift value of protein backbone atoms '*i*' for '*j*' kind of amino acid in '*k*' type of second structure.

We statistically computed all the amino acid residues of 1,552 proteins using Eq. 10, then found that each of the 20 native amino acid residues has different average chemical shifts and varies regularly with the secondary structure. Thus, the second structure of a protein can be represented by its ACS.

For a certain protein sequence *P*, we obtained the second structure from Porter (http://distill.ucd.ie/porter/) (Pollastri and McLysaght 2005; Pollastri et al. 2007), which is a server for predicting the protein's second structure. Every amino acid in the sequence is replaced by its ACS. Then *P* is expressed as:

$$P = [\mathrm{C}_1^i, \mathrm{C}_2^i \ldots \mathrm{C}_L^i]\ \left(i = {}^{15}\mathrm{N}, {}^{13}\mathrm{C}_\alpha, {}^{1}\mathrm{H}_\alpha, {}^{1}\mathrm{H}_N\right) \qquad (11)$$

Similar to PsePSSM, we selected $\lambda = 12$ and $i = {}^{1}\mathrm{H}_\alpha$, $^{1}\mathrm{H}_N$, then the PseACS would be expressed as:

$$P_{\mathrm{PseACS}} = [\varphi_1^0, \varphi_1^1, \ldots, \varphi_1^{12}, \varphi_2^0, \varphi_2^1, \ldots, \varphi_2^{12}] \qquad (12)$$

$$\varphi_i^\lambda = \frac{1}{L-\lambda}\sum_{k=1}^{L-\lambda}[\mathrm{C}_k^i - \mathrm{C}_{k+\lambda}^i]^2 (i = {}^{1}\mathrm{H}_\alpha, {}^{1}\mathrm{H}_\mathrm{N}; \lambda < L) \qquad (13)$$

then P_{PseACS} was inputted into ID according to the protein backbone atoms '*i*,' and the output of ID was selected as the parameter of PseACS, which was a 3 × 2 = 6D vector.

In order to better use the PseACS, we also established a user-friendly Web server, PseACS (http://wlxy.imu.edu.cn/college/biostation/fuwu/PseACS/index.asp), which is accessible to the public, and the $\omega_i^k(j)$ can be downloaded freely.

Methods

Increment of diversity

In a state space of *d* dimension, n_i indicates the absolute frequency of the *i*th state. The standard diversity measure for diversity source X:{n_1, n_2, …, n_i, …, n_d} is defined as (Li and Lu 2001):

$$D(X) = N\log N - \sum_{i=1}^{d} n_i \log_b n_i \qquad (14)$$

where $N = \sum_{i=1}^{d} n_i$, $\log(0) = 0$ if $n_i = 0$.

In general, for two sources of diversity in the same parameter space of *d* dimensions X:{n_1, n_2, …, n_i, …, n_d} and Y: {m_1, m_2, …, m_i, …, m_d}, the increment of diversity (ID), denoted by ID(*X*, *Y*), is defined as:

$$\mathrm{ID}(X, Y) = D(X + Y) - D(X) - D(Y) \qquad (15)$$

where $D(X + Y)$ is the measure of diversity of the sum of two diversity sources called the combination diversity source space.

ID is the method for measuring the similarity level of two diversity sources. If *X* is similar to *Y*, then ID(*X*, *Y*) will be small, especially if $X = Y$, ID(*X*, *Y*) = 0.

Support vector machine

SVM is a machine learning algorithm based on statistical learning theory (Vapnik 1998), which can be widely used for classification. In recent years, the SVM-based machine learning algorithm has also been used for predicting the membrane protein type (Cai et al. 2003b, 2004b), protein subcellular location (Chou and Cai 2002; Matsuda et al. 2005), protein structural class (Cai et al. 2002d; Ding et al. 2007), specificity of GalNAc-transferase (Cai et al. 2002c), HIV protease cleavage sites in protein (Cai et al. 2002b), β-turn types (Cai et al. 2002a), protein signal sequences and their cleavage sites (Cai et al. 2003a), α-turn types (Cai et al. 2003c), and catalytic triads of serine hydrolases (Cai et al. 2004a), among many others.

