Bioinformatics, 34(24), 2018, 4223–4231
doi: 10.1093/bioinformatics/bty522
Advance Access Publication Date: 27 June 2018
Original Paper

OXFORD

Sequence analysis

Quokka: a comprehensive tool for rapid and accurate prediction of kinase family-specific phosphorylation sites in the human proteome

Fuyi Li[1], Chen Li[1,2], Tatiana T. Marquez-Lago[3], André Leier[3], Tatsuya Akutsu[4], Anthony W. Purcell[1], A. Ian Smith[1,5], Trevor Lithgow[6], Roger J. Daly[1,*], Jiangning Song[1,7,*] and Kuo-Chen Chou[8,*]

[1]Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Clayton, VIC 3800, Australia, [2]Department of Biology, Institute of Molecular Systems Biology, ETH Zurich, Zurich 8093, Switzerland, [3]Department of Genetics, School of Medicine, University of Alabama at Birmingham, Birmingham, AL 35294, USA, [4]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Kyoto 611-0011, Japan, [5]ARC Centre of Excellence in Advanced Molecular Imaging, Monash University, Melbourne, VIC 3800, Australia, [6]Biomedicine Discovery Institute and Department of Microbiology and [7]Monash Centre for Data Science, Monash University, Clayton, VIC 3800, Australia and [8]Gordon Life Science Institute, Boston, MA 02478, USA

*To whom correspondence should be addressed.

Associated Editor: John Hancock

Received on February 12, 2018; revised on June 6, 2018; editorial decision on June 22, 2018; accepted on June 26, 2018

Abstract

Motivation: Kinase-regulated phosphorylation is a ubiquitous type of post-translational modification (PTM) in both eukaryotic and prokaryotic cells. Phosphorylation plays fundamental roles in many signalling pathways and biological processes, such as protein degradation and protein-protein interactions. Experimental studies have revealed that signalling defects caused by aberrant phosphorylation are highly associated with a variety of human diseases, especially cancers. In light of this, a number of computational methods aiming to accurately predict protein kinase family-specific or kinase-specific phosphorylation sites have been established, thereby facilitating phosphoproteomic data analysis.

Results: In this work, we present *Quokka*, a novel bioinformatics tool that allows users to rapidly and accurately identify human kinase family-regulated phosphorylation sites. *Quokka* was developed by using a variety of sequence scoring functions combined with an optimized logistic regression algorithm. We evaluated *Quokka* based on well-prepared up-to-date benchmark and independent test datasets, curated from the Phospho.ELM and UniProt databases, respectively. The independent test demonstrates that *Quokka* improves the prediction performance compared with state-of-the-art computational tools for phosphorylation prediction. In summary, our tool provides users with high-quality predicted human phosphorylation sites for hypothesis generation and biological validation.

Availability and implementation: The *Quokka* webserver and datasets are freely available at http://quokka.erc.monash.edu/.

Contact: roger.daly@monash.edu or jiangning.song@monash.edu or kcchou@gordonlifescience.org

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

© The Author(s) 2018. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

1 Introduction

Protein phosphorylation is a major post-translational modification (PTM), occurring when a phosphate group bonds with specific amino acids (such as serine, threonine and tyrosine) (Johnson and Barford, 1993). Numerous experimental studies have demonstrated that phosphorylation is involved in regulation of a variety of fundamental cellular processes, such as protein-protein interaction (Nishi *et al.*, 2011), protein degradation (Swaney *et al.*, 2013), signal transduction (McCubrey *et al.*, 2000) and signalling pathways (Duan and Walther, 2015). On the other hand, aberrant phosphorylation, usually introduced by mutations, is frequently causative of human diseases, including cancers (Fleuren *et al.*, 2016; Karaca *et al.*, 2015). Therefore, it is crucial to accurately identify human phosphorylation sites and to further characterise their biological functions.

Recent technical advances in mass spectrometry have significantly facilitated high-throughput analysis of entire proteomes and identification of specific types of PTMs; however, determining specific kinase(s) that are associated with PTM sites still remains a challenging task (Horn *et al.*, 2014). Kinases acting on phosphorylation sites are often unknown, making experimental validation of kinase-specific phosphorylation events even more challenging. In addition, a portion of phosphorylation sites cannot be identified as these modifications occur at very low levels in the cell (Boersema *et al.*, 2009). Therefore, development of computational methods that are capable of predicting kinase family-specific or kinase-specific phosphorylation sites is urgently needed. Such methods can assist biologists by providing high-quality predicted phosphorylation sites, guiding experimental design and complementing experimental efforts validating uncharacterized phosphorylation events.

To date, a number of computational methods have been established for this purpose. These tools can be classified into two groups: **(i)** Sequence-scoring function based methods, including PhoScan (Li *et al.*, 2007), GPS 2.0 (Xue *et al.*, 2008) and GPS 3.0 (http://gps.biocuckoo.org/). PhoScan is a statistical scoring function-based method, which uses the log-odds ratio to predict potential phosphorylation sites. GPS 2.0 is another statistical scoring function-based method that employs the Group-based Phosphorylation Scoring (GPS) approach to predict phosphorylation sites for a number of kinases from different groups/families/subfamilies. GPS 3.0 (i.e. the latest version of GPS; http://gps.biocuckoo.org/) is currently online available for academic use; **(ii)** Machine-learning based methods, including NetPhos 3.1 (Blom *et al.*, 2004), KinasePhos 2.0 (Wong *et al.*, 2007), Musite (Gao *et al.*, 2010), PhosphoPICK (Patrick *et al.*, 2015) and PhosphoPredict (Song *et al.*, 2017). NetPhos 3.1 is the most up-to-date version of the well-established NetPhos program. Benefiting from the artificial neural network (ANN) algorithm, NetPhos 3.1 can predict phosphorylation sites of 17 kinases in total. KinasePhos 2.0 is another machine-learning based method, developed based on a support vector machine (SVM) trained with a combination of sequence-derived features and solvent accessibility. Musite is an SVM-based bioinformatics tool that can predict both generic and kinase-specific phosphorylation sites (covering 13 kinase families) by integrating sequence similarities, protein disordered score and amino acid frequencies as the input features. PhosphoPICK utilizes cellular context and protein sequence information for phosphorylation site prediction trained by information extracted from three species: human (107 kinases), mouse (24 kinases) and yeast (26 kinases). PhosphoPredict is a random forest-based tool, which combines protein sequence-derived and functional features to predict kinase-specific phosphorylation sites for 12 human kinases and kinase families.

