Briefings in Bioinformatics, 18(4), 2017, 647–658

doi: 10.1093/bib/bbw041
Advance Access Publication Date: 28 May 2016
Paper

Computational prediction of methylation types of covalently modified lysine and arginine residues in proteins

Wankun Deng, Yongbo Wang, Lili Ma, Ying Zhang, Shahid Ullah and Yu Xue

Corresponding author. Yu Xue, Key Laboratory of Molecular Biophysics of Ministry of Education, College of Life Science and Technology and the Collaborative Innovation Center for Brain Science, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China. Tel.: +86-27-87793903; Fax: +86-27-87793172, E-mail: xueyu@hust.edu.cn

Abstract

Protein methylation is an essential posttranslational modification (PTM) mostly occurs at lysine and arginine residues, and regulates a variety of cellular processes. Owing to the rapid progresses in the large-scale identification of methylation sites, the available data set was dramatically expanded, and more attention has been paid on the identification of specific methylation types of modification residues. Here, we briefly summarized the current progresses in computational prediction of methylation sites, which provided an accurate, rapid and efficient approach in contrast with labor-intensive experiments. We collected 5421 methyllysines and methylarginines in 2592 proteins from the literature, and classified most of the sites into different types. Data analyses demonstrated that different types of methylated proteins were preferentially involved in different biological processes and pathways, whereas a unique sequence preference was observed for each type of methylation sites. Thus, we developed a predictor of GPS-MSP, which can predict mono-, di- and tri-methylation types for specific lysines, and mono-, symmetric di- and asymmetrical di-methylation types for specific arginines. We critically evaluated the performance of GPS-MSP, and compared it with other existing tools. The satisfying results exhibited that the classification of methylation sites into different types for training can considerably improve the prediction accuracy. Taken together, we anticipate that our study provides a new lead for future computational analysis of protein methylation, and the prediction of methylation types of covalently modified lysine and arginine residues can generate more useful information for further experimental manipulation.

Key words: protein methylation; post-translational modification; methyllysine; methylarginine; methylation type

Wankun Deng is a PhD student at Huazhong University of Science and Technology. His major research interest is focused on the big data analysis of functional post-translational modifications (PTMs).
Yongbo Wang is a PhD student at Huazhong University of Science and Technology. He developed a database of EKPD for kinases and phosphatases in eukaryotes, and systematically analyzed genetic variations that potentially change protein phosphorylation.
Lili Ma is a PhD student at Huazhong University of Science and Technology. Her research interest is focusing on computational and experimental analysis of PTM regulations in cell death and mitosis.
Ying Zhang is a PhD student at Huazhong University of Science and Technology. She mainly focused on computational and experimental analysis of functional PTMs in cell autophagy.
Shahid Ullah is a PhD student at Huazhong University of Science and Technology. He focused on the collection and integration of phosphorylation sites.
Yu Xue is a professor at Huazhong University of Science and Technology. He is interested in using both computational and experimental approaches to elucidate how various PTMs can be functional and precisely regulate biological processes, such as mitosis, autophagy and tumorigenesis, by predicting PTM sites in proteins, re-constructing PTM-regulated networks, identifying PTM-associated genetic variations and providing experimental evidence for the predictions.
Submitted: 29 January 2016; **Received (in revised form):** 28 March 2016

Introduction

Protein methylation is one of the most important reversible posttranslational modifications (PTMs) [1–4]. Although it has been more than half a century since the discovery of lysine methylation [4], protein methylation attracts less attention than other PTMs such as phosphorylation, ubiquitination and acetylation. Difficulties in the development of new experimental techniques and reagents greatly hampered the research progress in the identification of protein methylation during the past years [5]. However, as a key regulatory mechanism, protein methylation is involved in a broad spectrum of biological and physiological processes, such as transcriptional and epigenetic regulation, cell metabolism and the development of human diseases [1, 6–9]. Protein methylation can occur on several types of amino acid residues, such as lysine (K), arginine (R), proline (P), histidine (H), alanine (A) and asparagine (N) in the form of *N*-methylation [2, 10, 11]. Furthermore, *O*-methylation of glutamic acids (E) and *S*-methylation of cysteine (C) and methionine (M) residues were also reported [12, 13]. Currently, most studies have been focused on *N*-methylations of lysine and arginine residues because the two PTMs are predominant types of protein methylation, with particularly importance [1, 14].

Lysine methylation was classified into three types, including mono-, di- and tri-methylation according to the numbers of hydrogen atoms on amino group substituted by methyl groups, while arginine methylation was also classified into three types, such as mono-, symmetric di- and asymmetrical di-methylation [1]. For methyllysines, all substitutions occur on the ε-*N* atom under the catalysis of protein lysine methyltransferases (PKMTs), while S-adenosyl-L-methionine provides methyl group, receives hydrogen atom and turns into S-adenosyl-L-homocysteine [1]. The substitutions can occur one, two or three times which result in mono-, di- or tri-methylation, respectively. In contrast, specific arginine residues can be mono- or di-methylated by protein arginine methyltransferases (PRMTs) [1]. Because there are two guanidino groups in an arginine residue, double substitutions of methyl groups can occur either on a single guanidino group or both guanidino groups, resulting in asymmetrical or symmetric di-methylation, respectively [1].

Different PKMTs or PRMTs catalyze different types of protein methylations [1, 15–17]. For example, PRMTs were classified into type I or type II families according to whether they carry out asymmetrical or symmetric di-methylation of arginine residues [1]. G9A, also called as EHMT2, is a H3K9 PKMT that catalyzes both mono- and di-methylation of H3K9 (H3K9me1 and H3K9me2) in euchromatin, whereas the G9A-dependent H3K9me1 actives the serine-glycine biosynthetic pathway and prompts cancer cell proliferation [15, 16]. In addition, EZH2 specifically performs the tri-methylation of H3K27 (H3K27me3), and mediates in ataxia-telangiectasia (A-T) neurodegeneration [17]. Also, different types of protein methylations are associated with distinct biological functions [6, 18–20]. For example, the mono-methylation of H3 lysine 56 (H3K56me1) catalyzed by G9a/KMT1C provides a docking site for the interaction with proliferating cell nuclear antigen (PCNA), to regulate DNA replication in mammals [6]. Previously, it was demonstrated that the N-terminus of trithorax group (trxG) protein binds and actives promoter regions tri-methylated on H3K4 (H3K4me3), while the majority of polycomb group (PcG) recruiter binding sites, but not its binding sites, are associated with H3K4me3 [18]. The recruitment of PcG or trxG proteins to inactive or active promoter regions counteractively determines the gene expression profiles during the embryonic development [18]. In glioblastoma, the tumor suppressor gene ST7 was found to be silenced by PRMT5, which catalyzes symmetric di-methylation of arginine residues on histone tails [20].

In contrast to labor-intensive and expensive experiments, accurately computational prediction of methylation sites in proteins has also emerged to be an alternative approach [5, 21–24]. Here we first introduced and summarized the current progresses in computational prediction of general protein methylation sites. Also, we collected and integrated 1521 methyllysines and 3900 methylarginines from the scientific literature. We further classified these methylation sites into different types based on the experimental evidence. From the data set, we observed that different types of methylated proteins prefer to participate in different biological processes and pathways, while distinct sequences preferences were observed around different type of methylation sites. Thus, we proposed the prediction of specific methylation types for modified residues might be more helpful against a general prediction. A previously developed algorithm, Group-based Prediction System (GPS) [25], was adopted and considerably improved for the training and prediction, whereas the robustness and accuracy were also carefully evaluated. Then we developed a computational tool of GPS-MSP (Methyl-group Specific Predictor) for the prediction of general protein methylation sites and methylation types of methyllysines and methylarginines. By comparison, GPS-MSP exhibited a competitive performance to other existing tools for the prediction of general methylation sites, while the classification of methylation sites into different types for training can further improve the accuracies. Taken together, we proposed that GPS-MSP can be a highly useful tool for the computational analysis of protein methylation. The online service and local packages of GPS-MSP can be freely accessed for academic research at http://msp.biocuckoo.org/.

