Prediction of the subcellular location of apoptosis proteins based on approximate entropy

Chaohong Song Feng Shi* Corresponding author Xuan Ma
College of Science, Huazhong Agricultural University, Wuhan, Hubei province, 430070, China.
chh_song@mail.hzau.edu.cn, shifeng66@gmail.com, ma_xuan@mail.hzau.edu.cn
doi: 10.4156/jcit.vol4.issue4.17

Abstract

A novel approach based on the concept of approximate entropy is proposed here to predict apoptosis protein subcellular localization. By using of support vector machine and tested on a known dataset which includes 317 apoptosis proteins, the overall prediction accuracy is 85.2%, which is better more than the results of the method which used the hydrophathy distribution of protein sequences and increment of diversity to predict apoptosis protein subcellular localization. In order to validate the feasibility of the method ulteriorly, we do the same work on a new dataset including 1551 apoptosis proteins, the overall prediction accuracy is 78.9%, these results show that the approximate entropy representation of protein sequences are very useful for predicting subcellular location of apoptosis protein.

Keywords

Apoptosis Protein, Subcellular Location, Support Vector Machine, Approximate Entropy.

1. Introduction

Apoptosis is a fundamental process controlling normal tissue homeostasis by regulating a balance between cell proliferation and death[1]. it is considered to have a key role in some several devastating diseases, and also provide many targets for therapeutic intervention. To understand the apoptosis mechanism and functions of various apoptosis proteins, it would be helpful to obtain information about their subcellular location, because the subcellular location of apoptosis proteins is closely related to their function. Scientists usually deal with a number of protein sequences already known belonging to apoptosis proteins. However, it is both time-consuming and costly to determine which specific subcellular location a given apoptosis protein belongs to. Confronted with such a situation, can we develop a fast and effective way to predict the subcellular location for a given apoptosis protein based on its amino acid sequence? Chou [2] proposed a new method in which the covariant discriminate algorithm was augmented to incorporate the quasi-sequence order effect, in order to take into account the sequence-order effects and improve the prediction quality, Chou [3] had further incorporated the quasi-sequence order effect and introduced the concept of “pseudo-amino-acid composition”, Feng [4] proposed a new representation of unified attribute vector, all of proteins have their representative points on the surface of the 20-D globe. The representative points of the proteins in the same family or with the higher sequence identity are closer on the surface, The overall predictive accuracy could be improved from 3% to 5% for different databases with this simply modification of the usage of the amino acid composition. Zhou et al [5] attempted to identify the subcellular location of apoptosis proteins according to their sequences by means of the covariant discriminant function, which was established based on the Mahalanobis distance and Chou’s invariance theorem. A series of new powerful approaches have been developed by many authors [6~9], recently, Chen and Li [10~11] utilized the measure of diversity and increment of diversity to predict the subcellular location of apoptosis proteins. In their research, 317 apoptosis proteins are selected from the SWISS-PROT database and classified into the six subcellular locations. The overall predictive success rate is 82.7% and 84.2% by the jackknife tests respectively. These results indicated that the subcellular location of apoptosis proteins are predictable to a considerably accurate extent if a good vector representation of protein can be established. It is expected that, with a continuous improvement of vector representation methods by incorporating amino acid properties, and by using more powerful mathematics methods, some theory predictive method might eventually become a useful tool in this area.

In this paper, we would like to use a different strategy, the support vector machines, combining with the approximate entropy representation of the protein...
sequences, to predict subcellular location of apoptosis protein. In order to compare with other approach, The dataset we used is the same as that of [10]. By jackknife test, the overall predictive success rates are 85.2%, which is enhanced more than 3%. In order to validate the feasibility of the method ulteriorly, we do the same work on a new dataset which includes 1551 apoptosis proteins, the overall prediction accuracy is 78.9%. These results show that the approximate entropy representation of protein sequences are very useful for predicting subcellular location of apoptosis protein. our approach can play a complementary role to the subcellular location of apoptosis protein.