There has been outstanding progress in the development of useful methods for kinase family-specific or kinase-specific phosphorylation sites prediction; however, several issues still remain in most of the current methods that need to be addressed: (i) The datasets used for training are relatively out of date. For example, the methods aforementioned, including PhoScan, NetPhos 3.1, GPS 2.0, KinasePhos 2.0 and Musite, were all trained using the data extracted from older versions of the Phospho.ELM database (i.e. versions 3.0, 6.0 and 8.2). Since then, a larger number of experimentally verified novel kinase-specific phosphorylation sites have been identified and therefore should be incorporated into the training datasets to enhance the predictive power of the constructed models; (ii) The statistical techniques applied in the sequence-scoring function based methods are usually straightforward and prompt; however, as the performance of statistical scoring-based methods mainly depends on consensus patterns derived from the training data, the performance of these methods is usually less accurate than that of advanced machine learning-based methods (Miller and Blom, 2009; Song *et al.*, 2017) and (iii) Machine learning based methods usually require considerable computation time and resources to calculate heterogeneous and high-dimensional biological features for model training, although they generally outperform the sequence-scoring function based methods. Therefore, such machine learning-based methods are not practically suitable for high-throughput post translational modification (PTM) prediction. Based on these shortcomings, it becomes necessary to develop a novel computational method that is capable of predicting kinase family-specific phosphorylation sites both rapidly and accurately, allowing high-throughput and cost-effective kinase family-specific phosphorylation site prediction at the proteomic scale. This have strongly motivated us to develop the *Quokka* tool.

In this study, we introduce *Quokka* (*Qua*ntitative predictor of *k*inase family-specific *k*inome and phosphoryl*a*tion sites), to fill the above gaps and improve the performance of kinase family-specific phosphorylation sites prediction. *Quokka* includes a variety of sequence scoring functions for high-throughput phosphorylation prediction. Importantly, we built a logistic regression (LR) model by integrating the outcomes of sequence scoring functions for each kinase family. Experimental studies on phosphorylation sites of 11 kinase families using both benchmark and independent test datasets illustrate that *Quokka's* LR models performed best amongst all. When compared with the prediction performance of currently available predictors mentioned above, *Quokka* achieved highest AUC values for phosphorylation sites regulated by nine (out of 11) kinase families. Therefore, we subsequently built LR models for all 65 kinase families, which we also collected in this study to expand the predictive capacity of *Quokka*.

2 Materials and methods

2.1 Overall framework

We applied the Chou's 5-step rule (Chou, 2011) to construct and evaluate *Quokka*, as shown in Figure 1. These steps include data collection and pre-processing, sequence scoring, model construction and optimization, performance comparison and webserver construction. In the first step, the benchmark and independent tests datasets were collected from Phospho.ELM (version 9.0) (Dinkel *et al.*, 2011) and UniProt database (release June 07, 2017) (Pundir *et al.*, 2017) separately. In the second step, a variety of sequence scoring functions and their combinations were utilized to calculate scores for each protein, which were then used as the input features of the

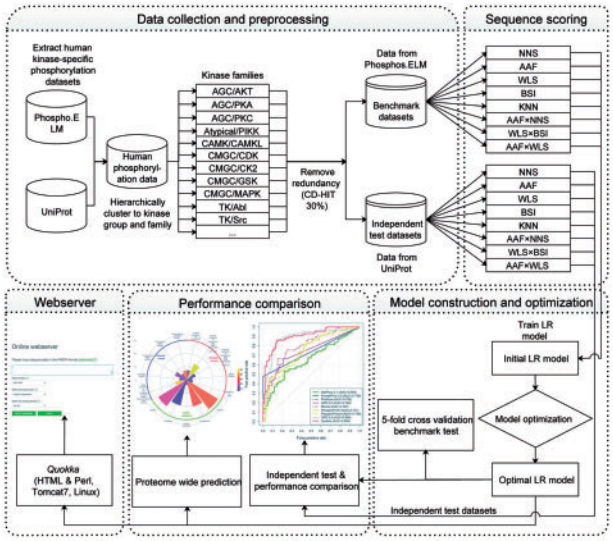

Fig. 1. Schematic framework of *Quokka*

LR model. An optimization algorithm was performed in the third step, based on the benchmark datasets to train the optimal LR for each kinase family. These optimized LR models were subsequently used for five-fold cross-validation test. In the fourth step, the optimized LR models were validated using the independent test datasets and the performance comparison with other existing methods was conducted. Using the optimized LR models, we further performed human proteome-wide prediction of phosphorylation sites. The *Quokka* server, based on the optimized LR models, was constructed in the final step, to facilitate public high-throughput prediction of phosphorylation sites.

2.2 Dataset collection and protein kinase classification

We extracted experimentally verified human kinase-specific phosphorylation sites from Phospho.ELM (Version 9.0) (Dinkel *et al.*, 2011) and UniProt (release June 07, 2017) (Pundir *et al.*, 2017), which are two well-curated public databases for protein post-translational modifications. The two resulting datasets were utilized as the benchmark and independent test datasets, respectively. To further expand the coverage of our curated datasets for human kinase-specific phosphorylation, we investigated the PhosphoSitePlus database and identified some novel kinase-specific phosphorylation sites using CD-HIT, which we added into the benchmark and independent test datasets. The statistics of augmented benchmark and independent test datasets are shown in Supplementary Table S1, respectively. All substrate proteins with experimentally validated phosphorylation sites were subsequently hierarchically clustered, according to the corresponding kinase group and family, adopting the same kinase nomenclature that links commonly used kinase names to a unified naming scheme (Eid *et al.*, 2017). We did not further sub-categorize kinase families down into kinase subfamilies and kinases, due to the relatively lower numbers of experimentally validated substrates within each subfamily and kinase.

To build accurate models and objectively assess the prediction performance, we removed sequence redundancy from all collected proteins using CD-HIT (Fu *et al.*, 2012) with a stringent sequence identity threshold of 30% for each kinase family. This ensured that any two protein sequences in the benchmark dataset and independent test dataset had a sequence identity of less than 30%. Finally, we collected a total of 43 serine/threonine and 22 tyrosine kinase families (each kinase family contained at least 5 phosphorylation sites) for training the *Quokka* models. Among this, nine serine/threonine kinase families and two tyrosine kinase families (each kinase family contained at least 20 phosphorylation sites) were selected for performance evaluation.

2.3 Sequence scoring functions

Quokka provides five sequence scoring functions, four combinations of individual scoring functions and an optimized logistic regression model for kinase family-specific phosphorylation prediction. The scoring functions used by *Quokka* are described as follows.

2.3.1 Nearest neighbour similarity (NNS)

NNS describes the similarity between two motifs $A(m, n)$ and $B(m, n)$ and is defined as:

$$\text{Similarity}(A, B) = \sum_{i=1,\ m+n+1} \text{Score}(A[i], B[i]) \quad (1)$$

where m and n denote the numbers of amino acids flanking both upstream and downstream of the centred phosphorylation site, respectively, while Score(A[i], B[i]) denotes the substitution score between amino acids A[i] and B[i] in the BLOSUM62 matrix (Henikoff and Henikoff, 1992). Therefore, $(m+n+1)$ denotes the total length of the sequence segment of a potential phosphorylation site, which was set to 15 (i.e. $m=n=7$) in our study.

