Tissue Specificity based Isoform Function Prediction

Guoxian Yu, Qiuyue Huang, Xiangliang Zhang, Maozu Guo and Jun Wang

Abstract—Alternative splicing enables a gene spliced into different isoforms and hence protein variants. Identifying individual functions of these isoforms help deciphering the functional diversity of proteins. Although much efforts have been made for automatic gene function prediction, few efforts have been moved toward computational isoform function prediction, mainly due to the unavailable (or scanty) functional annotations of isoforms. Existing efforts directly combine multiple RNA-seq datasets without account of the important tissue specificity of alternative splicing. To bridge this gap, we introduce a novel approach called TS-Isofun to predict the functions of isoforms by integrating multiple functional association networks with respect to tissue specificity. TS-Isofun firstly constructs tissue-specific isoform functional association networks using multiple RNA-seq datasets from tissue-wise. Next, TS-Isofun assigns weights to these networks and models the tissue specificity by selectively integrating them with adaptive weights. It then introduces a joint matrix factorization-based data fusion model to leverage the integrated network, gene-level data and functional annotations of genes to infer the functions of isoforms. To achieve coherent weight assignment and isoform function prediction, TS-Isofun jointly optimizes the weights of individual networks and the isoform function prediction in a unified objective function. Experimental results show that TS-Isofun significantly outperforms state-of-the-art methods and the account of tissue specificity contributes to more accurate isoform function prediction.

Keywords—Isoform function prediction, Tissue specificity, Joint matrix factorization, Data fusion, Alternative splicing

1 INTRODUCTION

Functional annotations of human genes are crucial importance for understanding biological mechanisms of living body and various genetic diseases [1], [2]. Predicting the functions of genes has been studied for decades and diverse machine learning methods have been developed [3], [4], such as Support Vector Machine (SVM) [5], network-based label propagation [6], [7], deep

neural networks [8], [9], data fusion models [10]–[14], and so on. Existing computational function prediction solutions still focus on the gene-level and collectively assign all functions of gene products to the same gene. In practice, a gene can execute diverse functions mainly caused by the isoforms alternatively spliced from the same gene, these isoforms and their translated proteoforms are actual executors for diverse biological functions [15]. It is recognized that > 90% human multi-exon genes undergo alternative splicing, which significantly increases the transcriptome and proteome complexity [16]. The proteoforms (or protein variants) translated from different isoforms of the same gene have different amino acid sequences and structures, and thus may have different (even opposite) functions [17]. For example, ErbB4 can be alternatively spliced into two juxtamembranes (JM-a and JM-b), JM-a demonstrates enhanced autophosphorylation, growth, and survival. In contrast, over-expressed JM-b undergoes starvation-induced death [18]. Increasing studies show that alternative splicing plays key roles in developmental abnormality and is closely related with many human diseases [19], [20]. Therefore, identifying functional annotations of isoforms contributes to a deeper understanding of both the molecular basis of diverse genetic diseases and the evolution of phenotype complexity [21].

Compared with the widely-studied gene/protein function prediction, how to predict the functions of isoforms is still a not well studied topic. The main bottleneck is the *lack* of functional annotation at the *isoform-level*. Existing functional databases (i.e., Gene Ontology [22] and KEGG [23]) collectively record the functional information of gene products at the gene-level. In other words, isoform-level functional annotations are not available in these databases. Due to the lack of ground-truth annotations of isoforms, typical gene function prediction methods can not be directly adopted for isoform function prediction. Given the principle that Gene Ontology (GO) follows the convention to collectively record the functions of gene products to the same gene, and each functional annotation of a gene should be triggered by at least one gene product, some efforts tailor multi-instance learning (MIL) [24], [25] to identify the functions of isoforms [26]–[32]. MIL models the scenario where each bag (gene) is composed with a set of instances

G. Yu and J. Wang are with the School of Software and Joint SDU-NTU Centre for Artificial Intelligence Research, Shandong University, Jinan 250101, China. Email: {gxyu, kingjun}@sdu.edu.cn

Q. Huang is with the College of Computer and Information Sciences, Southwest University, Chongqing 400715, Email: qiuyuehuang@email.swu.edu.cn. M. Guo is with the School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, and Beijing Key Laboratory of Intelligent Processing for Building Big Data, Beijing, 100044, China. Email: guomaozu@bucea.edu.cn

X. Zhang is with Computer, Electrical and Mathematical Science and Engineering Division, King Abdullah University of Science and Technology, Thuwal, SA. Email: xiangliang.zhang@kaust.edu.sa.

M. Guo and J. Wang are the corresponding authors, Email: kingjun@sdu.edu.cn, guomaozu@bucea.edu.cn

Manuscript received December 12, 2019; Revised January 3, 2021 and April 13, 2021; Accepted June 23, 2021

(isoforms) and the labels of a bag are derived from the labels of its instances. This scenario coincides with the convention of functional annotations of genes. Another bottleneck is the lack of interactome maps of isoforms [33], [34], and existing molecule-interaction databases (e.g., BioGRID [35]) still record the interaction between proteins at the gene-level. Thanks to the advent of RNA-seq techniques, we can obtain an unprecedented amount of isoform-level data and detect the alternative splicing events at a much deeper and larger level.

However, these MIL-based isoform function prediction methods typically concatenate the quantified expression profile features derived from multiple RNA-seq datasets of different tissues into a single feature vector [26], [28]–[32] or simply add up the isoform functional association networks of different RNA-seq datasets [27]. They do not well account for the important *tissue specificity* of alternative splicing [36]. It is recognized that tissue-specific alternative splicing remodels protein-protein interaction networks and hence diversifies the functions of proteoforms [37], [38]. As a result, such concatenation or add-up may override the important tissue specific pattern of the isoform functional association networks of different tissues and compromise the performance of isoform function prediction.

In this paper, we investigate whether it is possible to integrate multiple isoform functional association networks with tissue specificity to avoid such override and to boost the prediction performance. For this purpose, we introduce an approach called TS-Isofun, which predicts isoform function with account of tissue specificity by selectively integrating multiple networks derived from different tissues. The main idea of TS-Isofun and its schema are illustrated in Figure 1. Particularly, TS-Isofun firstly constructs five isoform-level functional association networks from five human tissues (brain, heart, liver, lung, stomach) based on the expression profile features of isoforms extracted from 298 RNA-seq datasets, and then initially combines these networks with equal weights into a composite network. Next, it introduces a joint matrix factorization based objective function to integrate the gene-interaction network with the composite network, and to replenish the functional annotations of genes, dispatch the annotations of genes to isoforms, and reversely aggregate the inferred annotations of isoforms to the hosting genes in a coherent way. To avoid the trivial weight assignment, it simultaneously optimizes the weights and function prediction in the unified objective. The empirical study shows that TS-Isofun achieves better results across various evaluation metrics than state-of-the-art isoform function prediction methods [26], [27], [30]–[32]. The ablation study confirms that the account of tissue specificity can further boost the performance of isoform function prediction.

2 RELATED WORK

The key for predicting isoform function is how to leverage gene-level functional annotations, gene-isoform

associations and isoform-level data to identify the functions of individual isoforms of the hosting gene. Recently, several multiple instance learning (MIL) [25] based approaches have been adapted for isoform function prediction [26]–[29]. In MIL, a bag is positive for a label if at least one of its instances is positively annotated with that label; on the other hand, the bag is negative for a label if all its instances are not annotated with that label. Given that, we can model a gene as a bag and isoforms spliced from the gene as instances. Eksi *et al.* [26] used multiple instance support vector machine (miSVM) [39] 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 derive an isoform-level maximum margin classifier. However, the optimal solution of miSVM is sensitive to the initial annotations of isoforms inherited from positive genes and a threshold that represents a degree of strictness for assigning labels. Given that, Luo *et al.* [28] proposed a novel approach (WLRM) to differentiate the functions of isoforms by integrating sparse simplex projection as a nonconvex sparsity-induced regularizer within the MIL framework. WLRM specifically takes the genes annotated with the function under consideration as positive bags and the genes without the function as negative bags, and then learns a mapping that embeds the original bag space to a discriminative feature space by sparse projections onto simplex, and further develops an efficient block coordinate descent algorithm to solve the resulting highly nontrivial non-convex and non-smooth optimization problem. Li *et al.* [27] proposed the instance-oriented multi-instance label propagation (iMILP) to predict functions of isoforms. iMILP firstly constructs 29 isoform functional association networks from 29 human RNA-seq datasets, each of which has at least six experiments. Next, it adopts a greedy sequential forward strategy to find the best subset of networks and combines them into a single network with equal weight. After that, it applies the instance-oriented label propagation on the combined network to update the annotations of isoforms and to find the best subset of networks in a wrapper network selection way. These MIL based methods might not be the best due to the lack of sufficient labeled training data. To alleviate this issue, Shaw *et al.* [29] proposed a deep learning based method (DeepIsoFun) that combines MIL with domain adaption to provide additional labeled training genes and to transfer the knowledge of gene functions to the prediction of isoform functions.

These aforementioned methods implicitly assume that the functional annotations of genes are complete, and only focus on multiple RNA-seq datasets. Yu *et al.* [30] recently introduced an approach (IsoFun) to predict isoform functions using a tailored label propagation on a heterogeneous network. IsoFun firstly constructs an isoform functional association network based on the expression profile values of isoforms collected from multiple RNA-seq datasets. Next, it uses the available

Fig. 1. A schema of TS-Isofun. The predicted gene function annotations data matrix $\mathbf{F}$ are factorized into three matrices (Λ , $\mathbf{R}_{12}$ and $\mathbf{Z}$). Λ is a diagonal matrix, $\Lambda(i, i) = 1/n_i$, n_i represents the number of distinct isoforms spliced from the i -th gene, $\mathbf{R}_{12}$ is the gene-isoform associations matrix, $\mathbf{Z}$ represents the predicted isoform function annotations matrix. TS-Isofun forces the predicted annotations of genes $\mathbf{F}$ being consistent with the known annotations of genes $\mathbf{Y}$, and further introduces the protein interaction network and multiple tissue-based isoform functional associations networks to further coordinate the optimization of $\mathbf{F}$ and $\mathbf{Z}$, and thus to predict isoform functions $\mathbf{Z}$.

