BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 6 2013, pages 686–694
doi:10.1093/bioinformatics/btt031

Sequence analysis

Advance Access publication January 22, 2013

Computational phosphorylation site prediction in plants using random forests and organism-specific instance weights

Brett Trost* and Anthony Kusalik

Department of Computer Science, University of Saskatchewan, Saskatoon, SK S7N 5C9, Canada

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Phosphorylation is the most important post-translational modification in eukaryotes. Although many computational phosphorylation site prediction tools exist for mammals, and a few were created specifically for *Arabidopsis thaliana*, none are currently available for other plants.

Results: In this article, we propose a novel random forest-based method called PHOSFER (PHOsphorylation Site FindER) for applying phosphorylation data from other organisms to enhance the accuracy of predictions in a target organism. As a test case, PHOSFER is applied to phosphorylation sites in soybean, and we show that it more accurately predicts soybean sites than both the existing *Arabidopsis*-specific predictors, and a simpler machine-learning scheme that uses only known phosphorylation sites and non-phosphorylation sites from soybean. In addition to soybean, PHOSFER will be extended to other organisms in the near future.

Availability: PHOSFER is available via a web interface at http://saphire.usask.ca.

Contact: brett.trost@usask.ca

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

Received on September 11, 2012; revised on December 12, 2012; accepted on January 14, 2013

1 INTRODUCTION

Kinase-mediated protein phosphorylation is a critical mechanism for the regulation of virtually all cellular processes in eukaryotes (Bu *et al.*, 2010; Kim and Lee, 2011; Lian *et al.*, 2010; Ressurreição *et al.*, 2011; Uddin *et al.*, 2003; Wang *et al.*, 2010; Wood *et al.*, 2009; Zhang and Johnson, 2000). To fully understand signaling mechanisms in an organism of interest, it is necessary to identify both its protein kinases and the sites that those kinases phosphorylate. Although the protein kinase complement of many organisms is known (Manning *et al.*, 2002), many phosphorylation sites have yet to be identified, particularly in less well-studied organisms.

Although mass spectrometry enables phosphorylation sites to be detected in a high-throughput manner, most laboratories do not have access to the instruments and expertise required to use this technique. As a result, computational methods for predicting phosphorylation sites have become increasingly popular. Dozens of predictors are now available; to review these, see Xue *et al.* (2010) and Trost and Kusalik (2011).

*To whom correspondence should be addressed

Most current predictors focus on human phosphorylation sites. However, the protein kinase complements in various organisms differ significantly both in quantity and in kind (Diks *et al.*, 2007); for example, the plant *Arabidopsis thaliana* encodes twice as many protein kinases as does human (Champion *et al.*, 2004), but it seems to lack any that are similar to classical human tyrosine kinases. This makes most current predictors suboptimal for predicting phosphorylation sites in non-human organisms. Although three tools—PhosPhAt (Durek *et al.*, 2010; Heazlewood *et al.*, 2008), PlantPhos (Lee *et al.*, 2011) and an unnamed tool developed by Gao *et al.* (2009a)—are specific to *Arabidopsis*, predictors are lacking for other plants.

This article describes PHOSFER (PHOsphorylation Site FindER), a phosphorylation site prediction tool for plants, particularly those for which little phosphorylation site data are available. As a test case, we use soybean (*Glycine max*), an economically important crop in many areas of the world. We use a novel strategy for using phosphorylation site data from other organisms to boost predictive performance. Specifically, Basic Local Alignment Search Tool (BLAST) searches are used to determine the degree of conservation between phosphorylation sites in soybean and those in several other organisms for which known phosphorylation sites are available. A machine-learning scheme is used in which a specific training instance from organism X is given a weight proportional to the level of phosphorylation site conservation between soybean and X, with greater weights implying more influence on the learning process. We show that the resultant predictors outperform the aforementioned *Arabidopsis*-specific tools when applied to soybean, and they also outperform a simpler machine-learning technique that uses only known phosphorylation sites from soybean. In the near future, PHOSFER will be extended to predict phosphorylation sites in other organisms.

2 METHODS

2.1 Data

2.1.1 Proteomes The human (*Homo sapiens*), mouse (*Mus musculus*), cow (*Bos taurus*), *Caenorhabditis elegans* and *Drosophila melanogaster* proteomes were obtained from UniProt (Apweiler *et al.*, 2004; UniProt Consortium, 2008, 2012). The *Arabidopsis* proteome was downloaded from The *Arabidopsis* Information Resource (TAIR) (Swarbreck *et al.*, 2008), whereas the yeast (*Saccharomyces cerevisiae*) proteome was downloaded from the *Saccharomyces* Genome Database (Cherry *et al.*, 1998; Engel *et al.*, 2010). Finally, the proteomes for rice (*Oryza sativa*) and soybean were retrieved from the Phytozome project (Goodstein *et al.*, 2012).

2.1.2 Positive phosphorylation site data Phosphorylation sites that have been experimentally characterized using mass spectrometry or low-throughput biological techniques were gathered from online databases for each of the nine organisms aforementioned. Known sites from *C.elegans* were gathered from Phospho.ELM (Diella *et al*., 2004, 2008; Dinkel *et al*., 2011). Both Phospho.ELM and PhosphoSitePlus (Hornbeck *et al*., 2004, 2012) contained sites from human, mouse, cow and *Drosophila*. Known sites from rice, soybean and *Arabidopsis* were downloaded from P^3DB (Gao *et al*., 2009b). Finally, sites from *S.cerevisiae* were obtained from PhosphoGRID (Stark *et al*., 2010).

Although disagreement exists over the optimal peptide length for representing phosphorylation sites in a machine-learning model (Trost and Kusalik, 2011), a few studies have proposed lengths between 9 and 15 (Biswas *et al*., 2010; Blom *et al*., 1999; Miller *et al*., 2008). In this study, phosphorylation sites were represented as peptides of length 15, with the phosphorylated residue in the center and seven amino acids on either side. When a particular phosphorylated residue was too close to the beginning or end of the protein to have seven residues on either side, the missing residues were represented by gap (−) characters. The handling of gaps with respect to the machine-learning features is described in Section 2.2.3. Peptides containing one or more ambiguous amino acids were removed.

