PLOS | ONE

RESEARCH ARTICLE

Kinase Identification with Supervised Laplacian Regularized Least Squares

Ao Li[1,2☯], **Xiaoyi Xu**[1☯], **He Zhang**[1], **Minghui Wang**[1,2*]

1 School of Information Science and Technology, University of Science and Technology of China, 443 Huangshan Road, Hefei 230027, Anhui, China, **2** Centers for Biomedical Engineering, University of Science and Technology of China, 443 Huangshan Road, Hefei 230027, Anhui, China

☯ These authors contributed equally to this work.
* mhwang@ustc.edu.cn

OPEN ACCESS

Citation: Li A, Xu X, Zhang H, Wang M (2015) Kinase Identification with Supervised Laplacian Regularized Least Squares. PLoS ONE 10(10): e0139676. doi:10.1371/journal.pone.0139676

Editor: Dinesh Gupta, International Centre for Genetic Engineering and Biotechnology (ICGEB), INDIA

Received: January 8, 2015

Accepted: September 16, 2015

Published: October 8, 2015

Data Availability Statement: The data can be obtained from Phospho.ELM database by downloading the file "ALL phosphorylation instances in tab-delimited format," and from the paper of Newman et al. (20.Newman RH, Hu J, Rho H-S, Xie Z, Woodard C, Neiswinger J, Cooper C, Shirley M, Clark HM, Hu S (2013) Construction of human activity-based phosphorylation networks. Mol System Biology, 9(1):655. doi: 10.1038/msb.2013.12.).

Funding: This work was supported by grants from National Natural Science Foundation of China (61471331, 61101061).

Abstract

Phosphorylation is catalyzed by protein kinases and is irreplaceable in regulating biological processes. Identification of phosphorylation sites with their corresponding kinases contributes to the understanding of molecular mechanisms. Mass spectrometry analysis of phosphor-proteomes generates a large number of phosphorylated sites. However, experimental methods are costly and time-consuming, and most phosphorylation sites determined by experimental methods lack kinase information. Therefore, computational methods are urgently needed to address the kinase identification problem. To this end, we propose a new kernel-based machine learning method called Supervised Laplacian Regularized Least Squares (SLapRLS), which adopts a new method to construct kernels based on the similarity matrix and minimizes both structure risk and overall inconsistency between labels and similarities. The results predicted using both Phospho.ELM and an additional independent test dataset indicate that SLapRLS can more effectively identify kinases compared to other existing algorithms.

Introduction

Protein phosphorylation is one of the most pervasive posttranslational modifications and plays an important role in regulating nearly all types of cellular processes in organisms, including signal transduction, translation and transcription [1,2,3,4,5]. Phosphorylation is catalyzed by protein kinases [6], which regulate most cellular processes. More than one-third of proteins can be phosphorylated, and half of the protein kinases have intimate relationships with cancer and diseases [7]. Each protein kinase specifically catalyzes a certain subset of substrates, and deficiencies in protein kinases often cause diseases and cancers [6]. In this regard, identifying potential phosphorylation sites and their corresponding protein kinases is beneficial for elucidating molecular mechanisms.

Conventional experimental methods such as high-throughput biological technique mass spectrometry [8] were developed to identify phosphorylation sites. Although these experimental methods provide the foundation for understanding the mechanisms underlying phosphorylation,

Competing Interests: The authors have declared that no competing interests exist.

they are often costly and time-consuming. Additionally, although mass spectrometry methods can generate a large number of phosphorylated sites, most of these sites lack kinase information and the kinase that catalyzes the site is unknown. For example, Phospho.ELM, which is a verified phosphorylation site database, contains 37,145 phosphorylation sites, but only a small number of them (3,599 items) have corresponding kinase information. Due to the limitation of experimental methods, computational methods are required to identify protein kinases for specific phosphorylation sites based on data verified by experimental methods.

Currently, many computational methods have been developed for protein phosphorylation prediction. The first computational method proposed by Blom [9] was based on an artificial neural network algorithm using peptide sequences. Since then, a large number of methods have been developed, such as PPSP [10] and Musite [11]. PPSP adopts the Bayesian decision theory (BDT), which is based on an assumption that all flanking residues are independent of each other, to construct a classifier. Musite calculates the distance between two peptide sequences using the distance calculator Blosum62, which is a matrix reflecting the relationship between amino acids, and then constructs the classifier with support vector machine (SVM). Due to the increasing demand for kinase identification, a few machine learning-based methods have been developed in recent years. Among them, NetworKIN [12] uses consensus sequence motifs and a probabilistic protein association network. IGPS [13] is based on peptide sequence similarity and uses protein-protein interaction (PPI) information to control the false positive rate.

Despite the success achieved by these computational approaches, most of them neglect the geometry of the probability distribution [14], thereby hampering the improvement of prediction accuracy. For example, SVM only focuses on structural risk minimization and the quadrature encoder and thus ignores the intrinsic relationship between different amino acids. Additionally, the distance of two peptide sequences defined in Musite may fail to fulfill the triangle inequality [11]. To solve these problems, Belkin et al. [14] proposed a framework exploiting the geometry of the probability distribution; the test results showed that the proposed framework efficiently addressed the classification problems.

In this study, we propose a kernel-based supervised learning algorithm called Supervised Laplacian Regularized Least Squares (SLapRLS), which incorporates a new kernel construction method and brings together the spectral graph theory, regularization and the geometry of the probability distribution for kinase identification. In SLapRLS, reasonable translations are performed on a similarity matrix to force it to act as a kernel matrix [15]. Additionally, we introduce the overall inconsistencies between sample similarities and labels for each class [16] and minimize both the inconsistency and the structure risk. To compare the proposed algorithm with existing algorithms, we perform a 10-fold cross-validation using data retrieved from Phospho.ELM and compare SLapRLS with three classical algorithms: SVM, BDT and the k-nearest neighbor (KNN). To confirm the effectiveness and superiority of SLapRLS, an additional independent test dataset is used to compare SLapRLS and two other kinase identification tools: iGPS and NetworKIN. The results show that SLapRLS is more effective than the competitive algorithms and that the kernel matrix construction method is useful for the identification of kinases corresponding to known phosphorylation sites.

