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, P.R.China*

^b*School of Computational Science and Engineering, University of South Carolina,
Columbia, USA*

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. 14-3-3 σ 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 1,000 peptide motifs with experimental binding affinity values. In this paper, we present a novel method for identifying peptide motifs binding to 14-3-3 σ 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 1,000 peptide motifs binding to seven 14-3-3 isoforms. On the 14-3-3 σ 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

*Corresponding author

Email address: fguo@tju.edu.cn (Fei Guo)

C-terminal sublibrary. We identify phosphopeptides that preferentially bind to 14-3-3 σ over other isoforms. Several positions on peptide motifs have the same amino acid as experimental substrate specificity of phosphopeptides binding to 14-3-3 σ . Our method is a fast and reliable computational method that can be used in peptide-protein binding identification in proteomics research.

Keywords: similarity-based undersampling, SMOTE-like oversampling, locally weighted regression, physicochemical property, auto-cross covariance, 14-3-3 isoforms,

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: β , ϵ , η , γ , σ , τ , ζ [1, 2]. Phosphate can bind to all of the 14-3-3 proteins and therefore being
5 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
10 proliferation, cell cycle control, and human tumorigenesis. With roles of different 14-3-3 isoforms in a wide variety of signal transduction processes, 14-3-3 σ is the only isoform directly linked to cancer in epithelial cells, which is regulated by major tumor suppressor gene [5, 6, 7].

In order to identify phosphopeptides that preferentially bind to 14-3-3 σ
15 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 14-3-3 σ , 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
20 $P_{-3}P_{-2}P_{-1} - p(S/T) - P_{+1}P_{+2}P_{+3}$. The $(+/-)$ represents relative position of $p(S/T)$, and $P_{+/-}$ represents ten or five individual amino acids. The phospho-

peptide library is synthesized to get 14-3-3 σ -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
25 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 14-3-3 σ -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
30 the physical-chemical information to construct a computational model to extrapolate 14-3-3 σ -specific binders from experimental data.

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
35 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. In 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, 13, 14],
40 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.
45 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.
50 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.

55 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.
60 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
65 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
70 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
75 is a continuous number. Without any priori knowledge, it's 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 nonlinear model [26]. We use locally weighted regression
80 [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 1,000 peptide motifs binding to seven distinct but highly

homologous 14-3-3 isoforms. On 14-3-3 σ isoform, our method has overall PCC
85 and RMSE values of 0.83 and 258.31 for N -terminal sublibrary, and 0.80 and
250.89 for C -terminal sublibrary. It demonstrates the rationality of our com-
putational method. We identify phosphopeptides that preferentially bind to
14-3-3 σ over other isoforms. Several positions on peptide motifs have the same
amino acid with experimental substrate specificity of phosphopeptides binding
90 to 14-3-3 σ .

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 physicochemi-
cal properties of amino acids to describe each peptide motif, and also use auto
95 cross covariance to extract correlative properties of amino acids in any two posi-
tions. 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 lin-
ear regression to predict 14-3-3 phosphopeptide binding specificity. The method
100 flow is shown in Figure 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 1,000 peptide motifs (500 N -terminal and C -terminal sublibraries).
105 These peptide motifs are formed as a phosphopeptide library. In a prede-
fined manner, they use a total of ten different amino acid building blocks
 $(R, E, F, L, Q, A, G, V, K, P)$ for $P_{+/-1}$ and $P_{+/-2}$ positions, and a total of five
different amino acid building blocks (R, E, F, L, P) for $P_{+/-3}$ position.

With respect to each N -terminal and C -terminal, there are $5 \times 10 \times 10$
110 possibilities. For each 14-3-3 isoform, we have 1,000 peptide motifs with exper-
imental binding affinity values. In order to test the effectiveness of a predictor,

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

we choose Leave-one-out cross-validation, which involves using one peptide motif as the validation set and the remaining peptides as the training set. For each peptide motifs, there are three amino acid positions for one terminal and
115 a corresponding affinity value. We should produce a predictor based on known peptide motifs, that can identify affinity values of the target peptide sequences binding to seven 14-3-3 isoforms.

2.2. Feature Extraction

Based on relevant peptide motifs, we extract a set of features from peptide sequences. There are two kinds of features in this study: one extracts
120 nine physicochemical properties for each position and this produces 27 features; the other extracts correlation of amino acids in any two positions by auto-cross covariance, nine features for every two positions, thus leads to another 27 features[28].

