A Dynamic-Time Distance Based on Wavelet Decomposition for Subcellular Localization Classification

CUIFANG GAO1,2, WANQIANG SHEN1,2, AND YANGLIJUN ZHANG1

1School of Science, Jiangnan University, Wuxi 214122, China
2Wuxi Engineering Research Center for Biocomputing, Jiangnan University, Wuxi 214122, China

Corresponding author: Wanqiang Shen (wq_shen@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61402202 and Grant 61772013, and in part by the China Postdoctoral Science Foundation under Grant 2015M581724.

ABSTRACT The use of bioinformatics to predict protein subcellular localization is a significant method to study protein function. In the present study, a novel dynamic-time distance measurement based on wavelet packet decomposition (WPD) was proposed for protein subcellular localization prediction. The protein sequence was firstly converted into multiple numerical signals according to physical/chemical properties. Following signal decomposition into multiple subsignals by wavelet packet decomposition, a comprehensive dynamic-time distance was obtained by the dynamic time warping (DTW) algorithm. Finally, the expected classification can be produced based on the new distance measurement. By introducing DTW, the present algorithm can overcome the shortcoming that traditional methodologies can not measure the similarity of unequal-length sequences. Additional, multi frequency band subsignals can retain more information to achieve accurate results. It was suggested that our algorithm provides superior classification recognition by setting traditional methods as a comparative experiment in the E.coli dataset. It can distinguish cytoplasm and cell membrane proteins that are particularly difficult to be identified by the traditional methodologies.

INDEX TERMS Dynamic time warping, feature extraction, subcellular localization, wavelet decomposition.

I. INTRODUCTION The development of genome sequencing has enabled the determination of a large number of protein and DNA sequence information. Currently biological researchers face the dilemma of the study method and use of the large amount of data generated by sequencing methods [1]. The computer-based approaches exert specific advantages while assimilating large amount of biological data, such as the high computational speed and low cost of experimentation. This plays an important role in the processing of biological information. The development of novel algorithms for biological sequences has become an important research interest.

Proteins are considered the material basis for living forms and constitute the basic organic matter of the cells. Subcellular location relates closely to the functioning of a protein, which is one of the key attributes of gene production. It is well known that the subcellular localization of proteins plays a crucial role in many aspects of their function. This involves the three essential features of the protein namely, biological objective, biochemical activity and cellular localization that includes active gene transcription [2]. However, each protein can only exert its specific function in a particular subcellular position. The localization of these proteins provides a gateway through which functional analysis may be explored. It is necessary to develop appropriate methods for the effective prediction of the subcellular localization of a protein.

Several studies have explored the application of bioinformatic methodologies in the identification of the subcellular localization of proteins. Nakashima and Nishikawa proposed the amino acid composition for the identification of intracellular and extracellular proteins in 1994 [3]. Bulashevska and Eils introduced a recursive algorithm namely, HensBC in 2006, which constructed a hierarchical ensemble of classifiers and improved the accuracy of the prediction of certain classes with few sequences [4]. Li et al. proposed an ensemble classifier of KNN and SVM algorithms in order to predict the subcellular localization of eukaryotic proteins based on a voting system in 2012 [5]. Mandal et al. utilized the multi-objective particle swarm optimization-based feature selection technique in order to predict the subcellular localization of proteins [6]. To date, various computational techniques such
C. Gao et al.: Dynamic-Time Distance Based on Wavelet Decomposition for Subcellular Localization Classification

TABLE 1. Information of E.coli Dataset.

| Dataset       | Subcellular localization | Number of sequence | Minimum length | Maximum length | Average length |
|---------------|--------------------------|--------------------|----------------|----------------|----------------|
| E.coli        | Cell                     | 658                | 1756           | 322            |
|               | Cytoplasm                | 121                | 1083           | 362            |
|               | membrane                 | 1635               | 19             | 1486           | 382            |

TABLE 2. Four Physical/Chemical Properties for Numerization Method.

