A New Machine Learning Approach for Protein Phosphorylation Site Prediction in Plants

Jianjiong Gao[1,2], Ganesh Kumar Agrawal[2,3], Jay J. Thelen[2,3],
Zoran Obradovic[4], A. Keith Dunker[5], and Dong Xu[1,2,*]

[1] Department of Computer Science
[2] C.S. Bond Life Sciences Center
[3] Department of Biochemistry, University of Missouri, Columbia, Missouri 65211
[4] Center for Information Science and Technology, Temple University,
Philadelphia, PA 19122
[5] Center for Computational Biology and Bioinformatics, Indiana University Schools of
Medicine and Informatics, Indianapolis, IN 46202
Tel.: +1 573-884-1887; Fax: +1 573-882-8318
xudong@missouri.edu

Abstract. Protein phosphorylation is a crucial regulatory mechanism in various organisms. With recent improvements in mass spectrometry, phosphorylation site data are rapidly accumulating. Despite this wealth of data, computational prediction of phosphorylation sites remains a challenging task. This is particularly true in plants, due to the limited information on substrate specificities of protein kinases in plants and the fact that current phosphorylation prediction tools are trained with kinase-specific phosphorylation data from non-plant organisms. In this paper, we proposed a new machine learning approach for phosphorylation site prediction. We incorporate protein sequence information and protein disordered regions, and integrate machine learning techniques of k-nearest neighbor and support vector machine for predicting phosphorylation sites. Test results on the PhosPhAt dataset of phosphoserines in *Arabidopsis* and the TAIR7 non-redundant protein database show good performance of our proposed phosphorylation site prediction method.

Keywords: Protein Phosphorylation, Phosphoproteomics, *Arabidopsis*, Protein Disorder, KNN, SVM.

1 Introduction

Reversible protein phosphorylation is one of the most pervasive posttranslational modification mechanisms, regulating diverse cellular processes in various organisms. It has been estimated that about 30% of all proteins in a cell are phosphorylated at any given time [1]. In recent years, publicly available protein phosphorylation data have rapidly accumulated due to large-scale, mass spectrometry studies of protein phosphorylation in different organisms [2-6] and the development of associated phosphorylation web resources [7-11].

[*] Corresponding author.

S. Rajasekaran (Ed.): BICoB 2009, LNBI 5462, pp. 18–29, 2009.

Protein phosphorylation can occur on serine, threonine and tyrosine residues, as well as histidine and aspartate residues in the case of two-component phosphorelays. However, O-linked phosphorylation, specifically on serine residues, is the most common form of phosphorylation in eukaryotes. Despite the increasing number of large-scale phosphorylation studies, experimental identification of phosphorylation sites is still a difficult and time-consuming task. Therefore, more efficient methods for predicting phosphorylation sites *in silico* are desirable. A number of phosphorylation site prediction tools have been developed, including Scansite 2.0 [12], NetPhosK [13], PredPhospho [14], DISPHOS [15], KinasePhos [16], PPSP [17], pkaPS [18], Predikin [19], GPS 2.0 [20], AutoMotif [21] and CRPhos [22]. However, these tools have limitations when predicting phosphorylation sites in plants for two major reasons: (1) they were trained mostly on phosphorylation data from non-plant—mainly mammalian organisms; (2) all of them except DISPHOS, were trained on kinase-specific phosphorylation data and aimed to predict kinase substrate specificities. Meanwhile, the phosphorylation data in plants are not as well annotated as those in mammals, with much less information available on the specificity of phosphorylation sites and their corresponding kinases. Therefore, there is a clear need to train a reliable phosphorylation predictor in plants given the increased frequency of protein kinases in plant genomes and the lack of knowledge about their substrate specificities. With the recently released PhosPhAt database, potential phosphoserines were predicted for the Arabidopsis protein database TAIR7 [23] by support vector machine (SVM) trained on the experimental data collected in the database [10]. Nevertheless, there is room for improvement in prediction accuracy.