In this work, we used the free software LIBSVM (Chang and Lin 2011) to predict submitochondria locations. A radial basis function (RBF) was chosen as the kernel function. For multi-classification, SVM uses a one-versus-one strategy, and construct $k \times (k-1)/2$ classifiers and voting strategy were used to assign the class for an unknown protein.

Hybrid model

There are some methods for combining the feature vector to improve the accuracy of the prediction. The simple one is to concatenate all the feature vectors into a single vector, then input the integrated vector into the classifier for training and prediction (Park and Kanehisa 2003; Bhasin and Raghava 2004; Du and Li 2006; Chen and Li 2007b; Shi et al. 2007; Gao et al. 2010; Wang and Geng 2011). It usually occurs when these vectors have likely characters and the dimension is not large. Others utilize the multiple classifiers for fusing (Cai and Chou 2003; Chou and Cai 2003, 2004; Chou and Shen 2006a, b, 2008; Li and Li 2008b). In these methods, several classifiers are trained for each kind of feature vector, and then the output of these classifiers were combined together into a vector as the input of SVM, KNN, or neural networks, etc., which serve as a fusion classifier (Reinhardt and Hubbard 1998; Cai and Chou 2000; Cai et al. 2000; Nair and Rost 2003; Chou and Shen 2006a, b, 2010a, b, c; Lee et al. 2008; Cai et al. 2010; Gu et al. 2010b). For the first method, it is simple to use, but it cannot be used for large dimensional vectors; otherwise, it will be time consuming and risk over-training. For the fusion method, the accuracy will be improved, but the robustness will be weak and depends on the databases. Nanni et al. (2010) evaluated several feature extraction approaches for representing proteins starting from their amino acid sequence as well as different feature descriptor combinations using an ensemble of classifiers in which more than ten different protein descriptors are compared using nine different data sets. Results show that the best method is different for different data sets, and the combined approaches seem to be more robust.

In this work, because of different kinds of feature vectors and large dimensions, we must first reduce the dimension. For the feature vector of DC, its dimension is 1200D, and after using the ID algorithm, the dimension is reduced to 9D. The feature vector of GO is 2103D, so we deleted the seldom-used dimensions to realize the aim of reducing the dimension, and finally it was reduced to 410D. Finally, six parts of the feature vector were combined together to form a 613D feature vector as in Eq. 16 and inputted into the SVM for training to select the best c and g for a classifier of predicting submitochondria locations. If a protein was predicted to be an inner membrane protein, then it was sent into the classifier for predicting membrane types. To provide an intuitive picture, a flowchart showed the process of the classifier's work as given in Fig. 2.

