Identification of signal peptide sequences based on mean value matrix image

Xuexuan ZHU\textsuperscript{1,a}, and Chuncai XIAO\textsuperscript{1,b,*}

\textsuperscript{1}School of Electromechanical Engineering Xinyu University, Jiangxi Xinyu 338004, China
\textsuperscript{a} 371734429@qq.com, \textsuperscript{b} xiaochuncai123@yeah.net

*corresponding Chun-cai XIAO

Keywords: Signal peptides, Secretory, Mean Value Matrix, Fuzzy K-nearest neighbor.

Abstract: Signal peptides are short peptide chains, which become a key tool in revising cells for gene therapy or developing new drugs. But for predicting the signal peptide sequences, the preliminary step is to distinguish whether the protein sequence is a secretory or non-secretory sequence. The sample with a pseudo amino acid (PseAA) composition representation of the protein sequence so as to incorporate a plentiful amount of protein sequence pattern information in order to increase the prediction precision for the classification. This article, which is based on the value of hydrophobicity, hydrophilicity, side-chain mass for sequence, we put forward a new method -- protein mean value matrix image(MVMI) to predict types from secretory and non-secretory protein sequences. Two geometric moments are on the base of the protein MVMI were collected from each of the protein sequences are made for their PseAA. It could testify by means of the jackknife cross-check method that the total successful rate are over 97%. The experimental results indicate that bioinformatics based on theory methodology can simplify and make experimental studies more intuitive.

1. Introduction

Signal peptide sequence is currently N-terminal extensions with 5–30 amino acids length despite it can also be situated inside a protein or at its N-terminal end[1]. Signal peptide sequence command the entry of almost all the pathway of secretory protein include in prokaryotes and eukaryotes[2,3]. As a matter of fact, signal peptide sequence has become a significant tool to pharmaceutical researchers who for producing effectively drugs will genetically amend animals, plants, and bacteria[4]. However, to reach that goal, the preliminary step is to distinguish the signal peptide for classifying it as secretory or non-secretory sequence.

In recent years, it was found that there was an obvious corelevance between structural classes of proteins and amino acid composition. Researchers put the algorithm developments as the goal followed by predicting the structural type of a protein according to amino acid composition alone[5]. It failed to contain any help of protein sequence information in a protein despite the amino acid composition model is very simple. The method of PseAA composition was raised to avoid the sequence information losing when using the amino acid composition model[6].

The method of PseAA composition was first put forward to improve the performance prediction of membrane protein type[7]. The expression of a known protein sample through a set of discrete decimal numbers, where the first 20 numbers show the 20 components of the traditional amino acid composition while other numbers show the PseAA composition, as defined by its expression way. In order to effectively forecast various attributes of proteins by the PseAA composition, how to extract the features from protein effectively, is the key step of the PseAA components processing methods. The current research was initiated in an attempt to present a totally different method, the known as “mean value matrix image” of protein to deal with this problem.
2. Method

The same data set built by Shen and Chou[8] and was used to indicate the present method for the purpose of facilitating the comparison between the other methods. A protein sequence is made up by 20 natural amino acids whose each character encoding is that: A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, and Y. It is extremely hard to discover its characteristic pattern detailedly when the sequence is interminable[9]. In order to overcome this limitation, we turn to the mean value matrix came out of the amino acid sequence with the help of the physical and chemical properties of amino acid. There were various kinds of physical and chemical properties of amino acids, which were given in Table 1. The characteristics of amino acid are inherently connected with atoms that form the amino acid, and the diversities among the molecules of 20 amino acids just because the side-chain mass. The other significant aspect we thought over seriously the hydrophilic amino acid, which influences the composition of protein sequences and is broadly used in a lot of magazines. Kellis et al. discover that the main reasons is that the hydrophobic of side-chains for the folding of proteins[10]; Chou apply the side-chain mass and the hydrophilic values of amino acids to make up the pseudo amino acid composition to improve the quality of prediction for protein cellular properties[11]. All of these results give effective assistances for the applying hydrophobic, hydrophilic and side-chain mass of protein to make up mean value matrix.