We select nine physicochemical properties of all 20 amino acid types to describe each peptide motif: hydrophobicity, hydrophilicity, volumes of side chains, polarity, polarizability, solvent-accessible surface area (SASA), net charge index (NCI) of side chains, mass, and hydrogen bond. Details are shown in Table 1. These nine physicochemical properties are normalized to zero mean and unit
130 standard deviation [28, 19], and the first kind of 27 features can be extracted by these normalized properties as follows.

$$P'_{i,j} = \frac{P_{i,j} - P_j}{S_j} \quad (1)$$

where P_j represents mean of j -th property, $P_{i,j}$ is j -th property of i -th amino acid, S_j is corresponding unit standard deviation.

We also use auto-cross covariance to extract correlation of amino acids in any
135 two positions. Auto-cross covariance (ACC) can get two kinds of variables, Auto cross (AC) between the same descriptor, and cross covariance (CC) between two different descriptors. In this study, we only use AC variables in order to avoid generating too large number of variants. We modify the AC variables to get

Table 1: Nine physicochemical properties for 20 amino acid types.

	Physicochemical Properties ^a								
	H_1	H_2	H_3	V	P_1	P_2	SASA	NCI	MASS
A	0.62	-0.5	2	27.5	8.1	0.046	1.181	0.007187	71.0788
C	0.29	-1	2	44.6	5.5	0.128	1.461	-0.03661	103.1388
D	-0.9	3	4	40	13	0.105	1.587	-0.02382	115.0886
E	-0.74	3	4	62	12.3	0.151	1.862	0.006802	129.1155
F	1.19	-2.5	2	115.5	5.2	0.29	2.228	0.037552	147.1766
G	0.48	0	2	0	9	0	0.881	0.179052	57.0519
H	-0.4	-0.5	4	79	10.4	0.23	2.025	-0.01069	137.1411
I	1.38	-1.8	2	93.5	5.2	0.186	1.81	0.021631	113.1594
K	-1.5	3	2	100	11.3	0.219	2.258	0.017708	128.1741
L	1.06	-1.8	2	93.5	4.9	0.186	1.931	0.051672	113.1594
M	0.64	-1.3	2	94.1	5.7	0.221	2.034	0.002683	131.1986
N	-0.78	2	4	58.7	11.6	0.134	1.655	0.005392	114.1039
P	0.12	0	2	41.9	8	0.131	1.468	0.239531	97.1167
Q	-0.85	0.2	4	80.7	10.5	0.18	1.932	0.049211	128.1307
R	-2.53	3	4	105	10.5	0.18	1.932	0.049211	156.1875
S	-0.18	0.3	4	29.3	9.2	0.062	1.298	0.004627	87.0782
T	-0.05	-0.4	4	51.3	8.6	0.108	1.525	0.003352	101.1051
V	1.08	-1.5	2	71.5	5.9	0.14	1.645	0.057004	99.1326
W	0.81	-3.4	3	145.5	5.4	0.409	2.663	0.037977	186.2132
Y	0.26	-2.3	3	117.3	6.2	0.298	2.368	0.023599	163.1760

^a H_1 , hydrophobicity; H_2 , hydrophilicity; H_3 , hydrogen bond; V , volumes of side chains; P_1 , polarity; P_2 , polarizability; SASA, solvent-accessible surface area; NCI, net charge index of side chains; MASS, average mass of amino acid.

correlation of amino acids in any two positions as follows.

$$AC_{(m,n,j)} = (X_{m,j} - \frac{1}{3} \sum_{i=1}^3 X_{i,j}) \times (X_{n,j} - \frac{1}{3} \sum_{i=1}^3 X_{i,j}) \quad (2)$$

140 where m, n are different position of a peptide and j is j -th property of residues, $X_{i,j}$ is j -th property of residue on i -th position.

2.3. Resampling Approach

We deeply study the known peptide motifs with affinity values, as showed in Figure 2. Each bar represents the number of peptides motifs whose binding
145 affinity is in a specific interval of binding affinity. The x-axis shows the interval of the binding affinity, the y-axis shows the number of the peptides motifs in this interval. Deeming each peptide motifs as a sample, the samples having affinity value more than 2,000 are rare, meanwhile, a large part of samples have affinity value around 0. However, our aim is to find the motif with rare extreme affinity
150 values, and this is much more important to improve the precision of the higher values rather than the whole values. An effective method to address this task is to change the distribution of given data set by resampling approaches [29, 30].

Our method considers two different categories of resampling approaches: the first method to undersample the large set to make it reach a small size, and the
155 second method to oversample the small set to make it more close to the large one.

2.3.1. Similarity-Based Undersampling Approach

The present undersampling method is mainly dependent on the random re-sampling majority class or resampling at the decision boundary[31, 32]. Random-
160 based undersampling randomly selects samples from the majority class without concerning the usefulness of samples, while, resampling at the decision boundary is sensitive to class overlapping. Here, we propose a similarity-based undersampling approach which effectively find interesting samples.

