Indonesian Journal of Life Sciences Vol. 01 | Number 01 | March (2019)
http://sso.i3l.ac.id/ojs/index.php/IJLS/index

REVIEW ARTICLE

Bioinformatics tools to Predict N-myristoylation Site: a Comparison Study

M. Z. Arifin[1], Arli A. Parikesit[1*]

[1] Department of Bioinformatics, Indonesia International Institute for Life Sciences, muhammad.zainul@i3l.ac.id
[1] Head of Bioinformatics Departement, Indonesia International Institute for Life Sciences,*

**Corresponding author. Email: arli.parikesit@i3l.ac.id*

ABSTRACT

Protein N-myristoylation is the covalent attachment of myrstate, via an amide bond, to the N-terminal glycine residue of a nascent polypeptide assisted by myristol-CoA protein: N-myristoltransferases (NMT). PROSITE motif describes 5 amino acid after glycine site that would give rise to myristoylation site. However, applied to whole database extract, this motif give too many false postive results. Therefore 2 new tools were developed for N-myristoylation prediction.Taking physical properties into consideration inceases prediction scores greatly. However, these algorithms still cannot predict myristoylated site correctly 100% of the time due to limited understanding of the real mechanims underlying N-myristoylation

Keywords: N-terminal myristoylation; In-silico; comparison

INTRODUCTION

Protein N-myristoylation is the covalent attachment of myrstate, via an amide bond, to the N-terminal glycine residue of a nascent polypeptide assisted by myristol-CoA protein: N-myristoltransferases (NMT) (Johnson, Bhatnagar, Knoll, & Gordon, 1994). This modification has only been observed in eukaryotes and appears to be irreversible.

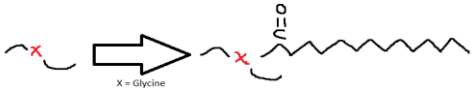

Figure 1. Myristate addition to N-terminal glycine

The first N-myrostoylated proteins were found in 1982 by Koiti Titatni et. al (Carr, Biemann, Shoji, Parmelee, & Titani, 1982) . Since then, many cellular N-myristolproteins have been identified. These proteins have diverse functions. Some examples include tyrosine kinases, serine/threonine kinases, phospoprotein phosphates, and kinase substrates. Other types of proteins involved in signal transduction cascades, and protein mediators and vesicular transport.

N-myrstolproteins have diverse intracellular destinations. Myristate is critical for protein-protein mediation and protein-membrane interactions for some N-myrostoylproteins.

The first N-myristoylation pattern was first described by PROSITE database of protein

domains. The pattern (PDOC00008) defines the amino acid sequence of five positions after the glycine site that would give rise to myristoylation (Towler, Gordon, Adams, & Glaser, 1988). The patterns are described below:

- The N-terminal residue must be glycine.
- In position 2, uncharged residues are allowed. Charged residues, proline and large hydrophobic residues are not allowed.
- In positions 3 and 4, most, if not all, residues are allowed.
- In position 5, small uncharged residues are allowed (Ala, Ser, Thr, Cys, Asn and Gly). Serine is favored.
- In position 6, proline is not allowed.

However, these pattern gives too many false positives when applied to whole database searching. Therefore, two tools were developed to reliably predict N-myristoylation in silico; the two tools are: NMT predictor* and Myristoylator**

* Available on:
http://mendel.imp.ac.at/myristate/SUPLpredictor.htm
** Available on:
https://web.expasy.org/myristoylator/

NMT Predictor

NMT predictor were developed by Maurer-Stroh et. al in 2002. They tried to predict N-myristoylation site based on the amino acid sequence (Maurer-Stroh, Eisenhaber, & Eisenhaber, 2002a). The system was based on 390 myristoylated proteins divided into three group: (1) 234 proteins were supposed to be myristoylated by similarity; (2) 56 proteins were potential candidate for myristoylation; (3) 100 proteins that were proven to be myristoylated.

They have refined the sequence motif for N-terminal N-myristoylation. Based on the in-depth study of amino acid sequence variability of substrates proteins, binding site analysis in X-ray structures analysis or 3D homology models for NMTs from various species, and biochemical data extracted from scientific literatures, they found an indication that, within a complete substrate protein, the first 17 N-terminal proteins residues experience different types of variability restrictions.

