Amino Acids (2014) 46:1069–1078
DOI 10.1007/s00726-014-1669-3

ORIGINAL ARTICLE

Prediction of protein kinase-specific phosphorylation sites in hierarchical structure using functional information and random forest

Wenwen Fan · Xiaoyi Xu · Yi Shen · Huanqing Feng · Ao Li · Minghui Wang

Received: 3 October 2013 / Accepted: 8 January 2014 / Published online: 23 January 2014

Abstract Reversible protein phosphorylation is one of the most important post-translational modifications, which regulates various biological cellular processes. Identification of the kinase-specific phosphorylation sites is helpful for understanding the phosphorylation mechanism and regulation processes. Although a number of computational approaches have been developed, currently few studies are concerned about hierarchical structures of kinases, and most of the existing tools use only local sequence information to construct predictive models. In this work, we conduct a systematic and hierarchy-specific investigation of protein phosphorylation site prediction in which protein kinases are clustered into hierarchical structures with four levels including kinase, subfamily, family and group. To enhance phosphorylation site prediction at all hierarchical levels, functional information of proteins, including gene ontology (GO) and protein–protein interaction (PPI), is adopted in addition to primary sequence to construct prediction models based on random forest. Analysis of selected GO and PPI features shows that functional information is critical in determining protein phosphorylation sites for every hierarchical level. Furthermore, the prediction results of Phospho.ELM and additional testing dataset demonstrate that the proposed method remarkably outperforms existing phosphorylation prediction methods at all hierarchical levels. The proposed method is freely available at http://bioinformatics.ustc.edu.cn/phos_pred/.

Keywords Phosphorylation · Hierarchical structure · Functional information · Random forest

Electronic supplementary material The online version of this article (doi:10.1007/s00726-014-1669-3) contains supplementary material, which is available to authorized users.

W. Fan · X. Xu · Y. Shen · H. Feng · A. Li · M. Wang
School of Information Science and Technology, University of Science and Technology of China, 443 Huangshan Road, Hefei 230027, China
e-mail: wwfan@mail.ustc.edu.cn

X. Xu
e-mail: xxyyy@mail.ustc.edu.cn

Y. Shen
e-mail: sltian@mail.ustc.edu.cn

H. Feng
e-mail: hqfeng@ustc.edu.cn

A. Li
e-mail: aoli@ustc.edu.cn

A. Li · M. Wang (✉)
Research Centers for Biomedical Engineering, University of Science and Technology of China, 443 Huangshan Road, Hefei 230027, China
e-mail: mhwang@ustc.edu.cn

Introduction

Protein phosphorylation, as a crucial dynamic and reversible post-translational modification, plays an essential role in multiple biological cellular processes, such as transcription, translation, cell cycle, signal transduction, DNA repair (Schafmeier et al. 2005; Singh et al. 2005; Lou et al. 2004; Pawson 2004; Wood et al. 2009). The regulation of phosphorylation can be catalyzed by protein kinases in various cellular processes and multiple pathophysiologic conditions. Statistically, more than one-third of proteins are phosphorylated (Ma et al. 2005) and about half of phosphorylation processes are related to diseases (Manning et al. 2002). As phosphorylation processes are crucial in studying diseases and drug design, it is urgent to identify potential phosphorylation sites with related protein kinases.

To this end, several experimental methods were developed to identify protein phosphorylation sites with substrate-specific kinases, such as the low-throughput biological technique ^{32}P-labeling (Aponte et al. 2009) and the high-throughput biological technique mass spectrometry (Beausoleil et al. 2006). However, the experimental identification methods are costly and labor-intensive. Due to the limitation of experimental techniques, a number of computational approaches have been developed recently.

Blom et al. (1999) developed the first phosphorylation prediction method, which adopted artificial neural network algorithm based on primary peptide sequences. After that, a number of machine learning methods were developed for protein phosphorylation site prediction, such as NetPhosK (Blom et al. 2004), PPSP (Xue et al. 2006), KinasePhos (Wong et al. 2007), GPS (Xue et al. 2008), Musite (Gao et al. 2010). PPSP was used to predict potential phosphorylation sites for about 70 protein kinase groups. KinasePhos incorporated protein coupling pattern and sequence profile for nearly 60 protein kinase datasets. However, most of the kinase-specific prediction methods clustered protein kinases into sub-groups according to sequence homology. For example, PPSP clustered six protein kinases with high sequence similarity into a unique group of S6K. To gain insights into kinase function and evolution, Manning et al. (2002) proposed a method to classify various protein kinases into a hierarchical structure with four levels including kinase, subfamily, family and group, according to sequence comparison of their catalytic domains, which can be used to deduce the functions of human protein kinases from their family members. This method provided an initial point to analyze protein phosphorylation comprehensively and by this way potential phosphorylation sites can be predicted at different levels. According to this clustering rule, Xue et al. (2008) proposed a novel tool GPS that can hierarchically predict protein phosphorylation sites. This inspired us to take full advantage of hierarchical structure in protein phosphorylation studies.

Although a number of computational approaches have been developed, most of them, such as PPSP, Musite, GPS, only considered primary sequence information, which hampered further improvement of the prediction performance as protein phosphorylation is a complicated process with different biological mechanisms involved. To solve this problem, Li et al. (2010) used primary sequence and other biological information, such as secondary structure and protein–protein interaction (PPI), to identify kinase-specific protein phosphorylation sites in human. The test results showed that the functional information, represented by input features of a prediction method, were important for phosphorylation site prediction at family level, however, they only focused on a part of kinase families and more importantly the performance on other hierarchical levels, i.e., kinase, subfamily and group, was not investigated. In addition, the statistical approach used in their work may not be sufficient to obtain the effective subset of functional features from the high-dimensional functional data.

