Research Article

Journal of Nutritional Biology

Multi-Site Protein Subcellular Localization Based on Deep Convolutional Neural Network

Cong H^{1,2,3,4}, Liu H^{1,2,*}, Chen Y^{3,4}, Zhao Y^{3,4} and Wang L^{3,4}

¹School of Information Science and Engineering, Shandong Normal University, Jinan, China

²Shandong Provincial Key Laboratory for Novel Distributed Computer Software Technology, Jinan, China

³School of Information Science and Engineering University of Jinan, Jinan, China

⁴Shandong Provincial Key Laboratory of Network Based Intelligent Computing University of Jinan, Jinan, China

***Correspondence:** Hong Liu, School of Information Science and Engineering, Shandong Normal University, No. 88, Wenhua East Road, Jinan City, China, E-mail: lhsdcn@126.com

Received: December 01, 2018; Accepted: January 17, 2019; Published: January 21, 2019

Abstract

Each part of internal structure of cells which is commonly mentioned as subcellular is highly ordered and interconnected has unique functions. The experiments show that deviated protein delivery to the corresponding subcellular causes of human disease. Studies of protein localization can clarify pathogenesis and find treatments. As protein subcellular localization has a very important position in the field of biology, the research in this area is extremely active. Most of the existing protein sub cellular localization methods are more suitable for single-site sub cellular localization. This paper proposed an algorithm based deep convolution neural network which is suit for multi-site protein subcellular localization and the algorithm is implemented on the human protein database to verify and analyze the performance. In order to further improve the classification result of the algorithm, it was combined ensemble learning and features fusion. It can be inferred from experiments that the proposed algorithm is effective in multi-site protein subcellular localization and the overall correct rate of classification is 59.13% which is higher than SAE, SVM and RF. The algorithm proposed in this paper is more uniform and less affected by the number of samples. When the data samples are different, the classification results will have a certain impact, but the overall classification is good. Besides ensemble learning and features fusion are effective for improving classification result.

Keywords: Multi-site subcellular localization, Deep convolution neural network, Ensemble learning, Features fusion

Introduction

The basic structure and functional unit of organism is cells. The internal structure of cells is complex and exquisite and each part of the structure which is highly ordered and interconnected has unique functions. It can be divided into many different kinds based on location and biological function, for example, nucleus, cell membrane, membrane, cytoskeleton, etc. The internal structure is commonly mentioned as subcellular [1]. Specific proteins can be found in subcellulars while they provide surroundings for proteins to fulfill their life activities and perform their functions. Proteins are

transported into designated subcellulars under the guidance of protein sorting signals after translation and synthesis. Proteins can participate in various life activities of cells through this process. It is called protein subcellular localization. The experiments show that the function of proteins closely related to their subcellular location and synthesized proteins must be in appropriate subcellulars to function properly [2]. Therefore, whether a protein can be correctly delivered to the corresponding subcellular is critical to its function. Deviated protein delivery sites are primary causes of human disease [3]. As medical research has found that abnormal subcellular localization

of proteins occurs in the process of cell pathogenesis [4], it can provide a theoretical basis for elucidating the mechanism of disease occurrence to understand the relationship between disease and protein subcellular abnormal location [5]. Further, molecular markers for early diagnosis of diseases can be obtained by abnormally located proteins using them as molecular defects in the design of new drugs to find a cure for diseases [6].

As protein subcellular localization has a very important position in the field of biology, the research in this area is extremely active. It is a time consuming task to determine the location of protein subcellulars relied solely on biological experiments and it is unrealistic with the rapid increase of data in protein databases [7]. In recent years, people have begun to use machine learning methods to achieve new protein subcellular localization [8-10]. The goal of this approach is to assign unknown protein sequences to known subcellular locations by machine learning and the essence is to learn the biological properties of unknown proteins by collecting information about known proteins.

Initial research focused on single-site protein subcellular localization, but it has been found that most proteins are located at multiple sites along with research in the field of biology [11]. In recent years, there have been studies on multi-site protein subcellular localization, but this research area is still in its infancy with great challenges. There are two main reasons. On the one hand, the research on the subcellular biological function of multi-site proteins is still limited and there are fewer protein sequences with multi-site subcellulars in the protein database. On the other hand, traditional machine learning methods, such as support vector machine (SVM), K-nearest neighbor model (KNN) and artificial neural network (ANN) [12] are more suitable for single-site protein subcellular localization. Although they can achieve a good classification result on single-site protein subcellular localization [13], the performance on multi-site protein subcellular localization is not satisfactory. This is because in the multi-site protein subcellular localization problem, the relationship between the predicted problem samples and the labels is no longer a traditional "one-to-one" relationship, but a "one-to-many" relationship. Existing machine learning methods are not directly applicable to multi-label classification problems. If these methods are applied to multi-site protein subcellular localization, they need to be changed. In addition, multi-site protein subcellular localization with the complex internal relationship between different sites

is more complicated than single-site protein subcellular localization [14]. In order to improve the accuracy of classification, the intrinsic relevance between different sites also needs to be properly considered.

The conventional approach is to treat protein sequences located at multi-sites as different protein sequences for processing in the research of extending protein subcellular localization problems to multiple sites. This treatment is relatively simple, but does not take into account the intrinsic relevance between proteins located at different sites. The hypothesis that the same protein localized to different subcellular locations is multiple proteins will inevitably has an impact on the classification results. In order to solve this problem, an ensemble learning algorithm that used the Randomk-labelsets (RAKEL) method [15] for multi-site protein subcellular classification was proposed in this paper. In addition, a method of different features fusion was adopted in order to improve the final classification results.

At present, the machine learning algorithms for multi-site protein subcellular localization include multi-label k-nearest neighbor algorithm (ML-kNN) [16], multi-label support vector machine algorithm (Rank-SVM) [17] and random forest algorithm (RF) [18]. Compared with these algorithms, the proposed algorithm has the following four characteristics. Firstly, it designed a deep convolutional neural network structure that is effective in multi-site protein subcellular localization. Secondly, the method applied RAKEL for protein subcellular localization for the first time and the integrated classifiers were formed to improve the accuracy. Thirdly, the features fusion was used to further improve the performance of algorithm. The fourth is to implement the proposed algorithm on the human protein database to verify and analyze the classification results.