They manage to identify three motif regions: region 1 (index 1-6) fitting the binding pocket, region 2 (index 7-10) interacting with the NMT's surface at the mouth of the catalytic activity, and region 3 (positions 11-17) containing a hydrophilic linker. Each region was characterized by physical requirements to single sequence positions or groups of positions in regard to polarity, backbone flexibility, volume, and other physical properties associated with amino acids (Maurer-Stroh, Eisenhaber, & Eisenhaber, 2002b).

From these information, they created a predictor that relies on a scoring system based on sensitive profile extraction, physical properties requirements, and compensatory effects among sequence positions as well as its validation. They follow the strategy that has been successfully applied for GPI-lipid anchors prediction (Eisenhaber, Bork, & Eisenhaber, 1998). The method is also assisted with false postives probability prediction; therefore, the tool facilitates large-scale database searching.

NMT Predictor Algorithm

From their previous analysis of substrate protein sequence variability, NMT sequences and their structures has revealed that the N-terminal 17 residues are characterized by amino acid type variability restriction and match a pattern of physical properties of amino acid side-chains. If compared with the motifs described PROSITE, the increased motif length together with physical properties consideration, including compensatory effects and multi-residue correlations, promises better discrimination between substrate proteins and non-myristoylated proteins.

Based on GPI-lipid anchor experiments, both amino acid type preferences and physical pattern of amino acid side chains can be implemented into a score function to determine how likely a protein can undergo N-myristoylation.

Heavy or moderate disagreement with many physical property pattern aspects or high deviation in one feature is sufficient to exclude many non-myristoylated proteins as possible post-translational modification target. Following the establised logic, a composite score function *S* was created:

$$S = S_{profile} + S_{ppt}$$

The weighted motif region based on protein profile $S_{profile}$ evaluates each amino from position 1 to 17 type preferences. It is calculated with the PSIC algorithm, a profiling technique that is applicable for sequence sets with redundant subsets which is the case for NMT substrate learning set; $S_{profile}$ scores can be either positive or negative (Eisenhabera, Borka, Yuanc, Löfflerb, & Eisenhaberb, 2000).

The second parameter, S_{ppt}, in essence is the sum of about a dozen of terms. Each of them is penalizing the deviation from the preffered physical pattern properties in the NMT motif. All physical property term results have negative scores based on the definition. Typically, a conserved physical properties such as polarity is assumed to follow a Gauss-like (normal distribution) among true substrates The parameters for this distribution were computed from the learning set. Physical properties differences from the corresponding learning set average can be converted into scores by Gaussian functions. Therefore, if non-conformity with the physical-chemical requirements is higher, the penalty score will be much higher.

The construction of the total score allows the model to interpret the prediction output physically, such as filtering negative terms that are responsible for potential substrate rejection. Protein queries with higher score should be more favored as substrate candidates.

However, this approach oversimplified the incompletely understood recognition process and cannot always reflect the naturally occuring different affinites to the enzyme relative to prediction function scores.

For the reason above, parameterization of the prediction function withouth the application of automated numerical optimization techniques to avoid overfitting to the data in learning set. The thresholds for NMT predictor were set close to the lowest score of experimentally verfied myristoylated protein. The reliable S score is S with the score of zero or above and twilight zone predictions ($0 > S \geq -2$).

The next step is the evaluation of false positve results which will be shown later in the "comparison result" section. In the real sequence, the correct motif can occur incidentally with a certain probabilty. For this to happen, there are two reasons: (1) the protein is a good NMT substrate candidate but, due to biological context, it can never be in contact with the enzyme; (2) the scoring function describes the motif incompletely and may produce false postive result. NMT predictor algorithm was implemented in C programming language.

To justify and vaildate their function, the authors performed several tests, but they also compare their results with experimental data. The tests were: (1) self-consistency test; (2) jack-knife test of whole score S; (3) jack-knife test of S_{ppt}; (4) scores for proteins that were known to have never give rise to myristoylation (5) Correlation analysis with experimental data on NMT binding kinetics of model substrates.

Myristoylator

2 years after the release of NMT predictor, Swiss-Prot group release a new N-myristoylation site predictor, Myristoylator. The models behind their system were different from that of NMT predictor. Their models used

several distinctive features: (1) the use of several neural networks and their combination (NNS); (2) The use of qualitative properties in the inputs; the use of both positive and negative sequences; (4) score definition that is closely related to probabilities; (5) an average speed at least 9 times faster (Bologna, Yvon, Duvaud, & Veuthey, 2004).

