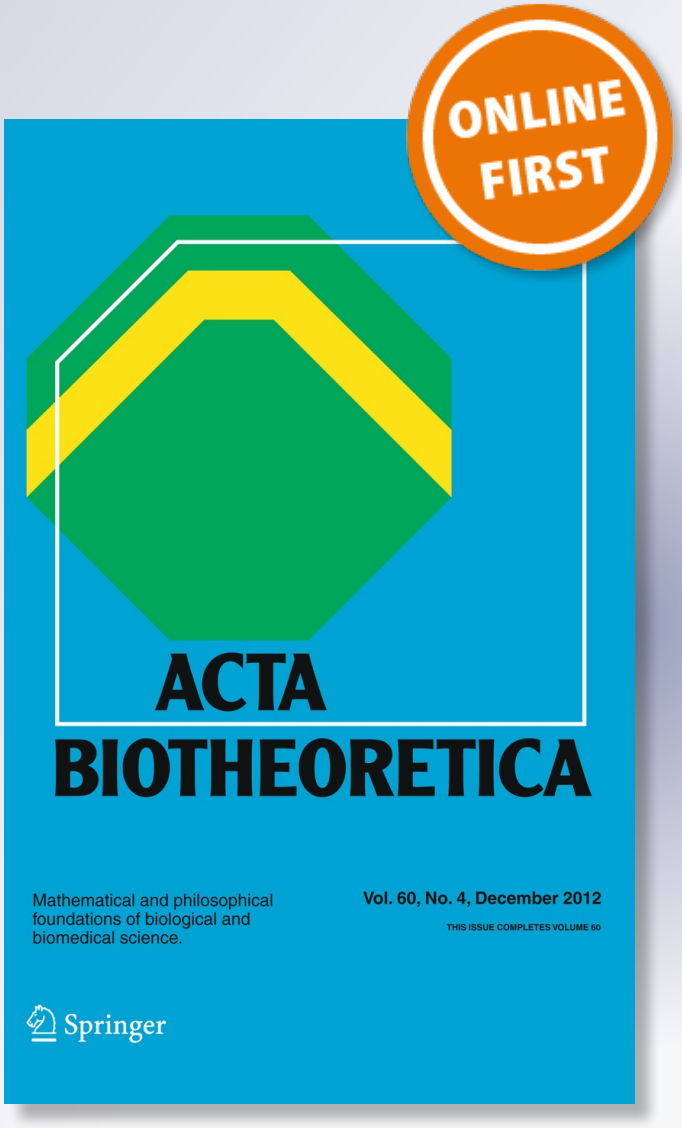
Using Over-Represented Tetrapeptides to Predict Protein Submitochondria Locations

Hao Lin, Wei Chen, Lu-Feng Yuan, Zi-Qiang Li & Hui Ding

Acta Biotheoretica
Mathematical and philosophical foundations of biological and biomedical science

ISSN 0001-5342

Acta Biotheor
DOI 10.1007/s10441-013-9181-9



Springer

Acta Biotheor
DOI 10.1007/s10441-013-9181-9

REGULAR ARTICLE

Using Over-Represented Tetrapeptides to Predict Protein Submitochondria Locations

Hao Lin · Wei Chen · Lu-Feng Yuan · Zi-Qiang Li · Hui Ding

Received: 3 July 2012 / Accepted: 23 February 2013

Abstract The mitochondrion is a key organelle of eukaryotic cell that provides the energy for cellular activities. Correctly identifying submitochondria locations of proteins can provide plentiful information for understanding their functions. However, using web-experimental methods to recognize submitochondria locations of proteins are time-consuming and costly. Thus, it is highly desired to develop a bioinformatics method to predict the submitochondria locations of mitochondrion proteins. In this work, a novel method based on support vector machine was developed to predict the submitochondria locations of mitochondrion proteins by using over-represented tetrapeptides selected by using binomial distribution. A reliable and rigorous benchmark dataset including 495 mitochondrion proteins with sequence identity $\leq$25 % was constructed for testing and evaluating the proposed model. Jackknife cross-validated results showed that the 91.1 % of the 495 mitochondrion proteins can be correctly predicted. Subsequently, our model was estimated by three existing benchmark datasets. The overall accuracies are 94.0, 94.7 and 93.4 %, respectively, suggesting that the proposed model is potentially useful in