Materials and Methods

Data description

In this work, we extracted 37,145 experimentally verified phosphorylation sites from humans, including 3,599 sites with corresponding kinase information, from the most recent version of Phospho.ELM [17]. Among the sites with kinase information, 2,398 unique phosphorylation sites with kinase information in 934 proteins are obtained after removing the duplicated data. To overcome the over-estimation aroused by homology bias and redundancy, we cluster the protein sequences using Blastclust with a threshold of 70%; only one representation of each cluster is reserved [18]. As a result, 2,289 sites in 889 proteins are employed for the analysis. There are 1,823 serine (S)/ threonine (T) phosphorylation sites and 446 tyrosine (Y) phosphorylation sites. For each kinase, the corresponding phosphorylation sites are treated as positive data, whereas sites phosphorylated by other kinases are treated as negative data. Several kinases that contain too few known phosphorylated substrates are excluded to achieve reliable results. Finally, 23 types of kinases are obtained for investigation after removing the kinases that contain less than 20 positive items.

Because iGPS and NetworKIN use data retrieved from the Phospho.ELM database for model training, the test dataset in this study at least partially includes the training dataset of these two methods. This factor would inevitably result in the overestimation of the prediction performance for iGPS and NetworKIN. To obtain a fair comparison result, an independent dataset is adopted in this work [19]. Similarly, protein kinases in the independent test dataset that contain less than 20 items are also excluded to ensure the reliability of the results. Finally, we select 6 kinases in the independent dataset: PKC alpha, Erk2, Erk1, P38a, SRC and SYK.

Algorithm

In this work, we propose that SLapRLS brings together spectral graph theory, regularization and the geometry of the probability distribution based on the regularized least squares (RLS) theory [14]. Similar to SVM, RLS is engaged in minimizing the structure risk [20]. SLapRLS is proposed based on the manifold assumption that similar samples tend to have similar results, and thus samples with the same label are predicted to have similar results. Therefore, the overall inconsistency between labels and pairwise similarities in the same class should be minimized. SLapRLS aims to minimize both structure risk and the overall inconsistency between labels and pairwise similarities.

Feature description

In this work, we take full advantage of sequence information in modeling. A 15 amino acid local sequence is used to represent a candidate phosphorylation site that has 7 amino acids upstream and downstream of the phosphorylation site (S, T or Y). Thus, a phosphorylation site can be denoted as $s = (s(1), s(2),\ldots, s(8),\ldots, s(15))$, where $s(i)$ represents the amino acid at the i_{th} position and $s(8)$ is the phosphorylation site.

Structure risk minimization and RLS

Structure risk minimization aims to minimize VC confidence and the summation of the empirical risk on each subset [21]. The square of the difference between the true label and the predicted result is often used as the loss function when calculating the empirical risk [22]. The optimization problem of RLS is shown as:

$$\min \frac{1}{l} \Sigma_{i=1}^{l} (y_i - f(x_i))^2 + \gamma_A || f ||_K^2 \tag{1}$$

where y_i and $f(x_i)$ represent the true label and the predicted result of the i_{th} sample, respectively.

Inconsistency between labels and pairwise similarities

A good predictor should predict similar data with similar results, and thus the overall inconsistency between labels and pairwise similarities should be minimized [16]. The inconsistency contains two parts: the first is the inconsistency in the positive dataset and the second is the inconsistency in the negative dataset. The overall inconsistency is minimized and shown as:

$$\min \frac{1}{\mathrm{p}^2} \Sigma_{\mathrm{i,j}=1}^{\mathrm{p}} (\mathrm{f}(\mathrm{x}_\mathrm{i}) - \mathrm{f}(\mathrm{x}_\mathrm{j}))^2 \mathrm{W}_{\mathrm{ij}} + \frac{1}{\mathrm{n}^2} \Sigma_{\mathrm{i,j}=\mathrm{p}+1}^{\mathrm{p}+\mathrm{n}} (\mathrm{f}(\mathrm{x}_\mathrm{i}) - \mathrm{f}(\mathrm{x}_\mathrm{j}))^2 \mathrm{W}_{\mathrm{ij}} \tag{2}$$

where p and n represent the number of positive data and negative results, respectively, and W_{ij} is the similarity between samples x_i and x_j.

Supervised LapRLS

Supervised LapRLS is based on the principle that both inconsistency between labels and pairwise similarities and structural risk should be minimized. The optimization problem aims to solve Eqs (1) and (2), which can be represented as Eq (3). This is a multiple objective optimization problem, and thus a weight parameter γ_I is introduced to weight the two objects [23].

$$\min \frac{1}{l} \Sigma_{i=1}^{j} (y_i - f(x_i))^2 + \gamma_A \| f \|_K^2 + \gamma_I \left(\frac{1}{p^2} \Sigma_{i,j=1}^{p} (f(x_i) - f(x_j))^2 W_{ij} + \frac{1}{n^2} \Sigma_{i,j=p+1}^{p+n} (f(x_i) - f(x_j))^2 W_{ij} \right) \tag{3}$$

Data imbalance is a common problem in bioinformatics, in which negative data often have larger numbers than positive data. However, few methods have been proposed to address this problem. In this paper, we assign different penalty coefficients to different samples [24]. Therefore, SLapRLS aims to solve the optimization problem as:

$$\min \frac{1}{l} \Sigma_{i=1}^{l} c_i (y_i - f(x_i))^2 + \gamma_A \| f \|_K^2 + \gamma_I \left(\frac{1}{P^2} \Sigma_{i,j=1}^{p} (f(x_i) - f(x_j))^2 W_{ij} + \frac{1}{N^2} \Sigma_{i,j=p+1}^{p+n} (f(x_i) - f(x_j))^2 W_{ij} \right) \tag{4}$$

where parameter $c = (c_1, c_{2,} . . . c_i, . . ., c_l)$ is the penalty coefficient and c_i represents the misclassification cost of the i_{th} data x_i. The misclassification cost contains two parts: the cost of misclassifying the positive samples as negative and the cost of misclassifying the negative samples as positive. Assuming that the number of class A is larger than class B, the model tends to classify the test data as class A. If the penalty of each class is equivalent, more samples in class B may be predicted as the wrong class. To solve the problem of data imbalance, the class with a smaller number is assigned a large penalty, while the class with a larger number is assigned a small penalty. The penalty coefficients for positive and negative data are set to $n/(p + n)$ and $p/(p + n)$ according to the numbers of positive and negative results. The two tuning parameters γ_A and γ_I in Eq (4) were selected from grid research in the range of $[10^{-5}, 10^5]$ via ten-fold cross-validation [25] [26]; the values of the selected γ_A and γ_I for each kinase are listed in S1 Table. Belkin *et al.* proved that optimization problems that share similar object functions to Eq (2) were all convex optimization problems [14] and thus shared the same form as the solutions shown in (5). By using K to represent the Kernel function, we can calculate $f^*(x)$ as follows:

