Bioinformatics, 36(1), 2020, 303–310
doi: 10.1093/bioinformatics/btz535
Advance Access Publication Date: 28 June 2019
Original Paper

OXFORD

Data and text mining

Isoform function prediction based on bi-random walks on a heterogeneous network

Guoxian Yu [1], Keyao Wang[1], Carlotta Domeniconi[2], Maozu Guo[3,4,*] and Jun Wang [1,*]

[1]College of Computer and Information Science, Southwest University, Chongqing, China, [2]Department of Computer Science, George Mason University, Fairfax, VA 22030, USA, [3]School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing, China and [4]Beijing Key Laboratory of Intelligent Processing for Building Big Data, Beijing, China

*To whom correspondence should be addressed.

Associate Editor: Janet Kelso

Received on September 12, 2018; revised on June 21, 2019; editorial decision on June 25, 2019; accepted on June 26, 2019

Abstract

Motivation: Alternative splicing contributes to the functional diversity of protein species and the proteoforms translated from alternatively spliced isoforms of a gene actually execute the biological functions. Computationally predicting the functions of genes has been studied for decades. However, how to distinguish the functional annotations of isoforms, whose annotations are essential for understanding developmental abnormalities and cancers, is *rarely* explored. The main bottleneck is that functional annotations of isoforms are generally unavailable and functional genomic databases universally store the functional annotations at the gene level.

Results: We propose *IsoFun* to accomplish Isoform Function prediction based on bi-random walks on a heterogeneous network. IsoFun firstly constructs an isoform functional association network based on the expression profiles of isoforms derived from multiple RNA-seq datasets. Next, IsoFun uses the available Gene Ontology annotations of genes, gene–gene interactions and the relations between genes and isoforms to construct a heterogeneous network. After this, IsoFun performs a tailored bi-random walk on the heterogeneous network to predict the association between GO terms and isoforms, thus accomplishing the prediction of GO annotations of isoforms. Experimental results show that IsoFun significantly outperforms the state-of-the-art algorithms and improves the area under the receiver-operating curve (AUROC) and the area under the precision-recall curve (AUPRC) by 17% and 44% at the gene-level, respectively. We further validated the performance of IsoFun on the genes ADAM15 and BCL2L1. IsoFun accurately differentiates the functions of respective isoforms of these two genes.

Availability and implementation: The code of IsoFun is available at http://mlda.swu.edu.cn/codes.php? name=IsoFun.

Contact: guomaozu@bucea.edu.cn or kingjun@swu.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Alternative splicing allows a multi-exon gene to produce multiple isoforms through mechanisms like exon skipping, mutual exclusion of exons, alternative 5′ donor site, alternative 3′ acceptor site and intron retention (Pan, 2008). Alternative splicing takes place in 90% of human multi-exon genes, and it significantly increases the transcriptome and proteome complexity in eukaryotic cells (Wang, 2008). 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 (Smith *et al.*, 2013). For example, two isoforms, Bcl-x(S) and Bcl-x(L) of B-cell lymphoma-x (BCL2L1) gene, have pro-apoptotic and anti-apoptotic biological functions, respectively (Revil *et al.*, 2007). The proteoforms actually carry out various biological functions and maintain the normality of living cells. Increasing evidence has shown that alternative splicing plays key roles in developmental abnormality and is closely related with many human diseases, such as breast cancer, colorectal cancer, spinal muscular atrophy and so on (Climente-Gonzalez *et al.*, 2017).

Predicting functional annotations of isoforms (or proteoforms), instead of genes, can contribute to a deeper understanding of both the molecular basis of diverse genetic diseases and the evolution of phenotypic complexity (Li *et al.*, 2014a, 2016b). However, existing computational solutions for function prediction are mainly at the gene level, or predict functions of the canonical isoforms, who are the most prevalent, best documented and often the longest. These solutions cannot differentiate the intrinsic functions of different isoforms. This is because existing functional databases [i.e. Gene Ontology (GO) (Gene Ontology Consortium, 2017) and KEGG (Kanehisa *et al.*, 2017)] universally store the functional knowledge of gene products at the gene-level, although some wet-lab experiments are actually conducted at the isoform-level. In other words, there are no (or scarce) functional annotations of isoforms in the functional databases. It is observed that interaction profiles of isoforms show tissue specific patterns (Ellis *et al.*, 2012), but existing databases still record the protein–protein interaction at the gene-level (Chatr-Aryamontri *et al.*, 2017; Szklarczyk *et al.*, 2014), and they do not record which isoforms are actually studied in the experiments. Furthermore, isoforms of a gene have subtle variances, and sequence-based features cannot provide reliable discriminant information. For these reasons, it is a *difficult challenge* to accurately predict the functional annotations of isoforms based on sequence and interaction data, which are prevalently used in gene function prediction (Jiang *et al.*, 2016). Some researchers employed expression tag (Neverov *et al.*, 2005), full length complementary DNA (Yura *et al.*, 2006) or exon array data (Emig *et al.*, 2010) to analyze isoform functions. However, these types of data generally have low coverage and different biases, which restrict their capability for accurate isoform function prediction.

