BIOINFORMATICS ORIGINAL PAPER Vol. 28 no. 7 2012, pages 962–969
doi:10.1093/bioinformatics/bts060

Systems biology

Advance Access publication February 1, 2012

Predicting kinase substrates using conservation of local motif density

Andy C. W. Lai, Alex N. Nguyen Ba and Alan M. Moses*

Department of Cell and Systems Biology, University of Toronto, Toronto, Canada M5S 3G5

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Protein kinases represent critical links in cell signaling. A central problem in computational biology is to systematically identify their substrates.

Results: This study introduces a new method to predict kinase substrates by extracting evolutionary information from multiple sequence alignments in a manner that is tolerant to degenerate motif positioning. Given a known consensus, the new method (ConDens) compares the observed density of matches to a null model of evolution and does not require labeled training data. We confirmed that ConDens has improved performance compared with several existing methods in the field. Further, we show that it is generalizable and can predict interesting substrates for several important eukaryotic kinases where training data is not available.

Availability and implementation: ConDens can be found at http://www.moseslab.csb.utoronto.ca/andyl/.

Contact: alan.moses@utoronto.ca

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

Received on October 12, 2011; revised on January 24, 2012; accepted on January 25, 2012

1 INTRODUCTION

Protein phosphorylation is a well studied class of post-translational modification that has powerful influence over the dynamics of biological systems. It is characterized by the addition of a phosphate group (PO_4) to an amino acid residue by a kinase enzyme (Berg, 2007). Proteins that are subjected to these events often undergo a biochemical change that can, in turn, affect biological pathways associated with them (Cohen, 1982).

Kinases often choose phosphorylation targets selectively and different kinases may favor residues with different local arrangements of amino acids (which are referred to as 'consensus sequences' or 'motifs') (Collins *et al.*, 2007). However, consensus sequences alone are often insufficient for kinase substrate prediction because they tend to be short and degenerate and matches are expected to occur frequently by chance in random sequences.

Computational kinase substrate predictors take advantage of a myriad of local biological information in generating their predictions. This includes amino acid arrangement (Blom *et al.*, 1999; Lam *et al.*, 2010; Obenauer *et al.*, 2003; Xue *et al.*, 2008, 2011; Zhou *et al.*, 2004), biochemical/structural property (Blom *et al.*, 1999; Iakoucheva *et al.*, 2004; Lam *et al.*, 2010) and quantity/density of consensus matches (Chang *et al.*, 2007; Gnad *et al.*, 2011; Lam *et al.*, 2010; Moses *et al.*, 2007). There are also some that incorporate interactome information to further enhance their predictions (Li *et al.*, 2010; Linding *et al.*, 2008).

The focus of our work is to develop a kinase substrate prediction method based on motif conservation over evolution. This idea is based on the work in a previous study (Budovskaya *et al.*, 2005) which illustrated that consensus sites for protein kinase A (or PKA) are more likely to be phosphorylated if preserved over longer evolutionary distances.

A challenge in using this approach is that patterns of molecular evolution in protein sequences are highly heterogeneous. In many cases, a match to a consensus sequence (which we will also refer to as 'match', 'motif match' or 'consensus match') can be considered as functionally conserved if it is well aligned in a multiple sequence alignment (or MSA) (Fig. 1A). However, in situations where the entire local sequence neighborhood is also conserved, it will be difficult to tell whether the matches are 'specifically' conserved due to a kinase substrate interaction or 'non-specifically' conserved as part of a larger domain with a different function (Fig. 1C) (Budovskaya *et al.*, 2005).

At the same time, functional phosphorylation motif matches do not necessarily align in a MSA. Previous studies have shown that phosphorylation sites are often not positionally conserved despite being located in the same local region in orthologous proteins (Ba and Moses, 2010; Chang *et al.*, 2007; Holt *et al.*, 2009; Moses *et al.*, 2007). For such examples (Fig. 1B), one can observe that the quantity of matches in the local sequence neighborhood is consistent among orthologous sequences (Ba and Moses, 2010; Holt *et al.*, 2009; Moses *et al.*, 2007). Nevertheless, all of these cases contrast with the patterns observed for randomly occurring matches to the consensus site (Fig. 1D) that typically show no consistent conservation patterns and disappear quickly.

Current conservation-based prediction methodologies are largely focused on the residue conservation of motif matches (Budovskaya *et al.*, 2005; Gnad *et al.*, 2011; Lam *et al.*, 2010). Since these strategies are dependent on the positional conservation of the matches of interest, they may be insensitive towards situations where matches are not positionally conserved (Fig. 1B). Furthermore, the lack of consideration for the local region's conservation can lead to false detection of matches located in highly conserved neighborhoods (Fig. 1C).

To address these issues, we designed a new kinase substrate prediction method (ConDens) that (i) considers the conservation of the number of motif matches in a local region of the protein, rather than the alignment of individual phosphorylation sites and (ii) uses an evolutionary model to account for the local sequence

*To whom correspondence should be addressed.

divergence to avoid detection of spurious matches in conserved domains.

Since the method uses a null model of evolution, it does not require a labeled set of bona fide phosphorylation sites and negative motif matches. Although it does require an input consensus sequence, we consider it to be unsupervised relative to other methods (Lam *et al.*, 2010; Xue *et al.*, 2006, 2011) that train their classification models based on the datasets of previously known phosphorylation sites.

We compared ConDens' ability to predict Cdc28 phosphorylation substrates against that of several methodologically different phospho-predictors in the field and showed it had an improved classification performance compared with these other methods. Finally, we used ConDens to scan the *Saccharomyces cerevisiae* genome for potential phosphorylation substrates of Mec1/Tel1, Prk1, Ipl1, PKA, CKII and Ime2 kinases. In all but one case, our results indicated a statistically significant enrichment of known kinase substrates among our predictions.

2 METHODS

2.1 ConDens, a new kinase–substrate prediction method

2.1.1 Overview ConDens is designed to assess the conservation of the number of matches in local regions of a protein of interest (Z), given the local sequence divergence and without relying on alignment of matches between orthologous sequences. ConDens first defines a reference local region around each motif match m and counts the number of motif matches located in corresponding regions within Z's orthologs. Then for each orthologous region, a 'pair-wise' *P*-value is computed to test the null hypothesis of observing the number of matches in Z's orthologs in the absence of specific selection to retain the motif matches (see Section 2.1.2). The score for a match is then derived by combining these pair-wise *P*-values using Fisher's method. Finally, ConDens assigns protein Z the most significant score of the matches in its primary sequence.

