Bioinformatics, 32(20), 2016, 3107–3115
doi: 10.1093/bioinformatics/btw377
Advance Access Publication Date: 26 June 2016
Original Paper

Sequence analysis

Accurate *in silico* prediction of species-specific methylation sites based on information gain feature optimization

Ping-Ping Wen[1], Shao-Ping Shi[1], Hao-Dong Xu[1], Li-Na Wang[1] and Jian-Ding Qiu[1,2,*]

[1]Department of Chemistry, Department of Mathematics, Nanchang University, Nanchang 330031, China and [2]Department of Materials and Chemical Engineering, Pingxiang University, Pingxiang 337055, China

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on March 27, 2016; revised on May 30, 2016; accepted on June 13, 2016

Abstract

As one of the most important reversible types of post-translational modification, protein methylation catalyzed by methyltransferases carries many pivotal biological functions as well as many essential biological processes. Identification of methylation sites is prerequisite for decoding methylation regulatory networks in living cells and understanding their physiological roles. Experimental methods are limitations of labor-intensive and time-consuming. While *in silicon* approaches are cost-effective and high-throughput manner to predict potential methylation sites, but those previous predictors only have a mixed model and their prediction performances are not fully satisfactory now. Recently, with increasing availability of quantitative methylation datasets in diverse species (especially in eukaryotes), there is a growing need to develop a species-specific predictor. Here, we designed a tool named PSSMe based on information gain (IG) feature optimization method for species-specific methylation site prediction. The IG method was adopted to analyze the importance and contribution of each feature, then select the valuable dimension feature vectors to reconstitute a new orderly feature, which was applied to build the finally prediction model. Finally, our method improves prediction performance of accuracy about 15% comparing with single features. Furthermore, our species-specific model significantly improves the predictive performance compare with other general methylation prediction tools. Hence, our prediction results serve as useful resources to elucidate the mechanism of arginine or lysine methylation and facilitate hypothesis-driven experimental design and validation.

Availability and Implementation: The tool online service is implemented by C# language and freely available at http://bioinfo.ncu.edu.cn/PSSMe.aspx.

Contact: jdqiu@ncu.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Many post-translational modifications (PTMs) of proteins provide the proteome with structural and functional diversity, and regulate cellular plasticity and dynamics (Mann and Jensen, 2003). Protein methylation is considered one of the most common and reversible PTMs (Paik and Kim, 1967), which is an important element in many significant biological functions (Bedford and Richard, 2005; Bannister *et al.*, 2005; Paik *et al.*, 2007). Meanwhile, researchers and clinicians have proved that protein methylation and their regulatory enzymes are involved in the process of a variety of human diseases such as cancer (Aleta *et al.*, 1998; Wang *et al.*, 2009), cardiovascular disease (Yang and Bedford, 2013), multiple sclerosis (Mastronardi *et al.*, 2006), rheumatoid arthritis (Suzuki *et al.*, 2007), as well as neuroses generative disorders (Longo and Kennedy, 2006). Therefore, annotation of methylation in proteomes would certainly provide very useful information or clues for drug discovery to study and analyze the mechanisms that govern these basic epigenetic phenomena.

Identification of methylated sites is the first step towards to understanding molecular mechanism of protein methylation. Although there have many conventional experimental methods to identify protein methylation sites, such as Chip-Chip (Johnson *et al.*, 2008), mass spectrometry (MS) (Ong *et al.*, 2004) and methylation-specific antibodies (Boisvert *et al.*, 2003), these experimental approaches are limited to labor-intensive and time-consuming. While *in silicon* prediction of methylation sites is much more desirable for its convenience and fast speed, and there have been several tools published for prediction protein methylated sites. Recently, our lab (Shi *et al.*, 2015) have systematically summarized current of computational prediction methylation sites method of the past ten years. Furthermore, we have also summarized characteristic of differences among several typical tools (the detailed information in Supplementary Table S1), including MeMo (Chen *et al.*, 2006), BPB-PPMS (Shao *et al.*, 2009), MASA (Shien *et al.*, 2009), PMeS (Shi *et al.*, 2012), PLMLA (Shi *et al.*, 2012), MethK (Lee *et al.*, 2014), iMethyl-PseAAC (Qiu *et al.*, 2014) and GPS-MSP (Deng *et al.*, 2016). For instance, Plewczynski *et al.* (2005) designed the first methylation sites predictor within their AutoMotif Server using regular expression technique. Subsequently, Shao *et al.* (2009) combined Bi-profile Bayes feature extraction with support vector machine (SVM) to predict arginine and lysine methylation. Based on an enhanced feature encoding scheme (composed of the sparse property coding, normalized van der Waals volume, position weight amino and accessible surface area), our lab developed a tool PMeS (Shi *et al.*, 2012) to predict arginine and lysine methylation. Meanwhile Shi *et al.* (2012) incorporated protein sequence information, secondary structure and amino acid properties to identify methylation and acetylation of lysine residues in whole protein sequences. Lee *et al.* (2014) used amino acid composition (AAC) and accessible surface area (ASA) coalition SVM to identify lysine-methylated sites on histones and non-histone proteins. Most recently, Deng *et al.* (2016) adopted GPS 3.0 algorithm and built GPS-MSP (Methyl-group Specific Predictor) for the prediction of general or type-specific methyllysine and methylarginine residues in proteins.

Although each of the aforementioned methods possesses its own merit and did play roles in stimulating the development of this area, they still have limitations when applied to whole proteomes and need improvement from one or more of the following aspects: (i) The benchmark dataset used by the previous investigators needs to updated by incorporating new experiment-confirmed data, because methylation data updated very quickly in recent years. For example, Lee *et al.* (2014) created MethK model just with 1306 methyllysine sites at 2013, whereas our work have found 3180 lysine methylation sites in total. (ii) The features used by the previous models were single and simple, which it is casing the performance of the several aforementioned predictors not fully satisfactory, so there is still room to improve the predictive accuracy (the detailed information see comparison with other general prediction tools in results and discussion). (iii) The third limitation is that the methylation specific classification (species or methyltransferase family) was not discussed in previous methods. The majority of existing studies except GPS-MSP disregarded the differences of species and considered all species methylation sites as generic sites to build a mixed model. However, sequences or structural patterns around the methylation sites may significantly differ in different organisms, so it highlighting necessity to develop species-specific predictors. (iv) Furthermore, some of these tools were only published as a method and did not provide a user-friendly web service or the web server does not work (Xu *et al.*, 2013; Zhang *et al.*, 2013). The tools with web services are useful for user to predict methylation sites when they want to experimental design and verify whether sequence is methylated, especially for large-scale predictions.

