Nucleic Acids Research, 2006, Vol. 34, Web Server issue ***W249–W253***
doi:10.1093/nar/gkl233

MeMo: a web tool for prediction of protein methylation modifications

Hu Chen, Yu Xue[1], Ni Huang, Xuebiao Yao[1,*] and Zhirong Sun*

Institute of Bioinformatics and Systems Biology, MOE Key Laboratory of Bioinformatics, State Key Laboratory of Biomembrane and Membrane Biotechnology, Department of Biological Sciences and Biotechnology, Tsinghua University, Beijing, China 100084 and [1]Laboratory of Cellular Dynamics, Hefei National Laboratory for Physical Sciences, and the University of Science and Technology of China, Hefei, China 230027

Received January 19, 2006; Revised February 9, 2006; Accepted March 28, 2006

ABSTRACT

Protein methylation is an important and reversible post-translational modification of proteins (PTMs), which governs cellular dynamics and plasticity. Experimental identification of the methylation site is labor-intensive and often limited by the availability of reagents, such as methyl-specific antibodies and optimization of enzymatic reaction. Computational analysis may facilitate the identification of potential methylation sites with ease and provide insight for further experimentation. Here we present a novel protein methylation prediction web server named MeMo, protein methylation modification prediction, implemented in Support Vector Machines (SVMs). Our present analysis is primarily focused on methylation on lysine and arginine, two major protein methylation sites. However, our computational platform can be easily extended into the analyses of other amino acids. The accuracies for prediction of protein methylation on lysine and arginine have reached 67.1 and 86.7%, respectively. Thus, the MeMo system is a novel tool for predicting protein methylation and may prove useful in the study of protein methylation function and dynamics. The MeMo web server is available at: http://www.bioinfo.tsinghua.edu.cn/~tigerchen/memo.html.

INTRODUCTION

In the post-genomic era, much attention has been paid to understanding the dynamics of the proteome, transcriptional regulation and post-translational modification of proteins (PTMs). Numerous PTMs supply the proteome with structural and functional diversity, and govern cellular plasticity and dynamics. Types of PTMs include phosphorylation (1,2), sumoylation (3), ubiquitination and methylation. Compared to well-known and extensively studied protein phosphorylation (1,2), protein methylation attracts much less attention, despite the fact that it was discovered nearly half a century ago (4). Protein methylation can modify the nitrogen atoms of either the backbone or side-chain (N-methylation) in several types of amino acids, such as lysine, arginine, histidine, alainine and asparagine, etc (5–11). Also, methylation occurs at cysteine residues as S-methylation (12). In this field, the predominant studies have focused on modifications of lysine and arginine residues.

Lysine residues can be mono-, di- or tri-methylated by histone lysine methyltransferases (HKMTs) (5,8,10,11). The methylation of lysine has been mostly studied in H3 and H4 histone proteins, which play essential roles in many biological processes, such as heterochromatin compaction, X-chromosome inactivation and transcriptional silencing or activation (5,10,11). Furthermore, the HKMTs also modify a variety of non-histone proteins with diverse functions (5,8,10,11). For example, Set9 methylates a transcription factor TAF10 to increase its interacting affinity for RNA polymerase II, which is implicated in transcriptional regulation of TAF10 target genes (13). In addition, methylation of p53 by Set9 *in vivo* increases its stability and regulates the expression of p53-dependent genes (14). Furthermore, the activity of lysine methylation of cytosolic Ezh2-containing methyltransferase complex is essential for receptor-induced actin organization and proliferation (15). Thus lysine methylation may also function in signaling processes.

