Machine Learning Prediction of Amino Acid Sequence Characterization in Protein N-Myristoylation

Ryo Okada\textsuperscript{1} \hspace{1cm} Chigusa Miyakawa\textsuperscript{1}
okada@ib.sci.yamaguchi-u.ac.jp \hspace{1cm} miyakawa@ib.sci.yamaguchi-u.ac.jp

Manabu Sugi\textsuperscript{2} \hspace{1cm} Hiroshi Matsuno\textsuperscript{1} \hspace{1cm} Satoru Miyano\textsuperscript{3}
manabu@yamaguchi-u.ac.jp \hspace{1cm} matsuno@sci.yamaguchi-u.ac.jp \hspace{1cm} miyano@ims.u-tokyo.ac.jp

\textsuperscript{1} Faculty of Science, Yamaguchi University, 1677-1, Yoshiida, Yamaguchi 753-8511, Japan
\textsuperscript{2} Media and Information Technology Center, Yamaguchi University, 1677-1, Yoshiida, Yamaguchi 753-8511, Japan
\textsuperscript{3} Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan

Keywords: protein N-myristoylation, amino acid sequence, machine learning

1 Introduction

Protein N-myristoylation is the lipid modification in which the saturated fatty acid of 14 carbon number binds covalently to N-terminal of virus-based and eukaryotic protein. N-myristoylated proteins have specific sequence at N-terminal called N-myristoylation signal sequence, and this sequence is probably composed of 6 to 9 (up to 20) numbers of amino acid from N-terminal [1][3]. In this study, we suggest an approach to predict the pattern of N-myristylyation signal using the machine learning system BONSAI [2], BONSAI finds rules in combination of a alphabet indexings and decision trees. The alphabet indexing groups letters in positive and negative examples by mapping these letters to fewer number of letters.

2 Positive and Negative Examples of N-Myristoylation to BONSAI

Different types of amino acid sequences which had been identified as N-myristoylation signal by biochemical experiments were given to BONSAI as positive examples. In addition, different types of amino acid sequences randomly selected from a human genome information [4] were provided to BONSAI as negative examples. A few N-myristoylation signal may be mixed in the selected negative examples, because it is known that no more than 0.5% of human proteins are probably N-myristoylated. The first Gly at sequences was omitted from the examples, because the first amino acid from N-terminal of any example is fixed to Gly.

3 Two Rules Characterizing N-Myristoylation Sequence

Figure 1 is the results of computational experiments of BONSAI. (Result1) Each letter in a string of example is substituted to either of the letter 0 or the letter 1, since the indexing parameter was set to 2. The decision tree

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{decision_tree.png}
\caption{Decision tree and Indexing}
\end{figure}
produced by BONSAl has the feature of positions at third and sixth residues: these residues have been classified to “0” by the alphabet indexing. This result is consistent with the fact from biochemical experiment which suggested that the positions of these residues are essential for N-myristoylation signal [3]. Specifically, the decision tree from BONSAl provided almost the same classification rule for N-myristoylation as the biological facts that showed the eleven residues which will place at the third position when the residue Ser comes at the six position in the N-myristoylation signal sequence (Figure 2).

Result 2 BONSAl provided the decision tree which shows the following pattern of N-myristoylation signal. For a sequence, when any of three residues Pro, Phe and Thr is located at just one of the positions 5th, 8th, 9th or 10th of the sequence, or any of these three residues is not located at any of the positions from 2th to 10th of the sequence, this sequence will be N-myristoylated. In other words, for a sequence, when any of Pro, Phe and Thr residues locates at two or more positions of 5th, 8th, 9th and 10th of the sequence, or any of Pro, Phe and Thr residues locates at positions of 2th, 3th, 4th, 6th and 7th of the sequence, this sequence will be N-myristoylated. These results will be supported by the biological fact that these residues are likely to change three-dimensional structure of protein. And, although it has been considered that any amino acid can locate at the 4th residue of N-myristoylation signal sequence, but this result suggests that some residues may be omitted for N-myristoylation signal sequences at the same position.

Acknowledgment
The authors thank to Prof. Toshihiko Utsumi at Yamaguchi University for insightful comments on this study.

References