All twenty amino acids can be divided into five categories [33]: amino acids
165 with positive charged side chains, amino acids with negative charged side chains,

sigma-left

sigma-right

Figure 2: The affinity value distribution of 14-3-3 σ phosphopeptide. Each bar represents the number of peptides motifs whose binding affinity is in a specific interval of binding affinity. The x-axis shows the interval of the binding affinity, the y-axis shows the number of the peptides motifs in this interval.

amino acids with polar uncharged side chains, amino acids with hydrophobic side chains and special cases. The details are shown in Table 2.

Table 2: Five categories of 20 amino acids

Category	Amino Acids ^a
Amino Acids with Positive Charged Side Chains	R , H , K
Amino Acids with Negative Charged Side Chains	D , E
Amino Acids with Polar Uncharged Side Chains	S , T , N , Q
Amino Acids with Hydrophobic Side Chains	A , I , L , M , F , W , Y , V
Special Cases	C , G , P

^a Standard abbreviations are used for all amino acids.

On each $P_{1/2/3}$ position, if two amino acids, one from each motif, belong to one common category, the amino acid similarity on this position is 1, otherwise is 0. We calculate the amino acid similarity between the sample S_A and the target sequence S_T as follows.

$$sim(S_A, S_T) = \varphi(A_1, T_1) + \varphi(A_2, T_2) + \varphi(A_3, T_3) \quad (3)$$

where A_i and T_i are corresponding amino acids on each position, $\varphi(\star, \star)$ represents that two amino acids belong to one common category or not.

Then, We propose a similarity-based undersampling approach to effectively reduce the size of large data set. For each target sequence motif, we will choose the samples which have at least two positions in the same category, and $sim(S_A, S_B) \geq 2$. Compared with other random-based undersampling approach, the similarity-based undersampling approach takes into more consideration informative samples in the data, and filter the irrelevant samples.

2.3.2. Weighted SMOTE-Based Oversampling Approach

As mentioned above, we filter some irrelevant peptides by their similarity with the target motif. Now, we use synthetic minority over-sampling technique (SMOTE) [22] to generate some new samples in the distribution of higher values. Torgo [30] used SMOTE in regression for the first time, and they changed SMOTE to generate new samples with continuous target values rather than the

discrete class labels. However, regardless of the original SMOTE and Torgo's SMOTE in regression, they are all k-nearest-neighbor-based method. By searching the k-nearest neighbors, a pair of sample can generate a new sample using SMOTE. But for this issue, with 54 features for each sample, a simple KNN
190 can not find the proper nearest neighbors. So, we propose a weighted k-nearest-neighbors-based synthetic minority over-sampling technique to solve the imbalanced distribution in high dimension regression. The novel KNN uses weighted
Euclidean distances to calculate the distance between two samples as follows:

$$dis(i_1, i_2) = \sum_{j=1}^p \phi_j (x_j^{i_1} - x_j^{i_2})^2 \quad (4)$$

where $x_j^{i_1}$ and $x_j^{i_2}$ are the j -th property of i_1 -th sample and i_2 -th sample, and
195 ϕ_j is the weight of the j -th feature. Considering highly accurate and robust performance, we use random forest as preprocessing to get the weight of each property.

In Algorithm 1, the new target value is generated by two seed samples, and it is a weighted average of these two seed samples. The distance between the
200 seed sample and the new sample is a weighted Euclidean distance.

2.4. Locally Weighted Linear Regression

We use locally weighted regression 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. Given feature vectors $X_1, \dots, X_p$ describing p features on each peptide sequence, we identify its corresponding value $f(X)$ to represent binding affinity value as follows.

$$f(X) = \beta_0 + \sum_{j=1}^p X_j \beta_j \quad (5)$$

The naive linear regression models calculate the predictor coefficient, through minimizing the residual sum of squares (RSS).

$$\arg \min_{\beta} \left\{ \sum_{i=1}^N (y^i - \beta_0 - \sum_{j=1}^p x_j^i \beta_j)^2 \right\} \quad (6)$$

Algorithm 1 The weighted KNN-based SMOTE for high dimension regression

Input:

- A data set, D ;
- The threshold for normal and extreme value, t ;
- The times of over-sampling, n_g ;
- The number of the nearest neighborhoods used in case generation, k ;
- The weighted vector of features, ω

Output:

