Send Orders for Reprints to reprints@benthamscience.ae

REVIEW ARTICLE

Predicting Protein Submitochondrial Locations: The 10th Anniversary

Pu-Feng Du*

School of Computer Science and Technology, Tianjin University, Tianjin 300350, China

ARTICLE HISTORY

Received: September 01, 2016
Revised: October 16, 2016
Accepted: November 02, 2016

DOI:
10.2174/1389202918666170228143256

Abstract: Predicting protein submitochondrial location has been studied for about ten years. A number of methods have been developed. The prediction performances have been improved to an almost perfect level. In this review, we introduce the background of this research topic. We also compare the methods, the performances and the datasets that have been used by these studies. Towards the end, we provide hints for the future directions of this research topic.

Keywords: Submitochondrial locations, Intermembrane space, Feature selection, DNA, Globular proteins, Mitochondria.

1. INTRODUCTION

A eukaryotic cell contains several different subcellular organelles. These organelles are usually enclosed by membranes. A mitochondrion is a subcellular organelle that is enclosed by two layers of membranes. A cell usually contains several mitochondria. A mitochondrion has its own genome, which is a circular DNA. Mitochondria are involved in many biological processes, such as energy metabolism, programmed cell death, and ionic homeostasis. The number of mitochondrial proteins is far more than the number of coding genes in the mitochondrial genome. Therefore, a number of mitochondrial proteins must be encoded by the nucleus genome, and must be transferred to mitochondria after or during translation.

The two layers of membranes, which enclose a mitochondrion, are called the inner membrane and the outer membrane. The outer membrane separate the internal space of a mitochondrion from the cytosol. The inner membrane separate the internal space of a mitochondrion into two parts. The space between the inner membrane and the outer membrane is called the intermembrane space. The space, which is surrounded by the inner membrane, contains mitochondrial matrix. The inner membrane fold into the matrix, which creates several cristae. Cristae are parts of inner membrane, which increase its surface area for biochemical reactions. There are four submitochondrial locations in a mitochondrion, the outer membrane, the intermembrane space, the inner membrane, and the matrix.

The functions of mitochondria proteins are related to their submitochondrial locations. The number of protein sequences in the UniProt database is increasing rapidly [1]. However, experimental approaches to determining exact locations of a protein is always costly and time consuming. Since protein subcellular localizations are related to their sequence, computational methods to predict protein localization from sequences are desired [2]. Computational prediction of protein subcellular locations have been extensively studied in the last two decades [3, 4]. Many different computational methods have been developed. Recently, the studies in predicting protein subcellular locations focused on four aspects: (1) predicting protein sub-subcellular locations, such as protein subnuclear locations [5], submitochondrial locations [6] and subchloroplast locations [7]; (2) predicting multiple subcellular locations for a single protein [8]; (3) predicting subcellular locations for proteins with specific structural topology, such as membrane proteins, globular proteins, and anchored proteins [9]; and (4) predicting mis-localized proteins in diseases, therapies and environmental stresses [10]. In this review, we will specifically focus on the progress of predicting protein submitochondrial locations.

The first report in predicting protein submitochondrial locations appeared in the year 2006. Du and Li proposed the SubMito predictor to assign submitochondrial locations to mitochondria proteins [6]. They also released the first benchmarking dataset in this research area. This dataset is currently known as the M3-317 dataset. Over the last decade, several studies were carried out in predicting protein submitochondrial locations. Nanni and Lumini introduced a genetic-algorithm-based method to select sequence-based protein descriptors [11]. Shi *et al.* developed the SubIdent method to further improve the prediction performance [12]. Fan and Li incorporated gene ontology annotations in a hybrid method with feature selection technology [13]. Zakeri *et al.* proposed the Mito-Loc method by fusing features of sequence, structures and annotations [14]. Lin *et al.* used over-represented tetra-peptides to predict the protein submitochondrial locations [15]. Du and Yu improved the prediction

*Address correspondence to this author at the School of Computer Science and Technology, Tianjin University, Tianjin 300350, China; Tel/Fax: +86-23689450; E-mail: PufengDu@gmail.com

1389-2029/17 $58.00+.00

performance by introducing a new concept, which is called Positional Specific Physicochemical Properties (PSPCP) [16]. Ahmad *et al.* applied Synthetic Minority Oversampling Technique (SMOTE) as a feature selection method [17]. Li *et al.* achieved extraordinary performance by using a hybrid method [18].

