BIOINFORMATICS ORIGINAL PAPER

Vol. 31 no. 3 2015, pages 382–389
doi:10.1093/bioinformatics/btu663

Systems biology

Advance Access publication October 9, 2014

PhosphoPICK: modelling cellular context to map kinase-substrate phosphorylation events

Ralph Patrick[1,*], Kim-Anh Lê Cao[2,3,4], Bostjan Kobe[1,4,5] and Mikael Bodén[1,4,*]

[1]School of Chemistry and Molecular Biosciences and [2]Queensland Facility for Advanced Bioinformatics, The University of Queensland, St Lucia 4072, [3]Translational Research Institute, The University of Queensland Diamantina Institute, Brisbane, St Lucia 4102, [4]Institute for Molecular Bioscience and [5]Australian Infectious Diseases Research Centre, The University of Queensland, St Lucia, 4072, Australia

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: The determinants of kinase-substrate phosphorylation can be found both in the substrate sequence and the surrounding cellular context. Cell cycle progression, interactions with mediating proteins and even prior phosphorylation events are necessary for kinases to maintain substrate specificity. While much work has focussed on the use of sequence-based methods to predict phosphorylation sites, there has been very little work invested into the application of systems biology to understand phosphorylation. Lack of specificity in many kinase substrate binding motifs means that sequence methods for predicting kinase binding sites are susceptible to high false-positive rates.

Results: We present here a model that takes into account protein–protein interaction information, and protein abundance data across the cell cycle to predict kinase substrates for 59 human kinases that are representative of important biological pathways. The model shows high accuracy for substrate prediction (with an average AUC of 0.86) across the 59 kinases tested. When using the model to complement sequence-based kinase-specific phosphorylation site prediction, we found that the additional information increased prediction performance for most comparisons made, particularly on kinases from the CMGC family. We then used our model to identify functional overlaps between predicted CDK2 substrates and targets from the E2F family of transcription factors. Our results demonstrate that a model harnessing context data can account for the short-falls in sequence information and provide a robust description of the cellular events that regulate protein phosphorylation.

Availability and implementation: The method is freely available online as a web server at the website http://bioinf.scmb.uq.edu.au/phosphopick.

Contact: m.boden@uq.edu.au

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

Received on February 7, 2014; revised on August 25, 2014; accepted on October 6, 2014

1 INTRODUCTION

Regulation of cellular processes occurs on multiple levels, with epigenetic modifiers and transcription factors (TFs) controlling gene expression, while various post-translational modifications regulate many protein functions (Blom *et al.*, 2004; Choudhary *et al.*, 2009; Hwang *et al.*, 2010). The most ubiquitous of post-translational modifications is phosphorylation, with at least 70% of human proteins estimated to be phosphorylation substrates (Olsen *et al.*, 2010). Phosphorylation is likely a significant factor in regulating the function of complex organisms, with a significant increase in the numbers of phosphorylation sites in eukaryotic compared with prokaryotic proteins (Gnad *et al.*, 2011). Phosphorylation is known to have numerous regulatory roles across the cell cycle, and specific kinases have been implicated in the regulation of G1 phase (Harbour *et al.*, 1999), the G1/S phase transition (Sherr and Roberts, 1999) and DNA replication and damage repair (Coverley *et al.*, 2002). Phosphorylation is particularly ubiquitous during mitosis where many complex operations such as spindle formation, centrosome maturation/separation and chromosome attachment to the spindle are controlled by kinases (Johnson, 2011).

While advanced phosphoproteomic technologies have succeeded in identifying thousands of phosphorylation sites across multiple proteomes (Huttlin *et al.*, 2010; Olsen *et al.*, 2010), there has been an ever widening gap between known phosphorylation sites and the kinases responsible for those sites (Diella *et al.*, 2004). Currently just over 10% of the phosphorylation sites recorded in the eukaryotic phosphorylation site database Phospho.ELM is annotated with a kinase. There have been examples of *in vitro* studies identifying kinase-substrate binding events (Mok *et al.*, 2010), and while these studies offer interesting insights into the consensus motifs of kinase binding sites, it is unknown whether the binding events observed *in vitro* would occur *in vivo*. Determining kinase-substrates *in vivo* is non-trivial however, though there have been promising results from combining *in vitro* kinase detection assays with *in vivo* phosphoproteomics (Xue *et al.*, 2012). As a result of the inherent difficulty in determining *in vivo* kinase substrates, there has been a great interest in developing computational tools to predict kinase-specific phosphorylation sites, with over 40 phosphorylation site-prediction methods published (Trost and Kusalik, 2011). While some methods aim only to predict phosphorylation sites (Blom *et al.*, 1999; Ingrell *et al.*, 2007), the majority predict kinase-specific phosphorylation sites.

Historically, phosphorylation site predictors have operated primarily on protein amino acid sequences, relying on the information contained in the sequence region surrounding phosphorylation sites. It has long been recognized that short sequence

*To whom correspondence should be addressed.

motifs alone are insufficient for achieving respectable accuracy in predicting kinase-specific phosphorylation sites. As a result, prediction methods have often complemented sequence information with other types of data such as knowledge of 3D structure (Saunders *et al*., 2008; Durek *et al*., 2009), sequence disorder (Gao *et al*., 2010) and kinase family similarity (Xue *et al*., 2011). While such additional data typically improve prediction performance to an extent, they do not reflect the wider cellular regulatory mechanisms that cause kinases to target their correct substrates—a protein with an appropriate kinase binding site will not necessarily come into contact with that kinase (Zhu *et al*., 2005).

