Gao *et al. J of Biol Res-Thessaloniki* 2016, 23(Suppl 1):S12
DOI 10.1186/s40709-016-0042-y

Journal of Biological Research-Thessaloniki

RESEARCH Open Access

PredPhos: an ensemble framework for structure-based prediction of phosphorylation sites

Yong Gao[1†], Weilin Hao[1,2†], Jing Gu[1], Diwei Liu[1], Chao Fan[1], Zhigang Chen[1] and Lei Deng[1,3*]

From 2014 International Conference on Intelligent Computing (ICIC2014)
Taiyuan, China. 3-6 August 2014

Abstract

Background: Post-translational modifications (PTMs) occur on almost all proteins and often strongly affect the functions of modified proteins. Phosphorylation is a crucial PTM mechanism with important regulatory functions in biological systems. Identifying the potential phosphorylation sites of a target protein may increase our understanding of the molecular processes in which it takes part.

Results: In this paper, we propose PredPhos, a computational method that can accurately predict both kinase-specific and non-kinase-specific phosphorylation sites by using optimally selected properties. The optimal combination of features was selected from a set of 153 novel structural neighborhood properties by a two-step feature selection method consisting of a random forest algorithm and a sequential backward elimination method. To overcome the imbalanced problem, we adopt an ensemble method, which combines bootstrap resampling technique, support vector machine-based fusion classifiers and majority voting strategy. We evaluate the proposed method using both tenfold cross validation and independent test. Results show that our method achieves a significant improvement on the prediction performance for both kinase-specific and non-kinase-specific phosphorylation sites.

Conclusions: The experimental results demonstrate that the proposed method is quite effective in predicting phosphorylation sites. Promising results are derived from the new structural neighborhood properties, the novel way of feature selection, as well as the ensemble method.

Keywords: Phosphorylation sites, Ensemble learning, Structural neighborhood properties

Background

Protein phosphorylation is one of the most prevailing post-translational modifications [1], playing a significant role in regulating almost every cellular process, including transcription [2], translation [3] and signal transductions [4], etc. It is estimated that around 30–50 % of proteins in eukaryotic cells are phosphorylated and abnormal phosphorylation is now recognized as a cause of human disease, especially cancer [5]. Considering its prominent role in biochemistry, researches in identifying phosphorylation sites are booming in recent years.

Historically, the experimental methods of phosphorylation site annotation have undergone several stages from low-throughput proteomics based on site-directed mutagenesis to high-throughput biological technique [6–12] with the advent of mass spectrometry. Providing a number of verified phosphorylation sites, experimental identification is pivotal in understanding the mechanism of phosphorylation dynamic and provides the guidance in biomedical drug design.

*Correspondence: leideng@csu.edu.cn
†Yong Gao and Weilin Hao contributed equally to this work
[1] School of Software, Central South University, No. 22 Shaoshan South RD., Changsha 410075, China
Full list of author information is available at the end of the article

Several databases have been established to store annotated phosphorylation sites. Swiss-Prot [13] is a widely used protein sequence and knowledge database, which provides plentiful information about the post-translational modification. PhosphoBase [14] is another database which specifically stores experimentally verified phosphorylation sites, collected from Swiss-Prot and PIR protein database, literature studies and experiments. Phospho.ELM [15] contains 8718 substrate proteins covering 3370 tyrosine, 31,754 serine and 7449 threonine instances. Information about the phosphorylated proteins and the exact position of known phosphorylated instances, the kinases responsible for the modification and links to bibliographic references can be gained from the entries. PhosphoPep [16] contains 12,756 assigned phosphorylation sites identified in *Drosophila melanogaster* Kc167 cells. Four species of phosphoproteome data are included in PhosphoPep, which are yeast (*Saccharomyces cerevisiae*), worm (*Caenorhabditis elegans*), fly (*D. melanogaster*) and human (*Homo sapiens*), respectively, and a novel function was implemented to analyze the conservation of the identified phosphorylation sites across species. PHOSIDA [17] is a database aims to manage post-translational modification sites of various species, including human, mouse, fly, worm and yeast proteins. Also, PHOSIDA provides a wide range of analysis tools. Under the demand for analyzing the structural features of experimentally verified phosphorylation sites, Phospho3D [18] was launched for storing information retrieved from Phospho.ELM and was enriched with structural information and annotation at the residue level.

Given a long list of candidate phosphorylation sites in protein of interest, efforts to verify all of them by time-consuming and resource-intensive biological techniques remain challenging [19]. Alternatively, computational approaches have become increasingly popular. Up to now, there have been around 40 phosphorylation site prediction tools being established, varying from one tool to another with respect to several particular attributes, including dataset construction, feature selection, training system design and so on. The prediction tools can be grouped into two categories: kinase-specific and non-kinase-specific tools. A kinase-specific prediction program requires as input both a protein sequence and the type of a kinase, and produces some measure of the likelihood that each S/T/Y residue in the sequence is phosphorylated by the chosen kinase. In contrast, a non-kinase-specific prediction tool requires only a protein sequence as input, and reports the likelihood that each S/T/Y residue is phosphorylated by any possible kinase. DISPHOS [20] and NetPhos [21] are two typical non-specific predictors. DISPHOS investigated more than 1500 experimentally determined phosphorylation sites in eukaryotic proteins deriving from Swiss-Prot combined with PhosphoBase. Position-specific amino acid frequencies and disorder information are two of its crucial features. As to another non-specific phosphorylation site predictor called NetPhos, 584 serine sites, 108 threonine sites and 210 tyrosine sites were extracted mainly from PhosphoBase. An artificial neural network method is used to predict phosphorylation site with both sequence and structure information. NetPhosK [22] is a kinase-specific phosphorylation site predictor selecting six serine/threonine kinases mainly from PhosphoBase, which are PKA, PKC, PKG, cdc2, CK-2 and CaM-II. Another state-of-art kinase-specific prediction tool named KinasePhos [23] identifies phosphorylation sites based on Hidden Markov Model (HMM) in KinasePhos 1.0 and support vector machine (SVM) in KinasePhos 2.0, using experimentally validated phosphorylation sites from both PhosphoBase and Swiss-Prot. GPS [24] gained totally more than 2000 phosphorylation sites mainly from Phospho.ELM and could predict more than 70 kinds of kinase-specific phosphorylation sites. Based on similar data set of GPS, PPSP [25] predicts kinase-specific phosphorylation sites for 68 kinds of kinases, implementing an algorithm of Bayesian decision theory (BDT). In addition, PPSP can be used for many novel kinases, such as TRK, mTOR, SyK, and MET/RON, etc.

