IsoDA: Isoform-disease association prediction by multi-omics data fusion

Qiuyue Huang[1,2], Jun Wang[2], Xiangliang Zhang[3], Maozu Guo[4]* and Guoxian Yu[1,2,3]*

[1] College of Computer and Information Science, Southwest University, Chongqing 400715, China
[2] School of Software, Shandong University, Jinan 250101, China
[3] CEMSE, King Abudullah University of Science and Technology, Thuwal, Saudi Arabia
[4] College of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing, China
*Corresponding author: guomaozu@buecea.edu.cn (Maozu Guo); gxyu@sdu.edu.cn (Guoxian Yu)

Abstract. A gene can be spliced into different isoforms by alternative splicing, which contributes to the functional diversity of protein species. Computationally predicting the gene-disease associations has been studied for decades. However, how to identify the isoform-disease associations at a large scale is rarely explored, which can decipher the pathology at a more granular level. The main bottleneck is the lack of isoform-disease associations in current databases and multi-level omics data fusion. To bridge this gap, we propose a computational approach called IsoDA to predict isoform-disease associations. Based on the relationship between a gene and its spliced isoforms, IsoDA firstly introduces a dispatch and aggregation term to dispatch gene-disease associations to individual isoforms, and reversely aggregate these dispatched associations to their hosting genes. At the same time, it fuses the genome, transcriptome and proteome data by joint matrix factorization to improve the prediction of isoform-disease associations. Experimental results show that IsoDA significantly outperforms the related state-of-the-art methods at the gene-level and isoform-level both, Case study further shows that IsoDA credibly identifies three isoforms spliced from APOE which have individual associations with Alzheimer disease, and two isoforms spliced from VEGFA that have different associations with Coronary Heart Disease. The codes of IsoDA are available at http://mlda.swu.edu.cn/codes.php?name=IsoDA.

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

1 Introduction

Understanding the genetic mechanism and pathology of diseases help the decipher of human genome and the development of life science [7, 51]. The discovery of genetic disease association is very important for disease prevention,

diagnosis and treatment. Wet-lab (clinical) based methods or high-throughput bio-technologies can help us identifying the candidate genes associated with a particular diseases, but they are still limited by a low throughput or coverage, but high costs. With the rapid accumulation of multi-omics data (*i.e.*, genomics, transcriptomics and proteomics) related with the gene products and disease phenotypes, diverse computational methods have been proposed [26, 29, 48, 57]. These computational solutions can save resources by excluding genes unlikely to be associated with diseases. These approaches build on different machine learning techniques [13, 46, 57], such as network propagation [17, 35, 48, 52], matrix factorization [29], data fusion [34] and deep neural networks [26, 55]. They mainly use gene-disease associations collected from public databases (i.e., DisGeNET [33] and OMIM [14]). The integration of multi-level omics data is essential for the development of high-precision predictive models. To achieve a better performance, researchers further fused protein-protein interaction data from BioGRID [44], Gene functional network from HumanNet [54], RNA-seq datasets and many others.

Existing computational solutions for predicting genetic disease associations still focus on gene-level. However, a gene can be associated with diverse diseases mainly caused by the isoforms alternatively spliced from the same gene. It is reported that more than 90% human multi-exon genes undergo alternative splicing [31, 49], which greatly increases the transcriptome and proteome complexity [43]. The preteoforms translated from different isoforms of the same gene have different amino acid sequences and structures, thus may have different associations with diverse diseases. Diverse complex diseases have been found to be associated with alternative splicing, such as autism spectrum disorders [42], ischemic human heart disease [30], Alzheimer's disease [15] and so on. Apolipoprotein E (apoE) is localized in the senile plaques, congophilic angiopathy, and neurofibrillary tangles of Alzheimer's disease. Strittmatter *et al.* [45] reported that the pathogenesis of Alzheimer's disease may be related with different bindings in apoE. They compared the difference of binding of synthetic amyloid beta (beta/A4) peptide to apoE4 and apoE3 (two commom isoforms of apoE), and observed that apoE4 is associated with the increased susceptibility to disease. Neagoe *et al.* [30] 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, but 32:68 in nonischemic transplants.

Identifying the isoform-disease associations enables a deeper view of the molecular basis of diverse genetic diseases, and helps exploring precise strategies and drugs to treat diverse complex diseases. However, available isoform-disease associations (IDAs) are mainly detected by biological experiments and there is no public database storing sufficient IDAs for training. Therefore, traditional machine learning methods can not be directly adopted for predicting IDAs. In fact, such bottleneck also exists in isoform function prediction. To overcome this difficulty, some researchers adapt multiple instance learning (MIL) [4, 28] for isoform function prediction. They model the gene as a bag and the isoforms spliced

from this gene as its instances, and then identify the individual functions of isoforms by leveraging the known gene-level functional annotations, gene-isoform relations, multiple RAN-seq datasets [6, 11, 23, 39, 50, 56]. These solutions mainly focus on using RNA-seq datasets, and/or genomic/proteomic data, without accounting for latent correlations between functional labels, or fuse two types of omics data only.

In this paper, we proposed a task of predicting IDAs (Isoform-Disease Associations). Compared with the canonical GDAs (Gene-Disease Associations) prediction task, the IDAs prediction task is more deeper and challenging, due to the lack of IDAs and complexity of alternative splicing. With the advance of RNA-seq technology, large-scale high-resolution transcript-level expression data can be easily collected [53] and the isoform expression can be quantified at a more precise level. Therefore, IsoDA integrates multiple RNA-seq datasets to identify IDAs. Particularly, IsoDA introduces a regularization term to distribute known GDAs of a gene to its isoforms and reversely aggregate IDAs to gene-level using the gene-isoform relations. Considering the incomplete GDAs, IsoDA leverages protein interaction data to replenish GDAs and constructs tissue-wise isoform co-expression networks using 298 RNA-seq datasets to account for the tissue specificity of alternative splicing. It further uses the isoform sequence data to construct another isoform functional association network, and then combines these networks with adaptive weights to induce a network regularized multi-label linear classifier to predict IDAs. In addition, IsoDA introduces an indicator matrix into the unified objective function to differentiate the observed GDAs from unobserved ones and thus alleviates the bias toward observed ones. This paper is an extension of our conference work [16], which as a showcase proposes the isoform-disease association prediction task and demonstrates the fusion of genomics and RNA-seq datasets enables the prediction of IDAs. In this extended version, we adopt a larger Human dataset with more genes, isoforms and diseases. We fuse more omics data (genomics, transcriptomics and proteomics), explicitly model the interrelationships between diseases, and give more details on optimizing the fusion of multi-level omics data, and conduct more comprehensive validations. Experimental results show that IsoDA achieves better results than other competitive approaches, including two approaches for predicting G-DAs [48, 57] and three solutions for predicting isoform functions [23, 50, 56].

2 Related Work

Due to the lack of IDAs in public repositories, there is almost no computational solution for identifying IDAs at a large scale. From the gene-isoform relations, the prediction of IDAs can be modeled as a multi-instance learning problem [4, 58], which has been extensively applied for isoform function prediction in recent years and has a close connection with the prediction of IDAs. Unlike the widely-studied gene/protein function prediction, the isoform function prediction is still a tough problem. The main difficulty is the lack of functional annotation at the isoform-level and the complex relation between genes and isoforms. Exist-

ing functional genome databases (*i.e.*, Gene Ontology [2] and KEGG [19]) only record the functional annotation of gene products at the gene-level and contemporary molecule-interaction databases (*i.e.*, BioGRID [5] and STRING [47]) still record the interaction between proteins at the gene-level.