GO annotations of genes, gene-level interactions, gene-isoform associations and GO hierarchy to construct a heterogeneous network. It then introduces a bi-random walk with clamp on the heterogeneous network to predict isoform functions. Chen *et al.* [32] presented Deep learning-based prediction of IsoForm Functions from Sequences and Expression (DIFFUSE) to predict isoform functions. DIFFUSE adopts a hybrid framework that first uses a deep neural network to predict the functions of isoforms from their genomic sequences and then refines the prediction by a conditional random field (CRF) based on co-expression relationship of isoforms. DIFFUSE further proposes an iterative semi-supervised learning algorithm to train both the DNN and CRF together to tackle the lack of annotations at the isoform-level. Wang *et al.* recently [31] proposed DisoFun to differentiate isoform functions with collaborative matrix factorization. DisoFun assumes the functional annotations of genes are aggregated from key isoforms. It collaboratively factorizes the isoform expression data matrix and gene-term association matrix (storing GO annotations) into low-rank matrices to simultaneously explore the latent key isoforms, and to identify the functions of isoforms.

However, none of the existing methods take into account the tissue specificity of alternative splicing. Each tissue has specific characteristics during alternative splicing that distinguish it from other tissues and some isoforms are abundant in one tissue, but not so (even

disappeared) in other tissue [40]. For example, among the large diversity of the AC isoforms, AC1 in neural origin, while AC5 in heart and striatum [41]. For another example, 5' AMP-activated protein kinase (AMPK), which performs different roles in human cells and has 3 subunits. (1) α , catalytic domain, has two isoforms: $\alpha1$ and $\alpha2$ which are encoded from PRKAA1 and PRKAA2; (2) β , regulatory domain, has two isoforms: $\beta1$ and $\beta2$ which are encoded from PRKAB1 and PRKAB2; (3) γ , regulatory domain, has three isoforms: $\gamma1$, $\gamma2$, and $\gamma3$ which are encoded from PRKAG1, PRKAG2, and PRKAG3. In human skeletal muscle, the preferred form is $\alpha2\beta2\gamma1$. But in human liver, the most abundant form is $\alpha1\beta2\gamma1$ [42]. In other words, two isoforms $\alpha1$ and $\alpha2$ express selectively among tissues. Therefore, the tissue specific pattern of isoform functional association networks of different tissues should be accounted for isoform function prediction.

To employ the tissue specificity for accurate isoform function prediction, TS-Isofun integrates isoform functional association networks (each is derived from RNA-seq datasets of the same tissue) with adaptive weights, along with the PPI network, to complete the functional annotations of genes and to identify the functions of individual isoforms. The experimental results show that TS-Isofun not only achieves a better performance than these related methods, but also prove the benefit to account for tissue specificity.

3 METHODOLOGY

3.1 Materials and data preprocessing

Suppose there are n genes, each gene has $n_i \geq 1$ 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 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 [27], [30], [31]. Particularly, we downloaded 596 RNA-seq runs (of total 298 samples from different tissues and conditions) of Human from the ENCODE project [43] (access date: 2018-11-10). These datasets are heterogeneous in terms of library preparation procedures and sequencing platforms. Following the pre-process done in [27], [31], for each tissue, we control the quality of these RNA-seq datasets and quantify the expression value of isoforms as follows.

(1) 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) [44], and A GTF annotation file of the same build with an option of no-novel-junction.

(2) Then, we use Stringtie(v.1.3.3b) [45] to calculate the relative abundance of the transcript as FPKM values. We separately compute the FPKM values of a total of 57,964 genes with 219,288 isoforms for each sample.

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

(4) To further control the quality of isoforms, we use known protein coding gene names to map those genes obtained in step (3). Due to the prohibitive runtimes on such a large number of isoforms and sufficient nonzero values in the expression vector are required to construct the functional association network, we re-filtered the data. Particularly, we set all FPKM values lower than 0.3 as 0, and then filtered out 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 hosted isoforms are removed also. Finally, we obtain 7,549 genes with 39,559 isoforms.

To construct the tissue-specific isoform functional association network $\mathbf{R}_{22}^{(t)} \in \mathbb{R}^{m \times m}$ (from the t -th tissue), we firstly select five representative tissues (brain, heart, liver, lung, stomach), and then separately extract tissue specific expression profile features of each tissue from the total 298 samples. After that, we construct a tissue-specific isoform-isoform functional association network based on the extracted expression profile features with cosine similarity. In this way, five tissue-specific isoform functional association networks are constructed.

We also download the ontology file and the contemporary GO annotation file of human (access date: 2019-06-02), and we directly use the available GO annotations and GO hierarchy to initialize the gene-term association matrix $\mathbf{Y} \in \mathbb{R}^{n \times c}$ between n genes and c GO terms

(functional labels). Specifically, if GO term s , or s 's descendant terms, provide(s) a positive annotation for gene i , then $\mathbf{Y}(i, s) = 1$. Otherwise, $\mathbf{Y}(i, s) = 0$.

We further collect the PPI data from BioGrid (<https://thebiogrid.org>), which is a curated biological database of protein-protein interactions, genetic interactions, chemical interactions, and post-translational modifications. 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 , and the weight of $\mathbf{R}_{11}(i, j)$ is determined by the interaction strength; $\mathbf{R}_{11}(i, j) = 0$ otherwise.

3.2 Isoform Function Prediction by Joint Matrix Factorization

The lack of functional annotations at the isoform-level prohibits the direct application of traditional gene function prediction methods to predict the functions of isoforms. Most efforts leverage the obtained gene-isoform associations $\mathbf{R}_{12}$ and functional annotations of genes $\mathbf{Y}$ within the MIL framework to identify the distinct functions of individual isoforms. Suppose $\mathbf{Z} \in \mathbb{R}^{m \times c}$ record the latent functions of m isoforms, given the known association between genes and isoforms, and motivated by the principle that the functional annotations of a gene are aggregated from the annotations of its individual isoforms, we can obtain the aggregated functions of a gene from its isoforms as follows:

$$\mathbf{F} = \mathbf{A}\mathbf{R}_{12}\mathbf{Z} \quad (1)$$

where $\mathbf{A} \in \mathbb{R}^{n \times n}$ is a diagonal matrix, $\mathbf{A}(i, i) = 1/n_i$, n_i represents the number of distinct isoforms spliced from the i -th gene. $\mathbf{F} \in \mathbb{R}^{n \times c}$ is the aggregated (approximated) the gene-term association matrix.

By substituting $\mathbf{F}$ with $\mathbf{Y}$ and taking $\mathbf{Z}$ as the to-be-predicted variable, we can reversely push the functional annotations of n genes to m isoforms and make the function prediction of isoforms. However, it is reported that functional annotations of genes are rather incomplete and biased [46]–[49], which may shroud the missing annotations of genes and lead to biased function prediction of genes and isoforms. Interacting genes (or proteins) are more likely to share the same functions [6]. Given that, we want to replenish the annotations of genes using the interactome of genes and extend Eq. (1) as follows:

$$\begin{aligned} \min H(\mathbf{F}, \mathbf{Z}) &= \|\mathbf{F} - \mathbf{A}\mathbf{R}_{12}\mathbf{Z}\|_F^2 + \sum_{i=1}^n \|\mathbf{F}_i - \mathbf{Y}_i\|_F^2 \\ &+ \frac{1}{2} \sum_{i,j=1}^n \|\mathbf{F}_i - \mathbf{F}_j\|_F^2 \mathbf{R}_{11}(i, j) \\ &= \|\mathbf{F} - \mathbf{A}\mathbf{R}_{12}\mathbf{Z}\|_F^2 + \text{tr}((\mathbf{F} - \mathbf{Y})^T(\mathbf{F} - \mathbf{Y})) \\ &+ \text{tr}(\mathbf{F}^T \mathbf{L}_{11} \mathbf{F}) \end{aligned} \quad (2)$$

where $\|\cdot\|_F^2$ is the Frobenius norm, $\text{tr}(\cdot)$ is the matrix trace operator, $\mathbf{L}_{11} = \mathbf{D}_{11} - \mathbf{R}_{11}$ is the graph Laplacian matrix, $\mathbf{D}_{11}$ is a diagonal matrix with $\mathbf{D}_{11}(i, i) =$

$\sum_{j=1}^n \mathbf{R}_{11}(i, j)$. The second term in Eq. (2) is to force the predicted annotations of genes being consistent with the known annotations of genes, and the third term is to force two genes with interaction having similar annotations.

By optimizing the above objective, we can not only complete the missing annotations of genes by gene-level interaction data, but also push the replenished and known annotations of genes to isoforms. We want to remark that the above equation still pushes the annotations to isoforms by matrix tri-factorization only, without using the readily available and important isoform-level expression data. In addition, a gene generates one or more isoforms by alternative splicing, and these isoforms have diverse expression patterns across tissues [40], [41]. Bossi *et al.* [50] also stated that if two genes are co-expressed in a tissue, then their products are quite probably to physically interact in that tissue. These observations suggest that the functional association networks of isoforms should be constructed from the tissue-level and more appropriate fusion of these networks can help to accurately identify the functions of individual isoforms. Unfortunately, existing isoform function prediction methods simply concatenate the expression profile features of isoforms derived from different tissues into a single feature vector and then construct a single functional association (or co-expression) network of isoforms [26], [28]–[32], or apply a tedious wrapper-based feature selection alike method to greedily find the best subset of networks and combine these networks with equal weight [27]. As such, these methods do not make full use of the tissue-specific patterns of multiple RNA-seq datasets to boost the performance of isoform function prediction.

