PERFORMANCE IMPROVEMENT IN PROTEIN *N*-MYRISTOYL CLASSIFICATION BY BONSAI WITH INSIGNIFICANT INDEXING SYMBOL

MANABU SUGII[1] RYO OKADA[2]
manabu@yamaguchi-u.ac.jp r-okada@hcu.co.jp
HIROSHI MATSUNO[3] SATORU MIYANO[4]
matsuno@sci.yamaguchi-u.ac.jp miyano@ims.u-tokyo.ac.jp

[1] *Media and Information Technology Center, Organization for Academic Information, Yamaguchi University, 1677-1 Yoshida, Yamaguchi 753-8511, Japan*
[2] *Network Solution Group, Hitachi Chugoku Solutions, Ltd., 11-10 motomachi, Hiroshima 730-0011, Japan*
[3] *Graduate School of Science and Engineering, Yamaguchi University, 1677-1 Yoshida, Yamaguchi 753-8511, Japan*
[4] *Human Genome Center, University of Tokyo, Tokyo 108-8639,Japan.*

Many *N*-myristoylated proteins play key roles in regulating cellular structure and function. In the previous study, we have applied the machine learning system BONSAI to predict patterns based on which positive and negative examples could be classified. Although BONSAI has helped establish 2 interesting rules regarding the requirements for *N*-myristoylation, the accuracy rates of these rules are not satisfactory. This paper suggests an enhancement of BONSAI by introducing an "insignificant indexing symbol" and demonstrates the efficiency of this enhancement by showing an improvement in the accuracy rates. We further examine the performance of this enhanced BONSAI by comparing the results of classification obtained the proposed method and an existing public method for the same sets of positive and negative examples.

Keywords: *N*-myristoylation; machine learning; alphabet indexing; protein classification.

1. Introduction

Protein *N*-myristoylation is a lipid modification of proteins, and many *N*-myristoylated proteins play key roles in regulating cellular structure and function such as the BH3-interacting domain death agonist (BID) which is involved in apoptosis that occurs via the alpha subunit of a G-protein localized on the cell membrane. *N*-myristoylated proteins have a specific sequence at the N-terminus called the *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 N-terminal sequence requirements for protein *N*-myristoylation, the amino acid sequences of *N*-myristoylated proteins have been examined [2, 3]. Most of the methods used by researchers predict the patterns for *N*-myristoylation based on the data obtained through biological experimen-

tations. However, the information on the amino acid sequences is very vast, and *N*-myristoylation is not based on one simple rule but many specific rules. Hence, computational techniques are essential for predicting the rules from a huge amount of data on the sequence required for *N*-myristoylation.

The machine learning system 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 [4] [5]. By using BONSAI, we carried out the computational experiment to establish new rules characterizing the difference between positive examples and negative examples of *N*-myristoylation sequences, and established the following 2 types of new rules: one, a rule that supports the existing *N*-myristoylation rule; the other, a rule that has not been discovered thus far [6]. Thus, the usefulness of BONSAI has been proved for the characterization of *N*-myristoylation sequences. However, the accuracy rates obtained using BONSAI are not sufficiently satisfactory for application in searching for new *N*-myristoylation sequences from real data. In addition, the difficulties of using our BONSAI-based method remained, specifically in terms of the complex decision trees and long processing time involved in obtaining rules.

In order to resolve with these, this paper introduces a modified BONSAI system called "BONSAI with insignificant indexing symbol." *Insignificant indexing symbol* is a special indexing symbol to which the sysytem assigns letters that do not concern with the rules of classification as either positive or negative. The results of the computational experiments show that this introduction improves the accuracy of decision trees particularly for positive examples, i.e., for sequences known to be *N*-myristoylated. In addition, this introduction allows BONSAI to generate decision trees that have smaller depth and fewer numbers of nodes than the decision trees produced by the original BONSAI [6]. We further report the results of a comparison between the proposed method and an existing public method, demonstrating that our method performs better than the existing method with respect to the accuracy of extraction of both positive and negative examples despite shorter extraction time.