Several teams tried to push the gene-level annotations to individual isoforms by adopting multi-instance learning [6, 12, 24, 27, 39, 50, 56]. These computational solutions model a gene as a bag and its spliced isoforms as instances. They typically follow the principle that a gene is positive for a functional label if at least one of its isoforms is positively annotated with that label, while the gene is negative for a label means none of its isoforms annotated with that label. To name a few, Eksi *et al.* [12] adopted the multiple instance support vector machine (miSVM) [1] to differentiate functions of isoforms on the Mouse RNA-seq data. miSVM leverages the functional annotations of genes, isoform expression data and gene-isoform associations to generate an isoform-level maximum margin classifier. Li *et al.* [24] developed the instance-oriented multi-instance label propagation (iMILP) to predict isoform functions. iMILP first constructs multiple isoform functional association networks, then uses GO annotations of a gene to universally initialize the annotations of isoforms, next it updates the annotations of isoforms based on the greedy combination of multiple networks and label propagation on the combined network. Luo *et al.* [27] proposed a novel sparse simplex projection based approach (WLRM) to differentiate the functions of isoforms within the MIL framework. WLRM specially takes the genes annotated with the function as positive bags and the genes without the function as negative ones, and then maps the original bag space to a different feature space. To alleviate the lack of ground-truth annotations at the isoform-level, Shaw *et al.* [39] proposed a deep learning based method (DeepIsoFun) that combines MIL with domain adaption to predict isoform functions, which provide additional labeled training data to transfer the knowledge of gene functions to the prediction of isoform functions from GO annotations and RNA-Seq data. Yu *et al.* [56] recently introduced an approach (IsoFun) to predict isoform functions based on bi-random walks on a heterogeneous network, which is composed of isoform functional association network, GO annotations of genes, gene-gene interaction network, and the gene-isoform relations. Chen *et. al.* [6] presented the Deep learning-based prediction of IsoForm Functions from Sequences and Expression (DIFFUSE). In the first stage, DIFFUSE designs a deep neural network (DNN) to capture features from isoform sequences and domains; in the second stage, it uses a conditional random field (CRF) to explore the relationship between isoforms and assigns GO annotations to isoforms based on initial scores computed by DNN. DIFFUSE trains both DNN and CRF together under a novel semi-supervised learning setting. Wang *et al.* recently [50] proposed DisoFun to differentiate isoform functions with collaborative matrix factorization. DisoFun complies with the main idea that the functional annotations of genes are aggregated from key isoforms, it jointly factorizes the isoform expression data matrix (derived from multiple RNA-seq datasets) and the gene-term association matrix (storing the Gene Ontology (GO) annotations of genes) into low-rank matri-

ces to explore the latent key isoforms, and pushes the annotations to isoforms by enforcing the aggregated annotations from isoforms being consistent with the known annotations of genes. DisoFun further leverages PPI networks and GO hierarchy structure to replenish the annotations of genes and those of key isoforms. These solutions mainly focus on using RNA-seq datasets [11, 24, 27, 39], some of them additional use genomic data [6], or protein-protein interactions [50, 56]. They neglect the important latent correlations between functional labels and simply fuse two types of omics data without differentiation.

Many studies reported that isoforms are indeed associated with many complex diseases [15, 22, 30], but the study of computational solution for isoform-disease associations are rarely reported, compared with the heavy study of gene-disease association prediction [26, 29, 48, 57]. The recent progress on isoform function prediction sheds light on how to infer IDAs. In this paper, we introduce a computational solution (IsoDA) by fusing multi-omics data and multi-instance learning in a principle way. IsoDA integrates multiple isoform-isoform association networks derived from multiple RNA-seq datasets and the sequence similarity work derived from nucleotides with adaptive weights. It takes advantage of the PPI network to replenish the missing GDAs, and then induce a linear classifier to push gene-level associations to individual isoforms in a coherent way. The experimental results show that IsoDA not only achieves a better performance than representative related GDAs prediction methods [48, 57], but also competitive isoform function prediction solutions [23, 50, 56]. Further case study again corroborates the effectiveness of IsoDA and advantages to these compared methods.

3 The Proposed Method

3.1 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 $m = \sum_{i=1}^{n} n_i$. $\mathbf{R}_{12} \in \mathbb{R}^{n \times m}$ is the relational data matrix between n genes and m isoforms, $\mathbf{R}_{12}(i, j) = 1$ if the the i-th gene hosts the j-th isoform, $\mathbf{R}_{12}(i, j) = 0$ otherwise.

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 [8] (access date: 2019-11-10). These datasets are heterogeneous in terms of library preparation procedures and sequencing platform. Following the pre-process done in [23, 50], 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) [21], 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) [32] 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). Finally, we obtain 15,204 genes with 137,910 isoforms. The expression values of these isoforms are stored in the data matrix $\mathbf{X}_1 \in \mathbb{R}^{m \times d_1}$. We further normalize $\mathbf{X}_1$ by $\mathbf{X}_1(:, j)_{nor} = \mathbf{X}_1(:, j)./max(\mathbf{X}_1(:, j))$. For convenience, we use the normalized $\mathbf{X}_1$ for subsequent experiments.

To get the available GDAs, we downloaded the GDAs file and the mappings file UMLS CUI to Disease Ontology (DO) [38] vocabularies from DisGeNET [33]. Then we directly use the available GDAs 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, $\mathbf{Y}(i, s) = 0$. In addition, we excluded the too sparse DO terms annotated to fewer than 30 genes, and the too general terms DO annotated to more than 300 genes.

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 $\mathbf{R}_{11} \in \mathbb{R}^{n \times n}$ encode the gene-level interaction, $\mathbf{R}_{11}(i, j) > 0$ if the gene i has a physical interaction with gene j, $\mathbf{R}_{11}(i, j) = 0$ otherwise, and the entry weight of $\mathbf{R}_{11}(i, j)$ is determined by the interaction strength.

We further collected the nucleotide sequences of isoform from NCBI Nucleotide database, and we adopted conjoint triad method [40] 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 sequence is composed by Adenine (A), Guanine (G), Cytosine (C), Thymine (T), three continuous bases were considered as a unit, thus a $4 \times 4 \times 4$-dimensional frequency vector were generated to represent the sequence information of each isoform. To handle the variable lengths of nucleotides of different isoforms, we further normalize the represented isoform sequence feature data matrix $\mathbf{X}_2 \in \mathbb{R}^{m \times d_2}$.

3.2 Isoform-Disease Associations Prediction

The lack of IDAs makes it difficult to directly apply the traditional supervised learning methods to predict IDAs. Within the MIL framework, we leverage the obtained gene-isoform relations $\mathbf{R}_{12}$ and gene-level disease associations to identify the distinct disease associations of individual isoforms. Suppose $\mathbf{Z} \in \mathbb{R}^{m \times c}$ stores the latent associations between m isoforms and c distinct DO terms, given the known associations $\mathbf{Y}$ between n genes and c diseases, and motivated by the principle that the labels of a bag is responsible by at least one instance of this bag [4, 28], a GDA should also be responsible by at least one isoform spliced from this gene. We can obtain the aggregated GDAs from its spliced isoforms

and distribute the collected gene-level GDAs (stored in $\mathbf{Y}$) to individual isoforms spliced from the genes 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. $\mathbf{Y}$ is the available gene-term association matrix. With this dispatch and aggregation objective, we can optimize $\mathbf{Z}$ and thus to predict the latent associations between m isoforms and c DO terms. Next, we can induce a linear predictor based on $\mathbf{Z}$ as follows:

$$\min \Omega(\mathbf{W},\mathbf{Z}) = \|\mathbf{Z} - \mathbf{X}\mathbf{W}\|_F^2 + \|\mathbf{Y} - \mathbf{\Lambda}\mathbf{R}_{12}\mathbf{Z}\|_F^2 + \lambda_1 \|\mathbf{W}\|_F^2 \tag{2}$$

where $\mathbf{X} \in \mathbb{R}^{m\times d}$ is the numeric features matrix of m isoforms, which is concatenated by isoform expression features matrix $\mathbf{X}_1$ and sequence features matrix $\mathbf{X}_2$, and $d = d_1 + d_2$. $\mathbf{W} \in \mathbb{R}^{d\times c}$ is the coefficient matrix for the linear predictor, which maps the numeric features matrix of m isoforms $\mathbf{X}$ onto c distinct DO terms. The scale parameter λ_1 are added to control the complexity of linear predictor.

