Evaluation of *N*-myristoylation Prediction Tool using Machine Learning

Sayaka Kado† Ryo Okada†† Manabu Sugii††† Hiroshi Matsuno† Satoru Miyano††††
†Graduate School of Science and Engineering, Yamaguchi University, Japan
††Network Solution Group, Hitachi Chugoku Solutions, Japan
†††Madia and Information Technology Center, Yamaguchi Unibersity, Japan
††††Human Genome Center, University of Tokyo, Japan
†1677-1 Yoshida, Yamaguchi 753-8513, Japan
††11-10, Motomachi, Hiroshima 730-0011, Japan
†††1677-1 Yoshida, Yamaguchi 753-8513, Japan
††††4-6-1 Shirokanedai, Minatoku, Tokyo 108-8639, Japan
†Tel/Fax: +81-83-933-5697
E-mail: manabu@yamaguchi-u.ac.jp

Abstract: Protein sequences constitute molecular complex in an organism. However it is difficult to find a sequence rule such as cascade reaction signals, post translational modification signals and so on. These sequence signals perform an essential role in regulating cellular structure and function. In previous study, we could find sequence rules of *N*-myristoylated proteins easily with computational approach. Subsequently, we have developed a CGI tool to predict *N*-myristoylated proteins with their sequence rules. In this study, we performed accuracy evaluation of our developed CGI tool. As a result, we show that developed CGI tool predict *N*-myristoylated proteins effectively with characteristics of *N*-myristoylated protein sequences.

Introduction

Protein *N*-myristoylation is a lipid modification of proteins. Figure 1 shows a reaction process of protein *N*-myristoylation. Protein *N*-myristoylation is a cotranslational protein modification catalyzed by two enzymes, methionime aminopeptidase and *N*-myristoyltransferase(NMT). The initial methionine(Met) is cleaved from *N*-terminus of amino acid sequences by a methionine aminopeptidase, and then the myristic acid is linked to exposed glycine via an amide bond by NMT. NMT catalyzes the transfer of myristic acid from myristoyl-CoA to the *N*-terminus glycine residue of substrate protein. Most of myristoylated proteins have a physiological activity such as cell signaling protein, expressing specific functions through binding organelle membrane. It is known that membrane binding reaction mediated by myristoylation is controlled variedly, and play a crucial role in functional regulation mechanisms of proteins in cell signaling pathway. *N*-myristoylated proteins have a specific sequence at the *N*-terminus called *N*-myristoylation signal sequence, and this sequence is probably composed of 6 to 9 amino acids (up to 17)[1]. However, it is very hard to predict the *N*-myristoylated proteins with the signal sequence among the genomic protein database, because the information on the amino acid sequences is very vast, and *N*-myristoylation is not based on one simple rule but many specific rules. We had applied the machine learning system BONSAI to predict patterns based on positive and negative examples to be classified. Figure 2 shows the mechanism of machine learning system BONSAI. Furthermore, Okada *et al.*[2] developed a CGI tool that predicts whether a given protein is *N*-myristoylated or not.

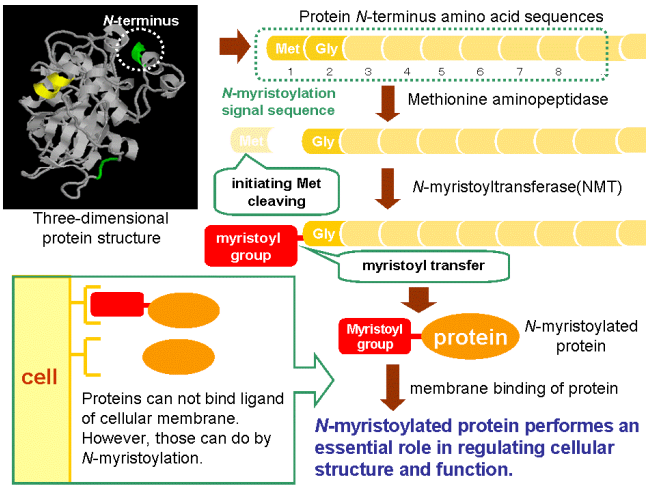

Figure 1: Reaction process of protein *N*-myristoylation.

In order to examine the performance of our CGI tool, we performed an experimentation using sample data of amino acid sequences that had been confirmed to be *N*-myristoylated or not by biochemical approach. Our CGI tool could identify *N*-myristoylated protein using predicted features of *N*-myristoylation signals and could classify sequences into positive and negative categories rapidly and correctly.