The phosphorylation of a target substrate by a kinase is not determined solely by its binding affinity, but by various context factors that determine how a kinase comes into contact with its substrates (Kobe *et al*., 2005). This is recognized by the NetworKIN predictor (Horn *et al*., 2014), which combines sequence-based scores with a score generated on the basis of a STRING network (Szklarczyk *et al*., 2011). Context factors can include cellular location (Scott and Pawson, 2009), mediating and activating proteins such as scaffold proteins (Good *et al*., 2011), cyclins (Lim and Kaldis, 2013) and cell cycle-specific expression of kinases and their substrates. Protein–protein interaction (PPI) data can certainly be used to represent such context factors; though while there is vast amounts of PPI data currently available in databases such as BioGRID (Chatr-aryamontri *et al*., 2013) and STRING, incomplete coverage and variable certainty means that the integration of context features into a model is non-trivial.

In this study, we explore a probabilistic model to accommodate missing values, seamless combination of protein interactions and cell-cycle expression, and to provide flexible options for querying potential kinase substrates. The model we present here, named PhosphoPICK (Phosphorylation in a Protein Interaction Context for Kinases), integrates known kinase-substrate relationships, PPI, and cell-cycle data to predict kinase substrates for 59 human kinases. PhosphoPICK shows high prediction accuracy, with a mean AUC of 0.86 across the 59 kinases. We then demonstrate how our method can boost the prediction accuracy of kinase-specific phosphorylation site prediction by combining PhosphoPICK predictions with the phosphorylation site predictions from three previously published methods. We find that PhosphoPICK improves kinase-specific phosphorylation site prediction for most comparisons made, though greater performance increases were noticed on CMGC kinases—in particular cyclin-dependant kinases (CDKs), where we observed substantial performance gains as measured by AUC50. We show that proteins predicted to be CDK2 substrates by PhosphoPICK have GO terms consistent with known CDK2 substrates, and investigate the functional overlap between known and predicted CDK2 substrates, and the targets of specific E2F TFs using ChIP-Seq data.

2 METHODS

2.1 Bayesian network model

We used a Bayesian network (BN) to design our model. BNs differ from machine learning tools that have previously been used for phosphorylation site prediction in several important ways. BNs are transparent, allowing for an understanding of how the variables in the model influence the final outcome (Patrick *et al*., 2012). Furthermore, the probabilistic nature of a BN means that even in the absence of missing data, the model can still infer the most likely value of the unknown variables on the basis of the known data (Bauer *et al*., 2011; Mehdi *et al*., 2011). We represent observations about protein interactions, kinase-specific phosphorylation events and cell-cycle profiles as Boolean variables in a BN model (Fig. 1). The model represents observations about a phosphorylation substrate—the kinases that bind to it, protein interactions, and whether it is upregulated during the cell-cycle phases. The kinase nodes are linked to PPI events that are believed to be relevant for the kinase to phosphorylate substrates. A latent variable is used to capture information from the cell-cycle data, and the kinase nodes are then conditioned on this latent variable.

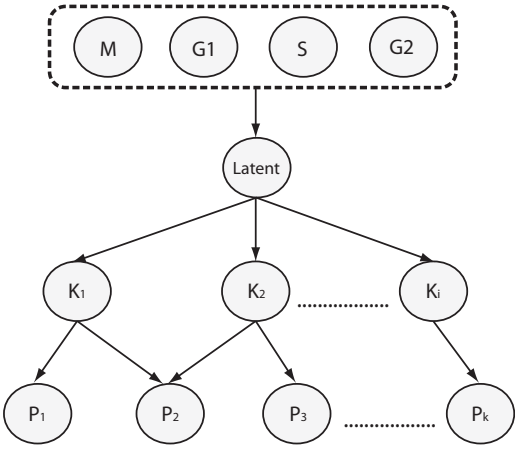

Fig. 1. The PhosphoPICK BN model. Each of the kinase (*K*) nodes, representing a phosphorylation event by that kinase, are conditioned on a latent variable incorporating protein abundance across four stages of the cell cycle: Mitosis (M), G1, S and G2. The 'leaf' nodes represent protein interaction (*P*) events between the proteins represented in the nodes and a potential substrate. These nodes are conditioned on relevant kinase-specific phosphorylation events

2.2 Data resources

2.2.1 Known kinase-substrate relationships We obtained kinase substrates from Phospho.ELM and HPRD, after converting HPRD IDs to Uniprot identifiers. To identify protein interactions between kinases and their substrates we selected kinases for which we found greater than 10 substrates. In total, we use 59 human kinases along with a total of 1210 substrates. Table 1 shows the numbers of substrates that were identified for each of the 59 kinases. The 1210 substrates contained 2964 unique phosphorylation sites that were annotated with at least one kinase.

2.2.2 PPI and association data To identify and model interaction networks of kinases and their substrates, we used PPI data. In cases where physical interaction data are unavailable, associations inferred on the basis of other sources such as gene coexpression or literature mining may be informative, and such information is available in the STRING database. PPI information was taken from the Biological General Repository for Interaction Datasets (BioGRID) (Chatr-aryamontri *et al*., 2013) by selecting entries that were of type 'direct interaction' or 'physical association'. As PPIs are represented in binary

format, this information was incorporated into the model as a Boolean value. The STRING database scores an association probability between two proteins, with a score of 0.4 defined as medium confidence. To convert this probability into a Boolean value we defined cutoff probabilities, such that given some cutoff θ, any association with a probability $\geq\theta$ was classified as true, and any association with a probability $<\theta$ was classified as false. We tested three cutoff probabilities, starting at the medium confidence level of 0.4 and increasing in increments of 0.2. We found that a cutoff probability of 0.6 provided the best overall performance (Supplementary Table S1), and is the cutoff used in this work.