RNA-Seq techniques can do massively parallel sequencing of the genome-wide transcript isoforms (translated into proteoforms) at a far higher resolution than ever before. The accumulated abundant RNA-Seq datasets in public databases (Edgar *et al.*, 2002) provide unprecedented amount of transcript-level data. In addition, some efficient methods have also been proposed to precisely quantify the expression levels of transcript isoforms with less bias (Patro *et al.*, 2014). They both pave the way for identifying alternative splicing events and predicting isoform function at large scale. Recently, some pioneers have explored RNA-Seq data for high-resolution isoform function prediction, and achieved successful predictions (Eksi *et al.*, 2013; Li *et al.*, 2014b; Luo *et al.*, 2017; Panwar *et al.*, 2016). These methods generally take each gene as a bag and isoforms of the gene as instances of the bag, and then perform function prediction under the multiple instance learning (MIL) framework (Dietterich *et al.*, 1997; Zhou *et al.*, 2012). 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 if all its instances are not annotated with that label. Eksi *et al.* (2013) developed a multiple instance support vector machine learning based solution (miSVM) (Andrews, 2003) to differentiate isoform functions based on RNA-seq data of mouse. miSVM uses the functional annotations at the gene level and generates classifying models at the isoform level instead of the gene level. Empirical study shows that miSVM can identify the 'responsible' isoform(s) that most likely carry the function of its originating gene and also predict novel functions of isoforms. Li *et al.* (2014b) proposed a novel multiple instance-based label propagation method (iMILP) to predict functions of isoforms. iMILP firstly constructs an isoform functional association network using the co-expression pattern derived from multiple RNA-seq datasets of humans, and uses available GO annotations of genes to initialize the functional annotations of isoforms. Next, it iteratively normalizes and updates labels in the isoform network, and allows all qualified isoforms to inherit positive gene's functions in a 'democratic' learning manner until convergence. In this way, iMILP achieves predictions for each isoform to be associated with a given functional label. The aforementioned MIL-based representative solutions treat each label separately. Given that the functional annotations of genes are rather imbalanced, and due to the large number of possible labels (more than 1000), these solutions not only have compromised performance, but also suffer from a heavy computational burden. This is because they ignore the hierarchical dependency between labels (or GO terms), which is captured by a direct acyclic graph, while its appropriate usage can significantly boost the prediction performance (Fu *et al.*, 2016; Gene Ontology Consortium, 2017; Yu *et al.*, 2018). Some researchers firstly constructed the isoform–isoform interaction network using multiple RNA-seq datasets and the guidance of gene-level interactions, and then identified functional modules or propagated labels on the network to accomplish function prediction (Li *et al.*, 2016a; Tseng *et al.*, 2015). However, the effectiveness of these network-based approaches is still restricted by incomplete gene-level interactions and by the fact that the GO hierarchy is still ignored.

In this study, we introduce an approach called IsoFun to predict isoform functions using 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, and assigns all the annotations of a gene to its isoforms. Next, it constructs a heterogeneous network composed of isoforms, genes and GO terms, to encode the relationships between genes and isoforms, the hierarchical relationship between GO terms and functional associations between isoforms. This heterogeneous network can achieve a synergy between the gene-level interactions, available GO annotations of genes and relationships between genes and isoforms, and thus reduce the impact of incomplete individual data sources. IsoFun then introduces a bi-random walk based label propagation on the constructed heterogeneous network to predict isoform functions. In addition, to ensure the known function of a gene is inherited by at least one of its isoforms, IsoFun clamps the known function to the most 'responsible' isoform in each iteration. We conduct experiments on 311 Human RNA-seq datasets collected from ENCODE (ENCODE Project Consortium, 2012), and find that IsoFun achieves significantly better results than other related and competitive approaches (Li *et al.*, 2014b; Tseng *et al.*, 2015; Wei *et al.*, 2017) across various evaluation metrics. A case study on two genes (ADAM15 and BCL2L1) confirms that IsoFun can differentiate the functional annotations of different isoforms of the same gene. In addition, we also prove that the gene-level interactions, and the hierarchical relationship between functional labels can improve the prediction performance and reduce the impact of class-imbalance in isoform function prediction.

2 Materials and methods

2.1 Heterogeneous network construction

Without loss of generality, suppose there are n genes, each gene has n_i isoforms, and the total number of isoforms is $m = \sum_{i=1}^{n} n_i$.

We denote the ith gene as $\mathcal{B}_i = \{\mathbf{x}_{i_1}, \mathbf{x}_{i_2}, \ldots, \mathbf{x}_{i_{n_i}}\}$, $\mathbf{x}_{i_j} \in \mathbb{R}^d$ is the expression profile feature vector of the jth isoform of the ith gene, $\mathcal{T}$ is the set of distinct GO terms (or functional labels) used to annotate the genes, and $\mathcal{A}_i^+/\mathcal{A}_i^-(\in \mathcal{T})$ are the known positive/negative annotations of the ith gene. Each GO annotation consists of an association between a gene and a GO term. A positive annotation means the gene carries the function described by the particular GO term, and a negative annotation means the gene does not carry the function. Since the functional annotations of isoforms are unknown, we initially set the isoform-term association matrix $\mathbf{A} \in \mathbb{R}^{m\times l}$ between m isoforms and l distinct GO terms ($l = |\mathcal{T}|$) as follows:

$$\mathbf{A}(k,t) = \begin{cases} 1, & \text{if } \mathbf{x}_k \in \mathcal{B}_i \text{ and } t \in \mathcal{A}_i^+ \\ -1, & \text{if } \mathbf{x}_k \in \mathcal{B}_i \text{ and } t \in \mathcal{A}_i^- \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

From Eq. (1) we can see that all the known positive (negative) annotations of a gene are initially inherited by its isoforms. Based on these initial annotations, iMILP (Li *et al.*, 2014b) iteratively propagates the inherited positive/negative annotations in an isoform functional association network to differentiate the functional annotations of different isoforms. At the same time, iMILP uses the MIL principle to pin the initial negative annotations to isoforms throughout the iterative process.

Existing computational solutions (Eksi *et al.*, 2013; Li *et al.*, 2014b; Luo *et al.*, 2017) solely depend on the isoform functional association network constructed from multiple RNA-seq datasets. As such, they may have a compromised performance, since the used network is hand-crafted and it does not refer to the raw-level but curated gene-level interaction network. Furthermore, they treat each GO term separately, and ignore the hierarchical relationship between them. Given this, we construct a heterogeneous network (as shown in Fig. 1) to achieve a synergy between the readily available gene-level interactions, the isoform functional network and the GO hierarchy for accurate isoform function prediction. The resulting heterogeneous network not only can reduce the impact of incomplete data sources, but can also utilize the relationship between genes and isoforms, and the dependency between GO terms.

