BIOINFORMATICS **REVIEW** Vol. 27 no. 21 2011, pages 2927–2935
doi:10.1093/bioinformatics/btr525

Sequence analysis Advance Access publication September 16, 2011

Computational prediction of eukaryotic phosphorylation sites

Brett Trost* and Anthony Kusalik

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

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Kinase-mediated phosphorylation is the central mechanism of post-translational modification to regulate cellular responses and phenotypes. Signaling defects associated with protein phosphorylation are linked to many diseases, particularly cancer. Characterizing protein kinases and their substrates enhances our ability to understand and treat such diseases and broadens our knowledge of signaling networks in general.

While most or all protein kinases have been identified in well-studied eukaryotes, the sites that they phosphorylate have been only partially elucidated. Experimental methods for identifying phosphorylation sites are resource intensive, so the ability to computationally predict potential sites has considerable value.

Results: Many computational techniques for phosphorylation site prediction have been proposed, most of which are available on the web. These techniques differ in several ways, including the machine learning technique used; the amount of sequence information used; whether or not structural information is used in addition to sequence information; whether predictions are made for specific kinases or for kinases in general; and sources of training and testing data.

This review summarizes, categorizes and compares the available methods for phosphorylation site prediction, and provides an overview of the challenges that are faced when designing predictors and how they have been addressed. It should therefore be useful both for those wishing to choose a phosphorylation site predictor for their particular biological application, and for those attempting to improve upon established techniques in the future.

Contact: brett.trost@usask.ca

Received on July 26, 2011; revised on September 2, 2011; accepted on September 12, 2011

1 INTRODUCTION

Phosphorylation is the most widespread post-translational modification in eukaryotes and plays a crucial role in the regulation of virtually every cellular behavior, including DNA repair (Wood *et al.*, 2009), environmental stress response (Wang *et al.*, 2010), regulation of transcription (Uddin *et al.*, 2003), apoptosis (Zhang and Johnson, 2000), cellular motility (Ressurreição *et al.*, 2011), immune response (Kim and Lee, 2011), metabolism (Bu *et al.*, 2010), and cellular differentiation (Lian *et al.*, 2010). Historically, novel phosphorylation sites have been discovered primarily through the use of low-throughput biological techniques. With the advent of site-directed mutagenesis, for instance, many labs started using this technique to characterize specific phosphorylation events (e.g. Meier *et al.*, 1997). Unfortunately, such techniques are time consuming, tedious and expensive to perform. More recently, a high-throughput technique—mass spectrometry—has greatly accelerated the identification of novel sites. For example, Huttlin *et al.* (2010) used mass spectrometry to map the phosphoproteome of nine different mouse tissues, identifying nearly 36 000 distinct phosphorylation sites. While useful, this technique has important limitations: it cannot identify the protein kinase(s) responsible for catalyzing the phosphorylation of a given site; many phosphorylation sites are modified at substoichiometric levels, with the unphosphorylated form sometimes preventing detection of the phosphorylated form; many interesting proteins are present at very low levels, making them difficult to detect through mass spectrometry; breaking open cells can place kinases together with substrates they would not normally encounter, potentially resulting in the detection of phosphorylation events that would not occur *in vivo*; technical challenges exist that sometimes make pinpointing exact phosphorylation sites difficult (Boersema *et al.*, 2009); and perhaps most importantly, mass spectrometry requires very expensive instruments and specialized expertise that are not available in typical laboratories.

Given the limitations associated with both low-throughput and high-throughput biological techniques for identifying novel phosphorylation sites, computational approaches have become increasingly popular. Such techniques require a protein sequence as input, and produce some numerical measure of the likelihood that each serine, threonine or tyrosine (S/T/Y) residue in that sequence is a phosphorylation site. For example, Slaugenhaupt *et al.* (2001) found that the mutation R696P in the protein encoded by the *IKBKAP* gene causes familial dysautonomia, with the hypothesized mechanism being disruption of the phosphorylation of T699—a site predicted to be phosphorylated by NetPhos (Blom *et al.*, 1999), the first phosphorylation site prediction tool. Prediction programs are often used to narrow down the list of possible phosphorylation sites in a protein of interest, with the predictions subsequently verified using biological experiments. For instance, Fan *et al.* (2009) used a combination of several predictors to identify seven putative sites phosphorylated by protein kinase C in the transient receptor potential vanilloid 4 (TRPV4) ion channel. When three of these sites were separately mutated to alanine, the resultant proteins exhibited markedly reduced activation in response to protein kinase C compared with wild-type TRPV4.

To the authors' knowledge, four review articles have previously been published that included significant discussion of computational phosphorylation site prediction. Kobe *et al.* (2005) provided a brief review of this field, along with a detailed discussion of the structural bases of protein kinase specificity. Hjerrild and Gammeltoft (2006) reviewed both computational and biological

*To whom correspondence should be addressed.

aspects of phosphoproteomics, while Miller and Blom (2009) briefly summarized some of the literature on phosphorylation site prediction and provided a protocol for the use of their NetPhos (Blom *et al.*, 1999, 2004; Hjerrild *et al.*, 2004) family of tools. Most recently, Xue *et al.* (2010) reviewed phosphorylation site databases, prediction tools and miscellaneous software associated with phosphorylation sites, and also compared the performance of a subset of available predictors.

Compared with the above reviews, here we concentrate more specifically on the methodologies employed by the various prediction tools, taking a comparative approach to examining the issues and challenges associated with computational phosphorylation site prediction. Section 2 of this review provides a brief overview of the available methods, while Section 3 compares and discusses the tools with respect to different aspects of their methodologies. Section 4 comments on some of the challenges that remain in the field, and Section 5 gives some concluding remarks.

2 AN OVERVIEW OF CURRENT TOOLS FOR PHOSPHORYLATION SITE PREDICTION

If significant updates to existing methods are treated separately, then there have been nearly 40 methods for the computational prediction of phosphorylation sites described since 1999. This total excludes tools that predict sites for more than one type of post-translational modification (e.g. Basu and Plewczyński, 2010; Schwartz *et al.*, 2009) and methods based on simple motifs (discussed by Xue *et al.*, 2010). Unlike Xue and co-authors, however, we include techniques for which no web implementation is available, as they can be valuable sources of ideas for developers of future tools.

A list of currently available phosphorylation site prediction tools is given in Table 1. Each tool is categorized with respect to several important attributes, including the machine learning technique(s) used (described further in Section 3.1); the number of residues surrounding the phosphorylation site that are taken into account (Section 3.2); whether the method uses only sequence information or also uses structural information (Section 3.3); whether the tool includes models specific to particular kinases or kinase families (Section 3.4); and the source(s) of known phosphorylation sites used for training and testing (Section 3.5). We avoid comparing the performance of each tool, for several reasons: some of the tools discussed do not have web implementations or have websites that are no longer accessible; different tools use different performance measures or were tested on different datasets; it would not be meaningful to compare kinase-specific with non-kinase-specific tools; and performance comparisons have been done elsewhere for some tools (Xue *et al.*, 2010). However, we do discuss two important issues regarding predictive performance—the creation of standardized testing datasets (Section 4.1) and balancing sensitivity and specificity (Section 4.3).

