BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Prediction model

Volume 6(5)

Prediction of N-myristoylation modification of proteins by SVM

Wei Cao[1]*, Kazuya Sumikoshi[1], Shugo Nakamura[1], Tohru Terada[2], Kentaro Shimizu[1]

[1]Department of Biotechnology, Graduate School of Agricultural and Life Sciences, The University of Tokyo, 1-1-1 Yayoi, Bunkyo-ku, Tokyo 113-8657, Japan; [2]Agricultural Bioinformatics Research Unit, The University of Tokyo, 1-1-1 Yayoi, Bunkyo-ku, Tokyo 113-8657, Japan; Wei Cao - Email: davecao@bi.a.u-tokyo.ac.jp; *Corresponding author

Received May 12, 2011; Accepted May 12, 2011; Published May 26, 2011

Abstract:
Attachment of a myristoyl group to NH_2-terminus of a nascent protein among protein post-translational modification (PTM) is called myristoylation. The myristate moiety of proteins plays an important role for their biological functions, such as regulation of membrane binding (HIV-1 Gag) and enzyme activity (AMPK). Several predictors based on protein sequences alone are hitherto proposed. However, they produce a great number of false positive and false negative predictions; or they cannot be used for general purpose (i.e., taxon-specific); or threshold values of the decision rule of predictors need to be selected with cautiousness. Here, we present novel and taxon-free predictors based on protein primary structure. To identify myristoylated proteins accurately, we employ a widely used machine-learning algorithm, support vector machine (SVM). A series of SVM predictors are developed in the present study where various scales representing physicochemical and biological properties of amino acids (from the AAindex database) are used for numerical transformation of protein sequences. Of the predictors, the top ten achieve accuracies of >98% (the average value is 98.34%), and also the area under the ROC curve (AUC) values of >0.98. Compared with those of previous studies, the prediction accuracies are improved by about 3 to 4%.

Keywords: SVM; Protein; NMT; Myristoylation; PTM; Prediction

Background:
Myristoylation is an irreversible post-translational protein modification (PTM), in which a myristoyl group is derived from myristic acid and covalently attached via an amide bond to the alpha-amino group of an NH_2-terminal residue of a nascent polypeptide. This PTM occurs most commonly on glycine residues exposed during co-translational NH_2-terminal methionine removal. The usual irreversible modification is absolutely essential for the biological functions of most myristoylated proteins. A cytosolic enzyme, N-myristoyl transferase (NMT; E.C. 2.3.1.97), is responsible for the transfer of myristate (from myristoyl-CoA) onto the NH_2-terminal residue of protein substrates. Therefore, NMT is a potential chemotherapeutic target in antifungal **[1]** and anticancer drug design **[2]**. The protein substrates of NMT (reviewed by Boutin **[3]**) are so diverse, including regular enzymes, such as the protein kinase A, protein kinase G, cytochrome b5 reductase, NO synthase, most of G protein α subunits, and key enzymes in transforming processes, such as non-receptor associated tyrosine kinase (fyn, lyn, src, hck), and structural proteins (gag, UL, etc) of retroviruses and other viruses which include poliovirus, polyma virus, herpes simplex, influenza virus and so on. Attachment of the myristoyl residue provides hydrophobicity to influence the partitioning of proteins to cellular membranes; it can serve to promote protein-protein interactions **[4]**. Since myristoylation has a central role in virus maturation and oncogenesis, specific NMT inhibitors might also lead to potent antivirus agents. The experimental procedure for identifying a myristoylated protein includes cellular labeling by [^{3}H] myristate, isolation of the protein, and recognition of the released fatty acid. The experimental techniques involve reverse phase high-performance liquid chromato- graphy (RP-HPLC), gas chromatography and mass spectrometry. The ionization techniques of mass spectrometry also involve fast-atom bomdartment, chemical ionization, electrospray and tandem mass spectrometry. The sensitivity of these techniques is insufficient and they often require much more biological materials. It is, therefore, desirable to identify precisely and reliably myristoylated proteins for protein functional annotations in the proteome-wide, especially when experimental measurements are unavailable.