2.1.3 Negative phosphorylation site data Negative phosphorylation sites (15mer peptides with S, T or Y central residues that are assumed not to be phosphorylated) for all organisms described earlier were gathered from their respective proteomes as follows. A given S/T/Y residue had to meet three criteria to be selected as a negative site. First, a potential negative site could not have been reported as a positive site. Second, as suggested by Neuberger *et al*. (2007), it had to be within a protein that contained known positive sites. The rationale for this criterion is that as proteins with several known phosphorylation sites have been well studied with respect to phosphorylation, sites in these proteins that are not known to be phosphorylated are more likely to be true negative sites. In this study, a potential negative site had to be in a protein containing at least three positive sites. Third, as suggested by Blom *et al*. (2004), a negative phosphorylation site had to be predicted as solvent-inaccessible; the rationale here is that residues buried in the core of a protein would not be accessible to any kinase. To predict solvent accessibility, the NetSurfP program (Petersen *et al*., 2009) was used. If a given S/T/Y residue was predicted as buried by NetSurfP, it was deemed to be a potential negative phosphorylation site.

2.1.4 Redundant sequence removal To remove redundant sites, all positive and negative sites from all nine organisms were combined into one dataset, which was then clustered using CD-HIT (Li and Godzik, 2006) at a sequence identity threshold of 65%. These clusters were processed using the following rules.

(1) If a cluster contained exactly one site, that site was retained.
(2) If a cluster contained multiple positive (or negative) sites from a single organism, then a single site was arbitrarily chosen to retain.
(3) Some clusters contained positive (or negative) sites from two or more organisms. To avoid redundancy, all but one of these sites were discarded. To choose the site to retain, the organism represented in the cluster with the highest level of phosphorylation site conservation with soybean (i.e. the highest value of C_{Bk}; see Section 2.2.2 for details) was determined. If there was only one site from that organism, that site was retained; otherwise, one of the sites was arbitrarily selected. Given this rule, a site from soybean was always selected if soybean was represented in the cluster.
(4) Because positive and negative data from different organisms were combined, a single cluster could contain both a positive site (from one organism) and a negative site (from a different organism). If a cluster contained at least one positive site and at least one negative site, then that sequence was considered to be a positive site (as the 'negative sites' in the other organisms are likely to be undiscovered positive sites). If the site was known to be positive in more than one organism, then the organism was selected according to the aforementioned rule 3.

2.1.5 Dataset imbalance correction In machine-learning problems, imbalanced datasets occur when one class has a significantly different number of instances than another class and can significantly affect the accuracy of some learning methods (Japkowicz and Stephen, 2002). In the context of phosphorylation site prediction, positive phosphorylation sites are vastly outnumbered by negative sites (Tang *et al*., 2007). To correct this imbalance, for each organism and for each site type (S, T or Y), the number of positive sites was determined, and an equal number of negative sites were randomly chosen from the list generated as described earlier. For example, if 123 positive sites were available for T sites in *Drosophila*, then 123 corresponding negative T sites were chosen.

2.2 Building the classifier

2.2.1 Random forests The random forest machine-learning technique (Breiman, 2001) was used as implemented in the data mining and machine-learning package Weka (Frank *et al*., 2004; Witten *et al*., 2011). This method involves building many decision trees, each of which is built using a number of randomly selected features. The more trees that predict that a given peptide contains a phosphorylation site, the more likely it is that this is indeed the case. Each model built for this study used 300 random trees; each built using 10 randomly selected features. Separate models were created for S, T and Y phosphorylation sites.

2.2.2 Organism-specific instance weights In this study, known phosphorylation sites both from soybean and from other organisms were used as training data. Each training instance with phosphorylated residue k ($k \in \{S, T, Y\}$) from organism B was assigned a weight based on (i) the degree of phosphorylation site conservation (specifically, conservation of 15mer peptides having phosphorylated residues in the center) between soybean and B and (ii) the number of instances of type k in organism B. Training instances from organisms whose phosphorylation sites were better conserved in soybean were given higher weights. Conversely, the more training instances of type k that were available for a given organism, the lower the weight given to each instance. The greater the weight assigned to a particular training instance, the more influence it had on the resultant model.

Formally, let T_{Bk} represent the set of positive training instances from organism B with phosphorylated residue k. The elements of a given set T_{Bk}, as well as the negative training instances for the same B and k, were each given an identical weight W_{Bk} according to the formula $W_{Bk} = 100 \times C_{Bk}/|T_{Bk}|$. The term C_{Bk}, which represents the degree of phosphorylation site conservation between organism B and soybean, is described in more detail later in the text. A scaling factor of 100 was applied to make the resulting numbers less unwieldy.

Each C_{Bk} was calculated as follows. Let A denote the soybean proteome, and let $(A \rightarrow B)_k$ represent the comparison in which all of the known phosphorylation sites from A (i.e. 15mer peptides with the phosphorylated residue in the center) of type k were used as BLAST queries against proteome B (which could be any of the proteomes described in Section 2.1.1, including soybean itself). This was done using all the positive phosphorylation sites for a given organism, not just the ones selected at the end of the filtering process described in Section 2.1.4. Note that for phosphorylated residues occurring within seven residues of the C- or N-terminus of a protein, the BLAST query was <15 residues, with the phosphorylated residue no longer in the middle. Specifically, let X be a known phosphorylation site from A, and let Y be its best BLAST match in B. Also, let X' and Y' denote the full-length proteins corresponding to

X and Y, respectively. X was deemed to be in the 'not conserved' category with respect to B if either X and Y, or X' and Y', were not conserved. X and Y were considered non-conserved if the E-value corresponding to Y was >100 when X was used as a BLAST query against B, or if the number of sequence differences between them was ≥7. X' and Y' were considered non-conserved if the E-value corresponding to Y' was $>10^{-3}$ when X' was used a BLAST query against B. If X was not in the 'not conserved' category according to the aforementioned criteria, then it was placed in the 'conserved' category. An analogous process was also done for the comparison $(B \rightarrow A)_k$ (i.e. in which proteins from some proteome B were used as query sequences, and the soybean proteome was used as the database).

Let H^1_{Bk} denote the percentage of known phosphorylation sites in the 'conserved' category for the comparison $(A \rightarrow B)_k$, and let H^2_{Bk} denote the same for $(B \rightarrow A)_k$. Then $C_{Bk} = (H^1_{Bk} + H^2_{Bk})/2$. For example, if 70% of sites were in the 'conserved' category for $(A \rightarrow B)_k$ and 80% were in the 'conserved' category for $(B \rightarrow A)_k$, then $C_{Bk} = (70 + 80)/2 = 75$. By definition, if B is soybean, then $C_{Bk} = 100$ for each k.

