Using position specific scoring matrix and auto covariance to predict protein subnuclear localization

Rong-Quan Xiao¹, Yan-Zhi Guo², Yu-Hong Zeng², Hai-Feng Tan¹, Xue-Mei Pu², Meng-Long Li¹,²*

¹College of Life Sciences, Sichuan University, Chengdu 610064. ²College of Chemistry, Sichuan University, Chengdu 610064, P.R. China. Correspondence should be addressed to Meng-Long Li(lml@scu.edu.cn). Tel: +86 28 89005151; Fax: +86 28 85412356.

Received September 8th, 2008; revised November 13th, 2008; accepted November 20th, 2008

ABSTRACT

The knowledge of subnuclear localization in eukaryotic cells is indispensable for understanding the biological function of nucleus, genome regulation and drug discovery. In this study, a new feature representation was proposed by combining position specific scoring matrix (PSSM) and auto covariance (AC). The AC variables describe the neighboring effect between two amino acids, so that they incorporate the sequence-order information; PSSM describes the information of biological evolution of proteins. Based on this new descriptor, a support vector machine (SVM) classifier was built to predict subnuclear localization. To evaluate the power of our predictor, the benchmark dataset that contains 714 proteins localized in nine subnuclear compartments was utilized. The total jackknife cross validation accuracy of our method is 76.5%, that is higher than those of the Nuc-PLoc (67.4%), the OET-KNN (55.6%), AAC based SVM (48.9%) and ProtLoc (36.6%). The prediction software used in this article and the details of the SVM parameters are freely available at http://chemlab.scu.edu.cn/predict_SubNL/index.htm and the dataset used in our study is from Shen and Chou’s work by downloading at http://chou.med.harvard.edu/bioinf/Nuc-PLoc/Data.htm.

Keywords: Position Specific Scoring Matrix; Auto Covariance; Support Vector Machine; Protein Subnuclear Localization Prediction

1. INTRODUCTION

The cell nucleus is complex, important subcellular organelle in eukaryotes cell. It organizes the comprehensive assembly of our genes and their corresponding regulatory factors [1]. Meanwhile, it also reflects various intricate biological activities, and controls various kinds of biologic processes [2]. Many proteins, from outside a nuclear, tend to be localized into specific subnuclear locations of the nucleus [3]. If proteins can not be correctly localized into its specific subnuclear locations in human, it will lead to genetic disease [4], cancer [5] or virally infected cells [6]. Thus, it’s desirable to get the knowledge of protein subnuclear localization for in-depth understanding cell biological processes and genomic regulation. However, it is costly and time-consuming to assay the subnuclear localization of proteins by biology experiments [7]. The number of protein sequences is increasing more rapidly than that of identified proteins [7]. So it is of great practical significance to develop computational approaches for identifying the protein subnuclear localizations in cell nucleus. At the same time, many lines of evidences have indicated that computational approaches, such as structural bioinformatics [8], molecular docking [9], pharmacophore modelling [10], QSAR [11,12,13], protein subcellular location prediction [7,14], identification of membrane proteins and their types [15], identification of enzymes and their functional classes [16], identification of proteases and their types [17], protein cleavage site prediction [18,19], and signal peptide prediction [20,21] can provide very useful information for both basic research and drug discovery in a timely manner. The present study is devoted to develop a new method for predicting protein subnuclear localization in hope to stimulate the development of the relevant areas.

Recently, many algorithms have already been developed for predicting protein subcellular localizations [22, 23,24,25,26,27,28,29,30,31,32,33], as reviewed by Chou [7]. Even several web sever have been constructed for predicting subcellular localization of various organisms [14,34,35,36,37]. However, there are only a few computational methods for predicting protein subnuclear localization [38,39,40,41], such as OET-KNN [42], ProtLoc [43], Nuc-PLoc [44], and AdaBoost classifiers [45]. Compared to the conventional amino acid composition (AAC), pseudo amino acid (PseAA) composition [46], originally introduced by Chou [47,48], can include the sequence-order information of sequences. Similarly, the PsePSSM was also proposed by Shen and Chou in order to incorporate the evolution information of proteins [44]. They built a new web server called Nuc-PLoc for predicting protein subnuclear localization by fusing PseAA composition and PsePSSM with a promising prediction result. In this study, we developed a new method by fus-
ing position specific scoring matrix (PSSM) and auto covariance (AC), so that this method can incorporate sequence-order information by AC and the evolutionary information by PSSM. A classifier based on SVM was constructed to predict protein subnuclear localization using jackknife test. The result indicates that our method has successfully enhanced accuracies of the existing methods for predicting protein subnuclear localization.