In this review, we collect and sort all types of related resources. The datasets, algorithms, performance values and software are collected and compared. We hope this review can serve as a useful resource to the readers. Although we cannot go into too much to the details of all algorithms, we compare these predictors in many aspects. We also discuss current stages in predicting protein submitochondrial locations, as well as the future.

2. DATASETS

The datasets for predicting protein submitochondrial locations have a naming convention. Although this is not an established standard, most existing studies are following. The name of a dataset is usually noted in the form Mx-y. The capital M stands for mitochondria. Sometimes, the capital M is written as SML, which are short for submitochondrial locations. The x is an integer, which indicates the number of submitochondrial locations in the dataset. The y is another integer, which indicates the number of protein sequences in the dataset. Besides the first dataset, M3-317, there are several other benchmarking datasets. These benchmarking datasets are M3-399, M3-495, M3-983 and M3-1105 (Table **1**).

The M3-317 dataset contains 317 proteins, including 131 inner membrane proteins, 41 outer membrane proteins and 145 matrix proteins [6]. The M3-399 dataset was curated by Zeng *et al*. [19]. It contains 171 inner membrane proteins, 62 outer membrane proteins and 166 matrix proteins. Lin *et al.* released the M3-495 dataset along with their TetraMito service [15]. It contains 254 inner membrane protein, 109 outer membrane proteins and 132 matrix proteins. Du and Yu introduced the M3-983 dataset, which contains 661 inner membrane proteins, 145 outer membrane proteins and 177 matrix proteins [16]. Fan and Li provided the M3-1105 dataset with 589 inner membrane proteins, 280 outer membrane proteins and 236 matrix proteins [13].

All the above datasets are collected from the UniProt database with several filtering steps. As they are curated in different years, the number of proteins in the dataset vary in a large range. Another reason of the number variation is whether to allow electronic annotations of submitochondrial locations. In the UniProt database, a number of proteins have submitochondrial locations that are annotated by electronic methods. If these annotations are not removed, the number of sequences in the dataset will increase. The impact of similarity cutoff should not be ignored also. Most of the authors used CD-HIT program [20] to control the sequence similarity to 40%, which is the minimal value of similarity cutoff that can be provided by the CD-HIT program. To achieve even lower similarity cutoff value, PISCES program was applied by Lin *et al*. [21]. Besides the PISCES, PSI-CD-HIT program can also control the sequence similarity at 25%.

Although recent works have included about 1000 sequences in the dataset, the number of submitochondrial locations is still three. As we have mentioned, there are four submitochondrial locations. The intermembrane space is always missing. The reason for excluding this location is always that the number of sequences is not sufficient. When the M3-317 dataset was released, the number of sequences in the intermembrane space is less than 10. However, when the M3-1105 dataset was released, the number is about 50. We believe it should be ready to include the fourth submitochondrial location in the next study.

Recently, predicting protein cellular attributes in multi-label context is popular [8]. However, there is still no multi-label submitochondrial location predictor. All existing datasets considers proteins with unique submitochondrial location. In fact, the proteins with multiple submitochondrial locations exist. Although the number of these proteins is still minimal, we believe it should also be considered in the next study.

3. METHODOLOGIES

All existing studies in predicting protein submitochondrial locations follow the research paradigm of predicting protein attributes. The protein sequences are represented with digital vectors. Machine learning algorithms are employed as predicting engines. We compare the existing studies on three aspects: (1) the features or information that have been encoded in the digital representations; (2) the learning algorithms; and (3) the feature selection methods.

Before we discuss the methods of every study, we should point out that most of the existing study were using the general form pseudo-amino acid compositions as their sequence

Table 1. Benchmarking datasets.