As an illustration of the entire step, suppose that 465 known threonine phosphorylation sites remained from rice after filtering ($|T_{Bk}| = 465$, where B is rice and k is T). Further, suppose that (before filtering) 27.7% of soybean T sites had conserved sites in the rice proteome, and 28.4% of rice T sites had conserved sites in the soybean proteome. Then $C_{Bk} = (27.7 + 28.4)/2 = 28.1$. The weight given to each training instance from rice (both positive and negative) would then be $W_{Bk} = 100 \times C_{Bk}/|T_{Bk}| = 100 \times 28.1/465 = 6.04$.

2.2.3 Features AAIndex (Kawashima *et al.*, 2008; Nakai *et al.*, 1988) is a database of 544 (as of release 9.1) amino acid properties gathered from the literature. Although useful for many bioinformatics tasks (Kawashima *et al.*, 2008), the sheer number of these properties could potentially cause both computational tractability and overfitting problems when used as features in a classification problem. Given that many of these properties are strongly correlated with one another, clustering them can produce a set that is substantially smaller than the full set, but nonetheless, it retains much of its information. Although this has been done by the authors of AAIndex itself using hierarchical clustering (Kawashima *et al.*, 2008; Tomii and Kanehisa, 1996), a more sophisticated method called consensus fuzzy clustering was recently developed by Saha *et al.* (2011). After deriving eight clusters using their technique, these authors identified a set of 24 'high-quality indices' consisting of three individual AAIndex indices from each cluster: the index at the center of the cluster (the medoid) and the two indices farthest from the medoid. The eight clusters roughly represent electric properties, hydrophobicity, alpha and turn propensities, physicochemical properties, residue propensity, composition, beta propensity and intrinsic propensities. A more detailed descriptions of each of these clusters can be found in the original article (Saha *et al.*, 2011); however, to give the reader a sense of these real-valued indices and how they relate (or do not relate) to an intuitive idea of the properties of each amino acid, Table 1 contains these values for 3 of the 24 high-quality indices. Except for the invariant middle residue, the 24 features were considered for each of the residues in a given 15-residue-long phosphorylation site, for $24 \times 14 = 336$ features. As described earlier in the text, missing residues (because of the phosphorylated residue being too close to the N- or C-terminus of the protein) were represented by gap characters (−) in the 15mer representation of phosphorylation sites. To make it as neutral as possible, for each index, the gap character was assigned a value equal to the average value for the 20 amino acids.

2.3 Performance evaluation

2.3.1 Methods compared As described in Section 1, there currently exist three methods for plant-specific phosphorylation site prediction—PhosPhAt (Durek *et al.*, 2010; Heazlewood *et al.*, 2008), PlantPhos (Lee *et al.*, 2011) and a method by Gao *et al.* (2009b). Unfortunately, no implementation is available for the latter technique; therefore, only PhosPhAt and PlantPhos were compared with PHOSFER. Although PhosPhAt and PlantPhos were trained only using data from *Arabidopsis* (and none from soybean), they are nonetheless the two most comparable tools with PHOSFER, with other phosphorylation site tools having been trained on mammalian data (Trost and Kusalik, 2011) (as these tools are not open-source, it was not possible to retrain them using soybean data). As PhosPhAt uses 13mers rather than 15mers (as PHOSFER does), the first and last residues were removed from each peptide before being input to PhosPhAt. PlantPhos uses 21mers; therefore, the 21mer corresponding to each 15mer site (three additional residues on either side) was used. Phosphorylated residues located too close to the beginning or end of the corresponding full protein sequence to make a full 13mer or 21mer could not be tested with PhosPhAt or PlantPhos, respectively. Also, PlantPhos did not return scores for a small portion of the sequences given as input; hence, these sites were considered to have been given a score lower than the minimum of the reported scores.

The performance of PHOSFER was also compared with those of several variants, which we have called PHOSFER-NC, PHOSFER-EW, PHOSFER-SO, PHOSFER-AO and PHOSFER-AO25. Each was identical to PHOSFER except for the following differences. PHOSFER-NC ('no conservation') used the weights $W_{Bk} = 100/|T_{Bk}|$—i.e. each training instance was weighted only according to the number of training instances for that organism and not also according to the phosphorylation site conservation between that organism and soybean. PHOSFER-EW ('equal weights') used equal weights for all training instances, regardless

Table 1. Value corresponding to each amino acid for three arbitrarily-selected high-quality indices from the clustering of amino acid properties performed by Saha *et al.* (2011)

Amino acid	Hydrophobicity	Composition	Physicochemical properties
A	16	0.3	89.3
C	168	0.72	102.5
D	−78	1.26	114.4
E	−106	1.33	138.8
F	189	1.2	190.8
G	−13	3.09	63.8
H	50	1.33	157.5
I	151	0.45	163
K	−141	0.71	165.1
L	145	0.96	163.1
M	124	1.89	165.8
N	−74	2.73	122.4
P	−20	0.83	121.6
Q	−73	0.97	146.9
R	−70	0.9	190.3
S	−70	1.16	94.2
T	−38	0.97	119.6
V	123	0.64	138.2
W	145	1.58	226.4
Y	53	0.86	194.6
—	24	1.19	143.4

Note: There are actually three indices corresponding to each of hydrophobicity, composition and physicochemical properties; the values listed are for an arbitrarily-selected index for each. The gap character represents missing residues in the 15mer peptides.

of the source organism. PHOSFER-SO ('soybean only') was trained exclusively using soybean data; thus, it did not involve the use of instance weights. PHOSFER-AO ('*Arabidopsis* only') was trained exclusively using *Arabidopsis* data; thus, it also did not involve instance weights. Finally, PHOSFER-AO25 ('*Arabidopsis* only 25%') was the same as PHOSFER-AO, except it used only 25% of the *Arabidopsis* data for training.

Evaluating the performance of these variants allows us to assess the contribution of various aspects of PHOSFER, including the machine-learning model (random forests using AAIndex-derived features), the use of data from other species, the use of instance weights and the use of species-specific instance weights. Specifically, comparing PHOSFER-AO with PhosPhAt and PlantPhos allowed us to compare the machine-learning model used here with those used by PhosPhAt and PlantPhos. As there are more *Arabidopsis* data currently available than there were at the time PhosPhAt and PlantPhos were developed, we also tested PHOSFER-AO25, which used fewer *Arabidopsis* training instances than PlantPhos for each type of phosphorylation site (Lee *et al*., 2011) [it is not clear how many sites were used in training the PhosPhAt predictor (Durek *et al*., 2010; Heazlewood *et al*., 2008)]. This allowed the impact of the machine-learning models used to be separated from the impact of a larger training set.