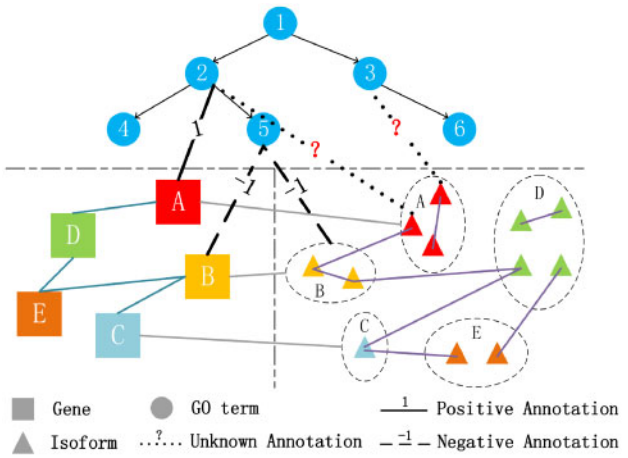

Fig. 1. Illustration of the heterogeneous network composed of isoforms, genes and GO terms. The solid line and segmented lines between subnetworks indicate the known positive and negative associations, and the dotted lines indicate missing associations

Suppose $\mathbf{W} \in \mathbb{R}^{(l+n+m)\times(l+n+m)}$ is the weighted adjacency matrix of the heterogeneous network. $\mathbf{W}$ can be presented using a block-wise notation as follows:

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_{gg} & \mathbf{W}_{gp} & \mathbf{W}_{gm} \\ \mathbf{W}_{pg} & \mathbf{W}_{pp} & \mathbf{W}_{pm} \\ \mathbf{W}_{mg} & \mathbf{W}_{mp} & \mathbf{W}_{mm} \end{bmatrix} \quad (2)$$

where $\mathbf{W}_{gg} \in \mathbb{R}^{l\times l}$, $\mathbf{W}_{pp} \in \mathbb{R}^{n\times n}$ and $\mathbf{W}_{mm} \in \mathbb{R}^{m\times m}$ correspond to the subnetworks of GO terms, genes and isoforms, respectively. $\mathbf{W}_{gp} \in \mathbb{R}^{l\times n}$, $\mathbf{W}_{gm} \in \mathbb{R}^{l\times m}$ and $\mathbf{W}_{pm} \in \mathbb{R}^{n\times m}$ store the known associations between GO terms and genes, GO terms and isoforms and genes and isoforms. $\mathbf{W}_{pg}$, $\mathbf{W}_{mg}$ and $\mathbf{W}_{mp}$ are the transposes of the three corresponding association matrices.

To construct the isoform functional association network $\mathbf{W}_{mm}$, we downloaded 588 RNA-seq runs (for a total of 311 samples) of humans from the ENCODE project (access date: 2017-12-15). These 311 samples are obtained from different tissues and conditions. Due to space limitation, the procedure of processing and controlling the quality of the original data is provided in Supplementary Section S1 of the Supplementary File. After the processing, we obtained a total of 8417 genes with 84 519 isoforms. Isoforms with similar expression profiles have high profile similarity and are more likely to have similar functions; as such we use the Pearson correlation coefficient of FPKM (Fragments Per Kilobase of exon model per Million mapped fragments) feature vectors to measure the functional association between isoforms. Since there are many small coefficient values, which might correspond to noise, we only retain the 100 largest association values for each isoform in the isoform functional association network.

For the inter-association subnetwork $\mathbf{W}_{gp}$ between genes and GO terms, we directly use the collected GO annotations and GO hierarchy to initialize the associations between l GO terms and n genes. Specifically, if the GO term t, or t's descendant terms, provide(s) a positive annotation for gene i, then $\mathbf{W}_{gp}(i,t) = 1$. On the other hand, if t, or its ancestor terms, give(s) a negative annotation to gene i, $\mathbf{W}_{gp}(i,t) = -1$. Otherwise, $\mathbf{W}_{gp}(i,t) = 0$. For the inter-association subnetwork $\mathbf{W}_{mg}$ between isoforms and GO terms, we initially set $\mathbf{W}_{mg} = \mathbf{A}$. For the inter-association subnetwork $\mathbf{W}_{pm}$ between isoforms and genes, if $\mathbf{x}_k \in \mathcal{B}_i$ (namely isoform $\mathbf{x}_k$ originates from gene i), then $\mathbf{W}_{pm}(i,k) = 1$; otherwise $\mathbf{W}_{pm}(i,k) = 0$.

$\mathbf{W}_{gg}$ and $\mathbf{W}_{pp}$ encode the hierarchical dependency between GO terms and the interaction between genes, respectively. They can be directly specified based on the GO file and interaction data, which is detailed in Supplementary Section S1 of the Supplementary File.

To eliminate the different scale effect and to make the transitional probability from a node to its connected nodes equal to 1, we further normalize $\tilde{\mathbf{W}}'_{mm}(i,j) = \mathbf{W}_{mm}(i,j)/\sum_{k=1}^{m}\mathbf{W}_{mm}(i,k)$ and $\tilde{\mathbf{W}}_{pp}(i,j) = \mathbf{W}_{pp}(i,j)/\sum_{k=1}^{n}\mathbf{W}_{pp}(i,k)$. Since $\mathbf{W}_{gg}$ is an asymmetric matrix, we normalize each row of $\mathbf{W}_{gg}$ as $\tilde{\mathbf{W}}_{gg}(s,t) = \mathbf{W}_{gg}(s,t)/\sum_{t'=1}^{l}\mathbf{W}_{gg}(s,t')$.

Given the bag-instance relation between a gene and its isoforms, and the lack of ground-truth isoform–isoform interactions, it is helpful to integrate gene interaction information to describe the functional associations between isoforms (Li *et al.*, 2016a; Tseng *et al.*, 2015). Here, to facilitate the follow-up random walks, we map the interactions between genes onto their isoforms. Specifically, if isoform $u \in \mathcal{B}_i$ and $v \in \mathcal{B}_j$, and $\mathbf{W}_{pp}(i,j) > 0$, then $\tilde{\mathbf{W}}_{mm}(u,v) = \tilde{\mathbf{W}}'_{mm}(u,v) + \tilde{\mathbf{W}}_{pp}(i,j)/(n_i \times n_j)$.

