IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 30, 2018, accepted January 15, 2019, date of publication January 23, 2019, date of current version February 8, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2894225

LPI-KTASLP: Prediction of LncRNA-Protein Interaction by Semi-Supervised Link Learning With Multivariate Information

CONG SHEN[1], YIJIE DING[2], JIJUN TANG[1,3], LIMIN JIANG[1], AND FEI GUO[1]
[1]School of Computer Science and Technology, College of Intelligence and Computing, Tianjin University, Tianjin 300350, China
[2]School of Electronic and Information Engineering, Suzhou University of Science and Technology, Suzhou 215009, China
[3]Department of Computer Science and Engineering, University of South Carolina, Columbia, SC 29208, USA

Corresponding author: Fei Guo (fguo@tju.edu.cn)

This work was supported in part by the grant from the National Natural Science Foundation of China under Grant NSFC 61772362, in part by the Tianjin Research Program of Application Foundation and Advanced Technology under Grant 16JCQNJC00200, and in part by the National Key Research and Development Program of China under Grant 2018YFC0910405 and Grant 2017YFC0908400.

ABSTRACT Long non-coding RNA, also known as lncRNA, is a series of single-stranded polynucleotides (no less than 200 nucleotides each), consisting of non-protein coding transcripts. LncRNA plays a crucial role in regulating gene expression, during the transcriptional, post-transcriptional, and epigenetic processes. This is achieved by lncRNA interacts with the corresponding RNA-binding proteins. It has been drawn to a lot of attention that the reduction of the excessive laboratory cost and the increase in speed and accuracy gains benefits from the employment of computational intelligence in lncRNA–protein interaction (LPI) identification. Although numerous pertinent in silico studies of LPI prediction have been proposed, there is still room for enhancing the accuracy of the existing LPI prediction methods. In this paper, we have proposed a novel method for identifying LPI with kernel target alignment based on semi-supervised link prediction (LPI-KTASLP), which adopts multivariate information to predict lncRNAs–proteins interactions. To integrate the heterogeneous kernels, kernel target alignment has been applied to deal with kernel fusion. We have calculated the low-rank approximation matrices of lncRNA and protein, where eigendecomposition is used to reduce computing pressure. The prediction model has been obtained by producing the ultimate LPI prediction matrix. Experimental results show that the prediction ability of the LPI-KTASLP algorithm has surpassed many other LPI prediction schemes. Our method of lncRNA–protein interaction prediction has been evaluated on a standard benchmark dataset of LPIs. We have observed that the highest AUPR of 0.6148 is obtained by our proposed model (LPI-KTASLP). This is superior to the integrated LPLNP (AUPR: 0.4584), the RWR (AUPR: 0.2827), the CF (AUPR: 0.2357), the LPIHN (AUPR: 0.2299), and the LPBNI (AUPR: 0.3302). It is very encouraging that most of the LPI predictions have been confirmed to be close to relevant concentrations.

INDEX TERMS LncRNA-protein interaction, kernel target alignment, low-rank approximation, multiple kernel learning, semi-supervised link prediction.

I. INTRODUCTION

Long non-coding RNA, also known as lncRNA, is a series of single-stranded polynucleotides (no less than 200 nucleotides each), consisting of non-protein coding transcripts [1]. Since the first group of lncRNAs were discovered two decades ago, researchers who specialize in biological sciences have corroborated the phenomenon, that non-coding RNAs can regulate ubiquitous gene expression during the transcriptional, post-transcriptional, and epigenetic processes [2]–[6]. It is realized by means of interactions between the corresponding RNA-binding proteins and lncRNAs per se. For example, a kind of lncRNA named lnc-Lsm3b, can refrain the activity of the receptor RIG-I by the induction of viruses during the regulation of immune response [7]. It has been drawn to a lot of researchers' attention that the reduction of the excessive laboratory cost and the increase in speed and accuracy, gains benefits from the employment of computational intelligence in LncRNA-Protein Interaction (LPI) identification [8].

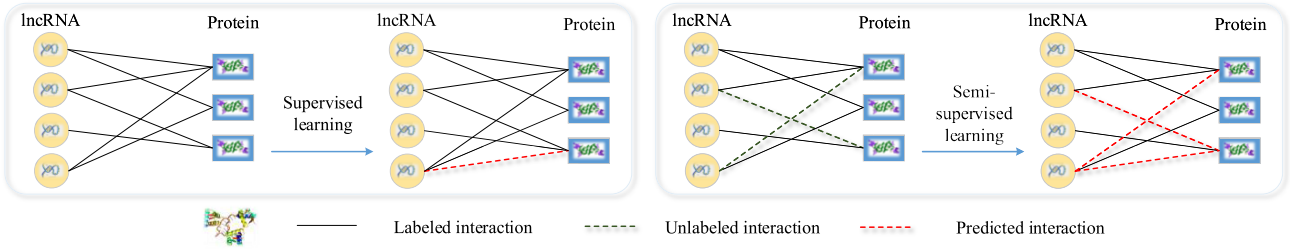

FIGURE 1. Comparison between supervised learning and semi-supervised learning for the prediction of LPI in bipartite network.

Dissections about the 3D structures with respect to a portion of microRNA genes are available. But currently, it is still rare to retrieve the available resources about the 3D structure of lncRNA. Consequently, several sequence-based approaches for identifying LPI have appeared in the past decade. Zhang *et al.* [9] have considered that all of these approaches can be classified into two categories. The first category has focused on viewing the prediction task as a binary classification under supervised techniques without known interactions [10]–[12]. For instance, Bellucci *et al.* [13] have proposed the catRAPID computational method of LPI prediction, which uses physicochemical properties and secondary structure as compound information. Muppirala *et al.* [14] have presented another well-known algorithm RPISeq, adopted Support Vector Machine (SVM) and Random Forest (RF) to reach the same goal. Wang *et al.* [15] have used both Naive Bayes (NB) and Extended NB (ENB) to predict LPI. Lu *et al.* [16] have encoded each LPI to a numeric code so that can form a vector, then applied matrix multiplication. Suresh *et al.* [17] have conceived RPI-Pred that builts an SVM model with the structure and sequence data of lncRNAs and proteins.

Different from the previous methods, another category has taken advantage of known interactions to forecast unknown lncRNA-protein interactions. Li *et al.* [18] have raised LPIHN that not only builds a heterogeneous profile, but also exploits a kind of random walk with a restart mechanism (RWR) on lncRNA-protein association network. Ge *et al.* [19] have developed LPBNI which makes use of a two-step scheme on a bipartite network. Recently, Hu *et al.* [20] have delineated a kind of Semi-Supervised Link Prediction called LPI-ETSLP, which also achieves outstanding performance. Also they have upgraded through the RWR and matrix factorization [21], [22].

Semi-supervised learning, which constructs predictors from datasets that contain both labeled and unlabeled samples, is a kind of an effective mechanism that can reduce the need of labeled data. Fig. 1 presents a comparison between supervised and semi-supervised learning. Because of the effectiveness and efficiency of similarity measurements, we introduce a semi-supervised learning approach based on similarity matrices for the LPI identification.

In this paper, we try to identify the lncRNAs and proteins that can interact with each other in statistical analysis. By trawling the literature resources, we have noticed that state-of-the-art models, such as Feature Extraction, Recursive Least Squares (RLS), Sparse Representation based Classifier (SRC), Multiple Kernel Learning (MKL), have been employed by several scenarios to speculate Drug-Target Interactions (DTIs) [23]–[25], Protein-Protein Interactions (PPIs) [26]–[31], drug-side effect associations [32], [33], MicroRNA-Disease [34] or LncRNA-Disease Associations [35], [36], and binding sites of biomolecules [37], [38] with remarkable performance. Matrix decomposition or factorization, is a popular technique in Machine Learning [39], [40]. In our essay, the low-rank approximation matrices of lncRNA and protein is computed, where eigendecomposition is wielded to reduce computing pressure. Moreover, due to cross validation is different from leave-one-out cross validation, we list the comparable methods to the best of our knowledge.

Our contributions can be summarized in threefolds: (I) We integrate a variety kinds of similarity matrices as kernels for LPI prediction; (II) We integrate the heterogeneous kernels from different molecular spaces through Kernel Target Alignment (KTA) to deal with kernels fusion; (III) In terms of the performance about the above predictors, we combine their advantages with Semi-supervised Link Prediction (SLP) [41], which utilizes MKL with matrix factorization and approximation.

II. METHODS

Technical flow chart of LPI-KTASLP is shown in Fig. 2. The multivariate information in predicting LPI which has been leveraged in this research is derived from the lncRNA expression, the local network and the sequence information. Differing from the state-of-the-art predictors, we propose an identification of LPI with Kernel Target Alignment based on Semi-Supervised Link Prediction (LPI-KTASLP), which utilizes matrix factorization and approximation. What's more, we carry out MKL by deploying KTA.

