Chemometrics and Intelligent Laboratory Systems 136 (2014) 97–103

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Effective identification of kinase-specific phosphorylation sites based on domain–domain interactions

Yun Zhong, Yanzhi Guo *, Jiesi Luo, Xuemei Pu, Menglong Li *

College of Chemistry, Sichuan University, Chengdu 610064, PR China

ARTICLE INFO

Article history:
Received 18 January 2014
Received in revised form 16 May 2014
Accepted 24 May 2014
Available online 29 May 2014

Keywords:
Kinase-specific phosphorylation site
Distinct residues
Domain–domain interaction (DDI)
Random forest (RF)
Difference analysis

ABSTRACT

Deciphering interactions between protein kinases (PKs) and the target substrates are fundamental for understanding the molecular mechanisms of phosphorylation. Although all PKs have been identified in eukaryotes, the sites that they phosphorylate are only partially elucidated. Experimental identification of phosphorylation sites is labor and resource intensive, so developing an effective method to computationally predict potential sites is increasingly important. Here, a novel method was proposed for the identification of kinase-specific phosphorylation sites based on domain–domain interactions (DDIs). Using difference analysis between phosphorylation sites and non-phosphorylation sites, the distinct neighbor residues around the phosphorylation sites were firstly identified in our study. The results of difference analysis by rank sum test indicate that 19, 26, 26 and 10 neighbor residues are distinctive for the phosphorylation site prediction of four major serine (S)/threonine (T) protein kinase families—CDK, CK2, PKA and PKC respectively. Then the correlation coefficients were computed to represent the interaction between PK domains and phosphorylation domains of the substrate proteins. Four random forest models (RF) were constructed to predict the potential sites, the CDK, CK2, PKA and PKC RF models yield an accuracy of 86.57%, 91.44%, 87.02% and 80.11% on the test sets respectively. Finally, the new substrate proteins in protein data bank (PDB) were extracted to verify the distinct residues around the phosphorylation sites at 3D-structural level and the results further demonstrate the reliability of our models, which indicate that our method will be a useful tool for elucidating dynamic interactions between PKs and their substrates.

1. Introduction

Phosphorylation is one of the most important post-translational modifications by adding a phosphate group (PO_4) to serine (S), threonine (T) or tyrosine (Y) in substrate proteins catalyzed by protein kinases (PKs) [1]. It regulates all aspects of biological process, including cell cycle [2], DNA repair [3], regulation of transcription [4], cellular motility [5], and so on. It has been proven that in human genome, about 2% of the genes are responsible for encoding protein kinases, while nearly 50% of them are linked to many diseases, particularly cancer [6]. Proteins that are phosphorylated often undergo biochemical changes that further affect the relative biological pathways [7]. In eukaryotic cells, 30%–50% proteins undergo phosphorylation [8]. While most or all PKs have been identified, the sites they phosphorylate have been just partially elucidated. Therefore, the identification of phosphorylation sites, especially kinase-specific phosphorylation sites is essential for understanding the functions of PKs and the molecular mechanisms of phosphorylation.

Experimental methods have been used for the identification of novel phosphorylation sites, such as low-throughput biological technique based on site-directed mutagenesis [9] and high-throughput technique of mass spectrometry [10]. However these methods are time consuming and expensive to perform, so it is of great significance to propose effective computational method to predict the potential sites. Several computational methods have been proposed to predict the phosphorylation sites. Originally, the so-called non-specific tool of NetPhos [11] was developed, regardless of the organism information. In fact, there may be different phosphorylation patterns in substrates of different organisms, so organism-specific methods were proposed including PhosPhat 3.0 [12], NetPhosBac [13] and NetPhosYeast [14]. Usually, one PK only selectively phosphorylates some of the substrates by recognizing the sequence or structural profiles around phosphorylation sites [15], so kinase-specific prediction tools have become increasingly popular, such as ScanSite [16] and KinasePhos [17,18]. At the same time, web servers on line to predict kinase-specific phosphorylation sites are available including GPS [19,20], PredPhospho [21] and PhoScan [22]. Recently, a known and predicted functional database about post-translational modifications (PTMs), PTMcode [23] was constructed to provide more valid data for computational methods.

Since these kinase-specific methods have obtained good performance, but they only take the phosphorylation domain information of substrates into account. It is known that the specific residues at certain positions are targeted by the catalytic domain of a particular kinase, so

* Corresponding authors. Tel.: +86 28 85413330; fax: +86 28 85412356.
E-mail addresses: yzguo@scu.edu.cn (Y. Guo), liml@scu.edu.cn (M. Li).

http://dx.doi.org/10.1016/j.chemolab.2014.05.009
0169-7439/© 2014 Elsevier B.V. All rights reserved.

we can expect that the performance of the phosphorylation site prediction would be further improved if the catalytic domain information is also considered. Moreover, the neighbor residues have been shown to be very important for the phosphorylation of one site, but the conversation varies from site to site and some residues with little conservation have no contributions to the site identification. So the neighbor residues that are distinctive for the site prediction need to be identified.

In this paper, we made a first attempt to identify the neighbor residues that are distinct for phosphorylation site and a new kinase-specific prediction method was proposed based on domain–domain interactions. Using difference analysis by rank sum test, out of 50 neighbor residues from position −25 to 25, 19, 26, 26 and 10 neighbors proved to be distinctive for phosphorylation site prediction of CDK, CK2, PKA and PKC respectively. Through computing correlation coefficients between the residues in PK domain and the distinct residues in the phosphorylation domain, the correlation vector was obtained for representing the DDI information. So the final RF models were constructed for the four kinase families and the prediction accuracy was higher than 80%. The promising prediction results on the testing dataset and the independent dataset demonstrate that our method will be useful in identifying the novel phosphorylation sites of different kinase families.

