Isoform-disease association prediction by data fusion

Qiuyue Huang[1], Jun Wang[1], Xiangliang Zhang[2] and Guoxian Yu[1,*]

[1] College of Comp. and Inf. Sci., Southwest Univ., Chongqing 400715, China
[2] CEMSE, King Abudullah Univ. of Sci. & Tech., Thuwal, Saudi Arabia
*Corresponding author: gxyu@swu.edu.cn (Guoxian Yu)

Abstract. Alternative splicing enables a gene spliced into different isoforms, which are closely related with diverse developmental abnormalities. Identifying the isoform-disease associations helps to uncover the underlying pathology of various complex diseases, and to develop precise treatments and drugs for these diseases. Although many approaches have been proposed for predicting gene-disease associations and isoform functions, few efforts have been made toward predicting isoform-disease associations in large-scale, the main bottleneck is the lack of ground-truth isoform-disease associations. To bridge this gap, we propose a multi-instance learning inspired computational approach called IDAPred to fuse genomics and transcriptomics data for isoform-disease association prediction. Given the bag-instance relationship between gene and its spliced isoforms, IDAPred introduces a dispatch and aggregation term to dispatch gene-disease associations to individual isoforms, and reversely aggregate these dispatched associations to affiliated genes. Next, it fuses different genomics and transcriptomics data to replenish gene-disease associations and to induce a linear classifier for predicting isoform-disease associations in a coherent way. In addition, to alleviate the bias toward observed gene-disease associations, it adds a regularization term to differentiate the currently observed associations from the unobserved (potential) ones. Experimental results show that IDAPred significantly outperforms the related state-of-the-art methods.

Keywords: Isoform-disease association, Alternative splicing, Data fusion, Multi-instance learning

1 Introduction

Deciphering human diseases and the pathology is one of key fundamental tasks in life science [4]. Thousands of genes have been identified as associated with a variety of diseases. Identifying gene-disease associations (GDA) contributes to decipher the pathology, which helps us to find new strategies and drugs to treat diverse complex diseases. Many computational solutions have been developed to predict GDAs in large-scale, such as network propagation [32, 35], literature mining [23], clustering analysis [31], data fusion [23], matrix completion [18], deep learning-based methods [16] and so on.

A single gene can produce multiple isoforms by alternative splicing, which greatly increases the transcriptome and proteome complexity [29]. More than 95% multi-exon genes in human genome undergo alternative splicing [20, 33]. In practice, a gene can be associated with diverse diseases mainly owing to its abnormally spliced isoforms [29]. Increasing studies confirm that alternative splicing is associated with diverse complex diseases, such as autism spectrum disorders [28], ischemic human heart disease [19], and Alzheimer disease [10]. Neagoe *et al.* [19] observed that a titin isoform switch in chronically ischemic human hearts with 47:53 average N2BA-to-N2B ratio in severely diseased coronary artery disease transplanted hearts, and 32:68 in nonischemic transplants. Long-term titin modifications can damage the ability of the heart. Apolipoprotein E (apoE) is localized in the senile plaques, congophilic angiopathy, and neurofibrillary tangles of Alzheimer disease. Strittmatter *et al.* [30] compared the difference of binding of synthetic amyloid beta (beta/A4) peptide to apoE4 and apoE3, which are two commom isoforms of apoE, and observed that apoE4 is associated with the increased susceptibility to disease. The results show that the pathogenesis of Alzheimer disease may be related to different bindings in apoE. Sanan *et al.* [24] observed the apoE4 isoform binds to a beta peptide more rapidly than apoE3. Holtzman *et al.* [10] found the expression of apoE3 and apoE4 in APPV717F transgenic (TG), no apoE mice resulted in fibrillar amyloid-β deposits and neuritic plaques by 15 months of age and substantially (>10-fold) more fibrillar deposits were observed in apoE4-expressing APPV717F TG mice. Lundberg *et al.* [15] demonstrated that FOXP3 in CD4+ T cells is associated with coronary artery disease and alternative splicing of FOXP3 is decreased in coronary artery disease.

Existing isoform-disease associations (IDAs) are mainly detected by wet-lab experiments (*i.e.*, gel electrophoresis and immunoblotting). To the best of authors knowledge, there is *no computational solution* for predicting IDAs at a large-scale. The main bottleneck is that there is *no public database* that stores sufficient IDAs, which are required for typical machine learning methods to induce a reliable classifier for predicting IDAs. In fact, such lack also exists in functional analysis of isoforms [13]. To bypass this issue, some researchers take a gene as a bag and its spliced isoforms as instances of that bag, and adapt multiple instance learning (MIL) [2, 17] to distribute the readily available functional annotations of a gene to its isoforms [3, 6, 14, 26, 34, 40].