To identify relevant connections between kinases and PPI events, the following steps were taken. Substrates were first grouped according to their kinase (one substrate could be assigned to multiple kinases). BioGRID was then searched for proteins that interacted both with a substrate and with its kinase—these proteins were added to a pool of potential protein interaction connections. For each kinase, the proteins in the pool were ranked in descending order according to the number of interactions that were observed with the kinase's substrates. An observation is defined as a substrate–protein interaction occurring in BioGRID and/or the STRING database. A count c was defined, so that for each kinase only the top c protein interactions were used to form connections. To ensure that there would be enough observations of substrate-protein interactions for setting model parameters, a lower bound of 10 was set such that for a given kinase, at least 10 substrate–protein observations were required for the protein to be considered as a connection to that kinase. We tested three different upper bounds of c: 25, 40 and 50 to determine the effect of varying sized interaction networks on prediction performance.

2.2.3 Protein cell-cycle data. In order to model the availability of substrates during the cell cycle, we used data obtained from the experiments by Olsen *et al.* (2010), who measured the abundance of proteins at six stages throughout the cell cycle—M phase, G1 phase, the transition between G1 and S phase (G1/S), early S phase, late S phase and G2 phase. An asynchronous population of cells was also measured, and the signal used to $\log_2$ normalize the measurements from the cells arrested during the six stages. A protein with a value of 0 during a stage of the cell cycle has an abundance equivalent to the asynchronous population, whereas a negative value indicates downregulation and a positive value indicates upregulation. To avoid fitting the model too strongly to data generated from a single cell type, we represented proteins' cell-cycle profiles in a simple binary format across four stages—M, G1, S and G2. We collapsed the G1 and G1/S stages in to the single variable 'G1' and the early S and late S stages into the variable 'S'. If a protein has a value greater than 0 that stage is labelled as true; otherwise it is labelled as false. The G1 and S variables were set to true if at least one of their respective collapsed stages had a value greater than 0.

2.3 Model parameters and training

The variables in the network were represented with two kinds of probability tables. A conditional probability table (CPT) represents all possible values that a variable X can take given the set of parents, $pa(X)$, it is conditioned on. Parameters are set during training by calculating the frequency of occurrence of all possible configurations of $pa(X)$. If X does not have parents, the CPT simply represents the observed frequency from training data of X being true.

For situations where a variable is conditioned on greater than six parents, we used a variation of the Noisy-OR approximation (Oniśko *et al.*, 2001). In order to set the parameters of the Noisy-OR table during training, each row (representing a parent variable) in the table was calculated as follows: each training sample where the parent is observed as being true was identified. A weighted frequency for each parent pa was calculated such that

$$\text{freq}(pa) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{t}{(t+f)pconf_i}\right),$$

where n is the number of configurations of parent variables where pa is observed to be true, $pconf_i$ is the number of parents set to true in configuration i, t is the count of the variable the Noisy-OR node is representing being true during the ith configuration of parents and f is the count of it being false.

For the latent variable, and variables that are conditioned on it, parameters are calculated using the expectation-maximization algorithm on a training set (Do and Batzoglou, 2008).

2.4 Evaluation and definition of negative test sets

A common problem to phosphorylation site prediction is that of defining a negative test set (Trost and Kusalik, 2011). However, as our model is not trained using sequence data, we were able to use a sequence scoring method to define negative test sets for each of the 59 kinases in the model. To score protein sequences for kinase binding sites, we used the Predikin web server (Ellis and Kobe, 2011) to obtain position weight matrices (PWMs) for 53 of the kinases in the model. For the remaining six we constructed PWMs using phosphorylation sites from curated data (Supplementary Methods S1.1). For a given kinase, we scored each substrate in the training data set by obtaining the highest scoring potential phosphorylation site. We then ranked the substrates based on the highest scoring site from lowest to highest, and assigned an equal number of positive and negative substrates for that kinase. As very low scores indicate a protein that the kinase cannot phosphorylate, this gives us a high-confidence negative test set for each of the kinases in the model.

We evaluated the model for each kinase for its ability to correctly predict known substrates compared with the negative set. To score the probability of a kinase phosphorylating a query protein, all nodes in the network were set according to the relevant data for the query protein except for the kinase that we were inferring. Model performance was evaluated using receiver operating characteristic (ROC) analysis by calculating the area under the ROC curve (AUC) (Baldi *et al.*, 2000). We used 15-fold cross validation, and performed the cross validation 10 times with different data set splits. To avoid the possibility of the model gaining information about the test data during training, we ensured that each protein interaction variable was only connected to a kinase if, within the training fold, there were 10 (our previously defined lower bound) or more kinase substrates interacting with that protein. The data sets used to train and test the model are available in the Supplementary material.

3 RESULTS

3.1 Model performance for predicting kinase substrates

We generated five BN models by grouping kinases according to their family similarities (Manning *et al.*, 2002): CMGC, AGC, TK, CAMK and a combined model that incorporated kinases from the CK1, STE, atypical and other families. We tested the ability of the model to classify kinase substrates with varying numbers of protein interaction connections, and under three conditions. To gauge the level of influence that substrate abundance during the cell cycle has on prediction performance, we evaluated a version of the model excluding the cell-cycle variables (PPI only model), and compared the performance to the full model. When making inferences about a kinase-substrate phosphorylation

event, the model relies on the knowledge of other potential kinases phosphorylating that substrate. However, for the majority of proteins there is little, if any, experimental information on any known kinase-specific phosphorylation events. Therefore, to determine whether the model could be reliably extended to the wider proteome, we tested model performance when setting non-query kinase nodes to false on the basis of their sequence binding motifs (Supplementary Methods S1.2).