Comparing PHOSFER-SO with PHOSFER-AO allowed us to compare the use of soybean-specific data with the use of *Arabidopsis*-specific data in predicting soybean phosphorylation sites. Comparing PHOSFER-EW with PHOSFER-SO allowed us to evaluate the impact of using, in addition to soybean data, data from organisms other than soybean. Comparing PHOSFER-NC with PHOSFER-EW allowed us to evaluate the impact of weighting the training instances based on the number of instances from each organism. Finally, comparing PHOSFER with PHOSFER-NC allowed us to determine whether there is value in also weighting the training instances based on the degree of phosphorylation site conservation between soybean and the source organism.

2.3.2 Training and testing Because PhosPhAt, PlantPhos, PHOSFER-AO and PHOSFER-AO25 were trained only using data from *Arabidopsis*, they were tested directly on the known soybean data, with no cross-validation needed. PHOSFER-SO was tested using 10-fold cross-validation, with each fold using 90% of the soybean data for training and the remaining 10% for testing. PHOSFER, PHOSFER-NC and PHOSFER-EW were evaluated in the same way, except all of the phosphorylation sites from the non-soybean organisms were used as training instances in each fold in addition to 90% of the soybean data. For completeness, in addition to 10-fold cross-validation, all performance evaluations were also done using leave-one-out cross-validation.

2.3.3 Evaluation criteria For each tool, receiver operating characteristic (ROC) curves were plotted, wherein the *y*-axis represents sensitivity, the *x*-axis represents 1 − specificity and each point represents the sensitivity and specificity of a given tool at a given scoring threshold. The score is the proportion of the 300 decision trees that classify a given residue as a phosphorylation site. Sensitivity was defined as TP/(TP + FN), where TP stands for true-positive sites and FN for false-negative sites. Specificity was defined as TN/(TN + FP), where TN is true-negative sites, and FP is false-positive sites. Each tool was evaluated based on the area under its ROC curve (A_{ROC}), where a value of 0.5 represents classification accuracy that is only as good as random guessing, and a value of 1 represents perfect discrimination. The ROCR package (Sing *et al*., 2005) for the R programming language was used to facilitate the ROC analysis.

Although an A_{ROC} value represents overall classification accuracy, a classifier with a higher A_{ROC} value than another does not necessarily make it more useful. In some applications, it is important to have good sensitivity at very high specificity. For example, suppose that a user wanted to scan an entire proteome for phosphorylation sites. Because there are so many potential sites, specificity must be very high to avoid getting large numbers of false-positive sites. Thus, the best tool for this situation would be the one having the highest sensitivity at very high specificity (say, 0.99). Another application might favor good sensitivity at somewhat lower specificity—for instance, specificities of 0.95 or 0.90 might be appropriate when scanning a limited number of proteins of interest. Given this, the tools were also evaluated according to their sensitivities at the practically useful specificity values of 0.99, 0.95 and 0.90. The Matthews Correlation Coefficient (MCC) was also calculated for each of those specificity values.

3 RESULTS

3.1 Phosphorylation site conservation and organism-specific instance weights

Positive and negative phosphorylation site data were gathered and filtered as described in Section 2.1. Table 2 shows the

Table 2. Summary data on the known phosphorylation sites used in this study

	$\lvert T_{Bk}\rvert$			H^1_{Bk}			H^2_{Bk}			W_{Bk}		
Organism	S	T	Y	S (%)	T (%)	Y (%)	S (%)	T (%)	Y (%)	S	T	Y
Arabidopsis	4738	1212	303	36.0	37.5	58.3	30.7	31.7	46.1	0.70	2.85	17.23
Cow	133[a]	30	17	2.6	4.9	3.6	1.1	3.2	5.3	1.43	13.53	25.98
C.elegans	848	204	25	2.6	2.2	3.6	3.3	4.9	17.1	0.35	1.74	41.43
Drosophila	1107	209	36	2.2	1.6	2.4	1.1	2.8	0.0	0.15	1.05	3.31
Soybean	332	108	49	100.0	100.0	100.0	100.0	100.0	100.0	30.12	92.59	204.08
Human	14246	4911	7251	1.6	3.8	3.6	1.7	4.7	7.3	0.01	0.09	0.08
Mouse	13532	2728	2207	3.3	4.3	3.6	1.7	4.3	9.0	0.02	0.16	0.28
Rice	3396	465	118	24.1	27.7	35.7	24.9	28.4	35.0	0.72	6.04	29.97
Yeast	3549	834	38	2.5	4.3	3.6	3.4	5.7	11.4	0.08	0.60	19.74

Note: All information is shown separately for each phosphorylated residue *k* (S, T and Y) and each organism *B*. Column headings correspond to the notation used in Section 2. $|T_{Bk}|$ indicates the number of training instances that remained after filtering. H^1_{Bk} indicates the percentage of phosphorylation sites in the soybean proteome that had conserved sites in the indicated organism's proteome. H^2_{Bk} indicates the percentage of phosphorylation sites in the indicated organism's proteome that had conserved sites in the soybean proteome. Finally, W_{Bk} indicates the weight given to each individual training instance according to the formula given in Section 2.2.2.

[a]Cow had only 92 negative S sites remaining after the filtering procedures described in Sections 2.1.3 and 2.1.4; hence, in this case, the number of negative sites was less than the number of positive sites.

number of positive S, T or Y sites from each organism (the quantities $|T_{Bk}|$ described earlier) following the removal of redundant sequences. The number of negative sites used for a given organism was made to be the same as the number of positive sites for that organism; however, because of the filtering steps performed in Sections 2.1.3 and 2.1.4, cow had too few negative S sites remaining to match the number of positive sites; in this case, all possible negative sites were used. Table 2 also shows the level of phosphorylation site conservation between each organism and soybean; these numbers were used to calculate the values C_{Bk}. Finally, Table 2 contains the instance weight W_{Bk} for each combination of B and k.