Based on the above analysis, we advocate to integrate multiple functional association networks of isoforms from the tissue-wise with weights, and further extend Eq. (2) as follows:

$$\begin{aligned}
 \min H(\mathbf{F}, \mathbf{Z}, \boldsymbol{\alpha}) &= \|\mathbf{F} - \boldsymbol{\Lambda} \mathbf{R}_{12} \mathbf{Z}\|_F^2 + tr((\mathbf{F} - \mathbf{Y})^T (\mathbf{F} - \mathbf{Y})) \\
 &\quad + tr(\mathbf{F}^T \mathbf{L}_{11} \mathbf{F}) \\
 &\quad + \frac{1}{2} \lambda \sum_{t=1}^v \sum_{i,j} \alpha_t^p \|\mathbf{Z}_i - \mathbf{Z}_j\|_F^2 \mathbf{R}_{22}^{(t)}(i, j) \\
 &= \|\mathbf{F} - \boldsymbol{\Lambda} \mathbf{R}_{12} \mathbf{Z}\|_F^2 + tr((\mathbf{F} - \mathbf{Y})^T (\mathbf{F} - \mathbf{Y})) \\
 &\quad + tr(\mathbf{F}^T \mathbf{L}_{11} \mathbf{F}) + \lambda \sum_{t=1}^v \alpha_t^p tr(\mathbf{Z}^T \mathbf{L}_{22}^{(t)} \mathbf{Z}) \\
 s.t. \quad &\alpha_t \geq 0, \boldsymbol{\alpha}^T \mathbf{1} = 1
 \end{aligned} \tag{3}$$

where $\lambda > 0$ is used to balance the scale difference between the first three terms (scale $n \times n$) and the fourth term (scale $m \times m$). $\mathbf{L}_{22}^{(t)} = \mathbf{D}_{22}^{(t)} - \mathbf{R}_{22}^{(t)}$, $\mathbf{D}_{22}^{(t)}$ is a diagonal matrix with $\mathbf{D}_{22}^{(t)}(i, i) = \sum_{j=1}^n \mathbf{R}_{22}^{(t)}(i, j)$. The fourth term on the right of Eq. (3) is motivated by the observation that two isoforms with the similar expression profiles (co-expressed or function associated) are more likely to

have similar biological functions, and hence minimizing this term can enforce functional associated isoforms to be predicted with similar functional annotations. To this end, Eq. (3) leverages both the gene-level and isoform-level biological data, along with the gene-isoform associations and gene-level annotations to make isoform function prediction in a coherent manner. $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \dots, \alpha_v]$ are the weights assigned to v functional association networks, each of which is derived from multiple RNA-seq datasets from the same tissue. As such, TS-Isofun can respect the specific alternative splicing pattern of each tissue and fuse these networks from tissue perspective to further boost the performance of isoform function prediction.

3.3 Optimization

Eq. (3) has three variables $\mathbf{F}$, $\mathbf{Z}$ and $\boldsymbol{\alpha}$, the solution of a single variable depends on those of other two. Following the principle of alternating direction method of multipliers [51], we alternatively optimize one of $\mathbf{F}$, $\mathbf{Z}$ and $\boldsymbol{\alpha}$, while fixing the other two as constants in an iterative manner. The detailed procedure is presented as follows.

The partial derivatives of $H(\mathbf{F}, \mathbf{Z}, \boldsymbol{\alpha})$ with respect to $\mathbf{F}$ and $\mathbf{Z}$ are:

$$\frac{\partial H(\mathbf{F}, \mathbf{Z}, \boldsymbol{\alpha})}{\partial \mathbf{F}} = 4\mathbf{F} - 2\boldsymbol{\Lambda} \mathbf{R}_{12} \mathbf{Z} + 2(\mathbf{D}_{11} - \mathbf{R}_{11})\mathbf{F} - 2\mathbf{Y} \tag{4}$$

$$\begin{aligned}
 \frac{\partial H(\mathbf{F}, \mathbf{Z}, \boldsymbol{\alpha})}{\partial \mathbf{Z}} &= -2\mathbf{R}_{12}^T \boldsymbol{\Lambda}^T \mathbf{F} + 2\mathbf{R}_{12}^T \boldsymbol{\Lambda}^T \boldsymbol{\Lambda} \mathbf{R}_{12} \mathbf{Z} \\
 &\quad + 2\lambda \sum_{t=1}^v \alpha_t^p (\mathbf{D}_{22}^{(t)} - \mathbf{R}_{22}^{(t)}) \mathbf{Z}
 \end{aligned} \tag{5}$$

We can then use the Karush-Kuhn-Tucker (KKT) conditions [52] for the nonnegativity of $\mathbf{F}$, $\mathbf{Z}$:

$$(2\mathbf{F} - \boldsymbol{\Lambda} \mathbf{R}_{12} \mathbf{Z} + \mathbf{D}_{11} \mathbf{F} - \mathbf{R}_{11} \mathbf{F} - \mathbf{Y})_{ij} [\mathbf{F}]_{ij} = 0 \tag{6}$$

$$\begin{aligned}
 &(-\mathbf{R}_{12}^T \boldsymbol{\Lambda}^T \mathbf{F} + \mathbf{R}_{12}^T \boldsymbol{\Lambda}^T \boldsymbol{\Lambda} \mathbf{R}_{12} \mathbf{Z} \\
 &\quad + \lambda \sum_{t=1}^v \alpha_t^p (\mathbf{D}_{22}^{(t)} - \mathbf{R}_{22}^{(t)})_{ij} [\mathbf{Z}]_{ij} = 0
 \end{aligned} \tag{7}$$

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

$$[\mathbf{F}]_{ij} = [\mathbf{F}]_{ij} \frac{\boldsymbol{\Lambda} \mathbf{R}_{12} \mathbf{Z} + \mathbf{R}_{11} \mathbf{F} + \mathbf{Y}}{2\mathbf{F} + \mathbf{D}_{11} \mathbf{F}} \tag{8}$$

$$[\mathbf{Z}]_{ij} = [\mathbf{Z}]_{ij} \frac{\mathbf{R}_{12}^T \boldsymbol{\Lambda}^T \mathbf{F} + \lambda (\sum_{t=1}^v \alpha_t^p \mathbf{R}_{22}^{(t)}) \mathbf{Z}}{\mathbf{R}_{12}^T \boldsymbol{\Lambda}^T \boldsymbol{\Lambda} \mathbf{R}_{12} \mathbf{Z} + \lambda (\sum_{t=1}^v \alpha_t^p \mathbf{D}_{22}^{(t)}) \mathbf{Z}} \tag{9}$$

When $\mathbf{F}$ and $\mathbf{Z}$ are fixed, Eq. (3) is equivalent as:

$$\begin{aligned} \arg \min_{\alpha} \lambda \sum_{t=1}^v \alpha_t^p \text{tr}(\mathbf{Z}^T \mathbf{L}_{22}^{(t)} \mathbf{Z}) \\ \text{s.t. } \sum_{t=1}^v \alpha_t = 1 \end{aligned} \quad (10)$$

Here, we adopt the Lagrange multiplier method to optimizing α :

$$H(\mathbf{Z}, \alpha, \eta) = \lambda \sum_{t=1}^v \alpha_t^p \text{tr}(\mathbf{Z}^T \mathbf{L}_{22}^{(t)} \mathbf{Z}) - \eta \left(\sum_{t=1}^v \alpha_t - 1 \right) \quad (11)$$

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

$$\frac{\partial H(\mathbf{Z}, \alpha, \eta)}{\partial \alpha_t} = \lambda p \alpha_t^{p-1} \text{tr}(\mathbf{Z}^T \mathbf{L}_{22}^{(t)} \mathbf{Z}) - \eta = 0 \quad (12)$$

$$\alpha_t = \left(\frac{\eta}{\lambda p \text{tr}(\mathbf{Z}^T \mathbf{L}_{22}^{(t)} \mathbf{Z})} \right)^{\frac{1}{p-1}} \quad (13)$$

Since $\sum_{t=1}^v \alpha_t = 1$ we can obtain:

$$\alpha_t = \frac{\left(\frac{1}{\text{tr}(\mathbf{Z}^T \mathbf{L}_{22}^{(t)} \mathbf{Z})} \right)^{\frac{1}{p-1}}}{\sum_{t=1}^v \left(\frac{1}{\text{tr}(\mathbf{Z}^T \mathbf{L}_{22}^{(t)} \mathbf{Z})} \right)^{\frac{1}{p-1}}} \quad (14)$$

By iteratively updating $\mathbf{F}$, $\mathbf{Z}$, α via Eq. (8), Eq. (9) and Eq. (14), we can obtain the local optimal of $\mathbf{F}$, $\mathbf{Z}$, α . **Algorithm 1** lists the above optimization procedure.

Algorithm 1 TS-Isofun: Tissue Specificity based Isoform Function Prediction

Input: $\mathbf{R}_{11}$, $\mathbf{R}_{12}$, $\{\mathbf{R}_{22}^t\}_{t=1}^v$, $\mathbf{Y}$, p , λ , $\maxIter$, tol .

Output: $\mathbf{Z}$, α .