- The new data set with generated new samples, D' ;
 - 1: Dividing the data set D into D_{norm} and D_{rare} by the threshold t ;
 - 2: $newCases \leftarrow \{\}$
 - 3: **for all** $case \in D_{rare}$ **do**
 - 4: $nns \leftarrow KNN_{weighted}(k, case, D_{rare} - \{case\})$; //get k-nns of case
 - 5: **for** i from 1 to n_g **do**
 - 6: $x \leftarrow$ randomly choose one from nns
 - 7: **for all** $a \in attributes$ **do**
 - 8: $diff \leftarrow case[a] - x[a]$
 - 9: $new[a] \leftarrow case[a] + RANDOM(0, 1) \times diff$
 - 10: **end for**
 - 11: $d_1 \leftarrow DIST(new, case)$ // DIST is a weighted Euclidean distances
 - 12: $d_2 \leftarrow DIST(new, x)$
 - 13: $new[Target] \leftarrow \frac{d_2 \times case[Target] + d_1 \times x[Target]}{d_1 + d_2}$
 - 14: $newCases \leftarrow newCases \cup \{new\}$
 - 15: **end for**
 - 16: **end for**
 - 17: **return** $newCases \cup D$;
-

The weighted least squares are used to label high weight for nearer samples, and label low weight for further samples. Through minimizing the weighted residual sum of squares, the value of regression function is then obtained by
205 evaluating the local linear model and using the explanatory variable. There are several weight functions that satisfy the properties listed in Cleveland [34]. Here, we choose a Gaussian kernel, as follows:

$$\omega^i = \exp\left(-\frac{\sum_{j=1}^p (x_j^i - x_j)^2}{2\gamma^2}\right) \quad (7)$$

where i is the i -th sample, j is the j -th feature, p is the number of features, x_j is target 14-3-3 binder, γ is a smoothing parameter which is used to control the
210 variation rate of weight.

As mentioned above, locally weighted linear regression is a k-nearest-neighbor-based regression model. We calculate a weighted Euclidean distances to obtain the weight of each sample. The weighted least squares in locally weighted linear regression with the weighted features can be calculated as follows:

$$\arg \min_{\beta} \left\{ \exp\left(-\frac{\sum_{j=1}^p \phi_j (x_j^i - x_j)^2}{2\gamma^2}\right) \sum_{i=1}^N (y^i - \beta_0 - \sum_{j=1}^p x_j^i \beta_j)^2 \right\} \quad (8)$$

215 where ϕ_j is weight of the j -th feature, which is the same weight as above in SMOTE.

3. Results

In this section, we have done three kinds of experiments. First, a series of experiments are designed to test the effectiveness of our resampling approaches
220 and the locally weighted linear regression as above. Then, our method verifies the 1,000 peptide motifs binding to seven distinct but highly homologous 14-3-3 isoforms. Finally, we identify phosphopeptides that preferentially bind to 14-3-3 σ over other isoforms.

3.1. Effectiveness of Resampling Approaches and Locally Weighted Linear Regression

225

In order to verify the effectiveness of our resampling methods as above, we select 1,000 peptide motifs binding to 14-3-3 σ . For each N -terminal and C -terminal, we use the linear regression without resampling approach and the linear regression with resampling approach to predict 14-3-3 isoforms binding affinity value, respectively. Then, a locally weighted linear regression with re-
 230 sampling approach are used to test effectiveness of locally weighted linear regression.

In Figure 3a and Figure 3d, we just use the simple linear regression without any resampling approach. In Figure 3b and Figure 3e, we filter samples that the similarities between them and the target sequence are less than 2 on the normal
 235 set, and generate two times new samples on the extreme rare set. Here, we choose the affinity value 1,000 as the threshold to divide the normal set and the rare set. In these two experiments, we use the same simple linear regression. In Figure 3c and Figure 3f, we use locally weighted linear regression combined with
 240 resampling approaches. As seen, the locally weighted linear regression produce better result than the simple linear regression.

In this paper, we aim to find peptide motif having 14-3-3 σ specificity, and also these peptide motifs must have high affinity values. Using a top rank number, we divide the 1,000 peptides to a set of motifs with high affinity values
 245 and a set of motifs with low affinity values, so a predictor that can accurately find the peptide motifs with high affinity values, is very important. This is the same as the classification problem, we define the motifs with high affinity values as positive cases, and the motifs with low affinity values as negative cases, so the precision and recall can be used to evaluate the predictor. In this experiment,
 250 we use F -measure as the main evaluation metric, which can be calculated with precision and recall as follows.

$$F = \frac{(\beta^2 + 1) * precision * recall}{\beta^2 * precision + recall} \quad (9)$$

Figure 3: Predicted values vs. True value for 14-3-3 σ phosphopeptide Binding Affinity. a) Simple linear regression without resampling methods for N -terminal of 14-3-3 σ . b) Simple linear regression with resampling methods for N -terminal of 14-3-3 σ . c) Locally weighted linear regression with resampling methods for N -terminal of 14-3-3 σ . d) Simple linear regression without resampling methods for C -terminal of 14-3-3 σ . e) Simple linear regression with resampling methods for C -terminal of 14-3-3 σ . f) Locally weighted linear regression with resampling methods for C -terminal of 14-3-3 σ .

where β is a parameter weighting the importance of precision and recall, here, we use $\beta = 1$.