H. Lin (✉) · L.-F. Yuan · H. Ding
Key Laboratory for NeuroInformation of Ministry of Education, Center of Bioinformatics, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, China
e-mail: hlin@uestc.edu.cn

W. Chen (✉)
Department of Physics, College of Sciences, Center for Genomics and Computational Biology, Hebei United University, Tangshan 063000, China
e-mail: chenwei_imu@yahoo.com.cn

Z.-Q. Li
School of Information and Engineering, Sichuan Agricultural University, Yaan 625014, China

Z.-Q. Li
School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China

Published online: 10 March 2013

the realm of mitochondrion proteome research. Based on this model, we built a predictor called TetraMito which is freely available at http://lin.uestc.edu.cn/server/TetraMito.

Keywords Submitochondria location · Tetrapeptide · Binomial distribution · Support vector machine

1 Introduction

Mitochondrion is one of the key organelles in eukaryotic cells, which provides the energy a cell needs to move, divide and produce secretory products etc. (Henze and Martin 2003). The mitochondrion mainly contains four distinct compartments i.e. the outer membrane, intermembrane space, inner membrane and the matrix. The outer membrane has large numbers of integral membrane proteins forming channels to allow small molecules and proteins with signal peptides diffusing into or out of mitochondrion. The intermembrane space is the space between the outer membrane and the inner membrane. The inner membrane contains many different polypeptides which are responsible for regulating metabolite and performing the redox reactions of oxidative phosphorylation. The matrix is the space enclosed by the inner membrane. The main function of the matrix is to produce ATP with the aid of the ATP synthase contained in the inner membrane.

The proteins located in these four submitochondria locations play distinct biological roles. For timely understanding protein functions, it needs to accurately identify the submitochondria location of mitochondrion proteins. Unfortunately, it is cost ineffective for experimental approach to confirm protein's location in mitochondrion. Phylogenetic tree is a traditional method for most experimental scholars to predict the sub-subcellular locations of proteins. Although this method is not particularly expensive, it is more time consuming than machine learning approaches. Furthermore, for the sequences which do not have homologous sequences in benchmark dataset, phylogenetic tree will provide ineffective, inexact and even wrong information. Therefore, it is a good choice to develop machine learning methods to predict the submitochondria location of mitochondrion proteins.

The protein sub-subcellular location prediction is a hot research area and has been studied by some scholars as reviewed in recent literature (Du et al. 2011). Due to the plenty of proteins in nuclear location, the prediction of protein subnuclear location is widely studied. Shen and Chou (2005, 2007) developed a method based on pseudo amino acid composition (PseAAC) to predict subnuclear location of proteins and achieved an overall accuracy of $\sim$65 %. Subsequently, other works improved the accuracy to $\sim$85 % by developing various methods (Lei and Dai 2005, 2006; Huang et al. 2007, 2008, 2009; Li and Li 2008; Jiang et al. 2008; Mei and Fei 2010). For the prediction of protein subchloroplast location, Du et al. (2009) developed a server, called SubChlo to predict subchloroplast locations of proteins. This server obtained an overall accuracy of 67.18 % in jackknife test. For improving the predictive accuracy, another work (Shi et al. 2011) used discrete wavelet transform to exact feature and achieved the accuracy of 89.31 %. The sub-Golgi

location of protein has also been attended by some researchers. van Dijk et al. (2008) have focused on the prediction of the sub-Golgi locations of type II membrane proteins. Recently, *cis*-Golgi and *trans*-Golgi proteins were studied (Ding et al. 2011) by using modified Mahalanobis discriminant.

