Machine Learning Prediction of Amino Acid Patterns in Protein N-myristoylation

Ryo Okada[1], Manabu Sugii[2], Hiroshi Matsuno[1], and Satoru Miyano[3]

[1] Graduate School of Science and Engineering
[2] Media and Information Technology Center,
Yamaguchi University, Yamaguchi 753-8511, Japan
[3] Human Genome Center, University of Tokyo, Tokyo 108-8639, Japan
okada@ib.sci.yamaguchi-u.ac.jp, manabu@yamaguchi-u.ac.jp,
matsuno@sci.yamaguchi-u.ac.jp, miyano@ims.u-tokyo.ac.jp

Abstract. Protein N-myristoylation is the lipid modification in which the 14-carbon saturated fatty acid binds covalently to N-terminal of virus-based and eukaryotic protein. In this study, we suggest an approach to predict the pattern of N-myristoylation signal using the machine learning system BONSAI. BONSAI finds rules in combination of an alphabet indexings and decision trees. Computational experiments with BONSAI classified amino acid residues depending on effect for N-myristoylation and found rules of the alphabet indexing. In addition, BONSAI suggested new requirements for the position of an amino acid in N-myristoylation signal.

1 Introduction

Protein *N*-myristoylation is the lipid modification, and many *N*-myristoylated proteins play key roles in regulating cellular structure and function such as the BID protein concerned with an apoptosis and the alpha subunit of the G-protein localized on the cell membrane. *N*-myristoylated proteins have a specific sequence at N-terminus called *N*-myristoylation signal sequence, and this sequence is probably composed of 6 to 9 amino acids (up to 17) [1].

In order to determine the amino-terminal sequence requirements for protein *N*-myristoylation, their sequences have been examined [2,3]. Most of methods used by researchers are those that predict patterns for *N*-myristoylation by biological experimentations based on their knowledge. However, information in the sequence is very rich, involving not only a simple rule but also many specific rules. Hence, computational techniques are essentially required to predict rules from huge amount of data involving the sequence prediction for *N*-myristoylation.

The machine learning system BONSAI is a system for knowledge acquisition from primary structural data [4]. BONSAI has discovered a rule which can classify amino acid sequences into transmembrane domains and other domains over 90% accuracy [4]. BONSAI finds the rules in the combination of alphabet indexing and decision tree from positive and negative examples of sequence.

J.C. Rajapakse, L. Wong, and R. Acharya (Eds.): PRIB 2006, LNBI 4146, pp. 4–14, 2006.

The alphabet indexing groups letters in positive and negative examples by mapping these letters to fewer numbers of letters. We have tried to predict the *N*-myristoylation signal sequence from amino acid sequences using BONSAI.

Section 2 describes features of protein *N*-myristoylation with the emphasis on the sequence requirement. Section 3 gives a brief description about BONSAI used to find rules for *N*-myristoylation. In Section 4, our computational experiments using BONSAI to find rules in amino acid sequences for *N*-myristoylation are described. Suggested results from the computational experiments are presented in Section 5. This section includes two interesting rules in the requirements for *N*-myristoylation sequence, discussing about the validity of the suggested results and giving biological interpretations of them.

2 Protein N-myristoylation

Protein *N*-myristoylation is the lipid modification in which the 14-carbon saturated fatty acid binds covalently to N-terminus of virus-based and eukaryotic protein. About 0.5% of human proteins are estimated to be *N*-myristoylated [1].

