Sun, Mengwei; Tiwari, Prayag; Qian, Yuqin; Ding, Yijie; Zou, Quan

MLapSVM-LBS: Predicting DNA-binding proteins via a multiple Laplacian regularized support vector machine with local behavior similarity

Published in:
KNOWLEDGE-BASED SYSTEMS

DOI:
10.1016/j.knosys.2022.109174

Published: 17/08/2022

Document Version
Publisher's PDF, also known as Version of record

Knowledge-Based Systems 250 (2022) 109174

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

MLapSVM-LBS: Predicting DNA-binding proteins via a multiple Laplacian regularized support vector machine with local behavior similarity

Mengwei Sun [a,c], Prayag Tiwari [b,*], Yuqin Qian [c], Yijie Ding [a,*], Quan Zou [a,d,*]

[a] Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, Quzhou, 324000, PR China
[b] Department of Computer Science, Aalto University, Espoo, Finland
[c] School of Electronic and Information Engineering, Suzhou University of Science and Technology, Suzhou 215009, PR China
[d] Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, 610054, PR China

ARTICLE INFO

Article history:
Received 18 February 2022
Received in revised form 27 May 2022
Accepted 28 May 2022
Available online 6 June 2022

Dataset link: https://github.com/prayagtiwari/MLapSVM-LBS.git, https://figshare.com/s/56182f00f08d7ef49b18

MSC:
00-01
99-00

Keywords:
DNA-binding proteins
Laplacian support vector machine
Multiple view
Protein feature extraction
Sequence classification

ABSTRACT

DNA-binding proteins (DBPs) are of great significance in many basic cellular processes. Experiment-based methods for identifying DBPs are costly and time-consuming. To deal with large-scale DBP identification tasks, a variety of computation-based methods have been developed. Inspired by previous work, we propose a multiple Laplacian regularized support vector machine with local behavior similarity (MLapSVM-LBS) to predict DBP. We serially combine three features that are extracted from protein sequences (including PsePSSM, GE, NMBAC) and feed them into MLapSVM-LBS. Based on human behavior learning theory, MLapSVM-LBS can better represent the relationship between samples through local behavior similarity. We introduce a new edge weight calculation method that takes label information into consideration. In addition, a local distribution parameter reflecting the underlying probability distribution of a sample's neighborhood is also employed. To further improve the robustness of the model, we utilize multiple Laplacian regularization to build a multigraph model in which five Laplacian graphs are constructed with local behavior similarity by changing the neighborhood size. To appraise the performance of our model, MLapSVM-LBS is trained and tested on the PDB186, PDB1075, PDB2272 and PDB14189 datasets. On two independent testing sets (PDB186 and PDB2272), our method reaches the accuracies of 0.887 and 0.712, respectively. The good results on both datasets demonstrate the reliable performance of our model.

1. Introduction

DNA-binding proteins consist of a large class of proteins that physically attach to DNA. These proteins play an important role in a number of major cellular processes, including DNA transcription, replication, recombination and transposition. There are several experiment-based biological methods for identifying DBPs, such as nuclear magnetic resonance (NMR), X-ray diffraction crystallography, filter-binding assays, chromatin immunoprecipitation and the yeast one-hybrid system (YIH). However, these traditional methods are highly costly and time-consuming. With the rapid development of biology, such methods cannot deal with large-scale DBP identification.

In the biomedicine and bioinformatics fields, machine learning methods have been widely used and have obtained good results. Examples include O-GlcNAcylation site prediction [1], electron transport protein identification [2], protein remote homology detection [3], protein crystallization prediction [4], protein subcellular localization detection [5,6], drug–target interaction prediction [7–9], drug–drug interaction identification [10,11], and potential disease-associated microRNA detection [12–17]. To solve the problems mentioned above, machine learning methods that can reduce the considerable consumption of resources and time are commonly implemented to detect DBP [18–20]. Some researchers employed structural information of proteins to identify DBP. Guy Nimrod et al. [21] built a random forest model for identifying DBP via the average surface electrostatic potential, amino acid conservation pattern information and dipole moment. Based on a support vector machine, Bhardwaj et al. [22] utilized three types of features (overall charge, surface patches and composition features) to develop a predictive model. By combining structural information and evolutionary information, Shahana Yasmin Chowdhury et al. [23] proposed a model called iDNAProt-ES to detect DBPs. Ahmad et al. [24] employed a neural network model to detect DBPs. Three types of features

* Corresponding authors.
E-mail addresses: smw010207@163.com (M. Sun), prayag.tiwari@aalto.fi (P. Tiwari), usts_qyq@qq.com (Y. Qian), wuxi_dyj@csj.uestc.edu.cn (Y. Ding), zouquan@nclab.net (Q. Zou).

https://doi.org/10.1016/j.knosys.2022.109174

are fed into the model, including the net charge of the protein, electric dipole moment and fourth-moment tensor.

The structural information of proteins is difficult to obtain, and most of them remain unknown. In reality, sequence-based methods are more effective in many practical applications. DNA methylation sites, recombination spots, and posttranslational modification (PTM) sites (protein) were detected by sequential methods. Proteins with similar sequences tend to have similar structural information. Consequently, it is highly possible to utilize sequence-based computational methods to identify DBPs. During the past decade, a massive number of sequence-based computational models have been developed to identify DBPs. Cai and Liu [25–27] extracted amino acid composition and pseudo amino composition (PseACC) from the protein sequence and fed them into a support vector machine. Based on the position specificity score matrix (PSSM), which was generated by PSI-BLAST software [28], Kumar and Liu [29] established a predictor named DNAbinder. Leyi Wei et al. [18] developed a predictive model named local-DPP by applying local conservation information of PSSM features to the random forest classifier. By means of Chou's five-step rule, Zou et al. [30] built a fuzzy kernel ridge regression model based on multiview sequence features (FKRR-MVSF) to detect DBPs. To take advantage of multiple protein sequence features, Ding et al. [31] proposed a multikernel support vector machine model through heuristic kernel alignment (MKSVM-HKA).

Inspired by previous work, we develop a novel DBP identification model called the multiple Laplacian regularized support vector machine with local behavior similarity (MLapSVM-LBS). Based on human behavior learning theory, we utilize local behavior similarity (LBS) to construct the adjacent matrix of samples. Specifically, we apply label information to the edge weight calculation and introduce a local distribution parameter to reflect the underlying probability distribution of a sample's neighborhood. Furthermore, we construct five Laplacian graphs by changing the number of nearest neighbors k to make the model less sensitive to the neighborhood size. Moreover, we combine three features extracted from the protein sequence, including PsePSSM, GE, and NMBAC, and feed them into the MLapSVM-LBS model. Compared with other methods, our model achieves better results reaching accuracies of 0.887 and 0.712 on PDB186 and PDB2272, respectively.