In this study, we conducted a systematic and hierarchical investigation of computational prediction of protein phosphorylation sites by adopting the clustering rule of protein kinases. We collected phosphorylation sites with related protein kinases from the latest version of Phospho.ELM data and grouped these protein kinases into four hierarchical levels. In addition to primary sequence, important functional information, including gene ontology (GO) terms and PPI, was also employed to improve prediction performance. We utilized a powerful feature selection algorithm called minimum-redundancy-maximum-relevance (mRMR) to extract useful functional information from high-dimensional GO and PPI data. After that, we generated hierarchy-specific predictive models based on random forest (RF) algorithm, which is a popular machine learning approach widely adopted in many areas of bioinformatics (Trost and Kusalik 2013; Teng et al. 2012; Yang 2009; Wang et al. 2010; Zou et al. 2012). The results demonstrated both GO and PPI information contributed to prediction performance remarkably at different hierarchical levels. Furthermore, by evaluating the performance of phosphorylation site prediction using Phospho.ELM and an additional testing dataset, we showed the proposed method outperformed other existing methods at all hierarchical levels.

Materials and methods

Data collection and pre-processing

The latest version of Phospho.ELM (9.0) (Dinkel et al. 2011) was selected as the phosphorylation dataset, which included 46,248 experimentally identified phosphorylation records. In this study, we focused on prediction of protein phosphorylation in human since most of the records (>80 %) were human proteins. The data was then carefully processed by the following steps:

1. After removing the redundant human phosphorylation entries, 3,151 non-redundant phosphorylation sites [1,881 serine (S) sites, 621 threonine (T) sites, and 649 tyrosine (Y) sites] within 934 protein sequences were selected with related protein kinase information.
2. To reduce the similarity of these 934 proteins, BLAST (version 2.2.9) (Dondoshansky and Wolf 2002) was employed with a sequence similarity threshold of 70 %. At this step, 889 phosphorylation proteins were selected.
3. Among the non-redundant protein sequences obtained in step 2, the phosphorylation sites annotated in

Phospho.ELM were considered as positive sites, while non-annotated were treated as negative sites. For each site, a phospho-peptide with length of 15, including the central residue and 7 residues upstream and downstream, was extracted from the corresponding protein sequence.

4. In previous studies, the homology reduction is only applied to the full-length protein sequences. To generate unbiased training data, the CD-HIT tool (Huang et al. 2010) was adopted to decrease the homology of phospho-peptides with length of 15 in each kinase with a similarity threshold of 70 %.
5. Next, we employed the method proposed by Manning et al. (2002) to cluster various protein kinases into a hierarchical structure including kinase, subfamily, family and group. For example, Figure S1 shows the hierarchal relationship of different kinases in CMGC group.
6. To ensure reliable results, only those protein kinases that contained more than 30 experimental phosphorylation sites were selected.

Based on this data processing procedure we finally obtained 54 protein kinases and the statistic of positive phospho-peptides is shown in Table S1. The negative data in phosphorylation dataset were randomly selected with a positive-to-negative ratio of 1:2.

Since GPS and Musite used Phospho.ELM database as training data, the phospho-peptides existed both in training and testing data would overestimate the prediction performance. To fairly compare these methods, we employed an additional testing dataset that included 4,417 kinase-specific phosphorylation sites in 652 substrates (Newman et al. 2013). After carefully removing the same phosphorylation sites within Phospho.ELM, we used the remaining phosphorylation sites as positive testing dataset. The negative testing dataset was randomly sampled from all other non-annotated S/T or Y sites in these substrates. The statistic of phosphorylation sites in this independent testing dataset is provided in Table S2.

Feature extraction

Besides protein primary sequences that are basic features to identify potential phosphorylation sites, in this study we also used GO terms and PPI as the functional features. The feature extraction procedure was described below:

Sequence features

The binary encoding scheme (Li et al. 2010) was adopted to transform each amino acid into a 21-dimensional binary vector. Then a peptide with length of 15 was transformed to a 315-dimensional vector.

GO features

GO terms were downloaded from gene ontology database (version 1.2) (Harris et al. 2004), which used three non-overlapping categories to describe the characteristics of gene or gene product: biological process (BP), cellular component (CC) and molecular function (MF). Totally 13,290 GO terms associated with the 889 human phosphorylation proteins were selected as the functional features of candidate substrates.

PPI features

The dataset of PPIs was retrieved from the STRING database (version 9.05), which contains both experimental and predicted interaction information (Von Mering et al. 2003). We obtained 16,025 proteins that had interactions with the selected phosphorylation proteins. For each PPI we used Gini index (Gastwirth 1972) to discretize PPI values. Suppose the sample set is split into two subsets with size $n_{1,t}$ and $n_{2,t}$ by a given threshold t, and the number of phosphorylation sites in each subset is $m_{1,t}$ and $m_{2,t}$, respectively. The definition of average Gini index is then calculated as:

$$\mathrm{Gini}(t) = \sum_{j=1}^{2} \frac{n_{j,t}}{N} \left(1 - \left(\frac{m_{j,t}}{n_{j,t}} \right)^2 - \left(\frac{n_{j,t} - m_{j,t}}{n_{j,t}} \right)^2 \right) \quad (1)$$

The optimal threshold t is determined as follow:

$$t = \arg\min_{t} \mathrm{Gini}(t), \quad t \in \{p_1, p_2, \ldots, p_i, \ldots p_l\} \quad (2)$$

where p_i is the ith value in the corresponding PPI feature vector.

Feature selection

In this study, we used mRMR (Peng et al. 2005) as the feature selection method, which is an informative filter method based on mutual information (Peng et al. 2013). The basic idea of mRMR considers both features related to sample classes and the redundancy of features simultaneously. A high mRMR score of a feature indicates that it is important to the class and less redundancy to other features. Suppose the whole feature set is represented by $V^{m \times n}$ and the sample class is c. We iteratively selected the feature with the highest mRMR score and based on previously selected feature set S_{k-1} with $k-1$ features, for the kth iteration the feature v_k is selected by:

$$v_k = \underset{\substack{v \in V^{m\times n} - S_{k-1} \\ k = 1,\ldots,n}}{\arg\max} \left[I(v,c) - \frac{1}{k-1} \sum_{v_i \in S_{k-1}} I(v,v_i) \right] \quad (3)$$

where I (v, c) represents mutual information between candidate feature v and class c, and I (v,v_i) represents mutual information between candidate feature v and selected feature v_i, respectively.