In this paper, we proposed a new machine learning approach for phosphorylation site prediction in plants, which integrates features from protein disorder information, nearest neighbors of known phosphorylation sites, and amino acid frequencies in the surrounding sequences of phosphorylation sites to train an SVM for phosphorylation site prediction. The key differences between our method and the previous study [10] are that we incorporated protein disorder prediction and nearest neighbor information in the prediction. A previous study demonstrated that disorder information significantly improved the discrimination between phosphorylation and non-phosphorylation sites [15]. With increasing volume of empirical phosphorylation sites, it is advantageous to use nearest neighbor information. Test results on the PhosPhAt [10] dataset of phosphoserines and the TAIR7 [23] non-redundant protein database indeed shows the remarkable performance of our proposed phosphorylation prediction method.

2 Materials and Methods

Phosphorylation site prediction can be formulated as a binary classification problem, namely each serine/threonine/tyrosine can be classified as either phosphorylation site or non-phosphorylation site. As with all general binary classification problems, there are three key issues: (1) a well-collected and curated dataset including positive and negative data; (2) a set of effective features to characterize the common patterns in each category and the differences between the two categories; (3) a classifier trained from the known data, capable of making reliable predictions for new data. In this study, datasets were extracted from the TAIR7 protein database and PhosPhAt phosphorylation database as discussed in Section 2.1. Outputs from a protein disorder predictor, outputs from k-nearest neighbor predictions and amino acid frequencies around the phosphorylation sites were taken as features as discussed in Section 2.2. We used SVM as the classifier.

2.1 Phosphorylation Dataset

Phosphorylation data in the model organism *Arabidopsis thaliana* collected in PhosPhAt [10] and the *Arabidopsis thaliana* protein database TAIR7 were utilized in this study. Sequences with high similarities were first removed from TAIR7 to build a non-redundant (NR) protein database using BLASTClust in the BLAST package version 2.2.19 with a sequence identity threshold of 30%. As a result, 12,018 representative proteins remain in the TAIR NR database. The PhosPhAt phosphorylation data were then incorporated resulting in 1152 phosphoproteins in the TAIR NR database, which contain 2050 phosphorylation sites, including 1818 phosphoserines, 130 phosphothreonines and 102 phosphotyrosines. We only study phosphoserine events in this paper because of the large number of available data for training and testing. However, the proposed method can be applied to all types of phosphorylation sites.

A 25-residue-long amino acid sequence surrounding each phosphoserine with the phosphoserine in the middle was extracted from each phosphoprotein in the TAIR NR database. Phosphoserines with upstream or downstream less than 12 residues were discarded. As a result, we retrieved a positive set with 1671 sequences surrounding phosphoserines. Similarly, the 433,744 sequences surrounding the non-phosphoserines (serines other than the phosphoserines) were assumed to be the negative set. Although not all these sites are necessarily true negatives, it is reasonable to believe that the vast majority of them are.

2.2 Feature Extraction and Selection

2.2.1 K-Nearest Neighbor Features

Both of the positive and negative sets are very diverse at the sequence level. However, clusters may exist in the positive set, since each phosphorylation site is the substrate of a specific protein kinase, and one kinase could target multiple substrates. It is well known that substrates of the same kinase may share similar patterns in sequence [24]. To take advantage of the cluster information when predicting phosphorylation for a new site (represented by its surrounding sequence), we extracted features from its similar sequences in both positive and negative sets retrieved by a k-nearest neighbor (KNN) algorithm as the following procedure.

i) For a new sequence s, find its k nearest neighbors (NN) in positive and negative sets respectively according to the sequence distance measure defined as follows. For two protein sequences $s_1=\{s_1(-w), s_1(-w+1),\ldots, s_1(w-1), s_1(w)\}$ and $s_2=\{s_2(-w), s_2(-w+1),\ldots, s_2(w-1), s_2(w)\}$, define the distance $Dist(s_1, s_2)$ between s_1 and s_2 as

$$Dist\left(s_1,s_2\right)=1-\frac{\sum_{i=-w}^{w} Sim\left(s_1\left(i\right),s_2\left(i\right)\right)}{2w+1} \tag{1}$$

where w is the length of left/right window (w=12) and *Sim*—amino acid similarity matrix—is derived from the normalized BLOSUM62 [25]:

$$Sim\left(a,b\right)=\frac{Blosum\left(a,b\right)-\min\left\{Blosum\right\}}{\max\left\{Blosum\right\}-\min\left\{Blosum\right\}} \tag{2}$$

where a and b are two amino acids, *Blosum* is the BLOSUM62 matrix, and max/min{*Blosum*} represent the largest/smallest number in the *Blosum* matrix.