Name	Year	Inner	Outer	Matrix	Total	Sim	E-anno
M3-317	2006	131	41	145	317	40%	Excluded
M3-399	2009	171	62	166	399	40%	Excluded
M3-495	2013	254	109	132	495	25%	Excluded
M3-983	2013	661	145	177	983	40%	Included
M3-1105	2012	589	280	236	1105	40%	Included

Inner stands for inner membrane. Outer stands for outer membrane. Sim stands for Similarity cutoff. E-anno indicates whether the electronic annotations are included.

representations [22]. The pseudo-amino acid compositions were invented by Chou in 2001 as a tool in predicting protein cellular attributes [22]. Recently, it was extended to represent DNA/RNA sequences [23]. A serial of online servers and programs were released to help converting biological sequence into the pseudo-amino acid compositions [24-28]. Especially, the PseKRAAC server, which represents the most recent advancement, provides the potentials to further improve the sequence representation abilities [29].

Du and Li proposed the SubMito method, in which only pseudo-amino acid compositions were applied [6]. Du and Li used SVM as the predicting engine. There is no feature selection process in their method.

In the GPLoc study, the features also encoded only sequence information [11]. The most creative part of the GPLoc method is that it uses genetic algorithm to generate optimal artificial features. These optimal features can be combinations of transformations of original features.

Shi *et al*. developed a wavelet-SVM based method, which is called SubIdent [12]. Although the original sequence representations were very simple, the power of the wavelet transformation makes the prediction accuracy much better than the SubMito and the GPLoc.

Zeng *et al*. introduced the Predict_subMITO method [19]. Their method used an augmented form of Chou's pseudo-amino acid compositions as the sequence representations. They also used SVM as the predicting engine. There is no feature selection in their method.

The above four studies can be called the first stage in predicting protein submitochondrial locations. They have two common characters: (1) only protein sequence information was encoded in the feature vector; and (2) the feature extraction methods generate artificial features rather than selecting features from original ones.

Zakeri *et al*. proposed the Mito-Loc method [14]. They included protein functional domain contents in their method. They used an SVM-based ensemble method as the predicting engine. Fan and Li applied various kinds of features in predicting protein submitochondrial locations [13]. They incorporated the gene ontology annotations, evolutionary information and average chemical shift in their work. They also proposed a new dataset, which is currently known as the M3-1105 dataset. Lin *et al*. released the TetraMito predictor [15]. This is the first time that the tetra-peptide compositions were used in predicting protein submitochondrial location. As the tetra-peptide compositions creates very high dimensional features. A Bayesian analysis based feature selection procedure was applied. They also proposed a new dataset, which is currently known as the M3-495 dataset.

Du and Yu developed a new concept, which is called the positional-specific physicochemical properties (PSPCP) [16]. It integrates evolutionary information into the pseudo-amino acid composition at a fundamental level. The most important advantage of PSPCP is that it is 100% compatible with the pseudo-amino acid compositions. The proteins, which are represented by PSPCP, can be mixed with those, which are represented by pseudo-amino acid compositions, without any modification. Besides the SubMito-PSPCP work, the concept of PSPCP has been applied in identifying Golgi-resident protein types [30, 31].

Ahmad *et al.* applied SMOTE technology [32] on split amino acid compositions. With various kinds of machine learning algorithms, they achieved very high overall accuracy. Li *et al.* applied wrapper-style SVM-based feature selection methods on a hybrid feature [18]. By sophisticated design and calibration, their method can provide extraordinary prediction performances.

The later six methods can be called the second stage in predicting protein submitochondrial locations. The methods in this stage commonly applied feature selection methods, which choose features from the original ones. This is different to the methods in the first stage, which uses original features to create artificial new ones. The methods in this stage also commonly incorporate information other than the sequence itself. The evolutionary information, gene ontology annotation, functional domain contents and many others were considered by these methods.