Based on the accumulated GDAs in public databases (*i.e.*, DisGeNET [22], OMIM (`www.omim.org`)) and inspired by the MIL-based isoform function prediction solutions, we kickoff a *novel* task of predicting IDAs, which is *more challenging* than traditional GDAs prediction, due to the lack of IDAs and the complex relationship between isoforms and genes. This task can provide a deeper understanding of the pathology of complex diseases. To combat this task, we introduce a computational solution (IDAPred) to predict IDAs in large scale by fusing genomic and transcriptome data and by distributing gene-disease associations to individual isoforms. IDAPred firstly introduces a dispatch and aggregation term to dispatch GDAs to individual isoforms and reversely aggregate these dispatched IDAs to affiliated genes based on the gene-isoform relations. To remedy incomplete GDAs, it fuses nucleic acid sequences and interactome of genes to further fulfil the to-be-dispatched GDAs. As well as that, it leverages multiple RNA-seq datasets to construct tissue-specific isoform co-expression networks and to induce a linear classifier to predict IDAs. In addition, it introduces an indicator matrix to differentiate the observed GDAs from the further fulfilled ones and thus to alleviate the bias toward observed ones. Finally, IDAPred merges these objectives into a unified objective function and predicts IDAs in a coherent way. Experimental results show that IDAPred achieves better results across various evaluation metrics than other competitive approaches that are introduced for predicting GDAs [32] or isoform functions [14, 34, 40].

2 Method

2.1 Materials and Pre-processing

Suppose there are n genes, the i-th gene produces $n_i \geq 1$ isoforms, and the total number of isoforms is $m = \sum_{i=1}^{n} n_i$. We adopt the widely-used Fragments Per Kilobase of exon per Million fragments mapped fragments (FPKM) values to quantify the expression of isoforms. Particularly, we downloaded 596 RNA-seq runs (of total 298 samples from different tissues and conditions) of Human from the ENCODE project [5] (access date: 2019-11-10). These datasets are heterogeneous in terms of library preparation procedures and sequencing platform. Following the pre-process done in [14, 34], for each tissue, we control the quality of these RNA-seq datasets and quantify the expression value of isoforms as follows:

(i) We firstly align the short-reads of each RNA-seq dataset of the Human genome (build GRCh38.90) from Ensemble using HISAT2(v.2-2.1.0) [12], and A GTF annotation file of the same build with an option of no-novel-junction.

(ii) Then, we use Stringtie(v.1.3.3b) [21] to calculate the relative abundance of the transcript as Fragments Per Kilobase of exon per Million fragments mapped fragments (FPKM). We separately compute the FPKM values of a total of 57,964 genes with 219,288 isoforms for each sample.

(iii) The FPKM values of very short isoforms are exceptionally higher. Therefore, we discard the isoforms with less than 100 nucleotides.

(iv) To further control the quality of isoforms, we use known protein coding gene names to map those genes obtained in step (iii). Due to the prohibitive runtime on such a large number of isoforms and sufficient nonzero values in the expression vector are required to induce a predictor, we refilter the data. Particularly, we set all FPKM values lower than 0.3 as 0, and then remove isoform with all FPKM values of 0. To ensure data filtered at the gene level, we do a further filtering: if an isoform of a gene is filtered, this gene and its all spliced isoforms are removed also. Finally, we obtain 7,549 genes with 39,559 isoforms, whose values are stored in the corresponding data matrix $\mathbf{X} \in \mathbb{R}^{m \times d}$. We further normalize $\mathbf{X}$ by $\mathbf{X}_{nor} = \mathbf{X}./max(\mathbf{X})$. We use the normalized $\mathbf{X}$ for subsequent experiments.

We downloaded the gene-disease associations file and the mappings file UMLS CUI to Disease Ontology (DO) [25] vocabularies from DisGeNET [22]. Next, we directly use the available gene-disease associations and DO hierarchy to specify the gene-term association matrix $\mathbf{Y} \in \mathbb{R}^{n \times c}$ between n genes and c DO terms. Specifically, if a DO term s, or s's descendant terms are positively associated with gene i, then $\mathbf{Y}(i,s) = 1$. Otherwise, $\mathbf{Y}(i,s) = 0$.

We collected the gene interaction data from BioGrid (https://thebiogrid.org), which is a curated biological database of genetic interactions, chemical interactions, and post-translational modifications of gene products. Let $\mathbf{S}_{11}^{(v)} \in \mathbb{R}^{n \times n}$ encode the gene-level interaction, $\mathbf{S}_{11}^{(1)}(i,j) > 0$ if the gene i has a physical interaction with gene j, $\mathbf{S}_{11}^{(1)}(i,j) = 0$ otherwise, and the weight of $\mathbf{S}_{11}^{(1)}(i,j)$ is determined by the interaction strength. We collected the gene sequence data from NCBI (https://www.ncbi.nlm.nih.gov/). We adopted conjoint triad method [27] to represent nucleic acid sequences by numeric features and then adopted cosine similarity to construct another gene similarity network $\mathbf{S}_{11}^{(2)} \in \mathbb{R}^{n \times n}$.