2. MATERIALS AND METHODS

2.1. Data Sets

In this paper, our dataset is obtained from article by Shen and Chou [44]. And anyone can freely download it at this page (http://chou.med.harvard.edu/bioinf/Nuc-PLoc/ Data.htm). This dataset consists of nine classes and 714 proteins in total. Details of this benchmark dataset are shown in Table 1. S, (i=1, 2, 9) is used to represent each of nine subsets and S represents the total dataset.

2.2. Feature Representations

2.2.1. Auto Covariance (AC)

We selected three common physicochemical properties, hydrophobicity [49], volumes of side chains of amino acids [50], and polarity [51], to represent the structure and function [52], the stereospecific blockade [53] and the electronic property [54] of residues in a protein respectively. These original values were taken from Guo et al. [55] and were first normalized to zero mean value and unit standard deviation (SD) by Equation (1):

$$ P_{i,j} = \frac{P_{i,j} - \bar{P}_j}{S_j} \quad (i=1, 2, 3; j=1, 2, 3, ..., 20.) $$

Where $P_{i,j}$ is the $i$-th descriptor value for $j$-th amino acid, $\bar{P}_j$ is the mean of the $j$-th descriptor of the 20 amino acids and $S_j$ is the value of SD. So each protein sequence was translated into three vectors with each amino acid represented by the normalized values.

Table 1. The benchmark dataset consists of 714 nuclear proteins classified into nine subnuclear localizations

<table>
<thead>
<tr>
<th>Subnuclear localization</th>
<th>Subset</th>
<th>No. of proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chromatin</td>
<td>S_1</td>
<td>99</td>
</tr>
<tr>
<td>Heterochromatin</td>
<td>S_2</td>
<td>22</td>
</tr>
<tr>
<td>Nuclear envelope</td>
<td>S_3</td>
<td>61</td>
</tr>
<tr>
<td>Nuclear matrix</td>
<td>S_4</td>
<td>29</td>
</tr>
<tr>
<td>Nuclear pore complex</td>
<td>S_5</td>
<td>79</td>
</tr>
<tr>
<td>Nuclear speckle</td>
<td>S_6</td>
<td>67</td>
</tr>
<tr>
<td>Nucleolus</td>
<td>S_7</td>
<td>307</td>
</tr>
<tr>
<td>Nucleoplasm</td>
<td>S_8</td>
<td>37</td>
</tr>
<tr>
<td>Nuclear PML body</td>
<td>S_9</td>
<td>13</td>
</tr>
<tr>
<td>Total</td>
<td>S</td>
<td>714</td>
</tr>
</tbody>
</table>

2.2.2. Position Specific Scoring Matrix (PSSM)

A PSSM is a Position Specific Scoring Matrix and is a commonly used representation of motifs (patterns) in biological sequences [64]. So far, this method has been used for predicting protein subcellular localization [65] and subnuclear localization [40,44].

For a protein sequence $P$ with $L$ amino acid residues, PSSM is obtained according to the following Equation [44].

$$ P_{\text{PSSM}} = \begin{bmatrix} P_{1 \rightarrow 1} & P_{1 \rightarrow 2} & \cdots & P_{1 \rightarrow j} & \cdots & P_{1 \rightarrow 20} \\ P_{2 \rightarrow 1} & P_{2 \rightarrow 2} & \cdots & P_{2 \rightarrow j} & \cdots & P_{2 \rightarrow 20} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ P_{i \rightarrow 1} & P_{i \rightarrow 2} & \cdots & P_{i \rightarrow j} & \cdots & P_{i \rightarrow 20} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ P_{L \rightarrow 1} & P_{L \rightarrow 2} & \cdots & P_{L \rightarrow j} & \cdots & P_{L \rightarrow 20} \end{bmatrix} $$

(4)