By taking $\mathbf{Z}$ as the to-be-predicted variable, we can reversely push the GDAs of n genes to m isoforms and achieve the prediction of IDAs at the isoform-level. However, the collected GDAs are rather incomplete and biased, which may miss some important GDAs and lead to biased prediction of IDAs. Given that, we attempt to replenish GDAs using the interactome of genes and extend the above equation as follows:

$$\begin{aligned}\min \Omega(\mathbf{W},\mathbf{Z},\mathbf{F}) &= \|\mathbf{Z} - \mathbf{X}\mathbf{W}\|_F^2 + \lambda_1 \|\mathbf{W}\|_F^2 + \|\mathbf{F} - \mathbf{\Lambda}\mathbf{R}_{12}\mathbf{Z}\|_F^2 + \|\mathbf{H}\odot(\mathbf{F}-\mathbf{Y})\|_F^2 \\ &\quad + \frac{1}{2}\sum_{i,j=1}^{n} \|\mathbf{F}(i,\cdot) - \mathbf{F}(j,\cdot)\|_F^2 \mathbf{R}_{11}(i,j) \\ &= \|\mathbf{Z} - \mathbf{X}\mathbf{W}\|_F^2 + \lambda_1 \|\mathbf{W}\|_F^2 + \|\mathbf{F} - \mathbf{\Lambda}\mathbf{R}_{12}\mathbf{Z}\|_F^2 + \|\mathbf{H}\odot(\mathbf{F}-\mathbf{Y})\|_F^2 \\ &\quad + tr(\mathbf{F}^T\mathbf{L}_{11}\mathbf{F})\end{aligned} \tag{3}$$

where $\mathbf{F} \in \mathbb{R}^{n\times c}$ stores the latent GDAs 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 GDAs 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. $tr(\mathbf{F}^T\mathbf{L}_{11}\mathbf{F})$ is introduced to replenish IDAs by introduce protein-level interaction data. Here, the $\mathbf{R}_{11}$ refers to the protein interaction network matrix (as stated in the data preprocess subsection). $\mathbf{L}_{11} = \mathbf{D}_{11} - \mathbf{R}_{11}$, $\mathbf{D}_{11}$ is a diagonal matrix with $\mathbf{D}_{11}(i,i) = \sum_{j=1}^{n}\mathbf{R}_{11}(i,j)$.

A gene generates one or more isoforms by alternative splicing, and these isoforms have diverse expression patterns across tissues [9,18]. Based on this observation, the association networks of isoforms should be constructed from the tissue-level and more appropriate fusion of these networks can help to accurately identify the IDAs. To make full use of the tissue-specific patterns of multiple

RNA-seq datasets, we advocate to integrate multiple isoform functional association networks from the tissue-wise with weights. In addition, sequence data also carry important information for the prediction of IDAs, so we also construct a sequence similarity based isoform functional association network of isoforms. To this end, we integrate multiple isoform functional association networks and extend Eq. (3) as follows:

$$\begin{aligned}\min \Omega(\mathbf{W},\mathbf{Z},\mathbf{F},\boldsymbol{\alpha}) = \|\mathbf{Z}-\mathbf{X}\mathbf{W}\|_F^2 + \sum_{v=1}^{V}\alpha_v^p tr(\mathbf{Z}^T\mathbf{L}_{22}^{(v)}\mathbf{Z}) + \lambda_1\|\mathbf{W}\|_F^2 \\ + \lambda_2(\|\mathbf{F}-\boldsymbol{\Lambda}\mathbf{R}_{12}\mathbf{Z}\|_F^2 + tr(\mathbf{F}^T\mathbf{L}_{11}\mathbf{F}) + \|\mathbf{H}\odot(\mathbf{F}-\mathbf{Y})\|_F^2) \\ s.t. \quad \alpha_v \geq 0, \boldsymbol{\alpha}^T\mathbf{1}=1\end{aligned} \tag{4}$$

$\mathbf{L}_{22}^{(v)} = \mathbf{D}_{22}^{(v)} - \mathbf{R}_{22}^{(v)}$, and $\mathbf{R}_{22}^{(v)} \in \mathbb{R}^{m\times m}$ encodes the co-expression strength induced from multiple RNA-seq datasets of the v-th tissue. Here $V = 10$, including 9 association networks from 9 different tissues, which are obtained by cosine similarity from isoform expression feature data and 1 network based on isoform sequence feature data. $\mathbf{D}_{22}^{(v)}$ is a diagonal matrix with $\mathbf{D}_{22}^{(v)}(i,i) = \sum_{j=1}^{m}\mathbf{R}_{22}^{(v)}(i,j)$. $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_V]$ are the weights assigned to V networks. λ_2 is introduced to balance the information sources from the gene-level and isoform-level.

An isoform can be associated with different diseases and these diseases have some latent correlations. For example, the diseases are hierarchical organized by a directed acyclic graph in the disease ontology. It is recognized that the account of such hierarchical information can boost the performance of isoform function prediction [50, 56]. Here, we introduce a latent disease-disease correlation matrix $\mathbf{S} \in \mathbb{R}^{c\times c}$ into our model. We adopt cosine similarity to construct the disease-disease associations network $\mathbf{S}$ from the available GDAs data. Since the initially estimated disease-disease correlations may be incomplete and unreliable, we further optimize $\mathbf{S}$ during the training of IsoDA and formalize the objective function of IsoDA as follows:

$$\begin{aligned}\min \Omega(\mathbf{W},\mathbf{Z},\mathbf{F},\mathbf{S},\boldsymbol{\alpha}) = \|\mathbf{Z}-\mathbf{X}\mathbf{W}\mathbf{S}\|_F^2 + \sum_{v=1}^{V}\alpha_v^p tr(\mathbf{Z}^T\mathbf{L}_{22}^{(v)}\mathbf{Z}) + \lambda_1\|\mathbf{W}\|_F^2 \\ + \lambda_2(\|\mathbf{F}-\boldsymbol{\Lambda}\mathbf{R}_{12}\mathbf{Z}\|_F^2 + tr(\mathbf{F}^T\mathbf{L}_{11}\mathbf{F}) + \|\mathbf{H}\odot(\mathbf{F}-\mathbf{Y})\|_F^2) \\ s.t. \quad \alpha_v \geq 0, \boldsymbol{\alpha}^T\mathbf{1}=1\end{aligned} \tag{5}$$