A. PROBLEM DESCRIPTION

Given the lncRNA-protein interactions that include n lncRNAs and m proteins. Specifically, $\mathbf{l} = \{l_1, l_2, \cdots, l_i, \cdots, l_n\}$ and $\mathbf{P} = \{P_1, P_2, \cdots, P_j, \cdots, P_m\}$ indicate the lncRNAs and proteins, respectively. Hence, the interactions between lncRNAs and proteins can be represented as an adjacency

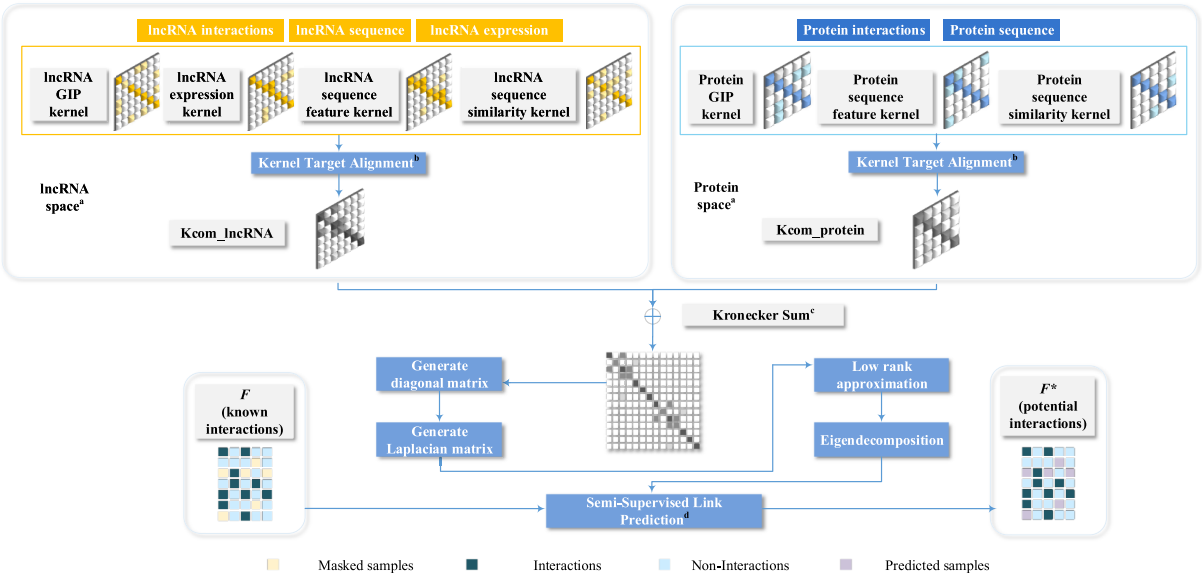

FIGURE 2. Technical flow chart of the proposed LPI prediction model. a. LncRNA and protein are two separated and independent spaces; b. Kernel Target Alignment (KTA) is applied in estimating the weight of each kernel for the corresponding space; c. Kronecker Sum is the technique about fusing different spatial data; d. Semi-supervised Link Prediction (SLP) is implemented according to (24).

matrix $\mathbf{F}$ with $n \times m$, which is formulated as follows:

$$\mathbf{F} = \begin{bmatrix} F_{1,1} & F_{1,2} & \cdots & F_{1,j} & \cdots & F_{1,m} \\ F_{2,1} & F_{2,2} & \cdots & F_{2,j} & \cdots & F_{2,m} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ F_{i,1} & F_{i,2} & \cdots & F_{i,j} & \cdots & F_{i,m} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ F_{n,1} & F_{n,2} & \cdots & F_{n,j} & \cdots & F_{n,m} \end{bmatrix}_{n \times m} \tag{1}$$

where $F_{i,j}$ denotes the corresponding value of element in matrix $\mathbf{F}$, $1 \leq i \leq n$, $1 \leq j \leq m$, and $m, n \in \mathbb{N}^*$. If lncRNA l_i interacts with protein P_j, the value of $F_{i,j}$ is marked as 1, otherwise it is marked as 0.

The interactions between lncRNAs and proteins can be abstractly represented as a bipartite network. Therefore, identification of new interactions between lncRNAs and proteins can be viewed as a recommender task, which can automatically seek out latent associated individuals. It needs to be emphasized that partial known links are used to predict potential interactions in the architecture, which are shown in Fig. 2. The recommender system for predicting lncRNA-protein interactions is achieved by means of MKL.

B. LNCRNA KERNELS AND PROTEIN KERNELS

LncRNA and protein feature spaces are derived from the interactions between two kinds of molecules. The lncRNA expression, lncRNA sequence and known interactions between one lncRNA and all proteins are examined in our framework. The information of lncRNA interactions can be extracted from training adjacency matrix $\mathbf{F}_{train}$ whereas interactions for each lncRNA corresponds to each row of $\mathbf{F}_{train}$. In addition, the training adjacency matrix $\mathbf{F}_{train}$ is obtained by masking the known information where partial known elements in matrix are set to 0 as validation set.

1) GAUSSIAN INTERACTION PROFILE KERNEL

The information of the interactions is the connectivity behavior in the subjacent network [23]. Due to the broad applicability of the Gaussian kernel, we use the Gaussian Interaction Profile kernel (GIP) to device interactions kernel defined for lncRNA l_i and l_k $(i, k = 1, 2, \cdots, n)$ and protein P_j and P_s $(j, s = 1, 2, \cdots, m)$ respectively. Each element value in GIP is calculated as follows:

$$\mathbf{K}_{GIP}^{lnc}(l_i, l_k) = exp(-\gamma_{lnc}\|\mathbf{F}_{l_i} - \mathbf{F}_{l_k}\|^2) \tag{2a}$$

$$\mathbf{K}_{GIP}^{pro}(P_j, P_s) = exp(-\gamma_{pro}\|\mathbf{F}_{P_j} - \mathbf{F}_{P_s}\|^2) \tag{2b}$$

where $\mathbf{F}_{l_i}$ and $\mathbf{F}_{l_k}$ are the information of interactions for vector lncRNA l_i and l_k, $\mathbf{F}_{P_j}$ and $\mathbf{F}_{P_s}$ are the information of interactions for vector protein P_j and P_s. In practice, the Gaussian kernel bandwidths γ_{lnc} and γ_{pro} are set to 1.

2) SEQUENCE SIMILARITY KERNEL

A sequence $\mathbf{S}$ with length q is an ordered list of characters. Inspired by the notions of [25], we use the normalized Smith-Waterman (SW) score to measure the sequence similarity between two sequences according to the following functions (3a) and (3b).

$$\mathbf{K}_{SW}^{lnc}(l_i, l_k) = \frac{SW(\mathbf{S}_{l_i}, \mathbf{S}_{l_k})}{\sqrt{SW(\mathbf{S}_{l_i}, \mathbf{S}_{l_i})}\sqrt{SW(\mathbf{S}_{l_k}, \mathbf{S}_{l_k})}} \tag{3a}$$

$$\mathbf{K}_{SW}^{pro}(P_j, P_s) = \frac{SW(\mathbf{S}_{P_j}, \mathbf{S}_{P_s})}{\sqrt{SW(\mathbf{S}_{P_j}, \mathbf{S}_{P_j})}\sqrt{SW(\mathbf{S}_{P_s}, \mathbf{S}_{P_s})}} \tag{3b}$$

where $SW(\cdot, \cdot)$ is Smith-Waterman score; $\mathbf{S}_{l_i}$ and $\mathbf{S}_{l_k}$ represent the information of sequences for vector lncRNA l_i and l_k,

respectively; $\mathbf{S}_{P_j}$ and $\mathbf{S}_{P_s}$ refer to the information of sequences for vector protein P_j and P_s, respectively.

3) SEQUENCE FEATURE KERNEL

Conjoint Triad (CT) [42] and Pseudo Position-Specific Score Matrix (Pse-PSSM) [43] are used to describe lncRNA and protein sequences, respectively. Both two Sequence Feature (SF) kernels $\mathbf{K}_{SF}^{lnc}$ and $\mathbf{K}_{SF}^{pro}$ are built by Radial Basis Function kernel (RBF) with bandwidth value equal to 1.

4) LNCRNA EXPRESSION KERNEL

We utilize expression profiles of lncRNAs in 24 cell types which are gleaned from the NONCODE database [44]. Hence, each lncRNA can be represented as a 24-dimensional expression profile vector. The kernel of the lncRNAs expression $\mathbf{K}_{EXP}^{lnc}$ is produced by a RBF, and the kernel bandwidth is also set to 1.