Protein methylation site prediction has not been as widely studied as other types of PTM (such as phosphorylation and acetylation) due to lack of methylation data. With techniques rapidly increasing in the past few years, the growing number of methylation datasets was found in diverse species (especially in eukaryotes). For example, Lott *et al.* (2014) found over 800 arginine methylated proteins in an early-branching eukaryote *Trypanosoma brucei*. Li *et al.* (2014) collected 6449 methylation sites with species information in 3237 experimentally validated methylated proteins in SysPTM 2.0. Protein methylation has attracted more and more attentions, there is a growing need to develop a species-specific predictor. In this work, we developed a new tool named PSSMe (prediction species-specific methylation sites) to overcome above limitations of the existing methods, which is specifically designed to predict both species-specific and general methylation sites. Reliable and large-scale experimental methylation sites data from multiple species were collected from several sources which used to train for building the species-specific models. In particular, our model first used a well-established feature selection method IG (Huang and Berger, 2008) coupled with a two stage to distinguish the importance and valuable dimension feature vector, which it is more simple and convenient comparison with other selection and optimization methods (such as the maximum relevance minimum redundancy (mRMR) method) (Peng *et al.*, 2005), The first-stage selects five features, including amino acid compositions (AAC), K-spaced amino acid pair composition, amino acids binary encoding (BE), K Nearest Neighbors features (KNN) and amino acid physicochemical property (PSP) features. The second-stage incorporates five features and finds most valuable dimension features based on information gain (IG) method, and then reconstitute a new orderly features to build the model for methylation sites prediction. After prediction of independent test dataset, the result showed that our prediction performance overall significantly enhance than other tools.

Furthermore, we discussed the model of methylation in different species; the analysis result shows the following: (i) the methylation patterns are significantly different in different species. (ii) IG method can effectively select valuable dimension features and reduce redundant features for improving the prediction performance. (iii) Based on IG score, it suggests that KNN feature is important and makes a great contribution for prediction methylation sites. (iv) Finally, the web service of PSSMe supports continuous stringency adjustment to meet the various confidence requirements of user's cross-validation tests and provides better probability information for prediction results by comparison with other prediction tools.

2 Methods

2.1 Data collection and preprocessing

Methylation is predominantly found on lysine and arginine residues, of which arginine methylation data were composed of three species: *H.sapiens* (human), *M.musculus* (mouse) and *R.norvegicus* (rat).

And lysine methylation data were also collected from three species: *H.sapiens* (human), *M.musculus* (mouse) and the others eukaryotes (i.e. yeast, cow, sheep and so on). The data were mainly extracted from several database sources including PhosphoSite (March 1, 2015) (Hornbeck *et al.*, 2004), UniProtKB/Swiss-Prot (March 1, 2015) (Chernorudskiy *et al.*, 2007), SYSPTM (Li *et al.*, 2014), dbPTM (Lee *et al.*, 2006), CPLM databases (Liu *et al.*, 2013) as well as the relevant literatures. From the literatures in recent years, we found 287 methylarginine sites and 57 methyllysine sites which have been experimentally verified and were not collected in databases. Totally, the original dataset of methylarginine contained 3550 proteins including 7198 methylated sites, of which 3603 sites were found from human, 2931 sites from mouse and 425 sites were from rat (detailed information in Supplementary Table S2). Similarly, the methyllysine dataset contained 3180 methylated sites from 1764 lysine proteins in different species. Among them, we found 2045 methylated sites from human, 289 methylated sites from mouse, while 480 methylated sites from the other eukaryotes in all. Because methylation dataset from several sources may own a high protein sequences homology, we need cluster the protein sequences from dataset with a threshold of 30% identity by CD-HIT (Li and Godzik, 2006) to eliminate homology protein sequences. Afterward, we got the non-redundant species methylation datasets, and then randomly selected 30 arginine and lysine proteins containing methylated sites, respectively, to use as the independent data for different species model testing. The remaining 2980, 2644, 321 arginine methylation sites as well as 1741, 186, 173 lysine methylation sites of various species were used to construct our training models, respectively. Finally, numbers of methylation sites for arginine and lysine different species model training and testing are shown in Supplementary Table S3 (the detailed training methylation set in Supplementary Data S1 and Data S2).

2.2 Features extraction and optimization

As heterogeneous features used to predict PTMs sites *in silicon* are often noisy and redundant leading to an adverse impact on model training, such as decreasing performance, a time-consuming train classifiers and possibly biased model prediction, thus features optimization is very necessary. In our study, a new method IG was adopted to select and optimize multi-features, which can effectively remove redundant features and significantly improve the predictive performance. In the first-stage, we selected some features from which have been evidenced good performances for prediction PTMs sites. After the preliminary evaluation testing of SVM (SVM information in Supplementary Ep1), we selected the five features which can effectively predictive methylation sites. These features are the amino acid composition (AAC), binary encoding (BE), K-spaced amino acid pair composition (K-spaced), K Nearest Neighbors (KNN) features and physicochemical properties (PSP). The detailed feature information is following (the K-spaced and BE feature information in Supplementary Ep2):

2.2.1 AAC

Amino acid composition feature is the most popular coding method and widely used for prediction PTMs sites (such as phosphorylation and acetylation) (Suo *et al.*, 2012), which reflects protein sequences amino acid occurrence frequencies information. In this work, AAC is the fraction of each type of amino acid in a sequence fragment, we calculated amino acid occurrence frequencies in the sequence surrounding the query site (the site itself is not counted). There are 20 types of amino acids, and thus 20 frequencies are calculated, the sum of which is 1.

2.2.2 KNN

Local sequence clusters often exist around methylation site because substrate sites of same methyltransferases usually share similar patterns in local sequence fragments (Kennelly and Krebs, 1995). We used the local sequence around a possible methylation site in a query protein and extracted features from similar sequences in both the positive and negative sets by a KNN algorithm. For example, two local sequences, $s_1 = [s_1\ (\mathrm{i})]$ and $s_2 = [s_2\ (\mathrm{i})]$ define the distance Dist (s_1, s_2) between s_1 and s_2 as:

$$\mathrm{Dist}(S_1, S_2) = 1 - \frac{\sum_{i=-p}^{p} \mathrm{Sim}(S_1(i), S_2(i))}{2p+1} \tag{1}$$

$$\mathrm{Sim}(a, b) = \frac{M(a, b) - \min\{M\}}{\max\{M\} - \min\{M\}} \tag{2}$$

where p represents the length of a protein sequence; Sim, the amino acid similarity matrix, is derived from the BLOSUM62 substitution matrix (Henikoff and Henikoff, 1992); a and b are two amino acids, M is the substitution matrix and max/min{M} represent the largest/smallest number in the matrix, respectively.

