Molecular BioSystems

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Mol. BioSyst.*, 2015, **11**, 3092

A novel method for predicting post-translational modifications on serine and threonine sites by using site-modification network profiles†

Minghui Wang,*[a,b] Yujie Jiang[a] and Xiaoyi Xu[a]

Post-translational modifications (PTMs) regulate many aspects of biological behaviours including protein–protein interactions and cellular processes. Identification of PTM sites is helpful for understanding the PTM regulatory mechanisms. The PTMs on serine and threonine sites include phosphorylation, *O*-linked glycosylation and acetylation. Although a lot of computational approaches have been developed for PTM site prediction, currently most of them generate the predictive models by employing only local sequence information and few of them consider the relationship between different PTMs. In this paper, by adopting the site-modification network (SMNet) profiles that efficiently incorporate *in situ* PTM information, we develop a novel method to predict PTM sites on serine and threonine. PTM data are collected from various PTM databases and the SMNet is built to reflect the relationship between multiple PTMs, from which SMNet profiles are extracted to train predictive models based on SVM. Performance analysis of the SVM models shows that the SMNet profiles play an important role in accurately predicting PTM sites on serine and threonine. Furthermore, the proposed method is compared with existing PTM prediction approaches. The results from 10-fold cross-validation demonstrate that the proposed method with SMNet profiles performs remarkably better than existing methods, suggesting the power of SMNet profiles in identifying PTM sites.

Received 8th June 2015,
Accepted 23rd August 2015

DOI: 10.1039/c5mb00384a

www.rsc.org/molecularbiosystems

Introduction

Post-translational modifications (PTMs) are chemical modifications of a protein after translation, and they play a crucial role in regulating the diversity of biological processes such as signal transduction, DNA repair, cell cycle control, protein–protein interactions and protein functions.[1–3] Among the PTMs, the most well studied PTM is phosphorylation on serine (S), threonine (T) or tyrosine (Y) sites. It has been reported that more than one-third of all proteins are phosphorylated, and many of them are related to diseases.[4] Besides phosphorylation, on serine (S) and threonine (T) sites there are other PTMs such as *O*-linked glycosylation and acetylation. *O*-linked glycosylation has two subtypes, namely *O*-*N*-acetylgalactosamine (*O*-GalNAc) and *O*-*N*-acetylglucosamine (*O*-GlcNAc),[5] and the latter is usually involved in regulating the pathways that are disrupted in diabetes mellitus. At the same time, acetylation is reported as a prevalent modification in enzymes that catalyse an intermediate metabolism.[6]

A lot of experimental approaches have been developed to identify PTM sites on S/T.[7,8] Historically, PTM sites have been discovered primarily through the use of low-throughput biological technology,[9] and with the recent development of high throughput mass spectrometry-based techniques,[10] experimentally determined PTM sites have exponentially increased. Since experimental methods are high-cost and labor-intensive,[11] considerable efforts have been devoted to the computational identification of PTM sites in recent years. With the development of machine learning algorithms, a number of PTM site prediction approaches have been proposed based on these methods including the support vector machine (SVM),[12,13] random forest (RF)[14–17] and artificial neural networks (ANNs).[18] Besides these methods, a lot of currently available tools are also developed. For example, Xue propose PPSP[19] and GPS[20] for identifying phosphorylation sites using the residue information around the potential sites. KinasePhos2.0[21] incorporates the protein coupling pattern and sequence profile to predict the phosphorylation sites, while Musite[22] employs a machine learning approach that integrates local sequence similarities for phosphorylation site prediction. *O*-GlcNAcPRED[23] adopts the SVM to build an *O*-GlcNAc site prediction model using a novel feature extraction method

[a] *School of Information Science and Technology, University of Science and Technology of China, Hefei AH230027, People's Republic of China*

[b] *Centers for Biomedical Engineering, University of Science and Technology of China, Hefei AH230027, People's Republic of China. E-mail: mhwang@ustc.edu.cn; Tel: +86 13805604436*

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c5mb00384a

based on the local sequences, and YinOYang[24] uses PTM local sequence information based on an artificial neural network (ANN) to predict the *O*-GlcNAc sites.

Despite the successes achieved by the aforementioned PTM site prediction methods, most of these computational approaches generate predictive models solely by employing protein local sequence information. While PTMs present very complicated processes and are usually involved in multiple biological mechanisms, only considering the local sequence information is not sufficient for accurate prediction and therefore more information should be included. Recently, several studies[25–27] have explored that functional associations (*i.e.*, PTM crosstalk[28,29]) exist between different PTMs occurring on the same S/T sites (thereafter called *in situ* PTMs). For example, phosphorylation can influence and regulate acetylation, *O*-linked glycosylation and many other *in situ* PTMs.[30–32] Therefore, the information on *in situ* PTMs presents potential functional associations between multiple PTMs, and therefore is helpful in PTM site prediction. However, such information cannot be simply adopted since the existing prediction methods are mainly focused on the analysis of a single type of PTM. Therefore, the performance of PTM prediction can be improved if the information on *in situ* PTMs is appropriately adopted.

In this study, we conduct a new, powerful computational approach for PTM site prediction by introducing novel site-modification network (SMNet) profiles that can efficiently incorporate *in situ* PTM information. By employing SMNet profiles as well as local sequence information, we generate SVM models for different PTMs on S/T sites that can accurately predict PTM sites for phosphorylation, *O*-GalNAc, *O*-GlcNAc and acetylation. To evaluate the performance of our method, the proposed approach is compared with SVM models that only adopt local sequence information and the results suggest that the SMNet profiles contribute to the remarkable prediction performance. Furthermore, we also compare the proposed method with other state of the art prediction tools such as GPS, PPSP, Musite and KinasePhos2.0, and further evaluation demonstrates that the proposed method significantly outperforms these existing PTM site prediction methods.

