Neurocomputing ∎ (∎∎∎∎) ∎∎∎–∎∎∎

Contents lists available at ScienceDirect

Neurocomputing

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

Learning from real imbalanced data of 14-3-3 proteins binding specificity

Zhao Li [a], Jijun Tang [a,b], Fei Guo [a,*]

[a] School of Computer Science and Technology, Tianjin University, 92 Weijin Road, Nankai District, Tianjin, PR China
[b] School of Computational Science and Engineering, University of South Carolina, Columbia, USA

ARTICLE INFO

Article history:
Received 31 October 2015
Received in revised form
27 February 2016
Accepted 1 March 2016

Keywords:
Similarity-based undersampling
SMOTE-like oversampling
Locally weighted regression
Physicochemical property
Auto-cross covariance
14-3-3 isoforms

ABSTRACT

The 14-3-3 proteins are a highly conserved family of homodimeric and heterodimeric molecules, expressed in all eukaryotic cells. In human cells, this family consists of seven distinct but highly homologous 14-3-3 isoforms. 14-3-3σ is the only isoform directly linked to cancer in epithelial cells, which is regulated by major tumor suppressor gene. For each 14-3-3 isoform, we have 1000 peptide motifs with experimental binding affinity values. In this paper, we present a novel method for identifying peptide motifs binding to 14-3-3σ isoform. First, we select nine physicochemical properties of amino acids to describe each peptide motif. We also use auto-cross covariance to extract correlative properties of amino acids in any two positions. Then, a similarity-based undersampling approach and a SMOTE-like oversampling approach are used to deal with imbalanced distribution of the known peptide motifs. Finally, we consider locally weighted regression to predict affinity values of peptide motifs, which combines the simplicity of linear least squares regression with the flexibility of nonlinear regression. Our method tests on the 1000 peptide motifs binding to seven 14-3-3 isoforms. On the 14-3-3σ isoform, our method has overall Pearson–product–moment correlation coefficient (PCC) and the root mean squared error (RMSE) values of 0.83 and 258.31 for *N*-terminal sublibrary, and 0.80 and 250.89 for *C*-terminal sublibrary, respectively. We identify phosphopeptides that preferentially bind to 14-3-3σ over other isoforms. Several positions on peptide motifs have the same amino acid as experimental substrate specificity of phosphopeptides binding to 14-3-3σ. Our method is a fast and reliable computational method that can be used in peptide–protein binding identification in proteomics research.

1. Introduction

The 14-3-3 proteins are a highly conserved family of molecules, which includes seven distinct but highly homologous 14-3-3 isoforms: $\beta, \epsilon, \eta, \gamma, \sigma, \tau, \zeta$ [1,2]. Phosphate can bind to all of the 14-3-3 proteins and therefore being present at high intracellular concentration [3,4]. As a key regulator of signal transduction, 14-3-3 isoforms participate in important cellular events including regulation of apoptosis, adhesion-dependent integrin signaling, cell cycle control, DNA damage, metabolism and transcriptional regulation. We have been particularly interested in understanding roles of different 14-3-3 isoforms in cell proliferation, cell cycle control, and human tumorigenesis. With roles of different 14-3-3 isoforms in a wide variety of signal transduction processes, 14-3-3σ is the only isoform directly linked to cancer in epithelial cells, which is regulated by major tumor suppressor gene [5–7].

