智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

融合多实例学习与注意力机制的异构体功能预测方法

郭茂祖, 周遨宇, 段然

引用本文:
郭茂祖, 周遨宇, 段然. 融合多实例学习与注意力机制的异构体功能预测方法[J]. 智能系统学报, 2025, 20(6): 1508-1519.
GUO Maozu, ZHOU Aoyu, DUAN Ran. Isoform function prediction based on attention mechanism and multiple instance learning[J]. *CAAI Transactions on Intelligent Systems*, 2025, 20(6): 1508-1519.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202410005

DOI: 10.11992/tis.202410005
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20250923.0958.002

融合多实例学习与注意力机制的异构体功能预测方法

郭茂祖[1,2], 周遨宇[1,2], 段然[1,2]
（1. 北京建筑大学 智能科学与技术学院, 北京 102616; 2. 北京建筑大学 城市建筑超级智能技术北京市重点实验室, 北京 102616）

摘　要: 基因功能的高分辨率注释是功能基因组学的核心任务。单个基因可变剪接产生的异构体（isoform）翻译出多种蛋白质变体, 为生物体提供了功能多样性。为实现异构体功能的高分辨率注释, 本文提出了一种方法 LossIsoFun。引入基因本体（gene ontology, GO）, 并利用图卷积神经网络（graph convolutional network, GCN）保留其层次结构和语义信息, 通过 GO 网络嵌入策略获得压缩的基因 GO 注释。融合异构体互作网络、共表达网络和序列相似性网络, 构建异构体功能网络, 并将异构体序列数据与功能网络输入 GCN, 获取异构体功能的低维表示。通过基因与异构体的关联关系, 得到基因功能的低维表示。提出一种基于注意力权重的损失函数, 通过最小化压缩的基因 GO 注释与基因功能低维表示之间的差异来训练模型。通过解压缩异构体的低维表示, 获得异构体的高分辨率注释。在人类基准数据集上的对比实验验证了 LossIsoFun 的有效性。

关键词: 基因功能; 高分辨率注释; 异构体功能; 图卷积神经网络; 基因本体嵌入; 异构体互作网络; 融合网络; 注意力权重; 损失函数
中图分类号: TP181　　**文献标志码:** A　　**文章编号:** 1673−4785(2025)06−1508−12

中文引用格式: 郭茂祖, 周遨宇, 段然. 融合多实例学习与注意力机制的异构体功能预测方法 [J]. 智能系统学报, 2025, 20(6): 1508–1519.
英文引用格式: GUO Maozu, ZHOU Aoyu, DUAN Ran. Isoform function prediction based on attention mechanism and multiple instance learning[J]. CAAI transactions on intelligent systems, 2025, 20(6): 1508–1519.

Isoform function prediction based on attention mechanism and multiple instance learning

GUO Maozu[1,2], ZHOU Aoyu[1,2], DUAN Ran[1,2]
(1. School of Intelligence Science and Technology, Beijing University of Civil Engineering and Architecture, Beijing 102616, China; 2. Beijing Key Laboratory of Super Intelligent Technology for Urban Architecture, Beijing University of Civil Engineering and Architecture, Beijing 102616, China)

Abstract: High-resolution annotation of gene functions is essential in functional genomics. Multiple isoforms are generated from a single gene via alternative splicing, thereby producing protein variants that contribute to functional diversity. This paper introduces LossIsoFun, a framework for high-resolution isoform function annotation. First, gene ontology (GO) and a graph convolutional network (GCN) are used to preserve hierarchical and semantic structures, producing compressed GO annotations. Then, isoform interaction, coexpression, and sequence similarity networks are integrated to construct an isoform functional network. The isoform sequence data and functional network are fed into a GCN to generate low-dimensional isoform representations. By leveraging gene–isoform relationships, gene function representations are derived. A novel loss function minimizes differences between compressed GO annotations and gene function representations. Finally, isoform functions are annotated by decompressing these representations. Validation on human benchmark datasets demonstrates that LossIsoFun effectively yields isoform function annotation.

