Amino Acids (2014) 46:1459–1469
DOI 10.1007/s00726-014-1711-5

ORIGINAL ARTICLE

# PhosphoSVM: prediction of phosphorylation sites by integrating various protein sequence attributes with a support vector machine

**Yongchao Dou · Bo Yao · Chi Zhang**

Received: 23 July 2013 / Accepted: 21 February 2014 / Published online: 13 March 2014


**Abstract** Phosphorylation is one of the most essential post-translational modifications in eukaryotes. Studies on kinases and their substrates are important for understanding cellular signaling networks. Because of the cost in time and labor associated with large-scale wet-bench experiments, computational prediction of phosphorylation sites becomes important and many computational tools have been developed in the recent decades. The prediction tools can be grouped into two categories: kinase-specific and non-kinase-specific tools. With more kinases being discovered by the new sequencing technologies, accurate non-kinase-specific prediction tools are highly desirable for whole-genome annotation in a wider variety of species. In this manuscript, a support vector machine is used to combine eight different sequence level scoring functions to predict phosphorylation sites. The attributes used by this work, including Shannon entropy, relative entropy, predicted protein secondary structure, predicted protein disorder, solvent accessible area, overlapping properties, averaged cumulative hydrophobicity, and *k*-nearest neighbor, were able to obtain better results than the previously used attributes by other similar methods. This method achieved AUC values of 0.8405/0.8183/0.7383 for serine (S), threonine (T), and tyrosine (Y) phosphorylation sites, respectively, in animals with a tenfold cross-validation. The model trained by the animal phosphorylation sites was also applied to a plant phosphorylation site dataset as an independent test. The AUC values for the independent test dataset were 0.7761/0.6652/0.5958 for S/T/Y phosphorylation sites, which compared favorably with those of several existing methods. A web server based on our method was constructed for public use. The server, trained model, and all datasets used in the current study are available at http://sysbio.unl.edu/PhosphoSVM.

**Keywords** Phosphorylation site prediction · Non-kinase-specific tool · Support vector machine

## Introduction

During protein phosphorylation, a phosphate group is added to the protein by a kinase. Phosphorylation is the most essential post-translational modification in eukaryotes and plays a crucial role in a wide range of cellular processes. Studies on kinases and their substrates are important for understanding signaling networks in cells, and helpful for developing new treatments to signaling defect diseases, such as cancer. The number of kinases was estimated to be around 500–1,000 in animals and plants (Caenepeel et al. 2004; Manning et al. 2002; Vlad et al. 2008), and they usually induce phosphorylation on serine (S), threonine (T), tyrosine (Y), and histidine residues in eukaryotic proteins. These phosphorylation sites have been experimentally discovered using techniques such as site-directed mutagenesis and mass spectrometry, in either low-throughput or high-throughput manners (Trost and Kusalik 2011). All experiments on phosphorylation site discovery are time consuming and expensive to perform, and even the high-throughput methods have their own limitations of high false-positive rate resulting from breaking open cells and of high false-negative rate resulting from the low level

**Electronic supplementary material** The online version of this article (doi:10.1007/s00726-014-1711-5) contains supplementary material, which is available to authorized users.

Y. Dou · B. Yao · C. Zhang (✉)
Center for Plant Science and Innovation, School of Biological Sciences, University of Nebraska, Lincoln, NE 68588, USA
e-mail: czhang5@unl.edu

of protein presentation, in addition to associated high cost. Therefore, computational prediction of protein phosphorylation sites, which is often used to narrow down the pool of potential phosphorylation sites in a given protein sequence, becomes increasingly popular as an important complementary approach in protein phosphorylation site studies.

There have been nearly 40 methods for the computational prediction of phosphorylation sites described in the literature since 1999 (Trost and Kusalik 2011). The prediction tools can be grouped into two categories: kinase-specific and non-kinase-specific tools. A kinase-specific prediction program requires as input both a protein sequence and the type of the kinase, and in the end produces some measure of likelihood of which S/T/Y residue in the sequence is phosphorylated by the chosen kinase. In contrast, a non-kinase-specific prediction tool requires only the protein sequence as input, and reports the likelihood of each S/T/Y residue being phosphorylated by any possible kinases. Two recent review articles comprehensively summarized existing methods in phosphorylation site prediction (Trost and Kusalik 2011; Xue et al. 2010). Trost and Kusalik summarized both kinase-specific and non-kinase-specific phosphorylation site prediction tools and provided an overview of the challenges that are faced when designing new prediction methods (Trost and Kusalik 2011). Xue et al. presented a comprehensive summary of kinase-specific phosphorylation sites and phospho-binding motifs prediction methods and a brief introduction of phosphorylation databases (Xue et al. 2010). Non-kinase-specific tools may be able to detect phosphorylation sites for which the associated kinase is unknown or the number of known substrate sequences of the associated kinase is few. With the development of sequencing technology, many genomes of non-model organisms have been sequenced, and more kinases in those species have been discovered, some of which have no sufficient substrate information to train the kinase-specific prediction algorithms. Thus, there is an increased demand for non-kinase-specific tools for a wider variety of species and high specificity for whole-genome annotation (Trost and Kusalik 2011).

Different machine learning methods have been applied to the prediction of post-translational modifications (Basu and Plewczynski 2010). For example, neural network were used for prediction of glycosylation (Julenius et al. 2005; Gupta and Brunak 2002), N-terminal myristoylation in protein sequences (Bologna et al. 2004), and cleavage sites (Blom et al. 1996; Duckert et al. 2004). Support vector machines (SVM) were used for the prediction of lysine acetylation sites (Li et al. 2009) and protein methylation site prediction (Shao et al. 2009). The random forest method was used for the prediction of glycosylation sites (Hamby and Hirst 2008). Some of them were also applied in phosphorylation site prediction. For example, neural networks were used by NetPhos (Blom et al. 1999) and NetPhosK (Hjerrild et al. 2004). The method of SVM was used by Swaminathan's method (Swaminathan et al. 2010) and PredPhospho (Kim et al. 2004). PPSP (Xue et al. 2006) applied Bayesian decision theory for the prediction of kinase-specific phosphorylation sites.

In this work, we report a development of a non-kinase-specific protein phosphorylation site prediction method that uses the SVM method to integrate eight different sequence level scores (PhosphoSVM). These sequence-based attributes are Shannon entropy (SE), relative entropy (RE), protein secondary structure (SS), protein disorder (PD), solvent accessible surface area (ASA), overlapping properties (OP), averaged cumulative hydrophobicity (ACH), and *k*-nearest neighbor profiles (KNN). For evaluation, the area under the receiver operating characteristic (ROC) curve (AUC) was used as the primary metric. With carefully optimized SVM parameters and sliding window size, this method achieved AUC values of 0.8405/0.8183/0.7383 for S/T/Y phosphorylation sites, respectively, in animals, and 0.7761/0.6652/0.5958 for S/T/Y phosphorylation sites, respectively, in plants as an independent test. The server, trained models, and all datasets used in the current study are available at http://sysbio.unl.edu/PhosphoSVM/.