C. KERNEL TARGET ALIGNMENT

The MKL model, uses 4 kernels in the lncRNA space including $\mathbf{K}_{GIP}^{lnc}$, $\mathbf{K}_{SW}^{lnc}$, $\mathbf{K}_{SF}^{lnc}$ and $\mathbf{K}_{EXP}^{lnc}$, and 3 kernels of protein space including $\mathbf{K}_{GIP}^{pro}$, $\mathbf{K}_{SW}^{pro}$, and $\mathbf{K}_{SF}^{pro}$. Consequently, we need to combine these kernels by means of linear combination in order to achieve the optimal ones. The optimal lncRNA kernel can be formulated according to (4a) and (4b).

$$\mathbf{K}_{lnc}^{*} = \sum_{a=1}^{4} w_a^{lnc} \mathbf{K}_a^{lnc}, \quad \mathbf{K}_a^{lnc} \in \Re^{n \times n} \tag{4a}$$

$$\mathbf{K}_{pro}^{*} = \sum_{a=1}^{3} w_a^{pro} \mathbf{K}_a^{pro}, \quad \mathbf{K}_a^{pro} \in \Re^{m \times m} \tag{4b}$$

where w denotes the weight of each kernel, and a represents the corresponding type of kernel.

In previous studies, both Qiu and Lane [45] and Gereke *et al.* [46] have employed Kernel Target Alignment (KTA) to estimate the corresponding weights. It is a fact that the score of KTA can be considered as the correlation between two kernels. The main idea is that larger alignment to $\mathbf{F}_{train}$ produces higher contribution to the combined kernel of a kernel matrix and vice versa. In this study, w_a^{lnc} that corresponds to the score between $\mathbf{K}_a^{lnc}$ and the ideal kernel matrix $\mathbf{K}_{ideal}$, is calculated as follows:

$$\mathbf{K}_{ideal} = \mathbf{F}_{train} \mathbf{F}_{train}^{\mathrm{T}} \quad \mathbf{F}_{train} \in \Re^{n \times m}. \tag{5}$$

The train set matrix $\mathbf{F}_{train}$ is obtained by masking the labels of the test set $\mathbf{F}$, setting all test labels to 0. Hence, we get the alignment score according to (6).

$$w_a^{lnc} = \frac{\left\langle \mathbf{K}_a^{lnc}, \mathbf{K}_{ideal} \right\rangle_F}{\|\mathbf{K}_a^{lnc}\|_F \|\mathbf{K}_{ideal}\|_F} \tag{6}$$

where $\left\langle \mathbf{K}_a^{lnc}, \mathbf{K}_{ideal} \right\rangle_F$ is Frobenius inner product $Trace(\cdot)$.

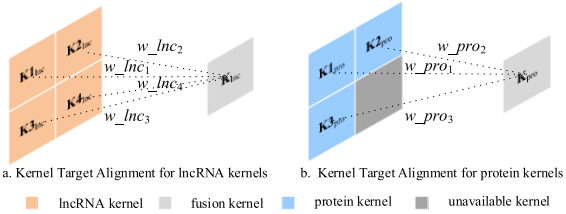

FIGURE 3. The schematic diagram of Kernel Target Alignment(KTA). The value of alignment score w_a is calculated according to the (6).

Actually, the lncRNA kernel weights in Fig. 3 are given by vector $\boldsymbol{w}_{lncRNA}$ whose values of elements are normalized by using (7).

$$\boldsymbol{w}_a^{lnc} = \frac{w_a^{lnc}}{\sum_{a=1}^{4} w_a^{lnc}} \tag{7}$$

Similarly, we can also obtain protein kernel weights vector $\boldsymbol{w}_{protein}$, so that $\boldsymbol{w}_{lncRNA}$ and $\boldsymbol{w}_{protein}$ can be used to conflate these kernels.

Finally, the SLP algorithm is used to estimate the link strength between lncRNAs and proteins.

D. SEMI-SUPERVISED LINK PREDICTION WITH APPROXIMATE LINK PROPAGATION

1) ORIGINAL MODEL

In order to extrapolate the link strength for the interactions with undetermined link state, the elements of adjacency matrix $\mathbf{F}$ are temporarily set to 0, and adjacency matrix $\mathbf{F}$ is treated as a training dataset in supervised learning, which is positive and unlabeled. We also respectively construct similarity matrices $\mathbf{K}_{lnc}$ and $\mathbf{K}_{pro}$ for each $F_{i,j}$ in matrix $\mathbf{F}$ (a bipartite network). As a further parenthetical explanation, these matrices are non-negatively symmetric.

Raymond and Kashima [41] developed a scenario of semi-supervised learning, which can be applied in predicting the link of a bipartite network. The basic assumption of SLP is that there is a high probability to have equal link strength when a pair of elements in $\mathbf{F}$ are similar. The general objective function of SLP is defined in (8).

$$\min_{\mathbf{F}^*} \quad \frac{\sigma}{2} vec(\mathbf{F}^*)^{\mathrm{T}} \mathbf{L} vec(\mathbf{F}^*) + \frac{1}{2} \|vec(\mathbf{F}^*) - vec(\mathbf{F})\|_2^2 \tag{8}$$

where $vec(\cdot)$ is a vectorization operator that can generate the arrangement of elements in one column, and $\mathbf{F}^*$ denotes the new link strength of the adjacency matrix $\mathbf{F}$ which can be estimated with SLP.

The first term of (8) is the similarity measurement that can justify whether two selected link strength values $F_{i,j}^*$ and $F_{p,q}^*$ for the corresponding two pairs can be viewed as neighbors. The second term in this equation represents loss function that aims to fit the predicting result $\mathbf{F}^*$ (partial known links $\mathbf{F}$ in network). Regularization parameter σ can balance two terms in (8).

2) GENERATE LAPLACIAN MATRIX

In this study, $\mathbf{L}$ represents Laplacian matrix of Kronecker kernel matrix $\mathbf{K}$, which can formulated as in (9).

$$\mathbf{L} = \mathbf{D} - \mathbf{K} \tag{9}$$

where $\mathbf{D}$ denotes a diagonal matrix with diagonal elements $\sum_j K_{ij}$. Thus we can naturally represent the Laplacian Matrices $\mathbf{L}_{lnc}$ and $\mathbf{L}_{pro}$ in a similar way.

In order to induce the statistical distribution of the uniform samples, the Laplacian matrices $\mathbf{L}_{lnc}$ and $\mathbf{L}_{pro}$ can be normalized as follows:

$$\mathbf{L}_{lnc} = \mathbf{I}_{lnc} - \mathbf{D}_{lnc}^{-\frac{1}{2}}\mathbf{K}_{lnc}\mathbf{D}_{lnc}^{-\frac{1}{2}}, \quad \mathbf{K}_{lnc} \in \mathbf{R}^{n\times n} \tag{10a}$$

$$\mathbf{L}_{pro} = \mathbf{I}_{pro} - \mathbf{D}_{pro}^{-\frac{1}{2}}\mathbf{K}_{pro}\mathbf{D}_{pro}^{-\frac{1}{2}}, \quad \mathbf{K}_{pro} \in \mathbf{R}^{m\times m} \tag{10b}$$

where $\mathbf{I}_{lnc}$ and $\mathbf{I}_{pro}$ denote identity matrices for lncRNA and protein, respectively.

Suppose that $\mathbf{X}$ is an $s \times t$ matrix and $\mathbf{Y}$ is a $u \times v$ matrix. The Kronecker sum is the $su \times tv$ block matrix. For the sake of combining multiple kernels as a whole, we use the $\oplus$ to surrogate Kronecker sum. Kronecker sum Laplacian can be formulated as

$$\mathbf{L} = \mathbf{D}_{pro} \oplus \mathbf{D}_{lnc} - \mathbf{K}_{pro} \oplus \mathbf{K}_{lnc} \tag{11a}$$

$$\mathbf{L} = (\mathbf{D}_{pro} - \mathbf{K}_{pro}) \oplus (\mathbf{D}_{lnc} - \mathbf{K}_{lnc}) \tag{11b}$$

where $\mathbf{D}_{pro}$ and $\mathbf{L}_{pro}$ are mutually similar matrices.

