Journal of Molecular Graphics and Modelling 56 (2015) 84–90

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

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

Using support vector machines to identify protein phosphorylation sites in viruses

Shu-Yun Huang[a], Shao-Ping Shi[b,c], Jian-Ding Qiu[a,b,*], Ming-Chu Liu[a,*]

[a] Department of Chemical Engineering, Pingxiang College, Pingxiang 337055, China
[b] Department of Chemistry, Nanchang University, Nanchang 330031, China
[c] Department of Mathematics, Nanchang University, Nanchang 330031, China

ARTICLE INFO

Article history:
Accepted 16 December 2014
Available online 24 December 2014

Keywords:
Phosphorylation site
Virus proteins
Support vector machine
Encoding scheme based on attribute grouping
Position weight amino acid composition

ABSTRACT

Phosphorylation of viral proteins plays important roles in enhancing replication and inhibition of normal host-cell functions. Given its importance in biology, a unique opportunity has arisen to identify viral protein phosphorylation sites. However, experimental methods for identifying phosphorylation sites are resource intensive. Hence, there is significant interest in developing computational methods for reliable prediction of viral phosphorylation sites from amino acid sequences. In this study, a new method based on support vector machine is proposed to identify protein phosphorylation sites in viruses. We apply an encoding scheme based on attribute grouping and position weight amino acid composition to extract physicochemical properties and sequence information of viral proteins around phosphorylation sites. By 10-fold cross-validation, the prediction accuracies for phosphoserine, phosphothreonine and phosphotyrosine with window size of 23 are 88.8%, 95.2% and 97.1%, respectively. Furthermore, compared with the existing methods of Musite and MDD-clustered HMMs, the high sensitivity and accuracy of our presented method demonstrate the predictive effectiveness of the identified phosphorylation sites for viral proteins.

1. Introduction

Protein phosphorylation is a ubiquitous post-translational modification (PTM) that controls a number of intracellular processes. It has been estimated that at least one-third of the cellular proteins are modified by phosphorylation [1]. In eukaryotic cells, phosphorylation occurs almost exclusively on serine, threonine or tyrosine residues [2]. Also for viruses, including vesicular stomatitis virus, human immunodeficiency virus type 1 (HIV-1), mosaic virus, and H1N1 influenza virus, protein phosphorylation has been shown to regulate vital processes such as virus transcription and replication, RNA binding activity, and virus assembly [3–7]. For instance, Polo-like kinase 1 (Plk1) can phosphorylate cyclin T1 at Ser564 and inhibit the kinase activity of cyclin T1/Cdk9 complex on phosphorylation of the C-terminal domain (CTD) of RNA polymerase II [8]. Hsiang et al. demonstrated that the only serine 42 (S42) phosphorylation of the NS1 protein catalyzed by protein kinase Cα(PKCα) regulated human influenza A virus replication [9]. Cheng et al. identified membrane-associated serine/threonine kinase-like protein from Nicotiana benthamiana involved in the cell-to-cell movement of Bamboo mosaic virus (BaMV) [10]. Therefore, investigating virus phosphorylation sites can provide useful clues for drug design and the treatment of various viral infections.

Phosphorylation site identification is usually experimentally determined by mass spectrometry-based techniques [11]. This has led to the establishment of several databases of phosphorylation sites, such as 'the Phosphorylation Site Database' [12], 'Phospho.ELM' [13], 'Phosphosite' [14], and 'PhosPhAT' [15]. While useful, mass spectrometry requires very expensive instruments and specialized expertise that are not available in typical laboratories [16]. At the same time, the identification of kinase specificity rules with mass spectrometry still remains a relatively slow and often inefficacious task. Thus, various computational methods for identifying protein phosphorylation sites have been proposed, including artificial neural networks (ANNs) [17,18], hidden Markov models (HMMs) [19,20], position-specific scoring matrices (PSSMs) [21–23], support vector machines (SVMs) [24–27], and more details can be found in recent reviews [28,29].

In virus phosphorylation prediction, Schwartz and Church used the scan-*x* tool to identify 329 phosphorylation sites in proteins from 52 human viruses [30]. However, it has not investigated the

* Corresponding author at: Department of Chemistry, Nanchang University, Nanchang 330031, China. Tel.: +86 791 83969518.
E-mail address: jdqiu@ncu.edu.cn (J.-D. Qiu).

http://dx.doi.org/10.1016/j.jmgm.2014.12.005
1093-3263/

various substrate motifs for viral protein phosphorylation sites [31]. More recently, Bretaña et al. employed maximal dependence decomposition (MDD) to investigate kinase substrate specificities in viral protein phosphorylation sites [31]. Although, the average accuracies of serine and threonine using the MDD-clustered HMMs were 84.93% and 78.05%, respectively, the number of phosphorylated serine sites was only 233, and 54 for phosphothreonine sites. As we all know, a small number of training set may be over-fitting. Hence, there is a need to develop a computational method in identifying enormous amount of viral protein phosphorylation data by selecting more informative feature descriptors.

In this paper, we presented a new approach to predict viral phosphorylation sites based on support vector machine. Physicochemical properties of amino acids and position weight amino acid compositions were utilized to extract sequence features of virus proteins. Our current work contained the following contents: (1) two types of features were analyzed, (2) SVM was employed to deal with the problem of binary classification, (3) ten-fold cross-validation method was chosen to evaluate the performance of SVM classifier, (4) the effect of window length was investigated, and (5) the independent testing data was used to compare with the existing models.