2.3.2 Amino acid frequency (AAF)

The AAF score of a sequence segment with the centered phosphorylation site is calculated as:

$$\text{Score}_{\text{Frequency}} = \sum_{i=\text{P}(-7),\ \text{P}(+7)} \text{NF}_i \quad (2)$$

where NF_i is the normalized relative amino acid frequency, and P denotes the amino acid position surrounding the potential phosphorylation site (i.e. $P=[-7, 7]$). NF_i is defined as:

$$\text{NF}_i = \frac{f_i}{f_{i,\max}} \quad (3)$$

Here $f_i = n_i/N$ represents the frequency value of the amino acid at position i, while $f_{i,\max}$ denotes the frequency value of the most common amino acid at the same position.

2.3.3 WebLogo-based sequence conservation (WLS)

WebLogo (Crooks *et al.*, 2004) is a widely-applied sequence logo generation tool based on the calculation of the sequence conservation score (W). Here, we used the conservation score generated by WebLogo to rank all the potential phosphorylation sites. The conservation score of a sequence segment can be calculated as:

$$\text{Score}_{\text{webLogo}} = \sum_{\text{i=P}(-7),\text{P}(+7)} W_i \quad (4)$$

2.3.4 BLOSUM62 substitution Index (BSI)

The BLOSUM62 substitution matrix has been employed to predict phosphorylation sites in previous studies [such as GPS2.0 (Xue *et al.*, 2008) and Musite (Gao *et al.*, 2010)]. Given a testing sequence segment *Test*, its BSI score can be calculated as:

$$\text{Score}_{\text{BLOSUM62}} = \frac{\sum_{P(+7)}^{P(-7)} \text{B(Test)}}{\sum_{P(+7)}^{P(-7)} \text{B(Known)}}. \quad (5)$$

$\sum_{P(+7)}^{P(-7)}$ B(Test) sums up the substitution score of the sequence segment *Test* against all known phosphorylation sequence segments for a given kinase family; $\sum_{P(+7)}^{P(-7)}$ B(Known) sums up the substitution score of all know phosphorylation segments. Note that the value of $\text{Score}_{\text{BLOSUM62}}$ ranges from 0 to 1, and a higher value indicates that the sequence segment *Test* has a higher similarity to known phosphorylation sites. Finally, the highest $\text{Score}_{\text{BLOSUM62}}$ is selected as the BSI score.

2.3.5 K-nearest neighbours (KNN)

KNN (Altman, 1992; Chen *et al.*, 2018) was previously used for scoring phosphorylation sites by Gao *et al.* (Gao *et al.*, 2010). KNN aims to find the *k* nearest neighbours of a potential phosphorylation site by calculating the distance between the testing site and those known phosphorylation sites in the datasets. The distance between two sequence segments $A(m, n)$ and $B(m, n)$ is defined as:

$$\text{Dist(A, B)} = 1 - \frac{\sum_{i=1}^{m+n+1} \text{Sim(A[i], B[i])}}{m+n+1}, \quad (6)$$

where m and n denote the numbers of amino acids flanking both upstream and downstream of the centred phosphorylation site, respectively. Moreover, if the length of sequence fragment in the upstream or downstream is shorter than m or n, '*' (representing gaps in the BLOSUM62 matrix) will be used and added to the corresponding position upstream or downstream. In the equation (6), the Sim() function calculates the similarity between two amino acids, defined as:

$$\text{Sim(a, b)} = \frac{\text{Matrix(a, b)} - \min\{\text{Matrix}\}}{\max\{\text{Matrix}\} - \min\{\text{Matrix}\}}, \quad (7)$$

where a and b denote amino acids from the sequence segments A and B, respectively. In Equation (7), the Matrix function denotes the BLOSUM62 substitution matrix; while max and min present the largest and smallest value of the element in the Matrix, respectively.

The KNN function describes the average distance of a given sequence segment with a potential phosphorylation site to the segments of all the positive/negative phosphorylation sites. The calculated distance values are subsequently sorted and the *k* nearest neighbours selected. Finally, the KNN score (the percentage of positive neighbours among the *k* nearest neighbours) is calculated. In this study, *k* was as 0.5% of the total number of positive and negative sites in the training dataset for calculating the KNN score in *Quokka*, and the value of *k* varied depending on the particular kinase family.

2.3.6 Combinations of individual scoring functions

In addition to the five scoring functions mentioned above, *Quokka* also provides four combinations of individual scoring functions for kinase-specific phosphorylation sites prediction. These include AAF×NNS, WLS×BSI, NNS×WLS and AAF×WLS.

2.4 Model training and evaluation

2.4.1 Logistic regression (LR)

The LR algorithm is designed to estimate the distribution P(Y|X) from the training data, where Y is a discrete value and $X=<X_1, \ldots, X_d>$ represents any feature vector containing discrete or continuous variables. In this study, given the prediction of a phosphorylation site is a binary classification task, we only considered Y∈{0, 1}, where Y = 1 and Y = 0 indicate a positive (phosphorylation site) and negative sample (non-phosphorylation site), respectively. The LR model can be defined as:

$$p(Y = 1X) = g\left(\theta^T X\right) = \frac{1}{1 + e^{-\theta^T X}}, \quad (8)$$

where $g(z) = \frac{1}{1+e^{-z}}$ denotes the logistic function and $\theta^T X = \theta_0 + \sum_{i=1}^{d} \theta_i X_i$. Both values of $g\left(\theta^T X\right)$ and p(Y|X) range from 0 to 1. Therefore, p(Y = 0|X) can be estimated using the following equation, given that the sum of the probabilities must be equal to 1:

$$p(Y = 0X) = 1 - g\left(\theta^T X\right) = \frac{e^{-\theta^T X}}{1 + e^{-\theta^T X}}. \quad (9)$$

Algorithm 1 Optimization procedure of the Logistic Regression models

Input:
Training feature set, ***T***;
Output:
Optimal Logistic Regression model, ***optimisedLR***.

```
initialLR = trainLR(T);
initialSummary = summary(initialLR);
T` = T;
initialAUC = getAUC(initialLR);
for each i∈[1, n] do
    Pvalues[i] = getCoefficientsPvalues(initialSummary, i);
    if Pvalues[i] >0.05
        T` = removeFeature(i, T`);
    end if;
end for;
reducedLR = trainLR(T`);
reducedAUC = getAUC(reducedLR);
chiSquareValue = chisqTest(initialLR, reducedLR);
if (reducedAUC ≧ initialAUC) && (chiSquareValue > 0.05)
    optimisedLR = reducedLR;
end if;
else
    optimisedLR = initialLR;
end else;
return optimisedLR;
```

In this study, the LR models were implemented using the R package and trained using the individual scoring functions and their combinations as the inputs. We also used an optimization algorithm to optimize the LR models for each kinase families. Algorithm 1 describes the detailed procedure of our optimization approach for the LR model. In the first step, an initial LR model *initialLR* is trained by using the training dataset *T*. In the second step, *summary*(*initialLR*) obtains the detailed model measures information of the *initialLR*. The information contains the *P*-values of each feature based on hypothesis test for the model. These *P*-values indicate whether the corresponding feature makes a significant contribution to the model. For example, if the *P*-value of a feature is equal or lower than 0.05, it makes a significant contribution to the model, and *vice versa*. In step 3 a copy of the training feature set ***T***, ***T`***, is created for the next optimization procedure.

