A novel method for apoptosis protein subcellular localization prediction combining encoding based on grouped weight and support vector machine

Zhen-Hui Zhang\textsuperscript{a,*}, Zheng-Hua Wang\textsuperscript{b}, Zhen-Rong Zhang\textsuperscript{c}, Yong-Xian Wang\textsuperscript{b}

\textsuperscript{a} Department of Mathematics and System Science, School of Science, National University of Defense Technology, 410073 Changsha, China
\textsuperscript{b} National Laboratory for Parallel and Distributed Processing, School of Computer, National University of Defense Technology, 410073 Changsha, China
\textsuperscript{c} School of Computer, Electronics and Information, Guangxi University, 530004 Nanning, China

Received 23 August 2006; revised 25 September 2006; accepted 6 October 2006

Available online 17 October 2006

Edited by Takashi Gojobori

Abstract Apoptosis proteins have a central role in the development and homeostasis of an organism. These proteins are very important for understanding the mechanism of programmed cell death. Based on the idea of coarse-grained description and grouping in physics, a new feature extraction method with grouped weight for protein sequence is presented, and applied to apoptosis protein subcellular localization prediction associated with support vector machine. For the same training dataset and the same predictive algorithm, the overall prediction accuracy of our method in Jackknife test is 13.2\% and 15.3\% higher than the accuracy based on the amino acid composition and instability index. Especially for the else class apoptosis proteins, the increment of prediction accuracy is 41.7 and 33.3 percentile, respectively. The experiment results show that the new feature extraction method is efficient to extract the structure information implicated in protein sequence and the method has reached a satisfied performance despite its simplicity. The overall prediction accuracy of EBGW\_SVM model on dataset ZD98 reach 92.9\% in Jackknife test, which is 8.2\%–20.4\% percentile higher than other existing models. For a new dataset ZW225, the overall prediction accuracy of EBGW\_SVM achieves 83.1\%. Those implied that EBGW\_SVM model is a simple but efficient prediction model for apoptosis protein subcellular location prediction.

Keywords: Apoptosis protein subcellular localization; Encoding based on grouped weight; Amino acid composition; Support vector machine; Component-coupled algorithm

1. Introduction

During the last decade, many theoretical methods were developed in an attempt to predict the subcellular location of a query protein according to its sequence information [1–33]. For a deeper comprehension can refer to the reviews [8,9,14]. To enhance the practical application value, we can see that: (i) the coverage scope in identifying the localization of a protein is gradually getting larger, such as from discriminating between only 2 subcellular locations [2], to among 5 locations [3].

Apoptosis, or programmed cell death, is a fundamental process controlling normal tissue homeostasis by regulating a balance between cell proliferation and death [34]. Cell death and renewal are responsible for maintaining the proper turnover of cells, which ensures a constant controlled flux of fresh cells. When apoptosis malfunctions, a variety of formidable disease can ensue: blocking apoptosis is associated with cancer, and autoimmune disease, whereas unwanted apoptosis can possibly lead to ischemic damage or neurodegenerative disease [35]. Obtaining information about subcellular location of apoptosis proteins will help us to understand the apoptosis mechanism and functions of proteins. This is because subcellular location of proteins are strongly correlated to their function. Thus, the study of subcellular location of apoptosis protein is very interesting in biology.

Zhou and Doctor [16] investigated the prediction of apoptosis protein subcellular location using amino acid composition and component-coupled algorithm. They constructed a training dataset of 98 apoptosis proteins, and the overall prediction accuracy in Jackknife test reach 72.5\%. Chen and Li [36] introduced dipeptide composition and algorithm of measure of diversity to predict apoptosis protein subcellular location. On the same dataset, the prediction accuracy was 84.7\%. Huang [37] applied the support vector machine by incorporating protein instability index to the same research, and the accuracy they got was 77.6\%. However, the prediction capacity of those three methods was unbalance. Especially, for other class proteins (exclude cytoplasmic, membrane and mitochondrial proteins), the prediction accuracy was not exceed 50\%.