2.2.3 PSP

Physicochemical property is the most intuitive feature for biochemical properties, the specificity and diversity of structure and function of proteins depend to a large extent on various properties of each of the amino acids. A large body of experimental and theoretical researches has been performed to characterize different kinds of properties of individual amino acids and represent them in terms of the numerical index (Tomii and Kanehisa, 1996). Version 9.1 of Amino Acid index database (AAindex) (Kawashima and Kanehisa, 2000) is containing total 544 amino acid indices physicochemical and biochemical properties of 20 amino acids. Furthermore, PSP has been successfully applied to predict such protein modification as phosphorylation, acetylation. Moreover, Shi *et al.* (2012) and Lee *et al.* (2014) also have shown that the feature set of PSP is one of best type of feature for predicting methylation sites. Thus, in our work, comparing the prediction accuracy of all PSP, then the top three were selected and defined as informative features for the prediction model. The detailed informative of PSP in different model is shown in Supplementary Table S4.

2.2.4 IG

The second-stage is selection and optimization aforementioned five features to find the most valuable dimension feature vectors for building the prediction model. In fact, compared to the single feature, the combination of features can reflect more protein sequence information leading to a certain improvement of prediction performance. However, the combination of complex features is highly dimensional, redundant, heterogeneous and noisy, which would lead to a time-consuming practice to train classifiers and possibly biased model prediction. In this stage, we analyzed all dimension feature vectors of five features based on IG method, then distinguished the importance and contribution features. We selected this valuable dimension feature vectors to reconstitute a new orderly features which employed for methylation sites prediction. The IG method could effectively select valuable dimension feature vectors and significantly improve the predictive performance. The detailed theory of IG method is following.

Information gain measures the decrease in entropy when a given feature is used to group values of another (class) feature. The entropy of a feature X is defined as

$$H(X) = -\sum_{i} P(x_i)\log_2(P(x_i)) \tag{3}$$

where $[x_i]$ is a set of values of X and $P\ (x_i)$ is the prior probability of x_i. The conditional entropy of X, given another feature Y is defined as

$$H(X|Y) = -\sum_{j} P(y_i)\sum_{i} P(x_i|y_i)\log_2(P(x_i|y_i)) \tag{4}$$

where $P\ (x_i \mid y_j)$ is the posterior probability of X given the value y_i of Y. The amount by which the entropy of X decreases reflects additional information about X provided by Y and is called information gain

$$IG(X|Y) = H(X) - H(X|Y) \tag{5}$$

According to the above theory, we can draw the conclusion that the larger the value of IG, the greater the impact of the corresponding features vector for prediction methylation site.

3 Results and discussion

3.1 Analysis of sequence information results

First, we used the sample logo to analyze sequence structural information to determine statistically significant residues surrounding methylation sites. In order to identify distinct patterns or conserved sequence motifs between methylation and non-methylation proteins, we applied Two Sample Logo tool (Vacic *et al.*, 2006) to generate sequence logos (Fig. 1) for the arginine/lysine methylation sequence with three different species based on the curated datasets. From the sequence logo, the amino acid residues that significantly enriched and depleted around arginine/lysine methylation sequence were easily identified. We find methylation sequence very similar in three different species, where ‘G’ (glycine) is primary acid residue and tends to appear across all positions. However, there is a slight difference in three species sequence logo. For example, ‘P’ (proline) at position +4 only found in human and mouse sequence, whereas residue is not favored in rat sequence. In contrast, residue‘P’ is favored at position −3 in rat, whereas not favored in human and mouse sequence. For the lysine methylation sequence logo, in Figure 1B, we observed that this sequence residue is more significant difference than arginine sequence in different species. In the human sequence, ‘R’ (arginine) residue tends to enrich at position −7, −6, −1, +1, +4 and +8, whereas it does not appear in mouse sequence and tends to deplet at position −4 and +7 in the others eukaryotes sequence. Another example is that at position −1 and +1 the residue diverse in three species sequence, the human sequence main residue is ‘R’ and the others eukaryotes sequence are ‘F’ (phenylalanine) and ‘Y’ (tyrosine) residues, whereas there are significant enriched ‘S’ (serine) and ‘G’ residue in mouse sequence. The sequence logo suggested that methylation and non-methylation fragments have a considerable difference among the species sequence. Altogether, the result highlights the necessity and significance of addressing the task of precise arginine and lysine methylation site recognition by developing species-specific predictors.

Second, we investigated the difference of the AAC between methylation and non-methylation sequence in different species. In Figure 2, from left to right, the upper amino acid residues mean that enriched in methylation sequence, whereas other underside amino acids are depleted. For arginine methylation sequences, in Figure 2A, with a slight tendency of distinctions among different species, the overall trends are very similar. Amino acids Gly, Phe, Met, Pro and Tyr are enriched in the methylation sequences, whereas Glu, Lys, Asp, Cys, Ile, Thr, Val and Trp are depleted. However, there are also some differences in diverse species. One example is that amino acids Pro and Tyr have significant difference between methylation and non-methylation sequences favored in rat, but have no distinction in human and mouse sequences. For lysine methylation, in Figure 2B, AAC distinctions are pretty obvious among three different species, and the amino acid residues of enriched and depleted were also easily identified. For instance, amino acid Cys and Tyr in mouse methylation sequences are depleted, while they are enriched in other two species sequences. Furthermore, amino acids Ala, Trp, Val, Pro, Lys, Asp and Arg are enriched or depleted in human and mouse sequences, whereas in the others eukaryotes model amino acids sequences is more probable opposite. These results show that AAC were obvious distinction between methylation and non-methylation sequence, which it could be a helpful feature for methylation prediction.