ii) The corresponding KNN feature is then extracted as follows

   a) Calculate the average distances from the new sequence s to the k nearest neighbors in the positive and negative sets, respectively.

   b) Calculate KNN score—the ratio of the average distance to the nearest neighbors in the positive set against that in the negative set.

iii) To take advantage of different properties of neighbors with different similarities, repeat (i) and (ii) for different k's to get multiple features for the phosphorylation predictor. In this paper, k was chosen to be 0.1%, 0.2%, 0.5%, 1%, 2%, 5% and 10% of the size of positive/negative sets, and thus 7 KNN scores were extracted as features for the phosphorylation prediction.

2.2.2 Protein Disorder Features

It was observed that sites of posttranslational modifications, including protein phosphorylation sites, are frequently located within disordered regions [15, 26]. In [15], the disorder prediction results for the phosphorylation sites were employed as features to construct a phosphorylation predictor—DISPHOS. In this study, we extracted the disorder information for all surrounding residues of each phosphorylation site and combined them to form a set of disorder features in SVM. The procedure is as follows:

i) For each protein in the TAIR NR database, predict its disordered region using VSL2B [27].

ii) Extract the disorder prediction scores for the surrounding residues in both positive and negative sets, and thus form a vector of 25 scores.

iii) Take the average scores surrounding the sites with different window sizes as features for the phosphorylation predictor. In this paper, we chose the window sizes to be 1, 9 and 25, and thus three disorder features were extracted for each sequence.

2.2.3 Amino Acid Frequency Features

In [15], Iakoucheva *et al.* analyzed the amino acid composition of the surrounding sequences of phosphorylation sites and found that rigid, buried, neutral amino acids (W, C, F, I, Y, V and L) are significantly depleted, while flexible, surface-exposed amino acids (S, P, E, K) are significantly enriched. This conclusion was confirmed by this study as illustrated in Section 3.3. This fact makes the amino acid frequencies good candidates as features for phosphorylation site prediction. In this paper, all 20 amino acid frequencies in each 25-residue sequence were extracted as features for the phosphorylation predictor.

3 Results and Discussions

3.1 KNN Scores as Features

The KNN scores were extracted as features according to the procedure described in Section 2.2.1. A KNN score for a sequence of interest actually compares its average distance (or dissimilarity) to the nearest neighbors (NNs) in the positive set with that in the negative set. A score smaller than 1 means the sequence is more similar to the positive set; a score larger than 1 means more similar to the negative set. The smaller the KNN score, the more similar the sequence is to known phosphorylation sites, and thus the more likely it contains a phosphorylation site.

Figure 1 compares the KNN scores of phosphoserines with non-phosphoserines. Overall the phosphoserines have smaller KNN scores than non-phosphoserines. All of the phosphoserines' average KNN scores with different sizes of NNs are smaller than 1, which means overall the sequences in the positive set are more similar to their NNs in the positive set as expected. It is worth mentioning that such similarities are not due to protein homology as there is no significant sequence similarity between any two proteins in our non-redundant dataset. This finding confirms that phosphorylation-related clusters may exist in the positive set as discussed in Section 2.2.1.

Interestingly, all of the non-phosphoserines' average KNN scores are around 1, which means overall the sequences in the negative set are not predominantly more similar to NNs in either the positive or negative sets. This is not surprising, since phosphorylation-related clusters are unlikely to exist in the negative set, and thus the sequences in the negative set have similar chance to find close neighbors in either positive or negative set.

In short, KNN scores capture the cluster information in phosphoserines, and hence distinguish them from non-phosphoserines. Therefore, KNN scores are suitable to serve as features for the phosphorylation site prediction. The prediction performance of KNNs scores will be demonstrated in Section 3.4.