Materials and methods

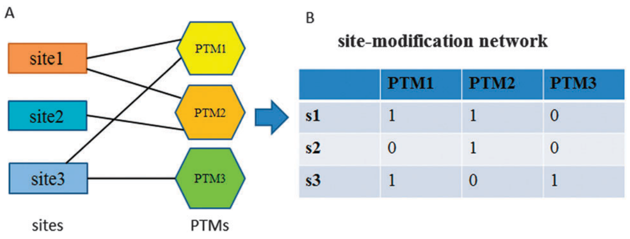
Data collection and pre-processing

Experimentally verified phosphorylation, *O*-GalNAc, *O*-GlcNAc and acetylation sites are extracted from comprehensive PTM resources including dbPTM version 3.0,[33] Phospho.ELM,[34] phosphositePlus,[35] dbOGAP[36] and SysPTM.[37] Since more than 80% of the known PTM proteins are human proteins, in this study we focus on the prediction of human PTMs. After removing the repetitive data, we collect 168 082 phosphorylation sites [116 161 S sites, 51 921 T sites], 2098 *O*-GalNAc sites [837 S sites, 1261 T sites], 408 *O*-GlcNAc sites [243 S sites, 165 T sites] and 536 acetylation sites [472 S sites, 64 T sites]. From 168 082 phosphorylation sites we only pick up the sites that are either phosphorylation related *in situ* PTMs or have kinase information for further analysis, as such information is necessary to construct the SMNet. In addition, the phosphorylation data with kinase group information are also obtained from Zou *et al.*[38] and are shown in Table S1 (ESI†). Together with 2098 *O*-GalNAc sites, 408 *O*-GlcNAc sites and 536 acetylation sites, we totally use 2990 PTM local sequences on S sites and 1961 on T sites to obtain the final SMNet. For a specific PTM, the positive dataset contains the sites in the SMNet that are known to be modified by this PTM, and the negative dataset includes the sites in the SMNet that are not known to be modified by this PTM. Therefore, for each PTM both positive and negative dataset are determined once an SMNet is constructed and cannot be changed freely as described by Biswas *et al.*[39] Local sequences that contain 10 residues upstream and 10 residues downstream from the PTM sites are extracted for further analysis. To infer kinase group information on the phosphorylation sites of *in situ* PTMs, we adopt the method proposed in iGPS[1] (*in vivo* GPS) by measuring the similarities of local sequences using the BLOSUM62 matrix.

Construction of the SMNet and predictive models

In order to quantitatively describe the known relationship between the sites and the PTMs, we construct the SMNet by using both single and *in situ* PTMs, which is represented by a bipartite graph (Fig. 1). The heterogeneous nodes in the SMNet correspond to either protein sites or different types of PTMs, and an edge is placed between a site and PTM if the site is modified by the PTM. Totally, there are 2243 and 3885 edges in the SMNet for T and S sites, respectively. Accordingly, we present the SMNet with a matrix N, in which each row represents a modified site and each column represents a phosphorylation kinase group, *O*-GalNAc, *O*-GlcNAc or acetylation. The entity N(i,j) in row i and column j is 1 if the site s(i) is associated with the PTM m(j), otherwise 0.

To incorporate SMNet information for prediction, we propose to extract SMNet profiles for each site, which is the modification data for site s(i). For example, the site (Q99717, T2) is with single PTM acetylation. We denote the SMNet profiles of site s(i) as the binary vector encoding the presence or absence of the relationship between site s(i) and each modification in the known PTM data, namely the ith row of the adjacency matrix N. The local



	PTM1	PTM2	PTM3
s1	1	1	0
s2	0	1	0
s3	1	0	1

Fig. 1 Construction of the SMNet profiles. (A) The bipartite graph shows the relationship between the protein sites and PTMs, the boxes and the hexagons represent the site nodes and the PTM nodes, respectively. (B) SMNet profiles are extracted from the bipartite graph and s1, s2 and s3 represent different sites. Each row is site s(i) and each column is PTM m(j), if s(i) is modified by m(j), the value is 1, otherwise 0.

sequences adopt the binary encoding scheme[40] to transform each amino acid into a 21-dimensional binary vector. For example, the local sequence LS = 'DFGLAREWHKTTQMSAAGTYA', G is expressed as a 21-dimensional vector [0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0] and Y is expressed as vector [0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0]. Therefore, the local sequence with a length of 21 is transformed to a 441-dimensional vector. To optimize the performance, the SMNet profiles that are employed as feature vectors with a length of 11 and the binary encoding of local sequences are combined to generate the feature vectors. For example, for site s_i the SMNet profiles are represented as a vector P_i, the binary encoding of local sequence is LS_i, so the combined feature vector $F_i = [LS_i\ P_i]$. For each type of PTM, the combined feature vectors generated from the training data are utilized to generate predictive models. This procedure is implemented by a public SVM library LIBSVM.[41] So given the training set $T = \{(x_1, y_1),(x_2, y_2),\ldots(x_N, y_N)\}$, $x_i \in R^n$, $x_i = F_i$, $y_i = \{1,0\}$, $i = 1, 2,\ldots N$. The model solves the following primal optimization problem.

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i \\ \text{s.t.} \quad & y_i(w\cdot x_i + b) \geq 1 - \xi_i, \quad i = 1, 2\ldots, N \\ & \xi_i \geq 0, \quad i = 1, 2\ldots, N \end{aligned} \tag{1}$$

where $C > 0$ is the regularization parameter. It is a convex quadratic programming problem and by solving this problem, the optimal ω^*, b^* can be obtained and the predictive model is shown as follows:

$$f(x) = \text{sign}(w^*\cdot x + b^*) \tag{2}$$

To evaluate the prediction performance, a 10-fold cross-validation is adopted, in which the dataset is randomly partitioned into ten subsamples, of the ten subsamples, each time one single subsample is retained as the validation data for testing the model, and the remaining nine subsamples are used as training data. The values of the SMNet profiles in the testing set are left out by setting their entries to 0. Therefore, no information about the relationship between sites and PTMs in the test part is leaked in this way. The source codes and the data in this study are available in the ESI.†

Performance evaluation

To evaluate the performance, the receiver operating characteristic (ROC) curve and the area under the curve (AUC) that represents the overall classification accuracy are also employed to assess the classifier we trained. In addition, we adopt commonly used measurements, namely sensitivity (Sn), specificity (Sp), accuracy (Acc), precision (Pre) and Matthew's correlation coefficient (MCC) which are defined as follows:

$$\text{Acc} = \frac{\text{TN} + \text{TP}}{\text{TN} + \text{TP} + \text{FN} + \text{FP}} \tag{3}$$

$$\text{Sn} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

$$\text{Sp} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{5}$$

$$\text{Pre} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{6}$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FN}) \times (\text{TP} + \text{FP}) \times (\text{TN} + \text{FN}) \times (\text{TN} + \text{FP})}} \tag{7}$$

TN and TP are the true negative and true positive, indicating the number of negative and positive sites that are truly predicted. FN and FP are the false negative and false positive, FN illustrates that the true positives are predicted as negatives, FP shows that the true negatives are predicted as positives. Sn and Sp represent the proportions of positive and negative sites that can be correctly identified. However, when the numbers of positive and negative sets are significantly imbalanced, MCC should be used to reflect the balance quality.