The contributions of our study include: (1) We employ an adjacent matrix calculation method named Local Behavior Similarity (LBS). Combined with human cognitive features, LBS can better characterize relationship between samples. (2) We propose Multiple Laplacian regularized LapSVM to make the model less sensitive to the value of nearest neighbors k. (3) An iterative optimization method is utilized to solve the objective function.

2. Method and materials

DBP prediction is a classical binary classification problem. To precisely detect DBPs, it is important to utilize methods of feature extraction and train a model that has good generalization. In our work, we utilize three feature extraction methods to represent protein features: pseudo-position specific scoring matrix (PsePSSM) [26], normalized Moreau–Broto autocorrelation (NMBAC) [32] and global encoding (GE) [33]. Based on human behavior learning theory, we utilize local behavior similarity(LBS), which applies label information to edge weight calculations and introduces a local distribution parameter to reflect the underlying distribution of a sample's neighborhood when constructing the adjacent matrix. To further enhance the performance of our model, multiple Laplacian regularization built with LBS is employed to develop the multigraph LapSVM model. The framework of our method is shown in Fig. 1.

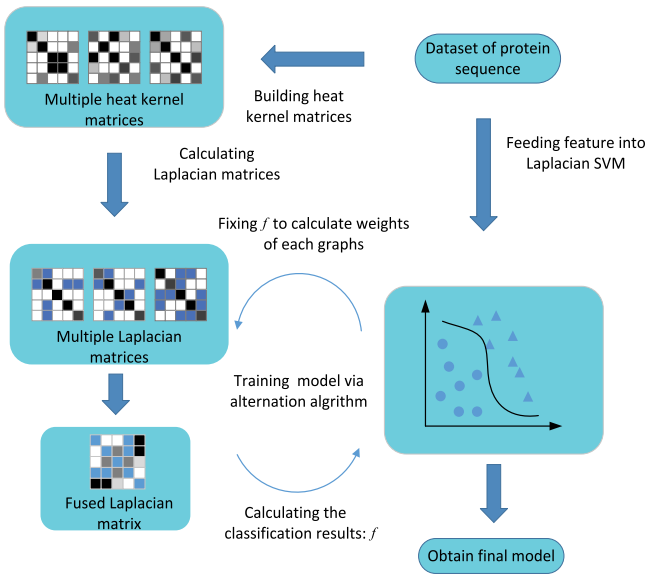

Fig. 1. The framework of MLapSVM-LBS.

Table 1
The details on four benchmark datasets.

Datasets	Number of negative	Number of positive
PDB1075	550	525
PDB186	93	93
PDB14189	7060	7129
PDB2272	1119	1153

2.1. Benchmark datasets

In our work, four datasets, including PDB1075, PDB186, PDB14189 and PDB2272 are employed to test the performance of our model. The details of the datasets are listed in Table 1. PDB1075 and PDB186 were obtained from Liu [25] and Lou [34]. PDB14189 and PDB2272 were constructed by Du [35]. All datasets were selected from the PDB bank.

2.2. Related work

The Laplacian support vector machine (LapSVM) [36] is a well-performing semi-supervised learning algorithm proposed by Mikhai Belkin. It successfully applies manifold regularization, which contains the geometric information of labeled and unlabeled samples, to the support vector machine(SVM) [37].

Let us consider the labeled sample set $\{(x_i, y_i)\}_{i=1}^{l}$ and unlabeled sample set $\{(x_i)\}_{i=l+1}^{l+u}$, where $x_i \in R^d$, labels $y_i \in \{+1, -1\}$. LapSVM uses reproducing kernel Hilbert space(RKHS) as hypothesis space. The objective function of LapSVM containing kernel function and loss function can be defined as follows:

$$f^* = \arg\min_{f \in H_K} \frac{1}{l} \sum_{i=1}^{l} (1 - y_i f(x_i)) + \gamma_A \|f\|_K^2 + \frac{\gamma_I}{(l+u)^2} f^T L f \qquad (1)$$

where $\frac{1}{l} \sum_{i=1}^{l} (1 - y_i f(x_i))$ denotes the hinge loss function of l labeled samples, $\|f\|_K^2$ is the standard regularization in RKHS aiming to maintain the smoothness of the solution, and L is the Laplacian matrix, γ_A and γ_I are the accommodation coefficients.

The solution to the formula can be shown as follows:

$$f^*(x) = \sum_{i=1}^{l+u} \alpha_i^* K(x, x_i) \qquad (2)$$

Therefore, the LapSVM optimization problem is equivalent to:

$$\min_{\alpha\in R^{l+u},\xi\in R^{l}} \frac{1}{l}\sum_{i=1}^{l}\xi_i + \gamma_A\alpha^T K\alpha + \frac{\gamma_I}{(l+u)^2}\alpha^T KLK\alpha \tag{3a}$$

$$s.t. \quad \left(\sum_{j=1}^{l+u}\alpha_j K\left(x_i, x_j\right) + b\right) \geq 1-\xi_i \tag{3b}$$

$$\xi_i > 0, i = 1, 2, \ldots, l, \tag{3c}$$

where $K\left(x_i, x_j\right)$ denotes the kernel function, the $(l+u)\times(l+u)$ matrix K is the kernel matrix of both labeled and unlabeled samples.

Introduce the Lagrangian multipliers β_i, ζ_i:

$$L(\alpha, \xi, b, \beta, \zeta) = \frac{1}{2}\alpha^T\left(2\gamma_A K + 2\frac{\gamma_I}{(l+u)^2}KLK\right)\alpha \tag{4a}$$

$$- \sum_{i=1}^{l}\beta_i\left(y_i\left(\sum_{j=1}^{l+u}\alpha_j K\left(x_i, x_j\right) + b\right) - 1 + \xi_i\right) \tag{4b}$$

$$+ \frac{1}{l}\sum_{i=1}^{l}\xi_i - \sum_{i=1}^{l}\xi_i\zeta_i \tag{4c}$$

Calculate the first-order partial derivative of b and ξ_i in Eq. (4):

$$\frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{i=1}^{l}\beta_i y_i = 0 \tag{5a}$$

$$\frac{\partial L}{\partial \xi_i} = 0 \Rightarrow \frac{1}{l} - \beta_i - \zeta_i = 0 \tag{5b}$$

$$\Rightarrow 0 \leq \beta_i \leq \frac{1}{l} \tag{5c}$$

We integrate Eq. (4) and Eq. (5) to obtain the simplified expression:

$$\begin{aligned} L_g(\alpha,\beta) = \ & \tfrac{1}{2}\alpha^T\left(2\gamma_A K + 2\tfrac{\gamma_I}{(u+l)^2}KLK\right)\alpha \\ & - \alpha^T KJ^T Y\beta + \sum_{i=1}^{l}\beta_i \end{aligned} \tag{6}$$

where $J = [I\ 0]$ is a $l\times(l+u)$ matrix, I is a $l\times l$ identity matrix, and $Y = diag(y_1, y_2, \ldots, y_l)$.