2. Protein N-Myristoylation

Protein *N*-myristoylation is the lipid modification of proteins in which the 14-carbon saturated fatty acid binds covalently to the N-terminus of viral and eukaryotic proteins. Approximately 0.5% of human proteins are estimated to be *N*-myristoylated [1]. Protein *N*-myristoylation is a cotranslational protein modification catalyzed by 2 enzymes, namely, methionine aminopeptidase and *N*-myristoyltransferase (NMT). It is estimated that for undergoing *N*-myristoylation, a protein must at least have a Met-Gly sequence on its N-terminus. The initial Met is removed cotranslationally by the Met aminopeptidase, and then the myristic acid is linked to the next Gly via an amide bond through catalysis 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 are involved in physiological activities such as cell

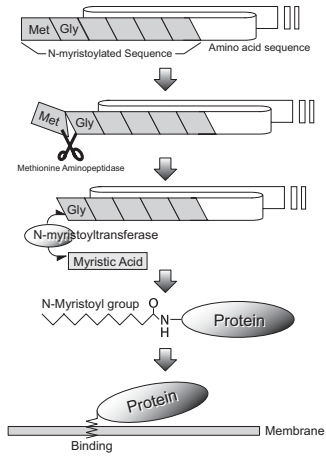

Fig. 1. Protein *N*-myristoylation is the lipid modification of proteins in which the 14-carbon saturated fatty acid binds covalently to the N-terminus. The initial Met is removed by methionine aminopeptidase. Gly is required at position 2 from the N-terminus for the formation of a bond with myristic acid through catalysis by *N*-myristoyltransferase.

signaling and exerting specific functions through binding with organelle membranes. It is known that the membrane binding mediated by myristoylation is controlled in various manners and plays a crucial role in the functional regulation mechanisms of proteins in cell signaling pathways and virus growth [7]. For example, the HIV-1 Gag protein is transferred to the plasma membrane via an *N*-myristoyl group and is involved in the formation and release of virus particles. Additionally, it is known that the apoptosis-inducing factor BID is digested by protease and that the new N-terminus of the digested peptide is also myristoylated [8].

N-myristoylated proteins have a specific sequence at the N-terminus called an *N*-myristoylation signal sequence. This sequence is probably composed typically of 6 to 9 amino acids, but this number can be as high as 17 [1]. The effect of the amino acid sequence on *N*-myristoylation depends on the distance and position from N-terminus; with the increase in the distance, this effect decreases. Table 1 shows examples of the N-terminus sequences in myristoylated proteins. Amino acids are usually denoted by 1-letter or 3-letter codes.

Biologists have revealed that the combination of amino acid residues at positions 3 and 6 constitutes a major determinant for the susceptibility to protein *N*-myristoylation. As shown in Fig. 1, when Ser is located at position 6, 11 amino acid residues (Gly, Ala, Ser, Cys, Thr, Val, Asn, Leu, Ile, Gln, His) may be located at position 3 to direct efficient protein *N*-myristoylation [2] [3]. Most of these 11 amino acids satisfy a rule that the radius of gyration of the residue is smaller than 1.80Å. In fact, other amino acids that have a radius of gyration larger than 1.80Åcannot be present at position 3. In addition to the restriction of the radius of gyration of the amino acid residues, it has been also revealed that the presence of negatively

Table 1. The sequences at the N-terminus of *N*-myristoyl proteins.

Protein	Amino acid sequence
GAG SIVM1	MGARNSVLSGKKADE
GAG MPMV	MGQELSQHERYVEQL
KCRF STRPU	MGCAASSQQTTATGG
Q26368	MGCNTSQELKTKDGA
GBAZ HUMAN	MGCRQSSEEKEAARR

charged residues (Asp and Glu) and a Pro residue at this position completely inhibited *N*-myristoylation. On the other hand, when Ala is located at position 6, 5 kinds of amino acid residues can occupy position 3 for *N*-myristoylation. When Thr or Phe is located at position 6, only 2 or 3 kinds of amino acid residues can occupy position 3 for *N*-myristoylation. In addition, some amino acid residues at position 7 dictate the amino acid requirement at position 3 for *N*-myristoylation. For example, although the presence of Ser at position 6 does not basically allow Lys to occupy position 3, the presence of Lys at position 7 alters to the requirement for amino acid residue at position 3; Lys can be present at position 3 [3].