| Amino acids | Hydrophobic value | Relative molecular mass | Standard Rf value | Isoelectric point |
|-------------|-------------------|-------------------------|-------------------|-------------------|
| A           | 0.62              | 89.09                   | 0.47              | 6.11              |
| R           | -1.73             | 174.2                   | 0.15              | 10.76             |
| D           | -1.05             | 133.1                   | 0.27              | 2.85              |
| L           | 0.29              | 121.2                   | 0.08              | 5.05              |
| Q           | -0.78             | 146.2                   | 0.46              | 5.65              |
| H           | -0.87             | 147.1                   | 0.35              | 3.15              |
| I           | -0.4              | 155.2                   | 0.11              | 7.6               |
| G           | 1.38              | 131.2                   | 0.77              | 6.05              |
| N           | 0.48              | 75.07                   | 0.3               | 6.06              |
| L           | -0.85             | 131.2                   | 0.01              | 5.41              |
| K           | -1.35             | 146.2                   | 0.12              | 9.6               |
| M           | 0.64              | 149.2                   | 0.64              | 5.74              |
| F           | 1.19              | 165.2                   | 0.73              | 5.49              |
| P           | 0.12              | 115.1                   | 0.48              | 6.7               |
| S           | -0.18             | 105.1                   | 0.28              | 5.68              |
| T           | -0.05             | 119.1                   | 0.37              | 5.6               |
| W           | 0.81              | 204.2                   | 0.64              | 5.89              |
| Y           | 0.26              | 181.2                   | 0.25              | 5.64              |
| V           | 1.08              | 117.2                   | 0.67              | 6                 |

as the hidden Markov model [7], the neural network [8],
the K-Nearest Neighbor [9], [10] and the Support Vector
Machine have been developed for the identification of the
subcellular protein localization prediction [11]–[13].

The aforementioned studies have shown that protein sub-
cellular localization is closely related to a variety of factors,
while the different length of protein sequences further renders
the similarity analysis more difficult, notably by the use of
the traditional similarity distance method. Dynamic Time
Warping (DTW) is a distance algorithm [14], which was
originally proposed by the Japanese scholar Itakura in the
1960s. DTW exhibits apparent advantages in the process of
signals, especially for the sequences of lack fragments and/or
dimension shift. Wavelet packet decomposition (WPD) is
also termed wavelet packet or subband tree. It is a very
sophisticated signal analysis method, which can provide spe-
cific signal decomposition [15]. WPD has a wide range of
applications in a majority of scientific fields, such as seismic
wave observation and oil and gas exploration.

In the present study, according to the characteristics of
protein sequences, a novel algorithm for the identification
of the subcellular protein localization based on Wavelet
Packet Decomposition and Dynamic Time Warping is pro-
posed. To obtain the comprehensive feature information in
the sequence, different physical/chemical properties were uti-
lized in the algorithm to get numerical signals; By WPD,
the sequence with various information were decomposed into
subsignals in different frequency bands, so as to reduces
the error from the comparison between different types
of sequences; DTW achieves warping matching between
unequal length sequences by stretching/shortening sequence
signals, which can more accurately calculate the similarity
between sequences. The algorithm can effectively classify the
protein sequences of the cytoplasmic and the cell membrane
proteins compared with the traditional methodologies.

II. MATERIALS AND METHODS

The dataset used in this article was obtained from the
UniProtKB/Swiss-Prot module in the UniProt database,
which is an expert-proven, high-quality, manual annotation,
non-redundant dataset consisting mainly of three modules:
UniProtKB/Swiss-Prot, UniProtKB/TrEMBL and UniParc.
According to the different subcellular localization, the E.coli
dataset was obtained. The specific information of the data set
is shown in TABLE 1.

A. PROTEIN SEQUENCE NUMERIZATION

The process of numerization the eigenvector from sequence
is directly related to the final classification results, since the
original protein sequence is represented by the letter symbol.
In the present study, we provide six numerical methods from
different aspects for the execution of the classification algo-

Method-AAC: Method-AAC indicates amino acid compo-
sition, which may be considered the easiest of all approaches.
The method enables the construction of the eigenvector based
on the ratio of 20 amino acids of the entire sequence, where
the first component is the percentage of alanine (A) in the
entire sequence and the second component is the percentage
of arginine (R).

Method-cwt: Method-cwt creates a wavelet feature vector
that is a method of digitizing data by using a continuous
wavelet transformation. This method provides the advantage
of characterizing the local characteristics of a signal based on
the parameters time and frequency. By selecting the option
low frequency, the method confers high resolution within
the time domain and low resolution within the frequency
domain. By selecting high frequency, the resolution is the
reverse. It automatically adapts to time-frequency signal anal-
ysis capabilities, making it possible to focus on any detail of
the signal. Method-cwt uses the continuous wavelet transfor-
mations in order to translate the original signal to a requisite
dimension eigenvector.