Methods

Data collection and preparation

First, we searched the PubMed with multiple key terms such as 'protein methylation', 'lysine methylation' and 'arginine methylation'. The related articles were carefully read by eyes and experimentally identified lysine or arginine methylation sites were manually curated. As previously described [25], we removed redundant or homologous sites to avoid the overestimation of the prediction accuracy. The CD-HIT (32) with a threshold of 40% sequence identity was used to single out homologous proteins. If two proteins were found to be methylated at the same position and to have >40% sequence identity, only one of the two proteins was preserved. Totally, the non-redundant data set contained 1521 methyllysines in 962 proteins, and 3900 methylarginines in 1751 proteins. In our data set, there were 947 methyllysines and 3843 methylarginines identified with the methylation type information. All methylated proteins were mapped to the primary sequences (PSs) downloaded from the UniProt databases [26], while the methylation sites were exactly pinpointed. We did not limit our search to any specific organism, and a detailed statistics of numbers of methylated proteins and sites was shown for each species (Supplementary Table S1).

For the general prediction of methyllysines (K.all) and methylarginines (R.all) in proteins, we adopted a previously published approach to prepare the training data sets [25]. As previously described [27], the experimentally identified lysine or arginine methylation sites were regarded as positive data (+), while all the other non-methylated lysine or arginine residues

in the same proteins were taken as negative data (–). Totally, the non-redundant lysine methylation data set used for training contained 1521 positive sites and 47 972 negative sites, while the training data set of arginine methylation had 3900 positive sites and 86 111 negative sites (Table 1). In this work, we also predicted methylation types of methyllysines and methylarginines. Here, we denoted the data sets of mono-, di- and tri-methylated lysine residues as K.mono, K.di and K.tri, and mono-, symmetric di- and asymmetrical di-methylated arginine residues as R.mono, R.s.di and R.a.di, respectively. Because a considerable number of methylarginines were only identified with di-methylation information, we additionally denoted the data set of all di-methylation of arginine sites as R.di. The corresponding statistics was shown in Table 1. The data set of known methylated substrates, together with UniProt accession numbers, protein sequences, methylated positions, methylation types, organisms and PMIDs of original references, can be downloaded at http://msp.biocuckoo.org/download.php.

Performance evaluation

As previously described [27], the four measurements of sensitivity (*Sn*), specificity (*Sp*), precision (*Pr*) and Mathew Correlation Coefficient (*MCC*) were adopted to evaluate the prediction performance. The four measurements were defined as shown below:

$$Sn = \frac{TP}{TP + FN}, Sp = \frac{TN}{TN + FP}, \Pr = \frac{TP}{TP + FP}$$

and

$$MCC = \frac{(TP \times TN)(FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}}.$$

The leave-one-out (LOO) validation and 4-, 6-, 8- and 10-fold cross-validations were performed. The Receiver Operating Characteristic (ROC) curves and AROC (area under ROC) values were also carried out.

Algorithm

During the past decade, we developed a series of GPS algorithms for the prediction of PTMs sites in proteins. In this work, we applied the previously released GPS 3.0 algorithm with a significant improvement [25]. The algorithm comprised two parts, including the scoring strategy and performance improvement.

The basic hypothesis of the scoring strategy is that similarly short peptides exhibit similar biochemical properties with similar functions. Thus, we defined a *methylation site peptide* MSP(*m*, *n*) as a methyllysine/methylarginine amino acid flanked by *m* residues upstream and *n* residues downstream. Then we used the amino acid substitution matrix BLOSUM62 to estimate the similarity between two MSP(*m*, *n*) peptides *A* and *B* as:

$$S(A, B) = \sum_{\le m \le i \le n} Score(A[i], B[i])$$

Score(*A*[i], *B*[i]) denotes the substitution score of the two amino acids of *A*[i] and *B*[i] in the BLOSUM62 at the position *i*. If S(*A*, *B*) < 0, we redefined it as S(*A*, *B*) = 0. A given MSP(*m*, *n*) is then compared with each of the experimentally identified methyllysine/methylarginine peptides in a pairwise manner to calculate the similarity score. The average value of the substitution scores is taken as the final score.

The performance improvement procedure comprises four distinct steps, including *k*-means clustering, motif length selection, weight training and matrix mutation.

(i) *k-means clustering*: Given two MSP(m, n) peptides *A* and *B*, the similarity was defined and measured as s(*A*, *B*) = N_s/N. The *N* is the number of all substitutions, whereas the N_s is the number of conserved substitutions with Score(a, b) > 0 in the BLOSUM62 matrix. The s(*A*, *B*) ranges from 0 ∼ 1. Thus, the distance between them can be defined as D(*A*, *B*) = 1/s(*A*, *B*), if s(*A*, *B*) = 0, D(*A*, *B*) = ∞. By exhaustively testing, we determined the cluster number with the best LOO result for each data set, while MSP(7, 7) was adopted. First, *N* methylation sites from the positive data (+) were randomly chosen as the centroids, while *N* (*N* = 5 in this work) is number of clusters. Second, the other positive sites were compared in a pairwise manner with the five centroids and clustered into groups with the highest similarity values. Third, the centroid of each cluster was updated with the highest average similarity (HAS). The second and third steps were iteratively repeated until the clusters did not change any longer. After the five clusters for the positive sites had been determined, we put each negative site into the cluster with the HAS.

(ii) *Motif length selection*: In this step, the optimal combination of MSP(*m*, *n*) was determined based on the highest LOO result.

Table 1. The LOO results of GPS-MSP

Type	Positive[a]	Negative[b]	High			Medium			Low		
			Pr (%)	Sn (%)	Sp (%)	Pr (%)	Sn (%)	Sp (%)	Pr (%)	Sn (%)	Sp (%)
K.all	1521	47 972	43	23.58	99.01	20.75	41.05	95.04	12.46	44.51	90.1
K.mono	592	14 416	56.97	30.21	99.04	24.95	39.24	95.01	16.52	46.63	90.05
K.di	316	6532	63.58	29.36	99.16	28.31	38.53	95.12	19.91	42.2	91.5
K.tri	268	4396	73.75	42.14	99.04	48.94	74.29	95.06	33.39	76.43	90.29
R.all	3900	86 111	46.09	18.89	99	30.21	47.78	95.01	22.28	63.31	90.01
R.mono	3318	83 095	37.37	14.96	99.01	25.82	45.5	95.01	18.67	58.14	90
R.s.di	72	1071	74.36	40.85	99.07	50	74.65	95.05	33.97	75.35	90.29
R.a.di	1300	29 325	41.33	15.78	99.01	28.99	45.96	95.01	22.12	63.97	90.03
R.di	1847	39 901	43.08	16.38	99	28.69	43.49	95.01	20.38	55.04	90.06

Note. Three thresholds including high, medium and low stringency were established based on the *Sp* values of ∼99%, ∼95 and ∼90%, respectively.
[a]Positive the number of positive sites
[b]Negative the number of negative sites.

The *Sp* value was fixed at 95% for each data set, while the combinations of MSP(m, n) ($m = 1, \ldots, 15$; $n = 1, \ldots, 15$) were exhaustively tested.

(iii) *Weight training*: The substitution score between two MSP(m, n) peptides A and B was refined as:

$$S'(A,B) = \sum_{\leq m \leq i \leq n} w_i Score(A[i], B[i])$$

Initially, the weight of each position in MSP(m, n) was taken as 1. The w_i value is the weight of position i. Again, if $S'(A, B) < 0$, we redefined it as $S'(A, B) = 0$. Then we randomly picked out the weight of any position for $+1$ or -1 and re-computed the LOO result. The *Sp* value was fixed at 95%. The manipulation was adopted when the *Sn* value was increased. This process was continued until the *Sn* value was not increased any longer.

(iv) *Matrix mutation*: The aim of this step is to generate an optimal or near-optimal scoring matrix. BLOSUM62 was chosen as the initial matrix, and the LOO performance was calculated. Again, we fixed the *Sp* value at 95% for each data set, to improve the *Sn* though randomly picking out an element of the BLOSUM62 matrix for $+1$ or -1. This process was repeated until convergence was reached.