3. BONSAI with Insignificant Indexing Symbol

BONSAI is a machine learning system for knowledge acquisition from positive and negative examples of strings (Fig. 2) [5]. A hypothesis generated by this system is presented using 2 kinds of classification of symbols called an alphabet indexing and a decision tree that classifies the given examples as either positives or negatives. *Alphabet indexing* (indexing, in short) 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 the case of amino acid residues,

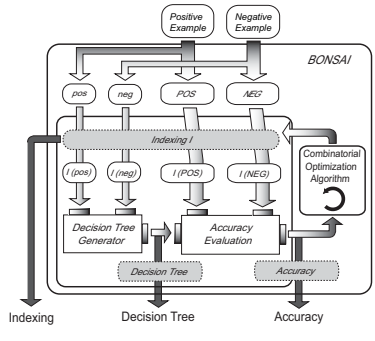

Fig. 2. For the *Positive Examples* and *Negative Examples* inputted, BONSAI computes indexings, decision trees, and accuracy. From the positive and negative examples randomly selected and transformed by indexing function I, *Decision Tree Generator* constructs decision trees. *Accuracy Evaluation* is used to search for a better indexing. With *Combinatorial Optimization Algorithm*, these are repeated until a locally optimal indexing and a decision tree are found.

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.

3.1. *Decision Tree for N-Myristoyl Sequences Generated using Original BONSAI*

Fig. 3 shows the result of BONSAI for some positive and negative examples of *N*-myristoylation. By analyzing binary patterns shown in the table in the Fig. 3, we found a rule that classifies the given positive and negative examples [6]. However, the accuracy of this rule is not high-61.1% for positive examples and 92.0% for negative examples.

A discriminative indexing pattern found by the original BONSAI is assigned at each node of the decision tree in Fig. 3. This decision tree classifies the given sequences by sequentially performing "OR operation" over the discriminative indexing patterns. This decision tree similar to a decision list has a large depth and small width, because the original BONSAI can only find such a decision tree if only poor rules exist naturally in positive and negative examples. According to the widely believed principle that "a smaller decision tree indicates essential knowledge," a tree with such a structure is not desirable.

3.2. *Introducing Insignificant Indexing Symbol*

This paper introduces a new concept of "insignificant indexing symbol" in BONSAI. *Insignificant indexing symbol* is a special indexing symbol to which the sysytem assigns letters that do not concern with the rules of classification as either positive or negative.

Insignificant indexing symbol can be realized by a simple modification to BONSAI as shown below.

(1) Choose 1 symbol from all indexing symbols as an insignificant indexing symbol, and

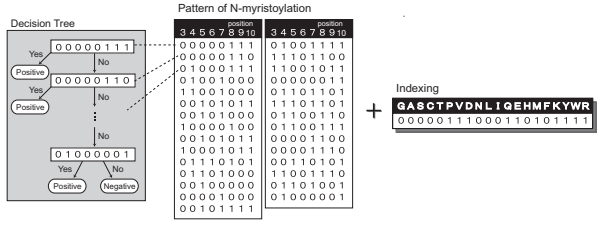

Fig. 3. Generation of decision tree and indexing for given positive and negative examples for *N*-myristoylation sequences using BONSAI [6].

(2) When evaluating a decision tree during the computation using BONSAI, the chosen insignificant indexing symbol is dealt with as "wildcard," that is, any single character can be matched at the locations of the insignificant indexing symbol.

In the following, we use sequential numbers i.e., 0, 1, 2,... for indexing symbols and assign a symbol 0 to function as the insignificant indexing symbol. Further, BONSAI$_{iis}$ (BONSAI with the insignificant indexing symbol) is also used Fig. 4 shows an example of the BONSAI$_{iis}$ process. The letters S, C, N, Q, H, M, Y, W and R are assigned as insignificant indexing symbols This implies that a more accurate decision tree can be obtained unless these letters are used in decision trees. In other words, if some of these letters are important for classifying positive and negative examples, these should be assigned to either of indexing symbol 1 or 2 in the case of Fig. 4.