2.1.2 The ConDens model In principle, a motif match can be conserved 'specifically' (due to the unique properties of the motif) or 'non-specifically' (due to being part of a conserved domain). In the absence of purifying selection specific for a kinase–substrate interaction, the number of associated phosphorylation motif matches in the sequence is expected to approach an equilibrium level over evolution as dictated by random mutation events and the local evolutionary rate. On the other hand, where there is purifying selection for a kinase–substrate interaction, the matches are likely to be conserved in quantity over evolution (Moses *et al.*, 2007) and thus deviate from an equilibrium level or approach this level more slowly.

We formulated a statistical model to detect selection on the number of matches to a consensus sequence, C, by rejecting the null hypothesis of evolution according to a local evolutionary model with no specific constraint to retain motif matches. For our purpose, we defined a consensus sequence C of width w as an array of character sets $C_1,\ldots,C_w$, where C_j represents the set of allowable amino acid residues in the j-th position in the consensus. As an example, the S/T-P consensus would be represented as [(S,T),(P)] and have $w=2$. Given peptide sequences S and S', we estimate the evolutionary distance, t, between S and S' and compute the distribution of matches in S' conditioned on S, an amino acid substitution model (SM) and t. This evolutionary distance t is estimated based on the model chosen. For instance, t would be the Jukes–Cantor distance if SM is the Jukes–Cantor substitution model.

To compute the distribution of the number of matches to C under the null hypothesis (or the 'null distribution'), we first determine the probability p_i that a match to C begins at the ith position in S after t evolutionary distance, assuming evolution of amino acids occurs as dictated by SM.

$$p_i = \Pr(\textit{match to } C \textit{ at position } i | S_i,\ldots S_{i+w-1}, SM, t)$$
$$= \prod_{j=1}^{w} \sum_{A\in C_j} \Pr(A | S_{i+j-1}, SM, t) \quad (1)$$

In this equation, S_x is the xth residue in sequence S, and A are the allowed amino acid residues at the jth position of consensus C.

At any time in evolution, each position can either match C or not. Therefore, Equation (1) can be used to approximate the null distribution of match counts in sequence S' as a sum of $N=|S|-w+1$ independent Bernoulli random variables with success probabilities $p_1,\ldots,p_N$. Since these Bernoulli random variables can have different parameters, we used a generalized form of the binomial distribution known as the 'Poisson binomial distribution' (Wang, 1993) to approximate the probability of observing n matches to the consensus sequence C after evolutionary distance t:

$$\Pr(\#\textit{ matches in } S' | S, SM, t) \approx PBN(n | p_1,\ldots p_N) \quad (2)$$

$$PBN(n|p_1,\ldots,p_N) = \sum_{X|\sum_j X_j = n} \prod_{i=1}^{N} p_i^{X_i}(1-p_i)^{1-X_i} \quad (3)$$

Where $PBN(n|p_1,\ldots,p_N)$ is the Poisson binomial probability density function (Wang, 1993), $X_i \in \{0,1\}$ is an indicator variable that equals 1 if the ith position matches the motif and 0 otherwise, and the sum is over all possible $X=[X_1,\ldots,X_N]$ with exactly n matches. We used dynamic programming to compute Equation (3) in $O(N^2)$ time. Details can be found in the Supplementary Material.

2.1.3 The ConDens algorithm We are given a protein of interest Z, a consensus C, a substitution model SM, a multiple sequence alignment of Z and orthologs, and a search radius parameter $d \in \mathbb{N}$. We compute a score for each consensus match m in Z, as follows (see Supplementary Fig. 1 for illustration):

Step 1: Find S, a region in Z that encompasses $\pm d$ positions around m's first position.

Step 2: Find the columns L in the multiple sequence alignment that corresponds to S.

Step 3: For each ortholog R in the multiple sequence alignment:

(i) Compute an evolutionary distance t based on the local sequence divergence between proteins Z and R at columns L

(ii) Find the position in R that aligns to the location of m in Z and then count the number of motif matches, n, within $\pm d$ residues around that position. If this position is a gap, choose the immediately preceding non-gapped position in R's primary sequence.

(iii) Compute the pair-wise P-value, $\Pr(\#matches \geq n | S, SM, t)$, by directly summing terms computed using Equation (3)

Step 4: Compute a score for match m by combining the pair-wise P-values acquired in Step 3 by using Fisher's method.

Once the scores for all m in the protein Z are calculated, the lowest of them is transformed to $-\log_e$ space and output as Z's score. Higher scores indicate a greater confidence that there is a match in the protein that is specifically conserved.

2.1.4 ConDens parameter and species choice In this study, the ConDens algorithm was applied to the *S. cerevisiae* proteome. The window radius d was chosen to be 20 because that would provide tolerance to the positional degeneracy of phosphorylated sites. Protein disordered region predictors (Linding *et al.*, 2003) use windows of this size, suggesting a length scale for the rapidly evolving regions that contain phosphorylation sites. The amino acid substitution model SM chosen for this study was $JC69_{AA}$, which is a derivative of the $JC69$ model (Jukes and Cantor, 1969) adapted to amino acids instead of nucleotides (i.e. 20 residue types instead of 4).