3.2 Protein Phosphorylation and Disorder

In this section, we will demonstrate that phosphoserines in the dataset we used are predominantly overrepresented in disordered regions, and hence confirm the effectiveness of the disorder scores as features for phosphorylation prediction. Figures 2(A) and 2(B) plot the histograms of the disorder scores of phosphoserines and non-phosphoserines' surrounding residues, respectively. From Fig. 2(A), the number of phosphoserines increases exponentially when the disorder score increases from 0 to 1; the number of phosphoserines with disorder scores larger than 0.9 is much higher than those in the other sub-ranges. In contrast, from Fig. 2(B), there is no such a pattern for the non-phosphoserines. The number of non-phosphoserines with disorder

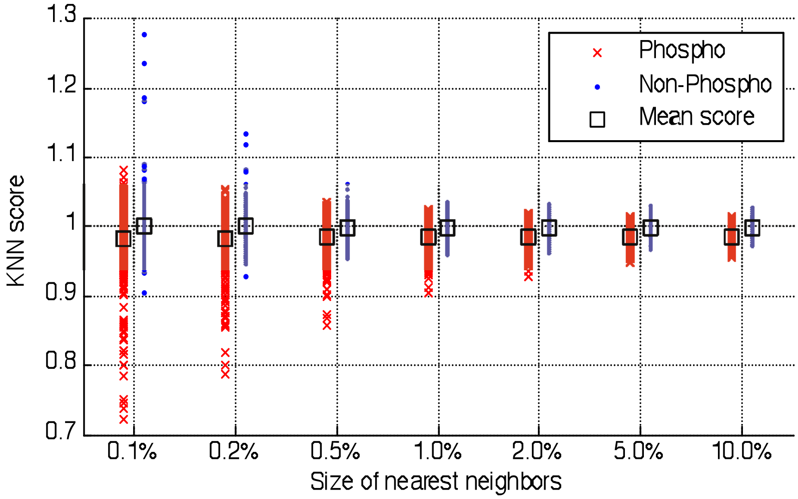

Fig. 1. Comparison of KNN scores in the positive set (1671 sequences around phosphoserines) and those in the negative set (randomly selected 1671 sequences around non-phosphoserines). The horizontal axis represents the size of nearest neighbors (in percentage relative to the size of positive/negative set). The vertical axis represents the KNN score. Each KNN score for one sequence is represented by 'x' (positive data) or dot (negative data). Each square symbol stands for the mean value of KNN scores for each size of NNs.

scores larger than 0.9 is slightly higher than those in the other sub-ranges. This may be because some phosphoserines were not discovered by the experiments in [10] and as a result were incorrectly classified as non-phosphoserines. Alternatively, this could also reflect the general preference of serine in disordered regions. In any case, it is clear that phosphoserines in this dataset are significantly overrepresented in disordered regions. In fact, the majority (~89%) of the phosphoserines have a disorder score larger than 0.5 (Note that VSL2B predicts a residue in the disordered region when its predicted value is larger than 0.5), while this percentage is only ~57% for non-phosphoserines.

3.3 Amino Acid Frequency Features

In this section, we will study the amino acid composition surrounding the phosphoserines. In Figure 3, from left to right, the amino acids vary from being depleted to being enriched in the surrounding sequences of phosphoserines. Similarly as observed in [15], amino acids C, W, Y, F, H, I, L are depleted around phosphoserines, while D, E, R, P and K are enriched. However, S is not significantly enriched around the phosphoserines in this dataset, in contrast to the previous study [15]. The different composition of amino acids surrounding phosphoserines and non-phosphoserines justifies the use of amino acid frequencies as features for the phosphorylation predictor.