Results

Evaluation of the SMNet profiles

To evaluate the importance of SMNet profiles in PTM prediction, the SVM models trained with both SMNet profiles as well as local sequences are assessed based on a 10-fold cross-validation, and the ROC curves of phosphorylation in Atypical, CAMK, STE kinase groups, *O*-GalNAc, *O*-GlcNAc and acetylation on S and T sites are plotted and shown in Fig. 2.

From the results it is evident that the proposed method has good prediction accuracy in predicting PTM sites. For kinase groups Atypical, CAMK and STE, the AUC values of SVM trained with only protein local sequences are 72.3%, 76.7%, and 71.4% on T sites and 74.5%, 70.8%, and 73.0% on S sites, respectively. By incorporating SMNet profiles, the corresponding AUC values are remarkably increased to 85.6%, 87.5%, and 94.1% on T sites and 92.3%, 89.3%, and 91.4% on S sites, respectively. In addition, for *O*-GalNAc, *O*-GlcNAc and acetylation, the AUC values of the proposed method are 19.2%, 19.8%, and 1.3% (T sites) and 9.8%, 13.7%, and 0.9% (S sites) higher than those obtained by using only local sequences. The ROC curves of the other kinase groups are plotted and shown in Fig. S1 (ESI†), which further suggest that the SMNet profiles contribute to the prediction performance improvement. Therefore, with the help of the SMNet profiles, the SMNet method outperforms the SVM trained with local sequences in all phosphorylation groups, *O*-GalNAc, *O*-GlcNAc and acetylation on S and T sites.

Besides the AUC values, other measurements such as Sn, Sp, Acc, Pre and MCC are also employed to evaluate the performance. To ensure a low false positive rate, we assess the performance at high and medium stringency levels that corresponding to Sp is 99.0% and 95.0% respectively. Table 1 shows the results obtained at the medium stringency level. On T sites, the values of Sn, MCC, Pre and Acc for the kinase group Atypical are 57.9%, 36.0%, 25.8% and 93.4%, which are increased by 36.8%, 24.1%, 14.6% and 0.5% compared with the models using

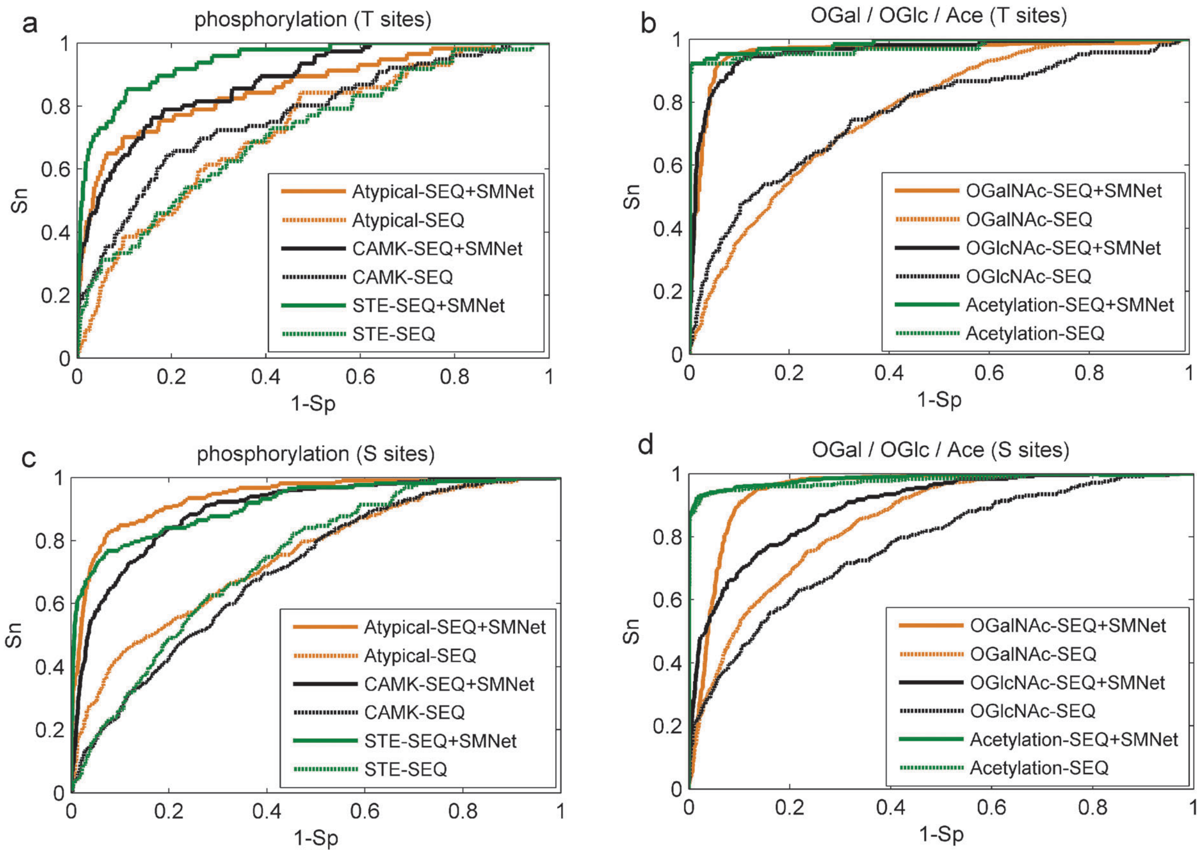


Fig. 2 ROC curves of different PTMs: phosphorylation, *O*-GalNAc, *O*-GlcNAc and acetylation. The subplots a and c represent the performance of phosphorylation prediction on T sites and S sites, respectively. The subplots b and d show the performance of *O*-GalNAc, *O*-GlcNAc and acetylation on T and S sites. OGal/OGlc/Ace represents *O*-GalNAc, *O*-GlcNAc and acetylation. The solid lines represent the SVM model constructed with SMNet profiles as well as local sequences, and the dotted lines represent the performance of the proposed method with only local sequences. SEQ and SMNet represent the sequence and site-modification network, respectively.