2. Material and Method

2.1 Data sets

In this work, the dataset 1 was provided by Chen and Li (denoted as CL317,see[10]). The dataset is generated by selecting the sequence length with more than 80 amino acids and there is a decided single subcellular location from 846 apoptosis proteins in SWISSPROT (version 49.0)[12]. The dataset contains 317 apoptosis proteins which was classified into the six subcellular locations (described as Table 1).

<table>
<thead>
<tr>
<th>Subcellular localization</th>
<th>Number of proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cytoplasmic</td>
<td>112</td>
</tr>
<tr>
<td>Membrane</td>
<td>55</td>
</tr>
<tr>
<td>Mitochondrial</td>
<td>34</td>
</tr>
<tr>
<td>Secreted</td>
<td>17</td>
</tr>
<tr>
<td>Nuclear</td>
<td>52</td>
</tr>
<tr>
<td>Endoplasmic reticulum</td>
<td>47</td>
</tr>
<tr>
<td>sum</td>
<td>317</td>
</tr>
</tbody>
</table>

In order to validate the effectiveness of the new prediction method, we also use the method into an expanded dataset, this dataset is generated by selecting the sequence length with more than 80 amino acids and from apoptosis proteins in SWISSPROT (version 56.9)[12]. At first, only those entries annotated with “apoptosis” in the ID (identification) fields were collected, then sequences annotated with ambiguous or uncertain words, such as “potential”, “probable”, “probably”, “maybe”, and “by similarity” were excluded, the protein which tagged two or more different subcellular location were also removed, and sequences with less than 80 amino acid residues and sequences annotated with “fragment” were removed because they might be just fragments. To avoid any homology bias, we reduce sequence redundancy to 90%. In order to keep balance, only those subcellular which with more than 60 proteins were hold. At last, we obtain the 1551 apoptosis proteins which are classified into the four subcellular locations: (1) 305 cytoplasmic proteins, (2) 550 membrane proteins, (3) 228 secreted proteins, (4) 468 nuclear proteins (see Tables 2).

<table>
<thead>
<tr>
<th>Subcellular localization</th>
<th>Number of proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td>cytoplasmic</td>
<td>305</td>
</tr>
<tr>
<td>membrane</td>
<td>550</td>
</tr>
<tr>
<td>secreted</td>
<td>228</td>
</tr>
<tr>
<td>nucleus</td>
<td>468</td>
</tr>
<tr>
<td>sum</td>
<td>1551</td>
</tr>
</tbody>
</table>

2.2. Schemes of feature parameters

2.2.1. The approximate entropy representation of protein sequences

The approximate entropy is a measure of system complexity [13], it has been widely used to deal with physiological signal [14]. A protein sequence can be assumed to be an short time series if every amino acid of it is replaced by the relevant value of hydrophobic amino acids, then the approximate entropy can be regarded as the feature of the complexity of protein sequences. The approximate entropy of the protein sequences is defined as follow:

\[
ApEn(m, r) = \phi^m(r) - \phi^{m+1}(r)
\]

Where,

\[
x(i) = (u(i), u(i+1) \cdots u(i+m-1)), \quad i = 1, 2 \cdots N - m + 1
\]

is protein sequences,

\[
\phi^m(r) = \frac{1}{N - m + 1} \sum_{i=1}^{N-m+1} \ln C^m_i(r),
\]

\[
C^m_i(r) = \sum_j \text{sgn}(r - d(x(i), x(j))),
\]

\[1 \leq i \leq N - m + 1\]

\[d(x(i), x(j)) = \max_{i \neq j} |u(i+k-1) - u(j+k-1)|
\]
Given \( N \) data points, it is need to choose the filter parameter \( r \) and \( m \), \( r \) is similarity criterion of the protein sequences, if the sequence is infinite, it approach zero, otherwise, the best value isn’t definitude. \( m \) is mode dimension, which defines length of the mode, different \( m \) expresses different length of amino acids pairs. Here we select \( m=2,3,4 \) and \( r=0.1,0.15,0.2,0.25 \), hence we can get 12 approximate entropy.