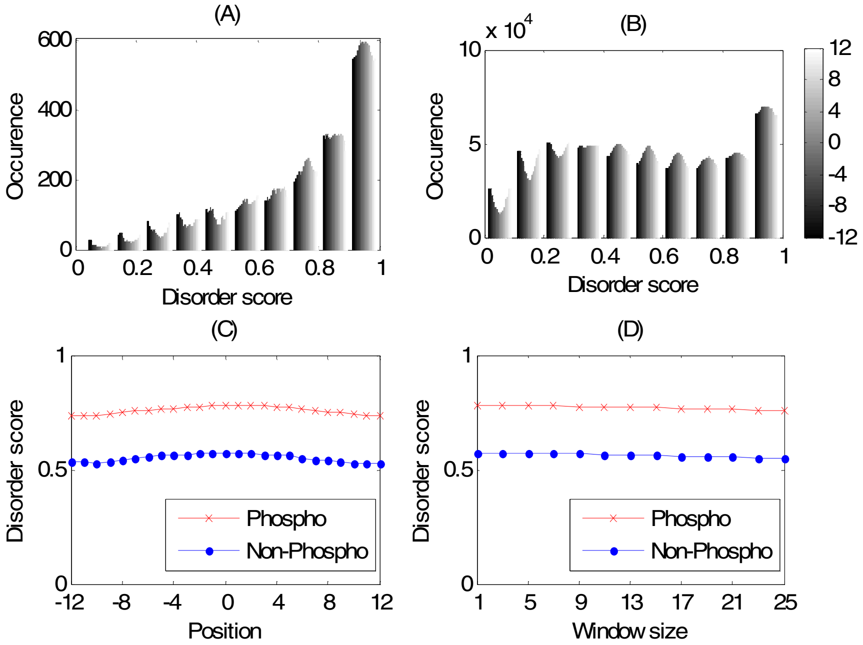

Fig. 2. Preference of phosphorylation sites (serines) in disordered regions. **(A)** Histogram of disorder scores of residues around phosphoserines (1671 in total). The horizontal axis represents the disorder score predicted by VSL2B, divided into 10 sub-ranges from 0 to 1; the vertical axis represents the occurrence (the number of sequences) in the corresponding disorder subrange. Different grayscale from dark to while in each bar stand for 25 different positions in the window from the upstream -12 to downstream 12 as indicated in the *color bar* on the top right. **(B)** Histogram of disorder scores of residues around non-phosphoserines (433,744 in total). **(C)** Disorder scores in the positive and negative sets. The horizontal axis represents the 25 positions (-12 to 12); the vertical axis represents the mean disorder score in the positive set ('x') or the negative set (dot). **(D)** Average disorder scores over windows of different lengths. The horizontal axis represents the window size over which to take average of the disorder scores for each surrounding sequence. The vertical axis represents the mean of those average scores.

3.4 SVM Training and Testing

In this study, an SVM was trained as the classifier between phosphoserines and non-phosphoserines. The SVM$^{\text{light}}$ Version 6.02 [28] was used. The parameters were optimized as '-t 2 -g 1 -c 10 –x 1', which means selecting the kernel as radial basis function with gamma equal to 1, setting C—the tradeoff between training error and margin to 10, and computing the leave-one-out estimate.

As mentioned in Section 2.1, there are 1671 serines in the positive set and 433,744 in the negative set. Testing of the proposed method was performed using the following procedure:

i) Randomly select 1671 samples from the negative set, together with the positive set, and form a balanced dataset of 3342 samples.

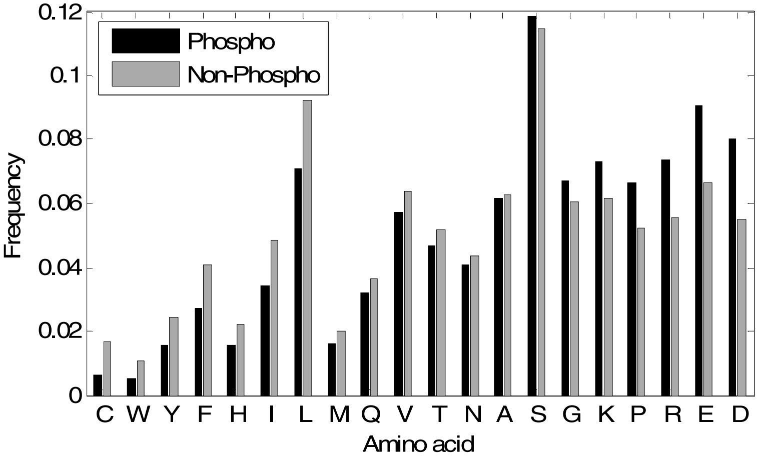

Fig. 3. Amino acid frequencies in the positive and negative sets (the serines in the middle of the 25 residues were excluded; all positive and negative data were used). The vertical axis represents the amino acid frequency. The horizontal axis represents the 20 amino acids sorted in ascending order by the ratio between the amino acid's frequency in the positive set (black) and that in the negative set (gray).

