DMIL-III: Isoform-isoform interaction prediction using deep multi-instance learning method

Jie Zeng
College of Computer and Information Science Southwest University
Chongqing, China
zengj1219@swu.edu.cn

Guoxian Yu*
College of Computer and Information Science Southwest University, Chongqing, China;
and, Computational Bioscience Research Center, CEMSE division, KAUST, Thuwal, SA
gxyu@swu.edu.cn

Jun Wang
College of Computer and Information Science Southwest University
Chongqing, China
kingjun@swu.edu.cn

Maozu Guo
School of Electrical and Information Engineering Beijing University of Civil Engineering and Architecture
Beijing, China
guomaozu@bucea.edu.cn

Xiangliang Zhang
Computational Bioscience Research Center, CEMSE division, KAUST, Thuwal, SA
xiangliang.zhang@kaust.edu.sa

***Abstract*—Alternative splicing modulates protein-protein and other ligand interactions, it results in proteoforms, translated from isoforms that are alternatively spliced from the same gene, to interact with different partners and have distinct or even opposing functions. Therefore, systematically identifying protein-protein interaction at the *isoform-level* is crucial to explore the function of proteoforms. Constructing the isoform-level interaction network currently is prohibited by the lack of a large golden set of experimentally validated interacting isoforms, which enable computationally predicting isoform-isoform interactions. In this paper, a deep convolution neural network based multi-instance learning approach called DMIL-III is proposed to predict isoform interactions. DMIL-III takes a gene pair as 'bag' and two isoforms of the pairwise genes as the 'instance' of the bag. DMIL-III follows the principle of multi-instance learning that at least one isoform-isoform interaction exists for a positive gene pair and none interacting isoforms occurs for a negative gene pair. DMIL-III integrates RNA-seq, nucleotide sequence, domain-domain interaction and exon array data. Experimental results indicate that DMIL-III achieves a superior performance with Accuracy of 93% on single-instance gene bags and of 94% on multi-instance gene bags, which are at least 14% and 29% higher than those of state-of-the-art methods. In addition, we further test DMIL-III on a set of experimentally confirmed isoform-isoform interactions and obtain an Accuracy of 65%, which is at least 10% higher than those of comparing methods at the isoform-level. All these results show the effectiveness of DMIL-III for predicting isoform-isoform interactions.**

***Keywords*-Isoform-Isoform interaction, Deep multi-instance learning, Data fusion, Alternative splicing**

I. Introduction

Protein-protein interactions (PPIs) play key roles in executing and regulating fundamental cellular processes. Constructing and analyzing large-scale protein interaction networks can strengthen the understanding of the interaction between proteins and the functions of proteins. Although significant advances have been made, the current studies on PPI still focus on the *gene-level* [1], [2], there are few studies on *isoform-level* PPIs in small scale [3], [4]. In actual fact, gene-level PPIs are typically referred to the canonical (or the longest) protein translated from a gene, most existing studies neglect the effect of alternative splicing events [5].

Alternative splicing is a very common mechanism of gene expression regulation. It is realized that alternative splicing can modulate protein-protein and other ligand interactions, cause the molecular interaction associated with certain key links of pathway being lost or increased [6]. As a consequence, proteoforms translated from different alternatively spliced isoforms of the same gene may interact with different partners and have distinct or even opposing functions [7]. Therefore, systematically identifying isoform-isoform interactions (IIIs) is crucial to explore the function of proteoforms and dissect the mechanism of molecular interactions.

Compared with the heavy study on predicting protein-protein interaction or gene-gene interaction [2], [8], [9], computationally predicting isoform-isoform interactions is much less studied. An isoform-isoform interaction database (IIIDB) [10] was developed to study PPIs at the isoform resolution, which uses a logistic regression model and integrates RNA-seq, domain-domain interaction data to predict IIIs. IIIDB does not account for multi-exon genes, each of which often produces more than one isoform. Li *et al.* proposed a Bayesian network-based multi-instance learning (SIB-MIL) algorithm to predict isoform-level interaction through integrating genomic and proteomic data, and extends the prediction of IIIs from single-isoform genes to multi-isoform genes. However, existing solutions on predicting IIIs in large scale still faces the following challenges:
i) Lack a large set of experimentally validated interacting

This work is supported by Natural Science Foundation of China (61872300, 61873214 and 61871020), Fundamental Research Funds for the Central Universities (XDJK2019B024), Natural Science Foundation of CQ CSTC (cstc2018jcyjAX0228). Corresponding Author: Guoxian Yu (gxyu@swu.edu.cn)

isoform pairs as a gold standard. At present, interacting isoform pairs validated by traditional biological methods are relatively rare. There are $< 5\%$ of the human genome being successfully tested for IIIs [11], and detecting IIIs by wet-lab experiments is very time consuming and labor intensive. That is the main bottleneck for predicting IIIs.

ii) Lack genome-wide research at the isoform level. Tseng *et al.* [10] only considered the case where a gene produces a single isoform. However, genes with only one isoform are very few. As we known, most genes in mammal species can produce a variety of different isoforms by alternative splicing.