$$f^*(x) = \Sigma_{i=1}^{l} \alpha_i K(x_i, x) \tag{5}$$

By solving the convex optimization problem shown in Eq (4), we can achieve $\boldsymbol{a}$ as follows:

$$\boldsymbol{\alpha} = \left(\boldsymbol{K} + \gamma_A l \boldsymbol{I} + \gamma_I l \left(\frac{1}{p^2} \boldsymbol{C}^{-1} \boldsymbol{L}_{\boldsymbol{P}} \boldsymbol{K}_{\boldsymbol{P},\boldsymbol{P}} + \frac{1}{n^2} \boldsymbol{C}^{-1} \boldsymbol{L}_N \boldsymbol{K}_{N,N} \right) \right)^{-1} \boldsymbol{Y} \tag{6}$$

Here, $\boldsymbol{K}$ is the kernel matrix with a size of $(p + n)\times(p + n)$ and can be denoted as $\begin{bmatrix} \boldsymbol{K_{P,P}} & \boldsymbol{K_{P,N}} \\ \boldsymbol{K_{N,P}} & \boldsymbol{K_{N,N}} \end{bmatrix}$, $\boldsymbol{K_{A,B}}$ is a kernel matrix between two datasets A and B, and P and N represent the positive dataset and negative dataset, respectively. $\boldsymbol{L_P}$ is the graph Laplacian of all positive data, which is given by

$$\mathrm{L_P} \ = \ \mathrm{D_P}^{-\frac{1}{2}}(\mathrm{D_P} - \mathrm{W_{P,P}})\mathrm{D_P}^{-\frac{1}{2}} \tag{7}$$

where the diagonal matrix $\boldsymbol{D_P}$ is given by:

$$\mathrm{DP_{i,i}} \ = \ \Sigma_{\mathrm{x_j}\in \mathrm{P}}\mathrm{W_{i,j}}, \mathrm{x_i} \in \mathrm{P} \tag{8}$$

Similarly, $\boldsymbol{L_N}$ is the graph Laplacian of all negative data, which is given by

$$\boldsymbol{L_N} \ = \ \boldsymbol{D_N}^{-\frac{1}{2}}(\boldsymbol{D_N} - \boldsymbol{W_{N,N}})\boldsymbol{D_N}^{-\frac{1}{2}} \tag{9}$$

where the diagonal matrix $\boldsymbol{D_N}$ is given by:

$$\mathrm{DN_{i,i}} \ = \ \Sigma_{\mathrm{x_j}\in \mathrm{N}}\mathrm{W_{i,j}}, \mathrm{x_i} \in \mathrm{N} \tag{10}$$

In (6), $\boldsymbol{C}$ is a diagonal matrix given by $\boldsymbol{C} = diag(c_1,c_2...c_l)$, $\boldsymbol{Y} = (y_1,y_2...y_l)$, $\boldsymbol{W_{P,P}}$ is the similarity matrix between data in the positive dataset and $\boldsymbol{W_{N,N}}$ is the similarity matrix between samples in the negative datasets.

Similarity among samples

Because SLapRLS is based on sample similarities, the method used to calculate the similarity can have a large impact. Blosum62 is a matrix that reflects the relationship among amino acids and has been proven to be efficient for calculating pairwise similarity [11]. Here, we assume Blosum62 as matrix B and use a and b to represent two amino acids. Then, the similarity $W_{i,\,j}$ between two samples s_i and s_j can be calculated as follows:

$$\mathrm{W_{i,\,j}} \ = \ \Sigma_{\mathrm{t}\,=\,1}^{\mathrm{w}}\mathrm{sim}\,(\mathrm{s_i}(\mathrm{t}), \mathrm{s_j}(\mathrm{t})) \tag{11}$$

where w is the window size of a local peptide sequence and is set to 15 in this study. $s_i(t)$ represents the amino acid located in the t_{th} position of s_i. Because the similarity between samples should be non-negative, we normalize B using:

$$sim(a, b) \ = \ \frac{B(a, b) - \min(B)}{\max(B) - \min(B)} \tag{12}$$

$$\mathrm{sim}(\mathrm{a},\ \mathrm{b}) > 0,$$

Because $sim(s_{i,\ sj})$ is non-negative, it is easy to come to the conclusion that $W_{i,\,j}$ is also non-negative.

Kernel matrix construction

Kernel-based algorithms embed the dataset into a Hilbert space, and the kernel matrix completely reflects the relative positions of the samples in the embedding space. Several mathematically defined kernel functions exist (i.e., Gaussian kernel and spline kernel); these functions have been widely utilized in many research fields. However, these mathematically defined kernels often require few parameters and cannot effectively reflect the relationship between objects in a certain field. For example, in the field of kinase identification the Gaussian kernel needs the pairwise distance. A common way to calculate the distance is to encode the peptide sequence using quadrature encoding; then, the distance is calculated based on the Euclidean distance, which assumes that each amino acid is independent of the others. However, close relationships exist between amino acids, and thus calculating the similarity using the Gaussian kernel may miss important information from the substrate sequence [11]. Because a kernel function can reflect pairwise similarity, a more reliable way to calculate the similarity is to use expert knowledge and other information rather than the kernel function [27]. A kernel matrix should be symmetric and positive definite [28]. In this regard, we can perform translation on the similarity matrix to make it fulfill these two properties (symmetric and positive definite). The similarity matrix calculated with Eq (11) is symmetric, and thus we only need to add a small multiple to the diagonal elements of the similarity matrix to force it to be positive definite; then, the translated similarity matrix can be treated as the kernel matrix [15]. The summary of SLapRLS is shown in Fig 1, and the procedure of this work is shown in Fig 2.