### 2.2.2. Feature vector

In this research, every protein is represented as a point or a vector in a 32-D space, the first 20 components of this vector were supposed to the occurrence frequencies of the 20 amino acids in the protein sequence, and the last 12 components of this vector were the 0.1 times approximate entropy of the protein sequence. Hence all proteins have their representative points in the 32-dimension Hilbert space.

### 2.3. Support Vector Machine

Support Vector Machine (SVM) is one type of learning machines based on statistical learning theory. Here input data is viewed as two sets of vectors in an n-dimensional space, in that space the SVM will construct a separating hyperplane to maximize the margin between the two datasets. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the neighboring data point of both classes.

Suppose \( v_i \in \mathbb{R}^n \) \( (i=1,2,\ldots,N) \) are the feature vectors representing protein sequences, the learning task of SVM is to solve the following convex quadratic programming (QP) problem:

\[
\max \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j x_i x_j k(v_i, v_j) \\
\text{Subject to} \quad \sum_{i=1}^{N} \alpha_i = 1 \\
\quad \alpha_i \geq 0, \quad 0 \leq \alpha_i \leq C \\
\quad \sum_{i=1}^{N} \alpha_i y_i = 0
\]

where the form of the decision function is

\[
f(x) = \text{sgn} \left( \sum_{i=1}^{N} \alpha_i x_i k(v_i, v_j) + b \right)
\]

Where labels \( y_i \in \{+1,-1\} \) \( i=1,2\ldots N \), standing respectively of the positive set and negative set. For the kernel, the radial basis function (RBF) was selected here because it is widely used by researchers. The RBF kernel is defined by the following equation:

\[
k(v_i, v_j) = \exp(-\gamma |v_i - v_j|^2)
\]

The parameter \( \gamma \) and regularization parameter \( C \) are adjusted in training to produce reliable performance. As \( \gamma \) becomes smaller, the decision boundary for discriminating positive and negative examples becomes smoother. \( C \) controls the trade-off between training error and margin. We determined the two parameters by trial and error. Other options of SVM software are set to their default.

A complete description to the theory of SVMs for pattern recognition is in Vapnik’s book [15]. SVMs have been used in a range of bioinformatics problems including protein subcellular location [16], protein–protein interactions prediction [17], protein secondary structure prediction [18].

In this paper, We used the OSU_SVM[19] which is an implementation of SVM for the problem of pattern recognition.

### 2.4. Evaluation of the Performance

The final correct prediction rate and reliability of the method is determined by measuring the sensitivity(Sn), Matthew’s Correlation Coefficient (MCC), and the total prediction accuracy(Ac), the Sn, MCC, and Ac are computed by the following equations, respectively.

\[
Ac = \frac{\sum_{i=1}^{k} p(i)}{N}
\]

\[
Sn(i) = \frac{p(i) / obs(i)}{p(i) + n(i) - p(i) - obs(i)}
\]

\[
MCC(i) = \frac{p(i) \cdot n(i) - u(i) \cdot o(i)}{\sqrt{(p(i) + u(i)) \cdot (p(i) + o(i)) \cdot (n(i) + u(i)) \cdot (n(i) + o(i))}}
\]

Where \( N \) is the total number of sequences, \( k \) is the class number, \( obs(i) \) is the number of sequences observed in localization \( i \), and \( p(i) \) is the number of correctly predicted sequences of localization \( i \), \( n(i) \) is the number of correctly predicted sequences not of localization \( i \), \( u(i) \) is the number of under- predicted sequences of localization \( i \) and \( o(i) \) is the number of over- predicted sequences of localization \( i \).

### 2.5. Performance test

Cross-validation test can reflect the extrapolating effectiveness of a prediction method. Among all cross-validation tests, the jackknife test is thought to be the most rigorous and objective one[10]. During jackknifing, the subcellular location of each apoptosis
protein is identified by the rule parameters derived using all the other apoptosis proteins except the one that is being identified. In this paper we use jackknife test on the dataset.