In step 4, *getAUC*(*initialLR*) extracted the AUC score, *initailAUC*, of the initial model. Steps 5–10 in Algorithm 1 describe our optimization strategy. For each feature i of feature set $\boldsymbol{T}$`, if the $P$-value of feature $i$ is insignificant (i.e. >0.05), this feature will be removed from $\boldsymbol{T}$` (step 7). Here in step 5, n denotes the dimension of the $\boldsymbol{T}$ and in the step 6, the P-value of feature i is extracted by the function *getCoefficientsPvalues*(*initialSummary*, i). This procedure will be halted till all the features in F has been evaluated. Afterwards, a new LR model (i.e. *reducedLR*) will be trained using optimized feature set $\boldsymbol{F}$ in step 11. Then, the AUC score, *reducedAUC*, of the *reducedLR* is obtained (step 12) and compared with the *initialLR* model (steps 13–15). A chi-square test was applied for testing whether the *reducedLR* model fits as well as the *initialLR* model. If the chi-square test result is bigger than 0.05 (not statistically significant), it means the *reducedLR* model fits as well as the *initialLR* model and *vice versa*. Therefore, the *reducedLR* model will be returned as the *optimisedLR* model if the chi-square test result is insignificant and the AUC score of the new model (*reducedLR*) is higher than the initial model (*initialLR*); otherwise, the initial LR model *initialLR* will be returned.

2.4.2 Performance evaluation

Throughout the literature, a set of four metrics directly taken from mathematics books are often used to quantitatively evaluate the prediction quality of a statistical predictor. They are (Chen *et al.*, 2007): Sn (sensitivity), Sp (specificity), ACC (overall accuracy) and MCC (Mathew's Correlation Coefficient) (Matthews, 1975). These metrics are not intuitive and hence biologists often find them difficult to understand. However, based on the Chou's symbols, originally introduced to study protein signal peptides (Chou, 2001a,b), a set of four very intuitive metrics were defined as given below (Chen *et al.*, 2013; Xu *et al.*, 2013):

$$\begin{cases} \mathrm{Sn} = 1 - \dfrac{N_-^+}{N^+} & 0 \le \mathrm{Sn} \le 1 \\ \mathrm{Sp} = 1 - \dfrac{N_+^-}{N^-} & 0 \le \mathrm{Sp} \le 1 \\ \mathrm{Acc} = \Lambda = 1 - \dfrac{N_-^+ + N_+^-}{N^+ + N^-} & 0 \le \mathrm{Acc} \le 1 \\ \mathrm{MCC} = \dfrac{1 - \left(\dfrac{N_-^+}{N^+} + \dfrac{N_+^-}{N^-}\right)}{\sqrt{\left(1 + \dfrac{N_+^- - N_-^+}{N^+}\right)\left(1 + \dfrac{N_-^+ - N_+^-}{N^-}\right)}} & -1 \le \mathrm{MCC} \le 1 \end{cases} \tag{10}$$

where N^+ represents the total number of positive samples investigated, while N_-^+ is the number of positive samples incorrectly predicted to be negative; N^- is the total number of negative samples investigated, while N_+^- is the number of the negative samples incorrectly predicted to be positive. With the set of formulations in Eq. 10, the meanings of Sn, Sp, Acc and MCC have become much clearer and easier to understand, as discussed in a series of recent studies in various biological areas (Ehsan *et al.*, 2018; Feng *et al.*, 2017, 2018; Lin *et al.*, 2014; Liu *et al.*, 2017a,b,c, 2018).

Moreover, we also plotted the Receiver-Operating Characteristic (ROC) curves and calculated the Area Under the Curve (AUC) values, as the primary measures to evaluate the predictive performance of *Quokka* and all the compared methods.

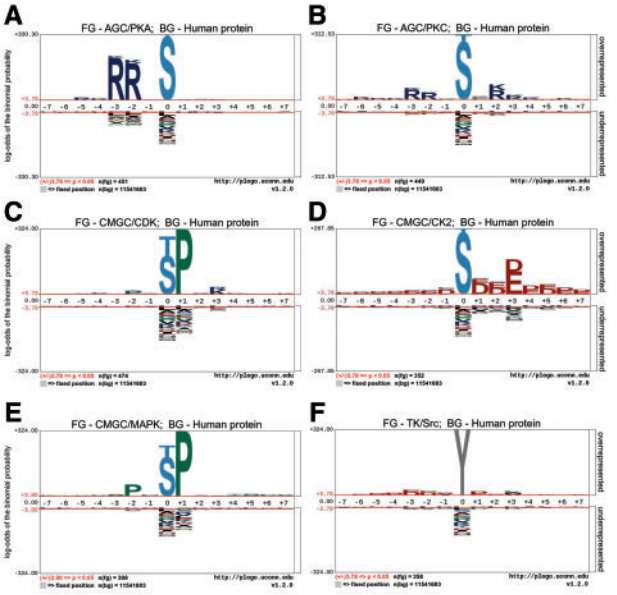

Fig. 2. Sequence logos showing the occurrences of amino acid residue types surrounding the kinase-specific phosphorylation sites for the six kinase families investigated in this study: (**A**) AGC/PKA; (**B**) AGC/PKC; (**C**) CMGC/CDK; (**D**) CMGC/CK2; (**E**) CMGC/MAPK and (**F**) TK/Src. FG, Foreground dataset; BG, Background dataset

3 Results and discussion

3.1 Amino acid specification and preference of phosphorylation in 11 kinase families

Based on the benchmark and independent test datasets, we analyzed the site specificity of 11 kinase-specific phosphorylation sites. Figure 2 and Supplementary Figure S1 show the sequence logos of the occurrences of amino acid residues located in both the upstream and downstream regions surrounding known phosphorylation sites for 11 kinase families (Only kinase families with more than 20 experimentally validated phosphorylation sites in the benchmark datasets were included in this analysis; see Supplementary Table S1 for more details). Not surprisingly, Figure 2 and Supplementary Figure S1 suggest that phosphorylation sites regulated by different kinases show different preferences in terms of the occurrences of flanking amino acids.

As expected, threonine (T) and serine (S) are the only two residue types occurring at the central position of sequence segments of phosphorylation sites of the serine/threonine kinases. Likewise, for phosphorylation sites of tyrosine kinases, the central position is predominated by tyrosine (Y). Nevertheless, other residue types at different positions also showed different amino acid preferences. For instance, a hallmark of the AGC/AKT kinase family is the requirement for a S or T residue at the central position with two arginine (R) in the upstream (-5 and -3 positions), as shown in Supplementary Figure S1A, i.e. the RXR phosphorylation site motif (Rust and Thompson, 2011).