Table 1 Performance comparison of SMNet profiles and local sequences on S/T sites at the medium stringency level (Sp = 95.0%)

	PTM	Features	Sn (%)	MCC (%)	Pre (%)	Acc (%)
T	Phosphorylation (Atypical)	SEQ	21.1	11.9	11.2	92.9
		SEQ + SMNet	**57.9**	**36.0**	**25.8**	**93.4**
	Phosphorylation (CAMK)	SEQ	29.0	19.6	19.0	92.5
		SEQ + SMNet	**51.3**	**35.6**	**29.3**	**93.3**
	Phosphorylation (STE)	SEQ	27.1	14.9	11.9	93.3
		SEQ + SMNet	**72.9**	**42.0**	**26.7**	**94.4**
	O-GalNAc	SEQ	21.3	21.5	88.5	47.6
		SEQ + SMNet	**87.4**	**80.0**	**96.9**	**90.1**
	Acetylation	SEQ	92.2	57.6	38.3	94.9
		SEQ + SMNet	**93.8**	**58.4**	**38.7**	**95.0**
S	Phosphorylation (Atypical)	SEQ	30.2	25.7	48.2	90.4
		SEQ + SMNet	**75.5**	**60.3**	**72.3**	**93.6**
	Phosphorylation (CAMK)	SEQ	16.9	13.8	20.6	88.5
		SEQ + SMNet	**56.6**	**49.1**	**63.0**	**91.8**
	Phosphorylation (STE)	SEQ	12.9	7.9	13.0	90.5
		SEQ + SMNet	**73.0**	**54.8**	**45.8**	**93.8**
	O-GalNAc	SEQ	33.8	38.4	78.6	77.9
		SEQ + SMNet	**65.4**	**65.5**	**78.0**	**86.7**
	Acetylation	SEQ	93.0	89.5	82.0	94.7
		SEQ + SMNet	**93.9**	**90.8**	**82.5**	**94.8**

only local sequences. And for phosphorylation in the kinase group CAMK, the measurements are also improved by 15.7%, 16.0%, 59.2% and 0.8%. In addition to phosphorylation, the results for other PTMs such as *O*-GalNAc and acetylation are also listed in Table 1, which show better performance of the proposed method. Taking *O*-GalNAc for instance, by incorporating SMNet profiles, the Sn, MCC, Pre and Acc values are 87.4%, 80.0%, 96.9%, and 90.1%, whereas the corresponding values with only local sequences are 21.3%, 21.5%, 88.5%, and 47.6% on T sites. Similarly, on S sites it can be seen that the SMNet profiles consistently contribute to better performance for different PTMs. For example, the phosphorylation group STE achieves the values of 73.0%, 54.8%, 45.8% and 93.8% for Sn, MCC, Pre and Acc, respectively, which are increased by 60.1%, 46.9%, 32.8% and 3.3% with Sp equal to 95.0%. The detailed results of different PTMs at the high stringency level are listed in Table S2 (ESI†). For the kinase group CAMK, the Sn, MCC, Pre and Acc values obtained using the SMNet method are 14.5%, 15.5%, 14.4%, and 0.5% better than the corresponding values with only local sequences on T sites. Similarly, the Sn, MCC, Pre and Acc of the SMNet method are also 15.7%, 26.5%, 42.4% and 1.3% better on S sites. These results indicate that the SMNet profiles make great contributions to improve the prediction of different

PTMs on S/T sites consistently for all different groups and highlight the importance of incorporating the SMNet profiles in identifying the PTM sites, and further reveal that the SMNet profiles are informative and helpful to improve the prediction performance of PTMs both on S and T sites.

Comparison with existing methods

To further evaluate our method, we compare it with four existing phosphorylation prediction tools, GPS (version 3.0),[20] PPSP,[19] Musite[22] and KinasePhos2.0.[21] We evaluate Musite for CMGC on T sites and CK1 on S sites. Meanwhile, KinasePhos2.0 is performed for CMGC on T sites and CK1 and CAMK on S sites. For the SMNet method, we adopt 2990 PTM local sequences on S sites and 1961 on T sites as training and testing data. It is noteworthy that KinasePhos2.0 and Musite do not report scores for the phosphorylated sites that are not predicted to be modified by phosphorylation. The scores of these sites are set at 0 to plot ROC curves by following the procedure demonstrated by Zou *et al.*,[38] which may sometimes lead to vertical ROC curves. The proposed method employs 10-fold cross-validation, so we implement the PPSP method under the same evaluation procedure. It should be noticed that 10-fold cross-validation is unavailable for GPS, Musite and KinasePhos2.0, so the PTM data in this study is only used as testing data to evaluate the performance, which may result in the over-estimation of the performance of GPS Musite and KinasePhos2.0. However, the proposed method still exhibits very competitive performance. Kinase groups CMGC and STE on T sites and CAMK and CK1 on S sites serve as examples and the phosphorylation prediction performance are displayed in Fig. 3. It indicates that our proposed method is outstanding and remarkable than PPSP, GPS, Musite and KinasePhos2.0. In addition, the detailed AUC values of the 10-fold cross-validation for kinase groups CMGC and STE on T sites and CAMK and CK1 on S sites are shown in Table 2. For CMGC on T sites, the AUC values are 96.0%, 88.3%, 84.1%, 91.3% and 89.9% for the SMNet-SVM, GPS, PPSP, Musite and KinasePhos2.0, respectively. For STE, the AUC value of the proposed method is 94.1%, which is 21.0% and 17.3% higher than GPS and PPSP on T sites, respectively. Meanwhile, on S sites the proposed method also achieves constantly better performance. For CAMK, the AUC value is 16.4%, 19.4%, and 21.2% better than GPS, PPSP and KinasePhos2.0, respectively. Therefore, the SMNet method significantly improves the PTM prediction performance and similar conclusion can also be obtained from the ROC curves for other kinase groups (Fig. S2, ESI†).