Protein methylation can also occur on the guanidino nitrogen atoms of arginine (6,7,9,10,16). Although arginine

*To whom correspondence should be addressed. Tel: +86 551 3606294; Fax: +86 551 3607141; Email: yaoxb@ustc.edu.cn
*Correspondence may also be addressed to Zhirong Sun. Tel: +86 10 62772237; Fax: +86 10 62772237; Email: sunzhr@mail.tsinghua.edu.cn

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

methylation can also modify the core histones and form 'histone code' together with lysine methylation (5,11), the substrates of PRMTs are much more diverse than HKMTs (6,7,10). Thus protein arginine methylation may be involved in more functional processes. Indeed, arginine methylation plays important roles in numerous cellular processes, including RNA processing, transcriptional regulation, signal transduction and DNA repair (6,7,10). For example, arginine methylation of SPT5 regulates its binding with RNA polymerase II to modulate the transcriptional elongation (17). And PMRT1 methylates NIP45, the nuclear factor of activated T cell (NFAT) cofactor protein, to play an essential role in cytokine gene transcription (18). In addition, as a potential role 'arginine protection', PRMTs may modify and protect the arginines against endogenous reactive methylglyoxal (9).

Protein methylation is a reversible type of PTM, just like phosphorylation and sumoylation. Recent study shows that LSD1 (lysine-specific demethylase 1) is responsible for the demethylation of histone H3 lysine 4 (5). Very recently, it has been verified that JHDM1 (JmjC domain-containing histone demethylase 1) is responsible for the demethylation of lysine 36 (19). Furthermore, peptidyl-arginine deiminase PAD4 is able to deiminate both unmodified arginine and monomethylarginine residues in histones into citrullines (6).

The full extent of regulatory roles of protein methylation is still elusive. Importantly, identification of methylated proteins with their sites will be a foundation of understanding the molecular mechanism of protein methylation. Besides the conventional experimental methods, such as mutagenesis of potential methylated residues, methylation-specific antibodies (20) and mass-spectrometry (21–23) have also been deployed. However, these experimental approaches are laborious and expensive. Therefore computational prediction of methylation sites is much more desirable for its convenience and fast speed. Unfortunately, although many methods with satisfying accuracies have been developed to predict phosphorylated protein sites (2,24), only one work, which focuses on only disordered regions of considered proteins, has been published on prediction of methylation sites (25).

In this work, we provide a novel online tool for protein methylation site prediction of MeMo, protein Methylation Modification prediction, employing the algorithm of Support Vector Machines (SVMs) (26) (http://www.csie.ntu.edu.tw/~cjlin/libsvm). We have collected all annotated methylated residues in SWISS-PROT version 48 (27). We have also manually mined literatures to extract experimentally verified methylated residues. Then we have combined the two datasets into an integrated positive (+) dataset for training after homology reducing. Limited by the available data, only methylated lysines and arginines have been considered. The suitable parameters have been used to train SVM models (26). A 7-fold cross validations have been performed to check the models' accuracies and generality. The accuracies of MeMo are 67.1% on lysines and 86.7% on arginines. The prediction results of MeMo remain to be experimentally verified. For convenience, we have implemented the prediction system in a web server, which is now available at: http://www.bioinfo.tsinghua.edu.cn/~tigerchen/memo.html.

IMPLEMENTATION

First, we obtained the dataset of methylation sites from the feature table of SWISS-PROT (version 48) (27). Only experimentally verified methylated sites were preserved. Potential methylated sites with keywords of 'By similarity', 'Potential' or 'Probable' in SWISS-PROT's comments were removed. In total we obtained 328 positive (+) sites, including lysines (148 items), arginines (76 items), histidines, asparagines and other kinds of residues (Supplementary Table 1). Then we searched the PubMed with the keywords of 'methylation lysine' and 'methylation arginine' for information on lysine and arginine methylation, respectively. From ~1700 scientific articles, we collected 107 and 264 experimentally verified methylation sites for lysine and arginine, respectively. In addition, we combined the manually curated data and the data from SWISS-PROT into an integrated positive (+) dataset. As a result, only lysines (227 items) and arginines (273 items) had enough data entries to train and test the SVM models. In the present work we focused on methylated lysine and arginine residues, and did not take other amino acids into consideration.