2.2 Isoform-Disease Associations Prediction

Owing to the lack of DO annotations of isoforms, traditional supervised learning cannot be directly applied to predict IDAs. A bypass solution is to distribute the collected gene-level GDAs (stored in $\mathbf{Y}$) to individual isoforms spliced from the genes using the readily available gene-isoform relations (stored in $\mathbf{R}_{12} \in \mathbb{R}^{n \times m}$, $\mathbf{R}_{12}(i,j) = 1$ if isoform j is spliced from gene i, $\mathbf{R}_{12}(i,j) = 0$ otherwise). Suppose $\mathbf{Z} \in \mathbb{R}^{m \times c}$ stores the latent associations between m isoforms and c distinct DO terms. Following the MIL principle that the labels of a bag is responsible by at least one instance of this bag [2, 17], a GDA should also be responsible by at least one isoform spliced from this gene. To concrete this principle, we define a dispatch and aggregation objective to push the gene-level associations to isoform-level and reversely aggregate the associations to gene-level in a compatible way as follows:

$$\mathbf{Y} = \mathbf{\Lambda} \mathbf{R}_{12} \mathbf{Z} \tag{1}$$

where $\mathbf{\Lambda} \in \mathbb{R}^{n \times n}$ is a diagonal matrix, $\mathbf{\Lambda}(i,i) = 1/n_i$, n_i represents the number of distinct isoforms spliced from the i-th gene. Given the known $\mathbf{Y}$, $\mathbf{\Lambda}$ and $\mathbf{R}_{12}$, we can optimize $\mathbf{Z}$ by minimizing $\|\mathbf{Y} - \mathbf{\Lambda}\mathbf{R}_{12}\mathbf{Z}\|_F^2$, and thus to predict the associations between m isoforms and c DO terms. Next, we can induce a linear predictor based on $\mathbf{Z}$ as follows:

$$\min \Omega(\mathbf{W}, \mathbf{Z}) = \|\mathbf{Z} - \mathbf{X}\mathbf{W}\|_F^2 + \|\mathbf{Y} - \mathbf{\Lambda}\mathbf{R}_{12}\mathbf{Z}\|_F^2 + \alpha \|\mathbf{W}\|_F^2 \tag{2}$$

where $\mathbf{W} \in \mathbb{R}^{d \times c}$ is the coefficient matrix for the linear predictor, which maps the numeric expression features $\mathbf{X}$ of isoforms onto c distinct DO terms. The Frobenius norm and scale parameter α are added to control the complexity of linear predictor.

The above equation can simultaneously distribute GDAs to individual isoforms and induce a classifier to predict IDAs. However, it ignores the important genomics data, which carry important information to boost the performance of isoform function prediction and to identify the genetic determinants of disease [3, 37]. Similarly, the incorporation of genomic data can also improve the performance of predicting IDAs. Furthermore, the collected GDAs are still incomplete. As a consequence, the distributed IDAs are also not sufficient to induce a reliable predictor and the predictor may be mislead by the collected GDAs, which are imbalanced and biased by the research interests of the community [8,9]. To alleviate these issues, we replenish GDAs by fusing gene-gene interactions and nucleic acid sequence data, and update the above equation as follows:

$$\begin{aligned}\min \Omega(\mathbf{W},\mathbf{Z},\mathbf{F}) &= \|\mathbf{Z}-\mathbf{X}\mathbf{W}\|_F^2 + \alpha\|\mathbf{W}\|_F^2 + \|\mathbf{F}-\mathbf{\Lambda}\mathbf{R}_{12}\mathbf{Z}\|_F^2 + \|\mathbf{H}\odot(\mathbf{F}-\mathbf{Y})\|_F^2 \\ &+ \frac{1}{2V_n}\sum_{v=1}^{V_n}\sum\nolimits_{i,j=1}^{n}\|\mathbf{F}(i,\cdot)-\mathbf{F}(j,\cdot)\|_F^2\mathbf{S}_n^{(v)}(i,j) \\ &= \|\mathbf{Z}-\mathbf{X}\mathbf{W}\|_F^2 + \alpha\|\mathbf{W}\|_F^2 + \|\mathbf{F}-\mathbf{\Lambda}\mathbf{R}_{12}\mathbf{Z}\|_F^2 + \|\mathbf{H}\odot(\mathbf{F}-\mathbf{Y})\|_F^2 \\ &+ \frac{1}{V_n}\sum_{v=1}^{V_n} tr(\mathbf{F}^T\mathbf{L}_n^{(v)}\mathbf{F}))\end{aligned} \tag{3}$$