Classification method

The RF algorithm adopted in this study is an efficient ensemble classifier consisting of multiple individual decision trees (Breiman 2001; Wang et al. 2009). A decision tree is a classifier with tree-like model, including one root node, several intermediate and terminal nodes. To construct each decision tree in RF, two essential random processes are performed. First, a bootstrap dataset is sampled with replacement from the original dataset, and the size of the bootstrap dataset is same with original dataset. The un-sampled dataset is treated as out-of-bag (OOB) dataset. Second, a subset of features is selected randomly at each node in the decision tree. These two steps are repeated several times to build a RF, and then the RF takes a majority vote to define class for each test sample. In this paper, Willows software package (Zhang et al. 2009) was adopted to build 5,000 decision trees, and square root of total number of features is used at each node.

Performance evaluation

To evaluate the prediction performance, several measurements are calculated: sensitivity (Sn) and specificity (Sp) are defined as the ratio of positive or negative sites which could be correctly predicted, reflecting the predictive ability for phosphorylation or non-phosphorylation sites, respectively; precision (Pre) indicates the percentage of true positive in predicted positive; accuracy (Acc) demonstrates the ratio of true prediction in the test datasets; Matthew's correlation coefficient (MCC) illustrates the correlation between true and predicted classes, reflecting the balance quality. The detailed definitions are shown below:

$$\mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \quad (4)$$

$$\mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \quad (5)$$

$$\mathrm{Pre} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \quad (6)$$

$$\mathrm{Acc} = \frac{\mathrm{TN} + \mathrm{TP}}{\mathrm{TN} + \mathrm{TP} + \mathrm{FN} + \mathrm{FP}} \quad (7)$$

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FN}) \times (\mathrm{TP} + \mathrm{FP}) \times (\mathrm{TN} + \mathrm{FN}) \times (\mathrm{TN} + \mathrm{FP})}} \quad (8)$$

where TN, FN, TP and FP represent true negative, false negative, true positive and false positive, respectively. In addition, to further evaluate the performance, receiver operating characteristic (ROC) curve and the area under ROC curve (AUC) are also adopted in this work.

Results

Evaluation of functional information

We first ranked GO features by mRMR and then the GO features with top-ranked scores were added iteratively for prediction. The AUC results of tenfold cross validation for each kinase at different levels are shown in Fig. 1. The prediction performance of most kinases is significantly improved by adding GO features to sequence features. For example, AUC values are increased by 0.7 $\sim$ 12.0, 1.8 $\sim$ 11.2, 1.3 $\sim$ 6.3 and 0.8 $\sim$ 5.8 % when adding ten GO features at kinase, subfamily, family and group level, respectively. The results also show that the prediction performance in a hierarchical structure is consistently increased. Using hierarchical structure SRC(kinase)-SrcA(subfamily)-Src(family)-TK(group) as an example, the prediction performance is increased by 12.0, 11.2, 6.3 and 2.0 %. Further improvement in performance, albeit not dramatic for some kinases, is also observed with additional GO features added. Taken together, these results indicate that GO features contribute to the prediction performance at different levels. The same procedure was adopted to examine the efficiency of PPI features and the results are shown in Figure S2. Similarly, the AUC values are increased by 1.8 $\sim$ 16.3, 2.6 $\sim$ 12.2, 2.0 $\sim$ 10.5 and 1.2 $\sim$ 11.4 % when ten PPI features are added to sequence features from kinase to group level, respectively. In addition, detailed information regarding the prediction performance improvement contributed by ten top-ranked GO and PPI features separately are shown in Tables S3 and S4, and the results demonstrate that both GO and PPI can generally improve the performance of phosphorylation site prediction at different hierarchical levels.

Since both GO and PPI features were important to prediction performance of protein phosphorylation sites, we combined sequence and functional features to enhance phosphorylation site prediction. The ROC curves of four levels are plotted and shown in Fig. 2. For hierarchical structure PKCa-Alpha-PKC-AGC, the AUC values of RF with sequence and functional features are 94.4, 94.4, 93.8 and 92.6 % at kinase, subfamily, family and group level,