The positive (+) dataset for training might contain some homologous sites from homologous proteins. If the testing data were highly homologous to the training data, the prediction accuracy would be overestimated. To avoid the overestimation, we clustered the protein sequences from positive(+) dataset with a threshold of 30% identity by BLASTCLUST, one program in the BLAST package (28). If two proteins were similar with $\geq$30% identity, we re-aligned the proteins with BL2SEQ, another program in the BLAST package (28), and checked the results manually. If two methylation sites from the two proteins were at the same position in alignment, only one item was kept while the other one was discarded. Thus, we obtained positive (+) data of high quality with 156 lysine and 250 arginine-methylated sites. As described previously (2,24), the negative (−) sites were taken from non-annotated lysine/arginine sites in the same proteins from which (+) sites were chosen. The homology reducing process was also carried out on (−) data.

LIBSVM (26) was employed to build SVM models. Suitable parameters were deployed to train SVM models. Finally, MeMo was implemented in PERL and hosted by Apache running on a Debian Linux system. A screenshot of MeMo is shown in Figure 1. More details about the algorithm and implementation are described in Supplementary Data. The web server of MeMo (version beta 1.0) has been available since Jan, 2005. The current version of MeMo is 2.0.

USAGE

To mimic queries from biologist users, we have randomly submitted to MeMo three proteins, PBX4 (Q9BYU1), Syntaxin 10 (O60499) and Sorting nexin (SNX)-17 (Q15036) from a large-scale experiment to identify the potential methylated proteins (20), as examples to demonstrate the simplicity and accuracy of MeMo (Table 1). With methyl-specific antibodies, there have been ~200 putatively arginine-methylated proteins to be identified (20). However, the precise arginine methylation sites on these substrates still remain to be verified. From the list we have blindly taken three proteins, which have

not been included in our training dataset, and predicted the potential arginine methylation sites in these proteins. Transcription factor PBX4 is a member of the Pbx family that is implicated in a variety of developmental processes including axial patterning, hindbrain development and organogenesis (29,30). Syntaxin 10 is a SNARE (soluble N-ethylmaleimide sensitive factor attachment protein receptor) protein. As a member of the syntaxin family, syntaxin 10 localizes to the trans-Golgi network (TGN) and play a potential role in regulating the expression profile of transferrin receptor (TfR) (31). SNX17 is a member of the SNX family that is involved in the sorting of transmembrane proteins (32). SNX17 associates with LDL receptor-related protein (LRP) to modulate its cell surface levels (33). In all three cases, the regulatory roles of arginine methylation remain to be elucidated. We have employed the MeMo web tool to predict the arginine methylation sites on these proteins.

We have retrieved the protein sequences of the three proteins in FASTA format, and pasted them in the INPUT form of MeMo. The prediction result is diagramed in Supplementary Figure 2. MeMo predicts R55, R57 and R63 of PBX4, R11 and R22 of Syntaxin 10 and R339, R341, R399 and R442 of Sorting nexin-17 as potentially positive hits (Table 2).

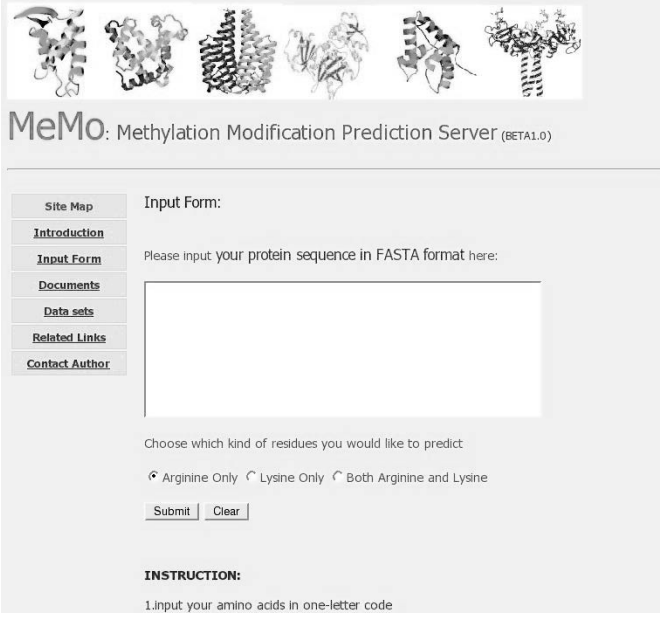

Figure 1. The screenshot of MeMo, Methylation Modification Prediction Server.