Meanwhile, to facilitate the dealing procedure, Raymond and Kashima [41] have used the normalized versions of the Laplacian matrices in (11a). The normalized Kronecker Laplacian sum matrix is given in (12a).

$$\mathbf{L} = 3\mathbf{I} - (\mathbf{D}_{pro}^{-\frac{1}{2}}\mathbf{K}_{pro}\mathbf{D}_{pro}^{-\frac{1}{2}} \oplus \mathbf{D}_{lnc}^{-\frac{1}{2}}\mathbf{K}_{lnc}\mathbf{D}_{lnc}^{-\frac{1}{2}}) \tag{12a}$$

$$\mathbf{L} = 3\mathbf{I} - (\tilde{\mathbf{K}}_{pro} \oplus \tilde{\mathbf{K}}_{lnc}) \tag{12b}$$

3) LOW RANK APPROXIAMTION

We have mentioned that $\mathbf{F}^*$ is the prediction score matrix. Here we give its form as in (13).

$$vec(\mathbf{F}^*) = ((1+3\sigma)\mathbf{I} - \sigma\tilde{\mathbf{K}}_{pro} \oplus \tilde{\mathbf{K}}_{lnc})^{-1} vec(\mathbf{F}) \tag{13}$$

Nevertheless, it will lead to $\mathcal{O}(nm \times nm)$ that calculates the inverse of matrix when using Kronecker sum. In order to avoid occupying too much physical memory during the accounting process, the SLP adopts the matrix approximation manner [41] to reduce the computing load. Thus, we can obtain the low-rank approximation of the kernel matrices $\mathbf{K}_{lnc}$ and $\mathbf{K}_{pro}$ as follows in (14a) and (14b).

$$\mathbf{K}_{lnc} \approx \mathbf{G}_{lnc}\mathbf{G}_{lnc}^{\mathrm{T}}, \quad \mathbf{G}_{lnc} \in \mathbf{R}^{n\times r_1} \tag{14a}$$

$$\mathbf{K}_{pro} \approx \mathbf{G}_{pro}\mathbf{G}_{pro}^{\mathrm{T}}, \quad \mathbf{G}_{pro} \in \mathbf{R}^{m\times r_2} \tag{14b}$$

where r_1 and r_2 are the parameters of the approximate matrices with $0 < r_1 < n$ and $0 < r_2 < m$, respectively.

In addition, the algebraic sum of each row in the approximate matrices for lncRNA can be obtained as in (15).

$$\mathbf{D}_{lnc} = diag(\mathbf{G}_{lnc}\mathbf{G}_{lnc}^{\mathrm{T}}\mathbf{l}) \tag{15}$$

And the formula of $\mathbf{D}_{pro}$ has similar form which only needs to exchange $\mathbf{G}_{lnc}$ and its transpose matrix with $\mathbf{G}_{pro}$ and $\mathbf{G}_{pro}^{\mathrm{T}}$, respectively. Therefore, normalized kernel matrices can be represented as (16) and (17).

$$\tilde{\mathbf{K}}_{lnc} \approx (\mathbf{D}_{lnc}^{-\frac{1}{2}}\mathbf{G}_{lnc})(\mathbf{G}_{lnc}^{\mathrm{T}}\mathbf{D}_{lnc}^{-\frac{1}{2}}) \tag{16a}$$

$$\tilde{\mathbf{K}}_{lnc} \approx \tilde{\mathbf{G}}_{lnc}\tilde{\mathbf{G}}_{lnc}^{\mathrm{T}} \tag{16b}$$

$$\tilde{\mathbf{K}}_{pro} \approx (\mathbf{D}_{pro}^{-\frac{1}{2}}\mathbf{G}_{pro})(\mathbf{G}_{pro}^{\mathrm{T}}\mathbf{D}_{pro}^{-\frac{1}{2}}) \tag{17a}$$

$$\tilde{\mathbf{K}}_{pro} \approx \tilde{\mathbf{G}}_{pro}\tilde{\mathbf{G}}_{pro}^{\mathrm{T}} \tag{17b}$$

4) EIGENDECOMPOSITION

The formulation of the diagonal matrices about lncRNA and protein are represented by $\bar{\Lambda}_{lnc}$ and $\bar{\Lambda}_{pro}$, so it is obvious that the sizes of these two matrices are r_1 and r_2, respectively. Therefore, the eigendecomposition of $\tilde{\mathbf{G}}_{lnc}^{\mathrm{T}}\tilde{\mathbf{G}}_{lnc}$ and $\tilde{\mathbf{G}}_{pro}^{\mathrm{T}}\tilde{\mathbf{G}}_{pro}$ can be easily obtained according to the eigenvalues for concrete forms as (18a) and (18b).

$$\tilde{\mathbf{G}}_{lnc}^{\mathrm{T}}\tilde{\mathbf{G}}_{lnc} = \bar{\mathbf{U}}_{lnc}\bar{\Lambda}_{lnc}\bar{\mathbf{U}}_{lnc}^{\mathrm{T}} \tag{18a}$$

$$\tilde{\mathbf{G}}_{pro}^{\mathrm{T}}\tilde{\mathbf{G}}_{pro} = \bar{\mathbf{U}}_{pro}\bar{\Lambda}_{pro}\bar{\mathbf{U}}_{pro}^{\mathrm{T}} \tag{18b}$$

Hence, the eigenvectors of the approximate kernel matrix $\tilde{\mathbf{K}}_{lnc}$ and $\tilde{\mathbf{K}}_{pro}$ can be obtained as (19a) and (19b).

$$\bar{\mathbf{V}}_{lnc} = \tilde{\mathbf{G}}_{lnc}\bar{\mathbf{U}}_{lnc}\bar{\Lambda}_{lnc}^{-\frac{1}{2}} \tag{19a}$$

$$\bar{\mathbf{V}}_{pro} = \tilde{\mathbf{G}}_{pro}\bar{\mathbf{U}}_{pro}\bar{\Lambda}_{pro}^{-\frac{1}{2}} \tag{19b}$$

In (13), $((1+3\sigma)\mathbf{I} - \sigma\tilde{\mathbf{K}}_{pro} \oplus \tilde{\mathbf{K}}_{lnc})^{-1}$ can be written as

$$\begin{aligned}&((1+3\sigma)\mathbf{I} - \sigma\bar{\mathbf{V}}diag(vec(\bar{\Lambda}))\bar{\mathbf{V}}^{\mathrm{T}})^{-1}\\ &= \frac{\mathbf{I}}{1+3\sigma} + \frac{\bar{\mathbf{V}}}{(1+3\sigma)^2}\left(\frac{diag(vec(\bar{\Lambda}))}{\sigma} - \frac{\mathbf{I}}{1+3\sigma}\right)^{-1}\bar{\mathbf{V}}^{\mathrm{T}}\end{aligned} \tag{20}$$

where $\bar{\mathbf{V}} = \bar{\mathbf{V}}_{pro} \oplus \bar{\mathbf{V}}_{lnc}$, and each element of the matrix $\bar{\Lambda}$ is defined as (21).

$$\bar{\Lambda}_{i,j} = \bar{\lambda}_{lnc}^{(i)} + \bar{\lambda}_{pro}^{(j)}, \quad 1 \le i \le r_1, \quad 1 \le j \le r_2 \tag{21}$$

Therefore, an approximate solution of the link prediction can be formulated as in (22).

$$vec(\mathbf{F}^*) = \frac{1}{1+3\sigma}vec(\mathbf{F}) + \frac{1}{(1+3\sigma)^2}\bar{\mathbf{V}}diag(vec(\bar{\mathbf{D}}))\bar{\mathbf{V}}^{\mathrm{T}}vec(\mathbf{F}) \tag{22}$$

where each element in $\bar{\mathbf{D}}$ is defined as

$$\bar{D}_{i,j} = \left(\frac{1}{\sigma\bar{\Lambda}_{i,j}} - \frac{1}{1+3\sigma}\right)^{-1} = \frac{\sigma(1+3\sigma)\bar{\Lambda}_{i,j}}{1+3\sigma-\sigma\bar{\Lambda}_{i,j}} \tag{23}$$

5) FINAL SEMI-SUPERVISED LINK PREDICTION FRAMEWORK

Recapitulating the aforementioned equations, $\mathbf{F}^*$ can be efficiently obtained as in (24).

$$\mathbf{F}^* = \frac{1}{1+3\sigma}\mathbf{F} + \frac{1}{(1+3\sigma)^2}\bar{\mathbf{V}}_{lnc}(\bar{\mathbf{D}} \odot (\bar{\mathbf{V}}_{lnc}^{T}\mathbf{F}\bar{\mathbf{V}}_{pro}))\bar{\mathbf{V}}_{pro}^{\mathbf{T}} \tag{24}$$

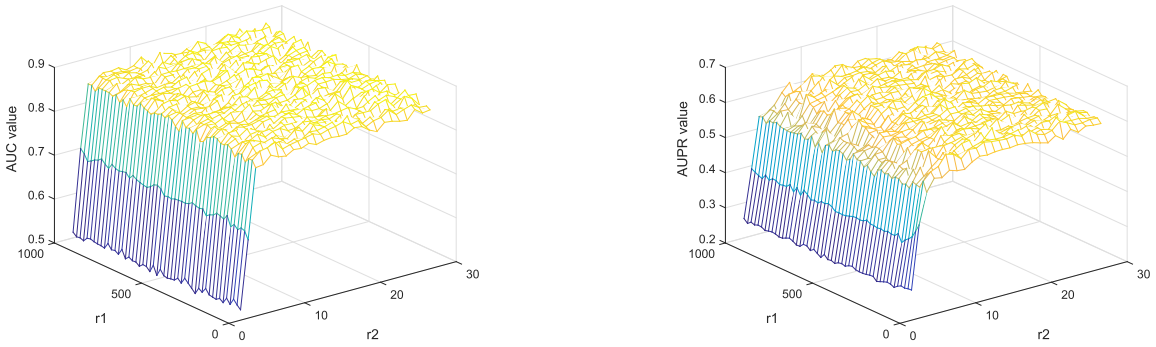

FIGURE 4. Grid search optimization for r_1 and r_2 on benchmark dataset.