Although much progress has been made, there still exist several difficulties which keep phosphorylation site prediction far away to be perfectly solved. Firstly, while one-dimension sequence information is proved to harbor most of the predictive power, recently published analyses pointed out that phosphorylation sites may be closely related with its structural conformations and, furthermore, affected by the specific three-dimensional spatial environment [26]. Secondly, various kinds of features containing either sequence or structural information seems to be too sufficient for a predictor to be trained and a superabundant feature set would reduce calculation efficiency and increase space complexity. Thirdly, and also most importantly, the imbalanced problem exists widely in phosphorylation site prediction because the number of phosphorylation sites of a protein is usually much smaller than that of non-phosphorylation sites. The imbalanced data tends to cause over-fitting and poor performance.

In this paper, we report a novel structure-based computational method, PredPhos, that combines three main sources of information, namely site, Euclidean and Voronoi features describing the properties of either the target residue or the target residue's structural neighborhood. A set of optimal features were selected from 153 site, Euclidean and Voronoi properties by a two-step feature selection method (Table 1). Also, PredPhos uses a hybrid approach, which incorporates bootstrap resampling technique, SVM-based fusion classifiers and majority voting strategy, to overcome the imbalanced problem. We have benchmarked PredPhos using a set of experimentally

Table 1 Performance of the two-step feature selection method

	AUC	Accuracy	Recall	Specificity	CC	F1
CK2						
Optimal features	0.877	0.963	0.433	0.992	0.433	0.429
All features	0.842	0.954	0.350	0.986	0.370	0.366
MAPK						
Optimal features	0.839	0.952	0.483	0.973	0.480	0.480
All features	0.833	0.959	0.400	0.985	0.424	0.423
PKA						
Optimal features	0.858	0.948	0.375	0.980	0.426	0.432
All features	0.846	0.926	0.335	0.959	0.279	0.310
PKC						
Optimal features	0.857	0.952	0.303	0.985	0.396	0.363
All features	0.821	0.948	0.226	0.984	0.282	0.274
SRC						
Optimal features	0.900	0.951	0.558	0.973	0.510	0.499
All features	0.890	0.946	0.241	0.985	0.317	0.294

verified phosphorylation sites and an independent dataset. Results show that PredPhos significantly outperforms the state of the art methods for both kinase-specific and non-kinase-specific phosphorylation site prediction.

Methods

Datasets

The experimentally verified phosphorylation sites were extracted from Phospho.ELM version 9.0 [15], PhosphoPOINT [27] and PhosphoSitePlus [28]. After removing the redundant sites among these databases, 44,663 phosphoproteins which had at least one phosphorylated site were exacted in the first step. Phosphorylation sites were mapped to the protein entries of Protein Data Bank (PDB) by using Blast [29] with sequence similarity $\geq$90 %. Redundant PDB sequences were removed with 90 % sequence identity through CD-HIT [30]. 981 phosphoprotein chains and 2404 phosphorylation sites were remained. Negative phosphorylation sites gathered from their respective proteins had to meet three criteria: (1) a potential negative site could not have been reported as a positive site; (2) it had to be within a protein that contained known positive sites; and (3) a negative phosphorylation site had to be solvent-inaccessible. We divided these data into a training set and an independent test set based on the date the phosphorylation sites deposited to the database. Phosphorylation sites deposited before the year of 2008 were used to construct the non-kinase-specific training set, and the remaining sites composed the independent non-kinase-specific test set.

For the kinase-specific evaluation, training sites and test sites in the family of PKA, PKC, CK2, SRC and MAP were selected from the non-kinase-specific training set and independent test set, respectively.

Evaluation measures

The performance of the proposed prediction method is evaluated using tenfold cross-validation. The training data set is randomly divided into ten subsets with an approximately equal number of residues. For each time, nine subsets are used as training data and the remaining subset is used as test data.

Several widely used measures are adopted in this study, including *sensitivity (recall), specificity, precision, correlation coefficient (*CC*),* F1*-score and* AUC *score.*

These measures are defined as follows:

$$Sensitivity/Recall = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN});$$

$$Specificity = \mathrm{TN}/(\mathrm{FP} + \mathrm{TN});$$

$$Precision = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP});$$

$$CC = \frac{(TP \times TN - FP \times FN)}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}};$$

$$\mathrm{F1} = 2 \times Precision \times Recall/(Precision + Recall)$$

Above, the *TP, FP, TN* and *FN* are abbreviations of the number of true positives, the number of false positives, the number of true negatives and the number of false negatives, respectively. The *AUC* score is the normalized area under the ROC curve. The ROC curve is plotted with *TP* as a function of *FP* for various classification thresholds.