3 COMPARING AND CONTRASTING THE AVAILABLE TOOLS

3.1 Machine learning methods

To provide tight control of cellular processes, a protein kinase catalyzes the phosphorylation of a given S/T/Y residue only if the amino acids around that residue fit a specific, yet flexible, pattern (Diks *et al.*, 2007). Sequence motifs that describe these patterns, such as those in the PROSITE database (Sigrist *et al.*, 2002), are neither sensitive (see Blom *et al.*, 1999) nor specific (the PROSITE motif for the protein kinase C recognition sequence, for instance, is [ST]-x-[RK], which would be expected to occur often at random). The poor specificity and sensitivity of motifs means that accurate prediction of phosphorylation sites requires the use of machine learning methods, which can identify more complex and subtle patterns. As Table 1 shows, many different machine learning methods have been used, including artificial neural networks (ANNs), decision trees, genetic algorithms, position-specific scoring matrices (PSSMs) and support vector machines (SVMs).

Perhaps the simplest machine learning technique is the PSSM, which is a matrix in which rows represent amino acids and columns represent positions in a multiple sequence alignment. In the simplest possible PSSM, a given matrix element would contain the frequency of a given amino acid in a given position, although more complex variations are usually developed in practice (e.g. Koenig and Grabe, 2004; Li *et al.*, 2008b). PSSMs are easy to understand and construct, but are unable to detect patterns in which combinations of amino acids are important (Blom *et al.*, 1999). PSSMs can, say, express the idea that proline promotes phosphorylation when found at position $+1$ (where position 0 is the phosphorylation site) and arginine promotes phosphorylation when found at -2, but cannot express the idea that both occurring at the same time prevents phosphorylation.

In contrast, machine learning techniques like ANNs and SVMs—two of the most popular methods used by phosphorylation site prediction tools—can capture more complex patterns (Blom *et al.*, 1999). This comes at the cost of added complexity. ANNs, in particular, are often regarded as 'black boxes' in which the classification function is essentially inscrutable. Some methods strike a balance between the simplicity of PSSMs and the opaqueness of ANNs. Xue *et al.* (2006), for example, proposed a method based on Bayesian probability that is more expressive than PSSMs, but is more easily interpreted biologically and mathematically than ANNs.

An interesting point of discussion is, what do the machine learning methods actually model? In other words, do they model the actual biological mechanisms underlying protein kinase recognition, or do they merely recognize patterns? For the majority of methods listed in Table 1, and certainly for those that consider only sequence information (Section 3.3), we would argue that it is the latter. This is not meant to denigrate these methods: clearly, recognizing patterns in sequence information is useful, both in the field of phosphorylation site prediction and elsewhere. Most tools utilizing structural information fall somewhere between the two categories mentioned above. For example, although DISPHOS (Iakoucheva *et al.*, 2004) uses secondary structure predictions as features, it would be better described as recognizing patterns than as modeling biological mechanisms. On the other end of the scale, pkaPS (Neuberger *et al.*, 2007) extensively models the kinase–substrate interaction. While pattern recognition has resulted in much success, it is plausible that more closely modeling the underlying biology of substrate recognition will result in the greatest gains in predictive performance.

3.2 Amount of sequence information used

Phosphorylation site prediction tools vary widely in the number of residues surrounding the phosphorylation site that are taken

Table 1. Currently available phosphorylation site prediction tools

Name	Technique	Residues	1D/3D	K-spec?	Data	Reference	Website
NetPhos	ANN	9–33	3D	No	PB	Blom *et al.* (1999)	cbs.dtu.dk/services/NetPhos
Scansite	PSSM	15	1D	Yes	PB	Yaffe *et al.* (2001)	scansite.mit.edu
Predikin 1.0	SA	7	3D	Yes	PB	Brinkworth *et al.* (2003)	predikin.biosci.uq.edu.au
rBBFNN	ANN, DT	9–11	1D	No	PB	Berry *et al.* (2004)	(no web implementation available)
DISPHOS	LR	25	3D	No	PB, SP	Iakoucheva *et al.* (2004)	www.dabi.temple.edu/disphos
NetPhosK	ANN	9–33	3D	Yes	Many	Hjerrild *et al.* (2004)	cbs.dtu.dk/services/NetPhos
PredPhospho	SVM	[a]	1D	Yes	PB, SP	Kim *et al.* (2004)	(website no longer accessible)
PHOSITE	PSSM	[a]	1D	Yes	PB	Koenig and Grabe (2004)	(website no longer accessible)
GPS 1.0	PSSM, MC	7	1D	Yes	P.ELM	Zhou *et al.* (2004)	gps.biocuckoo.org
[b]	Many	9	1D	No	PB	Senawongse *et al.* (2005)	(no web implementation available)
KinasePhos 1.0	HMM	9	1D	Yes	PB, SP	Huang *et al.* (2005a)	kinasephos.mbc.nctu.edu.tw
[b]	SVM	9	3D	Yes	SP	Plewczyński *et al.* (2005)	(no web implementation available)
PPSP	BP	9	1D	Yes	P.ELM	Xue *et al.* (2006)	ppsp.biocuckoo.org
pkaPS	SA	42	3D	Yes	P.ELM	Neuberger *et al.* (2007)	mendel.imp.ac.at/sat/pkaPS
[b]	STAT	2–4	1D	Yes	LIT, P.ELM	Moses *et al.* (2007)	(no web implementation available)
NetPhosYeast	ANN	[a]	1D	No	LIT, SP	Ingrell *et al.* (2007)	cbs.dtu.dk/services/NetPhosYeast
NetworKIN	ANN, PSSM	9–33	3D	Yes	P.ELM	Linding *et al.* (2007)	networkin.info
KinasePhos 2.0	SVM	9	1D	Yes	P.ELM	Wong *et al.* (2007)	kinasephos2.mbc.nctu.edu.tw
GANNPhos	Many	25	1D	No	P.ELM, SP	Tang *et al.* (2007)	(no web implementation available)
PHOSIDA	SVM	13	1D	No	PHOSIDA	Gnad *et al.* (2007)	phosida.de
PhosPhAt	SVM	13	1D	No	PPA	Heazlewood *et al.* (2008)	phosphat.mpimp-golm.mpg.de
IEPP	BP	31	1D	Yes	P.ELM	Wang *et al.* (2008a)	(no web implementation available)
AutoMotif	SVM	9	1D	Yes	SP	Plewczyński *et al.* (2008)	(website no longer accessible)
PhoScan	PSSM	25	1D	Yes	P.ELM	Li *et al.* (2008b)	bioinfo.au.tsinghua.edu.cn/phoscan
MetaPredPS	MP	[c]	[c]	Yes	Many	Wan *et al.* (2008)	metapred.biolead.org/MetaPredPS
SiteSeek	Many	[a]	1D	Yes	Many	Yoo *et al.* (2008)	(no web implementation available)
SMALI	PSSM	7	1D	Yes	IHE	Li *et al.* (2008a)	lilab.uwo.ca/SMALI.htm
Predikin 2.0	HMM, SA	7	3D	Yes	P.ELM, SP	Saunders *et al.* (2008)	predikin.biosci.uq.edu.au
GPS 2.0	PSSM, GA	15	1D	Yes	P.ELM	Xue *et al.* (2008)	gps.biocuckoo.org
CRPhos	CRF	9	1D	Yes	P.ELM	Dang *et al.* (2008)	www.ptools.ua.ac.be/CRPhos
Phos3D	SVM	13	3D	Yes	P.ELM	Durek *et al.* (2009)	phos3d.mpimp-golm.mpg.de
[b]	SVM	25	1D	No	PPA, TAIR	Gao *et al.* (2009)	(no web implementation available)
PostMod	PSSM	7–101	1D	Yes	P.ELM	Jung *et al.* (2010)	pbil.kaist.ac.kr/PostMod
PPRED	PSSM, SVM	7–15	1D	No	P.ELM	Biswas *et al.* (2010)	ashiskb.info/research/ppred
[b]	SVM	9–15	3D	No	P.ELM	Swaminathan *et al.* (2010)	(no web implementation available)
BAE	STAT	11	1D	Yes	P.ELM	Yu *et al.* (2010)	(no web implementation available)
PAAS	PSSM	[a]	1D	Yes	P.ELM	Sobolev *et al.* (2010)	(website no longer accessible)
[b]	STAT, SVM	9	1D	Yes	P.ELM	Li *et al.* (2010)	cmbi.bjmu.edu.cn/huphospho
Musite	SVM	[a]	1D	Yes	Many	Gao and Xu (2010)	musite.sourceforge.net
GPS 2.1	PSSM, GA	3–31	1D	Yes	P.ELM	Xue *et al.* (2011)	gps.biocuckoo.org