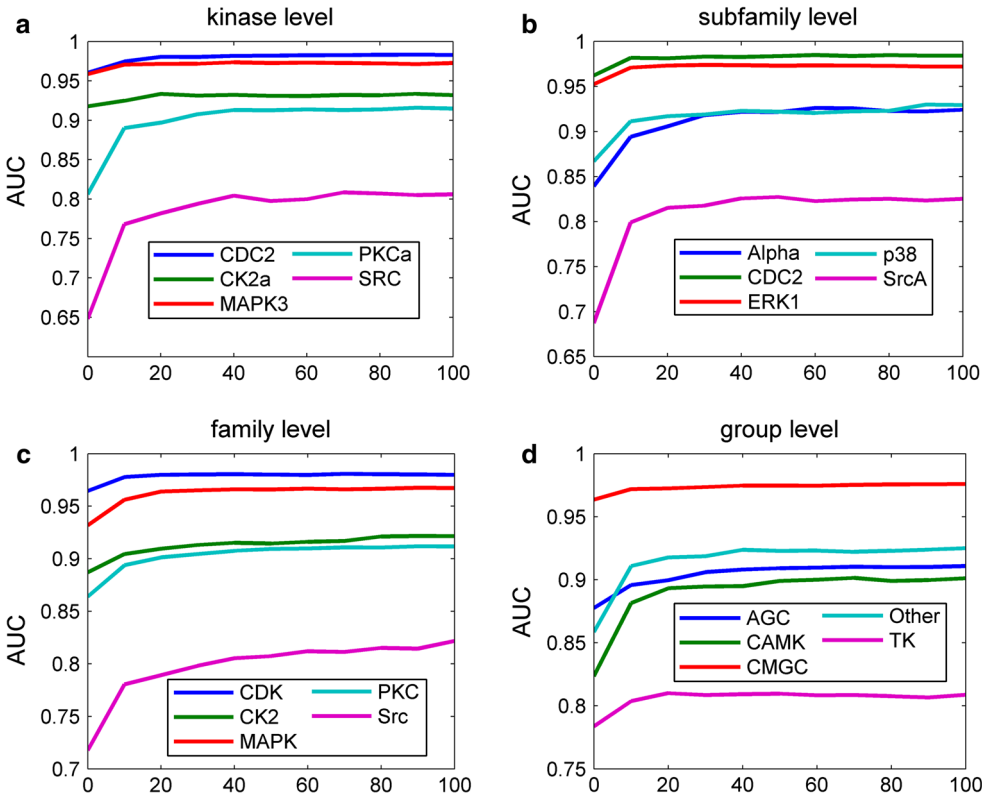

Fig. 1 Comparison of AUC values with different number of GO features. *x*-axis represents the number of ranked GO features added to sequence features, and *y*-axis represents the AUC of the corresponding models. The subplots *a*, *b*, *c* and *d* represent kinase, subfamily, family and group level, respectively. Each *line* represents the performance of one kinase data

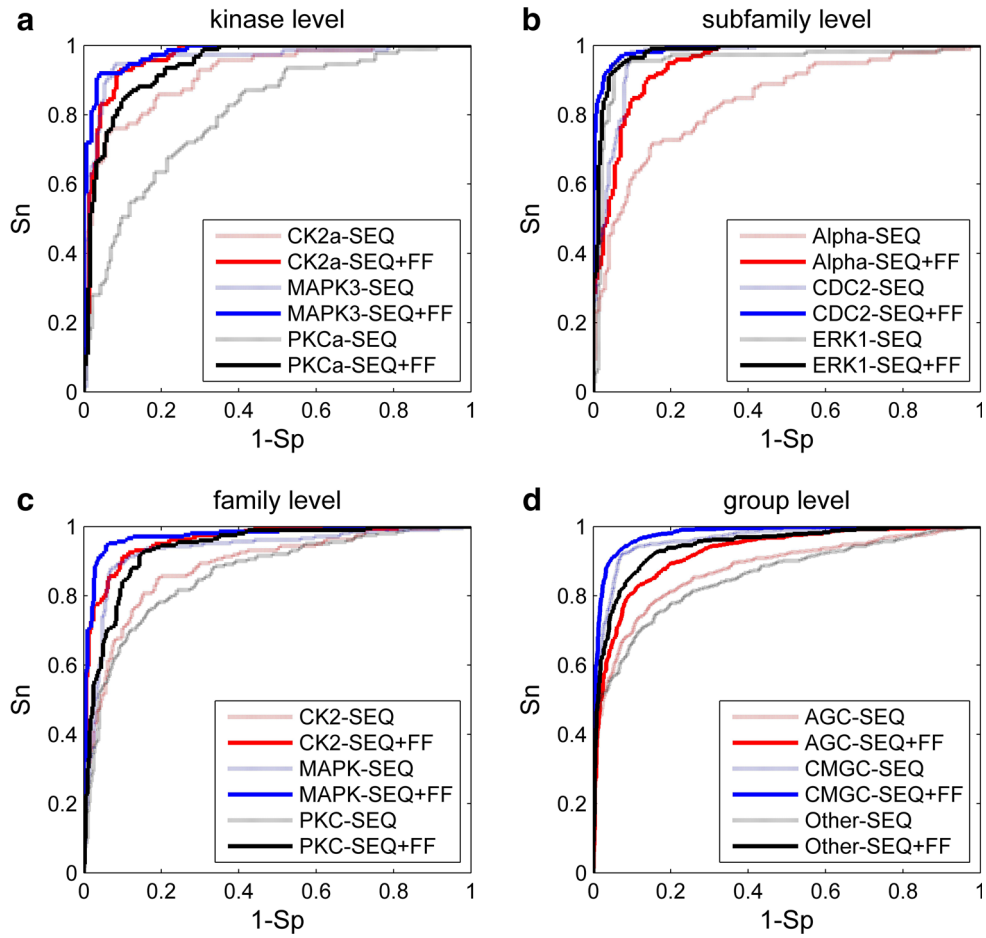

Fig. 2 ROC curves of kinase datasets with different features at all hierarchical levels. The *dotted lines* represent RF constructed with only sequence features and the solid lines represent RF built with sequence and functional features together. SEQ and FF represent sequence and functional features, respectively