3.2 Performance comparison among different scoring functions, combinations and the optimized LR model

We evaluated and compared the prediction performance of different individual scoring functions, combinations of individual scoring functions and the second-layer logistic regression models used by *Quokka* based on the kinase-specific phosphorylation site datasets for 11 kinase families. Each kinase family had at least 20

kinase-specific phosphorylation sites in both benchmark and independent test datasets (refer to Supplementary Table S1 for details). After applying this criterion, there were nine serine/threonine kinase families (AGC/AKT, AGC/PKA, AGC/PKC, Atypical/PIKK, CAMK/CAMKL, CMGC/CDK, CMGC/CK2, CMGC/GSK, CMGC/MAPK), and two tyrosine kinase families (TK/Abl and TK/Src) included in this analysis.

The percentages of true positives (TP) in the top 1, 3, 5, 10 and 20 ranking lists (Verspurten *et al.*, 2009) of predicted phosphorylation sites and the AUC, MCC, ACC, Sn and Sp are provided in Supplementary Table S2. As can be seen, the LR models performed best among all the compared scoring functions and their combinations in terms of AUC for all the 11 examined kinase families. In terms of the percentages of TP in the top *N* ranking lists, the LR models also performed best, except for CMGC/CDK [for which BSI performed best when predicting the True Positives (TP) in the Top 20 ranking list], CMGC/GSK (for which WLS×BSI performed best when predicting TPs in the Top 20 ranking list) and Atypical/PIKK (for which WLS performed best when predicting TPs in the Top 10 ranking list). In addition, BSI outperformed all other four scoring functions, except for the prediction of phosphorylation regulated by Atypical/PIKK. As for the three combinations of scoring functions, WLS×BSI outperformed the other two except for the AGC/AKT kinase family. In summary, BSI is the best scoring function among all five individual scoring functions, while the second-layer logistic regression model achieved the overall best performance. Accordingly, we used the LR model as the final model for performing phosphorylation site prediction in *Quokka*.

3.3 Performance comparison between *Quokka* and other prediction tools

In this section, we compared the prediction performance of *Quokka* against eight existing tools on the independent test datasets. Specifically, we compared *Quokka* against NetPhos 3.1 (Blom *et al.*, 2004), KinasePhos 2.0 (Wong *et al.*, 2007), PhoScan (Li *et al.*, 2007), GPS 2.0 (Xue *et al.*, 2008), Musite (Gao *et al.*, 2010), PhosphoPICK (Patrick *et al.*, 2015), PhosphoPredict (Song *et al.*, 2017) and GPS 3.0. To evaluate and compare the prediction performance, the protein sequences in FASTA format from the independent test datasets were submitted to each of these servers with default/recommended settings described by respective works. It is worth noting that some tools were specifically designed for certain kinases, indicating that they could be only used to predict the phosphorylation sites of the specific kinases they were designed for. As *Quokka* is designed for kinase family-specific phosphorylation prediction, we conducted performance comparisons of all the predictors at the kinase family level. To do so, for a given kinase family, we predicted the phosphorylation sites regulated by the kinase family. Then, the prediction result that received the best prediction score for each phosphorylation site was used to represent the final prediction outcome for this kinase family. For example, there are a number of kinases in the CMGC/CDK family, including CDK1, CDK2, CDK3, etc. PhosphoPICK is a computational tool that can predict phosphorylation sites regulated by eight types of CDK kinases (i.e. CDK1, CDK2, CDK3, CDK4, CDK5, CDK6, CDK7 and CDK9). To assess the performance of PhosphoPICK, we submitted the substrate sequences of the CMGC/CDK family to its webserver and selected all these eight CDK kinases to perform the prediction. For each protein sequence, PhosphoPICK calculated a score for each potential phosphorylation site (S/T). As a result, eight prediction scores of each S/T(/Y) site for eight kinases were generated. Then, for the

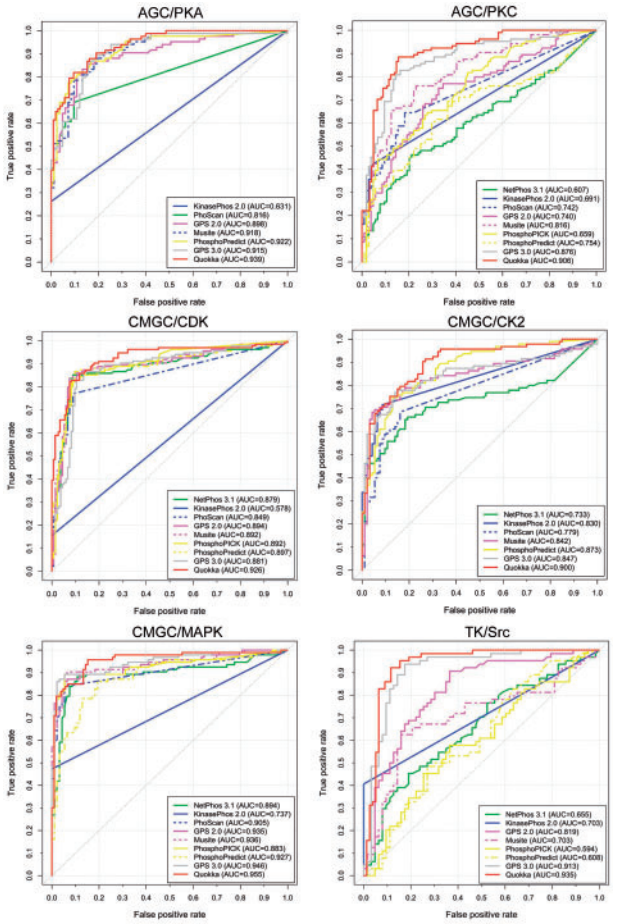

Fig. 3. Performance comparison results between different kinase-specific phosphorylation site predictors in terms of the AUC value, for AGC/PKA, AGC/PKC, CMGC/CDK, CMGC/CK2, CMGC/MAPK and TK/Src. The following phosphorylation site prediction tools were included and compared: *Quokka* (based on the logistic regression model), NetPhos 3.1, KinasePhos 2.0, PhoScan, GPS 2.0, Musite, PhosphoPICK, PhosphoPredict and GPS 3.0

sites which were labelled as positive samples in the independent test datasets, the best prediction score among these eight values was selected to represent the prediction outcome for such phosphorylation site regulated by the CMGC/CDK family. In contrast, for the sites which were labelled as negative samples, the minimum scores were selected as the prediction results to calculated the performance metrics. We summarize the performance comparison results between these methods for the 11 kinase-family-specific phosphorylation sites in Supplementary Table S3 and also present the ROC curves in Figure 3 and Supplementary Figure S2.

Figure 3 and Supplementary Figure S2 show the ROC curves of *Quokka* and other existing methods on the independent test datasets. As can be seen, *Quokka* (red line) outperformed the existing tools on nine out of 11 kinase families, with an exception of Atypical/PIKK and CMGC/GSK, for which GPS 3.0 performed best in terms of the AUC score (GPS 3.0, AUC = 0.925 versus *Quokka*, AUC = 0.921 for Atypical/PIKK and GPS 3.0, AUC = 0.886 versus *Quokka*, AUC = 0.870 for CMGC/GSK). The complete performance comparison results (as evaluated using AUC, MCC, ACC, Sn and Sp measures) on the independent tests are shown in Supplementary Table S3. Altogether, these results indicate *Quokka* is a powerful bioinformatics tool that provides a favourable prediction performance of kinase-specific phosphorylation prediction.