2.2 A bi-random walk on the heterogeneous network

A straightforward solution to infer the associations between isoforms and GO terms is to apply random walks on the heterogeneous

network. Random walk based approaches have been widely adopted in various bioinformatics problems due to their simplicity and effectiveness (Codling *et al.*, 2008). However, our constructed heterogeneous network has different types of edges and nodes, and each type of edges has a different meaning. Furthermore, each subnetwork has a different structure. For example, the GO term subnetwork is a directed acyclic graph, and the isoform functional association subnetwork can be viewed as an undirected network. Applying a global random walk on the whole network may enshroud the intrinsic structures of different networks and result in a compromised performance (Yu *et al.*, 2018). To account for the structural differences of the subnetworks and for the MIL principle, we introduce a tailored bi-random walk on the heterogeneous network to predict the GO annotations of isoforms. Our bi-random walk solution is also inspired by the observation that the functional association between isoforms and GO terms can be predicted in two ways: (i) A random walker first moves from an isoform to another isoform based on the functional association between them, and it stops at a GO term node based on the inherited/inferred isoform-term associations. (ii) A random walker first walks from an isoform node to a GO term node based on the updated isoform-term associations, and then moves downwards towards a descendant term of the current one based on the hierarchial relationship between them.

Based on the above analysis, we can formulate the two random walks as follows:

$$\mathbf{F}_1^{\tau}(i,t) = \sum_{k=1}^{m} \mathbf{W}_{mm}(i,k) * \mathbf{F}^{(\tau-1)}(k,t) \tag{3}$$

$$\mathbf{F}_2^{\tau}(i,t) = \sum_{t \in child(s)} \mathbf{F}^{(\tau-1)}(i,s) * \mathbf{W}_{gg}(s,t) \tag{4}$$

$$\mathbf{F}^{\tau} = (1_{m\times l} - \mathbf{H}) \odot (\mathbf{F}_1^{\tau} + \mathbf{F}_2^{\tau}) + \mathbf{H} \odot \mathbf{F}^0 \tag{5}$$

where $\mathbf{F}^{\tau} \in \mathbb{R}^{m\times l}$ is the predicted likelihood associations between m isoforms and l terms in the τth iteration, $\mathbf{F}^0 = \mathbf{A}$ and $\odot$ denotes the Hadamard product. $\mathbf{H} \in \{0,1\}^{m\times l}$ is introduced to pin the negative GO annotations of isoforms, and ensure the positive annotations of a single-isoform gene are inherited by its isoform. In particular, if isoform i is alternatively spliced from a gene which is negatively annotated by term t, or if this isoform is from a single-isoform gene, which is annotated with t, then $\mathbf{H}(i,t) = 1$. Otherwise $\mathbf{H}(i,t) = 0$. This setup is motivated by the bag-instance relation between gene and its isoforms, and also by the convention of GO annotations.

With the sequential application of Eqs. (3–5), we can iteratively update the association likelihoods between m isoforms and l GO terms. In each iteration, to ensure the annotated terms of a multi-isoform gene are inherited by its isoforms, we introduce a *clamp* process that assigns the term to its 'responsible' isoform as follows:

$$\mathbf{F}^{\tau}(k^*,t) = 1, \text{if } k^* = \underset{\mathbf{x}_k \in \mathcal{B}_i}{argmax}\, \mathbf{F}^{\tau}(k,t) \text{ and } t \in \mathcal{T}_i \tag{6}$$

where k^* is the isoform that has the maximum prediction score with respect to t. Since all the entries of $\mathbf{W}_{mp}$, $\mathbf{W}_{pg}$, $\mathbf{W}_{mm}$ and $\mathbf{W}_{gg}$ are at most one, and Eqs. (3, 4) are geometric sequences, $\mathbf{F}^{\tau}$ converges after a finite number of iterations. At convergence, we obtain the predicted association likelihoods between m isoforms and l terms.

3 Results and discussion

3.1 Experimental setup

To study the performance of IsoFun for function prediction, we collected two releases of GO annotation (GOA) files of Human archived in different years from the GO website (http://geneontology.org/ page/download-annotations/). The historical GOA file was archived on 2016-04-30, and the recent GOA file was archived on 2018-03-15. We train IsoFun on the historical (released in 2016) GOA file, and validate its predictions on the recent (released in 2018) GOA file with the new annotations archived between 2016 and 2018. To avoid the impact of GO structure changes, we also downloaded the contemporary GO files (http://geneontology.org/page/download-ontology/) and used the shared GO structure for the experiments. The biological functions of genes are divided into three branches by GO: Biological Process Ontology (BPO), Molecular Function Ontology (MFO) and Cellular Function Ontology (CCO). Each ontology structures GO terms via a direct acyclic graph to represent the hierarchical relationships between them. The GOA file stores known associations between genes and GO terms. To avoid circular prediction, direct annotations with evidence code 'IEA' (inferred from electronic annotations) were excluded. We filter out sparse GO terms that are associated to very few (<10) genes. The processed GO annotations of genes and isoforms are listed in Table 1. From this table, we can see that many new annotations of genes were appended during a two year interval.

The functional annotations of isoforms are generally unknown. To enable prediction evaluation, we need to aggregate the isoform-level predictions to the gene-level. For this aggregation, we simply summarize the predicted scores of all isoforms of a gene with respect to term t as the aggregated score as follows:

$$\mathbf{Y}(i,t) = \frac{\sum_{\mathbf{x}_k \in \mathcal{B}_i} \mathbf{F}^*(k,t)}{n_i} \tag{7}$$

where $\mathbf{F}^*$ is the finally predicted likelihoods obtained by Eqs. (3–5), and $\mathbf{Y} \in \mathbb{R}^{n\times l}$ is the aggregated likelihood scores between n genes and l terms. $\mathbf{Y}$ can then be used as a surrogate to evaluate the performance of IsoFun and of comparing methods at the gene level. The performance of gene function prediction can be evaluated by different evaluation metrics. To reach a comprehensive comparison, we use five representative evaluation metrics to evaluate the performance of the methods, namely *AUROC*, *AUPRC*, *Fmax*, *Smin* and *RankLoss*. We compare the performance of IsoFun against mi-SVM, MI-SVM (Eksi *et al.*, 2013), iMILP (Li *et al.*, 2014b), miFV and miVLAD (Wei *et al.*, 2017). The first three methods were reviewed in the Introduction. miFV and miVLAD are two efficient and scalable MIL algorithms, which learn new feature vector representations of bags and linear classifiers for bag-level prediction. We used the expression profile values of isoforms across all the collected RNA-seq datasets to construct one isoform functional association network for these comparing methods. For page limitation, details on the evaluation metrics and on the comparing methods are provided in Supplementary Section S2 of the Supplementary File.