## Results and discussions

Two datasets, Phospho.ELM (P.ELM) version 9.0 (Diella et al. 2008) and PhosPhAt (PPA) version 3.0 (Heazlewood et al. 2008; Durek et al. 2010; Zulawski et al. 2013), were used. All parameters of the SVM model were trained on the dataset of P.ELM. The optimal sets of window size and SVM parameters, $\gamma$ and $C$, were identified to be (21, 0.003, 4), (19, 0.003, 4), and (15, 0.007, 2) for S/T/Y phosphorylation sites, respectively. The optimal AUC values achieved were 0.8405, 0.8183, and 0.7383 for S, T, and Y, respectively. For different phosphorylation residues, the prediction performance was not equal. The performance for residue S was significantly better than the other two. The method with the optimal parameters (i.e., PhosphoSVM) was compared with other six prediction tools.

For the dataset of P.ELM, PhosphoSVM achieved the highest AUC values and Matthews correlation coefficient (MCC) values for all three types of phosphorylation sites, compared with the other six methods. For PhosphoSVM, Swaminathan's method (Swaminathan et al. 2010), and PPRED (Biswas et al. 2010), the results were obtained from a tenfold cross-validation. For NetPhosK, GPS 2.1, NetPhos, and Musite, the tenfold cross-validation was not applied; instead, the trained models from their package were used

**Table 1** Prediction results of seven methods for the dataset of P.ELM, and the tenfold cross-validation is applied to PhosphoSVM, Swaminathan's method (Swaminathan et al. 2010), and PPRED (Biswas et al. 2010)

| Methods | Residue = *S* | | | |
|---|---|---|---|---|
| | AUC | Sn (%) | Sp (%) | MCC |
| NetPhosK | 0.6346 ± 0.0074 | 50.90 | 67.84 | 0.0823 |
| GPS 2.1 | 0.7410 ± 0.0148 | 33.07 | 93.29 | 0.2014 |
| Swaminathan | 0.6965 ± 0.0186 | 31.26 | 88.70 | 0.1257 |
| NetPhos | 0.7019 ± 0.0141 | 34.14 | 86.73 | 0.1234 |
| PPRED | 0.7505 ± 0.0165 | 32.27 | 91.64 | 0.1686 |
| Musite | 0.8065 ± 0.0228 | 41.37 | 93.66 | 0.2492 |
| PhosphoSVM | 0.8405 ± 0.0158 | 44.43 | 94.04 | 0.2979 |
| Methods | Residue = *T* | | | |
| | AUC | Sn (%) | Sp (%) | MCC |
| NetPhosK | 0.6043 ± 0.0094 | 61.96 | 56.79 | 0.0712 |
| GPS 2.1 | 0.6952 ± 0.0126 | 38.10 | 92.30 | 0.2008 |
| Swaminathan | 0.7181 ± 0.0211 | 28.02 | 92.47 | 0.1391 |
| NetPhos | 0.6551 ± 0.0133 | 34.32 | 83.65 | 0.0901 |
| PPRED | 0.7262 ± 0.0179 | 30.31 | 90.99 | 0.1341 |
| Musite | 0.7846 ± 0.0123 | 33.84 | 94.76 | 0.2205 |
| PhosphoSVM | 0.8183 ± 0.0167 | 37.31 | 94.99 | 0.2508 |
| Methods | Residue = *Y* | | | |
| | AUC | Sn (%) | Sp (%) | MCC |
| NetPhosK | 0.6036 ± 0.0341 | 39.52 | 74.22 | 0.0805 |
| GPS 2.1 | 0.6107 ± 0.0266 | 34.49 | 78.86 | 0.0832 |
| Swaminathan | 0.6170 ± 0.0213 | 60.47 | 57.03 | 0.0911 |
| NetPhos | 0.6533 ± 0.0255 | 34.66 | 84.45 | 0.1322 |
| PPRED | 0.7019 ± 0.0222 | 43.04 | 82.65 | 0.1686 |
| Musite | 0.7201 ± 0.0155 | 38.42 | 86.74 | 0.1817 |
| PhosphoSVM | 0.7383 ± 0.0210 | 41.92 | 87.34 | 0.2088 |

AUC values, their standard deviations, Sensitivity (Sn), Specificity (Sp), and MCC are included

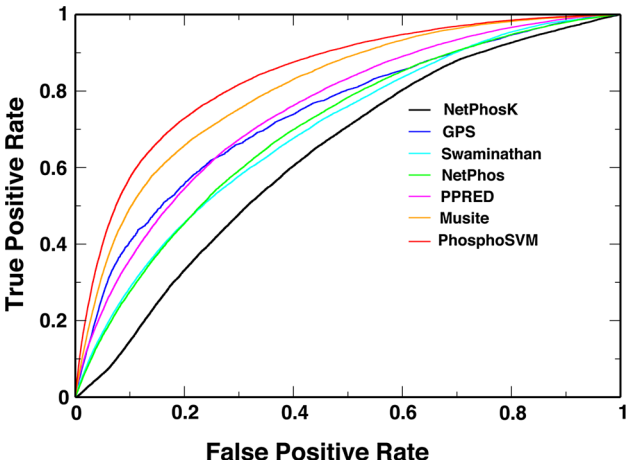


**Fig. 1** ROC curves of different methods on P.ELM, including NetPhosk (*black*), GPS (*blue*), Swaminathan (*cyan*), NetPhos (*green*), PPRED (*magenta*), Musite (*orange*), and PhosphoSVM (*red*) (color figure online)

without any modifications. These tools were applied to the same ten subsets of the data to calculate standard deviations. Table 1 lists in detail the outputs from all six methods on the P.ELM dataset, including the AUC values, sensitivity (Sn), specificity (Sp), and MCC with standard deviations. The results of these six methods we obtained here agreed well with those reported in their original literatures. Figure 1 shows the ROC curves of all methods on S-type phosphorylation sites in P.ELM. The AUC values of PhosphoSVM on all three types of sites were significantly higher than those of the other methods (*P* values: $2.72 \times 10^{-60}$, $4.06 \times 10^{-43}$, and $2.84 \times 10^{-4}$ for S/T/Y, respectively, for the comparison with Musite, which had the second best performance among all seven methods).