It has been validated on the Corominas data [3] that for an interacting gene pair, there must be at least one isoform pair is interacting [5]. Given that, we can formulate the task as multi-instance learning (MIL) [12] to explore the genome-wide isoform-isoform interactions using the abundance gene-level interactions in the public databases [13]. Particularly, a gene pair is considered as a 'bag', two isoforms spliced from respective genes in the bag form a pair and denote as an 'instance'. As illustrated in Figure 1, for a positive gene pair bag, there must be at least one interacting isoform pair; otherwise, none of its isoform pairs has an interaction. Diverse biological data can capture the information of isoforms and genes from different aspects, given the multiplicity of these data and inspired by the success of deep learning in complex data mining [14], [15], we proposed a deep multi-instance leaning approach (called DMIL-III) to predict isoform-isoform interactions. DMIL-III fuses RNA-seq, nucleotide sequence, domain-domain interaction and exon array. Experiments show that DMIL-III achieves an accuracy of 93% on single-isoform gene bags and 94% on multi-isoform gene bags at the gene-level, which is at least 14% and 29% higher than those of the start-of-the-art methods [5], [10]. In addition, we apply it on a set of experimentally confirmed IIIs and obtain an Accuracy of 65%, which is at least 10% higher than other methods.

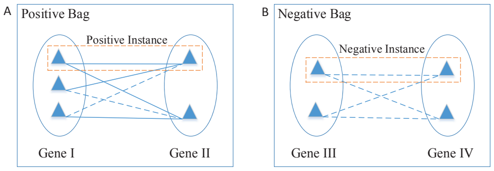

Fig. 1. A. An interacting gene pair is called positive bag, gene I with 3 isoforms and gene II with 2 isofroms. There are four isoform pairs with interactions (solid blue line) and are called positive instances (or witnesses), and two isoform pairs have no interactions (dashed blue line) and called as negative instances. B. A negative gene bag with no interaction between genes, each gene (gene III and IV) has two isoforms, and all isoform pairs have no interaction (dashed blue line), they are called negative instances.

This paper is organized as follows. In Section II, we introduce the preprocess of diverse data and deep multi-instance learning framework for predicting IIIs. Section III presents and discusses the results of DMIL-III and different comparing methods. Section IV gives the conclusion and future work.

II. Materials and Method

A. Gold Standard Construction

Due to the lack of ground-truth interacting isoform pairs to serve as the gold standard, our golden set is constructed at the gene level. We obtained PPI dataset from Pan's dataset [13], which contains 36, 630 positive pairs and 36, 480 negative pairs. The positive PPI dataset was downloaded from human protein reference database (HPRD) by removing duplicated interactions. The golden standard negative dataset was generated by pairing proteins with different cellular compartments. We first matched the RefSeq IDs in Pan's dataset with IDs of Ensembl database [16], genes whose RefSeq IDs can not find in Ensembl are removed. Each gene has at least one isoform. Gene pairs in the golden standard set are regarded as a bag, isoform pairs which are randomly paired by isoforms of two genes are regarded as the instances of the gene bag. Finally, we obtained 31,343 positive pairs and 25,775 negative pairs in our golden set with 18,809 isoforms of 7,651 genes, which includes 4,824 positive single-instance bags, 4,280 negative single-instance bags, 26,519 positive multi-instance bags and 20,955 negative multi-instance bags.

B. Feature Extraction

In this section, we describe the integrated data in our paper, our model takes a 233-dimensional isoform-level feature as input, which includes RNA-seq (43-dimensional), nucleotide sequence (128-dimensional), DDI (1-dimensional) and exon array (61-dimensional). Details for these data are shown below.

1) RNA-seq Datasets: High-throughput mRNA sequencing (RNA-seq) technology is a widely-used technique for transcript quantification of gene isoforms. RNA-Seq has the advantage of providing unprecedented amounts of transcript-level expression data in a deep level. We downloaded 51 tissue-specific data with 298 RNA-seq datasets of human from the ENCODE project [17]. Firstly, we used HISAT2 [18] to align the short-reads of each RNA-seq dataset against the human genome build GRCh38.90 from Ensembl database [16]. Then, StringTie [19] was employed to calculate the relative abundance of the transcript as FPKM (Fragments Per Kiolobase of exon per Million fragments mapped). Tissue-specific data which has at least 4 RNA-seq datasets is chosen to measure the correlation, thus we get a total of 43 tissue-specific data with 282 RNA-seq datasets. We calculated the FPKM value for all isoforms in these dataset. Finally, we calculated the Pearson correlation coefficients between all possible isoform pairs for each tissue-specific data. As a result, a 43-dimensional feature vector for each isoform pair was created.