In order to identify phosphopeptides that preferentially bind to 14-3-3σ over other isoforms, Yaffe et al. [3] identified a consensus hexapeptide binding motif, *RXXpSXP*, binding to all known 14-3-3 isoforms. Erik et al. [8] solved the X-ray crystal structure of 14-3-3σ, which provided structure information. Lu et al. [9] used fragment-based combinatorial peptide microarray platform, dividing whole library into *N*-terminal and *C*-terminal sublibraries $P_{-3}P_{-2}P_{-1} - p(S/T) - P_{+1}P_{+2}P_{+3}$. The $(+/-)$ represents relative position of $p(S/T)$, and $P_{+/-}$ represents ten or five individual amino acids. The phosphopeptide library is synthesized to get 14-3-3σ-specific binding peptide. For each binding peptide, they measured its binding affinity with each isoform of 14-3-3 family. The binding affinity describes the binding degree between a peptide and a 14-3-3 protein, the binding is more stable when the binding affinity is higher. They confirmed the previous consensus binding motif by Yaffe, and finally identified two 14-3-3σ-specific binders. However, their experimental methods are expensive and time consuming. Sequence variation at other positions near the phosphorylated site can cause differences in binding affinities, thus we can use the physical–chemical information to construct a computational model to extrapolate 14-3-3σ-specific binders from experimental data.

* Corresponding author.
E-mail address: fguo@tju.edu.cn (F. Guo).

http://dx.doi.org/10.1016/j.neucom.2016.03.093
0925-2312/

On the whole, there exist three categories of methods for Protein–Protein Interaction. First, some researchers use the evolution of information [10], which is extracted from multiple sequence alignment of homologous proteins. This is a novel but time-consuming approach. In addition, based on Natural Language Processing (NLP) [11] and relevant scientific literatures, the evidence for PPIs can be retrieved by some keywords. At the same time, some researchers propose PPIs prediction methods which only use protein sequence information. They extract useful features from the noisy amino acid sequences [12–14], and construct classification or regression models by some machine learning algorithms. There exist also some researchers who use three-dimensional structural information [15] and dug the hidden internal structure buried into noisy amino acid sequences. They predict protein interaction with a considerable accuracy and coverage that are superior to predictions based non-structural evidence. Recently, some web servers or stand alone tools were constructed to generate various features for DNA, RNA, and protein sequences, such as repDNA [16] and Pse-in-One [17].

Here, we propose a novel computational method to identify and analysis 14-3-3 phosphopeptide binding specificity, which is a fast and stable algorithm. For each 14-3-3 isoform, we have the peptide motifs with experimental binding affinity values, treated as known in this study. We should produce a predictor based these peptide motifs that can identify affinity values of the target peptide sequences binding to seven 14-3-3 isoforms. In this issue, there exist three significance challenges that we need to break down.

First, several effective computational methods identify protein binding sites, mainly based on three-dimensional structural information [15] and protein sequences information [18,13,14]. However, 14-3-3 phosphopeptide binders only have six meaningful positions in binding motif sequences, and the existing state-of-the-art binding sites prediction methods must be not suitable for this issue. So, how to dig the useful and important features is the first challenge. We select nine physicochemical properties of amino acids to describe each peptide motif, and also use auto cross covariance [19,20] to extract correlative properties of amino acids in any two positions.

Then, we deeply study the known peptide motifs with affinity values, which have an imbalanced distribution. There are many effective approaches in classification for imbalanced data sets [21]. Nitesh et al. [22] use synthetic minority over-sampling technique (SMOTE), which achieved better classifier performance. Therefore, several SMOTE-based methods were proposed, such as borderline-SMOTE [23], density-based SMOTE [24] and so on. However, we cannot find many works for imbalanced data sets in regression. So, how to improve the predicted precision of rare extreme values is the second challenge. We propose a similarity-based under-resampling method and a weighted SMOTE-based [25] over-resampling method.

Finally, we aim to predict the affinity value of different 14-3-3 isoforms, which is a continuous number. Without any prior knowledge, it is hard to choose a proper regression model, and the classical procedures cannot be effectively applied to have an accurate prediction. So, how to get the appropriate regression model is the third challenge. The regression algorithms are mainly divided into the linear model and the nonlinear model [26]. We use locally weighted regression [27] to predict affinity values of 14-3-3 isoforms binding peptide motifs, which combines the simplicity of linear least squares regression with the flexibility of nonlinear regression.