Studies of specificity of NMT protein substrates **[5]** led to a proposed model of the 8-residue NH_2-terminal consensus sequence (GNXXXXRR). In PROSITE **[6]**, a 6-residue sequence model, pattern PS00008 (G−{EDRKHPFYW}−X(2)−[STAGCN]−{P}), is presented as the consensus sequence of myristoylated proteins. In this motif, uncharged residues are allowed in position 2 while charged residues, proline and large hydrophobic residues are not allowed; most residues are allowed in position 3 and 4; small uncharged residues are allowed in position 5 and Serine is favored; proline is not allowed in position 6. Then, Johnson *et al*. **[4]** focused on the fifth position of the 8-residue model and proposed one which only Ser or Thr is allowed in the fifth position (GXXXS/TXXX). Later, sequence analysis of myristoylated proteins **[7]** suggests a protein motif that has three regions: positions 1-6 for fitting the binding pocket, positions 7-10 for interacting with the surface of N-myristoyltransferase at the mouth of the catalytic cavity and positions 11-17 for containing a hydrophilic linker. Many attempts have been done for developing predictors for myristoylated proteins, such as pattern search, scoring system, neural network (NN). Hitherto, the proposed predictors are based on protein sequences alone. Among them, PS00008 of PROSITE **[6]** constructed from a small dataset was available in 1990; however, it had not been updated since then, and was reported that it produces a great number of not only false positive but false negative predictions. Maurer-Stroh *et al*. **[8]** showed a taxon-specific and score-based predictor named as NMT predictor. Regarding predictions

with scores lower than the threshold, they gave a lower bound and defined them as "twilight zone" predictions. Although Boisson, Giglione and Meinnel (BGM) **[9]** attempted to modify its threshold parameter and improve identification for plant protein sequences, "twilight zone" predictions are still unsolved. Bologna *et al*. **[10]** suggested a rule-based model using average output scores generated from 25 NNs, and Podell and Gribskov **[11]** put forward a plant-specific hidden Markov model. However, the former one needs much more samples to optimize the rule set and the latter one is also taxon-specific. In the present study, we show new predictors, trained by using support vector machine (SVM), to improve the prediction of myristoylated proteins.

Methodology:

Training Datasets:

We retrieved 447 entries, each of which is clearly labeled with "N-myristoyl glycine" in the field of feature table, through online search with a keyword "myristate" against the UniProtKB database. Each sequence length of these proteins is larger than 100 residues; they cover a wide range of functions, including kinases, phosphatases, cytochrome c oxidase and so on. We noted that Maurer-Stroh's learning dataset and ours share 210 common entries. The negative dataset **[12]** consists of 429 entries, which were selected from GenBank by text-based searching. These entries contain four kinds of proteins: cytosolic proteins, secreted proteins, N-TM-C (transmembrane proteins with an NH_2-terminal export signal predicted by SignalP and a hydrophobic COOH-terminus), and transmembrane proteins. Homology similarity of sequences in the negative set was reduced to less than 50% by the employment of Smith-Waterman algorithm. Sequence-length requirement for these proteins in the negative dataset is also larger than 100 residues.

Feature vector construction:

According to the motif proposed by Maurer-Stroh *et al*. **[7]**, we used 17 residues at NH_2-terminus as input for each protein sequence. To transform these residues into a numerical format (called a feature vector), we applied a binary transform (20 binary values per residue), or the value Kyte-Doolittle (KD) scale, or their combination. Here, we exemplify the numerical transformation of a protein sequence with the feature vector named "KD+binary", which is a vector of 355 components. Of 355 components, the first 15 components are generated by applying the KD scale (window size of three residues; 15 windows in a 17-residues sequence), and the remaining 340 components (17×20) are generated from binary transformation. Also, we tried to combine KD with another scale representing physicochemical and biological properties of amino acids. As for this additional scale, we examined all the 543 scales in AAindex database **[13]**.