PredPhos framework

The framework of PredPhos is shown in Fig. 1. The computational approach used by PredPhos consists of three

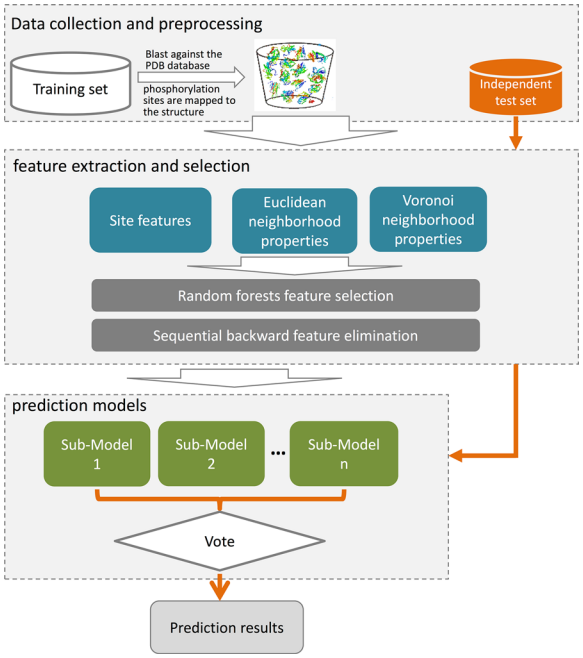

Fig. 1 The framework of PredPhos. Phosphorylation sites in the training set were mapped to the protein entries of Protein Data Bank (PDB) by using Blast. We encode each residue using 51 site features, 51 Euclidean neighborhood features and 51 Voronoi neighborhood features. The first step of feature selection is done by a random forest algorithm. Features are ranked in descending order by Z-Scores and the top 80 features are selected. The second step is performed using a wrapper-based feature selection. Features are evaluated by tenfold cross-validation with the SVM algorithm, redundant features are removed by sequential backwards elimination. Finally, an ensemble of n classifiers is built using different subsets, the final result is determined by majority votes among the outputs of the n classifiers

main component processes: (1) data collection and preprocessing: phosphorylation sites of both training and independent test set are mapped to the PDB structures with by using Blast [29] with sequence similarity ≥90 %; (2) feature extraction and selection: extract a wide variety of sequence, structural, and energy features, together with two types of structural neighborhoods, a two-step feature selection process that combines random forest and a sequential backward elimination; and (3) prediction models: ensemble classifiers are built for identifying phosphorylation sites based on the optimally selected features.

Feature extraction and selection

Site features

A large variety of 51 sequence, structural, and energy attributes are selected for the phosphorylation sites classification. Both conventional and new attributes were exploited in this kind of study, including PSSM (20) [29, 31, 32], evolutionary conservation score (1) [33], disorder (6) [34, 35], Solvent accessible area (ASA) (2) [36], pair potential (1) [37], atom and residue contacts (2) [38], Topographical index (1) [39], physicochemical features (6) [40], four-body statistical pseudo-potential (1) [41], local structural entropy (2) [42], side-chain energy (6) [41], Voronoi contacts (2) [43] and structural conservation score (1) [44]. The most interesting features are described below.

Four-body pseudo-potential

Given unique properties, the Delaunay tessellation [41] is an optimal choice when nearest neighbors should be objectively defined. Based on the Delaunay tessellation of proteins, the four-body pseudo-potential is defined as follows:

$$Q^{\alpha}_{ijkl} = \log\left[\frac{f^{\alpha}_{ijkl}}{p^{\alpha}_{ijkl}}\right] \quad (1)$$

Above, *i, j, k* and *l* represent the residue identities of the four amino acids (20 possibilities) in a Delaunay tetrahedron from the tessellation of the protein. Each residue is represented by a single point located at the centroid of the atoms in its side chain. Also, f^{α}_{ijkl} is the observed frequency of the residue composition *(ijkl)* in a tetrahedron of type a over a set of protein structures, while p^{α}_{ijkl} is the expected random frequency.

Local structural entropy

Each residue has its unique local structural entropy (LSE) [42], which can be calculated according to the protein sequence. The possibility of each candidate amino acid existing in eight secondary structure types (α-helices, π-helices, β-bridges, extended β-sheets, 3_{10}-helices, bends, turns and others) defined by DSSP is computed by averaging four sequential sequence windows along the protein sequence. A higher value of LSE indicates this amino acid is more likely to be found in these secondary structures.

In addition, an original attribute named ΔLSE is defined in order to estimate the distinction of LSE score between the wild-type protein and its mutants.

Side chain energy score

Each given residue of a protein has its own energy score which is originally applied for protein design [41]. A side chain energy score is a linear combination of various energetic terms, including buried hydrophilic solvent accessible surface between the current residue and the rest of the protein, buried hydrophobic solvent accessible surface, atom contact surface area, electrostatic interaction energy, hydrogen bonding energy, and overlap volume.

Structural conservation score

For a query protein, structural neighbors are obtained by using the structure alignment method-Ska [45]. Contact frequency maps are generated based on the mappings

between the neighbor's surface residues and the surface residues of the query protein [46]. Based on the contact frequency maps, conservation scores of each surface residue is calculated to evaluate degrees of interface conservation.

For all the site features, phosphorylation sites were represented as peptides of length 15, with the phosphorylated residue in the center and seven amino acids on either side. When a particular phosphorylated residue was too close to the beginning or end of the protein to have seven residues on either side, the missing residues were represented by gap (_) characters.

Structural neighborhood properties

Most of the conventional features such as physicochemical features, evolutionary conservation, and solvent accessible area describe only the properties of the current site itself, cannot represent the real situation efficiently, and thus are insufficient to predict phosphorylation sites with high accuracy. Here, we develop a new way to calculate two types of structural neighborhood properties using Euclidean distance and Voronoi diagram [47].

The Euclidean neighborhood is a group of residues located within a sphere of 10 Å defined by the minimum Euclidean distances between any heavy atoms of the surrounding residues and any heavy atoms from the central residue. The value of a specific residue-based feature *f* for neighbor *j* with regard to the target residue *i* is defined as

$$P_f(i,j) = \begin{cases} \text{the value of feature f for residue j} & \text{if } |\text{i} - \text{j}| \geq 1 \quad \text{and} \quad d_{i,j} \leq 10\,\text{Å} \\ 0 \qquad \text{otherwise} \end{cases} \tag{2}$$

Above, $d_{i,j}$ is the minimum Euclidean distance between any heavy atoms of residue *i* and any heavy atoms of residue *j*. The Euclidean neighborhood property of target residue *i* is defined as follows:

$$ENP_f(i) = \sum_{j=1}^{n} P_f(i,j) \tag{3}$$

where *n* is the total number of Euclidean neighbors.

