PLOS ONE

RESEARCH ARTICLE

PRmePRed: A protein arginine methylation prediction tool

Pawan Kumar, Joseph Joy, Ashutosh Pandey, Dinesh Gupta*

Translational Bioinformatics Group, ICGEB, New Delhi, India

* dinesh@icgeb.res.in

Abstract

Protein methylation is an important Post-Translational Modification (PTMs) of proteins. Arginine methylation carries out and regulates several important biological functions, including gene regulation and signal transduction. Experimental identification of arginine methylation site is a daunting task as it is costly as well as time and labour intensive. Hence reliable prediction tools play an important task in rapid screening and identification of possible methylation sites in proteomes. Our preliminary assessment using the available prediction methods on collected data yielded unimpressive results. This motivated us to perform a comprehensive data analysis and appraisal of features relevant in the context of biological significance, that led to the development of a prediction tool PRmePRed with better performance. The PRmePRed perform reasonably well with an accuracy of 84.10%, 82.38% sensitivity, 83.77% specificity, and Matthew's correlation coefficient of 66.20% in 10-fold cross-validation. PRmePRed is freely available at http://bioinfo.icgeb.res.in/PRmePRed/

Citation: Kumar P, Joy J, Pandey A, Gupta D (2017) PRmePRed: A protein arginine methylation prediction tool. PLoS ONE 12(8): e0183318. https://doi.org/10.1371/journal.pone.0183318

Editor: Bin Liu, Harbin Institute of Technology Shenzhen Graduate School, CHINA

Received: April 11, 2017

Accepted: August 2, 2017

Published: August 15, 2017

Data Availability Statement: All relevant data are within the paper, its Supporting Information files and the PRmePRed tool URL http://bioinfo.icgeb.res.in/PRmePRed/.

Funding: This work was financially supported by the Department of Biotechnology, Govt. of India (BT/PR6963/BID/7/427/2012 and BT/BI/25/066/2012) awarded to DG.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Protein arginine methylation (PRme) is an abundant post-translational modification (PTM) which affects several major cellular processes in eukaryotes. PRme has been implicated in several diseases and to such an extent that some eukaryotic viruses can take the liberty of host arginine methylation machinery for their own benefit. Any biological question which aims to investigate the role of PRme in a protein's function, stability, localization and its interactions initiates with steps that lead to prior identification and validation of the methylation event. In this regard, large-scale proteomics, bolstered by recent advancements in PRme labeling, enrichment, and mass spectrometry (MS) techniques, have contributed significantly towards identification of experimentally verified repertoire of arginine methylated proteins. MS-based proteomics employing in vivo metabolic labeling of methyl group (Heavy- methyl SILAC) offers the best credible identification results as opposed to label-free approaches, which may be fraught poor reproducibility and discovery of artifact sites. However, apart from being expensive, metabolic labeling cannot be done for all biological samples such as intraerythrocytic *in vitro* cell culture of intracellular parasites like *P. falciparum*. Additionally, it is a tedious task to confirm each methylation site independently from the thousands of sites identified from a label-free MS experiment. Another option is to go for high throughput screening in vitro

enzyme assays that use recombinant protein arginine methyltransferases (PRMTs; enzymes which catalyze arginine methylation) and protein/peptide substrates. However, any in vitro outcome cannot be considered as a natural event unless supported by in vivo evidence. Also, currently, one cannot perform such experiments for a very large number of arginine residues in any organism, for example nearly more than half a million arginine residues are present in human proteome (consisting of 20193 reviewed proteins from UniProt database [1] and excluding their isoforms) and that too with eleven different human PRMTs (again excluding their isoforms). Another limitation with any experiment involving a biological sample is that a cell, at any given time, usually never carries all the PTMs it can possibly acquire during its life cycle. Also, the specialized cell types in a multicellular organism produce their own distinct methylation profiles. Thus, due to several technical and analytical shortcomings, one cannot capture the entire spectrum of any particular PTM present in a cell/organism. Therefore, in such situations where it is difficult to perform reliable large-scale experimental studies for global PTM identification, one can use computational biology based approaches as an alternative strategy.

A computational tool called "FindMod" [2], which utilizes peptide mass fingerprinting data of individual proteins to identify methylated peptides, has been successfully applied in yeast proteome. However, this strategy has limited scope because it relies on peptide mass fingerprint (PMF) data of each protein which comes from single MS analysis and not tandem MS/MS. Therefore, reliability of assigned methylation sites is limited to only the peptides with no other PTM except methylation, and which only possess a single arginine and not any other amino acid capable of undergoing methylation (e.g. lysine) in their sequences. Another approach would be to find particular properties specific to methylation and use them to computationally identify potential PTM sites in whole proteomes. For example, one can employ a homology-based sequence search for evolutionarily conserved methylated sites present in evolutionarily conserved protein and domains. Histone proteins are highly conserved proteins in eukaryote kingdom; therefore, any characterized methylated arginine site in histone from one organism will most likely be methylated in other eukaryotes. Likewise, motif-based search can be employed if conserved motifs are known in the case of arginine methylation. In case of mammals particularly, it has been observed that several methylated sites lay in either glycine arginine-rich (GAR) or, arginine or proline-rich stretches. Unfortunately, there are no well-defined universal motifs in the case of arginine methylation. Hence, in such cases, machine learning based prediction models fits the choice of a universal method that can provide quick probing of large evolutionarily divergent proteomes to identify potential methylation sites. Consequently, fourteen machine-learning studies for prediction of arginine-methylated sites have been reported till date.