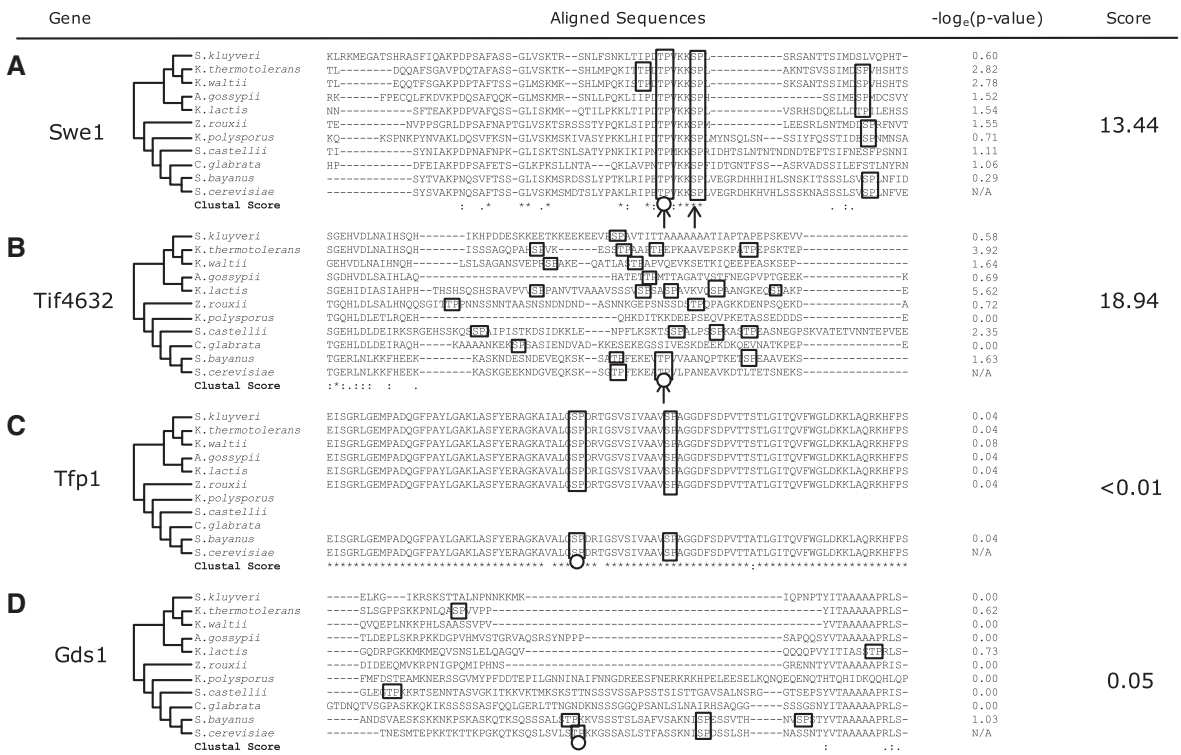

Fig. 1. Patterns of match conservation. Multiple sequence alignments are shown with S/T-P matches boxed and bona fide phosphorylation sites (Harvey *et al*., 2005; Holt *et al*., 2009;) labeled with arrows. Each alignment has a match of interest on the *S*. *cerevisiae* sequence marked with a circle. The ConDens pairwise *P*-values for these matches of interest are shown on the right of each ortholog ['-log$_e$ (*P*-value)'] and the overall scores for these matches are shown as large numbers ('Score'). (**A**) Swe1 and orthologs. Marked match starts at Thr 196 where the local rate of evolution is high but the consensus site itself is conserved. (**B**) Tif4632 and orthologs. Marked match starts at Thr 196 where the local quantity of S/T-P matches is conserved but not aligned. (**C**) Tfp1 and orthologs. Marked match starts at Ser 858 (Ser in first motif match) where local rate of evolution is low. Both matches in the *S*. *cerevisiae* are perfectly aligned but are not thought to be phosphorylated by Cdc28. (**D**) Gds1 and orthologs. Marked match starts at Thr 363 (Thr in first match) where local rate of evolution is high. Both matches in the *S*. *cerevisiae* protein are neither conserved nor thought to be phosphorylated by Cdc28.

The protein sequences of the *S*. *cerevisiae* proteome were a set of 5885 non-dubious open-reading frames (ORFs) in the *Saccharomyces* Genome Database (or 'SGD') (Cherry *et al*., 1998) on September 2010. Multiple sequence alignments were computed using MUSCLE (Edgar, 2004) and based on protein orthologs of *S*. *cerevisiae* in *Saccharomyces bayanus*, *Candida glabrata*, *Saccharomyces castellii*, *Kluyveromyces polysporus*, *Zygosaccharomyces rouxii*, *Kluyveromyces lactis*, *Ashbya gossypii*, *Kluyveromyces waltii*, *Kluyveromyces thermotolerans* and *Saccharomyces kluyveri*. The phylogenetic relationships between species were acquired from (Conde e Silva *et al*., 2009) and the homology relationships between their proteins were obtained from the Yeast Genome Order Browser (or 'YGOB') (Muffato *et al*., 2010).

The choice of species can affect the performance of our method. In absence of substantial sequence divergence, the conservation of motif matches will almost certainly be expected under the null hypothesis. On the other hand, motif matches are less likely to be conserved between very distantly-related species because the kinase substrate interaction may have diverged. Highly gapped alignments and repetitive sequences can also pose problems in some cases (Supplementary Fig. 2)

The effects of using different sets of budding yeast species in the multiple sequence alignments were studied and it was found that the AUC and AUC$_{50}$ of ConDens reached a plateau when seven or more species were included in the sequence alignments (Supplementary Table 1). A similar consistency of AUC and AUC$_{50}$ was observed when different values for the window radii parameter ($d = 10$, 15, 20, 25, 30) were used (Supplementary Table 2).

Table 1. Performance of supervised and unsupervised classifiers

Classifier	AUC$_{whole}$	AUC50$_{whole}$	AUC$_{sparse}$	AUC50$_{sparse}$
Unsupervised				
ConDens	**0.790**	**0.039**	**0.753**	**0.034**
Scansite	0.648	0.017	0.624	0.018
S_{LR}	0.555	0.021	0.498	0.015
Supervised				
GPS 2.1	0.743	0.030	0.710	0.028
MOTIPS	0.627	0.032	0.582	0.029
Random	0.500	0.008	0.500	0.009

Numbers in bold indicate the highest value among the five classifiers. The 'whole' subscript indicates the AUC or AUC50 scores are derived from the entire validation dataset. The 'sparse' subscript indicates the AUC or AUC50 scores are derived from proteins in the validation dataset that have a sparse spatial distribution of S/T-P matches. Associated ROC plots can be found in Supplementary Figure 4.

2.1.5 ConDens implementation and usage Computational implementation of ConDens and details of its use can be found at http://www.moseslab.csb.utoronto.ca/andyl. For the purpose of substrate prediction, we recommend users to either follow a ConDens cut-off of nine

Table 2. Performance of supervised and unsupervised classifiers

Kinase(s)	Consensus	Hits	$Enrichment_{whole}$	$Enrichment_{motif}$
Mec1/Tel1	S/T-Q	46	**29.9-fold**	**21.4-fold**
Prk1	L/I/V/M-X_4-T-G	9	**326.9-fold**	**82.6-old**
Ipl1	R/K-X-S/T-L/I/V	40	**36.8-fold**	**22.6-fold**
PKA	R-R/K-X-S	52	**16.6-fold**	**3.3-fold**
CKII	S/T-D/E-X-D/E	170	**10.1-fold**	**5.1-fold**
Ime2	R-P-X-S/T	6	0-fold	0-fold

$Enrichment_{whole}$ denotes the enrichment of known kinase substrates in the hits relative to all non-dubious ORFs in the *S. cerevisiae* genome. $Enrichment_{motif}$ denotes the enrichment of known targets in the hits relative to all other non-dubious ORFs in the *S. cerevisiae* genome that have at least one match to the consensus and at least one ortholog annotated by YGOB. The enrichment values in bold denote statistically significant ($P < 0.05$) enrichment based on a one-tailed Fisher's exact test.

(see Section 3.3) or to choose the top x proteins from the results (with x being the number of proteins the user would like to examine).

To facilitate manual verification of our predictions, a browser was also provided to allow users to view the multiple sequence alignments of individual predictions.