We also use Voronoi diagram/Delaunay triangulation to define neighbor residues in 3D protein structures. For a protein structure, Voronoi tessellation partitions the 3D space into Voronoi polyhedra around individual atoms. Delaunay triangulation is the dual graph of Voronoi diagram, a group of four atoms whose Voronoi polyhedra meet at a common vertex form a unique Delaunay tetrahedra. In the context of Voronoi diagram (Delaunay triangulation), a pair of residues are said to be neighbors when at least one pair of heavy atoms of each residue have a Voronoi facet in common (in the same Delaunay tetrahedra). The definition of neighbors is based on geometric partitioning other than the use of an absolute distance cutoff, and hence is considered to be more robust. Voronoi/Delaunay polyhedra are calculated using the Qhull package that implements the Quickhull algorithm developed by Barber et al. [48]. Figure 1 illustrates an example of Voronoi/Delaunay neighbors (green) of a target residue (red).

Given the target residue *i* and its neighbors $\{j = 1,\ldots,n\}$, for each site feature *f*, a Voronoi/Delaunay neighborhood property is defined as follows:

$$VDP_f = \sum_{j=1}^{n} P_f(j) \tag{4}$$

where $P_f(j)$ is the value of the site feature *f* for residue *j*.

Two-step feature selection

In this paper, we propose a two-step feature selection method, as summarized in Algorithm 1, to select a subset of features that contribute the most in the classification.

In the first step, we assess the feature vector elements using the mean decrease Gini index (MDGI) calculated by the RF package in R [49, 50]. MDGI represents the importance of individual feature vector element for correctly classifying an interface residue into phosphorylation sites and non-phosphorylation sites. The mean MDGI Z-Score of each vector element is defined as

$$MDGI\ \text{Z} - \text{Score} = \frac{x_i - \bar{x}}{\sigma} \tag{5}$$

where is the mean MDGI of the *i-th* feature, $\bar{x}$ is the mean value of all elements of the feature *x*, and σ is the standard deviation (SD). Here, we select the top 80 features.

The second step is performed using a wrapper-based feature selection where features are evaluated by tenfold cross-validation performance with the SVM algorithm, and redundant features are removed by sequential backward elimination (SBE). The SBE scheme sequentially removes features from the whole feature set till an optimal feature subset is obtained. Each removed feature is the one whose removal maximizes the performance of the predictor. The ranking criterion $R_c(i)$ represents the prediction performance of the predictor, which is built on a subset features exclusive of feature *i*, and is defined as follows:

$$R_c(i) = \frac{1}{k}\sum_{j=1}^{k}\{AUC_j + Accu_j + Sen_j + Spe_j\} \quad (6)$$

where k is the repeat times of tenfold cross validation; AUC_j, $Accu_j$, Sen_j and Spe_j represent the values of AUC score, accuracy, sensitivity and specificity, respectively.

classifiers is a SVM. Here the LIBSVM package 2.8 [1] is used with radial basis function (RBF) as the kernel. Finally, a simple majority voting method is adopted in the fusion procedure, and the final result is determined by majority votes among the outputs of the n classifiers.

Algorithm 1 Two-step feature selection of PredPhos

Input:

Given the training set T = $\{(x_i, y_i)\}_{i=1}^{N}$, $x_i \in S^n$, y ∈ {-1,1}

Output:

Initialize S = [1, 2, ... n] to the subset of surviving features; ranked feature set R = [ ] ;
Train a random forest with features in set S as input variables;
Calculate MDGI Z-Score for each feature in S;
Features with MDGI Z-Score smaller than 2.5 are removed from S;
while $S \neq [\,]$ do
    for each variable *i* in S ,do
        Train SVM classifiers on feature set $S^{'}$, a subset of S exclusive of variable *i*;
        Evaluate the ranking criterion *Rc(i)* of variable *i*;
    end for
    Rank the variable that maximizes *Rc*;
    best = arg $\max_i$ *Rc* ;
    R = [*best R*];
    Remove the variable best from the set S ;
end
Return the ranked feature set R .

Prediction models

PredPhos uses an ensemble of n classifiers and decision fusion technique on the training datasets. An asymmetric bootstrap resampling approach is adopted to generate subsets. It performs random sampling with replacement only on the majority class so that its size is equal to the number of minority samples, and keeps the entire minority samples in all subsets.

First, the majority class of phosphorylation sites is under-sampled and split into n groups by random sampling with replacement, where each group has the same or similar size as the minority class of interaction sites. After the sampling procedure, we obtain n new datasets from the set of non-phosphorylation sites. Each of the new dataset and the set of phosphorylation sites are combined into n new training datasets. Then, we train n sub-models by using the n new training datasets as input. Each of these

Results and discussion

Selection of optimal features

We implemented tenfold cross-validation using two distinctive feature sets, namely full set of features (SVM-F) and sub-selected feature set (SVM-Sub). The comparison result is summarized in Table 2 and illustrated in Fig. 2. The performance of each model is measured by six metrics: area under curve (AUC), accuracy (Acc), sensitivity (Sn), specificity (Sp), CC, and F1-score.

As to accuracy and specificity, SVM-sub performed marginally worse than SVM-F in the family of MAPK (−0.07, −1.2 %) and SRC (+5.3, −1.2 %), but still outperformed in the family of CK2 (+0.9, +0.6 %), PKA (+2.4, +2.2 %) and PKC (+0.4, +0.1 %). It can be observed from Table 2 and Fig. 2 that SVM-sub shows dominant advantages over SVM-F in the other four metrics: AUC, sensitivity, CC, and F1-score for all five families. The improvement derived from the two-step feature selection is so obvious that we can also