2) Nucleotide Sequences: Amino acid sequences has been proved to be sufficient enough to detect protein interactions [8]. Therefore, it is necessary to use sequences information for predicting isoform-isoform interactions. Here, nucleotide sequences of isoforms were used to predict IIIs. Conjoint triad (CT) [20] was employed to extract the numeric feature of nucleotide sequence, which considers three continuous bases as a unit and calculates the frequency of each triad type. The nucleotide sequences of isoform were downloaded from NCBI Nucleotide databse. For nucleotide sequence composed by Adenine (A), Guanine (G), Cytosine (C), Thymine (T), three continuous bases were considered as a unit, thus $4\times4\times4$ frequencies were calculated. After computing the number of each triad type and normalizing it, a 64-dimensional feature vector was generated to represent the sequence information of each isoform. As for isoform pair, it was represented by a 128-dimensional feature vector by concatenating features of two isoforms.

3) Domain-Domain Interaction Data: Domain plays a vital role in the molecular interaction process. It is domain-domain interaction that results in protein-protein interaction. Therefore, it is essential to study domain-domain interactions for predicting isoform-isoform interactions. The domain-domain interaction (DDI) data was downloaded from DIMA (Domain Interaction MAp) database [21], where each domain pair is assigned with a score between 1 and 5 as the interaction score. Due to lack of domain annotation on isoforms, isoform from single-instance bag is assumed to have all domains of its gene. As for isoforms from multi-instance gene bag, domain information for each isoform was obtained from [11]. By this way, we got domain annotation on all isoforms in our dataset. Then, domains of an isoform pair were respectively randomly paired as domain pairs. If a domain pair has domain-domain interaction score in the DIMA database, we take the score as the feature of this domain pair; otherwise, this domain pair is assigned a score of 0. DDI feature of an isoform pair was quantified by its domain pair score. For an isoform pair with several DDI scores, the highest score was taken. Finally, we took the score as the feature of isoform pairs.

4) Exon Array: The Gene Chip Exon Array provides a great deal of expression data for all transcripts of a gene, which is necessary for us to utilize exon array data for predicting IIIs. Firstly, we downloaded 61 exon array datasets of human from the NCBI GEO (Gene Expression Omnibus) database. Then, the R package MEAP (version 2.0.1) was employed to calculate the expression of transcripts. All isoforms in our dataset were calculated the expression value for each exon array dataset. Next, we calculated the Pearson correlation between isoform pairs for each dataset. Finally, a 61-dimensional vector was used to encode the features of each isoform pair.

C. Deep Multi-Instance Learning

In this subsection, we present DMIL-III for learning deep representation of isoforms. Based on the multi-instance learning methodology, we unify the deep learning with MIL framework to predict IIIs in a weakly-supervised manner.

1) Problem Formulation: Different from classical supervised learning methods, in multi-instance learning, data are organized as bags $\boldsymbol{X}_i \in \mathbb{R}^{n_i \times d}$, each of which contains a number of instances $\boldsymbol{x}_{ij} \in \mathbb{R}^d$ [12]. Only the label of bag Y_i is available, labels of instances are unknown, that is in accordance with our benchmark dataset. Considering a gene pair G_1 and G_2 as a 'bag', isoform pairs generated by Cartesian product from isoforms of G_1 and from G_2 are called as 'instance'. Each bag has at least one instance. In general, a gene bag of n isoform pairs is denoted by $\boldsymbol{X}_i = \{\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, ..., \boldsymbol{x}_{in_i}\}$ and $\boldsymbol{x}_{ij} \in \mathbb{R}^d$ is a feature vector corresponding to the *j*-th isoform pair of the *i*-th gene pair. The class label of the *i*-th gene pair is denoted as Y_i, which is associated with the entire gene bag. Based on the typical principle of MIL: if a gene pair with gene-gene interaction, then at least one isoform pair is interacting; if a gene pair with no interaction, then none of its isoform pair is interacting. The aim of our model is to identify the true (or responsible) interacting isoform pairs of the interacting gene pairs (positive bags).

2) Deep Multi-instance Neural Network: Given the multiplicity of isoform-isoform interactions and the success of deep learning for protein-protein interaction prediction [2], we proposed a deep neural network (shown in Figure 2) to predict IIIs.