Because the original training process is too time-consuming, here we employed the simulated annealing (SA) algorithm to optimize the parameters for the steps of Weight Training and Matrix Mutation [28]. Such a procedure greatly improved the efficiency for training.

Statistical analysis

To analyze the functional distribution of human protein methylation, we downloaded Gene Ontology (GO) (version 125, released on 18 September 2013) [29] annotation files from the EBI Web site (http://www.ebi.ac.uk/GOA). There were 45 530 human proteins annotated with at least one GO term, including 531 lysine-methylated proteins and 750 arginine-methylated proteins. Here we defined the following:

N = number of proteins in human proteome annotated by at least one GO term.
n = number of proteins in human proteome annotated by the GO term t.
M = number of proteins in human methylated proteins annotated by at least one GO term.
m = number of proteins in human methylated proteins annotated by the GO term t.

Then the enrichment ratio (E-ratio) of the GO term t was calculated, and the p-value was calculated with the hypergeometric distribution [25] as below:

$$E\_ratio = \frac{\frac{m}{M}}{\frac{n}{N}}$$

$$p - value = \sum_{m'=m}^{n} \frac{\binom{M}{m'}\binom{N-M}{n-m'}}{\binom{N}{n}} (Enrichment\_ratio \geq 1) \text{ or}$$

$$p - value = \sum_{m'=0}^{m} \frac{\binom{M}{m'}\binom{N-M}{n-m'}}{\binom{N}{n}} (Enrichment\_ratio < 1)$$

In this work, we only considered the over-represented GO groups with E-ratio ≥ 1. Furthermore, we purchased a KEGG (Kyoto Encyclopedia of Genes and Genomes) FTP subscription for personal use [30], and mapped all human UniProt proteins to KEGG pathways if available. Totally, there were 6195 human proteins annotated with at least one KEGG entry, including 147 lysine and 236 arginine methylated proteins respectively. Analogously, we also performed similar analyses to identify statistically over-represented KEGG pathways that were associated with protein methylation.

The implementation of the web service and local packages

For a convenient usage, we constructed the online service in an easy-to-use manner, with a user-friendly interface using PHP and JavaScript. Also, IUPred [31] and NetSurfP [32] were integrated into the web service to predicted potential protein structural features, such as disorder regions, secondary structures (SSs) and surface accessibilities. Such predictions will be helpful for further experimental consideration. The Web site of GPS-MSP was extensively tested on various web browsers including Internet Explorer, Mozilla Firefox and Google Chrome to provide a robust service. In addition, for the prediction of large sequence data sets, the stand-alone packages were implemented in JAVA and supported for three major operation systems including Windows, Linux and Mac OS. For convenience, the online service and local packages of GPS-MSP were implemented in JAVA and freely available at http://msp.biocuckoo.org/.

Results

Current progresses in the prediction of protein methylation sites

Previously, most studies were mainly focused on histone methylations. However, owing to recent advances in the development of high-throughput techniques, such as the immunoaffinity enrichment of methylated peptides and the large-scale identification of methylation sites using mass spectrometry, more and more attention has been taken to the methylation of non-histone proteins [3, 14, 33]. Besides experimental approaches, a number of computational tools were developed for identifying potential methylation sites in proteins, and the predictions can greatly narrow down potential candidates for further experimental consideration [5, 21–24] Table 2.

In 2006, we collected 227 methyllysines and 273 methylarginines from the literature, and constructed a non-redundant positive data set (+) containing 156 lysine and 250 arginine methylation sites, by clearing the redundant sites which generated bias for the prediction accuracy [5] (Table 2). The negative data sets (−) were prepared as non-methylated lysine or arginine sites taken from the same methylated proteins. The PS features of amino acid compositions and frequencies around methylation sites were considered, and then we constructed the first web server of MeMo for predicting protein methylation sites [5]. Using MeMo, we predicted potential arginine methylation sites for three known methylated proteins, and the predictions were highly consistent with experimental evidence [5].

Table 2. A summary of currently available tools for predicting protein methylation sites

Predictor	Type[a]	Feature[b]	Training data[c]	
			Positive	Negative
MeMo [5]	K/R	PS	K (156); R (250)	
BPB-PPMS [21]	K/R	PS	K (188); R (216)	K (2157); R (1980)
MASA [22]	K/R/E/N	PS, ASA, SS	K (460); R (303); E (45); N (22)	K (6237); R (1216); E (885); N (375)
PLMLA [23]	K	PS, SS, PP	K (546)	K (2842)
PMeS [24]	K/R	PS, ASA, SS, PP	K (322); R (355)	K (4126); R (3960)
GPS-MSP	K/R	PS	K (1,521); R (3,900)	K (47,972); R (86,111)

Note. PS = primary sequence; ASA = solvent-accessible surface area; SS = secondary structure; PP =, ysicochemical property.
[a]The type of predictable methylation residue
[b]The features used for the prediction.
[c]For each predictor, the number of positive or negative sites in the training data set is shown in brackets.

Later, two predictors of BPB-PPMS [21] and MASA [22] were reported. For training the computational models, BPB-PPMS mainly used the PS features [21], while MASA further included the features of solvent-accessible surface area (ASA) and SS around the methylation sites [22]. Owing to the data limitation, the homologous or redundant methylation sites across different proteins were not cleared [21, 22] (Table 2). Both MeMo and BPB-PPMS can only predict methyllysines and methylarginines [5, 21], whereas MASA can predict potential methylation states for K, R, E and N residues [22] (Table 2). Recently, Shi *et al.* [23] integrated 546 lysine methylation sites and developed PLMLA specifically for predicting of methyllysines. Besides PS features, they also considered the SS features and physicochemical properties (PPs) of amino acids, such as hydrophobicity and charge [23] (Table 2). Later, they prepared a non-redundant data set containing 322 methyllysines and 355 methylarginines, and further included more features, including ASA and normalized van der Waals volume (VDWV) of 20 types of amino acids. The VDWV is also a PP feature. Then based on the features of PS, ASA, SS and PP, a highly useful tool of PMeS were constructed [24] (Table 2). Interestingly, all above-mentioned tools adopted the algorithm of Support Vector Machines for training the computational models [5, 21–24].

By summarizing all available tools for the prediction of methylation sites, we observed that including more features considerably but not dramatically increased the accuracy against the PS-based prediction [5, 21–24]. Also, because more features were considered, both the training and predicting processes are complicated and time-consuming. In this regard, we developed GPS-MSP by merely using PS features to balance the training time and prediction performance.

The functional distribution of different types of methylated proteins

Previously, experimental studies demonstrated that different types of protein methylations are differentially associated with distinct biological processes [6, 18–20]. However, a systematic analysis of the functional distribution of known methylated proteins remained to be performed to evaluate the correctness of experimental observations. In our data set, there were 962 lysine methylated proteins including 350 K.mono, 196 K-di and 173 K.tri substrates, and 1751 arginine methylated substrates including 1624 R.mono, 38 R.s.di and 560 R.a.di proteins. Using the data sets, we first performed an enrichment analysis of GO terms with the hypergeometric test (Figure 1, Supplementary Table S2, $P < 0.05$). For each type of methylated proteins, the top five most significantly over-represented biological processes, molecular functions and cellular components were shown (Figure 1). For lysine methylation, we observed that different types of methylated proteins are significantly associated with different GO terms. For example, K.mono but not K.di and K.tri proteins are statistically enriched in gene expression (GO:0010467) and RNA splicing (GO:0000398, GO:0008380), whereas the GO term of defense response to bacterium (GO:0042742) was only over-represented in K.di proteins (Figure 1A). Again, the GO term of adenine transport (GO:0015853) was only significantly associated with K.tri but not K.mono or K.di proteins. From the results, only two GO terms, including nucleosome assembly (GO:0006334) and nucleosome (GO:0000786), were significantly over-represented in all three types of lysine methylated proteins (Figure 1A). For arginine methylation, we got a similar result and different arginine methylated proteins also preferentially participated in distinct biological pathways. For example, both R.mono and R.a.di but not R.s.di proteins are statistically associated with gene expression (GO:0010467), whereas R.s.di proteins are significantly enriched with GO terms of viral process (GO:0016032), DSB repair (GO:0006302) and base-excision repair (GO:0006284) (Figure 1B).