In Equation (4), where $i \rightarrow j$ describes $i$-th amino acid residue of the protein sequence $P$ being mutated to amino acid type $j$ in the biology evolution process, $P_{i \rightarrow j}$ is the score of this mutation and $L$ is the length of the sequence $P$. Here we used the numerical codes 1, 2, 3… 20 to represent the single character of ordered 20 native amino acids types in Equation (4). To get the $L \times 20$ scores of the $P_{\text{PSSM}}$ in the Equation (4), we used three iterations of PSI-BLAST [66] with default threshold (the default E-value is 0.001) to search the Swiss-Prot database (version 54.4, released on 25 Oct. 2007) for multiple sequence alignment against the protein $P$. Then, the value of $P_{i \rightarrow j}$ is standardized by Equation (5), as given below.

$$ P_{i \rightarrow j}^o = P_{i \rightarrow j} \cdot \frac{1}{20} \frac{1}{20} \sum_{i=1}^{20} P_{i \rightarrow j}^o $$

(5)
(i = 1, 2, 3...L; j = 1, 2, 3...20)

Where $P^{*}$ is the original scores generated by PSI-BLAST, $P_{i,j}$ is a zero mean value over the 20 native amino acids and the value is between -1 and 1. However, because of proteins with different lengths $L$, the matrices of the PSSM descriptor in Equation (4) have different numbers of rows. To gain the uniform matrix for protein sequences of different lengths, we converted the PSSM of protein $P$ to a uniform vector through the Equation (6) [44].

$$P_{PSSM} = \left[ \overline{P_1}, \overline{P_2}, \ldots, \overline{P_j}, \ldots, \overline{P_n} \right] (j = 1,2,\ldots,20) \quad (6)$$

Where $T$ is the transpose operator, $\overline{P_j}$ is the average score over $j$-th column in Equation (4).

Finally, the $P_{PSSM}$ describes the evolutionary information of a protein sample, and AC variables contain the interaction information between two amino acid residues of a sequence. So each protein sequence was converted into a numerical vector by concatenating PSSM and AC. Here, each AC variable was appended a weight factor of 0.05.

2.2.3. Accuracy and Matthew’s Correlation Coefficient (MCC)

To evaluate the performance of this method, two parameters, accuracy and Matthew’s correlation coefficient (MCC), were selected in this article. They are calculated by Equation (7) and Equation (8), respectively.

$$\text{Accuracy} = \frac{TP}{TP + FN}$$

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN) \times (TN + FP) \times (TN + FN) \times (TP + FP)}} \quad (8)$$

Where TP represents the true positive; TN, the true negative; FP, the false positive and FN, the false negative.

3. RESULTS AND DISCUSSION

In statistical prediction, the following three cross-validation methods are often used to examine a predictor for its effectiveness in practical application: independent dataset test, subsampling test, and jackknife test [67]. However, as elucidated in [14] and demonstrated by Eq.50 of [7], among the three cross-validation methods, the jackknife test is deemed the most objective that can always yield a unique result for a given benchmark dataset, and hence has been increasingly used by investigators to examine the accuracy of various predictors (see, e.g., [7,33,68,69,70,71,72,73,74,75,76,77,78,79,80, 81,82]). So in this paper, the jackknife test was chosen to validate the current algorithm. Because the benchmark dataset used has nine subsets, the one-to-one multiclass classification system led to $9^9(9-1)/2=36$ SVM models for one single encoding methods. Meanwhile, for AC variables, the value of $lg$ was optimized as 13 through a series of control experiments, and the value of $p$ is 3. So, the number of AC variables, $D$, is 39 ($D = lg \times p = 13 \times 3 = 39$) according to Equation (3).

Amino acid composition (AAC) has been widely used for predicting subcellular localizations [7,14,22,23,24, 25,26,27,28,30,31,32,34,35,36,37,83,84,85], so it was also used as a substitution model in our study. And thus, three SVM models based on AAC, AC and PSSM, were respectively constructed.

The results according to jackknife test are listed in Table 2. As can be seen from Table 2, the prediction accuracy of PSSM based model is nearly equal to that of AAC based model. However, AC based model gives the lower accuracy of 64.13%. Then we constructed models by fusing the three substitution models, so four fused classifier were built. Table 2 shows that the accuracies of the four fused models are higher than those of the three anterior models. Among those four fused models, the accuracy of the model combining PSSM, AAC and AC is lower than that of PSSM and AC based model that obtains the best performance with an accuracy of 76.45%. So the final SVM model was built based on PSSM and AC. The kernel function of SVM is radio basis function (rbf), and the parameters of C and $\gamma$ are listed in the table by downloading at http://chemlab.scu.edu.cn/predict_ SubNl/index.htm.