Performance evaluation

To evaluate the performance of the classifiers, we calculate the specificity (*Sp*), sensitivity (*Sn*), accuracy (*Acc*), precision (*Pre*) and Matthews correlation coefficient (*Mcc*). *Sp* and *Sn* represent the ratio of correctly predicted negative and positive sites, *Acc* indicates the percentage of truly predicted sites, and *Pre* indicates the ratio of true positive sites over predicted positive sites. *Mcc* reflects the balance quality between the true and predicted classes and illustrates the correlation between the true and predicted class. The definitions of *Sn*, *Sp*, *Acc*, *Pre* and *Mcc* are shown in Eqs (13), (14), (15), (16) and (17), respectively. The receiver operating characteristic (ROC) curve is widely used to evaluate the performance of a classifier in machine learning and plots (*1-Sp*, *Sn*) using each predicted value as the threshold. The corresponding area under the ROC curve (*AUC*) represents the overall accuracy of a classifier.

$$\mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{13}$$

$$\mathrm{Sp} = \frac{\mathrm{TP}}{\mathrm{TN} + \mathrm{FP}} \tag{14}$$

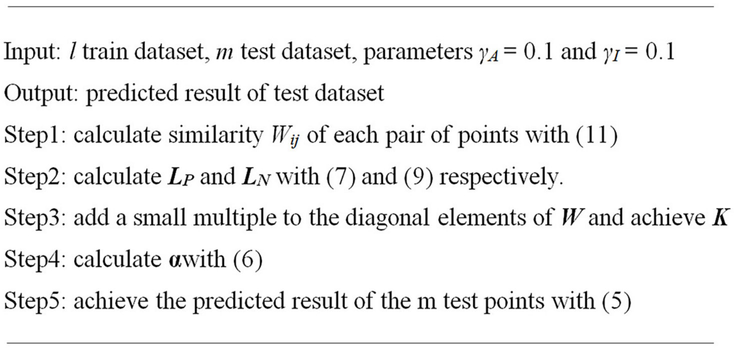
$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{15}$$

$$\mathrm{Pre} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{16}$$

$$\mathrm{Mcc} = \frac{\mathrm{TP} \times \mathrm{FP} - \mathrm{TN} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FN}) \times (\mathrm{TP} + \mathrm{FP}) \times (\mathrm{TN} + \mathrm{FN}) \times (\mathrm{TN} + \mathrm{FP})}} \tag{17}$$

where *TP*, *FP*, *TN* and *FN* represent the number of true positives, false positives, true negatives and false negatives, respectively.

AUC is used as an overall performance measurement for comparison with other algorithms, and *Acc*, *Pre* and *Mcc* are utilized to evaluate the performance when the *Sp* is extremely high. Notably, we cannot use *AUC* as the evaluation for comparison with existing tools because the prediction scores are not available for these tools. In this situation, we make the comparison using the corresponding *Sn* value with a comparable *Sp* value by selecting a suitable threshold.



```
Input: $l$ train dataset, $m$ test dataset, parameters $\gamma_A = 0.1$ and $\gamma_I = 0.1$
Output: predicted result of test dataset
Step1: calculate similarity $W_{ij}$ of each pair of points with (11)
Step2: calculate $\boldsymbol{L_P}$ and $\boldsymbol{L_N}$ with (7) and (9) respectively.
Step3: add a small multiple to the diagonal elements of $W$ and achieve $\boldsymbol{K}$
Step4: calculate $\boldsymbol{\alpha}$with (6)
Step5: achieve the predicted result of the m test points with (5)
```

Fig 1. A summary of SLapRLS.

doi:10.1371/journal.pone.0139676.g001

Results and Discussion

Comparison with other algorithms

We first compare our method with three existing algorithms (SVM, BDT and KNN) with a 10-fold cross validation using local peptide sequence information. When using SVM, peptide sequences are coded into numeric features using a quadrature encoder. In this study, we adopt the LIBSVM with an RBF kernel function [29], and parameters C andγin SVM are chosen by cross-validation. For BDT, the method introduced in PPSP [10] is adopted in this work. In KNN, the parameter K is set to 11 and the Blosum 62 matrix is employed to calculate the distance d among samples. Assuming S_{i+} is the total similarity between sample s_i and the top K nearest positive samples and S_{i-} is total similarity between sample s_i and the top K nearest negative samples, the final predicted score of sample s_i is denoted as the ratio of S_{i+} and S_{i-} [11].

The ROC curve is utilized to compare these four algorithms. The ROC curves for the Erk2, Erk1, CDC2 and PKC alpha kinases are shown in Fig 3, with the red line, blue line, yellow line and cyan line representing SLapRLS, SVM, BDT and KNN, respectively. As shown in Fig 3, the red line outperformed the other three lines, indicating that SLapRLS achieved better overall performance than the other algorithms.

To illustrate the robustness of our proposed method, we repeat the 10-fold cross validation five times and then compare the $AUCs$. Detailed results are listed in Table 1. As expected,

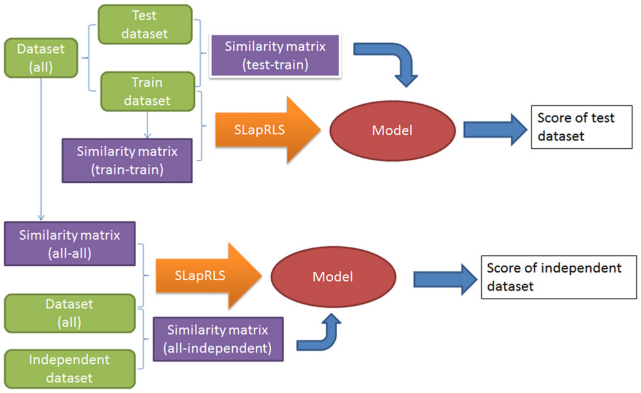

Test dataset
Dataset (all)
Train dataset
Similarity matrix (test-train)
SLapRLS
Model
Score of test dataset
Similarity matrix (train-train)
Similarity matrix (all-all)
SLapRLS
Model
Score of independent dataset
Dataset (all)
Similarity matrix (all-independent)
Independent dataset

Fig 2. Procedure of this work. Firstly, label dataset are derived from Phospho.ELM, and it is split into train dataset and test dataset. Secondly, the model is developed using train dataset and its similarity matrix with SLapRLS, with which the predicted result of test dataset is achieved. Additionally, an independent test dataset is used. The model that predicts the independent dataset is developed with all the label dataset.