Bag construction: Inspired by natural language processing [22], we take the instances of a gene bag as words, a gene pair as a document. Each instance of a gene bag maps a k-dimensional vector $\boldsymbol{x_{ij}}$, a gene bag with n_i instances can be presented by a matrix

$$\boldsymbol{X_i} = [\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \ldots, \boldsymbol{x}_{in}] \tag{1}$$

where $\boldsymbol{X}_i \in \mathbb{R}^{n \times k}$. Since the number of isoform pairs in each gene bag is not equal, we fix the maximum number of isoform pairs in each gene bag to n. Gene bags whose number of isoform pairs larger than n are excluded. If a gene bag has the isoform pairs fewer than n, this bag is padded with $n - n_i$ zero vectors.

Convolution layer: The DMIL-III model takes matrices of gene bags $\boldsymbol{X}_i$ as input. Each convolution layer performs convolution operation with a kernel $\boldsymbol{w} \in \mathbb{R}^{1 \times h}$ and stride $s = 1$ on $\boldsymbol{x}_{ij}$ to extract a high-level abstraction for each instance of a gene bag. The convolution operation is defined as:

$$\boldsymbol{c}_{ijg} = f(\boldsymbol{w} * \boldsymbol{x}_{ijg:g+h}), \ g \in (1, k-h) \tag{2}$$

where $*$ is a convolution operation, f is a non-linear function. The new feature vector of $\boldsymbol{c}_{ij}$ is defined as:

$$\boldsymbol{c}_{ij} = [\boldsymbol{c}_{ij1}, \boldsymbol{c}_{ij2}, \ldots, \boldsymbol{c}_{ijp}] \tag{3}$$

where $p = k - h + 1$.

After the convolution operation on all instances of a gene bag, we can obtain a new feature map of gene bag

$$\boldsymbol{C}_i = [\boldsymbol{c}_{i1}; \boldsymbol{c}_{i2}, \ldots, \boldsymbol{c}_{in}] \tag{4}$$

with $\boldsymbol{C}_i \in \mathbb{R}^{n \times p}$. In this way, one feature is extracted from one filter. To get multiple features, the model uses multiple different filters with the same size.

Our model is consisted with four convolution layers, the first three layers adopt the rectified linear unit (ReLU) as activation function with a recommended kernel size of 1×7, numbers of filters respectively fixed to 32, 32, 16 to extract features of each instance in a gene bag. The last layer uses the sigmoid activation function with one filter, which quantifies the probability that an isoform pair is interacting or not.

Batch Normalization layer: In our model, each of the first two convolution layers is followed by a batch normalization layer to normalize the feature map generated by convolution layer to obey the normal distribution, which can not only speed up the convergence, but also avoid the disappeared gradient.

Pooling layer: A max pooling layer follows by the batch normalization layer. Similarity to convolution operation, the

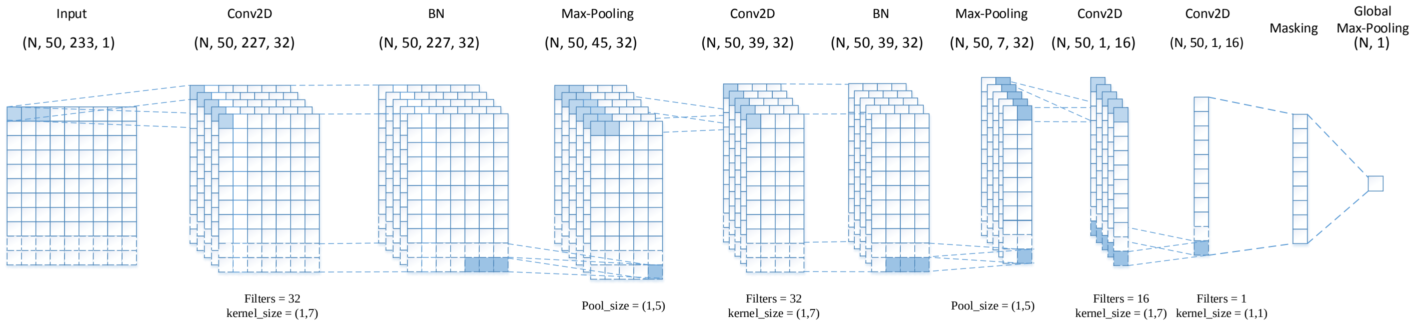

Fig. 2. The network architecture of DMIL-III. Our model takes a $N \times 50 \times 233 \times 1$ matrix as input, where N refers the number of bags in a batch. Isoform pairs of each bag are fixed as 50, which is represented by a 233-dimensional feature vector.