Since Gene Ontology (GO) annotations contain the cellular components information, it was thought to be unfair to use GO annotations in predicting protein submitochondrial locations. However, several existing studies have pointed out that this is an unnecessary concern [33, 34]. GO annotations are safe to be applied as features in predicting protein submitochondrial locations, as well as other protein cellular attributes.

A more comprehensive comparison of the methods, including the sequence representations, the machine-learning algorithms, and the feature selection methods, can be found in (Table **2**).

Table 2. Features in representing protein sequences.

Methods	Seq	Evo	FuncDom	GO	FS
SubMito[6]	Yes	No	No	No	No
GPLoc[11]	Yes	No	No	No	Yes
SubIdent[12]	Yes	No	No	No	Yes
Predict_SubMito[19]	Yes	No	No	No	No
MitoLoc[14]	Yes	No	Yes	No	Yes
Fan and Li[13]	Yes	Yes	No	Yes	Yes
TetraMito[15]	Yes	No	No	No	Yes
SubMito-PSPCP[16]	Yes	Yes	No	No	No
Ahmad *et al.*[17]	Yes	No	No	No	Yes
Li *et al.*[18]	Yes	Yes	No	Yes	Yes

Seq: Sequence-based features, including amino acid compositions, dipeptide compositions, physicochemical properties, average chemical shift and *etc*; Evo: Evolutionary information, including positional-specific scoring matrix and PSI-BLAST/BLAST search; FuncDom: Functional domain information; GO: Gene Ontology annotations; FS: Feature selection methods.

4. PERFORMANCE COMPARISONS

Three main cross-validation methods have been widely applied in evaluating performances of bioinformatics predic-

tors [8, 35-38]. Since all existing methods reported jackknife test results on M3-317 dataset, we focus on comparing prediction performances of these methods on the M3-317 dataset. We collect the jackknife test results in (Table **3**). Three commonly applied performance measures [6, 12-14, 17, 39-41] can be defined as follows:

$$ACC_k = \frac{TP_k}{TP_k + FN_k}, \quad (1)$$

$$MCC_k = \frac{TP_k TN_k - FP_k FN_k}{\sqrt{(TP_k + FP_k)(TP_k + FN_k)(TN_k + FP_k)(TN_k + FN_k)}}, \quad (2)$$

$$ACC = \frac{1}{n}\sum_{k=1}^{q} TP_k, \quad (3)$$

where ACC_k is the prediction accuracy for the k-th locations, MCC_k the Mathew's correlation coefficients of the k-th location, ACC the overall accuracy, TP_k, TN_k, FP_k and FN_k the number of true positives, true negatives, false positives and false negatives of the k-th location, and n the total number of proteins in a dataset. As we are discussing M3-317 dataset, $k \in \{1, 2, 3\}$, $q = 3$, and $n = 317$.

Table 3. Prediction performances comparison.

Methods	Inner	Matrix	Outer	Overall
SubMito	85.5%[0.79]	94.5%[0.77]	51.2%[0.64]	85.2%
GPLoc	83.2%[0.80]	97.2%[0.85]	78.1%[0.77]	89.0%
SubIdent	91.6%[0.86]	97.3%[0.79]	82.9%[0.88]	93.1%
Predict_SubMito	91.8%[0.79]	96.4%[0.79]	66.1%[0.63]	89.7%
MitoLoc	97.7%[0.94]	99.0%[0.93]	68.3%[0.81]	94.7%
Fan and Li	94.7%[0.91]	99.3%[0.96]	80.5%[0.84]	94.9%
TetraMito	100.0%[0.90]	96.6%[0.95]	65.9%[0.79]	94.0%
SubMito-PSPCP	98.6%[0.92]	93.9%[0.89]	70.7%[0.79]	93.1%
Ahmad *et al.*	94.0%[0.90]	93.3%[0.89]	98.7%[0.90]	95.2%
Li *et al.*	100.0%[0.99]	98.6%[0.99]	100.0%[1.00]	99.4%

The performance in this table are presented in ACC[MCC] form. ACC is accuracy, as defined in eq (1). MCC is Mathew's Correlations Coefficients, as defined in eq (2). Overall accuracy is defined in eq (3). All methods names have the same indication as Table **2**.