doi:10.1371/journal.pone.0139676.g002

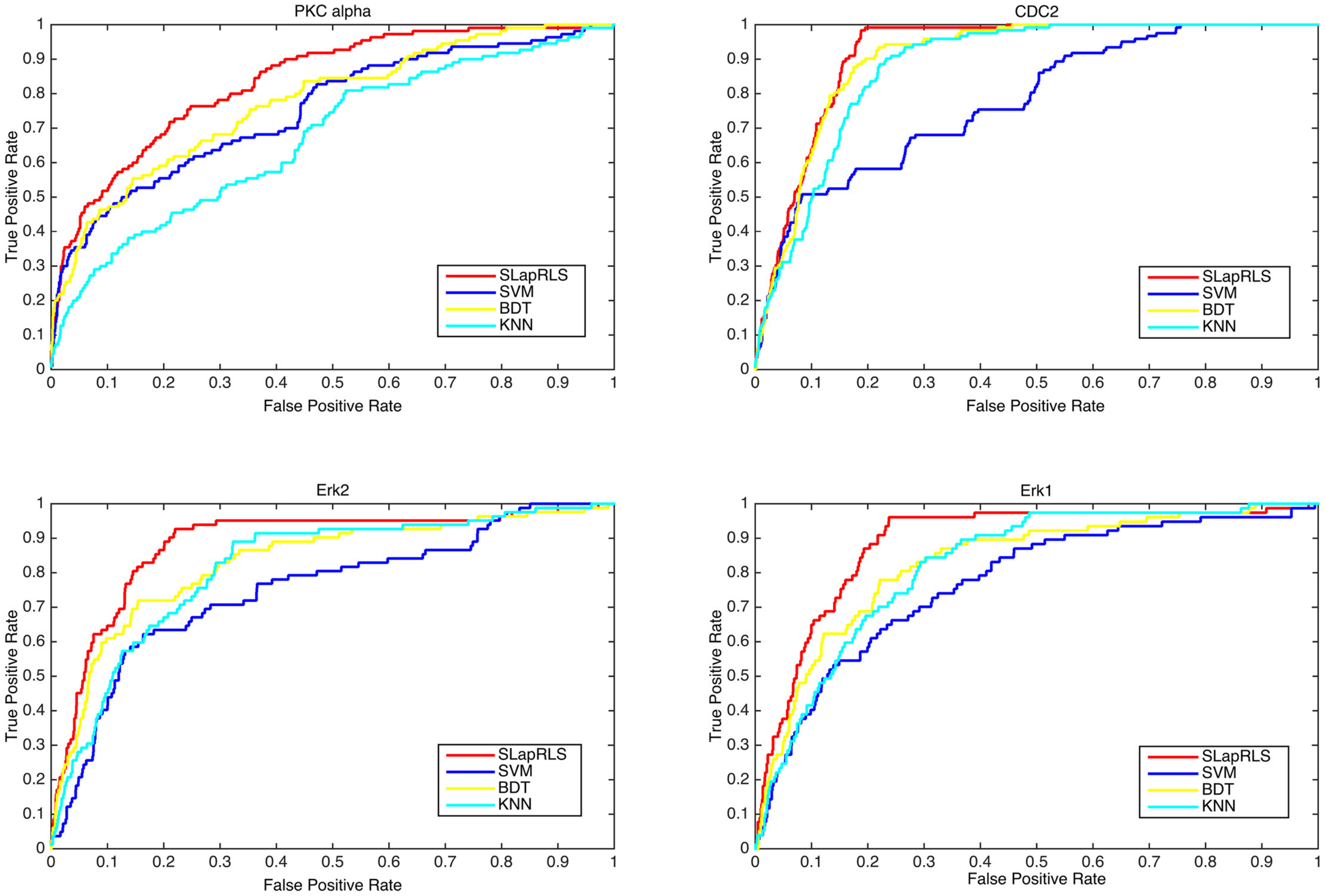


Fig 3. ROC curves of different algorithms. ROC curves of kinase Erk2, Erk1, CDC2 and PKC alpha achieved by four different algorithms are plotted. The red line, blue line, yellow line and cyan line represent SLapRLS, SVM, BDT and KNN, respectively.

doi:10.1371/journal.pone.0139676.g003

SLapRLS achieves better performance than the other three algorithms on S/T/Y substrate kinases. For example, the average AUCs achieved by SLapRLS on kinase PKC alpha are 7.7%, 5.7% and 14.2% higher than SVM, BDT and KNN, respectively.

Additionally, *Sn*, *Acc*, *Pre* and *Mcc* are utilized to evaluate the performance of the four algorithms at a high stringency level (*Sp* = 0.99). The phosphorylated S/T and Y site kinases are divided into two groups (i.e., S/T substrate kinases and Y substrate kinases), and the performance of the two kinase groups are plotted in Fig 4. The results show that SLapRLS achieves higher *Sn*, *Pre*, *Acc* and *Mcc* and a slightly higher *Sp*. For example, the average *Sn* and *Pre* achieved by SLapRLS on Y substrate kinases are more than 2% and 9% higher than the other algorithms. Table 1 and Fig 4 show that SLapRLS also achieves better performance in S/T substrate kinases than Y substrate kinases.

Because SLapRLS relies on the similarity between samples, a good similarity calculator is essential to achieve a satisfying performance. In this work, the similarity between samples is represented by peptide similarity, which is calculated using sequence information. The conservation is strong in the S/T substrate kinases but weak in the Y substrate kinases [30]. Fig 5 shows that the amino acid distributions of two S/T substrate kinases (ATM and ck2_alora2) have stronger conservation than two Y substrate kinases (INSR and EGFR). As shown in Fig 5,

Table 1. Compared AUC values of the four algorithms: SLapRLS, SVM, BDT and KNN.