Using different values of top rank number, the F -score metric is showed in Table 3. In our method, almost all of F -scores get better regardless of the top rank number. The results provide clear evidence of the advantage of our resampling methods and regression model.

Table 3: F -score on different values of top rank number.

	N-terminal				C-terminal			
	Top-100	Top-50	Top-40	Top-30	Top-100	Top-50	Top-40	Top-30
LR No-Resample	0.760	0.700	0.675	0.667	0.640	0.620	0.600	0.567
LWLR Resample	0.770	0.720	0.700	0.600	0.680	0.660	0.625	0.700

3.2. Verification on 1,000 peptide motifs

Our method verifies 1,000 peptide motifs binding to seven 14-3-3 isoforms. We using the 999 peptide motifs with experimental binding affinity values as training data, removing the predicted peptide sequence. When only selecting 'relevant' data for building the predictor, about 300 peptide motifs are selected as training data each time on average. The pearson-product-moment correlation coefficient (PCC) and the root mean squared error (RMSE) [35] are used to evaluate performance as follows:

$$PCC = \sqrt{1 - \frac{\sum_{i=1}^N (e_i - p_i)^2}{\sum_{i=1}^N (e_i - \bar{e})^2}} \quad (10)$$

and

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (e_i - p_i)^2}{|D|}} \quad (11)$$

where D contains all of relevant binding motifs, $\bar{e}$ is the average binding affinity, e_i denotes experimental binding affinity value of the i -th peptide sequence, p_i denotes the predicted affinity value of the i -th peptide sequence. An accurate predictor will get $PCC = 1$, $RMSE = 0$.

270 Details on identifying peptide motifs binding to 14-3-3 isoforms are shown
in Table 4. On the 14-3-3 σ 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. It yields a considerable PCC in all seven isoforms,
and the results clearly highlight the effectiveness of our method. At the same
275 time, the RMSE values vary in different isoforms, because of several extra large
values of affinity and imbalanced peptide distribution between diverse values in
different isoforms.

Table 4: Details on predicting peptide motifs binding to 14-3-3 isoforms.

	N-terminal		C-terminal	
	PCC	RMSE	PCC	RMSE
σ	0.83	258.31	0.80	250.89
β	0.79	205.46	0.71	225.49
ϵ	0.88	426.23	0.82	409.08
η	0.84	218.76	0.75	240.47
γ	0.88	440.94	0.83	427.84
τ	0.85	539.63	0.78	608.44
ζ	0.87	2100.10	0.81	2339.94

We produce a position-specific scoring matrix [36] on the top 50 motifs identified from each N -terminal and C -terminal sublibrary against each individual
280 14-3-3 isoform, to reflect position specialty for each amino acid, as shown in Figure 4. The height of each letter represents weighted contribution of that amino acid to overall peptide binding. Our method is compared with the experimental results from Yaffe [3] and Lu [9], as summarized in Table 5. Our method obtains accurate results, in accord with previous experimental works on 14-3-3
285 isoforms binding peptide motifs. We get relative binding potency of all seven 14-3-3 isoforms across six permuted positions, as shown in Figure 5. Each bar represents the frequency of a particular amino acid. This conforms highly homologous feature of 14-3-3 isoforms, and highly agrees with consensus binding motif, $RXXpSXP$.

Figure 4: Position-specific scoring matrix on top 50 motifs identified from 1,000 peptide sequences against individual 14-3-3 isoforms.

Table 5: 14-3-3 preferences determined with different methods

	Position Relative to p(S/T)					
	P_{-3}	P_{-2}	P_{-1}	P_{+1}	P_{+2}	P_{+3}
Yaffe	RK	FSYWHHA	RKH	WYFLA	PG	X
H.S. Lu	R	AFPR	RK	AVFL	PA	FLP
Our Method ^a	R	APFVRL	KRAV	AVFL	PAG	FLPR

^a Results on 1,000 peptide motifs binding to seven 14-3-3 isoforms by our method.

Figure 5: Binding affinity of seven 14-3-3 isoforms across six positions from top-50 peptides from both N- and C-terminal sublibrary.

290 3.3. Specificity of 14-3-3 σ binding peptide motifs

On the 1,000 peptide motifs, we identify the top 100 peptide motifs, irrespective of *N*-terminal or *C*-terminal, binding each 14-3-3 isoform. We filter and identify consensus sequences present in all seven isoforms, giving a total of 52 unique peptide motifs, as shown in Table 6. Compared with Lu [9], 33 peptide motifs of our results are the same with experimental 46 binding sequences, which are represented by the $\star$ label.