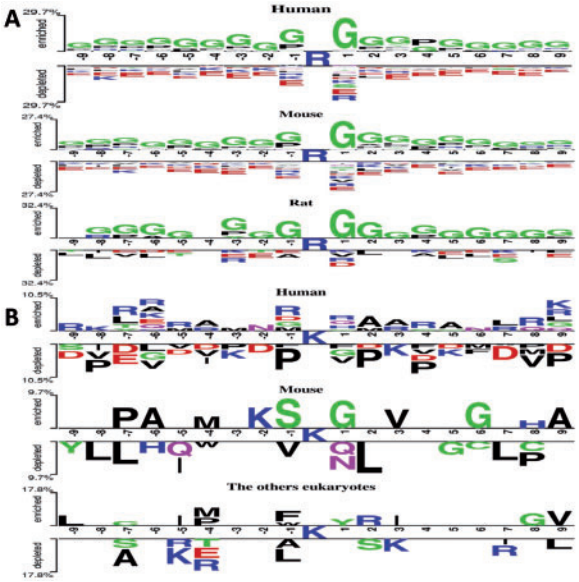

Fig. 1. Sequence logo illustration generated by Two Sample Logo for methylation sites sequence information in different species. (**A**) Arginine methylation logo for human, mouse and rat. (**B**) Lysine methylation logo for human, mouse and the other eukaryotes (Color version of this figure is available at *Bioinformatics* online.)

3.2 Analysis of evolutionary features results

Evolutionary information is an important characteristic of protein, in the field, KNN feature was used to quantify the evolutionary conservative information. The KNN feature has been successfully used for prediction PTM sites such as phosphorylation (Gao *et al.*, 2010) and ubiquitylation (Chen *et al.*, 2013) sites. KNN score measures whether the local sequence surrounding a query site is more similar to the sequences containing methylation sites in the positive set or those with non-methylation sites in the negative set. When the score greater than 0.5 means the query site is more likely to the positive set, while score smaller than 0.5 means it is more similar to the negative set. The larger the KNN score, the more similar the site is to some known methylation sites, and thus, the more likely it is a

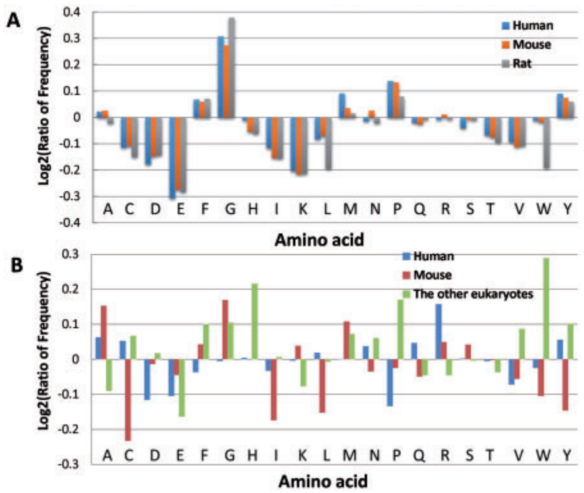

Fig. 2. Comparisons of AAC in positive and negative datasets. The vertical axis represents the log2 ratio of amino acid frequencies surrounding methylation sites and non-methylation sites. The horizontal axis represents the 20 amino acids sorted in descending order by the mean log2 ratio (**A**) arginine sequence; (**B**) lysine sequence (Color version of this figure is available at *Bioinformatics* online.)

methylation site. In Figure 3, we compared the KNN scores of methylation sites with those of non-methylation sites (due to the gap of data, we selected different *k* values and comparison sets; Supplementary Table S5). As can be seen in Figure 3A, the KNN score has significant difference among various values of *k* in human methylarginine and non-methylation sequences, the methylation site scores are within 0.6–0.7, whereas non-methylation scores are around 0.4–0.5. For KNN scores of three different species in methylarginine, the average KNN scores of methylation sites with different sizes of nearest neighbors are within 0.6–0.75 for all three organisms. For non-methylation, the average KNN scores are around 0. 35–0.5. The same to methyllysine, from in Figure 3B, we found the lysine methylation average KNN scores are within 0.5–0.7 for all three organisms and the average scores of non-methylation are around 0.35–0.5, respectively. Overall, methylation sites have larger KNN scores than non-methylation sites. These analyses show that the local sequences surrounding known methylation sites are more similar to their nearest neighbors in the positive set (excluding self-matches) on average as expected, but the sequences in the negative set are not predominantly more similar to nearest neighbors in either the non-methylation or negative set. These similarities are not due to protein homology or the global sequence similarity, because any two proteins in our comparison datasets are either insignificant or low. Also, the result confirms that methylation-related clusters exist in regional sequences around methylation sites. For non-methylation sites, the average KNN scores are minimal, because methylation-related sequence clusters are unlikely to exist in the negative set. Thus, the sequences in the negative set have a similar chance to find close neighbors in either the positive or negative set. In short, KNN scores capture the cluster information in the local sequence around methylation sites and hence distinguish them from the background. Therefore, KNN scores are suitable to be used as a feature for prediction methylation site.

KNN feature is very effective for predicting both general and species-specific methylation sites. For methylation site predictions, KNN feature relies on similarity of local sequence substrates of a common methylation-related enzyme family to automatically distinguish whether site is methylation or not. Although most known methylation sites have no annotation about their sequence substrates of corresponding enzymes family, KNN features can still use the inherent cluster information in them. Oftentimes, one enzyme corresponds to multiple local sequence motifs, and using a single sequence profile may not be as effective as KNN, which better handles diverse sequence clusters.

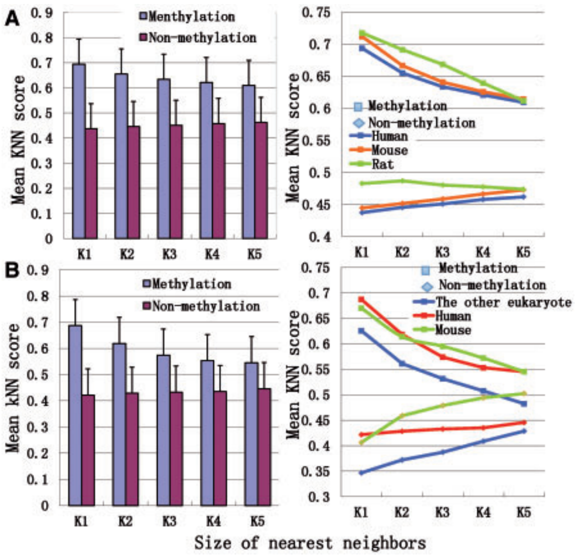

Fig. 3. Comparison of KNN scores between methylation sequences and non-methylation sequences. (**A**) Histogram of KNN scores of five thresholds (K1, K2, K3, K4 and K5) in arginine human and comparison of mean KNN scores between the arginine methylation in three species. (**B**) The KNN scores of five thresholds in lysine human and comparison of mean KNN scores between the lysine methylation in three species (the detailed comparison data information see Supplementary Table S5) (Color version of this figure is available at *Bioinformatics* online.)