Materials and Methods

Dataset

The human protein dataset used in the experiment was obtained from UniProtKB [19] and it was screened by the team of Professor Zou Quan of Tianjin University. In order to ensure the concentration of data statistics, the dataset counted the frequency of subcellular positions of the remaining protein sequences after eliminating the repeated protein sequences and only retained 10 subcellular locations of the most frequently occurring from nearly 200 protein subcellular positions. The final dataset contains only 10 subcellular positions of the human protein sequences. The subcellular positions

in the dataset are cytoplasm, nucleus, cell membrane, membrane, secreted, cytoskeleton, cell projection, endoplasmic reticulum membrane, cell junction and mitochondrion. The most frequently occurring protein subcellular location is the nucleus. There are 3,520 times. In addition, the dataset reduced sequence similarity by CD-HIT program in order to avoid homology differences. The final dataset contains a total of 9858 human protein sequences and two of the proteins in the dataset have $\leq 70\%$ sequence similarity. The length of protein sequences in the dataset varies greatly. The shortest protein sequence contains 16 amino acid residues and the longest protein sequence contains 34,450 amino acid residues. The number of protein sequences with length below 1000 is 8,396, accounting for 85.17% of whole dataset. The protein sequences in the dataset may contain more than one subcellular position and they contain up to seven protein subcellular positions. There are 3013 sequences containing two or more protein subcellular positions in the dataset and this ratio reaches 30.56% of the entire dataset.

In this paper, multi-site protein subcellular localization was studied, but it should be noted that this does not mean that all protein sequences in the dataset must contain two or more subcellular locations. At the beginning of the experiment, the number of protein subcellular locations specifically contained in the protein sequences was not known. In traditional methods, a protein sequence is considered to be N protein sequences if it may occur at N different subcellular locations. The algorithm mentioned in this paper is more difficult compared with the above mentioned methods, but it considers the actual situation to obtain more practical classification results.

The number of subcellular locations in the protein sequences of the human protein dataset used in this paper is as follows: 6845 proteins contain 1 protein subcellular location, 2209 proteins contain 2 protein subcellular locations, 583 proteins contain 3 protein subcellular locations, 158 proteins contain 4 protein subcellular locations, 48 proteins contain 5 protein subcellular locations, 11 proteins contain 6 protein subcellular locations and 4 proteins contain 7 protein subcellular locations. The number of subcellular locations corresponding to these protein sequences is shown in table 1.

Deep convolutional neural network

An important characteristic of the deep convolutional neural network (DCNN) is that complex nonlinear

transformations can be used to select different categories of features without specifying features [20-23]. This characteristic differs from the shallow machine learning algorithm which requires specific requirements for the extracted features. Due to the complex linkages between different classes in multi-site protein subcellular localization, it is difficult to find a feature extraction method that is fully applicable to the classification algorithm using traditional shallow machine learning algorithms which resulting in unsatisfactory classification results. The characteristics of deep convolutional neural network make it more competent for complex classifications and can achieve better results in multi-label classification.

Deep convolutional neural network is similar in structure to traditional convolutional neural network, but it has its own unique characteristics. DCNN shares the same weight between nodes in the same layer. This characteristic makes it easier to select network parameters and better handle large-scale data. There are two difficulties that need to be solved in the classifier using neural networks [24]. The first is to determine the structure of the neural network which is often selected by the specific problem to be solved. The more complex the structure of the neural network, the closer it is to complex classification problems while the computational complexity will increase rapidly. Compared with traditional convolutional neural network, deep convolutional neural network has more network layers and the number of neural network layers directly affects the final classification results. The second is the selection of main parameters. If the adjustment speed of the network parameters is too fast in the process of neural network training, or the adjustment direction is wrong, it will affect the classification results of the final classifier. In general, we hope that the adjustment speed of parameters is fast at training begins and it can be slowed as training time increases to avoid missing the optimal solution.

In the structural design of DCNN, the nodes of the input layer are closely related to the dimension of the features. Several different features fusion methods were adopted in this paper and the different number of input nodes were selected respectively. As the positioning of 2 sites was performed each time in this experiment, the number of output layer nodes of the DCNN was selected to be 2. These two structures are determined based on specific problems easily. In the design of other layers of the neural network, the convolutional layer and the pooled layer were spaced apart which is also a characteristic of

the DCNN. The convolutional layer has a local perceptual effect and the use of it can focus on extracting more detailed features to better distinguish small differences between features. In addition, the characteristics of multi-core convolution layer give different weights to the features which help to highlight important features. The main role of the pooling layer in the algorithm proposed in this paper is to reduce the dimension while retaining the main features and avoid the over-fitting caused by too many parameters.

The difference between the output value and the actual value was measured by the quadratic cost in the experiment and can be expressed as follows:

$$C = \frac{1}{2N} \sum_x \|y(x) - a(x)\|^2 \quad (1)$$

Where C is the cost function, x is the sample, $y(x)$ is the actual value, $a(x)$ is the output value, and N is the total number of samples.

Use the gradient descent method to adjust the weight parameters:

$$\Delta\omega = (a(x) - y(x))\sigma'(z)x \quad (2)$$

$$\Delta b = (a(x) - y(x))\sigma'(z) \quad (3)$$

Where z is the input of the neuron, σ is the activation function, $\Delta\omega$ is the weight change and Δb is the offset change.

$$\omega^{(n+1)} = \omega^{(n)} - \frac{\eta}{N} * \Delta\omega \quad (4)$$

$$b^{(n+1)} = b^{(n)} - \frac{\eta}{N} * \Delta b \quad (5)$$

Where η is the learning rate, $\omega^{(n)}$ is the weight of the nth iteration, and $b^{(n)}$ is the offset of the nth iteration.

Random k-labelsets (RAkEL)

The multi-site protein subcellular localization is essentially a multi-label classification problem. Unlike the single-label classification problem, it is necessary to consider the intrinsic association between the individual labels in the multi-label classification and use the correlation between them to improve the classification accuracy [25]. At present, the multi-label classification methods can be summarized into two categories. One is to convert multi-label classification problem to single-label classification problem to process. The other is to modify the single-label classification classifier to suitable

for multi-label classification problem [26].

In the research of protein subcellular localization, most of algorithms are still limited to single-site protein subcellular localization. When dealing with multi-site protein subcellular localization, the traditional treatment is to directly process the protein sequences, that is, a protein sequence is dealt to some repetitive protein sequences when it belongs to different sites. This simple treatment neglects the intrinsic link when the protein sequences belong to different sites, which inevitably reduces the accuracy of the final classification results.

The Randomk-labelsets (RAkEL) algorithm is a good multi-label classification method [27] which adopts the idea of ensemble learning. The algorithm is as follows:

A set of labels L is provided, all possible k-label subsets are represented by L^k , where $L^k(q)$ represents the qth k-tag subset, and $L^k(q) \subseteq L, |L^k(q)| = k$. In the training process, the multi-label training set is first constructed as a multi-category training set.

$$D_{L^k(q)} = \left\{ \left(x_i, \sigma_{L^k(q)}(y_i \cap L^k(q)) \right) \mid 1 \leq i \leq n \right\} \quad (6)$$

Where $\sigma_{L^k(q)}$ is a mapping function that maps each subset of the label subset $L^k(q)$ to a different category, i represents the sample number in the training set, and n represents the number of samples in training set.