Table 1. The various kinds of physical and chemical characteristics of amino acids

<table>
<thead>
<tr>
<th>Amino acid</th>
<th>Symbol</th>
<th>Hydrophobic</th>
<th>Hydrophilic</th>
<th>Side-chain mass</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alanine</td>
<td>A</td>
<td>0.62</td>
<td>-0.5</td>
<td>15.0</td>
</tr>
<tr>
<td>Cysteine</td>
<td>C</td>
<td>0.29</td>
<td>-1.0</td>
<td>47.0</td>
</tr>
<tr>
<td>Aspartic acid</td>
<td>D</td>
<td>-0.90</td>
<td>3.0</td>
<td>66.0</td>
</tr>
<tr>
<td>Glutamic acid</td>
<td>E</td>
<td>-0.74</td>
<td>3.0</td>
<td>73.0</td>
</tr>
<tr>
<td>Phenylalanine</td>
<td>F</td>
<td>1.19</td>
<td>2.5</td>
<td>91.0</td>
</tr>
<tr>
<td>Glycine</td>
<td>G</td>
<td>0.48</td>
<td>0.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Histidine</td>
<td>H</td>
<td>-0.40</td>
<td>-0.5</td>
<td>82.0</td>
</tr>
<tr>
<td>Isoleucine</td>
<td>I</td>
<td>1.38</td>
<td>1.8</td>
<td>57.0</td>
</tr>
<tr>
<td>Lysine</td>
<td>K</td>
<td>-1.5</td>
<td>3.0</td>
<td>73.0</td>
</tr>
<tr>
<td>Leucine</td>
<td>L</td>
<td>1.06</td>
<td>-1.8</td>
<td>57.0</td>
</tr>
<tr>
<td>Methionine</td>
<td>M</td>
<td>0.64</td>
<td>-1.3</td>
<td>75.0</td>
</tr>
<tr>
<td>Asparagine</td>
<td>N</td>
<td>-0.78</td>
<td>0.2</td>
<td>58.0</td>
</tr>
<tr>
<td>Proline</td>
<td>P</td>
<td>0.12</td>
<td>0.0</td>
<td>42.0</td>
</tr>
<tr>
<td>Glutamine</td>
<td>Q</td>
<td>-0.85</td>
<td>0.2</td>
<td>72.0</td>
</tr>
<tr>
<td>Arginine</td>
<td>R</td>
<td>-2.53</td>
<td>3.0</td>
<td>101.0</td>
</tr>
<tr>
<td>Serine</td>
<td>S</td>
<td>-0.18</td>
<td>0.3</td>
<td>31.0</td>
</tr>
<tr>
<td>Threonine</td>
<td>T</td>
<td>0.03</td>
<td>-0.4</td>
<td>45.0</td>
</tr>
<tr>
<td>Valine</td>
<td>V</td>
<td>1.08</td>
<td>-1.5</td>
<td>43.0</td>
</tr>
<tr>
<td>Tyrosine</td>
<td>Y</td>
<td>0.81</td>
<td>-3.4</td>
<td>130.0</td>
</tr>
</tbody>
</table>

We have a hypothesis to make the meaning more clear, N amino acids make up a protein P, the following notation:

\[ P = R_1 R_2 \cdots R_N \]  \hspace{1cm} (1)

Where \( R_1 \) refers to the first amino acids, \( R_2 \) refers to the second amino acids, and so on. In order to represent a protein sequence from a English letter as a coordinates in space, protein sequences are defined as follows:

\[ B = \{ Coor_{1,1}, Coor_{1,2}, \cdots, Coor_{N,1} \} \]  \hspace{1cm} (2)

Where \( Coor_{n,i} \) (i=1,2,3) refers to space coordinates \( Coor_{n,1}, Coor_{n,2}, Coor_{n,3} \) of the Nth amino acids of the protein sequence, \( Coor_{n,1} \) refers to hydrophobic value of the Nth amino acids of the protein.
sequence, $\text{Coor}_{n,2}$, refers to hydrophilic value of the $N^{th}$ amino acids of the protein sequence, and $\text{Coor}_{n,3}$ refers to side-chain mass value of the $N^{th}$ amino acids of the protein sequence. The mean value matrix is defined as follows:

$$\text{DM} = \begin{cases} \text{dm}(p, q) = \frac{1}{3} \sum_{j=1}^{3} (\text{Coor}_{p,j} + \text{Coor}_{q,j}) \\ \end{cases}$$  \hspace{1cm} (3)$$

Where $(1 \leq p, q \leq N)$.