pooling operation is performed with a sliding window $1 \times h^{'}$, $h^{'}$ is the width of the pooling windows, which is set to 5 as the recommended parameter. The max pooling is used to remain the most significant feature of a sliding window as follows:

$$f_{ijg} = max(\boldsymbol{c}_{ijg}, \ldots, \boldsymbol{c}_{ijg+h^{'}}),\ g \in (1, k - h^{'}) \qquad (5)$$

In addition, a global max pooling layer is used as the MIL layer in the last layer of our model:

$$P_i = F(p_{i1}, p_{i2}, ..., p_{in}) \qquad (6)$$

where p_{ij} corresponds to the output of the last convolution layer for each instances $\boldsymbol{x}_{ij}$, P_i is the probability that a gene bag $\boldsymbol{X}_i$ is positive, F is the aggregation function. Typically, both the mean pooling and max pooling function can be utilized in MIL, while the latter is mostly used [23], [24]. Our model adopts the maximum pooling. In this way, the instance of positive bag, which is the most likely to be positive, is labeled as 'positive'. Meanwhile, instances of negative bags, which have the maximum probabilities, are labeled as negative samples for training. Therefore, we can get an instance-level prediction.

It is important to exclude the prediction on the padded zero vectors before pooling operation. Otherwise, it may get the maximal probability instance in a negative gene bag. Therefore, a masking layer is added before the global max pooling layer to exclude the probability generated by padded instances as shown in Figure 3. Through the last Conv2D layer, we can compute the probability score for all isoform pairs of a gene bag. Then, a masking layer is used to mask the probability score of padded instances. However, the pooling layer can not support the mask operation in Keras [25], so we modify the global max pooling layer of Keras by setting "supports_masking =True". In this way, we obtain the probability scores of all real instances filtered by masking layer. Finally, we get the maximum probability of all real instances in a bag through the global max pooling operation.

The loss in our model is defined by binary cross-entropy.

To minimize the loss, the stochastic gradient descent was employed for optimization through back propagation.

III. Results and Analysis

We compare DMIL-III against IIIDB [10], SIB-MIL [5], SVM [26] and mi-SVM [27]. All the input parameters are kept the same as the author reported or optimized in the suggested ranges. The best hyper-parameter configuration of deep learning model is dependent by its data and application. The increase in hyper-parameter results in an exponential growth of configuration, which makes it impossible to try all of configurations in practice. It is recommended to optimize important hyper-parameters. In our paper, these hyper-parameters are set as follows: learning_rate=0.00005, epoch=1000, L1_regularization=0.001. To verify that the integrated data and proposed method, input of the comparing methods [5], [10] adopts the integrated feature in its paper, SVM and mi-svm take the same input as DMIL-III.

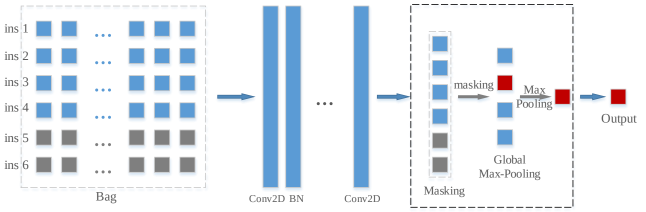

Fig. 3. The details of masking the global max pooling. The blue squares are the features of real instances in the bags, and gray squares are padded instances. The blue circles represent the probability scores of real instances, and the gray circles refer to the probability scores of padded instances.

A. Results of Predicting IIIs on Single-isoform Gene Pairs

There are a small proportion of genes produce only single-isoforms in the benchmark dataset. Following the evaluation principle in [10], we first use single-instance gene bags to evaluate the performance of our model through 5-fold cross validation and report the results of DMIL-III and other comparing methods in Table I. From Table I, we can see that the results of DMIL-III on single-isoform gene pairs are highly accurate with an Accuracy of 93%. Compared with the Accuracy of IIIDB and SIB-MIL, DMIL-III is 33% and 35% higher than IIIDB and SIB-MIL, respectively. This comparison suggests that DMIL-IIIs can utilize deep multi-instance neural networks to better fuse diverse and essential biological data than IIIDB and SIB-MIL, and it can more credibly find the triggering isoform-isoform interactions of the positive bag (protein-protein interaction).

To further validate our analysis, we apply IIIDB and SIB-MIL on the integrated data used by DMIL-III and also report their results (IIIDB2 and SIB-MIL2) in Table I. We find that the performance of IIIDB2 is better, with accuracy as 72%, AUC as 80% and AUPRC as 83%, which are 12%, 15% and

TABLE I
RESULTS OF FIVE-FOLD CROSS VALIDATION ON PREDICTING IIIS OF SINGLE-ISOFORM GENE BAGS.