Calculate the derivative of α :

$$\frac{\partial L_g}{\partial \alpha} = \left(2\gamma_A K + 2\frac{\gamma_I}{(l+u)^2}KLK\right)\alpha - KJ^T Y\beta \tag{7}$$

We can obtain the optimal solution:

$$\alpha^* = \left(2\gamma_A I + 2\frac{\gamma_I}{(u+l)^2}LK\right)^{-1} J^T Y\beta^* \tag{8}$$

Substitute α^* in Eq. (6), and the Lagrangian dual problem can be obtained:

$$\beta^* = \max_{\beta\in R^l}\sum_{i=1}^{l}\beta_i - \frac{1}{2}\beta^T Q\beta \tag{9a}$$

$$s.t. \quad \sum_{i=1}^{l} y_i\beta_i = 0 \tag{9b}$$

$$0 \leq \beta_i \leq \frac{1}{l} \quad i = 1, 2, \ldots, l \tag{9c}$$

where $Q = YJK\left(2\gamma_A I + 2\frac{\gamma_I}{(l+u)^2}LK\right)^{-1}J^T Y$. Note that LapSVM is equal to traditional SVM when $\gamma_I = 0$. The overview of LapSVM is shown in Algorithm 1.

Algorithm 1 Procedure of LapSVM.

Input:

Training set $\{x_i, y_i\}_{i=1}^{l}$, test set $\{x_i\}_{i=l+1}^{l+u}$;
Coefficients of γ_A and γ_I;

1: Use $W_{ij} = e^{-d\|x_i - x_j\|^2/4t}$ to calculate the edge weight, construct the adjacent graph of labeled and unlabeled samples;

2: Calculate Laplacian matrix L, $L = D - W$;
3: Select kernel function $K\left(x_i, x_j\right)$ and build kernel matrix K for both labeled and unlabeled samples.
4: Utilize standard SVM algorithm to solve quadratic programming, obtain α^*;

Output:

The decision function $f^*(x) = \sum_{i=1}^{l+u}\alpha_i{}^* K(x, x_i)$;

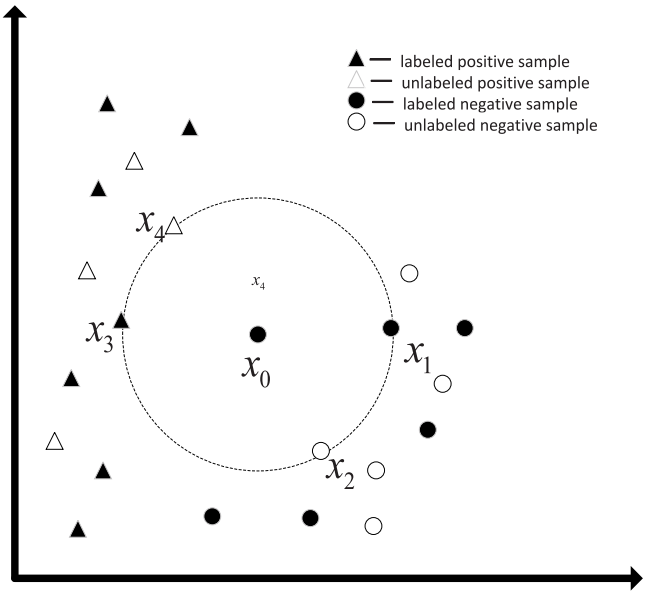

Fig. 2. Construction of the data adjacent graph.

2.3. Local behavior similarity

In many practical applications, people tend to comprehensively utilize their experience including supervised experience obtained from teachers and passive experience obtained from nature to make classification decisions. During the learning process, people think about the concept mechanism and the learning task is terminated when they determine the concept category. This is in line with the central concept of the semi-supervised learning algorithm. Consequently, it makes sense to determine how humans build the conceptual dividing line based on both labeled and unlabeled samples [38].

According to the LapSVM solving process, a critical step is to work out the Laplacian graph L, which is equal to calculating the adjacent matrix W in essence. Accordingly, the quality of W has a decisive impact on the classification performance and efficiency. Traditional LapSVM employs the heat kernel function, $W_{ij} = e^{-\|x_i - x_j\|^2/4t}$, to calculate W. However it has several shortcomings that degrade the model performance. Assume that Fig. 2 is a dataset with two kinds of samples. Triangles and circles represent samples of different classes. Filled dots and hollow dots represent labeled and unlabeled samples, respectively. x_1, x_2, x_3, x_4 are all in the neighborhood of x_0, and they share the same distance from x_0. Thus in LapSVM, the edge weight between x_1, x_2, x_3, x_4 and

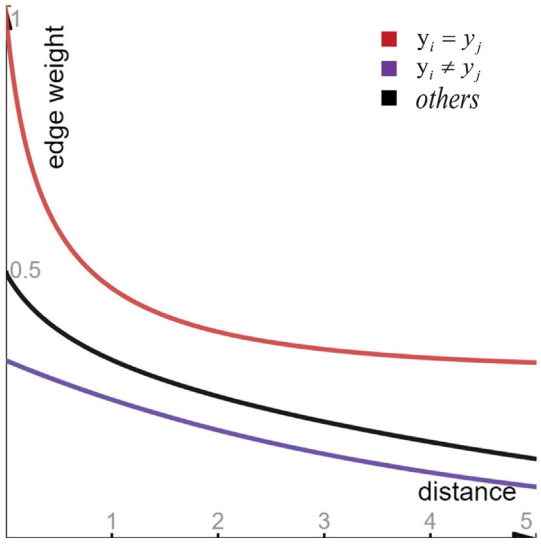

Fig. 3. Edge weight calculation based on Behavior Similarity.

x_0 is the same. Obviously, it does not conform to humans' characteristics of cognition. People prioritize the label information of data when making decisions by nature. We prefer to place objects of identical labels in the same category even if their features are not so similar and distinguish things of different classes although they bear a strong resemblance in features. From Fig. 2, the label of x_0 is the same as x_1 and different from x_2; we can naturally conclude that x_0 and x_1 have a higher similarity. The density and distance information of the feature space distribution are implied in the sample labels. Thus, we can better depict the distribution of samples by introducing label information to the adjacent matrix construction. Combined with human cognition characteristics, the edge weight based on behavior similarity is defined as follows:

$$W_{ij}^{BS} = \begin{cases} 1/\left(3 \times \sqrt{\frac{10}{9} - W_{ij}}\right) & y_i = y_j \\ 1/\left(3 \times \sqrt{\frac{1}{W_{ij}}}\right) & y_i \neq y_j \\ 2/\left(3 \times \left(\sqrt{1 - W_{ij}} + \sqrt{\frac{1}{W_{ij}}}\right)\right) & others \end{cases} \tag{10}$$

where $W_{ij} = e^{-d\|x_i - x_j\|^2/2}$. Fig. 3 is graphical presentation of Eq. (10), where red line denotes $y_i = y_j$, purple line represents $y_i \neq y_j$, black represents other situations. As shown in Fig. 3, on the assumption that the distance between samples is same, edge weight is clearly divided into three areas based on label information. The intent that edge weight of samples with identical label is larger than those belong to different class is well achieved.