It is extremely difficult to discover key characteristics by directly reading information from these mean value matrixes. Nevertheless, if the mean value matrixes can be converted to corresponding patterns, their some significant characteristics can also be automatically provided and easier to visualize the results. When transforming the mean value matrix into a visual image with corresponding software, the basic two-dimensional image is chosen because its characteristic is quickly disposed. By this means, software can transform each matrix element into corresponding an image pixels, the values are equal to the gray value. Figure 1 shows the mean value matrix image of non-secretory and secretory sequence.

![Image](non-secretory sequence) ![Image](secretory sequence)

Figure 1. The mean value matrix image of non-secretory and secretory sequence

Geometric moment invariants are extremely common forms of moments and have been widely used in the digital image processing field. The geometric moment $M_{pq}$ about coordinates order $(p+q)$ for a pair of two-dimensional image $f(x, y)$ is defined as follows:

$$M_{pq} = \sum_{x=1}^{N-1} \sum_{y=1}^{N-1} x^p y^q f(x, y) \hspace{1cm} (p, q=0, 1, 2 \cdots)$$  \hspace{1cm} (4)$$

The zero order geometric moment is defined as follows:

$$M_{00} = \sum_{x=1}^{N-1} \sum_{y=1}^{N-1} f(x, y)$$  \hspace{1cm} (5)$$

The first geometric moment is defined as follows:

$$M_{01} = \sum_{x=1}^{N-1} \sum_{y=1}^{N-1} yf(x, y)$$  \hspace{1cm} (6)$$

$$M_{10} = \sum_{x=1}^{N-1} \sum_{y=1}^{N-1} xf(x, y)$$  \hspace{1cm} (7)$$

The image centroid is defined as follows:
When the geometric moment $M_{pq}$ in Eq. (4) is mentioned the center $(x_0, y_0)$ of the image. The central moment $U_{pq}$ is defined as follows:

$$U_{pq} = \sum_{x=1}^{N-1} \sum_{y=1}^{N-1} (x-x_0)^p (y-y_0)^q f(x, y)$$

(9)

The central moment $U_{pq}$ is invariant to any images with displacement and rotation. In order to make automatically these invariant satisfied to the area[18], let us use the form of standardizationas defined by Hu[19]:

$$U^0_{pq} = \frac{U_{pq}}{U_{00}^{(p+q+2)/2}} \quad (p,q=0,1,2\ldots)$$

(10)

According to Eq. (10), Hu defines the following several RTS (rotation, translation, size-scaling) invariant functions, which are described as follows:

$$\varphi_1 = U^0_{20} + U^0_{02}$$

(11)

$$\varphi_2 = (U^0_{20} - U^0_{02})^2 + 4(U^0_{11})^2$$

(12)

$$\varphi_3 = (U^0_{30} - 3U^0_{12})^2 + (3U^0_{21} - U^0_{03})^2$$

(13)

$$\varphi_4 = (U^0_{30} + U^0_{12})^2 + (U^0_{03} + U^0_{21})^2$$

(14)

$$\varphi_5 = (U^0_{30} - 3U^0_{12})(U^0_{30} + U^0_{12})[(U^0_{03} + U^0_{12})^2 - 3(U^0_{21} + U^0_{03})^2]$$

(15)

$$+ (3U^0_{21} - U^0_{03})(U^0_{21} + U^0_{03})[3(U^0_{30} + U^0_{12})^2 - (U^0_{21} + U^0_{03})^2]$$

$$\varphi_6 = (U^0_{20} - U^0_{02})[(U^0_{30} + U^0_{12})^2 - (U^0_{21} + U^0_{03})^2]$$

(16)

$$+ 4U^0_{11}(U^0_{30} + U^0_{12})(U^0_{21} + U^0_{03})$$

$$\varphi_7 = (3U^0_{21} - U^0_{03})(U^0_{30} + U^0_{12})[(U^0_{30} + U^0_{12})^2 - 3(U^0_{03} + U^0_{21})^2]$$

(17)

$$-(3U^0_{12} - U^0_{30})(U^0_{03} + U^0_{21})[3(U^0_{30} + U^0_{12})^2 - (U^0_{03} + U^0_{21})^2]$$

This all about equations are mentioned, only two moments as given by Eqs. (5) and Eqs. (12) are selected as the PseAA components. This is due to initial experiments manifested that containment of the other geometric moments did not work out better consequences because of that the two moments may already carry sufficient information and the other moments would be surplus for the present research.