Method-Hy: Method-Hy digitizes the amino acid sequence
according to its hydrophobicity. Unlike method-AAC, the
dimension of the eigenvector constructed by method-Hy is
not fixed and it varies according to the length of the amino
cid sequence. For example, if the specific numerical process
is the amino acid X in the sequence, then X is replaced by X’s
hydrophobic value x. The hydrophobic values of the 20 amino
acids are shown in TABLE 2.

Method-MW: Method-MW digitizes the amino acid sequence
according to its relative molecular mass. In the
in order to match the signal spectrum and ultimately improve the time-frequency resolution. The excellent time-frequency characteristics of the WPD can improve the wavelet transform. On the basis of wavelet decomposition, the discrete detail is decomposed in order to yield the binary tree decomposition of the signal. The incomplete WPD can continue to decompose or terminate the decomposition on any node. The signal band is then divided according to the requirements of the actual signal analysis. Equation (1) is the function of the orthogonal wavelet packet decomposition.

\[
\begin{align*}
    u_{2t}(t) &= \sqrt{2} \sum_{k \in Z} h(k) u_{t}(2t - k) \\
    u_{2t+1}(t) &= \sqrt{2} \sum_{k \in Z} g(k) u_{t}(2t - k)
\end{align*}
\]  

(1)

In (1), \(g(k)\) and \(h(l-k)\) is a pair of orthogonal mirror filters. When \(n=0\), \(u_0(t)\) and \(u_1(t)\) are the scale function \(\psi(t)\) and the wavelet function \(\varphi(t)\).

The WPD diagram is shown in Fig. 1 and the three-tier decomposition is used as an example. Specifically, WPD was used in order to convert the original signal \(S\). The signal was divided into two types of signals: low frequency and high frequency respectively, denoted as \(A_1\) and \(D_1\). Subsequently, the two signals were decomposed again in order to yield low frequency of \(A_1\), high frequency of \(A_1\), low frequency of \(D_1\) and high frequency of \(D_1\). The aforementioned parameters were denoted as \(AA_2, DA_2, AD_2\) and \(DD_2\) respectively. Similarly, WPD was used in order to additionally convert the four signals. As a result eight signals of high and low frequency were obtained, which were denoted as \(AAA_3, DAA_3, ADA_3, ADD_3, AAD_3, DAD_3, ADD_3\) and \(DDD_3\). In the present study, the wavelet packet decomposition was used that is based on the haar wavelet. This indicates that the signal dimension can be strictly divided into one-half. Consequently, the 2\(k\) signals of the \(k\)th layer exhibit the \(2^k\) dimension of the original signal \(S\). For example, an original signal \(S\) with dimension \(N=800\), following 3 times of decomposition, is further decomposed into \(2^3=8\) signal fragments, and the dimension of each signal fragment is reduced to 100.

Following protein sequence numerization, the digital signal was decomposed into a number of smaller signal segments according to a particular rule by using WPD. Consequently, the different information in a sequence was decomposed into specific fragments, which is conducive in order to explore the new distance measure, and finally enhance the recognition accuracy.
C. DYNAMIC TIME WARPING

Equation (2) gives the mathematical formula of the Euclidean distance.

\[ d(X, Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (2) \]

It should be noted that X and Y both are feature vectors whose dimension is n in the aforementioned equation. This indicates that the Euclidean distance can be calculated based on the precondition that the dimension of the two vectors is the same. But for unequal length signals, it is not feasible to use the traditional Euclidean distance.

In this aspect, the DTW exhibits certain advantages. Dynamic Time Warping (DTW) is a distance algorithm, which is suitable for unequal length signal. The unknown sequence is stretched and/or shortened until it coincides with the length of the reference template. In this process, the time axis of the unknown sample is distorted or bent so that its feature corresponds to the standard pattern. A total of two vectors with any dimension can identify the dynamic-time distance by DTW. This results in the distortion of the two samples to the standard model in the time domain, and the enhancement of the classification accuracy.

Fig. 2 indicates the general process of DTW. Initially, the corresponding points of the two sequences and the warp sequence were identified using the standard model and subsequently the regular distance can be estimated according to the corresponding points.