- 1: Initialize $\alpha_r = 1/v$, $iter = 1$, $tol = 10^{-2}$, $\maxIter = 20$
 - 2: $\mathbf{F} = \mathbf{Y}$
 - 3: $\mathbf{Z} = \mathbf{R}_{12}^T \mathbf{F}$
 - 4: $loss^{iter} = H(\mathbf{F}, \mathbf{Z}, \alpha)$
 - 5: While $iter < \maxIter$ and $|\delta| > tol$
 - 6: Update $\mathbf{F}$ using Eq. (8)
 - 7: Update $\mathbf{Z}$ using Eq. (9)
 - 8: Update α using Eq. (14)
 - 9: $loss^{iter+1} = H(\mathbf{F}, \mathbf{Z}, \alpha)$
 - 10: $\delta \leftarrow loss^{iter+1} - loss^{iter}$
 - 11: $iter = iter + 1$
 - 12: End While
-

3.4 Isoform/gene function prediction

Suppose $\mathbf{F}^*$, $\mathbf{Z}^*$ are optimized variables. $\mathbf{F}^*$ is the predicted gene-term association likelihood matrix and $\mathbf{Z}^*$ is the predicted isoform-term association likelihood matrix. However, the functional annotations of isoforms are generally unknown. To enable a surrogate evaluation, we need to aggregate the isoform-level predictions to

the gene-level. For this surrogate evaluation, we recall Eq. (1) to approximate the gene-term association matrix as follows:

$$\mathbf{Y}^* = \mathbf{A} \mathbf{R}_{12} \mathbf{Z}^* \quad (15)$$

4 EXPERIMENT RESULTS AND ANALYSIS

4.1 Experimental setup

To investigate the performance of TS-Isofun for function prediction, we collected the GO file and GO annotation (GOA) file of Human archived on 2019-06-02 from the GO website <http://geneontology.org/>. The GO file includes the hierarchical relationship between GO terms, and is organized in three branches: Biological Process Ontology (BPO), Molecular Function Ontology (MFO) and Cellular Component Ontology (CCO). The GOA file records associations between GO terms and genes, which indicate the relevant genes that carry out the biological functions described by the corresponding GO terms. To avoid circular prediction, direct annotations with evidence code 'IEA' (inferred from electronic annotations) were excluded. In addition, following the screening criteria of compared methods [26], [30]–[32], we excluded the too sparse terms annotated to fewer than 30 genes, and the too general terms annotated to more than 300 genes.

After that, the numbers of GO terms used for the experiments are 260 (CCO), 258 (MFO) and 1548 (BPO). The processed GO annotations of genes and isoforms of the genes are listed in Table 1.

TABLE 1

Statistics of GO annotations of Human. 'annotations' is the number of used annotations of genes for experiment.

genes(n)	7,549		
isoforms(m)	39,559		
	BPO	CCO	MFO
terms(c)	1,548	260	258
annotations	132,156	23,275	22,112

Many evaluation metrics were used to quantify the performance of gene function prediction. Here, we adopt four widely-used metrics, AUROC, AUPRC, Fmax, and Smin, these metrics were also used in Critical Assessment of protein Function Annotation algorithms [3]. AUROC computes the area under receiver operating curve of each GO term at first, and then takes the average value of these areas to quantify the overall performance. 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. Fmax is the global maximum harmonic mean of recall and precision across all possible thresholds on $\mathbf{F}$. Smin computes the information theoretic analogs of precision and recall by referring to the GO hierarchy and takes the minimum semantic distance between the predictions and ground-truths across all possible thresholds on $\mathbf{F}$. In addition,

since gene function prediction can be solved as a multi-label learning problem [53]–[55], we additionally adopt the RankingLoss, a widely-used multi-label evaluation metric that computes the average fraction of incorrectly predicted annotations ranking ahead of the ground-truth annotations, to measure the prediction results. Except Smin and RankingLoss, the higher the value of these metrics, the better the performance is. The opposite trend holds for Smin and RankingLoss. The formal definition of these metrics can be found in [3], [53]. These metrics quantify the prediction results from different perspectives, and it is difficult for one method to consistently perform better than another one across all these metrics.

To comparatively study the performance of TS-Isofun, we take six state-of-the-art isoform function prediction methods, miSVM [26], MISVM [26], iMILP [27], IsoFun [30], DisoFun [31] and DIFFUSE [32], as comparing methods. These methods were detailed in Related Work Section. Particularly, miSVM identifies the top 25% isoforms as the ‘responsible’ isoforms of a gene with respect to a positive annotation, while MISVM only takes the isoform with the maximum score as the ‘responsible’ one.

For TS-Isofun, we choose λ in $\{10^{-7}, 10^{-6}, \dots, 10^{-2}, 10^{-1}\}$ and p in $\{2, 3, \dots, 6\}$. We then select $p = 3$ and $\lambda = 10^{-3}$ for the following experiments, and fix the number of iterations to optimize $\mathbf{F}$ and $\mathbf{Z}$ as 20. The input parameters of these comparing methods are fixed or optimized as the authors suggested in the paper or shared codes. In addition, we also approximately evaluate TS-Isofun at the isoform-level using its predictions on single-isoform genes (as done in [27], [30], [31]), whose functional annotations directly inherit from the hosting genes.

4.2 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 set the test part of $\mathbf{Y}$ to 0 in Eq. (3). We initialize the isoform-term association matrix $\mathbf{Z}$ by the gene-term association matrix $\mathbf{F}$. In other words, if a term is annotated to the gene, the term is also annotated to all isoforms of the hosting gene. Table 2 reports the results of TS-Isofun and of the comparing methods in five-fold cross validation. The functional annotations of genes in the validation set are considered as unknown during training and prediction. Moreover, to speed up the training of miSVM and MISVM, we select negative genes twice as positive ones for each term to train the model.

TS-Isofun gives significantly better results than the comparing methods across most evaluation metrics. More specifically, the performance margin between TS-Isofun and other comparing methods is very prominent on AUROC, AUPRC, Fmax and 1-RankingLoss. AUROC and AUPRC are GO term-centric metrics, Fmax and 1-RankingLoss are gene-centric metrics. This prominent

improvement shows that TS-Isofun can more reliably predict the functions of genes (isoforms) from both the gene and GO term perspectives. There are two causes for this improvement. TS-Isofun utilizes the PPI network to complete the gene-term associations, and other comparing methods (except IsoFun and DisoFun) exclude the important PPI data, and they generally assume the gene-term associations are complete. TS-Isofun accounts for tissue specificity and selectively fuse functional association networks of different tissues. In contrast, miSVM, MISVM, IsoFun, DIFFUSE and DisoFun concatenate the isoform expression profile features of different tissues into a single feature vector, and iMILP combines the greedily selected isoform functional networks with equal weight. iMILP sometimes obtains a smaller (better) Smin than other comparing methods, but it always loses to TS-Isofun on the other evaluation metrics. That is possible due to different preferences of these evaluation metrics. DIFFUSE manifests a higher AUROC than TS-Isofun, since it uses an iterative semi-supervised learning algorithm to train both the DNN (deep neural networks) and CRF (conditional random field) together to tackle the lack of annotations at the isoform-level, and it separately optimizes the DNN and CRF for each term, while AUROC is a term-centric metric. For the same reason, comparing with other methods, miSVM and MISVM also have much lower performance with respect to gene (or isoform) centric metrics (Fmax, Smin and 1-RankingLoss) than with respect to term-centric metrics (AUROC and AUPRC). Besides the RNA-seq data, IsoFun and DisoFun additionally uses the PPI network data, DIFFUSE utilizes the sequence data by complex DNN. Even though, they all lose to TS-Isofun, which also additionally uses the PPI data but accounts for the tissue specificity. This comparison indicates the contribution of tissue specificity for isoform function prediction.

We also find that MISVM often works better in the BPO than miSVM, but not so in the other two subontology. That is because Biological Process Ontology (BPO) has more processed terms and the ontology structure is more complex, compared with Molecular Function Ontology (MFO) and Cellular Component Ontology (CCO). In practice, miSVM identifies the top 25% isoforms as the ‘responsible’ isoforms of a gene with respect to a positive annotation, while MISVM only takes the isoform with the maximum score as the responsible one. In addition, these evaluation metrics quantify the performance of function prediction from different perspectives. For these reasons, MISVM does not consistently work better than miSVM across all the three subontology.

To further investigate the contribution of tissue specificity, we introduced two variants of TS-Isofun: (i) TS-Isofun(E) assigns equal weight ($\alpha_t = 1/5$) to each individual isoform functional association network; (ii) TS-Isofun(A) concatenates the isoform expression profile feature vectors of five different tissues into a single one, and then directly constructs a single isoform functional

TABLE 2

Experimental results of five-fold cross-validation at the gene-level. $\downarrow$ means the lower the value, the better the performance is. $\bullet/\circ$ indicates TS-Isofun performing significantly better/worse than the other comparing method, where significance is measured by a pairwise t -test at 95% level.

