Journal of Theoretical Biology 344 (2014) 78–87

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Prediction of posttranslational modification sites from amino acid sequences with kernel methods

Yan Xu [a], Xiaobo Wang [b], Yongcui Wang [b], Yingjie Tian [c], Xiaojian Shao [d], Ling-Yun Wu [e], Naiyang Deng [b,*]

[a] Department of Information and Computer Science, University of Science and Technology Beijing, Beijing 100083, China
[b] Department of Applied Mathematics, College of Science, China Agricultural University, Beijing 10083, China
[c] Research Center on Fictitious Economy and Data Science, Chinese Academy of Sciences, Beijing 100190, China
[d] Department of Mathematics and Information Science, BinZhou University, BinZhou 256603, China
[e] Institute of Applied Mathematics, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

HIGHLIGHTS

- In this paper, a novel encoding method PSPM (position-specific propensity matrices) is developed.
- Then a support vector machine (SVM) with the kernel matrix computed by PSPM is applied to predict the PTM sites.
- The prediction software can be freely downloaded from http://www.aporc.org/doc/wiki/PTMPred.

ARTICLE INFO

Article history:
Received 28 June 2013
Received in revised form
13 September 2013
Accepted 16 November 2013
Available online 27 November 2013

Keywords:
Kinase-specific
O-glycosylation
Phosphorylation
Support vector machine

ABSTRACT

Post-translational modification (PTM) is the chemical modification of a protein after its translation and one of the later steps in protein biosynthesis for many proteins. It plays an important role which modifies the end product of gene expression and contributes to biological processes and diseased conditions. However, the experimental methods for identifying PTM sites are both costly and time-consuming. Hence computational methods are highly desired. In this work, a novel encoding method PSPM (position-specific propensity matrices) is developed. Then a support vector machine (SVM) with the kernel matrix computed by PSPM is applied to predict the PTM sites. The experimental results indicate that the performance of new method is better or comparable with the existing methods. Therefore, the new method is a useful computational resource for the identification of PTM sites. A unified standalone software PTMPred is developed. It can be used to predict all types of PTM sites if the user provides the training datasets. The software can be freely downloaded from http://www.aporc.org/doc/wiki/PTMPred.

1. Introduction

The posttranslational modification (PTM) of proteins is a common biological mechanism which regulates protein functions. The PTM sites are covalent processing events that change the properties of a protein by proteolytic cleavage or adding a modifying group to one or more amino acids (Mann and Jensen, 2003) and occur on almost proteins (Blom et al., 2004). Protein PTM sites can determine proteins' activity state, localization, turnover, and interactions with other proteins. For example, kinase cascades are turned on or off by the reversible adding or removing phosphate groups in signaling. Ubiquitination marks cyclins for destruction at defined time points in the cell cycle.

The study of PTM sites has been restricted due to lack of suitable methods. Some high-throughput experimental technologies, including mass spectrum (Kraft et al., 2003), peptide microarray (Rychlewski et al., 2004), and phosphor-specific proteolysis (Knight et al., 2003) have been applied to study PTM sites. However, such methods are usually expensive and time-consuming. A vital question is that the functions of proteins may be hampered or altered outside the living organisms (Wang et al., 2008). For these reasons, the computational methods to predict PTM sites are urgently needed.

Many studies have indicated that sequence-based prediction approaches, such as protein subcellular location prediction (Chou and Elrod, 1999; Chou and Cai, 2002), identification of membrane proteins and their types (Cai et al., 2003; Chou and Shen, 2007a),

* Corresponding author. Tel.: +86 1062736265.
E-mail address: dengnaiyang@cau.edu.cn (N. Deng).

0022-5193/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jtbi.2013.11.012

identification of enzymes and their functional classes (Cai and Chou, 2005; Chou and Cai, 2004), identification of GPCR (G-protein-coupled receptor) and their types (Xiao et al., 2009; Lin et al., 2009), protein cleavage site prediction (Chou and Shen, 2009; Chou, 1993, 1996), signal peptide prediction (Chou and Shen, 2009, 2007b), and protein 3D structure prediction based on sequence alignment (Chou, 2004), can timely provide very useful information and insights for both basic research and drug design.

The support vector machine (SVM) has been efficiently applied into the pattern recognition problems in the fields of computational biology and bioinformatics including predicting protein subcellular location (Chou and Cai, 2002), membrane protein type (Cai et al., 2003), posttranslational modification (Xu et al., 2013), GalNAc-transferase (Chou, 1995) and so on. SVM also has good performances in the prediction of PTM sites. Kim et al. (2004) first used SVM with the standard binary encoding scheme to predict phosphorylation sites and obtained better outcomes than all previous methods. Wong et al. (2007) applied SVM based on coupling patterns to predict kinase-specific phosphorylation sites. Shao et al. (2009a) used SVM with bi-profile Bayes feature to predict methylation sites and Chang et al. (2009) applied SVM combined with structural features to predict protein tyrosine sulfation sites. Gao et al. (2010) used SVM to predict general and kinase-specific phosphorylation sites. Lately, Zhao et al. (2012) predicted protein phosphorylation sites via SVM by using the composition of k-spaced amino acid pairs. Although many advanced methods have been exploited in the prof PTM sites, there are some room to improve the accuracy.