The AUC results (Table 1) for 10 cross-validation runs evaluated on all 59 kinases in the model for the three different conditions demonstrate that the prediction accuracy of the full model is quite high, with most kinases having median AUCs surpassing 0.8. The average AUC over all of the kinases is 0.86. The generally low-standard deviation indicates that these results are consistent regardless of the breakup of training/test data that is presented to the model. We tested three different values for maximum number of protein interactions that could be connected to a kinase variable (25, 40 and 50), but found that increasing the number of protein interaction events connected to the kinase variables had very little effect on the performance of the model (Supplementary Table S2), indicating the model's ability to make classifications based on a relatively small number of connections to the individual kinase nodes.

When comparing the performance of the PPI only model to the full model, we found that on average the inclusion of protein abundance data collected across the cell cycle offered modest improvements to prediction performance. For some kinases there was a greater performance improvement—for example a 15% increase for PRKC kinases, a nearly 10% increase for the tyrosine kinase CSK—but for many other kinases the inclusion of cell-cycle data seemed to have little effect. This demonstrates that while the PPI data provide the main source of information for the model, the use of cell-cycle data can offer improved prediction performance for some kinases. This performance increase occurs despite the fact that we only have cell-cycle data for less than half of the substrates in our set: the model infers the cell-cycle profiles for the remaining proteins. This indicates that the model, when trained on cell-cycle data, can still be applied to query proteins that have no associated cell-cycle data.

Table 1 also shows a comparison between setting the kinase nodes with database ("Full model") versus sequence information ("Seq. approx."). We found that for many kinases, using sequence to set the kinase nodes actually resulted in an increase in performance. The median AUC for the CMGC kinases went from 0.87 using database information to 0.91 using sequence, and the median AUC for the tyrosine kinases increased from 0.88 to 0.94.

The possibility was raised that as the kinase-substrate data from HPRD and Phospho.ELM is sourced from the literature, and the STRING database also includes text mining from the literature, a system of circular logic could be inflating the performance values. To determine whether such an effect was occurring, we re-ran our simulations for each kinase with the text mining information for that kinase removed (Supplementary Methods S1.3). We found that while for some kinases this information appeared to have a large impact on prediction capability, it was not the case for the majority of kinases (Supplementary Table S3).

Table 1. Evaluation of model performance with median AUC on all kinases in the model, as tested under three different conditions: interactions only (int. only), full model and kinase variables approximated using sequence data (seq. approx.)

Kinase	Substrates	Int. only	Full model	Seq. approx.
CMGC				
CDK	247	0.88 ± 0.011	0.87 ± 0.015	**0.91 ± 0.01**
GSK3B	58	0.81 ± 0.01	0.82 ± 0.009	**0.88 ± 0.005**
MAPK	136	0.84 ± 0.016	0.88 ± 0.016	**0.92 ± 0.015**
AGC				
AKT1	79	0.89 ± 0.007	0.89 ± 0.004	**0.91 ± 0.001**
GRK2	14	0.86 ± 0.022	0.87 ± 0.035	**0.87 ± 0.01**
PDPK1	23	**0.95 ± 0.011**	0.94 ± 0.011	0.91 ± 0.021
PRKACA	154	0.94 ± 0.003	0.93 ± 0.006	**0.96 ± 0.002**
PRKC	394	0.73 ± 0.005	**0.86 ± 0.006**	0.82 ± 0.006
PRKG1	26	0.86 ± 0.014	0.86 ± 0.009	**0.90 ± 0.01**
ROCK1	21	**0.80 ± 0.006**	0.80 ± 0.01	0.79 ± 0.011
RSK1	27	0.91 ± 0.027	0.89 ± 0.019	**0.93 ± 0.008**
RSK2	22	0.67 ± 0.012	**0.77 ± 0.027**	0.71 ± 0.036
TK				
ABL1	40	0.89 ± 0.017	0.88 ± 0.014	**0.97 ± 0.006**
BTK	14	0.79 ± 0.056	**0.83 ± 0.091**	0.69 ± 0.11
CSK	18	0.87 ± 0.012	**0.95 ± 0.034**	0.91 ± 0.036
EGFR	38	0.84 ± 0.01	0.84 ± 0.016	**0.95 ± 0.001**
FYN	38	0.83 ± 0.033	0.85 ± 0.049	**0.96 ± 0.01**
HCK	16	0.94 ± 0.01	**0.96 ± 0.032**	0.95 ± 0.046
INSR	23	0.92 ± 0.011	**0.96 ± 0.012**	0.93 ± 0.002
JAK1	11	0.65 ± 0.12	0.69 ± 0.078	**0.76 ± 0.098**
JAK2	17	0.95 ± 0.013	0.95 ± 0.026	**0.97 ± 0.036**
LCK	23	0.93 ± 0.004	0.94 ± 0.004	**0.96 ± 0.011**
LYN	39	0.76 ± 0.028	0.77 ± 0.028	**0.87 ± 0.02**
RET	16	0.60 ± 0.11	**0.82 ± 0.07**	0.69 ± 0.096
SRC	125	0.85 ± 0.01	0.87 ± 0.009	**0.89 ± 0.003**
SYK	27	**1.00 ± 0.0**	**1.00 ± 0.0**	0.98 ± 0.004
ZAP70	12	**0.95 ± 0.064**	0.92 ± 0.019	0.94 ± 0.059
CAMK				
CAMK1A	12	0.22 ± 0.074	**0.75 ± 0.075**	0.56 ± 0.083
CAMK2A	41	0.84 ± 0.014	**0.89 ± 0.022**	0.81 ± 0.021
CAMK2G	26	0.97 ± 0.006	0.96 ± 0.001	**0.98 ± 0.012**
CHK1	11	0.88 ± 0.045	0.52 ± 0.05	**0.91 ± 0.038**
LKB1	17	0.86 ± 0.023	**0.90 ± 0.054**	0.88 ± 0.03
MAPKAPK2	21	0.91 ± 0.008	0.93 ± 0.025	**0.93 ± 0.01**
Combined				
ATM	46	**0.99 ± 0.001**	**0.99 ± 0.001**	0.98 ± 0.003
ATR	14	**0.99 ± 0.029**	0.98 ± 0.018	0.92 ± 0.047
AURKB	16	0.94 ± 0.004	**0.95 ± 0.017**	0.91 ± 0.03
CSNK1A1	25	0.88 ± 0.014	**0.89 ± 0.011**	0.86 ± 0.017
CSNK1D	13	0.64 ± 0.13	**0.69 ± 0.074**	0.63 ± 0.147
CSNK2A1	135	0.87 ± 0.002	**0.9 ± 0.004**	0.89 ± 0.005
CSNK2A2	67	**0.96 ± 0.001**	**0.96 ± 0.005**	0.95 ± 0.004
CSNK2B	20	0.86 ± 0.005	**0.88 ± 0.017**	0.87 ± 0.012
PAK1	27	**0.59 ± 0.03**	0.58 ± 0.043	0.49 ± 0.029
PAK2	12	0.21 ± 0.13	**0.53 ± 0.12**	0.40 ± 0.115
PLK1	23	**0.92 ± 0.005**	**0.92 ± 0.006**	**0.92 ± 0.006**
PRKDC	11	0.74 ± 0.064	0.76 ± 0.053	**0.81 ± 0.068**