Table 2 Comparison of AUC values with different methods for phosphorylation kinase groups on T and S sites

	T		S	
Methods	CMGC (%)	STE (%)	CAMK (%)	CK1 (%)
SMNet-SVM	**96.0**	94.1	89.7	88.6
GPS	88.3	73.1	73.3	68.0
PPSP	84.1	76.8	70.3	70.5
Musite	91.3	—	—	63.8
KinasePhos2.0	89.9	—	68.5	70.1

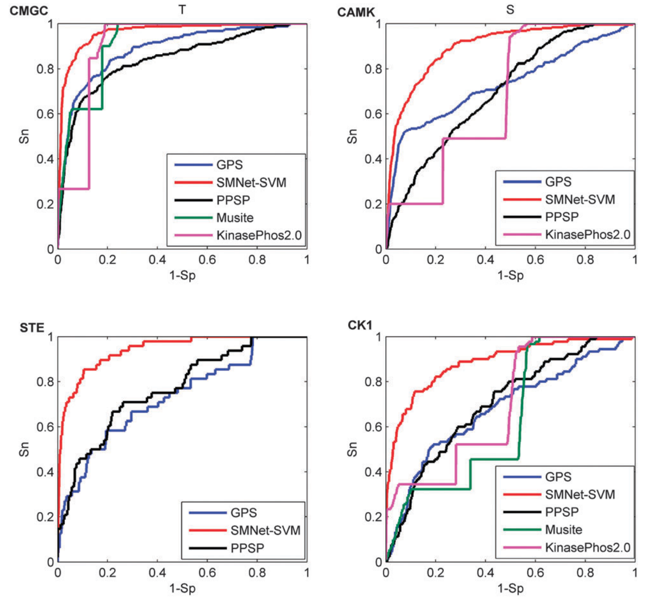


Fig. 3 Performance of phosphorylation ROC curves in kinase groups CMGC, CAMK, STE and CK1 with different methods. The red lines represent the performance of the SMNet method, and the blue, black, purple and green lines show the GPS, PPSP, KinasePhos2.0 and Musite, respectively. The kinase groups CMGC and STE are in response to T sites, and kinase groups CAMK and CK1 are in response to S sites.

To further evaluate our prediction method, we also plotted the Sn-Acc-Pre-MCC bar graph of the five methods to assess the detailed performance for kinase groups CMGC and STE on T sites and CAMK and CK1 on S sites according to the high and medium stringency levels, as shown in Fig. 4. It suggests that the four kinase groups achieve the best performance in almost all circumstances. For example, at the high stringency level with Sp of 99.0%, the 10-fold cross-validation Sn, Acc, Pre, MCC values of the kinase group CK1 are increased by 32.2%, 1.0%, 45.2%, and 38.7% compared with GPS and 33.3%, 1.0%, 48.4%, and 40.6% compared with PPSP on S sites respectively. What is more, for the kinase group CK1 the proposed method also performs better compared with Musite and KinasePhos2.0. For the kinase group CAMK at the high stringency level, the Sn, Acc, Pre, and MCC values are improved by 15.7%, 1.3%, 42.4%, and 26.5% as compared with PPSP and 9.7%, 0.8%, 18.1%, and 14.2% compared with GPS on T sites, respectively. Compared with KinasePhos2.0, it also increased by 2.4%, 0.2%, 4.2%, and 3.4%, respectively. At the medium stringency level, using STE as an example, the proposed method outperforms GPS with about 43.7%, 1.0%, 14.0%, and 25.8% higher Sn, Acc, Pre and MCC values, respectively. Likewise, the proposed method outperforms PPSP by increasing 43.7%, 1.0%, 14.0%, and 25.8% on Sn, Acc, Pre and MCC, respectively. Tables S3 and S4 (ESI†) show the detailed comparative analysis of other kinase groups including Sn, Acc, Pre and MCC values at high and medium stringency levels, respectively. And the proposed method achieves better or comparable performance in phosphorylation groups.

To further assess the prediction performance of our method, besides phosphorylation, we also study other PTMs.

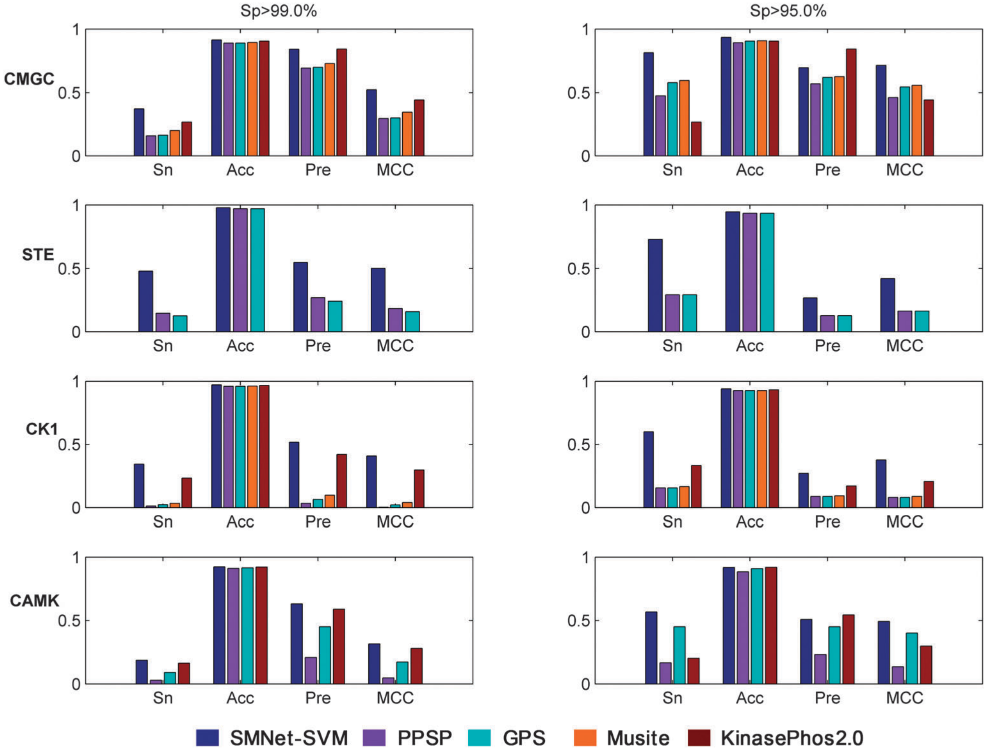


Fig. 4 The Sn, Acc, Pre, and MCC value comparison with different methods for kinase groups TKL and AGC on T sites and CMGC and STE on S sites at two stringency levels. The left part is at a specificity of 99.0%, the measurements in the right one are at a specificity of 95.0%. The horizontal axis represents sensitivity, accuracy, precision and Matthew's correlation coefficient, respectively. SMNet-SVM: the SMNet based on the SVM method.