Myristoylator Algorithm

For the dataset, the authors searched for proteins containing N-terminal glycine in Swiss-Prot. They obtained 327 proteins that were proven experimentally to have never given rise to myristoylation. The new data set was created by combining Maurer-Stroh et. al's positive dataset with negative set. In total, they obtained 717 proteins. 16 amino acids after glycine are the determinant for N-myristoylation for this model.

Machine learning models, artifical neural networks (NNs) and decision trees (DTs) for example, are able to discriminate concepts, by viewing examples repetitively. For the experiment, the usage NNs and DTs were compared to determine which model is the best for prediction.

In general, DTs are built by a recursive function splitting the input spaces by axis-parallel hyperplanes. At each step of the algorithm, a criterion to determine the best split is used, also known as "Divide and Conquer" approach. The most popular DT models is C4.5 (Quinlan, 1993).

During the training phase, the strategy of NNs and DTs are different. DT might miss combinations of several variables which are weakly predictive separately, but become strong predictive if combined. On the other hand, NN might fail to differentiate a strongly relevant variables among several irrelevant ones.

The discretized interpretable multilayer perceptron (DIMLP) network is a special multilayer perceptron for which symbolic rules are generated to explain the information embedded within the connections and the activation of neurons (Haykin, 1994).

Moreover, the computational complexity of the rule extraction algorithm scales in polynomial time with the dimensionality of the problem, the number of training examples, and the size of the network.

In the DIMLP model there are an input layer, one or more hidden layers (also known as intermediate layers), and an output layer. Fig. 2 illustrates a DIMLP network with two hidden layers. The activation functions of the neurons are the "staircase" (first intermediate layer) and the sigmoid function (second intermediate layer and output layer). Bias units represented in the left side are special neurons with constant activity; they are useful for several technical reasons. The adaptable parameters of the model are the weights denoted by symbol *w*, while symbols *x* and *h* represent input vectors and activations of intermediate neurons, respectively. Learning is achieved by determining the values of the weights which classify the training examples in the correct classes. Weights are adapted based by an optimization algorithm based on the back-propagation gradient.

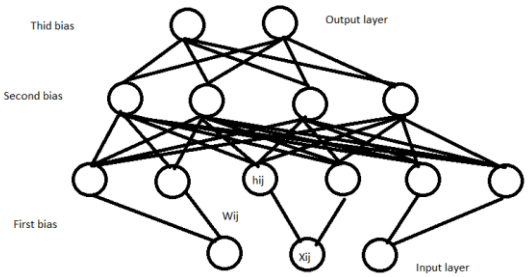

Figure2. DIMLP network with an input layer, two intermediate layers, and an output layer. The activations of the neurons are given by the first intermediate layer and the sigmoid function (second intermediate layer and output layer). Bias units represented in the left are special neurons with constant activity. The adaptable parameters of the model are the weights denoted by symbol *w*.

It was demonstrated that under several hypothesis, the output neurons of the perceptron converge towards the probability of the class given the observation (Haykin, 1994). Therefore, in this model S is defined as a score that is related to probabilities as described below.

$$S = O_{pos} + OS_{neg}$$

O_{pos} and O_{neg} are the output neurons that describe the prensence of N-myristoylated proteins and non-myristoylated proteins, respectively. Since O_{pos} and O_{neg} are probabilites, they reflect a measure of confident in the range of (-1 to +1). A score closer to +1 indicates high confidence of myristoylation. On the other hand, a score closer -1 indicates the absence of myristoylation. Finally, a score between 0 and 0.5 for both positive and negative score is the twilight zone where no decision can be made. Below are some of the rules that were applied to DIMLP. Not all of the rules were published.

Rule 1. IF (P_2 = NOT LARGE) AND (P_5 = S) THEN NMT
Rule 2. IF (P_5 = S) AND (P_6 = POSITIVE) THEN NMT
Rule 3. IF (P_2 = A) AND (P_5 = NEUTRAL) AND (P_5 = NOT LARGE THEN NMT
Rule 4. IF (P_5 = S) AND (P_{13} = NEGATIVE) THEN NMT
Rule 5. IF (P_2 = NEUTRAL) AND (P_4 = LARGE) AND (P_5 = TINY) AND (P_6 = POSITIVE) THEN NMT
Rule 6. IF (P_2 = N) AND (P_5 = TINY) THEN NMT
Rule 7. IF (P_2 = N) AND (P_7 = HYDROPHPOBIC) AND (P_{17} = NOT ALIPHATIC) THEN NMT
Rule 8. IF (P_5 = T) AND (P_{17} = NOT HYDROPHOB-IC) THEN NMT

Neural networks input vectors of amino acids were encoded by "sparse coding". Each amino acid was transformed into a vector of 20 input neurons with a particular "1" at a position and "0" at the others. Because 16 amino acids after glycine were considered, a total of 320 (16 x 20) input neurons were obtained. The second input series, the authors added the properties of each amino acid. Since there are 20 binary possibilites for each 16 amino acid, another vector of 320 inputs were obtained.