Table 2 Performance comparison on the independent test dataset

Tools	Kinase family	Sn	Sp	Pre	CC	F1
PPSP	PKA	1.000	0.540	0.096	0.228	0.176
	PKC	0.400	0.527	0.031	−0.028	0.058
	CK2	0.500	0.390	0.038	−0.047	0.071
	SRC	0.538	0.859	0.286	0.304	0.373
	MAPK	0.571	0.380	0.043	−0.021	0.081
Kinasephos	PKA	0.125	0.877	0.048	0.001	0.069
	PKC	0.200	0.863	0.053	0.034	0.083
	CK2	0.500	0.976	0.500	0.476	0.500
	SRC	0.115	0.960	0.231	0.103	0.154
	MAPK	0.571	0.937	0.308	0.381	0.400
NetphosK	PKA	0.375	0.914	0.176	0.204	0.240
	PKC	0.200	0.802	0.037	0.001	0.063
	CK2	0.500	0.805	0.111	0.158	0.182
	SRC	0.038	1.000	1.000	0.187	0.074
	MAPK	0.286	0.979	0.400	0.311	0.333
GPS	PKA	0.500	0.871	0.160	0.222	0.242
	PKC	0.600	0.695	0.070	0.119	0.125
	CK2	0.500	0.854	0.143	0.202	0.222
	SRC	0.462	0.871	0.273	0.265	0.343
	MAPK	0.571	0.789	0.118	0.182	0.195
PredPhos	PKA	0.571	0.779	0.100	0.164	0.170
	PKC	0.824	0.870	0.452	0.544	0.583
	CK2	1.000	0.659	0.176	0.341	0.300
	SRC	0.789	0.802	0.234	0.356	0.361
	MAPK	0.375	0.986	0.600	0.452	0.462

easily get the intuitive comparison directly from Fig. 2. Concretely, for the family of CK2, after the operation of random forest feature selection based on the full set of 153 features (SVM-F), the value of AUC, sensitivity, CC, and F1-score increased by about 4.2, 23.7, 17 and 17.2 %, respectively. For the family of MAPK, the value of AUC, sensitivity, CC, and F1-score increased by about 0.7, 20.8, 13.2 and 13.5 %, respectively. For the family of PKA, the value of AUC, sensitivity, CC, and F1-score increase by about 1.4, 11.9, 52.7 and 39.4 %, respectively. For the family of PKC, the value of AUC, sensitivity, CC, and F1-score increased by about 4.4, 34, 40 and 32.5 %, respectively. As to the family of SRC, the value of AUC, sensitivity, CC, and F1-score increase by about 1.1, 131.5, 60.9 and 69.7 %, respectively.

Taken the family of PKC as an example, the size of its optimal feature set is 25, although shrank by about 84 % compared with the original size of 153, the prediction performance significantly improved, indicating that our two-step feature selection method can effectively improve the prediction performance with less computational cost and reduce the risk of over-fitting.

We investigated three types of features including site, Euclidean, and Voronoi features. The proportions of the three types of features on the top 10 list ranked by the two-step feature selection method for 5 families are presented in Fig. 3. From Fig. 3, we can find that for all families except SRC and MAPK, structural neighborhood properties (Euclidean or Voronoi) dominated the top 10 list. To be more specific, CK2 are mainly influenced by Euclidean features while Voronoi features are the most prominent features to PKA and PKC, suggesting that structural neighborhood properties are more predictive for those four families. Opposed to the former 3 families, Fig. 3 indicated that the residue-based features dominated top 10 list for SRC and MAPK. As to MAPK, considering about its large size (112) after feature selection and the poor performance in accuracy and specificity, maybe it suggests that adding its neighborhood properties as feature vectors has few benefit for phosphorylation site prediction.

Performance comparison with the state of the art approaches on the kinase-specific datasets

In this section, the proposed method (PredPhos) was benchmarked against PPSP, NetPhosK 1.0, KinasePhos 1.0 and GPS 2.1, four widely used kinase-specific