SVM and Cross-validation assessment:

In this study, 1-norm soft margin optimization of SVM **[14]** and Radial basic Function (RBF), as the kernel function, were employed for our task. A *K*-fold cross validation (CV) scheme was used to prevent the over-fitting problem and $K = 5$ was adopted here, i.e. the whole data set was divided at random into five equal subsets, each of them served as the test set in turn and the remaining subsets used for training. Regarding CV assessment, the most frequently used performance measure, the area under the ROC curve (AUC) that provides us a single numeric value, was also used in this study; the closer AUC is to 1.00, the better performance of the predictor is; AUC of 0.50 means the predictor that predicts the class at random.

Discussion:

Since myristoylation has a central role in virus maturation and oncogenesis, NMT inhibitors would be potent antivirus and anticancer agents. Hence, the ability to computationally recognize NMT substrate has important implications. Apart from accelerating NMT protein substrate annotation in the current database, it would be helpful for designing NMT inhibitors. The assessments of various SVM predictors are summarized in **Table 1 (see Supplementary material)**. As it's shown, the "Binary" predictor shows better performance than the "KD" or the "KD+binary" predictor; however, the remaining predictors in Table 1 improve the performance of the "Binary". Owing to addition of properties of amino acids, the prediction accuracies are all over 98% (the average value is 98.34%) and corresponding AUC values are also over 0.98, which is close to the perfect value (1.0). Regarding the top ten of 543 predictors whose features were generated by utilizing KD with other scales (the remains of them are not listed here), these ten scales representing properties of amino acid could be roughly classified into three categories: preference of protein secondary structure (AURR980117, RICJ880115, MAXF760104, MAXF760105, TANS770107 and ISOY800108), relative stability (ZHOH040102) and geometry property (FAUJ880104, FAUJ880106 and LEVM760102). The motif search by PS00008 of PROSITE, reported by Maurer-Stroh *et al*. **[8]** and Bologna *et al*. **[10]**, generates a great number of the false positive hits as applied to large database scan. Therefore, we first compared our results with those of NMT predictor developed by Maurer-Stroh *et al*. **[8]**. They showed the jack-knife tests for their learning dataset show that for the fungi dataset, the prediction accuracy is 95.9% (353/368) and is 95.5% (21/22) for the eukaryotic and viral sequences. Comparing with these results, the average prediction accuracy (98.34%) produced by our SVM predictors is higher and improved by 3%. The NMT predictor is based on a statistical scoring scheme and is taxon-specific, i.e., one has to choose the parameter set according to the taxonomy of the query protein sequences. There are two taxon-specific fields: one is fungi, the other is eukaryota and their viruses (non-fungal). In contrast, our SVM predictors are taxon-free and have few parameters. Further, the classification rule of the NMT predictor generated is based on positive sequences. For screening of unknown proteins, however, it is difficult to balance between false positive and false negative errors.

Second, we also compared our predictors with the NN predictor **[10]**. In the NN predictor, Bologna et al. used 16 amino acids after the NH_2-terminal glycine. NN input vectors of amino acids were encoded by sparse coding, i.e., binary values. Each amino acid was represented by 40 binary values. The first 20 binary values stand for amino acid types and the last 20 binary values stand for properties of each amino acid which are based on taxonomy of 20 amino acids, such as aromatic, aliphatic, neutral, charge, and so on. The prediction accuracy produced by the NN method is 93.8%. Compared with those in **Table 1** (except for "KD"), the prediction accuracy is improved by 4.12% (97.92%) for "Binary", 3.53% (97.15%) for "KD+binary" and 4.54% (average 98.34%) for the remaining predictors. Compared with predictors proposed in the previous studies, the current predictors trained in this work show higher performance. We must note that since there is a subtle difference among the conditions used in the different methods, it could not make us do a complete fair comparison.

Conclusion:

In this study, we presented new predictors based on the SVM for myristoylated protein prediction. The ten SVM predictors, whose features were generated by applying the KD scale and other scales, showed average accuracies of 98.34%. Roughly compared with those of previous studies, the prediction accuracies were improved by about 3 to 4%. Achievement of the higher prediction accuracies of our SVM predictors may imply that, except for hydrophobicity of amino acids represented by the KD scale, preference of protein secondary structure, relative stability and geometry property represented by the other ten scales might also be important to myristolation. Our SVM predictors have remarkable generalization ability by virtue of support vector machine algorithm, and we believe that our new predictors will be helpful to identify myristolated proteins on the proteomic scale. A crystal structure (PDB code: 1IID) shows that protein substrate in active site prefers to α helix and coil rather than β sheet since β sheet needs more space. Regarding myristoylation of proteins, the further investigations of the secondary structure and space requirement also need be done in future.