Due to the special function of mitochondrion, the submitochondria localization of proteins has attracted many bioinformatics scholars. Du and Li (2006) presented a PseAAC-based method to predict protein submitochondria locations. Total of 317 mitochondria proteins with sequence identity less than 40 % were constructed and the overall accuracy achieved 85.2 % in jackknife cross-validation. Subsequently, a genetic programming-based method was developed by Nanni and Lumini (2008) to predict protein submitochondria locations. The jackknife test accuracy increased to 89 %. Zakeri et al. (2011) increased the overall accuracy to 94.7 % by using feature fusion. Shi et al. (2011) proposed to use discrete wavelet transform to extract features and achieved the overall accuracy of 93.38 % in jackknife cross-validation. Mei (2012) used GO information to improve the accuracy of submitochondria location prediction of proteins. Another dataset including 399 mitochondria proteins with sequence identity less than 40 % was constructed by Zeng et al. (2009). By using augmented PseAAC as parameters, the overall accuracy achieved 89.7 % in jackknife cross-validation. Fan and Li (2012) constructed the third dataset containing 1,105 mitochondria proteins with sequence identity less than 40 %. By using the pseudo-average chemical shift and GO information, they obtained an overall accuracy of 93.57 %. Although these models have their respective merits, all of them have a common limit: the success rate is very low when the query protein has less than 25 % sequence identity to proteins with known locations (Chou and Shen 2007, 2008).

In this study, we constructed a very stringent benchmark dataset in which none of the proteins have $\geq$25 % sequence identity with any other proteins with the same submitochondria location. The binomial distribution was used to optimize the tetrapeptide words. The support vector machine (SVM) was proposed to perform prediction. Results of jackknife test showed that the overall accuracy of 91.1 % was achieved with the average accuracy of 88.0 %. Our method was also examined on other three benchmark datasets and achieved the accuracies of 94.0, 94.7 and 93.4 %, respectively. These results show that the proposed method can be efficiently and accurately used to annotate submitochondria locations for new mitochondria proteins.

2 Materials and Methods

2.1 Dataset

The mitochondria proteins were extracted from Universal Protein Resource (Uniprot) (UniProt Consortium 2012). To construct a reliable benchmark dataset, the following steps were used to prepare high quality datasets. (1) Although proteins with multiple submitochondria locations have some special biological functions, we collected the proteins with only one mitochondria location because the number of

proteins with multiple submitochondria locations is too small to have statistical signification. (2) Proteins with ambiguous protein existence annotations, such as 'uncertain', 'predicted' and 'inferred from homology' were excluded because they lack confidence. (3) Only those proteins with experimental confirmed submitochondria location were included because they can provide validate information. (4) Sequences which are fragment of other proteins were excluded because their information is redundant and not integrity. (5) Sequences containing nonstandard letters, such as 'B', 'X' or 'Z', were excluded because their meanings are ambiguous. (6) To avoid any homology bias, the protein sequence with $\geq$25 % sequence identity to any other proteins in the same subset was excluded by using PISCES (Wang and Dunbrack 2005). After strictly following the above procedures, we finally obtained 495 mitochondria proteins called M495, which includes 254 inner membrane proteins, 132 matrix proteins and 109 outer membrane proteins. The data can be freely downloaded from http://lin.uestc.edu.cn/server/subMito/data. The method that divides benchmark dataset into training set and test set is strict and objective for evaluating the performance of proposed method. However, in this study we did not use such method because the currently available data do not allow us to do so. Otherwise, the number of proteins for some subsets would be too few to have statistical significance.

To facilitate comparison with previous studies, three benchmark datasets were used. The dataset M317 constructed by Du and Li (2006) contains 131 inner membrane proteins, 145 matrix proteins and 41 outer membrane proteins. The second dataset M399 constructed by Zeng et al. (2009) contains 171 inner membrane proteins, 166 matrix proteins and 62 outer membrane proteins. The third dataset M1105 constructed by Fan and Li (2012) contains 589 inner membrane proteins, 280 matrix proteins and 236 outer membrane proteins. Each of the three benchmark datasets has the sequence identity of <40 %.

2.2 Tetrapeptide Words

Informative parameters play a key role in machine learning problem. Secondary structure information of protein has been widely used in protein structure and function prediction. It has been proved that the predicted secondary structure information of protein can be used for predicting submitochondria location of proteins by calculating average chemical shift (Fan and Li 2012). Rackovsky (1993) has estimated that 60–70 % of tetrapeptides encode the specific structure. Feng and Luo (2008) have used tetrapeptide signals to predict secondary structure of proteins. Thus, in this study, tetrapeptide words were utilized to represent the sample of mitochondria proteins. The following processes were performed to define the tetrapeptide words.