The first prediction tools developed by Daily et al. [3] and Shien et al. [4] introduced most of the key features that formed the backbone of the future methods. Subsequent tools focused more towards refinement of feature encoding, extraction and selection methods; resolving data imbalance and adoption of different classification approaches. The collection of arginine methylated sites employed by all the reported prediction based studies including the most recent ones, was restricted to few hundreds of methylation sites (about 200) which mostly were acquired from the UniProt database. A major surge in repertoire of identified arginine methylated sites came only post 2012 owing to several large-scale proteomic studies; however, these sites are yet to be incorporated into UniProt. Hence, we generated a database of the PRme data, which includes more than five thousand unique methylation sites. Of the 15 reported studies, only six provided access to user-friendly web server applications, whereas few others offer downloadable models, which unfortunately, we were unable to operate upon. Our preliminary assessment using each of the web server prediction application on our

collected data yielded unimpressive results. This motivated us to perform our own comprehensive data analysis, appraisal of feature relevance in the context of biological significance (similar to Daily et al.) that led to the development of a prediction tool with better performance than the rest. We have also tried to offer an in-depth insight into the current problems faced in development PRme prediction methods, and possible areas of improvement.

Machine learning is a branch of artificial intelligence which has been successfully used for providing solutions to classification problems related to biological datasets. Amongst several machine learning algorithms, support vector machine (SVM), artificial neural networks (ANN), decision trees random forest (RF) and LibD3C are have been effectively used in bioinformatics [5–9]. Most of the available arginine methylation site prediction methods are based on SVMs, using different amino acid based features and feature selection methods. The training model developed in the study was trained on different machine learning algorithms for comparison and selection of the best training model for PRmePRed server.

Methods

Datasets for classifier generation

We collected experimentally verified *in vivo* methylated arginine sites from literature along with those reported in UniProt database (release 2015_06). Search terms like "arginine", "methylation", "methylation sites", were used for database and literature searches. Peptides/proteins mentioned in the relevant publications (PubMed search performed in June-December 2015) were included in the study dataset only after close scrutiny. We did not consider any *in vitro* reported methylated sites with no credible evidence of *in vivo* existence. We removed sites/proteins with ambiguities such as those containing nonstandard amino acids, site mismatches, very small protein fragments (less than 30 aa) and obsolete protein entries. The extracted dataset contains 6754 methylation sites from 2077 protein sequences. We did not include any methylation sites from PhosphoSitePlus database [10], since it did not provide the exact experimental source and other supporting information for verifying PTM evidence. However, majority of our methylation data did match with the ones they reported to have extracted from the literature.

It is assumed that local environment around methylated arginine, dictated by adjacent flanking residues, plays a major role in substrate selectivity and catalysis by PRMTs. These assumptions arise from the observations in which PRMT active site and certain substrate features complement each other, though not always. For instance, in one substrate, positive flanking residues were shown to affect substrate binding and catalysis by PRMT active site [11]. This is supported by the fact that the surface surrounding active site in few PRMTs have grooves that are acidic in nature. Additionally, many of the known methylated arginine sites hail from either glycine-arginine rich (GAR) or arginine-rich and proline/serine-rich regions, which favor arginine methylation. In order to assess the role of flanking residues, we generated symmetric peptide datasets of varying window lengths (7, 11, 15, 19, 23, 27, 31 and 35) all of which were centered on methylated arginine. Since we adopted position specific feature encoding for model building, therefore it was necessary to fill the ends of peptides which lacked symmetry with arbitrary "X" residue that has been the generally accepted norm in some previous prediction classifiers as well [12].

We followed the conventional practice of generating a negative set from those sites which are not reported to be methylated in the methylated proteins. Briefly, we first created an unlabeled class of all the arginine sites, which are not methylated from the respective methylated proteins. We termed the set as unlabeled because they may contain potential sites, which could be methylated but has not been established yet. Using CD-HIT-2d [13] with 40% identity cut-

off, we created a negative set from this unlabeled set by removing sequences which were similar to positive set.

There are chances that data will contain highly similar peptide sequences (since 2/3 of data belongs to human and mouse proteome, and also multiple adjacently placed arginine residues are methylated in sequences which are arginine rich such as those hailing from GAR peptides). Since most of our features are calculated position wise thus to reduce any biases especially during feature assessment with training set, we removed similar sequences from both positive and pseudo-negative sets using CD-HIT with 40% identity cut-off. We found that the pseudo-negative sets of window lengths 7, 11 and 15 were far lower than positive set and thus excluded from the model-building task. Dataset information (after CD-HIT) of different residues window length, chosen for model training are given in the supporting information S1 Table.