2. Material and methods

2.1. Dataset preparation

The most common S/T kinases are found in the four families: CDK, CK2, PKA and PKC. Kinases in these four families are responsible for about half of the known S/T kinase reactions taking place in eukaryotic organisms [24]. The information of other kinase families is limited, so only the four major kinase families were considered in this paper. The experimentally verified kinase–substrate interaction data are from the two public databases, PhosphoPIONT [25] and PhosphoSitePlus [26]. From PhosphoPOINT, category 4 included 1911 PPIs and 1280 PPIs were extracted from PhosphoSitePlus. After further verifying them in HPRD (Human Protein Reference Database), there are 556, 258, 264 and 460 PPIs for the four major S/T kinase families of CDK, CK2, PKA and PKC respectively according to the kinase name index table in KinBase and Kinomer [27].

Proteins consist of one or multiple domains thought as functional units of protein and interactions between proteins typically involve binding between specific domains [28]. Therefore, DDIs can be key supporting evidences for protein interaction mechanisms. In our paper, we used DDI to represent the interaction between a kinase and its substrate. So a DDI denotes the interaction between a PK domain and a phosphorylation domain. The data of phosphorylation domains of substrates were retrieved from the phosphoELM version 9.0 [29] and Uniprot release 2013_12. Since the kinase-specific phosphorylation sites can not be found in many substrates, there are only 390 sites for CDK, 272 for CK2, 201 for PKA and 380 for PKC, respectively. In order to investigate the residues surrounding the phosphorylation sites, the sequence fragments were extracted by a window size of 51 centered on S/T. The window size consists of 51 residues placed from position −25 to 25. Fragments with a phosphorylated S/T on position 0 are deemed as the phosphorylation domains (positive data) while those centered on non-phosphorylated S/T are the negative data. The PK domain is a structurally conserved domain containing the catalytic function of PKs [30–32]. In Pfam database [33], the PK domain information for each kinase has been available. So a positive DDI is composed of a PK domain and the target phosphorylation domain and a negative DDI includes a PK domain and the non-phosphorylation domain.

In order to avoid homology bias, the commonly used multiple sequence alignment tool of CD-HIT program [34] was used to remove redundancy. Using a low sequence identity threshold of 30%, CD-HIT program was performed on the phosphorylation domains and non-phosphorylation domains respectively. So, the four datasets of DK, CK2, PKA and PKC include 180 positive and 1765 negative DDIs, 121 positive and 1067 negative DDIs, 162 positive and 1927 negative DDIs, and 169 positive and 1773 negative DDIs, respectively. Considering the very large amount of negative samples, the size of the negative set was set to be 1.5 times that of the positive set. The final, balanced datasets for CDK, CK2, PKA and PKC consist of 180 positive and 270 negative samples, 121 positive and 182 negative samples, 162 positive and 243 negative samples and 169 positive and 254 negative samples, respectively.

2.2. Feature extraction

It is important to describe protein sequence quantitatively. PPIs can be defined as four interaction modes: electrostatic interaction, steric interaction, hydrophobic interaction and hydrogen bond [35,36]. According to the four interaction modes, four feature groups including 31 electronic properties, 248 steric properties, 110 hydrophobic properties and 6 hydrogen bond properties were manually selected from 544 natural amino acid properties from AAindex [37]. The original features are listed in Supplementary Table S1.

First, the original properties were normalized to zero mean and unit standard deviation (SD) to eliminate the unit difference. Considering the redundancy of features, the properties that have >90% correlation identity to one another were removed in each group. As a result, four original variable matrices were obtained, including 23 electronic properties, 134 steric properties, 32 hydrophobic properties and 5 hydrogen bond properties. Then the four variable matrices were processed by principal component analysis (PCA) respectively. Accounting for ≥90% variance of the original information, the top 7, 10, 5 and 3 significant principal components were obtained for electrostatic, steric, hydrophobic and hydrogen bond properties, respectively. Multiplying the scoring coefficient of each principal component by the original features, 25 significant principal component scores were yielded as a new amino acid descriptor on behalf of the original data and the information loss was insignificant (<10%).

2.3. Correlation vectors for DDIs

The interaction information of the two interacting proteins is generally characterized by the relationship between the residues in the two different domains. So Chou's work [38], we used correlation coefficients (CCs) to numerically represent the interactions between residues in PK domains and phosphorylation domains. Given a phosphorylation domain with 51 residues and a PK domain with N' residues, CC variables can be calculated by the following equation:

$$\left.\begin{array}{ll} CC_{ij} = \dfrac{\sum_{k=1}^{N'} \left(P_{ij} - P'_{kj}\right)^2}{N'} & (1) \\ Vcc = \left[CC_{ij}\right]^T & (2) \end{array}\right\} \qquad \begin{array}{l} i \in [-25, 25] \\ j \in [1, 25] \end{array}$$

where P_{ij} represents the value of j-th property for i-th residue in a phosphorylation domain and P'_{kj} is the value of j-th property for k-th residue in a PK domain. ^{T} denotes the matrix transposition. In this way, each DDI was converted into a numerical vector and the dimension is the 51∗25.

2.3. Model construction and evaluation

As a good classification and regression method [39,40], Random forest (RF) has been successfully used in many fields [e.g. 41–43]. Also it has proven to perform well on the prediction of kinase–substrate interactions [44]. So in this paper, we used RF to construct the model for predicting kinase-specific phosphorylation sites. In this study, the RF algorithm was implemented by the RF package in *R* language.