where $\mathbf{F}\in\mathbb{R}^{n\times c}$ stores the latent IDAs between n genes and c DO terms. $\mathbf{H}=\mathbf{Y}$, $\odot$ means the element-wise multiplication. $\|\mathbf{H}\odot(\mathbf{F}-\mathbf{Y})\|_F^2$ is introduced to enforce latent IDAs being consistent with the collected ones and also to differentiate the observed ones from latent ones, and thus to reduce the bias toward observed ones. $\frac{1}{V_n}\sum_{v=1}^{V_n} tr(\mathbf{F}^T\mathbf{L}_n^{(v)}\mathbf{F}))$ is introduced to replenish IDAs by fusing diverse gene-level data, and V_n is the number of genomic data sources. Here, we specify the elements of $\mathbf{S}_n^{(v)}$ using the gene interaction network and nucleic acid sequences (as stated in the data preprocess subsection). $\mathbf{L}_n^{(v)}=\mathbf{D}_n^{(v)}-\mathbf{S}_n^{(v)}$, $\mathbf{D}_n^{(v)}$ is a diagonal matrix with $\mathbf{D}_n^{(v)}(i,i)=\sum_{j=1}^{n}\mathbf{S}_n^{(v)}(i,j)$.

The co-expression pattern of isoforms also carry important information about the functions of isoforms [3, 40], whose usage also boosts the prediction of IDAs. In addition, the expression of isoforms has tissue specificity [7, 38]. To make use of tissue-specific co-expression patterns, we update the objective function of IDAPred as follows:

$$\begin{aligned}\min \Omega(\mathbf{W},\mathbf{Z},\mathbf{F}) = \|\mathbf{Z}-\mathbf{X}\mathbf{W}\|_F^2 + \frac{1}{V_m}\sum_{v=1}^{V_m} tr(\mathbf{Z}^T\mathbf{L}_m^{(v)}\mathbf{Z}) + \alpha\|\mathbf{W}\|_F^2 \\ +\beta(\|\mathbf{F}-\mathbf{\Lambda}\mathbf{R}_{12}\mathbf{Z}\|_F^2 + \frac{1}{V_n}\sum_{v=1}^{V_n} tr(\mathbf{F}^T\mathbf{L}_n^{(v)}\mathbf{F}) + \|\mathbf{H}\odot(\mathbf{F}-\mathbf{Y})\|_F^2)\end{aligned} \tag{4}$$

where V_m counts the number of tissues that are used to construct the expression profile feature vectors of m isoforms in $\mathbf{X}$, $\mathbf{L}_m^{(v)}=\mathbf{D}_m^{(v)}-\mathbf{S}_m^{(v)}$, and $\mathbf{S}_m^{(v)}\in\mathbb{R}^{m\times m}$ encodes the co-expression patterns of m isoforms from the v-th tissue. $\mathbf{D}_m^{(v)}$ is a diagonal matrix with $\mathbf{D}_m^{(v)}(i,i)=\sum_{j=1}^{m}\mathbf{S}_m^{(v)}(i,j)$. β is introduced to balance the information sources from the gene-level and isoform-level.

The optimization problem in Eq. (4) is non-convex with respect to $\mathbf{W}$, $\mathbf{Z}$ and $\mathbf{F}$ altogether. It is difficult to seek the global optimal solutions for them at the same time. We follow the idea of alternating direction method of multipliers (ADMM) [1] to alternatively optimize one variable by fixing the other two variables in an iterative way. IDAPred often converges in 60 iterations on our used datasets. The optimization detail is omitted here for page limit.

3 Experiment results and analysis

3.1 Experimental setup

To assess the performance of IDAPred for predicting IDAs, we collect multiple RNA-Seq datasets from ENCODE project, gene-disease associations data from DisGeNET, gene interaction data from BioGrid, sequence data of genes from NCBI. We only consider the genes within all the four types of data for experiments. The pre-processed GDAs and isoforms of the genes are listed in Table 1.

Table 1. Statistics of isoforms and collected GDAs. 'associations' is the number of GDAs for experiment.

genes(n)	isoforms(m)	terms(c)	associations
2,482	14,484	2,949	73,515

To comparatively study the performance of IDAPred, we take the state-of-the-art isoform function prediction methods (iMILP [14], IsoFun [40], Disofun [34]) and gene-disease association prediction method (PRINCE [32]) as comparing methods. The input parameters of these comparing methods are fixed/optimized as the original papers or shared codes. For IDAPred, we choose α and β in $\left\{10^{-4}, 10^{-3}, \ldots, 10^{3}, 10^{4}\right\}$. Due to the lack of IDAs, we surrogate the evaluation by aggregating the predicted IDAs to affiliated genes, this approximate evaluation was also adopted in isoform function prediction [14, 40]. In addition, we further compare IDAPred against its degenerated variants to further study the contribution components of IDAPred.

