Supplementary materials - MeMo: A web tool for prediction of protein methylation modifications

Hu Chen[1†], Yu Xue[2†], Ni Huang[1], Xuebiao Yao[2,*] and Zhirong Sun[1,*]

[1]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;

[2]Laboratory of Cellular Dynamics, Hefei National Laboratory for Physical Sciences, and the University of Science and Technology of China, Hefei, China 230027

Running title: Protein methylation sites prediction

*To whom correspondence should be addressed.

Zhirong Sun: Tel: +86-10-62772237; Fax: +86-10-62772237; Email: sunzhr@mail.tsinghua.edu.cn

Xuebiao Yao: Tel: +86-551-3606294; Fax: +86-551-3607141; Email: yaoxb@ustc.edu.cn

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

Data Processing

(+) data set

We denote the amino acid residues that undergo methylation modification as positive samples (+), while those non-methylatable residues are designated as negative samples (-). Firstly, we obtained the data set of methylation sites from the feature table of SWISS-PROT (version 48) (1). Only experimentally verified methylation sites were selected. Potential methylation 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 residues (see in Table S1). We then searched the PubMed with the keywords of “methylation lysine” and “methylation arginine” for information on lysine and arginine methylation, respectively. From ~1,700 scientific articles, we collected 107 and 264 unambiguously and experimentally verified methylation sites for lysine and arginine, separately. Finally, we combined the newly curated data and the data derived from SWISS-PROT into an integrated positive (+) data set. Since only lysines (227 items) and arginines (273 items) had enough data entries to train and test the SVM models, we focused on the methylation of lysine and arginine residues and did not include other residues into consideration. The statistics of the (+) data processing is listed in Table S2.

(-) data set

As previously described (2-4), the (-) sites are composed of non-annotated lysine/arginine sites in the same proteins from which (+) sites are taken, instead of using proteins randomly picked from the SWISS-PROT database. Thus, both (+) and (-) sites are extracted from the same protein pool, making our test more strict. Clearly the (-) sites may contain some false negative samples – these lysine/arginine sites in fact undergo methylation but are not known so far. As a result, the SVMs’ performance measurements will overestimate the false positive rates. However, without a reliable standard (-) set, this overestimation is inevitable. The detailed information of both (+) and (-) data used are listed in Table S2.

Algorithm design and validation

Sequence coding

We employed a traditional sliding window strategy to represent methylation sites. The considered window size was 14 symmetrical residues because preliminary tests show that 14 is the minimum size to achieve good performance. A fragment of 14 amino acids centering on methylated residues was adopted to represent the considered methylation site. Since there is always K or R in the considered methylation site we didn't include the center methylation site into the encoding fragment. We chose orthogonal binary coding scheme to transform protein sequences into numeric vectors. For example, glycine was designated as 00000000000000000001, alanine designated as 00000000000000000010, and so on. The length of final vector representing the methylated site is 7×2×20=280.

SVM and Parameter search:

The support vector machine (SVM) is a new machine learning method, which has been applied for many kinds of pattern recognition problems. The principle of the SVM method is to transform the samples into a high dimension Hilbert space and seek a separating hyperplane in the space. The separating hyperplane, which is called the optimal separating hyperplane, is chosen in such a way as to maximize its distance from the closest training samples. As a supervised machine learning technology, SVM is well founded theoretically on Statistical Learning Theory[30, 31]. The SVM usually outperforms other traditional machine learning technologies, including the neural network and the k-nearest neighbor classifier. Recently, SVM has been successfully adopted to solve many biological problems, such as predicting protein subcellular locations (5), protein secondary structures (5,6), tumor classification (7) and phosphorylated sites (2).

In this work, we have employed LIBSVM (8) to build SVM models. The considered parameters include: kernel function types (RBF and polynomial), gamma, the penalty parameter C. The parameters combination used for training is shown in Table 1.

Selection of (-) sites and 7-fold cross-validation

Obviously there are many more (-) sites than (+) sites in our data sets. The SVMs trained with all these (-) sites will overweigh (-) sites and subsequently predict all sites as (-) sites. Hence we have employed a strategy, which is usually named "under-sampling" and has been used in previous work (2,9), to overcome the imbalance between (+) sites and (-) sites. At first, all the (+) sites and (-) sites were combined and then divided equally into seven parts, keeping the same distribution of (+) and (-) sites in each part. Then six parts were merged into a training data set while the seventh part was taken as a test data set. Since there are more under-sampling (-) sites in the training data than those of (+) sites, we reduced the number of (-) sites to keep (+) sites and (-) sites balanced. SVM models were then trained on the balanced training set and tested on the test data set. Seven-fold cross validation was carried out. The average accuracy of cross validation was used to estimate the performance.