Column headings are as follows: name, the name of the tool; technique, the machine learning technique used; residues, the number of residues flanking (and including) the phosphorylated residue that are used in the tool's predictions; 1D/3D, whether only sequence information is used (1D) or structural information is used as well (3D); K-spec, yes if the tool makes predictions for specific kinases or kinase families, and no otherwise; data, the source of the known phosphorylation sites used for training and testing; reference, the paper describing that tool; website, the address of that tool's web implementation (if applicable). Due to space considerations, only one reference per tool is included in the table; tools described by more than one paper include ScanSite [also described in Obenauer *et al.* (2003)], NetPhosK (Blom *et al.*, 2004), PredPhospho (Ryu *et al.*, 2009), GPS 1.0 (Xue *et al.*, 2005), KinasePhos 1.0 (Huang *et al.*, 2005b), PHOSIDA (Gnad *et al.*, 2011), PhosPhAt (Durek *et al.*, 2010), Predikin 2.0 (Saunders and Kobe, 2008) and Musite (Gao *et al.*, 2010). BP, Bayesian probability; CRF, conditional random fields; DT, decision tree; GA, genetic algorithm; IHE, in-house experiments; LIT, literature; LR, logistic regression; MC, Markov clustering; MP, meta-predictor; PB, PhosphoBase; P.ELM, Phospho.ELM; PPA, PhosPhAt database; PS, PhosphoSitePlus; SA, structural analysis; SP, Swiss-Prot (or UniProt); STAT, statistical method; TAIR, The *Arabidopsis* Information Resource database.

[a] Exact range of lengths not explicitly stated.

[b] No name was given to these tools by their authors.

[c] Varies depending on individual predictors used (see text).

into account. At one extreme, PostMod (Jung *et al.*, 2010) was designed to consider up to 101 residues (between positions −50 and +50), whereas Predikin 1.0 (Brinkworth *et al.*, 2003) considers just seven. The number of residues considered is important because too few means information useful for making predictions gets ignored, while too many will decrease the signal-to-noise ratio. Using many residues can also make some machine learning methods computationally intractable (Biswas *et al.*, 2010).

Several strategies have been used to determine the optimum number of residues. First, it has been argued that the optimum should be consistent with the number of residues in physical contact with the kinase (Blom *et al.*, 1999). An early report stated that 9–12 residues surrounding the phosphorylation site are likely to physically contact the kinase (Songyang *et al.*, 1994), an estimate consistent with the number of residues used by many prediction methods. Depending on the 3D structure of the substrate, however, the 9–12 residues contacted by the kinase may not be the same as the 9–12 residues surrounding the phosphorylation site in the linear sequence. Residues not in contact with the kinase may also affect its binding by influencing the charge or hydrophobicity of the microenvironment, or by affecting the conformation of residues in contact with the kinase. As such, the number of residues that physically contact the kinase may not reliably indicate the number of residues that should be used for making predictions.

Second, some authors have empirically tested various numbers of residues, and then chosen the number that gives the best predictive performance. The authors of PostMod (Jung *et al.*, 2010) tried between 7 and 101 residues, and found that 41 resulted in the best accuracy. Other authors reported much smaller optima, with Blom *et al.* (1999) suggesting between 9 and 11 and Biswas *et al.* (2010) reporting 15. Given that reported optima are inconsistent, researchers should use caution when applying previously reported empirical optima for developing future methods.

Third, Neuberger *et al.* (2007) examined how residues around phosphorylation sites compare with residues in general proteins with respect to two properties—hydrophobicity and flexibility. Figure 1 of their paper plots position (40 residues upstream and downstream of the phosphorylation site) versus deviation from baseline values, and shows that each property deviates substantially from baseline values near phosphorylation sites, and then gradually returns as one gets farther from a given site. The authors found that both properties deviate significantly between positions −18 and +23, and thus advocate the use of these 42 residues. Because this experiment was done only for protein kinase A, it is not known whether its results generalize to other protein kinases.

The lack of agreement among the three strategies described above may be due to differences among kinases (some kinases may use more residues as a recognition sequence than others) and/or among machine learning methods (some methods may handle greater dimensionality—in other words, longer sequences—better than others). The lack of agreement could also be related to effect size. For example, a residue at position −20 may have a real, but very small, effect on phosphorylation—an effect that might be ignored by some authors, but not others. As the most appropriate number of residues remains unclear, a rigorous investigation of this issue would be invaluable for developers of future tools, especially if consideration was given to the particular machine learning technique used and the particular kinase under consideration.