The task of predicting IDAs can be evaluated alike gene function prediction [11, 39], and multi-instance multi-label learning by taking each gene as bag, the spliced isoforms as instances and associated diseases (DO terms) as distinct labels [36,41]. Given that, we adopt five evaluation metrics $MicroF1$, $MacroF1$, $1 - RankLoss$, $Fmax$ and $AUPRC$, which are widely-used in gene function prediction and multi-label learning. $MicroF1$ computes the F1-score on the predictions of different DO terms as a whole; $MacroF1$ calculates the F1-score of each term, and then takes the average value across all DO terms; $RankLoss$ computes the average fraction of incorrectly predicted associations ranking ahead of the ground-truth associations. $Fmax$ is the global maximum harmonic mean of recall and precision across all possible thresholds. $AUPRC$ calculates the area under the precision-recall curve of each term, and then computes the average value of these areas as the overall performance. The higher the value of $MicroF1$, $MacroF1$, $1 - RankLoss$, $Fmax$ and $AUPRC$, the better the performance is. We want to remark that these five metrics quantify the prediction results from different aspects, and it is difficult for one method to always outperform another one across all these metrics.

3.2 Results evaluation at gene-level

We adopt five-fold cross-validation at the gene-level for experiment. For each test fold, we randomly initialize the test part of $\mathbf{F}$ and $\mathbf{Y}$ in Eq. (4). We initialize the isoform-term association matrix $\mathbf{Z}$ by the gene-term association matrix $\mathbf{F}$, say all the diseases associated with a gene are also initialized as temporarily associated with its spliced isoforms. The GDAs in the validation set are considered as unknown during training and prediction, and only used for validation. Table 2 reports the results of IDAPred and of compared methods.

Table 2. Experimental results of five-fold cross-validation. ●/○ indicates IDAPred performing significantly better/worse than the other comparing method, with significance assessed by pairwise t-test at 95% level.

	PRINCE	iMILP	IsoFun	Disofun	IDAPred
MicroF1	0.3122±0.0274●	0.2349±0.0273●	0.2829±0.0195●	0.3306±0.0092●	0.8248±0.0118
MacroF1	0.2863±0.0341●	0.0645±0.0269●	0.1232±0.0254●	0.0398±0.0046●	0.4250±0.0241
1-RankLoss	0.8591±0.0434●	0.0836±0.0473●	0.6877±0.0536●	0.8699±0.0016●	0.9966±0.0003
Fmax	0.3281±0.0097●	0.1559±0.0750●	0.2140±0.0143●	0.2250±0.0109●	0.6795±0.0068
AUPRC	0.3596±0.0025●	0.0092±0.0051●	0.0413±0.0031●	0.0430±0.0049●	0.4782±0.0067

IDAPred gives significantly better results than the compared methods across all the five evaluation metrics. $MicroF1$, $MacroF1$ and $AUPRC$ are disease term-centric metrics, while $1 - Rankloss$ and $Fmax$ are gene-centric metrics. The significant improvement shows that IDAPred can more reliably predict the GDAs (IDAs) from both the gene (isoforms) and DO term perspectives. Three factors contribute to this improvement. (i) IDAPred fuses the gene sequence and interaction data to complete GDAs, along with the isoform expression data, while these compared methods either use only the interaction data and/or the expression data. (ii) IDAPred accounts for tissue specificity and fuses co-expression networks of different tissues, while IsoFun and Disofun concatenate the expression profiles of different tissues into a single feature vector and then construct a single co-expression network; as a result, they do not make use of the important tissue specificity patterns of alternative splicing. (iii) IDAPred models the incompleteness of the gene-term associations and introduces the indicator matrix $\mathbf{H}$ to enforce latent IDAs being consistent with the collected ones, and to differentiate the observed ones from latent ones.

PRINCE directly predicts GDAs based on the topology of gene interaction networks, and it outperforms most comparing methods (except our proposed

IDAPred). One explanation is that the evaluation is approximately made at the gene-level, not the targeted isoform-level, and these compared methods more focus on using the transcriptomics expression data. Last but not least, we want to remark that IDAPred is an inductive approach that can directly predict the associations between diseases and a new isoform, whereas these compared methods can only work in transductive setting, they have to include this isoform for retraining the model and then to make the prediction.

Overall, these comparisons indirectly prove the effectiveness of IDAPred in predicting the associations between isoforms and diseases.

3.3 Further analysis

Ablation study To further study the contribution components, we introduce five variants of IDAPred, which are IDAPred(L), IDAPred(P), IDAPred(S), IDAPred(A) and IDAPred(H). IDAPred(L) removes the $\frac{1}{V_n}\sum_{v=1}^{V_n} tr(\mathbf{F}^T\mathbf{L}_n^{(v)}\mathbf{F})$ in Eq. (4), namely both the gene sequence and interaction data are excluded; IDAPred(P) only uses the gene interaction data; IDAPred(S) only utilizes the gene sequence data; IDAPred(A) concatenates the isoform expression profile feature vectors of different tissues into a single one, and then directly constructs a single isoform co-expression network using cosine similarity also. IDAPred(H) removes the indicator $\mathbf{H}$ in $\|\mathbf{H}\odot(\mathbf{F}-\mathbf{Y})\|_F^2$) in Eq. (4), say it does not consider the bias toward the observed GDAs. Figure 1 reports the performance results of IDAPred and of its variants. The experimental settings are the same as the evaluation at the gene-level.