The heat kernel function simply focuses on the corresponding sample, ignoring its neighbors. Apparently, it is contrary to humans' cognitive features. When people measure the interrelationship of objects, they do not merely contemplate the thing itself; the surrounding environment also makes indispensable differences in decision-making. People often take the local distribution of the feature space into account to make correct decisions. Based on such cognitive characteristics, the definition of the local view distance from x_i to x_j is as follows:

$$d\left(x_i, x_j\right)/\rho_i \tag{11}$$

where $\rho_i = \frac{1}{N_k}\sum_{k=1}^{N_k} d\left(x_i, x_k\right)$ denotes the local distribution parameter of x_i; in other words, ρ_i is the average distance between x_i and its N_k nearest neighbors $\left(x_1, x_2, \ldots, x_{N_k}\right)$, where N_k is the number of neighbors. The local view distance is based on human cognitive features, and it can better represent neighborhood distribution.

The local view distance from sample x_j to sample x_i is $d\left(x_j, x_i\right)/\rho_j$. Integrating the local view distance of both, the mutual distance between x_i and x_j can be obtained as follows:

$$\frac{d\left(x_i, x_j\right) d\left(x_j, x_i\right)}{\rho_i \rho_j} = \frac{d^2\left(x_i, x_j\right)}{\rho_i \rho_j} \tag{12}$$

Replace the distance measurement and kernel parameter t in the heat kernel function:

$$W_{ij} = \exp\left(\frac{-d^2\left(x_i, x_j\right)}{\rho_i \rho_j}\right) \tag{13}$$

2.4. MLapSVM-LBS

In the case of local behavior similarity, if the dataset is fixed, the key factor that changes the edge weight is the value of nearest neighbors k. However, the value of k relies heavily on the neighborhood distribution of a sample. Moreover, it is difficult to manually set the proper k for different datasets. A value of k that is too small may lead to useful neighborhood information being insufficient, whereas outliers are included in the k nearest neighbors if k far exceeds the local neighborhood size of the samples. As a result, it is meaningful to combine the information of various neighborhood sizes rather than tuning k as a fixed value.

In this paper, we construct five Laplacian graphs via different numbers of nearest neighbors (ranging from 2, 4, 8, 16, 32) to make the model less sensitive to the neighborhood size and apply the multi-Laplacian regularization to the LapSVM framework.

2.4.1. Formulation

Different graphs represent various distribution information in the neighborhood of samples, each of which makes diverse contribution to the MLapSVM-LBS model. $\alpha_v = [\alpha_1, \alpha_2, \ldots, \alpha_v]^T \in R^{v\times 1}$ is the nonnegative weight vector to integrate Laplacian matrices, and v is the number of matrices. The formula of the fused Laplacian matrix is as follows:

$$L^* = \sum_{v=1}^{V} \eta_v L_v \tag{14a}$$

$$s.t. \sum_{v=1}^{V} \eta_v = 1, \eta_v \in [0, 1] \tag{14b}$$

Integrating the multigraph regularization, the optimization problem of MLapSVM-LBS is as follows:

$$\arg\min_{f^*, \eta_v} \frac{1}{l}\sum_{i=1}^{l}\left(1 - y_i f\left(x_i\right)\right) + \gamma_A \|f\|_K^2 \tag{15a}$$

$$+ \frac{\gamma_I}{(l+u)^2} f^T \left(\sum_{v=1}^{V} {\eta_v}^e L_v\right) f \tag{15b}$$

$$s.t. \quad \sum_{v=1}^{V} \eta_v = 1, \eta_v \in [0, 1] \tag{15c}$$

In certain situations where the weight of a Laplacian matrix approaches 0 or 1, the model cannot find the complement of multiple graphs. To prevent the above problem, we employ a technique in which the η_v is replaced with ${\eta_v}^e$ and $e > 1$.

2.4.2. Optimization

To solve the objective function in Eq. (15), an effective alternation algorithm is utilized. The overview of MLapSVM-LBS is listed in Algorithm 2.

First, we fix $\eta_v = 1/v$ and optimize the variant f^*. Given $L^* = \sum_{v=1}^{V} \eta_v L_v$, the objective function is the same as that of the traditional LapSVM. We can obtain the solution f^* according to the solution to LapSVM.

Second, we fix the variant f^* to optimize the variant η_v. The objective function is related to:

$$\arg\min_{\eta_v} tr\left(\left(f^*\right)^T\left(\sum_{v=1}^{V}\eta_v{}^e L_v\right)\left(f^*\right)\right) \tag{16a}$$

$$s.t. \quad \sum_{v=1}^{V}\eta_v = 1, \eta_v \in [0,1] \tag{16b}$$

Introduce the Lagrange multiplier ξ and convert the problem above to the Lagrange function:

$$Lag\left(\eta_v, \xi\right) = tr\left(\left(f^*\right)^T\left(\sum_{v=1}^{V}\eta_v{}^e L_v\right)\left(f^*\right)\right) - \xi\left(\sum_{v=1}^{V}\eta_v - 1\right) \tag{17}$$

Set the derivative of η_v and ξ to 0:

$$\begin{cases} e\eta_v{}^{e-1} tr\left(\left(f^*\right)^T\left(\sum_{v=1}^{V}\eta_v{}^e L_v\right)\left(f^*\right)\right) - \xi = 0 \\ \sum_{v=1}^{V}\eta_v - 1 = 0 \end{cases} \tag{18}$$

Thus, η_v can be obtained by the following equation:

$$\eta_v = \frac{\left(\frac{1}{tr\left(\left(f^*\right)^T L_v f^*\right)}\right)^{\frac{1}{e-1}}}{\sum_{v=1}^{V}\left(\frac{1}{tr\left(\left(f^*\right)^T L_v f^*\right)}\right)^{\frac{1}{e-1}}} \tag{19}$$

Algorithm 2 Algorithm of MLapSVM-LBS.