Table 1. Statistics of GO annotations of *Human*

♯genes(n)	8714		
♯isoforms(m)	84 519		
Dimensions of isoforms(d)	311		
	BPO	MFO	CCO
History	396 936	73 851	168 084
Recent	491 729	88 287	202 274
♯terms(l)	3357	658	537

Note: 'history' is the number of positive and negative annotations in the historical GOA file (archived date: 2016-04-30), and 'recent' is the number of positive and negative annotations in the recent GOA file (archived date: 2018-03-15).

3.2 Network contribution analysis

To investigate the contribution of accounting for dependencies between functional labels and gene-level interactions, we introduce two variants of IsoFun: IsoFun(P) and IsoFun(G). IsoFun(P) propagates functional annotations on the isoform association network and on the gene–gene interaction network, and it does not account for the dependency between GO terms. As such, Eq. (4) is not used in this case. IsoFun(G) only propagates functional annotations on the isoform subnetwork and GO term subnetwork. This means that the gene-level interactions are disregarded and are not mapped onto the isoform functional association network. The differences between these two variants and iMILP are summarized in Table 2. Figure 2 reports the *Fmax* and *Smin* values of these comparing methods under the *historical to recent* experimental protocol. This protocol is adopted by CAFA (Jiang *et al.*, 2016) and is more challenging than the widely adopted cross-validation protocol.

IsoFun has consistently larger *Fmax* and smaller *Smin* values than its two variants and iMILP; in turn, the two variants are superior to iMILP with respect to both *Fmax* and *Smin* values. Given the performance margin achieved between IsoFun(G) and iMILP, and the margin between IsoFun(P) and iMILP, we can conclude that both the gene-level interactions and the GO hierarchy should be used for accurate isoform function prediction. The improvement of IsoFun(P) against iMILP is smaller than that of IsoFun(G) against iMILP. This is because the isoform-level network can describe the interactions between genes/proteins with higher resolution, and has overlaps with gene–gene interactions. If there is an interaction between two genes, there exists at least one interaction between the respective isoforms of the two genes. But the interactions between the respective isoforms are rarely known. Although both IsoFun(P) and iMILP computationally construct the isoform network from multiple RNA-seq datasets and do not use the dependency between GO terms, IsoFun(P) still obtains a better performance than iMILP. This is because IsoFun(P) additionally uses the gene-level interaction in constructing the isoform functional association network. Another reason is that the newly gathered GO annotations of genes often provide more specific functional knowledge of genes. The related GO terms of these new annotations correspond to descendants of the terms already annotated to genes, and the downward random process in Eq. (4) can take advantage of this pattern. For these reasons, IsoFun(G) achieves better results than IsoFun(P), and in turn IsoFun achieves a better performance than IsoFun(G). The performance margin between IsoFun and IsoFun(G) is more obvious in the BPO, which includes more GO terms than MFO and CCO. The reason is that the GO terms are annotated to genes/isoforms in a rather imbalanced way, and the imbalance effect is more serious in BPO. This observation corroborates the fact that IsoFun can leverage the dependency between GO terms to reduce the impact of class imbalance in isoform function prediction.

Table 2. Differences between IsoFun, iMILP and the variants of IsoFun

	iMILP	IsoFun(P)	IsoFun(G)	IsoFun
Isoform–Isoform network	√	√	√	√
Gene–Gene network	×	√	×	√
GO hierarchy	×	×	√	√

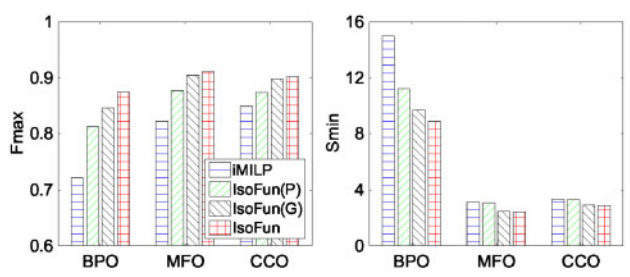

Fig. 2. Results comparison on the new archived GO annotations between 2016 and 2018

The experiments are conducted on archived GO annotations in different years, without random partition of the training set and testing set, so there is no standard deviation to report. To statistically compare the performance of the methods, we use the Wilcoxon signed-rank test (Wilcoxon, 1945) to assess the difference in performance between IsoFun and the other methods across the evaluation metrics and ontologies; the test has shown that all the *P*-values are smaller than 0.031. In summary, these results confirm that both gene-level interactions and dependency between functional labels should be considered in isoform function prediction.

Since we only selected the $k = 100$ most correlated isoforms of each isoform to construct the network, we conducted additional experiments to investigate the sensitivity of IsoFun to k. The obtained results show that an effective k can be easily selected from a wide range of values. Due to page limitation, the experimental results and analysis are provided in Supplementary Figure S1 and Supplementary Section S3 of the Supplementary File.

3.3 Comparison results at the gene-level

Following the experimental protocol used in Eksi *et al.* (2013) and Li *et al.* (2014b), we conduct fivefold cross validation experiments on the *recent* GOA data to study the performance of IsoFun. Due to the prohibitive runtimes of miFV, miVLAD, mi-SVM and MI-SVM on such a large number of isoforms and of functional labels, we re-filtered the data. Particularly, we set all FPKM values less 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 did a further filtering: if an isoform of a gene is filtered, this gene and its all isoforms will be filtered out. We exclude the terms annotated to fewer than 30 genes, and the too general terms annotated to more than 300 genes. After that, the numbers of genes, isoforms, GO terms used for the experiments are 4738, 30 251, 204 (CCO), 210 (MFO) and 1113 (BPO), respectively. For a comprehensive comparison, we introduce other two variants of IsoFun, IsoFun(Y) and IsoFun(M). IsoFun(Y) is similar to IsoFun, but it does not manually *clamp* the positive annotations of a gene to its most 'responsible' isoform in each iteration. As such, Eq. (6) is not used in this case. IsoFun(M) is also similar to IsoFun, but it only selects the isoform with the *maximum* score as the 'responsible' isoform of a gene for a function. Table 3 lists the results of the comparing methods.