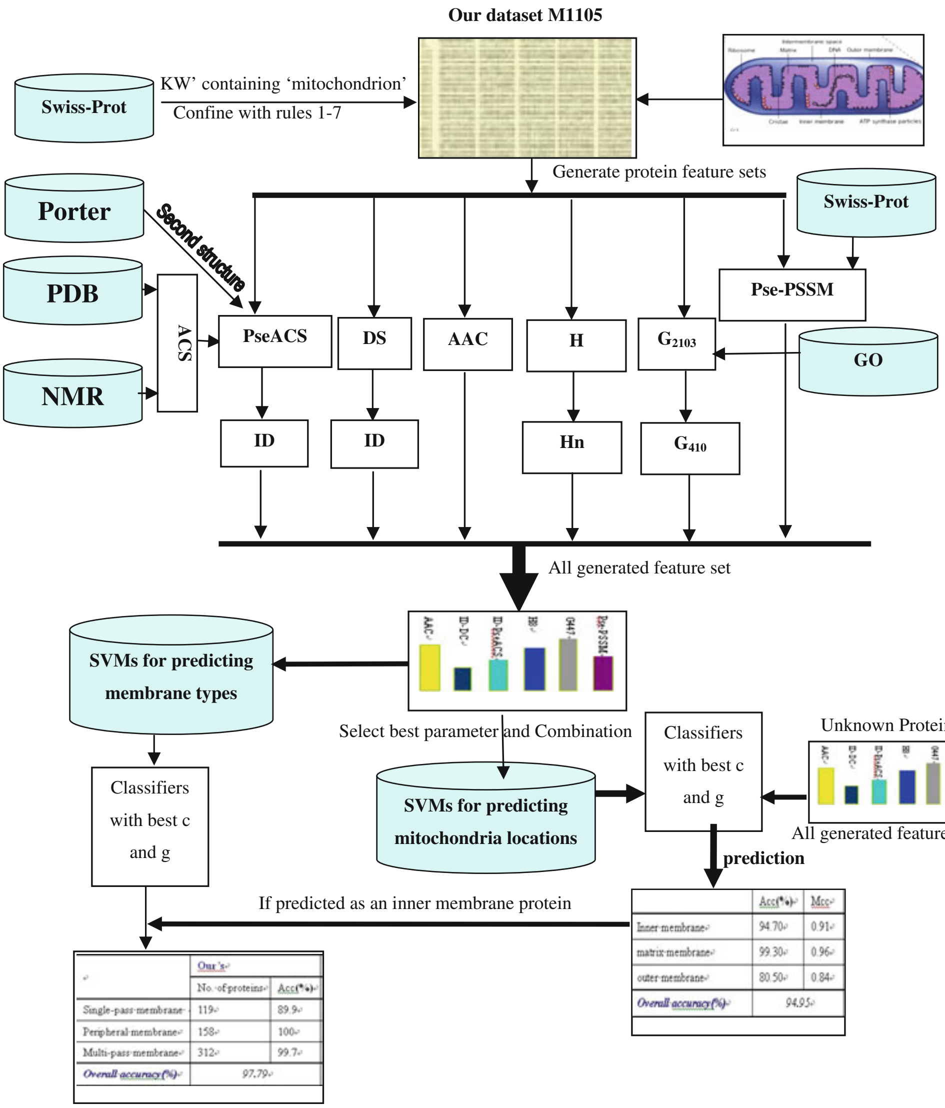
$$\vec{V} = [\vec{V}_{i1}, \vec{V}_{i2}, \vec{H}_n, \vec{G}_{410}, \vec{P}_{\mathrm{PsePSSM}}, \vec{P}_{\mathrm{PseACS}}]. \tag{16}$$

Results and discussion

Evaluation methods

In statistical prediction, the following three cross-validation methods are often used to examine a predictor for its effectiveness in practical application: independent data set test, sub-sampling test, and jack-knife test (Chou and Zhang 1995). However, as elucidated by Chou and Shen (2008) and demonstrated in Chou and Shen (2007), among the three cross-validation methods, the jackknife test is deemed the most objective one (Feng 2002) and can always yield a unique result for a given benchmark data set; hence, it has been increasingly used by investigators to examine the accuracy of various predictors (Zhou 1998; Zhou and Assa-Munt 2001; Zhou and Doctor 2003; Zhou et al. 2007; Jiang et al. 2008a, b; Li and Li 2008b; Lin 2008; Lin et al. 2008; Zhang and Fang 2008; Zhang et al. 2008; Bi et al. 2011; Ding et al. 2011; Hayat and Khan 2011; Hu et al. 2011; Joshi and Sekharan 2010; Kandaswamy et al. 2010; Kandaswamy et al. 2011; Lin and Ding 2011; Liu et al. 2010; Zakeri et al. 2011). During the jackknife test process, each protein is singled out in turn as a test sample; the remaining proteins are used as a training set to calculate the test sample's membership and predict the class.

The prediction performance was evaluated by the sensitivity (S_n), specificity (S_p) (Schaffer et al. 2001), positive predictive value (PPV), accuracy (Scharfe et al. 2000), and Mathew's correlation coefficient (MCC) (Matthews 1975), defined as follows:


Our dataset M1105
Swiss-Prot
KW' containing 'mitochondrion'
Confine with rules 1-7
Generate protein feature sets
Porter
Second structure
PDB
NMR
ACS
PseACS
DS
AAC
H
G_{2103}
Swiss-Prot
Pse-PSSM
GO
ID
ID
Hn
G_{410}
All generated feature set
AAC
ID-DC
ID-EbACS
HB
G447
Ex-PSSM
SVMs for predicting membrane types
Select best parameter and Combination
Classifiers with best c and g
Unknown Proteins
All generated feature set
SVMs for predicting mitochondria locations
Classifiers with best c and g
prediction
If predicted as an inner membrane protein

	Acc(%)	Mcc
Inner membrane	94.70	0.91
matrix membrane	99.30	0.96
outer membrane	80.50	0.84
Overall accuracy(%)	94.95	