Input:

Training set $\{x_i, y_i\}_{i=1}^{l}$, test set $\{x_i\}_{i=l+1}^{l+u}$;
The maximum number of iterations $t_{\max}$;
Coefficients of γ_A and γ_I;

1: Use Eq. 10 and Eq. 13 to calculate the edge weight, $W_{ij}{}^{BS}$, via different neighborhood size k;
2: Calculate Laplacian matrix L_v, $v = 1, 2, \ldots, V$, $L_v = D_v - W_v$;
3: Initialize $\eta_v{}^{(0)} = 1/V$, $v = 1, 2, \ldots, V$,
4: **for** $t = 1$ to $t_{\max}$ **do**
5: Calculate $L^{(t)}$, $L^{(t)} = \sum_{v=1}^{V}\eta_v{}^{(t-1)} L_v$;
6: Calculate $f^{*(t)}$ according to the solution to LapSVM;
7: Update $\eta_v{}^{(t)}$, $v = 1, 2, \ldots, V$ via Eq. 19;
8: **end for**

Output:

The decision function $f^*(x) = \sum_{i=1}^{l+u}\alpha_i{}^* K(x, x_i)$;

3. Result

3.1. Evaluation measurements

In this work, we employ accuracy (ACC), Matthew's correlation coefficient (MCC), specificity (SP), sensitivity (SN), and area under the ROC curve (AUC) to measure the performance of our method. ACC, SP, SN and MCC are calculated as follows:

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{20a}$$

$$SN = \frac{TP}{TP + FN} \tag{20b}$$

$$SP = \frac{TN}{TN + FP} \tag{20c}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}} \tag{20d}$$

where TP and TN denotes the number of true positive and true negative samples. FN, FP is the number of false negative and false positive samples. Moreover, we also employ the area under the receiver operating characteristic curve (AUC) to evaluate our model.

Table 2
Comparison among different feature combination by LapSVM on PDB1075 (LOOCV).

Methods	ACC	MCC	SN	SP	AUC
GE	0.6567	0.34	0.6567	0.4286	0.777
MCD	0.6902	0.40	0.4990	**0.8727**	0.786
PSSM-AB	0.7591	0.41	0.7010	0.8145	0.834
PsePSSM	0.7516	0.51	0.6552	0.8436	**0.845**
PSSM-DWT	0.7312	0.46	0.6686	0.7909	0.829
NMBAC	0.6874	0.40	0.4876	0.8782	0.807
GE+NMBAC+PsePSSM	**0.7598**	**0.52**	0.7213	0.8537	0.842
Fusion of all features	0.7427	0.48	**0.7435**	0.7323	0.843

3.2. Feature combination

In our study, we utilize six types of features extracted from protein sequence, which includes GE, NMBAC, MCD, PSSM-AB, PsePSSM and PSSM-DWT. To obtain best feature combination, single feature and the feature fusion are compared on PDB1075 under LOOCV by traditional LapSVM. The comparison results are listed in Table 2. As shown in Table 2, LapSVM achieve best performance (ACC: 0.7598, MCC: 0.52) with fused feature (GE, NMBAC, PsePSSM). Consequently, we utilize GE, NMBAC, PsePSSM to train and test our model.

3.3. Parameter selection

In our work, the grid search method is utilized to obtain optimal parameters. There are four parameters (C, γ, γ_A, $a = \frac{\gamma_I}{(l+u)^2}$) in our model and we test these parameters on PDB1075 by fivefold cross-validation (5-CV). The range of values for C and γ are from 2^{-5} to 2^5 with step 2^1. We set the range of values for γ_A and a from 0.1 to 0.9 with step 0.1. For MLapSVM-LBS, we obtain the optimal parameters C, γ, γ_A under 8, 0.5, 0.9 and 1, respectively.

3.4. Performance analysis on PDB1075

We appraise different models (different neighborhood sizes k) on PDB1075 under leave-one-out cross-validation (LOOCV). The comparison is listed in Table 3 and Fig. 4. Different numbers of neighbors lead to different results of prediction. LapSVM-LBS with k=8 (ACC:0.7691) performs better than LapSVM-LBS with k=2 (ACC: 0.7274) and k=32 (ACC: 0.7115). The moderate neighborhood size (k=8) achieves better results than too small (k=2) or too large size (k=32). Our method (MLapSVM-LBS) obtains the best MCC(0.67), ACC(0.8323), SN(0.7941) and AUC(0.908). Apparently, the fusion of multiple neighborhood sizes performs better.

To further prove the effectiveness of our model, we also evaluated it on PDB1075 and compared its result with traditional

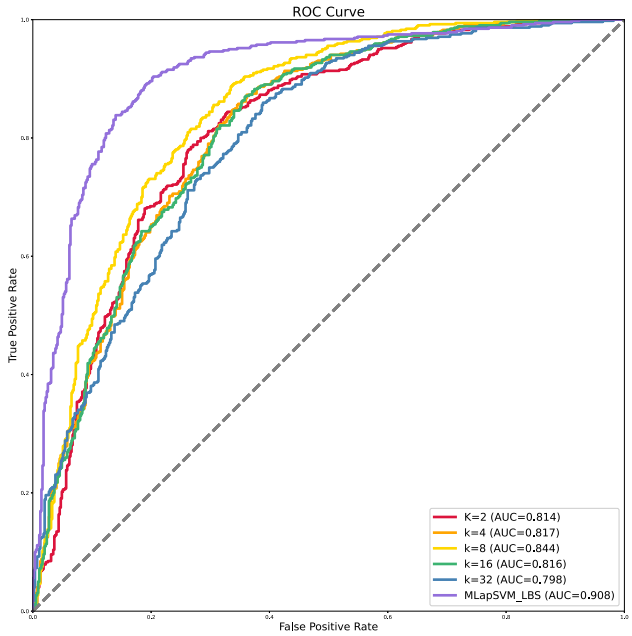

Fig. 4. The ROC curve of LapSVM-LBS with different k and MLapSVM-LBS on PDB1075(LOOCV).

Table 3
Comparison between LapSVM-LBS with different k and MLapSVM-LBS on PDB1075 (LOOCV).

Methods	ACC	MCC	SN	SP	AUC
LapSVM-LBS with k=2	0.7274	0.42	0.7112	0.8143	0.814
LapSVM-LBS with k=4	0.7368	0.48	0.6834	0.8324	0.817
LapSVM-LBS with k=8	0.7691	0.52	0.7663	**0.8618**	0.844
LapSVM-LBS with k=16	0.7359	0.41	0.5027	0.8318	0.816
LapSVM-LBS with k=32	0.7115	0.39	0.4834	0.8113	0.798
MLapSVM-LBS	**0.8323**	**0.67**	**0.7941**	0.8534	**0.908**

Table 4
Comparison among LapSVM, Multi-graph LapSVM, LapSVM-LBS and MLapSVM-LBS on PDB1075 (LOOCV).

Methods	ACC	MCC	SN	SP	AUC
LapSVM	0.7598	0.52	0.7213	0.8537	0.842
LapSVM-LBS with k=8	0.7691	0.52	0.7663	**0.8618**	0.844
Multi-graph LapSVM	0.7601	0.52	0.7514	0.8314	0.843
MLapSVM-LBS	**0.8323**	**0.67**	**0.7941**	0.8534	**0.908**

LapSVM, LapSVM-LBS (k=8), Multi-graph LapSVM. The multiple graph is built by rbf kernel with different γ. The results is shown in Table 4. LapSVM-LBS preformed better than LapSVM, from which we can draw a conclusion that Local Behavior Similarity make a difference to the performance. Moreover, MLapSVM-LBS achieves better results than single graph model (LapSVM-LBS with k=8).