Algorithm 1 Semi-Supervised Link Prediction Algorithm With Approximate Link Propagation in Predicting LPI

Input: Similarity matrices $\mathbf{K}_{lnc} \in \Re^{n\times n}$ and $\mathbf{K}_{pro} \in \Re^{m\times m}$; the bipartite network $\mathbf{F} \in \Re^{n\times m}$.

Output: The new link strength of the bipartite network $\mathbf{F}^*$.

1: Calculate the low-rank approximation matrices of $\mathbf{K}_{lnc}$ and $\mathbf{K}_{pro}$ by (14), and adjust the parameters of approximate matrices r_1 and r_2;
2: Calculate the normalized matrices $\tilde{\mathbf{G}}_{lnc}$ and $\tilde{\mathbf{G}}_{pro}$ by (16) and (17), and also analyze the eigendecomposition of $\tilde{\mathbf{G}}_{lnc}^T\tilde{\mathbf{G}}_{lnc}$ (18a) and $\tilde{\mathbf{G}}_{pro}^T\tilde{\mathbf{G}}_{pro}$ (18b), and then obtain the eigenvectors $\bar{\mathbf{V}}_{lnc}$ and $\bar{\mathbf{V}}_{pro}$ by (19a) and (19b);
3: Calculate the elements of the matrix $\bar{\mathbf{D}}$ by (23), and adjust the regularization parameter σ;
4: Calculate the elements of $\mathbf{F}^*$ by (24).

where $\odot$ denotes the Hadamard product of two matrices, i.e. element-wise multiplication, also known as the Schur product [47].

The objective function about the SLP is illustrated by elucidating in Algorithm 1.

III. RESULTS

Here we use the benchmark dataset to evaluate our approach, in order to ensure fairness and objectivity and to conduct an independent analysis of the single kernel performance. Moreover, the way of KTA is not only compared with a mean weighted model but also has been assessed in a parallel comparison to well established algorithms. Additionally, we utilize the case study to evaluate our method in predicting unknown lncRNA-protein interactions. The results are available at https://github.com/6gbluewind/LPI_KTASLP.

A. BENCHMARK DATASET

NPInter database stores experimentally verified interactions between non-coding RNAs and other biomolecules such as genomic DNAs and proteins. In addition, NONCODE [44] is an integrated knowledge database, which has collected non-coding RNAs. Zhang *et al.* [9] obtained experimentally determined lncRNA-protein interactions with 1114 lncRNAs and 96 proteins from NPInter V2.0 [48]. The sequence information of proteins is collected from the SUPERFAMILY database [49]. To facilitate computation, Zhang removed lncRNAs and proteins whose expression or sequence information is unavailable. Those lncRNAs and proteins with only one interaction were also removed. Finally, they collected a dataset with 4158 lncRNA-protein interactions which contains 990 lncRNAs and 27 proteins.

B. EVALUATION MEASUREMENTS

To test the stability of our model, treatments such as randomly selecting training set and test set, model-building and model-evaluating, have been applied in five-fold Cross Validation (5-fold CV). The Area Under ROC curve (AUC) and Area Under the Precision-Recall curve (AUPR) measures have been used to evaluate our method. Due to the sparsity of true lncRNA-protein interactions, AUPR is more significant than AUC as a quality measurement.

C. EXPERIMENTAL ENVIRONMENT

In this paper, our developed predictor has been implemented by using MATLAB to carry out. All programs have been validated on a computer with 3.8 GHz 4-core CPU, 20 GB memory and Windows operating system. The optimal regularization parameter σ is set as 0.125, which is obtained by enumerating the best value from number set $\{2^{-5}, 2^{-4}, \cdots, 2^{0}, \cdots, 2^{5}\}$.

D. PARAMETER OPTIMIZATION

Grid search schema is adopted to get the optimized parameters of approximate matrices r_1 and r_2. The range of r_1 is from 20 to 980 and each step is 20. Similarly, r_2 is from 2 to 27 and each step is 1. We have selected the optimal values of r_1 and r_2 by the highest AUPR value and the lowest values of r_1 and r_2, because the smaller the values of r_1 and r_2, the less the running time of the algorithm. We find that $r_1 = 160$ and $r_2 = 17$ are the best parameters (AUPR: 0.6148) as shown in Fig. 4.

E. PERFORMANCE ANALYSIS

In this subsection we analyze the greatest contribution of different kinds of kernel matrices, the use of a single kernel, the mean weighted kernels and the weighted kernels with KTA conducted, respectively. By testing these kernels

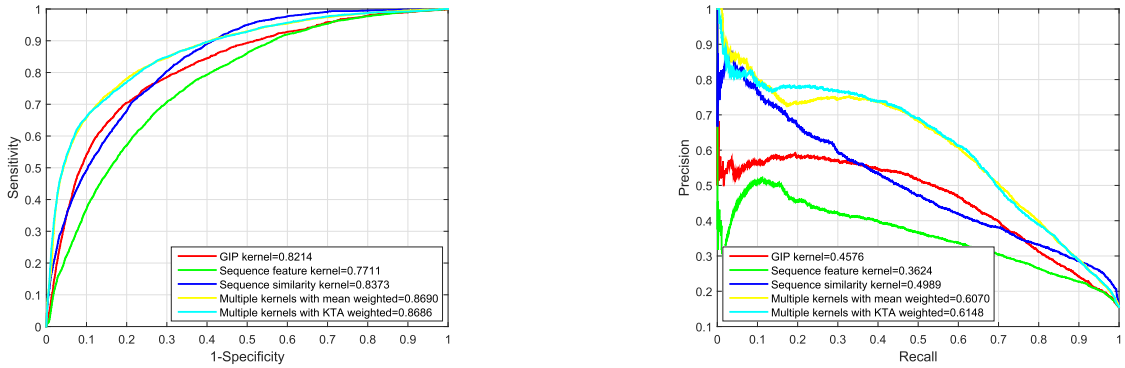

FIGURE 5. The ROC and PR curve of different kernels in 5-fold CV on benchmark dataset.

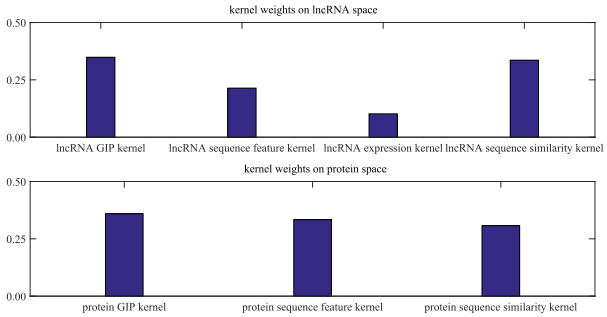

FIGURE 6. The kernel weights of LPI-KTASLP on benchmark dataset.

on the benchmark dataset in Fig. 5, we have obtained the following: The AUPRs of the GIP kernel, the sequence feature kernel and the sequence similarity kernel are 0.4576, 0.3624 and 0.4989, respectively. The AUPRs of the sequence similarity kernel is higher than the AUPRs of other single kernels. Moreover, the AUPR of the multiple kernels of the mean weighted model has reached a value of 0.6070 that is better than all single kernels. Multiple kernels with KTA weighted model, achieve AUPR equal to 0.6148, which is an outstanding performance. In Fig. 5, we can see that the KTA performs better than the other models. Obviously, KTA is helpful for improving the performance of predicting lncRNA-protein interactions.

In addition, Fig.6 shows the weight of each kernel, including lncRNA space and protein space. Obviously, weights of the GIP kernel obtain the largest values on the whole space. It is clearly that lncRNA expression kernel has a low weight in lncRNA space which has been shown in Fig.6. It is illustrated that the weighting strategy has reached its goal to select the optimal combination of kernels. Expression information has not played a significant role, but it indeed has a certain weight to contribute the result as an advantage.