	Our's	
	No. of proteins	Acc(%)
Single-pass membrane	119	89.9
Peripheral membrane	158	100
Multi-pass membrane	312	99.7
Overall accuracy(%)	97.79	

Fig. 2 Flowchart shows the construction of a classifier and how it works

$$S_n = TP/(TP + FN) \quad (17)$$

$$S_p = TN/(TN + FP) \quad (18)$$

$$PPV = TP/(TP + FP) \quad (19)$$

$$ACC = (TP + TN)/(TP + FN + TN + FP) \quad (20)$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FN) \times (TN + FN) \times (TP + FP) \times (TN + FP)}} \quad (21)$$

where TP denotes the numbers of the correctly predicted positives, FN denotes the numbers of the positives predicted as negatives, FP denotes the numbers of the negatives predicted as positives, and TN denotes the numbers of correctly predicted negatives.

Results of leave-one-out tests for M1105

Two kinds of SVMs are constructed for three submitochondria locations and three types of mitochondria inner membrane protein using six kinds of feature vectors, which reduced the dimension by the ID algorithm and cutoff value method. Since we chose the RBF as the kernel function, the grid-search approach was used to find the best parameters of C and γ for each SVM. The parameter optimizations are listed in Table 2. The prediction results for submitochondria locations of the data set M1105 are shown in Table 3, and the predictive accuracy for three membrane protein types is also shown in Table 4.

From Tables 3 and 4, we can see that the prediction performance is quite good; the total accuracy is 93.57% for submitochondria locations and 97.79% for three membrane

Table 2 The parameters of classifiers

Classifiers	*C*	*γ*
Submitochondria locations	8	0.03125
Three types of mitochondria inner membrane protein	2	0.03125

Table 3 The predictive accuracy for submitochondria locations in the data set M1105

Submitochondria locations	TP	TN	FP	FN	ACC (%)	MCC
Inner membrane	566	479	37	23	96.1	0.891
Matrix membrane	263	800	25	17	93.9	0.901
Outer membrane	205	860	9	31	86.9	0.890
Overall accuracy (%)	93.57					

Table 4 The predictive accuracy for three types of mitochondria inner membrane protein in the data set M1105

Membrane protein types	TP	TN	FP	FN	ACC (%)	MCC
Single-pass membrane	107	470	0	12	89.9	0.936
Peripheral membrane	158	427	4	0	100.0	0.983
Multi-pass membrane	311	268	9	1	99.7	0.966
Overall accuracy (%)	97.79					

protein types. For prediction of three membrane protein types in mitochondria inner membranes, 576 out of 589 were correctly predicted, and only 13 of them were predicted to be wrong protein types. For different membrane types, 107 out of 119 single-pass membranes and 311 out of 312 multi-pass membranes were correctly predicted; for 158 peripheral membranes, the predictive accuracy was 100%.

Comparison with other methods

In order to assess the performance of our predictor, we applied our method to the data set constructed by Du and Li. In Table 5, the results predicted by Submito, GP-Loc, and Predict_subMITO were compared with our method. Using the ID and SVM algorithm and combining several feature vectors, 94.95% accuracy was obtained in the jackknife test. It was 5.2% higher than the best predictor (Predict_subMITO). In Table 6, for the three types of mitochondrial inner membranes, the overall accuracy reached 97.79% when the leave-one-out (LOO) cross-validation was used. From the results, we can see that performance of our predictor is best because of its high accuracy and strong robustness. The contributions of each feature vector to submitochondria locations of M317 and M1105 are listed in Tables 7 and 8. The comparison of PseACS with AAC, DC, and PseAAC is also listed.

Table 5 Comparison of predictive accuracy for submitochondria locations of M317 with other methods

	Ours		Submito		GP-Loc		Predict_subMITO	
	ACC (%)	MCC	ACC (%)	MCC	ACC (%)	MCC	ACC (%)	MCC
Inner membrane	94.70	0.91	85.50	0.79	83.21	0.80	91.80	0.79
Matrix membrane	99.30	0.96	94.50	0.77	97.24	0.85	96.40	0.79
Outer membrane	80.50	0.84	51.20	0.64	78.05	0.77	66.10	0.63
Overall accuracy (%)	94.95		85.20		89.00		89.70	

Table 6 Comparison of predictive accuracy for three membrane protein types in the mitochondria inner membrane with other methods