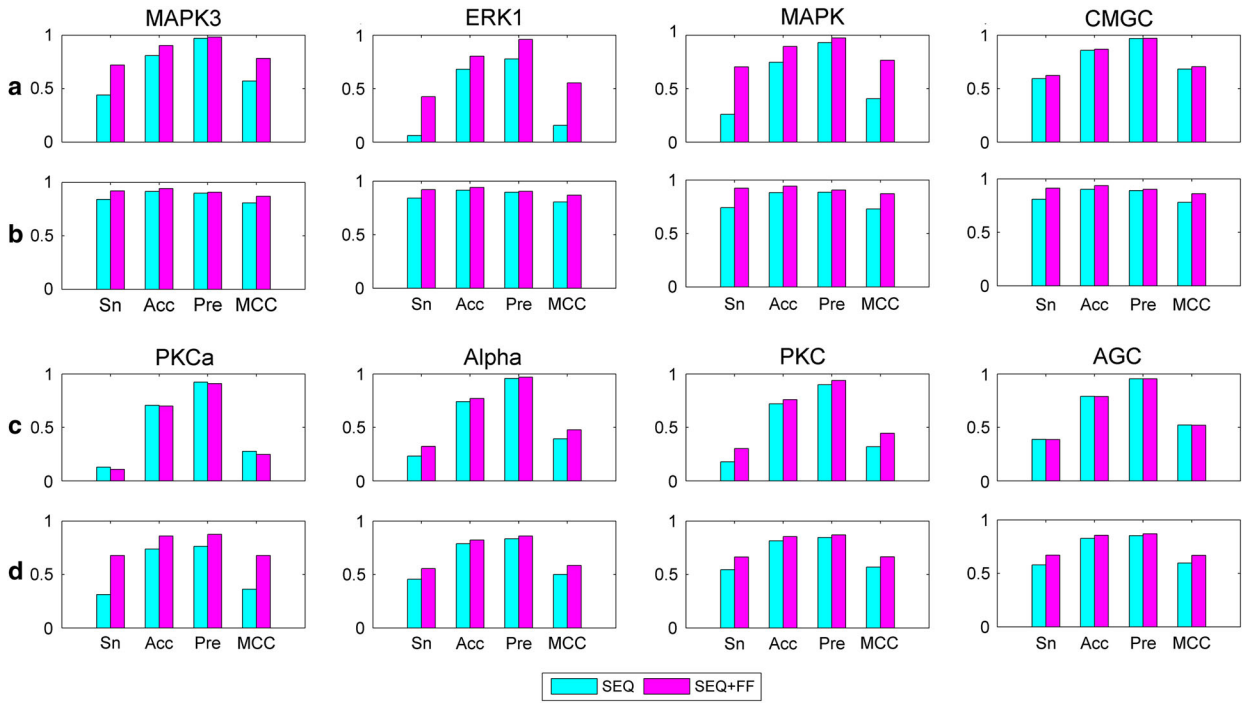

Fig. 3 Performance comparison of different features in two hierarchical structures. The rows *a* and *b* represent the measurements of hierarchical structure MAPK3-ERK1-MAPK-CMGC at specificities of 0.99 and 0.95, respectively. The rows *c* and *d* represent the measurements of hierarchical structure PKCa-Alpha-PKC-AGC at specificities of 0.99 and 0.95, respectively

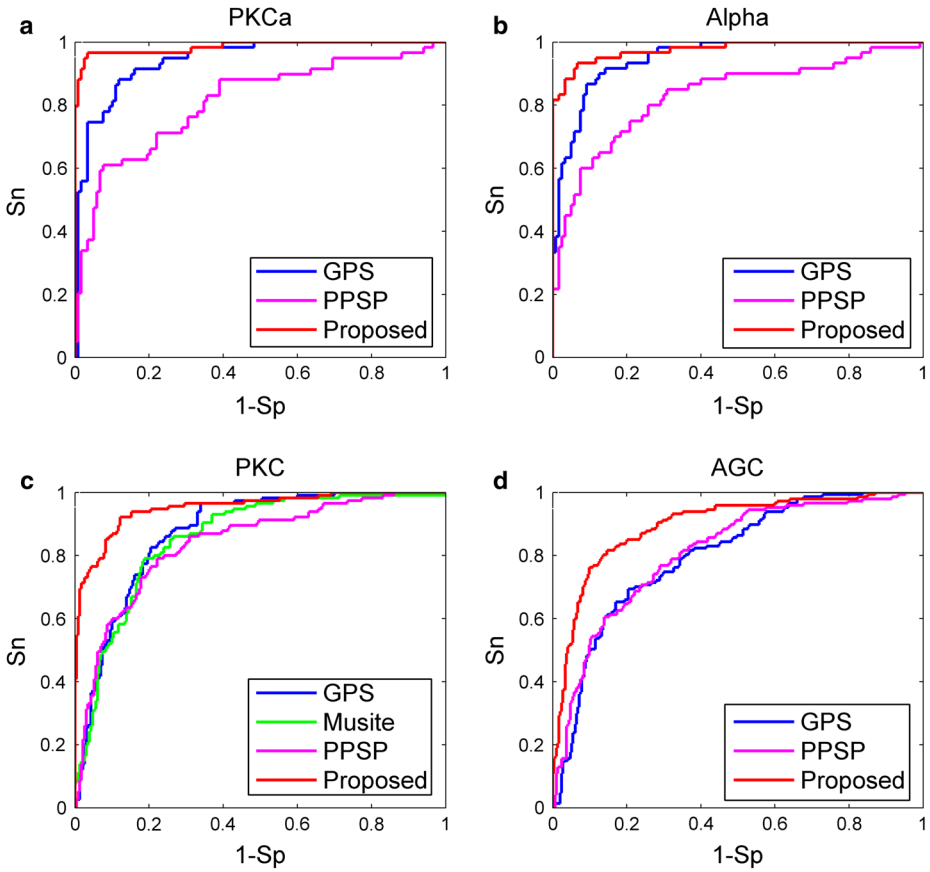

Fig. 4 ROC curves of different methods using additional testing data

Table 1 Performance comparison of different methods at different hierarchical levels