Our method verifies 1000 peptide motifs binding to seven distinct but highly homologous 14-3-3 isoforms. On 14-3-3σ isoform, our method has overall PCC and RMSE values of 0.83 and 258.31 for *N*-terminal sublibrary, and 0.80 and 250.89 for *C*-terminal sublibrary, respectively. It demonstrates the rationality of our computational method. We identify phosphopeptides that preferentially bind to 14-3-3σ over other isoforms. Several positions on peptide motifs have the same amino acid with experimental substrate specificity of phosphopeptides binding to 14-3-3σ.

2. Materials and methods

Here, we propose a novel computational method to identify and analysis 14-3-3 phosphopeptide binding specificity. First, we select nine physicochemical properties of amino acids to describe each peptide motif, and also use auto cross covariance to extract correlative properties of amino acids in any two positions. Then, we use a similarity-based under-resampling method and a weighted SMOTE-based over-resampling method to change the imbalanced distribution of 14-3-3 isoform binding affinity values. Finally, we consider locally weighted linear regression to predict 14-3-3 phosphopeptide binding specificity. The method flow is shown in Fig. 1.

2.1. Data set

Lu [9] proposed a fragment-based combinatorial peptide microarray, which enables sufficient coverage of all $(P_{-3}P_{-2}P_{-1} - p(S/T) - P_{+1}P_{+2}P_{+3})$ sequences with only 1000 peptide motifs (500 *N*-terminal and *C*-terminal sublibraries). These peptide motifs are formed as a phosphopeptide library. In a predefined

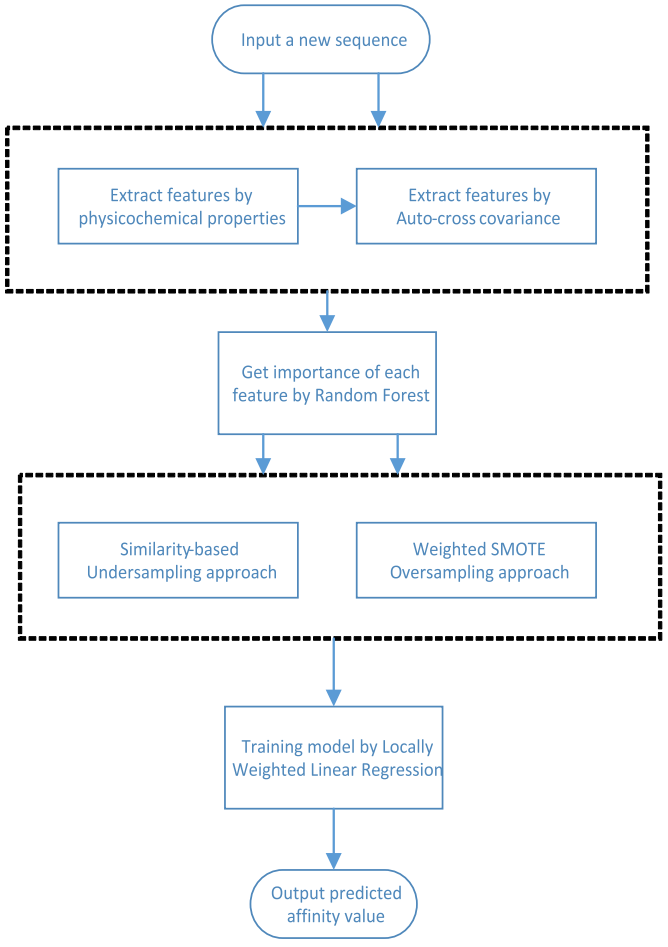

Fig. 1. The architecture of our computational approach to identify 14-3-3 proteins phosphopeptide-binding specificity.

Please cite this article as: Z. Li, et al., Learning from real imbalanced data of 14-3-3 proteins binding specificity, Neurocomputing (2016), http://dx.doi.org/10.1016/j.neucom.2016.03.093