F. COMPARISON TO EXISTING PREDICTORS

Our approach is also compared with other existing methods on the benchmark dataset, showed in Table 1. We observe that the highest AUPR of 0.6148 is obtained by our proposed method, which is superior to: Integrated LPLNP (AUPR: 0.4584) [9], RWR (AUPR: 0.2827) [50], CF (AUPR: 0.2357) [51], LPIHN (AUPR: 0.2299) [18] and LPBNI (AUPR: 0.3302) [19]. There are two possible reasons for the satisfied performance. Firstly, the KTA effectively combines multivariate information by exploiting Multiple Kernel Learning. Simultaneously, LPI-KTASLP is an effective semi-supervised link prediction algorithm which employs Kronecker sum to fuse lncRNA and protein feature spaces. Since the imbalance of the number of lncRNAs and proteins can lead to prediction difficulties, PRC is more effective than ROC on highly imbalanced datasets. Therefore, under the situation that we acquire competitive AUC value among the state-of-the-art schemes, our method can be applied in the extrapolation of LPI.

TABLE 1. Comparison to the existing methods in 5-fold CV on benchmark dataset.

Methods	AUPR	AUC
LPI-KTASLP	**0.6148**	0.8686
Integrated LPLNP*	0.4584	**0.9104**
RWR*	0.2827	0.8134
CF*	0.2357	0.7686
LPIHN*	0.2299	0.8451
LPBNI*	0.3302	0.8569

* Results are derived from [9].

G. COMPARISON TO THE OTHER SEQUENCE FEATURE EXTRACTION METHODS

We have extracted CT and PsePSSM as features to build sequence feature kernel. Because we want to illustrate the outcome of the comparison with respect to sequence feature extraction methods, including CT, PsePSSM, Ngram [52], HOG [53] and AVBLOCK [54], we show the results in Table 2. We notice that CT-PsePSSM gets AUPR of 0.6148, which is higher than the results of Ngram-HOG (AUPR: 0.5735) and Ngram-AVBLOCK (AUPR: 0.5869).

H. COMPARISON BETWEEN KTA AND OTHER KERNEL FUSION METHODS

To compare KTA with other kernel fusion methods, we also find several other weighted models, including

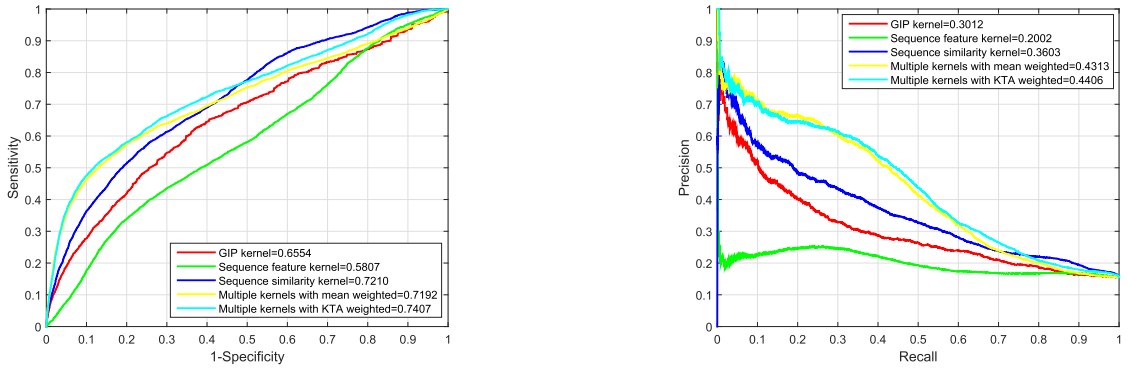

FIGURE 7. The ROC and PR curve by local LOOCV on benchmark dataset.

TABLE 2. Comparison between CT/PsePSSM and other feature extraction methods on benchmark dataset.

Type of sequence feature	AUPR	AUC
LPI-KTASLP (CT/PsePSSM)	**0.6148**	**0.8686**
N-gram/HOG	0.5735	0.8584
N-gram/AVBLOCK	0.5869	0.8684

TABLE 3. Comparison between KTA and other kernel fusion methods on benchmark dataset.

Type of sequence feature	AUPR	AUC
LPI-KTASLP (KTA)	**0.6148**	0.8686
FastKL	0.5418	0.8493
mas_UMKL	0.4696	0.8146
sparse_UMKL	0.6095	**0.8754**

fast kernel learning (FastKL) [55], mas_UMKL [56] and sparse_UMKL [57]. The outcome of the comparison are listed in Table 3. We can see clearly that KTA achieves the highest AUPR value, which gets 0.6148, and is superior to FastKL (0.5418), mas_UMKL (0.4696) and sparse_UMKL (0.6095). Simultaneously, we also notice that KTA and sparse_UMKL have obtained competitive performance both on AUPR and AUC.

I. CASE STUDY

Local Leave-One-Out Cross-Validation (LOOCV) used to evaluate the predictive performance. The aim of local LOOCV is to verify the capability of the algorithm that deals the samples without any known associations, while global LOOCV assesses the capability of a model that for predicting a known specific association. Compared with global LOOCV, local LOOCV is better for assessing a model. Consequently, we adopt this kind of strategy to draw as the evaluation scenario. Local LOOCV masks the relationships between one protein and all lncRNAs. Our model is trained by the rest of the known information, no matter whether they interacted or not and whether they were tested on masked relationships. For a protein not appearing in trial, our method can estimate the strength of interactions between this protein and gross 990 lncRNAs in the experiment. Then, we rank

TABLE 4. Top 20 novel interactions on protein ENSP00000309558.

lncRNA ID	Protein ID	Ranks	Confirm
NONHSAT135828	ENSP00000309558	1	Confirmed
NONHSAT135822	ENSP00000309558	2	–
NONHSAT021830	ENSP00000309558	3	Confirmed
NONHSAT137541	ENSP00000309558	4	Confirmed
NONHSAT135796	ENSP00000309558	5	Confirmed
NONHSAT001511	ENSP00000309558	6	Confirmed
NONHSAT041921	ENSP00000309558	7	Confirmed
NONHSAT134595	ENSP00000309558	8	–
NONHSAT104991	ENSP00000309558	9	Confirmed
NONHSAT027070	ENSP00000309558	10	Confirmed
NONHSAT009703	ENSP00000309558	11	Confirmed
NONHSAT084827	ENSP00000309558	12	–
NONHSAT078782	ENSP00000309558	13	Confirmed
NONHSAT138142	ENSP00000309558	14	Confirmed
NONHSAT011652	ENSP00000309558	15	Confirmed
NONHSAT104639	ENSP00000309558	16	Confirmed
NONHSAT007698	ENSP00000309558	17	–
NONHSAT054716	ENSP00000309558	18	–
NONHSAT002344	ENSP00000309558	19	Confirmed
NONHSAT079374	ENSP00000309558	20	–

these strength of the interactions in a descending order, since the high ranking with high interaction possibility. In Fig. 7, we can see the performances of single kernel, the mean weighted kernels and the weighted kernels with KTA. Multiple kernels with KTA weighted model also gain the best performance with the AUPR value equal to 0.4406 and the AUC value equal to 0.7407.

As it is shown in Tables 4 and 5, two cases including proteins ENSP00000309558 and ENSP00000401371 of the top 20 interactions are extrapolated by the LPI-KTASLP. We have checked them up in the masked relationships between one protein and all lncRNAs. Our approach achieves identification ratio of 14/20 and 12/20 on proteins ENSP00000309558 and ENSP00000401371, respectively.

J. EVALUATION ON ZHENG DATASET

In order to measure the stability of our model from experimental point of view, we further employ another dataset, which is mentioned in a recent publication of Zheng *et al.* [58] to corroborate the capability of prediction. The size of this dataset is bigger than the benchmark one, especially the protein number, which illustrated in Table 6.

TABLE 5. Top 20 novel interactions on protein ENSP00000401371.

lncRNA ID	Protein ID	Ranks	Confirm
NONHSAT021830	ENSP00000401371	1	Confirmed
NONHSAT078782	ENSP00000401371	2	–
NONHSAT137541	ENSP00000401371	3	Confirmed
NONHSAT135828	ENSP00000401371	4	–
NONHSAT135796	ENSP00000401371	5	Confirmed
NONHSAT135822	ENSP00000401371	6	–
NONHSAT041921	ENSP00000401371	7	Confirmed
NONHSAT030153	ENSP00000401371	8	–
NONHSAT084827	ENSP00000401371	9	–
NONHSAT002344	ENSP00000401371	10	Confirmed
NONHSAT101154	ENSP00000401371	11	–
NONHSAT022115	ENSP00000401371	12	Confirmed
NONHSAT079548	ENSP00000401371	13	Confirmed
NONHSAT027070	ENSP00000401371	14	Confirmed
NONHSAT001953	ENSP00000401371	15	Confirmed
NONHSAT104991	ENSP00000401371	16	Confirmed
NONHSAT009703	ENSP00000401371	17	Confirmed
NONHSAT080206	ENSP00000401371	18	–
NONHSAT145960	ENSP00000401371	19	Confirmed
NONHSAT104639	ENSP00000401371	20	–

TABLE 6. The information of two datasets in the experiment.

dataset	number of lncRNAs	number of proteins	LPIs
Benchmark*	990	27	4158
Zheng*	1050	84	4467

* The benchmark dataset and the Zheng dataset come from the paper of Zhang et al. [9] and Zheng et al. [59], respectively.