Furthermore, we performed an additional enrichment analysis of KEGG pathways for different types of lysine (Figure 2A) and arginine methylated proteins (Figure 2B), while the detailed results were shown in Supplementary Table S3. The statistical analysis of R.s.di proteins was not performed, owing to the data limitation. Again, the results suggested that different types of methylated proteins play a different role in distinct pathways. Interestingly, we observed that K.mono, K.di and K.tri proteins are significantly associated with systemic lupus erythematosus (SLE, KEGG ID: hsa05322), a systemic autoimmune disease occurred after environmental triggering of genetically susceptible individuals [34]. Previous experiments demonstrated that H3K9me3 and H3K27me3 are highly associated with SLE [35, 36], whereas our results proposed a potentially more general mechanism that non-histone protein methylation might also participate in SLE.

Motif-based analysis of sequence preferences around different types of methylation sites

Protein lysine and arginine methylations are catalyzed by a variety of PKMTs and PRMTs, respectively [1, 15–17]. Thus, the unique sequence and 3D structure of a PKMT or PRMT will determine the recognition specificity of substrates. Because different types of methylated proteins prefer to be involved in

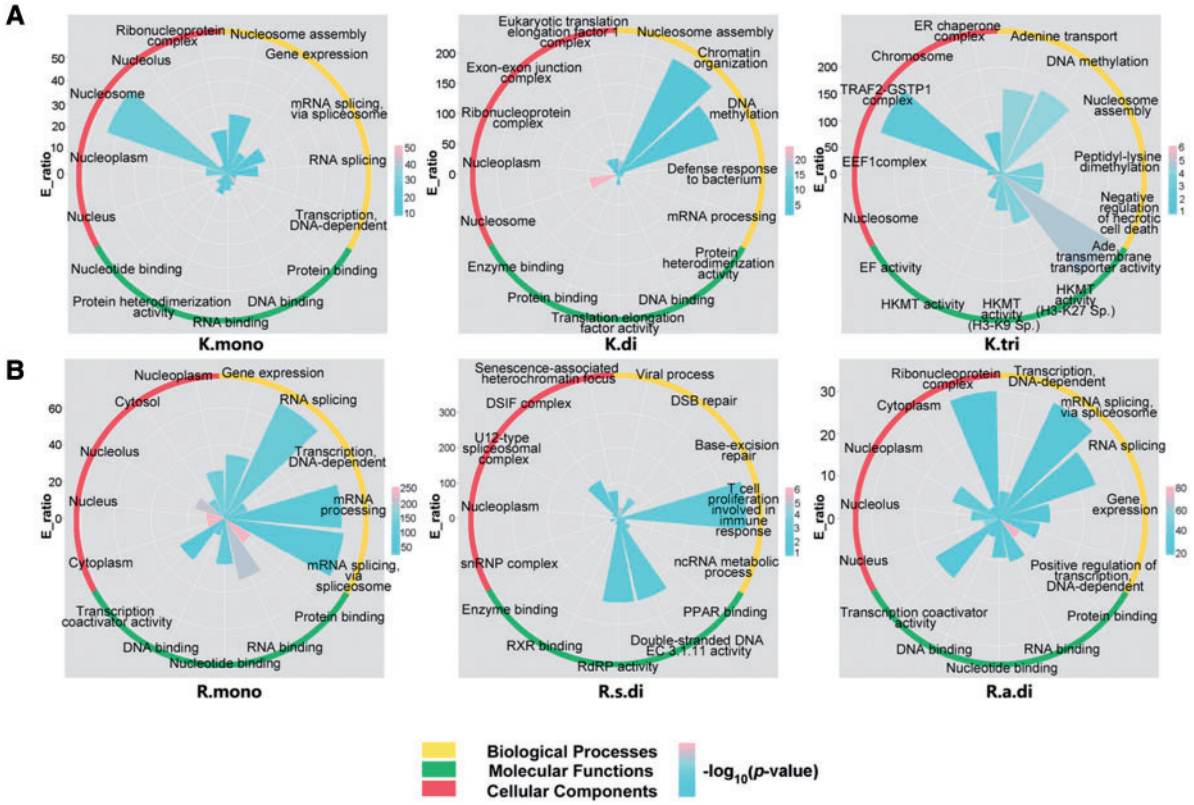

Figure 1. The functional distribution of different types of (**A**) lysine and (**B**) arginine methylated proteins, respectively. Three classes of GO terms including biological processes, molecular functions and cellular components were adopted, while the statistical enrichment analysis of GO terms for methylated proteins were performed with the hypergeometric distribution [25]. A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

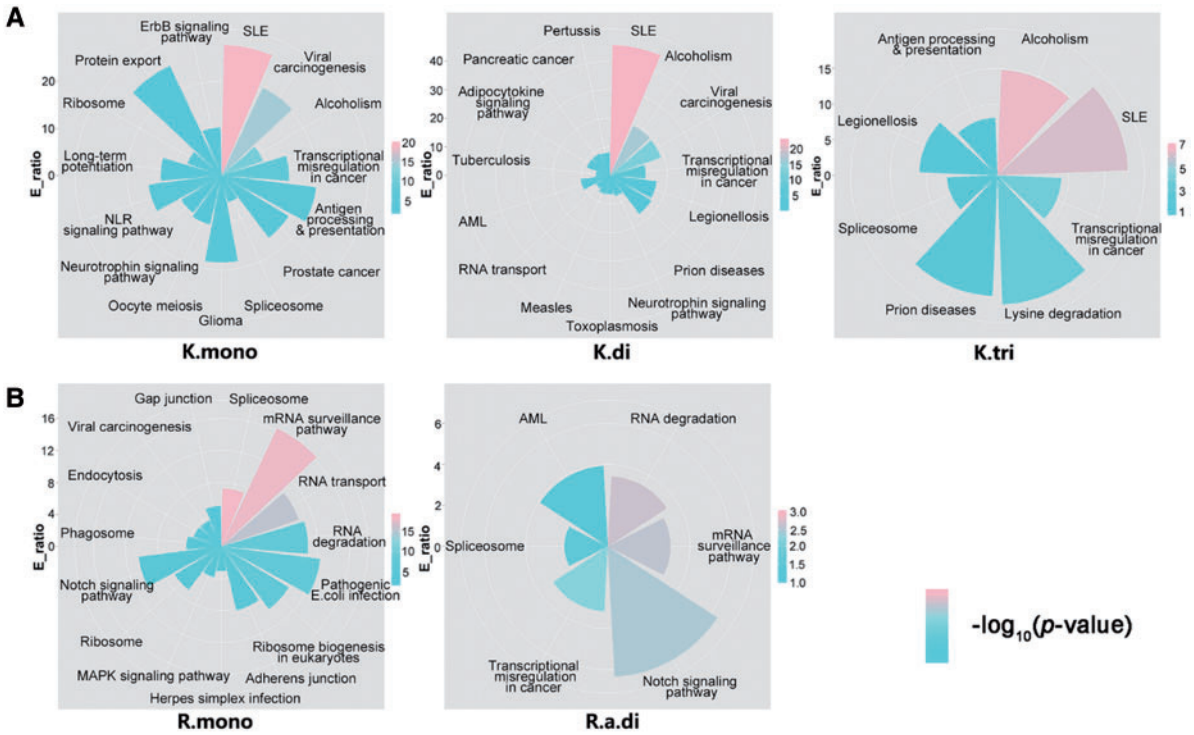

Figure 2. The enrichment analysis of KEGG pathways for different types of (**A**) lysine and (**B**) arginine methylated proteins, respectively. The analysis of R.s.di was not performed owing to the data limitation. A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