The set of categories covered by the multi-class data set $D_{L^k(q)}$ is:

$$\Gamma(D_{L^k(q)}) = \left\{ \sigma_{L^k(q)}(y_i \cap L^k(q)) \mid 1 \leq i \leq n \right\} \quad (7)$$

Get the corresponding multi-label classifier on the multi-class data set:

$$h_{D_{L^k(q)}} : \rightarrow \Gamma(D_{L^k(q)}) \quad (8)$$

The RAkEL algorithm combines the idea of ensemble learning to process multi-label data. It first establishes an integration framework containing m multi-classifiers and then determines the final classification result by means of classifier voting. The specific process of the algorithm is as follows: Firstly, the corresponding multi-classifiers $h_{D_{L^k(q)}}(\bullet)$ are obtained on m random k-tag subsets $L^k(q) (1 \leq q \leq m)$ respectively. According to the selection method of the k-tag subset $L^k(q)$, there are **RAkEL_dRAkEL_d** which has the disjoint label power set and **RAkEL_oRAkEL_o** which has the overlapping label power set. The fundamental difference between the two methods is whether the labels contained in

$L^k(q)$ overlap when q is not the same. This paper used **RAkEL** method.

Then, when facing an unknown sample, the following two quantities need to be calculated for each l_j tag:

$$\tau(x, l_j) = \sum_{q=1}^m [l_j \in L^k(q)] \quad (9)$$

$$\mu(x, l_j) = \sum_{q=1}^m \left[l_j \in \sigma_{L^k(q)}^{-1} \left(h_{D_{L^k(q)}} \right) \right] (1 \leq j \leq c) \quad (10)$$

Where c is the number of tags included in the tag set L . $\tau(x, l_j)$ calculates the maximum number of votes received by the tag l_j in the integration framework, that is, the sum of the occurrences of l_j tag. $\mu(x, l_j)$ is the actually number that received in the integration framework of tag l_j . Therefore, the subset of prediction labels for this sample is:

$$y = \left\{ l_j \mid \frac{\mu(x, l_j)}{\tau(x, l_j)} > 0.5, 1 \leq j \leq c \right\} \quad (11)$$

It can be seen that the RAKEL classification algorithm promotes the training of the classification model by establishing an ensemble learning method utilizing the high-order correlation between multiple tags in the tag subset. The algorithm based on RAKEL proposed in this paper not only considers the difference between protein subcellulars belonging to different sites, but also uses the RAKEL classification algorithm to train the classifiers with the intrinsic relationship between different sites of protein subcellulars to improve the positioning accuracy of multiple sites.

Features fusion method

The feature extraction process can be divided into three aspects of characteristic statistics according to the different methods of protein sequence feature extraction. One is the frequency of occurrence of amino acid composition. The second is the sequence information of amino acid composition. The third is the physicochemical properties of amino acid composition [28]. Different feature extraction methods reflect the properties of different aspects of the protein sequences. According to the different protein dataset, the selection of appropriate feature extraction methods is of great significance for the experimental results [29]. The human protein dataset studied in this paper contains a large number of sequences, the lengths between the different sequences vary widely and the same sequence may be located at multiple sites. The complexity of the dataset

makes it difficult to find a feature extraction method that fully reflects the characteristics of protein sequences belonging to different sites.

Since a single feature extraction method representing the information of protein sequences has limitations [30], the features fusion method was used in this paper. Obviously, choosing the suitable features to fuse in the right way is a very critical issue [31]. In this paper, the features extracted by different feature extraction methods were added by dimension and the fusion features were obtained. Through experiments, it had been known that proteins at different sites have different characteristics regardless of the frequency of their amino acid appearance or physical and chemical properties. The significance of features fusion is to fully reflect these differences. Therefore, the features fusion method used in this paper is effective to represent complete protein sequence characteristics. However, the features fusion by the way of dimension addition brings two obvious problems [32]. Firstly, the size of the data to be processed is greatly increased. Secondly, there is a large amount of redundancy between the features extracted by different methods and the redundancy may reduce the accuracy of the classification results.

In order to solve the two problems mentioned above, this paper compared the classification results of different features in the algorithm. The four features mainly used were dipeptide, pseudo amino acid, physicochemical properties of amino acids and entropy density.

Dipeptide [33] reflects the nature of a protein sequence by calculating the frequency of occurrence of amino acids. The eigenvectors can be expressed as:

$$V = [f_1, f_2, \dots, f_{400}]^T \quad (12)$$

Where $f_i = \sum_{i=1}^n R_i / (N-1)$, $i=1, 2, \dots, 400$, R_i represents the frequency of occurrence of the amino acid pair, and N is the amino acids number of the protein sequence.

The advantage of the dipeptide is that it reflects the frequency of amino acid occurrence and takes into account the positional information of adjacent amino acids. Its shortcoming is that there are many dimensions and there may be a large amount of redundant information. The redundant information is very easy to cause the complexity of the operation time and may affect the classification accuracy.

The pseudo amino acid (Pseaac) [34] not only considers the composition information of the protein sequence, but also contains the positional order information of the sequence. Its eigenvector is expressed as:

$$Y = [y_1, y_2, \dots, y_{20}, y_{20+1}, \dots, y_m]^T \quad (m < N) \quad (13)$$

Where N represents the number of amino acids in the protein sequence.

The n -level positional information between amino acid residues is expressed as:

$$\lambda_n = \frac{1}{L-n} \sum_{i=1}^{L-n} X_{i,i+n} \quad (n < L) \quad (14)$$

The $X_{i,i+n}$ in the formula can be calculated from:

$$(15) \quad X_{i,i+n} = \frac{1}{6} \left\{ \begin{aligned} &[H_1(R_{i+n}) - H_1(R_i)]^2 + [H_2(R_{i+n}) - H_2(R_i)]^2 + [Pk_1(R_{i+n}) - Pk_1(R_i)]^2 \\ &+ [Pk_2(R_{i+n}) - Pk_2(R_i)]^2 + [PI(R_{i+n}) - PI(R_i)]^2 + [M(R_{i+n}) - M(R_i)]^2 \end{aligned} \right.$$

Where $H_1(R_i)$, $H_2(R_i)$, $Pk_1(R_i)$, $Pk_2(R_i)$, $PI(R_i)$ and $M(R_i)$ represent the properties of the i th amino acid residue R_i in the protein sequence. They represent respectively the hydrophobic value, the hydrophilic value, $Pk_1(-COOH)$, $Pk_2(-NH_3)$, PI and side chain molecular weight values.