4. Verification of the Effect of a Modified Algorithm for BONSAI

We have examined the performance of BONSAI$_{iis}$ with the same positive and negative examples as the experiment in section 3. The positive examples include 78 myristoylated amino acid sequences, and the negative examples include 800 amino acid sequences randomly selected from the human protein database. The indexing size was set to 3, and the length of the input sequences is 9 or 8, which excludes the first amino acid Met or the both of the first and second amino acid from the N-terminus.

Fig. 5 shows accuracy rates of classifications using 10 decision trees obtained from the original BONSAI and BONSAI$_{iis}$. We used 10 decision trees because BONSAI generally creates different trees with the same input data. The vertical axis indicates the accuracy rate of classification, where the white bar represents positive examples and the black bar, negative examples.

Fig. 5(a) clearly indicates that decision trees obtained using BONSAI can classify the example data more accurately than those obtained using the original BONSAI.

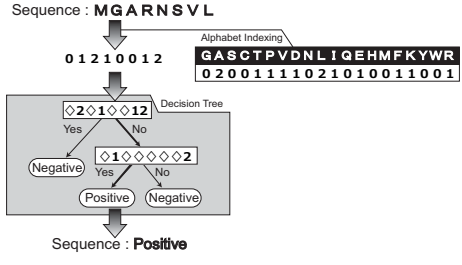

Fig. 4. Example of the BONSAI$_{iis}$ process. A given sequence MGARNSVL is converted to the sequence 01210012 according to the alphabet indexing table. The decision tree classifies this converted sequence to Positive. In order to clearly express that this symbol 0 works as a wildcard, the symbol ◇ is used instead of the previous symbol in the decision tree.

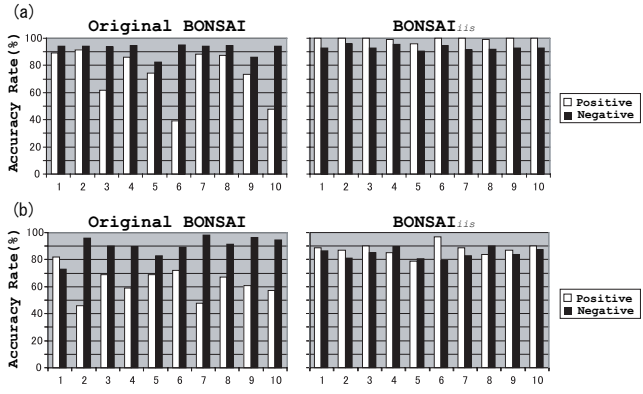

Fig. 5. Accuracy rates of classifications using 10 decision trees obtained from the original BONSAI (left) and BONSAI$_{iis}$ (right) with 9 amino acids sequences eliminating the first Met (a) and with 8 amino acids sequences eliminating the first Met and second Gly (b). The vertical axis indicates the accuracy rate of classification, the white bar indicates positive examples; the black bar, negative examples.

The accuracy rate of BONSAI$_{iis}$ is 96.3%, which is superior to 83.1% of original BONSAI. Hence, the decision trees obtained using BONSAI$_{iis}$ can more accurately provide a signal sequence required for *N*-myristoylation. A comparison of the results obtained from the original BONSAI and BONSAI$_{iis}$, shows that BONSAI$_{iis}$ shows more stable performance than the original BONSAI. Fluctuation in the accuracy rates of the original BONSAI depends on the structure of a decision tree, as shown in Fig. 3. This structure of the decision tree obtained using BONSAI$_{iis}$ also contributes to finding decision trees with higher accuracy rates.