Protein *N*-myristoylation is a cotranslational protein modific ation catalyzed by two enzymes, methionine aminopeptidase and *N*-myristoyltransferase (NMT). The estimated *N*-myristoylation protein has the sequence Met-Gly on its N-terminus at least. The initial Met is removed cotranslationally by the methionine

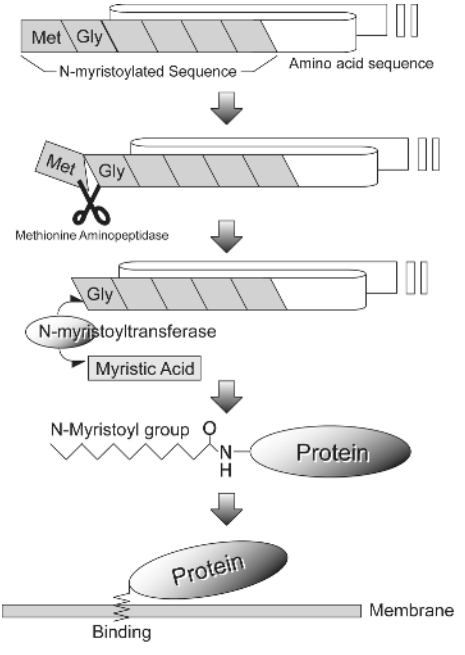

Fig. 1. Protein *N*-myristoylation

Table 1. Example of myristoylated sequence

Protein	Amino Acid Sequence
GAG SIVM1	MGARNSVLSGKKADE
KCRF STRPU	MGCAASSQQTTATGG
Q26368	MGCNTSQELKTKDGA
GBAZ HUMAN	MGCRQSSEEKEAARR
COA2 POVM3	MGAALTILVDLIEGL
RASH RRASV	MGQSLTTPLSLTLDH

Three Letter Code	Gly	Ala	Ser	Cys	Thr	Pro	Val	Asp	Asn	Leu	Ile	Gln	Glu	His	Met	Phe	Lys	Tyr	Trp	Arg
Single Letter Code	G	A	S	C	T	P	V	D	N	L	I	Q	E	H	M	F	K	Y	W	R

Fig. 2. Correspondence between amino acids in one letter and three letters

aminopeptidase, and then the myristic acid is linked to the next Gly via an amide bond by NMT. NMT catalyzes the transfer of myristic acid from myristoyl-CoA to the N-terminus Gly residue of the substrate protein (Fig. 1).

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 and process of virus growth [5,6]. For example, HIV-1 Gag protein transfer to the plasma membrane by using *N*-myristoyl group, and is involved in the formation of virus particle and emission. Additionally, it is known that the apoptosis-inducing factor Bid is digested by protease, and the new N-terminus of digested peptide is also myristoylated [7].

N-myristoylated proteins have a specific sequence at the *N*-terminus called a *N*-myristoylation signal sequence. This sequence is probably composed typically of 6 to 9, but can be as many as 17 amino acids [1]. The effect of the amino acid sequence on *N*-myristoylation depends on the distance from N-terminus; with the increase of the distance, this effect is decrease. Table 1 shows examples of N-terminus sequence of myristoylated protein. Amino acids are usually written in one letter or three letters. Fig. 2 shows the correspondence of them.

Researchers in biology have revealed that the combination of amino acid residues at positions 3 and 6 constitute a major determinant for the susceptibility to protein *N*-myristoylation. As shown in Fig. 3, when Ser is located at position 6, 11 amino acid residues (Gly, Ala, Ser, Cys, Thr, Val, Asn, Leu, Ile, Gln, His) are permitted locating at position 3 to direct efficient protein *N*-myristoylation [2,3]. Most of these 11 amino acids have a rule that the radius of gyration of residue is smaller than 1.80Å. Actually other amino acids that have radius of gyration is larger than 1.80Å, being not allowed at position 3. In addition to the restriction by the radius of gyration of the amino acid residues, it has been also revealed that the presence of negatively charged residues (Asp and Glu) and Pro residue at this position completely inhibited the *N*-myristoylation reaction.