For each window length, positive dataset was split randomly into a training set and test set in the ratio of 4:1. We also split negative dataset into training and test set (size of the negative test set equal to positive test set). For window length 19 onward we had a larger proportion of negative training set with respect to a positive training set. Thus to overcome class imbalance issue we opted for under-sampling and created equal subsets of negative training set in 1:1 ratio with a positive training set by random sampling. For computational timesaving, we restricted the size of negative training subsets to 5 for each window length. During the course of our work, we accumulated more instances of arginine-methylated proteins from recent studies and separately prepared an independent dataset for final evaluation and comparison.

Feature collection, encoding, and evaluation

An extensive literature survey implicated PRme with the amino acid composition; physico-chemical properties such as positive charge, hydrophilicity, isoelectric point; and structural properties including ASA and disorder. We finally collected the following features:

Atchley factors [14]. Since the distinct physico-chemical properties of amino acids reported in AAIndex [15] were too large to computationally handle in our analysis, therefore instead we relied on the reduced and transformed AAIndex feature subsets represented by the five Atchley factors (AF), namely, AF-I, AF-II, AF-III, AF-IV, and AF-V. Factor I represents residue polarity, hydrophobicity, and surface accessibility. Factor II captures secondary structure information whereas factor III relates to molecular size or volume. Factor IV reflects relative amino acid composition in various proteins and codon diversity. Factor V refers to electrostatic charge with high coefficients on isoelectric point and net charge. The PSE-in-One [16] features for protein are similar to AAIndex features, hence we did not consider them separately.

AA frequency. We generated amino acid composition features from position-wise amino acid frequency of each amino acid from the non-redundant positive peptide list. The values were normalized and a table of 21x n was created for each window, where n denotes window length.

ASA. ASA has been used as a feature by previous tools such as MASA [4] and PMeS [12], using RVP-NET for prediction of ASA values for amino acid residues, based on protein sequences. To evaluate the margin of error in these predictions, we compared the predicted values versus actual values calculated by NACCESS from PDB structures. For the sake of convenience, we considered only the methylated arginine sites from those protein sequences, which are represented by experimentally, solved PDB structures with greater than 80% sequence coverage and 100% identity.

Disorder [17]. Predicted protein intrinsic disorder was calculated for full length methylated protein sequences, using VSL2b standalone package. The output file for each protein

sequence contained disorder scores for each residue. The predicted results of methylated proteins were compared with their respective experimental disorder information available in the DisProt database [18].

Hydrophobicity [19]. Hydrophobicity values for amino acids were obtained from Kyte and Doolittle hydrophobicity scales. The grand average of hydropathy (GRAVY) for a given peptide instance was calculated as sum average of hydrophobicity value of individual amino acids in the peptide.

Van der Waal's volume. Van der Waal's volume for each residue was calculated from scale reported by Darby and Creighton [20]. The average Van der Waals volume for each peptide was calculated as sum average of individual VDWV values.

Total charge and isoelectric point pI. Total charge and isoelectric point for each peptide were calculated using pyteomics, a python package [21].

For a given peptide instance, the following features Atchley factors, ASA, disorder, hydrophobicity, van der waal's volume and AA frequency were encoded for individual residues in position wise manner whereas average VDWV, GRAVY, total charge, and pI were calculated for the entire peptide. Thus in total, we obtained feature sizes of 194, 234, 274, 314 and 354 for window lengths 19, 23, 27, 31 and 35 respectively.

Feature relevance assessment was performed by InfoGain (Information Gain) analysis on training sets in WEKA [22]. InfoGain selects the feature that has the best potential to separate the instances into individual classes. The value of InfoGain is lies between 0 and 1. A feature with a high information gain is said to be "relevant". InfoGain is evaluated independently for each feature and the features with the top scores are selected as the relevant features.

The irrelevant features with a score of 0 were removed from total feature set and thus did not form part of feature selection. The removed features were indeed irrelevant as most of them belonged to a zeroth position which corresponded with central arginine thus corroborating that InfoGain analysis was correct. After removing irrelevant features (having value 0) from the total feature set, features set rearranged on the basis of relevance.

Classifier

Support Vector Machines (SVMs), developed by Vladimir Vapnik and co-workers [23], is a useful technique for data classification. SVM is rigorously based on statistical learning theory. For linearly separable problems SVM employs a maximum margin hyper-plane for separating examples belonging to two different classes and for non-linearly separable problems, SVM first transforms the data into a higher dimensional feature space and subsequently employs a maximum margin linear hyper plane. There are four basic kernels that can be used in SVM.

$$\text{Linear}: \ K(x_i, x_j) = x^T{}_i x_j.$$

$$\text{Polynomial}: \ K(x_i, x_j) = (\gamma x_i{}^T x_j + r)^d, \ \gamma > 0.$$

$$\text{Radial basis function (RBF)}: \ K(x_i, x_j) = \exp(-\gamma ||x_i - x_j||^2), \ \gamma > 0.$$

$$\text{Sigmoid}: \ K(x_i, x_j) = \tanh(\gamma x_i{}^T x_j + r).$$

Where, $K(xi, xj) \equiv \varphi(x_i)^T \varphi(x_j)$ that is, the kernel function, represents a dot product of input data points mapped into the higher dimensional feature space by transformation.