3.3 Optimization

The optimization problem in Eq. (5) is non-convex with respect to $\mathbf{W}$, $\mathbf{Z}$, $\mathbf{F}$, $\mathbf{S}$ and $\boldsymbol{\alpha}$ 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) [3] to alternatively optimize one variable by fixing the other four

variables in an iterative way. The detailed procedure is presented as follows: The partial derivatives of $\Omega(\mathbf{W}, \mathbf{Z}, \mathbf{F}, \mathbf{S}, \boldsymbol{\alpha})$ with respect to $\mathbf{W}$, $\mathbf{Z}$, $\mathbf{F}$, $\mathbf{S}$ are:

$$\frac{\partial \Omega(\mathbf{W}, \mathbf{Z}, \mathbf{F}, \mathbf{S}, \boldsymbol{\alpha})}{\partial \mathbf{W}} = 2\lambda_1 \mathbf{W} - 2\mathbf{X}^T \mathbf{Z}\mathbf{S} + 2\mathbf{X}^T \mathbf{X}\mathbf{W}\mathbf{S}\mathbf{S} \tag{6}$$

$$\begin{aligned}\frac{\partial \Omega(\mathbf{W}, \mathbf{Z}, \mathbf{F}, \mathbf{S}, \boldsymbol{\alpha})}{\partial \mathbf{Z}} &= 2\mathbf{Z} - 2\mathbf{X}\mathbf{W}\mathbf{S} + 2\sum_{v=1}^{V} \alpha_v^p \mathbf{D}_{22}^{(v)} \mathbf{Z} - 2\sum_{v=1}^{V} \alpha_v^p \mathbf{R}_{22}^{(v)} \mathbf{Z} \\ &\quad - 2\lambda_2 \mathbf{R}_{12}^T \boldsymbol{\Lambda}\mathbf{F} + 2\lambda_2 \mathbf{R}_{12}^T \boldsymbol{\Lambda}\boldsymbol{\Lambda}\mathbf{R}_{12}\mathbf{Z}\end{aligned} \tag{7}$$

$$\begin{aligned}\frac{\partial \Omega(\mathbf{W}, \mathbf{Z}, \mathbf{F}, \mathbf{S}, \boldsymbol{\alpha})}{\partial \mathbf{F}} &= 2\lambda_2 \mathbf{F} - 2\lambda_2 \boldsymbol{\Lambda}\mathbf{R}_{12}\mathbf{Z} + 2\lambda_2 \mathbf{H}.*\mathbf{F}.*\mathbf{H} - 2\lambda_2 \mathbf{H}.*\mathbf{Y}.*\mathbf{H} \\ &\quad + 2\lambda_2 \mathbf{D}_{11}\mathbf{F} - 2\lambda_2 \mathbf{R}_{11}\mathbf{F}\end{aligned} \tag{8}$$

$$\frac{\partial \Omega(\mathbf{W}, \mathbf{Z}, \mathbf{F}, \mathbf{S}, \boldsymbol{\alpha})}{\partial \mathbf{S}} = -2\mathbf{W}^T \mathbf{X}^T \mathbf{Z} + 2\mathbf{W}^T \mathbf{X}^T \mathbf{X}\mathbf{W}\mathbf{S} \tag{9}$$

We can then use the Karush-Kuhn-Tucker(KKT) conditions [3] for the nonnegativity of $\mathbf{W}$, $\mathbf{Z}$, $\mathbf{F}$ and $\mathbf{S}$:

$$(\lambda_1 \mathbf{W} - \mathbf{X}^T \mathbf{Z}\mathbf{S} + \mathbf{X}^T \mathbf{X}\mathbf{W}\mathbf{S}\mathbf{S})_{ij}[\mathbf{W}]_{ij} = 0 \tag{10}$$

$$\begin{aligned}&(\mathbf{Z} - \mathbf{X}\mathbf{W}\mathbf{S} + \sum_{v=1}^{V} \alpha_v^p \mathbf{D}_{22}^{(v)} \mathbf{Z} - \sum_{v=1}^{V} \alpha_v^p \mathbf{R}_{22}^{(v)} \mathbf{Z} - \lambda_2 \mathbf{R}_{12}^T \boldsymbol{\Lambda}\mathbf{F} \\ &+ \lambda_2 \mathbf{R}_{12}^T \boldsymbol{\Lambda}\boldsymbol{\Lambda}\mathbf{R}_{12}\mathbf{Z})_{ij}[\mathbf{Z}]_{ij} = 0\end{aligned} \tag{11}$$

$$(\mathbf{F} - \boldsymbol{\Lambda}\mathbf{R}_{12}\mathbf{Z} + \mathbf{H}.*\mathbf{F}.*\mathbf{H} - \mathbf{H}.*\mathbf{Y}.*\mathbf{H} + \mathbf{D}_{11}\mathbf{F} - \mathbf{R}_{11}\mathbf{F})_{ij}[\mathbf{F}]_{ij} = 0 \tag{12}$$

$$(-\mathbf{Z} + \mathbf{X}\mathbf{W}\mathbf{S})_{ij}[\mathbf{S}]_{ij} = 0 \tag{13}$$

These nonnegative constraints give the fixed point relationship that the solution must satisfy. As such, we can update $\mathbf{W}$, $\mathbf{Z}$, $\mathbf{F}$ and $\mathbf{S}$ using the following update rules:

$$[\mathbf{W}]_{ij} = [\mathbf{W}]_{ij} \frac{\mathbf{X}^T \mathbf{Z}\mathbf{S}}{\lambda_1 \mathbf{W} + \mathbf{X}^T \mathbf{X}\mathbf{W}\mathbf{S}\mathbf{S}} \tag{14}$$

$$[\mathbf{Z}]_{ij} = [\mathbf{Z}]_{ij} \frac{\mathbf{XWS} + \sum_{v=1}^{V} \alpha_v^p \mathbf{R}_{22}^{(v)} \mathbf{Z} + \lambda_2 \mathbf{R}_{12}^T \mathbf{\Lambda F}}{\mathbf{Z} + \sum_{v=1}^{V} \alpha_v^p \mathbf{D}_{22}^{(v)} \mathbf{Z} + \lambda_2 \mathbf{R}_{12}^T \mathbf{\Lambda \Lambda R}_{12} \mathbf{Z}} \tag{15}$$

$$[\mathbf{F}]_{ij} = [\mathbf{F}]_{ij} \frac{\mathbf{\Lambda R}_{12} \mathbf{Z} + \mathbf{H}. * \mathbf{Y}. * \mathbf{H} + \mathbf{R}_{11} \mathbf{F}}{\mathbf{F} + \mathbf{H}. * \mathbf{F}. * \mathbf{H} + \mathbf{D}_{11} \mathbf{F}} \tag{16}$$

$$[\mathbf{S}]_{ij} = [\mathbf{S}]_{ij} \frac{\mathbf{Z}}{\mathbf{XWS}} \tag{17}$$

When $\mathbf{W}$, $\mathbf{Z}$, $\mathbf{F}$ and $\mathbf{S}$ are fixed, Eq. (5) is equivalent as:

$$\begin{aligned} \arg\min_{\boldsymbol{\alpha}} \lambda \sum_{v=1}^{V} \alpha_v^p tr(\mathbf{Z}^T \mathbf{L}_{22}^{(t)} \mathbf{Z}) \\ s.t. \sum_{v=1}^{V} \alpha_v = 1 \end{aligned} \tag{18}$$

Here, we adopt the Lagrange multiplier method to optimizing $\boldsymbol{\alpha}$:

$$H(\mathbf{Z}, \boldsymbol{\alpha}, \eta) = \lambda \sum_{v=1}^{V} \alpha_v^p tr(\mathbf{Z}^T \mathbf{L}_{22}^{(t)} \mathbf{Z}) - \eta(\sum_{v=1}^{V} \alpha_v - 1) \tag{19}$$

where η is the Lagrange multiplier. We can take the partial derivative of $H(\mathbf{Z}, \boldsymbol{\alpha}, \eta)$ respect to α_v and set it to 0 as follows:

$$\frac{\partial \Omega(\mathbf{W}, \mathbf{Z}, \mathbf{F}, \mathbf{S}, \boldsymbol{\alpha})}{\partial \alpha_v} = \lambda p \alpha_v^{p-1} tr(\mathbf{Z}^T \mathbf{L}_{22}^{(v)} \mathbf{Z}) - \eta = 0 \tag{20}$$

$$\alpha_v = \left(\frac{\eta}{\lambda p tr\left(\mathbf{Z}^T \mathbf{L}_{22}^{(v)} \mathbf{Z} \right)} \right)^{\frac{1}{p-1}} \tag{21}$$

Since $\sum_{v=1}^{V} \alpha_v = 1$ we can obtain:

$$\alpha_v = \frac{\left(\frac{1}{tr\left(\mathbf{Z}^T \mathbf{L}_{(22)}^{(v)} \mathbf{Z} \right)} \right)^{\frac{1}{p-1}}}{\sum_{v=1}^{V} \left(\frac{1}{tr\left(\mathbf{Z}^T \mathbf{L}_{22}^{(v)} \mathbf{Z} \right)} \right)^{\frac{1}{p-1}}} \tag{22}$$

By iteratively updating $\mathbf{W}$, $\mathbf{Z}$, $\mathbf{F}$ and $\mathbf{S}$ and $\boldsymbol{\alpha}$ via Eq. (14), Eq. (15), Eq. (16), Eq. (17) and Eq. (22), we can obtain the local optimal of $\mathbf{W}$, $\mathbf{Z}$, $\mathbf{F}$ and $\mathbf{S}$ and $\boldsymbol{\alpha}$. **Algorithm 1** lists the above optimization procedure, and IsoDA often converges in 50 iterations on our used dataset.