		miSVM	MISVM	iMILP	IsoFun	DIFFUSE	Disofun	TS-Isofun
AUROC	CCO	0.5556 $\pm$ 0.0147 $\bullet$	0.5487 $\pm$ 0.0196 $\bullet$	0.5372 $\pm$ 0.0249 $\bullet$	0.7116 $\pm$ 0.0125 $\bullet$	0.7392 $\pm$ 0.0000 $\bullet$	0.7207 $\pm$ 0.0067 $\bullet$	0.7461 $\pm$ 0.0096
	MFO	0.5324 $\pm$ 0.0119 $\bullet$	0.5159 $\pm$ 0.0108 $\bullet$	0.5385 $\pm$ 0.0267 $\bullet$	0.6407 $\pm$ 0.0076 $\bullet$	0.7476 $\pm$ 0.0000 $\circ$	0.6494 $\pm$ 0.0205 $\bullet$	0.6768 $\pm$ 0.0157
	BPO	0.5395 $\pm$ 0.0129 $\bullet$	0.5560 $\pm$ 0.0321 $\bullet$	0.5358 $\pm$ 0.0337 $\bullet$	0.6887 $\pm$ 0.0033	0.7060 $\pm$ 0.0000 $\circ$	0.6659 $\pm$ 0.0092 $\bullet$	0.6910 $\pm$ 0.0107
AUPRC	CCO	0.0191 $\pm$ 0.0030 $\bullet$	0.0195 $\pm$ 0.0044 $\bullet$	0.0196 $\pm$ 0.0065 $\bullet$	0.0514 $\pm$ 0.0044 $\bullet$	0.0415 $\pm$ 0.0000 $\bullet$	0.0562 $\pm$ 0.0053 $\bullet$	0.0671 $\pm$ 0.0042
	MFO	0.0134 $\pm$ 0.0014 $\bullet$	0.0122 $\pm$ 0.0010 $\bullet$	0.0134 $\pm$ 0.0019 $\bullet$	0.0320 $\pm$ 0.0028 $\bullet$	0.0337 $\pm$ 0.0000	0.0333 $\pm$ 0.0054	0.0380 $\pm$ 0.0033
	BPO	0.0239 $\pm$ 0.0043 $\bullet$	0.0272 $\pm$ 0.0013 $\bullet$	0.0361 $\pm$ 0.0039 $\bullet$	0.0473 $\pm$ 0.0032	0.0421 $\pm$ 0.0000 $\bullet$	0.0393 $\pm$ 0.0040 $\bullet$	0.0472 $\pm$ 0.0066
Fmax	CCO	0.0555 $\pm$ 0.0038 $\bullet$	0.0627 $\pm$ 0.0200 $\bullet$	0.1631 $\pm$ 0.0709 $\bullet$	0.2148 $\pm$ 0.0124 $\bullet$	0.1114 $\pm$ 0.0000 $\bullet$	0.2236 $\pm$ 0.0087 $\bullet$	0.3155 $\pm$ 0.0098
	MFO	0.0416 $\pm$ 0.0023 $\bullet$	0.0378 $\pm$ 0.0015 $\bullet$	0.1733 $\pm$ 0.0790 $\bullet$	0.1573 $\pm$ 0.0064 $\bullet$	0.1090 $\pm$ 0.0000 $\bullet$	0.1433 $\pm$ 0.0164 $\bullet$	0.2114 $\pm$ 0.0145
	BPO	0.0318 $\pm$ 0.0135 $\bullet$	0.0389 $\pm$ 0.0087 $\bullet$	0.1848 $\pm$ 0.0880 $\bullet$	0.1746 $\pm$ 0.0037 $\bullet$	0.0748 $\pm$ 0.0000 $\bullet$	0.1580 $\pm$ 0.0049 $\bullet$	0.2273 $\pm$ 0.0075
Smin $\downarrow$	CCO	0.9264 $\pm$ 0.0284 $\bullet$	0.9285 $\pm$ 0.0381 $\bullet$	0.8553 $\pm$ 0.0359 $\bullet$	0.8618 $\pm$ 0.0371 $\bullet$	0.8695 $\pm$ 0.0000 $\bullet$	0.8910 $\pm$ 0.0380 $\bullet$	0.8305 $\pm$ 0.0241
	MFO	1.0716 $\pm$ 0.0387 $\bullet$	1.0871 $\pm$ 0.0458 $\bullet$	0.9796 $\pm$ 0.0580 $\bullet$	1.0523 $\pm$ 0.0393 $\bullet$	2.3354 $\pm$ 0.0000 $\bullet$	1.0568 $\pm$ 0.0684 $\bullet$	1.0263 $\pm$ 0.0583
	BPO	4.4214 $\pm$ 0.0201 $\bullet$	4.3750 $\pm$ 0.0371 $\bullet$	3.9950 $\pm$ 0.3692 $\circ$	4.2813 $\pm$ 0.1589 $\bullet$	10.0548 $\pm$ 0.0000 $\bullet$	4.3350 $\pm$ 0.1539 $\bullet$	4.1794 $\pm$ 0.2199
1-RankingLoss	CCO	0.5610 $\pm$ 0.0268 $\bullet$	0.5540 $\pm$ 0.0585 $\bullet$	0.0864 $\pm$ 0.0495 $\bullet$	0.7544 $\pm$ 0.0066 $\bullet$	0.6867 $\pm$ 0.0000 $\bullet$	0.7649 $\pm$ 0.0065 $\bullet$	0.8161 $\pm$ 0.0085
	MFO	0.5395 $\pm$ 0.0302 $\bullet$	0.5223 $\pm$ 0.0186 $\bullet$	0.0941 $\pm$ 0.0535 $\bullet$	0.7207 $\pm$ 0.0025 $\bullet$	0.6750 $\pm$ 0.0000 $\bullet$	0.7129 $\pm$ 0.0200 $\bullet$	0.7673 $\pm$ 0.0056
	BPO	0.5280 $\pm$ 0.0215 $\bullet$	0.5582 $\pm$ 0.0393 $\bullet$	0.1014 $\pm$ 0.0574 $\bullet$	0.7475 $\pm$ 0.0026 $\bullet$	0.6808 $\pm$ 0.0000 $\bullet$	0.7143 $\pm$ 0.0086 $\bullet$	0.7760 $\pm$ 0.0041

association network using cosine similarity. Fig. 2 reports the AUROC, AUPRC and Fmax values of TS-Isofun and of its variants.

Fig. 2. Results of TS-Isofun, TS-Isofun(E) and TS-Isofun(A). TS-Isofun(E) combines the functional association networks with equal weight, and TS-Isofun(A) works on the concatenated expression profiles of isoforms across all tissues.

TS-Isofun gives improved AUROC, AUPRC and Fmax values than TS-Isofun(E) and TS-Isofun(A). This improvement justifies our motivation to fuse functional association networks of different tissues with different weights. TS-Isofun(A) has much lower AUPRC values than TS-Isofun and TS-Isofun(E). This fact proves that the isoform expression profile features of different tissues should be differently used for function prediction. By referring to the results in Table 2, TS-Isofun(E) and TS-Isofun(A) also give a better performance than other comparing methods on most evaluation metrics. This superiority confirms the advantage of our joint matrix factorization on leveraging gene-level and isoform-level data for isoform function prediction.

We further visualize the weights (α_t ($t \in \{1, 2, \dots, 5\}$)) of individual functional association networks ($\mathbf{R}_{22}^{(t)}$) of five different tissues under different input values of λ (p fixed as 3 in CCO) in Fig. 3. From this figure, we observe that when $\lambda = 10^{-1}$, the functional association network of Brain has a weight close to 1, while that of other tissues are close to zero. For the configuration of $\lambda = 10^{-1}$ and $p = 3$, the AUROC of TS-Isofun is 0.6852. When $\lambda = 10^{-5}$ (or $\lambda = 10^{-7}$) and $p = 3$, TS-Isofun

assigns slightly different weights to these five networks, and obtains an AUROC as 0.7295 (or 0.7212). When $\lambda = 10^{-3}$ and $p = 3$, TS-Isofun assigns quite different weights to these networks and obtains a much higher AUROC as 0.7360 than other cases. This fact confirms that it is necessary to assign different weights to tissue-specific isoform functional association networks, and the account of tissue specificity boosts the performance of isoform function prediction.

Fig. 3. The weights of five tissues under different λ .

4.3 Results evaluation at isoform-level

In this subsection, we further assess the performance of TS-Isofun at isoform-level. Since isoforms lack ground-truth annotations, we take 1338 single-isoform genes that produce one isoform only, in our collected data as the test set, and the rest as the training set. The obtained Fmax, Smin values of miSVM, MISVM, iMILP, IsoFun, Disofun, and TS-Isofun are given in Fig. 4.

TS-Isofun again manifests a better performance than other comparing methods at the isoform-level function prediction. iMILP only propagates annotations of isoforms on the isoform functional association network, and it does not consider the tissue specificity and the PPI data. For similar reasons, miSVM and MISVM loses to TS-Isofun. Alike TS-Isofun, IsoFun and Disofun also use the PPI data, but they do not account for the tissue specificity of multiple RNA-seq data. As a result, they

obtain better results than other comparing methods but still lose to TS-Isofun. The performance margin between IsoFun and TS-Isofun is more obvious with respect to Fmax than to Smin. That is because Isofun additionally uses the GO hierarchy for isoform function prediction and Smin is defined with respect to the GO hierarchy. Even though, TS-Isofun still holds a comparable Smin with Isofun.

Fig. 4. Results of miSVM, MISVM, iMILP, IsoFun, DisoFun, and TS-Isofun at the isoform-level.

4.4 Parameter sensitivity analysis

There are two input parameters (p and λ) involved with TS-Isofun. λ balances the importance of isoform functional association networks, PPI data and GO annotations of genes, p controls the weight decay rate of individual networks. To investigate the sensitivity of p and λ , following the experimental setup in Section 4.2, we test the performance of TS-Isofun by varying λ in $\{10^{-7}, 10^{-6}, \dots, 10^{-2}, 10^{-1}\}$, and p in $\{2, 3, \dots, 6\}$, and report the AUROC, AUPRC, Fmax values of TS-Isofun in CCO in Fig. 5.

TS-Isofun firstly gives a sharply increased performance as λ decreased from 10^{-1} to 10^{-3} , and then shows a gradually decreased performance as λ further decrease. This pattern is explainable. A too large λ over-weights the functional association networks of isoforms and overrides the complementary PPI network data, but a too small λ under-weights the important functional association networks of isoforms. This trend shows that both the gene-level and isoform-level data should be leveraged for isoform function prediction. On the other hand, TS-Isofun is not so sensitive to p and it has a slightly better results when a small p value is used. As to the evaluation metric Fmax, TS-Isofun gives a relative stable performance across a wide range of combinations of λ and p . The possible cause is that Fmax is the maximum harmonic mean of precision and recall across all thresholds of the predicted gene-term association matrix. From above analysis, we adopt $p = 3$ and $\lambda = 10^{-3}$ for experiments.

In the data preprocessing Section 3.1, we took a threshold value that set FPKM values lower than 0.3 as 0, and then filtered out 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 isoforms are removed also. To study the sensitivity of this threshold value, we separately

set the value as 0.1, 0.2, 0.5, 0.6 and 0.7 for five-fold cross-validation experiments, and report the prediction results in TABLE 3. We can observe that TS-Isofun with threshold fixed as 0.3 generally gives better results than other threshold values, and a threshold ≥ 0.5 presents gradually decreased results, while threshold < 0.3 sometimes gives slightly better results. That is because a larger threshold discards too many isoforms and genes, which distort the underlying relationships between human isoforms/genes, while a lower threshold can keep more genes and isoforms, but also bring in more noisy features in the isoform expression data. For these reasons, we adopt the threshold value as 0.3 for experiments.