3. Results and discussion

Table 3. Comparison of prediction performance for different methods on the 317 apoptosis proteins dataset with jackknife test

<table>
<thead>
<tr>
<th>Location</th>
<th>Diversity incrementa</th>
<th>Diversity incrementb</th>
<th>(c = 52, c = 23)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sn</td>
<td>MCC</td>
<td>Sn</td>
</tr>
<tr>
<td>Cy</td>
<td>81.3</td>
<td>0.80</td>
<td>91.1</td>
</tr>
<tr>
<td>Me</td>
<td>81.8</td>
<td>0.77</td>
<td>89.1</td>
</tr>
<tr>
<td>Mi</td>
<td>85.3</td>
<td>0.74</td>
<td>79.4</td>
</tr>
<tr>
<td>Se</td>
<td>88.2</td>
<td>0.68</td>
<td>58.8</td>
</tr>
<tr>
<td>Nu</td>
<td>82.7</td>
<td>0.73</td>
<td>73.1</td>
</tr>
<tr>
<td>En</td>
<td>83.0</td>
<td>0.90</td>
<td>87.2</td>
</tr>
</tbody>
</table>

Ac(%) 82.7  84.2  85.2  

Table 4. The performance of the method on the 1551 apoptosis proteins dataset with jackknife test

<table>
<thead>
<tr>
<th>Location</th>
<th>SVM(γ = 82, c = 68)</th>
<th>Sn(%)</th>
<th>MCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>cytoplasmic</td>
<td>69.5</td>
<td>0.62</td>
<td></td>
</tr>
<tr>
<td>membrane</td>
<td>85.3</td>
<td>0.79</td>
<td></td>
</tr>
<tr>
<td>secreted</td>
<td>71.5</td>
<td>0.71</td>
<td></td>
</tr>
<tr>
<td>nuclear</td>
<td>81.0</td>
<td>0.74</td>
<td></td>
</tr>
<tr>
<td>Ac(%)</td>
<td>78.9</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

From Table 3, we can see that for dataset CL317, based on our method for jackknife test, the sensitivity (Sn), and Mathew Correlation Coefficient (MCC) are 84.8% and 0.83 for cytoplasmic proteins, 89.1% and 0.90 for Membrane proteins, 70.6% and 0.79 for mitochondrial proteins, 70.6% and 0.80 for Secreted proteins, 92.3% and 0.90 for Nuclear proteins, 91.5% and 0.92 for Endoplasmic reticulum proteins, respectively. The overall prediction accuracy (Ac) is 85.2%. Compare with the results of Chen and Li (2007a,b), the sensitivity (Sn), the Mathew’s Correlation Coefficient (MCC) and the overall prediction accuracy (Ac) of the method are increased greatly.

From Table 4, it has correctly identified 212 out of 305 cytoplasmic proteins, 469 out of 550 Membrane proteins, 163 out of 228 Secreted proteins, 379 out of 468 Nuclear proteins, and yield an accuracy of 69.5% for cytoplasmic proteins, 85.3% for Membrane proteins, 71.5% for Secreted proteins, 81.0% for Nuclear proteins, respectively. The total prediction accuracy (Ac) is 78.9%.

The higher predictive rates than the other method and successful prediction on the new dataset indicates that the amino acid composition accompany with approximate entropy representation of the protein sequences is helpful for subcellular location prediction of apoptosis proteins.

4. Acknowledgments

The work was partly supported by Youth Fund of college of science, Huazhong Agriculture University Research Launching Funds (06033) and the Discipline crossing Research Foundation of Huazhong Agricultural University (2008XKJC006).

5. References

Prediction of the subcellular location of apoptosis proteins based on approximate entropy
Chaohong Song, Feng Shi, Xuan Ma


[12] www.ebi.ac.uk/swissprot.