ii) Perform a 10-fold cross validation test: the dataset was partitioned into 10 subsets; a single subset was retained as validation data and the other 9 sets as training data; the cross-validation process is then repeated 10 times, with each subset used exactly once as the validation data. The 10 results were then combined to produce an average estimation.

iii) Note: in each training/test, the disorder and frequency features remained the same. However, the KNN features of each training or validation needed to be re-extracted from the training data, and every time the training data was changed.

The above testing procedure was performed on each separate set of features (amino acid features only, disorder features only, or KNN features only) and combined features (all three sets of features together) 10 times each. Table 1 shows the area under receiver operating characteristic (ROC) curve (AUC) for each test of each set of features, and also the mean AUCs and the standard deviations. Figure 4 shows the mean ROC curves for these tests.

Table 1 and Figure 4 show that all of the three sets of features provide certain predictive powers, but the combined features gave the best test results with the smallest variance (standard deviation) among the 10 random tests. This indicates that combining various features yields more accurate and robust prediction. When testing the features separately, the disorder features were not performed as accurately as the KNN features and frequency features. This may be partially due to fact that all the data came from the same species (*Arabidopsis*). It is unclear whether similar performance can be maintained for cross-species prediction (e.g., training with *Arabidopsis*

Table 1. Prediction performance (AUC) for 10 random tests for different sets of features

Test	Frequency only	Disorder only	KNN only	Combined
1	0.754	0.722	0.816	0.840
2	0.769	0.707	0.806	0.825
3	0.768	0.729	0.812	0.825
4	0.758	0.723	0.796	0.830
5	0.769	0.727	0.813	0.830
6	0.764	0.730	0.794	0.823
7	0.765	0.733	0.813	0.829
8	0.734	0.719	0.819	0.827
9	0.771	0.715	0.816	0.828
10	0.759	0.715	0.793	0.817
Mean	0.761	0.722	0.808	0.827
Standard Deviation	0.011	0.008	0.010	0.006

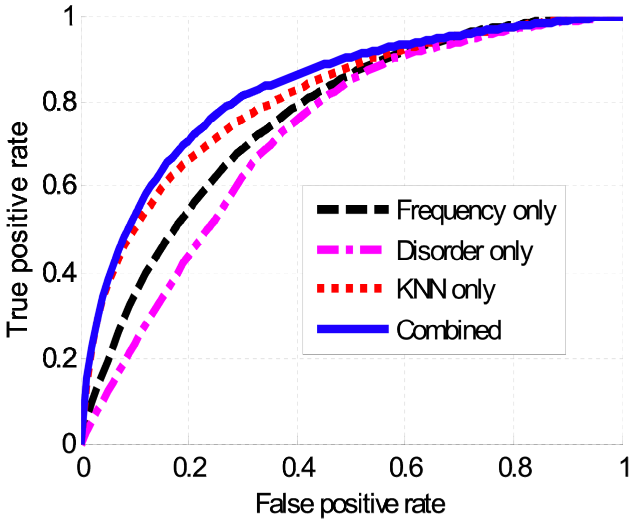

Fig. 4. Mean receiver operating characteristic curves of 10 random tests for different sets of features. The *horizontal axis* represents the false positive rate (the fraction of misclassified samples in the randomly selected negative set); the *vertical axis* represents the true positive rate (the fraction of correctly detected samples in the positive set).

data and predicting phosphorylation sites in soybean). There, the disordered information may be more generic and species-independent.

The phosphoserine predictor in [10] gave a performance of AUC around 0.81 on the redundant *Arabidopsis* TAIR7 protein dataset. It is worth mentioning that for the redundant dataset, the test results of our method achieved 0.84-0.85 on AUC, as KNN may find sequence neighbors in close homologs of the query protein.