Firstly, by sliding a window of four residues with step of one residue along mitochondria protein sequences, we calculated the occurrence frequency (n_{ij}) of the i-th tetrapeptide in the j-th submitochondria location, here $i = 1, 2, \ldots$ 160,000 and $j = 1, 2, 3$ respectively for inner membrane protein, matrix protein and outer membrane protein.

Secondly, for a stochastic event, when one observes the i-th tetrapeptide occurring in the j-th submitochondria location, there are two possible outcomes: occurrence and not occurrence in the j-th submitochondria location. Each outcome has a fixed probability, the same from trial to trial. Thus, the probability p_j, also called prior probability, occurring in the j-th submitochondria location is defined as:

$$p_j = \sum_{i=1}^{160,000} n_{ij} \Big/ \sum_{j=1}^{3} \sum_{i=1}^{160,000} n_{ij} \tag{1}$$

where $\sum_{j=1}^{3} \sum_{i=1}^{160,000} n_{ij}$ is the total occurrence number of all tetrapeptides in the benchmark dataset. $\sum_{i=1}^{160,000} n_{ij}$ denotes the occurrence number of all tetrapeptides in the j-th submitochondria location proteins. According to this definition, the prior probability p_j correlates with the dimensions of benchmark dataset and sub-dataset. Correspondingly, the probability not occurring in the j-th submitochondria location is $q_j = 1 - p_j$.

Thirdly, we denoted the frequency of the i-th tetrapeptide occurring the benchmark dataset as N_i which formulated as $N_i = \sum_{j=1}^{3} n_{ij}$. That is to say that, under the condition of the prior probability p_j, one performs trial or observation with N_i times. Then, the probability (P_{ij}) of the i-th tetrapeptide occurring in the j-th submitochondria location n_{ij} or more times obeys the binomial distribution and can be defined by:

$$P_{ij} = 1 - CL_{ij} = \sum_{n=n_{ij}}^{N_i} \frac{N_i!}{n!(N_i - n)!} p_j^n (1 - p_j)^{N_i - n} \tag{2}$$

where CL_{ij} is called the confidence level (CL) of the i-th tetrapeptide in the j-th submitochondria location.

Fourthly, there are three submitochondria locations in the current study, namely $j = 1, 2, 3$. Therefore, according to Eqs. (1) and (2), for an arbitrary tetrapeptide i, it will have three confidence levels ($CL_{i\ inner}$, $CL_{i\ inner\ matrix}$ and $CL_{i\ outer}$) which describe the probabilities the i-th tetrapeptide occurring in the three classes of submitochondria proteins, respectively. Then the confidence level of tetrapeptide i in benchmark dataset can be defined as follows:

$$CL_i = \max\{CL_{i\,inner},\ CL_{i\,matrix},\ CL_{i\,outer}\} \tag{3}$$

Finally, if there are m tetrapeptides whose CL_i values are larger than a given cutoff CL_o, the frequencies of the m tetrapeptides are selected as optimized features. Accordingly, a protein in the benchmark dataset can be formulated by a discrete vector $\boldsymbol{F}$ as given by

$$F_m = [f_1, f_2, \ldots, f_i, \ldots, f_m]^T \tag{4}$$

where f_i ($i = 1, 2, \ldots, m$) are the frequencies of the m tetrapeptides in a protein and T is the transposing operator. If CL_o is set to zero, 160,000 tetrapeptides are all selected. If $CL_o > 1$, no tetrapeptides are selected. Based on confidence level (Eq. 1), high-dimensional data can be projected into low-dimensional space. The parameter m or CL_o can be chosen by use of cross-validation.

2.3 Support Vector Machine

Support vector machine (SVM) is a wonderful and popular machine learning method, which has been widely applied in bioinformatics. In this study, we used the software LIBSVM to implement SVM (Fan et al. 2005). For multi-class problems, we adopts one-versus-one (OVO) strategy for classification. The radial basis function (RBF) was chosen as the kernel function. The grid search program was applied to optimize the regularization parameter C and kernel parameter γ by using five-fold cross-validation.