3.3 Analysis of physicochemical properties results

Physicochemical property is the most intuitive feature applied in PTMs prediction studies. From the previous work, Shi *et al.* (2012) and Lee *et al.* (2014) used solvent accessible surface area (ASA) (one of physicochemical properties) coding feature and have been evidenced PSP plays an important role for prediction methylation sites. Because the physicochemical property has a total of 544 (if we selected all PSP of features that would highly dimensional, heterogeneous, noisy and time-consuming classification of the training model), we compared the prediction accuracy of all PSP and only selected top three to define as informative features for the prediction model. As can be seen from Supplementary Figure S2, one of the specific examples of informative PSP in arginine mouse model was provided. We calculated the mean values of physicochemical property ‘Thermodynamic beta sheet propensity’ (KIMC930101) at each position in methylation and non-methylation sequences. Interestingly, we found mean values of this property in methylation sequences is higher than in the non-methylation at all the positions, especially at the near the central sites (the positions −1, +1 and +2) the phenomenon is more remarkable. This property features could easily distinguish the methylation or non-methylation sequences, which is suitable for prediction methylation sites (compare with other PSP, this feature have higher prediction accuracy). Furthermore, the selected optimal features of PSP differed, depending on the species of interest. The detailed information of selected top three PSP in different species see Supplementary Table S3.

3.4 Analysis of feature optimization results

3.4.1 The features optimization results

The selected five features (AAC, BE, K-spaced, KNN feature and PSP) were independent coding tests for prediction methylation sites, which the performances are not fully satisfactory. However, compare with single feature, the combination of different five features would provide a more powerful predictor. If we combine all of five features that have totally 2245-dimensional features vector (detailed five features dimensional information in Supplementary Table S6). The highly dimensional features which will lead to a time-consuming classification of the training model, some of features vector may unwanted noise and possibly biased model prediction performance. Accordingly, not all features are equivalently important for the performance of the trained model, so optimization of incorporative features is generally necessary to reduce dimensionality but keep important one. In this work, we first apply IG method to select the valuable dimension feature vectors to reconstitute a new pseudo amino acid composition feature. After first-stage of feature selection, we combined all of five features which totally have 2245-dimensional feature vectors. Subsequently, we find that the IG score have a quite different in each dimension. We know a higher IG score means that it is a more valuable vector based on the IG theories. Then we select IG score top 400 vectors from 2245-dimensional feature vectors and reconstitute it a new orderly feature (from low to high of IG score) named as the 'Final feature'. The prediction performance of used single five features and the 'Final feature' for prediction methylation sites are shown in Figure 4, and we found the 'Final feature' have better prediction performance than other single feature. One of example that the accuracy of mouse lysine methylation prediction model used single five features (AAC, BE, K-spaced, KNN feature and PSP) are 66.39%, 61.02%, 67.47%, 68.01% and 65.86%, respectively (The other species of five features prediction performance detailed information in Supplementary Tables S7 and S8). While when we applied 'Final feature' for prediction, the accuracy is 81.18%. Comparing with single five features, which improved accuracy are 14.79%, 20.16%, 13.71%, 13.17% and 15.32%, respectively. The result shows that the IG method could effectively extract those important valuable dimension vectors from multi-features for methylation sites prediction, which have significantly improvement of prediction accuracy about 15%.

3.4.2 Analysis of feature importance and contribution

As mentioned previously, the five features sets (AAC, BE, K-spaced, KNN feature and PSP) were selected to predict methylation sites. We do not know the five features make how much contribution and which is more importance than others. However, we adopted IG method to analyze features importance and contribution, and evaluated which dimensions features vectors are valuable to prediction model. For example, for human arginine methylation (the IG score of 2245-dimensional feature vector is shown in Supplementary Fig. S3), we found that IG score of KNN 5-dimensions is significantly higher than other four features dimensions vector, suggesting that KNN features play significant roles in determining the prediction performance of the model and make a great contribution to predict methylation sites. In fact, methylation histones possess highly conserved sequences in most species. The KNN feature is related to conserved residues and protein evolutionary information, which has good prediction performance in this model. In contrast, most of IG score values of K-spaced and BE features dimensions vector are low, which suggest that most of dimension features is unwanted and this two features are not importance as KNN feature for the model.

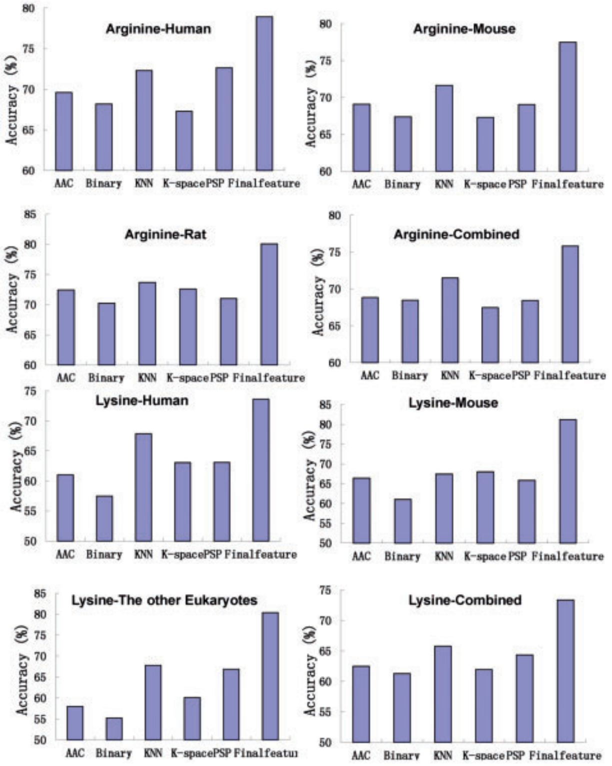

Fig. 4. Prediction performance of using single five features and the 'Final feature' for prediction methylation sites in different species models

We additionally discuss the constitution of finally dimensions features vector, which is various at different species model. These analyses show that the number of optimal dimensional features selected from five features was differed, depending on the species of interest. Detailed information of our results is presented in Supplementary Table S9. Taking human lysine methylation as an example, the reconstitution of 400-dimension new feature from aforementioned five features (AAC, BE, K-spaced, KNN feature and PSP), and the rations of selected dimension feature vectors belonging to the five features group are 0.7 (14/20), 0.09 (38/399), 0.16 (289/1746), 1 (5/5) and 0.95 (54/57), respectively. From these results, we discover that most of dimension feature vectors from AAC, PSP and KNN three features are valuable and make a contribution to prediction model, while the other two features BE and K-spaced have many unwanted or redundant dimension vectors. These suggest that compared with BE and K-spaced features, the other three features are more importance to this model. Furthermore, considering the features optimization results in different species, therefore, we expect that IG method fully consider the importance and contribution of each dimension feature vectors, which can obtain the higher prediction accuracy.