Method	Proposed				GPS		PPSP		Musite	
	High		Medium							
	Sn (%)	Sp (%)	Sn (%)	Sp (%)	Sn (%)	Sp (%)	Sn (%)	Sp (%)	Sn (%)	Sp (%)
Kinase level										
CK2a	68.1	99.1	83.4	97.2	72.4	96.0	49.7	96.6	–	–
GSK3B	68.9	100	82.2	95.6	53.3	94.4	28.9	93.3	–	–
MAPK1	71.5	99.1	83.5	96.2	65.8	96.8	38.6	95.9	82.9	95.9
MAPK3	80.3	99.1	88.0	98.3	53.8	96.2	35.0	96.6	63.2	92.3
MAPK14	83.8	100	89.2	97.3	56.8	94.6	27.0	95.9	–	–
PLK1	84.2	100	84.2	97.4	42.1	97.4	21.1	92.1	–	–
LCK	86.7	100	86.7	96.7	66.7	93.3	53.3	93.3	–	–
LYN	94.1	100	94.1	97.1	64.7	97.1	35.3	91.2	–	–
SYK	85.0	100	90.0	95.0	50.0	90.0	75.0	92.5	–	–
Subfamily level										
ERK1	76.3	99.3	87.9	96.4	57.0	94.4	13.5	98.6	–	–
JNK	48.4	99.4	60.9	97.2	55.3	96.9	18.6	96.6	–	–
Family level										
CDK	37.6	99.0	63.4	95.1	55.6	94.9	23.4	98.8	34.6	98.8
MAPK	63.9	99.0	82.1	97.2	58.5	97.0	25.3	98.7	25.3	98.7
Src	70.0	100	82.5	96.3	82.5	87.5	57.5	97.5	55.0	90.0
Group level										
CAMK	54.8	100	66.7	97.6	23.8	91.7	26.2	95.2	–	–
CMGC	37.5	99.8	72.0	97.1	55.5	97.1	61.9	95.6	–	–
TK	46.2	100	63.7	97.8	59.3	89.0	34.1	96.2	–	–

–, Not applicable

respectively, whereas the corresponding values with only sequence features are 80.6, 83.9, 86.4 and 87.8 %. Furthermore, the AUC values of all 54 kinases were illustrated by the box plots (Figure S3). With the help of functional features, the median AUC values are increased by 13.7, 21.0, 17.8 and 13.1 % at kinase, subfamily, family and group level, respectively. In addition to AUC, other performance measurements, such as Sn, Sp, Acc, Pre and MCC, were also employed to evaluate the performance. We evaluated the performance at high and medium stringency levels that correspond to $Sp > 99.0$ and 95.0 %, respectively. Figure 3 shows the performance for two hierarchical structures (MAPK3-ERK1-MAPK-CMGC and PKCa-Alpha-PKC-AGC) and it can be seen that the contribution of functional features is remarkable and consistent at different levels. Taken MAPK for instance, the values of Sn, Acc, Pre and MCC are increased by 44.5, 14.8, 4.3 % and 0.360 with Sp equal to 99.0 %, and 18.2 %, 6.1 %, 2.0 % and 0.142 with Sp equal to 95.2 %. These results further suggest that functional features are essential to improve the prediction performance of phosphorylation sites at different levels.

Performance comparison with existing methods

To further evaluate the prediction performance of our method, we compared it with other existing kinase-specific phosphorylation site prediction methods: GPS (version 2.1), Musite and PPSP. For the proposed method, we adopted the 889 phosphorylation proteins as training data and performed an unbiased tenfold cross validation. At the same time, we re-implemented the PPSP method and adopted the same evaluation procedure mentioned above. Since cross validation was unavailable for GPS and Musite, we simply used these proteins as testing data. It should be noticed that by this way the performance of GPS and Musite were over-estimated, as the phosphorylation sites were also used as training data for these two methods. However, the proposed method still achieved very competitive performance. As an example, Figure S4 shows the

ROC curves of the hierarchical structure PKCa-Alpha-PKC-AGC. At different levels, our proposed method consistently outperformed the competing methods.

To make better performance evaluation, the additional testing dataset was further employed for performance assessment and the corresponding ROC curves for the same hierarchal structure PKCa-Alpha-PKC-AGC are shown in Fig. 4. The performance of the proposed method is consistently better than other three methods at different levels. For PKC, at the high stringency level with Sp of 99.1 %, the Sn of the proposed method is 60.9 %, which is 58.3, 49.6 and 55.7 % higher than GPS, Musite and PPSP, respectively. We also listed the detailed performance obtained from different methods in Table 1, and the proposed method achieves the best performance at all hierarchal levels.

Interpretation of selected features

Since GO term is an ontology representing function information of a gene or gene product, we investigated the protein kinases and selected GO terms for functional analysis. We analyzed the selected GO terms for kinase CDC2 that is well known as the important role in modulating the cell cycle and the onset of mitosis. Intriguingly, we found some of the top-ranked GO terms, such as GO:0000075 ("cell cycle checkpoint"), GO:0007067 ("mitosis") and GO:0006260 ("DNA replication"), directly reflect the function of the catalytic kinase. A possible explanation is that CDC2 and its substrates may have similar functional relationship as they take part in the same signaling pathways to regulate cell cycle and mitosis. Therefore, the substrate GO terms related to the function of CDC2 are helpful in determining the CDC2-mediated protein phosphorylation. This result provides further support for the utility of GO information in phosphorylation prediction.

To better understand the biological meaning of selected PPI features, we also analyzed the interaction between the proteins associated with selected PPI features and the corresponding substrates. Take kinase LCK for instance, the interaction network diagram is given in Fig. 5, including the kinase LCK, protein substrates and the proteins associated with selected PPI features. The results show all experimentally identified protein substrates were correctly predicted as targets of LCK. Meanwhile, Fig. 5 demonstrates that the proteins associated with selected PPIs, such as CD19, CD79B, IL7 and LIME1, have intimate interactions with these protein substrates. One possible explanation is that these proteins bind both LCK and protein substrates and thereby play an important role in the phosphorylation progress. For example, it was reported that the CD19 receptor was physically associated with the rapid activation of tyrosine kinase LCK and significantly