In addition, to investigate if *Quokka* is able to provide kinase-specific phosphorylation prediction, we selected eight kinases according to the numbers of available phosphorylation sites specific for these kinases (each kinase has more than 30 specific phosphorylation sites) from the independent test datasets, including AGC/AKT/AKT1, AGC/PKA/PKACA, AGC/PKC/PKCA, Atypical/PIKK/ATM, CMGC/CDK/CDK1, CMGC/CK2/CK2A1, TK/Abl/Abl1 and TK/Src/Src. We then conducted experiments to compare the performance of *Quokka* with existing methods at the kinase-specific level. The performance comparison results (in terms of AUC, MCC, ACC, Sn and Sp) are shown in Supplementary Table S4. We can see that *Quokka* outperformed the other eight tools on six out of eight tested kinases in terms of AUC, with the exception of AGC/AKT/AKT1 for which PhosphoPICK achieved the best performance (PhosphoPICK, AUC = 0.944 versus *Quokka*, AUC = 0.936) and CMGC/CK2/CK2A1 for which Musite achieved the best performance (Musite, AUC = 0.918 versus *Quokka*, AUC = 0.906).

3.4 Web server implementation

We constructed an open-access web server for *Quokka* using HTML and Perl programming languages, which we have made available at http://quokka.erc.monash.edu/. The *Quokka* web server is managed by Tomcat7 and resides on a Linux server, equipped with an 8-core CPU, 4 TB hard disk and 16 GB memory. Several steps need to be conducted to perform kinase-regulated phosphorylation prediction using *Quokka*. First, users need to provide *Quokka* proteins of interest in the FASTA format. An 'Example' link has been provided to assist users with the acceptable input format. Note that, to guarantee the prediction efficiency, we allow users to submit no more than 1000 sequences simultaneously. In the second step, users will need to select a specific kinase family for phosphorylation prediction. In total, *Quokka* provides prediction for 43 serine/threonine and 22 tyrosine kinase families. In the third step, a scoring function/model will be selected for *Quokka* to employ, overall, *Quokka* is able to conduct prediction for phosphorylation sites using nine functions/models. Based on our experimental results, the logistic regression models performed best amongst all scoring functions. However, the sequence scoring functions usually return the prediction result more rapidly than the logistic regression model. Therefore, the selection of appropriate scoring functions is left at the users' discretion, based on their computational requirement and the complexity of the question at hand. In the last step, users can choose the number of top-ranking predicted phosphorylation sites (N; $N = 1$, 3, 5, 10 and 20) to be displayed in the result webpage. An example of a typical input for *Quokka* is demonstrated in Supplementary Figure S3, and an online instruction explaining the input, prediction parameters and result interpretation are detailed on the *Quokka* website.

A strength of *Quokka* is that it can rapidly return the prediction result, thus facilitating high-throughput prediction. Our test suggested that, on average, it takes eight minutes to finish processing 1000 sequences and to return the prediction results. This efficiency is a result of *Quokka* not calculating high-dimensional feature sets for the input, unlike most other machine-learning based methods. After *Quokka* completes the prediction, the outcomes of all submitted sequences are returned to the result webpage. Each result table contains the prediction scores for each protein (Supplementary Fig. S3). To comprehensively demonstrate the prediction results, six sortable columns, including 'Rank', 'Position', 'Site', 'Motif', 'Score', 'Phosphorylation site?' and 'Kinase family', are provided within each table. All the prediction results can be easily exported to widely used file formats, including CSV, Excel® spreadsheet and PDF. In addition, there are two buttons 'Phosphorylation Sites Distribution' and 'Visualize' at the top of each result table for prediction results visualization. The corresponding results are shown in Supplementary Figures S5 and S6.

3.5 Proteome-wide prediction of kinase-specific phosphorylation sites

We further applied *Quokka* to perform proteome-wide prediction of kinase family-specific phosphorylation sites in the human proteome (containing 20 198 proteins extracted from the UniProt database; release June 07, 2017) for the 11 kinase families. We briefly summarize the results in this section. To obtain high-confidence predicted phosphorylation sites, we applied the prediction threshold at the 99% specificity (Gao *et al.*, 2010; Li *et al.*, 2015, 2016; Song *et al.*, 2018a,b,c). The statistical summary of the predicted phosphorylated proteins and phosphorylation sites for the 11 kinase families are shown in Supplementary Table S5. A complete list of the predicted phosphorylated proteins and their phosphorylation sites are available on the *Quokka*, and can be downloaded and analyzed on a local computer (http://quokka.erc.monash.edu/#proteome).

3.6 Gene ontology enrichment analysis

Furthermore, we performed an in-depth analysis of enriched gene ontology (GO) terms including cellular component (CC), biological process (BP), molecular function (MF) and pathways of predicted phosphorylated proteins at the proteome-scale, in an effort to provide an insight into biological annotations and functional roles of those predicted phosphorylated proteins. For this purpose, we carried out two-sided hypergeometric tests, to identify significantly enriched (P-value ≤ 0.05) GO terms and pathways against the background protein dataset composed by the entire human proteome. The P-value of a given term t can be defined as:

$$\text{P-value} = \text{F}_{\text{hypergeom}}(m, n, N, M), \tag{15}$$

where m is the number of predicted phosphorylated proteins annotated by the term t, n denotes the number of human proteins annotated by term t, N is the number of human proteins annotated by at least one term, while M is the number of phosphorylated proteins annotated by at least one term. All the statistical results of GO term and pathway analysis are listed in Supplementary Tables S5 and S6, respectively. Figure 4 and Supplementary Figure S4 highlights the top five significantly over-represented BP, CC and MF terms of the predicted phosphorylated proteins for the 11 kinase families at the proteome-scale.

Importantly, as shown by Figure 4 and Supplementary Figure S4, phosphorylated proteins by different kinase families are associated with different GO terms. The sectorial area for a GO term indicates the number of human proteins annotated by this term (n) while the different colors of the sectorial area represents the P-value for the corresponding GO term.