2.4 Performance Evaluation

The jackknife cross-validation was used to evaluate the performance of the proposed model (Chou and Shen 2007). Three important parameters: sensitivity (Sn), overall accuracy (OA) and Matthews correlation coefficient (MCC) were calculated as the following formulas:

$$Sn = TP/(TP + FN) \tag{5}$$

$$OA = (TP + TN)/(TP + TN + FP + FN) \tag{6}$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP) \times (TN + FN) \times (TP + FN) \times (TN + FP)}} \tag{7}$$

where TP, TN, FP and FN denote true positives, true negatives, false positives and false negatives, respectively.

3 Results

For the benchmark dataset M495, the over-represented tetrapeptides can be obtained by using Eq. 1. In our statistics, the tetrapeptides with $N_i < 3$ in the dataset are eliminated, since these tetrapeptides do not prefer to occur in mitochondria proteins ($p < 0.0001$). Generally, the tetrapeptides with high confidence level give more reliable information for classification. However, the number of these tetrapeptides is too small to afford enough information, which deduces the poor predictive accuracy. For example, using >99.9 % as confidence level, we can achieve 13 tetrapeptides. But the overall accuracy is only 56.8 % by using five-fold cross-validation (Table 1). In contrast, the tetrapeptides with low confidence contains too many components. But it would reduce the cluster-tolerant capacity so as to lower down the cross-validation accuracy. For instance, 20,781 tetrapeptides with >50.0 % of confidence level can only produce the overall accuracy of 60.4 % by using five-fold cross-validation (Table 1).

Therefore, using appropriate tetrapeptides would yield a prediction with higher accuracy. By changing the cutoff of confidence level, we can obtain a series of tetrapeptide sets and examine the accuracies of these sets. For economizing time and improving efficiency, the five-fold cross-validation was used to optimize the

Table 1 The accuracies of different confidence levels by using five-fold cross-validation

Confidence level (%)	Number of features	Overall accuracy (%)
>99.9	13	56.8
>50.0	20,781	60.4
>89.9	1,302	89.9

regularization parameter C and kernel parameter γ. By examining all features subsets, we found that when the cutoff of CL is set to >96.5 %, the maximum overall accuracy reaches its maximum in jackknife cross-validation. A total of 91.1 % mitochondrial proteins can be correctly predicted by using 1,302 optimized tetrapeptides (Table 2). The *Sn*s and *MCC*s are 98.8 % and 0.84, 86.4 % and 0.89 as well as 78.9 % and 0.85, respectively for inner membrane proteins, matrix proteins and outer membrane proteins.

To verify the advantage of the proposed method, we repeated the process of feature selection and prediction on another three benchmark datasets: M317, M399 and M1105. When the cutoffs of CL were selected as 90.0, 93.9 and 93.2 %, the overall accuracies reached their maximums. The results were recorded in Table 2. It shows that the proposed method can achieve $\sim$94 % overall accuracy for the three benchmark datasets. These are almost as high as the best accuracies obtained by other methods, indicating that our method can be used for the prediction of proteins submitochondria location. Especially, our model can achieve the highest accuracy for the prediction of inner membrane proteins among other methods.

We noticed that the gene ontology (GO)-based model developed by Fan and Li (2012) obtained the best results on M317 and M1105. GO database was established based on the molecular function, biological process and cellular component. If the molecular function, biological process or cellular component of one protein has

Table 2 Performance of different methods on different datasets

Dataset	References	Inner membrane proteins		Matrix proteins		Outer membrane proteins		Overall accuracy (%)
		Sn (%)	*MCC*	*Sn* (%)	*MCC*	*Sn* (%)	*MCC*	
M495	This work	98.8	0.84	86.4	0.89	78.9	0.85	91.1
M317	Du and Li (2006)	85.5	0.79	94.5	0.77	51.2	0.64	85.2
	Nanni and Lumini (2008)	83.2	0.80	97.2	0.85	78.1	0.77	89.0
	Shi et al. (2011)	91.6	0.86	97.3	0.79	82.9	0.88	93.4
	Zakeri et al. (2011)	97.7	0.94	99.3	0.93	68.3	0.81	94.7
	Fan and Li (2012)	94.7	0.91	99.3	0.96	80.5	0.84	94.9
	This work	100	0.90	96.6	0.95	65.9	0.79	94.0
M399	Zeng et al. (2009)	91.8	0.79	96.4	0.79	66.1	0.63	89.7
	This work	99.4	0.90	99.4	0.98	69.4	0.81	94.7
M1105	Fan and Li (2012)	96.1	0.89	93.9	0.90	86.9	0.89	93.6
	This work	99.8	0.88	90.0	0.93	81.4	0.88	93.4