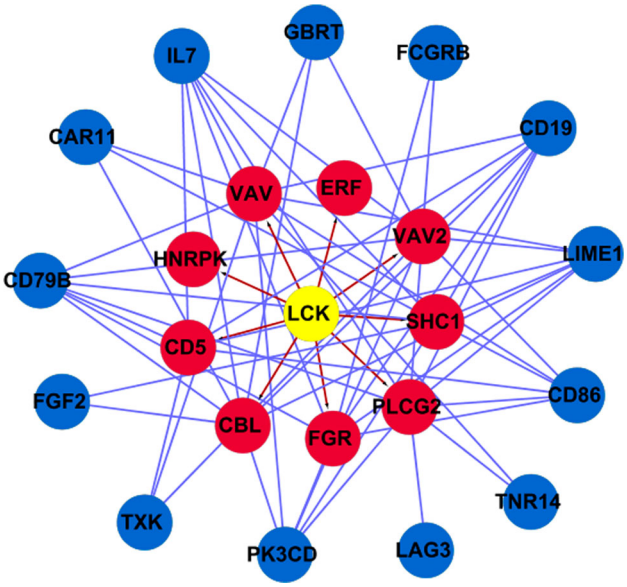

Fig. 5 Interactions among kinase LCK, the substrates and the selected proteins associated with PPI features. *Yellow*, *red* and *blue round* represent kinase, substrate and selected proteins associated with PPI features, respectively. *Red lines* mean the relationships between kinases and substrates, and *blue lines* mean relationships between substrates and selected proteins (color figure online)

enhanced tyrosine phosphorylation (Waddick et al. 1993). This suggests that the selected PPIs contain useful interaction information for phosphorylation prediction.

Discussions and conclusions

To overcome the shortcomings of phosphorylation site identification using experimental techniques, it is urgent to develop effective computational approaches. Although a number of computational approaches have been developed for kinase-specific phosphorylation site prediction, hierarchical structure has not been paid enough attention in these studies. Another issue is that most of these approaches use only local sequence information, which is not effective to identify phosphorylation sites. To solve these problems, in this study, we performed systematic investigation of protein phosphorylation prediction with respect to hierarchical levels of protein kinases. In addition, the functional information, including GO and PPI features, is incorporated to enhance the phosphorylation site prediction performance in all hierarchical structures. This finding was consistent with the study of Li et al. (2010), in which the performance improvement was mainly depended on functional information. For example, in their study functional information contributed most to GSK with more than 10 % improvement in Acc, and in this study, we also observed that the Acc of GSK family was improved by 15.5 % when functional information was adopted. Moreover, we also

evaluated the performance of structure features, such as predicted protein secondary structure and solvent accessibility, at all hierarchical levels. Compared with functional features, structure features did not significantly contribute to performance improvement (data not shown), which may lie in the fact that the predicted structure information is not accurate enough to provide sufficient discrimination power for phosphorylation site prediction.

In this study, we also examined the relevance of functional information at different levels of a hierarchical structure. For example, we analyzed the selected GO terms for CDC2 (kinase)-CDC2 (subfamily)-CDK (family) hierarchy, which is known to play an important role in cell cycle regulation. Intriguingly, we found totally 40 GO terms were concordantly selected for three hierarchical levels (Table S5), such as GO:0000075 ("cell cycle checkpoint"), GO:0000076 ("DNA replication checkpoint") in biological processes and GO:0000228 ("nuclear chromosome"), GO:0005654 ("nucleoplasm") in cellular components. This indicates the biological functions associated with these hierarchical levels are similar and therefore the selected GO terms are useful to gain insights into the functions of related kinases.

Protein phosphorylation is an important process that is related to many complex biological mechanisms. Since phosphorylation site recognition is benefited by biological information implicated in intrinsic phosphorylation mechanism, the aim of this study is to enhance the prediction performance of phosphorylation sites with relevant information at all hierarchical levels. Although the proposed method shows superior performance in phosphorylation site prediction, further improvement can still be obtained from various perspectives. For example, more biological information such as evolutionary information could be considered for further performance improvement. In addition, this study only focus on phosphorylation site prediction in human, and leaves large room for further study in other organisms. It also should be pointed out that since most of human phosphorylation data are generated from various cell sources/line or insufficiently controlled conditions, the method proposed in this study in fact disregards great biological differences that may greatly influence protein phosphorylation significantly, such as taxonomy/phylogeny, the basic living paradigm (nutrition, sex, age) (Maeshima et al. 2007; Lagranha et al. 2010; Jung et al. 2013) and the experimental methods to generate phosphorylation data (Fang et al. 2010). In fact, this critical issue has been widely ignored in current computational study of protein phosphorylation, partly because the corresponding experimental information of the phosphorylation data is not recorded in most phosphorylation databases. Therefore, further efforts should be made to address this issue in future studies of protein phosphorylation.

Acknowledgments This work was supported by National Natural Science Foundation of China (61101061, 31100955), Fundamental Research Funds for the Central Universities (WK2100230011).

Conflict of interest The authors declare that they have no conflict of interest.

References

Aponte AM, Phillips D, Harris RA, Blinova K, French S, Johnson DT, Balaban RS (2009) ³² P labeling of protein phosphorylation and metabolite association in the mitochondria matrix. Methods Enzymol 457:63–80

Beausoleil SA, Villén J, Gerber SA, Rush J, Gygi SP (2006) A probability-based approach for high-throughput protein phosphorylation analysis and site localization. Nat Biotechnol 24(10): 1285–1292

Blom N, Gammeltoft S, Brunak S (1999) Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. J Mol Biol 294(5):1351–1362

Blom N, Sicheritz-Pontén T, Gupta R, Gammeltoft S, Brunak S (2004) Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. Proteomics 4(6):1633–1649

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Dinkel H, Chica C, Via A, Gould CM, Jensen LJ, Gibson TJ, Diella F (2011) Phospho. ELM: a database of phosphorylation sites—update 2011. Nucleic Acids Res 39 (suppl 1):D261–D267

Dondoshansky I, Wolf Y (2002) Blastclust (NCBI Software Development Toolkit). NCBI, Bethesda

Fang B, Haura EB, Smalley KS, Eschrich SA, Koomen JM (2010) Methods for investigation of targeted kinase inhibitor therapy using chemical proteomics and phosphorylation profiling. Biochem Pharmacol 80(5):739–747

Gao J, Thelen JJ, Dunker AK, Xu D (2010) Musite, a tool for global prediction of general and kinase-specific phosphorylation sites. Mol Cell Proteomics 9(12):2586–2600