BONSAI$_{iis}$ found a more accurate decision tree than the original BONSAI. However, BONSAI$_{iis}$ attempts to search for an *N*-myristoylation signal whether the second position of the pattern in the decision tree is occupied by Gly. This is not desirable for our research to find a new rule for N-myristoylation. Thus, we further examined the performance of the 2 BONSAI systems with the 8 amino acid sequences eliminating the first Met and second Gly to find new rules while excluding those already established for known myristoylation signals.

The result is shown in Fig. 5(b). Both the BONSAI systems show low accuracy rates compared to above experiments; this is because of the lack of the second Gly that is indispensable for *N*-myristoylation. But BONSAI$_{iis}$ has an advantage in that it provides accurate classification, with an accuracy rate of 86.1%, while the accuracy rate of the original BONSAI is 76.6%. Moreover, BONSAI$_{iis}$ maintains stable and high performance and the smaller difference in the accuracy rates between positives and negatives.

Fig. 6 shows 2 typical decision trees-one is chosen from the decision trees of the original BONSAI and the other, from those of BONSAI$_{iis}$. The average depth indicated in this figure for each of 10 decision trees obtained by the original BONSAI or

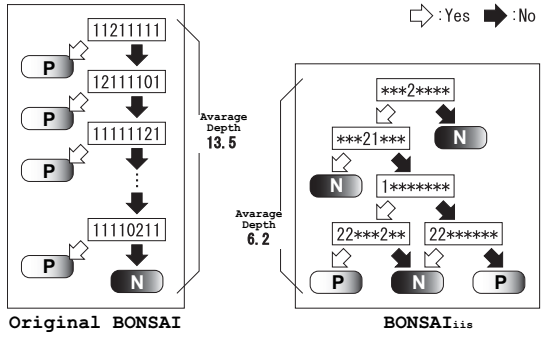

Fig. 6. Typical decision trees from the original BONSAI and BONSAI$_{iis}$. The left tree has been selected from decision trees obtained by the original BONSAI and the right one obtained by BONSAI$_{iis}$.

BONSAI$_{iis}$. We can see that the tree depth is large for a decision tree obtained by the original BONSAI, while it is small for a decision tree obtained by BONSAI$_{iis}$. At the same time, the widths of these trees are different. The decision tree obtained by BONSAI$_{iis}$ is more desirable since this decision tree is more compact than the tree from the original BONSAI. In addition, this decision tree provides a more clear representation of rules classifying positive and negative examples.

5. Comparison with Results on an Existing Website for Predicting *N*-Myristoylation

Currently, a website predicts whether a given sequence will be *N*-myristoylated or not [9]. The prediction function on this website comprises terms evaluating amino acid type preferences at sequences that are close to the N-terminus as well as terms indicate deviations from the pattern of the physical properties of amino acid side-chains encoded in a multi-residue correlation within the motif sequence [10]. The underlying biological facts for determining the scores of the prediction function are described in the paper [1]. We have compared the method used in that paper [10] with our method by using the same amino acid sequence set for both of these methods.

Table 2 shows the result of performance comparison of these 2 methods. Seventy-eight and 88 sequences were selected as positive and negative examples, respectively. Positive examples were the same sequences as those used in section 4, while negative examples were sequences presented in the literature [2, 3] as sequences that are not *N*-myristoylated sequences.

The classification results for the sequences used in our method (BONSAI$_{iis}$) are expressed as probabilities ranging from 0% to 100%. On the other hand, the classification results on the website [9] (NMT) are expressed as RELIABLE, TWILIGHT ZONE, and NO, which indicate that *N*-myristoylation of a given sequence will occur, can not be judged, and will not occur, respectively. Hence, we have derived relation-

Table 2. Performance comparison between the proposed method (BONSAI$_{iis}$) and the method used in the website [9] (NMT). Symbols used for the classification results are as follows: P=*N*-myristoylated, U=unknown, N=not *N*-myristoylated.