Non-kinase-specific prediction tools are expected to discover new phosphorylation sites for unknown kinases. There is a big difference between animal and plant kinases; for example, plants lack AMP activation of peptide kinase (Mackintosh et al. 1992). The performance of non-kinase-specific predictors can be assessed by applying the model trained by the animal dataset (P.ELM) onto a plant (therefore independent) dataset (PPA). After all the physics and chemistry behind the phosphorylation are same, albeit there is a big difference of phosphorylation sites at the sequence level between plants and animals. Therefore, all methods, including PhosphoSVM, were applied onto the PPA dataset for testing. Table 2 shows the detailed results of all methods tested on the PPA dataset. The AUC values of PhosphoSVM for S, T, and Y types of phosphorylation sites were 0.7761, 0.6652, and 0.5958, respectively. Although the AUC values of PhosphoSVM for PPA were lower than those obtained on P.ELM, they remained the best results among all compared methods. According to a generalized *u* test, the AUC value of PhosphoSVM was significantly higher than that of Musite (the second best performer among all seven compared methods) with the combined model for S- and T-type sites (*P* value $<10^{-16}$). For Y-type phosphorylation sites, PhosphoSVM had a higher value of AUC than Musite, but was not statistically significant. Although all MCC values were not very high, the MCC values of the PhosphoSVM's results were also the best ones in the corresponding result group of each phosphorylation site type. The performances of all methods decreased when they were applied onto the plant sequences after being trained on P.ELM, a dataset of primarily animal proteins. This could possibly come from the difference of amino acid compositions between animal and plant phosphorylation sites. The kinase-specific methods, NetPhosK

**Table 2** Prediction results of seven methods to PPA

| Methods | Residue = *S* | | | |
|---|---|---|---|---|
| | AUC | Sn (%) | Sp (%) | MCC |
| NetPhosK | 0.5916 ± 0.0148 | 35.96 | 75.67 | 0.0467 |
| GPS 2.1 | 0.6702 ± 0.0154 | 22.20 | 95.26 | 0.1352 |
| Swaminathan | 0.6341 ± 0.0168 | 25.84 | 88.29 | 0.0749 |
| NetPhos | 0.6437 ± 0.0170 | 28.55 | 87.23 | 0.0805 |
| PPRED | 0.6763 ± 0.0159 | 21.32 | 94.00 | 0.1077 |
| Musite | 0.7269 ± 0.0178 | 28.60 | 95.21 | 0.1827 |
| PhosphoSVM | 0.7761 ± 0.0170 | 34.01 | 95.90 | 0.2371 |
| Methods | Residue = T | | | |
| | AUC | Sn (%) | Sp (%) | MCC |
| NetPhosK | 0.5451 ± 0.0233 | 42.76 | 65.54 | 0.0342 |
| GPS 2.1 | 0.5720 ± 0.0235 | 13.48 | 94.51 | 0.0670 |
| Swaminathan | 0.5541 ± 0.0305 | 31.35 | 76.87 | 0.0380 |
| NetPhos | 0.5548 ± 0.0156 | 27.02 | 80.66 | 0.0380 |
| PPRED | 0.5782 ± 0.0181 | 26.43 | 83.51 | 0.0521 |
| Musite | 0.6220 ± 0.0178 | 15.56 | 95.36 | 0.0976 |
| PhosphoSVM | 0.6652 ± 0.0149 | 21.79 | 93.41 | 0.1155 |
| Methods | Residue = *Y* | | | |
| | AUC | Sn (%) | Sp (%) | MCC |
| NetPhosK | 0.5094 ± 0.0431 | 75.59 | 26.97 | 0.0139 |
| GPS 2.1 | 0.5528 ± 0.0350 | 47.93 | 60.83 | 0.0430 |
| Swaminathan | 0.5120 ± 0.0301 | 66.27 | 36.82 | 0.0154 |
| NetPhos | 0.5549 ± 0.0313 | 63.91 | 46.10 | 0.0483 |
| PPRED | 0.5393 ± 0.0296 | 42.01 | 65.08 | 0.0395 |
| Musite | 0.5876 ± 0.0417 | 28.85 | 81.71 | 0.0647 |
| PhosphoSVM | 0.5958 ± 0.0353 | 28.55 | 84.39 | 0.0840 |

All methods were trained by the dataset of P.ELM. AUC values, their standard deviations, Sensitivity (Sn), Specificity (Sp), and MCC are included

(Hjerrild et al. 2004) and GPS 2.1 (Xue et al. 2011), were significantly affected by this difference. Moreover, some prediction methods are very sensitive to the training data. For example, Swaminathan's method (Swaminathan et al. 2010) showed a significant performance degradation; its MCC value decreased 40 %. Though its performance worsened, among all methods, PhosphoSVM was shown to be the most stable. Currently, more and more genomes of non-model organism have been sequenced, and therefore, PhosphoSVM can be used for whole-genome annotation of phosphorylation sites.

It is worth noting that the performance of kinase-specific methods was inferior to that of some non-kinase-specific methods in terms of AUC or MCC. The application of kinase-specific methods on non-kinase-specific prediction generates more false-positive predictions possibly because the raw scores of a given site predicted by different family models were not normalized and hence not comparable.

**Table 3** Performance of different models by removing one attribute or two attributes

| | AUC | Sn (%) | Sp (%) | MCC |
|---|---|---|---|---|
| ΔSE | 0.8398 ± 0.0147 | 44.01 | 94.09 | 0.2960 |
| ΔRE | 0.8403 ± 0.0141 | 45.46 | 93.74 | 0.2976 |
| ΔSS | 0.8375 ± 0.0140 | 44.38 | 93.95 | 0.2951 |
| ΔPD | 0.8399 ± 0.0145 | 46.37 | 93.48 | 0.2976 |
| ΔASA | 0.8399 ± 0.0156 | 45.57 | 93.72 | 0.2978 |
| ΔOP | 0.8336 ± 0.0160 | 46.83 | 93.22 | 0.2949 |
| ΔKNN | 0.8331 ± 0.0153 | 44.18 | 93.68 | 0.2871 |
| ΔACH | 0.8391 ± 0.0152 | 47.18 | 93.25 | 0.2980 |
| ΔSS + ΔOP | 0.8303 ± 0.0160 | 45.12 | 93.53 | 0.2911 |
| ΔSE + ΔRE | 0.8400 ± 0.0140 | 45.19 | 93.82 | 0.2978 |
| ΔPD + ΔASA | 0.8388 ± 0.0155 | 42.31 | 94.43 | 0.2926 |
| PhosphoSVM | 0.8405 ± 0.0158 | 44.43 | 94.04 | 0.2979 |

All models were tested with tenfold cross-validation on the dataset of P.ELM. AUC values, their standard deviations, Sensitivity (Sn), Specificity (Sp), and MCC are included

High sensitivity is beneficial when predicting phosphorylation sites in a single protein because, in wet-bench studies, experimental biologists may select some candidates from the predicted sites for further experimental design. However, a method with a higher specificity is suitable for whole-genome annotation (Trost and Kusalik 2011). At the maximal F-measure point, PhosphoSVM achieved 94.04 and 95.90 % Sp for S-type phosphorylation sites in P.ELM and PPA, respectively. However, the performance for Y sites was not as satisfactory; the values of Sp were 87.34 and 84.39 % for P.ELM and PPA, respectively. Currently, most methods showed high Sp values, i.e., >80 %, but it does not mean all methods were good in terms of specificity because the ratios of positive to negative sites were very low, at only about 0.04, and specificity was likely to be high for any prediction methods.