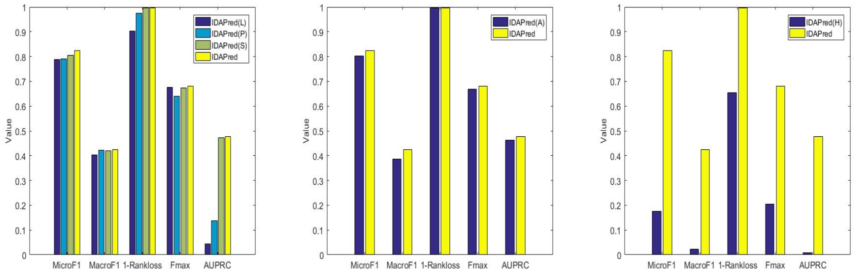

Fig. 1. Performance results of IDAPred and its variants, which fuse fewer data or do not alleviate the bias toward observed GDAs.

It is easy to observe that IDAPred manifests the highest performance among its variants. IDAPred(L) has much lower performance values than IDAPred. This fact corroborates our assumption that the observed GDAs are incomplete, and also the contribution of fusing gene interaction and sequence data to complete the GDAs, which then improve the prediction of IDAs. IDAPred(P) and

IDAPred(S) manifest better results than IDAPred(L), but they both are outperformed by IDAPred. This comparison not only shows that gene interaction network data and gene sequence data can help to replenish GDAs, but also expresses the joint benefit of fusing gene interaction and sequence data. IDAPred(P) has a lower performance than IDAPred(S), this facts the gene sequence data is more positively related with the isoform/gene-disease associations than the incomplete gene interaction data. IDAPred(A) also loses to IDAPred, which proves the necessity of combining isoform co-expression patterns from tissue-wise, instead from sample-wise. There is a big performance margin between IDAPred and IDAPred(H), which expresses the importance to explicitly account for the incompleteness of observed GDAs and to alleviate the bias toward observed GDAs, which is overlooked by most compared methods.

In summary, the ablation study also confirms the effectiveness of our unified objective function in fusing genomics and trascriptomics data, and in handling the difficulty of predicting IDAs.

Parameter sensitivity analysis There are two input parameters (α and β) involved with IDAPred. α controls the complexity of linear predictor, and β balances the information sources from the gene-level and isoform-level. We vary α and β in the grid of $\{10^{-4}, 10^{-3}, \cdots, 10^{3}, 10^{4}\}$, and visualize the results of IDAPred under different combinations of α and β in Figure 2.

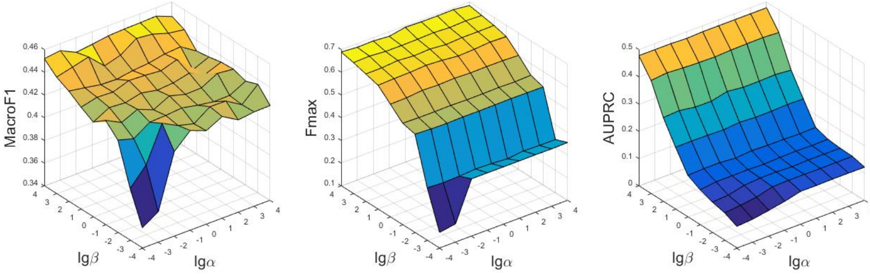

Fig. 2. Performance results vs. α and β.

We observe that IDAPred firstly has a clearly increased performance as α growing from 10^{-4} to 10^{-2}, and then holds a relatively stable performance as α further growing. As β growing from 10^{-4} to 10^{-1}, IDAPred also shows a sharply increased performance trend, and a slowly increased trend as β further growing from 10^{-1} to 10^{4}. This trend again confirms that the gene-level data should be leveraged for predicting IDAs. We also find β playing more important role than α. That is because α only controls the complexity of predictor, while the complexity is also inherently controlled by the simple linear classifier. When both α and β are fixed with too small values, IDAPred has the lowest performance. This observation again expresses the effectiveness of the unified objective function for handling the difficulty of predicting IDAs. Based on these results, we adopt $\alpha = 10^{-2}$ and $\beta = 10^4$ for experiments.

4 Conclusion

In this paper, we proposed an approach called IDAPred to computationally predict isoform-disease associations by data fusion. IDAPred makes use of multi-instance learning to bypass the lack of the ground-truth isoform-disease associations and to push gene-disease associations onto individual isoforms. It fuses nucleic acid sequences and interactome of genes to further fulfil the incomplete GDAs. In addition, it leverages multiple RNA-seq datasets to construct tissue-specific isoform co-expression networks and to induce a linear classifier to predict IDAs. Experimental results show that IDAPred significantly outperforms related comparing methods, which target to identify gene-disease associations or isoform functions.