Methods	SLapRLS	SVM	BDT	KNN
		S/T		
PKC alpha	**0.833LSBDT**	0.756LSBDT	0.776LSBDT	0.691LSBDT
ATM	**0.964LSBDT**	0.964LSBDT	0.905LSBDT	0.876L0.006
CDC2	**0.918L0.006**	0.714L0.006	0.894L0.006	0.880L0.006
Erk2	**0.882L0.006**	0.733L0.006	0.825L0.006	0.816L0.006
Erk1	**0.888L0.006**	0.760L0.006	0.831L0.006	0.815L0.006
AurA	**0.782L0.006**	0.652L0.006	0.724L0.006	0.718L0.006
BARK1	**0.747L0.006**	0.659L0.006	0.566±0.023	0.644±0.023
CaMK2a	**0.862a0.023**	0.682a0.023	0.688a0.023	0.757a0.023
CDK2	**0.887a0.023**	0.783a0.023	0.780a0.023	0.833a0.023
GSK3B	**0.806a0.023**	0.694a0.023	0.758a0.023	0.704a0.023
Ck2 a1ora2	**0.95420.023**	0.93220.023	0.93820.023	0.93020.023
MAPKAPK2	0.651PK2023	0.502PK2023	0.500PK2023	**0.716PK2023**
PDK1	**0.854PK2023**	0.787PK2023	0.811PK2023	0.842PK2023
P38a	**0.838PK2023**	0.708PK2023	0.781PK2023	0.732PK2023
PLK1	**0.734PK2023**	0.655PK2023	0.606PK2023	0.594PK2023
		Y		
ABL1	0.560PK2023	**0.609PK2023**	0.467PK2023	0.519±0.018
EGFR	0.559±0.018	0.460±0.018	**0.594±0.018**	0.530±0.018
FYN	**0.730±0.018**	0.629±0.018	0.650±0.018	0.683±0.018
INSR	**0.547±0.018**	0.537±0.018	0.427±0.018	0.506±0.018
LCK	**0.638±0.018**	0.573±0.018	0.589±0.018	0.634±0.018
LYN	0.619±0.027	0.641 0.046	0.560±0.022	**0.672±0.022**
SRC	0.564±0.022	**0.570±0.022**	0.509±0.022	0.548±0.022
SYK	0.718±0.022	**0.726±0.022**	0.677±0.022	0.658±0.022

doi:10.1371/journal.pone.0139676.t001

substrates of kinases with good performances tend to exhibit strong conservation, whereas the conservation of kinases with bad performances is weak. Therefore, local peptide sequence similarity may not effectively reflect the similarity between samples for kinases that have weak sequence conservation, and thus all four algorithms tend to achieve better performance for the S/T substrate kinases than the Y substrate kinases. It should also be noted that although only sequence information is used as an input feature, SLapRLS could actually address any type of data that can be used to calculate similarities between samples.

Comparison with existing tools

To evaluate the performance of SLapRLS, we compare it with two existing kinase identification tools: iGPS and NetworKIN. Because cross validation is not available for iGPS and NetworKIN, we adopt an independent test dataset. We also compare *Sn* at the same high stringency *Sp* level using different algorithms. Comparison results are shown in Table 2. Although SLapRLS only uses sequence information as a feature while both iGPS and NetworKIN use protein-protein interaction information to filter the results, SLapRLS still achieves a satisfying performance (5 out of 6 kinases have a better performance compared with iGPS and NetworKIN). For instance, iGPS achieves an *Sp* of 0.780 and *Sn* of 0.525 for PKC alpha. To make a reasonable comparison with iGPS, we set the threshold to ensure that SLapRLS has a comparable *Sp* (0.782) and then calculate the corresponding *Sn* (0.983). The results show that the *Sn* of SLapRLS is 46% higher than *iGPS*. Thus, SLapRLS can effectively identify the corresponding kinase of the new site.

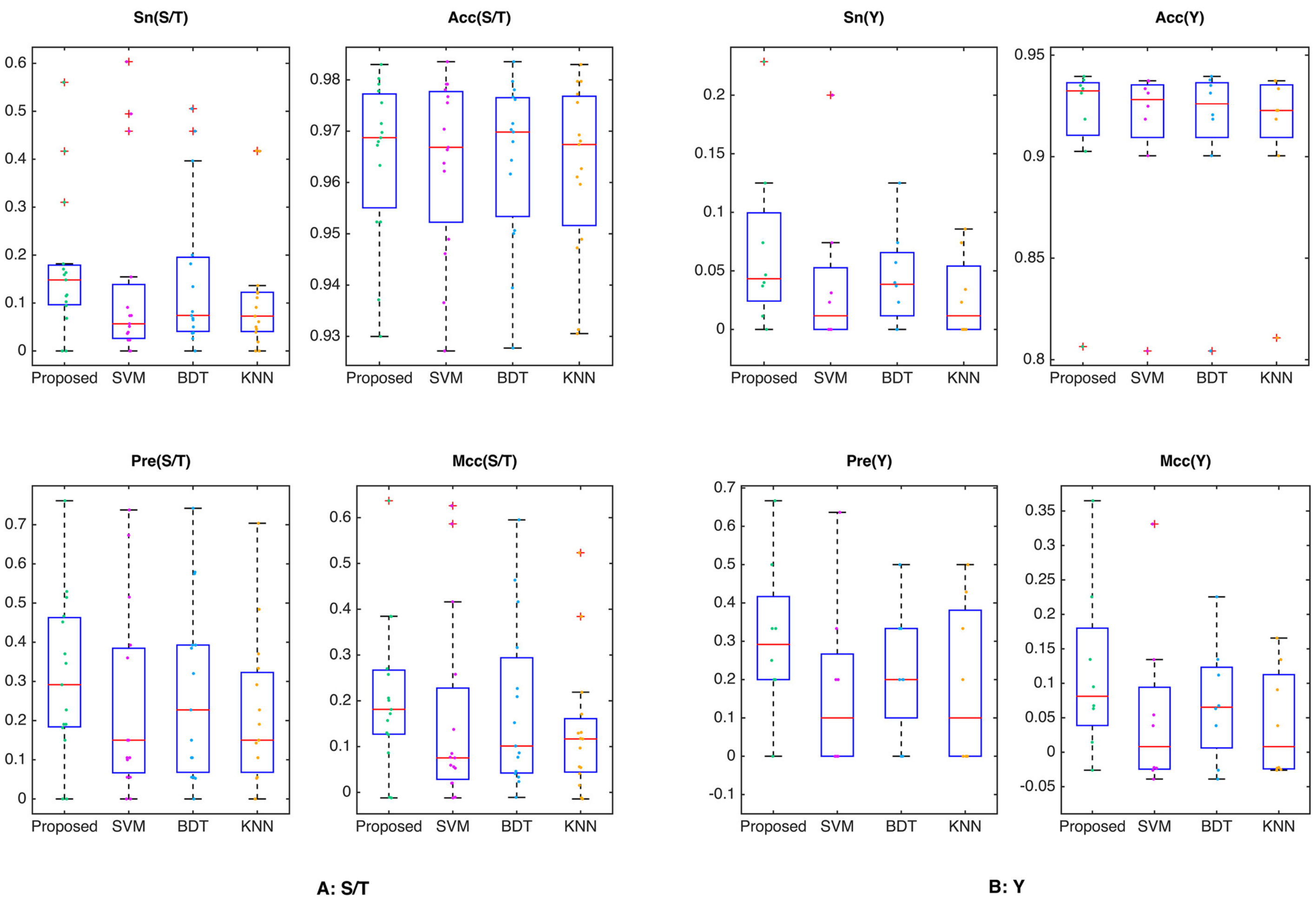


Fig 4. Comparison of the four algorithms at high stringency level (*sp = 0.99*). The *Sn*, *Sp*, *Acc* and *Mcc* values at high stringency level (*sp = 0.99*) of four algorithms on the S/T and Y kinases.

