RECOGNITION OF OCCURRENCE AND LOCALIZATION OF CLEAVAGE SITE IN SIGNAL PEPTIDES

*1 Zagoruiko N.G., 1 Kutnenko O.A., 2 Nikolaev S.V., **2,3 Ivanisenko V.A.

1 Institute of Mathematics, SB RAS, Novosibirsk, Russia
2 Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
3 State Research Center of Virology and Biotechnology “VECTOR”, Koltsovo, Novosibirsk region, Russia
*e-mail: zag@math.nsc.ru, ** salix@bionet.nsc.ru

Key words: signal peptide, signal anchor, cleavage site, informative characteristics, recognition

Resume

Motivation: Automatic recognition of signal peptides and cleavage sites in proteins is a topical issue for both detection of their cellular localization and solving of applied medical and biotechnological problems. The available recognition methods utilize either amino acid substitution matrices (von Heijne, 1986) or neuronal network algorithms using a 20-letter amino acid code (Nielsen et al., 1997).

In this work, the feasibility of using physicochemical characteristics of amino acids for recognizing cleavage sites in signal peptides is studied. The algorithm AddDel (Zagoruiko, 1999) was applied for selecting the significant characteristics. Cleavage sites were recognized by a sliding test technique. The rule of ”k nearest neighbors” at $k = 1$ was used as a decisive rule.

Results: A method for selecting the informative subset of characteristics was developed and the decisive rule based on the membership function was constructed. Testing of the method proposed using a large amount of experimental data has demonstrated that it detects the cleavage sites correctly and localizes at a rate of 85%.

Introduction

Signal peptides are N-terminal markers of proteins. Eukaryotic proteins carrying these markers are transported through the endoplasmic reticulum membrane; prokaryotic, through the internal membrane. Upon transportation, signal peptidase cleaves the signal peptide. The signal peptides display a common structure: a short positively charged N-terminal region, central hydrophobic region, and a more polar C-terminal region, containing the site where the peptide bond is cleaved.

X-ray structure analysis of signal peptidase demonstrated that the spatial matching of the enzyme in question and a signal peptide required that the amino acid residues at position $-1$ and $-3$ with respect to the cleavage point were small (von Heijne, 1998). N-Terminal regions of type II membrane proteins are signal anchors, providing the integration of these proteins with the membrane. In their physicochemical properties, signal anchors are close to signal peptides; however, unlike the latter, they are not cleaved by signal peptidases (Sakaguchi et al., 1992).

The goal of this work was to study the feasibility of using physicochemical characteristics of amino acids for recognizing signal peptides and signal anchors. The signal peptides were recognized according to the presence of cleavage sites using the method of $k$ nearest neighbors.

Methods and Algorithms

Forming the learning sample. The learning sample used was represented by three sets of fragments of eukaryotic proteins, namely, (1) the fragments containing cleavage sites, (2) the fragments containing anchors, and (3) the fragments of nuclear and cytoplasmic containing neither sites nor anchors. All the necessary data were extracted from http://www.cbs.dtu.dk/services/SignalP/.

Physicochemical properties of amino acids. Two set of amino acid characteristics were used: (1) a Kidera’s set of 10 properties (Kidera et al., 1985), that is, noncorrelating linear combinations of amino acid structural and physicochemical properties, and (2) 434 structural and physicochemical properties from the database http://www.genome.ad.jp/dbget/aaindex.html.

Selecting the window size and decisive rules. At the first stage, we studied the accuracy of cleavage site recognition depending on the window size, that is, on the number of symbols taken into account to the left and right of the cleavage point. The Kidera’s set of properties was used as a character; the window size was changed from 6 to 36 symbols.

At this stage, the sample was formed of fragments with a length L containing cleavage site at the center. The negative sample was formed of a random protein fragments lacking the cleavage site. Thus, a data array was formed of representatives of two images as a table comprising 1012 lines (objects) and $L \times 10$ columns (characters).
A method of sliding test was used for recognition; method of k nearest neighbors, as the decisive rule. The limit window size of 18 symbols was selected due to a limited data volume; all the further studies were performed with this window size.

Selecting the descriptors. The informative amino acid characteristics were selected from each set of physicochemical properties. As the size of the negative sample exceeded considerably that of the positive sample, 7 learning samples containing the same 253 elements from the positive sample and 7 different sets of 253 elements each from the negative sample were formed. The data for the control recognition were not used in the learning samples.

Seven most informative characteristics were selected from the Kidera’s set of ten properties (properties Nos. 3, 5, and 10 were excluded) using the window of a size L = 18. The reliability of recognition in the learning sample using these characteristics amounted to 87.6%.