5 Acknowledgements

This research is supported by NSFC (61872300), Fundamental Research Funds for the Central Universities (XDJK2019B024 and XDJK2020B028), Natural Science Foundation of CQ CSTC (cstc2018jcyjAX0228).

References

1. Boyd, S., Vandenberghe, L.: Convex optimization. Cambridge University Press (2004)
2. Carbonneau, M.A., Cheplygina, V., Granger, E., Gagnon, G.: Multiple instance learning: A survey of problem characteristics and applications. Pattern Recognition **77**, 329–353 (2018)
3. Chen, H., Shaw, D., Zeng, J., Bu, D., Jiang, T.: Diffuse: predicting isoform functions from sequences and expression profiles via deep learning. Bioinformatics **35**(14), i284–i294 (2019)
4. Claussnitzer, M., Cho, J.H., Collins, R., Cox, N.J., Dermitzakis, E.T., Hurles, M.E., Kathiresan, S., Kenny, E.E., Lindgren, C.M., MacArthur, D.G., et al.: A brief history of human disease genetics. Nature **577**(7789), 179–189 (2020)
5. Consortium, E.P., et al.: An integrated encyclopedia of dna elements in the human genome. Nature **489**(7414), 57 (2012)
6. Eksi, R., Li, H.D., Menon, R., Wen, Y., Omenn, G.S., Kretzler, M., Guan, Y.: Systematically differentiating functions for alternatively spliced isoforms through integrating rna-seq data. PLoS Computational Biology **9**(11), e1003314 (2013)
7. Ellis, J.D., Barrios-Rodiles, M., Çolak, R., Irimia, M., Kim, T., Calarco, J.A., Wang, X., Pan, Q., O'Hanlon, D., Kim, P.M., et al.: Tissue-specific alternative splicing remodels protein-protein interaction networks. Molecular Cell **46**(6), 884–892 (2012)

8. Gaudet, P., Dessimoz, C.: Gene ontology: pitfalls, biases, and remedies. In: The Gene Ontology Handbook, pp. 189–205. Humana Press, New York, NY (2017)
9. Holman, L., Head, M.L., Lanfear, R., Jennions, M.D.: Evidence of experimental bias in the life sciences: why we need blind data recording. PLoS Biology **13**(7) (2015)
10. Holtzman, D.M., Bales, K.R., Tenkova, T., Fagan, A.M., Parsadanian, M., Sartorius, L.J., Mackey, B., Olney, J., McKeel, D., Wozniak, D., et al.: Apolipoprotein e isoform-dependent amyloid deposition and neuritic degeneration in a mouse model of alzheimer's disease. Proceedings of the National Academy of Sciences **97**(6), 2892–2897 (2000)
11. Jiang, Y., Oron, T.R., Clark, W.T., Bankapur, A.R., D¡Andrea, D., Lepore, R., Funk, C.S., Kahanda, I., Verspoor, K.M., Ben-Hur, A., et al.: An expanded evaluation of protein function prediction methods shows an improvement in accuracy. Genome Biology **17**(1), 184 (2016)
12. Kim, D., Langmead, B., Salzberg, S.L.: Hisat: a fast spliced aligner with low memory requirements. Nature Methods **12**(4), 357 (2015)
13. Li, H.D., Menon, R., Omenn, G.S., Guan, Y.: The emerging era of genomic data integration for analyzing splice isoform function. Trends in Genetics **30**(8), 340–347 (2014)
14. Li, W., Kang, S., Liu, C.C., Zhang, S., Shi, Y., Liu, Y., Zhou, X.J.: High-resolution functional annotation of human transcriptome: predicting isoform functions by a novel multiple instance-based label propagation method. Nucleic Acids Research **42**(6), e39–e39 (2014)
15. Lundberg, A.K., Jonasson, L., Hansson, G.K., Mailer, R.K.: Activation-induced foxp3 isoform profile in peripheral cd4+ t cells is associated with coronary artery disease. Atherosclerosis **267**, 27–33 (2017)
16. Luo, P., Li, Y., Tian, L.P., Wu, F.X.: Enhancing the prediction of disease–gene associations with multimodal deep learning. Bioinformatics **35**(19), 3735–3742 (2019)
17. Maron, O., Lozano-Pérez, T.: A framework for multiple-instance learning. In: NeurIPS. pp. 570–576 (1998)
18. Natarajan, N., Dhillon, I.S.: Inductive matrix completion for predicting gene–disease associations. Bioinformatics **30**(12), i60–i68 (2014)
19. Neagoe, C., Kulke, M., del Monte, F., Gwathmey, J.K., de Tombe, P.P., Hajjar, R.J., Linke, W.A.: Titin isoform switch in ischemic human heart disease. Circulation **106**(11), 1333–1341 (2002)
20. Pan, Q., Shai, O., Lee, L.J., Frey, B.J., Blencowe, B.J.: Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nature Genetics **40**(12), 1413 (2008)
21. Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T., Salzberg, S.L.: Stringtie enables improved reconstruction of a transcriptome from rna-seq reads. Nature Biotechnology **33**(3), 290 (2015)
22. Piñero, J., Ramírez-Anguita, J.M., Saüch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., Furlong, L.I.: The disgenet knowledge platform for disease genomics: 2019 update. Nucleic Acids Research **48**(D1), D845–D855 (2020)
23. Pletscher-Frankild, S., Pallejà, A., Tsafou, K., Binder, J.X., Jensen, L.J.: Diseases: Text mining and data integration of disease–gene associations. Methods **74**, 83–89 (2015)
24. Sanan, D.A., Weisgraber, K.H., Russell, S.J., Mahley, R.W., Huang, D., Saunders, A., Schmechel, D., Wisniewski, T., Frangione, B., Roses, A.D.: Apolipoprotein e associates with beta amyloid peptide of alzheimer's disease to form novel monofibrils.