doi:10.1371/journal.pone.0139676.g004

A case study

To illustrate the usefulness of SLapRLS and to better elucidate the biological mechanism underlying phosphorylation, we perform an enrichment analysis on phosphorylation sites and their corresponding kinases. We adopt the kinase PKC alpha to illustrate the capability of SLapRLS to discover new phosphorylation sites. A total of 110 sites on 66 proteins phosphorylated by PKC alpha are extracted from the Phospho.ELM database. Additionally, we perform cross-validation on PKC alpha with SLapRLS and predict 56 candidate sites with the top 100 predicted

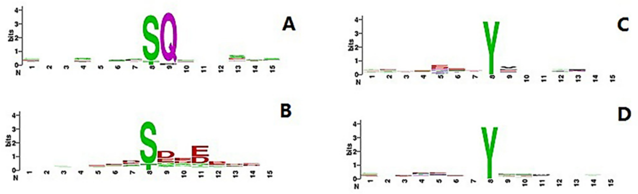


Fig 5. Weblogos of S/T substrate kinases and Y substrate kinases. A and C are the Weblogos of kinase ATM and ck2 alora2 and B and D are the Weblogos of kinase EGFR and INSR.

doi:10.1371/journal.pone.0139676.g005

Table 2. Comparison among SLapRLS, iGPS and NetworKIN on independent test data.

Methods	iGPS		SLapRLS		NetworKIN		SLapRLS	
	Sp	*Sn*	*Sp*	*Sn*	*Sp*	*Sn*	*Sp*	*Sn*
PKC alpha	0.780	0.525	**0.782**	**0.983**	0.997	0.475	**0.997**	**0.559**
Erk2	0.466	0.709	**0.471**	**0.993**	0.865	0.278	**0.870**	**0.329**
Erk1	0.508	0.709	**0.510**	**0.974**	0.939	0.222	**0.942**	**0.247**
P38a	0.367	0.703	**0.369**	**0.865**	**1.000**	0.027	0.933	**0.054**
SRC	0.300	**0.875**	**0.310**	0.867	0.300	0.100	**0.310**	**0.867**
SYK	0.283	0.850	**0.301**	**0.850**	0.830	0.300	**0.849**	**0.400**

doi:10.1371/journal.pone.0139676.t002

Table 3. Pathway enrichment analysis of PKC alpha.

Terms	Count	P-value	Benjamini P-values
Regulation of phosphorylation	14(6)	1.4e-7	9.2e-5
Cell migration	10(3)	4.2e-6	7.7e-4
Learning or memory	5(2)	1.7e-3	1.8e-2
Regulation of heart contraction	3(2)	2.4e-3	3.6e-2

doi:10.1371/journal.pone.0139676.t003

scores. Enrichment analysis of the combined known and predicted proteins is performed using DAVID [31] to identify enriched pathways. As shown in Table 3, 6 KEGG pathways are highly enriched, with the most significant pathway related to the regulation of phosphorylation, and the corresponding Benjamin P-value is 9.2E-5. Additionally, 6 proteins in this pathway are predicted as substrates of PKC alpha by SLapRLS but are not included in Phospho.ELM, indicating that SLapRLS is able to identify potential corresponding kinases for known phosphorylation sites.

Conclusion

Phosphorylation plays an important role in multiple biological processes, and protein kinases have a tight relationship with many kinds of diseases. Thus, identifying the corresponding kinases for known phosphorylation sites is important. To overcome shortcomings such as costly and time-consuming experimental methods, the development of computational methods for kinase identification is urgently needed. At present, existing phosphorylation prediction-related computational methods neglect the geometry of the data distribution, and most kernel-based methods are based on distance. These distance-based methods often assume that the amino acids are independent when using local protein sequence information to calculate distances, while the connections between amino acids, which are also very important in expressing the relationships among samples, are neglected.

In this work, we propose the kernel-based algorithm SLapRLS that relies on similarity rather than distance and introduce the inconsistency between label and pairwise similarity to reflect the geometric distribution of the data. Instead of optimizing a single objective function, SLapRLS optimizes two functions: minimizing structure risk and the overall inconsistency between label and pairwise similarity. Because the kernel function reflects the closeness of two samples, we translate the kernel matrix from the similarity matrix instead of any famous kernel functions. The results show that SLapRLS outperforms three other algorithms (SVM, BDT and KNN) and two existing kinase identification tools (iGPS and NetworKIN).

It should to be noted that SLapRLS is based on a similarity matrix. Although only local sequence information is used as a feature in this work, SLapRLS is able to address any type of data as a feature, including characteristic types and numerical types. Although SLapRLS can solve kinase identification problems efficiently, there is also room for further improvement. For example, we only focus on local sequence information and disregard all other biological information. However, it has been proven that protein function and structural information is also useful for phosphorylation predictions [30, 32]. Therefore, such information could be utilized for kinase identification in a future work. To this end, a combination regulation should be introduced to incorporate different types of features.

Supporting Information

S1 Table. The selected parameters γ_A and γ_I for each kinase.
(XLSX)

Acknowledgments

This work was supported by grants from National Natural Science Foundation of China (61471331,61101061).

Author Contributions

Conceived and designed the experiments: AL MW. Performed the experiments: AL XX HZ. Analyzed the data: XX HZ. Contributed reagents/materials/analysis tools: XX HZ. Wrote the paper: ZH XX AL.