Algorithm 1 IsoDA: Isoform-disease association prediction by multi-omics data fusion

Input: $\mathbf{X}$, $\mathbf{R}_{11}$, $\mathbf{\Lambda}$, $\mathbf{R}_{12}$, $\{\mathbf{R}_{22}^v\}_{v=1}^V$, $\mathbf{Y}$, p, λ_1, λ_2, $maxIter$, tol.
Output: $\mathbf{W}$, $\mathbf{Z}$, $\mathbf{F}$, $\mathbf{S}$, $\boldsymbol{\alpha}$.

Initialize $\boldsymbol{\alpha}_V = 1/V$, $iter = 1$, $tol = 10^{-2}$, $maxiIter = 60$
Initialize $\mathbf{W}$ randomly
specify $\mathbf{S}$ as the disease-disease correlation matrix by cosine similarity
Initialize $\mathbf{F} = \mathbf{Y}$
Initialize $\mathbf{Z} = \mathbf{R}_{12}^T\mathbf{F}$
$loss^{iter} = \Omega(\mathbf{W}, \mathbf{Z}, \mathbf{F}, \mathbf{S}, \boldsymbol{\alpha})$
While $iter < maxIter$ and $|\delta| > tol$
    Update $\mathbf{W}$ using Eq. (14)
    Update $\mathbf{Z}$ using Eq. (15)
    Update $\mathbf{F}$ using Eq. (16)
    Update $\mathbf{S}$ using Eq. (17)
    Update $\boldsymbol{\alpha}$ using Eq. (22)
    $loss^{iter+1} = \Omega(\mathbf{W}, \mathbf{Z}, \mathbf{F}, \mathbf{S}, \boldsymbol{\alpha})$
    $\delta \leftarrow loss^{iter+1} - loss^{iter}$
    $iter = iter + 1$
End While

3.4 Isoform/gene disease associations prediction

Suppose $\mathbf{Z}^*$ is the optimized variable. However, the disease associations of isoforms are generally unknown. To enable a surrogate evaluation, we need to aggregate the isoform-disease associaitons to the gene-level. For this surrogate evaluation, we recall Eq. (23) to approximate the gene-disease associations matrix as follows:

$$\mathbf{Y}^* = \mathbf{\Lambda}\mathbf{R}_{12}\mathbf{Z}^* \tag{23}$$

4 Experiment results and analysis

4.1 Experimental setup

In our article, we collect multiple RNA-Seq datasets from ENCODE project, gene-disease associations data from DisGeNET, gene interaction data from BioGrid, sequence data of isoforms from NCBI for assessing the performance of IsoDA for predicting IDAs. The pre-processed GDAs and isoforms of the genes are listed in Table 1.

Table 1. Statistics of genes, isoforms and GDAs for experiments.

genes(n)	isoforms(m)	diseases(c)	GDAs
12,371	26,866	3,883	673,046

To comparatively study the performance of IsoDA, we take the state-of-the-art isoform function prediction methods (iMILP [23], IsoFun [56], Disofun [50]) and two gene-disease association prediction methods (PRINCE [48], KnowGENE [57]) as compared methods. The input parameters of these comparing methods are fixed/optimized as the original papers or shared codes. For IsoDA, we choose λ_1 and λ_2 in $\{10^{-4}, 10^{-3}, \ldots, 10^3, 10^4\}$, $p = 2$. 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 [23, 56]. In addition, we compare IsoDA against its degenerated variants to study the contribution components of IsoDA. We further use isoforms with collected isoform-disease to prove the effectiveness of IsoDA.

We adopt six evaluation metrics $MicroF1$, $MacroF1$, $1-RankLoss$, $Fmax$, $AUPRC$ and $AUROC$ to evaluate the performance of IsoDA. $MicroF1$ 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. $AUROC$ computes the area under receiver operating curve of each term at first, and then takes the average value of these areas to quantify the overall performance. The higher the value of $MicroF1$, $MacroF1$, $1 - RankLoss$, $Fmax$, $AUPRC$ and $AUROC$, the better the performance is. We want to remark that these six metrics quantify the prediction results from different aspects, and it is difficult for one method to always outperform another one across all these metrics.

4.2 Results evaluation at gene-level

We adopt five-fold cross-validation at the gene-level for experiment. 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 IsoDA and of compared methods. IsoDA achieves better results than the compared methods across all the six evaluation metrics. $MicroF1$, $MacroF1$, $AUPRC$ and $AUROC$ are disease-centric metrics, while $1 - Rankloss$ and $Fmax$ are gene/isoform-centric metrics. These results confirm clearly that IsoDA can more accurately predict GDAs (IDAs) from both the genes (isoforms) and DO term perspectives. IsoDA takes tissue specificity and isoform sequence data into account, and fuses multiple isoform functional association networks constructed from RNA-seq datasets and isoform sequence data with adaptive weights. In contrast, iMILP only fuses functional association networks derived from RNA-seq datasets without adaptive weights. IsoFun and DisoFun concatenate the isoform expression profiles of different tissues into a single feature vector and ignore the tissue specificity. For these reasons, IsoDA gives better results than these multi-instance learning based isoform function prediction methods. IsoDA, IsoFun and DisoFun all

incorporate the important PPI data to complete GDAs, but IsoDA adds an additional indicator matrix $\mathbf{H}$ to separately model the seen GDAs and unseen ones. As a result, the optimized GDAs being consistent with the collected ones and IsoDA is less biased toward seen ones than IsoFun and DisoFun. In addition, the completed GDAs can be distributed to individual isoforms, and thus boost the prediction of IDAs. These advantages will be further confirmed by ablation study.

We also compare the performance of IsoDA with two GDAs prediction methods (PRINCE [48] and Know-GENE [57]). PRINCE uses a network propagation strategy to predict causal genes and protein complexes that are involved in a disease of interest. Know-GENE firstly quantifies gene-gene mutual information using the co-occurrence of genes in GDAs data and then combines the mutual information with PPI networks via a boosted tree regression method to predict GDAs. Compared with PRINCE, Know-GENE makes a better use of GDAs, it integrates gene-gene mutual information calculated from GDAs and the available PPIs to predict GDAs in a knowledge driven way, so Know-GENE outperforms PRINCE by a large margin. For the similar reasons, the performance margin between IsoDA and Know-GENE is smaller than those between IsoDA and other compared methods. From Table 2, we can observe that some isoform function prediction methods sometimes lose to the two GDAs prediction methods. This is because isoform-level methods more focus on utilizing transcriptomics data, while the surrogate evaluations are made at the aggregated gene-level, instead of the target isoform-level. We want to highlight that our IsoDA is an inductive approach that can directly predict the associations between diseases and a new isoform, whereas these compared methods are transductive solutions, they need this new isoform being included for training before the prediction.

We further applied signed-rank test [10] to compare the results of IsoDA against those of compared methods across the six evaluation metrics, all the p-values are smaller than 0.0313. In summary, these results indicate the effectiveness of IsoDA in identifying IDAs.