To identify which one in all eight different scores played a more important role in prediction, each attribute had been removed from the system, and the same training (except parameters) and testing procedures were conducted to S-type sites with the dataset P.ELM. The data are shown in Table 3. The absence of any one attribute led to some decrease of the AUC value, but none was significant. The largest changes occurred when OP or KNN attributes were turned off, whereas the RE attribute only caused a slight change, 0.0002, of AUC. We also tested the removal of different attribute pairs. Three pairs of attributes, SE + RE, SS + OP, and PD + ASA, were removed and the same procedures were repeated. The changes of AUCs for double-attribute-absence were lower than that of one-

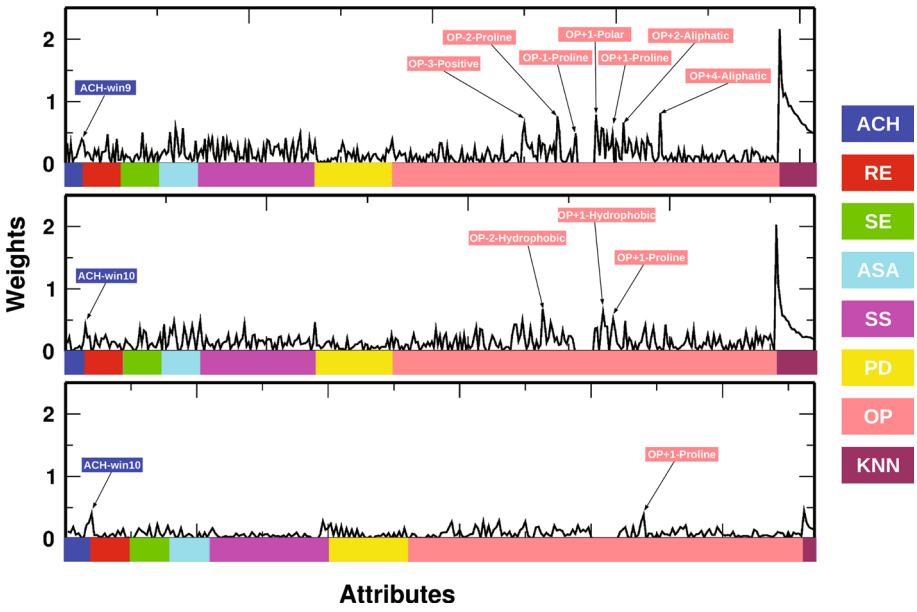


**Fig. 2** Weight vectors of attributes in the trained models for S (*top*), T (*middle*), and Y (*bottom*). Eight attributes are shown in different colors on *x*-axis, and each attribute is shown as a certain length of feature vector. For example, S-type sites have 408 features in total

attribute-absence, though not dramatic. For ΔSE + ΔRE, the AUC was even higher than that of ΔSE. These results indicate that the eight attributes we selected for this algorithm each made their contribution to the overall accuracy.

The weighting vectors of the trained models were extracted for P.ELM. Figure 2 (top panel) shows the weight profiles of the trained model for S-type phosphorylation sites. The weights of KNN profiles were higher than the weights of other scores, which indicates that the KNN attribute is important for prediction and many S-type sites have similar sequence patterns. There were three peaks of OP Proline at +1, −1, −2 positions around the phosphorylation sites. The occurring frequency of Proline at +1 position in all known phosphorylation sites is 31.48 % for the P.ELM dataset and 21.95 % for the PPA dataset. In the background of all proteins, the occurring frequency of Proline is only 3.44 % (Chou and Fasman 1974), which is significantly smaller than the value for Proline at +1 position of phosphorylation sites. This agrees with the previous discovery (Kreegipuu et al. 1998) that Proline at +1 position of S-type phosphorylation site is conserved. Many Prolines around the phosphorylation residue S indicate the coil and/or in a disorder region and correspondingly there were many peaks of weights for SS and PD scores, which is consistent with the previous discovery by Iakoucheva et al. (2004). The other peaks for the OP attribute were Polar residue at +1 position and aliphatic residues at +2 and +4 positions. For the residue T-type of phosphorylation site, the profile of weight vector (Fig. 2 middle panel) was similar to that of residue S. The KNN scores also had high peaks. The Proline at +1 position of a phosphorylation site was assigned a large weight as well. The occurring frequency of Proline at +1 position in all known phosphorylation sites for residue T is 38.37 % for the P.ELM dataset. Interestingly, there were two peaks of OP at position +1 and −2 for hydrophobic amino acids, which was different from the S-type sites. The situation for Y-type phosphorylation sites, however, was different from S and T; the weights for all scores were small and there were fewer peaks (Fig. 2 bottom panel). The overall performance of all existing prediction methods on the prediction of residue Y-type phosphorylation sites was poor. This might be because the sequences of Y-type phosphorylation sites are very diverse, and, on the other hand, the number of known phosphorylation sites for residue Y is too small to train a proper model. Comparing weight profiles, all three types of phosphorylation sites had a peak of ACH with window size = 9 or 10, which indicates that the mean hydrophobicity of 10 residues around the phosphorylation residues is important for phosphorylation site recognition.

The performance of all methods on S-type phosphorylation sites was always better than that for other types. An obvious difference between the S-type dataset and the other two types is the number of known phosphorylation sites. Therefore, one may hypothesize that the performance of the prediction method is correlated with the size of dataset. To test this hypothesis, two subsets of S-type phosphorylation sites were randomly selected from P.ELM datasets so that the numbers of positive sites of subsets were similar to those of datasets of the corresponding T or Y sites. With the same testing procedure, the performance of PhosphoSVM on those subsets was similar to that obtained on the original dataset. For example, the AUC values for two subsets of S sites were 0.8403 and 0.8404, which were similar to the AUC value on the original S-type dataset and

significantly higher than the AUC values for the T- and Y-type datasets. The same procedure was also applied to T-type dataset; a subset of T-type sites was randomly selected to have the same size as the dataset of Y-type sites. The AUC of PhosphoSVM on the subset of T-type sites was 0.8209, which was also similar to the value on the original set and higher than that on the Y-type sites. Therefore, the conclusion is that the performance of predictor is independent of the dataset size in a certain range and that the better prediction performance on S-type phosphorylation sites compared with the other two types is not because of the larger number of known S-type sites.

## Conclusion

A non-kinase-specific protein phosphorylation site prediction method, PhosphoSVM, was developed by applying a SVM on eight sequence attributes. PhosphoSVM achieved AUC values 0.8405/0.8183/0.7383 for S/T/Y phosphorylation sites, respectively, in P.ELM with a tenfold cross-validation. The model trained with P.ELM dataset was applied to an independent PPA dataset, and the AUC values were 0.7761/0.6652/0.5958 for S/T/Y, respectively. In terms of AUC, this method achieved significantly better performance than six existing methods compared in the study. The prediction performance of PhosphoSVM was independent of the size of dataset for a certain range. The server, trained models, and all datasets used in the current study are available at http://sysbio.unl.edu/PhosphoSVM. The score assigned to each site by the web server is the raw score from the SVM.