Table 6: List of 52 consensus top binders from 1,000 peptide sequences against all seven 14-3-3 isoforms.

No.		N-terminal ^a	No.		N-terminal ^a	No.		C-terminal ^a
1	*	RAApS/TXXX	21	*	RPKpS/TXXX	37		XXXpS/TAAF
2	*	RAFpS/TXXX	22	*	RPLpS/TXXX	38		XXXpS/TAGF
3	*	RAKpS/TXXX	23	*	RPQpS/TXXX	39		XXXpS/TAGP
4	*	RALpS/TXXX	24	*	RPRpS/TXXX	40	*	XXXpS/TAPF
5	*	RAQpS/TXXX	25		RPVpS/TXXX	41	*	XXXpS/TAPL
6	*	RARpS/TXXX	26	*	RRApS/TXXX	42	*	XXXpS/TAPP
7	*	RAVpS/TXXX	27	*	RRFpS/TXXX	43		XXXpS/TAPR
8	*	RFApS/TXXX	28		RRLpS/TXXX	44	*	XXXpS/TFPF
9	*	RFKpS/TXXX	29		RRRpS/TXXX	45	*	XXXpS/TFPL
10	*	RFRpS/TXXX	30	*	RRVpS/TXXX	46		XXXpS/TFPP
11		RFVpS/TXXX	31	*	RVApS/TXXX	47	*	XXXpS/TLPF
12		RGApS/TXXX	32	*	RVFpS/TXXX	48	*	XXXpS/TLPL
13		RLApS/TXXX	33	*	RVKpS/TXXX	49		XXXpS/TLPP
14		RLFpS/TXXX	34		RVLpS/TXXX	50	*	XXXpS/TVPF
15		RLKpS/TXXX	35	*	RVRpS/TXXX	51	*	XXXpS/TVPL
16		RLLpS/TXXX	36		RVVpS/TXXX	52	*	XXXpS/TVPP
17		RLRpS/TXXX						
18		RLVpS/TXXX						
19	*	RPApS/TXXX						
20	*	RPFpS/TXXX						

^a The motif with label $\star$ is the same with experimental binding sequences of H.S. Lu.

We identify four peptide motifs that have 14-3-3 σ specificity, as shown in Table 7. They belong to the top 100 sequences binding 14-3-3 σ , but not being part of the top 100 sequences binding other 14-3-3 isoforms. Compared with two 14-3-3 σ preferable binders of Lu, B1:LFGpSLLR and B2:LFGpSLVR, one motif is in the same amino acid category on P_{-3} position, two motif have the same

amino acid on P_{-2} and P_{+1} positions, as shown in Table 2. It demonstrates that our predictions are in excellent agreement with previous work using degenerate peptide libraries.

Table 7: List of four preferable binders of 14-3-3 σ from 1,000 peptide sequences.

No.	N-terminal	No.	C-terminal
1	FFKpS/TXXX	4	XXXpS/TLGF
2	PAFpS/TXXX		

305 4. Conclusions

We present a novel method for identifying peptide motifs binding to 14-3-3 isoforms. 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, we propose
310 a similarity-based undersampling approach and a SMOTE-like oversampling approach 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.

315 Our method verifies the 1,000 peptide motifs binding to seven distinct but highly homologous 14-3-3 isoforms, and the relative binding potency across six permuted positions highly agree with the experimental value. We identify phosphopeptides that preferentially bind to 14-3-3 σ over other isoforms. Several of positions on peptide motifs have the same amino acid as experimental
320 substrate specificity of phosphopeptides binding to 14-3-3 σ . It indicates that, regardless of how the data are analyzed, 14-3-3 σ consensus binding motifs derived from our experiments are in excellent agreement with previous work. Our method is designed as a generalized method that can be used to accurately predict the binding affinity for protein-peptide interaction in proteomics research.

325 Since user-friendly and publicly accessible web-servers represent the future di-
rection for developing practically more useful models, simulated methods, or
predictors as pointed out in [37, 38], we shall make efforts in our future work to
provide a web-server for the method presented in this paper.

5. Availability

330 Test set of peptide sequences and prediction results are available for down-
load from <https://github.com/Victor-LiZhao/1433Sigma.git>.

Acknowledgment

This work is supported by a grant from National Science Foundation of
China (NSFC 61402326).