Also shown is the number of substrates (positive test set) that were identified for each kinase. Results are shown for 15-fold cross validation across 10 data set splits. The best result for each kinase is highlighted in bold. CDK, MAPK and PRKC represent a family of kinases—the average values of their family members are included in the table. Kinases are listed according to the family-specific BN that they were incorporated into, where the 'combined' model contained kinases from the CK1, STE, atypical and other families of kinases

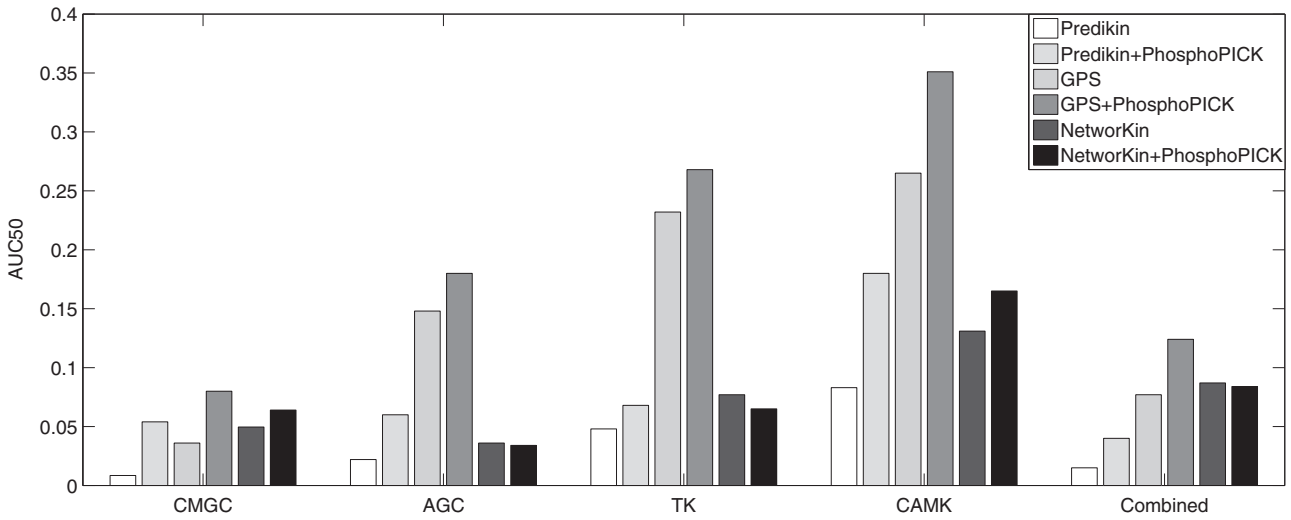

Fig. 2. Comparison between predicting kinase-specific phosphorylation sites with three alternative scoring methods, and when the methods are informed by PhosphoPICK. AUC50 was calculated for each kinase as a measure of the predictive performance at low false-positive levels. Shown here are the average values for each individual BN. The comparison is made by normalizing the scores of the alternative methods to a value between 0 and 1, then summing this value with the PhosphoPICK prediction for a substrate

3.2 Improving sequence-based prediction of phosphorylation sites

For the remainder of this article, the results were generated using the full model, with a PPI count of 25, and setting non-query kinase nodes on the basis of their sequence binding motifs. We tested the ability of PhosphoPICK to complement two phosphorylation site predictors that operate on sequence data: Predikin (Ellis and Kobe, 2011) and GPS (Xue *et al*., 2011). We also tested NetworKIN (Horn *et al*., 2014), which combines sequence scores with a context score generated on the basis of STRING associations. Comparisons were made by normalizing the values of the methods being tested against, and summing the PhosphoPICK prediction (Supplementary Methods S1.4). Figure 2 shows the AUC50 (the AUC obtained when calculating ROC up to the fiftieth false positive) comparison for Predikin, GPS and NetworKIN across the five BNs, where the highest false-positive rate for serine/threonine kinases was 0.0005, and the highest for tyrosine kinases was 0.002. Individual results for each kinase are shown in Supplementary Tables S4–S6. The results show that across all kinase families, there is an average increase in performance when the Predikin and GPS scores are complemented with PhosphoPICK predictions, with largest performance increases observed with kinases from the CMGC family. We found that the performance of GPS improved by 2-fold for predicting CMGC sites when combined with PhosphoPICK, and that the performance of Predikin was improved by over 6-fold. The smallest performance increases were observed with tyrosine kinases, where we found a 15% performance increase for GPS and a 40% increase for Predikin.