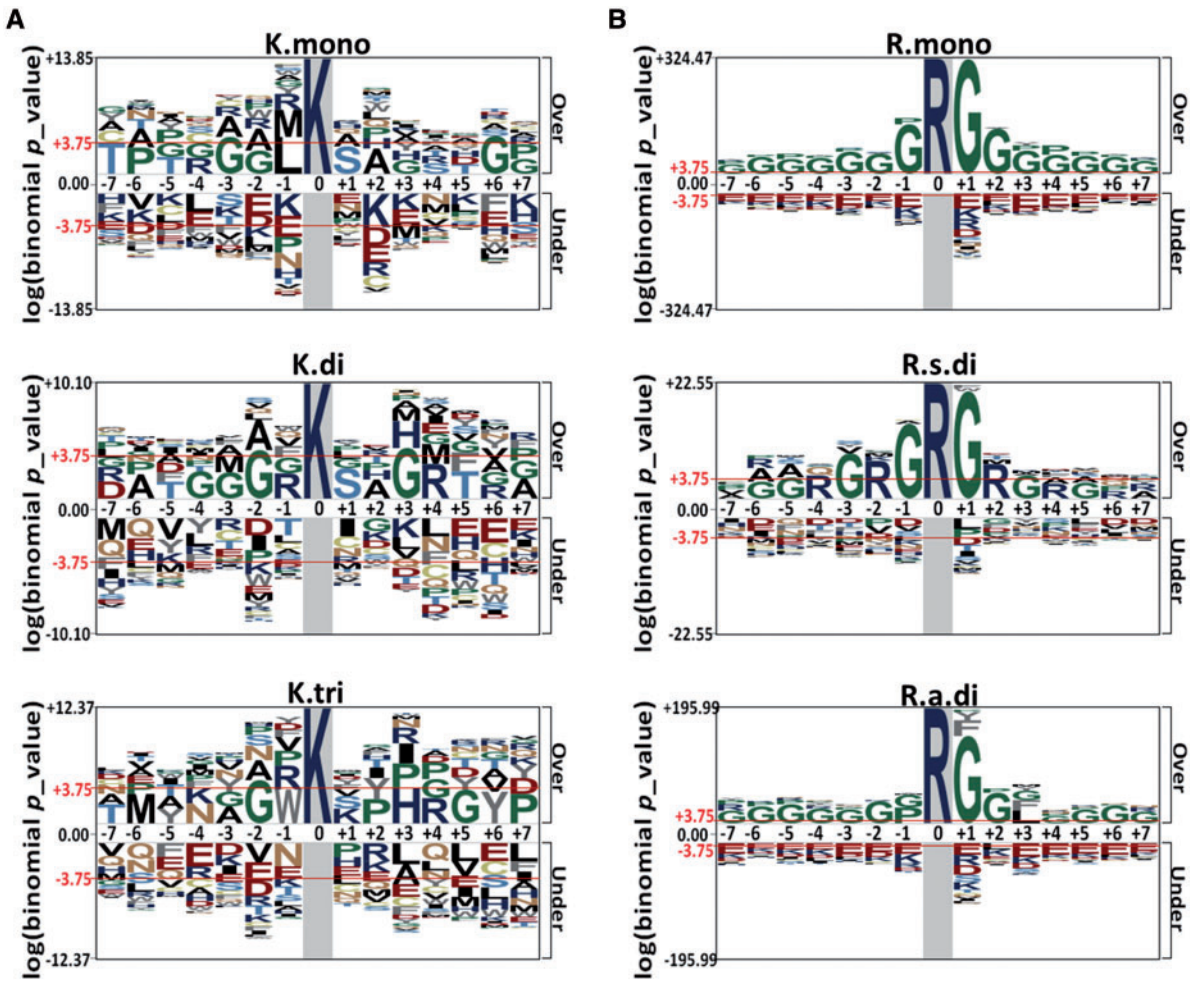

Figure 3. The amino acid frequencies of different types of (**A**) methyllysines and (**B**) methylarginines were analyzed and visualized by pLogo [37]. A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

distinct biological processes and pathways, here we speculated that there are potentially different sequence preferences around different types of methylation sites. To address this concern, we used pLogo [37], a highly useful tool to visualize sequence logos, to analyze the amino acid occurrence around methyllysines (Figure 3A) and methylarginines (Figure 3B), respectively.

Generally, the sequence preferences of lysine methylations are significantly different from arginine methylations (Figure 3). For all three types of methylarginines, there is a much informative and over-represented glycine/G residue at −1 position (Figure 3B). However, different types of arginine methylations still follow distinct sequence patterns. For example, the G residues are significantly enriched in all positions of the flanking region of R.mono, whereas a proline/P at +1 and an L at −3 positions are also over-represented for R.a.di (Figure 3B). However, the R residues prefer to occur at positions of +4, +2, −2 and −4 of R.s.di sites, which follow an RG repeat, and the result was consistent with experimental observations [38]. In contrast with methylarginines, the sequence profiles of lysine methylations are more complicated. For example, in position of +1, the L residue was mostly significant for K.mono, while the R and tryptophan/W residues were mostly enriched for K.di and K.tri, respectively (Figure 3A). Taken together, our results demonstrated that different types of methylation sites preferentially follow distinct sequence profiles.

Development of GPS-MSP for the prediction of methylation types of methyllysines and methylarginines

In this work, besides the general prediction of lysine and arginine methylation sites, GPS-MSP can also predict methylation types of covalently modified lysine and arginine residues. From the scientific literature, we totally collected 1521 methyllysines and 3900 methylarginines. We classified these sites based on their identified methylation types into seven sub-types, including K.mono, K.di, K.tri, R.mono, R.s.di, R.a.di and R.di (Table 1). Based on a previously developed algorithm of GPS 3.0 [25], we further adopted the SA algorithm to rapidly determine the optimal parameters of the computational model for each data set. As the first tool to predict methylation type-specific sites for lysine and arginine residues, GPS-MSP was developed in an easy-to-use manner, while both online service and stand-alone packages were provided.

For the usage of GPS-MSP web server, here we used the protein sequence of human p53 as an example (Supplementary Figure S1). The input of the web service contained three parts, including the methylation types, the protein sequences and the thresholds (Supplementary Figure S1A). The methylation types can be selected by clicking the checkboxes, while four threshold options including 'High', 'Medium' and 'Low' and 'All' were provided in the lower panel. The 'High', 'Medium' and 'Low' options

were selected with *Sp* values of ~99%, ~95 and 90%, respectively. The ''All' option will provide all the predicted results, with no stringency. One or multiple protein sequences could be input by the direct 'copy and paste' or upload of a sequence file in FASTA format. Furthermore, if the annotations of SS and surface accessibility for the inputted protein are needed, user could transfer to 'comprehensive' mode by clicking the 'here ≫' link. To ensure the stability of webserver, the input of protein sequences was limited with < 2M, while the large-scale computation could be performed through installing the stand-alone software packages locally.

After starting the prediction, the Web site will be redirected to a waiting page and then transferred to the result page (Supplementary Figure S1B). The results of p53 contained four sequential parts, including the list of 18 predicted methylation sites with the type information, the potential annotations of surface accessibilities and disorder regions, predicted SSs and the summarization of the results. All the results could be downloaded through clicking the 'Download Zip' button. To avoid the potential long waiting, if users submit multiple protein sequences, the prediction will be performed one by one and the prediction of next protein could be triggered by clicking the 'Next protein≫' link. Alternatively, users can download local packages for a more convenient and rapid prediction (Supplementary Figure S1C).