Taking *O*-GlcNAc as an example, the SMNet method is compared with *O*-GlcNAcPRED, YinOYang and PPSP. As illustrated in Fig. 5, the AUC value of the proposed method is 95.9%, while the AUC values of *O*-GlcNAcPRED, YinOYang and PPSP are 60.2%, 67.4% and 73.6% on T sites, respectively. Furthermore, on S sites the AUC values are also improved by 30.7%, 21.4% and 13.0% compared with *O*-GlcNAcPRED, YinOYang and PPSP. Therefore, it can be seen that the proposed method significantly improves the performance compared with *O*-GlcNAcPRED, YinOYang and PPSP on S/T sites. The comparison of Sn and Sp with four methods at a high and medium specificity on S and T sites is listed in Table 3. From Table 3, it can be noted

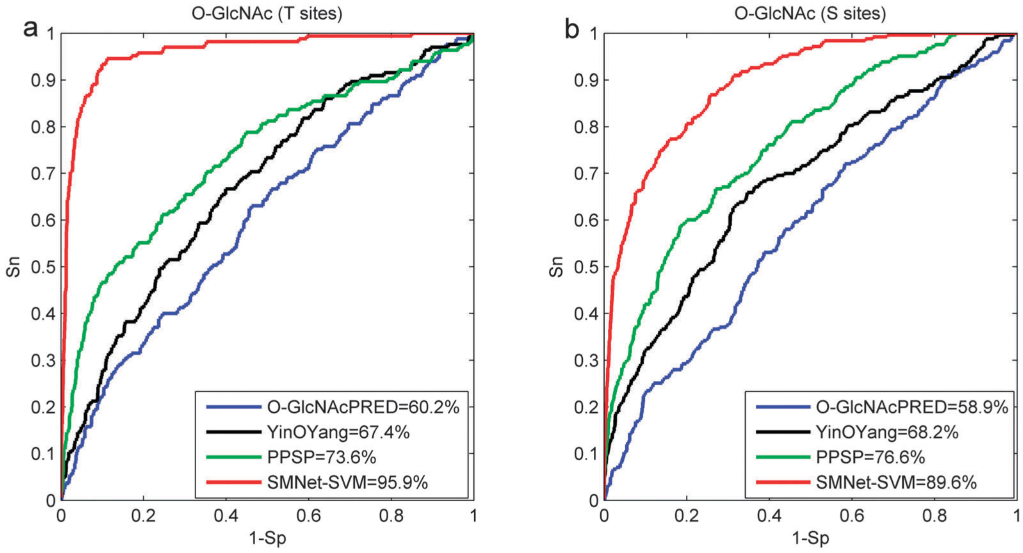


Fig. 5 Performances of four *O*-GlcNAc prediction methods on S and T sites. The red lines represent the proposed SMNet method, and they are compared with the green lines (PPSP), the black lines (YinOYang) and the blue lines (*O*-GlcNAcPRED). The legends in the figure list the AUC values of different methods.

Table 3 Performance comparison of different methods for *O*-GlcNAc site prediction

	Method	Sp (%)	Sn (%)	Sp (%)	Sn (%)
T	*O*-GlcNAcPRED	99.0	3.03	95.0	11.5
	YinOYang		4.85		15.8
	PPSP		10.9		32.7
	SMNet-SVM		**39.4**		**84.2**
S	*O*-GlcNAcPRED	99.0	3.29	95.0	10.3
	YinOYang		10.3		22.2
	PPSP		17.3		32.1
	SMNet-SVM		**25.5**		**56.4**

that the performance of *O*-GlcNAcPRED, PPSP and YinOYang are rather lower than the SMNet method, respectively. For example, on T sites, the proposed method achieves a sensitivity of 39.4% at the high stringency level, which is 36.4%, 34.6% and 28.5% higher than *O*-GlcNAcPRED, YinOYang and PPSP. On S sites the Sn value is increased by 46.1%, 34.2% and 24.3% compared with the other three prediction methods, respectively. We also list other measurements such as MCC, Pre and Acc obtained from different methods in Table S5 (ESI†) and the proposed approach gains the best performance. In summary, the proposed method achieves excellent performance in predicting different types of PTM sites on S/T.

Functional enrichment analysis for different PTMs

For phosphorylation kinase groups CMGC and CK1, *O*-GlcNAc and acetylation, the functional enrichment analysis of the top 100 ranked predicted sites on serine is performed using DAVID[42] and the results are listed in Table S6 (ESI†). There are several important terms enriched in the results. For the kinase group CMGC we find that several GO terms and KEGG pathways are significantly enriched, the GO terms such as ''intracellular signalling cascade'' (*p*-value = 8.93×10^{-8}), ''regulation of cell cycle'' (*p*-value = 6.90×10^{-5}) and ''regulation of cell death'' (*p*-value = 9.84×10^{-5}) are found in the results. Furthermore, the kinase group CMGC is also enriched in the SP_PIR_KEYWORDS term ''phosphoprotein'' (*p*-value = 2.13×10^{-25}). For *O*-GlcNAc, the UP_SEQ_FEATURE term ''glycosylation site: *O*-linked (GlcNAc)'' (*p*-value = 5.92×10^{-9}) is enriched in the results. At the same time there are cellular component related GO terms in the results, including ''nucleus'' (*p*-value = 1.32×10^{-11}), ''cytoskeleton'' (*p*-value = 4.57×10^{-8}) and ''cytoplasm'' (*p*-value = 2.36×10^{-5}). For the predicted acetylation sites, it can be seen from the functional analysis that the SP_PIR_KEYWORDS term ''acetylation'' (*p*-value = 4.1×10^{-20}) is significantly enriched in the top ranked proteins. Some GO terms such as ''protein–DNA complex'' (*p*-value = 1.26×10^{-9}), ''chromatin'' (*p*-value = 1.60×10^{-8}) and ''DNA packaging'' (*p*-value = 4.71×10^{-8}) are also highly enriched.