4.5 Complexity and convergence analysis

We also empirically study the convergency trend of TS-Isofun on the three branches (BPO, MFO, CCO). From Fig. 6, we can find the objective function loss of TS-Isofun decreases dramatically in the first five iterations and decreases slowly in the next iterations, and it reaches to the convergency within twenty iterations. This observation shows our alternative optimization procedure for solving the objective function of TS-Isofun can quickly converge.

Fig. 6. The convergence trend of TS-Isofun as the number of iterations increasing.

The time complexity of TS-Isofun is mainly dominated by matrix multiplication. In each iterative optimization, the update of $\mathbf{F}$, $\mathbf{Z}$ and α separately costs $\mathcal{O}(n^2m + nmc)$, $\mathcal{O}(n^2m + m^2(n + c))$ and $\mathcal{O}(cm^2 + c^2m)$. Suppose τ is the number of iterations for the convergence of TS-Isofun, the overall complexity of TS-Isofun is $\mathcal{O}(\tau * (n^2m + nmc + m^2n + m^2c + c^2m))$. Here n , m and c are the number of genes, isoforms and distinct GO terms.

We also record the runtime of TS-Isofun and other comparing methods in Table 4 (Platform: Intel E5-2678, Tesla K40c GPU with 256GB RAM). From Table 4, we can find that iMILP and IsoFun have less runtime than other methods. This is because iMILP and IsoFun are based on efficient label propagation. miSVM and MISVM use multiple instance support vector machine for each GO term, and they select twice negative genes as positive ones for each term to speedup the model training, so miSVM and MISVM have less runtime than DIFFUSE, which also makes binary classification but with the time

Fig. 5. AUROC, AUPRC and Fmax of TS-Isofun under different input values of p and λ .

TABLE 3

Experimental results of TS-Isofun under different threshold values for filtering isoforms and genes. $\downarrow$ means the lower the value, the better the performance is. $\bullet/\circ$ indicates TS-Isofun with threshold value 0.3 performing significantly better/worse than the other threshold value, where significance is measured by a pairwise t -test at 95% level.

Threshold Values		0.1	0.2	0.5	0.6	0.7	0.3 (TS-Isofun)
AUROC	CCO	0.7522 $\pm$ 0.0154 $\circ$	0.7409 $\pm$ 0.0175 $\bullet$	0.7263 $\pm$ 0.0159 $\bullet$	0.7209 $\pm$ 0.0132 $\bullet$	0.6919 $\pm$ 0.0211 $\bullet$	0.7461 $\pm$ 0.0096
	MFO	0.6899 $\pm$ 0.0215 $\circ$	0.6801 $\pm$ 0.0129	0.6513 $\pm$ 0.0208 $\bullet$	0.6545 $\pm$ 0.0085 $\bullet$	0.6552 $\pm$ 0.0145 $\bullet$	0.6768 $\pm$ 0.0157
	BPO	0.6904 $\pm$ 0.0131	0.6881 $\pm$ 0.0239	0.6784 $\pm$ 0.0227 $\bullet$	0.6640 $\pm$ 0.0025 $\bullet$	0.6533 $\pm$ 0.0138 $\bullet$	0.6910 $\pm$ 0.0107
AUPRC	CCO	0.0505 $\pm$ 0.0028 $\bullet$	0.0665 $\pm$ 0.0034	0.0652 $\pm$ 0.0053	0.0592 $\pm$ 0.0067 $\bullet$	0.0516 $\pm$ 0.0048 $\bullet$	0.0671 $\pm$ 0.0042
	MFO	0.0307 $\pm$ 0.0036 $\bullet$	0.0381 $\pm$ 0.0024	0.0339 $\pm$ 0.0012	0.0360 $\pm$ 0.0026	0.0355 $\pm$ 0.0017	0.0380 $\pm$ 0.0033
	BPO	0.0309 $\pm$ 0.0087 $\bullet$	0.0471 $\pm$ 0.0026	0.0419 $\pm$ 0.0042 $\bullet$	0.0427 $\pm$ 0.0014	0.0448 $\pm$ 0.0053	0.0472 $\pm$ 0.0066
Fmax	CCO	0.3005 $\pm$ 0.0126 $\bullet$	0.3169 $\pm$ 0.0233	0.3107 $\pm$ 0.0280	0.3019 $\pm$ 0.0118 $\bullet$	0.3023 $\pm$ 0.0059 $\bullet$	0.3155 $\pm$ 0.0098
	MFO	0.1941 $\pm$ 0.0018 $\bullet$	0.2039 $\pm$ 0.0042 $\bullet$	0.2109 $\pm$ 0.0139	0.2161 $\pm$ 0.0043	0.2091 $\pm$ 0.0073	0.2114 $\pm$ 0.0145
	BPO	0.2305 $\pm$ 0.0103	0.2307 $\pm$ 0.0136	0.2237 $\pm$ 0.0106	0.2182 $\pm$ 0.0059 $\bullet$	0.2148 $\pm$ 0.0039 $\bullet$	0.2273 $\pm$ 0.0075
Smin $\downarrow$	CCO	0.8144 $\pm$ 0.0153 $\circ$	0.8414 $\pm$ 0.0285 $\bullet$	1.1398 $\pm$ 0.0364 $\bullet$	1.1652 $\pm$ 0.0273 $\bullet$	1.2989 $\pm$ 0.1024 $\bullet$	0.8305 $\pm$ 0.0241
	MFO	0.9519 $\pm$ 0.0304 $\circ$	0.9915 $\pm$ 0.0356 $\circ$	1.1172 $\pm$ 0.0352 $\bullet$	1.3953 $\pm$ 0.0593 $\bullet$	1.2593 $\pm$ 0.0668 $\bullet$	1.0263 $\pm$ 0.0583
	BPO	4.0066 $\pm$ 0.1832 $\circ$	4.4080 $\pm$ 0.1713 $\bullet$	4.4742 $\pm$ 0.2016 $\bullet$	4.4302 $\pm$ 0.1893 $\bullet$	4.8354 $\pm$ 0.1981 $\bullet$	4.1794 $\pm$ 0.2199
1-RankingLoss	CCO	0.8217 $\pm$ 0.0181 $\circ$	0.8137 $\pm$ 0.0280	0.8037 $\pm$ 0.0302 $\bullet$	0.7996 $\pm$ 0.0120 $\bullet$	0.7980 $\pm$ 0.0086 $\bullet$	0.8161 $\pm$ 0.0085
	MFO	0.7610 $\pm$ 0.0295 $\bullet$	0.7640 $\pm$ 0.0205	0.7555 $\pm$ 0.0218 $\bullet$	0.7460 $\pm$ 0.0129 $\bullet$	0.7346 $\pm$ 0.0045 $\bullet$	0.7673 $\pm$ 0.0056
	BPO	0.7727 $\pm$ 0.0169	0.7705 $\pm$ 0.0031 $\bullet$	0.7625 $\pm$ 0.0269 $\bullet$	0.7554 $\pm$ 0.0102 $\bullet$	0.7432 $\pm$ 0.0058 $\bullet$	0.7760 $\pm$ 0.0041

consuming deep learning. Both Disofun and TS-Isofun are matrix factorization based methods, but TS-Isofun takes less runtime than Disofun. This is because TS-Isofun utilizes tissue specificity and converges faster than Disofun, which concatenates all expression profile features of isoforms across tissues for joint matrix factorization. In summary, TS-Isofun holds a moderate runtime but achieves the best performance among these state-of-the-art solutions.

TABLE 4

The runtime of TS-Isofun and other comparing methods.

	miSVM	MISVM	iMILP	IsoFun	DIFFUSE	Disofun	TS-Isofun
BPO	6407.5	6155.0	1690.5	827.5	1928160.0	8875.8	4369.5
MFO	1203.8	1178.4	243.2	167.0	338564.0	8353.4	2377.1
CCO	1290.8	1556.5	227.4	139.0	329664.0	8414.3	2326.5

4.6 Case study

To further study the capability of TS-Isofun in identifying the functions of isoforms spliced from the same gene, we take four exemplar multiple-isoform genes (LMNA, ADAM15, CFLAR and BCL2L1), whose isoform-level isoforms were studied by wet-lab experiments, and report the prediction results of TS-Isofun and of four compared methods (DisoFun, IsoFun, iMILP and miSVM) in

Table 5. DIFFUSE is too time consuming and excluded for experiments.

Lopez-Mejia *et al.* [56] reported the Lamin C expression is mutually exclusive with the splicing of lamin A and progerin isoforms, and occurs by alternative polyadenylation. LMNA encodes functionally distinct isoforms that have opposing effects on energy metabolism and lifespan in mammals. Particularly, lamin C expressing mice live longer, show decreased energy metabolism, increased weight gain, and reduced respiration. In contrast, progerin-expressing mice show increased energy metabolism and are lipodystrophic. Based on this, we used three GO terms (GO:0006612: 'Protein targeting to membrane', GO:0034446: 'Substrate adhesion-dependent cell spreading', and GO:2001257: 'Regulation of cation channel activity') related to the three isoforms to check the reliability of TS-Isofun. In Table 5(a), we can see that TS-Isofun can identify the functions of isoforms almost perfectly, and the overall accuracy on the three isoforms is 88.89% (8/9).

Zhong *et al.* [57] showed that the expression of ADAM15A and ADAM15B isoforms in MDA-MB-435 cells had differential effects on cell morphology, with adhesion, migration, and invasion enhanced by expression of ADAM15A, whereas ADAM15B led to reduced adhesion. Therefore, we used two GO terms (GO: 0010810:

‘regulation of cell-substrate adhesion’, and GO: 0045785: ‘positive regulation of cell adhesion’) related to cell adhesion to test TS-Isofun. In Table 5(b), we can see that TS-Isofun can precisely differentiate the functions of these two isoforms, and the overall accuracy on the two isoforms is 100% (4/4).