In this paper, based on the concept of coarse-grained description and grouping, a new encoding method with grouped weight for protein sequence (encoding based on grouped weight, named as EBGW) was presented. Integrating the new scheme (EBGW) with the support vector machine algorithm, it show that the overall prediction accuracy of apoptosis protein subcellular location was significantly improved. Furthermore, the methodology presented here might be useful for other studies of protein structure and function.
2. Materials and methods

2.1. Datasets

To have a critical comparison between the different approaches, the datasets constructed by Zhou [16] and Chen [36] were adopted in our work. Proteins in those datasets were extracted from Swiss-Prot [38]. The ZD98 consisted of 98 apoptosis protein sequences, 43 of which are cytoplasmic proteins, 30 plasma membrane-bound proteins, 13 mitochondrial proteins and 12 else proteins. The CL151 consisted of 151 proteins, 52 nuclear proteins, 45 cytoplasmic proteins, 21 mitochondrial proteins and 33 membrane proteins. The accession numbers can be referred to [16,36].

A new much larger dataset, ZW225, was also constructed to further test the prediction model. The new dataset included 225 apoptosis proteins in four subcellular localizations with 41 nuclear proteins, 70 cytoplasmic proteins, 25 mitochondrial proteins and 89 membrane proteins. All the proteins in this dataset were selected from Swiss-Prot [38] release 50.3 using the same selection rule as ZD98 and CL151.

2.2. EBGW of protein sequence

For many quite different things, we can treat them as one if they have some same characters. This is the main idea of coarse-grained. It is well known that the three-dimensional structure of protein is more conservative than its protein sequence. In the process of folding, the insertion, deletion or permutation of single amino acid residue may not destroy the three-dimensional structure. The most important influencing factor of protein folding is the unique character of amino acid residue. Thus, in the following, we present a new encoding scheme of amino acid sequence based on the different character of amino acid residue and coarse-grained idea.

Considering the hydrophobicity and charged character, we can divide the 20 amino acid residues into four different classes as follows [39]:

- neutral and non-polarity residue
- neutral and polarity residue
- acidic residue
- basic residue

Thus, we can get three combinations, each of which can partition the 20 amino acid residues into four different classes as follows:

2. C2 = {Q, N, S, T, Y, C}
3. C3 = {D, E}
4. C4 = {H, K, R}

Definition 1 (Characteristic sequence). Let \( A(n) = a_1, a_2, \ldots, a_n \) be a protein sequence, we can transform it into three binary sequences by three homomorphic maps \( \Phi(A(n)) = \Phi_1(a_1), \Phi_2(a_2), \ldots, \Phi_n(a_n) \) where \( i = 1, 2, 3 \) which are defined as follows:

\[
\Phi_1(a) = \begin{cases} 
1 & \text{if } a \in C_1 \cup C_2 \\
0 & \text{if } a \in C_3 \cup C_4
\end{cases} \quad (j = 1, 2, \ldots, n)
\]

\[
\Phi_2(a) = \begin{cases} 
1 & \text{if } a \in C_1 \cup C_3 \\
0 & \text{if } a \in C_2 \cup C_4
\end{cases} \quad (j = 1, 2, \ldots, n)
\]

\[
\Phi_3(a) = \begin{cases} 
1 & \text{if } a \in C_1 \cup C_4 \\
0 & \text{if } a \in C_2 \cup C_3
\end{cases} \quad (j = 1, 2, \ldots, n)
\]

Denote \( H(n) = \Phi(A(n)) = h_1, h_2, \ldots, h_n \) be a characteristic sequence, the weight of \( H(n) \) is defined as the number of occurrences of digit 1 in \( H(n) \).

For simplicity, in the following text we denote \( H(n) = h_1, h_2, \ldots, h_n \) as any characteristic sequence of three defined above.

Definition 2 (Weight). Let \( H(n) = h_1, h_2, \ldots, h_n \) be a characteristic sequence, the weight of \( H(n) \) is defined as the number of occurrences of digit 1 in \( H(n) \).