IsoFun significantly outperforms the other methods across different evaluation metrics, except *Smin*. Also IsoFun(Y) and IsoFun(M) frequently achieve a better performance than the competing methods. Both IsoFun and iMILP often have a larger *Smin* value than other comparing methods. The reason is that both IsoFun and iMILP are label propagation based solutions. The negative associations between gene and GO terms, and those between isoforms and terms, are over-propagated in the random walk process, and thus expand the semantic distance between the predictions and ground-truths, whereas miFV, miVLAD, mi-SVM and MI-SVM adopt binary classifiers and the negative annotations can enhance the discriminant ability of the classifiers. Furthermore, the

Table 3. Results of different comparing methods on fivefold cross validation at the gene-level

		MiFV	miVLAD	mi-SVM	MI-SVM	iMILP	IsoFun(Y)	IsoFun(M)	IsoFun
AUROC	BPO	0.5775 ± 0.0018•	0.5514 ± 0.0012•	0.5585 ± 0.0023•	0.5421 ± 0.0017•	0.5476 ± 0.0037•	0.5700 ± 0.0018•	0.5924 ± 0.0012•	**0.7143 ± 0.0023**
	MFO	0.5519 ± 0.0034•	0.5490 ± 0.0021•	0.5475 ± 0.0050•	0.5388 ± 0.0052•	0.5256 ± 0.0049•	0.5608 ± 0.0066•	0.5736 ± 0.0061•	**0.6456 ± 0.0043**
	CCO	0.5943 ± 0.0019•	0.5848 ± 0.0096•	0.5777 ± 0.0047•	0.5588 ± 0.0046•	0.5937 ± 0.0035•	0.5810 ± 0.0073•	0.6250 ± 0.0089•	**0.6704 ± 0.0050**
AUPRC	BPO	0.0550 ± 0.0093•	0.0517 ± 0.0097•	0.0503 ± 0.0051•	0.0501 ± 0.0035•	0.2063 ± 0.0127•	0.2644 ± 0.0310•	0.2380 ± 0.0323•	**0.3769 ± 0.0222**
	MFO	0.1214 ± 0.0237•	0.0982 ± 0.0126•	0.1129 ± 0.0181•	0.1012 ± 0.0073•	0.2086 ± 0.0275•	0.2607 ± 0.0319•	0.2185 ± 0.0289•	**0.3156 ± 0.0194**
	CCO	0.1641 ± 0.0096	0.1283 ± 0.0128•	0.1297 ± 0.0034•	0.1061 ± 0.0179•	0.1170 ± 0.0106•	0.1153 ± 0.0194•	0.1298 ± 0.0236•	**0.1651 ± 0.0137**
Fmax	BPO	0.0847 ± 0.0001•	0.0834 ± 0.0000•	0.0872 ± 0.0039•	0.0842 ± 0.0016•	0.2718 ± 0.0168•	0.2875 ± 0.0199•	0.2989 ± 0.0175•	**0.3583 ± 0.0462**
	MFO	0.2615 ± 0.0006•	0.2178 ± 0.0001•	0.2481 ± 0.0072•	0.2473 ± 0.0072•	0.3855 ± 0.0138•	0.4264 ± 0.0324•	0.4233 ± 0.0168•	**0.4810 ± 0.0286**
	CCO	**0.2654 ± 0.0007**°	0.2355 ± 0.0008°	0.2622 ± 0.0040°	0.2515 ± 0.0187°	0.1642 ± 0.0110•	0.1736 ± 0.0122•	0.1806 ± 0.0117•	0.2018 ± 0.0257
Smin↓	BPO	8.9196 ± 0.0006°	**8.6903 ± 0.1499**°	9.3121 ± 0.4070°	9.0030 ± 0.3884°	34.2571 ± 2.0432	33.4458 ± 1.9697	34.4172 ± 2.0171	33.2775 ± 1.8894
	MFO	6.9408 ± 0.0009°	**5.5150 ± 0.0233**°	6.7031 ± 0.1671°	6.6805 ± 0.1803°	11.4297 ± 0.4245	11.0617 ± 0.4199	11.2767 ± 0.4242	11.0009 ± 0.3636
	CCO	5.5684 ± 0.0019•	4.4631 ± 0.1331•	5.4993 ± 0.0786•	5.2463 ± 0.4037•	3.8218 ± 0.2345	3.8193 ± 0.2352	3.8419 ± 0.2417	**3.8146 ± 0.2302**
Rankloss↓	BPO	0.4108 ± 0.0459•	0.4342 ± 0.0353•	0.4713 ± 0.0122•	0.4980 ± 0.0158•	0.2198 ± 0.0149•	0.1062 ± 0.0097•	0.1451 ± 0.0108•	**0.0542 ± 0.0040**
	MFO	0.4194 ± 0.0303•	0.4437 ± 0.0055•	0.4651 ± 0.0402•	0.4834 ± 0.0356•	0.3061 ± 0.0228•	0.1494 ± 0.0122•	0.2062 ± 0.0142•	**0.0966 ± 0.0079**
	CCO	0.3675 ± 0.0538•	0.4124 ± 0.0502•	0.4387 ± 0.0209•	0.4860 ± 0.0087•	0.1034 ± 0.0092•	0.0664 ± 0.0064•	0.0664 ± 0.0052•	**0.0358 ± 0.0029**

Note: ↓ means the lower the value, the better the performance is. •/° indicates IsoFun performing significantly better/worse than the other comparing method, where significance is measured using a pairwise *t*-test at 95% level.

used metrics evaluate the prediction performance from different perspectives, and it's unlikely for an approach to outperform another solution across all the metrics. For example, the performance margin between IsoFun and other comparing methods is very prominent on *Rankloss*, which is also a gene-centric metric and it computes the average fraction of incorrectly predicted annotations ranking ahead of ground-truth annotations. For the two GO term-centric metrics (AUROC and AUPRC), IsoFun consistently performs better than other comparing methods. The advantage of IsoFun and its variants is mainly due to the fact that IsoFun fuses both the gene-level and isoform-level data, and accounts for the dependency between GO terms, whereas the comparing methods solely utilize isoform-level data, although they resort to different machine learning techniques to predict the annotations of isoforms. This observation supports our motivation of constructing a heterogeneous network to achieve a synergy between gene-level interactions, GO hierarchy and isoform-level interactions for function prediction.

iMILP does not give comparable AUROC values as reported in the original paper, that is because it directly works on a composite isoform functional association networks derived from all RNA-seq datasets, without time-consuming multiple network selection, which was done in each iteration of the original paper. Another cause is that we test iMILP and other comparing methods on both single-isoform and multi-isoform genes, whereas iMILP was tested on single-isoform genes. Although we filtered out the gene once its isoform (if any) was filtered out, we still did not get a comparable AUROC value of mi-SVM and MI-SVM as the authors reported. The possible cause is that the adopted processing toolkits are different from those used by mi-SVM; and original mi-SVM filters out gene with cutoff less than 0.5, but IsoFun does not. The samples with cutoff less than 0.5 are usually obtained by single-end sequencing with low quality.