	Ours		Submito		Predict_subMITO	
	No. of proteins	ACC (%)	No. of proteins	ACC (%)	No. of proteins	ACC (%)
Single-pass membrane	119	89.9	–	–	14	64.3
Peripheral membrane	158	100	–	–	–	–
Multi-pass membrane	312	99.7	101	83.2	127	98.4
Overall accuracy (%)	97.79		80.9		93.6	

Table 7 The contribution of each feature vector for submitochondria locations of M317 and the comparison of PseACS with AAC, DC, and PseACC

Feature vector	PSSM	GO	AAC	DC-ID	Pse-ACS-ID	Hn	PseAAC
Predictive accuracy (%)	93.1	93.3	82.3	84.2	85.5	76.3	84.5

Table 8 The contribution of each feature vector for submitochondria locations of M1105 and the comparison of PseACS with AAC, DC, and PseACC

Feature vector	PSSM	GO	AAC	DC-ID	Pse-ACS-ID	Hn	PseAAC
Predictive accuracy (%)	87.9	85.9	74.5	73.2	76.3	65.5	75.9

Conclusion

In this work, a benchmark data set was constructed that has about 3.5 times more proteins than the data set constructed by Du and Li. The data set presented here contains more information on proteins located in the mitochondria and can be used to perform much detailed research about submitochondria, such as submitochondria locations, the protein function of different submitochondria locations, and the interaction of submitochondria protein, etc.

The various features of submitochondrial protein are considered, and an ID algorithm and dimension reduced methods were used to construct the classifier. By using this method, we obtained 93.57% with the jackknife test of our data set and 94.95% with the data set of Du and Li, which is better than the best approach in the literature.

Among the feature vectors, a novel constructed feature PseACS is proposed. From Table 7, we can see that the performance of PseACS is also excellent, and the predictive accuracy of the submitochondria locations reaches 85.5% when using only the PseACS feature. The result was better than that of Du and Li, which used Chou's PseAAC. Therefore, PseACS can be an effective tool for future proteomic studies.

Since user-friendly and publicly accessible Web servers represent the future direction for developing more practically useful models, simulated methods, or predictors (Chou and Shen 2009), we will make efforts in future work to develop a Web server for the method presented in this article.

Acknowledgments The authors would like to thank the reviewers for their helpful comments on our manuscript. This work was supported by a grant from the National Natural Science Foundation of China (61063016, 31160188), The Research Fund for the Doctoral Program of Higher Education of China (no.20101501110004), The Project for '211' Innovative Talents of Inner Mongolia University (no. 2-1.2.1_035), and the Inner Mongolia University Fund for Young Scholars (no. 208152).

Conflict of interest The authors declare that they have no conflict of interest.

References

Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P (2002) Molecular biology of the cell, 4th edn. Garland, New York

Andrade MA, O'Donoghue SI, Rost B (1998) Adaption of protein surface to subcellular location. J Mol Biol 276:517–525

Ashburner M, Ball CA et al (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25–29

Berman HM, Westbrook J et al (2000) The protein data bank. Nucleic Acids Res 28:235–242

Bhasin M, Raghava GP (2004) ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. Nucleic Acids Res 32:W414–W419 (Web Server issue)

Bi J, Yang H, Yan H, Song R, Fan J (2011) Knowledge-based virtual screening of HLA-A*0201-restricted CD8(+) T-cell epitope peptides from herpes simplex virus genome. J Theor Biol 281:133–139

Cai YD, Chou KC (2000) Using neural networks for prediction of subcellular location of prokaryotic and eukaryotic proteins. Mol Cell Biol Res Commun 4:172–173

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 et al (2000) Support vector machines for prediction of protein subcellular location. Mol Cell Biol Res Commun 4:230–233

Cai YD, Liu XJ et al (2002a) Support vector machines for the classification and prediction of β-turn types. J Pept Sci 8:297–301

Cai YD, Liu XJ, Xu XB, Chou KC (2002b) Support vector machines for predicting HIV protease cleavage sites in protein. J Comput Chem 23:267–274

Cai YD, Liu XJ, Xu XB, Chou KC (2002c) Support vector machines for predicting the specificity of GalNAc-transferase. Peptides 23:205–208