For \( X = (x_1, x_2, \ldots, x_n) \), \( Y = (y_1, y_2, \ldots, y_m) \), the mathematical expression of dynamic time warping is provided by (3).

\[ D(X, Y) = \min_{W = w_1, \ldots, w_k, \ldots, w_L} \left[ \frac{\sum_{k=1}^{K} (x_{w_k} - y_{i,j})^2}{K} \right] \quad (3) \]

In (3), \( i=1,2,\ldots,n; j=1,2,\ldots,m \). \( D(X,Y) \) represents the distance of the vector X and Y. W is a regular path. The regular path is a set containing multiple path points of pairwise \( w_k = (i,j) \), and always originates with \((1,1)\) and ends with \((n,m)\). A total number of \( L \) paths was selected for the process of the dynamic-time distance calculation of X and Y. \( D(X,Y) \) is the optimal path, so as to indicate the dynamic-time distance of the shortest path. Different paths exhibited different number of path points \( (K) \). The path point \( w_k = (i,j) \) indicates that the distance between the component \( x_i \) of X vector and the component \( y_j \) of Y vector, and their average value calculated into the cumulative distance. Generally, the dynamic-time distance for a vector of smaller dimensions is always smaller due to the fact that the path points on the optimal path are always smaller. Therefore, the cumulative distance was divided by \( K \) in order to calculate the average distance as \( D(X,Y) \).

Another way to explain the process of DTW is shown below. For the same parameters namely, \( X = (x_1, x_2, \ldots, x_n) \), \( Y = (y_1, y_2, \ldots, y_m) \), a distance matrix \( D \) of \( n \times m \) was constructed initially, where \( d(i,j) \) in the matrix represents an estimate of the square of the distance between \( x_i \) and \( y_j \). The cumulative distance matrix was denoted as \( D \) and \( D(n,m)/K \) represented the dynamic-time distance of X and Y. The conditions for the construction of the cumulative matrix D were as follows:

1) \( D(1,1) = d(1,1) \)
2) \( D(i,j) = D(i,j-1) + d(i,j) \)
3) \( D(i,1) = D(i-1,1) + d(i,1) \)
4) \( D(i,j) = \min(D(i-1,j), D(i,j-1), D(i-1,j-1)) + d(i,j) \)

In the aforementioned equations, \( i=2,3,\ldots,n; j=2,3,\ldots,m \). \( D(i,j) \) may be \( d(i,j) \) plus any one of \( D(i-1,j), D(i,j-1), D(i-1,j-1) \). The smallest value is also calculated by its left, top and/or upper left elements. The path from \( D(1,1) \) to the end of \( D(n,m) \) is the optimal path and the elements in the cumulative matrix \( D \) are the corresponding points.

The signals analyzed in the present study are biological signals, and their source organisms are usually mutated. Biological mutations frequently lead to biological signal changes. Specific examples include omission of certain amino acid molecules, chromosome fragments exchange and/or protein sequence length changes. Therefore, the use of DTW in the biological signal processing exhibits specific advantages compared with the traditional distance algorithm.

D. THE PREDICTION ALGORITHMS

The key point of the proposed method is the combination of WPD and DTW. Firstly, the numerical signals of the test samples and the training set are divided into 16-segment subsignals from low to high by four-layer WPD. This results in preprocessing of the information. Subsequently, the DTW compares the test sample with the 16 subsignals of the training sample in order to obtain 16 distances. The sum of the 16 distances is further compared as the dynamic-time distance between them. Finally, the k-nearest neighbor method is used to determine the partition of the test sample.

In the present study, we used six numerical methods to yield eight different algorithms and compare their performance. The specific algorithms are shown in Fig. 3:

The second row in Fig. 3 represents the six aforementioned numerical methods. For any protein sequence, the first four numerical methods create features with unfixed dimension, which change according to the length of the protein sequence. The latter two numerical methods are used as contrast in the experiments to create 20-dimensional features. The third
row indicates the selection of WPD based on the previous eigenvectors. In the present study, we used the four-layer wavelet packet decomposition. The fourth row indicates the application of the DTW algorithm based on WPD.

The fifth row demonstrates the classification algorithm that was used in the present study. KNN and SVM are well-known supervised classifiers, they are widely-used for their good performance, easy to understand and implement. Considering that our method can estimate the distance between test samples and training samples, KNN and SVM are suitable for the prediction task.