TABLE 7. The AUPR and AUC of different methods on Zheng dataset.

Methods	AUPR	AUC
LPI-KTASLP	**0.7173**	**0.9152**
PPSNs	–*	0.9098

* AUPR indicator is not exploited in the experiment of Zheng et al. [9].

Homo sapiens lncRNA-protein interactions come from NPInter (v2.0), while protein sequences are obtained from the UniProt database and lncRNA database is NONCODE (v4.0). There are totally 4, 467 lncRNA-protein interactions in the dataset of Zheng *et al.* [58], involving 1, 050 lncRNAs and 84 proteins. We have adopted 5-fold CV on the the Zheng dataset, and have compared LPI-KTASLP with the results of Zheng *et al.* in Table 7. The value of AUC in the LPI-KTASLP is 0.9152, which is higher than the PPSNs. Moreover, the application on the Zheng dataset (AUPR=0.7173) represents the stability of the LPI-KTASLP on an imbalanced dataset.

IV. CONCLUSIONS AND DISCUSSIONS

In this paper, we have proposed a novel prediction method of lncRNAs-protein interaction by using a semi-supervised MKL learning approach. LPI-KTASLP employs the operator of Kronecker sum to fuse lncRNA and protein spaces. Then, semi-supervised learning is used to estimate the strength of interactions between lncRNAs and proteins. Towards the application of five-fold cross validation (5-fold CV) on benchmark dataset, LPI-KTASLP can achieve better results on the benchmark dataset. Furthermore, a comparison with state-of-the-art methods, proves that LPI-KTASLP achieves satisfactory performance and demonstrates the robustness of our model.

Because the topic of ncRNA is so fashionable, that currently attracts a lot of people continuously pursue the cutting edge. For instance, the AUPR of method IRWNRLPI [22], which is higher than our result, is due to integration strategy. IRWNRLPI combines random walk and neighborhood regularized logistic matrix factorization to avoid the unsatisfactory result of using one of the two methods alone. Moreover, IRWNRLPI adopts global LOOCV to assess the capability of a model that for predicting a known specific association, while local LOOCV in our experiment is used to verify the capability of the algorithm that deals the samples without any known associations.

In the future, we can further improve the prediction by adding more related information, such as available 3D structure data, or take the integration strategy. More similarity matrices can be constructed by employing various types of classical distance measures. And we can also construct a mini-repository for kernels, through considering the selections including of the RBF kernel subtypes and of the conditional definition kernels.

REFERENCES

[1] M. Guttman and J. L. Rinn, "Modular regulatory principles of large non-coding RNAs," *Nature*, vol. 482, no. 7385, pp. 339–346, 2012.
[2] X. Chen, D. Xie, Q. Zhao, and Z.-H. You, "MicroRNAs and complex diseases: From experimental results to computational models," *Briefings Bioinf.*, vol. 18, no. 4, p. 558, 2017.
[3] X. Chen, D. Xie, L. Wang, Q. Zhao, Z.-H. You, and H. Liu, "BNPMDA: Bipartite network projection for miRNA–disease association prediction," *Bioinformatics*, vol. 34, no. 18, pp. 3178–3186, 2018.
[4] X. Chen, L. Huang, D. Xie, and Q. Zhao, "EGBMMDA: Extreme gradient boosting machine for MiRNA-disease association prediction," *Cell Death Disease*, vol. 9, no. 1, p. 3, 2018.
[5] X. Chen and L. Huang, "LRSSLMDA: Laplacian regularized sparse subspace learning for MiRNA-disease association prediction," *PLOS Comput. Biol.*, vol. 13, no. 12, 2017, Art. no. e1005912.
[6] H. Hu *et al.*, "HLPI-Ensemble: Prediction of human lncRNA-protein interactions based on ensemble strategy," *RNA Biol.*, vol. 15, no. 6, pp. 797–806, 2018.
[7] M. Jiang *et al.*, "Self-recognition of an inducible host lncRNA by RIG-I feedback restricts innate immune response," *Cell*, vol. 173, no. 4, pp. 906–919, 2018.
[8] S. Jalali, S. Kapoor, A. Sivadas, D. Bhartiya, and V. Scaria, "Computational approaches towards understanding human long non-coding RNA biology," *Bioinformatics*, vol. 31, no. 14, pp. 2241–2251, 2015.
[9] W. Zhang, Q. Qu, Y. Zhang, and W. Wang, "The linear neighborhood propagation method for predicting long non-coding RNA–protein interactions," *Neurocomputing*, vol. 273, pp. 526–534, Jan. 2017.
[10] F. Guo, S. C. Li, Z. Wei, D. Zhu, C. Shen, and L. Wang, "Structural neighboring property for identifying protein-protein binding sites," *BMC Syst. Biol.*, vol. 9, no. 5, pp. 1–9, 2015.
[11] F. Guo, S. C. Li, Y. Fan, and L. Wang, "Identifying protein-protein binding sites with a combined energy function," *Current Protein Peptide Sci.*, vol. 15, no. 6, pp. 540–552, 2014.
[12] F. Guo, S. C. Li, P. Du, and L. Wang, "Probabilistic models for capturing more physicochemical properties on protein–protein interface," *J. Chem. Inf. Model.*, vol. 54, no. 6, pp. 1798–1809, 2014.
[13] M. Bellucci, F. Agostini, M. Masin, and G. G. Tartaglia, "Predicting protein associations with long noncoding RNAs," *Nature Methods*, vol. 8, no. 6, pp. 444–445, 2011.
[14] U. K. Muppirala, V. G. Honavar, and D. Dobbs, "Predicting RNA-protein interactions using only sequence information," *BMC Bioinf.*, vol. 12, no. 1, p. 489, 2011.
[15] Y. Wang *et al.*, "De novo prediction of RNA–protein interactions from sequence information," *Mol. Biosyst.*, vol. 9, no. 1, pp. 133–142, 2012.
[16] Q. Lu *et al.*, "Computational prediction of associations between long non-coding RNAs and proteins," *BMC Genomics*, vol. 14, no. 1, p. 651, 2013.

[17] V. Suresh, L. Liu, D. Adjeroh, and X. Zhou, "RPI-Pred: Predicting ncRNA-protein interaction using sequence and structural information," *Nucleic Acids Res.*, vol. 43, no. 3, pp. 1370–1379, 2015.

[18] A. Li, M. Ge, Y. Zhang, C. Peng, and M. Wang, "Predicting long noncoding RNA and protein interactions using heterogeneous network model," *Biomed Res. Int.*, vol. 2015, 2015, Art. no. 671950.

[19] M. Ge, A. Li, and M. Wang, "A bipartite network-based method for prediction of long non-coding RNA–protein interactions," *Genomics, Proteomics Bioinf.*, vol. 14, no. 1, pp. 62–71, 2016.

[20] H. Hu *et al.*, "LPI-ETSLP: LncRNA–protein interaction prediction using eigenvalue transformation-based semi-supervised link prediction," *Mol. Biosyst.*, vol. 13, no. 9, pp. 1781–1787, 2017.

[21] H. Liu *et al.*, "LPI-NRLMF: LncRNA-protein interaction prediction by neighborhood regularized logistic matrix factorization," *Oncotarget*, vol. 8, no. 61, pp. 103975–103984, 2017.

[22] Q. Zhao, Y. Zhang, H. Hu, G. Ren, W. Zhang, and H. Liu, "IRWNRLPI: Integrating random walk and neighborhood regularized logistic matrix factorization for lncRNA-protein interaction prediction," *Frontiers Genet.*, vol. 9, p. 239, Jul. 2018.

[23] T. V. Laarhoven, S. B. Nabuurs, and E. Marchiori, "Gaussian interaction profile kernels for predicting drug–target interaction," *Bioinformatics*, vol. 27, no. 21, pp. 3036–3043, 2011.

[24] C. Shen, Y. Ding, J. Tang, X. Xu, and F. Guo, "An ameliorated prediction of drug–target interactions based on multi-scale discrete wavelet transform and network features," *Int. J. Mol. Sci.*, vol. 18, no. 8, p. 1781, 2017.

[25] A. C. A. Nascimento, R. B. C. Prudêncio, and I. G. Costa, "A multiple kernel learning algorithm for drug-target interaction prediction," *BMC Bioinf.*, vol. 17, no. 1, pp. 46–61, 2016.

[26] J. Zeng, D. Li, Y. Wu, Q. Zou, and X. Liu, "An empirical study of features fusion techniques for protein-protein interaction prediction," *Current Bioinf.*, vol. 11, no. 1, pp. 4–12(9), 2016.