2.2 Performance evaluation

2.2.1 Cdc28 dataset We assembled a 'Cdc28 dataset' consisting of proteins phosphorylated by the Cdc28 kinase ('positives') and proteins not phosphorylated by the Cdc28 kinase ('negatives'). This data was drawn from two Cdc28 phosphorylation studies—an *in vitro* kinase assay by Ubersax *et al.* (2003) and an *in vivo* genome-wide mass spectrometry experiment by Holt *et al.* (2009). Proteins with one or more Cdc28 phosphorylation site in the *in vivo* study were considered to be positives and proteins not discovered to be Cdc28 targets in both studies were considered to be 'negatives' (Supplementary Fig. 3). We made a special exception to Cln2 despite it being not a hit in both studies, since it was a well known Cdc28 substrate (Deshaies *et al.*, 1995; Lanker *et al.*, 1996) and was the only false negative prediction discussed by the *in vitro* study that was not recovered by the *in vivo* study.

We explicitly excluded Cdc28 targets found in the *in vitro* study that were not confirmed in the *in vivo* study, since the conditions used in *in vitro* studies may not reflect the characteristics of *in vivo* environments.

To define a subset of the Cdc28 dataset with sparsely distributed S/T-P matches, we used the S_{LR} algorithm (Moses *et al.*, 2007) to compute a cluster score that measures the 'clusteredness' of S/T-P matches. For our purposes, we consider proteins that score <3 to have a sparse spatial distribution of S/T-P matches (or 'unclustered'). The command line operation we used was `perl compute_SLR.pl <ORF FILE> '[ST]P'`, where <ORF File> is a FASTA file of all non-dubious *S. cerevisiae* protein sequences from SGD. It is important to note that this is different to the S_{LR} score, which uses the strong Cdc28 consensus (S/T-P-X-R/K) in addition to the S/T-P consensus.

2.2.2 Phosphorylation consensus sequences In our experiments, we used an S/T-P consensus to detect the Cdc28 targets. Although the kinase is widely reported to have a more stringent S/T-P-X-R/K consensus (Friedman *et al.*, 1996) (which we refer to as a 'strong' Cdc28 consensus), we chose not to use it due to many known target sites not having the R/K at the +3 position (Holt *et al.*, 2009).

Based on published literature, we also derived consensus sequences for Mec1/Tel1 (Schwartz *et al.*, 2002), Prk1 (Huang *et al.*, 2003), Ipl1 (Cheeseman *et al.*, 2002), Ime2 (Holt *et al.*, 2007), CKII (Meggio and Pinna, 2003; Niefind *et al.*, 2007) and PKA (Budovskaya *et al.*, 2005; Kemp and Pearson, 1990; Townsend *et al.*, 1996) (Table 2).

2.2.3 Known targets of other kinases There are generally no large datasets for other *S.* cerevisiae kinases on the scale of the Cdc28 dataset. As a result, we obtained known targets of these kinases from the PhosphoGRID database (Stark *et al.*, 2010) and Kinase Interaction Database (KID) (Sharifpoor *et al.*, 2011). Specifically, we took substrates with two or more references in PhosphoGRID and substrates that met the high confidence threshold (6.60) in KID. Information from these databases was retrieved on Dec 2011.

2.2.4 Collection of classification data The following are instructions for obtaining Cdc28 prediction data from various protein phosphorylation classifiers. For methods that were specific to human kinases, data from Cdc2 (human ortholog of Cdc28) was collected instead.

Motif Analysis Pipeline: Protein sequences from the whole S. cerevisiae genome were submitted to the Motif analysis pipeline (MOTIPS) web interface (http://motips.gersteinlab.org/) (Lam *et al.*, 2010). Scores for each protein were taken directly from the program output. The parameters used for the program are same as default; we used the Cdc28 dataset as its training data and provided it with a PWM determined from the Cdc28-phosphorylated sites denoted in (Holt *et al.*, 2009).

Group-based Prediction System 2.1 (GPS): Protein sequences from the whole *S.* cerevisiae genome were submitted to the GPS 2.1 program (Zhou *et al.*, 2004). The 'Threshold' parameter was set to 'All'. Matches were assigned scores for the human Cdc2 kinase from program output. Each protein was then assigned a score equivalent to the highest given to any of their constituent matches.

Scansite 2.0: Protein sequences from the whole *S. cerevisiae* genome were submitted to Scansite web interface (http://scansite.mit.edu/) (Obenauer *et al.*, 2003) using the low stringency option. Matches were assigned scores for the human Cdc2 kinase from program output. Each protein was then assigned a score equivalent to the lowest given to any of their constituent matches.

S_{LR}: The S_{LR} program was run iteratively on every protein sequence in the *S. cerevisiae* genome. As per recommendation by the paper (Moses *et al.*, 2007), we used both the S/T-P and S/T-P-X-R/K consensuses for predicting Cdc28 targets. The command line operation we used was `perl compute_SLR.pl <ORF File> '[ST]P.[RK]' '[ST]P'`, where <ORF File> is a FASTA file of all non-dubious *S. cerevisiae* protein sequences from SGD. We then parsed the resulting score for each individual protein from the output.

2.2.5 Measures of classification performance The performance of a classifier was evaluated using the area under receiver operator characteristic curve (AUC) (Fawcett, 2004) and AUC50 (Bauer *et al.*, 2011; Gribskov and Robinson, 1996), which is the AUC measured for the first 50 false positives. In measuring the AUC, proteins for which no prediction was made were considered 'missing data' and were excluded. Data-points with identical scores were ranked lexically by their ORF identifiers. For plotting precision and recall and the score distribution histogram (shown in Fig. 2) 'missing data' were included in the lowest bin.

2.2.6 Biological analysis using FunSpec FunSpec (Robinson *et al.*, 2002) was used to analyze the biological functions of individual proteins. The parameters we used are the same as the default settings with Bonferroni correction turned on. Emphasis was given to results in 'MIPS Functional Classification', 'GO Biological Process', and 'GO Molecular Function'.