Performance evaluation of MeMo

Performance measurements

We have adopted four frequently used measurements: *accuracy*, *specificity*, *sensitivity* and *Mathew correlation coefficient* (*MCC*), to evaluate our prediction system's performance. *Accuracy* represents the correct ratio among both positive and negative data sets, while *sensitivity* and *specificity* illustrate the correct prediction ratios of positive and negative data sets respectively. But when the number of positive data and negative data differ too much from each other, the *Mathew correlation coefficient* (*CC*) should be calculated to assess the prediction performance. The value of *MCC* ranges from -1 to 1, and a larger *MCC* stands for better prediction performance.

Among the data with positive predictions by MeMo, the real positives are defined as *true positives* (*TP*), while the others are defined as *false positives* (*FP*). Among the data with negative predictions by MeMo, the real positives are defined as *false negatives* (*FN*), while the others are defined as *true negatives* (*TN*).

Then the measurements are defined as follows:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad sensitivity = \frac{TP}{TP + FN} \qquad specificity = \frac{TN}{TN + FP}$$

$$MCC = \frac{(TP \square TN) - (FN \square FP)}{\sqrt{(TP + FN) \square (TN + FP) \square (TP + FP) \square (TN + FN)}}$$

Performance Comparison

The parameter combinations and more details about the accuracy of MeMo are shown in Table 1. Performance comparison of MeMo to previous work (Daily *et al.*) (10) is shown in Table S3. To predict methylation sites, Daily *et al.* used numerous features including amino acid frequencies, aromatic content, flexibility scale, net charge, hydrophobic moment, beta entropy, disorder information and PSI-BLAST profiles. A principal component analysis was applied to reduce dimensionality. Compared with Daily's method, MeMo uses only sequence information and doesn't need dimensionality reduction. On the same dataset "From SWISS-PROT", MeMo's performance is comparable to Daily's. Adding the manually collected data from literatures improves the performance of MeMo to a great extent. There are two corresponding reasons: first, integrating more data means more information to SVM, resulting in a more accurate model; second, manually mined data are experimental verified and reported by the literature. Thus, they are more qualified compared to those from SWISS-PROT database.

Sequence logos

For MeMo, the sensitivity on lysine and arginine sites is nearly identical. And the better accuracy for arginine sites over lysine is entirely due to increased specificity. Why is the performance of our SVM models on arginine sites much better than on lysines sites? We suppose that in the current available data, sequences profiles of the flanking regions of methylated arginine sites might be of higher specificity. To validate this hypothesis we utilize the WebLogo program (11) to generate sequence logos, which represent residue compositions in an intuitive way. The methylated arginine sites (coded by “R”) are often in R-G rich regions which are much different from non-methylated arginine sites (Figure S2). In contrast, the methylated lysine sites (coded by “K”) are less conservative (Figure S3). Thus, the sequence pattern of methylated arginine sites is more conservative with higher specificity than methylated lysine sites. And this will lead MeMo to think unmethylated lysine sites are in fact methylated to a greater extent than arginine sites. Another potential reason might be that there exist many more methylated lysine sites than arginine sites that remain to be experimentally detected. Continuously more comprehensive experimental analyses remain to be performed to address this issue.

www server

Based on the trained SVM models a web server interface is built up, which is freely available at http://www.bioinfo.tsinghua.edu.cn/~tigerchen/memo.html. The data sets are available upon request. The screenshots of Memo are shown in Figure 1 and Figure S1.

Functional analysis of Methylated Proteins

In order to determine which types of proteins will be methylated, we search for Gene Ontology from QuickGO (http://www.ebi.ac.uk/ego/). The Table S4 and Table S5 show top five Gene Ontology (GO) categories of biological processes, molecular functions and cellular components for lysine methylated proteins and arginine methylated proteins, separately. In our non-redundant data set, there are 61 lysine methylated and 92 arginine methylated proteins, respectively.