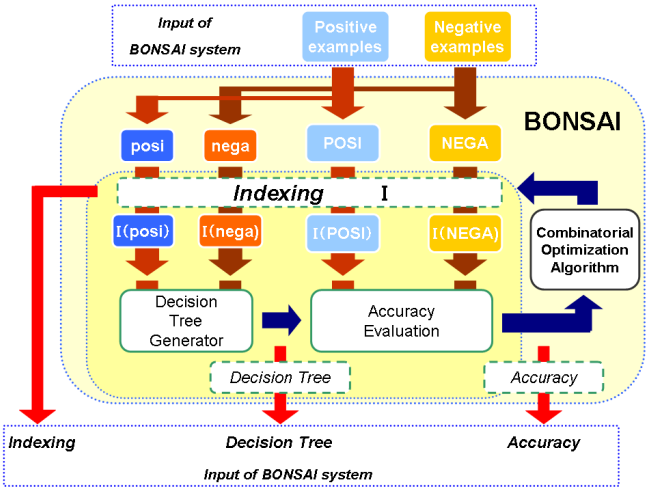

Figure 2: Machine learning system BONSAI.

A Prediction Tool with Computational Approach

Computational techniques are useful for predicting rules from a huge amount of data on the sequence required for *N*-myristoylation. The machine learning system BONSAI, which finds rules in the combination of alphabet indexings and decision trees, is applied to this CGI tool. BONSAI is a system for knowledge acquisition based on the theory of Probably Approximately Correct Learning (PAC learnability) and uses the method of local search[3]. The alphabet indexing groups letters in positive and negative example by mapping these letters to fewer numbers of letters. This CGI tool predicted *N*-myristoylated proteins with a focus on the *N*-myristoylation signal sequence important to be *N*-myristoylated. This CGI tool is composed with 4 decision trees that were provided with positive and negative examples by BONSAI and has high accuracy rate of the classification. Figure 3 shows one of 4 decision trees used our developed CGI tool. The positive examples include 78 myristoylated amino acid sequences, and the negative examples include 800 amino acid sequences randomly selected from the human protein in NCBI GenBank database[4]. The reason that we used randomly selected protein sequences as negative example is that no protein sequence being not *N*-myristoylated protein has been known. As a result, the CGI tool could identify *N*-myristoylated protein using predicted features of *N*-myristoylation signals and could classify sequences under two categories rapidly and correctly.

Figure 4 shows a system flowchart of our developed CGI tool. Users input the data of *N*-terminus amino acid sequences to our tool. And, users select one or more decision trees for the prediction. Our CGI tool removes the initial metchionine and glycine from the *N*-terminus of given sequences because *N*-myristoylated proteins surely have these two amino acids at the *N*-terminus. Subsequently, our CGI tool convert alphabets of amino acid sequences into numbers based on alphabet indexings(indexing, in short), provided at each decision tree . Alphabet indexing is the transformation of symbols to reduce the number of letters assigned to positive and negative examples without omitting important information in the original data in order to find the pattern of *N*-myristoylation signals. In the case of amino acid residues, alphabet indexing can be regarded as a classification of 20 kinds of amino acid residues to a few categories. Indexing contributes not only quicken the computations involved in finding rules but also to simplify expression patterns assigned at the nodes of decision trees.

After that, our CGI tool classifies proteins into *N*-myristoylated proteins and others according to rules of decision trees with indexed amino acid sequences. Finally, users obtain prediction results that gives the classification. Our CGI tool can predict the *N*-myristoylation without a machine learning process because we provided decision trees beforehand, which can classify the *N*-myrisoylated protein more exactly and adapted the these decision trees to our CGI tool. This contributes to reduce the time for processing the given inputs.

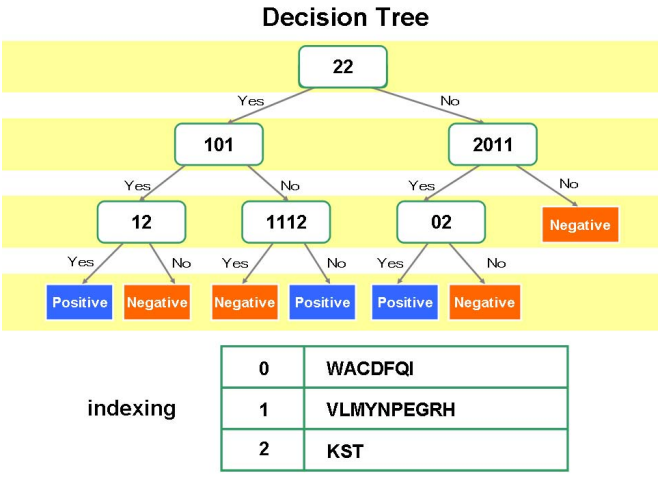

Figure 3: One of 4 decision trees used our CGI tool.