Here, γ, r, and d are kernel parameters.

The RBF is by far the most popular choice of kernel types used in Support Vector Machines. This is mainly because RBF kernel non-linearly maps samples into a higher dimensional space so it, unlike the linear kernel, can handle the case when the relation between class labels and attributes is nonlinear.

LIBSVM (A Library for Support Vector Machines) [24] is currently one of the most widely used SVM software. A typical use of LIBSVM involves two steps: first, training a data set to obtain a model and second, using the model to predict information of a testing data set.

Here we used C-SVC from LIBSVM package with RBF kernel to build the classifier. C (cost) and g (gamma) optimized by grid search strategy using 10 fold cross validation with AUCROC as an evaluation function.

Major evaluation parameters: Accuracy (Acc), Sensitivity (Sn), Specificity (Sp) and Matthews Correlation Coefficient (MCC).

$$Sn = \frac{TP}{\mathrm{TP} + \mathrm{FN}}$$

$$Sp = \frac{TN}{\mathrm{TN} + \mathrm{FP}}$$

$$Acc = \frac{TP + TN}{\mathrm{TP} + \mathrm{FN} + \mathrm{TN} + \mathrm{FP}}$$

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

where, TP represents the number of correctly predicted methylated arginine sites by the SVM-predictor, TN represents the number of correctly predicted arginine non-methylated sites, FP represents the incorrectly predicted methylated arginine sites, and FN represents the incorrectly predicted arginine non-methylated sites. Further description of the terms is available elsewhere [25].

Results and discussion

Selection of feature subset and window size

Incremental feature selection was performed with various feature subsets in an incremental fashion for each window length. The evaluation parameters were compared with training data test and test set.

For the arginine methylation prediction problem, best accuracy achieved by window length 19 with a subset of 150 features (Fig 1A), best sensitivity achieved by window length 19 with subset of 100 features (Fig 1B), best specificity achieved by window length 35 with subset of 100 features (Fig 1C) and the best MCC achieved by window length 19 with a subset of 100 features (Fig 1D). Considering all the evaluation parameters (Acc, Sn, Sp, and MCC), window length 19 with subsets of 100 features selected by information gain perform better (Table 1). The details of predictive performance of model trained with different features subset for window lengths 19, 23, 27, 31 and 35 may be obtained from supporting information S2–S6 Tables, respectively.