For the 61 lysine methylated proteins, we have observed 213 distinct GO groups. And here we provide the top five GO items of biological processes, molecular functions and cellular components, respectively (Figure S4). The most abundant GO item of biological process in which lysine methylated proteins are implicated is “transport” (9 proteins). The other four biological processes are “chromosome organization and biogenesis (sensu Eukaryota)” (8 proteins), “nucleosome assembly” (7 proteins), “protein biosynthesis” (6 proteins) and “electron transport” (4 proteins). The most enriched GO group of molecular function is “nucleotide binding” (12 proteins). And the most frequent GO entry of cellular component is “nucleus” (10 proteins).

For the 92 arginine methylated proteins, we have observed 324 distinct GO groups. And here we provide the top five GO items of biological processes, molecular functions and cellular components, respectively (Figure S5). The top five GO items of biological process are “mRNA processing” (15 proteins), “transport” (15 proteins), “transcription” (13 proteins), “regulation of transcription, DNA-dependent” (13 proteins) and “nuclear mRNA splicing, via spliceosome” (7 proteins). The most enriched GO of molecular function is “protein binding” (12 proteins). The other four GO categories are “RNA binding” (33 proteins), “nucleotide binding” (30 proteins), “nucleic acid binding” (29 proteins) and “DNA binding” (17 proteins). And the most frequent GO of cellular component is “nucleus” (41 proteins).

Taken together, the functional analysis of the methylated proteins proposes that the functions of these proteins are quite diverse. Thus, the data set is suitable for our prediction work as training data.