3.3 Use and non-use of structural information

The structural basis of protein kinase-catalyzed phosphorylation has been examined in numerous studies. For example, Dunker *et al.* (2002) reported that phosphorylation sites are frequently found in disordered regions [a fact exploited by the authors of the NETPHOS prediction tool (Iakoucheva *et al.*, 2004)], while Kitchen *et al.* (2008) described the degree to which electrostatic interactions stabilize phosphorylated residues. Further, a review by Kobe *et al.* (2005) examined the structural determinants of protein kinase specificity.

The degree to which the substrate's 3D structure affects kinase specificity is unclear. Studies that find correlations between these two variables, like those cited above, suggest that the substrate's structure is important in the recognition process. Conversely, short peptides containing known phosphorylation sites can be recognized with similar kinase-catalyzed kinetics as the corresponding intact protein (Kemp *et al.*, 1977; Zetterqvist *et al.*, 1976; see also Houseman *et al.*, 2002 and Löwenberg *et al.*, 2005), suggesting a minor role for structure. Despite this, it seems plausible that the use of structural information can play at least some role in increasing the accuracy of phosphorylation site prediction tools. Although lack of structural data remains an obstacle, the amount of structural information about phosphorylation sites is growing rapidly. The most recent version of the Phospho3D database (Zanzoni *et al.*, 2011), for instance, contains structural information for over 1700 sites, nearly 11 times the number contained in the previous version (Zanzoni *et al.*, 2007).

Table 1 shows that approximately one-quarter of tools utilize information regarding the 3D structure of the kinase and/or its substrate, whereas the other tools use only primary sequence information. Blom *et al.* (1999), in addition to devising a method based only on primary sequence, superimposed the structures of 12 different tyrosine phosphorylation sites, and found that nine of them had a common conformation, while the other three shared a second conformation. In contrast, non-phosphorylated tyrosine residues exhibited a wide range of conformations. They also determined that phosphorylated residues were generally more flexible than average, consistent with the hypothesis that high flexibility would be required to fit into a kinase's active site. While conformation and flexibility thus seemed like two structural features that could increase prediction accuracy, the authors' sequence-based method outperformed their structure-based method, although the latter did make more accurate predictions for a few atypical tyrosine phosphorylation sites. In contrast, Durek *et al.* (2009) found that, for several different kinase families, adding structural information to a sequence-only model resulted in a modest but consistent increase in predictive performance, showing that the use of structural information can add discriminatory power.

Given that sequence-only methods, by definition, ignore information about the kinase–substrate interaction, the upper limit to their accuracy is likely less than the upper limit of structure-based methods. As structural information becomes available for more and more phosphorylation sites, structure-based methods will continue to improve.

3.4 Kinase-specific versus non-kinase-specific tools

Most phosphorylation site prediction programs are kinase-specific, as they require as input both a protein sequence and the name of a protein kinase, and produce some measure of the likelihood that each S/T/Y residue in the sequence is phosphorylated by the chosen kinase. In contrast, a few tools require only a protein sequence as input, and report the likelihood that each S/T/Y residue is phosphorylated by any kinase. Kinase-specific tools can be further divided based on whether they make predictions for individual kinases [e.g. NetPhosK (Blom *et al.*, 2004) and pkaPA (Neuberger *et al.*, 2007)] or for kinase families [e.g. SiteSeek (Yoo *et al.*, 2008) and PAAS (Sobolev *et al.*, 2010)]. Part of the motivation for making predictions for kinase families is that some individual kinases have very few target sites known, making the training component of machine learning difficult. As kinases from the same family will likely have similar recognition sequences (Kim *et al.*, 2004), their known target sites can be combined, resulting in a model that utilizes much more information than if kinases from the same family were modeled separately.

How do the accuracies of non-kinase-specific tools compare with those of kinase-specific tools? Given the issues involved in comparing the performance of different tools (see Section 3.5.3), this question is more difficult to answer than it would appear. It has been claimed that, since there is no ‘average’ phosphorylation site, only kinase-specific predictors should be able to achieve good accuracy (Neuberger *et al.*, 2007)—an argument with considerable logical appeal. Indeed, most users will likely be interested in particular biological pathways (and thus particular kinases), making kinase-specific tools an ideal choice. For applications in which the specific kinase is not a concern, the user could still take advantage of the higher accuracy of kinase-specific tools by aggregating the results from many kinase-specific predictions to make a general list of phosphorylation sites in the protein(s) of interest. On the other hand, non-kinase-specific tools may be able to detect phosphorylation sites for which the associated kinase is unknown—an advantage that may be of interest to some users. Additionally, non-kinase-specific tools have reported respectable performance, with accuracy rates approaching 80% (Swaminathan *et al.*, 2010).

3.5 Training and testing data

Both positive (actual phosphorylated residues) and negative (actual non-phosphorylated residues) data are required for training and testing a particular prediction tool. Section 3.5.1 discusses sources of positive data, while Section 3.5.2 describes the problem of obtaining negative data. Finally, the issue of fair performance comparisons is discussed in Section 3.5.3.

3.5.1 Positive data Several sources of known phosphorylation sites have been used. Most early prediction tools used either PhosphoBase (Blom *et al.*, 1998; Kreegipuu *et al.*, 1999), a database solely containing known phosphorylation sites, or Swiss-Prot, for which the annotation of a given protein includes its known phosphorylation sites. Authors using Swiss-Prot (e.g. Iakoucheva *et al.*, 2004; Plewczyński *et al.*, 2005) generally discard sites described as ‘hypothetical’, ‘predicted’ or ‘by similarity’, choosing instead only experimentally confirmed sites. In 2004, the information from PhosphoBase was integrated into a new database called Phospho.ELM (Diella *et al.*, 2004, 2008; Dinkel *et al.*, 2011). Most tools developed after 2004 have used Phospho.ELM, although there are exceptions: other databases that have been used (some of which are specialized in nature) are PhosphoSitePlus (Hornbeck *et al.*, 2004), The *Arabidopsis* Protein Phosphorylation Site Database (PhosPhAt) (Durek *et al.*, 2010; Heazlewood *et al.*, 2008), The *Arabidopsis* Information Resource (TAIR) (Swarbreck *et al.*, 2008) and PHOSIDA (Gnad *et al.*, 2007, 2011). Finally, a few authors searched the literature for known phosphorylation sites (e.g. Hjerrild *et al.*, 2004; Moses *et al.*, 2007).