4 Conclusion and Future Work

In this paper, we developed a new approach for predicting protein phosphorylation sites in plants. We treated phosphorylation site prediction as a binary classification problem, and employed machine learning techniques to solve it. Multiple features were first extracted from the dataset, including features from nearest neighbors, protein disordered regions and amino acid frequencies. We demonstrated that phosphoserines in the PhosPhAt dataset are predominantly overrepresented in disordered regions. An SVM was then trained based on these features, and used to predict phosphorylation sites in new data. Our method combined both KNN to take advantage of similar known sequence fragments around phosphorylation sites to query protein sequences and SVM to account for other generic features. Test results show good performance of this proposed phosphorylation prediction method. As more phosphorylation sites are experimentally identified, the accuracy of our method is expected to increase automatically.

In future work, we plan to apply our method on phosphothreonines and phosphotyrosines, as well as to the whole proteomes of *Arabidopsis* and other plant species. We will also develop a standalone application and a web service based on this work.

Acknowledgments. This work was supported by the funding from the National Science Foundation-Plant Genome Research Program [grant number DBI-0604439 awarded to JJT] and the National Institute of Health [grant number R21/R33 GM078601 awarded to DX]. The authors wish to thank Dr. Predrag Radivojac, Dr. Jingfen Zhang, and Zhiquan He for helpful discussion and technical assistance.

References

1. Steen, H., Jebanathirajah, J.A., Rush, J., Morrice, N., Kirschner, M.W.: Phosphorylation analysis by mass spectrometry: myths, facts, and the consequences for qualitative and quantitative measurements. Mol. Cell Proteomics 5(1), 172–181 (2006)
2. Olsen, J.V., Blagoev, B., Gnad, F., Macek, B., Kumar, C., Mortensen, P., Mann, M.: Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. Cell 127, 635–648 (2006)
3. Villén, J., Beausoleil, S.A., Gerber, S.A., Gygi, S.P.: Large-scale phosphorylation analysis of mouse liver. Proc. Natl. Acad. Sci. USA 104, 1488–1493 (2007)
4. Chi, A., Huttenhower, C., Geer, L.Y., Coon, J.J., Syka, J.E., Bai, D.L., Shabanowitz, J., Burke, D.J., Troyanskaya, O.G., Hunt, D.F.: Analysis of phosphorylation sites on proteins from Saccharomyces cerevisiae by electron transfer dissociation (ETD) mass spectrometry. Proc. Natl. Acad. Sci. USA 104, 2193–2198 (2007)
5. Benschop, J.J., Mohammed, S., O'Flaherty, M., Heck, A.J., Slijper, M., Menke, F.L.: Quantitative Phosphoproteomics of Early Elicitor Signaling in Arabidopsis. Mol Cell Proteomics 6, 1198–1214 (2007)
6. Sugiyama, N., Nakagami, H., Mochida, K., Daudi, A., Tomita, M., Shirasu, K., Ishihama, Y.: Large-scale phosphorylation mapping reveals the extent of tyrosine phosphorylation in Arabidopsis. Mol. Syst. Biol. 4, 193 (2008)