The SubMito method, which is the first study in predicting protein submitochondrial location, had 85.2% overall accuracy. Li *et al*. work, which is the latest one, provided 99.4% overall accuracy. In the last ten years, the prediction accuracy increased nearly 14%.

The TetraMito method and Li *et al*. method achieved 100% accuracy in inner membrane. However, the TetraMito method has only 65.9% accuracy in outer membrane, which makes its overall accuracy less than Li *et al.*

Generally, Li *et al*. provided the highest overall accuracy with almost perfect results in every location. Especially, they achieved 100% accuracy in outer membrane, which is usually the lowest performed location. The only imperfect prediction of Li *et al*. work is the matrix location. They achieved 98.6% accuracy, which is slightly lower than Fan and Li.

The superior prediction performance of Li *et al*. method should be the result of carefully designed sequence representations and feature selection schemes. The power of feature selection methods have been demonstrated in many studies [42]. The original features contain many different types of features, which cover almost every possible feature that has been applied in other studies. The wrapper-style feature selection method can efficiently select those most relevant features. We believe that their method has a great potential in predicting other protein cellular attributes.

The second stage methods usually have higher prediction performances than the first stage ones. The methods, which uses various kinds of information, usually performed better than those using only sequence features. The methods with feature selections usually performed better. For the methods that are not using feature selection methods, the PSPCP method is the best one. It provides comparable performances to almost every other methods, even to those ones using feature selection methods.

It should be noticed that, besides the most recent work, the performance in outer membrane is not as high as the other two locations. This may be due to the imbalance in M3-317 dataset, as the outer membrane is just the location that contain least number of proteins. This observation conforms to the superior performances in Ahmad *et al*., where over-sampling technology was applied to solve the imbalanced dataset problem.

As the performance of Li *et al*. is almost perfect, we believe that the race of prediction performance on M3-317 dataset has reached its destination. However, there are still problems that should be discussed in predicting protein submitochondrial locations. For example, if the intermembrane space proteins are included in the dataset, can existing methods perform as well as now?

5. RESOURCES AND AVAILABILITY

Although existing methods have been extensively tested and rigorously compared, it is also very important that whether these methods can be practically applied. We compare the availability of existing methods in two aspects: (1) the availability of software; and (2) the availability of benchmarking datasets.

The M3-317 dataset was curated in the year 2006. At first, this dataset was not provided online. Instead, it can be required by emails. The software SubMito has both online and local versions. It is still working recently. The M3-399 dataset was provided along with the Predict_subMITO service. However, it is not accessible by using the URL that was provided in the paper.

Fan and Li released the M3-1105 dataset. However, the URL that was provided in their paper is not working recently. They did not provided an online service or local program as the implementation of their method.

The TetraMito service works very well in a recent test. The dataset M3-495, which was released along with

Table 4. Resources and availability.

Methods	Year	Online Service	Local Program	Dataset	Web URL
SubMito	2006	Yes	Yes	M3-317	http://bioinfo.au.tsinghua.edu.cn/subMito/
GPLoc	2008	No	No	M3-317	NA
SubIdent	2011	No	No	M3-317	NA
Predict_SubMito	2009	No	No	M3-317; M3-399	NA
MitoLoc	2011	No	No	M3-317	NA
Fan and Li	2012	No	No	M3-317; M3-1105	NA
TetraMito	2013	Yes	No	M3-317; M3-495	http://lin.uestc.edu.cn/server/TetraMito
SubMito-PSPCP	2013	Yes	No	M3-317; M3-983	http://www.pufengdu.org/bioinfo/submito-pspcp/
Ahmad *et al*.	2016	No	No	M3-317; M3-983	NA
Li *et al*.	2015	No	No	M3-317; M3-1105	NA

NA: Not available; All methods names have the same indications as Table **2**.

TetraMito can be downloaded smoothly from the server. The SubMito-PSPCP services cannot be accessed in its original URL. However, it can be accessed by a new URL, as shown in (Table **4**). The M3-983 dataset and the M3-317 dataset can be downloaded from the SubMito-PSPCP website.