Interestingly, six of nine predicted sites are consistent with the methyl-specific antibody epitopes (20). The examples used here also shows that MeMo could be an important and useful computational tool for further experimental work. The prediction results still remain to be experimentally verified.

DISCUSSION

MeMo reaches the accuracies of 67.1% on lysines and 86.7% on arginines. Our approach can be comparable with previous method (25). More details about performance comparison are shown and discussed in Supplementary Data (Supplementary Table 3).

Our results point to several paths for future research. Firstly, the prediction systems are greatly hampered by lack of data. The known methylated protein residues (Supplementary Table 1) are still far fewer than known phosphorylated residues (1,2). In addition, the accuracies may also be affected by a lack of training data. However, as proteomic techniques improve, more and more methylation sites will be identified. We can expect that the prediction systems could be expanded to other kinds of methylated residues besides arginine and lysine. The accuracies will also improve with more training data. In addition, a powerful predictor of methylation sites in a methyltransferase family-specific fashion is also desirable. Moreover, some other machine learning methods could be applied, i.e. Group-based Prediction and Scoring algorithm (GPS) (2), artificial neural networks and hidden Markov models. These approaches could be used separately or combined together to build potentially better models. Furthermore, evolutionary information, e.g. phylogenetic conservation between human and mouse (3), could be integrated into the prediction system to improve its accuracy. Finally, the sequence patterns and structural specificities, which facilitate the binding between methylation sites and methyltransferases, remain to be dissected.

CONCLUSIONS

Here we have developed a high-performance protein methylation predictor using the SVMs. Due to the data limitation our system focuses on methylated arginine and lysine sites. The accuracies for lysine and arginine methylation reach 67.1 and 86.7%, respectively. We have implemented our method in an accessible web server, and expect that the Memo may serve as a useful tool to experimentalists, who study protein methylation function and dynamics.

Table 1. The best parameters and accuracy measurements

MeMo	Parameter used					Accuracy of MeMo			
	Window length	Kernel type	Degree[a]	Gamma[b]	C[c]	Accuracy	Sensitivity	Specificity	MCC
Lysine	14	Polynomial[d]	3	0.004	10	67.10%	69.20%	66.70%	0.29
Arginine	14	RBF[e]	—	0.001	10	86.70%	69.60%	89.20%	0.54

[a]Degree: degree in polynomial kernel function.
[b]Gamma: gamma value in the kernel functions.
[c]C: trade-off between training error and margin.
[d]Polynomial: $(\text{gamma}*u'*v + \text{coef0})^{\text{degree}}$.
[e]RBF: radial basis function, $\exp(-\text{gamma}*|u-v|^2)$.

Table 2. The predicted arginine methylation sites of PBX4 (Q9BYU1), Syntaxin 10 (O60499) and Sorting nexin-17 (Q15036) with the methyl-specific antibody epitopes in these protein sequences

Methylated proteins	Methyl-specific antibody epitope	Predicted arginine methylation site
PBX4	52-PEKRGRGG-59	55, 57, 63
Syntaxin 10	11-RGEVQKAVNTARGLYQRWCE-30	11, 22
Sorting nexin-17	331-SGSTSSPGRGRGEVRLELAF-350	339, 341, 399, 442

MeMo predicts R55, R57 and R63 of PBX4, R11 and R22 of Syntaxin 10 and R339, R341, R399 and R442 of Sorting nexin-17 as potentially positive hits.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Andrew Shaw and Gordon Leitch for critical reading on the manuscript. The authors are grateful to Dr Chih-Chung Chang and Dr Chih-Jen Lin for providing the LIBSVM software. Special thanks go to the two anonymous reviewers, whose suggestions greatly improved the presentations of our manuscript. This work is supported by National Nature Science Grant (90408019, 90303017), Chinese 863 projects (2002AA234041, 2002AA231031) and Chinese 973 projects (2003CB715900) to Z.S., and Chinese Natural Science Foundation (39925018, 30270654, 30270293 and 90508002), Chinese Academy of Science (KSCX2-2-01), Chinese 973 project (2002CB713700), Chinese 863 project (2001AA215331), and Chinese Minister of Education (20020358051) to X.Y. X.Y. is a Cheung Kong Scholar. Funding to pay the Open Access publication charges for this article was provided by Chinese 863 projects (2002AA234041, 2002AA231031) to Z.S.