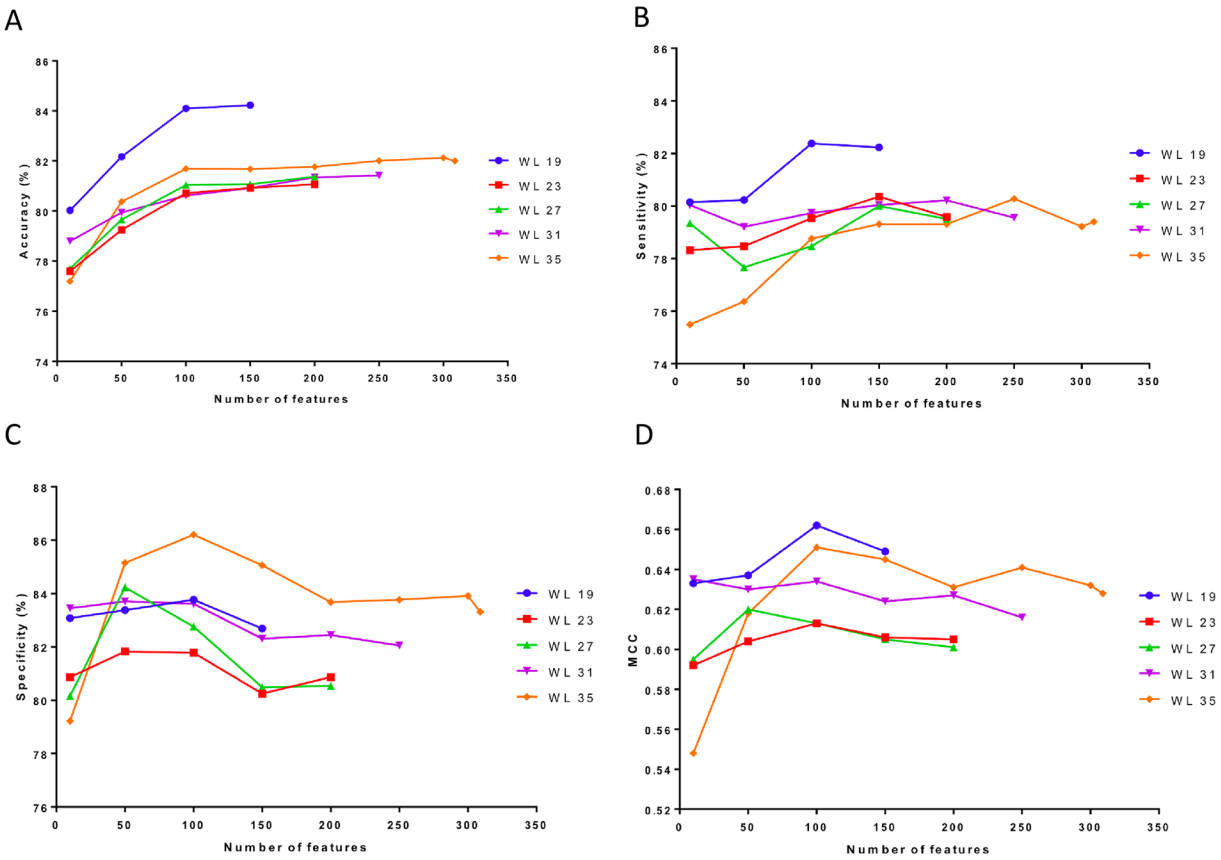

Fig 1. The relationship between different evaluation parameters and feature subsets. A) The relationship between the Accuracy and number of features. B) The relationship between the Sensitivity and number of features. C) The relationship between the Specificity and number of features. D) The relationship between the MCC and number of features.

https://doi.org/10.1371/journal.pone.0183318.g001

Comparisons with existing methods

To further evaluate the prediction performance of the PRmePRed impartially, we made comparisons with other existing PRme prediction tools. Generally, to perform a comparison between distinct machine learning prediction methods, either cross-validation experiment or an independent dataset test is used. For cross-validation experiment, identical training dataset is required. As described in the Methods section PRmePRed training dataset is not similar to previous methods. Therefore, a comparison between distinct machine learning prediction methods through cross-validation performance is irrelevant. Here, we used independent

Table 1. Comparisons with best models of different window lengths.

Window Length (Features subset)	MCC	Accuracy	Sensitivity	Specificity
WL_19 (100)	0.662	84.10%	82.38%	83.77%
WL_23 (150)	0.606	80.93%	80.36%	80.25%
WL_27 (150)	0.605	81.07%	80.00%	80.49%
WL_31 (200)	0.629	81.35%	80.22%	82.45%
WL_35 (250)	0.641	82.01%	80.28%	83.77%

https://doi.org/10.1371/journal.pone.0183318.t001

Table 2. Comparison of PRmePRed with other prediction methods.

Method (yr. developed)	Algorithm	MCC	Accuracy	Sensitivity	Specificity
MeMo (Chen et al. 2006)[26]	SVM	0.462	0.6839	0.3811	0.987
MASA (Shien et al. 2009)[4]	SVM	0.411	0.6503	0.3095	0.991
BPB-PPMS (Shao et al. 2009)[27]	SVM	0.253	0.5601	0.1202	1.000
PMeS (Shi et al. 2012)[12]	SVM	0.159	0.5756	0.4253	0.726
iMethyl-PseAAC (2014)[28]	SVM	0.302	0.5866	0.1768	0.997
PSSMe (Wen et al. 2016) [29]	SVM	0.444	0.7162	0.6003	0.832
MePred-RF (Wei et al. 2017) [30]	RF	0.462	0.6908	0.4095	0.972
PRmePRed (2017)	SVM	0.737	0.8683	0.8709	0.866

https://doi.org/10.1371/journal.pone.0183318.t002

dataset test to evaluate the performance of PRmePRed to compare it with other PRme prediction tools.

There are three major differences between our approach and previously reported methods. First, we used experimentally verified in vivo methylated arginine sites. Second to avoid any biases, we used CD-HIT (40%) on peptides rather than removing redundancy in protein sequences. Finally, rather than defining a broad range of parameters to describe the peptides, we used most relevant parameter for the methylation process. We used independent dataset to evaluate the performance of PRmePRed with comparison to other prediction tools (Table 2).