Cai YD, Liu XJ et al (2002d) Prediction of protein structural classes by support vector machines. Comput Chem 26:293–296

Cai YD, Lin S, Chou KC (2003a) Support vector machines for prediction of protein signal sequences and their cleavage sites. Peptides 24:159–161

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

Cai YD, Feng KY, Li YX, Chou KC (2003c) Support vector machine for predicting α-turn types. Peptides 24:629–630

Cai YD, Zhou GP, Jen CH, Lin SL, Chou KC (2004a) Identify catalytic triads of serine hydrolases by support vector machines. J Theor Biol 228:551–557

Cai YD, Pong-Wong R, Feng K, Jen JCH, Chou KC (2004b) Application of SVM to predict membrane protein types. J Theor Biol 226:373–376

Cai YD, Ricardo PW et al (2004c) Application of SVM to predict membrane protein types. J Theor Biol 226:373–376

Cai YD, Lu L et al (2010) Predicting subcellular location of proteins using integrated-algorithm method. Mol Divers 14:551–558

Chang CC, Lin CJ (2011) LIBSVM: a library for support vector machines. ACM Transact Intell Syst Technol 2:27:1–27:27. doi: 10.1145/1961189.1961199. http://www.csie.ntu.edu.tw/~cjlin/libsvm

Chen YL, Li QZ (2007a) Prediction of apoptosis protein subcellular location using improved hybrid approach and pseudo-amino acid composition. J Theor Biol 248:377–381

Chen YL, Li QZ (2007b) Prediction of the subcellular location of apoptosis proteins. J Theor Biol 245:775–783

Chen C, 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

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

Chou KC (2009) Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. Curr Proteomics 6:262–274

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

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

Chou KC, Cai YD (2003) A new hybrid approach to predict subcellular localization of proteins by incorporating gene ontology. Biochem Biophys Res Commun 311:743–747

Chou KC, Cai YD (2004) Prediction of protein subcellular locations by GO-FunD-PseAA predictor. Biochem Biophys Res Commun 320:1236–1239

Chou KC, Cai YD (2005) Using GO-PseAA predictor to identify membrane proteins and their types. Biochem Biophys Res Commun 327:845–847

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 (2007) Recent progress 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, Shen HB (2009) Review: recent advances in developing web-servers for predicting protein attributes. Nat Sci 2:63–92 (openly accessible at http://www.scirp.org/journal/NS/)

Chou KC, Shen HB (2010a) Cell-PLoc2.: a improved package of Web servers for predicting subcellular localization of proteins in various organisms. Nat Sci 2:1090–1103

Chou KC, Shen HB (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 KC, Shen HB (2010c) Plant-mPLoc: a top-down strategy to augment the power for predicting plant protein subcellular localization. PLoS One 5:e11335

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

Chou KC, Wu ZC, Xiao X (2011) iLoc-Euk: a multi-label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins. PLoS One 6:e18258 (50th Anniversary Year Review)

Cotter D, Guda P et al (2004) MitoProteome: mitochondrial protein sequence database and annotation system. Nucleic Acids Res 32:D463–D467 (Database issue)

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

Ding H, Luo L, Lin H (2009) Prediction of cell wall lytic enzymes using Chou's amphiphilic pseudo amino acid composition. Protein Pept Lett 16:351–355

Ding H, Liu L, Guo FB, Huang J, Lin H (2011) Identify Golgi protein types with modified mahalanobis discriminant algorithm and pseudo amino acid composition. Protein Pept Lett 18:58–63

Du P, Li YD (2006) Prediction of protein submitochondria locations by hybridizing pseudo-amino acid composition with various physicochemical features of segmented sequence. BMC Bioinforma 7:518–525

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

Feng ZP (2002) An overview on predicting the subcellular location of a protein. In Silico Biol 2:291–303

Fyshe A, Liu Y et al (2008) Improving subcellular localization prediction using text classification and the gene ontology. Bioinformatics 24:2512–2517

Gao QB, Ye XF et al (2010) Improving discrimination of outer membrane proteins by fusing different forms of pseudo amino acid composition. Anal Biochem 398:52–59

Georgiou DN, Karakasidis TE, Nieto JJ, Torres A (2009) Use of fuzzy clustering technique and matrices to classify amino acids and its impact to Chou's pseudo amino acid composition. J Theor Biol 257:17–26