Conflict of interest statement. None declared.

REFERENCES

1. Beausoleil,S.A., Jedrychowski,M., Schwartz,D., Elias,J.E., Villen,J., Li,J., Cohn,M.A., Cantley,L.C. and Gygi,S.P. (2004) Large-scale characterization of HeLa cell nuclear phosphoproteins. *Proc. Natl Acad. Sci. USA*, **101**, 12130–12135.
2. Xue,Y., Zhou,F., Zhu,M., Ahmed,K., Chen,G. and Yao,X. (2005) GPS: a comprehensive www server for phosphorylation sites prediction. *Nucleic Acids Res.*, **33**, W184–W187.
3. Zhou,F., Xue,Y., Lu,H., Chen,G. and Yao,X. (2005) A genome-wide analysis of sumoylation-related biological processes and functions in human nucleus. *FEBS Lett.*, **579**, 3369–3375.
4. Ambler,R.P. and Rees,M.W. (1959) Epsilon-N-Methyl-lysine in bacterial flagellar protein. *Nature*, **184**, 56–57.
5. Bannister,A.J. and Kouzarides,T. (2005) Reversing histone methylation. *Nature*, **436**, 1103–1106.
6. Bedford,M.T. and Richard,S. (2005) Arginine methylation an emerging regulator of protein function. *Mol. Cell*, **18**, 263–272.
7. Boisvert,F.M., Chenard,C.A. and Richard,S. (2005) Protein interfaces in signaling regulated by arginine methylation. *Sci. STKE, re2*.
8. Cheng,X., Collins,R.E. and Zhang,X. (2005) Structural and sequence motifs of protein (histone) methylation enzymes. *Annu. Rev. Biophys. Biomol. Struct.*, **34**, 267–294.
9. Fackelmayer,F.O. (2005) Protein arginine methyltransferases: guardians of the Arg? *Trends Biochem. Sci.*, **30**, 666–671.
10. Lee,D.Y., Teyssier,C., Strahl,B.D. and Stallcup,M.R. (2005) Role of protein methylation in regulation of transcription. *Endocr. Rev.*, **26**, 147–170.
11. Martin,C. and Zhang,Y. (2005) The diverse functions of histone lysine methylation. *Nature Rev. Mol. Cell. Biol.*, **6**, 838–849.
12. Lapko,V.N., Cerny,R.L., Smith,D.L. and Smith,J.B. (2005) Modifications of human betaA1/betaA3-crystallins include S-methylation, glutathiolation, and truncation. *Protein Sci.*, **14**, 45–54.
13. Kouskouti,A., Scheer,E., Staub,A., Tora,L. and Talianidis,I. (2004) Gene-specific modulation of TAF10 function by SET9-mediated methylation. *Mol. Cell*, **14**, 175–182.
14. Chuikov,S., Kurash,J.K., Wilson,J.R., Xiao,B., Justin,N., Ivanov,G.S., McKinney,K., Tempst,P., Prives,C., Gamblin,S.J. *et al.* (2004) Regulation of p53 activity through lysine methylation. *Nature*, **432**, 353–360.
15. Su,I.H., Dobenecker,M.W., Dickinson,E., Oser,M., Basavaraj,A., Marqueron,R., Viale,A., Reinberg,D., Wulfing,C. and Tarakhovsky,A. (2005) Polycomb group protein ezh2 controls actin polymerization and cell signaling. *Cell*, **121**, 425–436.
16. Lee,J., Sayegh,J., Daniel,J., Clarke,S. and Bedford,M.T. (2005) PRMT8, a new membrane-bound tissue-specific member of the protein arginine methyltransferase family. *J. Biol. Chem.*, **280**, 32890–32896.