We found that in most cases, PhosphoPICK was unable to improve the performance of the NetworKIN predictions. As Figure 2 shows, the differences in AUC50 between classifying phosphorylation sites with NetworKIN alone and NetworKIN + PhosphoPICK are minor. However, as the NetworKIN score is already a combination of a STRING and sequence-based score, it is possible that a simple summing of scores cannot yield further performance increases.

3.3 Understanding E2F and CDK2 regulation

To evaluate the ability of the predictions made by PhosphoPICK to provide biological insights, we used CDK2 as a case study for a proteome-wide analysis. To determine whether predictions were consistent with what is known about CDK2, several GO enrichment analyses were performed (Supplementary Methods S1.5), comparing significantly overrepresented GO terms (Fisher's exact test, Bonferroni correction, E-value < 0.05) obtained for known CDK2 substrates with those obtained for the predicted substrates. We found that the known CDK2 substrates were enriched most strongly in various terms related to the G1/S transition of the cell cycle, such as DNA damage response and DNA repair (Supplementary Table S7). This is consistent with the role of CDK2 in the regulation of the transition from G1 to S phase in response to DNA damage (Coverley *et al*., 2002).

To investigate the agreement of PhosphoPICK predictions with known CDK2 substrates, we performed a proteome-wide scoring for CDK2 and took the top 300 novel predictions, excluding known CDK2 substrates from the set of predicted substrates. We again performed a GO enrichment analysis, and compared the values of the prediction terms with the significant terms that were found during the analysis on the known substrates. Supplementary Table S7 shows the GO terms found to be significantly overrepresented among known CDK2 substrates, ranked from most significant to least significant. Over half of the terms (31/59) were found to be significantly overrepresented among the novel substrates predicted by PhosphoPICK.

CDK2 is known to be a regulator of the TF E2F1 (Lorna Morris and Thangue, 2000), a member of the E2F family, that is known to play a role in the G1/S transition and DNA

replication during the S phase (Attwoll *et al*., 2004; Lammens *et al*., 2009; Biswas and Johnson, 2012). The E2F family is comprised of three classes of TFs: transcriptional activators, retinoblastoma (Rb)-dependent repressors, and Rb-independent repressors. What is currently lacking is an understanding of what specific roles in the S phase are controlled at transcriptional level by E2F, and at the post-translational level by CDK2.

In order to investigate what overlapping functions may exist between E2F-regulated transcription and CDK2-mediated phosphorylation, we took ChIP-Seq data (ENCODE Project Consortium, 2012) for E2F1 (activator), E2F4 (Rb-dependent repressor) and E2F6 (Rb-independent repressor). Supplementary Figure S1 shows a Venn diagram of the unique and overlapping gene targets that exist among the three TFs (Supplementary Methods S1.6). It has been shown previously that overlapping targets between E2F1 and E2F4 are enriched in DNA replication and repair GO terms (Lee *et al*., 2011). We found that the overlapping targets of all three TFs are also enriched in such GO terms.

We then combined our set of predicted CDK2 substrates with known CDK2 substrates and identified proteins from this combined set of substrates that were in the unique and overlapping groups of E2F targets. GO enrichment tests were performed with the CDK2 substrates as the foreground and the remainder of the TF target group as the background. This allowed us to detect what role CDK2 plays within these TF target groups. Supplementary Tables S8–S14 contain the GO terms found to be significantly overrepresented (E-value <0.05) among CDK2 substrates in the TF target groups. While we found significantly overrepresented GO terms in all TF target groups, we noticed a larger number of process-specific terms among the unique E2F1 targets (Supplementary Table S8), unique E2F6 targets (Supplementary Table S9) and the overlapping targets among all three TFs (Supplementary Table S11). We found that CDK2 substrates among unique E2F1 targets and unique E2F6 targets were enriched in several terms relating to the regulation of apoptosis, as well as ubiquitination. Substrates in the overlapping group of targets of all three TFs were enriched in terms relating to DNA replication and DNA damage repair.

4 DISCUSSION

Protein phosphorylation is a highly regulated process, being controlled by the binding specificity to the protein kinase catalytic site, as well as various cellular processes that further enhance the kinase-substrate fidelity (Kobe *et al*., 2005; Zhu *et al*., 2005). We have demonstrated how a probabilistic model of PPIs and cell-cycle data can be used to accurately classify kinase substrates. Importantly, we found that our model, when combined with sequence-operating methods, was able to improve the accuracy of kinase-specific phosphorylation site prediction at false-positive levels below 0.002 for tyrosine kinases and below 0.0005 for serine/threonine kinases.