After calculating the m -level correlation factor, the pseudo-amino acid characteristics can be derived from the following formula:

$$y_j = \begin{cases} \frac{f_j}{\sum_{i=1}^{20} f_i + \omega \sum_{n=1}^m \lambda_n} & (1 < j < 20) \\ \frac{\omega \lambda_{j-20}}{\sum_{i=1}^{20} f_i + \omega \sum_{n=1}^m \lambda_n} & (20+1 < j < 20+m) \end{cases} \quad (16)$$

Where ω is weight factor, generally taking a value of 0.05.

The physicochemical properties of amino acids [35] are characterized by the classification of seven physical and chemical properties of amino acids into three classes: polar, neutral and hydrophobic. For a certain protein sequence, the eigenvectors of the model are obtained by statistically calculating the frequency of occurrence of polar amino acids, neutral amino acids, and hydrophobic amino acids in the protein sequence. The frequency of polar amino acids, neutral amino acids, or hydrophobic amino acids appears in a certain protein sequence can be obtained by the following formula:

$$f_{i,polar} = \frac{N_{polar}}{N} \quad (17)$$

$$f_{i,neutral} = \frac{N_{neutral}}{N} \quad (18)$$

$$f_{i,hydrophobi} = \frac{N_{hydrophobi}}{N} \quad (19)$$

Where the value of i from 1 to 7, and N represents the total number of amino acid residues in the protein sequence. The model ultimately yields a 21-dimensional eigenvector.

The entropy density [36] is obtained by calculating the information entropy of 20 amino acid residues and is also the composition information of the protein sequence. The eigenvector can be expressed as:

$$S = [s_1, s_2, \dots, s_{20}]^T \quad (20)$$

Where s_i can be expressed as:

$$s_i = -\frac{1}{H(S)} f_i \log f_i \quad (21)$$

$$H(S) = -\sum_{i=1}^{20} f_i \log f_i \quad (22)$$

Where f_i is the frequency at which amino acid residues appear in the protein sequence?

Operation algorithm

The multi-site protein subcellular localization is restricted by imperfect protein sequence annotation in the protein database which affects the accuracy of the final classification results and the generalization of the classification. The selection of a suitable dataset is very important. The algorithm proposed in this paper was implemented on a multi-site human protein dataset and the experimental results were analyzed.

The proposed algorithm mainly includes four parts as shown in figure 1. They are protein sequence data accessing, features extracting, deep convolutional

neural network constructing and multi-label classifier integrating. Protein sequence data accessing mainly includes the setting up a dataset of protein sequences, reading data, and storage of classification results. The features extracting part mainly includes extracting the features of the protein sequences and obtaining the same dimension features from different protein sequences. There are many methods to extract the features of protein sequences. In this paper, four feature extraction methods were selected and the fusion features were obtained by different combinations. The constructing of deep convolutional neural network (DCNN) is a primary part of the algorithm. It involves the design of deep convolutional neural network structure and the determination of hyperparameters. Multi-label classifier uses multiple classifiers to form an ensemble learning framework. The result of the judgment is obtained by voting.

Figure 1: Algorithm framework.

The above four parts were combined in three ways in the experiments. The specific implementation process of the proposed algorithm is as follows.

Multi-site protein subcellular localization based on DCNN

- Step 1: Establish and store protein sequence dataset;
- Step 2: Extract features of the protein sequence in the training set;
- Step 3: Build and train a deep convolutional neural network suitable for multi-site protein subcellular localization;
- Step 3.1: Set the input layer, output layer and hidden layer nodes of the deep convolutional

neural network according to the problem size; Initialize randomly the DCNN parameters and build a multi-label classification ensemble framework;

- Step 3.2: Input the features of the protein sequence in the training set one by one and train the deep convolutional neural network to obtain the positioning result of each protein sequence;
- Step3.3: Use the cost function to determine whether the obtained result meets the requirements. If the requirements are met, the training process ends. If the requirements are not met, first determine whether the maximum number of times is reached. If the maximum number of times is reached, you need to go back to step 3.1 to re-establish the setting of the deep convolutional neural network. If the maximum number of times is not reached, you need to return to step 3.2 to adjust the hyperparameters of the deep convolutional neural network;
- Step 4: Extract features of the protein sequence in the test set and input the extracted features into the trained deep convolutional neural network classifier to predict the protein subcellular position.

Multi-site protein subcellular localization based on DCNN and ensemble

- Step 1: Establish different classifier combinations based on different subcellular locations;
- Step 2: Obtain the classification results of each classifier according to steps 1 to 4 in 2.5.1;
- Step 3: According to the algorithm of RAKEL and the results of each classifier, establish a multi-label classification ensemble framework and the final classification result is obtained.

Multi-site protein subcellular localization based on DCNN and features fusion

- Step 1: Establish and store protein sequence dataset;
- Step 2: Extract and fuse features of the protein sequence in the training set;
- Step 3: Build and train a deep convolutional neural network as Step3 in 2.5.1;
- Step 4: Extract and fuse features of the protein sequence in the test set; Input the fused features into the trained deep convolutional neural network classifier to predict the protein subcellular position.

The extraction method of features has a great influence on the final classification results. Whether

the selected features fully reflect the characteristic of protein sequences is crucial for achieving good positioning results. However, the most important to the final classification results is the design of the classifier. Choosing the appropriate classifier is the pivotal to improving the accuracy of the classification results.

Experiments

In the process of multi-site protein subcellular localization, the 2-label RAKEL classification method was used in this experiment that is two sites were located each time.

Multi-site protein subcellular localization based on DCNN

Datasets selection: In the experiment, the method of 10-fold cross validation was adopted. A protein dataset containing 9858 sequences was randomly divided into 10 subsets. In each experiment, 9 subsets were taken as the training set and the remaining 1 subset was used as the test set. In addition, 30% of the sequences in the entire protein dataset were randomly extracted as a validation set for the training process.

As the classification results will be poor in the case of unbalanced samples, samples balance was considered in the selection of the dataset. In the process of randomly dividing the data subset and extracting the verification set, different kinds of samples were extracted corresponding proportions. For example, in the process of located the class-0 cytoplasm and class-1 nucleus, there might be 4 cases, class-0 and class-1 do not exist, class-0 exists and class-1 does not exist, class-0 does not exist and class-1 exists, class-0 and class-1 exist. In the experiment, each of them was randomly divided into 10 subsets. One of the subsets was extracted to constitute test set and the remaining 9 subsets constituted train set. The same principle was applied to the process of extracting the validation set.

Based on the initial training set, the number of categories with the largest number of the four classification categories was used as a standard and the category with smaller number of samples was resample. This was also for the purpose of sample balance. In general, the most samples were two sites do not exist.

The architecture and parameters of DCNN:

The number of input layer nodes of the DCNN is directly related to the selected feature. In this set of experiments, PseAAC feature was selected. Since the length of the

shortest amino acid sequence in the protein dataset was 16, the m was selected to be 15 in the process of calculating the PseAAC and finally formed a 35-dimensional PseAAC feature. The number of input layer nodes for DCNN was selected to be 35.