Performance evaluation and a comparison with other existing tools

To evaluate the robustness and accuracy of GPS-MSP, both of the LOO validation and 4-, 6-, 8- and 10-fold cross-validations were performed on each data set (Figure 4). For methyllysine prediction, the AROC values of LOO results are 0.697, 0.748, 0.686 and 0.870 for K.all, K.mono, K.di and K.tri, respectively (Figure 4). For the prediction of methylarginines, the AROC results of LOO validations are 0.848, 0.806, 0.848, 0.859 and 0.772 for R.all, R.mono, R.s.di, R.a.di and R.di, respectively (Figure 4). The detailed results under the high, medium and low cutoff values were shown in Table 1. From the results, we also observed that the prediction of methylation types of methyllysines and methylarginines considerably increased the accuracy against the general predictions (Figure 4). In addition, the results of 4-,

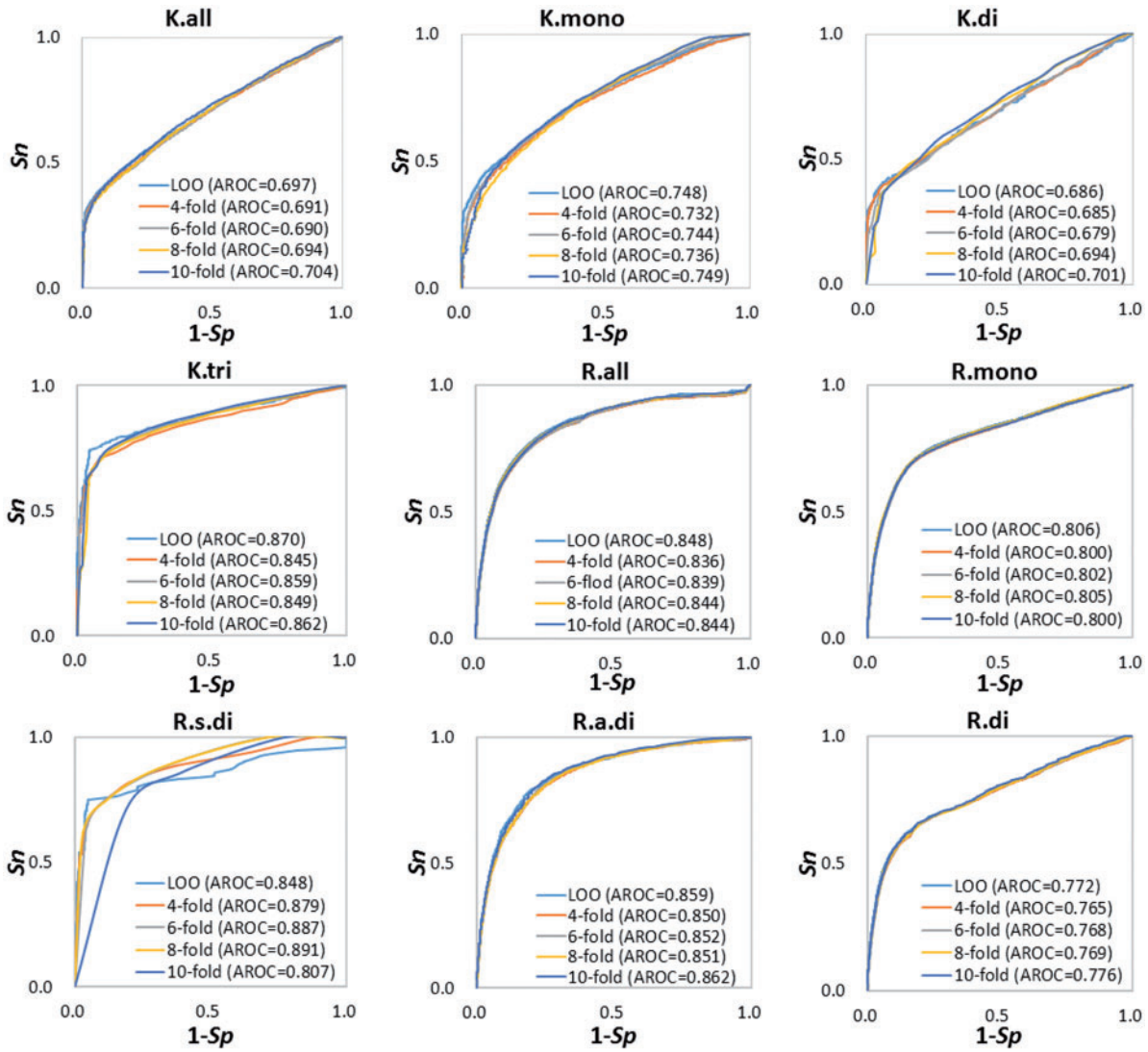

Figure 4. The LOO validation and 4-, 6-, 8-, 10-fold cross-validations were performed for all types of protein methylations, including K.all, K.mono, K.di, K.tri, R.all, R.mono, R.s.di, R.a.di and R.di. The AROC values were calculated. A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

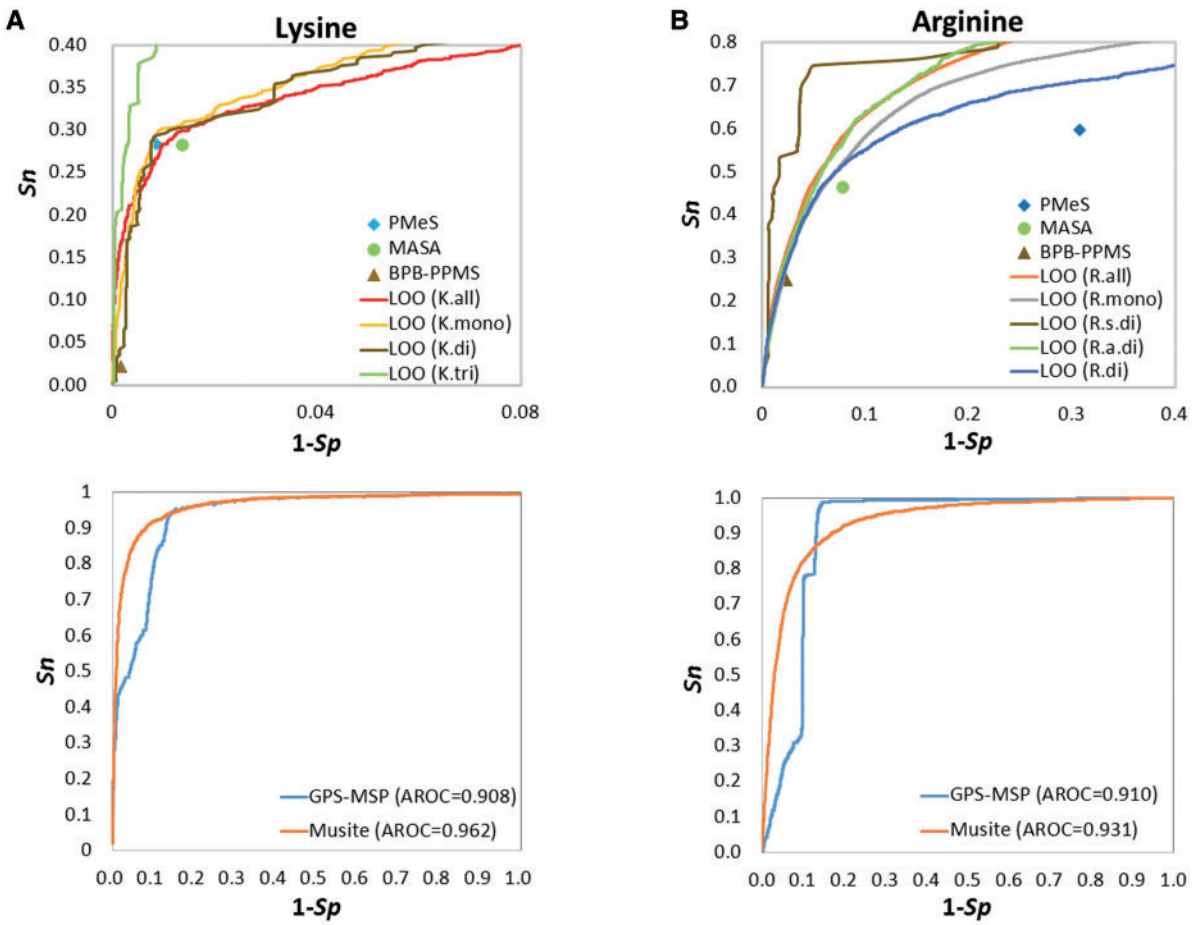

Figure 5. The comparison of GPS-MSP with other existing tools, including BPB-PPMS [21], MASA [22] and PMeS [24]. We directly submitted the training data set to these tools for the prediction, while the LOO results of GPS-MSP were used for the comparison. (**A**) For methyllysines; (**B**) For methylarginines. A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

6-, 8- and 10-fold cross-validations are highly similar with the LOO validations. Thus, GPS-MSP is a stable and robust predictor with a satisfying performance.