IsoFun(Y) does not apply the clamp step in the bi-random walk process. In other words, it does not force the isoforms of a gene to inherit the known functions of the gene. As a result, the MIL principle, which states that if a gene is annotated with a GO term, then at least one of its isoform should be annotated with that term, might be violated. For this reason, it always loses to IsoFun. IsoFun(M) selects the isoform with the maximum score as the 'responsible' isoform of a gene for a function, and disregards other isoforms. It then aggregates the annotations of these 'responsible' isoforms to the gene-level. IsoFun(M) is significantly outperformed by IsoFun. This observation suggests that a function can be simultaneously inherited by different isoforms, and aggregating the annotations of a gene from all its isoforms is more effective. In fact, the GO follows the convention to collectively annotate the functions of gene products to the gene. For this reason, IsoFun has a better performance than IsoFun(M) across all the metrics. Both mi-SVM and MI-SVM lose to IsoFun, since they separately aggregate the top 25% isoforms or the most 'responsible' isoform for each GO term.

In addition, we adopted the filtering process used by iMILP to filter the data, and finally obtained 7069 genes with 15 826 isoforms. We reported the results of iMILP and IsoFun on this dataset in Supplementary Figure S2 of the Supplementary File. IsoFun has clearly higher values of AUROC and Fmax, and lower values of Smin than those of iMILP. These extended experiments prove the effectiveness of IsoFun under different data filtering protocols. We also recorded the runtime costs of all the methods, and IsoFun has significantly reduced cost compared to the other methods. These results and analysis are provided in Supplementary Section S5 and Supplementary Table S1 of the Supplementary File.

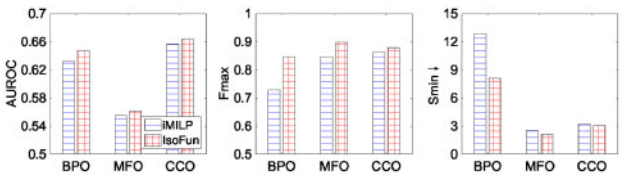

Fig. 3. IsoFun versus iMILP on single-isoform genes

3.4 Comparison results at the isoform-level

In this subsection, we further assess the performance of IsoFun at the isoform-level. Since the ground-truth annotations of isoforms are unknown, we take 422 *single-isoform* genes as the test set, and the annotations in the historical GOA file as the training set. This surrogate assessment was also used by iMILP (Li *et al.*, 2014b). The obtained *AUROC*, *Fmax* and *Smin* values of IsoFun and iMILP are given in Figure 3. We do not report the results of other comparing methods, since the experiment was conducted in the historical to recent protocol and these methods cannot be applied in this setting.

IsoFun again obtains a better performance than iMILP in the isoform-level prediction. iMILP only combines multiple RNA-seq datasets, it does not account for interrelations between GO terms and the closely related interaction data at gene-level, whereas IsoFun takes advantage of all these information sources. The improvement of IsoFun with respect to iMILP is more obvious in the BPO, since BPO has a larger number of hierarchically organized GO terms, and these terms are annotated to genes in a rather unbalanced way. This improvement suggests the GO hierarchy should be considered in isoform function prediction, and this hierarchy knowledge can reduce the impact of class-imbalance. From these experiments, we can conclude that IsoFun can achieve a performance superior to other competitive methods at both gene- and isoform-level.

To further investigate the capability of IsoFun in differentiating the functions of isoforms originated from the same gene, we select two multi-isoform genes, 'ADAM15' (ADAM Metallopeptidase Domain 15) and 'BCL2L1' (B-cell lymphoma-2 like 1), whose isoforms have been studied in wet-lab experiments. ADAM15 is a type I transmembrane glycoprotein known to be involved with cell adhesion cloned, and it has characterized alternatively spliced forms related with human breast cancers (Zhong *et al.*, 2008). ADAM15 has two isoforms (ADAM15A and ADAM15B), which are associated with poorer relapse-free survival in node-negative patients. These two isoforms have different effects on cell morphology. The expression of ADAM15A enhances the adhesion, migration and invasion, whereas ADAM15B reduces adhesion. IsoFun correctly predicted the associations between ADAM15A and GO terms (GO: 0045785, positive regulation of cell adhesion; GO: 0010810, regulation of cell-substrate adhesion). The predicted association values between ADAM15A and GO: 0045785 and GO: 0010810 are much higher than the average value. On the other hand, the predicted association values between ADAM15B and these two terms are far below the average.

BCL2L1, as a protein coding gene, has a vital effect in apoptotic (Boise *et al.*, 1993). BCL2L1 has two isoforms, Bcl-x(S) and Bcl-x(L). Studies have shown that Bcl-x(S) and Bcl-x(L) have pro-apoptotic (GO: 0043065) and anti-apoptotic (GO: 0043066) functions (Revil *et al.*, 2007), respectively. The result of IsoFun fully reflects the functional information of Bcl-x(S) and Bcl-x(L). IsoFun gives the larger association value between Bcl-x(S) and 'GO: 0043065' (positive regulation of apoptotic process) than all the other isoforms. In contrast, the predicted association value for Bcl-x(L) and GO: 0043065 is lower than the average value. On the other hand, the association value between Bcl-x(L) and 'GO: 0043066' (negative regulation of apoptotic process) is twice than that between Bcl-x(S) and 'GO: 0043066', and the former value is higher than the average, and the latter is lower than the average.