3.5 Species-specific methylation site prediction of PSSMe

The proposed PSSMe predictor trained and tested for arginine and lysine methylation has four models based on IG method optimizations features, respectively. To evaluate the performance of PSSMe for species-specific methylation site prediction, we performed a 10-fold cross-validation test in each species. Sensitivity at different specificity levels in each cross-validation were calculated according to Supplementary Ep3. By taking different thresholds, then we plotted the receiver operating characteristic (ROC) curves and calculated the AUC as shown in Supplementary Figure S4. Three arginine methylation models of human, mouse and rat are achieved under ROC curves (AUC) of 0.846, 0.864 and 0.853, respectively. Similarly, three lysine methylation models of human, mouse and the other eukaryotes achieved AUC of 0.844, 0.856 and 0.867, respectively. From the ROC curves, we knew that those models have good confident predictions with high specificities, especially for the arginine mouse model, the prediction sensitivities reached 58% at the 90% specificity level, and 30% at the 95% specificity level. Furthermore, the AUC of models of arginine combined and lysine combined achieved of 0.851 and 0.834, respectively. The PSSMe model was testing by 10-fold cross-validation, which has stabled prediction performance. In contrast, we take a testing for cross-species prediction for methylation sites, but the results of prediction is not fully satisfactory and achieved lower specificities and AUC scores. In summary, these justify our PSSMe general and species-specific predictor model have stabilization and good prediction performance.

3.6 Comparison with other prediction tools

To further evaluate the performance of PSSMe prediction methylation site, we compared it with some existing widely used arginine and lysine methylation prediction tools. Although there have been reported many tools for predicting protein methylation sites since 2005, considering the web server tools whether do work or not, four general tools and one species-specific predictor GPS-MSP were selected to compare the arginine and lysine models, respectively. The comparison of general arginine methylation predictors are MeMo, BPB-PPMS, PMeS and iMethyl-PseAAC, while the comparison of lysine methylation tools are MeMo, PLMLA, PMeS and MethK. Because the training data used to build these tools were collected from different databases, the prediction performance varies greatly among them. To avoid any bias, in this work, the additional independent test dataset was applied in the cross-species prediction evaluation. We submitted each species test dataset to MeMo, BPB-PPMS, PMeS and iMethyl-PseAAC tools for prediction arginine methylation site. From the results (Table 1), PSSMe achieves significantly higher AUC values than other four tools (the ROC performance comparison of our model with other tools in Supplementary Figs. S5 and S6). For arginine methylation model, AUC values of the prediction performance of PSSMe are 0.870, 0.893 and 0.978 in human, mouse and rat model, respectively. Comparing with MeMo (the AUC values is 0.683), our three model improved the AUC values are 0.187, 0.21 and 0.295, respectively. In addition to the AUC value comparison, we also calculated the MCC, sensitivity and specificity, as shown in Table 1. BPB-PPMS and iMethyl-PseAAC tools can achieve high specificity along with sacrificing sensitivity, which would case a low MCC and AUC values. PMeS have a good balance between specificity and sensitivity, but the MCC and AUC of prediction performance are not fully satisfactory. However, PSSMe not only offers good sensitivity as well as high specificity, but also have higher AUC and MCC values. One of the examples for human arginine model, the prediction performance is sensitivity of 83.3%, specificity of 91.0%, accuracy of 87.1%, MCC of 0.745 and AUC of 0.870, while other four approaches have lower values. It shows that species-specific PSSMe outperforms general tools.

For the lysine methylation tools, we made a comparison of prediction to the species of human, mouse and the others eukaryotes. The performance of PSSMe and other four tools predict independent test methylation dataset are shown in Table 2. We found that MeMo and PMeS tools have high specificity meanwhile sacrificing sensitivity, the AUC and MCC are not good when prediction human and mouse methylation sites. PLMLA and MethK have a good balance of prediction performance, but comparison with PSSMe, which our prediction result is better. For example, the performances of MethK for prediction human lysine methylation are sensitivity of 72.7%, specificity of 96.9%, accuracy of 84.8%, MCC of 0.718 and AUC of 0.858, while the performances of our tool are sensitivity of 98.4%, specificity of 96.9%, accuracy of 97.7%, MCC of 0.954 and AUC of 0.969. Meanwhile, MethK have not good performances when used to predict the other eukaryotes methylated sites, while PSSMe have significantly improvement of sensitivity and AUC values for species-specific methylation site prediction. In addition, we compared PSSMe arginine and lysine combined models with other four tools for methylation sites, respectively. The detailed

Table 1. Prediction performance comparison between our method (arginine methylation model of different species) and other tools (SN: sensitivity, SP: specificity, ACC: accuracy, MCC: Matthew correlation coefficient)

Organism	Tools	SN	SP	ACC	MCC	AUC
Human	MeMo	0.384	0.897	0.641	0.328	0.683
	BPB-PPMS	0.115	**1.00**	0.557	0.247	0.746
	iMethyl-PseAAC	0.166	0.974	0.570	0.239	0.732
	PMeS	0.512	0.705	0.609	0.222	0.604
	GPS-MSP	0.807	0.897	0.852	0.708	0.841
	Our work	**0.833**	0.910	**0.871**	**0.745**	**0.870**
Mouse	MeMo	0.478	0.753	0.615	0.241	0.621
	BPBPPMS	0.188	0.898	0.543	0.123	0.596
	iMethyl-PseAAC	0.275	**0.913**	0.594	0.244	0.621
	PMeS	0.608	0.521	0.565	0.130	0.555
	GPS-MSP	0.841	0.754	0.797	0.597	0.804
	Our work	**0.927**	0.840	**0.884**	**0.771**	**0.893**
Rat	MeMo	0.460	0.857	0.658	0.345	0.683
	BPBPPMS	0.269	**0.984**	0.627	0.362	0.763
	iMethyl-PseAAC	0.285	0.968	0.627	0.347	0.744
	PMeS	0.539	0.730	0.634	0.274	0.625
	GPS-MSP	0.362	0.921	0.659	0.377	0.700
	Our work	**1.00**	0.952	**0.976**	**0.953**	**0.978**

The best evaluation parameters for each test model are highlighted in bold.