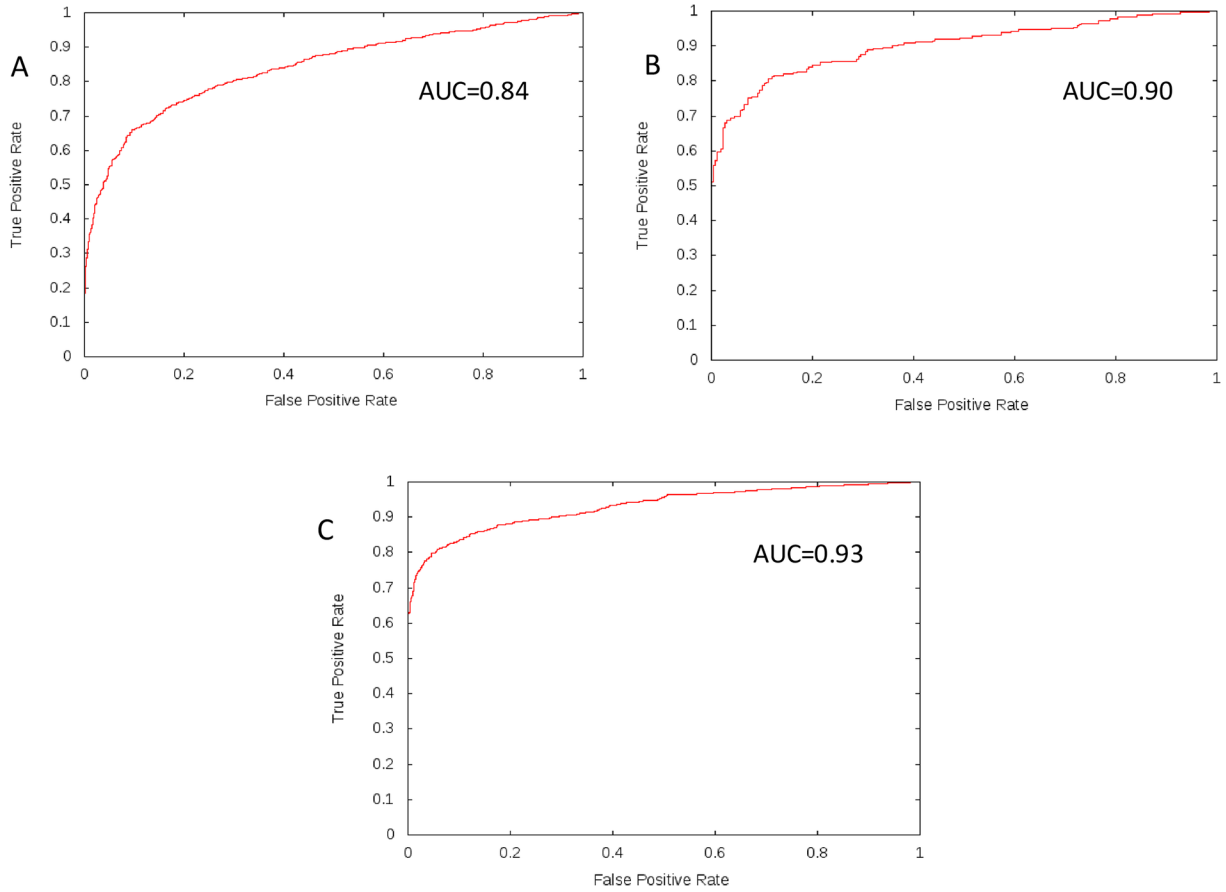

Fig 2. ROC curve for SVM classifier with different datasets. A) ROC curve for SVM classifier with training set. B) ROC curve for SVM classifier with test set. C) ROC curve for SVM classifier with independent set.

https://doi.org/10.1371/journal.pone.0183318.g002

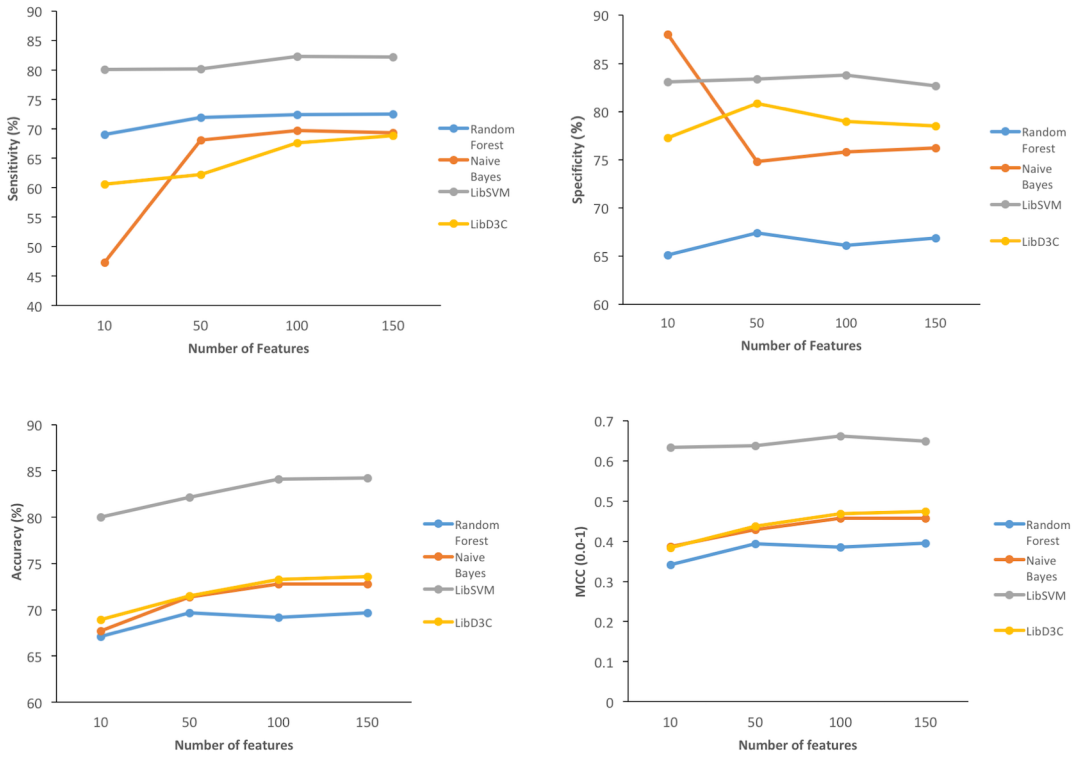

Fig 3. Comparisons with other classifiers based on evaluation parameters.