## Materials and methods

### Datasets

P.ELM version 9.0 (Diella et al. 2008) and PPA version 3.0 (Heazlewood et al. 2008; Durek et al. 2010; Zulawski et al. 2013) were used in this study. Phosphorylation sites in P.ELM were extracted from the scientific literature and phosphoproteomic analyses, while data in PPA were measured by mass spectrum experiments. PPA also provides the results predicted by computer algorithms, but only experimentally measured phosphorylation sites in PPA were used by this project. The sites identified as phosphorylation sites by the computational way were not considered as either positives or negatives. Protein sequences in these two datasets were clustered using BLASTClust (Altschul et al. 1997) with the cutoff of 30 % sequence identity and redundant sequences were removed. Amino acid residues around phosphorylated amino acid residues were kept as the known subsequences of phosphorylation sites, and the length of this subsequence (window size) was a to-be-determined parameter. The similarity between any two subsequences of phosphorylation sites was also checked to ensure the sequence identity was smaller than 30 %. The flowchart of this dataset-construction procedure is shown in Fig. S1 in the supplementary material. Detailed information about these non-redundant data is shown in Table 4. The known phosphorylation sites in P.ELM are mainly for animals, 62 % for *Homo sapiens*, 16 % for *Mus musculus*, 13 % for *Drosophila melanogaster*, and 7 % for *Caenorhabditis elegans*, whereas the PPA set is only for *Arabidopsis thaliana,* a model organism of plants. Therefore, the P.ELM and PPA datasets are relatively independent of each other. For example, for S-type phosphorylation sites, there are 3,037 protein sequences in PPA after applying the filter, but only 168 of them have sequence identities more than 30 % with ones in P.ELM. For T- and Y-type sites, the number of plant protein sequences having similarity more than 30 % with the animal sequences was only 46 and 21, respectively.

**Table 4** The numbers of protein sequences and known phosphorylation sites for P.ELM and PPA datasets

| Dataset | Residue | # of sequences | # of sites |
|---|---|---|---|
| P.ELM | S | 6,635 | 20,964 |
| | T | 3,227 | 5,685 |
| | Y | 1,392 | 2,163 |
| PPA | S | 3,037 | 5,437 |
| | T | 1,359 | 1,686 |
| | Y | 617 | 676 |

In both datasets, experimentally identified phosphorylation sites were considered as positive sites and a subset of the other S/T/Y sites were used as negative ones. In the P.ELM dataset, the ratio of positive to negative sites for S/T/Y is 4.65, 3.77, and 7.66 %, respectively, compared with 3.14, 4.19, and 6.55 % for the PPA dataset. For phosphorylation site prediction model training, it has been shown that the optimal ratio of positive to negative sites is one (Biswas et al. 2010). Therefore, a subset of negative sites was randomly selected for model training so that there were the same numbers of positive and negative phosphorylation sites. All negative sites were ensured to have sequence identity <30 % with any other negative sites and all positive sites.

### Support vector machine setup

#### *Feature vectors*

For classification, we used a SVM classifier, which is a machine learning algorithm with supervised learning

models (Vapnik 1998, 2000). During the training procedure, a SVM constructs a hyperplane by finding support vectors in the high-dimensional space, which is the space for feature vectors (Vapnik 1998, 2000). For a given amino acid residue, a subsequence with all residues adjacent to it in a certain window size was used to create the feature vector for the SVM. A window size, e.g., 7, meant that the given residue had three neighbors on each side in the subsequence. This subsequence was encoded with a multidimensional vector based on eight sequence-based attributes, which were SE, RE, SS, PD, ASA, OP, ACH, and KNN. In the following, each attribute is described in details.

*Shannon entropy*

Shannon entropy score is one of most commonly used sequence conservation measures, and is commonly used for prediction of functionally important residues (Capra and Singh 2007; Mihalek et al. 2004). We used the SE score to quantify the conservation of phosphorylation sites. Shannon entropy was calculated by weighted observed percentages (WOP), which was extracted by PSI-BLAST (Altschul et al. 1997). For a given full-length protein sequence that could potentially have phosphorylation sites, PSI-BLAST was applied to it against the NCBI BLAST Non-redundant protein database. The WOP vector for a position represents the position-specific distribution of 20 amino acids. The SE score for the given position is defined as:

$$\mathrm{SE} = -\sum_{i=1}^{20} p_i \log(p_i) \quad (1)$$

where $p_i = a_i/\Sigma a_j$, $a_j$ is the $j$th value in the WOP vector for this given position. If a position has complete conservation, the SE score has the smallest value, 0.

*Relative entropy*

Relative entropy measures the conservation of amino acids compared with the background distribution, and the deviation from a background distribution is also important for functionally important amino acid residues (Johansson and Toh 2010). Calculating RE scores requires the WOP matrix from PSI-BLAST as well. The RE score of a type of amino acid is calculated as:

$$\mathrm{RE} = \sum_{i=1}^{20} p_i \log\left(\frac{p_i}{p_0}\right) \quad (2)$$

where $p_i = a_i/\Sigma a_j$, $a_j$ is the $j$th value in the WOP vector for this given position and $p_0$ is the protein BLOSUM62 background distribution.

*Secondary structure*

Protein functions are dependent on their structures, and phosphorylation sites are enriched in some specific secondary structures (2004). The secondary structure (SS) attribute describes the structural environment of a phosphorylation site and its surrounding amino acid residues. The most accurate way to obtain the information of secondary structure would be from the 3D structures of proteins, but for a given protein sequence without known 3D structures, currently, the secondary structures can only be predicted. In this manuscript, the SS attribute of each residue has three bits to show the possibility scores of three types of secondary structures (H, E, and C), which were predicted using PSIPRED (McGuffin et al. 2000).

*Protein disorder*

Protein disorder is important for understanding protein function (Iakoucheva et al. 2004). Previous works suggested that PD information is helpful to improve the discrimination between phosphorylation and non-phosphorylation sites (Iakoucheva et al. 2004). In this study, PD was predicted using DISOPRED (Ward et al. 2004), because it is free for downloading and convenient to use. For each residue, the prediction result provides two scores, each between 0 and 1, which correspond to either a structured or disordered status.

*Accessible surface area*

All phosphorylation sites are on the surface of substrate proteins, and hence large solvent accessibility is also an important feature for the catalytic residues. To improve the prediction accuracy, the ASA information of each residue was included into the algorithm as well. For the same reason as for the case of SS, the ASA attribute is also predicted with protein sequences. In this study, RVP-net (Ahmad et al. 2003) was used to predict the relative solvent ASA for each residue in a give protein sequence. Each amino acid residue has a real value in (0, 1) for the ASA attribute. The prediction of ASA does not have high resolution, and the phosphorylation sites that become accessible upon protein conformational changes cannot be evaluated by current existing methods.