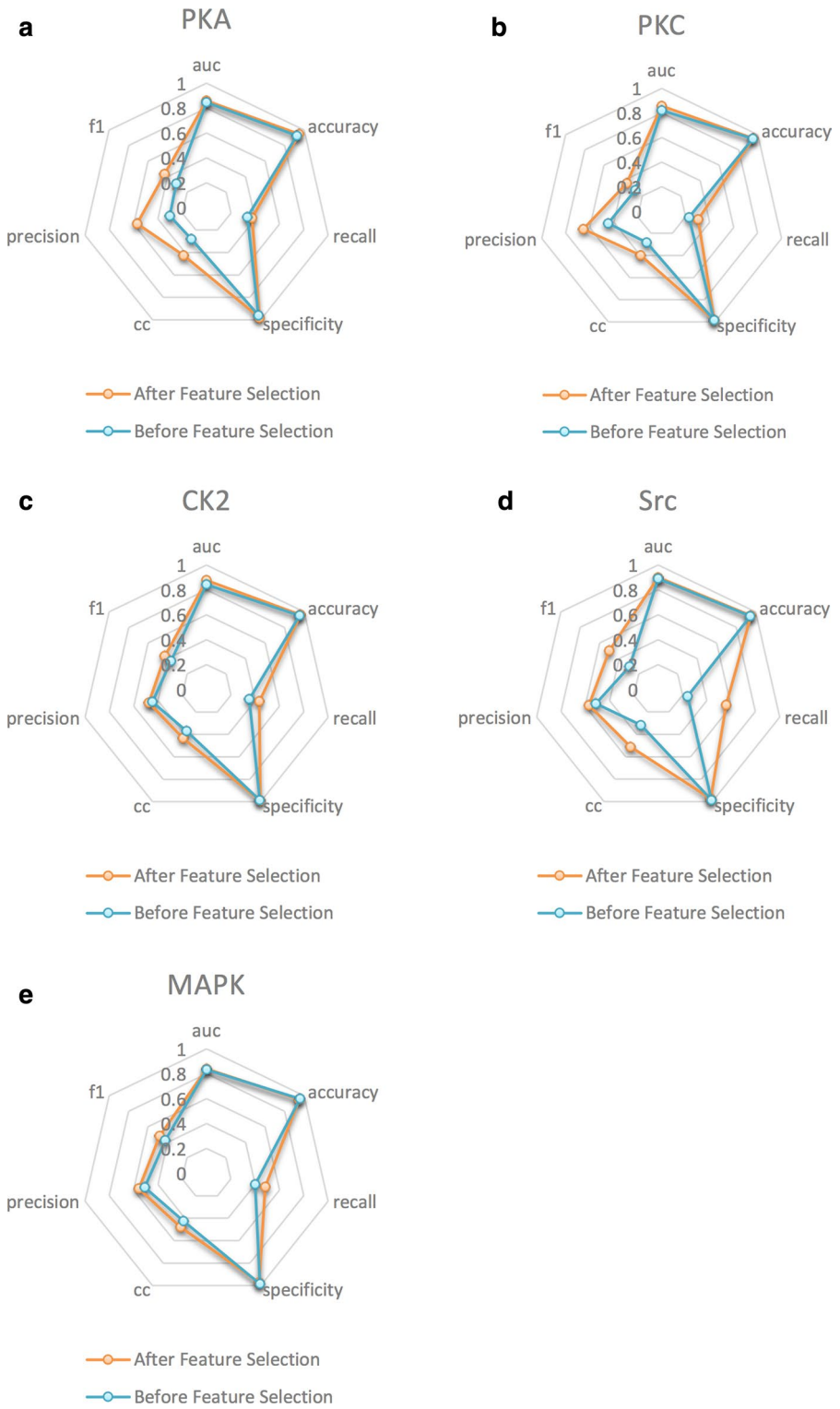

Fig. 2 Performance comparison on feature selection and non-feature selection. The performances of PKA, PKC, CK2, SRC and MAPK are shown in **a**, **b**, **c**, **d** and **e**, respectively

predictors on the independent dataset with 171, 136, 43, 274 and 149 phosphorylation sites of family of PKA, PKC, CK2, SRC and MAPK, respectively.

PPSP is implemented in an algorithm of Bayesian decision theory (BDT). KinasePhos is a predictor applying Profile Hidden Markov Model (HMM) for learning to each group of sequences surrounding to the phosphorylation residues. NetPhosK is an artificial neural network method that predicts phosphorylation sites in independent sequences with sensitivity in the range from 69 to 96 %. GPS 2.1 adopted a PK classification established by Manning et al. [5] as the standard rule to cluster the human PKs into a hierarchical structure with four levels, including group, family, subfamily and single PK. To have a performance comparison, we submitted the substrate sequence into the above tools for prediction. KinasePhos has three cut-off values for prediction specificity, as 90, 95 and 100 %. As for PPSP, NetPhosK and GPS, we also adopt different parameters to perform. We choose the best results with different parameters of other four tools to compare with PredPhos. Comparison results of the independent test are presented in Table 2.

It should be noted that without available training methods of most tools, it is nearly an impossible task to compare our predictor with the rest by running cross-validations. Benchmark test is an alternative solution to test and compare our method with others by using the same test set. However, unfair comparison may generate if our test data are included in the training set of other tools, and thus leading a fake high performance of other existing ones and underestimation of ours.

High sensitivity is beneficial when predicting phosphorylation sites in a single protein because, in wet-bench studies, experimental biologists may select some candidates from the predicted sites for further experimental design. However, a method with a high specificity is useful for whole-genome annotation. According to Table 2, the Sn values of PredPhos for the families of PKC, CK2 and SRC were 0.824, 1.000, and 0.789, respectively. PredPhos outperformed all the predictors with high Sn values of the most kinase families. Although all MCC values were not very high, the MCC values of the PredPhos results were also the best ones among other predictors. For the family of PKC, PredPhos has the highest recall (0.824), specificity (0.870), precision (0.452), mcc (0.544) and F1-score (0.583). As shown in Table 2, for the family of MAPK, although having a lower recall (0.375) than GPS (0.571), PredPhos did make a better balance between the positive dataset and negative dataset, and thus, acquired an outperformance in comprehensive strength (the sum of recall (0.375), specificity (0.986), cc (0.600) and F1-score (0.462)) compared with other prediction tools. Note that for SRC, GPS only considers about Y, while S, T and Y are all taken into account in our method, which may lead to a less-explicate prediction result compared with GPS. In any case, the prediction performance of our method is at least comparable with other kinase-specific prediction tools.

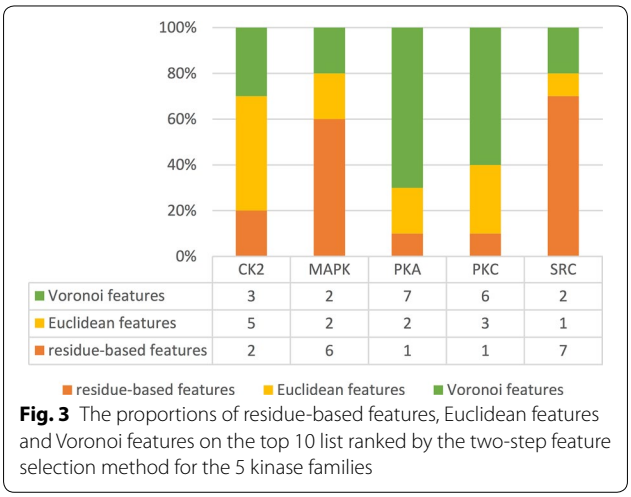

Fig. 3 The proportions of residue-based features, Euclidean features and Voronoi features on the top 10 list ranked by the two-step feature selection method for the 5 kinase families

Performance comparison on the non-kinase-specific dataset

To further evaluate the performance of the proposed PredPhos, a widely used non-kinase-specific prediction method, Netphos [21], is evaluated on the independent test set. Netphos used artificial neural networks with both sequence-based and structural-based features. We can see that our PredPhos approach substantially outperforms the Netphos method in six performance metrics (accuracy, recall, specificity, precision, CC and F1 score) (Table 3).