BCL2L1 has two isoforms, Bcl-x(S) and Bcl-x(L), which ‘positively/negatively regulate the apoptotic process’ (GO:0043065/GO:0043066), respectively [58]. CFLAR has two well-annotated isoforms, cFLIP-L and cFLIP-S. Krueger *et al.* [59] reported that the experimental functional validation of both cFLIP-L and cFLIP-S are the inhibitors of apoptotic proteins. Moreover, cFLIP-L is also a promoter of apoptosis [60]. In Table 5(c), for the two genes, the overall accuracy of TS-Isofun is 87.50% (7/8), much higher than those of other compared methods.

In summary, TS-Isofun can more accurately (90.48%=19/21) identify the functions of isoforms than Disofun (85.71%=18/21), IsoFun (71.43%=15/21), iMILP (52.38%=11/21) and miSVM (61.90%=13/21). These case studies again support the superiority of TS-Isofun to Disofun, IsoFun, iMILP and miSVM.

TABLE 5

The annotated (predicted) GO annotations of isoforms of LMNA, ADAM15, BCL2L1 and CFLAR. The results (Prediction and Accuracy) are order by TS-Isofun, DisoFun, IsoFun, iMILP and miSVM. ✓ indicates that the GO term is annotated (predicted) to the isoform, and × means the opposite. When the predicted association probability between a GO term and an isoform is in the top 20% of total isoforms of the gene, the isoform is annotated with this term.

(a) Gene: LMNA				
Isoform		GO:0006612	GO:0034446	GO:2001257
lamin C	Annotation	×	×	✓
	Prediction	×/×/×/×/×	×/×/×/×/×	✓/✓/✓/✓/✓
lamin A	Annotation	✓	×	×
	Prediction	✓/✓/✓/×/✓	×/✓/×/×/✓/✓	×/×/×/×/×
progerin	Annotation	×	✓	×
	Prediction	×/×/×/×/×	×/✓/×/×/×	×/×/✓/✓/✓
Accuracy		88.89% / 88.89% / 77.78% / 55.56% / 66.67%		

(b) Gene: ADAM15				
		GO:0010810	GO:0045785	
ADAM15A	Annotation	✓	✓	
	Prediction	✓/✓/✓/×/×	✓/✓/✓/✓/✓	
ADAM15B	Annotation	×	×	
	Prediction	×/×/×/×/×	×/×/✓/✓/✓	
Accuracy		100.00% / 100.00% / 75.00% / 50.00% / 50.00%		

(c) Gene: BCL2L1 and CFLAR				
		GO:0043065	GO:0043066	
BCL2L1	Bcl-x(L)	Annotation	×	✓
		Prediction	×/×/×/×/×	✓/×/×/×/×
	Bcl-x(S)	Annotation	✓	×
		Prediction	×/×/×/×/×	×/×/×/✓/✓
CFLAR	cFLIP-L	Annotation	✓	✓
		Prediction	✓/✓/✓/✓/✓	✓/✓/✓/✓/✓
	cFLIP-S	Annotation	×	✓
		Prediction	×/×/×/×/×	✓/✓/×/×/✓
Accuracy		87.50% / 75.00% / 62.50% / 50.00% / 62.50%		

5 CONCLUSION

Most multi-exon genes can generate multiple alternatively spliced isoforms, which significantly increase the transcriptome and proteome complexity. However, computational isoform function prediction faces with the challenge of unavailable functional annotations at the isoform-level. Multiple-instance learning based solutions have been proposed to bypass the issue and to transfer the gene-level functional annotations to individual isoforms. These solutions still overlook the important tissue specificity of alternative splicing and fuse multiple isoform-level data at a coarse manner. In this paper, we introduce an isoform function prediction approach based on tissue-specificity, which can fuse multiple isoform functional association networks derived from different tissues at a meticulous manner. Experimental results on multiple RNA-seq datasets of Human from five tissues show that TS-Isofun outperforms other related and representative solutions. These results also confirm the significance of tissue specificity for improving the performance of isoform function prediction.

6 ACKNOWLEDGEMENTS

This work is supported by National Natural Science Foundation of China (61872300, 62031003 and 62072380) and Qilu Scholar Fund of Shandong University. The authors appreciate the reviewers for valuable comments on significantly improving the work.

REFERENCES

- [1] J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle *et al.*, “Gencode: the reference human genome annotation for the encode project,” *Genome Research*, vol. 22, no. 9, pp. 1760–1774, 2012.
- [2] P. Radivojac, W. T. Clark, T. R. Oron, A. M. Schnoes, T. Wittkop, A. Sokolov, K. Graim, C. Funk, K. Verspoor, A. Ben-Hur *et al.*, “A large-scale evaluation of computational protein function prediction,” *Nature Methods*, vol. 10, no. 3, p. 221, 2013.
- [3] Y. Jiang, T. R. Oron, W. T. Clark, A. R. Bankapur, D. DAndrea, R. Lepore, C. S. Funk, I. Kahanda, K. M. Verspoor, A. Ben-Hur *et al.*, “An expanded evaluation of protein function prediction methods shows an improvement in accuracy,” *Genome Biology*, vol. 17, no. 1, p. 184, 2016.
- [4] Y. Zhao, J. Wang, J. Chen, X. Zhang, M. Guo, and G. Yu, “A literature review of gene function prediction by modeling gene ontology,” *Frontiers in Genetics*, vol. 11, p. 400, 2020.
- [5] K. M. Borgwardt, C. S. Ong, S. Schöner, S. Vishwanathan, A. J. Smola, and H.-P. Kriegel, “Protein function prediction via graph kernels,” *Bioinformatics*, vol. 21, no. S1, pp. i47–i56, 2005.
- [6] R. Sharan, I. Ulitsky, and R. Shamir, “Network-based prediction of protein function,” *Molecular Systems Biology*, vol. 3, no. 1, 2007.
- [7] G. Yu, H. Rangwala, C. Domeniconi, G. Zhang, and Z. Yu, “Protein function prediction with incomplete annotations,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 11, no. 3, pp. 579–591, 2014.
- [8] M. Kulmanov, M. A. Khan, and R. Hoehndorf, “Deepgo: predicting protein functions from sequence and interactions using a deep ontology-aware classifier,” *Bioinformatics*, vol. 34, no. 4, pp. 660–668, 2017.
- [9] G. Zhou, J. Wang, X. Zhang, M. Guo, and G. Yu, “Predicting functions of maize proteins using graph convolutional network,” *BMC Bioinformatics*, vol. 21, no. S16, p. 420, 2020.
- [10] G. Yu, H. Rangwala, C. Domeniconi, G. Zhang, and Z. Zhang, “Predicting protein function using multiple kernels,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 12, no. 1, pp. 219–233, 2015.