The number of output layer nodes of the DCNN is related to the type of samples to be determined. As two sites was located each time in this experiment, the number of output layer nodes of the DCNN was selected to be 2.

Protein subcellular localization is the process of classifying data sequences. In the experiment, it is hoped that the convolutional layer of DCNN can be used as much as possible to further amplify the difference between different samples. Therefore, the iterative structure of three convolutional layers to connect one pooling layer was used. The specific architecture is shown in table 2.

According to the data distribution characteristics, the convolution layer selected the lrelu as activation function, the non-output forward layer selected the tanh as activation function and the output layer selected the sigmoid as activation function.

The cost function used the least mean squared cost function. For the specific calculation method, see equation (1). The learning rate was changed in a discontinuous manner with 100,000 iterations per iteration and it was changed from large number to small number.

Measurements: To quantify the final prediction results, three metrics were used. ACC_i^j is the accuracy of the j-th experiment of i-site and $ACC(i)$ is accuracy of the i-point after 10-fold cross validation. OA is the final accuracy which is the average accuracy of all the points. The calculation formulas of the above three metrics are as follows:

$$ACC_i^j = \frac{C_i^j}{N_i^j} \quad (23)$$

$$ACC(i) = \frac{\sum_{j=1}^{10} C_i^j}{\sum_{j=1}^{10} N_i^j} \quad (24)$$

$$OA = \frac{\sum_{i=1}^{10} \sum_{j=1}^{10} C_i^j}{\sum_{i=1}^{10} \sum_{j=1}^{10} N_i^j} \quad (25)$$

Where C_i^j represents the correct number in the jth experiment which should be one for i-site and N_i^j is the

number of i-site that should be one in the jth experiment.

Multi-site protein subcellular localization based on DCNN and ensemble learning

Datasets selection: The classification results obtained using the two-label RAKEL method will be not exactly the same when the combination of the labels is different. This is due to the impact of different data distribution. In this set of experiments, the focus was on the study of the effects of different combinations of labels on the classification results. In order to comprehensively measure the classification results of a classifier and eliminate the difference of the dataset, the whole protein dataset was selected as the test set. The entire protein dataset was resampled using the largest number of categories as a standard to form the training set. Thirty percent of the entire dataset was randomly selected as the validation set for the training process.

RAKEL: In order to eliminate the influence of different integration methods on the classification results, the RAKEL multi-label classification method was adopted. The classification result of a certain point was obtained by a separate classifier firstly and then the final result was determined by integrating the classifier to vote. This is actually a way of ensemble learning.

Multi-site protein subcellular localization based on DCNN and features fusion

Datasets selection: In this group of experiments, the focus was on studying the effects of different amino acid feature selections on classification results. The two-label RAKEL method was also used in the experiment. On the selection of the two sites, class-0 cytoplasm and class-1 nucleus were selected. Because the two categories have the largest number of samples and the numbers are close. On the selection of the dataset, the entire protein dataset was selected as the test set. The entire protein dataset was resample using the largest number of categories as a standard to form the training set. Thirty percent of the entire dataset was randomly selected as the validation set for the training process.

Features selection: The features of protein sequences are divided into the frequency of amino acids, the physical and chemical properties of amino acids, amino acid sequence information and annotation information. Different kinds of features reflect the characteristics of protein sequences from different aspects. This group of experiments mainly studied the effects of different sequence features and multiple feature fusions on the

classification results. In the experiment, four features were selected and they were dipeptide, pseudo amino acid, physicochemical properties of pseudo amino acid and entropy density. Among them, the dipeptide reflects the frequency of amino acid pairs appearing in the sequence. The pseudo amino acid includes both the frequency of occurrence of amino acid residues and their physical and chemical properties. The physicochemical properties of amino acids contain the physicochemical properties of seven amino acid residues. The entropy density uses the concept of entropy to measure the frequency of occurrence of amino acid residues.

Results

Multi-site protein subcellular localization based on DCNN

For the purpose of balancing the samples, the two sites selected in the classifier are close in number. Since the rounding was used in the process of randomly dividing the subset, the number of samples contained in the final test set was somewhat different from the number of samples in the protein dataset. For example, for class-0, its number in the overall protein dataset is 3374 and the number of tests performed is 3368. This is because the number was rounded down during the division of 10 subsets. In some cases, the number of tested proteins was greater than the number of such categories in the protein dataset, such as class-9. Its number in the protein database is 392 and the number of test proteins is 440, because the number was rounded up and resampled during the division of 10 subsets. The specific results are shown in figure 2 and table 3.

Multi-site protein subcellular localization based on DCNN and ensemble learning

In the experiment, three classifier integration and five classifier integration were carried out respectively. The specific results are shown in figure 3. Integrating three classifiers contained classifier-1, classifier-2 and classifier-3 in the figure. Integrating five classifiers contained classifier-1, classifier-2, classifier-3, classifier-4 and classifier-5 in the figure.

Multi-site protein subcellular localization based on DCNN and features fusion

In the experiment, tests of individual features and fusion features were performed separately. The fusion features contain a fusion of two to four features, the specific results are shown in figure 4 and table 4.

Figure 2: Test results of 10-fold cross validation of multi-site protein subcellular localization based on DCNN.

Figure 3: Test results of multi-site protein subcellular localization based on DCNN and ensemble learning.

Figure 4: Localization based on DCNN and ensemble learning.

Table 1: Protein subcellular location distribution of locative proteins in the dataset.

Classes	Subcellular locations	Number of proteins
Class-0	Cytoplasm	3374
Class-1	Nucleus	3520
Class-2	Cell membrane	1611
Class-3	Membrane	1677
Class-4	Secreted	1090
Class-5	Cytoskeleton	800
Class-6	Cell projection	521
Class-7	Endoplasmic reticulum membrane	550
Class-8	Cell junction	443
Class-9	Mitochondrion	392
Total		13978

Table 2: The details of proposed DCNN architecture for this work.