It is sad that the remaining works did not provide website or local software package for their method. Although some of these methods have almost perfect prediction performances, without software availability, it is hard to be applied in other studies.

To facilitate a resource, we collect all the URLs in (Table **4**). If the readers need to access one of these services, they can simply click the link.

CONCLUSION

Predicting protein submitochondrial locations has been studied for about ten years. As we have mentioned, a number of methods have been developed in this regard. Recent studies have provided almost perfect prediction performances on the benchmarking dataset. Therefore, we believe the race in the prediction performances should have a result now. However, as long as the mechanism of protein localization has not been fully deciphered, the machine learning-based methods will be continually introduced to predict protein localizations, not only the submitochondrial locations, but also the others.

As we have mentioned, the intermembrane space is a location that is always missing in existing studies. Ten years before, this is due to the limited number of known proteins in this locations. Recently, the number of proteins in this location has increased. Although the number is still not large, we believe that it is time to incorporate this location. Therefore, we expect to see the next study in predicting protein submitochondrial locations with the consideration of four locations. Moreover, since the number of proteins with multiple submitochondrial locations has also increased, maybe it is time to incorporate these proteins also.

Since most of the mitochondrial proteins are encoded by the nucleus genome, the mechanism that transport the proteins into the mitochondrion is important. Several experimental studies have determined that the proteins have to be modified in many ways while importing to the mitochondria [43]. However, the information of the importing mechanism has not been considered in existing studies. Using the importing mechanism information will not only improve the prediction performance, but also improve the interpretability of the prediction results. We believe that predicting protein submitochondrial location still worth further study. With the accumulation of known proteins, a new stage is coming in this research topic.

CONFLICT OF INTEREST

The author declares no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

We would like to present this paper for the 80th birthday of Prof. Yanda Li, who had directed the SubMito study. We would like to thank the editor of *Current Genomics* for the rare opportunity to share this survey with the research community. We would like to thank all the colleagues that have been involved in this research topic in the last ten years. We would also thank the anonymous reviewers for their useful comments for improving this review paper. No funding support was utilized in writing this review.

REFERENCES

[1] UniProt Consortium. UniProt: A hub for protein information. *Nucleic Acids Res.*, **2015**, *43*, D204-212.
[2] Imai, K.; Nakai, K. Prediction of subcellular locations of proteins: where to proceed? *Proteomics*, **2010**, *10*, 3970-3983.
[3] Du, P.; Xu, C. Predicting multisite protein subcellular locations: progress and challenges. *Expert Rev. Proteomics*, **2013**, *10*, 227-237.
[4] Wang, Z.; Zou, Q.; Jiang, Y.; Ju, Y.; Zeng, X. Review of protein

subcellular localization prediction. *Curr. Bioinformat.*, **2014**, *9*, 331-342.

[5] Shen, H.-B.; Chou, K.-C. Nuc-PLoc: A New web-server for predicting protein subnuclear localization by fusing pseaa composition and pseppssm. *Protein Eng. Des. Sel.*, **2007**, *20*, 561-567.

[6] Du, P.; Li, Y. Prediction of protein submitochondria locations by hybridizing pseudo-amino acid composition with various physicochemical features of segmented sequence. *BMC Bioinformat.*, **2006**, *7*, 1.

[7] Wang, X.; Zhang, W.; Zhang, Q.; Li, G.-Z. MultiP-SChlo: multi-label protein subchloroplast localization prediction with chou's pseudo amino acid composition and a novel multi-label classifier. *Bioinformatics*, **2015**, *31*, 2639-2645.

[8] Chou, K.-C. Some remarks on predicting multi-label attributes in molecular biosystems. *Mol. Biosyst.*, **2013**, *9*, 1092-1100.

[9] Pierleoni, A.; Martelli, P.L.; Casadio, R. MemLoci: Predicting subcellular localization of membrane proteins in eukaryotes. *Bioinformatics*, **2011**, *27*, 1224-1230.