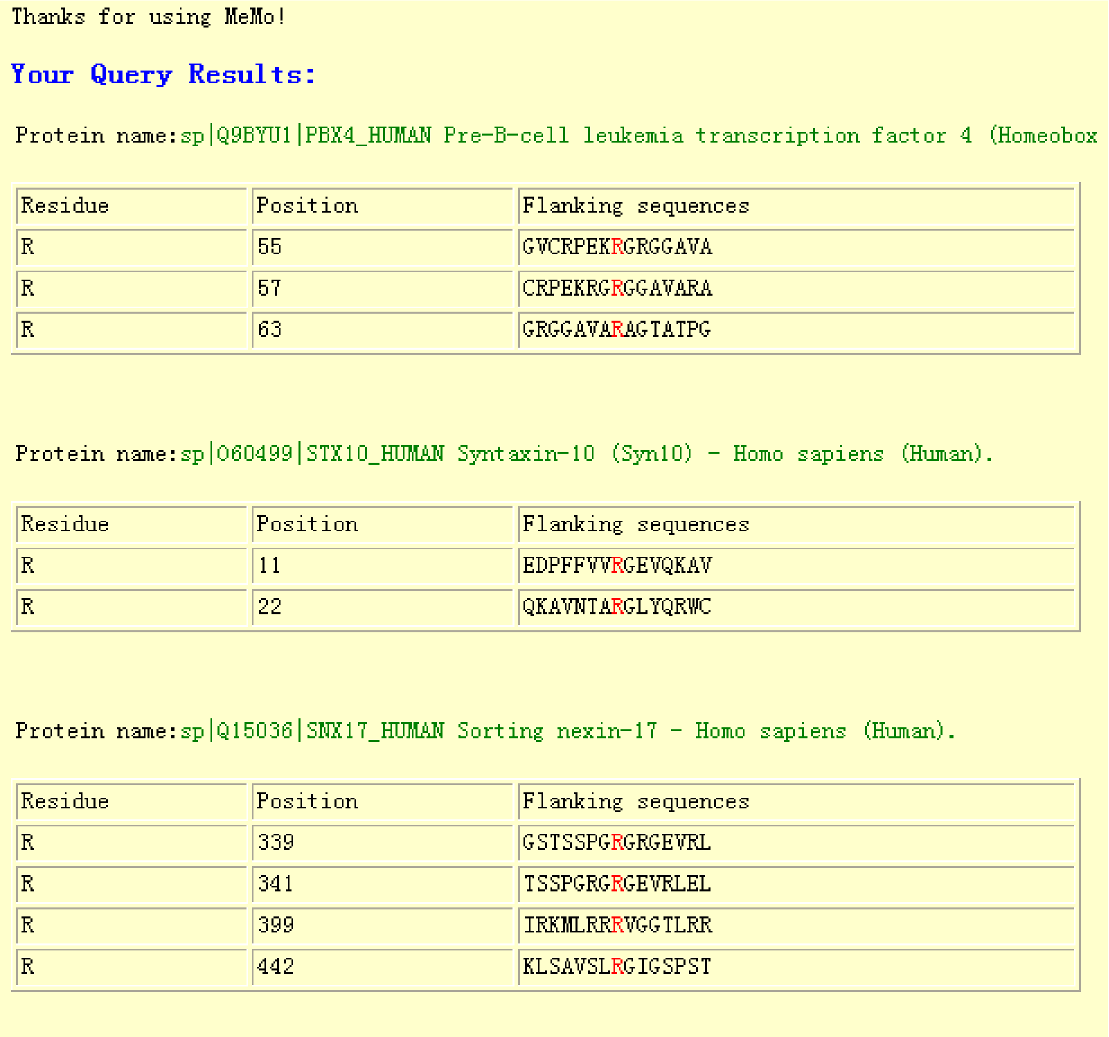
REFERENCES

1. 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.
2. Kim, J.H., Lee, J., Oh, B., Kimm, K. and Koh, I. (2004) Prediction of phosphorylation sites using SVMs. *Bioinformatics*, **20**, 3179-3184.
3. 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**, W184-187.
4. Zhou, F.F., Xue, Y., Chen, G.L. and Yao, X. (2004) GPS: a novel group-based phosphorylation predicting and scoring method. *Biochem Biophys Res Commun*, **325**, 1443-1448.
5. Hua, S. and Sun, Z. (2001) A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. *J Mol Biol*, **308**, 397-407.
6. Guo, J., Chen, H., Sun, Z. and Lin, Y. (2004) A novel method for protein secondary structure prediction using dual-layer SVM and profiles. *Proteins*, **54**, 738-743.
7. Lee, Y. and Lee, C.K. (2003) Classification of multiple cancer types by multicategory support vector machines using gene expression data. *Bioinformatics*, **19**, 1132-1139.
8. Chang, C.-C. and Lin, C.-J. (2001). Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.
9. Gomez, S.M. and Rzhetsky, A. (2002) Towards the prediction of complete protein--protein interaction networks. *Pac Symp Biocomput*, 413-424.
10. Daily, K.M., Radivojac, P. and Dunker, A.K. (2005), *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB 2005*, San Diego, California, U.S.A., pp. 475-481.
11. Crooks, G.E., Hon, G., Chandonia, J.M. and Brenner, S.E. (2004) WebLogo: a sequence logo generator. *Genome Res*, **14**, 1188-1190.
12. Mulder, N.J., Apweiler, R., Attwood, T.K., Bairoch, A., Barrell, D., Bateman, A., Binns, D., Biswas, M., Bradley, P., Bork, P. *et al.* (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res*, **31**, 315-318.

SUPPLEMENTARY LENGEND

SUPPLEMENTARY FIGURES

FIGURE S1- The screenshot of MeMo, showing prediction results



Thanks for using MeMo!

Your Query Results:

Protein name:sp|Q9BYU1|PBX4_HUMAN Pre-B-cell leukemia transcription factor 4 (Homeobox

Residue	Position	Flanking sequences
R	55	GVCRPEKRGRGGAVA
R	57	CRPEKRGRGGAVARA
R	63	GRGGAVARAGTATPG

Protein name:sp|O60499|STX10_HUMAN Syntaxin-10 (Syn10) - Homo sapiens (Human).