Under LOOCV, MLapSVM-LBS and other methods are also tested on the PDB1075 dataset. The comparison results are listed in Table 5. Existing methods include iDNA-Prot|dis, PseDNA-Pro, IDNA-Prot, DNA-Prot, DNAbinder, iDNAPro-PseAAC, Kmerl+ACC, Local-DPP, MKSVM with MKL-CKA [39] and MKL-HSIC with H-LapSVM [40].

Compared with iDNA-Prot|dis(MCC:0.54, ACC:0.7730), PseDNA-Pro (MCC:0.53, ACC:0.7655), iDNA-Prot(MCC:0.50, ACC:0.7540), DNA-Prot(MCC:0.44, ACC:0.7255), DNAbinder (MCC:0.48, ACC: 0.7395), iDNAPro-PseAAC(MCC:0.53, ACC:0.7656), Kmerl+ACC (MCC:0.53, ACC:0.7523), Local-DPP (MCC:0.59, ACC:0.7910) and MKL-HSIC with H-LapSVM(MCC:0.65, ACC:0.8205), We find that MLapSVM-LBS(MCC:0.67, ACC:0.8323) has the better performance.

Table 5
Comparison between different classification models on PDB1075 (LOOCV).

Methods	ACC	MCC	SN	SP	AUC
DNA-Prot	0.7255	0.44	0.8267	0.5976	0.789
iDNA-Prot	0.7540	0.50	0.8381	0.6437	0.761
iDNA-Prot|dis	0.7730	0.54	0.7940	0.7527	–
Local-DPP	0.7910	0.59	0.8480	0.736	–
DNAbinder	0.7395	0.48	0.6857	0.7909	0.815
PseDNA-Pro	0.7655	0.53	0.7961	0.7363	–
Kmerl+ACC	0.7523	0.53	0.7676	0.7376	–
iDNAPro-PseAAC	0.7656	0.53	0.7562	0.7745	0.839
MKSVM with MKL-CKA	**0.8419**	**0.68**	0.8591	0.8255	**0.914**
MSFBinder	0.8353	0.67	0.838	0.832	–
Adilina's work	0.7021	0.41	0.6100	0.7970	–
MKSVM-HKA	0.8130	0.63	0.8229	0.8036	–
FKRR-MVSF	0.8326	0.67	**0.8571**	0.8091	–
MKLL-HSIC with H-LapSVM	0.8205	0.65	0.7486	**0.8891**	0.901
MLapSVM-LBS	0.8323	0.67	0.7941	0.8534	0.908

[a] "–" represents that the value is not available.

3.5. Performance analysis on PDB186

PDB1075 and PDB186 are employed as the training and test sets, respectively, to further test the reliability of our model. The sequence similarity between the protein in the test set and the protein in the training set has a great influence on the prediction results. We removed proteins in PDB1075 which bear more than 25% similarity with protein in PDB186 and rebuilt the model on reduced PDB1075. The test results are listed in Table 6.

Our method (MLapSVM-LBS) achieves 0.887, 0.76, 0.903, and 0.870 on ACC, MCC, SN, and SP, respectively. According to the comparison results, we can conclude that our model works well in DBP prediction. MKSVM-HKA(ACC: 0.812, MCC:0.65), Adilina's work(ACC: 0.823, MCC:0.67), FKRR-MVSF(ACC: 0.817, MCC:0.68), MKSVM with MKL-CKA(ACC: 0.837, MCC:0.69), KKDPP(ACC: 0.812, MCC:0.66), and MKL-HSIC with H-LapSVM(ACC: 0.871, MCC:0.75) also obtain good performance on PDB186. In Adilina's work [41], seven types of feature selection methods were employed to construct the model. FKRR-MVSF [30] and MKSVM-HKA [31] utilized the MKL algorithm to integrate different features. MSFBinder [42] developed a model with a stacking framework via several features. KKDPP [43] built a random forest model with fused PSSM features. From the results, methods based on multiple feature fusion achieve better performance.

3.6. Performance analysis on PDB14189

To evaluate the robustness of our method, we also test our model on dataset of large size (PDB14189) by five-fold cross validation. The comparison results are shown in Table 7. Compared with evaluated methods, MLapSVM-LBS achieves best results. The value of ACC, SN, SP, MCC, AUC are 0.8451, 0.8312, 0.8681, 0.65, 0.917, respectively. The better performance shows the effectiveness of proposed model.

3.7. Performance analysis on PDB2272

We also remove proteins in PDB14189 with more than 25% similarity to any protein in PDB2272 and rebuild the model on reduced PDB14189. The comparison results between MLapSVM-LBS

Table 6
The results of comparison between MLapSVM-LBS and previous methods on PDB186 (independent test of reduced PDB1075).

Methods	ACC	MCC	SN	SP	AUC
DNA-Prot	0.618	0.24	0.699	0.538	–
iDNA-Prot	0.672	0.34	0.677	0.667	–
iDNA-Prot|dis	0.720	0.45	0.795	0.645	–
Kmerl+ACC	0.710	0.43	0.828	0.591	–
iDNAPro-PseAAC	0.715	0.44	0.828	0.602	0.778
DNAbinder	0.608	0.22	0.570	0.645	0.607
Local-DPP	0.790	0.63	0.925	0.656	–
Adilina's work	0.823	0.67	0.950	0.699	–
MKSVM-HKA	0.812	0.65	0.946	0.677	0.887
FKRR-MVSF	0.817	0.68	**0.989**	0.645	0.901
DPP-PseACC	0.774	0.55	0.839	0.710	0.799
PseDNA-Pro	0.715	0.243	0.828	0.602	–
MSFBinder	0.817	0.64	0.893	0.742	–
MKLL-HSIC with H-LapSVM	0.871	0.75	0.914	0.828	0.931
MKSVM with MKL-CKA	0.837	0.69	0.936	0.742	0.899
MsDBP	0.801	0.61	0.860	0.742	0.875
KKDPP	0.812	0.66	0.978	0.645	–
MLapSVM-LBS	**0.887**	**0.76**	0.903	**0.870**	**0.957**

[a] "–" represents that the value is not available.
[b] Proteins in PDB1075 with more than 25% similarity to any protein in PDB186 are removed.

Table 7
Comparison between our model and existing methods on the PDB14189 dataset (Five-Fold Cross Validation).

Methods	ACC	MCC	SN	SP	AUC
MsDBP	0.8094	0.62	0.8087	0.7972	0.889
MKLL-HSIC with H-LapSVM	0.8347	0.64	**0.8465**	0.8342	0.904
MLapSVM-LBS	**0.8451**	**0.65**	0.8312	**0.8681**	**0.917**

Table 8
Comparison between our model and existing methods on the PDB2272 dataset (Independent test of reduced PDB14189).