Gastwirth JL (1972) The estimation of the Lorenz curve and Gini index. Review Econ Stat 54(3):306–316

Harris M, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C (2004) The Gene Ontology (GO) database and informatics resource. Nucleic Acids Research 32 (Database issue):D258–2D61

Huang Y, Niu B, Gao Y, Fu L, Li W (2010) CD-HIT suite: a web server for clustering and comparing biological sequences. Bioinformatics 26(5):680–682

Jung H-J, Kim Y-J, Eggert S, Chung KC, Choi KS, Park SA (2013) Age-dependent increases in tau phosphorylation in the brains of type 2 diabetic rats correlate with a reduced expression of p62. Exp Neurol 248:441–450

Lagranha CJ, Deschamps A, Aponte A, Steenbergen C, Murphy E (2010) Sex differences in the phosphorylation of mitochondrial proteins result in reduced production of reactive oxygen species and cardioprotection in females. Circ Res 106(11):1681–1691

Li T, Du P, Xu N (2010) Identifying human kinase-specific protein phosphorylation sites by integrating heterogeneous information from various sources. PLoS One 5(11):e15411

Lou Y, Yao J, Zereshki A, Dou Z, Ahmed K, Wang H, Hu J, Wang Y, Yao X (2004) NEK2A interacts with MAD1 and possibly functions as a novel integrator of the spindle checkpoint signaling. J Biol Chem 279(19):20049–20057

Ma L, Chen Z, Erdjument-Bromage H, Tempst P, Pandolfi PP (2005) Phosphorylation and functional inactivation of TSC2 by Erk:

implications for tuberous sclerosis and cancer pathogenesis. Cell 121(2):179–193
Maeshima Y, Fukatsu K, Kang W, Ueno C, Moriya T, Saitoh D, Mochizuki H (2007) Lack of enteral nutrition blunts extracellular-regulated kinase phosphorylation in gut-associated lymphoid tissue. Shock 27(3):320–325
Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S (2002) The protein kinase complement of the human genome. Science 298(5600):1912–1934
Newman RH, Hu J, Rho H-S, Xie Z, Woodard C, Neiswinger J, Cooper C, Shirley M, Clark HM, Hu S (2013) Construction of human activity-based phosphorylation networks. Mol Syst Biol 9(1):655. doi:10.1038/msb.2013.12
Pawson T (2004) Specificity in signal transduction: from phosphotyrosine-SH2 domain interactions to complex cellular systems. Cell 116(2):191–203
Peng H, Long F, Ding C (2005) Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. Pattern Anal Mach Intell IEEE Trans 27(8): 1226–1238
Peng C, Wang M, Shen Y, Feng H, Li A (2013) Reconstruction and analysis of transcription factor–miRNA co-regulatory feed-forward loops in human cancers using filter-wrapper feature selection. PLoS One 8(10). doi:10.1371/journal.pone.0078197
Schafmeier T, Haase A, Káldi K, Scholz J, Fuchs M, Brunner M (2005) Transcriptional feedback of neurospora circadian clock gene by phosphorylation-dependent inactivation of its transcription factor. Cell 122(2):235–246
Singh CR, Curtis C, Yamamoto Y, Hall NS, Kruse DS, He H, Hannig EM, Asano K (2005) Eukaryotic translation initiation factor 5 is critical for integrity of the scanning preinitiation complex and accurate control of GCN4 translation. Mol Cell Biol 25(13): 5480–5491
Teng S, Luo H, Wang L (2012) Predicting protein sumoylation sites from sequence features. Amino Acids 43(1):447–455
Trost B, Kusalik A (2013) Computational phosphorylation site prediction in plants using random forests and organism-specific instance weights. Bioinformatics 29(6):686–694
Von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B (2003) STRING: a database of predicted functional associations between proteins. Nucleic Acids Res 31(1):258–261
Waddick KG, Chae HP, Tuel-Ahlgren L, Jarvis LJ, Dibirdik I, Myers DE, Uckun FM (1993) Engagement of the CD19 receptor on human B-lineage leukemia cells activates LCK tyrosine kinase and facilitates radiation-induced apoptosis. Radiat Res 136(3): 313–319
Wang M, Chen X, Zhang M, Zhu W, Cho K, Zhang H (2009) Detecting significant single-nucleotide polymorphisms in a rheumatoid arthritis study using random forests. In: BMC proceedings. BioMed Central Ltd, p S69
Wang M, Chen X, Zhang H (2010) Maximal conditional Chi square importance in random forests. Bioinformatics 26(6):831–837
Wong Y-H, Lee T-Y, Liang H-K, Huang C-M, Wang T-Y, Yang Y-H, Chu C-H, Huang H-D, Ko M-T, Hwang J-K (2007) KinasePhos 2.0: a web server for identifying protein kinase-specific phosphorylation sites based on sequences and coupling patterns. Nucleic acids research 35 (suppl 2):W588–W594
Wood CD, Thornton TM, Sabio G, Davis RA, Rincon M (2009) Nuclear localization of p38 MAPK in response to DNA damage. Int J Biol Sci 5(5):428
Xue Y, Li A, Wang L, Feng H, Yao X (2006) PPSP: prediction of PK-specific phosphorylation site with Bayesian decision theory. BMC Bioinform 7(1):163
Xue Y, Ren J, Gao X, Jin C, Wen L, Yao X (2008) GPS 2.0, a tool to predict kinase-specific phosphorylation sites in hierarchy. Mol Cell Proteomics 7(9):1598–1608
Yang ZR (2009) Predicting sulfotyrosine sites using the random forest algorithm with significantly improved prediction accuracy. BMC Bioinform 10(1):361
Zhang H, Wang M, Chen X (2009) Willows: a memory efficient tree and forest construction package. BMC Bioinform 10(1):130
Zou L, Huang Q, Li A, Wang M (2012) A genome-wide association study of Alzheimer's disease using random forests and enrichment analysis. Sci China Life Sci 55(7):618–625