References

1. Lou Y, Yao J, Zereshki A, Dou Z, Ahmed K, Wang H, et al. (2004) NEK2A interacts with MAD1 and possibly functions as a novel integrator of the spindle checksample signaling. J Biol Chem 279 (19):20049–20057. PMID: 14978040
2. Schafmeier T, Haase A, Kaldi K, Scholz J, Fuchs M, Brunner M (2005) Transcriptional feedback of neurospora circadian clock gene by phosphorylation-dependent inactivation of its transcription factor. Cell 122(2):235–246. PMID: 16051148
3. Singh CR, Curtis C, Yamamoto Y, Hall NS, Kruse DS, He H, et al. (2005) Eukaryotic translation initiation factor 5 is critical for integrity of the scanning preinitiation complex and accurate control of GCN4 translation. Mol Cell Biol 25(13):5480–5491. PMID: 15964804
4. Pawson T (2004) Specificity in signal transduction: from phospho-tyrosine-SH2 domain interactions to complex cellular systems. Cell 116(2):191–203 PMID: 14744431
5. Wood CD, Thornton TM, Sabio G, Davis RA, Rincon M (2009) Nuclear localization of p38 MAPK in response to DNA damage. Int J Biol Sci 5(5):428. PMID: 19564926
6. Ma L, Chen Z, Erdjument-Bromage H, Tempst P, Pandolfi PP (2005) Phosphorylation and functional inactivation of TSC2 by Erk: implications for tuberous sclerosisand cancer pathogenesis. Cell 121: 179–193. PMID: 15851026
7. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S (2002). The protein kinase complement of the human genome. Science. 298(5600):1912–1934. PMID: 12471243
8. Beausoleil SA, Ville´ n J, Gerber SA, Rush J, Gygi SP (2006) A probability-based approach for high-throughput protein phos-phorylation analysis and site localization. Nat Biotechnol 24(10): 1285–1292. PMID: 16964243
9. Blom N, Gammeltoft S, Brunak S (1999) Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. J Mol Biol 294(5):1351–1362. PMID: 10600390
10. Xue Y, Li A, Wang L, Feng H, Yao X (2006) PPSP: Prediction of PK-specific phosphorylation site with Bayesian decision theory. BMC Bioinform 7(1):163.
11. Gao J, Thelen JJ, Dunker AK, Xu D (2010) Musite, a tool for global prediction of general and kinase-specific phosphorylation sites. Mol Cell Proteomics 9(12):2586–2600. doi: 10.1074/mcp.M110.001388 PMID: 20702892

12. Linding R, Jensen LJ, Pasculescu A, Olhovsky M, Colwill K, Bork P (2008) NetworKIN: a resource for exploring cellular phosphorylation networks. Nucleic acids research 36: D695–D699. PMID: 17981841

13. Song C, Ye M, Liu Z, Cheng H, Jiang X, Han G (2012) Systematic analysis of protein phosphorylation networks from phosphoproteomic data. Molecular & Cellular Proteomics 11: 1070–1083. doi: 10.1016/j.theriogenology.2015.08.009

14. Belkin M, Niyogi P, Sindhwani V (2006) Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. The Journal of Machine Learning Research 7: 2399–2434.

15. van Laarhoven T, Nabuurs SB, Marchiori E (2011) Gaussian interaction profile kernels for predicting drug–target interaction. Bioinformatics 27: 3036–3043. doi: 10.1093/bioinformatics/btr500 PMID: 21893517

16. Mallapragada PK, Jin R, Jain AK, Liu Y (2009) Semiboost: Boosting for semi-supervised learning. Pattern Analysis and Machine Intelligence, IEEE Transactions on 31: 2000–2014.

17. Dinkel H, Chica C, Via A, Gould CM, Jensen LJ, Gibson TJ (2011) Phospho. ELM: a database of phosphorylation sites—update 2011. Nucleic acids research 39: D261–D267. doi: 10.1093/nar/gkq1104 PMID: 21062810

18. Kim JH, Lee J, Oh B, Kimm K, Koh I (2004) Prediction of phosphorylation sites using SVMs. Bioinformatics 20: 3179–3184. PMID: 15231530

19. Newman RH, Hu J, Rho HS, Xie Z, Woodard C, Neiswinger J (2013) Construction of human activity-based phosphorylation networks. Molecular systems biology 9: 655. doi: 10.1038/msb.2013.12 PMID: 23549483

20. Zhang P, Peng J. SVM vs regularized least squares classification; 2004. IEEE. pp. 176–179.

21. Sewell Martin, Structural Risk Minimization, 2008.

22. Kadri H, Rabaoui A, Preux P, Duflos E, Rakotomamonjy A (2013) Functional Regularized Least Squares Classification with Operator-valued Kernels. arXiv preprint arXiv:13012655.

23. Miettinen K, Ruiz F, Wierzbicki AP (2008) Introduction to multiobjective optimization: interactive approaches. Multiobjective Optimization: Springer. pp. 27–57.

24. Wu J-S, Zhou Z-H (2013) Sequence-Based Prediction of microRNA-Binding Residues in Proteins Using Cost-Sensitive Laplacian Support Vector Machines. Computational Biology and Bioinformatics, IEEE/ACM Transactions on 10: 752–759.

25. Gómez-Chova L, Camps-Valls G, Munoz-Mari J, Calpe J (2008) Semisupervised image classification with Laplacian support vector machines. Geoscience and Remote Sensing Letters, IEEE 5: 336–340.

26. Yang L, Yang S, Jin P, Zhang R (2014) Semi-supervised hyperspectral image classification using spatio-spectral laplacian support vector machine. Geoscience and Remote Sensing Letters, IEEE 11: 651–655.

27. Lu F, Keleş S, Wright SJ, Wahba G (2005) Framework for kernel regularization with application to protein clustering. Proceedings of the National Academy of Sciences of the United States of America 102: 12332–12337. PMID: 16109767

28. Lanckriet GRG, Cristianini N, Bartlett P, Ghaoui LE, Jordan MI (2004) Learning the kernel matrix with semidefinite programming. The Journal of Machine Learning Research 5: 27–72.

29. Chang C-C, Lin C-J (2011) LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) 2: 27.

30. Xu X, Li A, Zou L, Shen Y, Fan W, Wang M (2014) Improving the performance of protein kinase identification via high dimensional protein–protein interactions and substrate structure data. Molecular BioSystems 10: 694–702. doi: 10.1039/c3mb70462a PMID: 24448631

31. Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane H.C (2003) DAVID: database for annotation, visualization, and integrated discovery. Genome biol 4: P3. PMID: 12734009

32. Fan W, Xu X, Shen Y, Feng H, Li A, Wang M (2014) Prediction of protein kinase-specific phosphorylation sites in hierarchical structure using functional information and random forest. Amino acids 46: 1069–1078. doi: 10.1007/s00726-014-1669-3 PMID: 24452754