4 Conclusions

Differentiating the functions of alternatively spliced isoforms can pave the way for explaining the proteome complexity and various complex diseases in a higher resolution than at the gene-level. Compared with the widely studied gene function prediction, isoform function prediction is rarely studied. The major challenge is that functional annotations of isoforms are generally unavailable and functional genomic data are universally recorded at the gene-level. To attack this challenge, we develop a data integration model called IsoFun. IsoFun firstly constructs a heterogeneous network to encode gene-level interactions, GO terms, isoforms and inter and intra-associations between them. It then introduces a tailored bi-random walk on the heterogeneous network to predict novel associations between isoforms and GO terms, and ensures that the known annotations of a gene are inherited by at least one isoform of the gene. Experimental results show that IsoFun outperforms other related and representative solutions. The study also confirms that integrating the gene-level data and using GO hierarchy can significantly improve the prediction performance.

There are several avenues to further improve the performance of IsoFun, such as fusing multiple gene-level and transcript-level heterogeneous data sources, and taking into account the tissue specific pattern of isoform–isoform interaction network.

Funding

This work was supported by Natural Science Foundation of China (61872300, 61873214, 61871020, 61571163 and 61532014), Fundamental Research Funds for the Central Universities (XDJK2019B024), the National Key Research and Development Plan Task of China (Grant No. 2016YFC0901902), Natural Science Foundation of CQ CSTC (cstc2018jcyjAX0228 and cstc2016jcyjA0351).

Conflict of Interest: none declared.

References

Andrews,S. (2003) Support vector machines for multiple-instance learning. *Adv. Neural Inf. Process. Syst.*, 577–584.

Boise,L.H. *et al.* (1993) bcl-x, a bcl-2-related gene that functions as a dominant regulator of apoptotic cell death. *Cell*, **74**, 597–608.

Chatr-Aryamontri,A. *et al.* (2017) The BioGRID interaction database: 2017 update. *Nucleic Acids Res.*, **45**, D369–D379.

Climente-Gonzalez,H. *et al.* (2017) The functional impact of alternative splicing in cancer. *Cell Rep.*, **20**, 2215–2226.

Codling,E.A. *et al.* (2008) Random walk models in biology. *J. R. Soc. Interface*, **5**, 813–834.

Dietterich,T.G. *et al.* (1997) Solving the multiple instance problem with axis-parallel rectangles. *Artif. Intell.*, **89**, 31–71.

Edgar,R. *et al.* (2002) Gene Expression Omnibus: nCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.

Eksi,R. *et al.* (2013) Systematically differentiating functions for alternatively spliced isoforms through integrating RNA-seq data. *PLoS Comput. Biol.*, **9**, e1003314.

Ellis,J. *et al.* (2012) Tissue-specific alternative splicing remodels protein–protein interaction networks. *Mol. Cell*, **46**, 884–892.

Emig,D. *et al.* (2010) AltAnalyze and DomainGraph: analyzing and visualizing exon expression data. *Nucleic Acids Res.*, **38**, W755–W762.

ENCODE Project Consortium. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57.

Fu,G. *et al*. (2016) NegGOA: negative GO annotations selection using ontology structure. *Bioinformatics*, **32**, 2996–3004.

Gene Ontology Consortium. (2017) Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res*., **45**, D331–D338.

Jiang,Y. *et al*. (2016) An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biol*., **17**, 184.

Kanehisa,M. *et al*. (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res*., **45**, D353–D361.

Li,H. *et al*. (2014a) The emerging era of genomic data integration for analyzing splice isoform function. *Trends Genet*., **30**, 340–347.

Li,W. *et al*. (2014b) High-resolution functional annotation of human transcriptome: predicting isoform functions by a novel multiple instance-based label propagation method. *Nucleic Acids Res*., **42**, e39.

Li,H. *et al*. (2016a) A network of splice isoforms for the mouse. *Sci. Rep*., **6**, 24507.

Li,W. *et al*. (2016b) Pushing the annotation of cellular activities to a higher resolution: predicting functions at the isoform level. *Methods*, **93**, 110–118.

Luo,T. *et al*. (2017) Functional annotation of human protein coding isoforms via non-convex multi-instance learning. In: ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 345–354.

Neverov,A. *et al*. (2005) Alternative splicing and protein function. *BMC Bioinformatics*, **6**, 266.

Pan,Q. *et al*. (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet*., **40**, 1413–1415.

Panwar,B. *et al*. (2016) Genome-wide functional annotation of human protein-coding splice variants using multiple instance learning. *J. Proteome Res*., **15**, 1747–1753.

Patro,R. *et al*. (2014) Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat. Biotechnol*., **32**, 462.

Revil,T. *et al*. (2007) Protein kinase C-dependent control of Bcl-x alternative splicing. *Mol. Cell. Biol*., **27**, 8431–8441.

Smith,M. *et al*. (2013) Proteoform: a single term describing protein complexity. *Nat. Methods*, **10**, 186.

Szklarczyk,D. *et al*. (2014) STRING v10: protein Cprotein interaction networks, integrated over the tree of life. *Nucleic Acids Res*., **10**, D447–D452.

Tseng,Y. *et al*. (2015) IIIDB: a database for isoform–isoform interactions and isoform network modules. *BMC Genomics*, **16**, S10.

Wang,E.T. *et al*. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470–476.

Wei,X.S. *et al*. (2017) Scalable multi-instance learning. *IEEE Trans. Neural Networks Learn. Syst*., **28**, 975–987.

Wilcoxon,F. (1945) Individual comparisons by ranking methods. *Biometrics Bull*., **1**, 80–83.

Yu,G. *et al*. (2018) NewGOA: predicting new GO annotations of proteins by bi-random walks on a hybrid graph. *IEEE/ACM Trans. Comput. Biol. Bioinform*., **13**, 1390–1402.

Yura,K. *et al*. (2006) Alternative splicing in human transcriptome: functional and structural influence on proteins. *Gene*, **380**, 63–71.

Zhong,J.L. *et al*. (2008) Distinct functions of natural ADAM-15 cytoplasmic domain variants in human mammary carcinoma. *Mol. Cancer Res*., **6**, 383–394.

Zhou,Z.H. *et al*. (2012) Multi-instance multi-label learning. *Artif. Intell*., **176**, 2291–2320.