3 RESULTS AND DISCUSSION

3.1 Predicting targets of Cdc28

In principle, computational predictors can predict phosphorylation sites, phosphorylated proteins or both. ConDens can serve either purpose. We decided to focus on substrate prediction at a protein-level because we have a comprehensive dataset of Cdc28 substrates

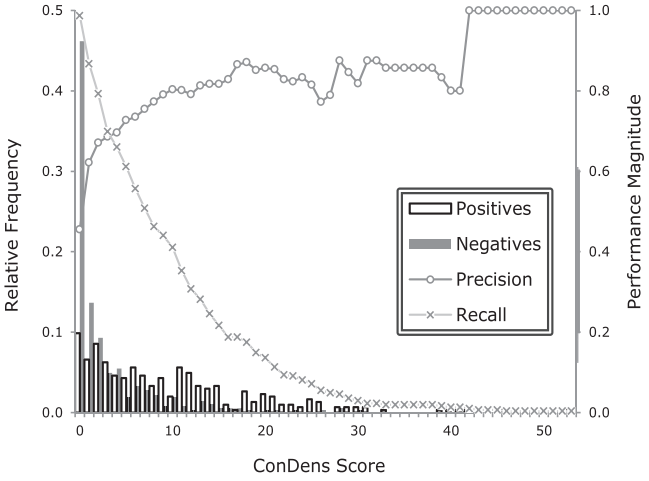

Fig. 2. Classification power of the ConDens score. Distribution of ConDens scores among Cdc28 targets ('positives', white bars) and non-targets ('negatives', gray bars). The bin size for the histogram was 2, and the left vertical axis shows the frequency. Also plotted is the precision (circles) and recall (crosses) of ConDens' binary classification over a range of score cut-offs and is indicated on the right vertical axis.

in *S. cerevisiae*, as well as a confident set of proteins that are unlikely to be substrates of this kinase (see Section 2.2). We also studied the method's utility in Cdc28 phosphorylation site prediction and results are illustrated in Supplementary Table 3.

To test ConDens' classification power at a protein-level, we computed a ConDens score (see Section 2.1) for every protein in the Cdc28 dataset (see Section 2.2). The known targets (or 'positives') and non-targets (or 'negatives') distributed differently over this score spectrum (Fig. 2), with the positives being noticeably shifted toward higher scores (due to having lower pair-wise *P*-values). Because both the positives and negative proteins contain matches to the Cdc28 consensus, these results indicate that ConDens scores will differentiate bona fide kinase targets from other motif-containing proteins. This Cdc28 dataset was also used to benchmark ConDens against two other unsupervised kinase substrate predictors: Scansite (Obenauer *et al*., 2003) and S_{LR}. (Moses *et al*., 2007). Scansite is a method that finds good matches to a position weight matrix and S_{LR} is a method that uses the spatial distribution of motif matches in the primary amino acid sequence.

The performances of these classifiers were compared using the area under ROC curves (AUC). To assess the classifiers' utility in guiding experimental kinase substrate discovery, we also computed the AUC50 (Bauer *et al*., 2011; Gribskov and Robinson, 1996), see Section 2.2.5.

Overall, ConDens' AUC (0.790) was substantially higher than that of S_{LR} (0.555), Scansite (0.648) and the expectation for a random classifier (0.500) (Table 1). The same could be said about the AUC50 scores, although the difference between AUC50s of ConDens (0.039) and S_{LR} (0.021) was not as large.

Since S_{LR} was based on the detection of motif match clusters, it is intrinsically unsuited to predict kinase substrates that have a sparse spatial distribution of motif matches. To test whether or not ConDens suffered from the same shortcoming, we repeated the same classification analysis on proteins in the dataset that do not possess spatially clustered S/T-P matches (see Section 2.2.1).

Under this circumstance, S_{LR}'s AUC was no better than the random expectation. On the other hand, ConDens' AUC only decreased by 5% (0.753, Table 1), which indicates that spatial clustering of matches had little effect on ConDens' classification performance. Taken together, the results indicated ConDens had a superior predictive performance compared with these other unsupervised methods, even when spatial distribution of matches is sparse.

3.2 Comparison with supervised predictors

Although ConDens is an 'unsupervised' method insofar as it does not require a labeled set of positive and negative examples for training of parameters, we could also compare its performance with supervised methods. We repeated the experiments in Section 3.1 using GPS 2.1 (Zhou *et al*., 2004), trained on Cdc2 targets (the human homolog of Cdc28) and MOTIPS (Lam *et al*., 2010), which we trained on our Cdc28 dataset. The AUCs of the supervised methods were much closer to those achieved by ConDens. Remarkably, ConDens' AUC50s were notably higher (Table 1), than these supervised methods.

This is important because ConDens does not require training data, but can still obtain strong classification results. We therefore suggest that ConDens will be particularly useful to identify substrates for kinases when training sets of characterized substrates unavailable, thus retaining the portability of unsupervised methods.

3.3 Substrate prediction for other *S. cerevisiae* kinases

ConDens' generalizability was tested on several additional kinases in *S. cerevisiae* (Table 2). For validation, experimentally verified substrates from PhosphoGRID (Stark *et al*., 2010) and Kinase Interaction Database (KID) (Sharifpoor *et al*., 2011) were selected as 'known kinase targets'. While these databases may not be exhaustive sources of information, they presented a quick and tractable means of keeping track of the ever-expanding phosphoproteome.

We performed a binary classification experiment where proteins with a ConDens score greater than 9 were considered as predicted kinase substrates or 'hits'. The threshold was decided based on classification performance with the Cdc28 dataset (Fig. 2). The enrichment of the known kinase targets among the hits was assessed.

Encouragingly, we found statistically significant ($P < 0.05$, Fisher's Exact Test) enrichment of known targets (or true positives) among hits (Table 2) for the additional consensus sequences tested. In all, 323 predictions were made for the 6 kinases with 26 being found in PhosphoGRID and KID as bona fide substrates and the remaining 297 being 'novel predictions'. Our hits were, on average, 10 times more enriched in known targets than what would be expected from a random sample of consensus-containing *S. cerevisiae* proteins and 25 times more enriched in known targets than what would be expected from a random sample of *S. cerevisiae* proteins. The biological functions of each set of predictions were analyzed using FunSpec (see sections below).

We also performed a similar test on other unsupervised methods (Scansite and S_{LR}) to compare their relative predictive power for these kinases with respect to ConDens. Unfortunately, the differences in classification power were not statistically significant (see Supplementary Fig. 5), which we believe to be due to the small number of known kinase targets.