*Overlapping properties*

Taylor's OP, reflecting the amino acid groups with common physicochemical properties, were used for identification of protein motifs (Wu and Brutlag 1995), and prediction of T-cell epitopes (Gok and Ozcerit 2012), and prediction of functionally important protein residues (Wu

et al. 2012; Dou et al. 2010, 2012), etc. These properties are: Polar {NQSDECTKRHYW}, Positive {KHR}, Negative {DE}, Charged {KHRDE}, Hydrophobic {AGCTIVLKHFYWM}, Aliphatic {IVL}, Aromatic {FYWH}, Small {PNDTCAGSV}, Tiny {ASGC} and Proline {P} (Taylor 1986). Amino acid residues are encoded using 10-bit vectors where the dimensions of the corresponding properties are set to 1 and remaining positions are 0, i.e., A (0000100010), ……, V (0000110100).

*Average cumulative hydrophobicity*

Average cumulative hydrophobicity (ACH) has been demonstrated to be an important attribute for protein functional important residues (Dou et al. 2012; Zhang et al. 2008; Wang and Brown 2006), because it quantifies the propensity of the propensity of a phosphorylation site amino acid and its surrounding residues to be exposed to solvents. The attribute was quantified by computing the average of the cumulative hydrophobicity indices around the central amino acid residue of a candidate subsequence over the sliding windows with sizes of 3, 5, 7, …, 21, respectively. Therefore, we used ten bits in the feature vector for ACH scores for one given candidate subsequence. Hydrophobicity index proposed by Sweet and Eisenberg (1983) was used in this manuscript, where 20 standard amino acids (A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y) have the values of (0.62, 0.29, −0.90, −0.74, 1.19, 0.48, −0.40, 1.38, −1.50, 1.06, 0.64, −0.78, 0.12, −0.85, −2.53, −0.18, −0.05, 1.08, 0.81, 0.26), respectively.

*K-nearest neighbor profiles (KNN)*

Similar patterns often appear in the local sequences of phosphorylation sites, and this information is helpful for phosphorylation prediction, especially for kinase-specific phosphorylation site prediction. To quantify this kind of information, the KNN score was introduced. A KNN score for one given sequence is the portion of positive phosphorylation sites in its $k$-nearest neighbors in the training set, where the distance between two sequences is proportional to their sequence similarity; a pair of similar sequences has a short distance. In this study, the BLOSUM62 substitution matrix was used to calculate similarities between amino acids, and the sequence similarity was defined as the sum of all amino acid substitution scores. To get a global view of the distribution of similar sequences in the training set, several different $k$ parameters were used to calculate KNN profiles for a given sequence. For S- and T-type sites, the parameter $k$ of KNN was (0.25, 0.5, …, 5 %) of the training dataset, and thus the KNN profile attribute had 20 bits. For Y-type sites, the parameter $k$ was (1, 2, …, 5 %) for five bits because the size of the training set for Y-type sites is small. A KKN attribute vector actually quantifies the neighborhood of a phosphorylation site in the similarity network of all known sites. This feature has been used and described in detail by Musite for phosphorylation site prediction (Gao et al. 2010).

If a residue on a sequence terminus is the central bit of a sliding window, zeros were used to fill in blanks on one side of the window. Attributes KNN and ACH were just applied to the central amino acid of a give subsequence, and hence independent of the size of the windows.

We employed PSIPRED (McGuffin et al. 2000), DISOPRED (Ward et al. 2004), and RVP-net (Ahmad et al. 2003) to predict protein SS, PD, and ASA, and hence their training protein sequences could potentially affect our prediction results if they had high sequence identity with the protein sequences in our dataset. To assess this aspect, we compared the training protein sequences for RVP-net (Ahmad et al. 2003) with our data. Only 0.5 % of the sequences in our data had sequence identity more than 30 % with sequences from RVP-net training dataset. The training protein sequences for PSIPRED (McGuffin et al. 2000) and DISOPRED (Ward et al. 2004) are not available, but we know that the number of proteins that they used as training data was small. For example, DISOPRED (Ward et al. 2004) collected only 750 proteins from PDB as their training sequences. Therefore, the overlap between the datasets we used for phosphorylation site prediction and the training data of these two tools is very small, and hence the effect from the training data of these two tools to the accuracy of phosphorylation site prediction is very limited.

Training/prediction procedure

In this study, we used an SVM package, LIBSVM (Fan et al. 2005), obtained from http://www.csie.ntu.edu.tw/~cjlin/libsvm. The parameters, the window size and parameters of C, the cost, and $\gamma$ for RBF kernel in SVM were optimized on the P.ELM dataset with a tenfold cross-validation. All protein sequences were grouped into ten sets using BLASTClust (Altschul et al. 1997) again, and it ensures the most similar phosphorylation sites in the same set. The reason we did like this is because the filtered known subsequences of phosphorylation sites still have some redundancy. To obtain a sound estimate of the performance of a prediction method, we need to ensure that the overlap between training and evaluation data is minimized. For one round, nine of those ten sets were used to train the model. For training, all positive sites of proteins were used and the same number of negative sites was randomly selected as the positive sites on each protein sequence. All positive and negative sites on proteins in the 10th group were scored by the trained model. After 10

rounds, all positive and negative sites in the whole dataset obtained prediction scores for analysis. The optimal set of parameters resulting in the highest AUC values were obtained by a grid search in the interval of (1, 25) in steps of 2 for the window size (0.001, 0.01) in steps of 0.001 for $\gamma$, and ($2^{-5}$, $2^{4}$) in steps of $\times 2$ for C. Since this method was found to be sensitive to parameter C, an additional fine linear search in (1, 6) in steps of 1 for parameter C was conducted, while the other parameters kept the same grid sizes as before. For the application on the independent dataset and the online server, the prediction model was obtained by training the whole P.ELM dataset with the same number of positive and negative sites, and this model is free for downloading to use on other applications of phosphorylation prediction.

## Evaluation methods

The statistical terms, Specificity (Sp), Sensitivity (Sn), Precision, and MCC are defined in the following equations:

$$S_{\mathrm{p}} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{TP}} \times 100\,\%$$
$$S_{\mathrm{n}} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \times 100\,\%$$
$$P = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \times 100\,\%$$
$$\mathrm{MCC} = \frac{(\mathrm{TP})(\mathrm{TN}) - (\mathrm{FP})(\mathrm{FN})}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \quad (3)$$

where TP, TN, FP, and FN stand for true positive, true negative, false positive, and false negative, respectively. To compare among different algorithms, all Sp and Sn were calculated at the point with the maximal F-measure, which is the balance point of specificity and sensitivity on the ROC curve (Lasko et al. 2005). The *F*-measure is defined in the following equation:

$$F = \frac{2 \times P \times \mathrm{Sn}}{P + \mathrm{Sn}} \quad (4)$$

where $P$ and Sn are Precision and Sensitivity, respectively. The ROC curve represents the interplay of Sn and (1 − Sp). To obtain the ROC curve, all sites in a dataset are sorted with their prediction scores and points of (Sn, 1 − Sp) calculated by increasing the number of returned sites in steps of one site a time. The area under ROC curve (AUC) was used as the primary evaluation metric. A java program available at http://pages.cs.wisc.edu/~richm/programs/AUC/ was used to calculate the AUC. The online tool StAR (DeLong et al. 1988; Vergara et al. 2008) was used to test whether the difference between the ROC curves resulting from two methods is statistically significant. Since, we used tenfold cross-validation, there were ten different subsets of P.ELM data. The standard deviation of AUC values for a method was calculated based on AUC values from these ten subsets. For PPA data, we randomly grouped them into ten subsets to calculate standard deviation of AUC values.