	Accuracy	Precision	Recall	MCC	F_1	AUC	AUPRC
DMIL-III	**0.93±0.01**	**0.96±0.01**	**0.91±0.01**	**0.87±0.02**	**0.94±0.01**	**0.97±0.01**	**0.98±0.00**
IIIDB	0.60±0.01	0.64±0.01	0.59±0.01	0.21±0.01	0.61±0.01	0.65±0.00	0.71±0.01
SIB-MIL	0.58±0.01	0.63±0.02	0.50±0.00	0.18±0.03	0.56±0.01	0.63±0.01	0.66±0.01
SVM	0.79±0.01	0.80±0.01	0.80±0.01	0.57±0.02	0.80±0.01	0.87±0.01	0.88±0.00
IIIDB2	0.72±0.01	0.74±0.01	0.74±0.01	0.44±0.01	0.74±0.01	0.80±0.01	0.83±0.01
SIB-MIL2	0.60±0.01	0.82±0.02	0.31±0.01	0.29±0.01	0.45±0.01	0.71±0.01	0.76±0.01

12% higher than those of IIIDB with RNA-seq and DDI data only. SIB-MIL2 also manifests an increased performance than SIB-MIL, which are 2%, 8% and 10% higher in terms of Accuracy, AUC and AUPRC. This study suggests that our integrated data contributes to a better performance in predicting IIIs, which are 21%, 17% and 15% higher than those of IIIDB2, and 33%, 26% and 22% higher than those of SIB-MIL2 on metrics of Accuracy, AUC and AUPRC. Therefore, we can safely say that our DMIL-III has a better performance on extracting more useful information for predicting IIIs.

In addition, we compare the performance of SVM on our integrated data. From Table I, we can find that the Accuracy of DMIL-III is 14% higher than that of SVM. The performance margin between DMIL-III and SVM confirms the necessity of adapting deep multi-instance learning to model the complexity of isoform-isoform interaction and to achieve a better performance.

B. Results of Predicting IIIs on Multi-isoform Gene Pairs

There is a large proportion of genes generating more than one isoform. Therefore, we further apply DMIL-III to predict IIIs of multi-isoform gene pairs. Through analyzing the distribution of the number of isoform pairs of gene bags, we find that the number of isoform pairs in most gene bags is between 1 and 50, gene bags with two isoform pairs have the largest portion. Only a few gene bags have more than 50 isoform pairs. Therefore, we set the *maximal* number of isoform pairs in a gene bag to $n = 50$. Gene pairs (or bags) with more than n isoform pairs in our golden standard sets are discarded. After this process, we get a total of 26, 344 positive gene bags and 20, 910 negative gene bags, corresponding to 177,456 positive and 130,138 negative isoform pairs, respectively.

Since the isoform-level information is not available, we train and evaluate our DMIL-III using the gene level interaction. Each gene pair is assigned with a score as the maximum probability of all its isoform pairs under the hypotheses that there is at least one isoform pair interacting if its originating gene pair has an interaction. Gene-level prediction results are provided in Table II. In addition, two multi-instance learning based methods, SIB-MIL and mi-SVM [27] are taken as the comparing methods. IIIDB cannot apply on multi-isoform genes, so its results are not reported.

From Table II, we can see that DMIL-III achieves the Accuracy of 94%, AUC of 98%, and AUPRC of 99%, which are 40%, 40%, 36% higher than those of SIB-MIL, and 30%, 25% and 27% higher than those of mi-SVM, although they all follow the principle of multi-instance learning. The performance superiority is because our integrated data provides more important information, and because DMIL-III adapts the convolution operation and pooling layer, the former can extract more useful information for isoform-isoform interactions, and the later extracts the most significant features by remaining the maximum value of a pooling window.