[27] Y. Ding, J. Tang, and F. Guo, "Identification of protein–protein interactions via a novel matrix-based sequence representation model with amino acid contact information," *Int. J. Mol. Sci.*, vol. 17, no. 10, p. 1623, 2016.

[28] Y. Ding, J. Tang, and F. Guo, "Identification of drug-target interactions via multiple information integration," *Inf. Sci.*, vols. 418–419, pp. 546–560, Dec. 2017.

[29] Y. Ding, J. Tang, and F. Guo, "Predicting protein-protein interactions via multivariate mutual information of protein sequences," *BMC Bioinf.*, vol. 17, no. 1, p. 398, 2016.

[30] F. Guo, Y. Ding, S. C. Li, C. Shen, and L. Wang, "Protein–protein interface prediction based on hexagon structure similarity," *Comput. Biol. Chem.*, vol. 63, pp. 83–88, Aug. 2016.

[31] Y. Ding, J. Tang, and F. Guo, "Identification of residue-residue contacts using a novel coevolution-based method," *Current Proteomics*, vol. 13, no. 2, pp. 122–129, 2016.

[32] Y. Ding, J. Tang, and F. Guo, "Identification of drug-side effect association via multiple information integration with centered kernel alignment," *Neurocomputing*, vol. 325, pp. 211–224, Jan. 2018.

[33] Y. Ding, J. Tang, and F. Guo,, "Identification of drug-side effect association via semi-supervised model and multiple kernel learning," *IEEE J. Biomed. Health Inform.*, Nov. 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8550713

[34] L. Jiang, Y. Ding, J. Tang, and F. Guo, "MDA-SKF: Similarity kernel fusion for accurately discovering miRNA-disease association," *Frontiers Genet.*, vol. 9, p. 618, Dec. 2018.

[35] Q. Zou *et al.*, "Prediction of microRNA-disease associations based on social network analysis methods," *Biomed Res. Int.*, vol. 14, no. 10, 2015, Art. no. 810514.

[36] Y. Liu, X. Zeng, Z. He, and Q. Zou, "Inferring microRNA-disease associations by random walk on a heterogeneous network with multiple data sources," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 14, no. 4, pp. 905–915, Jul./Aug. 2016.

[37] Y. Ding, J. Tang, and F. Guo, "Identification of protein–ligand binding sites by sequence information and ensemble classifier," *J. Chem. Inf. Model.*, vol. 57, no. 12, pp. 3149–3161, 2017.

[38] C. Shen, Y. Ding, J. Tang, J. Song, and F. Guo, "Identification of DNA–protein binding sites through multi-scale local average blocks on sequence information," *Molecules*, vol. 22, no. 12, p. 2079, 2017.

[39] W. Zhang, X. Liu, Y. Chen, W. Wu, W. Wang, and X. Li, "Feature-derived graph regularized matrix factorization for predicting drug side effects," *Neurocomputing*, vol. 287, pp. 154–162, Apr. 2018.

[40] W. Zhang *et al.*, "Predicting drug-disease associations by using similarity constrained matrix factorization," *BMC Bioinf.*, vol. 19, no. 1, p. 233, Jun. 2018.

[41] R. Raymond and H. Kashima, "Fast and scalable algorithms for semi-supervised link prediction on static and dynamic graphs," in *Proc. ECML PKDD*, Athens, Greece, 2010, pp. 131–147.

[42] J. Shen *et al.*, "Predicting protein–protein interactions based only on sequences information," *Proc. Nat. Acad. Sci.*, vol. 104, no. 11, pp. 4337–4341, 2007.

[43] K.-C. Chou and H.-B. Shen, "MemType-2L: A Web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM," *Biochem. Biophys. Res. Commun.*, vol. 360, no. 2, pp. 339–345, 2007.

[44] C. Xie *et al.*, "NONCODEv4: Exploring the world of long non-coding RNA genes," *Nucleic Acids Res.*, vol. 42, no.D 1, pp. D98–D103, 2014.

[45] S. Qiu and T. Lane, "A framework for multiple kernel support vector regression and its applications to siRNA efficacy prediction," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 6, no. 2, pp. 190–199, Jun. 2009.

[46] N. Cristianini, J. Shawe-Taylor, A. Elisseeff, and J. S. Kandola, "On kernel-target alignment," in *Proc. Adv. Neural Inf. Process. Syst.*, 2001, pp. 367–373.

[47] C. Davis, "The norm of the Schur product operation," *Numer. Math.*, vol. 4, no. 1, pp. 343–344, 1962.

[48] J. Yuan, W. Wu, C. Xie, G. Zhao, Y. Zhao, and R. Chen, "NPInter v2.0: An updated database of ncRNA interactions," *Nucleic Acids Res.*, vol. 42, no. D1, pp. D104–D108, 2014.

[49] J. Gough, K. Karplus, R. Hughey, and C. Chothia, "Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure," *J. Mol. Biol.*, vol. 313, no. 4, pp. 903–919, 2001.

[50] M. Gan, "Walking on a user similarity network towards personalized recommendations," *Plos One*, vol. 9, no. 12, 2014, Art. no. e114662.

[51] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-based collaborative filtering recommendation algorithms," in *Proc. 10th Int. Conf. World Wide Web*, Hong Kong, China, 2001, pp. 285–295

[52] L. R. Murphy, A. Wallqvist, and R. M. Levy, "Simplified amino acid alphabets for protein fold recognition and implications for folding," *Protein Eng.*, vol. 13, no. 3, pp. 149–152, 2000.

[53] O. Ludwig, Jr., D. Delgado, V. Goncalves, and U. Nunes, "Trainable classifier-fusion schemes: An application to pedestrian detection," in *Proc. ITSC*, St. Louis, MO, USA, Oct. 2009, pp. 1–6.

[54] X. Lin and X. W. Chen, "On position-specific scoring matrix for protein function prediction," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 8, no. 2, pp. 308–315, Mar. 2011.

[55] J. He, S.-F. Chang, and L. Xie, "Fast kernel learning for spatial pyramid matching," in *Proc. IEEE CVPR*, Anchorage, Ak, USA, Jun. 2008, pp. 1–7.

[56] J. Zhuang, J. Wang, S. C. H. Hoi, and X. Lan, "Unsupervised multiple kernel learning," in *Proc. ACML*, vol. 20. South Garden Hotels and Resorts, Taoyuan, Taiwain: PMLR, Nov. 2011, pp. 129–144.

[57] J. Mariette and N. Villa-Vialaneix, "Unsupervised multiple kernel learning for heterogeneous data integration," *Bioinformatics*, vol. 34, no. 6, pp. 1009–1015, 2018.

[58] X. Zheng *et al.*, "Fusing multiple protein-protein similarity networks to effectively predict lncRNA-protein interactions," *BMC Bioinf.*, vol. 18, no. 12, p. 420, 2017.

[59] X. Zheng *et al.*, "Fusing multiple protein-protein similarity networks to effectively predict lncRNA-protein interactions," *BMC Bioinf.*, vol. 18, no. 12, p. 420, 2017.

CONG SHEN received the B.Sc. degree in computer science and technology from the Tianjin University of Technology, in 2010, and the M.Sc. degree from the Hebei University of Technology, in 2015. He is currently pursuing the Ph.D. degree with the School of Computer and Technology, College of Intelligence and Computing, Tianjin University. His research interests include bioinformatics and machine learning.

YIJIE DING received the B.Sc. and M.Sc. degrees from the Tianjin University of Science and Technology, in 2009 and 2012, respectively, and the Ph.D. degree from Tianjin University, in 2018. He is currently a Lecturer with the School of Electronic and Information Engineering, Suzhou University of Science and Technology. His research interests include bioinformatics and machine learning.

JIJUN TANG received the Ph.D. degree from The University of New Mexico, in 2004. He is currently a Professor with the Department of Computer Science and Engineering, University of South Carolina. He is also an Adjunct Professor with the School of Computer and Technology, College of Intelligence and Computing, Tianjin University. His main research interest includes computational biology, with focus in algorithms' development on phylogenetic reconstruction from genome rearrangement data.

LIMIN JIANG received the B.Sc. and M.Sc. degrees from the Hebei University of Engineering, in 2014 and 2017, respectively. She is currently pursuing the Ph.D. degree with the School of Computer and Technology, College of Intelligence and Computing, Tianjin University. Her research interests include bioinformatics and machine learning.

FEI GUO received the B.Sc. and Ph.D. degrees from Shandong University, in 2007 and 2012, respectively. She was a Postdoctoral Fellow with the City University of Hong Kong, between 2012 and 2013. She is currently an Associate Professor with the School of Computer and Technology, College of Intelligence and Computing, Tianjin University. Her research interests include bioinformatics and computational biology.