3.5.2 Negative data An ever-present difficulty in the field of phosphorylation site prediction concerns negative training and testing data. While experiments can verify that a particular residue can be phosphorylated, it would be difficult to prove definitively that a particular residue is not phosphorylated under any conditions. Thus, while databases such as Phospho.ELM and PhosphoSitePlus contain thousands of known phosphorylation sites, they do not contain sites known not to be phosphorylated.

To circumvent this problem, most authors make the assumption that any S/T/Y residue that has not been shown to be phosphorylated is a negative. While some of these residues will likely turn out to be positives as more phosphorylation sites are discovered, the majority of these are probably actual negatives, making this a reasonable, if imperfect, approach. Some authors (e.g. Neuberger *et al.*, 2007) have gone a step further, requiring that the residue not be found in any phosphorylation site database and that its encompassing protein contains at least one residue known to be phosphorylated by the kinase of interest. The assumption here is that, if a protein has at least one residue that is known to be phosphorylated, then the phosphorylation of that protein has been studied in at least some detail, making it less likely that its other S/T/Y residues are undiscovered phosphorylation sites.

Another approach is to use, as negative data, S/T/Y residues that are buried in the core of a particular protein (Blom *et al.*, 2004). This strategy relies on the assumption that buried residues would not be physically accessible to any kinase, thus reducing the number of so-called negatives that later turn out to be positives. A disadvantage of this approach is that it requires knowledge of the protein’s tertiary structure, and only a small portion of proteins currently have solved structures (although the use of structure prediction programs could partially compensate for this). More importantly, however, this method’s underlying assumption may not be entirely valid. In a detailed analysis of experimentally verified phosphorylation sites, Jiménez *et al.* (2007) found that while phosphorylation sites are more solvent-exposed than the average residue, close to 15% have little solvent accessibility. Moreover, a site can be buried in one structure of a given protein, but unburied in another (Durek *et al.*, 2009; Zhou *et al.*, 2006). Despite these caveats, choosing solvent-inaccessible residues as negatives currently seems like the most reliable approach to obtaining negatives for training and testing.

3.5.3 Performing fair comparisons of performance While new prediction tools can improve upon previous ones in various ways, developers must usually show that a new tool offers an improvement in predictive performance. To perform a fair comparison, both the new method and existing methods must be tested using the same data. When testing existing tools, typically one has access only to the already-trained versions that are available on the web (Dang *et al.*, 2008), and since data that were used to train a given tool should not be used to test it (Dang *et al.*, 2008), it can be difficult to identify suitable testing data.

Some authors have simply ignored this problem, comparing their tools’ performance numbers (sensitivity, specificity, etc.) directly with those given in the papers describing previous tools, even though the testing data used may have been different. While having some value, such comparisons are certainly less informative than they could be.

Positive data appropriate for comparing new and existing tools can be obtained by collecting known phosphorylation sites added to a database after the publication of all existing methods (Wan *et al.*, 2008). If access is available to known phosphorylation sites that have not yet been deposited in the databases, they could be used as well. Note that while new known sites are required for comparing performance with existing methods, older known sites can still be used for training a new method.

Given how negative data are obtained (see Section 3.5.2), obtaining negative data appropriate for testing seems harder than for positive data (and interestingly, has been given little or no attention in the literature). Suppose that, as many have done, developer *A* focuses exclusively on predicting phosphorylation in humans. Since the entire human proteome is known, he might use all S/T/Y sites not known to be phosphorylated as negative data for training his method. If developer *B* later wishes to compare his new method to that of *A*, there would be no negative data available that were not used to train *A*'s method, making a fair comparison impossible.

While requiring coordination among those in the phosphorylation site database and prediction community, possible solutions to these problems do exist, some of which are suggested in Section 4.1.

3.6 Other differences among the available tools

While Table 1 categorizes the tools in terms of important properties for which they vary, these categories do not capture all their differences, and there are several tools that deviate from the norm in a notable way. For instance, Musite (Gao and Xu, 2010) is unique in that it is an open-source platform that allows the creation of a customized predictor, with the user able to choose different training and testing data, features, stringency thresholds and so on. Other examples of tools that differ from the norm are given below.

While most tools were trained using known phosphorylation sites, ScanSite's (Obenauer *et al.*, 2003; Yaffe *et al.*, 2001) authors created an oriented peptide library, and then incubated it with a given protein kinase. Phosphorylated peptides were separated from those that were not phosphorylated, and the former sequenced to determine the abundance of each amino acid at each position. The ScanSite program uses this information to output the likelihood that a given S/T/Y residue in its input sequence can be phosphorylated by that protein kinase. Although known phosphorylation sites were not used for training, known sites from PhosphoBase (Blom *et al.*, 1998; Kreegipuu *et al.*, 1999) were used for testing. More recently, Li *et al.* (2008a) developed SMALI, a tool which is similar to ScanSite but claims to have better accuracy.

Most tools require protein sequences as input, and produce as output score indicating the likelihood that a given S/T/Y residue is a phosphorylation site. In contrast, Predikin 1.0 (Brinkworth *et al.*, 2003) takes the sequence of an uncharacterized protein kinase as input, and reports a 7-mer predicted to be its optimal recognition sequence. Predikin 2.0 (Saunders and Kobe, 2008; Saunders *et al.*, 2008) improved upon the original tool's ability to output optimal kinase recognition sequences and also added the conventional functionality of scoring potential phosphorylation sites.

Finally, MetaPredPS (Wan *et al.*, 2008) is currently the only meta-predictor, which is a type of tool that combines the classifications from several individual predictors in the hope of achieving better accuracy. MetaPredPS uses a weighted voting strategy to combine predictions from GPS 1.0, KinasePhos 1.0, NetPhosK, PPSP, PredPhospho and Scansite (see Table 1 for references). Meta-predictors have also been successfully applied to other bioinformatics-related classification problems, including subcellular localization prediction (Liu *et al.*, 2007; Shen *et al.*, 2007), major histocompatibility complex-binding prediction (Karpenko *et al.*, 2008; Trost *et al.*, 2007; Wang *et al.*, 2008b) and protein structure prediction (Ginalski *et al.*, 2003). Given that many different strategies can be used to combine the output of individual predictors, and that there exist dozens of individual tools for phosphorylation site prediction, there is likely room for additional work on meta-predictors in this field.

4 FUTURE DIRECTIONS

In some respects, the field of phosphorylation site prediction is mature. As Table 1 shows, many different machine learning methods have been utilized; widely varying amounts of information (in terms of number of residues surrounding the phosphorylation site) have been incorporated into predictive models; many methods have been proposed in both the structure-based and sequence-based categories; several tools exist for both kinase-specific and non-kinase-specific predictions; and many sources of training and testing data have been utilized. In other respects, however, the field is immature. Three challenges that remain (to rigorously determine the optimum number of residues surrounding the phosphorylation site, to develop improved structure-based methods and to develop additional meta-predictors) were discussed earlier in this review. A number of others were described by Xue *et al.* (2010). Four additional challenges, which we feel are of particular significance, are described below.