## Six other prediction methods for comparison

The methods being tested against PhosphoSVM in the manuscript included Netphos (Blom et al. 1999), NetPhosK (Hjerrild et al. 2004), Swaminathan's method (Swaminathan et al. 2010), PPRED (Biswas et al. 2010), GPS 2.1 (Xue et al. 2011), and Musite (Gao et al. 2010). These methods were chosen because they have good performance, and also because they either have free downloadable programs or provide user-friendly web servers. All of these methods were obtained from their official websites, except Swaminathan's method (Swaminathan et al. 2010) and PPRED (Biswas et al. 2010), which were implemented by us locally. Though the different version of P.ELM database was used, the local implemented software had the similar results as that from the original articles with the test procedure described in their original articles. Two of them, NetPhosK (Hjerrild et al. 2004) and GPS 2.1 (Xue et al. 2011), are kinase-specific methods, and the others are non-kinase-specific methods. Both NetPhos (Blom et al. 1999) and NetPhosK (Hjerrild et al. 2004) are neural network-based methods for predicting potential phosphorylation in protein sequences. GPS is a group-based phosphorylation site prediction method and the latest version (i.e., v2.1) was used for testing (Xue et al. 2011). Swaminathan's method uses SVM classifier and three sets of sequence-based attributes (Swaminathan et al. 2010). PPRED method uses only PSSM but uses a sliding window to incorporate information of neighboring residues (Biswas et al. 2010). Musite employs the aggregation of multiple SVMs on three attributes: disorder scores, KNN profiles, and amino acid frequency scores (Gao et al. 2010). For the kinase-specific method, all types of kinase family prediction rules were applied to a given peptide sequence, and only the prediction result with the highest score used for analysis. Netphos (Blom et al. 1999) was trained using PhosphoBase, an early version of P.ELM (Kreegipuu et al. 1999). Following their original manuscript, the training sets of Swaminathan's method (Swaminathan et al. 2010) and PPRED (Biswas et al. 2010) were extracted from P.ELM by us. Musite (Gao et al. 2010) extracted species-specific training data from P.ELM and PPA databases, and all data were combined to train a model for generic eukaryotes. The models trained by the combined data, including data from PPA, were used for Musite.

**Acknowledgments** This project was supported by funding under CZ's startup funds from University of Nebraska, Lincoln, NE. The manuscript was written through contributions of all authors. YD designed the study and implemented the algorithm. BY and CZ built the web servers. CZ supervised the whole project. All authors read and approved the final manuscript.

**Conflict of interest** The authors declare that they have no conflict of interest.

## References

Ahmad S, Gromiha MM, Sarai A (2003) RVP-net: online prediction of real valued accessible surface area of proteins from single sequences. Bioinformatics 19(14):1849–1851

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25(17):3389–3402

Basu S, Plewczynski D (2010) AMS 3.0: prediction of post-translational modifications. BMC Bioinforma 11:210. doi:10.1186/1471-2105-11-210

Biswas AK, Noman N, Sikder AR (2010) Machine learning approach to predict protein phosphorylation sites by incorporating evolutionary information. BMC Bioinforma 11:273. doi:10.1186/1471-2105-11-273

Blom N, Hansen J, Blaas D, Brunak S (1996) Cleavage site analysis in picornaviral polyproteins: discovering cellular targets by neural networks. Protein Sci 5(11):2203–2216. doi:10.1002/pro.5560051107

Blom N, Gammeltoft S, Brunak S (1999) Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. J Mol Biol 294(5):1351–1362. doi:10.1006/jmbi1999.3310

Bologna G, Yvon C, Duvaud S, Veuthey AL (2004) N-terminal myristoylation predictions by ensembles of neural networks. Proteomics 4(6):1626–1632. doi:10.1002/pmic.200300783

Caenepeel S, Charydczak G, Sudarsanam S, Hunter T, Manning G (2004) The mouse kinome: discovery and comparative genomics of all mouse protein kinases. Proc Natl Acad Sci USA 101(32):11707–11712. doi:10.1073/pnas.0306880101

Capra JA, Singh M (2007) Predicting functionally important residues from sequence conservation. Bioinformatics 23(15):1875–1882. doi:10.1093/bioinformatics/btm270

Chou PY, Fasman GD (1974) Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. Biochemistry 13(2):211–222

DeLong ER, DeLong DM, Clarke-Pearson DL (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44(3):837–845

Diella F, Gould CM, Chica C, Via A, Gibson TJ (2008) Phospho.ELM: a database of phosphorylation sites—update 2008. Nucleic Acids Res 36(Database issue):D240–D244. doi:10.1093/nar/gkm772

Dou Y, Zheng X, Yang J, Wang J (2010) Prediction of catalytic residues based on an overlapping amino acid classification. Amino Acids 39(5):1353–1361. doi:10.1007/s00726-010-0587-2

Dou Y, Wang J, Yang J, Zhang C (2012) L1pred: a sequence-based prediction tool for catalytic residues in enzymes with the L1-logreg classifier. PLoS One 7(4):e35666. doi:10.1371/journal.pone.0035666

Duckert P, Brunak S, Blom N (2004) Prediction of proprotein convertase cleavage sites. Protein Eng Des Sel 17(1):107–112. doi:10.1093/protein/gzh013

Durek P, Schmidt R, Heazlewood JL, Jones A, MacLean D, Nagel A, Kersten B, Schulze WX (2010) PhosPhAt: the *Arabidopsis thaliana* phosphorylation site database. An update. Nucleic Acids Res 38(Database issue):D828–D834. doi:10.1093/nar/gkp810

Fan RE, Chen PH, Lin CJ (2005) Working set selection using second order information for training support vector machines. J Mach Learn Res 6:1889–1918

Gao J, Thelen JJ, Dunker AK, Xu D (2010) Musite, a tool for global prediction of general and kinase-specific phosphorylation sites. Mol Cell Proteomics 9(12):2586–2600. doi:10.1074/mcp.M110.001388

Gok M, Ozcerit AT (2012) Prediction of MHC class I binding peptides with a new feature encoding technique. Cell Immunol 275(1–2):1–4. doi:10.1016/j.cellimm.2012.04.005

Gupta R, Brunak S (2002) Prediction of glycosylation across the human proteome and the correlation to protein function. Pac Symp Biocomput 7:310–322