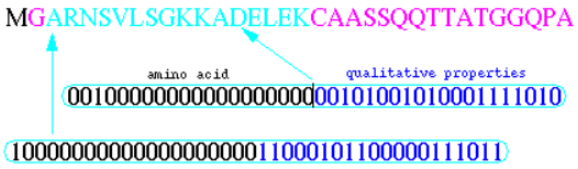

Figure3. The first series define the amino acid sequence. Each amino acid has their own sparse coding. On the other hand, the second series define the physical properties of each amino acid, e.g polarity, hydrophobicity, size, etc. There are 640 unique inputs

DTs also used the same input series. The authors define two inputs to both DT and DIMLP. The first is prediction only based on amino acid sequence (aa), whereas the second input are both the amino acid sequence and physical properties of each amino acid (aa + prop). Finally, all the models were trained with default learning parameter(Quinlan, 1993).

Comparison between PROSITE motif, NMT Predictor, and Myristoylator

Comparison study was performed by Swiss-Prot group to compare their model, for both DTs and DIMLP, with PROSITE motif and NMT predictor. A total of 717 proteins samples were used, more specifically 390 positive protein samples and 327 negative protein samples. Summary table was provided below.

Table 1. Summary table of comparison study
SENS: sensitivity; SPEC: specificity; aa: amino acid; prop: physical properties; Truepos: true positive

	SENS %	*SPEC %*	*C1 %*	*C2 %*
DIMLP- (aa)	86.7	95.4	86.6	81.7
C4.5 (aa)	89.7	94.8	84.1	83.8
DIMLP (aa + prop)	93.8	97.9	91.4	91.2
C4.5 (aa + prop.)	82.6	92.4	80.0	74.6
DIMLP (aa + prop + Truepos)	77.0	97.9	84.4	79.6

PROSITE	93.6	77.7	-	72.7
NMT predictor	95.9	97.2	-	92.9

Sensitivity is the number of true positives. Specificity is the number of true negativs. C1 and C2 are Mathews correlation with binary and non-binary response, respectively. Bolded value shows the highest percentage relative to others.

For sensitivity (true positives), NMT predictor developed by Maurer-Stroh et. al performed the best with the value of 95.9 %. For specificity, both DIMLP (aa + prop) and DIMLP (aa + prop + Truepos) performed the best with the value 0f 97.9 %. Matthew correlation coefficient is used in machine learning as a measure of the binary classifications quality; the higher the score the better (Matthews, 1975). The highest scores for both C1 and C2 are shown in DIMLP (aa + prop) with the score of 91.4 % and 91.2 %, respectively.

Based on the result, taking amino acid physical properties into consideration increases both sensitivty and specificity greatly. Interestingly, DIMLP (aa + prop + Truepos) shows the poorest sensitivity even lower than PROSITE motif. For sensitivty, PROSITE motif and DIMLP (aa + prop) shows very similar result with 93.8 and 93.6, respectively. For specificity, NMT predictor also shows similar with DIMLP (aa + prop) with value of 97.9 and 972. respectively.

Aside from this, the authors of Myristoylator also perfomed speed test online to compare myristoylator and NMT predictor. 10 randomly selected different protein sequences were used for this experiment.

Table 2. Time for predictor to run

Protein acc. No.	*Protein ID*	*NMT Pred (s)*	*Myristoylator (s)*
P49702	ARF5_CHICK	29	3
Q07085	EST2_CAEEL	29	3
Q00743	GBA1_EMENI	29	3
P19627	GBAZ_RAT	28	3
P08239	GB01_BOVIN	28	3
P26201	CD36_BOVIN	28	3
P00015	CYC2_MOUSE	28	3
P02097	HBG_MACNE	28	3
P16050	LOX1_HUMAN	28	2
P39080	PGQ_XENLA	28	3

The top 5 proteins are proteins that were positively tested for myristoylation, whereas the bottom 5 were proteins that have never been known to give rise to N-myristoylation. It turns out, Myristoylator is 9 times faster relative to NMT predictor.