The superiority of adding the PseAA components is that they can include some key characteristics of a protein sequence by means of a discrete mode as stated above[20]. In this way, a protein sequence can be represented by a vector with the help of the Chou’s PseAA composition, i.e.,
\[ X = [x_1, x_2, \cdots, x_{20}, x_{21}, x_{22}]^T \] 

(18)

Where \( T \) is the inversion operator.

\[
X_k = \begin{cases} 
\frac{f_k}{\sum_{i=1}^{20} f_i + \sum_{j=1}^{2} w_j p_j}, & (1 \leq k \leq 20) \\
\frac{w_{(k-20)} p_{(k-20)}}{\sum_{i=1}^{20} f_i + \sum_{j=1}^{2} w_j p_j}, & (21 \leq k \leq 22) 
\end{cases}
\]

(19)

where \( f_i \) (\( i = 1, 2, \ldots, 20 \)) are the percentages of the 20 natural amino acids in each protein, arranged on the basis of their single alphabetic codes in alphabetical order, \( p_j \) (\( j=1, 2 \)) are come from geometric moments of the mean value matrix of protein sequences and provided by Eqs. (5) and Eqs. (12), also the weight coefficients \( w_1 = 0.75 \times 10^{-4}, w_2 = 1 \times 10^6 \). Meanwhile the Fuzzy K-nearest neighbor algorithm was used to create the classification.

3. Conclusions

Among the methods for examining the effects of prediction, the independent sample test, secondary sample procedures test, and jackknife test, which are employed frequently to checkout the precision of a statistical forecasting technique[15], the jackknife test was regarded as the best objective in practical use. As a consequence, we employ the jackknife cross-validation to checkout our way. The success rates of jackknife test are got with the present HPV-MVMI predictor in recognizing secretory protein sequences types are provided in Table 2.

Table 2: Success rates are got by jackknife test in recognizing the secretory proteins and non-secretory proteins.

<table>
<thead>
<tr>
<th>Organism</th>
<th>PrediSi</th>
<th>Signal-3L</th>
<th>Mean value matrix image</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>91.1</td>
<td>92.3</td>
<td>97.68</td>
</tr>
<tr>
<td>Plant</td>
<td>93.6</td>
<td>95.8</td>
<td>98.89</td>
</tr>
<tr>
<td>Animal</td>
<td>93.2</td>
<td>95.7</td>
<td>97.94</td>
</tr>
<tr>
<td>Eukaryotic</td>
<td>92.1</td>
<td>94.0</td>
<td>97.39</td>
</tr>
<tr>
<td>Gram-positive</td>
<td>94.6</td>
<td>98.1</td>
<td>98.88</td>
</tr>
<tr>
<td>Gram-negative</td>
<td>91.2</td>
<td>94.4</td>
<td>99.03</td>
</tr>
</tbody>
</table>

(The weight coefficients (see Eq. (5) and (12)) for the two PseAA components are \( w_1 = 0.75 \times 10^{-4}, w_2 = 1 \times 10^{-6} \), respectively.)

Using the geometric moments are on the base of mean value matrix images of protein sequence as their PseAA components can contain availably the relevant protein sequence patterns, producing a higher total success rate in forecasting types of six subsets. These types of interrelated patterns are contained in a cluster of tanglesome sequences. Therefore, the types of secretory or non-secretory are just a example for verification. It is promoting significance to indicate that the present innovation method can also be adopted to forecast a range of other protein properties, such as the G protein-coupled receptor functional class, subclass classification of nuclear receptor, protein structural classes, and so on.
References