[10] Lee, K.; Byun, K.; Hong, W.; Chuang, H.-Y.; Pack, C.-G.; Bayarsaikhan, E.; Paek, S.H.; Kim, H.; Shin, H.Y.; Ideker, T.; Lee, B. Proteome-wide discovery of mislocated proteins in cancer. *Genome Res.*, **2013**, *23*, 1283-1294.

[11] Nanni, L.; Lumini, A. Genetic programming for creating chou's pseudo amino acid based features for submitochondria localization. *Amino Acids*, **2008**, *34*, 653-660.

[12] Shi, S.-P.; Qiu, J.-D.; Sun, X.-Y.; Huang, J.-H.; Huang, S.-Y.; Suo, S.-B.; Liang, R.-P.; Zhang, L. Identify submitochondria and subchloroplast locations with pseudo amino acid composition: approach from the strategy of discrete wavelet transform feature extraction. *Biochim. Biophys. Acta*, **2011**, *1813*, 424-430.

[13] Fan, G.-L.; Li, Q.-Z. Predicting protein submitochondria locations by combining different descriptors into the general form of chou's pseudo amino acid composition. *Amino Acids*, **2012**, *43*, 545-555.

[14] Zakeri, P.; Moshiri, B.; Sadeghi, M. Prediction of protein submitochondria locations based on data fusion of various features of sequences. *J. Theor. Biol.,* **2011**, *269*, 208-216.

[15] Lin, H.; Chen, W.; Yuan, L.-F.; Li, Z.-Q.; Ding, H. Using over-represented tetrapeptides to predict protein submitochondria locations. *Acta Biotheor.*, **2013**, *61*, 259-268.

[16] Du, P.; Yu, Y. SubMito-PSPCP: Predicting protein submitochondrial locations by hybridizing positional specific physicochemical properties with pseudoamino acid compositions. *BioMed Res. Intl.*, **2013**, *2013*, 263829

[17] Ahmad, K.; Waris, M.; Hayat, M. Prediction of protein submitochondrial locations by incorporating dipeptide composition into chou's general pseudo amino acid composition. *J. Membrane Biol.*, **2016**, *249*, 293-304.

[18] Li, L.; Yu, S.; Xiao, W.; Li, Y.; Hu, W.; Huang, L.; Zheng, X.; Zhou, S.; Yang, H. Protein submitochondrial localization from integrated sequence representation and svm-based backward feature extraction. *Mol. BioSyst.*, **2014**, *11*, 170-177.

[19] Zeng, Y.; Guo, Y.; Xiao, R.; Yang, L.; Yu, L.; Li, M. Using the augmented chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach. *J. Theor. Biol.,* **2009**, *259*, 366-372.

[20] Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **2012**, *28*, 3150-3152.

[21] Wang, G.; Dunbrack, R.L. PISCES: a protein sequence culling server. *Bioinformatics*, **2003**, *19*, 1589-1591.

[22] Chou, K.C. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins*, **2001**, *43*, 246-255.

[23] Chen, W.; Lin, H.; Chou, K.-C. Pseudo nucleotide composition or pseknc: an effective formulation for analyzing genomic Sequences. *Mol. Biosyst.*, **2015**, *11*, 2620-2634.

[24] Liu, B.; Liu, F.; Fang, L.; Wang, X.; Chou, K.-C. repDNA: A python package to generate various modes of feature vectors for dna sequences by incorporating user-defined physicochemical properties and sequence-order effects. *Bioinformatics*, **2015**, *31*, 1307-1309.

[25] Liu, B.; Liu, F.; Fang, L.; Wang, X.; Chou, K.-C. repRNA: a web server for generating various feature vectors of rna sequences. *Mol. Genet. Genomics*, **2016**, *291*, 473-481.

[26] Liu, B.; Liu, F.; Wang, X.; Chen, J.; Fang, L.; Chou, K.-C. Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Res.*, **2015**, *43*, W65-71.