4.1 Creating standardized testing datasets

Perhaps the most important challenge involves the development of standardized testing datasets. As described in Section 3.5.3, it is currently extremely difficult to properly compare the accuracy of different prediction tools—either by reading the papers describing them or by testing them anew. Given the large number of phosphorylation site prediction programs already available, it is critical that authors of newly developed tools be able to show a clear improvement in performance compared with older ones. A dataset containing both positive and negative data, half of which is designated for training (only) and half of which is designated for testing (only), would be an invaluable resource, as it would provide a fair, standardized benchmark by which each tool could be judged. Such a database could be created if a laboratory were to use mass spectrometry to identify as exhaustively as possible the phosphorylation sites in an organism for which few sites are currently known. The 'buried residue method' (see Section 3.5.2) could then be used to identify negatives. Unfortunately, this solution has an important limitation: mass spectrometry does not give information about the kinase that phosphorylates each site—information required by tools making kinase-specific predictions. Another solution would be for curators of phosphorylation site databases to designate a portion of all future data collected (from low-throughput or high-throughput sources) as 'testing data', and for the developers of future tools to voluntarily refrain from using these data for training. This strategy could substantially improve the ability to compare phosphorylation site prediction methods.

4.2 Developing tools for a wider variety of organisms

Both the quantity of protein kinases and the types represented (tyrosine kinases, calmodulin-dependent kinases, etc.) differ substantially in different eukaryotes (Diks *et al.*, 2007). For instance, *Arabidopsis* encodes around twice as many protein kinases as does human (Champion *et al.*, 2004; Manning *et al.*, 2002), but does not encode any classical tyrosine kinases. In addition, the lower eukaryote *Plasmodium falciparum* encodes only a few dozen protein kinases, but some of these are of a type observed in few other organisms (Ward *et al.*, 2004). The disparate nature of different organisms' kinomes means that prediction programs designed for human kinases (the majority of the tools currently available) are less useful for organisms like plants. While a few plant-specific prediction tools have been developed (Durek *et al.*, 2010; Gao *et al.*, 2009; Heazlewood *et al.*, 2008), further work needs to be done both for plants and for other non-human organisms. While such work is challenging due to the smaller number of phosphorylation sites that are known for these organisms, further progress can be made as such data become more plentiful, and as structure-based methods for phosphorylation site prediction become more refined.

4.3 Making high specificity predictions for whole-genome annotations

As with other classification problems, predicting phosphorylation sites involves a trade-off between sensitivity and specificity. Greater sensitivity might be beneficial when predicting sites in a single protein, whereas greater specificity may be desirable when identifying sites in an entire proteome. This trade-off is illustrated well in Table 5 of Xue *et al.* (2010), which shows that different tools can achieve very high specificity, but only by greatly sacrificing sensitivity (and vice versa). When sensitivity and specificity are balanced, the most accurate tools can achieve rates for both simultaneously of ∼90%—a rate likely to be satisfactory when predicting sites in a limited number of proteins, but that would yield an unacceptable number of false positives when applied to an entire proteome. Unfortunately, using current prediction tools in genome annotation pipelines would therefore result in too many false positives (or too many false negatives, depending on the threshold selected). As such, the field of phosphorylation site prediction will not be truly mature until tools are developed that offer good sensitivity combined with very high specificity.

4.4 Making use of evolutionary information

Many types of functional sites in proteins and nucleic acids are known to be evolutionarily conserved, such as transcription factor binding sites (Berezikov *et al.*, 2004), mRNA splice junctions (Shapiro and Senapathy, 1987), microRNA target sites (Friedman *et al.*, 2009) and surface residues that participate in protein–protein interfaces (Caffrey *et al.*, 2004). The evolutionary conservation of phosphorylation sites has also been examined in numerous studies. For instance, the phosphorylation site Ser2 is conserved in versions of the small ubiquitin-like modifier (SUMO) protein in species as distantly related as human, *Saccharomyces cerevisiae* and *Drosophila melanogaster* (Matic *et al.*, 2008). Significant conservation of phosphorylation sites also occurs among different species of plants (Maathuis, 2008; Nakagami *et al.*, 2010). Some degree of conservation even extends to prokaryotes—although signaling via the phosphorylation of S/T/Y residues was once thought to be limited to eukaryotes, several such sites have been identified in *Escherichia coli* (Macek *et al.*, 2008) and *Bacillus subtilis* (Macek *et al.*, 2007).

As evolutionary information is valuable for many bioinformatics-related tasks, including protein structure prediction, gene finding, genome annotation and sequence assembly, it should prove valuable for phosphorylation site prediction as well. For example, Jalal *et al.* (2009) developed a protocol that uses known human phosphorylation sites to identify putative bovine sites. While not involving machine learning, the success of this approach shows the value of using evolutionary information in order to identify novel sites. Strangely, evolutionary information has largely been ignored in the context of identifying phosphorylation sites using machine learning. In one exception, Gnad *et al.* (2007) used information concerning phosphorylation site conservation to improve the accuracy of their SVM-based predictor. In future, evolutionary conservation of protein kinases (rather than, or in addition to, phosphorylation sites) may also prove useful in leveraging knowledge about one organism to predict phosphorylation sites in a second organism. Given the considerable predictive power of evolutionary information, its more widespread incorporation into future prediction tools has the potential to greatly increase accuracy.

5 CONCLUSION

There has already been a great deal of success in applying different methodologies to the problem of phosphorylation site prediction. Addressing the challenges outlined above, as well as those described by Xue *et al.* (2010), will require much coordination and effort, but would constitute significant steps forward for the field.

ACKNOWLEDGEMENTS

Special thanks to Chantel Krakowetz and Scott Napper for critical reading and advice, and to the reviewers for many helpful suggestions.

Funding: Natural Sciences and Engineering Research Council of Canada (NSERC), Vanier Canada Graduate Scholarship (to B.T.) and NSERC Discovery Grant (to A.K.).

Conflict of Interest: none declared.

REFERENCES

Basu,S. and Plewczyński,D. (2010) AMS 3.0: prediction of post-translational modifications. *BMC Bioinformatics*, **11**, 210.

Berezikov,E. *et al.* (2004) CONREAL: conserved regulatory elements anchored alignment algorithm for identification of transcription factor binding sites by phylogenetic footprinting. *Genome Res.*, **14**, 170–178.

Berry,E.A. *et al.* (2004) Reduced bio basis function neural network for identification of protein phosphorylation sites: comparison with pattern recognition algorithms. *Comput. Biol. Chem.*, **28**, 75–85.