However, some similar GO terms can be found not only in the same kinase group but are shared by different families. For instance, in the AGC group (kinase families AGC/AKT, AGC/PKA and AGC/PKC), the phosphorylated proteins of both AGC/PKA and AGC/PKC families were found to be enriched with the GO BP terms 'Interferon-gamma-mediated signalling pathway (GO: 0002479)' and certain immune/anti-viral terms including 'Antigen processing and presentation of exogenous peptide antigen via MHC class I (TAP-dependent) (GO: 0002479)', 'Antigen processing and presentation of exogenous peptide antigen via MHC class I (TAP-independent) (GO: 0002480)',

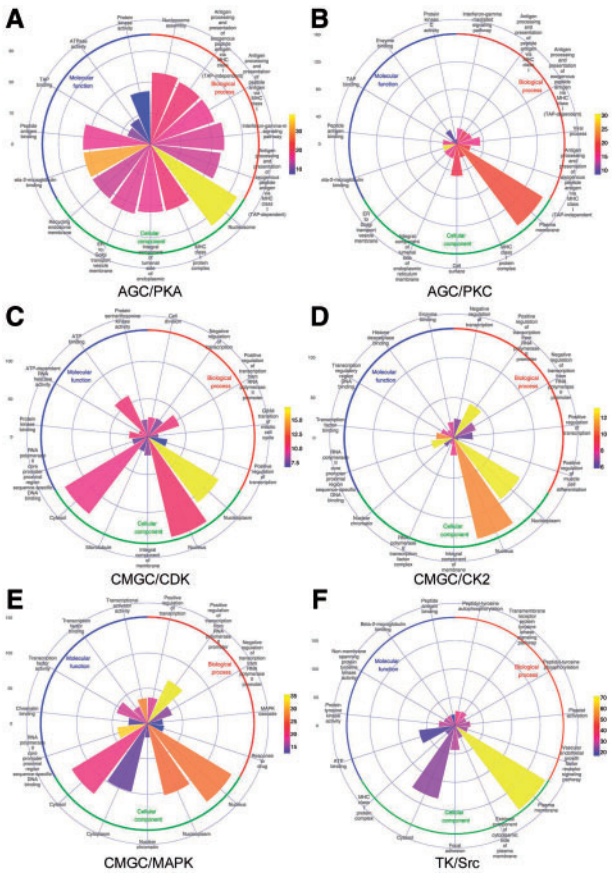

Fig. 4. The GO term distributions of predicted phosphorylated proteins regulated by 6 kinase families, including (**A**) AGC/PKA; (**B**) AGC/PKC; (**C**) CMGC/CDK; (**D**) CMGC/CK2; (**E**) CMGC/MAPK and (**F**) TK/Src. The enrichment analyses of GO terms, including biological process, molecular functions and cellular components for phosphorylated proteins were performed with the hypergeometric distribution

'Antigen processing and presentation of peptide antigen via MHC class I (GO: 0002474)'. This observation is consistent with previous experimental studies of the activities of these two kinase families (Burke *et al.*, 1989; David-Watine and Yaniv, 1996; Kirshner, 2000; Lv *et al.*, 2015). The GO_MF term 'Protein kinase activity (GO: 0004672)' was over-represented in phosphorylated proteins of AGC/AKT and AGC/PKA families, and 'Enzyme binding (GO: 0019899)' was over-represented in phosphorylated proteins of AGC/AKT and AGC/PKC families. In the case of the CMGC group (family CMGC/CDK, CMGC/CK2, CMGC/GSK and CMGC/MAPK), the GO_CC term 'Nucleus (GO: 0005634)' and the GO_BF term 'Positive regulation of transcription from RNA polymerase II promoter (GO: 0045944)' were over-represented in phosphorylated proteins for all these four kinase families.

In addition, our analysis also identified significantly enriched GO terms which were in good agreement with several previous experimental studies. For instance, the phosphorylated substrate of the Atypical/PIKK family was found to be enriched in the biological process of 'Cellular response to DNA damage stimulus'. This is not surprising considering the fact that ATM and ATR have been identified as important members of this family due to their key roles in DNA damage response (Cortez *et al.*, 1999; Zhou and Elledge, 2000). Similarly, the CMGC/CDK kinase family was characterized to be significantly enriched in the biological process of 'cell division'. For example, the cyclin-dependent kinases (CDKs) from the CMGC/CDK family, have been previously demonstrated to be strongly associated with the process of cell division (Ortega *et al.*, 2003; Wang *et al.*, 2000). In addition, our analysis identified that the GO CC term 'Focal adhesion' was enriched for the TK/Src, consistent with the known regulatory role for Src at this subcellular location (Frame, 2004). Overall, the GO term enrichment analysis suggests that kinase-specific phosphorylated proteins play multi-faceted and fundamentally important roles in a variety of biological processes.

4 Conclusion

In this study, we present *Quokka*, a novel computational tool for rapid and accurate prediction of kinase family-specific phosphorylation sites in the human proteome. *Quokka* combines various sequence-scoring functions, and a logistic regression optimization algorithm is used to train the optimal model for each kinase family. Our empirical studies, based on cross-validation and independent tests, demonstrate *Quokka* logistic regression models' competitiveness by outperforming existing tools, including NetPhos 3.1, KinasePhos 2.0, PhoScan, GPS 2.0, Musite, PhosphoPICK, PhosphoPredict and GPS 3.0. The improved performance of *Quokka* can be attributed to three major factors: i) Extraction of the most-recent experimental datasets that provide up-to-date knowledge on kinase-specific phosphorylation; ii) Inclusion of a variety of sequence-scoring functions and their combinations for calculating scores for each phosphorylation site, which were subsequently used as input features of the logistic regression models. These scoring functions themselves can be used as phosphorylation site prediction methods. Thus, *Quokka* is a two-level prediction system. The sequence-scoring functions are the first-level, while the subsequent logistic regression models are the second-level of this predictive system; iii) Use of an optimization algorithm to build the optimized logistic regression models which showed robust predictive power for each kinase family. The *Quokka* web server further provides a user-friendly interface and allows customizable prediction of kinase family-specific phosphorylation sites. The predicted human phosphoproteomic data by *Quokka* is also provided for further biological validation and analysis. We anticipate *Quokka* will be a valuable addition to efforts in developing next-generation bioinformatics tools for phosphorylation site identification and phosphoproteomic data analysis. We intend to apply this two-level prediction system to the prediction of kinase-specific phosphorylation sites of other species in our future work.

Funding

This work was supported by grants from the Australian Research Council (ARC) (LP110200333 and DP120104460), National Health and Medical Research Council of Australia (NHMRC) (490989), the National Institute of Allergy and Infectious Diseases of the National Institutes of Health (R01 AI111965) and a Major Inter-Disciplinary Research (IDR) Grant Awarded by Monash University. CL is currently supported by an NHMRC CJ Martin Early Career Fellowship (1143366). AL and TML were partially supported by Informatics start-up packages through the UAB School of Medicine. TL is an ARC Australia Laureate Fellow (130100038). RJD is an NHMRC Principal Research Fellow (1058540).

Conflict of Interest: none declared.

References

Altman,N.S. (1992) An introduction to kernel and nearest-neighbor nonparametric regression. *Am. Stat.*, **46**, 175–185.

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

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

Burke,T. *et al*. (1989) Phosphorylation of Class I but not Class II MHC molecules by membrane-localized protein kinase C. *Mol. Immunol*., **26**, 1095–1104.

Chen,J. *et al*. (2007) Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. *Amino Acids*, **33**, 423–428.

Chen,W. *et al*. (2013) iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. *Nucleic Acids Res*., **41**, e68.