3.2 Performance of PHOSFER, the PHOSFER variants, PhosPhAt and PlantPhos

As mentioned earlier, the performance of the primary classifier (PHOSFER) was tested, along with those of a number of variants: PHOSFER-NC (trained using instance weights that take into account only the number of training instances from a given organism, and not phosphorylation site conservation with soybean), PHOSFER-EW (trained using equal instance weights for all organisms), PHOSFER-SO (trained using only soybean phosphorylation sites), PHOSFER-AO (trained using only *Arabidopsis* sites) and PHOSFER-AO25 (trained using only 25% of the available *Arabidopsis* sites). Figure 1 contains ROC curves illustrating the performance of PhosPhAt and PlantPhos, both of which were trained only using data from *Arabidopsis*, and compares them to PHOSFER-AO and PHOSFER-AO25. Figure 2 contains ROC curves for the first four PHOSFER variants as aforementioned. These data were a result of using 10-fold cross-validation; results using leave-one-out cross-validation can be found in the Supplementary Materials. In addition, Table 3 contains the A_{ROC} value for each tool and each site type, as well as sensitivity at various practically useful specificity values.

In Section 2.3.1, the purpose of including each of the PHOSFER variants was explained. The results given in Figure 1, Figure 2 and Table 3 allow the comparisons mentioned in that section to be made.

Figure 1 and Table 3 show that, for all site types, both PHOSFER-AO and PHOSFER-AO25 outperformed PhosPhAt and PlantPhos. All four of these tools were trained using data from *Arabidopsis* and then tested on soybean data, and although PHOSFER-AO had the advantage of a greater amount of training data than PlantPhos (as mentioned earlier in the text, it is unclear how many training instances were used for PhosPhAt), PHOSFER-AO25 had fewer training instances than PlantPhos for all three site types. This implies that the model used here, which uses random forests and AAIndex-derived features, compares favorably with the models used by PlantPhos (and probably PhosPhAt).

As expected, Figure 2 and Table 3 show that PHOSFER-SO had the lowest performance among the variants of PHOSFER that used soybean data for training. This was the case for all three types of phosphorylation sites. PHOSFER-EW, which was trained using equally weighted data from soybean and other organisms, exhibited comparable performance with PHOSFER-SO for S sites, but it greatly improved performance for T and Y sites, for which less data were available from soybean. PHOSFER-NC and PHOSFER had comparable A_{ROC} values with PHOSFER-EW for S and T sites and improved A_{ROC} values for Y sites. Finally, PHOSFER and PHOSFER-NC had comparable A_{ROC} values, but PHOSFER generally had slightly to moderately better sensitivity at high specificity than PHOSFER-NC.

Perhaps the most surprising observation from Table 3 is that the performance of PHOSFER-AO rivaled (and sometimes bettered) that of PHOSFER, both in terms of A_{ROC} values and in terms of sensitivity at high specificities. This observation is discussed further in Sections 4.3 and 4.4.

3.3 The relationship between improvements in performance and the amount of available data

Given the previous observations, it seems that using data from other species provide substantial benefit when few known phosphorylation sites from the organism of interest are available (T and Y sites in this case), but a more modest benefit when many sites are available (S sites). To more explicitly examine this phenomenon, three subvariants of PHOSFER (PHOSFER75, PHOSFER50 and PHOSFER25) and PHOSFER-SO (PHOSFER-SO75, PHOSFER-SO50 and PHOSFER-SO25) were created, which used 75, 50 or 25%, respectively, of the available soybean data. The performance of each tool was evaluated using 10-fold cross-validation. The results are presented in Table 4, which suggests that the aforementioned conjecture may be at least partially incorrect. We expected the performance of PHOSFER-SO to degrade when using smaller amounts of soybean training data, but the performance of PHOSFER to remain essentially the same; however, the performance of *both* tools remained essentially unchanged even when using only 25% of the soybean data. This may indicate that the greater improvement in performance between PHOSFER and PHOSFER-SO for T and Y sites relative to S sites cannot be attributed solely to the greater amount of soybean data available for S sites. Although we cannot pinpoint with confidence an alternative explanation for this difference, it is possible that the patterns governing T and Y site recognition are more complex than those governing S site recognition, and thus benefit more from the cross-species phosphorylation site data used by PHOSFER. In any case, the fact that PHOSFER-SO, PHOSFER-SO75, PHOSFER-SO50 and PHOSFER-SO25 performed similarly shows that the machine-learning model used here (random forests using AAIndex indices as features) is robust in the face of different amounts of training data.

4 DISCUSSION

4.1 Phosphorylation site conservation

As shown in Table 2, the different organisms varied greatly in the degree to which their phosphorylation sites were conserved in soybean, and vice versa. Although the numbers varied somewhat depending on the exact organism, in general, the percentage of phosphorylation sites shared between soybean and another organism was ~10 times higher in the plants (*Arabidopsis* and rice) than in the non-plant organisms. For example, 24.1% of S sites

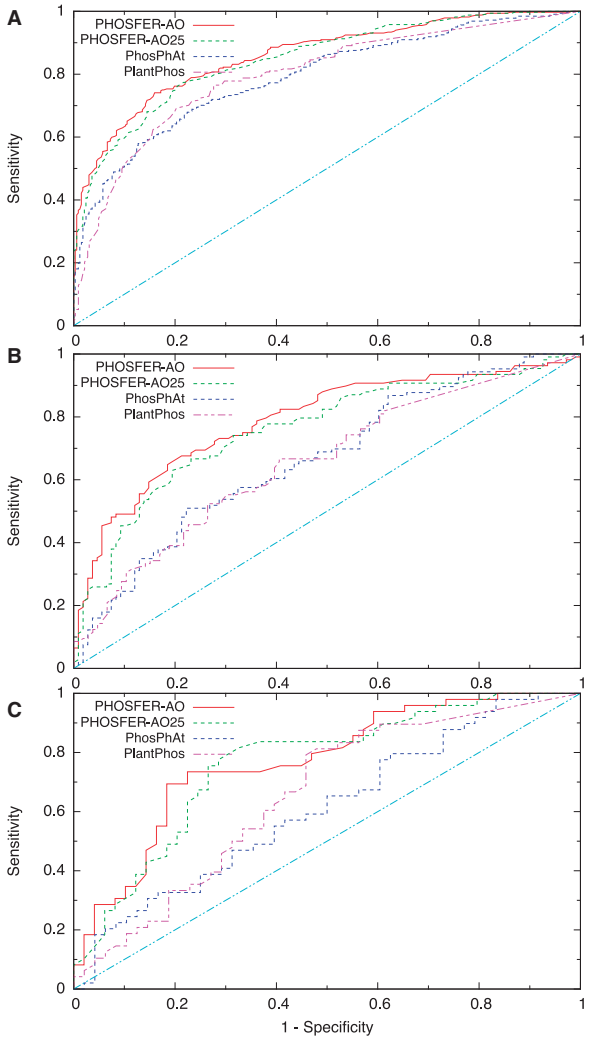

Fig. 1. ROC curves for PHOSFER-AO, PHOSFER-AO25, PhosPhAt and PlantPhos for (**A**) S phosphorylation sites, (**B**) T phosphorylation sites and (**C**) Y phosphorylation sites. The diagonal line denotes the expected performance of a tool that uses random guessing