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.* (1998) PhosphoBase: a database of phosphorylation sites. *Nucleic Acids Res.*, **26**, 382–386.

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.

Boersema,P.J. *et al.* (2009) Phosphopeptide fragmentation and analysis by mass spectrometry. *J. Mass Spectrom.*, **44**, 861–878.

Brinkworth,R.I. *et al.* (2003) Structural basis and prediction of substrate specificity in protein serine/threonine kinases. *Proc. Natl Acad. Sci. USA*, **100**, 74–79.

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.

Caffrey,D.R. *et al.* (2004) Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Sci.*, **13**, 190–202.

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

Dang,T.H. *et al.* (2008) Prediction of kinase-specific phosphorylation sites using conditional random fields. *Bioinformatics*, **24**, 2857–2864.

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.

Dunker,A.K. *et al.* (2002) Intrinsic disorder and protein function. *Biochemistry*, **41**, 6573–6582.

Durek,P. *et al.* (2009) Detection and characterization of 3D-signature phosphorylation site motifs and their contribution towards improved phosphorylation site prediction in proteins. *BMC Bioinformatics*, **10**, 117.

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

Fan,H.-C. *et al.* (2009) Activation of the TRPV4 ion channel is enhanced by phosphorylation. *J. Biol. Chem.*, **284**, 27884–27891.

Friedman,R.C. *et al.* (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.*, **19**, 92–105.

Gao,J. and Xu,D. (2010) The Musite open-source framework for phosphorylation-site prediction. *BMC Bioinformatics*, **11** (Suppl. 12), S9.

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

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

Ginalski,K. *et al.* (2003) 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics*, **19**, 1015–1018.

Gnad,F. *et al.* (2007) PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. *Genome Biol.*, **8**, R250.

Gnad,F. *et al.* (2011) PHOSIDA 2011: the posttranslational modification database. *Nucleic Acids Res.*, **39**, D253–D260.

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.

Hjerrild,M. and Gammeltoft,S. (2006) Phosphoproteomics toolbox: computational biology, protein chemistry and mass spectrometry. *FEBS Lett.*, **580**, 4764–4770.

Hjerrild,M. *et al.* (2004). Identification of phosphorylation sites in protein kinase A substrates using artificial neural networks and mass spectrometry. *J. Proteome Res.*, **3**, 426–433.

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

Houseman,B.T. *et al.* (2002) Peptide chips for the quantitative evaluation of protein kinase activity. *Nat. Biotechnol.*, **20**, 270–274.

Huang,H.-D. *et al.* (2005a) Incorporating hidden Markov models for identifying protein kinase-specific phosphorylation sites. *J. Comput. Chem.*, **26**, 1032–1041.

Huang,H.-D. *et al.* (2005b) KinasePhos: a web tool for identifying protein kinase-specific phosphorylation sites. *Nucleic Acids Res.*, **33**, W226–W229.

Huttlin,E.L. *et al.* (2010) A tissue-specific atlas of mouse protein phosphorylation and expression. *Cell*, **143**, 1174–1189.

Iakoucheva,L.M. *et al.* (2004) The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res.*, **32**, 1037–1049.

Ingrell,C.R. *et al.* (2007) NetPhosYeast: prediction of protein phosphorylation sites in yeast. *Bioinformatics*, **23**, 895–897.

Jalal,S. *et al.* (2009) Genome to kinome: species-specific peptide arrays for kinome analysis. *Sci. Signal.*, **2**, pl1.

Jiménez,J.L. *et al.* (2007) A systematic comparative and structural analysis of protein phosphorylation sites based on the mtcPTM database. *Genome Biol.*, **8**, R90.

Jung,I. *et al.* (2010) PostMod: sequence based prediction of kinase-specific phosphorylation sites with indirect relationship. *BMC Bioinformatics*, **11** (Suppl. 1), S10.

Karpenko,O. *et al.* (2008) A probabilistic meta-predictor for the MHC class II binding peptides. *Immunogenetics*, **60**, 25–36.

Kemp,B.E. *et al.* (1977) Role of multiple basic residues in determining the substrate specificity of cyclic AMP-dependent protein kinase. *J. Biol. Chem.*, **252**, 4888–4894.

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.

Kim,J.H. *et al.* (2004) Prediction of phosphorylation sites using SVMs. *Bioinformatics*, **20**, 3179–3184.

Kitchen,J. *et al.* (2008) Charge environments around phosphorylation sites in proteins. *BMC Struct. Biol.*, **8**, 19.

Kobe,B. *et al.* (2005) Substrate specificity of protein kinases and computational prediction of substrates. *Biochim. Biophys. Acta*, **1754**, 200–209.

Koenig,M. and Grabe,N. (2004) Highly specific prediction of phosphorylation sites in proteins. *Bioinformatics*, **20**, 3620–3627.

Kreegipuu,A. *et al.* (1999) PhosphoBase, a database of phosphorylation sites: release 2.0. *Nucleic Acids Res.*, **27**, 237–239.

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

Linding,R. *et al.* (2007) Systematic discovery of in vivo phosphorylation networks. *Cell*, **129**, 1415–1426.

Liu,J. *et al.* (2007) Meta-prediction of protein subcellular localization with reduced voting. *Nucleic Acids Res.*, **35**, e96.

Li,L. *et al.* (2008a) Prediction of phosphotyrosine signaling networks using a scoring matrix-assisted ligand identification approach. *Nucleic Acids Res.*, **36**, 3263–3273.

Li,T. *et al.* (2008b) Prediction of kinase-specific phosphorylation sites with sequence features by a log-odds ratio approach. *Proteins*, **70**, 404–414.

Li,T. *et al.* (2010) Identifying human kinase-specific protein phosphorylation sites by integrating heterogeneous information from various sources. *PLoS One*, **5**, e15411.

Löwenberg,M. *et al.* (2005) Rapid immunosuppressive effects of glucocorticoids mediated through Lck and Fyn. *Blood*, **106**, 1703–1710.

Maathuis,F.J. (2008) Conservation of protein phosphorylation sites within gene families and across species. *Plant Signal. Behav.*, **3**, 1011–1013.

Macek,B. *et al.* (2007) The serine/threonine/tyrosine phosphoproteome of the model bacterium Bacillus subtilis. *Mol. Cell Proteomics*, **6**, 697–707.

Macek,B. *et al.* (2008) Phosphoproteome analysis of E. coli reveals evolutionary conservation of bacterial Ser/Thr/Tyr phosphorylation. *Mol Cell Proteomics*, **7**, 299–307.

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

Matic,I. *et al.* (2008) Phosphorylation of SUMO-1 occurs in vivo and is conserved through evolution. *J. Proteome Res.*, **7**, 4050–4057.

Meier,R. *et al.* (1997) Mitogenic activation, phosphorylation, and nuclear translocation of protein kinase Bbeta. *J. Biol. Chem.*, **272**, 30491–30497.