Therefore, the final classification algorithm adopts the k-nearest neighbor method, and classifies the test sample according to the largest number of training samples. For wavelet features and amino acid composition features, the support vector machine (LIBSVM) was used in order to perform the classification. LIBSVM exhibits great impact and high classification effect. In addition, 10-cross validation was used that ensured division of the dataset into ten subsets. Each routine sample test was conducted as a test set and the remaining sets were used for the training set in order to calculate the final result. The average of ten results was used as the final classification accuracy. The sixth row indicates the abbreviation of the different methods.

The entire process of the prediction is shown in Fig. 4.

It should be noted that all the steps in Fig. 4 are required for KNN method, but the steps in dashed box are not need for SVM method.

The mathematical basis behind the KNN-Hy method includes numerical method, wavelet analysis, similarity distance and statistic-based learning. The dimension of the eigenvector constructed by hydrophobicity varies with the length of the protein sequences, DTW is a distance algorithm which is suitable for unequal length sequences. Wavelet analysis has been viewed as a “mathematical microscope” so that subsignals from WPD contain different structure information of protein sequences. KNN is a machine learning algorithms based on data statistics. Therefore, KNN-Hy exhibits certain advantages by combining these mathematical processes. Algorithm 1 and Algorithm 2 give the pseudo codes of the KNN-Hy prediction method.

In DTW_matrix, the jth row-vector indicates the distances between the jth test sample and all training samples. We can use the function knn_min to get the prediction result of the jth test sample according to KNN algorithm.

Where, wpd16 is a function to decompose a signal into 16 segments. dtwd is a function to get the DTW distance between WP1 and WP2, which is the sum of one-to-one distances between 16 segments in pairs.

![Figure 3. Eight different algorithms.](image)

![Figure 4. Flow sheet of the prediction.](image)

**Algorithm 1 Pseudo Code of KNN-Hy Method**

```plaintext
input DATA // symbol sequences
Vect_D ← numerical (DATA)
Subset ← group (Vect_D, 10)
for i = 1 to 10
    test_D is the Subset[i];
    train_D is the remain Subset
    DTW_matrix ← dtw_dis(test_D, train_D)
    // the size of DTW_matrix is test_num * train_num
    for j = 1 to test_num
        result_D ← knn_min(DTW_matrix[j])
    end for
end for
output result_D // prediction results
```
Algorithm 2 Pseudo Code of the Function dtw_dis

```
input test_D, train_D // test samples and train samples
    test_num is the number of samples in test_D
    train_num is the number of samples in train_D
for jj = 1 to train_num
    WP2 ← wpd16(train_D[jj])
for ii = 1 to test_num
    WP1 ← wpd16(test_D[ii])
    DTW_matrix ← dtwd(WP1, WP2)
end for
end for
output DTW_matrix // Distances between all training // sequences and test sequences
```

III. EXPERIMENTS AND ANALYSIS

The data sets used in this study are based on a different subcellular localization that is downloaded from UniProt. The eight algorithms were respectively used in the experiment with the E.coli dataset, and the result is shown in Fig. 5.

The results of wavelet and amino acid composition on the E.coli dataset can be estimated accurately based on the new method proposed in the current study. The optimal classification is KNN-Hy and the accuracy rate can increase up to 83.76%. The lowest accuracy rate was noted for SVM-cwt and was estimated to 62.95%. The new method was applied on wavelet features and the amino acid composition features, and the classification accuracy was considerably improved. The wavelet features were increased from 62.95% to 75.84%, while the amino acid compositions from 64.67% to 76.26%. The use of WPD in combination with DTW improved the subcellular localization classification are shown in Fig. 6.

The classification accuracy rates were 97.22%, 98.98% and 97.41% respectively, with the exception of the cytoplasmic locations that were estimated to 3.95%. Consequently, the overall classification of KNN-pl is the lowest and is estimated to 66.19%. Specific experiments for the detection of the localization of cytoplasmic proteins were designed in order to further analyze the performance of KNN-pl.