Layers	Type	Number of feature maps	Number of neurons in the layer	The size of the kernel involves to form each feature map	Stride	Activation function
0	Input	–	1×35	–	–	tanh
1	Convolution	4	1×128	1×4	1×1	Irelu
2	Convolution	4	1×500	1×4	1×1	Irelu
3	Convolution	4	1×1988	1×4	1×1	Irelu
4	Pooling	4	1×994	1×2	1×1	none
5	Convolution	4	1×3972	1×2	1×1	Irelu
6	Convolution	4	1×15884	1×2	1×1	Irelu
7	Convolution	4	1×63532	1×2	1×1	Irelu
8	Pooling	4	1×31766	1×2	1×1	none
9	Convolution	2	1×63530	1×2	1×1	Irelu
10	Convolution	2	1×127058	1×2	1×1	Irelu
11	Convolution	2	1×254114	1×2	1×1	Irelu
12	Pooling	2	1×254114	1×1	1×1	none
13	Fully-connected	–	1×32	–	–	tanh
14	Fully-connected	–	1×16	–	–	tanh
15	Output	–	1×2	–	–	sigmoid

Figure 4: Test results for different categories of multi-site protein subcellular localization based on DCNN and features fusion

In figure 4 and table 4, the Pseaac represents pseudo amino acid feature. The Entropy represents the entropy density feature. The Physicochemical represents the physicochemical properties of amino acids feature. The “dipeptide + Pseaac” feature represents the fusion of dipeptide and pseudo amino acid features, and so on. The correct number of subcellular localization is the number of samples in the class-0 cytoplasm and the class-1 cell

nucleus simultaneously located correctly.

Discussion

Through the first set of experiments, the results of the “10-fold cross validation” of ten categories were respectively obtained and the following conclusions can be obtained. Firstly, the proposed algorithm is effective in multi-site protein subcellular localization. The overall correct rate of classification is 59.13%, which is higher than SAE, SVM and RF [37], as shown in figure 5. Specific to

Table 3: Test results of multi-site protein subcellular localization based on DCNN.

Classes	Subcellular locations	Number of proteins	Number of correctly predicted proteins	ACC(%)
Class-0	Cytoplasm	3368	2446	72.62
Class-1	Nucleus	3518	2299	65.36
Class-2	Cell membrane	1610	786	48.82
Class-3	Membrane	1670	783	46.89
Class-4	Secreted	1150	650	56.52
Class-5	Cytoskeleton	870	523	60.11
Class-6	Cell projection	700	266	38
Class-7	Endoplasmic reticulum membrane	730	382	52.33
Class-8	Cell junction	490	231	47.14
Class-9	Mitochondrion	440	235	53.41
Total		14546	8601	59.13

Table 4: Test results of multi-site protein subcellular localization based on DCNN and features fusion.

Feature	The correct number of subcellular localization	ACC(%)
Dipeptide	4186	42.46
Pseaac	4768	48.37
Entropy	4462	45.26
Physicochemical	3738	37.92
Dipeptide+Pseaac	4716	47.84
Entropy+Pseaac	5082	51.55
Physicochemical+Pseaac	4774	48.43
Physicochemical+Entropy	4592	46.58
Dipeptide+Entropy	4841	49.11
Dipeptide+Physicochemical	4972	50.44
Physicochemical+Entropy+Pseaac	4442	45.06
Dipeptide+Physicochemical+Entropy	4652	47.19
Dipeptide+Physicochemical+Pseaac	4971	50.43
Dipeptide+Entropy+Pseaac	5087	51.6
Dipeptide+Physicochemical+Entropy+Pseaac	5017	50.89

each class, it is better than the SAE. In particular, SAE has a poor classification results in class-5, class-6 and class-8 with a small sample size. The algorithm classification results in these classes are much better than SAE. The Statistical significance of correct rate between the proposed algorithm and SAE is 0.000331. The algorithm proposed in this paper is more uniform for each class and less affected by the number of samples, as shown in table 5. Secondly, when the data samples are different, the classification results will have a certain impact, but the overall classification is good, as shown in figure 6.

Table 5: Test results of DCNN and SAE.

Classes	Subcellular locations	DCNN ACC(%)	SAE ACC(%)
Class-0	Cytoplasm	72.62	45.9
Class-1	Nucleus	65.35	53.6
Class-2	Cell membrane	48.82	31.2
Class-3	Membrane	46.89	30.6
Class-4	Secreted	56.52	47.8
Class-5	Cytoskeleton	60.11	4.9
Class-6	Cell projection	38	0.38
Class-7	Endoplasmic reticulum membrane	52.33	21.3
Class-8	Cell junction	47.14	1.35
Class-9	Mitochondrion	53.41	23
Total		59.13	37.4

Figure 5: Overall accuracy of different classifiers.

The second set of experiments was to select a different combination of categories for the RAKEL classification of the two labels. In the process of classification, the RAKEL multi-label classification not only distinguishes the case where a protein sequence should exist at a certain site, but also distinguishes the case where the protein sequence should not exist at the site. The accuracy of the classification needs to be considered in both cases. Firstly, through the experimental results, it can be seen that the choice of different combinations of categories directly affects the final classification results, which reflects the impact of sample distribution on the classification results. Secondly, increasing the number of integrated classifiers helps to improve the accuracy of classification. For each category, the results of using five classifiers are better than those using three classifiers, as shown in figure 7.

Figure 7: Comparison of test results of three classifiers integration and five classifiers integration. **Figure 6:** Ten-fold cross validation results for each category.

The third set of experiments was to select four different features for features fusion and the following conclusions were obtained. First, different features have an impact on the final classification results which is obviously related to whether the selected features can fully reflect the characteristics of each category. For example, for class-0 cytoplasm, the best classification accuracy is 70.44% through the fusion of dipeptide, physicochemical properties and pseudo-amino acid features, and the classification accuracy obtained by physicochemical properties is only 53.96%. For class-1 nucleus, the best classification accuracy obtained by fusion of dipeptide, entropy density and pseudo amino acid is 73.84% and the worst classification rate by dipeptide is only 60.74%. Second, different features fusion methods have different performance for each category. For example, the features fusion mode of dipeptide, physicochemical properties and entropy density is better for classification of class-0 cytoplasmic, but for class-1 nucleus, the performance is poor. Third, the results obtained by using individual features for classification are generally

inferior to those obtained using fusion features. For example, in the classification of class-0 cytoplasm, the worst two classifications results using individual features. Fourth, when features are fused, not the more features, the more sufficient they reflect the characteristics of a certain category. For example, in the classification results of the category-1 nucleus, the performance of three features fusion is worse than the classification results obtained by the fusion of two features. In this group of experiments, the results of two sites were correctly judged were considered as correct for the samples. The highest accuracy of sample judgment is the result of the fusion feature of dipeptide, entropy density and pseudo-amino acid, and it is 51.60%. The worst accuracy of samples judgment is obtained by the physical and chemical properties and it is 37.92%. The specific results are shown in figure 8. The PseAAC represents pseudo amino acid feature. The Entropy represents the entropy density feature. The Physicochemical represents the physicochemical properties of amino acids feature. The “dipeptide + PseAAC” feature represents the fusion of dipeptide and pseudo amino acid features, and so on.

Figure 8: ACC based on DCNN and features fusion.