Methods	ACC	MCC	SN	SP	AUC
DPP-PseACC	0.581	0.163	0.566	0.596	0.610
PseDNA-Pro	0.618	0.243	0.753	0.481	–
MsDBP	0.643	0.340	0.707	0.632	0.738
MKLL-HSIC with H-LapSVM	0.694	0.401	**0.721**	0.561	0.772
MLapSVM-LBS	**0.712**	**0.424**	0.716	**0.708**	**0.798**

[a] "–" represents that the value is not available.
[b] Proteins in PDB14189 with more than 25% similarity to any protein in PDB2272 are removed.

and other existing methods are listed in Table 8. MLapSVM-LBS achieves 0.712, 0.424, 0.716, and 0.708 on ACC, MCC, SN, and SP, respectively. According to Table 8, MLapSVM-LBS achieves best MCC(0.424), which is better than MsDBP(0.340) and MKL-HSIC with H-LapSVM(0.401). DPP-PseAAC [44] achieved MCC:0.163, which is lower than PseDNA-Pro. We find that MLapSVM-LBS has better generalization performance than other methods on PDB2272.

4. Conclusion and discussion

In this work, we utilize three protein features, PsePSSM, NMBAC and GE, to represent proteins and built the MLapSVM-LBS model for detecting DBPs. Compared with standard LapSVM, the Laplacian matrix of MLapSVM-LBS is built with local behavior similarity (LBS) which is inspired by human behavior learning theory. In detail, MLapSVM-LBS takes advantage of label information when calculating edge weight and introduces a local distribution parameter to reflect the underlying probability distribution of a sample's neighborhood. In addition, to improve the predictive performance, MLapSVM-LBS applies multiple Laplacian regularization by changing the neighborhood size which can make full use of the geometric distribution of samples. Our method reaches the accuracy of 0.887 and 0.712 on PDB186 (independent test set of PDB1075) and PDB2272 (independent test set of PDB14189), respectively.

Despite the fact that a number of computation-based methods have been proposed to detect DBPs, the performance of existing methods can still improve. Similar to many other sequence-based approaches, our method does not consider noise. In future work, we will utilize other graph-based model [45,46], density-based methods and fuzzy theory to improve the performance of our model.

CRediT authorship contribution statement

Mengwei Sun: Initial methodology, Experimental part, Writing – original draft, Proofreading. **Prayag Tiwari:** Methodological part, Experimental part, Writing – original draft, Proofreading. **Yuqin Qian:** Improved the methodological part, Experimental part, Writing – original draft, Proofreading. **Yijie Ding:** Improved the methodological part, Experimental part, Writing – original draft, Proofreading. **Quan Zou:** Improved the methodological part and helped in the experimental part, Writing – original draft, Proofreading.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The source code is available at https://github.com/prayagtiwari/MLapSVM-LBS.git, or https://figshare.com/s/56182f00f08d7ef49b18.

Acknowledgments

This work is supported by a grant from the National Natural Science Foundation of China (NSFC 62172076, 61902271), the Academy of Finland (grants 336033, 315896), Business Finland (grant 884/31/2018), EU H2020 (grant 101016775), the Natural Science Research of Jiangsu Higher Education Institutions of China (19KJB520014) and the Municipal Government of Quzhou (Grant Number 2020D003 and 2021D004).

References

[1] Cangzhi Jia, Yun Zuo, Quan Zou, O-GlcNAcPRED-II: an integrated classification algorithm for identifying O-GlcNAcylation sites based on fuzzy undersampling and a K-means PCA oversampling technique, Bioinformatics 34 (12) (2018) 2029–2036.
[2] Xiaoqing Ru, Lihong Li, Quan Zou, Incorporating distance-based top-n-gram and random forest to identify electron transport proteins, J. Proteome Res. 18 (7) (2019) 2931–2939.
[3] Bin Liu, Shuangyan Jiang, Quan Zou, HITS-PR-HHblits: protein remote homology detection by combining PageRank and hyperlink-induced topic search, Brief. Bioinform. 21 (1) (2020) 298–308.
[4] Yubo Wang, Yijie Ding, Jijun Tang, Yu Dai, Fei Guo, CrystalM: a multi-view fusion approach for protein crystallization prediction, IEEE/ACM Trans. Comput. Biol. Bioinform. (2019).