We can see that the weight of characteristic sequence is dependent on the sequence length. So it could not be applied to the comparison or analysis of sequences with different lengths.

Definition 3 (Normalized weight). Let \( H(n) = h_1, h_2, \ldots, h_n \) be a characteristic sequence, the normalized weight \( w(n) \) is defined as the frequency of digit 1 occurs in \( H(n) \), that is \( w(n) = \frac{h_1}{n} \), where \( n \) is the weight of \( H(n) \).

Definition 4 [Encoding based on grouped weight]. Let \( H(n) = h_1, h_2, \ldots, h_n \) be a characteristic sequence, assume \( L \) is a positive integer, we can partition \( H(n) \) into \( L \) pieces of subsequence. The length of each subsequence is progressive increase. Let \( H([n/L], L) \) \( (k = 1, 2, \ldots, L) \) be subsequences of \( H(n) \) whose length are \( \lfloor n/L \rfloor \), \( \lfloor (n+1)/L \rfloor \), \( \ldots \), \( \lfloor (n+L-1)/L \rfloor \) where \( \lfloor \cdot \rfloor \) is the operation returning a number down to the nearest integer, and \( w([n/L], L) \) \( (k = 1, 2, \ldots, L) \) be the normalized weight of \( H([n/L], L) \) \( (k = 1, 2, \ldots, L) \), we can get \( W = \{w([n/L], L), w([2n/L], L), \ldots, w([Ln/L], L)\} \) which we call as the EBGW string of characteristic sequence \( H(n) \).

Thus, given a protein sequence \( A(n) = a_1, a_2, \ldots, a_n \) we can transform it into three characteristic sequences \( H(n)^1, H(n)^2, H(n)^3 \) by using Definition 1. For each characteristic sequence \( H(n)^i \) \( (i = 1, 2, 3) \), it can be encoded into a \( L \)-dimension vector \( W^i \) \( (i = 1, 2, 3) \) with Definition 4. That is to say, we can transform a protein sequence into a \( 3L \)-dimension vector \( X = [W^1, W^2, W^3] = [x_1, x_2, \ldots, x_{3L}] \), we call \( X \) as the EBGW string of protein sequence \( A \).

In EBGW approach, characteristic sequence is introduced based on the concept of coarse-grained. It reflects the distribution of residues with the same unique characteristic and portrays the essence of protein sequence. Although the amino acid composition is very convenient to calculate, the information contained in the protein sequence is reduced considerably. In EBGW approach, grouped presentation can contain more information in the protein sequence. If grouping based on the amino acid composition, a protein sequence can be transformed into a \( 2L \)-dimension vector, where \( L \) is the number of groups. However, grouping based on characteristic sequence, a protein sequence can be transformed into a \( 3L \)-dimension vector. The computational complexity is largely decreased. From Definition 4, we know that the larger the value of \( L \) used, the more information of EBGW approach contained, and the higher accuracy of test reached. On the other hand, information may be less when \( L \) equals the length of protein sequence. So the optimal value of \( L \) should be carefully chosen for different datasets.

2.3. Support vector machine

The support vector machine (SVM) is a new machine learning method, which has been used for many kinds of pattern recognition problems. The basic idea of applying SVM to pattern classification can be stated briefly as follows. First, map the input vectors into one feature space (possible with a higher dimension). Then within this feature space, construct a hyperplane which can separate two classes. The mapping function will involve only the relatively low-dimensional vectors in the input space and dot products in the feature space. SVM training always seeks a globally optimized solution and avoids over-fitting, so it is of the ability to deal with a large number of features. More detailed descriptions of the SVM method can be found in Vapnik’s publications [40].