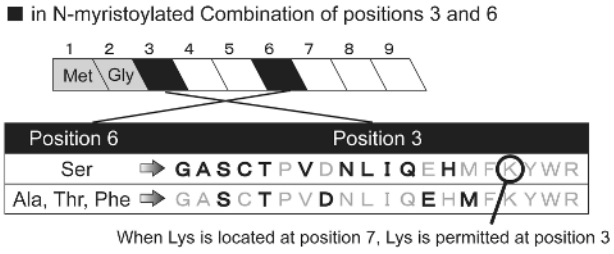

Fig. 3. Protein *N*-myristoylation rule

On the other hand, when Ala is located at position 6, 5 kinds of amino acid residues are permitted locating at position 3 for *N*-myristoylation. When Thr or Phe is located at position 6, only 2 or 3 kinds of amino acid residues are permitted locating at position 3 for *N*-myristoylation. In addition to the amino acids at position 6, there is a case that some amino acid residues at position 7 affects amino acid requirement at position 3 for *N*-myristoylation. For example, although location of Ser at position 6 does not basically allow Lys to locate at position 3, location of Lys at position 7 makes a changes to the requirement for amino acid residue at position 3; Lys can be located at position 3 [2].

3 Machine Learning System BONSAI

BONSAI is a machine learning system for knowledge acquisition from positive and negative examples of strings (Fig. 4) [4]. A hypothesis generated by the system is given as a pair of a classification of symbols called an alphabet indexing

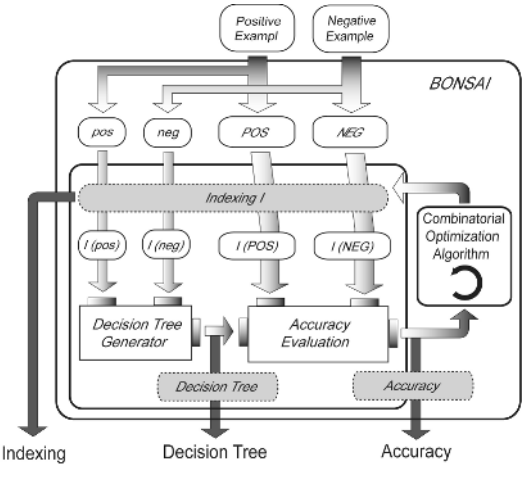

Fig. 4. Behavior of BONSAI

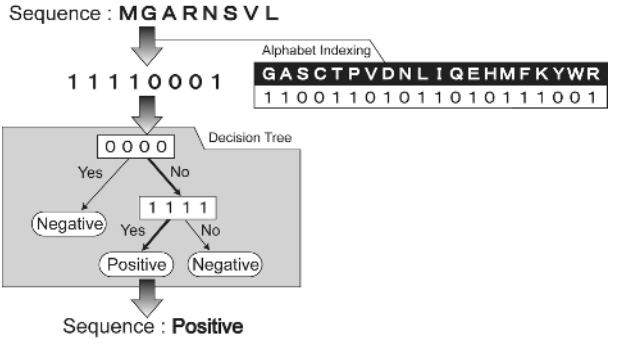

Fig. 5. Indexing

and a decision tree that classifies given examples to either positives or negatives (Fig. 5).

An alphabet indexing (indexing for short) is a transformation of symbol to reduce the size of the alphabet for positive and negative examples, without missing important information in original data. In the case of amino acid residues, the alphabet indexing can be regarded as a classification of 20 kinds of amino acid residues to a few categories. Indexing contributes not only to speed up computations in finding rules but also to simplify expression patterns assigned at nodes of decision trees.

It has been reported that BONSAI has discovered knowledge which can classify amino acid sequences of transmembrane domains and randomly chosen amino acid sequence with over 90% accuracy [4]. In the experiment, this system has found an indexing that is nearly the same as the hydropathy index of Kyte and Doolittle [8], without any knowledge on the hydropathy index.