7. Diella, F., Gould, C.M., Chica, C., Via, A., Gibson, T.J.: Phospho.ELM: a database of phosphorylation sites–update 2008. Nucleic Acids Res. 36(Database issue), D240–D244 (2008)
8. Gnad, F., Ren, S., Cox, J., Olsen, J.V., Macek, B., Oroshi, M., Mann, M.: PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. Genome Biol. 8, R250 (2007)
9. Tchieu, J.H., Fana, F., Fink, J.L., Harper, J., Nair, T.M., Niedner, R.H., Smith, D.W., Steube, K., Tam, T.M., Veretnik, S., Wang, D., Gribskov, M.: The PlantsP and PlantsT Functional Genomics Databases. Nucleic Acids Res. 31, 342–344 (2003)
10. Heazlewood, J.L., Durek, P., Hummel, J., Selbig, J., Weckwerth, W., Walther, D., Schulze, W.X.: PhosPhAt: a database of phosphorylation sites in Arabidopsis thaliana and a plant-specific phosphorylation site predictor. Nucleic Acids Res. 36(Database issue), D1015–D1021 (2008)
11. Gao, J., Agrawal, G.K., Thelen, J.J., Xu, D.: P3DB: a plant protein phosphorylation database. Nucleic Acids Res. 37(Database issue), D960–D962 (2009)
12. Obenauer, J.C., Cantley, L.C., Yaffe, M.B.: Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. Nucleic Acids Res. 31(13), 3635–3641 (2003)
13. Blom, N., Sicheritz-Ponten, T., Gupta, R., Gammeltoft, S., Brunak, S.: Proteomics. Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence 4(6), 1633–1649 (2004)
14. Kim, J.H., Lee, J., Oh, B., Kimm, K., Koh, I.: Prediction of phosphorylation sites using SVMs. Bioinformatics 20(17), 3179–3184 (2004)
15. Iakoucheva, L.M., Radivojac, P., Brown, C.J., O'Connor, T.R., Sikes, J.G., Obradovic, Z., Dunker, A.K.: The importance of intrinsic disorder for protein phosphorylation. Nucleic Acids Res. 32(3), 1037–1049 (2004)
16. Huang, H.D., Lee, T.Y., Tzeng, S.W., Horng, J.T.: KinasePhos: a web tool for identifying protein kinase-specific phosphorylation sites. Nucleic Acids Res. 33(Web Server issue), W226–W229 (2005)
17. Xue, Y., Li, A., Wang, L., Feng, H., Yao, X.: PPSP: prediction of PK-specific phosphorylation site with Bayesian decision theory. BMC Bioinformatics 7, 163 (2006)
18. Neuberger, G., Schneider, G., Eisenhaber, F.: pkaPS: prediction of protein kinase A phosphorylation sites with the simplified kinase substrate binding model. Biol. Direct. 2, 1 (2007)
19. Saunders, N.F., Kobe, B.: The Predikin webserver: improved prediction of protein kinase peptide specificity using structural information. Nucleic Acids Res. 36(Web Server issue), W286–W290 (2008)
20. Xue, Y., Ren, J., Gao, X., Jin, C., Wen, L., Yao, X.: GPS 2.0, a tool to predict kinase-specific phosphorylation sites in hierarchy. Mol. Cell Proteomics 7(9), 1598–1608 (2008)
21. Plewczynski, D., Tkacz, A., Wyrwicz, L.S., Rychlewski, L., Ginalski, K.: AutoMotif Server for prediction of phosphorylation sites in proteins using support vector machine: 2007 update. J. Mol. Model 14(1), 69–76 (2008)
22. Dang, T.H., Van Leemput, K., Verschoren, A., Laukens, K.: Prediction of kinase-specific phosphorylation sites using conditional random fields. Bioinformatics 24(24), 2857–2864 (2008)
23. Swarbreck, D., Wilks, C., Lamesch, P., Berardini, T.Z., Garcia-Hernandez, M., Foerster, H., Li, D., Meyer, T., Muller, R., Ploetz, L., Radenbaugh, A., Singh, S., Swing, V., Tissier, C., Zhang, P., Huala, E.: The Arabidopsis Information Resource (TAIR): gene structure and function annotation. Nucleic Acids Res. 36(Database issue), D1009–D1014 (2008)

24. Kennelly, P.J., Krebs, E.G.: Consensus sequences as substrate specificity determinants for protein kinases and protein phosphatases. J. Biol. Chem. 266, 15555–15558 (1991)
25. Henikoff, S.: Amino acid substitution matrices from protein blocks. Proc. Natl. Acad Sci. USA 89, 10915–10919 (1992)
26. Dunker, A.K., Oldfield, C.J., Meng, J., Romero, P., Yang, J.Y., Chen, J.W., Vacic, V., Obradovic, Z., Uversky, V.N.: The unfoldomics decade: an update on intrinsically disordered proteins. BMC Genomics 9(Suppl. 2), S1 (2008)
27. Obradovic, Z., Peng, K., Vucetic, S., Radivojac, P., Dunker, A.K.: Exploiting heterogeneous sequence properties improves prediction of protein disorder. Proteins 61(suppl 7), 176–182 (2005)
28. Joachims, T.: SVM^{light} Version 6.0.2 (2008), `http://svmlight.joachims.org`