The human protein dataset used in this experiment is relatively new. As the number of amino acid sequences contained in the dataset is more than commonly used dataset and the length of the amino acid sequence is different, the human protein dataset is selected to conducive the ability of DCNN to process large-scale data. In order to further verify the generalization ability of DCNN to process different data, this paper verifies on the commonly used Gneg-mPLOC dataset. The classification results were compared with MLKNN and ICNN which achieved good results in multi-site protein subcellular localization on the dataset. The Dataset of Gneg-mPLOC contains 1392 different protein sequences and 1456 locative proteins. The specific results are shown in table 6.

The experimental results show that the proposed algorithm has higher correct rate of classification than MLKNN and CNN. The other two algorithms have a particularly low recognition rate for certain subcellular locations. The algorithm proposed in this paper has an average correct rate of classification for each subcellular location.

Conclusion

This paper proposed a multi-site protein subcellular algorithm based on deep convolutional neural network. Existing protein subcellular localization methods generally require manual extraction of features and they are generally process multi-site protein sequences separately as single-site protein. The algorithm proposed in this paper can better automatically select features of protein sequences and directly classify multi-site protein sequences.

In order to verify the performance of the algorithm, this paper has done three sets of experiments on the human protein dataset. The first one was to verify the average classification accuracy of the algorithm. For the ten subcellular positions of the proteins contained in the dataset, experiments were carried out by means of “10-fold cross-validation”. The experimental results showed that although the number of samples was affected, the algorithm maintained good classification accuracy in each category and the algorithm was robust. The second one was to use the ensemble learning method for classification. It was found that different classifiers have different performance and ensemble learning can achieve better classification results. The third one was to use the method of features fusion to classify. It was found that the way of features fusion helps to improve the classification results, but it is not the more the fusion features, the better classification accuracy will get.

Overall, the algorithm proposed in this paper has achieved a good classification result on protein subcellular localization. During the experiment, it was found that the small number of samples and the uneven distribution of samples greatly affected the final classification results. In the following study, we will further consider how to eliminate or reduce the impact on the final classification results in the case of insufficient sample size. In addition, the use of different structures of DCNN and hyperparameters setting can have a large impact on the classification results. The structure of DCNN and hyperparameters setting mentioned in

Table 6: Test results in Gneg-mPLoc dataset. ‘

Classes	Subcellular locations	MLKNN ACC(%)	ICNN ACC(%)	DCNN ACC(%)
Class-0	Cell inner membrane	84.38	79.71	83.12
Class-1	Cell outer membrane	45.97	52.42	89.51
Class-2	Cytoplasm	60.24	70	85.37
Class-3	Extracellular	45.11	40.6	96.99
Class-4	Fimbrium	0	40.63	90.63
Class-5	Flagellum	100	100	100
Class-6	Nucleoid	87.5	87.5	100
Class-7	Periplasm	0	37.78	91.11
Total		58.59	65.25	86.95

this paper was based on experiments. In the following study, further consideration needs to be given to how to automatically construct a suitable structure of DCNN and select appropriate hyperparameters on different datasets to achieve better classification results.

Declarations

Copyright: © Hanhan Cong et al. This is an Open Access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Competing interests

The authors declare that they have no competing interests.

Author contributions

Hanhan Cong implemented the algorithms and wrote the manuscript, Hong Liu and Yuehui Chen directed the algorithms and edited the manuscript, Yaou Zhao and Lei Wang analyzed the data.

Acknowledgements

This research is supported by the National Key Research and Development Program of China (No. 2016YFC0106000).

This research was partially supported by the National Natural Science Foundation of China (61876102, 61472232, 61272094).

References

1. Qiu J-D, Xu H-D, Suo S-B, Shi S-P, Chen X. Proteomic analysis and prediction of human phosphorylation sites in subcellular level reveal subcellular specificity. *Bioinformatics*. 2015;31(2):194-200. doi: <http://dx.doi.org/10.1093/bioinformatics/btu598>
2. Jiang W, Zuo Y, Li T, Lv Y, Yang L. Human proteins characterization with subcellular localizations. *J Theor Biol*. 2014;358:61-73. doi: <http://dx.doi.org/10.1016/j.jtbi.2014.05.008>
3. Nair R, Rost B. Protein subcellular localization prediction using artificial intelligence technology. *Methods Mol Biol*. 2008;484:435-463. doi: http://dx.doi.org/10.1007/978-1-59745-398-1_27
4. Xu Y-Y, Yang F, Zhang Y, Shen H-B. An image-based multi-label human protein subcellular localization predictor (iLocator) reveals protein mislocalizations in cancer tissues. *Bioinformatics*. 2013;29(16):2032-2040. doi: <http://dx.doi.org/10.1093/bioinformatics/btt320>
5. Wan S, Mak M-W, Kung S-Y. Gram-LocEN: Interpretable prediction of subcellular multi-localization of Gram-positive and Gram-negative bacterial proteins. *Chemometrics and Intelligent Laboratory Systems*. 2017;162:1-9. doi: <http://dx.doi.org/10.1016/j.chemolab.2016.12.014>
6. Zhou H, Yang Y, Shen H-B. Hum-mPLoc 3.0: prediction enhancement of human protein subcellular localization through modeling the hidden correlations of gene ontology and functional domain features. *Bioinformatics*. 2017;33(6):843-853. doi: <http://dx.doi.org/10.1093/bioinformatics/btw723>
7. Chou KC, Ding Y, Zhang T. Prediction of protein subcellular location using hydrophobic patterns of amino acid sequence. *Comput Biol Chem*. 2006;30(5):367-371. doi: <http://dx.doi.org/10.1016/j.compbiolchem.2006.08.003>
8. Mak M-W, Guo J, Kung S-Y. PairProSVM: Protein Subcellular Localization Based on Local Pairwise Profile Alignment and SVM. *IEEE/ACM Trans Comput Biol Bioinformatics*. 2008;5(3):416-422. doi: <http://dx.doi.org/10.1109/TCBB.2007.70256>