References:

[1] Cardenas ME *et al. Clin Microbiol Rev.* 1999 **12**: 583 [PMID: 10515904]
[2] Farazi TA *et al. J Biol Chem.* 2001 **276**: 39501 [PMID: 11527981]
[3] Boutin JA. *Cell Signal.* 1997 **9**: 15 [PMID: 9067626]
[4] Johnson DR *et al. Annu Rev Biochem.* 1994 **63**: 869 [PMID: 7979256]
[5] Towler DA *et al. Annu Rev Biochem.* 1988 **57**: 69 [PMID: 3052287]
[6] Hulo N *et al. Nucleic Acids Res.* 2006 **34**: D227 [PMID: 16381852]
[7] Maurer-Stroh S *et al. J Mol Biol.* 2002 **317**: 523 [PMID: 11955007]
[8] Maurer-Stroh S *et al. J Mol Biol.* 2002 **317**: 541 [PMID: 11955008]
[9] Boisson B *et al. J Biol Chem.* 2003 **278**: 43418 [PMID: 12912986]
[10] Bologna G *et al. Proteomics.* 2004 **4**: 1626 [PMID: 15174132]
[11] Podell S & Gribskov M. *BMC Genomics.* 2004 **5**: 37 [PMID: 15202951]
[12] Fankhauser N & Maser P. *Bioinformatics.* 2005 **21**: 1846 [PMID: 15691858]
[13] Kawashima S & Kanehisa M. *Nucleic Acids Res.* 2000 **28**: 374 [PMID: 10592278]
[14] http://www.csie.ntu.edu.tw/~cjlin/libsvm

Edited by P Kangueane

Citation: Cao *et al.* Bioinformation 6(5): 204-206 (2011)

BIOINFORMATION

open access

Supplementary material:

Table 1: Performance assessment of myristoylated protein prediction by SVM

Name of SVM predictors[a]	Win[b]	Total length[c]	ACC(%)[d]	AUC[e]
Binary	1	17	97.92	0.983
KD	3	17	90.16	0.919
KD + binary	(3,1)	17	97.15	0.977
KD + TANS770107	(3,3)	17	98.64	0.989
KD + RICJ880115	(3,3)	17	98.50	0.988
KD + FAUJ880104	(3,3)	17	98.48	0.988
KD + MAXF760105	(3,3)	17	98.41	0.987
KD + LEVM760102	(3,3)	17	98.37	0.987
KD + FAUJ880106	(3,3)	17	98.36	0.986
KD + MAXF760104	(3,3)	17	98.35	0.987
KD + ISOY800108	(3,3)	17	98.22	0.986
KD + ZHOH040102	(3,3)	17	98.05	0.984
KD + AURR980117	(3,3)	17	98.02	0.984

[a]SVM predictors: named after the types of components of the input feature vectors. The top ten of 543 predictors whose features were generated by utilizing KD with other scales are shown here. KD: Kyte-Doolittle scale; TANS770107, normalized frequency of left-handed helix; RICJ880115, relative preference value at C-cap; FAUJ880104, length of the side chain; MAXF760105, normalized frequency of zeta L; LEVM760102, distance between Cα and centroid of side chain; FAUJ880106, mini- mum width of the side chain; MAXF760104, normalized frequency of left-handed α-helix; ISOY800108, normalized relative frequency of coil; ZHOH040102, the relative stability scale extracted from mutation experiments; AURR980117, normalized positional residue frequency at helix termini C′; [b]Win: the size of sliding window; [c] Total length: the total number of residues at NH_2-terminus for input; [d]ACC: average prediction accuracy of 100 runs of 5-fold CV test; [e]AUC: the value of the area under the ROC curve derived from 100 runs of 5-fold CV test.