17. Kwak,Y.T., Guo,J., Prajapati,S., Park,K.J., Surabhi,R.M., Miller,B., Gehrig,P. and Gaynor,R.B. (2003) Methylation of SPT5 regulates its interaction with RNA polymerase II and transcriptional elongation properties. *Mol. Cell*, **11**, 1055–1066.
18. Mowen,K.A., Schurter,B.T., Fathman,J.W., David,M. and Glimcher,L.H. (2004) Arginine methylation of NIP45 modulates cytokine gene expression in effector T lymphocytes. *Mol. Cell*, **15**, 559–571.
19. Tsukada,Y., Fang,J., Erdjument-Bromage,H., Warren,M.E., Borchers,C.H., Tempst,P. and Zhang,Y. (2006) Histone demethylation by a family of JmjC domain-containing proteins. *Nature*, **439**, 811–816.
20. Boisvert,F.M., Cote,J., Boulanger,M.C. and Richard,S. (2003) A proteomic analysis of arginine-methylated protein complexes. *Mol. Cell. Proteomics*, **2**, 1319–1330.
21. MacCoss,M.J., McDonald,W.H., Saraf,A., Sadygov,R., Clark,J.M., Tasto,J.J., Gould,K.L., Wolters,D., Washburn,M., Weiss,A. *et al.* (2002) Shotgun identification of protein modifications from protein complexes and lens tissue. *Proc. Natl Acad. Sci. USA*, **99**, 7900–7905.
22. Ong,S.E., Mittler,G. and Mann,M. (2004) Identifying and quantifying *in vivo* methylation sites by heavy methyl SILAC. *Nature Meth.*, **1**, 119–126.
23. Wu,C.C., MacCoss,M.J., Howell,K.E. and Yates,J.R.III (2003) A method for the comprehensive proteomic analysis of membrane proteins. *Nat. Biotechnol.*, **21**, 532–538.
24. Kim,J.H., Lee,J., Oh,B., Kimm,K. and Koh,I. (2004) Prediction of phosphorylation sites using SVMs. *Bioinformatics*, **20**, 3179–3184.
25. Daily,K.M., Radivojac,P. and Dunker,A.K. Intrinsic disorder and protein modifications: building an SVM predictor for methylation. *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB 2005,* San Diego, California, USA, November 2005, pp. 475–481.
26. Chang,C.C. and Lin,C-.J. LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.
27. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
28. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
29. Wagner,K., Mincheva,A., Korn,B., Lichter,P. and Popperl,H. (2001) Pbx4, a new Pbx family member on mouse chromosome 8, is expressed during spermatogenesis. *Mech. Dev.*, **103**, 127–131.
30. Waskiewicz,A.J., Rikhof,H.A. and Moens,C.B. (2002) Eliminating zebrafish pbx proteins reveals a hindbrain ground state. *Dev. Cell*, **3**, 723–733.

31. Wang,Y., Tai,G., Lu,L., Johannes,L., Hong,W. and Luen Tang,B. (2005) Trans-Golgi network syntaxin 10 functions distinctly from syntaxins 6 and 16. *Mol. Membr. Biol.*, **22**, 313–325.
32. Knauth,P., Schluter,T., Czubayko,M., Kirsch,C., Florian,V., Schreckenberger,S., Hahn,H. and Bohnensack,R. (2005) Functions of sorting nexin 17 domains and recognition motif for P-selectin trafficking. *J. Mol. Biol.*, **347**, 813–825.
33. van Kerkhof,P., Lee,J., McCormick,L., Tetrault,E., Lu,W., Schoenfish,M., Oorschot,V., Strous,G.J., Klumperman,J. and Bu,G. (2005) Sorting nexin 17 facilitates LRP recycling in the early endosome. *EMBO J.*, **24**, 2851–2861.