9. Casadio R, Pierleoni A, Martelli PL, Alhamdoosh M, Fariselli P, Savojardo C. Improving the prediction of disulfide bonds in Eukaryotes with machine learning methods and protein subcellular localization. *Bioinformatics*. 2011;27(16):2224-2230. doi: <http://dx.doi.org/10.1093/bioinformatics/btr387>
10. Wan S, Mak MW. *Machine Learning for Protein Subcellular Localization Prediction*. Boston: De Gruyter; 2015.
11. Carrie C, Kühn K, Murcha MW, et al. Approaches to defining dual-targeted proteins in Arabidopsis. *Plant J*. 2009;57(6):1128-1139. doi: <http://dx.doi.org/10.1111/j.1365-3113X.2008.03745.x>
12. Mooney C, Wang Y-H, Pollastri G. SCLpred: protein subcellular localization prediction by N-to-1 neural networks. *Bioinformatics*. 2011;27(20):2812-2819. doi: <http://dx.doi.org/10.1093/bioinformatics/btr494>
13. Lu Z, Szafron D, Greiner R, et al. Predicting subcellular localization of proteins using machine-learned classifiers. *Bioinformatics*. 2004;20(4):547-556. doi: <http://dx.doi.org/10.1093/bioinformatics/bth026>
14. Chou K-C. Some remarks on predicting multi-label attributes in molecular biosystems. *Mol Biosyst*. 2013;9(6):1092-1100. doi: <http://dx.doi.org/10.1039/c3mb25555g>
15. Cai X, Wang H, Huang H, Ding C. Joint stage recognition and anatomical annotation of drosophila gene expression patterns. *Bioinformatics*. 2012;28(12):i16-i24. doi: <http://dx.doi.org/10.1093/bioinformatics/bts220>
16. Xiao X, Wu Z-C, Chou K-C. iLoc-Virus: a multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites. *J Theor Biol*. 2011;284(1):42-51. doi: <http://dx.doi.org/10.1016/j.jtbi.2011.06.005>
17. Wan S, Mak M-W, Kung S-Y. mGOASVM: Multi-label protein subcellular localization based on gene ontology and support vector machines. *BMC Bioinformatics*. 2012;13:290. doi: <http://dx.doi.org/10.1186/1471-2105-13-290>
18. Li Z-C, Lai Y-H, Chen L-L, et al. Identifying subcellular localizations of mammalian protein complexes based on graph theory with a random forest algorithm. *Mol Biosyst*. 2013;9(4):658-667. doi: <http://dx.doi.org/10.1039/c3mb25451h>
19. Boutet E, Lieberherr D, Tognolli M, Schneider M, Bairoch A. UniProtKB/Swiss-Prot. *Methods Mol Biol*. 2007;406:89-112.
20. Wang Y, Ding Y, Guo F, Wei L, Tang J. Improved detection of DNA-binding proteins via compression technology on PSSM information. *PLoS One*. 2017;12(9). doi: <http://dx.doi.org/10.1371/journal.pone.0185587>
21. Wei L, Xing P, Shi G, Ji Z-L, Zou Q. Fast prediction of protein methylation sites using a sequence-based feature selection technique. *IEEE/ACM Trans Comput Biol Bioinform*. February 2017. doi: <http://dx.doi.org/10.1109/TCBB.2017.2670558>
22. Wu Y, Cai S, Lu M, Krishnan S. An Artificial-Neural-Network-Based Multiple Classifier System for Knee-Joint Vibration Signal Classification. In: Wu Y, ed. *Advances in Computer, Communication, Control and Automation*. Lecture Notes in Electrical Engineering. Springer Berlin Heidelberg; 2012:235-242.
23. Wei L, Fei Guo, Su R, Xing P. Identifying N⁶-methyladenosine sites using multi-interval nucleotide pair position specificity and support vector machine. *Scientific Reports*. 2017;7:46757. doi: <http://dx.doi.org/10.1038/srep46757>
24. Wen M, Zhang Z, Niu S, et al. Deep-Learning-Based Drug-Target Interaction Prediction. *J Proteome Res*. 2017;16(4):1401-1409. doi: <http://dx.doi.org/10.1021/acs.jproteome.6b00618>
25. Xu C, Du P. Predicting multisite protein subcellular locations: progress and challenges. *Expert Rev Proteomics*. 2013;10(3):227-237. doi: <http://dx.doi.org/10.1586/epr.13.16>
26. Tsoumakas G, Katakis I. Multi-label classification: An overview. *Int J Data Warehousing and Mining*. 2007;2007:1-13.
27. Tsoumakas G, Katakis I, Vlahavas I. Random k-Labelsets for Multilabel Classification. *IEEE Trans on Knowl and Data Eng*. 2011;23(7):1079-1089. doi: <http://dx.doi.org/10.1109/TKDE.2010.164>
28. Castellanos-Domínguez G, Jaramillo-Garzón JA, Arango-Argoty GA. Feature extraction by statistical contact potentials and wavelet transform for predicting subcellular localizations in gram negative bacterial proteins. *J Theor Biol*. 2015;364:121-130. doi: <http://dx.doi.org/10.1016/j.jtbi.2014.08.051>
29. Yuan JQ, Huang C. Predicting protein subchloroplast locations with both single and multiple sites via three different modes of Chou's pseudo amino acid compositions. *J Theor Biol*. 2013;335:205-212. doi: <http://dx.doi.org/10.1016/j.jtbi.2013.06.034>
30. Qu X, Wang D, Chen Y, Qiao S, Zhao Q. Predicting the Subcellular Localization of Proteins with Multiple Sites Based on Multiple Features Fusion. *IEEE/ACM*

- Trans Comput Biol Bioinform.* 2016;13(1):36-42. doi: <http://dx.doi.org/10.1109/TCBB.2015.2485207>
31. Wang L, Zhao Y, Chen Y, Wang D. The effect of three novel feature extraction methods on the prediction of the subcellular localization of multi-site virus proteins. *Bioengineered.* 2018;9(1):196-202. doi: <http://dx.doi.org/10.1080/21655979.2017.1373536>
 32. Rahman J, Mondal NI, Islam KB, Hasan AM. Feature Fusion Based SVM Classifier for Protein Subcellular Localization Prediction. *Journal of Integrative Bioinformatics.* 2016;13(1). doi: <http://dx.doi.org/10.1515/jib-2016-288>
 33. Khan A, Majid A, Hayat M. CE-PLoc: An ensemble classifier for predicting protein subcellular locations by fusing different modes of pseudo amino acid composition. *Computational Biology and Chemistry.* 2011;35(4):218-229. doi: <http://dx.doi.org/10.1016/j.compbiolchem.2011.05.003>
 34. Chou K-C. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins: Structure, Function, and Bioinformatics.* 2001;43(3):246-255. doi: <http://dx.doi.org/10.1002/prot.1035>
 35. Mei S. Multi-kernel transfer learning based on Chou's PseAAC formulation for protein submitochondria localization. *Journal of Theoretical Biology.* 2012;293:121-130. doi: <http://dx.doi.org/10.1016/j.jtbi.2011.10.015>
 36. Shen H-B, Chou K-C. Virus-mPLoc: a fusion classifier for viral protein subcellular location prediction by incorporating multiple sites. *J Biomol Struct Dyn.* 2010;28(2):175-186. doi: <http://dx.doi.org/10.1080/07391102.2010.10507351>
 37. Wei L, Ding Y, Su R, Tang J, Zou Q. Prediction of human protein subcellular localization using deep learning. *Journal of Parallel and Distributed Computing.* 2018;117:212-217. doi: <http://dx.doi.org/10.1016/j.jpdc.2017.08.009>

Copyright: © **Cong et al.** This is an Open Access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.