There are several parameters in the SVM, including the kernel function and regularization parameter \( C \). The basic SVM algorithm is designed for binary classification problems only. Nevertheless, there are several methods to extend the SVM for classifying multi-class problems, for example “One-Versus-Rest (OVR)”, “One-Versus-One (OVO)”, and DAGSVM. This paper used the “One-Versus-One” strategy. For a \( k \)-classification problem, the OVO strategy constructs \( k(k-1)/2 \) classifiers with each one trained with the data from two different classes. The software toolbox used to implement the SVM in this paper was LIBSVM by Chang and Lin [41]. The software toolbox can be downloaded from: http://www.csie.ntu.edu.tw/cjlin/libsvm/.

2.4. Test and assessment

In statistical prediction, the following three cross-validation tests are often used to examine the power of a predictor: independent dataset test, sub-sampling test, and Jackknife test. Of these three, the Jackknife test is thought the most rigorous and objective one (see [42] for a comprehensive review in this regard), and hence has been used by more and more investigators [1–32] in examining the power of various prediction methods. In our study, Jackknife test is performed. Each protein sequence in the dataset is singled out in turn as a test sample, and the remaining protein sequences are used as a training dataset to predict the test sample’s subcellular location. The overall prediction accuracy (Qv), individual sensitivity (Sen(i)) and Matthew’s correlation coefficient (MCC) are used to measure the prediction performance of our work.
Let denote $A = (a_{ij})_{C \times C}$ as confusion matrix, where $a_{ij}$ $(i,j = 1, 2, \ldots, \mu)$ is the number of proteins which belong to the class $i$, while predicted to the class $j$. Denote $ob_i$, and $M$ as the number of proteins in each class and the total number in dataset, respectively, we have

$$ob_i = \sum_{j=1}^{\mu} a_{ij} \quad (i, j = 1, 2, \ldots, \mu)$$

$$M = \sum_{i=1}^{C} \sum_{j=1}^{C} a_{ij} \quad (i, j = 1, 2, \ldots, \mu)$$

$$Q_i = \sum_{j=1}^{\mu} a_{ij}/M \quad (i = 1, 2, \ldots, \mu)$$

$$\text{Sen}(i) = a_{ii}/ob_i \quad (i = 1, 2, \ldots, \mu)$$

$$\text{MCC}_i = \frac{TP \times TN_i - FP \times FN_i}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

where $TP$ is number of correctly predicted sequences of location $i$, $TN_i$ the number of correctly predicted sequences not of location $i$, $FP$, the number of under-predicted sequences and $FN_i$ is the number of over-predicted sequences.

### 3. Results and discussion

The results with SVM and EBGW in the Jackknife test for three datasets are shown in Table 1. For the dataset of CL151 and ZW225, the fourth class is nuclear proteins, while for the dataset ZD98, it is else proteins. The overall prediction accuracy for ZD98, CL151 and ZW225 are 92.86%, 91.39% and 83.11%, respectively. The prediction performance for three datasets is variant. This may be owing to the discrepancy of dataset traits, e.g. the size, sequence homologous and unbalance of subset. And the sequence homologous will greatly affect the prediction performance.

#### 3.1. Effect of sequence homologous

To investigate the influence of the sequence homologous on prediction models, we established four subsets (ZW_90, ZW_70, ZW_50 and ZW_25). Those subsets are derived from the dataset ZW225 by a culling program PISCES[43] to winnow those sequences which have $\geq 90$, $\geq 70$, $\geq 50$ and $\geq 25$ sequence identity to any other in a same subset. The results of the four subsets and ZW225 are shown in Table 2. It is seen that when the homology of the dataset decreases, the performance of the classification system also lowers. That is to say, evaluating the performance of prediction model, the construction of dataset is vitally important.

#### 3.2. Effect of grouped number

From Section 2.1, we know that for different datasets and prediction models, the optimal choice of grouped number $L$ is different. The value of $L$ is from 1 to the length of protein sequence. Fig. 1 shows the effect of grouped number $L$ on CL151 and ZW225 in Jackknife test. From Fig. 1, we know that with the increment of $L$, the prediction accuracy ascend, but with a little surge. This is because when the $L$ is large, the zero elements (noises) in the feature extraction method will manifold, which will lead to the instability of prediction accuracy.