Gottlieb RA (2000) Programmed cell death. Drug News Perspect 13:471–476

Gu Q, Ding YS, Zhang TL (2010a) Prediction of G-protein-coupled receptor classes in low homology using chou's pseudo amino acid composition with approximate entropy and hydrophobicity patterns. Protein Pept Lett 17:559–567

Gu Q, Ding YS et al (2010b) Prediction of subcellular location apoptosis proteins with ensemble classifier and feature selection. Amino Acids 38:975–983

Hayat M, Khan A (2011) Predicting membrane protein types by fusing composite protein sequence features into pseudo amino acid composition. J Theor Biol 271:10–17

Hu L, Zheng L, Wang Z, Li B, Liu L (2011) Using pseudo amino acid composition to predict protease families by incorporating a series of protein biological features. Protein Pept Lett 18:552–558

Huang WL, Tung CW et al (2008) ProLoc-GO: utilizing informative Gene Ontology terms for sequence-based prediction of protein subcellular localization. BMC Bioinforma 9:80

Jassem W, Heaton ND (2004) The role of mitochondria in ischemia/reperfusion injury in organ transplantation. Kidney Int 66:514–517

Jiang X, Wei R, Zhang TL, Gu Q (2008a) Using the concept of Chou's pseudo amino acid composition to predict apoptosis proteins subcellular location: an approach by approximate entropy. Protein Pept Lett 15:392–396

Jiang X, Wei R et al (2008b) Using Chou's pseudo amino acid composition based on approximate entropy and an ensemble of AdaBoost classifiers to predict protein subnuclear location. Amino Acids 34:669–675

Joshi RR, Sekharan S (2010) Characteristic peptides of protein secondary structural motifs. Protein Pept Lett 17:1198–1206

Kandaswamy KK, Pugalenthi G, Moller S, Hartmann E, Kalies KU, Suganthan PN, Martinetz T (2010) Prediction of apoptosis protein locations with genetic algorithms and support vector machines through a new mode of pseudo amino acid composition. Protein Pept Lett 17:1473–1479

Kandaswamy KK, Chou KC, Martinetz T, Moller S, Suganthan PN, Sridharan S, Pugalenthi G (2011) AFP-Pred: a random forest approach for predicting antifreeze proteins from sequence-derived properties. J Theor Biol 270:56–62

Lee K, Chuang HY et al (2008) Protein networks markedly improve prediction of subcellular localization in multiple eukaryotic species. Nucleic Acids Res 36:e136

Li FM, Li QZ (2008a) Predicting protein subcellular location using Chou's pseudo amino acid composition and improved hybrid approach. Protein Pept Lett 15:612–616

Li FM, Li QZ (2008b) Using pseudo amino acid composition to predict protein subnuclear location with improved hybrid approach. Amino Acids 34:119–125

Li QZ, Lu ZQ (2001) The prediction of the structural class of protein: application of the measure of diversity. J Theor Biol 213:493–502

Li W, Jaroszewski L et al (2001) Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinformatics 17:282–283

Lin H (2008) The modified Mahalanobis discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition. J Theor Biol 252:350–356

Lin H, Ding H (2011) Predicting ion channels and their types by the dipeptide mode of pseudo amino acid composition. J Theor Biol 269:64–69

Lin H, Ding H et al (2008) Predicting subcellular localization of mycobacterial proteins by using Chou's pseudo amino acid composition. Protein Pept Lett 15:739–744

Liu T, Zheng X, Wang C, Wang J (2010) Prediction of subcellular location of apoptosis proteins using pseudo amino acid composition: an approach from auto covariance transformation. Protein Pept Lett 17:1263–1269

Luginbuhl P, Szyperski T, Wuthrich K (1995) Statistical basis for the use of ^{13}C a chemical shifts in protein structure determination. J Magn Reson B 109:229–233

Matsuda S, Vert JP et al (2005) A novel representation of protein sequences for prediction of subcellular location using support vector machines. Protein Sci 14:2804–2813

Matthews BW (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 405:442–451

Mielke SP, Krishnan VV (2003) Protein structural class identification directly from NMR spectra using averaged chemical shifts. Bioinformatics 19:2054–2064

Mohabatkar H (2010) Prediction of cyclin proteins using Chou's pseudo amino acid composition. Protein Pept Lett 17:1207–1214