In order to further examine the prediction power of the current classifier, the performance of this method was also compared with those of the existing methods on the same training dataset. The results obtained by several algorithms with different substitution models were summarized in Table 3. From Table 3, we can see that the accuracy obtained by Nuc-PLoc [44] is much higher than those of ProtLoc [43], AAC based SVM and OET-KNN [42]. When compared to Nuc-PLoc, our method obtains a better performance with the accuracy of 76.5%. It means our method is successful in predicting protein subnuclear localization only using primary sequences of proteins.
Table 3. Overall accuracy by jackknife tests with different algorithms on the benchmark dataset of Table 1

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Protein sample descriptor</th>
<th>Overall accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>ProtLoc^a</td>
<td>Amino acid composition</td>
<td>261/714=36.6%</td>
</tr>
<tr>
<td>SVM^a</td>
<td>Amino acid composition</td>
<td>349/714=48.9%</td>
</tr>
<tr>
<td>OET-KNN^b,d</td>
<td>PseAA Composition</td>
<td>397/714=55.6%</td>
</tr>
<tr>
<td>Nuc-PLoc^c,d</td>
<td>Fusion of PsePSSM and PseAA Composition</td>
<td>481/714=67.4%</td>
</tr>
<tr>
<td>Our method</td>
<td>Combination of PSSM and AC</td>
<td>546/714=76.5%</td>
</tr>
</tbody>
</table>

a: See Cedano et al. (1997)[86]  
b: See Shen and Chou (2005)[42]  
c: See Shen and Chou (2007)[44]  
d: The results were from Shen and Chou (2007)[44], and the original data could been seen in that article.

Table 4. The MCC values obtained by the jackknife tests with Nuc-PLoc and our method on the benchmark dataset of Table 1

<table>
<thead>
<tr>
<th>Subnuclear localization</th>
<th>Matthew’s correlation coefficient</th>
<th>Nuc-PLoc^a</th>
<th>Our method^b</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chromatin S1</td>
<td>0.60</td>
<td>0.55</td>
<td></td>
</tr>
<tr>
<td>Heterochromatin S2</td>
<td>0.52</td>
<td>0.58</td>
<td></td>
</tr>
<tr>
<td>Nuclear envelope S3</td>
<td>0.53</td>
<td>0.65</td>
<td></td>
</tr>
<tr>
<td>Nuclear matrix S4</td>
<td>0.52</td>
<td>0.61</td>
<td></td>
</tr>
<tr>
<td>Nuclear pore complex S5</td>
<td>0.70</td>
<td>0.72</td>
<td></td>
</tr>
<tr>
<td>Nuclear speckle S6</td>
<td>0.43</td>
<td>0.57</td>
<td></td>
</tr>
<tr>
<td>Nucleolus S7</td>
<td>0.57</td>
<td>0.57</td>
<td></td>
</tr>
<tr>
<td>Nucleoplasm S8</td>
<td>0.31</td>
<td>0.54</td>
<td></td>
</tr>
<tr>
<td>Nuclear PML body S9</td>
<td>0.32</td>
<td>0.51</td>
<td></td>
</tr>
</tbody>
</table>

a: The results were from Shen and Chou (2007)[44], and the original data could been seen in that article.  
b: The classifier fused PSSM and AC.

In addition, to evaluate the stability of our method, the values of the MCC for the nine subsets were compared based on Nuc-PLoc and our current predictor, respectively, as seen in Table 4. For nine subsets, our method yields a higher MCC than Nuc-PLoc, except the subset S1. So, compared to the existing methods, our classifier combined with PSSM and AC has further improved the prediction accuracy of protein subnuclear localization.

4. CONCLUSION

In this paper, a new classifier was developed by fusing PSSM and AC for predicting protein subnuclear localization only using the primary sequences of nuclear proteins. The SVM predictor was constructed based on PSSM and AC. AC variables represent the interactions between amino acids in protein sequences; PSSM describes the evolutionary information. So the method incorporated not only the evolution information, but also the sequence-order information. Compared with the current methods, this method successfully raises the prediction accuracy. Hence, it may be a good supplementary tool for protein function studies.

ACKNOWLEDGEMENT

The authors gratefully thank Shen and Chou for sharing the benchmark dataset. The work was funded by the National Natural Science Foundation of China (No. 20775052).

REFERENCES

for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. Biochemical and Biophysical Research Communications, 360, 339–345.