Chen,Z. *et al*. (2018) iFeature: a python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics*, doi: 10.1093/bioinformatics/bty140.

Chou,K.-C. (2001a) Prediction of signal peptides using scaled window. *Peptides*, **22**, 1973–1979.

Chou,K.-C. (2001b) Using subsite coupling to predict signal peptides. *Protein Eng*., **14**, 75–79.

Chou,K.C. (2011) Some remarks on protein attribute prediction and pseudo amino acid composition. *J. Theor. Biol*., **273**, 236–247.

Cortez,D. *et al*. (1999) Requirement of ATM-dependent phosphorylation of brca1 in the DNA damage response to double-strand breaks. *Science*, **286**, 1162–1166.

Crooks,G.E. *et al*. (2004) WebLogo: a sequence logo generator. *Genome Res*., **14**, 1188–1190.

David-Watine,B. and Yaniv,M. (1996) Two RAREs and an overlapping CRE are involved in the hepatic transcriptional regulation of the Q10 MHC class I gene. *Cell Death Differ*., **3**, 37–46.

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

Duan,G. and Walther,D. (2015) The roles of post-translational modifications in the context of protein interaction networks. *PLoS Comput. Biol*., **11**, e1004049.

Ehsan,A. *et al*. (2018) A Novel Modeling in Mathematical Biology for Classification of Signal Peptides. *Sci. Rep*., **8**, 1039.

Eid,S. *et al*. (2017) KinMap: a web-based tool for interactive navigation through human kinome data. *BMC Bioinformatics*, **18**, 16.

Feng,P. *et al*. (2017) iRNA-PseColl: identifying the occurrence sites of different RNA modifications by incorporating collective effects of nucleotides into PseKNC. *Mol. Therapy Nucleic Acids*, **7**, 155–163.

Feng,P. *et al*. (2018) iDNA6mA-PseKNC: identifying DNA N6-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC. *Genomics*, doi: 10.1016/j.ygeno.2018.01.005.

Fleuren,E.D.G. *et al*. (2016) The kinome 'at large' in cancer. *Nat. Rev. Cancer*, **16**, 83–98.

Frame,M.C. (2004) Newest findings on the oldest oncogene; how activated src does it. *J. Cell Sci*., **117**, 989–998.

Fu,L. *et al*. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.

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

Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA*, **89**, 10915–10919.

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

Johnson,L.N. and Barford,D. (1993) The effects of phosphorylation on the structure and function of proteins. *Annu. Rev. Biophys. Biomol. Struct*., **22**, 199–232.

Karaca,M. *et al*. (2015) Mutation of androgen receptor N-terminal phosphorylation site Tyr-267 leads to inhibition of nuclear translocation and DNA binding. *PLoS One*, **10**, e0126270.

Kirshner,S. (2000) Major histocompatibility class I gene transcription in thyrocytes: a series of interacting regulatory DNA sequence elements mediate thyrotropin/cyclic adenosine 3', 5'-monophosphate repression. *Mol Endocrinol*, **14**, 82–98.

Li,F. *et al*. (2016) GlycoMine(struct): a new bioinformatics tool for highly accurate mapping of the human N-linked and O-linked glycoproteomes by incorporating structural features. *Sci. Rep*., **6**, 34595.

Li,F. *et al*. (2015) GlycoMine: a machine learning-based approach for predicting N-, C- and O-linked glycosylation in the human proteome. *Bioinformatics*, **31**, 1411–1419.

Li,T. *et al*. (2007) Prediction of kinase-specific phosphorylation sites with sequence features by a log-odds ratio approach. *Proteins*, **70**, 404–414.

Lin,H. *et al*. (2014) iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. *Nucleic Acids Res*., **42**, 12961–12972.

Liu,B. *et al*. (2017a) iRSpot-EL: identify recombination spots with an ensemble learning approach. *Bioinformatics*, **33**, 35–41.

Liu,B. *et al*. (2017b) 2L-piRNA: a two-layer ensemble classifier for identifying piwi-interacting RNAs and their function. *Mol. Therapy Nucleic Acids*, **7**, 267–277.

Liu,L.M. *et al*. (2017c) iPGK-PseAAC: identify lysine phosphoglycerylation sites in proteins by incorporating four different tiers of amino acid pairwise coupling information into the general PseAAC. *Med. Chem*., **13**, 552–559.

Liu,B. *et al*. (2018) iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC. *Bioinformatics*, **34**, 33–40.

Lv,D. *et al*. (2015) Neuronal MHC class I expression is regulated by activity driven calcium signaling. *PLoS One*, **10**, e0135223.

Matthews,B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.

McCubrey,J.A. *et al*. (2000) Serine/threonine phosphorylation in cytokine signal transduction. *Leukemia*, **14**, 9–21.

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

Nishi,H. *et al*. (2011) Phosphorylation in protein-protein binding: effect on stability and function. *Structure*, **19**, 1807–1815.

Ortega,S. *et al*. (2003) Cyclin-dependent kinase 2 is essential for meiosis but not for mitotic cell division in mice. *Nat. Genet*., **35**, 25.

Patrick,R. *et al*. (2015) PhosphoPICK: modelling cellular context to map kinase-substrate phosphorylation events. *Bioinformatics*, **31**, 382–389.

Pundir,S. *et al*. (2017) UniProt Protein Knowledgebase. *Methods Mol. Biol*., **1558**, 41–55.

Rust,H.L. and Thompson,P.R. (2011) Kinase consensus sequences—A breeding ground for crosstalk. *ACS Chem. Biol*., **6**, 881.

Song,J. *et al*. (2017) PhosphoPredict: a bioinformatics tool for prediction of human kinase-specific phosphorylation substrates and sites by integrating heterogeneous feature selection. *Sci. Rep*., **7**, 6862.

Song,J. *et al*. (2018a) PROSPERous: high-throughput prediction of substrate cleavage sites for 90 proteases with improved accuracy. *Bioinformatics*, **34**, 684–687.

Song,J. *et al*. (2018b) PREvaIL, an integrative approach for inferring catalytic residues using sequence, structural, and network features in a machine-learning framework. *J. Theor. Biol*., **443**, 125–137.

Song,J. *et al*. (2018c) iProt-Sub: a comprehensive package for accurately mapping and predicting protease-specific substrates and cleavage sites. *Brief. Bioinf*, doi: 10.1093/bib/bby028.

Swaney,D.L. *et al*. (2013) Global analysis of phosphorylation and ubiquitylation cross-talk in protein degradation. *Nat. Methods*, **10**, 676–682.

Verspurten,J. *et al*. (2009) SitePredicting the cleavage of proteinase substrates. *Trends Biochem. Sci*., **34**, 319–323.

Wang,H. *et al*. (2000) Expression of the plant cyclin-dependent kinase inhibitor ICK1 affects cell division, plant growth and morphology. *Plant J*., **24**, 613–623.

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

Xu,Y. *et al*. (2013) iSNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins. *PeerJ*, **1**, e171.

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

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