Residue	Position	Flanking sequences
R	11	EDPFFVVRGEVQKAV
R	22	QKAVNTARGLYQRWC

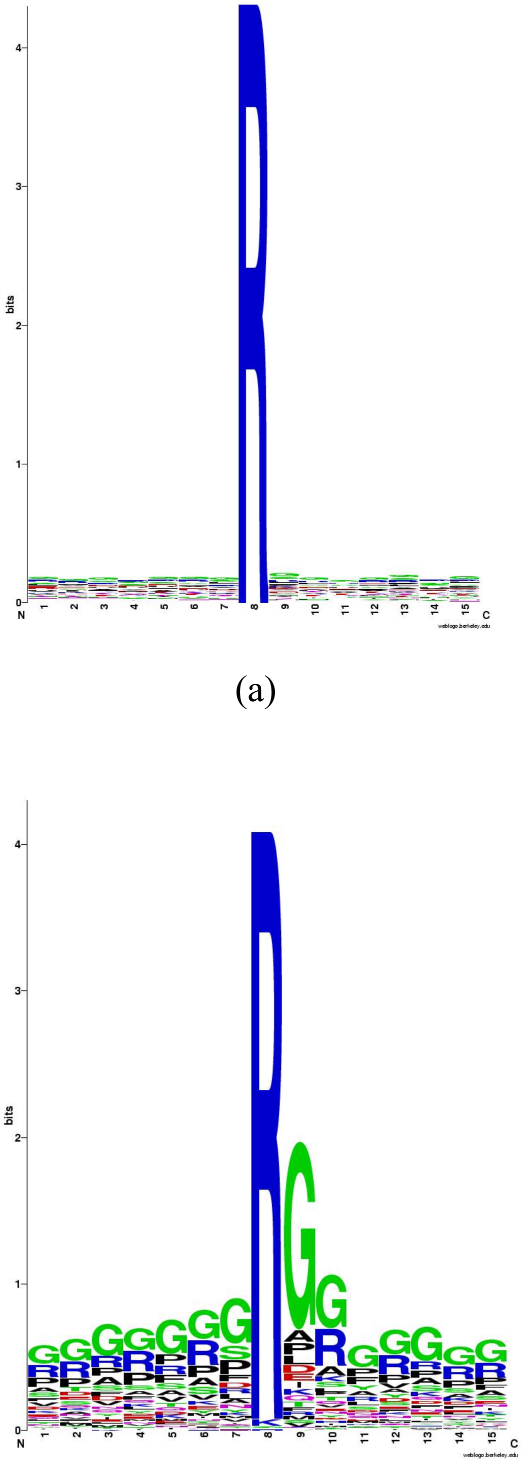
Protein name:sp|Q15036|SNX17_HUMAN Sorting nexin-17 - Homo sapiens (Human).

Residue	Position	Flanking sequences
R	339	GSTSSPGRGRGEVRL
R	341	TSSPGRGRGEVRLEL
R	399	IRKMLRRRVGGTLRR
R	442	KLSAVSLRGIGSPST

FIGURE S2. The sequence logos of arginine sites. A taller letter indicates that this kind of residue is more frequently used.

(a) The non-methylated arginine sites and their flanking sequences.

(b) The methylated arginine sites' pattern, rich of R and G.



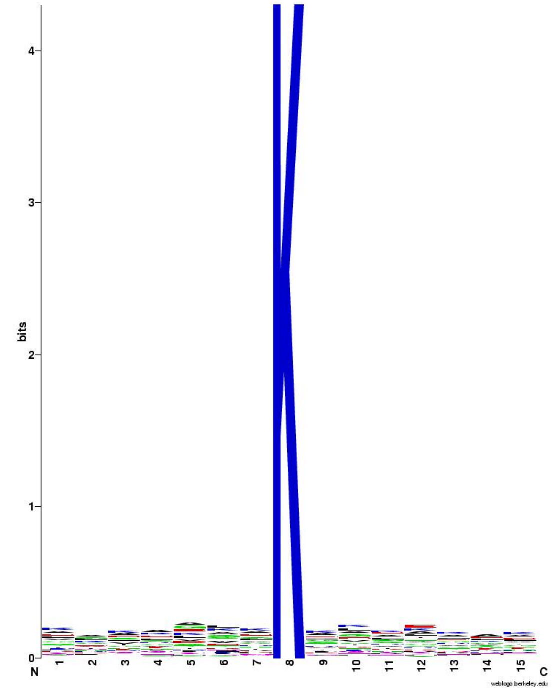
(a)

(b)

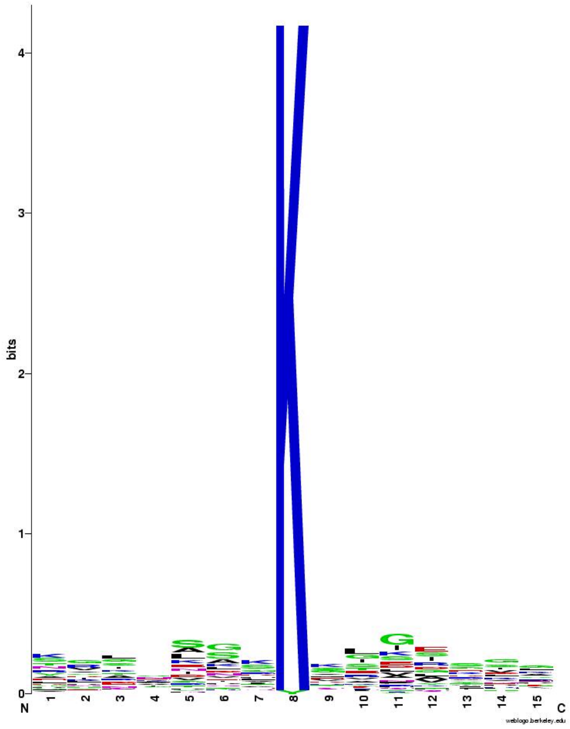
FIGURE S3. The sequence logos of lysine sites. A taller letter indicates that this kind of residue is more frequently used.