Table 2 Prediction performance comparison between our method (different species of lysine methylation model) and other tools

Organism	Tools	SN	SP	ACC	MCC	AUC
Human	MeMo	0.197	0.893	0.545	0.126	0.575
	PLMLA	0.439	0.636	0.537	0.077	0.547
	PMeS	0.303	0.969	0.636	0.365	0.733
	MethK	0.727	0.969	0.848	0.718	0.858
	GPS-MSP	0.767	**1.00**	0.849	0.731	0.882
	Our work	**0.984**	0.969	**0.977**	**0.954**	**0.969**
Mouse	MeMo	0.278	**0.934**	0.606	0.282	0.723
	PLMLA	0.508	0.606	0.557	0.115	0.582
	PMeS	0.377	0.950	0.663	0.400	0.723
	MethK	0.868	0.885	0.877	0.754	**0.885**
	GPS-MSP	0.328	0.918	0.623	0.306	0.682
	Our work	**0.946**	0.875	**0.910**	**0.823**	0.847
The others eukaryotes	MeMo	0.625	0.906	0.765	0.553	0.786
	PLMLA	0.546	0.546	0.546	0.093	0.558
	PMeS	0.703	0.890	0.796	0.604	0.793
	MethK	0.750	0.890	0.820	0.647	0.823
	Our work	**1.00**	**0.921**	**0.960**	**0.924**	**0.963**

The best evaluation parameters for each test model are highlighted in bold.

information of result is listed in Supplementary Table S10; PSSMe arginine combined models improved the AUC values of 0.244, 0. 331, 0.328 and 0.374 than MeMo, BPB-PPMS, PMeS and iMethyl-PseAAC, respectively. Similarly, PSSMe lysine combined models improved the AUC values of 0.233, 0.342, 0.154 and 0.039 than MeMo, PLMLA, PMeS and MethK, respectively. What is more, we also compared with other species-specific predictor GPS-MSP, which is developed by the author of MeMo. First, in order to make the results comparable, sensitivities of GPS-MSP and PSSMe were set to 0.5. Then, we submitted species-specific independent test dataset to GPS-MSP model for prediction methylation site, and the results are shown in Tables 1 and 2. We found that comparison with other general prediction tools, prediction performances of species-specific models PSSMe and GPS-MSP have outperformed. Take arginine human model as an example, PSSMe and GPS-MSP achieved significantly higher AUC values than other four general tools, it is evidenced that developing a species-specific predictors is necessary. Meanwhile, comparing with species-specific prediction methylation sites, PSSMe have outperformed GPS-MSP. For example, for lysine methylation rat model, our model could more accurately predict methylation than GPS-MSP. It is show that PSSMe has good prediction performance for species-specific prediction methylation sites.

The reason that PSSMe outperforms other predictors is mainly following: (i) the data sources of the experimental methylation sites for the previous tools was collected mainly from UniProtKB/Swiss-Prot and it may bias on some special species (most of earlier methylation data were collected from human protein and only a few hundred methylated proteins). The little data may case a bias classification of the training model and are not suitable for predicting all species. However, PSSMe integrated 8100 experimental methylation sites from multiple resources, which have been accumulated the methylation data from thousands of proteins in different species. (ii) The BPB-PPMS and iMethyl-PseAAC tools have low AUC that is partly attributed to using the single feature, which could not extract fully features information from methylation sequences. In contrast, PSSMe integrated five features to ensure the complete extraction of sequence information, further conducted multi-features selection and optimization based on IG method. Finally, analysis of each features contribution and importance found valuable dimensions feature vector to build the model for prediction methylation sites. (iii) When performing the comparisons, we used a prediction model that was trained from a dataset excluding the protein sequences in the independent test dataset. However, for comparison tools, some of the test proteins might have been included in their training processes, and thus prediction performances may be biased favorably toward these tools in the comparisons. This possibility implies that the performance improvement of over these tools might be underestimated.

4 Conclusion

This is the study which classifies proteins into species-specific to identify potential arginine and lysine methylation sites. We not only demonstrated that PSSMe both general and species-specific models have stabilization and good prediction performance, but also showed that our models significantly improved the prediction results compared to previous mixed model. Our analysis shows the methylation patterns are significantly different in different species, and feature optimization by using IG method indicates that KNN feature is important and makes a great contribution for prediction model. Furthermore, we have developed a user-friendly web server, PSSMe, to implement the described methylation site prediction, which could be especially useful for some hypothesis-driven experiments. However, our method still has certain limitations, which are common to almost all computational prediction tools. For example, the methylation data that we have found only labeled positive data, while the negative data have no any evidenced to labeled, which may affect the prediction performance. It will be resolved in the future with techniques rapidly increasing, we will develop some novel models for specific prediction methylation sites. In conclusion, we believe that PSSMe could serve as a powerful and complementary tool for *in vivo* or *in vitro* species-specific methylation site identification. Additionally, the combination of computational analyses with experimental verification could greatly speed up our systematic understanding of the methylation mechanisms and explore the corresponding regulatory networks in living cells.

Funding

This work was supported by the National Natural Science Foundation of China (21175064 and 21305062), the Program for New Century Excellent Talents in University (NCET-11-1002) and the Jiangxi Province Natural Science Foundation (20151BAB203022).

Conflict of Interest: none declared.

References

Aleta,J.M. *et al*. (1998) Protein methylation: a signal event in post-translational modification. *Trends Biochem. Sci*., **23**, 89–91.

Bannister,A.J., *et al*. (2005) Reversing histone methylation. *Nature*, **436**, 1103–1106.

Bedford,M.T. and Richard,S.S. (2005) Arginine methylation an emerging regulator of protein function. *Mol. Cell*, **18**, 263–272.

Boisvert,F.M. *et al*. (2003) A proteomic analysis of arginine-methylated protein complexes. *Mol. Cell Proteomics*, **2**, 1319–1330.