### 3.3. Comparison of feature extraction

In order to show the efficiency of EBGW, we compare it with other methods on the same dataset ZD98, with the same prediction algorithm in Jackknife test. These methods include the amino acid composition (AAC) [16] and instability index (Instab)[37]. Fig. 2a shows the prediction accuracies of EBGW and AAC with component-couple algorithm on ZD98. Fig. 2b shows the prediction results of EBGW and Instab with SVM on ZD98. From Fig. 2a, we know that the overall prediction accuracy of EBGW achieve 85.7% with component-coupled algorithm, which is 13.2 percentile higher than AAC with the same algorithm on the same dataset in the same test. Especially for the else class proteins, the increment is 41.7 percentile; the increment of prediction accuracy for membrane and mitochon-

### Table 1

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Sensitivity (%)</th>
<th>MCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZD98</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cyto</td>
<td>97.67</td>
<td>0.90</td>
</tr>
<tr>
<td>Memb</td>
<td>90.00</td>
<td>0.88</td>
</tr>
<tr>
<td>Mito</td>
<td>92.31</td>
<td>0.91</td>
</tr>
<tr>
<td>Else(nucl)</td>
<td>83.33</td>
<td>0.90</td>
</tr>
<tr>
<td>$Q_1$</td>
<td>92.86</td>
<td></td>
</tr>
<tr>
<td>CL151</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cyto</td>
<td>93.33</td>
<td>0.85</td>
</tr>
<tr>
<td>Memb</td>
<td>89.44</td>
<td>0.89</td>
</tr>
<tr>
<td>Mito</td>
<td>76.19</td>
<td>0.86</td>
</tr>
<tr>
<td>Else(nucl)</td>
<td>94.23</td>
<td>0.93</td>
</tr>
<tr>
<td>$Q_1$</td>
<td>91.39</td>
<td></td>
</tr>
<tr>
<td>ZW225</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cyto</td>
<td>90.00</td>
<td>0.80</td>
</tr>
<tr>
<td>Memb</td>
<td>93.26</td>
<td>0.83</td>
</tr>
<tr>
<td>Mito</td>
<td>60.00</td>
<td>0.62</td>
</tr>
<tr>
<td>Else(nucl)</td>
<td>63.41</td>
<td>0.66</td>
</tr>
</tbody>
</table>

### Table 2

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Sensitivity (%)</th>
<th>MCC</th>
<th>$Q_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZW225</td>
<td>90.00</td>
<td>93.26</td>
<td>90.00</td>
</tr>
<tr>
<td>ZW_90</td>
<td>88.24</td>
<td>90.14</td>
<td>50.00</td>
</tr>
<tr>
<td>ZW_70</td>
<td>84.91</td>
<td>79.07</td>
<td>27.78</td>
</tr>
<tr>
<td>ZW_50</td>
<td>83.33</td>
<td>68.75</td>
<td>29.41</td>
</tr>
<tr>
<td>ZW_25</td>
<td>83.33</td>
<td>59.09</td>
<td>37.50</td>
</tr>
</tbody>
</table>

Influence of sequence homologous on CL151 and ZW225 in Jackknife test. From Fig. 1, we know that with the increment of $L$, the prediction accuracy ascend, but with a little surge. This is because when the $L$ is large, the zero elements (noises) in the feature extraction method will manifold, which will lead to the instability of prediction accuracy.

Fig. 1. Effect of the grouped number $L$ on CL151 and ZW225 in Jackknife test.
drial proteins is 20.0% and 23.1%, respectively. From Fig. 2b, we also find that the prediction accuracies of EBGW are all higher than that of Instab method. The results show that EBGW is effective and helpful for prediction of apoptosis protein subcellular location because it can extract more structure and function information from protein sequence.

3.4. Comparison of prediction models

Table 3 listed the prediction results by different prediction models on ZD98 in Jackknife test. We denoted the model with amino acid composition and component-coupled algorithm [16] as AAC_CCA; model with protein instability index and SVM [37] as Instab_SVM; model with dipeptide composition and increment of diversity algorithm [36] as Dipep_Diver; model with EBGW and SVM as EBGW_SVM.