[5] Leyi Wei, Yijie Ding, Ran Su, Jijun Tang, Quan Zou, Prediction of human protein subcellular localization using deep learning, J. Parallel Distrib. Comput. 117 (2018) 212–217.
[6] Yijie Ding, Jijun Tang, Fei Guo, Human protein subcellular localization identification via fuzzy model on kernelized neighborhood representation, Appl. Soft Comput. 96 (2020) 106596.
[7] Yijie Ding, Jijun Tang, Fei Guo, The computational models of drug-target interaction prediction, Protein Peptide Lett. 27 (5) (2020) 348–358.
[8] Yijie Ding, Jijun Tang, Fei Guo, Identification of drug-side effect association via semisupervised model and multiple kernel learning, IEEE J. Biomed. Health Inf. 23 (6) (2018) 2619–2632.
[9] Yijie Ding, Jijun Tang, Fei Guo, Identification of drug–target interactions via dual laplacian regularized least squares with multiple kernel fusion, Knowl.-Based Syst. 204 (2020) 106254.
[10] Wen Zhang, Kanghong Jing, Feng Huang, Yanlin Chen, Bolin Li, Jinghao Li, Jing Gong, SFLLN: a sparse feature learning ensemble method with linear neighborhood regularization for predicting drug–drug interactions, Inform. Sci. 497 (2019) 189–201.
[11] Yifan Deng, Xinran Xu, Yang Qiu, Jingbo Xia, Wen Zhang, Shichao Liu, A multimodal deep learning framework for predicting drug–drug interaction events, Bioinformatics 36 (15) (2020) 4316–4322.
[12] H. Liu, G. Ren, H. Chen, Q. Liu, Y. Yang, Q. Zhao, Predicting lncRNA–miRNA interactions based on logistic matrix factorization with neighborhood regularized, Knowl.-Based Syst. 191 (2019) 105261.
[13] Xiangxiang Zeng, Li Liu, Linyuan Lü, Quan Zou, Prediction of potential disease-associated microRNAs using structural perturbation method, Bioinformatics 34 (14) (2018) 2425–2432.
[14] Limin Jiang, Yongkang Xiao, Yijie Ding, Jijun Tang, Fei Guo, FKL-Spa-LapRLS: an accurate method for identifying human microRNA-disease association, BMC Genomics 19 (10) (2018) 11–25.
[15] Wen Zhang, Zhishuai Li, Wenzheng Guo, Weitai Yang, Feng Huang, A fast linear neighborhood similarity-based network link inference method to predict microRNA-disease associations, IEEE/ACM Trans. Comput. Biol. Bioinform. 18 (2) (2019) 405–415.
[16] Yuchong Gong, Yanqing Niu, Wen Zhang, Xiaohong Li, A network embedding-based multiple information integration method for the MiRNA-disease association prediction, BMC Bioinformatics 20 (1) (2019) 1–13.
[17] Jia Qu, Xing Chen, Jun Yin, Yan Zhao, Zheng-Wei Li, Prediction of potential miRNA-disease associations using matrix decomposition and label propagation, Knowl.-Based Syst. 186 (2019) 104963.
[18] Leyi Wei, Jijun Tang, Quan Zou, Local-DPP: An improved DNA-binding protein prediction method by exploring local evolutionary information, Inform. Sci. 384 (2017) 135–144.
[19] Bin Liu, Jinghao Xu, Xun Lan, Ruifeng Xu, Jiyun Zhou, Xiaolong Wang, Kuo-Chen Chou, iDNA-Prot| dis: identifying DNA-binding proteins by incorporating amino acid distance-pairs and reduced alphabet profile into the general pseudo amino acid composition, PLoS One 9 (9) (2014) e106691.
[20] Yubo Wang, Yijie Ding, Fei Guo, Leyi Wei, Jijun Tang, Improved detection of DNA-binding proteins via compression technology on PSSM information, PLoS One 12 (9) (2017) e0185587.
[21] Guy Nimrod, Maya Schushan, András Szilágyi, Christina Leslie, Nir Ben-Tal, iDBPs: a web server for the identification of DNA binding proteins, Bioinformatics 26 (5) (2010) 692–693.
[22] Nitin Bhardwaj, Robert E. Langlois, Guijun Zhao, Hui Lu, Kernel-based machine learning protocol for predicting DNA-binding proteins, Nucleic Acids Res. 33 (20) (2005) 6486–6493.
[23] Shahana Yasmin Chowdhury, Swakkhar Shatabda, Abdollah Dehzangi, iDNAProt-ES: Identification of DNA-binding proteins using evolutionary and structural features, Sci. Rep. 7 (1) (2017) 1–14.
[24] Shandar Ahmad, Akinori Sarai, Moment-based prediction of DNA-binding proteins, J. Mol. Biol. 341 (1) (2004) 65–71.
[25] Bin Liu, Jinghao Xu, Shixi Fan, Ruifeng Xu, Jiyun Zhou, Xiaolong Wang, PseDNA-Pro: DNA-binding protein identification by combining Chou's PseAAC and physicochemical distance transformation, Mol. Inform. 34 (1) (2015) 8–17.
[26] Bin Liu, Shanyi Wang, Xiaolong Wang, DNA binding protein identification by combining pseudo amino acid composition and profile-based protein representation, Sci. Rep. 5 (1) (2015) 1–11.
[27] Yu-dong Cai, Shuo Liang Lin, Support vector machines for predicting rRNA-, RNA-, and DNA-binding proteins from amino acid sequence, Biochimica Et Biophysica Acta (BBA)-Proteins and Proteomics 1648 (1–2) (2003) 127–133.
[28] Stephen F Altschul, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, David J Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (17) (1997) 3389–3402.
[29] Manish Kumar, Michael M. Gromiha, Gajendra P.S. Raghava, Identification of DNA-binding proteins using support vector machines and evolutionary profiles, BMC Bioinformatics 8 (1) (2007) 1–10.
[30] Yi Zou, Yijie Ding, Jijun Tang, Fei Guo, Li Peng, FKRR-MVSF: a fuzzy kernel ridge regression model for identifying DNA-binding proteins by multi-view sequence features via Chou's five-step rule, Int. J. Mol. Sci. 20 (17) (2019) 4175.
[31] Yijie Ding, Feng Chen, Xiaoyi Guo, Jijun Tang, Hongjie Wu, Identification of DNA-binding proteins by multiple kernel support vector machine and sequence information, Current Proteomics 17 (4) (2020) 302–310.
[32] Zhi-Ping Feng, Chun-Ting Zhang, Prediction of membrane protein types based on the hydrophobic index of amino acids, J. Protein Chem. 19 (4) (2000) 269–275.
[33] Xi Li, Bo Liao, Yu Shu, Qingguang Zeng, Jiawei Luo, Protein functional class prediction using global encoding of amino acid sequence, J. Theoret. Biol. 261 (2) (2009) 290–293.
[34] W.C. Lou, X.Q. Wang, Y.X. Chen, Zhang Jiang, Sequence based prediction of DNA-binding proteins based on hybrid feature selection using random forest and Gaussian naive Bayes, PLOS ONE 9 (1) (2014) 2014.
[35] Xiuquan Du, Yanyu Diao, Heng Liu, Shuo Li, MsDBP: exploring DNA-binding proteins by integrating multiscale sequence information via chou's five-step rule, J. Proteome Res. 18 (8) (2019) 3119–3132.
[36] Mikhail Belkin, Partha Niyogi, Vikas Sindhwani, Manifold regularization: A geometric framework for learning from labeled and unlabeled examples, J. Mach. Learn. Res. 7 (11) (2006).
[37] C. Cortes, Support-vector networks, Mach. Learn. 20 (1995).
[38] Z. Liu, Yang Jun-An, H. Liu, W. Wang, Laplacian support vector machine by local behavioral similarity, J. Chin. Comput. Syst. (2016).
[39] Yuqing Qian, Limin Jiang, Yijie Ding, Jijun Tang, Fei Guo, A sequence-based multiple kernel model for identifying DNA-binding proteins, BMC Bioinformatics 22 (3) (2021) 1–18.
[40] Yuqing Qian, Hao Meng, Weizhong Lu, Zhijun Liao, Yijie Ding, Hongjie Wu, Identification of DNA-binding proteins via hypergraph based laplacian support vector machine, Current Bioinform. 17 (1) (2022) 108–117.
[41] Sheikh Adilina, Dewan Md Farid, Swakkhar Shatabda, Effective DNA binding protein prediction by using key features via Chou's general PseAAC, J. Theoret. Biol. 460 (2019) 64–78.
[42] Xiu-Juan Liu, Xiu-Jun Gong, Hua Yu, Jia-Hui Xu, A model stacking framework for identifying DNA binding proteins by orchestrating multi-view features and classifiers, Genes 9 (8) (2018) 394.
[43] Yuran Jia, Shan Huang, Tianjiao Zhang, KK-DBP: A multi-feature fusion method for DNA-binding protein identification based on random forest, Front. Genetics (2021) 2458.
[44] M Saifur Rahman, Swakkhar Shatabda, Sanjay Saha, Mohammad Kaykobad, M Sohel Rahman, Dpp-pseaac: A dna-binding protein prediction model using chou's general pseaac, J. Theoret. Biol. 452 (2018) 22–34.
[45] R. Yin, K. Li, G. Zhang, J. Lu, A deeper graph neural network for recommender systems, Knowl.-Based Syst. 185 (2019) 105020.
[46] Z. Kang, L.J. Wen, W.Y. Chen, Z.L. Xu, Low-rank kernel learning for graph-based clustering, Knowl.-Based Syst. 163 (2019) 510–517.