https://doi.org/10.1371/journal.pone.0183318.g003

ROC curve

ROC curve is graphical display true positive rate (sensitivity) on y-axis and false positive rate (1 –specificity) on x-axis for varying cut-off points of test values. The area under the curve (AUC) is an effective and combined measure of sensitivity and specificity for assessing the inherent validity of a classification test. Maximum AUC = 1 and it means classification test is perfect in differentiating positive with negative class. This implies both sensitivity and specificity are one and both errors—false positive and false negative—are zero. This can happen when the distribution of methylated and non-methylated test values do not overlap. This is extremely unlikely to happen in practice. ROC curve of training model represent in Fig 2A (AUC = 0.8411), ROC curve of test data on training model represent in Fig 2B (AUC = 0.9000) and ROC curve of independent data on training model represent in Fig 2C (AUC = 0.9299). A result of $0.8 <= AUC <= 0.95$ represent excellent ability to discriminate between of methylated and non-methylated arginine sites.

We evaluated SVM, RF, Naïve Bayes and LibD3C algorithms for PRmePRed development, and found that SVM performs comparatively better for the same set of features (See Fig 3).

Conclusion

We have developed an arginine methylation predictor based on sequence and structure derived features, using SVMs. Dataset used to build the predictor is not biased and has experimentally verified entries only. Moreover, the PRmePRed shows better performance as compared with existing tools (Table 2). We believe that PRmePRed is a useful, reliable and rapid prediction tool for arginine methylation sites in proteins.

Supporting information

S1 Table. Dataset information of different residues window length.
(DOC)

S2 Table. The predictive performance of model trained with different features subset for window length 19.
(DOC)

S3 Table. The predictive performance of model trained with different features subset for window length 23.
(DOC)

S4 Table. The predictive performance of model trained with different features subset for window length 27.
(DOC)

S5 Table. The predictive performance of model trained with different features subset for window length 31.
(DOC)

S6 Table. The predictive performance of model trained with different features subset for window length 35.
(DOC)

Acknowledgments

We acknowledge Mr. Rajan Pandey for help and suggestions regarding manuscript figures.

Author Contributions

Conceptualization: Joseph Joy, Dinesh Gupta.

Data curation: Joseph Joy.

Formal analysis: Pawan Kumar, Ashutosh Pandey.

Funding acquisition: Dinesh Gupta.

Investigation: Pawan Kumar.

Methodology: Pawan Kumar, Ashutosh Pandey, Dinesh Gupta.

Project administration: Dinesh Gupta.

Software: Pawan Kumar, Ashutosh Pandey.

Supervision: Dinesh Gupta.

Validation: Pawan Kumar.

Visualization: Pawan Kumar, Joseph Joy.

Writing – original draft: Joseph Joy, Dinesh Gupta.

Writing – review & editing: Dinesh Gupta.

References

1. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, et al. UniProt: the Universal Protein knowledgebase. Nucleic acids research. 2004; 32(Database issue):D115–9. https://doi.org/10.1093/nar/gkh131 PMID: 14681372.