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

	PRINCE	Know-GENE	iMILP	IsoFun	Disofun	IsoDA
MicroF1	0.2146±0.0253•	0.4804±0.0075•	0.1954±0.0148•	0.2561±0.0205•	0.2879±0.0158•	0.6739±0.0076
MacroF1	0.1759±0.0204•	0.2930±0.0142•	0.0782±0.0194•	0.1152±0.0238•	0.1035±0.0115•	0.3187±0.0132
1-RankLoss	0.8243±0.0262•	0.9355±0.0133•	0.3764±0.0291•	0.7268±0.0109•	0.8914±0.0037•	0.9465±0.0022
Fmax	0.2854±0.0186•	0.3269±0.0061•	0.1492±0.0163•	0.2267±0.0162•	0.2450±0.0128•	0.5437±0.0059
AUPRC	0.3053±0.0197•	0.3741±0.0084•	0.0152±0.0034•	0.0745±0.0118•	0.0806±0.0065•	0.3816±0.0087
AUROC	0.5846±0.0139•	0.6365±0.0068•	0.5149±0.0095•	0.6077±0.0060•	0.6223±0.0082•	0.6471±0.0046

4.3 Results evaluation at isoform-level

In this subsection, we further assess the performance of IsoDA at the isoform-level. Due to the lack of ground-truth isoform-disease associations, we take 5568 single-isoform genes, each of which produces only one isoform within our used dataset as the testbed, and take the rest genes and isoforms as the training set. We follow the same setting as previous experiments and report the results in Figure 1. PRINCE and Know-GENE do not consider the gene-isoform relations, they can only predict the associations between genes and diseases, so their results are excluded.

IsoDA again achieves a better performance than three compared methods (iMILP, IsoFun and Disofun) at the isoform-level disease associations prediction. iMILP universally distributes GDAs to all isoforms of a gene, then only propagates IDAs on the isoform co-expression network, so it has the lowest performance. IsoFun and Disofun leverage the protein interaction data alike IsoDA, but they do not consider the tissue specificity of multiple RNA-seq datasets and isoform sequence data, so they both lose to IsoDA. By referring to Table 2 and Figure 1, we can conclude that IsoDA is indeed effective in fusing genomics, transcriptomics and proteomic data to handle the multiplicity of predicting IDAs at the isoform-level.

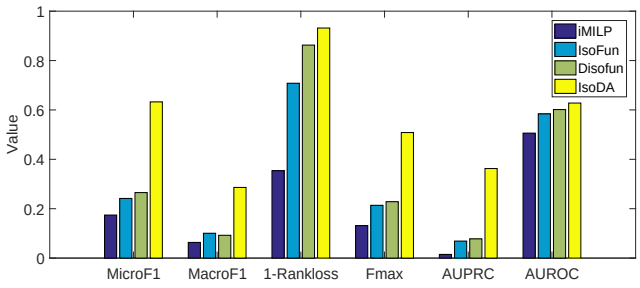

Fig. 1. Performance results of IsoDA and of compared methods on predicting IDAs, whose isoforms spliced from single-isoform genes.

4.4 Case study

To further explore the reliability of IsoDA in predicting IDAs, we collect some IDAs from Pubmed literature [25,37,45]. APOE2, APOE3 and APOE4 are three alternatively spliced isoforms of APOE, and Alzheimer disease (AD) is associated with different bindings with APOE [45]. It is recognized that the expression of APOE4 increases the risk of AD, while APOE2 decreases the risk. Accumulated evidences suggest the detrimental effect of APOE4, and APOE2 protects

against AD through both amyloid-β (Aβ)-dependent and independent mechanisms [25]. APOE performs neuroprotective and neurotrophic functions in the normal, aging brain, while APOE2 and APOE3 execute these functions more efficiently than APOE4. Therefore, individuals without APOE2 or APOE3 are at the risk for AD [37]. We report the prediction results of IsoDA and three compared methods (iMILP, IsoFun and Disofun) with respect to APOE in Table 3. We observe that IsoDA correctly differentiates individual IDAs of APOE, while iMILP incorrectly predicts the association between APOE3 and AD, iMILP and IsoFun wrongly predict associations between APOE4 and AD. Particularly, IsoDA predicts APOE2 less positively associated with AD than APOE3, and this fact also agrees with the founding that APOE2 can more decrease the risk of AD than APOE3 [25].

Table 3. The isoform-disease associations of APOE (Apolipoprotein E) and VEGFA (Vascular Endothelial Growth Factor A). $\surd$ indicates the disease known (or predicted) being associated with the isoform, and $\times$ means the opposite. When the predicted association probability between a disease and an isoform is in the top 3 of the total isoforms of the gene, this isoform is associated with this disease.

Gene	Isoform	Disease	Association	iMILP	IsoFun	Disofun	IsoDA
APOE	APOE2	Alzheimer's disease	$\times$	$\times$	$\times$	$\times$	$\times$
	APOE3		$\times$	$\surd$	$\times$	$\times$	$\times$
	APOE4		$\surd$	$\times$	$\times$	$\surd$	$\surd$
VEGFA	$VEGFA_{121}$	Coronary heart disease	$\times$	$\surd$	$\surd$	$\times$	$\times$
	$VEGFA_{165}b$		$\surd$	$\times$	$\surd$	$\times$	$\surd$
Accuracy				20%	60%	80%	100%

We further investigate IDAs of the gene VEGFA (Vascular endothelial growth factor A) with respect to CHD (Coronary Heart Disease). VEGFA undergoes extensive alternative splicing, and encodes isoforms with both angiogenic and anti-angiogenic potential through the differential use of an alternative splice site with exon 8 [36]. Some researches [20, 22, 36, 41] found that two isoforms ($VEGFA_{165}b$ and $VEGFA_{165}b$) of VEGFA exert the opposite effects of antiangiogenesis and proangiogenesis. The antiangiogenic isoform $VEGFA_{165}b$ was found to be associated with CHD [22]. From the results in Table 3, we can find that IsoDA more credibly predicts the individual associations between CHD and isoforms spliced from the same gene.

Based on these case results, we can conclude that IsoDA has the potential to accurately identify IDAs of isoforms spliced from the same gene.

4.5 Ablation study

To further investigate the contribution components, we design seven variants of IsoDA, which are configured as follows:
(i) IsoDA(nS) removes the disease correlation matrix $\mathbf{S}$ from $\|\mathbf{Z} - \mathbf{XWS}\|_F^2$ in Eq. (5), that means disease-disease associations network is disregarded.
(ii) IsoDA(cS) uses the initial disease correlation matrix $\mathbf{S}$ without update in

the iterative optimization process.
(iii) IsoDA(RNA) only uses the isoform expression data derived from multiple RNA-seq datasets.
(iv) IsoDA(Seq) only utilizes the isoform sequence data.
(v) IsoDA(RNA+Seq) concatenates the isoform expression profile feature vectors of different tissues and the isoform sequence feature vectors into a single one, and then directly constructs a single isoform functional association network using cosine similarity.
(vi) IsoDA(nα) integrates multiple isoform functional association networks with equal weight.
(vii) IsoDA(nH) removes the indicator matrix $\mathbf{H}$ in $\|\mathbf{H} \odot (\mathbf{F} - \mathbf{Y})\|_F^2$) in Eq. (5), say it does not consider the bias toward the observed GDAs.
All the other configurations of these variants are kept the same as IsoDA, unless extra specified. Figure 2, Figure 3 and Figure 4 reports the performance results of IsoDA and its variants. The experimental settings are the same as the evaluation at the gene-level and we can easily observe that IsoDA achieves a better performance than its variants.