Hamby SE, Hirst JD (2008) Prediction of glycosylation sites using random forests. BMC Bioinforma 9:500. doi:10.1186/1471-2105-9-500

Heazlewood JL, Durek P, Hummel J, Selbig J, Weckwerth W, Walther D, Schulze WX (2008) PhosPhAt: a database of phosphorylation sites in *Arabidopsis thaliana* and a plant-specific phosphorylation site predictor. Nucleic Acids Res 36(Database issue):D1015–D1021. doi:10.1093/nar/gkm812

Hjerrild M, Stensballe A, Rasmussen TE, Kofoed CB, Blom N, Sicheritz-Ponten T, Larsen MR, Brunak S, Jensen ON, Gammeltoft S (2004) Identification of phosphorylation sites in protein kinase A substrates using artificial neural networks and mass spectrometry. J Proteome Res 3(3):426–433

Iakoucheva LM, Radivojac P, Brown CJ, O'Connor TR, Sikes JG, Obradovic Z, Dunker AK (2004) The importance of intrinsic disorder for protein phosphorylation. Nucleic Acids Res 32(3):1037–1049. doi:10.1093/nar/gkh253

Johansson F, Toh H (2010) A comparative study of conservation and variation scores. BMC Bioinforma 11:388. doi:10.1186/1471-2105-11-388

Julenius K, Molgaard A, Gupta R, Brunak S (2005) Prediction, conservation analysis, and structural characterization of mammalian mucin-type *O*-glycosylation sites. Glycobiology 15(2):153–164. doi:10.1093/glycob/cwh151

Kim JH, Lee J, Oh B, Kimm K, Koh I (2004) Prediction of phosphorylation sites using SVMs. Bioinformatics 20(17):3179–3184. doi:10.1093/bioinformatics/bth382

Kreegipuu A, Blom N, Brunak S, Jarv J (1998) Statistical analysis of protein kinase specificity determinants. FEBS Lett 430(1–2):45–50

Kreegipuu A, Blom N, Brunak S (1999) PhosphoBase, a database of phosphorylation sites: release 2.0. Nucleic Acids Res 27(1):237–239

Lasko TA, Bhagwat JG, Zou KH, Ohno-Machado L (2005) The use of receiver operating characteristic curves in biomedical informatics. J Biomed Inform 38(5):404–415. doi:10.1016/j.jbi.2005.02.008

Li S, Li H, Li M, Shyr Y, Xie L, Li Y (2009) Improved prediction of lysine acetylation by support vector machines. Protein Pept Lett 16(8):977–983

Mackintosh RW, Davies SP, Clarke PR, Weekes J, Gillespie JG, Gibb BJ, Hardie DG (1992) Evidence for a protein kinase cascade in higher plants. 3-Hydroxy-3-methylglutaryl-CoA reductase kinase. Eur J Biochem 209(3):923–931

Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S (2002) The protein kinase complement of the human genome. Science 298(5600):1912–1934. doi:10.1126/science.1075762

McGuffin LJ, Bryson K, Jones DT (2000) The PSIPRED protein structure prediction server. Bioinformatics 16(4):404–405

Mihalek I, Res I, Lichtarge O (2004) A family of evolution-entropy hybrid methods for ranking protein residues by importance. J Mol Biol 336(5):1265–1282. doi:10.1016/j.jmb.2003.12.078

Shao J, Xu D, Tsai SN, Wang Y, Ngai SM (2009) Computational identification of protein methylation sites through bi-profile Bayes feature extraction. PLoS One 4(3):e4920. doi:10.1371/journal.pone.0004920

Swaminathan K, Adamczak R, Porollo A, Meller J (2010) Enhanced prediction of conformational flexibility and phosphorylation in proteins. Adv Exp Med Biol 680:307–319. doi:10.1007/978-1-4419-5913-3_35

Sweet RM, Eisenberg D (1983) Correlation of sequence hydrophobicities measures similarity in three-dimensional protein structure. J Mol Biol 171(4):479–488

Taylor WR (1986) The classification of amino acid conservation. J Theor Biol 119(2):205–218

Trost B, Kusalik A (2011) Computational prediction of eukaryotic phosphorylation sites. Bioinformatics 27(21):2927–2935. doi:10.1093/bioinformatics/btr525

Vapnik VN (1998) Statistical learning theory. Adaptive and learning systems for signal processing, communications, and control. Wiley, New York

Vapnik VN (2000) The nature of statistical learning theory. Statistics for engineering and information science, 2nd edn. Springer, New York

Vergara IA, Norambuena T, Ferrada E, Slater AW, Melo F (2008) StAR: a simple tool for the statistical comparison of ROC curves. BMC Bioinforma 9:265. doi:10.1186/1471-2105-9-265

Vlad F, Turk BE, Peynot P, Leung J, Merlot S (2008) A versatile strategy to define the phosphorylation preferences of plant protein kinases and screen for putative substrates. Plant J 55(1):104–117. doi:10.1111/j.1365-313X.2008.03488.x

Wang L, Brown SJ (2006) BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. Nucleic Acids Res 34(Web server issue):W243–W248. doi:10.1093/nar/gkl298

Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J Mol Biol 337(3):635–645. doi:10.1016/j.jmb.2004.02.002

Wu TD, Brutlag DL (1995) Identification of protein motifs using conserved amino acid properties and partitioning techniques. Proc Int Conf Intell Syst Mol Biol 3:402–410

Wu CY, Hwa YH, Chen YC, Lim C (2012) Hidden relationship between conserved residues and locally conserved phosphate-binding structures in NAD(P)-binding proteins. J Phys Chem B. doi:10.1021/jp3014332

Xue Y, Li A, Wang L, Feng H, Yao X (2006) PPSP: prediction of PK-specific phosphorylation site with Bayesian decision theory. BMC Bioinforma 7:163. doi:10.1186/1471-2105-7-163

Xue Y, Gao X, Cao J, Liu Z, Jin C, Wen L, Yao X, Ren J (2010) A summary of computational resources for protein phosphorylation. Curr Protein Pept Sci 11(6):485–496

Xue Y, Liu Z, Cao J, Ma Q, Gao X, Wang Q, Jin C, Zhou Y, Wen L, Ren J (2011) GPS 2.1: enhanced prediction of kinase-specific phosphorylation sites with an algorithm of motif length selection. Protein Eng Des Sel 24(3):255–260. doi:10.1093/protein/gzq094

Zhang T, Zhang H, Chen K, Shen S, Ruan J, Kurgan L (2008) Accurate sequence-based prediction of catalytic residues. Bioinformatics 24(20):2329–2338. doi:10.1093/bioinformatics/btn433

Zulawski M, Braginets R, Schulze WX (2013) PhosPhAt goes kinases–searchable protein kinase target information in the plant phosphorylation site database PhosPhAt. Nucleic Acids Res 41(Database issue):D1176–D1184. doi:10.1093/nar/gks1081