C. Results of Predicting Experimentally Validated IIIs

To evaluate the isoform-level prediction performance of DMIL-III, a set of experimentally verified isoform-isoform interaction data were collected. Firstly, we downloaded four isoform interacotme datasets [4], [28]–[30] from HuRI project (http://interactome.baderlab.org/). After removing redundant isoform pairs, we get a total of 7,605 interacting isoform pairs corresponding to 4,891 isoforms and 2,640 genes. For each interacting isoform pair, genes of each isoform pair are paired as a positive gene bag, and the remaining isoforms respectively from the gene pair are randomly paired as instances of this positive gene bag. Negative gene bags are constructed by the subcellular localizations based on the setting that genes with different subcellular localizations are not interacting. By this way, we get a number of 519, 071 non-interacting gene pairs. Then, we randomly select non-interacting gene pairs with the same number of positive bags as negative samples. Similar to positive instances, two isoforms from two different genes of a negative pair are paired as a negative instance. Finally, we get a dataset with 7, 605 positive gene bags and 29, 014 instances, and 7, 605 negative gene bags and 20, 503 negative instances. In the positive gene bags, isoform pairs existed in the raw dataset are labeled as positive. There are still some isoform pairs can not find in the raw dataset, since isoform pairs of a gene bag are not all identified through experiments or some isoform pairs are non-interacting, which result in label missing in these isoform pairs. In our experiment, we consider the label-missing isoform pairs as negative. Experiment is carried out on data with isoform-level interactions through 5-fold cross validation and results are revealed in Table III. We can see that our model achieves a good performance with Accuracy as 65%, AUC as 73% and AUPRC as 32%, which are 20%, 15% and 7% higher than those of SIB-MIL, and 10%, 5% and 1% higher than those of mi-svm. From Table III, we can also find that the recall of mi-svm is higher than DMIL-III. It is because that features of different isoforms from the same gene usually have strong similarities, which make it difficult to differentiate isoform pairs from the same gene bags. Thus, mi-svm tends to classify negative samples in positive gene bags as positive samples, which results in a higher recall. Compared with other methods, DMIL-III can extract more critical features through convolution and max pooling operation. As a result, it can more credibly distinguish different isoform pairs from the same gene bag and thus obtain a higher performance.

TABLE II

Results of five-fold cross validation on IIIs of multi-instance gene bags.

	Accuracy	Precision	Recall	MCC	F_1	AUC	AUPRC
DMIL-III	**0.94±0.01**	**0.96±0.03**	**0.93±0.02**	**0.87±0.02**	**0.94±0.01**	**0.98±0.00**	**0.99±0.00**
SIB-MIL	0.54±0.00	0.63±0.00	0.41±0.01	0.11±0.01	0.50±0.00	0.58±0.00	0.63±0.00
mi-SVM	0.64±0.01	0.61±0.02	0.83±0.03	0.31±0.03	0.70±0.02	0.73±0.02	0.72±0.02

TABLE III

Results of predicting IIIs on experimentally validated IIIs.

Method	Accuracy	Precision	Recall	MCC	F_1	AUC	AUPRC
SIB-MIL	0.45±0.00	0.21±0.00	0.68±0.00	0.06±0.00	0.32±0.00	0.58±0.01	0.25±0.00
mi-SVM	0.55±0.00	0.27±0.00	**0.80±0.01**	0.23±0.01	0.40±0.00	0.68±0.01	0.31±0.01
DMIL-III	**0.65±0.00**	**0.28±0.00**	0.74±0.00	**0.28±0.00**	0.40±0.00	**0.73±0.00**	**0.32±0.00**

IV. Conclusions

Isoform-isoform interaction prediction can provide a deeper granularity of interactions between proteins, but compared with the traditional and widely-studied gene-level protein interaction prediction, it is less studied for the lack of isoform-level interactions. In this study, we proposed a deep multi-instance learning solution called DMIL-III to predict isoform-level interactions by integrating RNA-seq, nucleotide sequence, domain-domain interaction and exon array data. Extensive results confirm that DMIL-III outperforms other related solutions with remarkable performance on the benchmark dataset.

References

[1] L. Zhang, G. Yu, M. Guo, and J. Wang, "Predicting protein-protein interactions using high-quality non-interacting pairs," *BMC Bioinformatics*, vol. 19, no. S19, p. 525, 2018.

[2] L. Zhang, G. Yu, D. Xia, and J. Wang, "Protein–protein interactions prediction based on ensemble deep neural networks," *Neurocomputing*, vol. 324, pp. 10–19, 2019.

[3] R. Corominas, X. Yang, G. N. Lin, S. Kang, Y. Shen, L. Ghamsari, M. Broly, M. Rodriguez, S. Tam, S. A. Trigg *et al.*, "Protein interaction network of alternatively spliced isoforms from brain links genetic risk factors for autism," *Nature Communications*, vol. 5, p. 3650, 2014.

[4] X. Yang, J. Coulombe-Huntington, Kang *et al.*, "Widespread expansion of protein interaction capabilities by alternative splicing," *Cell*, vol. 164, no. 4, pp. 805–817, 2016.

[5] H.-D. Li, R. Menon, R. Eksi, A. Guerler, Y. Zhang, G. S. Omenn, and Y. Guan, "A network of splice isoforms for the mouse," *Scientific Reports*, vol. 6, p. 24507, 2016.

[6] J. D. Ellis, M. Barrios-Rodiles, R. Çolak, M. Irimia, T. Kim, J. A. Calarco, X. Wang, Q. Pan, D. O'Hanlon, P. M. Kim *et al.*, "Tissue-specific alternative splicing remodels protein-protein interaction networks," *Molecular Cell*, vol. 46, no. 6, pp. 884–892, 2012.