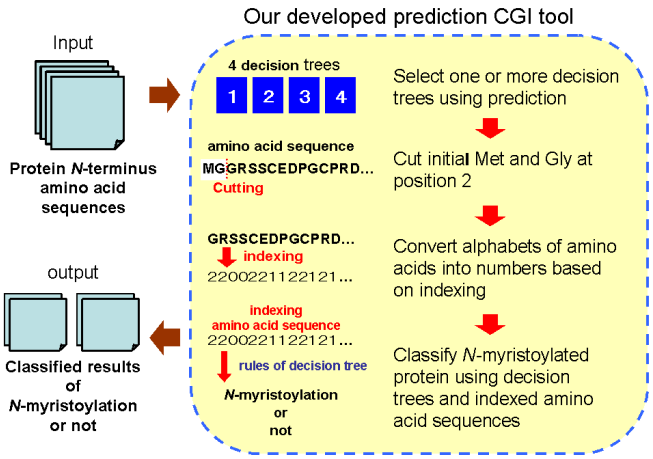

Figure 4: System flowchart of our CGI tool.

Experiment

In this study, we performed accuracy evaluation of the developed CGI tool. We used amino acid sequences that are confirmed to induce *N*-myristoylation biologically, and conducted accuracy comparison with other predicted tool predicting protein *N*-myristoylation called "NMT-The predictor"[5]. NMT-The predictor classifies proteins into *N*-myristoylated and others searching a motif of protein *N*-myristoylation with analysis of substrate sequences and kinetic data of predicted proteins. We focused on the results of two tool's predictions and the features of the pattern whose sample sequences were classified into positive and negative categories using decision trees. We used 78 protein amino acid sequences of *eukaryota* and 44 protein amino acid sequences of *A.Thaliana* that have been confirmed to be induced *N*-myristoylation biologically, and 88 protein amino acid sequences of TNF(Tumor necrosis factor) which have been confirmed not to be induced *N*-myristoylation biologically.

Additionally, our developed CGI tool and NMT-The predictor need to configure some parameters to predict. On our developed tool, users need to make one or more selections of using decision trees for the prediction. It is possible for users to configure 15 paremeters in all. On the other hand, users need to select one of 2 parameters to identify species of organisms, *eukaryote* parameter and *fungi* parameter, on NMT-The predictor. Accordingly, we try to evaluate an accuracy rate between two systems with all of 17 parameters.

We classified prediction results into 3 categories. "Induce *N*-myristoylation" is a rate of all protein sequences that were predicted to induce *N*-myristoylation. "Can not classify *N*-myristoylated or not" is a rate of all protein sequences that could not be classified a *N*-myristoylated proteins and others clearly. "Non-induce *N*-myristoylation" is a rate of all protein sequences that were predicted not to be induced *N*-myristoylation.

Result and Discussion

Our *N*-myristoylation prediction tool showed high accuracy rate of classification for *N*-myristoylation and almost the same accuracy rate of classification for most of the 210 protein sequences in this evaluation experiment (Fig.7, 8, 9).

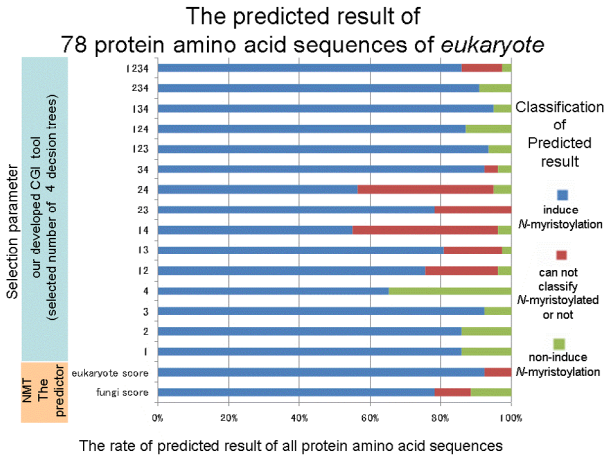

Figure 5: Prediction result of 78 myristoylated amino acid sequences of *eukaryote*.