- [11] G. Yu, H. Zhu, C. Domeniconi, and M. Guo, "Integrating multiple networks for protein function prediction," *BMC Systems Biology*, vol. 9, no. 1, p. S3, 2015.
- [12] V. Glorigrijević, M. Barot, and R. Bonneau, "deepnf: deep network fusion for protein function prediction," *Bioinformatics*, vol. 34, no. 22, pp. 3873–3881, 2018.
- [13] G. Yu, G. Fu, J. Wang, and H. Zhu, "Predicting protein function via semantic integration of multiple networks," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 13, no. 2, pp. 220–232, 2015.
- [14] Y. Zhao, J. Wang, M. Guo, X. Zhang, and G. Yu, "Cross-species protein function prediction with asynchronous-random walk," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 18, no. 1, pp. 1–12, 2021.
- [15] L. M. Smith and N. L. Kelleher, "Proteomes as the next proteomics currency," *Science*, vol. 359, no. 6380, pp. 1106–1107, 2018.
- [16] Q. Pan, O. Shai, L. J. Lee, B. J. Frey, and B. J. Blencowe, "Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing," *Nature Genetics*, vol. 40, no. 12, p. 1413, 2008.
- [17] L. M. Smith, N. L. Kelleher, M. Linial, D. Goodlett, P. Langridge-Smith, Y. A. Goo, G. Safford, L. Bonilla, G. Kruppa, R. Zubarev *et al.*, "Proteome: a single term describing protein complexity," *Nature Methods*, vol. 10, no. 3, p. 186, 2013.
- [18] M. Sundvall, V. Veikkolainen, K. Kurppa, Z. Salah, D. Tvorogov, E. J. van Zoelen, R. Aqilan, and K. Elenius, "Cell death or survival promoted by alternative isoforms of *erbb4*," *Molecular Biology of the Cell*, vol. 21, no. 23, pp. 4275–4286, 2010.
- [19] H. Climente-González, E. Porta-Pardo, A. Godzik, and E. Eyras, "The functional impact of alternative splicing in cancer," *Cell Reports*, vol. 20, no. 9, pp. 2215–2226, 2017.
- [20] Q. Huang, J. Wang, X. Zhang, and G. Yu, "Isoform-disease association prediction by data fusion," in *International Symposium on Bioinformatics Research and Applications*, 2020, pp. 44–55.
- [21] H.-D. Li, R. Menon, G. S. Omenn, and Y. Guan, "The emerging era of genomic data integration for analyzing splice isoform function," *Trends in Genetics*, vol. 30, no. 8, pp. 340–347, 2014.
- [22] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig *et al.*, "Gene ontology: tool for the unification of biology," *Nature Genetics*, vol. 25, no. 1, p. 25, 2000.
- [23] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima, "Kegg: new perspectives on genomes, pathways, diseases and drugs," *Nucleic Acids Research*, vol. 45, no. D1, pp. D353–D361, 2016.
- [24] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial Intelligence*, vol. 89, no. 1–2, pp. 31–71, 1997.
- [25] Z.-H. Zhou, M.-L. Zhang, S.-J. Huang, and Y.-F. Li, "Multi-instance multi-label learning," *Artificial Intelligence*, vol. 176, no. 1, pp. 2291–2320, 2012.
- [26] R. Eksi, H.-D. Li, R. Menon, Y. Wen, G. S. Omenn, M. Kretzler, and Y. Guan, "Systematically differentiating functions for alternatively spliced isoforms through integrating rna-seq data," *PLoS Computational Biology*, vol. 9, no. 11, p. e1003314, 2013.
- [27] W. Li, S. Kang, C.-C. Liu, S. Zhang, Y. Shi, Y. Liu, and X. J. Zhou, "High-resolution functional annotation of human transcriptome: predicting isoform functions by a novel multiple instance-based label propagation method," *Nucleic Acids Research*, vol. 42, no. 6, pp. e39–e39, 2014.
- [28] T. Luo, W. Zhang, S. Qiu, Y. Yang, D. Yi, G. Wang, J. Ye, and J. Wang, "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*, 2017, pp. 345–354.
- [29] D. Shaw, H. Chen, and T. Jiang, "Deepisofun: a deep domain adaptation approach to predict isoform functions," *Bioinformatics*, vol. 35, no. 15, pp. 2535–2544, 2019.
- [30] G. Yu, K. Wang, C. Domeniconi, M. Guo, and J. Wang, "Isoform function prediction based on bi-random walks on a heterogeneous network," *Bioinformatics*, vol. 36, no. 1, pp. 303–310, 2020.
- [31] K. Wang, J. Wang, C. Domeniconi, X. Zhang, and G. Yu, "Differentiating isoform functions with collaborative matrix factorization," *Bioinformatics*, vol. 36, no. 6, pp. 1864–1871, 2020.
- [32] H. Chen, D. Shaw, J. Zeng, D. Bu, and T. Jiang, "Diffuse: predicting isoform functions from sequences and expression profiles via deep learning," *Bioinformatics*, vol. 35, no. 14, pp. i284–i294, 2019.
- [33] Y.-T. Tseng, W. Li, C.-H. Chen, S. Zhang, J. J. Chen, X. J. Zhou, and C.-C. Liu, "Iidb: a database for isoform-isoform interactions and isoform network modules," *BMC Genomics*, vol. 16, no. 2, p. S10, 2015.
- [34] J. Zeng, G. Yu, J. Wang, M. Guo, and X. Zhang, "Dmil-iii: Isoform-isoform interaction prediction using deep multi-instance learning method," in *IEEE International Conference on Bioinformatics and Biomedicine*, 2019, pp. 171–176.
- [35] A. Chatr-Aryamontri, R. Oughtred, L. Boucher, J. Rust, C. Chang, N. K. Kolas, L. O'Donnell, S. Oster, C. Theesfeld, A. Sellam *et al.*, "The biogrid interaction database: 2017 update," *Nucleic Acids Research*, vol. 45, no. D1, pp. D369–D379, 2017.
- [36] E. T. Wang, R. Sandberg, S. Luo, I. Khrebtkova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth, and C. B. Burge, "Alternative isoform regulation in human tissue transcriptomes," *Nature*, vol. 456, no. 7221, p. 470, 2008.
- [37] J. D. Ellis, M. Barrios-Rodiles, R. Çolak, M. Irimia, T. Kim, J. A. Calarco, X. Wang, Q. Pan, D. O'Hanlon, P. M. Kim *et al.*, "Tissue-specific alternative splicing remodels protein-protein interaction networks," *Molecular Cell*, vol. 46, no. 6, pp. 884–892, 2012.
- [38] X. Yang, J. Coulombe-Huntington, S. Kang, G. M. Sheynkman, T. Hao, A. Richardson, S. Sun, F. Yang, Y. A. Shen, R. R. Murray *et al.*, "Widespread expansion of protein interaction capabilities by alternative splicing," *Cell*, vol. 164, no. 4, pp. 805–817, 2016.
- [39] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Advances in Neural Information Processing Systems*, 2003, pp. 577–584.
- [40] G. Kandoi and J. A. Dickerson, "Tissue-specific mouse mrna isoform networks," *Scientific reports*, vol. 9, no. 1, pp. 1–24, 2019.
- [41] N. Defer, M. Best-Belpomme, and J. Hanoune, "Tissue specificity and physiological relevance of various isoforms of adenylyl cyclase," *American Journal of Physiology-Renal Physiology*, vol. 279, no. 3, pp. F400–F416, 2000.
- [42] B. Dasgupta and R. R. Chhipa, "Evolving lessons on the complex role of ampk in normal physiology and cancer," *Trends in Pharmacological Sciences*, vol. 37, no. 3, pp. 192–206, 2016.
- [43] E. P. Consortium *et al.*, "An integrated encyclopedia of dna elements in the human genome," *Nature*, vol. 489, no. 7414, p. 57, 2012.
- [44] D. Kim, B. Langmead, and S. L. Salzberg, "Hisat: a fast spliced aligner with low memory requirements," *Nature Methods*, vol. 12, no. 4, p. 357, 2015.
- [45] M. Pertea, G. M. Pertea, C. M. Antonescu, T.-C. Chang, J. T. Mendell, and S. L. Salzberg, "Stringtie enables improved reconstruction of a transcriptome from rna-seq reads," *Nature Biotechnology*, vol. 33, no. 3, p. 290, 2015.
- [46] P. D. Thomas, V. Wood, C. J. Mungall, S. E. Lewis, J. A. Blake, G. O. Consortium *et al.*, "On the use of gene ontology annotations to assess functional similarity among orthologs and paralogs: a short report," *PLoS Computational Biology*, vol. 8, no. 2, p. e1002386, 2012.
- [47] G. Fu, J. Wang, B. Yang, and G. Yu, "Neggoa: negative go annotations selection using ontology structure," *Bioinformatics*, vol. 32, no. 19, pp. 2996–3004, 2016.
- [48] P. Gaudet and C. Dessimoz, "Gene ontology: pitfalls, biases, and remedies," in *The Gene Ontology Handbook*. Humana Press, New York, NY, 2017, pp. 189–205.
- [49] G. Yu, G. Fu, J. Wang, and Y. Zhao, "Newgoa: Predicting new go annotations of proteins by bi-random walks on a hybrid graph," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 15, no. 4, pp. 1390–1402, 2018.
- [50] A. Bossi and B. Lehner, "Tissue specificity and the human protein interaction network," *Molecular systems biology*, vol. 5, no. 1, p. 260, 2009.
- [51] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein *et al.*, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.
- [52] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge University Press, 2004.
- [53] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 8, pp. 1819–1837, 2013.
- [54] G. Yu, H. Rangwala, C. Domeniconi, G. Zhang, and Z. Yu, "Protein function prediction using multi-label ensemble classification," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 10, no. 4, pp. 1045–1057, 2013.

- [55] G. Yu, K. Wang, G. Fu, M. Guo, and J. Wang, "Nmfgo: Gene function prediction via nonnegative matrix factorization with gene ontology," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 17, no. 1, pp. 238–249, 2020.
- [56] I. C. Lopez-Mejia, M. de Toledo, C. Chavey, L. Lapasset, P. Cavelier, C. Lopez-Herrera, K. Chebli, P. Fort, G. Beranger, L. Fajas *et al.*, "Antagonistic functions of Imna isoforms in energy expenditure and lifespan," *EMBO Reports*, vol. 15, no. 5, pp. 529–539, 2014.
- [57] J. L. Zhong, Z. Poghosyan, C. J. Pennington, X. Scott, M. M. Handsley, A. Warn, J. Gavrilovic, K. Honert, A. Krüger, P. N. Span *et al.*, "Distinct functions of natural adam-15 cytoplasmic domain variants in human mammary carcinoma," *Molecular Cancer Research*, vol. 6, no. 3, pp. 383–394, 2008.
- [58] L. H. Boise, M. González-García, C. E. Postema, L. Ding, T. Lindsten, L. A. Turka, X. Mao, G. Nuñez, and C. B. Thompson, "bcl-x, a bcl-2-related gene that functions as a dominant regulator of apoptotic cell death," *Cell*, vol. 74, no. 4, pp. 597–608, 1993.
- [59] A. Krueger, I. Schmitz, S. Baumann, P. H. Krammer, and S. Kirchhoff, "Cellular flce-inhibitory protein splice variants inhibit different steps of caspase-8 activation at the cd95 death-inducing signaling complex," *Journal of Biological Chemistry*, vol. 276, no. 23, pp. 20 633–20 640, 2001.
- [60] D. W. Chang, Z. Xing, Y. Pan, A. Algeciras-Schimmich, B. C. Barnhart, S. Yaish-Ohad, M. E. Peter, and X. Yang, "c-flip1 is a dual function regulator for caspase-8 activation and cd95-mediated apoptosis," *The EMBO journal*, vol. 21, no. 14, pp. 3704–3714, 2002.

Maozu Guo is a professor at the College of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing, China. He received the Ph.D. degree in Computer Science and Technology from Harbin Institute of Technology. His research interests include bioinformatics, machine learning, and data mining.

Guoxian Yu is a Professor at the School of Software, Shandong University, Jinan, China. He received the Ph.D. in Computer Science from South China University of Technology, Guangzhou, China in 2013. His current research interests include data mining and bioinformatics. He serves as reviewers for KDD, ICML, IJCAI, AAAI, TKDE, TNNLS, TCBB, Bioinformatics and other prestigious conferences and journals.

Qiuyue Huang is an MPhil student at the College of Computer and Information Sciences, Southwest University, Chongqing, China. She received B.Sc. degree in Information and Computing Science from China West Normal University, Nanchong, China. Her research interests include data mining and bioinformatics.

Jun Wang is a Professor with the Joint SDU-NTU Centre for Artificial Intelligence Research, Shandong University. She received B.Sc. degree in Computer Science, M.Eng. degree in Computer Science and Ph.D. in Artificial Intelligence from Harbin Institute of Technology, Harbin, China in 2004, 2006 and 2010, respectively. Her current research interests include data mining and their applications in bioinformatics.

Xiangliang Zhang is an Associate Professor and directs the Machine Intelligence and Knowledge Engineering (MINE) Laboratory in King Abdullah University of Science and Technology (KAUST). She earned her Ph.D. degree in computer science with great honors from INRIA-University Paris-Sud 11, France, in 2010. Her main research interests and experiences are in diverse areas of machine learning and data mining.