isoform apoe4 associates more efficiently than apoe3. Journal of Clinical Investigation **94**(2), 860–869 (1994)
25. Schriml, L.M., Arze, C., Nadendla, S., Chang, Y.W.W., Mazaitis, M., Felix, V., Feng, G., Kibbe, W.A.: Disease ontology: a backbone for disease semantic integration. Nucleic Acids Research **40**(D1), D940–D946 (2012)
26. Shaw, D., Chen, H., Jiang, T.: Deepisofun: a deep domain adaptation approach to predict isoform functions. Bioinformatics **35**(15), 2535–2544 (2019)
27. Shen, J., Zhang, J., Luo, X., Zhu, W., Yu, K., Chen, K., Li, Y., Jiang, H.: Predicting protein–protein interactions based only on sequences information. Proceedings of the National Academy of Sciences **104**(11), 4337–4341 (2007)
28. Skotheim, R.I., Nees, M.: Alternative splicing in cancer: noise, functional, or systematic? International Journal of Biochemistry & Cell Biology **39**(7-8), 1432–1449 (2007)
29. Smith, L.M., Kelleher, N.L.: Proteoforms as the next proteomics currency. Science **359**(6380), 1106–1107 (2018)
30. Strittmatter, W.J., Weisgraber, K.H., Huang, D.Y., Dong, L.M., Salvesen, G.S., Pericak-Vance, M., Schmechel, D., Saunders, A.M., Goldgaber, D., Roses, A.D.: Binding of human apolipoprotein e to synthetic amyloid beta peptide: isoform-specific effects and implications for late-onset alzheimer disease. Proceedings of the National Academy of Sciences **90**(17), 8098–8102 (1993)
31. Sun, P.G., Gao, L., Han, S.: Prediction of human disease-related gene clusters by clustering analysis. International Journal of Biological Sciences **7**(1), 61 (2011)
32. Vanunu, O., Magger, O., Ruppin, E., Shlomi, T., Sharan, R.: Associating genes and protein complexes with disease via network propagation. PLoS computational biology **6**(1), e1000641 (2010)
33. Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., Burge, C.B.: Alternative isoform regulation in human tissue transcriptomes. Nature **456**(7221), 470 (2008)
34. Wang, K., Wang, J., Domeniconi, C., Zhang, X., Yu, G.: Differentiating isoform functions with collaborative matrix factorization. Bioinformatics **36**(6), 1864––1871 (2020)
35. Wang, X., Gulbahce, N., Yu, H.: Network-based methods for human disease gene prediction. Briefings in Functional Genomics **10**(5), 280–293 (2011)
36. Xing, Y., Yu, G., Domeniconi, C., Wang, J., Zhang, Z., Guo, M.: Multi-view multi-instance multi-label learning based on collaborative matrix factorization. In: AAAI. pp. 5508–5515 (2019)
37. Xiong, H.Y., Alipanahi, B., Lee, L.J., Bretschneider, H., Merico, D., Yuen, R.K., Hua, Y., Gueroussov, S., Najafabadi, H.S., Hughes, T.R., et al.: The human splicing code reveals new insights into the genetic determinants of disease. Science **347**(6218), 1254806 (2015)
38. Yeo, G., Holste, D., Kreiman, G., Burge, C.B.: Variation in alternative splicing across human tissues. Genome Biology **5**(10), R74 (2004)
39. Yu, G., Rangwala, H., Domeniconi, C., Zhang, G., Yu, Z.: Protein function prediction using multilabel ensemble classification. IEEE/ACM Transactions on Computational Biology and Bioinformatics **10**(4), 1045–1057 (2013)
40. Yu, G., Wang, K., Domeniconi, C., Guo, M., Wang, J.: Isoform function prediction based on bi-random walks on a heterogeneous network. Bioinformatics **36**(1), 303–310 (2020)
41. Zhou, Z.H., Zhang, M.L., Huang, S.J., Li, Y.F.: Multi-instance multi-label learning. Artificial Intelligence **176**(1), 2291–2320 (2012)