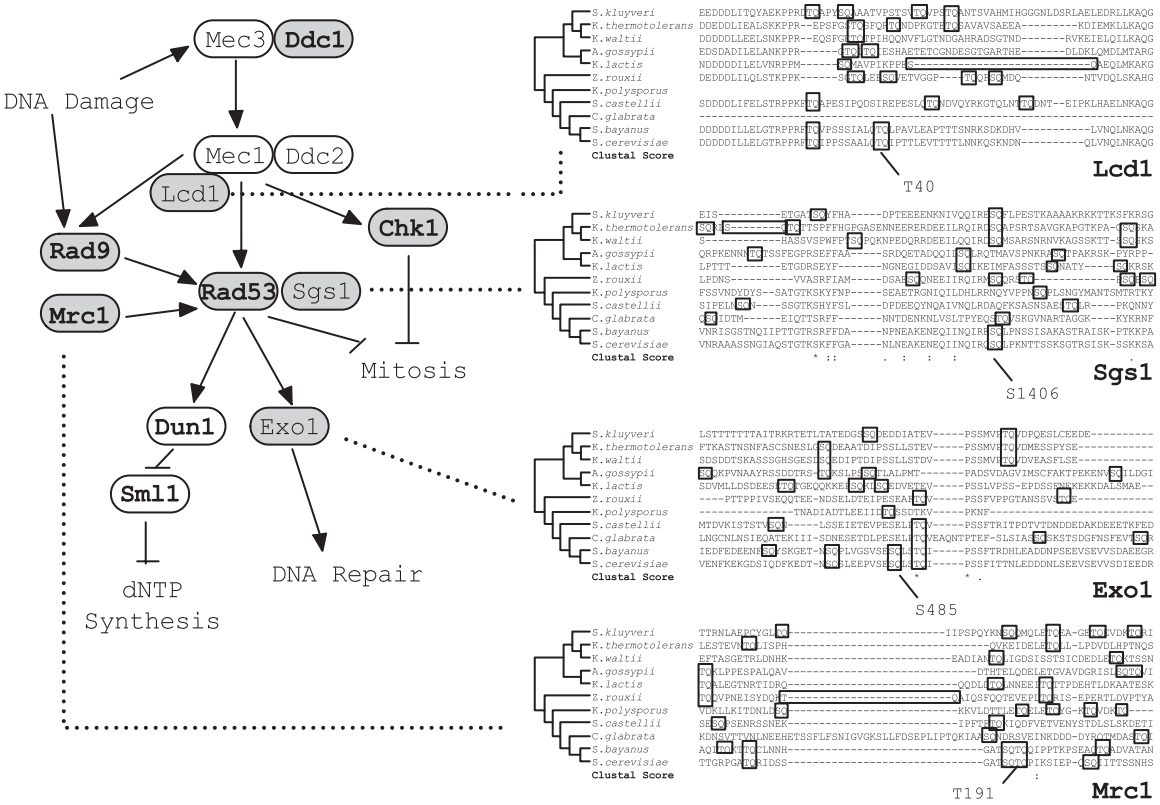

Fig. 3. An illustration of the DNA damage signaling pathway in *S. cerevisiae* (Anderson and Sinclair, 2000; Clarke *et al.*, 2000; D'Amours and Jackson, 2002; Segurado and Tercero, 2009; Zewail *et al.*, 2003). Shaded bubbles indicate proteins predicted by ConDens to be Mec1/Tel1 targets. Bolded bubbles indicate proteins that are among our list of known Mec1/Tel1 targets. Alignments of novel predictions (Lcd1, Sgs1 and Exo1) and top hit (Mrc1) are shown on the right. Matches to the S/T-Q consensus are surrounded by gray boxes.

3.4 Mec1, Tel1, Prk1 and CKII

The Mec1/Tel1 hits were statistically significantly ($P < 0.05$) enriched in functional categories such as cell-cycle checkpoint, DNA damage response, DNA repair and chromosome organization. These functions were all relevant to the roles of the Mec1 and Tel1 kinases as DNA damage sensors (Anderson and Sinclair, 2000; Clarke *et al.*, 2000; D'Amours and Jackson, 2002; Segurado and Tercero, 2009; Zewail *et al.*, 2003).

We further examined the novel predictions by determining whether or not they were involved in the DNA damage signaling pathway (Anderson and Sinclair, 2000; Clarke *et al.*, 2000; D'Amours and Jackson, 2002; Segurado and Tercero, 2009; Zewail *et al.*, 2003) (Fig. 3). Based on this analysis, we identified three predicted substrates (Sgs1, Lcd1 and Exo1) that function in this pathway, but were not found in the databases of known substrates. In particular, Sgs1 was found to have a human ortholog phosphorylated by the human Mec1/Tel1 (ATM/ATR) (Davies *et al.*, 2007; Friedel *et al.*, 2009) and therefore we consider it to be a very promising prediction.

Like the Mec1/Tel1 hits, the Prk1 hits appeared to be functionally associated with their kinase's biological role. Prk1 is a kinase known for regulating actin organization and endocytosis (Zeng and Cai, 1999), and the Prk1 hits were statistically significantly enriched ($P < 0.05$) in proteins related to actin, endocytosis and cell polarity. Among the four novel predictions, Prk1 (the kinase itself) and YAP1802 belonged in at least two of the aforementioned biological processes. YAP1802 is a paralog of a known Prk1 target (YAP1801) and Prk1 was previously reported to autophosphorylate (Huang *et al.*, 2009).

Although CKII is involved in a large number of biological processes, the CKII hits were remarkably enriched in ribosome-related functions, especially for rRNA processing and ribosome biosynthesis ($P < 10^{-14}$). The connection between CKII and the aforementioned processes were supported by a number of literature articles. A study on pre-rRNA processing and ribosome synthesis suggested that CKII was biologically related to the novel predictions Ifh1 and Fhl1 with the former also being an *in vitro* CKII target (Rudra *et al.*, 2007). Another study (Meier, 1996) also suggested a role of CKII in ribosome synthesis through Srp40 phosphorylation. Interestingly, although Srp40 was noted as being involved in 'pre-ribosome assembly' in SGD, it was not actually listed under rRNA processing or ribosome biosynthesis in the FunSpec results.

3.5 PKA and Ipl1

For the remaining kinases (PKA, Ipl1 and Ime2), we were unable to find any statistically significant functional enrichment in their hits. However, our top two predictions for PKA, Tod6 and Dot6 were both previously found to be functionally related to PKA in the ribosome biogenesis pathway. (Deminoff *et al.*, 2006; Lippman and Broach, 2009) (Supplementary Fig. 6). Whereas Dot6 has recently been

confirmed as a direct target of PKA, Tod6 has not. The fact that Tod6 was predicted to be a target of PKA by ConDens strongly suggested that PKA also inhibits the activity of this repressor by direct phosphorylation. Another of our interesting predictions for PKA is Cyr1 (previously known as Cdc35), which is an adenylyl cyclase that regulates PKA (Guarente and Kenyon, 2000) (Supplementary Fig. 6).

Our top Ipl1 prediction (Tid3, formerly known as Ndc80) was previously reported to be an *in vitro* and *in vivo* Ipl1 substrate (Cheeseman *et al.*, 2002), but this protein has not yet been included in the databases of known substrates. The remaining predictions appear to be unrelated to Ipl1 functions. A closer inspection of these results suggests some of them may in fact be detected for other reasons. For example, we found three closely related flippases (Dnf1, Dnf2 and Dnf3) among the Ipl1 predictions. One of these (Dnf1) was reported as an *in vitro* substrate of a flippase kinase known as Fpk1. Remarkably, the same study (Roelants *et al*., 2010) also proposed a consensus for the Fpk1 kinase that greatly overlapped with the Ipl1 consensus we used. As a result, it is possible that the Ipl1 predictions also included substrates of Fpk1 or other related kinases.