Analysis of potential *in situ* PTMs

In this study, we examine potential *in situ* PTMs with at least two different PTMs at the same site in the results. We rank all 2990 S sites and 1961 T sites for acetylation according to the scores obtained by the 10-fold cross-validation, respectively. Fig. S3 (ESI†) shows the distribution of 5 kinds of potential *in situ* PTMs, namely Pho-Ace(phosphorylation-acetylation), OGal-Ace(*O*-GalNAc-acetylation), OGlc-Ace(*O*-GlcNAc-acetylation), Pho-OGal-Ace(phosphorylation-*O*-GalNAc-acetylation) and Pho-OGlc-Ace(phosphorylation-*O*-GlcNAc-acetylation) corresponding to the top 100 ranked candidate sites for acetylation on S sites. Pho-Ace, OGal-Ace, OGlc-Ace, Pho-OGal-Ace and Pho-OGlc-Ace mean the sites that are both modified by two or three PTMs. The majority of potential *in situ* PTMs are Pho-Ace with the corresponding proportion of 79%, while OGal-Ace and OGlc-Ace only contribute to 5% and 1% of the top 100 ranked candidate sites. At the same time, we also find that there are totally 15 sites with known *in situ* PTMs (Pho-OGal, Pho-OGlc) that are also potentially modified by acetylation. Furthermore, the details of the top 10 potential *in situ* PTMs and the information related to the corresponding proteins are provided in Table 4. These highly-ranked *in situ* PTMs all belong to Pho-Ace and occur in different proteins. The potential *in situ* PTMs with the largest score (0.800) is Ser2 of POLR2F, a common component of RNA polymerases I, II, and III. Interestingly, we find according to previous studies that this site can be phosphorylated,[34] the records in the UniprotKB database (http://www.uniprot.org/uniprot/P61218#ptm_processing) also show a recent study confirming the acetylation of this site.[43] In addition, in Table S7 (ESI†) we also list the top 10 ranked *in situ* PTMs for acetylation on T sites. In protein EBP, there are two sites Thr2 and Thr3 with already known *in situ* PTMs of OGal-OGlc, and the results suggest additional modifications on these sites. Notably, in the UniprotKB manual assertion on the acetylation of Thr2 has been inferred based on the sequence similarity to P70245 (EBP_MOUSE) (http://www.uniprot.org/uniprot/Q15125#ptm_processing).

Table 4 Information on top 10 potential *in situ* PTMs for acetylation on S sites

Ranking	UniProt ID	Protein name	Position	Known PTM	Potential *in situ* PTMs	Score
1	P61218	POLR2F	2	Phosphorylation	Pho-Ace	0.800
2	P01375	TNF	2	Phosphorylation	Pho-Ace	0.787
3	P67870	CSNK2B	3	Phosphorylation	Pho-Ace	0.739
4	P23528	CFL1	3	Phosphorylation	Pho-Ace	0.667
5	P17931	LGALS3	6	Phosphorylation	Pho-Ace	0.649
6	Q12888	TP53BP1	6	Phosphorylation	Pho-Ace	0.647
7	P22314	UBA1	4	Phosphorylation	Pho-Ace	0.628
8	P06748	NPM1	4	Phosphorylation	Pho-Ace	0.619
9	P67870	CSNK2B	3	Phosphorylation	Pho-Ace	0.595
10	P35611	ADD1	5	Phosphorylation	Pho-Ace	0.590

Conclusions and discussion

In order to overcome the high-cost and labor-intensive shortcomings of PTM site identification using experimental techniques, it is urgent to develop effective computational approaches. Although a number of computational approaches have been proposed, PTM prediction methods usually use local sequence information or functional information for one single PTM without considering the relationship between different PTMs. In this work, we employ the SMNet profiles that make use of the relationship between substrate sites and PTMs by simultaneously borrowing information from multiple PTMs that are both functionally related and statistically dependent. Upon 10-fold cross-validation with the PTM data on S/T sites, the performance of the SMNet method is superior to simply adding the sequence information or the existing tools, indicating that the SMNet profiles can be very helpful in identifying the PTM sites. Furthermore, through the analysis of highly ranked results, we find some important functional enrichment results for different PTMs and we also examine potential *in situ* PTMs that may have intrinsic functional associations. It is anticipated that the predicted results of the SMNet method could be useful for biomedical research and to guide the related experimental validations by providing important clues of the PTM mechanism.

Despite the proposed method showing a good performance, there is still much room for improvement. First, for a completely new candidate protein site that has no prior PTM information, the SMNet method is only able to learn with the local sequence information and hence its performance is limited due to the lack of network information. Second, in this paper we only combine the sequence binary encoding with the SMNet profiles, and other biological information such as protein–protein interaction can be further combined with the SMNet profiles in the future. Finally, currently available *in situ* PTMs are still very limited, leading to comparatively sparse edges in the SMNet. Therefore, it is expected that the performance of the SMNet method will be further improved when more *in situ* PTM data becomes available in the future.

Acknowledgements

This work was supported by National Natural Science Foundation of China (No. 61471331 and 61101061).

References

1 C. Song, M. Ye, Z. Liu, H. Cheng, X. Jiang, G. Han, S. Zhou, Y. Tan, H. Wang, J. Ren, Y. Xue and H. Zou, *Mol. Cell. Proteomics*, 2012, **11**, 1070–1083.

2 X. Xu, A. Li, L. Zou, Y. Shen, W. Fan and M. Wang, *Mol. BioSyst.*, 2014, **10**, 694–702.

3 L. Zhu and N. Li, *Front. Recent Dev. Plant Sci.*, 2013, **3**, 302.

4 G. Manning, D. B. Whyte, R. Martinez, T. Hunter and S. Sudarsanam, *Science*, 2002, **298**, 1912–1934.

5 P. Van den Steen, P. M. Rudd, R. A. Dwek and G. Opdenakker, *Crit. Rev. Biochem. Mol. Biol.*, 1998, **33**, 151–208.

6 S. Zhao, W. Xu, W. Jiang, W. Yu, Y. Lin, T. Zhang, J. Yao, L. Zhou, Y. Zeng, H. Li, Y. Li, J. Shi, W. An, S. M. Hancock, F. He, L. Qin, J. Chin, P. Yang, X. Chen, Q. Lei, Y. Xiong and K.-L. Guan, *Science*, 2010, **327**, 1000–1004.