Miller,M.L. and Blom,N. (2009) Kinase-specific prediction of protein phosphorylation sites. *Methods Mol. Biol.*, **527**, 299–310.

Moses,A.M. *et al.* (2007) Clustering of phosphorylation site recognition motifs can be exploited to predict the targets of cyclin-dependent kinase. *Genome Biol.*, **8**, R23.

Nakagami,H. *et al.* (2010) Large-scale comparative phosphoproteomics identifies conserved phosphorylation sites in plants. *Plant Physiol.*, **153**, 1161–1174.

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

Obenauer,J.C. *et al.* (2003) Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res.*, **31**, 3635–3641.

Plewczyński,D. *et al.* (2005) A support vector machine approach to the identification of phosphorylation sites. *Cell Mol. Biol. Lett.*, **10**, 73–89.

Plewczyński,D. *et al.* (2008) AutoMotif Server for prediction of phosphorylation sites in proteins using support vector machine: 2007 update. *J. Mol. Model*, **14**, 69–76.

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

Ryu,G.-M. *et al.* (2009) Genome-wide analysis to predict protein sequence variations that change phosphorylation sites or their corresponding kinases. *Nucleic Acids Res.*, **37**, 1297–1307.

Saunders,N.F.W. and Kobe,B. (2008) The Predikin webserver: improved prediction of protein kinase peptide specificity using structural information. *Nucleic Acids Res.*, **36**, W286–W290.

Saunders,N.F.W. *et al.* (2008) Predikin and PredikinDB: a computational framework for the prediction of protein kinase peptide specificity and an associated database of phosphorylation sites. *BMC Bioinformatics*, **9**, 245.

Schwartz,D. *et al.* (2009) Predicting protein post-translational modifications using meta-analysis of proteome scale data sets. *Mol. Cell Proteomics*, **8**, 365–379.

Senawongse,P. *et al.* (2005) Predicting the phosphorylation sites using hidden Markov models and machine learning methods. *J. Chem. Inf. Model.*, **45**, 1147–1152.

Shapiro,M.B. and Senapathy,P. (1987) RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. *Nucleic Acids Res.*, **15**, 7155–7174.

Shen,H.-B. *et al.* (2007) Euk-PLoc: an ensemble classifier for large-scale eukaryotic protein subcellular location prediction. *Amino Acids*, **33**, 57–67.

Sigrist,C.J.A. *et al.* (2002) PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief. Bioinformatics*, **3**, 265–274.

Slaugenhaupt,S.A. *et al.* (2001) Tissue-specific expression of a splicing mutation in the IKBKAP gene causes familial dysautonomia. *Am. J. Hum. Genet.*, **68**, 598–605.

Sobolev,B. *et al.* (2010) Functional classification of proteins based on projection of amino acid sequences: application for prediction of protein kinase substrates. *BMC Bioinformatics*, **11**, 313.

Songyang,Z. *et al.* (1994) Use of an oriented peptide library to determine the optimal substrates of protein kinases. *Curr. Biol.*, **4**, 973–982.

Swaminathan,K. *et al.* (2010) Enhanced prediction of conformational flexibility and phosphorylation in proteins. *Adv. Exp. Med. Biol.*, **680**, 307–319.

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.

Trost,B. *et al.* (2007) Strength in numbers: achieving greater accuracy in MHC-I binding prediction by combining the results from multiple prediction tools. *Immunome Res.*, **3**, 5.

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

Wan,J. *et al.* (2008) Meta-prediction of phosphorylation sites with weighted voting and restricted grid search parameter selection. *Nucleic Acids Res.*, **36**, e22.

Wang,M. *et al.* (2008a) Prediction of PK-specific phosphorylation site based on information entropy. *Sci. China C Life Sci.*, **51**, 12–20.

Wang,P. *et al.* (2008b) A systematic assessment of MHC class II peptide binding predictions and evaluation of a consensus approach. *PLoS Comput. Biol.*, **4**, e1000048.

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

Ward,P. *et al.* (2004) Protein kinases of the human malaria parasite Plasmodium falciparum: the kinome of a divergent eukaryote. *BMC Genomics*, **5**, 79.

Wong,Y.-H. *et al.* (2007) KinasePhos 2.0: a web server for identifying protein kinase-specific phosphorylation sites based on sequences and coupling patterns. *Nucleic Acids Res.*, **35**, W588–W594.

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.* (2005) GPS: a comprehensive WWW server for phosphorylation sites prediction. *Nucleic Acids Res.*, **33**, W184–W187.

Xue,Y. *et al.* (2006) PPSP: prediction of PK-specific phosphorylation site with Bayesian decision theory. *BMC Bioinformatics*, **7**, 163.

Xue,Y. *et al.* (2008) GPS 2.0, a tool to predict kinase-specific phosphorylation sites in hierarchy. *Mol. Cell Proteomics*, **7**, 1598–1608.

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

Xue,Y. *et al.* (2011) GPS 2.1: enhanced prediction of kinase-specific phosphorylation sites with an algorithm of motif length selection. *Protein Eng. Des. Sel.*, **24**, 255–260.

Yaffe,M.B. *et al.* (2001) A motif-based profile scanning approach for genome-wide prediction of signaling pathways. *Nat. Biotechnol.*, **19**, 348–353.

Yoo,P.D. *et al.* (2008) SiteSeek: post-translational modification analysis using adaptive locality-effective kernel methods and new profiles. *BMC Bioinformatics*, **9**, 272.

Yu,Z. *et al.* (2010) Identifying protein-kinase-specific phosphorylation sites based on the Bagging-AdaBoost ensemble approach. *IEEE Trans. Nanobioscience*, **9**, 132–143.

Zanzoni,A. *et al.* (2007) Phospho3D: a database of three-dimensional structures of protein phosphorylation sites. *Nucleic Acids Res.*, **35**, D229–D231.

Zanzoni,A. *et al.* (2011) Phospho3D 2.0: an enhanced database of three-dimensional structures of phosphorylation sites. *Nucleic Acids Res.*, **39**, D268–D271.

Zetterqvist,O. *et al.* (1976) The minimum substrate of cyclic AMP-stimulated protein kinase, as studied by synthetic peptides representing the phosphorylatable site of pyruvate kinase (type L) of rat liver. *Biochem. Biophys. Res. Commun.*, **70**, 696–703.

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.

Zhou,F.-F. *et al.* (2004) GPS: a novel group-based phosphorylation predicting and scoring method. *Biochem. Biophys. Res. Commun.*, **325**, 1443–1448.

Zhou,T. *et al.* (2006) Docking interactions induce exposure of activation loop in the MAP kinase ERK2. *Structure*, **14**, 1011–1019.