This result brings up an important aspect of ConDens' design. Since the method is motif-based, it has no actual conception of what is a kinase. As a result, it can experience difficulty in differentiating substrates of kinases that have substantially similar (or identical) consensus sequences (as we suspected to be the case for the Fpk1 and Ipl1 kinases). Under these circumstances, other information such as subcellular localization and physical interactions (Linding *et al*., 2008) are required for ambiguity resolution.

3.6 Ime2

The Ime2 hits were different to the hits of the other kinases in that they were very few (a total of six) and devoid of known Ime2 targets. We examined the cause of this negative result by inspecting the local alignments of the known Ime2 phosphorylation sites in *S. cerevisiae*. Remarkably, none of the known Ime2 targets showed a strong conservation of R-P-X-S/T motif matches among the fungal orthologs. The weak conservation patterns observed was also not clade-specific and varied from protein to protein.

A likely explanation of this unexpected observation is that the functional regulation of Ime2 had diverged over evolution. Indeed, recent comparative studies on *S. cerevisiae* and other more distant fungal species (Hutchison and Glass, 2010; Irniger, 2011) suggested that Ime2 may have functionally diversified over evolution. If this phenomenon also held for the more closely-related fungal species considered here, then it could explain our inability to detect Ime2 targets through conservation analysis.

3.7 Biologically important T-G motifs in Nup?

Since we were unsatisfied with the small number of novel predictions for Prk1, we ran ConDens on the *S. cerevisiae* proteome again using a 'relaxed' consensus (T-G, as opposed to L/I/V/M-X_4-T-G) to obtain a larger and possibly even more interesting set of predictions. In contrast to what we hoped for, the majority of novel predictions for this relaxed consensus did not appear to be biologically related to Prk1. As a result, we suspect they were not actually targets of the Prk1 kinase.

However, as with our analysis of the Ipl1 results (where they contained targets of another kinase), some of these T-G-based predictions appeared to be biologically important. As an example, we detected an enrichment of conserved T-G motif matches in three nuclear pore proteins (Nup57, Nup100 and Nup116). This enrichment occurs close to the F-G matches that are characteristic of many nuclear pore proteins (Yang, 2011) (Supplementary Fig. 7). Although it may seem unlikely for Prk1 to be associated with the Nup proteins due to vast differences in their biological functions, these T-G's might be functionally important and associated with the F-G's. This observation is also an indication of this method's general applicability. Even though this study placed a strong emphasis on phosphorylation motifs, the ConDens algorithm was based on the principle that functionally-important motif matches are conserved over evolution and this should be applicable to any type of short linear motifs.

4 CONCLUSION

In all, we offer a new method to predict kinase substrates based on evolutionary conservation of phosphorylation site recognition motifs. The requirement for accurate phosphorylation site alignment was circumvented by using the local retention of motif density as the measure of conservation. Since the new method is based on a statistical model of molecular evolution, a labeled training set is not required. Furthermore, we demonstrated the method's utility in mining substrates for kinases with some information on substrate specificity but few characterized *in vivo* substrates. ConDens should be applicable to a wide variety of model organisms due to available databases such as Ensembl (Flicek *et al*., 2010) and INPARANOID (Ostlund *et al*., 2010) that provide homologous sets of proteins.

ACKNOWLEDGEMENTS

We thank Dr David Guttman, Dr Sergio Peisajovich, Dr Anthony Bonner, Dr Peter McCourt, Dr Philip Kim, Dr Christian Landry, Dr Annabelle Haudry, Dr Amin Zia and Louis-Francois Handfield for useful advice.

Funding: This work was supported by National Science and Engineering Research Council through grants to A.M.M. and scholarship to A.N.N.B. Infrastructure for this research was supported by grants from the Canadian Foundation for Innovation grant to A.M.M.

REFERENCES

Anderson,R.M. and Sinclair,D.A. (2000) Yeast RecQ helicases: clues to DNA repair, genome stability and aging. In, *Madame Curie Bioscience Database*. Landes Bioscience and Springer Science+Business Media.

Ba,A.N.N. and Moses,A.M. (2010) Evolution of characterized phosphorylation sites in budding yeast. *Mol. Biol. Evol.*, **27**, 2027–2037.

Bauer,D.C. *et al*. (2011) Sorting the nuclear proteome. *Bioinformatics*, **27**, i7–i14.

Berg,J. (2007) *Biochemistry*. W.H. Freeman, New York.

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

Budovskaya,Y.V. *et al*. (2005) An evolutionary proteomics approach identifies substrates of the cAMP-dependent protein kinase. *Proc. Natl Acad. Sci. USA*, **102**, 13933–13938.

Chang,E.J. *et al*. (2007) Prediction of cyclin-dependent kinase phosphorylation substrates. *PLoS ONE*, **2**, e656.

Cheeseman,I.M. *et al*. (2002) Phospho-regulation of kinetochore-microtubule attachments by the Aurora kinase Ipl1p. *Cell*, **111**, 163–172.

Chen,S.-H. and Zhou,H. (2009) Reconstitution of Rad53 activation by Mec1 through adaptor protein Mrc1. *J. Biol. Chem.*, **284**, 18593–18604.

Cherry,J.M. *et al*. (1998) SGD: Saccharomyces Genome Database. *Nucleic Acids Res*., **26**, 73–80..

Clarke,D.J. *et al*. (2000) DNA damage-independent checkpoints from yeast to man. In *Madame Curie Bioscience Database*. Landes Bioscience and Springer Science+Business Media.

Cohen,P. (1982) The role of protein phosphorylation in neural and hormonal control of cellular activity. *Nature*, **296**, 613–620.

Collins,M.O. *et al*. (2007) Analysis of protein phosphorylation on a proteome-scale. *Proteomics*, **7**, 2751–2768.

Conde e Silva,N. *et al*. (2009) KlAft, the *Kluyveromyces lactis* ortholog of Aft1 and Aft2, mediates activation of iron-responsive transcription through the PuCACCC Aft-type sequence. *Genetics*, **183**, 93–106.

Davies,S.L. *et al*. (2007) Role for BLM in replication-fork restart and suppression of origin firing after replicative stress. *Nat. Struct. Mol. Biol.*, **14**, 677–679.

Deminoff,S.J. *et al*. (2006) Using substrate-binding variants of the cAMP-dependent protein kinase to identify novel targets and a kinase domain important for substrate interactions in Saccharomyces cerevisiae. *Genetics*, **173**, 1909–1917.

Deshaies,R.J. *et al*. (1995) Ubiquitination of the G1 cyclin Cln2p by a Cdc34p-dependent pathway. *EMBO J.*, **14**, 303–312.

D'Amours,D. and Jackson,S.P. (2002) The mre11 complex: at the crossroads of dna repair and checkpoint signalling. *Nat. Rev. Mol. Cell Biol.*, **3**, 317–327.

Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

Fawcett,T. (2004) ROC graphs?: notes and practical considerations for researchers. *HP Laboratories*, **31**, 1–38.

Flicek,P. *et al*. (2010) Ensembl 2011. *Nucleic Acids Res.*, **39**, D800–D806.

Friedel,A.M. *et al*. (2009) ATR/Mec1: coordinating fork stability and repair. *Curr. Opin. Cell Biol.*, **21**, 237–244.

Friedman,D.B. *et al*. (1996) The 110-kD spindle pole body component of Saccharomyces cerevisiae is a phosphoprotein that is modified in a cell cycle-dependent manner. *J. Cell Biol.*, **132**, 903–914.

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

Gribskov,M. and Robinson,N.L. (1996) Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. *Comput. Chem.*, **20**, 25–33.

Guarente,L. and Kenyon,C. (2000) Genetic pathways that regulate ageing in model organisms. *Nature*, **408**, 255–262.

Harvey,S.L. *et al.* (2005) Cdk1-dependent regulation of the mitotic inhibitor Wee1. *Cell*, **122**, 407–420.

Holt,L.J. *et al*. (2007) Evolution of Ime2 phosphorylation sites on Cdk1 substrates provides a mechanism to limit the effects of the phosphatase Cdc14 in meiosis. *Mol. Cell*, **25**, 689–702.

Holt,L.J. *et al*. (2009) Global analysis of Cdk1 substrate phosphorylation sites provides insights into evolution. *Science*, **325**, 1682–1686.

Huang,B. *et al*. (2003) Identification of novel recognition motifs and regulatory targets for the yeast actin-regulating kinase Prk1p. *Mol. Biol. Cell*, **14**, 4871–4884.

Huang,B. *et al*. (2009) Negative regulation of the actin-regulating kinase Prk1p by patch localization-induced autophosphorylation. *Traffic*, **10**, 35–41.

Hutchison,E.A. and Glass,N.L. (2010) Meiotic regulators Ndt80 and ime2 have different roles in Saccharomyces and Neurospora. *Genetics*, **185**, 1271–1282.

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

Irniger,S. (2011) The Ime2 protein kinase family in fungi: more duties than just meiosis. *Mol. Microbiol.*, **80**, 1–13.

Jukes,T.H. and Cantor,C.R. (1969) Evolution of protein molecules. *New York: Academic Press*, 21–132.

Kemp,B.E. and Pearson,R.B. (1990) Protein kinase recognition sequence motifs. *Trends Biochem. Sci.*, **15**, 342–346.

Lam,H.Y.K. *et al*. (2010) MOTIPS: automated motif analysis for predicting targets of modular protein domains. *BMC Bioinformatics*, **11**, 243.

Lanker,S. *et al*. (1996) Rapid degradation of the G1 cyclin Cln2 induced by CDK-dependent phosphorylation. *Science*, **271**, 1597–1601.

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

Linding,R. *et al*. (2003) Protein disorder prediction: implications for structural proteomics. *Structure*, **11**, 1453–1459.

Linding,R. *et al*. (2008) NetworKIN: a resource for exploring cellular phosphorylation networks. *Nucleic Acids Res.*, **36**, D695–D699.

Lippman,S.I. and Broach,J.R. (2009) Protein kinase A and TORC1 activate genes for ribosomal biogenesis by inactivating repressors encoded by Dot6 and its homolog Tod6. *Proc. Natl Acad. Sci. USA*, **106**, 19928–19933.

Meggio,F. and Pinna,L.A. (2003) One-thousand-and-one substrates of protein kinase CK2? *FASEB J.*, **17**, 349–368.

Meier,U.T. (1996) Comparison of the rat nucleolar protein nopp140 with its yeast homolog SRP40. Differential phosphorylation in vertebrates and yeast. *J. Biol. Chem.*, **271**, 19376–19384.

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

Moses,A.M. *et al*. (2007b) Regulatory evolution in proteins by turnover and lineage-specific changes of cyclin-dependent kinase consensus sites. *Proc. Natl Acad. Sci. USA*, **104**, 17713–17718.

Muffato,M. *et al*. (2010) Genomicus: a database and a browser to study gene synteny in modern and ancestral genomes. *Bioinformatics*, **26**, 1119–1121.

Niefind,K. *et al*. (2007) Evolved to be active: sulfate ions define substrate recognition sites of CK2alpha and emphasise its exceptional role within the CMGC family of eukaryotic protein kinases. *J. Mol. Biol.*, **370**, 427–438.

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.

Ostlund,G. *et al*. (2010) InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res.*, **38**, D196–D203.

Robinson,M.D. *et al*. (2002) FunSpec: a web-based cluster interpreter for yeast. *BMC Bioinformatics*, **3**, 35.

Roelants,F.M. *et al*. (2010) A protein kinase network regulates the function of aminophospholipid flippases. *Proc. Natl Acad. Sci. USA*, **107**, 34–39.

Rudra,D. *et al*. (2007) Potential interface between ribosomal protein production and pre-rRNA processing. *Mol. Cell Biol.*, **27**, 4815–4824.

Schwartz,M.F. *et al*. (2002) Rad9 phosphorylation sites couple Rad53 to the Saccharomyces cerevisiae DNA damage checkpoint. *Mol. Cell*, **9**, 1055–1065.

Segurado,M. and Tercero,J.A. (2009) The S-phase checkpoint: targeting the replication fork. *Biol. Cell*, **101**, 617–627.

Sharifpoor,S. *et al*. (2011) A quantitative literature-curated gold standard for kinase-substrate pairs. *Genome Biol.*, **12**, R39.

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.

Townsend,R.R. *et al*. (1996) Identification of protein kinase A phosphorylation sites on NBD1 and R domains of CFTR using electrospray mass spectrometry with selective phosphate ion monitoring. *Protein Sci.*, **5**, 1865–1873.

Ubersax,J.A. *et al*. (2003) Targets of the cyclin-dependent kinase Cdk1. *Nature*, **425**, 859–864.

Wang,Y.H. (1993) On the number of successes in independent trials. *Statistica Sinica*, **3**, 295–312.

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*. (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.

Yang,W. (2011) "Natively unfolded" nucleoporins in nucleocytoplasmic transport: Clustered or evenly distributed? *Nucleus*, **2**, 10–16.

Zeng,G. and Cai,M. (1999) Regulation of the actin cytoskeleton organization in yeast by a novel serine/threonine kinase Prk1p. *J. Cell Biol.*, **144**, 71–82.

Zewail,A. *et al*. (2003) Novel functions of the phosphatidylinositol metabolic pathway discovered by a chemical genomics screen with wortmannin. *Proc. Natl Acad. Sci. USA*, **100**, 3345–3350.

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