Then, the informative characteristics were selected out of the 434 physicochemical properties plus the 10 Kidera’s properties. The algorithm AddDel (Zagoruiko, 1999) was used for this purpose. This algorithm combines the concepts of “successive addition of most valuable”(Addition) and “successive deletion of least valuable”(Deletion) properties. It appeared that comparatively small number of properties—from 7 to 30—gave the best results. Overall, 91 property of 444 tested were included into these 7 sets. The Kidera’s properties failed to display their advantages—only one character was included into one of the sets.

These seven sets formed of 444 characteristics were selected as sets of descriptors.

Recognition in the control sequence. The window with a size of 18 symbols was moved along the protein chain with a shift of 1 symbol. Overall, 296,202 control regions were distinguished by this technique; of them, 252 fragments contained the cleavage site and 295,950 were without the site. The decision on the presence or absence of the cleavage site was made at each window position. These two images were recognized according to the seven sets of characteristics described above in parallel. The decision in favor of either first or second image was made by a majority of these seven votes.

When 252 fragments containing the cleavage site were subjected to recognition, the correct decisions rate amounted to 213 or 84.5%. The number of correct decisions while studying the fragments lacking the site amounted 232,403 (78.5%).

Localization of the cleavage site. Estimation of the probability of the presence or absence of the cleavage site in the sliding window was calculated using a modified rule of the k closest neighbors. The distances r1 and r2 to two closest neighbors, one from each image, were found for each control object y. The function of membership in a certain image was specified as $f = 1 - 2*r/k$ if $f \geq 0$, then the object y belongs to the first image; in the opposite case, to the second image. The overall value of the function $F$ was obtained by a mere averaging of the membership functions $f$ obtained for each descriptor set. The mean reliability of recognizing the two images appeared equal to 82.5%.

Recognition of the signal peptides containing cleavage site and of signal anchors. The technique developed was applied to solving the problem of recognition of the three images: Signal Peptide, Anchor, and Negative (fragments of cytoplasmic and nuclear peptides). The signal peptides were recognized according to the presence of cleavage site; the signal anchor, according to the C-terminal boundary of the membrane region in protein. The method of pairwise comparison (Zagoruiko, 2002) was used for recognizing these images. For each pair of images, an individual “competent” space, where these images differ maximally from one another, was formed. Information on the frequencies of all the symbols of the alphabet used at the first two odd positions of the window was used in addition to physicochemical properties. While performing the pairwise comparisons, the solution was made according to the value of the membership function. The overall decision on the membership of control object in one of the three images was made basing on the results of the pairwise comparison.

**Results and Discussion**

Symmetric and asymmetric windows of various sizes were analyzed. It appeared that the window with four odd symbols, two at either side from the center, gave the best results. Thus, the window of size $L$ equalling eight symbols was used in the further studies.

*Localization of cleavage site.* The cleavage site recognition function averaged over 252 control proteins containing the cleavage site is plotted in Fig. 1. The protein fragments are positioned with respect to the cleavage site position (vertical line). The plot in Fig. 2 demonstrates the same function obtained for 252 randomly selected cytoplasmic and nuclear proteins lacking the cleavage site. The length of the fragments amounted to 84 symbols. It is evident from Figs. 1 and 2 that the cleavage site is recognized and localized well.
Recognition of the signal peptides containing cleavage site and of signal anchors.

The results of recognizing signal peptides and signal anchors are listed in Table 1. The control sample comprised 65,607 objects, including 705 signal peptides, 47 signal anchors, and 64,855 objects of the negative sample. These results demonstrate a satisfactory capability of discriminating between signal peptides and signal anchors.

<table>
<thead>
<tr>
<th>Presented</th>
<th>Recognized</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Signal peptide</td>
<td>549</td>
<td>50</td>
</tr>
<tr>
<td>Anchor</td>
<td>14</td>
<td>25</td>
</tr>
<tr>
<td>Negative</td>
<td>11873</td>
<td>12540</td>
</tr>
</tbody>
</table>

The experiments on recognizing signal peptides and signal anchors have demonstrated that the additional information on amino acid frequencies at certain position within the fragments increases the accuracy of recognition.

As an example, plots of changes in type I and II errors versus the threshold value of the membership function $F$ while recognizing signal anchors (47 objects) with and without the data on amino acid frequencies are shown in Fig. 3. The negative set comprised 50,111 objects.

Testing the “oddness” hypothesis. Significance of the positions even and odd with respect to the cleavage site while its recognition was studied. The symbols were numbered from left to right starting from 0 to L-1. The experiments have demonstrated that positions with odd numbers contribute to the recognition to a greater degree, thereby complying with the known $(-1; -3)$ rule (von Heijne, 1998).
Acknowledgements

The work was supported by Russian Foundation for Basic Research (grants № 02 01-00082 to N.G.Zagoruiko and O.A.Kuntenko and № 01-07-90376 to V.A.Ivanisenko and S.V.Nikolaev) and Siberian Branch of the Russian Academy of Sciences (Integration Project № 65 Simulation of Basic Genetic Processes and Systems).

References