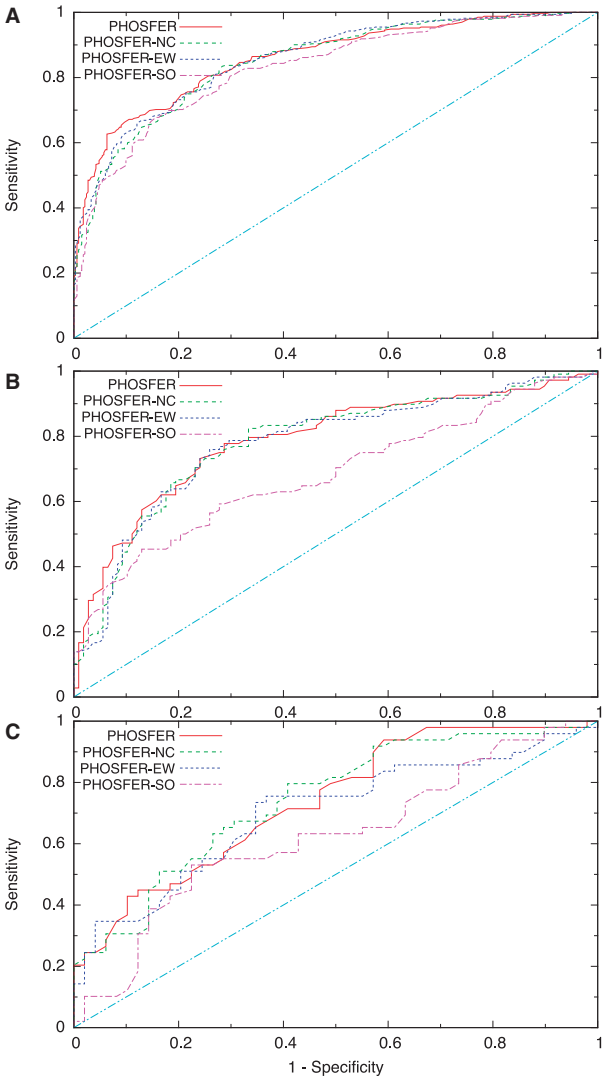

Fig. 2. ROC curves for PHOSFER and variants for (**A**) S phosphorylation sites, (**B**) T phosphorylation sites and (**C**) Y phosphorylation sites. The diagonal line denotes the expected performance of a tool that uses random guessing

in rice had a conserved site in soybean compared with just 2.6% of cow sites.

The huge disparity in phosphorylation site conservation among the different organisms means that the information provided by a training instance from one organism (e.g. human) may not be as relevant to the decision problem as one from another organism (e.g. *Arabidopsis*). This was the motivation behind the use of the C_{Bk} terms when calculating instance weights for PHOSFER.

4.2 Kinase specificity

Although PHOSFER provides respectable accuracy for predicting phosphorylation sites in soybean, its accuracy is still less than that of most predictors that focus on human sites (Xue *et al.*, 2010). A portion of this underperformance could be attributed to the smaller number of known phosphorylation sites in soybean relative to human. However, a more important factor is likely the lack of information regarding the kinases responsible for phosphorylating soybean phosphorylation sites. From years of low-throughput laboratory experiments, the kinases responsible for phosphorylating many human sites are known. For example, there are currently 4985 human sites in the PhosphoSitePlus database (Hornbeck *et al.*, 2004, 2012) for which the corresponding kinase is known. This information allows the creation of kinase-specific predictors [the majority of current human predictors (Trost and Kusalik, 2011)], which typically have greater accuracy than non-specific predictors (Neuberger *et al.*, 2007). Individual kinases, as well as families of kinases, have characteristic recognition patterns, and it is likely easier to model such recognition patterns than those of kinases in general. In contrast to human, all currently known soybean phosphorylation sites were determined using mass spectrometry—a high-throughput technique that, although cheaper and faster than traditional

Table 3. Performance data for PHOSFER and its variants, as well as for the comparison tools PhosPhAt and PlantPhos

Site	Tool	A_{ROC}	Sensitivity at specificity			MCC at specificity		
			0.99	0.95	0.9	0.99	0.95	0.9
S	PHOSFER	0.860	0.337	0.545	0.663	0.434	0.544	0.583
	PHOSFER-NC	0.850	0.271	0.512	0.599	0.378	0.512	0.521
	PHOSFER-EW	0.857	0.307	0.482	0.630	0.409	0.491	0.551
	PHOSFER-SO	0.830	0.208	0.470	0.551	0.313	0.481	0.482
	PHOSFER-AO	0.859	0.367	0.530	0.633	0.458	0.531	0.553
	PHOSFER-AO25	0.849	0.304	0.509	0.596	0.406	0.510	0.522
	PhosPhAt	0.792	0.199	0.399	0.508	0.313	0.417	0.444
	PlantPhos	0.796	0.129	0.341	0.508	0.238	0.371	0.448
T	PHOSFER	0.788	0.167	0.324	0.472	0.278	0.358	0.409
	PHOSFER-NC	0.782	0.111	0.204	0.454	0.214	0.238	0.393
	PHOSFER-EW	0.778	0.139	0.167	0.481	0.247	0.195	0.418
	PHOSFER-SO	0.683	0.139	0.269	0.380	0.247	0.305	0.325
	PHOSFER-AO	0.789	0.185	0.352	0.491	0.297	0.383	0.414
	PHOSFER-AO25	0.760	0.111	0.259	0.454	0.214	0.296	0.393
	PhosPhAt	0.666	0.019	0.160	0.245	0.041	0.188	0.190
	PlantPhos	0.656	0.086	0.143	0.305	0.179	0.163	0.249
Y	PHOSFER	0.738	0.204	0.245	0.429	0.337	0.292	0.370
	PHOSFER-NC	0.744	0.204	0.245	0.306	0.337	0.292	0.253
	PHOSFER-EW	0.701	0.143	0.347	0.347	0.277	0.387	0.293
	PHOSFER-SO	0.624	0.020	0.102	0.122	0.102	0.119	0.032
	PHOSFER-AO	0.770	0.082	0.286	0.347	0.206	0.331	0.293
	PHOSFER-AO25	0.763	0.082	0.143	0.306	0.206	0.177	0.253
	PhosPhAt	0.609	0.000	0.184	0.245	0.000	0.224	0.185
	PlantPhos	0.655	0.042	0.104	0.188	0.146	0.120	0.118

Note: A_{ROC} values are shown, as well as sensitivity and MCC at various specificity values.