The incorrect classification of the protein sequence is shown in blue, and the correct classification in orange. The majority of the cytoplasm sequences are incorrectly classified as membrane sequences by the KNN-pI algorithm (Fig. 8). Consequently, the use of isoelectric point numericalized protein sequences cannot effectively distinguish between cytoplasmic and cell membrane protein sequences, and this eventually results in lower classification accuracy. It is worth mentioning that the classification accuracy of the cytoplasm is 0% as determined by the SVM-AAC and SVM-cwt methods, and more than 98% are partitioned to the membrane sequence. Although the majority of the membrane can be assigned to the right partition, this does not ensure that the algorithm can identify the cell membrane sequence. In contrast to this hypothesis, KNN-Hy, KNN-MW, KNN-Rf, KNN-cwt and KNN-AAC were effective in distinguishing between the cytoplasmic and cell membrane regions. With regard to the KNN-pl method, the classification results of the cytoplasm and the cell membrane using the KNN-Hy method are shown in Fig. 9.
FIGURE 8. Diagram of which categories protein sequences are divided into while using isoelectric point of amino acid as characteristics (the y-axis represents the classification accuracy). The left figure is the cytoplasmic sequence, the right is the cell membrane sequence.

FIGURE 9. Diagram of which categories protein sequences are divided into while using hydrophobicity of amino acid as characteristics (the y-axis represents the classification accuracy). The left figure is the cytoplasmic sequence, the right is the cell membrane sequence.

FIGURE 10. The frequency distribution of dynamic-time distance (the x-axis represents the distance measure, the y-axis represents the density).

Furthermore we analyzed the dynamic-time distance using the KNN-Hy algorithm. By calculating the dynamic-time distance of each sample in the dataset, we estimated the distance, and constructed the histogram of the distance (Fig. 10). The maximum density of the distance was between 20 and 25. Its frequency was estimated by the multiplication of the density with the width of the histogram. This was approximately 0.09×5=0.45. According to the four types of subcellular localization, the distances were classified into two categories of protein sequences with ten indicators. These were the following: four intra-class distance, and six inter-class distance. In theory, the purpose of the present analysis was to design a classification algorithm that reduces the intra-class distance and increases the inter-class distance as much as possible. The intra-class distance of the membrane sequences was slightly larger compared to the other classes (Fig. 11).

This may be due to the other sequences that exert special class characteristics compared with the membrane sequence. Therefore, a part of the aforementioned algorithm exhibits poor performance in the distinction of the membrane sequences, as opposed to other categories.

The accuracy of all eight algorithms is provided in detail in TABLE 3.

It should be noted that the KNN-Hy algorithm exhibited optimal performance on the prediction of the subcellular localization of the E.coli dataset. The algorithm was capable of discriminating the cell membrane and cytoplasmic protein sequences that is considered a particularly difficult task.

The average classification accuracy by KNN and SVM are respectively shown in Fig. 12 and Fig. 13.

As can be seen in Fig. 12 and Fig. 13, the SVM method can not distinguish cytoplasm from cell membrane, and the classification accuracy by KNN method is not high. Aiming at the sequences in cytoplasm and cell membrane, we further analyze their confusion matrix, as shown in TABLE 4:

Without considering cell fluid and quasi nuclear, the classification accuracy of cytoplasm increases slightly, but there are still some overlap and confusion samples between cytoplasm and cell membrane. According to Fig. 11, their intra class distance and inter class distances are very close, which may result in the unclear boundaries between the two types of data. In addition, the large sample size also increases the difficulty of identification.
a numerical method, which can effectively reflect various physical and chemical properties of biological sequences, can improve the classification performance significantly. In our experiments, the proteins with classification label from UniProtKB are used to verify the performance of the algorithm. In theory, as a similarity measurement, dynamic-time distance based on WPD can also be introduced into some unsupervised clustering algorithms based on distance [16, 17], which is a very promising method to achieve the subcellular localization prediction for unknown proteins in the future.

**REFERENCES**

[1] O. Emanuelsson, H. Nielsen, S. Brunak, and G. von Heijne, “Predicting subcellular localization of proteins based on their N-terminal amino acid sequence,” J. Mol. Biol., vol. 300, no. 4, pp. 1005–1016, Jul. 2000.

[2] B. Niu, Y.-H. Jin, K.-Y. Feng, W.-C. Lu, Y.-D. Cai, and G.-Z. Li, “Using AdaBoost for the prediction of subcellular location of prokaryotic and eukaryotic proteins,” Mol. Diversity, vol. 12, no. 1, pp. 41–45, May 2008.

[3] H. Nakashima and K. Nishikawa, “Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies,” J. Mol. Biol., vol. 238, no. 1, pp. 54–61, Apr. 1994.

[4] A. Bulashevska and R. Elis, “Predicting protein subcellular locations using hierarchical ensemble of Bayesian classifiers based on Markov chains,” BMC Bioinf., vol. 7, no. 1, p. 1–13, Jun. 2006.