Conclusions

In this work, we presented a novel phosphorylation site prediction approach. Experimental results revealed that the proposed method outperformed most existing kinase-specific and non-kinase-specific prediction methods. Three

Table 3 Performance comparison on the non-kinase-specific dataset

Methods	Accuracy	Recall	Specificity	Precision	CC	F1
Netphos	0.66	0.51	0.68	0.14	0.11	0.21
PredPhos	0.77	0.60	0.82	0.38	0.23	0.45

key factors are responsible for our success. First, the wide exploitation of heterogeneous information, i.e. sequence-based, structure-based and energetic features, together with two types of structural neighborhood (Euclidian and Voronoi), provides more important clues for phosphorylation identification. A total of 153 features, including 51 site properties, 51 Euclidian neighborhood properties and 51 Voronoi neighborhood properties, have been investigated. Second, significant lower computational cost and lower risk of over-fitting was achieved by a two-step feature selection. Third, the ensemble classifiers with resampling technique alleviated the imbalanced problem and improved the prediction accuracy. A limitation of structure-based phosphorylation site prediction is that, proteins without structures can't be predicted well. However, reliable homology models of a large number of sequences can be generated on the residue level, the overall structural coverage of proteins has increased to 40 % [51].

As for the future work, major existing phosphorylation site prediction methods, including NetPhos and GPS, are considered to be integrated into the PredPhos method to further improve the prediction performance by using Bayesian networks.

Authors' contributions

YG and WH carried out the literature study, developed the new ensemble method and drafted the manuscript. JG, DL and CF participated in several independent tests. LD participated in its design and coordination. LD and ZC helped to draft and revise the manuscript. All authors read and approved the final manuscript.

Author details

[1] School of Software, Central South University, No. 22 Shaoshan South RD., Changsha 410075, China. [2] School of Electronics Engineering and Computer Science, Peking University, No. 5 Yiheyuan Road, Beijing 100871, China. [3] Shanghai Key Laboratory of Intelligent Information Processing, No. 220 Handan Road, Shanghai 200433, China.

Acknowledgements

This work was supported by National Natural Science Foundation of China under Grants No. 61309010 and No. 61379057, China Postdoctoral Science Foundation under Grant No. 2015T80886, Specialized Research Fund for the Doctoral Program of Higher Education of China under Grant No. 20130162120073 and Shanghai Key Laboratory of Intelligent Information Processing under Grant No. IIPL-2014-002.

Competing interests

The authors declare that they have no competing interests.

Declarations

The publication charges for this article were funded by National Natural Science Foundation of China under grant No. 61309010.
This article has been published as part of Journal of Biological Research—Thessaloniki, Volume 23, Supplement 1, 2016: Proceedings of the 2014 International Conference on Intelligent Computing. The full contents of the supplement are available online at http://www.jbiolres.biomedcentral.com/articles/supplements/volume-23-supplement-1.

Published: 4 July 2016

References

1. Steen H, Jebanathirajah J, Rush J, Morrice N, Kirschner M. Phosphorylation analysis by mass spectrometry: myths, facts, and the consequences for qualitative and quantitative measurements. Mol Cell Proteomics. 2006;5:172–81.
2. Schafmeier T, Haase A, Káldi K, Scholz J, Fuchs M, Brunner M. Transcriptional feedback of *Neurospora* circadian clock gene by phosphorylation-dependent inactivation of its transcription factor. Cell. 2005;122:235–46.
3. Delom F, Chevet E. Phosphoprotein analysis: from proteins to proteomes. Proteome Sci. 2006;4:15.
4. Pawson T. Specificity in signal transduction: from phosphotyrosine-sh2 domain interactions to complex cellular systems. Cell. 2004;116:191–203.
5. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S. The protein kinase complement of the human genome. Science. 2002;298:1912–34.
6. Olsen JV, Blagoev B, Gnad F, Macek B, Kumar C, Mortensen P, et al. Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. Cell. 2006;127:635–48.
7. Villén J, Beausoleil SA, Gerber SA, Gygi SP. Large-scale phosphorylation analysis of mouse liver. Proc Natl Acad Sci USA. 2007;104:1488–93.
8. Chi A, Huttenhower C, Geer LY, Coon JJ, Syka JE, Bai DL, et al. Analysis of phosphorylation sites on proteins from *Saccharomyces cerevisiae* by electron transfer dissociation (ETD) mass spectrometry. Proc Natl Acad Sci USA. 2007;104:2193–8.
9. Munton RP, Tweedie-Cullen R, Livingstone-Zatchej M, Weinandy F, Waidelich M, Longo D, et al. Qualitative and quantitative analyses of protein phosphorylation in naive and stimulated mouse synaptosomal preparations. Mol Cell Proteomics. 2007;6:283–93.
10. Sugiyama N, Nakagami H, Mochida K, Daudi A, Tomita M, Shirasu K, et al. Large-scale phosphorylation mapping reveals the extent of tyrosine phosphorylation in *Arabidopsis*. Mol Syst Biol. 2008;4:193.
11. Zhai B, Villén J, Beausoleil SA, Mintseris J, Gygi SP. Phosphoproteome analysis of *Drosophila melanogaster* embryos. J Proteome Res. 2008;7:1675–82.
12. Boersema PJ, Foong LY, Ding VM, Lemeer S, van Breukelen B, Philp R, et al. In-depth qualitative and quantitative profiling of tyrosine phosphorylation using a combination of phosphopeptide immunoaffinity purification and stable isotope dimethyl labeling. Mol Cell Proteomics. 2010;9:84–99.
13. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. 2003;31:365–70.
14. Blom N, Kreegipuu A, Brunak S. Phosphobase: a database of phosphorylation sites. Nucleic Acids Res. 1998;26:382–6.
15. Dinkel H, Chica C, Via A, Gould C, Jensen L, Gibson T, et al. Phospho.ELM: a database of phosphorylation sites–update 2011. Nucleic Acids Res. 2011;39D:261–7.
16. Bodenmiller B, Campbell D, Gerrits B, Lam H, Jovanovic M, Picotti P, et al. PhosphoPep–a database of protein phosphorylation sites in model organisms. Nat Biotechnol. 2008;26:1339–40.
17. Gnad F, Gunawardena J, Mann M. PHOSIDA 2011: the posttranslational modification database. Nucleic Acids Res. 2011;39D:253–60.
18. Zanzoni A, Carbajo D, Diella F, Gherardini P, Tramontano A, Helmer-Citterich M, et al. Phospho3D 2.0: an enhanced database of three-dimensional structures of phosphorylation sites. Nucleic Acids Res. 2011;39D:268–71.
19. Trost B, Kusalik A. Computational prediction of eukaryotic phosphorylation sites. Bioinformatics. 2011;27:2927–35.
20. Iakoucheva LM, Radivojac P, Brown CJ, O'Connor TR, Sikes JG, Obradovic Z, et al. The importance of intrinsic disorder for protein phosphorylation. Nucleic Acids Res. 2004;32:1037–49.
21. Blom N, Gammeltoft S, Brunak S. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. J Mol Biol. 1999;294:1351–62.
22. Blom N, Sicheritz-Pontén T, Gupta R, Gammeltoft S, Brunak S. Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. Proteomics. 2004;4:1633–49.
23. Huang H-D, Lee T-Y, Tzeng S-W, Horng J-T. Kinasephos: a web tool for identifying protein kinase-specific phosphorylation sites. Nucleic Acids Res. 2005;33:226–9.
24. Xue Y, Ren J, Gao X, Jin C, Wen L, Yao X. Gps2.0, a tool to predict kinase-specific phosphorylation sites in hierarchy. Mol Cell Proteomics. 2008;7:1598–608.