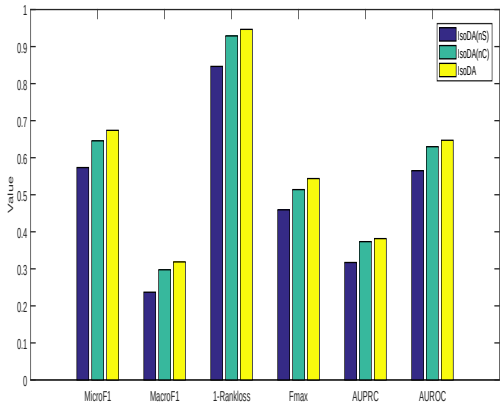

Fig. 2. Performance results of IsoDA(nS), IsoDA(cS) and IsoDA. IsoDA(nS) disregards the disease correlations, and IsoDA(cS) directly uses disease correlations estimated from initial GDAs.

In Fig. 2, IsoDA(nS) has clear lower performance values than IsoDA, which considers the disease correlations. This facts that exploring latent correlations between diseases can boost the performance of IDAs. IsoDA(cS) incorporates the estimated disease correlation $\mathbf{S}$ but does not iteratively refine $\mathbf{S}$, it gives a better performance than IsoDA(nS) but loses to IsoDA. This trend not only confirms the necessity of incorporating the disease correlations, but also the necessity

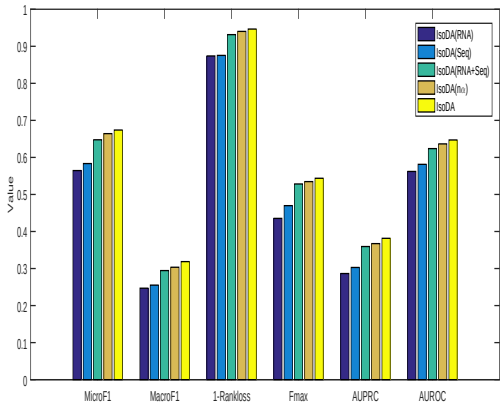

Fig. 3. Performance results of IsoDA(RNA), IsoDA(Seq), IsoDA(RNA+Seq), IsoDA(nα) and IsoDA. IsoDA(RNA) only uses RNA-seq datasets, IsoDA(Seq) only uses the isoform nucleotide data, while IsoDA(RNA+Seq) combines these two types of data into a single feature vector. IsoDA(nα) integrates multiple isoform functional association networks with equal weights.

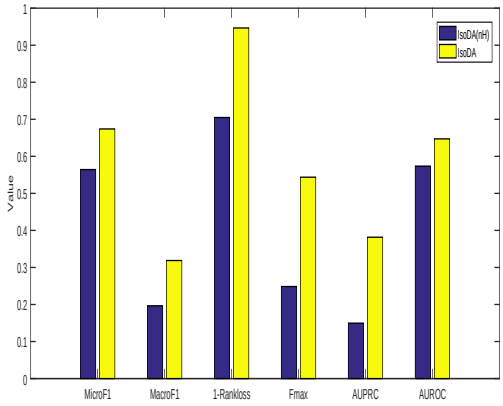

Fig. 4. Performance results of IsoDA(nH) and IsoDA. IsoDA(nH) implicitly assumes the observed GDAs are complete.

of refining the coarse disease correlations (estimated from known GDAs) using additional biological data.

In Fig. 3, IsoDA(RNA) loses to IsoDA(Seq), which tells the isoform sequence data makes more contributions for identifying IDAs(GDAs) than RNA-seq datasets, since the sequence data include important functional sites and domains of isoforms. Both IsoDA(RNA) and IsoDA(Seq) have lower performance values than IsoDA(RNA+Seq), and not to mention IsoDA. This fact shows that fusing RNA-seq data and sequence data can boost the prediction of IDAs(GDAs). IsoDA(RNA+Seq) has lower performance values than IsoDA(nα), which considers the tissue specicity of alternative splicing and combines isoform functional association networks with equal weight from tissue-wise. This pattern proves the necessity of combining multiple isoform functional associations networks from tissue-level. However, IsoDA(nα) gives lower performance values than IsoDA, which not only considers the tissue specificity but also integrates multiple isoform functional association networks with adaptive weights. This contrast supports the effectiveness of adaptive weights and rationality of combining multi-omics data.

From Figure 4, we can see that IsoDA has an obvious improvement to IsoDA(nH), which implicitly assumes the observed GDAs are complete. In contrast, IsoDA considers the incompleteness of observed GDAs and enforces latent IDAs being consistent with the collected ones. At the same time, it differentiates the currently observed associations from the unobserved (potential) ones. As a result, IsoDA is less biased toward observed ones. In practice, the incomplete associations are implicitly ignored by most GDAs prediction methods and isoform function prediction methods. As a result, these compared methods and IsoDA(H) lose to IsoDA.

In conclusion, the ablation study confirms the effectiveness of IsoDA on fusing genomics, transcriptomics and proteomic data to more accurately predict IDAs. It also supports the importance to specifically consider the incomplete GDAs. IsoDA models these important factors and thus obtains clearly better performance results than these variants.

4.6 Parameter sensitivity analysis

There are two input parameters (λ_1 and λ_2) in IsoDA. λ_1 controls the complexity of the linear predictor, and λ_2 is a balance factor for the information sources from gene-level to isoform-level. We vary λ_1 and λ_2 in $\left\{10^{-4}, 10^{-3}, \cdots, 10^{3}, 10^{4}\right\}$ and present the results of IsoDA under different combinations of λ_1 and λ_2 in Figure 5.

We observe that IsoDA increases clearly in $Fmax$, $AUPRC$, $AUROC$ when λ_1 growing from 10^{-4} to 10^{-2}, and then the performance of IsoDA decreases with a slight trend as λ_1 further growing. Alike λ_1, IsoDA firstly has an obviously increased performance as λ_2 growing from 10^{-4} to 10^{3} and then decreases a little as λ_2 growing to 10^{4}. These results confirm that it is important to fuse the gene-level data and the gene-isoform relations for predicting IDAs. Meanwhile, we find that λ_2 is more positively related with the performance of IsoDA than

λ_1. The reason is that λ_1 only controls the complexity of the multi-label linear predictor, but λ_2 balances the information sources from gene-level to isoform-level, which plays a more important role in fusing genomics, transcriptomics and proteomics data to improve the performance of predicting IDAs. Moreover, when both λ_1 and λ_2 are fixed to too small values, IsoDA shows the lowest performance, this phenomenon expresses the superiority of the unified objective function for predicting IDAs, it also corroborates the necessity of fusing multi-omics data. Based on above analysis, we adopt $\lambda_1 = 10^{-2}$ and $\lambda_2 = 10^3$ for experiments.

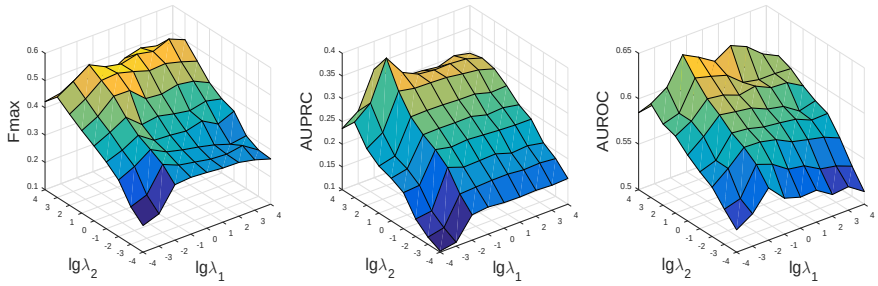

Fig. 5. Performance results of IsoDA under different input values of λ_1 and λ_2.