[7] L. M. Smith, N. L. Kelleher, M. Linial, D. Goodlett, P. Langridge-Smith, Y. A. Goo, G. Safford, L. Bonilla, G. Kruppa, R. Zubarev *et al.*, "Proteoform: a single term describing protein complexity," *Nature Methods*, vol. 10, no. 3, p. 186, 2013.

[8] S. Hashemifar, B. Neyshabur, A. A. Khan, and J. Xu, "Predicting protein–protein interactions through sequence-based deep learning," *Bioinformatics*, vol. 34, no. 17, pp. i802–i810, 2018.

[9] H. Li, X.-J. Gong, H. Yu, and C. Zhou, "Deep neural network based predictions of protein interactions using primary sequences," *Molecules*, vol. 23, no. 8, p. 1923, 2018.

[10] Y.-T. Tseng, W. Li, C.-H. Chen, S. Zhang, J. J. Chen, X. J. Zhou, and C.-C. Liu, "Iiidb: a database for isoform-isoform interactions and isoform network modules," in *BMC Genomics*, vol. 16, no. 2, 2015, p. S10.

[11] M. A. Ghadie, L. Lambourne, M. Vidal, and Y. Xia, "Domain-based prediction of the human isoform interactome provides insights into the functional impact of alternative splicing," *PLoS Computational Biology*, vol. 13, no. 8, p. e1005717, 2017.

[12] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial Intelligence*, vol. 89, no. 1-2, pp. 31–71, 1997.

[13] X.-Y. Pan, Y.-N. Zhang, and H.-B. Shen, "Large-scale prediction of human protein- protein interactions from amino acid sequence based on latent topic features," *Journal of Proteome Research*, vol. 9, no. 10, pp. 4992–5001, 2010.

[14] K. Lan, D. T. Wang, S. Fong, L. S. Liu, and N. Dey, "A survey of data mining and deep learning in bioinformatics," *Journal of Medical Systems*, vol. 42, no. 8, p. 139, 2018.

[15] S. Min, B. Lee, and S. Yoon, "Deep learning in bioinformatics," *Briefings in Bioinformatics*, vol. 18, no. 5, p. 851, 2017.

[16] T. Hubbard, D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down *et al.*, "The ensembl genome database project," *Nucleic Acids Research*, vol. 30, no. 1, pp. 38–41, 2002.

[17] E. P. Consortium *et al.*, "The encode (encyclopedia of dna elements) project," *Science*, vol. 306, no. 5696, pp. 636–640, 2004.

[18] D. Kim, B. Langmead, and S. L. Salzberg, "Hisat: a fast spliced aligner with low memory requirements," *Nature Methods*, vol. 12, no. 4, p. 357, 2015.

[19] M. Pertea, D. Kim, G. M. Pertea, J. T. Leek, and S. L. Salzberg, "Transcript-level expression analysis of rna-seq experiments with hisat, stringtie and ballgown," *Nature Protocols*, vol. 11, no. 9, p. 1650, 2016.

[20] J. Shen, J. Zhang, X. Luo, W. Zhu, K. Yu, K. Chen, Y. Li, and H. Jiang, "Predicting protein–protein interactions based only on sequences information," *Proceedings of the National Academy of Sciences*, vol. 104, no. 11, pp. 4337–4341, 2007.

[21] Q. Luo, P. Pagel, B. Vilne, and D. Frishman, "Dima 3.0: domain interaction map," *Nucleic Acids Research*, vol. 39, no. S1, pp. D724–D729, 2010.

[22] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1746–1751.

[23] W. Zhu, Q. Lou, Y. S. Vang, and X. Xie, "Deep multi-instance networks with sparse label assignment for whole mammogram classification," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2017, pp. 603–611.

[24] P. O. Pinheiro and R. Collobert, "From image-level to pixel-level labeling with convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1713–1721.

[25] F. Chollet *et al.*, "Keras," https://keras.io, 2015.

[26] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, p. 27, 2011.

[27] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Advances in Neural Information Processing Systems*, 2003, pp. 577–584.

[28] J.-F. Rual, K. Venkatesan, T. Hao *et al.*, "Towards a proteome-scale map of the human protein–protein interaction network," *Nature*, vol. 437, no. 7062, p. 1173, 2005.

[29] H. Yu, L. Tardivo, S. Tam *et al.*, "Next-generation sequencing to generate interactome datasets," *Nature Methods*, vol. 8, no. 6, p. 478, 2011.

[30] T. Rolland, M. Taşan, B. Charloteaux *et al.*, "A proteome-scale map of the human interactome network," *Cell*, vol. 159, no. 5, pp. 1212–1226, 2014.