Mohabatkar H, Beigi MM, Esmaeili A (2011) Prediction of GABA (A) receptor proteins using the concept of Chou's pseudo-amino acid composition and support vector machine. J Theor Biol 281:18–23

Nair R, Rost B (2003) Better prediction of sub-cellular localization by combining evolutionary and structural information. Proteins 53:917–930

Nanni L, Lumini A (2008) Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization. Amino Acids 34:653–660

Nanni L, Brahnam S, Lumini A (2010) High performance set of PseAAC and sequence based descriptors for protein classification. J Theor Biol 266:1–10

Park KJ, Kanehisa M (2003) Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. Bioinformatics 19:1656–1663

Pollastri G, McLysaght A (2005) Porter: a new, accurate server for protein secondary structure prediction. Bioinformatics 21:1719–1720

Pollastri G, Martin AJ et al (2007) Accurate prediction of protein secondary structure and solvent accessibility by consensus combiners of sequence and structure information. BMC Bioinforma 8:201

Qiu JD, Huang JH, Shi SP, Liang RP (2010) Using the concept of Chou's pseudo amino acid composition to predict enzyme family classes: an approach with support vector machine based on discrete wavelet transform. Protein Pept Lett 17:715–722

Reinhardt A, Hubbard T (1998) Using neural networks for prediction of the subcellular location of proteins. Nucleic Acids Res 26:2230–2236

Schaffer AA, Aravind L et al (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res 29:2994–3005

Scharfe C, Zaccaria P et al (2000) MITOP, the mitochondrial proteome database: 2000 update. Nucleic Acids Res 28:155–158

Seavey BR, Farr EA et al (1991) A relational database for sequence-specific protein NMR data. J Biomol NMR 1:217–236

Shi JY, Zhang SW et al (2007) Prediction of protein subcellular localization by support vector machines using multi-scale energy and pseudo amino acid composition. Amino Acids 33:69–74

Sibley AB, Cosman M, Krishnan VV (2003) An empirical correlation between secondary structure content and averaged chemical shifts in proteins. Biophys J 84(2):1223–1227

Spera S, Bax A (1991) Empirical correlation between protein backbone conformation and C_α and C_β ^{13}C nuclear magnetic resonance chemical shifts. J Am Chem Soc 113:5490–5492

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

Wang W, Geng XB et al (2011) Predicting protein subcellular localization by pseudo amino acid composition with a segment-weighted and features-combined approach. Protein Pept Lett (e-pub ahead of print)

Wishart DS, Sykes BD, Richards FM (1991) Relationship between nuclear magnetic resonance chemical shift and protein secondary structure. J Mol Biol 222:311–333

Wu CH, Apweiler R et al (2006) The universal protein resource (UniProt): an expanding universe of protein information. Nucleic Acids Res 34:D187–D191 (Database issue)

Xiao X, Wu ZC, Chou KC (2011a) A multi-label classifier for predicting the subcellular localization of gram-negative bacterial proteins with both single and multiple sites. PLoS One 6:e20592

Xiao X, Wu ZC, Chou KC (2011b) iLoc-Virus: a multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites. J Theor Biol 284:42–51

Yu L, Guo Y, Li Y, Li G, Li M, Luo J, Xiong W, Qin W (2010) SecretP: identifying bacterial secreted proteins by fusing new features into Chou's pseudo-amino acid composition. J Theor Biol 267:1–6

Zakeri P, Moshiri B, Sadeghi M (2011) Prediction of protein submitochondria locations based on data fusion of various features of sequences. J Theor Biol 269:208–216

Zeng YH, Guo YZ et al (2009) Using the augmented Chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach. J Theor Biol 259:366–372

Zhang GY, Fang BS (2008) Predicting the cofactors of oxidoreductases based on amino acid composition distribution and Chou's amphiphilic pseudo-amino acid composition. J Theor Biol 253:310–315

Zhang GY, Li HC et al (2008) Predicting lipase types by improved Chou's pseudo-amino acid composition. Protein Pept Lett 15:1132–1137

Zhao Y, Alipanahi B et al (2010) Protein secondary structure prediction using NMR chemical shift data. J Bioinform Comput Biol 8:867–884

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

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

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

Zhou XB, Chen C et al (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