7 S. A. Beausoleil, J. Villen, S. A. Gerber, J. Rush and S. P. Gygi, *Nat. Biotechnol.*, 2006, **24**, 1285–1292.

8 A. M. Aponte, D. Phillips, R. A. Harris, K. Blinova, S. French, D. T. Johnson and R. S. Balaban, *Methods Enzymol.*, 2009, **457**, 63–80.

9 B. Trost and A. Kusalik, *Bioinformatics*, 2011, **27**, 2927–2935.

10 O. Nørregaard Jensen, *Curr. Opin. Chem. Biol.*, 2004, **8**, 33–41.

11 Z. Songyang, S. Blechner, N. Hoagland, M. F. Hoekstra, H. Piwnica-Worms and L. C. Cantley, *Curr. Biol.*, 1994, **4**, 973–982.

12 J. H. Kim, J. Lee, B. Oh, K. Kimm and I. S. Koh, *Bioinformatics*, 2004, **20**, 3179–3184.

13 S. J. Li, B. S. Liu, R. Zeng, Y. D. Cai and Y. X. Li, *Comput. Biol. Chem.*, 2006, **30**, 203–208.

14 Z. R. Yang, *BMC Bioinf.*, 2009, **10**, 361.

15 B. Trost and A. Kusalik, *Bioinformatics*, 2013, **29**, 686–694.

16 W. Fan, X. Xu, Y. Shen, H. Feng, A. Li and M. Wang, *Amino Acids*, 2014, **46**, 1069–1078.

17 S. E. Hamby and J. D. Hirst, *BMC Bioinf.*, 2008, **9**, 500.

18 M. Hjerrild, A. Stensballe, T. E. Rasmussen, C. B. Kofoed, N. Blom, T. Sicheritz-Ponten, M. R. Larsen, S. Brunak, O. N. Jensen and S. Gammeltoft, *J. Proteome Res.*, 2004, **3**, 426–433.

19 Y. Xue, A. Li, L. R. Wang, H. Q. Feng and X. B. Yao, *BMC Bioinf.*, 2006, **7**, 163.

20 Y. Xue, J. Ren, X. Gao, C. Jin, L. Wen and X. Yao, *Mol. Cell. Proteomics*, 2008, **7**, 1598–1608.

21 Y.-H. Wong, T.-Y. Lee, H.-K. Liang, C.-M. Huang, T.-Y. Wang, Y.-H. Yang, C.-H. Chu, H.-D. Huang, M.-T. Ko and J.-K. Hwang, *Nucleic Acids Res.*, 2007, **35**, W588–W594.

22 J. Gao, J. J. Thelen, A. K. Dunker and D. Xu, *Mol. Cell. Proteomics*, 2010, **9**, 2586–2600.

23 C.-Z. Jia, T. Liu and Z.-P. Wang, *Mol. BioSyst.*, 2013, **9**, 2909–2913.

24 R. Gupta and S. Brunak, *Pac. Symp. Biocomput., 17th*, 2002, 310–322.

25 L. Wells, L. K. Kreppel, F. I. Comer, B. E. Wadzinski and G. W. Hart, *J. Biol. Chem.*, 2004, **279**, 38466–38470.

26 P. Minguez, I. Letunic, L. Parca and P. Bork, *Nucleic Acids Res.*, 2013, **41**, D306–D311.

27 M. Peng, A. Scholten, A. J. R. Heck and B. van Breukelen, *J. Proteome Res.*, 2014, **13**, 249–259.

28 T. Hunter, *Mol. Cell*, 2007, **28**, 730–738.

29 Z. Pan, Z. Liu, H. Cheng, Y. Wang, T. Gao, S. Ullah, J. Ren and Y. Xue, *Sci. Rep.*, 2014, **4**, 7331.

30 X.-J. Yang and E. Seto, *Mol. Cell*, 2008, **31**, 449–461.

31 C. Choudhary, C. Kumar, F. Gnad, M. L. Nielsen, M. Rehman, T. C. Walther, J. V. Olsen and M. Mann, *Science*, 2009, **325**, 834–840.

32 Y. Zhao, J. R. Brickner, M. C. Majid and N. Mosammaparast, *Trends Cell Biol.*, 2014, **24**, 426–434.

33 C.-T. Lu, K.-Y. Huang, M.-G. Su, T.-Y. Lee, N. A. Bretana, W.-C. Chang, Y.-J. Chen, Y.-J. Chen and H.-D. Huang, *Nucleic Acids Res.*, 2013, **41**, D295–D305.

34 F. Diella, S. Cameron, C. Gemund, R. Linding, A. Via, B. Kuster, T. Sicheritz-Ponten, N. Blom and T. J. Gibson, *BMC Bioinf.*, 2004, **5**, 75.

35 P. V. Hornbeck, J. M. Kornhauser, S. Tkachev, B. Zhang, E. Skrzypek, B. Murray, V. Latham and M. Sullivan, *Nucleic Acids Res.*, 2012, **40**, D261–D270.

36 J. Wang, M. Torii, H. Liu, G. W. Hart and Z.-Z. Hu, *BMC Bioinf.*, 2011, **12**, 91.

37 H. Li, X. Xing, G. Ding, Q. Li, C. Wang, L. Xie, R. Zeng and Y. Li, *Mol. Cell. Proteomics*, 2009, **8**, 1839–1849.

38 L. Zou, M. Wang, Y. Shen, J. Liao, A. Li and M. Wang, *BMC Bioinf.*, 2013, **14**, 247.

39 A. K. Biswas, N. Noman and A. R. Sikder, *BMC Bioinf.*, 2010, **11**, 273.

40 T. Li, P. Du and N. Xu, *PLoS One*, 2010, **5**, e15411.

41 C.-C. Chang and C.-J. Lin, *Acm Transactions on Intelligent Systems and Technology*, 2011, **2**, 27.

42 G. Dennis, B. T. Sherman, D. A. Hosack, J. Yang, W. Gao, H. C. Lane and R. A. Lempicki, *Genome Biol.*, 2003, **4**, P3.

43 J. V. Olsen, M. Vermeulen, A. Santamaria, C. Kumar, M. L. Miller, L. J. Jensen, F. Gnad, J. Cox, T. S. Jensen, E. A. Nigg, S. Brunak and M. Mann, *Sci. Signaling*, 2010, **3**, ra3.