To further exhibit the superiority of our method, here we compared the performance of GPS-MSP with other existing tools, including BPB-PPMS [21], MASA [22] and PMeS [24]. Because these tools can only predict general methylation sites in proteins, we directly submitted the training data sets of K.all and R.all for the prediction, while the LOO results of GPS-MSP were used for a comparison (Figure 5). For the general prediction of methyllysines, only the performance of PMeS but not BPB-PPMS and MASA was slightly higher than the LOO result of GPS-MSP (Figure 5A). In contrast, GPS-MSP was much better than other tools on methylarginine predictions (Figure 5B). The promising performance of GPS-MSP might be owing to either better methodology or the larger data set for training. Thus, we compared GPS-MSP with an existing tool of Musite [39–41], using the same data set for an unbiased comparison. Originally, Musite was reported as a PS-based tool mainly for the prediction of general and kinase-specific phosphorylation sites [39–41]. Recently, the new version of Musite further integrated more features such as SS properties, and can train predictive models from customized data sets (Unpublished, personal communications). The training data set of GPS-MSP was used for training models of Musite. Because the training procedure of Musite was quite time-consuming, here we only performed the self-consistency validations to compare GPS-MSP and Musite for the prediction of methyllysines (Figure 5C) and methylarginines (Figure 5D), respectively. It was shown that the Musite algorithm was slightly better than GPS-MSP, as GPS-MSP only used PS features. Taken together, our results demonstrated that the accuracy of GPS-MSP can be better or at least comparative with previously reported predictors, whereas the methylation type-specific predictions can considerably improve the performance.

Discussion

The past decade has witnessed a rapid progress in the identification of protein methylation. In 2006, we only obtained 500 lysine and arginine methylation sites from the literature [5]. However, we currently collected and integrated over 5000 experimentally identified methyllysines and methylarginines in proteins, with a >10-fold increase. Because more and more sites were characterized, our understandings on protein methylation were greatly advanced, while accumulative evidence demonstrated that the aberrant methylation is highly associated with human diseases [1, 6–9]. In this regard, the analysis of site-specific protein methylation will provide important hints for further understanding regulatory mechanisms of the cellular signaling, and generate potential drug targets for biomedical usage.

Owing to the data accumulation, here we compiled a much larger data set of experimentally identified protein methylation

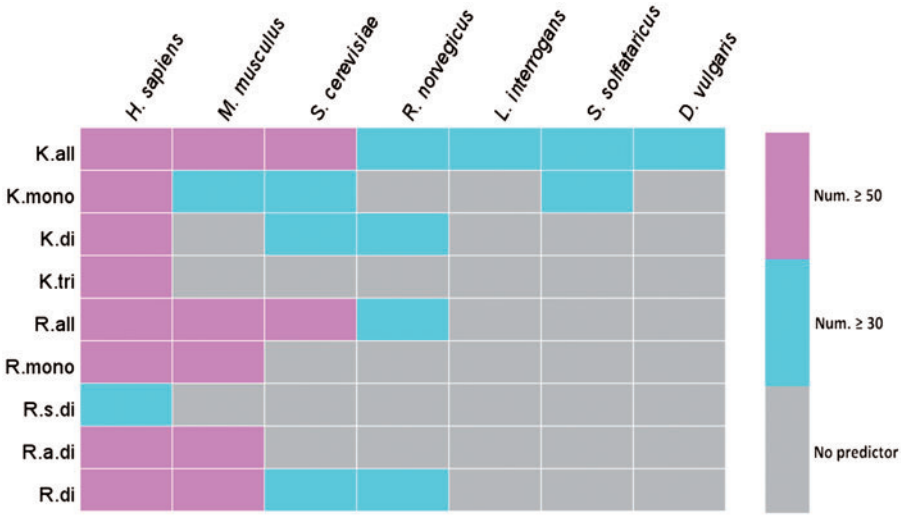

Figure 6. A summary of organism-specific predictors with ≥50 or ≥30 known methylation sites for training. A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

sites. We took all known methylation sites as positive data (+), and all non-methylated lysines or arginines in the same proteins as negative data (−). The negative data set can also be prepared with other methods [42]. For example, Musite included unphosphorylated sites from both phosphorylated and unphosphorylated proteins [40]. However, a previous analysis demonstrated that different training data construction methods generated similar and consistent results for the performance evaluation [42]. By analyzing the data set, we observed that different types of methylated proteins prefer to be significantly enriched in distinct biological processes and pathways, and there is a strong sequence preference for each type of methylation sites. Then we used the data set for training computational models, and developed GPS-MSP for the prediction of methylation sites. Besides the general prediction of methyllysines and methylarginines, GPS-MSP can also predict potential methylation types for modified sites. The robustness and performance were critically evaluated, and GPS-MSP was compared with other existing tools. We observed that the classification of methylation sites into distinct types can considerably improve the prediction accuracy, whereas the satisfying accuracy proposed that GPS-MSP can be a useful predictor for analyzing protein methylation.

In our data set, known methylation sites were collected from up to 106 organisms (Supplementary Table S1). Although the number of methylation sites was limited in most of species, we still constructed 15 and 13 organism-specific predictors with ≥50 and ≥30 sites, for the prediction of general or type-specific methylation sites, in *Homo sapiens*, *Mus musculus*, *Saccharomyces cerevisiae*, *Rattus norvegicus*, *Leptospira interrogans*, *Sulfolobus solfataricus* and *Desulfovibrio vulgaris*, respectively (Figure 6). Also, we performed the LOO validation and 4-, 6-, 8- and 10-fold cross-validations for organism-specific predictors with ≥50 sites of *H. sapiens* (Supplementary Figure S2), *M. musculus* and *S. cerevisiae* (Supplementary Figure S3).

For the future prediction of protein methylation sites, we proposed that currently available tools including GPS-MSP should be maintained and improved for academic research. If available, newly identified protein methylation sites will be continuously collected and integrated into computational models, for a better prediction. Also, analogous to protein phosphorylation which is catalyzed by numerous protein kinases (PKs) [43], lysine and arginine methylations are also performed by various PKMTs and PRMTs, respectively. Because different PKs exhibit different specificities for the recognition of substrates, we believe that protein methylation is also carried out in a PKMT- or PRMT-specific manner. However, owing to the data limitation, the prediction of PKMT- or PRMT-specific methylation sites is still not available in the current stage. Also, although 208 human methyltransferases were computationally identified [44], the exact numbers of PKMTs and PRMTs in most of the organisms were not known. In this regard, both the collection of methyltransferase-specific sites and the characterization of PKMTs and PRMTs in eukaryotes will be important challenges for future studies. In addition, although several thousands of protein methylation sites were identified, the biological functions and regulatory roles of most of sites were not reported. Thus, combining both computational predictions and experimental validations will generate more useful information, and propel the study of protein methylation into a new phase.

Key Points

- Different types of methylated proteins are preferentially involved in distinct biological processes and pathways.
- Different types of methyllysines and methylarginines have different sequence profiles for the modification.
- The GPS-MSP is a computational tool for the prediction of different types of methyllysines and methylarginines, besides the general prediction of methylation sites in proteins.
- The classification of methylation site into different types for training can considerably improve the prediction accuracy.

Supplementary data

Supplementary data are available online at http://bib.oxfordjournals.org/.

Funding

This work was supported by grants from the National Basic Research Program (973 project) (2013CB933900, 2012CB910101 and 2011CB910600), Natural Science

Foundation of China (31171263, 81272578 and J1103514) and International Science & Technology Cooperation Program of China (2014DFB30020).