Limitation for Both Models

The best way to determine whether a protein can be myristoylated is through experimental methods. However, wet-lab methods are expensive and very time consuming. Therefore, in silico prediction is highly prefered.

The two models consider not only amino acid sequence but also the physical properties. But even then, these models still cannot predict N-myrsitoylation site correctly all the time. This is because the whole algorithm is oversimplifying the natural condition for N-myristoylation to occur. several conditions for N-myristoylation might also still be unknown to us.

Lastly, the two models can only predict myristoylation that happen on N-terminal glycine. Although it is rare, myristoylation can happen in the middle of a protein sequence. They cannot detect myristoylation that happens in the middle of the sequence because they only consider the first 16 amino acid after glycine

According to Table 1, the result of the real data is similar with the simulated data, meaning that the allelic dropout event is well-estimated under the model. The model also successfully corrected the bias in heterozygosity estimation with very low standard deviation. However, it must be taken into account that the model is constructed based on the Native American dataset (i.e. the simulated data is the same with

the real data); its performance might be altered when the heterozygosity mechanism of our dataset is different with the Native Amrican one.

REFERENCE

Bologna, G., Yvon, C., Duvaud, S., & Veuthey, A. L. (2004). N-terminal myristoylation predictions by ensembles of neural networks. *Proteomics*, *4*(6), 1626–1632. https://doi.org/10.1002/pmic.200300783

Carr, S. A., Biemann, K., Shoji, S., Parmelee, D. C., & Titani, K. (1982). n-Tetradecanoyl is the NH2-terminal blocking group of the catalytic subunit of cyclic AMP-dependent protein kinase from bovine cardiac muscle. *Proceedings of the National Academy of Sciences of the United States of America*, *79*(20), 6128–6131. https://doi.org/10.1073/pnas.79.20.6128

Eisenhaber, B., Bork, P., & Eisenhaber, F. (1998). Sequence properties of GPI-anchored proteins near the omega-site: constraints for the polypeptide binding site of the putative transamidase. *Protein Engineering Design and Selection*, *11*(12), 1155–1161. https://doi.org/10.1093/protein/11.12.1155

Eisenhabera, B., Borka, P., Yuanc, Y., Löfflerb, G., & Eisenhaberb, F. (2000). Automated annotation of GPI anchor sites: Case study C. elegans. *Trends in Biochemical Sciences*. https://doi.org/10.1016/S0968-0004(00)01601-7

Haykin, S. (1994). Neural networks-A comprehensive foundation. *New York: IEEE Press. Herrmann, M., Bauer, H.-U., & Der, R*. https://doi.org/10.1017/S0269888998214044

Johnson, D. R., Bhatnagar, R. S., Knoll, L. J., & Gordon, J. I. (1994). Genetic and Biochemical Studies of Protein N-Myristoylation. *Annual Review of Biochemistry*, *63*(1), 869–914. https://doi.org/10.1146/annurev.bi.63.070194.004253

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *BBA - Protein Structure*, *405*(2), 442–451. https://doi.org/10.1016/0005-2795(75)90109-9

Maurer-Stroh, S., Eisenhaber, B., & Eisenhaber, F. (2002a). N-terminal N-myristoylation of proteins: Prediction of substrate proteins from amino acid sequence. *Journal of Molecular Biology*, *317*(4), 541–557. https://doi.org/10.1006/jmbi.2002.5426

Maurer-Stroh, S., Eisenhaber, B., & Eisenhaber, F. (2002b). N-terminal N-myristoylation of proteins: Refinement of the sequence motif and its taxon-specific differences. *Journal of Molecular Biology*, *317*(4), 523–540. https://doi.org/10.1006/jmbi.2002.5425

Quinlan, J. R. (1993). C4.5:Programs for Machine Learning. *Morgan Kaufmann ,San Mateo*, *1*. https://doi.org/10.1145/2536536.2536556

Towler, D. a, Gordon, J. I., Adams, S. P., & Glaser, L. (1988). The biology and enzymology of eukaryotic protein acylation. *Annual Review of Biochemistry*, *57*, 69–99. https://doi.org/10.1146/annurev.biochem.57.1.69