In the theme of using machine learning methods to predict PTM sites, the encoding scheme (i.e. the construction of input feature vectors) is very important. In this work, two encoding schemes are introduced for predicting PTM sites, one is PSPM (position-specific propensity matrices) which is developed by us, the other is constructed by Tang et al. (2007) and named as PSAAP (position-specific amino acid propensity). The proposed encoding scheme PSPM characterizes the position-specific amino acid pairs propensity surrounding the potential PTM sites and PSAAP reflects amino acids in different positions surrounding the corresponding PTM sites. In addition, PSPM considers the sequence-order effects which would influence the PTM sites, however PSAAP ignores this important situation. The novel algorithms are mainly applied in two reversible PTM sites, phosphorylation and O-glycosylation. The results show that the performance of the new algorithms is better than or comparable with previous methods. Take into account the performance of the two encoding schemes, PSPM is better than PSAAP. The dimensions of PSPM and PSAAP are absolutely low comparing with other encoding schemes such as conventional binary encoding and coupling pattern. To facilitate the biological community, a executable software "PTMPred" is developed, which can be used for proteome-wide PTM sites prediction. More and more PTM sites will be confirmed experimentally, our "PTMPred" allows the users to submit their own training datasets to obtain different predictors. In order to examine the behavior of our PTMPred for general PTM sites, it is used to predict sulfated proteins. The PTMPred obtain good performance.

As pointed out by a comprehensive review (Chou, 2011a) and demonstrated by a series of recent publications (Chen et al., 2013, 2012b; Xu et al., 2013), to develop a useful statistical predictor for a protein or peptide system, we need to engage in the following procedures: (i) construct or select a valid benchmark dataset to train and test the predictor; (ii) formulate the protein or peptide samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted; (iii) introduce or develop a powerful algorithm (or engine) to operate the prediction; (iv) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor; (v) establish a user-friendly web-server for the predictor that is accessible to the public. Below, let us describe how to deal with these steps one by one. Fig. 1 is the flow chart of the PTM sites prediction.

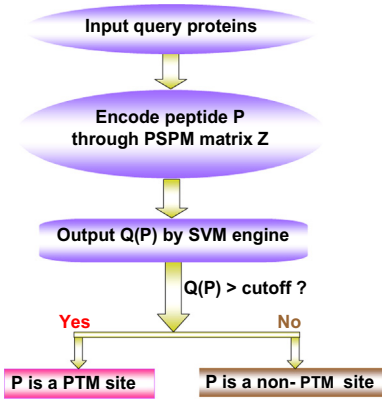

Fig. 1. System flow chart of the PTM site prediction.

2. Methods

2.1. PSPM encoding scheme

To successfully use SVM as a powerful classifier, the key is how to effectively define a feature-vector to formulate the statistical samples concerned. According to Eq. (6) of Xu et al. (2013), the feature vector for any protein, peptide, or biological sequence is none but a general form of pseudo amino acid composition (Chou, 2011a, 2005) can be formulated as

$$P=[\psi_1\ \ \psi_2\ \ \ldots\psi_u\ \ \ldots\psi_\Omega]^{\mathrm{T}} \tag{1}$$

where T is a transpose operator, the components $\psi_1,\psi_2,\ldots$ will depend on how to extract the desired information from the statistical samples concerned, while the subscript Ω is an integer representing the dimension of the feature vector P. Below, let us describe how to extract useful information from the training dataset to define the peptide samples concerned via Eq. (1).

A new feature encoding scheme called position-specific propensity matrix (PSPM), which uses position-specific amino acid pairs to construct input features, is proposed in this section. The feature vector P is based on a propensity matrix Z which is constructed as follows:

- In the first step, considering 21 types of amino acids (20 native and one dummy amino acid X), there are 441 kinds of dipeptide. Suppose that the positive training dataset M_{pos} consists of l positive sequence fragments and the length of every sequence fragment is m. The value of m is empirically determined. In this work, we set the m as 13 for phosphorylation site prediction and 41 for Mucin-type O-glycosylation prediction after some trials (see Tables 1 and 5). A position specific dipeptide composition matrix A^+ is constructed based on M_{pos}. The j-th column of A^+ is defined as follows:

$$A_j^+=(a_{1,j}^+,a_{2,j}^+,\ldots,a_{441,j}^+)^T,\quad j=1,2,\ldots,m-1, \tag{2}$$

  where $a_{i,j}^+$ represents the frequency of i-th dipeptide in the j-th position of sequence fragments in M_{pos}. For example, $a_{1,j}^+$ denotes the frequency of the first dipeptide (AA) in the j-th position. Obviously, the size of the matrix A^+ is $441\times(m-1)$.
- In the second step, similarly, the position specific dipeptide composition matrix A^- can be computed for the negative training dataset M_{neg}. However, there are much more