335 References

- [1] J. D. M. J. Aitken A., Howell S., P. Y, 14-3-3 α and δ are the phospho-
rylated forms of raf-activating 14-3-3 β and ζ in vivo stoichiometric phos-
phorylation in brain at a ser-pro-glu-lys motif, The Journal Of Biological
Chemistry 270 (2005) 5706–5709.
- 340 [2] E. Wilker, M. B. Yaffe, 14-3-3 proteins - a focus on cancer and human
disease, Journal of Molecular and Cellular Cardiology 37 (2004) 633–642.
- [3] M. B. Yaffe, K. Rittinger, S. Volinia, P. R. Caron, A. Aitken, H. Leffers,
S. J. Gamblin, S. J. Smerdon, L. C. Cantley, The structural basis for 14-3-
3:phosphopeptide binding specificity, Cell 91 (1997) 961–71.
- 345 [4] N. N. Sluchanko, N. A. Chebotareva, N. B. Gusev, Modulation of 14-3-
3/phosphotarget interaction by physiological concentrations of phosphate
and glycerophosphates, Plos One 8 (2013) 8.
- [5] H. Hermeking, The 14-3-3 cancer connection, Nature Reviews Cancer 3
(2003) 931–943.

- 350 [6] Y. Zhang, Y. Li, C. Lin, J. Ding, G. Liao, B. Tang, Aberrant upregulation of 14-3-3sigma and ezh2 expression serves as an inferior prognostic biomarker for hepatocellular carcinoma, *PloS one* 9 (2014) e107251.
- [7] Y. J. Qi, M. Wang, R. M. Liu, H. Wei, W. X. Chao, T. Zhang, Q. Lou, X. M. Li, J. Ma, H. Zhu, Z. H. Yang, H. Q. Liu, Y. F. Ma, Downregulation of 14-3-3 sigma correlates with multistage carcinogenesis and poor prognosis of esophageal squamous cell carcinoma, *Plos One* 9 (2014) 11.
- 355 [8] E. W. Wilker, R. A. Grant, S. C. Artim, M. B. Yaffe, A structural basis for 14-3-3 sigma functional specificity, *Journal of Biological Chemistry* 280 (2005) 18891–18898.
- 360 [9] C. H. S. Lu, H. Y. Sun, F. B. Abu Bakar, M. Uttamchandani, W. Zhou, Y. C. Liou, S. Q. Yao, Rapid affinity-based fingerprinting of 14-3-3 isoforms using a combinatorial peptide microarray, *Angewandte Chemie-International Edition* 47 (39) (2008) 7438–7441.
- [10] D. D. F. V. Alfonso, Emerging methods in protein co-evolution, *Nature Reviews Genetics* 14(4) (2013) 249–261.
- 365 [11] H. J. L. L. P. Lee, S. Park, Finding the evidence for protein-protein interactions from pubmed abstracts, *Bioinformatics* 22(14) (2006) e220e226.
- [12] X. Y. Pan, Y. N. Zhang, H. B. Shen, Large-scale prediction of human protein-protein interactions from amino acid sequence based on latent topic features, *Journal of Proteome Research* 9 (2010) 4992–5001.
- 370 [13] Z. H. You, Y. K. Lei, L. Zhu, J. F. Xia, B. Wang, Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis, *Bmc Bioinformatics* 14 (2013) 11.
- 375 [14] Z. H. You, K. C. C. Chan, P. W. Hu, Predicting protein-protein interactions from primary protein sequences using a novel multi-scale local feature representation scheme and the random forest, *Plos One* 10 (2015) 19.

- [15] Q. C. Zhang, D. Petrey, L. Deng, L. Qiang, Y. Shi, C. A. Thu, B. Bisikirska, C. Lefebvre, D. Accili, T. Hunter, T. Maniatis, A. Califano, B. Honig, Structure-based prediction of protein-protein interactions on a genome-wide scale, *Nature* 490 (2012) 556–+.
- [16] B. Liu, F. L. Liu, L. Y. Fang, X. L. Wang, K. C. Chou, repdna: a python package to generate various modes of feature vectors for dna sequences by incorporating user-defined physicochemical properties and sequence-order effects, *Bioinformatics* 31 (8) (2015) 1307–1309.
- [17] B. Liu, F. Liu, X. Wang, J. Chen, L. Fang, K.-C. Chou, Pse-in-one: a web server for generating various modes of pseudo components of dna, rna, and protein sequences, *Nucleic Acids Research* 43 (W1) W65–W71.
- [18] N. Zaki, S. Lazarova-Molnar, W. El-Hajj, P. Campbell, Protein-protein interaction based on pairwise similarity, *Bmc Bioinformatics* 10 (2009) 12.
- [19] Y. Z. Guo, L. Z. Yu, Z. N. Wen, M. L. Li, Using support vector machine combined with auto covariance to predict proteinprotein interactions from protein sequences, *Nucleic Acids Research* 36 (9) (2008) 3025–3030.
- [20] V. S. Mathura, D. Kolippakkam, Apdbase: Amino acid physico-chemical properties database, *Bioinformation* 1 (2005) 2–4.
- [21] C. Lin, W. Q. Chen, C. Qiu, Y. F. Wu, S. Krishnan, Q. Zou, Libd3c: Ensemble classifiers with a clustering and dynamic selection strategy, *Neurocomputing* 123 (2014) 424–435.
- [22] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, Smote: Synthetic minority over-sampling technique, *Journal of Artificial Intelligence Research* 16 (2002) 321–357.
- [23] H. Han, W. Y. Wang, B. H. Mao, Borderline-smote: A new over-sampling method in imbalanced data sets learning, *Advances in Intelligent Computing, Pt 1, Proceedings* 3644 (2005) 878–887.