4 Discovery of Amino Acid Patterns with Locations

We have used the following two sets of sequences as the positive and negative examples for BONSAI.

positive examples 78 sequences identified as sequences of *N*-myristoylation by the biological experiments [1] and sequences verified as *N*-myristoylation sequences presented in [6], and

negative examples sequences randomly selected from all amino acid sequences among human proteins in the NCBI database [11]. This random selection of amino acid sequences for negative examples is assured by the fact that only 0.5% of all human proteins are estimated to satisfy the requirements for *N*-myristoylation [1].

Computational experiments with BONSAI have been performed with varying the length of an amino acid sequence and the number of indexing in order to identify the proper values of them. It seems that the result is not affected by the 0.5% non-negative example in the negative examples. Because BONSAI can

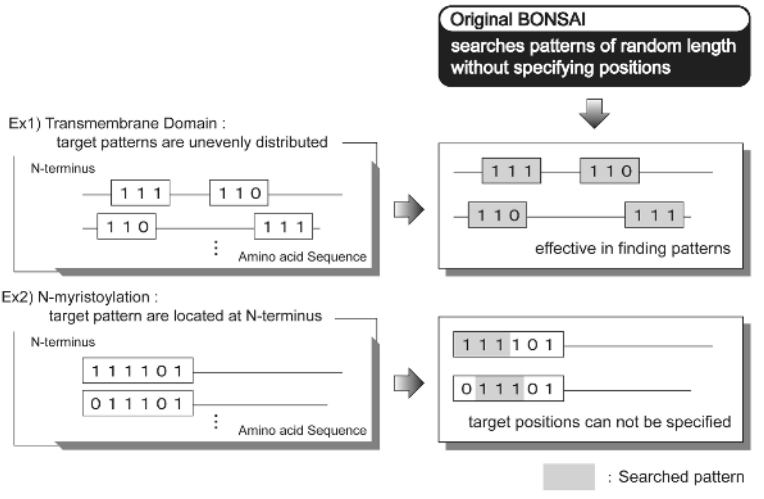

Fig. 6. Pattern search by original BONSAI

find the pattern which classifies whole given examples into either positives or negatives correctly best, even if examples contain a few exceptions. The symbol M (Met) at the N-terminus was removed from any of sequences since all the sequences of positive and negative examples have the symbol M at the N-terminus.

We modified the program of BONSAI so that BONSAI find patterns of nodes at a decision tree whose lengths are equal to the lengths of amino acid sequences inputted. Although original BONSAI finds a decision tree with indexing which can decide whether specific patterns exist in given sequences or not, it does not provide any information to identify the locations of these specific patterns. Hence, as shown in Fig. 6, the original BONSAI works well in finding transmembrane domains of amino acid sequences [4], but it can not be used to find patterns with these locations in given sequences such as patterns for *N*-myristoylation. For example, even if the original BONSAI would find a rule for the existence of successive amino acid residues Met and Gly which locate at the first and second position of the *N*-myristoylation sequence, respectively, we could not know these locations of these two amino acids by the original BONSAI.

Hence, with the modified BONSAI, we have employed the following strategy to find patterns for the *N*-myristoylation classification with amino acid locations.

1. Fix the length of sequences given to BONSAI.
2. Produce decision trees; pattern length at any node of the tree is the same as that fixed by the above procedure. We modified the program of BONSAI for this purpose.

By this strategy, we can find rules that classify sequence patterns for *N*-myristoylation with all the positions of amino acids in the patterns. Fig. 7 shows a case when the length of sequences for BONSAI is fixed to 6 and the lengths of patterns from BONSAI are restricted to the same number 6.