been experimentally defined, it is easy to guess the subcellular location of this protein. For example, Mei (2012) used this parameter to predict submitochondria protein location and achieved >99 % accuracy for M317. Thus it is not strange to obtain high accuracies by Fan and Li's (2012) model. However, there is a serious issue for using GO information to predict. The percentage (<50 %) of the protein entries with subcellular annotations in GO database is lower than that (>50 %) in the Uniprot database (Chou and Shen 2008). If a query protein has not been annotated in GO database and no homologous can be found in GO database, their model can not perform prediction. Our model predicts the location of mitochondria proteins only using primary sequence information, suggesting that our model is more neatly and freely. Furthermore, no matter which dataset was used to evaluate it, our model always achieved >91 % accuracy, suggesting the model is robust.

4 Discussion

Using tetrapeptides to recode protein sequences play a key role for predicting submitochondria location of proteins. Because frequencies of tetrapeptides occurrence in random sequence are very low (1/160,000), particular tetrapeptides tend to be present within a protein because of their contributions to the particular functional role of that protein and not as the result of some random choice (Stuart et al. 2002). Tripeptides or larger peptides could be used in prediction. However, tripeptides appear about 20 times more frequently than tetrapeptides; hence, they would bring more noise into prediction. For larger peptides, the size of the feature dimension is so large that the three problems: over-fitting, information redundancy and dimension disaster, would appear in computation. Moreover, some studies have proved that several mitochondrial intermembrane space proteins share tetrapeptide motifs (Verhagen et al. 2007; Polianskyte et al. 2009; Shi 2002). Furthermore, study has shown that tetrapeptides can product comprehensive gene and species phylogenies for mitochondrial genomes and also serve to identify correlated peptides as motifs (Stuart et al. 2002). According to these analyses, tetrapeptides are suitable for the mitochondrial protein data.

In addition, it is widely accepted that a protein sequence determines its structure and its structure determines its function. Rackovsky (1993) have used entropy to investigate the local coding properties of protein sequences and estimated that 60–70 % of tetrapeptides encode the specific structures. The Feng and Luo's (2008) results that the accuracy of 80 % was achieved by these tetrapeptides in the prediction of protein secondary structure have further demonstrated the relationship between sequence and structure. They suggested that these tetrapeptide signals can be regarded as the protein folding code in the protein structure prediction. Furthermore, the results of Fan and Li (2012) have proved that the predictive secondary structure information of protein can improve the accuracy for predicting submitochondria location of proteins. Therefore, using tetrapeptides directly to predict submitochondria location of proteins is a feasible approach. Our results also suggested that the optimized tetrapeptides are informative and can reflect inherent properties of mitochondria proteins.

The tetrapeptide words occurring once or twice in benchmark dataset do not prefer to occur in mitochondria proteins, thus we ignored them to guarantee the reliability of feature selection. In our statistics, we achieved 1,302 over-represented tetrapeptides which is much larger than the size (495) of the dataset. However, total of 188,940 tetrapeptides occurring in the dataset can guarantee the statistical significant of the over-represented tetrapeptides. Furthermore, the jackknife cross-validation (Chou and Zhang 1995) that can always yield a unique outcome was used to evaluate our method. Thus our results are credible.

For the convenience of the vast majority of experimental scientists, we constructed an on-line server, called TetraMito, which can be freely available at http://lin.uestc.edu.cn/server/TetraMito. The server may become a useful vehicle for in-depth studying mitochondria proteins, or at least a complementary tool to the existing methods in this area.

5 Conclusions

In this study, we developed a feature selection-based method to predict the submitochondria locations of mitochondria proteins using primary sequence information. Results demonstrate that the proposed method has the capability to predict and annotate the submitochondria locations of mitochondria proteins. Based on this model, we have constructed a free online server TetraMito. The current study will become an important progress in the prediction of the submitochondria protein locations and promote the study in the related areas.