- 405 [24] C. Bunkhumpornpat, K. Sinapiromsaran, C. Lursinsap, Dbmsote: Density-based synthetic minority over-sampling technique, *Applied Intelligence* 36 (2012) 664–684.
- [25] L. Torgo, R. P. Ribeiro, B. Pfahringer, P. Branco, Smote for regression, *Progress in Artificial Intelligence, Epia 2013* 8154 (2013) 378–389.
- 410 [26] F. Stulp, O. Sigaud, Many regression algorithms, one unified model: A review, *Neural Networks* 69 (2015) 60–79.
- [27] V. Centner, D. L. Massart, Optimization in locally weighted regression, *Analytical Chemistry* 70 (1998) 4206–4211.
- [28] Z. H. You, Y. K. Lei, L. Zhu, J. F. Xia, B. Wang, Prediction of protein-protein interactions from amino acid sequences with ensemble extreme
415 learning machines and principal component analysis, *Bmc Bioinformatics* 14 (2013) 11.
- [29] L. Torgo, R. Ribeiro, Precision and recall for regression, *Discovery Science, Proceedings* 5808 (2009) 332–346.
- 420 [30] L. Torgo, P. Branco, R. P. Ribeiro, B. Pfahringer, Resampling strategies for regression, *Expert Systems* 32 (2015) 465–476.
- [31] Q. Zou, Z. Wang, X. J. Guan, B. Liu, Y. F. Wu, Z. Y. Lin, An approach for identifying cytokines based on a novel ensemble classifier, *Biomed Research International* (2013) 11.
- 425 [32] L. Song, D. P. Li, X. X. Zeng, Y. F. Wu, L. Guo, Q. Zou, ndna-prot: identification of dna-binding proteins based on unbalanced classification, *Bmc Bioinformatics* 15 (2014) 10.
- [33] I. Wagner, H. Musso, New naturally occurring amino acids, *Angewandte Chemie International Edition in English* 22 (1983) 816–828.
- 430 [34] W. S. Cleveland, Robust locally weighted regression and smoothing scatter-plots, *Journal of the American Statistical Association* 74 (1979) 829–836.

- [35] S. Giguere, M. Marchand, F. Laviolette, A. Drouin, J. Corbeil, Learning a peptide-protein binding affinity predictor with kernel ridge regression, *Bmc Bioinformatics* 14 (2013) 16.
- 435 [36] C. J. B. S. Crooks GE, Hon G, Weblogo: A sequence logo generator, *Genome Research* 14(6) (2004) 1188–1190.
- [37] B. Liu, L. Fang, R. Long, X. Lan, K.-C. Chou, ienhancer-2l: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition, *Bioinformatics (Oxford, England)* 32 (3) (2016) 362–9.
- 440 [38] Q. Zou, Q. H. Hu, M. Z. Guo, G. H. Wang, Halign: Fast multiple similar dna/rna sequence alignment based on the centre star strategy, *Bioinformatics* 31 (15) (2015) 2475–2481.

Zhao Li received the bachelor's degree from Inner Mongolia University, in 2014. He is currently a Mphil with Prof. Jijun Tang in the College of Computer Science and Technology, Tianjin University. His research interest is bioinformatics.

Fei Guo received the bachelor's degree and the PhD degree from Shandong University, in 2007 and 2012. She was a postdoctoral fellow with Prof. Lusheng Wang at City University of Hong Kong between 2012 and 2013. She is currently a lecturer in the College of Computer Science and Technology, Tianjin University. Her research interests include bioinformatics and computational biology.

Jijun Tang received the PhD degree from the University of New Mexico in 2004. He is currently an associate professor in the Department of Computer Science and Engineering, University of South Carolina. He is also an adjunct professor in the School of Computer Science and Technology, Tianjin University, China. His main research interests include computational biology, with a focus on algorithm development in phylogenetic reconstruction from genome rearrangement data.

Photo of Zhao Li(author)

Photo of Fei Guo(author)

Photo of Jijun Tang(author)