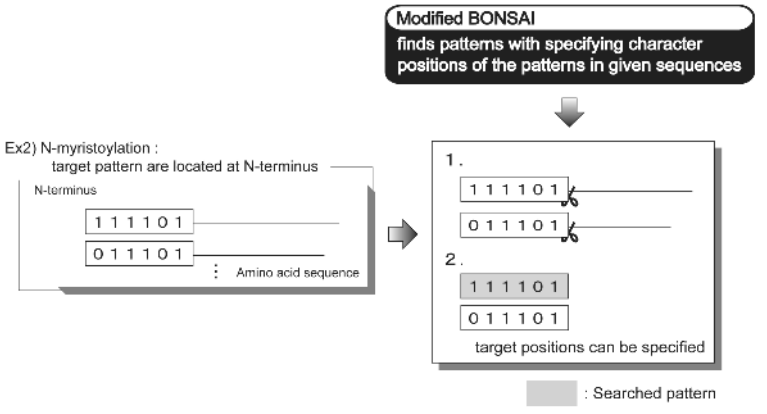

Fig. 7. Pattern search by modified BONSAI

5 Obtained Two Rules for Amino Acid Patterns in N-Myristoylation

BONSAI has presented two rules in the form of decision tree with indexing as shown in Fig. 8 and 10. Although one rule is a known fact confirmed by the biological experiment [2], the other rule suggests new amino acid sequence patterns for *N*-myristoylation.

5.1 Rule 1: Identification of Amino Acid Residue at Position 3 (Existing Rule)

Confirmed sequences of *N*-myristoylation whose *N*-myristoylation was experimentally verified in the recent report [1] and sequences presented in the literature [6] have been provided to BONSAI as positive examples. As negative examples,

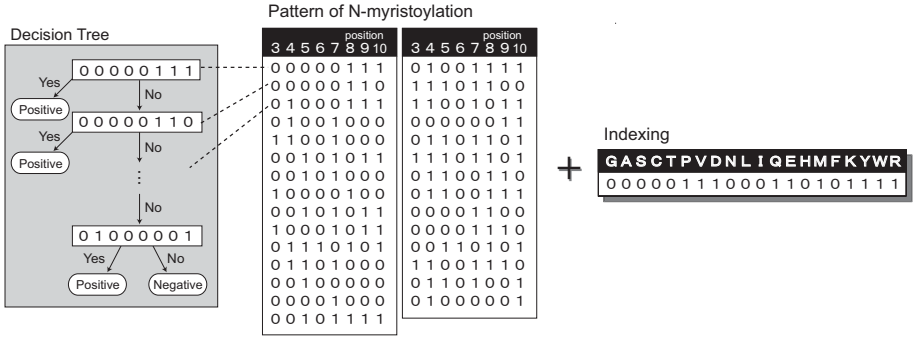

Fig. 8. Decision tree and indexing at Result1

Fig. 9. Indexing of Rule1

we used 800 human protein sequences that have been randomly selected from NCBI database [11]. This number of 800 negative examples was determined under the consideration of the tradeoff between the preciseness of produced rules from BONSAI and the processing time of BONSAI; much examples produce more precise rules, while the processing is required more. The first symbol M was removed from sequences of both of the positive and negative examples, namely all sequences had the length 9.

Fig. 8 shows a rule produced by BONSAI. The decision tree of the rule has a simple structure as shown in the figure, in which binary patterns (b-patterns for short) of the length 8 such as 00000111 is assigned to each node. These b-patterns were obtained by replacing amino acid residue symbols with each of the symbol 0 or 1 according to the indexing table in the figure. All of such 29 b-patterns are listed in the table in Fig. 8.

In the table, of 29 b-patterns of Fig. 8, we can find characteristics across two positions of them; 23 b-patterns have 0 at position 3 (79%) and 27 b-patterns have 0 at position 6 (93%). By noting that most of positive examples inputted to BONSAI has Ser at position 3 and the result of indexing that assigned the symbol 0 to Ser, we can see the reason that 93% of b-patterns at the position 6 were occupied by the symbol 0.

Fig. 9 summarizes a relationship between the amino acid pattern dependency at the position 3 on Ser at position 6 and the result of indexing from BONSAI. Eleven amino acid residues, which are biologically determined to be located at position 3 under the existence of Ser at the position 6 [2], are marked with black circles in the figure. By comparing the black circles pattern and the result of indexing, we can see that, out of these 11 amino acid residues, 9 amino acid residues (except Val and Gln) have been classified to the symbol 0. This means that BONSAI have worked well in finding requirements for *N*-myristoylation in given amino acid sequences.