(a) 78 *N*-myristoylated sequences

	P	U	N	accuracy
BONSAI$_{iis}$	74	1	3	94.9%
NMT	72	6	0	92.3%

(b) 88 not *N*-myristoylated sequences

	P	U	N	accuracy
BONSAI$_{iis}$	9	6	73	83.0%
NMT	6	13	69	74.0%

ships for these 2 different expressions as follows: RELIABLE $= 55\% \leq p$, TWILIGHT ZONE $= 45\% \leq p < 55\%$, and NO $= p < 45\%$ for probability p provided by BONSAI$_{iis}$. From Table 2, we can see that

- the 2 methods express almost the same accuracy rates for positive examples (*N*-myristoylated sequences), but BONSAI$_{iis}$ expresses a higher accuracy rate than NMT for negative sequences, and
- the number of false positives in case of NMT is less than that obtained by BONSAI$_{iis}$ for both positive and negative examples.

Based on these results, we cannot emphasize that our BONSAI$_{iis}$ method is superior to the method used in NMT in terms of the accuracy rate. However, our method offers a great advantage over the NMT method with respect to computational time owing to the structure of the decision tree used in our method for classification rules. False positives in the NMT method were less in number than in BONSAI$_{iis}$, because the algorithm used in the NMT method [10] is more complex than decision trees. The number of false positives and negatives will increase as an error when only poor rules exist naturally in examples because BONSAI creates decision trees using the local optimum solution by the local search method. Thus, if BONSAI$_{iis}$ uses other system such as the database used in the NMT method, the accuracy rate of classification will be higher and the number of false positives will be reduced. BONSAI$_{iis}$ can find desirable rules in addition to reducing the process time without requiring complex algorithms.

6. Conclusion

In the previous paper [6], we modified BONSAI in order that it enables the assigning of positions of amino acids from the N-terminus. When sequences that occupy the low selective positions for amino acid were given, the modified BONSAI in [6] have produced decision trees with large depths, similar to a decision list, such as the tree

in Fig. 3, in which all conditions for classifying positive and negative examples are reflected as node labels. We reported 2 types of new rules in the previous paper [6]; however, it is difficult to interpret rules for *N*-myristoylated sequences with a decision tree having such a large depth.

Futher, taking into account the fact that *N*-myristoylated sequences have the low selective positions for amino acid, we have further modified BONSAI by introducing a new concept called an "insignificant indexing symbol." The insignificant indexing symbol will be assigned to amino acid symbols unimportant for *N*-myristoylation. This introduction allows BONSAI to distinguish letters that do not concern with the rules of classification as either positive and negative examples. We have not yet found a new rule regarding the node patterns in decision trees obtained using BONSAI$_{iis}$, although several known biological rules were confirmed.

However BONSAI is based on the theory that PAC learnability can search for the local optimum solution using local search, the local optimum solution found using BONSAI does not necessarily represent the global optimum solution. Other learning systems based on other algorithms such as support vector machine may be expected to improve the accuracy rate of classification and to find new rules for *N*-myristoylation.

References

[1] Maurer-Stroh, S., Eisenhaber, B., and Eisenhaber, F., N-terminal *N*-myristoylation of proteins: refinement of the sequence motif and its taxon-specific differences, *J. Mol. Biol.*, 317, 523-540, 2002.

[2] Utsumi, T., Sato, M., Nakano, K., Takemura, D., Iwata, H., and 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, 10505-10513, 2001.

[3] Utsumi, T., Nakano, K., Funakoshi, T., Kayano, Y., Nakao, S., Sakurai, N., Iwata, H., and 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, 863–874, 2004.

[4] Valiant, L. G., A theory of the learnable, *CACM* 27(11), 1134–1142, 1984.

[5] 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, 2009-2018, 1994.

[6] Okada, R., Sugii, M., Matsuno, H., and Miyano, S., Machine learning prediction of amino acid patterns in protein N-myristoylation, *Pattern Recognition in Bioinformatics (LNBI)*, 4146, 4-14, 2006.

[7] Farazi, T.A., Waksman, G., and Gordon, L.I., The biology and enzymology of protein *N*-myristoylation, *J. Biol. Chem.*, 276, 39501-39504, 2001.

[8] 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, 1761-1765, 2000.

[9] `http://mendel.imp.ac.at/myristate/`

[10] 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, 541-557, 2002.