Acknowledgments We are grateful to Dr. Loris Nanni for his help. This work was supported by the National Nature Scientific Foundation of China (No. 61202256, 61100092), the Project of Education Department in Sichuan (12ZA112), the Fundamental Research Funds for the Central Universities (ZYGX2012J113) and the Scientific Research Startup Foundation of UESTC.

References

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

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 Y (2006) Prediction of protein submitochondria locations by hybridizing pseudo-amino acid composition with various physicochemical features of segmented sequence. BMC Bioinform 7:518

Du P, Cao S, Li Y (2009) SubChlo: predicting protein subchloroplast locations with pseudo-amino acid composition and the evidence-theoretic K-nearest neighbor (ET-KNN) algorithm. J Theor Biol 261:330–335

Du P, Li T, Wang X (2011) Recent progress in predicting protein sub-subcellular locations. Expert Rev Proteomics 8:391–404

Fan GL, Li QZ (2012) Predicting protein submitochondria locations by combining different descriptors into the general form of Chou's pseudo amino acid composition. Amino Acids 43:545–555

Fan RE, Chen PH, Lin CJ (2005) Working set selection using the second order information for training SVM. J Mach Learn Res 6:1889–1918

Feng Y, Luo L (2008) Use of tetrapeptide signals for protein secondary-structure prediction. Amino Acids 35:607–614

Henze K, Martin W (2003) Evolutionary biology: essence of mitochondria. Nature 426:127–128

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 90:573–581

Huang WL, Tung CW, Ho SW, Hwang SF, Ho SY (2008) ProLoc-GO: utilizing informative Gene Ontology terms for sequence-based prediction of protein subcellular localization. BMC Bioinform 9:80

Huang WL, Tung CW, Huang HL, Ho SY (2009) Predicting protein subnuclear localization using GO-amino-acid composition features. Biosystems 98:73–79

Jiang X, Wei R, Zhao Y, Zhang T (2008) 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

Lei Z, Dai Y (2005) An SVM-based system for predicting protein subnuclear localizations. BMC Bioinform 6:291

Lei Z, Dai Y (2006) Assessing protein similarity with Gene Ontology and its use in subnuclear localization prediction. BMC Bioinform 7:491

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

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

Mei S, Fei W (2010) Amino acid classification based spectrum kernel fusion for protein subnuclear localization. BMC Bioinform Suppl 1:S17

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

Polianskyte Z, Peitsaro N, Dapkunas A, Liobikas J, Soliymani R, Lalowski M, Speer O, Seitsonen J, Butcher S, Cereghetti GM, Linder MD, Merckel M, Thompson J, Eriksson O (2009) LACTB is a filament-forming protein localized in mitochondria. Proc Natl Acad Sci USA 106:18960–18965

Rackovsky S (1993) On the nature of protein folding code. Proc Natl Acad Sci USA 90:644–648

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 (2007) Nuc-PLoc: a new web-server for predicting protein subnuclear localization by fusing PseAA composition and PsePSSM. Protein Eng Des Sel 20:561–567

Shi Y (2002) A conserved tetrapeptide motif: potentiating apoptosis through IAP-binding. Cell Death Differ 9:93–95

Shi SP, Qiu JD, Sun XY, Huang JH, Huang SY, Suo SB, Liang RP, Zhang L (2011) Identify submitochondria and subchloroplast locations with pseudo amino acid composition: approach from the strategy of discrete wavelet transform feature extraction. Biochim Biophys Acta 1813:424–430

Stuart GW, Moffett K, Leader JJ (2002) A comprehensive vertebrate phylogeny using vector representations of protein sequences from whole genomes. Mol Biol Evol 19:554–562

UniProt Consortium (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Res 40:D71–D75

van Dijk AD, Bosch D, ter Braak CJ, van der Krol AR, van Ham RC (2008) Predicting sub-Golgi localization of type II membrane proteins. Bioinformatics 24:1779–1786

Verhagen AM, Kratina TK, Hawkins CJ, Silke J, Ekert PG, Vaux DL (2007) Identification of mammalian mitochondrial proteins that interact with IAPs via N-terminal IAP binding motifs. Cell Death Differ 14:348–357

Wang G, Dunbrack RL Jr (2005) PISCES: recent improvements to a PDB sequence culling server. Nucleic Acids Res 33:W94–W98

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, Xiao RQ, Yang L, Yu LZ, Li ML (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