Fig. 8 shows also a relationship between positions 3 and 7; if the symbol at position 3 is '1', the symbol at position 7 is '1'. This will reflect the fact that Lys can locate at position 3 under the existence of Lys at position 7, but otherwise Lys can not [2].

5.2 Rule 2: New Rules of Amino Acid Requirements Predicted by BONSAI

Confirmed 78 sequences of *N*-myristoylation have been provided to BONSAI as positive examples. As negative examples, we used 100 sequences randomly selected from NCBI database in order to avoid taking a long processing time

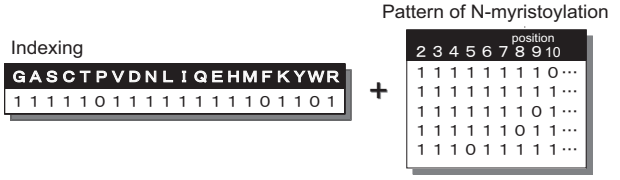

Fig. 10. Binary pattern of nodes in decision tree and indexing of Rule2

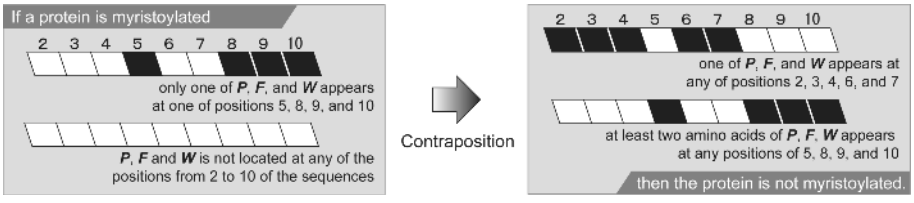

Fig. 11. Biological Interpretation of Rule2

by BONSAI. We extracted sequences of the length 20 from these positive and negative examples with removing the first symbol M from them.

With the sequences of the length 19 for these positive and negative examples, BONSAI suggested the rule as shown in Fig. 10. The decision tree is not described in the figure since it has the same structure as the one in Fig. 8. In addition, according to the biological observation that amino acid sequences up to 10 will affect *N*-myristoylation, only the parts from positions 2 to 10 of b-patterns are presented in the table. We extracted the following rule from the result of BONSAI.

- **if** a protein is *N*-myristoylated **then** the sequence of the protein satisfies the following condition;
  - only one of three amino acid residues Pro, Phe, and Try is allowed to appear at one of four positions 5, 8, 9, and 10 in the sequence, **or**
  - none of these three residues appears at any position from 2 to 10 in the sequence.

By taking the contraposition of the above rule, we can get the following (Fig. 11);

(Proposition from BONSAI)

- **if** the sequence of a protein satisfies the following condition;
  - one of these three residues appears at any of positions 2, 3, 4, 6, and 7 in the sequence, **or**
  - the sequence has more than one residue of Pro, Phe, and Try at any position of 5, 8, 9, and 10

then the protein is not *N*-myristoylated.

In the following, we will consider biological meaning of **(Proposition from BONSAI)**.

First, there has been no biological examination of amino acid requirement for positions 8, 9, and 10, and it has been biologically confirmed that amino acid residue at position 5 does not affect *N*-myristoylation [9,10]. However, the first part of "if the sequence of a protein has more than one amino acid residue of Pro, Phe, and Try at any positions of 5, 8, 9, 10, then the protein is not *N*-myristoylated" in the **(Proposition from BONSAI)** suggests the possibility that a protein which has more than one Pro at positions 5, 8, 9, and 10 will not be *N*-myristoylated. That is, Pro at position 5 of a protein may affect *N*-myristoylation of the protein, which has not been stated in any literature.