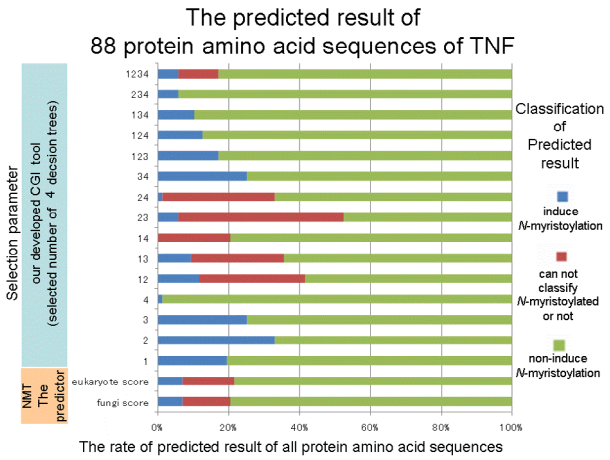

Figure 6: Prediction result of 88 not myristoylated amino acid sequences of TNF.

And, we showed more a beneficial result of our tool than that of NMT-The predictor. NMT-The predictor could not obtain high accuracy rate of classification using 44 protein amino acid sequences of *A.Thaliana*. We considered that the parameter setting of NMT-The predictor is not applied to

the prediction of protein *N*-myristoylation on plant species. On the other hand, our CGI tool provided the high accuracy rate of classification using all sequence samples (Fig.9). Our CGI tool used positive and negative examples of protein sequences of several species for the machine learning process in which the decision trees developed by Okada *et al.* are used. Therefore, our CGI tool could identify common rules for *N*-myristoylation that do not depend on species of organisms.

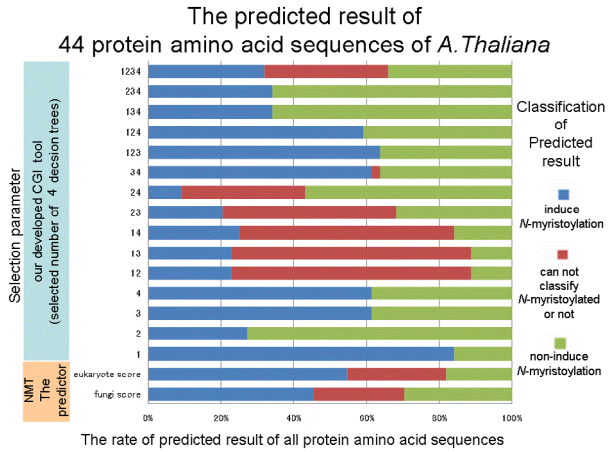

Figure 7: Prediction result of 44 myristoylated amino acid sequences of *A.Thaliana*.

In addition, we investigated what sequence was recognized by decision trees adopted our CGI tool in order to classify the protein sequence. Our CGI tool could predict *N*-myristoylation signal sequence according to features of the amino acid at position 6 from the *N*-terminus that is important position for *N*-myristoylation. Consequently, we suggest that our CGI tool can predict *N*-myristoylated proteins effectively using characteristics of amino acid sequences that are important for protein *N*-myristoylation.

We considered that NMT-The predictor have been developed base on *N*-myristoylation signal sequences as its own database, and can predict *N*-myristoylation signal sequences correctly if the database has the sequence information about species of oranisms of query. However our developed tool can predict *N*-myristoylation signal effectively without the sequence information, our developed tool has the advantage of classfying unknown protein correctly.

References

[1] Maurer-Stroh, S., Eisenhaber, B., and Eisenhaber, F., "*N*-terminal *N*-myristoylation of proteins: prediction of substrate proteins from amino acid sequence", *J. Mol. Biol.*, 317(4):541-557, 2002

[2] Sugii, M., Okada, R., Matsuno, H., Mayano, S., "Performance Improvement in Protein *N*-Myristoyl Classification by BONSAI with Insignificant Indexing Symbol", *Genome Inform.*, Vol.32, pp.277-286, 2007

[3] Shimozono, S., Shinohara, A., Shinohara, T., Miyano, S., Kuhara, S., and Arikawa, S., "Knowledge Acquisition from Amino Acid Sequences by Machine Learning System BONSAI", *Trans. Inform. Process. Soc. Japan*, Vol.35, pp.2009-2018

[4] NCBI : *ftp://ftp.ncbi.nih.gov/*

[5] NMT : *http://mendel.imp.ac.at/myristate*