2. Wilkins MR, Gasteiger E, Gooley AA, Herbert BR, Molloy MP, Binz PA, et al. High-throughput mass spectrometric discovery of protein post-translational modifications. J Mol Biol. 1999; 289(3):645–57. https://doi.org/10.1006/jmbi.1999.2794 PMID: 10356335.
3. Daily KM, Radivojac P, Dunker AK, editors. Intrinsic disorder and prote in modifications: building an SVM predictor for methylation. 2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology; 2005: IEEE.
4. Shien DM, Lee TY, Chang WC, Hsu JBK, Horng JT, Hsu PC, et al. Incorporating structural characteristics for identification of protein methylation sites. Journal of computational chemistry. 2009; 30(9):1532–43. https://doi.org/10.1002/jcc.21232 PMID: 19263424
5. Chen J, Guo M, Wang X, Liu B. A comprehensive review and comparison of different computational methods for protein remote homology detection. Brief Bioinform. 2016. Epub 2016/11/25. https://doi.org/10.1093/bib/bbw108 PMID: 27881430.
6. Liu B, Zhang D, Xu R, Xu J, Wang X, Chen Q, et al. Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection. Bioinformatics. 2014; 30(4):472–9. Epub 2013/12/10. https://doi.org/10.1093/bioinformatics/btt709 PMID: 24318998.
7. Jagga Z, Gupta D. Classification models for clear cell renal carcinoma stage progression, based on tumor RNAseq expression trained supervised machine learning algorithms. BMC Proc. 2014; 8(Suppl 6 Proceedings of the Great Lakes Bioinformatics Confer):S2. Epub 2014/11/07. https://doi.org/10.1186/1753-6561-8-S6-S2 PMID: 25374611.
8. Kalita MK, Nandal UK, Pattnaik A, Sivalingam A, Ramasamy G, Kumar M, et al. CyclinPred: a SVM-based method for predicting cyclin protein sequences. PLoS One. 2008; 3(7):e2605. Epub 2008/07/04. https://doi.org/10.1371/journal.pone.0002605 PMID: 18596929.
9. Zou Q, Wang Z, Guan X, Liu B, Wu Y, Lin Z. An approach for identifying cytokines based on a novel ensemble classifier. Biomed Res Int. 2013; 2013:686090. Epub 2013/09/13. https://doi.org/10.1155/2013/686090 PMID: 24027761.
10. Hornbeck PV, Kornhauser JM, Tkachev S, Zhang B, Skrzypek E, Murray B, et al. PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. Nucleic acids research. 2011:gkr1122.
11. Osborne TC, Obianyo O, Zhang X, Cheng X, Thompson PR. Protein arginine methyltransferase 1: positively charged residues in substrate peptides distal to the site of methylation are important for substrate binding and catalysis. Biochemistry. 2007; 46(46):13370–81. https://doi.org/10.1021/bi701558t PMID: 17960915
12. Shi S-P, Qiu J-D, Sun X-Y, Suo S-B, Huang S-Y, Liang R-P. PMeS: prediction of methylation sites based on enhanced feature encoding scheme. PloS one. 2012; 7(6):e38772. https://doi.org/10.1371/journal.pone.0038772 PMID: 22719939
13. Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics. 2010; 26(5):680–2. https://doi.org/10.1093/bioinformatics/btq003 PMID: 20053844
14. Atchley WR, Zhao J, Fernandes AD, Drüke T. Solving the protein sequence metric problem. Proceedings of the National Academy of Sciences of the United States of America. 2005; 102(18):6395–400. https://doi.org/10.1073/pnas.0408677102 PMID: 15851683
15. Kawashima S, Ogata H, Kanehisa M. AAindex: amino acid index database. Nucleic acids research. 1999; 27(1):368–9. PMID: 9847231
16. Liu B, Liu F, Wang X, Chen J, Fang L, Chou KC. Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. Nucleic Acids Res. 2015; 43(W1): W65–71. Epub 2015/05/11. https://doi.org/10.1093/nar/gkv458 PMID: 25958395.
17. Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z. Length-dependent prediction of protein intrinsic disorder. BMC bioinformatics. 2006; 7(1):1.
18. Vucetic S, Obradovic Z, Vacic V, Radivojac P, Peng K, Iakoucheva LM, et al. DisProt: a database of protein disorder. Bioinformatics. 2005; 21(1):137–40. Epub 2004/08/18. https://doi.org/10.1093/bioinformatics/bth476 PMID: 15310560.
19. Cid H, Bunster M, Canales M, Gazitúa F. Hydrophobicity and structural classes in proteins. Protein engineering. 1992; 5(5):373–5. PMID: 1518784
20. Darby NJ. Protein structure / Darby N.J. and Creighton T.E.. Creighton TE, editor. Oxford; New York: IRL Press at Oxford University Press; 1993.
21. Goloborodko AA, Levitsky LI, Ivanov MV, Gorshkov MV. Pyteomics—a Python framework for exploratory data analysis and rapid software prototyping in proteomics. Journal of the American Society for Mass Spectrometry. 2013; 24(2):301–4. Epub 2013/01/08. https://doi.org/10.1007/s13361-012-0516-6 PMID: 23292976.

22. Frank E, Hall M, Trigg L, Holmes G, Witten IH. Data mining in bioinformatics using Weka. Bioinformatics. 2004; 20(15):2479–81. https://doi.org/10.1093/bioinformatics/bth261 PMID: 15073010

23. Vapnik V, Cortes C. Support Vector Networks, machine learning 20, 273–297. Kunwer Acedemic Publisher; 1995.

24. Chang C-C, Lin C-J. LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST). 2011; 2(3):27.

25. Chen J, Liu H, Yang J, Chou KC. Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. Amino Acids. 2007; 33(3):423–8. Epub 2007/01/26. https://doi.org/10.1007/s00726-006-0485-9 PMID: 17252308.

26. Chen H, Xue Y, Huang N, Yao X, Sun Z. MeMo: a web tool for prediction of protein methylation modifications. Nucleic acids research. 2006; 34(suppl 2):W249–W53.

27. Shao J, Xu D, Tsai S-N, Wang Y, Ngai S-M. Computational identification of protein methylation sites through bi-profile Bayes feature extraction. PloS one. 2009; 4(3):e4920. https://doi.org/10.1371/journal.pone.0004920 PMID: 19290060

28. Qiu W-R, Xiao X, Lin W-Z, Chou K-C. iMethyl-PseAAC: Identification of protein methylation sites via a pseudo amino acid composition approach. BioMed research international. 2014;2014.

29. Wen P-P, Shi S-P, Xu H-D, Wang L-N, Qiu J-D. Accurate in silico prediction of species-specific methylation sites based on information gain feature optimization. Bioinformatics. 2016:btw377.

30. Wei L, Xing P, Shi G, Ji ZL, Zou Q. Fast prediction of protein methylation sites using a sequence-based feature selection technique. IEEE/ACM Trans Comput Biol Bioinform. 2017. Epub 2017/02/22. https://doi.org/10.1109/TCBB.2017.2670558 PMID: 28222000.