(a) The non-methylated lysine sites and their flanking sequences.

(b) The methylated lysine sites and their flanking sequences. There are not large differences between (a) and (b).



(a)



(b)

SUPPLEMENTARY TABLES

TABLE S1 - The numbers of methylated sites on different types of residues in SWISS-PROT.

Residue type	Number of methylated sites
Lysine	148
Arginine	76
Histidine	14
Asparagine	17
Cysteine	5
Others	68
In total	328

TABLE S2. Summary of (+) and (-) sites of lysine and arginine from both SWISS-PROT and manually collected data.

Data Set	Lysine		Arginine	
	(+) sites	(-) sites	(+) sites	(-) sites
From SWISS-PROT	148	-*	76	-*
From manual collection	107	-*	264	-*
In total	227	661	273	1395
After Homology-reduced (30%)	145	579	247	1211

*: the numbers are unavailable. We have firstly collected the positive (+) sites, after all the (+) sites are collected and merged together. Then we have retrieved non-annotated lysine/arginine sites as (-) sites from which the same proteins (+) sites were chosen, instead of randomly picking other proteins from the SWISS-PROT database.

TABLE S3. Performance comparison between MeMo and the previous work (10).

Method	Data set	Residue Type	Accuracy	Sensitivity	Specificity	MCC
Daily et al, 2005 (10)	From SWISS-PROT	Arginine	77.9%	73.6%	82.2%	0.40*
		Lysine	63.1%	65.9%	60.4%	0.13*
MeMo	From SWISS-PROT	Arginine	76.0%	70.6%	81.5%	0.37
		Lysine	60.6%	66.2%	55.5%	0.11
MeMo	From SWISS-PROT + Manual collection	Arginine	86.7%	69.6%	89.2%	0.54
		Lysine	67.1%	69.2%	66.7%	0.29

*. Daily et al (10) didn’t give out MCC values. MCC values shown here are calculated based on Sensitivity (*Sn*) and Specificity (*Sp*) values in their article (10).

TABLE S4. Top five Gene Ontology (GO) categories of biological processes, molecular functions and cellular components in lysine methylated proteins.

GO Symbol	Name of Gene Ontology	No. of Proteins
Top five biological process		
GO:0006810	transport	9
GO:0007001	chromosome organization and biogenesis (sensu Eukaryota)	8
GO:0006334	nucleosome assembly	7
GO:0006412	protein biosynthesis	6
GO:0006118	electron transport	4
Top five molecular function		
GO:0000166	nucleotide binding	12
GO:0003677	DNA binding	11
GO:0005515	protein binding	8
GO:0005525	GTP binding	8
GO:0046872	metal ion binding	7
Top five cellular component		
GO:0005634	nucleus	10
GO:0005694	chromosome	8
GO:0005737	cytoplasm	8
GO:0000786	nucleosome	7
GO:0016020	membrane	7

TABLE S5. Top five Gene Ontology (GO) categories of biological processes, molecular functions and cellular components in arginine methylated proteins.

GO Symbol	Name of Gene Ontology	No. of Proteins
Top five biological processes		
GO:0006397	mRNA processing	15
GO:0006810	transport	15
GO:0006350	transcription	13
GO:0006355	regulation of transcription, DNA-dependent	13
GO:0000398	nuclear mRNA splicing, via spliceosome	7
Top five molecular functions		
GO:0005515	protein binding	34
GO:0003723	RNA binding	33
GO:0000166	nucleotide binding	30
GO:0003676	nucleic acid binding	29
GO:0003677	DNA binding	17
Top five cellular components		
GO:0005634	nucleus	41
GO:0016020	membrane	26
GO:0016021	integral to membrane	18
GO:0030529	ribonucleoprotein complex	16
GO:0005737	cytoplasm	13