One potential point of concern in this approach is that we only had access to cell-cycle data for a single cell type, and whether this could result in a tissue-specific influence that impeded predictions in some cell types. However, as the phosphorylation site data we obtained from Phospho.ELM originates from multiple cell types [including, for example, HeLa cells (Wells and Hickson, 1995; Kraft *et al*., 2003), HEK 293T cells (Johnson *et al*., 1999), MELN cells (Medunjanin *et al*., 2005) and T98G glioma cells (Hansen *et al*., 2001)], the performance of the model across these varying cell types validates the appropriateness of the data we used. We attribute this largely to the simple representation of the cell-cycle data as four Boolean variables, which would be unlikely to result in a cell type-specific bias. Somewhat counterintuitively, we found that the cell-cycle data did not improve prediction performance for the CDKs—kinases whose activity is strongly linked to cell-cycle progression—while offering performance increase to other kinases. Though this study focussed on the use of protein abundance data for representing protein cell-cycle profiles, we note that dynamic gene-expression data across the cell-cycle also exists for human proteins (Gauthier *et al*., 2010). Further study could investigate what influence dynamic gene-expression data can provide to kinase-substrate prediction.

We observed some variations among the performance evaluations for the individual kinases, indicating that the model works better on certain kinases. However, we found that the performance for prediction of kinase substrates (Table 1) was not necessarily an indicator of what improvement would be seen when applying the model to phosphorylation site prediction. For example, the PhosphoPICK algorithm had excellent performance when classifying tyrosine kinase substrates—in several cases with AUCs greater than 0.9. However, when predicting phosphorylation sites of tyrosine kinases using Predikin and GPS, we found that the prediction of tyrosine kinase phosphorylation sites benefitted the least from the addition of the PhosphoPICK score, and the score appeared to be detrimental to predictions made by NetworKIN.

The kinase family where PhosphoPICK consistently demonstrated the most powerful prediction performance was the CMGC family—principally CDK and MAPK kinases. We found that PhosphoPICK generally improved the prediction of phosphorylation sites for the CMGC kinases as tested across each of the three methods, though there were some cases where PhosphoPICK resulted in a decrease in the accuracy of NetworKIN predictions. As the kinases in these families have very similar binding patterns, it is likely that mediating proteins captured by the PhosphoPICK model make a greater contribution in the correct assignment of a kinase to a substrate. These results lend support to the intuitive notion that the addition of context information would support sequence-based predictions most powerfully when the kinase binding patterns are less specific, or are very similar among family members—or both, as is the case with CDK and MAPK kinases.

It was interesting to note that the putative CDK2 substrates within the overlapping E2F1, E2F4 and E2F6 targets groups were overrepresented with GO terms related to DNA replication and DNA damage repair. Considering this group of genes was itself already enriched in such terms (when compared with the proteome), this underscores the importance that CDK2 has in regulating DNA replication and DNA damage repair (Deans *et al*., 2006; Satyanarayana and Kaldis, 2009). There are several potential responses to DNA damage, but in some cases cells may undergo apoptosis (Zhou and Elledge, 2000; Bakkenist and Kastan, 2004). We also noticed that putative CDK2 substrates within the unique E2F1 and E2F6 target groups were both overrepresented with terms relating to the regulation of

apoptosis and ubiquitination. These are both processes that CDK2 has previously been implicated in (Hayami *et al*., 2005; Huang *et al*., 2006), and ubiquitination is also known to play an important role in regulating apoptotic proteins (Zhang *et al*., 2004). E2F1 is known to be a regulator of apoptosis (DeGregori *et al*., 1997), and similarly E2F6 can negatively regulate apoptosis (Yang *et al*., 2006), so it was interesting to find that the putative CDK2 substrates within the unique E2F1 and E2F6 target groups were enriched in apoptosis and ubiquitination GO terms. These results seem to suggest a dynamic regulatory interplay between the E2F family at the transcriptional level, and the CDK2 kinase at the post-translational level.

ACKNOWLEDGEMENTS

K.A.L.C. is supported, in part, by the Wound Management Innovation CRC (established and supported under the Australian Government's Cooperative Research Centres Program) and the Australian Cancer Research Foundation (ACRF) for the Comprehensive Cancer Genomics Facility at The University of Queensland Diamantina Institute. B.K. is a National Health and Medical Research Council Senior Research Fellow (APP1003325).

Conflict of interest: none declared.

REFERENCES

Attwoll,C. *et al.* (2004) The E2F family: specific functions and overlapping interests. *EMBO J*., **23**, 4709–4716.

Bakkenist,C.J. and Kastan,M.B. (2004) Initiating cellular stress responses. *Cell*, **118**, 9–17.

Baldi,P. *et al.* (2000) Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, **16**, 412–424.

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

Biswas,A.K. and Johnson,D.G. (2012) Transcriptional and nontranscriptional functions of E2F1 in response to DNA damage. *Cancer Res*., **72**, 13–17.

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.

Chatr-aryamontri,A. *et al.* (2013) The BioGRID interaction database: 2013 update. *Nucleic Acids Res*., **41**, D816–D823.

Choudhary,C. *et al*. (2009) Lysine acetylation targets protein complexes and co-regulates major cellular functions. *Science*, **325**, 834–840.

Coverley,D. *et al*. (2002) Distinct roles for cyclins E and A during DNA replication complex assembly and activation. *Nat. Cell. Biol*., **4**, 523–528.

Deans,A.J. *et al*. (2006) Cyclin-dependent kinase 2 functions in normal DNA repair and is a therapeutic target in BRCA1-deficient cancers. *Cancer Res*., **66**, 8219–8226.

DeGregori,J. *et al*. (1997) Distinct roles for E2F proteins in cell growth control and apoptosis. *Proc. Natl Acad. Sci. USA*, **94**, 7245–7250.

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

Do,C.B. and Batzoglou,S. (2008) What is the expectation maximization algorithm. *Nat. Biotechnol*., **26**, 897–899.

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 Bioinform*., **10**, 117.