Table 4. Performance comparison of PHOSFER and PHOSFER-SO when using different amounts of soybean data

Tool	A_{ROC}	Sensitivity at specificity			MCC at specificity		
		0.99	0.95	0.9	0.99	0.95	0.9
PHOSFER75	0.870	0.305	0.566	0.671	0.409	0.561	0.586
PHOSFER-SO75	0.839	0.205	0.474	0.602	0.319	0.485	0.531
PHOSFER50	0.892	0.337	0.578	0.741	0.428	0.571	0.653
PHOSFER-SO50	0.826	0.229	0.452	0.584	0.333	0.466	0.507
PHOSFER25	0.896	0.386	0.518	0.675	0.468	0.521	0.594
PHOSFER-SO25	0.837	0.277	0.518	0.614	0.401	0.521	0.527

Note: PHOSFER75 and PHOSFER-SO75 were the same as PHOSFER and PHOSFER-SO, respectively, except that they used only 75% of the soybean training data, and similarly for the tools numbered 50 (50% of the soybean training data) and 25 (25%).

methods for studying kinase substrates, does not provide any information on the kinases that catalyze the phosphorylation of a given site. As such, it is currently impossible to create a kinase-specific predictor for soybean, likely creating a ceiling on the accuracy of future soybean predictors—as well as predictors for any other organism for which kinase-specific information is unavailable.

4.3 Phosphorylation site conservation and kinase recognition patterns

It was surprising that although PHOSFER generally exhibited improved performance over the other PHOSFER variants that used data from soybean (PHOSFER-NC, PHOSFER-EW and PHOSFER-SO), its performance was rivaled—and in some cases exceeded—by PHOSFER-AO. This is particularly interesting given the level of phosphorylation site conservation between *Arabidopsis* and soybean, which—although the highest of the organisms tested—was only ~50% for Y sites and significantly less than that for S and T sites. How, then, can using *Arabidopsis* data to predict soybean sites result in high predictive accuracy? One possibility is that, although cellular signaling pathways and processes may be only partially conserved between the two plants (thus explaining the proportion of conserved phosphorylation sites), the patterns dictating kinase recognition of those sites

are more similar. If this is the case, it certainly validates the use of machine-learning tools for predicting phosphorylation sites in an organism of interest (e.g. soybean), instead of (or in addition to) simply finding conserved sites using known phosphorylation data from a related organism (e.g. *Arabidopsis*). It would make interesting future work to statistically characterize the patterns found in the phosphorylation sites of various organisms, with statistical measures indicating their similarity or dissimilarity.

4.4 Testing the efficacy of simpler cross-species models

Given the comparable performance of PHOSFER with PHOSFER-AO, as additional future work, it would be valuable to further investigate the performance of simpler (than PHOSFER) cross-species models. For example, would using only rice sites as training data result in similar predictive performance (relative to PHOSFER-AO) on soybean testing instances? More generally, it would be worthwhile to determine the relationship between conservation of phosphorylation sites for each organism, as shown in Table 2, and the efficacy of using those data as training instances for predicting soybean sites.

4.5 Applicability to other organisms

In addition to soybean, the technique used by PHOSFER should be applicable to other plants for which few known phosphorylation sites are available. For example, the number of known phosphorylation sites in economically important crops like corn, canola and wheat and in scientifically important model organisms like *Medicago truncatula*, are currently comparable with, or less than, that of soybean (Dinkel *et al*., 2011; Gao *et al*., 2009b; Hornbeck *et al*., 2012). In addition, the technique used by PHOSFER need not be restricted to plants; the accuracy of predicting phosphorylation sites for virtually any organism of interest could be enhanced by using data from related organisms.

4.6 Availability

Using the Galaxy platform (Goecks *et al*., 2010), we have made PHOSFER available on the web. The user simply needs to browse to http://saphire.usask.ca and upload a multi-FASTA file. Three tab-delimited output files will be created: one containing the score given to each 15mer with S at its center and similarly for T and Y. Each file contains four columns: the name of the source protein, the 15mer sequence, the position of that 15mer sequence in the full protein and the predicted score. Sequences with higher scores are more likely to be phosphorylation sites.

5 CONCLUSION

In this article, we have described a novel machine-learning model for predicting phosphorylation sites in soybean using known phosphorylation sites from both soybean and other organisms. The use of data from other species resulted in a large improvement in predictive accuracy for T and Y sites and a more modest improvement for S sites. The species-specific instance weights generally imparted a modest, but noticeable, increase in predictive performance over similar models that lacked them; however, surprisingly a model using only *Arabidopsis* data for training was able to achieve roughly equivalent performance. We hope that the techniques outlined here—random forests, AAIndex features and the use of cross-species training data—can be used as the basis for even more accurate phosphorylation site predictors, especially for organisms having few experimentally characterized phosphorylation sites.

Funding: Natural Sciences and Engineering Research Council of Canada (NSERC).

Conflict of Interest: none declared.

REFERENCES

Apweiler,R. *et al.* (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.*, **32**, D115–D119.

Biswas,A.K. *et al.* (2010) Machine learning approach to predict protein phosphorylation sites by incorporating evolutionary information. *BMC Bioinformatics*, **11**, 273.

Blom,N. *et al.* (1999) Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J. Mol. Biol.*, **294**, 1351–1362.

Blom,N. *et al.* (2004) Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics*, **4**, 1633–1649.

Breiman,L. (2001) Random Forests. *Mach. Learn.*, **45**, 5–32.

Bu,Y.H. *et al.* (2010) Insulin receptor substrate 1 regulates the cellular differentiation and the matrix metallopeptidase expression of preosteoblastic cells. *J. Endocrinol.*, **206**, 271–277.