Chen,H. *et al*. (2006) MeMo: a web tool for prediction of protein methylation modifications. *Nucleic Acids Res*., **34**, W249–W253.

Chen,X. *et al*. (2013) Incorporating key position and amino acid residue features to identify general and species-specific Ubiquitin conjugation sites. *Bioinformatics*, **29**, 1614–1622.

Chernorudskiy,A.L. *et al*. (2007) UbiProt: a database of ubiquitylated proteins. *BMC Bioinformatics*, **8**, 126.

Daily,K.M. *et al*. (2005) Intrinsic disorder and protein modifications building an SVM predictor for methylation. *IEEE*, 475–481.

Deng,W.-K. *et al*. (2016) Computational prediction of methylation types of covalently modified lysine and arginine residues in proteins. *Brief. Bioinf*, doi: 10.1093/bib/bbw041.

Gao,J. *et al*. (2010) Musite, a tool for global prediction of general and kinase-specific phosphorylation sites. *Mol. Cell Proteomics*, **9**, 2586–2600.

Henikoff,S. and Henikoff,J.G. (1992) Amino-acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U. S. A*., **89**, 10915–10919.

Hornbeck,P.V. *et al*. (2004) PhosphoSite: a bioinformatics resource dedicated to physiological protein phosphorylation. *Proteomics*, **4**, 1551–1561.

Huang,J. and Berger,S.L. (2008) The emerging field of dynamic lysine methylation of non-histone proteins. *Curr. Opin. Genet. Dev*., **18**, 152–158.

Johnson,D.S. *et al*. (2008) Systematic evaluation of variability in ChIP-chip experiments using predefined DNA targets. *Genome Res*., **18**, 393–403.

Kawashima,S. and Kanehisa,M. (2000) AAindex: amino acid index database. *Nucleic Acids Res*., **44**, 325–340.

Kennelly,P.J. and Krebs,E.G. (1995) Consensus sequences as substrate specificity determinants for protein kinases and protein phosphatases. *J. Biol. Chem*, **266**, 15555–15558.

Lee,T.Y. *et al*. (2014) Identification and characterization of lysine-methylated sites on histones and non-histone proteins. *Comput. Biol. Chem*., **50**, 11–18.

Lee,T.Y. *et al*. (2006) dbPTM: an information repository of protein post-translational modification. *Nucleic Acids Res*., **34**, 323–625.

Li,J. *et al.* (2014) SysPTM 2.0: an updated systematic resource for post-translational modification. *Database*, **2014**. doi:10.1093/database/bau025.

Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

Liu,Z.-X. *et al.* (2013) CPLM: a database of protein lysine modifications. *Nucleic Acids Res.*, **42**, D531–D536.

Longo,V.D. and Kennedy,B.K. (2006) Sirtuins in aging and age-related disease. *Cell*, **126**, 257–268.

Lott,K. *et al.* (2014) Functional interplay between protein arginine methyltransferases in *Trypanosoma brucei*. *MicrobiologyOpen*, **3**, 595–609.

Mann,M. and Jensen,O.N. (2003) Proteomic analysis of post-translational modifications. *Nat. Biotechnol.*, **21**, 255–261.

Mastronardi,F.G. *et al.* (2006) Increased citrullination of histone H3 in multiple sclerosis brain and animal models of demyelination: a role for tumor necrosis factor-induced peptidylarginine deiminase 4 translocation. *J. Neurosci.*, **26**, 11387–11396.

Ong,S.E. *et al.* (2004) Identifying and quantifying in vivo methylation sites by heavy methyl SILAC. *Nat. Methods*, **1**, 119–126.

Paik,K.W. and Kim,S. (1967) Enzymatic methylation of protein fractions from calf thymus nuclei. *Biochem. Biophys. Res. Commun.*, **29**, 14–20.

Paik,W.K. *et al.* (2007) Historical review: the field of protein methylation. *Trends Biochem. Sci.*, **32**, 146–152.

Peng,H. *et al.* (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.*, **27**, 1226–1238.

Plewczynski,D. *et al.* (2005) AutoMotif server: prediction of single residue post-translational modifications in proteins. *Bioinformatics*, **21**, 2525–2527.

Qiu,W.R. *et al.* (2014) iMethyl-PseAAC: identification of protein methylation sites via a pseudo amino acid composition approach. *Biomed. Res. Int.*, **2014**, 947416.

Shao,J. *et al.* (2009) Computational identification of protein methylation sites through bi-profile Bayes feature extraction. *Plos One*, **4**, e4920.

Shi,S.P. *et al.* (2012) PLMLA: prediction of lysine methylation and lysine acetylation by combining multiple features. *Mol. Biosyst.*, **8**, 1520–1527.

Shi,S.P. *et al.* (2015) Progress and challenges in predicting protein methylation sites. *Mol. Biosyst.*, **11**, 2610–2619.

Shi,S.P. *et al.* (2012) PMeS: prediction of methylation sites based on enhanced feature encoding scheme. *Plos One*, **7**, e38772.

Shien,D.M. *et al.* (2009) Incorporating structural characteristics for identification of protein methylation sites. *J. Comput. Chem.*, **30**, 1532–1543.

Suo,S.-B. *et al.* (2012) Position-specific analysis and prediction for protein lysine acetylation based on multiple features. *Plos One*, **7**, e49108.

Suzuki,A. *et al.* (2007) Citrullination by peptidylarginine deiminase in rheumatoid arthritis. *Ann. N. Y. Acad. Sci.*, **1108**, 323–339.

Tomii,K. and Kanehisa,M. (1996) Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. *Protein Eng.*, **9**, 27–36.

Vacic,V. *et al.* (2006) Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. *Bioinformatics*, **22**, 1536–1537.

Wang,Z. *et al.* (2009) Targeted metabolomic evaluation of arginine methylation and cardiovascular risks: potential mechanisms beyond nitric oxide synthase inhibition. *Arterioscler. Thromb. Vasc. Biol.*, **29**, 1383–1391.

Xu,Y. *et al.* (2013) Prediction of protein methylation sites using conditional random field. *Protein Pept. Lett.*, **20**, 71–77.

Yang,Y. and Bedford,M.T. (2013) Protein arginine methyltransferases and cancer. *Nat. Rev. Cancer*, **13**, 37–50.

Zhang,W. *et al.* (2013) Prediction of methylation sites using the composition of K-spaced amino acid pairs. *Protein Pept. Lett.*, **2020**, 911–917.