5 Conclusion

In this paper, we study how to computationally identify isoform-disease associations, which is an interesting, important but largely unexplored topic that can uncover the disease pathology at a deeper level than the well-studied gene-disease associations analysis. Our proposed approach IsoDA leverages genome, transcriptome and proteome data and multi-instance learning to bypass the lack of isoform-disease associations and to distribute gene-disease associations to individual isoforms. IsoDA considers the incompleteness of available GDAs and incorporates PPI data and indicator matrix to complete GDAs. It further takes into account the tissue specificity of alternative splicing and adaptively combines multiple isoform functional association networks induced from multiple RNA-seq datasets at the tissue-level. IsoDA performs significantly better than related competitive methods that target to identify gene-disease associations or isoform functions.

6 Acknowledgements

This research is supported by National Natural Science Foundation of China (61872300 and 62031003).

References

1. Andrews, S., Tsochantaridis, I., Hofmann, T.: Support vector machines for multiple-instance learning. In: Neural Information Processing Systems. pp. 577–584 (2003)
2. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al.: Gene ontology: tool for the unification of biology. Nature Genetics **25**(1), 25 (2000)
3. Boyd, S., Vandenberghe, L.: Convex optimization. Cambridge University Press (2004)
4. 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)
5. Chatr-Aryamontri, A., Oughtred, R., Boucher, L., Rust, J., Chang, C., Kolas, N.K., O'Donnell, L., Oster, S., Theesfeld, C., Sellam, A., et al.: The biogrid interaction database: 2017 update. Nucleic Acids Research **45**(D1), D369–D379 (2017)
6. 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)
7. 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)
8. Consortium, E.P., et al.: An integrated encyclopedia of dna elements in the human genome. Nature **489**(7414), 57 (2012)
9. Defer, N., Best-Belpomme, M., Hanoune, J.: Tissue specificity and physiological relevance of various isoforms of adenylyl cyclase. American Journal of Physiology-Renal Physiology **279**(3), F400–F416 (2000)
10. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research **7**, 130 (2006)
11. 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)
12. 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)
13. Frasca, M.: Gene2disco: Gene to disease using disease commonalities. Artificial Intelligence in Medicine **82**, 34–46 (2017)
14. Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A., McKusick, V.A.: Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. Nucleic Acids Research **33**(S1), D514–D517 (2005)
15. 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)
16. Huang, Q., Wang, J., Zhang, X., Yu, G.: Isoform-disease association prediction by data fusion. In: International Symposium on Bioinformatics Research and Applications. pp. 44–55 (2020)
17. Jiang, R.: Walking on multiple disease-gene networks to prioritize candidate genes. Journal of Molecular Cell Biology **7**(3), 214–230 (2015)

18. Kandoi, G., Dickerson, J.A.: Tissue-specific mouse mrna isoform networks. Scientific Reports **9**(1), 1–24 (2019)
19. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., Morishima, K.: Kegg: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Research **45**(D1), D353–D361 (2016)
20. Kikuchi, R., Nakamura, K., MacLauchlan, S., Ngo, D.T.M., Shimizu, I., Fuster, J.J., Katanasaka, Y., Yoshida, S., Qiu, Y., Yamaguchi, T.P., et al.: An antiangiogenic isoform of vegf-a contributes to impaired vascularization in peripheral artery disease. Nature Medicine **20**(12), 1464–1471 (2014)
21. Kim, D., Langmead, B., Salzberg, S.L.: Hisat: a fast spliced aligner with low memory requirements. Nature Methods **12**(4), 357 (2015)
22. Latorre, E., Pilling, L.C., Lee, B.P., Bandinelli, S., Melzer, D., Ferrucci, L., Harries, L.W.: The vegfa156b isoform is dysregulated in senescent endothelial cells and may be associated with prevalent and incident coronary heart disease. Clinical Science **132**(3), 313–325 (2018)
23. 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)
24. 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)
25. Li, Z., Shue, F., Zhao, N., Shinohara, M., Bu, G.: Apoe2: protective mechanism and therapeutic implications for alzheimers disease. Molecular Neurodegeneration **15**(1), 1–19 (2020)
26. 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)
27. Luo, T., Zhang, W., Qiu, S., Yang, Y., Yi, D., Wang, G., Ye, J., Wang, J.: Functional annotation of human protein coding isoforms via non-convex multi-instance learning. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 345–354 (2017)
28. Maron, O., Lozano-Pérez, T.: A framework for multiple-instance learning. In: Neural Information Processing Systems. pp. 570–576 (1998)
29. Natarajan, N., Dhillon, I.S.: Inductive matrix completion for predicting gene–disease associations. Bioinformatics **30**(12), i60–i68 (2014)
30. 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)
31. 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)
32. 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)
33. 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)
34. 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)

35. Qian, Y., Besenbacher, S., Mailund, T., Schierup, M.H.: Identifying disease associated genes by network propagation. BMC Systems Biology **8**(S1), S6 (2014)
36. Qiu, Y., Hoareau-Aveilla, C., Oltean, S., Harper, S.J., Bates, D.O.: The anti-angiogenic isoforms of vegf in health and disease. Biochemical Society transactions **37**(6), 1207 (2009)
37. Rebeck, G.W., Kindy, M., LaDu, M.J.: Apolipoprotein e and alzheimer's disease: the protective effects of apoe2 and e3. Journal of Alzheimer's Disease **4**(3), 145–154 (2002)
38. 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)
39. Shaw, D., Chen, H., Jiang, T.: Deepisofun: a deep domain adaptation approach to predict isoform functions. Bioinformatics **35**(15), 2535–2544 (2019)
40. 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)
41. Shiying, W., Boyun, S., Jianye, Y., Wanjun, Z., Ping, T., Jiang, L., Hongyi, H.: The different effects of vegfa121 and vegfa165 on regulating angiogenesis depend on phosphorylation sites of vegfr2. Inflammatory Bowel Diseases **23**(4), 603–616 (2017)
42. 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)
43. Smith, L.M., Kelleher, N.L.: Proteoforms as the next proteomics currency. Science **359**(6380), 1106–1107 (2018)
44. Stark, C., Breitkreutz, B.J., Reguly, T., Boucher, L., Breitkreutz, A., Tyers, M.: Biogrid: a general repository for interaction datasets. Nucleic Acids Research **34**(suppl_1), D535–D539 (2006)
45. 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)
46. 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)
47. Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K.P., et al.: String v10: protein–protein interaction networks, integrated over the tree of life. Nucleic Acids Research **43**(D1), D447–D452 (2015)
48. 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)
49. 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)
50. Wang, K., Wang, J., Domeniconi, C., Zhang, X., Yu, G.: Differentiating isoform functions with collaborative matrix factorization. Bioinformatics **36**(6), 1864–1871 (2020)
51. Wang, X., Gong, Y., Yi, J., Zhang, W.: Predicting gene-disease associations from the heterogeneous network using graph embedding. In: IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 504–511 (2019)

52. Wang, X., Gulbahce, N., Yu, H.: Network-based methods for human disease gene prediction. Briefings in Functional Genomics **10**(5), 280–293 (2011)
53. Wang, Z., Gerstein, M., Snyder, M.: Rna-seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics **10**(1), 57 (2009)
54. Wu, G., Feng, X., Stein, L.: A human functional protein interaction network and its application to cancer data analysis. Genome Biology **11**(5), R53 (2010)
55. Yang, K., Wang, R., Liu, G., Shu, Z., Wang, N., Zhang, R., Yu, J., Chen, J., Li, X., Zhou, X.: Hergepred: heterogeneous network embedding representation for disease gene prediction. IEEE Journal of Biomedical and Health Informatics **23**(4), 1805–1815 (2018)
56. 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)
57. Zhou, H., Skolnick, J.: A knowledge-based approach for predicting gene–disease associations. Bioinformatics **32**(18), 2831–2838 (2016)
58. Zhou, Z.H., Zhang, M.L., Huang, S.J., Li, Y.F.: Multi-instance multi-label learning. Artificial Intelligence **176**(1), 2291–2320 (2012)