Second, the part of "if the sequence of a protein has Pro, Phe, and Try at any of positions 2, 3, 4, 6, and 7, then the protein is not *N*-myristoylated" involves the biologically confirmed fact that Pro is not allowed to locate at positions 2, 3, 6, and 7 [9,10]. For position 4, furthermore, **(Proposition from BONSAI)** suggests the new possibility that Pro, Phe, and Try can be located at position 4, while it has been considered that any of these amino acid residues can not be located at position 4.

6 Conclusion

With the increase of sequences such as amino acid sequences and base sequences produced from biological experiments, computational techniques for pattern identifications in these sequences will become more important. Using a machine learning system BONSAI, this paper examined the requirement of amino acid patterns for protein *N*-myristoylation. Suggested amino acid positions for *N*-myristoylation include not only the known positions but also positions which have not been biologically confirmed. We will proceed to the next stage to verify the new suggestion with the help of researchers in biology.

Acknowledgments. The authors thank to Prof. Toshihiko Utsumi at Yamaguchi University for insightful comments on this study. The work was partially supported by Grand-in-Aid for Scientific Research on Priority Areas "Systems Genomics" from the Ministry of Education, Culture, Sports, Science, and Technology, Japan.

References

1. Maurer-Stroh, S., Eisenhaber, B., Eisenhaber, F.: N-terminal N-myristoylation of proteins: refinement of the sequence motif and its taxon-specific differences. J. Mol. Biol. **317** (2002) 523–540
2. Utsumi, T., Nakano, K., Funakoshi, T., Kayano, Y., Nakao, S., Sakurai, N., Iwata, H., Ishisaka, R.: Vertical-scanning mutagenesis of amino acid in a model N-myristoylation motif reveals the major amino-terminal sequence requirements for protein N-myristoylation. Eur. J. Mol. Biochem. **271** (2004) 863–874

3. Utsumi, T., Sato, M., Nakano, K., Takemura, D., Iwata, H., Ishisaka, R.: Amino Acid Residue Penultimate to Amino-terminal Gly Residue Strongly Affects Two Cotranslational Protein Modifications, N-Myristoylation and N-Acetylation. J. Biol. Chem. **276** (2001) 10505–10513
4. Shimozono, S., Shinohara, A., Shinohara, T., Miyano, S., Kuhara, S., Arikawa, S.: Knowledge Acquisition from Amino Acid Sequences by Machine Learning System BONSAI. Trans. Inform. Process. Soc. Japan **35** (1994) 2009–2018
5. Farazi, T.A., Waksman, G., Gordon, J.I.: The biology and enzymology of protein N-myristoylation. J. Biol. Chem. **276** (2001) 39501–39504
6. Resh, M.D.: Fatty acylation of proteins: new insights into membrane targeting of myristoylated and palmitoylated proteins. Biochim. Biophys. Acta **1451** (1999) 1–16
7. Zha, J., Weiler, S., Oh, K.J., Wei, M.C., Korsmeyer, S.J.: Posttranslational N-myristoylation of BID as a molecular switch for targeting mitochondria and apoptosis. Science **290** (2000) 1761–1765
8. Kyte, J., Doolittle, R.F.: A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. **157** (1982) 105–132
9. Towler, D.A., Adams, S.P., Eubanks, S.R., Towery, D.S., Jackson-Machelski, E., Glaser, L., Gordon, J.I.: Purification and characterization of yeast myristoyl CoA:protein N-myristoyltransferase. Proc. Natl. Acad. Sci. USA **84** (1987) 2708–2712
10. Rocque, W.J., McWherter, C.A., Wood, D.C., Gordon, J.I.: A comparative analysis of the kinetic mechanism and peptide substrate specificity of human and Saccharomyces cerevisiae myristoyl-CoA:protein N-myristoyltransferase. J. Biol. Chem. **268** (1993) 9964–9971
11. NCBI: ftp://ftp.ncbi.nih.gov/