2. Materials and methods

2.1. Data collection and statistics

All training datasets were extracted from the NCBI RefSeq and the Plant Protein Phosphorylation Database (P^3DB) [32] databases, as presented in Fig. 1. Firstly, we obtained 327 proteins with 2793 experimental phosphorylation sites by searching information containing "phosphorylation" and "virus" from the NCBI RefSeq. The P^3DB is one of the most significant in vivo data resources for studying plant phosphoproteomics. According to the keyword of virus, we obtained 363 proteins covering 1274 experimental phosphorylation sites from the P^3DB. Secondly, the sliding window strategy was used to extract positive and negative datasets from protein sequences, which were represented by peptide sequences with serine, threonine and tyrosine symmetrically surrounded by flanking residues. If the candidate phosphorylation sites were near the N- or C-terminus, we used the letter "O" instead of the absent letters. We respectively designated peptide sequences of experimentally validated phosphoserine, phosphothreonine and phosphotyrosine as positive datasets. It would be difficult to prove definitively that a particular serine/threonine/tyrosine residue is not phosphorylated under any conditions. Almost all of researchers of phosphorylation prediction made the assumption that any serine/threonine/tyrosine residue that is not marked by any phosphorylation information on the same protein is a non-phosphorylated site [25,31,33]. Besides, Radivojac et al. have concluded that the choosing of negative samples upon this assumption did not significantly influence prediction performance through comparing with that of using the validated negative samples [34]. So we adopt this assumption that negative samples were the serine/threonine/tyrosine residues that were not marked by any phosphorylation information on the same proteins, the rational of which is that the resulting negative samples are more likely to be non-phosphorylation sites than those obtained by random as these proteins were experimentally investigated. Although not all these sites are necessarily true negatives, it is reasonable to believe that a large majority of them are [35]. Moreover, the redundancy reducing process was also carried out on training datasets. For example, for two phosphorylated serine peptide sequences with 100% identity, when the phosphoserine sites in the two proteins were in the same positions, only one was kept. After strictly following the above procedures, we attained 2444 high quality positive sites for phosphoserine, 635 positive sites for phosphothreonine, and 268 positive sites for phosphotyrosine, as shown in Supplementary materials (see Tables S1–S3).

Meanwhile, in order to further evaluate the performance of our method and compare it with existing methods, an independent testing set was extracted from the viral posttranslational modification (virPTM) database (http://virptm.hms.harvard.edu/), which includes 230 phosphoserine sites and 2494 non-phosphorylated serine sites, 61 phosphothreonine sites and 1211 non-phosphorylated threonine sites, 14 phosphotyrosine sites and 57 non-phosphorylated tyrosine sites from 111 human virus proteins shown in Fig. 1. Finally, the ratio of positive and negative samples was 1:1 and three negative training sets were obtained by randomly extracting from the negative datasets, with expectation to ensure unbiased and objective results.

2.2. Feature encoding

2.2.1. Encoding based on attribute grouping

Previously, Fan and Zhang have detected that the serine and threonine acceptor site microenvironment is depleted in nonpolar and hydrophobic amino acids. Whereas the tyrosine acceptor site microenvironment is characterized by only one enriched property, namely the charge, and is depleted in cysteine (C) and proline (P), which are neutral residues [36].

Thus, we adopted an encoding scheme of protein sequences considering the hydrophobicity and charged character of amino acid residues. The encoding method based on attribute grouping (named as EBAG) divides the 20 amino acid residues into four different classes according to their physicochemical property: the hydrophobic group C1 = [A, F,G, I, L, M, P, V, W], the polar group C2 = [C, N, Q, S, T, Y], the acidic group C3 = [D, E], and the basic group C4 = [H, K, R] [37,38].

Given a protein sequence p fragment with $2L+1$ amino acid residues, we used the above classification to transform it into four binary sequences as follows:

$$\begin{aligned} &H1_p(j) = 1 \quad \text{if } p(j) \in C1 \quad \text{else } H1_p(j) = 0 \\ &H2_p(j) = 1 \quad \text{if } p(j) \in C2 \quad \text{else } H2_p(j) = 0 \\ &H3_p(j) = 1 \quad \text{if } p(j) \in C3 \quad \text{else } H3_p(j) = 0 \quad j = -L, \ldots, L \\ &H4_p(j) = 1 \quad \text{if } p(j) \in C4 \quad \text{else } H4_p(j) = 0 \end{aligned} \tag{1}$$

2.2.2. Position weight amino acid composition

To reveal the sequence-order information around phosphorylation sites, we used position weight amino acids composition (PWAA) to extract the sequence position information of amino acid residues. Given an amino acid residue a_i ($i = 1, 2, \ldots, 20$), we can express the position information of amino acid a_i in the protein sequence fragment p with $2L+1$ amino acids by following formula:

$$C_i = \frac{1}{L(L+1)} \sum_{j=-L}^{L} x_{i,j} \left(j + \frac{|j|}{L} \right), \quad j = -L, \ldots, L \tag{2}$$

where L denotes the number of upstream residues or downstream residues from the central site in the protein sequence fragment p, $x_{i,j} = 1$ if a_i is the jth position residue in protein sequence fragment p, otherwise $x_{i,j} = 0$. Finally, a protein sequence fragment p is defined as 20 dimension feature vectors.

2.3. Model learning and evaluation

SVM is a supervised learning method for classification and regression designed by Cortes and Vapnik [39]. The principle of