25. Xue Y, Li A, Wang L, Feng H, Yao X. PPSP: prediction of PK-specific phosphorylation site with Bayesian decision theory. BMC Bioinformatics. 2006;7:163.
26. Su MG, Lee TY. Incorporating substrate sequence motifs and spatial amino acid composition to identify kinase-specific phosphorylation sites on protein three-dimensional structures. BMC Bioinformatics. 2013;14:S2.
27. Yang CY, Chang CH, Yu YL, Lin TC, Lee SA, Yen CC, et al. PhosphoPOINT: a comprehensive human kinase interactome and phospho-protein database. Bioinformatics. 2008;24i:14–20.
28. Hornbeck PV, Kornhauser JM, Tkachev S, Zhang B, Skrzypek E, Murray B, et al. PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. Nucleic Acids Res. 2011;40:D261–70.
29. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI_BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389–402.
30. Huang Y, Niu B, Gao Y, Fu L, Li W. Cd-hitsuite: a webserver for clustering and comparing biological sequences. Bioinformatics. 2010;26:680–2.
31. Yu HJ, Huang DS. Normalized feature vectors. A novel alignment-free sequence comparison method based on the numbers of adjacent amino acids. IEEE/ACM Trans Comput Biol Bioinf. 2013;10:457–67.
32. Xia JF, Zhao XM, Song J, Huang DS. APIS: accurate prediction of hot spots in protein interfaces by combining protrusion index with solvent accessibility. BMC Bioinformatics. 2010;11:174.
33. Armon A, Graur D, Ben-Tal N. ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. J Mol Biol. 2001;307:447–63.
34. Obradovic Z, Peng K, Vucetic S, Radivojac P, Dunker AK. Exploiting heterogeneous sequence properties improves prediction of protein disorder. Proteins. 2005;61(suppl. 7):176–82.
35. Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z. Length-dependent prediction of protein intrinsic disorder. BMC Bioinformatics. 2006;7:208.
36. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983;22:2577–637.
37. Keskin O, Bahar I, Badretdinov AY, Ptitsyn OB, Jernigan RL. Empirical solvent-mediated potentials hold for both intra-molecular and inter-molecular inter-residue interactions. Protein Sci. 1998;7:2578–86.
38. Cho KI, Kim D, Lee D. A feature-based approach to modeling protein–protein interaction hotspots. Nucleic Acids Res. 2009;37:2672–87.
39. Assi SA, Tanaka T, Rabbitts TH, Fernandez-Fuentes N. PCRPi: presaging critical residues in protein interfaces, a new computational tool to chart hot spots in protein interfaces. Nucleic Acids Res. 2010;38:e86.
40. Deng L, Guan J, Dong Q, Zhou S. Prediction of protein-protein interaction sites using an ensemble method. BMC Bioinformatics. 2009;10:426.
41. Liang S, Grishin NV. Effective scoring function for protein sequence design. Proteins. 2004;54:271–81.
42. Chan CH, Liang HK, Hsiao NW, Ko MT, Lyu PC, Hwang JK. Relationship between local structural entropy and protein thermostability. Proteins. 2004;57:684–91.
43. Zimmer R, Wöhler M, Thiele R. New scoring schemes for protein fold recognition based on Voronoi contacts. Bioinformatics. 1998;14:295–308.
44. Zhang QC, Deng L, Fisher M, Guan J, Honig B, Petrey D. PredUs: a web server for predicting protein interfaces using structural neighbors. Nucleic Acids Res. 2011;39W:283–7.
45. Yang AS, Honig B. An integrated approach to the analysis and modeling of protein sequences and structures. I. protein structural alignment and a quantitative measure for protein structural distance. J Mol Biol. 2000;301:665–78.
46. Zhang QC, Petrey D, Norel R, Honig BH. Protein interface conservation across structure space. Proc Natl Acad Sci USA. 2010;107:10896–901.
47. Deng L, Guan J, Wei X, Yi Y, Zhang Q, Zhou S. Boosting prediction performance of protein-protein interaction hot spots by using structural neighborhood properties. J Comput Biol. 2013;20:878–91.
48. Barber CB, Dobkin DP, Huhdanpaa H. The quickhull algorithm for convex hulls. ACM Trans Math Softw. 1996;22(4):469–83.
49. Liaw A, Wiener M. Classification and regression by random forest. R News. 2002;2:18–22.
50. Liu KH, Huang DS. Cancer classification using rotation forest. Comput Biol Med. 2008;38:601–10.
51. Khafizov K, Madrid-Aliste C, Almo SC, Fiser A. Trends in structural coverage of the protein universe and the impact of the Protein Structure Initiative. Proc Natl Acad Sci USA. 2014;111:3733–8.