Acknowledgement

We are grateful for Prof. Dong Xu (UMC), Dr Jianjiong Gao and Qiuming Yao, for providing us the new version of Musite, for training the customized models to predict protein methylation sites.

References

1. Paik WK, Paik DC, Kim S. Historical review: the field of protein methylation. *Trends Biochem Sci* 2007;**32**:146–52.
2. Bannister AJ, Kouzarides T. Reversing histone methylation. *Nature* 2005;**436**:1103–6.
3. Herz HM, Garruss A, Shilatifard A. SET for life: biochemical activities and biological functions of SET domain-containing proteins. *Trends Biochem Sci* 2013;**38**:621–39.
4. Ambler RP, Rees MW. Epsilon-N-Methyl-lysine in bacterial flagellar protein. *Nature* 1959;**184**:56–7.
5. Chen H, Xue Y, Huang N, *et al*. MeMo: a web tool for prediction of protein methylation modifications. *Nucleic Acids Res* 2006;**34**:W249–53.
6. Yu Y, Song C, Zhang Q, *et al*. Histone H3 lysine 56 methylation regulates DNA replication through its interaction with PCNA. *Mol Cell* 2012;**46**:7–17.
7. Brinkmann SJ, de Boer MC, Buijs N, *et al*. Asymmetric dimethylarginine and critical illness. *Curr Opin Clin Nutr Metab Care* 2014;**17**:90–7.
8. Li T, Chen H, Li W, *et al*. Promoter histone H3K27 methylation in the control of IGF2 imprinting in human tumor cell lines. *Hum Mol Genet* 2014;**23**:117–28.
9. Zhang C, Gao S, Molascon AJ, *et al*. Quantitative proteomics reveals histone modifications in crosstalk with H3 lysine 27 methylation. *Mol Cell Proteomics* 2014;**13**:749–59.
10. Bedford MT, Richard S. Arginine methylation an emerging regulator of protein function. *Mol Cell* 2005;**18**:263–72.
11. Lee DY, Teyssier C, Strahl BD, *et al*. Role of protein methylation in regulation of transcription. *Endocr Rev* 2005;**26**:147–70.
12. Predel R, Brandt W, Kellner R, *et al*. Post-translational modifications of the insect sulfakinins: sulfation, pyroglutamate-formation and O-methylation of glutamic acid. *Eur J Biochem* 1999;**263**:552–60.
13. Lapko VN, Cerny RL, Smith DL, *et al*. Modifications of human betaA1/betaA3-crystallins include S-methylation, glutathiolation, and truncation. *Protein Sci* 2005;**14**:45–54.
14. Guo A, Gu H, Zhou J, *et al*. Immunoaffinity enrichment and mass spectrometry analysis of protein methylation. *Mol Cell Proteomics* 2014;**13**:372–87.
15. Kubicek S, O'Sullivan RJ, August EM, *et al*. Reversal of H3K9me2 by a small-molecule inhibitor for the G9a histone methyltransferas. *Mol Cell* 2007;**25**:473–81.
16. Ding J, Li T, Wang X, *et al*. The Histone H3 Methyltransferase G9A epigenetically activates the Serine-Glycine synthesis pathway to sustain Cancer cell survival and proliferation. *Cell Metab* 2013;**18**:896–907.
17. Li J, Hart RP, Mallimo EM, *et al*. EZH2-mediated H3K27 trimethylation mediates neurodegeneration in ataxia-telangiectasia. *Nat Neurosci* 2013;**16**:1745–53.
18. Schuettengruber B, Ganapathi M, Leblanc B, *et al*. Functional anatomy of polycomb and trithorax chromatin landscapes in Drosophila embryos. *PLoS Biol* 2009;**7**:e13.
19. Han HS, Jung CY, Yoon YS, *et al*. Arginine methylation of CRTC2 is critical in the transcriptional control of hepatic glucose metabolism. *Sci Signal* 2014;**7**:ra19.
20. Yan F, Alinari L, Lustberg ME, *et al*. Genetic validation of the protein arginine methyltransferase PRMT5 as a candidate therapeutic target in glioblastoma. *Cancer Res* 2014;**74**:1752–65.
21. Shao J, Xu D, Tsai SN, *et al*. Computational identification of protein methylation sites through bi-profile Bayes feature extraction. *PLoS One* 2009;**4**:e4920.
22. Shien DM, Lee TY, Chang WC, *et al*. Incorporating structural characteristics for identification of protein methylation sites. *J Comput Chem* 2009;**30**:1532–43.
23. Shi SP, Qiu JD, Sun XY, *et al*. PLMLA: prediction of lysine methylation and lysine acetylation by combining multiple features. *Mol Biosyst* 2012;**8**:1520–7.
24. Shi SP, Qiu JD, Sun XY, *et al*. PMeS: prediction of methylation sites based on enhanced feature encoding scheme. *PLoS One* 2012;**7**:e38772.
25. Xue Y, Liu Z, Gao X, *et al*. GPS-SNO: computational prediction of protein S-nitrosylation sites with a modified GPS algorithm. *PLoS One* 2010;**5**:e11290.
26. UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res* 2015;**43**:D204–12.
27. Liu Z, Ma Q, Cao J, *et al*. GPS-PUP: computational prediction of pupylation sites in prokaryotic proteins. *Mol Biosyst* 2011;**7**:2737–40.
28. Kirkpatrick S, Gelatt CD, Jr, Vecchi MP. Optimization by simulated annealing. *Science* 1983;**220**:671–80.
29. Barrell D, Dimmer E, Huntley RP, *et al*. The GOA database in 2009–an integrated Gene Ontology Annotation resource. *Nucleic Acids Res* 2009;**37**:D396–403.
30. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;**28**:27–30.
31. Destiny Z, Csizmok V, Tompa P, *et al*. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* 2005;**21**:3433–4.
32. Petersen B, Petersen TN, Andersen P, *et al*. A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC Struct Biol* 2009;**9**:51.
33. Wei H, Mundade R, Lange KC, *et al*. Protein arginine methylation of non-histone proteins and its role in diseases. *Cell Cycle* 2014;**13**:32–41.
34. Wahren-Herlenius M, Dorner T. Immunopathogenic mechanisms of systemic autoimmune disease. *Lancet* 2013;**382**:819–31.
35. van Bavel CC, Dieker JW, Kroeze Y, *et al*. Apoptosis-induced histone H3 methylation is targeted by autoantibodies in systemic lupus erythematosus. *Ann Rheum Dis* 2011;**70**:201–7.
36. Zhao M, Wu X, Zhang Q, *et al*. RFX1 regulates CD70 and CD11a expression in lupus T cells by recruiting the histone methyltransferase SUV39H1. *Arthritis Res Ther* 2010;**12**:R227.
37. O'Shea JP, Chou MF, Quader SA *et al*. pLogo: a probabilistic approach to visualizing sequence motifs. *Nat Methods* 2013;**10**:1211-12.
38. Thandapani P, O'Connor TR, Bailey TL, *et al*. Defining the RGG/RG motif. *Mol Cell* 2013;**50**:613–23.

39. Gao J, Xu D. The Musite open-source framework for phosphorylation-site prediction. *BMC Bioinformatics* 2010;**11(Suppl. 12)**:S9.
40. Gao J, Thelen JJ, Dunker AK, *et al*. Musite, a tool for global prediction of general and kinase-specific phosphorylation sites. *Mol Cell Proteomics* 2010;**9**:2586–600.
41. Yao Q, Gao J, Bollinger C, *et al*. Predicting and analyzing protein phosphorylation sites in plants using musite. *Front Plant Sci* 2012;**3**:186.
42. Gong H, Liu X, Wu J, *et al*. Data construction for phosphorylation site prediction. *Brief Bioinform* 2014;**15**: 839–55.
43. Xue Y, Ren J, Gao X, *et al*. GPS 2.0, a tool to predict kinase-specific phosphorylation sites in hierarchy. *Mol Cell Proteomics* 2008;**7**:1598–608.
44. Petrossian TC, Clarke SG. Uncovering the human methyltransferasome. *Mol Cell Proteomics* 2011;**10**:M110 000976.