Champion,A. *et al.* (2004) *Arabidopsis* kinome: after the casting. *Funct. Integr. Genomics.*, **4**, 163–187.

Cherry,J.M. *et al.* (1998) SGD: *Saccharomyces* genome database. *Nucleic Acids Res.*, **26**, 73–79.

Diella,F. *et al.* (2004) Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins. *BMC Bioinformatics*, **5**, 79.

Diella,F. *et al.* (2008) Phospho.ELM: a database of phosphorylation sites–update 2008. *Nucleic Acids Res.*, **36**, D240–D244.

Diks,S.H. *et al.* (2007) Evidence for a minimal eukaryotic phosphoproteome? *PLoS One*, **2**, e777.

Dinkel,H. *et al.* (2011) Phospho.ELM: a database of phosphorylation sites–update 2011. *Nucleic Acids Res.*, **39**, D261–D267.

Durek,P. *et al.* (2010) PhosPhAt: the *Arabidopsis thaliana* phosphorylation site database. An update. *Nucleic Acids Res.*, **38**, D828–D834.

Engel,S.R. *et al.* (2010) *Saccharomyces* genome database provides mutant phenotype data. *Nucleic Acids Res.*, **38**, D433–D436.

Frank,E. *et al.* (2004) Data mining in bioinformatics using Weka. *Bioinformatics*, **20**, 2479–2481.

Gao,J. *et al.* (2009a) A new machine learning approach for protein phosphorylation site prediction in plants. *Lect. Notes Comput. Sci.*, **5462/2009**, 18–29.

Gao,J. *et al.* (2009b) P3DB: a plant protein phosphorylation database. *Nucleic Acids Res.*, **37**, D960–D962.

Goecks,J. *et al.* (2010) Galaxy Team. (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.*, **11**, R86.

Goodstein,D.M. *et al.* (2012) Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.*, **40**, D1178–D1186.

Heazlewood,J.L. *et al.* (2008) PhosPhAt: a database of phosphorylation sites in *Arabidopsis thaliana* and a plant-specific phosphorylation site predictor. *Nucleic Acids Res.*, **36**, D1015–D1021.

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

Hornbeck,P.V. *et al.* (2012) PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res.*, **40**, D261–D270.

Japkowicz,N. and Stephen,S. (2002) The class imbalance problem: a systematic study. *Intell. Data Anal.*, **6**, 429–449.

Kawashima,S. *et al.* (2008) AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res.*, **36**, D202–D205.

Kim,S.H. and Lee,C.E. (2011) Counter-regulation mechanism of IL-4 and IFN-signal transduction through cytosolic retention of the pY-STAT6:pY-STAT2:p48 complex. *Eur. J. Immunol.*, **41**, 461–472.

Lee,T.Y. *et al.* (2011) PlantPhos: using maximal dependence decomposition to identify plant phosphorylation sites with substrate site specificity. *BMC Bioinformatics*, **12**, 261.

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

Lian,I. *et al.* (2010) The role of YAP transcription coactivator in regulating stem cell self-renewal and differentiation. *Genes Dev.*, **24**, 1106–1118.

Manning,G. *et al.* (2002) The protein kinase complement of the human genome. *Science*, **298**, 1912–1934.

Miller,M.L. *et al.* (2008) Linear motif atlas for phosphorylation-dependent signaling. *Sci. Signal.*, **1**, ra2.

Nakai,K. *et al.* (1988) Cluster analysis of amino acid indices for prediction of protein structure and function. *Protein Eng.*, **2**, 93–100.

Neuberger,G. *et al.* (2007) pkaPS: prediction of protein kinase A phosphorylation sites with the simplified kinase-substrate binding model. *Biol. Direct.*, **2**, 1.

Petersen,B. *et al.* (2009) A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC Struct. Biol.*, **9**, 51.

Ressurreição,M. *et al.* (2011) A role for p38 MAPK in the regulation of ciliary motion in a eukaryote. *BMC Cell Biol.*, **12**, 6.

Saha,I. *et al.* (2011) Fuzzy clustering of physicochemical and biochemical properties of amino Acids. *Amino Acids*, **43**, 583–594.

Sing,T. *et al.* (2005) ROCR: visualizing classifier performance in R. *Bioinformatics*, **21**, 3940–3941.

Stark,C. *et al.* (2010) PhosphoGRID: a database of experimentally verified in vivo protein phosphorylation sites from the budding yeast *Saccharomyces cerevisiae*. *Database*, **2010**, bap026.

Swarbreck,D. *et al.* (2008) The *Arabidopsis* information resource (TAIR): gene structure and function annotation. *Nucleic Acids Res.*, **36**, D1009–D1014.

Tang,Y.R. *et al.* (2007) GANNPhos: a new phosphorylation site predictor based on a genetic algorithm integrated neural network. *Protein Eng. Des. Sel.*, **20**, 405–412.

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

Trost,B. and Kusalik,A. (2011) Computational prediction of eukaryotic phosphorylation sites. *Bioinformatics*, **27**, 2927–2935.

Uddin,S. *et al.* (2003) Role of Stat5 in type I interferon-signaling and transcriptional regulation. *Biochem. Biophys. Res. Commun.*, **308**, 325–330.

UniProt Consortium. (2008) The universal protein resource (UniProt). *Nucleic Acids Res.*, **36**, D190–D195.

UniProt Consortium. (2012) Reorganizing the protein space at the universal protein resource (UniProt). *Nucleic Acids Res.*, **40**, D71–D75.

Wang,Y.Y. *et al.* (2010) Hydrogen peroxide stress stimulates phosphorylation of FoxO1 in rat aortic endothelial cells. *Acta. Pharmacol. Sin.*, **31**, 160–164.

Witten,I.H. *et al.* (2011) *Data Mining: Practical Machine Learning Tools and Techniques,* 3rd edn. Morgan Kaufmann, Burlington, MA.

Wood,C.D. *et al.* (2009) Nuclear localization of p38 MAPK in response to DNA damage. *Int. J. Biol. Sci.*, **5**, 428–437.

Xue,Y. *et al.* (2010) A summary of computational resources for protein phosphorylation. *Curr. Protein Pept. Sci.*, **11**, 485–496.

Zhang,J. and Johnson,G.V. (2000) Tau protein is hyperphosphorylated in a site-specific manner in apoptotic neuronal PC12 cells. *J. Neurochem.*, **75**, 2346–2357.