Ellis,J.J. and Kobe,B. (2011) Predicting protein kinase specificity: Predikin update and performance in the DREAM4 challenge. *PLoS ONE*, **6**, e21169.

ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

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

Gauthier,N.P. *et al.* (2010) Cyclebase.org: version 2.0, an updated comprehensive, multi-species repository of cell cycle experiments and derived analysis results. *Nucleic Acids Res*., **38** (Suppl 1), D699–D702.

Gnad,F. *et al.* (2011) PHOSIDA 2011: the posttranslational modification database. *Nucleic Acids Res*., **39** (Suppl 1), D253–D260.

Good,M.C. *et al.* (2011) Scaffold proteins: hubs for controlling the flow of cellular information. *Science*, **332**, 680–686.

Hansen,K. *et al.* (2001) Phosphorylation-dependent and -independent functions of p130 cooperate to evoke a sustained G1 block. *EMBO J*., **20**, 422–432.

Harbour,J. *et al.* (1999) Cdk phosphorylation triggers sequential intramolecular interactions that progressively block Rb functions as cells move through G1. *Cell*, **98**, 859–869.

Hayami,R. *et al.* (2005) Down-regulation of BRCA1-BARD1 ubiquitin ligase by CDK2. *Cancer Res*., **65**, 6–10.

Horn,H. *et al.* (2014) KinomeXplorer: an integrated platform for kinome biology studies. *Nat. Methods*, **11**, 603–604.

Huang,H. *et al.* (2006) CDK2-dependent phosphorylation of FOXO1 as an apoptotic response to DNA damage. *Science*, **314**, 294–297.

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

Hwang,C.-S. *et al.* (2010) N-terminal acetylation of cellular proteins creates specific degradation signals. *Science*, **327**, 973–977.

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

Johnson,L.N. (2011) Substrates of mitotic kinases. *Sci. Signal*., **4**, pe31.

Johnson,T.K. *et al.* (1999) Phosphorylation of B-Myb regulates its transactivation potential and DNA binding. *J. Biol. Chem*., **274**, 36741–36749.

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

Kraft,C. *et al.* (2003) Mitotic regulation of the human anaphase-promoting complex by phosphorylation. *EMBO J*., **22**, 6598–6609.

Lammens,T. *et al.* (2009) Atypical E2Fs: new players in the E2F transcription factor family. *Trends Cell Biol*., **19**, 111–118.

Lee,B.-K. *et al.* (2011) Wide-ranging functions of E2F4 in transcriptional activation and repression revealed by genome-wide analysis. *Nucleic Acids Res*., **39**, 3558–3573.

Lim,S. and Kaldis,P. (2013) Cdks, cyclins and CKIs: roles beyond cell cycle regulation. *Development*, **140**, 3079–3093.

Lorna Morris,E.A. and Thangue,N.B.L. (2000) Regulation of E2F transcription by cyclin E-Cdk2 kinase mediated through p300/CBP co-activators. *Nat. Cell. Biol*., **2**, 232–239.

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

Medunjanin,S. *et al.* (2005) Glycogen synthase kinase-3 interacts with and phosphorylates estrogen receptor α and is involved in the regulation of receptor activity. *J. Biol. Chem*., **280**, 33006–33014.

Mehdi,A. *et al.* (2011) A probabilistic model of nuclear import of proteins. *Bioinformatics*, **27**, 1239–1246.

Mok,J. *et al.* (2010) Deciphering protein kinase specificity through large-scale analysis of yeast phosphorylation site motifs. *Sci. Signal*., **3**, ra12.

Olsen,J.V. *et al.* (2010) Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis. *Sci. Signal*., **3**, ra3.

Oniśko,A. *et al.* (2001) Learning Bayesian network parameters from small data sets: application of noisy-or gates. *Int. J. Approx. Reason*., **27**, 165–182.

Patrick,R. *et al.* (2012) Mapping the stabilome: a novel computational method for classifying metabolic protein stability. *BMC Syst. Biol*., **6**, 60.

Satyanarayana,A. and Kaldis,P. (2009) A dual role of Cdk2 in DNA damage response. *Cell Div*., **4**, 9.

Saunders,N. *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.

Scott,J.D. and Pawson,T. (2009) Cell signaling in space and time: where proteins come together and when they're apart. *Science*, **326**, 1220–1224.

Sherr,C.J. and Roberts,J.M. (1999) CDK inhibitors: positive and negative regulators of g1-phase progression. *Genes Dev*., **13**, 1501–1512.

Szklarczyk,D. *et al.* (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res*., **39** (Suppl 1), D561–D568.

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

Wells,N.J. and Hickson,I.D. (1995) Human topoisomerase iiα is phosphorylated in a cell-cycle phase-dependent manner by a proline-directed kinase. *Eur. J. Biol. Chem.*, **231**, 491–497.

Xue,L. *et al.* (2012) Sensitive kinase assay linked with phosphoproteomics for identifying direct kinase substrates. *Proc. Natl Acad. Sci. USA*, **109**, 5615–5620.

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.-W. *et al.* (2006) E2F6 negatively regulates ultraviolet-induced apoptosis via modulation of BRCA1. *Cell Death Differ.*, **14**, 807–817.

Zhang,H.-G. *et al.* (2004) Regulation of apoptosis proteins in cancer cells by ubiquitin. *Oncogene*, **23**, 2009–2015.

Zhou,B.-B.S. and Elledge,S.J. (2000) The DNA damage response: putting checkpoints in perspective. *Nature*, **408**, 433–439.

Zhu,G. *et al.* (2005) Protein kinase specificity: a strategic collaboration between kinase peptide specificity and substrate recruitment. *Cell Cycle*, **4**, 52–56.