[27] Du, P.; Wang, X.; Xu, C.; Gao, Y. PseAAC-Builder: A Cross-platform stand-alone program for generating various special chou's pseudo-amino acid compositions. *Analyt. Biochem.*, **2012**, *425*, 117-119.

[28] Du, P.; Gu, S.; Jiao, Y. 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.*, **2014**, *15*, 3495-3506.

[29] Zuo, Y.; Li, Y.; Chen, Y.; Li, G.; Yan, Z.; Yang, L. PseKRAAC: A flexible web server for generating pseudo k-tuple reduced amino acids composition. *Bioinformatics*, **2017**, *33*(1), 122-124.

[30] Jiao, Y.-S.; Du, P.-F. Predicting golgi-resident protein types using pseudo amino acid compositions: approaches with positional specific physicochemical properties. *J. Theoret. Biol.*, **2016**, *391*, 35-42.

[31] Jiao, Y.-S.; Du, P.-F. Predicting golgi-resident protein types using pseudo amino acid compositions: approaches with positional specific physicochemical properties. *J. Theor. Biol.,* **2016**, *391*, 35-42.

[32] Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res. (JAIR)*, **2002**, *16*, 321-357.

[33] Lin, W.-Z.; Fang, J.-A.; Xiao, X.; Chou, K.-C. iLoc-Animal: A Multi-Label learning classifier for predicting subcellular localization of animal proteins. *Mol. BioSyst.*, **2013**, *9*, 634-644.

[34] Du, P.; Wang, L. Predicting human protein subcellular locations by the ensemble of multiple predictors *via* protein-protein interaction network with edge clustering coefficients. *PLoS One*, **2014**, *9*, e86879.

[35] Song, L.; Li, D.; Zeng, X.; Wu, Y.; Guo, L.; Zou, Q. nDNA-Prot: identification of dna-binding proteins based on unbalanced classification. *BMC Bioinformat.*, **2014**, *15*, 298.

[36] Lin, C.; Zou, Y.; Qin, J.; Liu, X.; Jiang, Y.; Ke, C.; Zou, Q. Hierarchical classification of protein folds using a novel ensemble classifier. *PLoS ONE*, **2013**, *8*, e56499.

[37] Zou, Q.; Li, X.; Jiang, Y.; Zhao, Y.; Wang, G. BinMemPredict: a web server and software for predicting membrane protein types. *Curr. Proteomics*, **2013**, *10*, 2-9.

[38] Li, D.; Ju, Y.; Zou, Q. Protein folds prediction with hierarchical Structured SVM http://chinesesites.library.ingentaconnect.com/content/ben/cp/2016/00000013/00000002/art00003 (accessed Nov 5, 2016).

[39] Zuo, Y.; Lv, Y.; Wei, Z.; Yang, L.; Li, G.; Fan, G. iDPF-PseRAAAC: A web-server for identifying the defensin peptide family and subfamily using pseudo reduced amino acid alphabet composition. *PLoS ONE*, **2015**, *10*, e0145541.

[40] Zuo, Y.-C.; Su, W.-X.; Zhang, S.-H.; Wang, S.-S.; Wu, C.-Y.; Yang, L.; Li, G.-P. Discrimination of membrane transporter protein types using k-nearest neighbor method derived from the similarity distance of total diversity measure. *Mol. Biosyst.*, **2015**, *11*, 950-957.

[41] Zuo, Y.-C.; Peng, Y.; Liu, L.; Chen, W.; Yang, L.; Fan, G.-L. Predicting peroxidase subcellular location by hybridizing different descriptors of Chou' Pseudo amino acid patterns. *Anal. Biochem.*, **2014**, *458*, 14-19.

[42] Zou, Q.; Zeng, J.; Cao, L.; Ji, R. A novel features ranking metric with application to scalable visual and bioinformatics data classification. *Neurocomputing*, **2016**, *173*, *Part 2*, 346-354.

[43] Dudek, J.; Rehling, P.; van der Laan, M. Mitochondrial protein import: Common principles and physiological networks. *Biochimica et Biophysica Acta (BBA) - Mol. Cell Res.*, **2013**, *1833*, 274-285.