[5] L. Li, Y. Zhang, L. Zou, C. Li, B. Yu, X. Zheng, and Y. Zhou, “An ensemble classifier for eukaryotic protein subcellular location prediction using gene ontology categories and amino acid hydrophobicity,” PLoS ONE, vol. 7, no. 1, Jan. 2012, Art. no. e31057.

[6] M. Mandal, A. Mukhopadhyay, and U. Maulik, “Prediction of protein subcellular localization by incorporating multiobjective PSO-based feature subset selection into the general form of Chou’s PseAAC,” Med. Biol. Eng. Comput., vol. 53, no. 4, pp. 331–344, Jan. 2015.

[7] T.-h. Lin, R. F. Murphy, and Z. Bar-Joseph, “Discriminative motif finding for predicting protein subcellular localization,” IEEE/ACM Trans. Comput. Biol. Bioinf., vol. 8, no. 2, pp. 441–451, Mar. 2011.

[8] L. Zou, Z. Wang, and J. Huang, “Prediction of subcellular localization of eukaryotic proteins using position-specific profiles and neural network with weighted inputs,” J. Genet. Genomics, vol. 34, no. 12, pp. 1080–1087, Dec. 2007.

[9] X. Xiao, Z.-C. Wu, and K.-C. Chou, “Loc-virus: A multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites,” J. Theor. Biol., vol. 284, no. 1, pp. 42–51, Sep. 2011.

[10] X. Xiao, Z.-C. Wu, and K. C. Chou, “A multi-label classifier for predicting the subcellular localization of gram-negative bacterial proteins with both single and multiple sites,” PLoS ONE, vol. 6, no. 6, Jun. 2011, Art. no. e20592.

[11] Y.-D. Cai, X.-J. Liu, X.-B. Xu, and K.-C. Chou, “Support vector machines for prediction of protein subcellular location by incorporating quasidependence-order effect,” J. Cellular Biochem., vol. 84, no. 2, pp. 343–348, Jan. 2002.

[12] R.-P. Liang, S.-Y. Huang, S.-P. Shi, X.-Y. Sun, S.-B. Suo, and J.-D. Qiu, “A novel algorithm combining support vector machine with the discrete wavelet transform for the prediction of protein subcellular localization,” Comput. Biol. Med., vol. 42, no. 2, pp. 180–187, Feb. 2012.

[13] T. Liu, P. Tao, X. Li, Y. Qiu, and C. Wang, “Prediction of subcellular location of apoptosis protein combination tri-gram encoding based on PSSM and recursive feature elimination,” J. Theor. Biol., vol. 366, pp. 8–12, Feb. 2015.

[14] R. J. Kate, “Using dynamic time warping distances as features for improved time series classification,” Data Mining Knowl. Discovery, vol. 30, no. 2, pp. 283–312, Mar. 2016.

[15] C. Wang, M. Gan, and C. Zhu, “Fault feature extraction of rolling element bearings based on wavelet packet transform and sparse representation theory,” J. Intell. Manuf., vol. 29, no. 4, pp. 937–951, Apr. 2018.
[16] A. Rodriguez and A. Laio, “Clustering by fast search and find of density peaks,” *Science*, vol. 344, no. 6191, pp. 1492–1496, Jun. 2014.

[17] T. Liu, H. Li, and X. Zhao, “Clustering by search in descending order and automatic find of density peaks,” *IEEE Access*, vol. 7, pp. 133772–133780, Sep. 2019.

**CUIFANG GAO** received the B.S. degree from Sun Yat-sen University, in 1998, and the M.S. and Ph.D. degrees from Jiangnan University, in 2007 and 2011, respectively. She is currently an Associate Professor with the School of Science, Jiangnan University. Her research interests include pattern recognition, computational intelligence, and bioinformatics.

**WANQIANG SHEN** was born in Jiangsu. She received the M.S. and Ph.D. degrees from Zhejiang University, in 2005 and 2010, respectively. She is currently an Associate Professor with the School of Science, Jiangnan University. Her research interests include pattern recognition, computer aided design, and computer graphics.

**YANGLIJUN ZHANG** received the M.S. degree in applied mathematics from Jiangnan University, in 2018. His thesis was on researches on protein sequence similarity based on dynamic time warping. His research interests include pattern recognition and bioinformatics.

---

***