From Table 3, we can find that the prediction capacity of EBGW_SVM is stronger than that of other existing models. The overall prediction accuracy of EBGW_SVM is 20.4, 8.2 and 15.3 percentile higher than that of AAC_CCA, Dipep_Diver and Instab_SVM, respectively. Especially for the else class proteins, the prediction accuracy of other models not exceed 50.0%, while that of EBGW_SVM achieves 83.3%. From Table 3, we also find that among the existing three models, the prediction capacity of Dipep_Diver is relatively higher, so we further compare it with EBGW_SVM on CL151 and ZW225.

Fig. 3 shows the comparison of EBGW_SVM and Dipep_Diver on CL151. We can see that the overall prediction accuracy and accuracies of each class of EBGW_SVM are all higher than those of Dipep_Diver. Especially for the nuclear class proteins, the increment is about 26.9 percentile.

For the new dataset ZW225, the prediction results of two models are listed in Table 4. The overall prediction accuracy of EBGW_SVM reach 83.11%, which is 16.00 percentile higher than that of Dipep_Diver. Especially for the membrane proteins, the prediction accuracy of EBGW_SVM is 93.26%, 41.57% higher than that of Dipep_Diver model. But we also notice that for the nuclear and mitochondrial proteins, the prediction accuracy of EBGW_SVM are lower than those of Dipep_Diver model. Those results show that the prediction

Table 3
Prediction results with different models on ZD98 in Jackknife test

<table>
<thead>
<tr>
<th>Model</th>
<th>Sensitivity for each class (%)</th>
<th>Q1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cyto</td>
<td>Memb</td>
</tr>
<tr>
<td>AAC_CCA [16]</td>
<td>42/43 = 97.7</td>
<td>22/30 = 73.3</td>
</tr>
<tr>
<td>Dipep_Diver</td>
<td>38/43 = 88.4</td>
<td>27/30 = 90.0</td>
</tr>
<tr>
<td>Instab_SVM</td>
<td>33/43 = 76.8</td>
<td>25/30 = 83.3</td>
</tr>
<tr>
<td>EBGW_SVM</td>
<td>42/43 = 97.7</td>
<td>27/30 = 90.0</td>
</tr>
</tbody>
</table>

Table 4
Comparison of EBGW_SVM and Dipep_Diver on ZW225 in Jackknife test

<table>
<thead>
<tr>
<th>Model</th>
<th>Sensitivity for each class (%)</th>
<th>Q1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Nucl</td>
<td>Cyto</td>
</tr>
<tr>
<td>Dipep_Diver</td>
<td>29/41 = 70.73</td>
<td>57/70 = 81.43</td>
</tr>
<tr>
<td>EBGW_SVM</td>
<td>26/41 = 63.41</td>
<td>63/70 = 90.00</td>
</tr>
</tbody>
</table>
capacity of different models are complementary, if we can better joint them together, the prediction accuracies will be further improved.

4. Conclusion

Based on coarse-grained description and grouped, a new feature extraction method named EBGW is presented, and applied to apoptosis protein subcellular location prediction with SVM and component-coupled algorithm on three datasets. Compared with other feature extraction methods, EBGW is shown more effectively in representing the protein structure information from protein sequences. Moreover, EBGW_SVM prediction model is stronger in predicting apoptosis protein subcellular locations than other existing prediction models. The experiment results show that EBGW approach is convenient to calculate and provides an effective tool to extract valuable information from protein sequences, which may be a useful tool in other assignment problems in proteomics and genome research.

Acknowledgements: The authors thank Dr. Zhou G.P., Prof. Li Q.Z. and Dr. Chen Y.L. for providing the datasets and their helpful comments. We thank Mr. Chih-Chung Chang and Chih-Jen Lin for providing the LIBSVM software. We are also grateful to anonymous reviewers for their constructive comments. This work was supported in part by the National Natural Science Foundation of China (No. 60603054).

References