Keywords: gene functions; high-resolution annotation; isoform functions; graph convolutional network; gene ontology embedding; isoform interaction network; fusion network; attention-weighted; loss function

异构体是指由同一个基因通过不同的选择性剪接、不同的转录起始位点, 或不同的翻译起始位点生成的多种核糖核酸（ribonucleic acid, RNA）

收稿日期: 2024−10−09.　**网络出版日期:** 2025−09−23.
基金项目: 国家自然科学基金重点项目（62031003）; 国家自然科学基金青年基金项目（62301021）.
通信作者: 段然. E-mail: duanran@bucea.edu.cn.

变体。尽管这些变体来自同一个基因, 它们的序列、结构以及功能可能有所不同。异构体在基因表达的调控中发挥重要作用, 能够赋予同一基因在不同细胞类型、发育阶段或环境条件下执行多种功能的能力[1], 影响超过 90% 的人类基因[2]。这一过程生成多种蛋白质, 为生物体提供了功能多样性[3]。选择性剪切的变化对细胞功能有重大影响, 并与多种疾病相关[4]。研究发现, 同一基因的异构体可能在功能上存在显著差异[5-6], 甚至相反[7]。例如, FGFR2 基因通过选择性剪切产生 FGFR2-IIIb 和 FGFR2-IIIc 两种异构体, 前者主要在上皮细胞中表达, 参与皮肤和外胚层组织的发育, 后者在间充质细胞中表达, 涉及骨骼和肌肉的发育。这些异构体对不同的成纤维细胞生长因子具有不同的结合特异性, 从而在不同组织和器官中发挥不同的功能[8]。类似地, VEGFA 基因的异构体 VEGF165 和 VEGF121 分别具有高效和低效的血管生成作用[9], 而 CD44 基因的异构体在细胞黏附、迁移和信号传导中表现出不同的作用[10]。Bcl-x 基因则通过选择性剪切产生抗凋亡蛋白 Bcl-xL 和促凋亡蛋白 Bcl-xS, 它们分别在细胞凋亡中具有对立的功能[11]。因此, 精确发现异构体的功能对于揭示基因和蛋白质功能的分子基础至关重要。

GO 是一个重要的生物信息学工具[12], 旨在为所有物种的基因和基因产物的功能属性提供统一的描述。GO 由 GO 联盟(gene ontology consortium)开发和维护, 目前已包含超过 45 000 个术语, 分为生物过程(biological process, BP)、分子功能(molecular function, MF)和细胞成分(cellular component, CC)3 个子本体。GO 术语通过有向无环图(directed acyclic graph, DAG)组织, 每个术语都有一个唯一的标识符; 在生物信息学中, DAG 结构允许术语拥有多个父术语和子术语, 灵活地表达复杂的层次关系。例如, 在 GO 中, GO:0051171(氮化合物代谢过程的调控)同时是 GO:0019222(代谢过程的调控)和 GO:0006139(含核碱基化合物代谢过程)的子术语, 反映了生物学概念的复杂性和层次性。GO 在功能注释、预测、数据整合和富集分析中广泛应用。早期方法通常将 GO 术语视为平面标签, 通过二元或多类分类方法预测基因产物的 GO 注释。近年来, 研究者开始利用 GO 的层次结构对异构体的功能进行预测。例如, Zhao 等[13] 利用 GO 的层次结构执行异步随机游走以预测蛋白质与 GO 术语之间的关联, 从而提高异构体功能预测准确性; Zhao 等[14] 使用层次保留哈希技术来保持 GO 术语之间的层次顺序, 以预测基因功能; Yu 等[15] 采用矩阵分解技术将 GO 术语压缩到低维空间中, 这些工作通过压缩大量 GO 术语以高效预测异构体功能。

为了提高异构体功能预测的准确率, 一些研究人员提出了基于多实例学习(multiple instance learning, MIL)的方法来预测异构体功能[16]。例如, DIFFUSE 结合深度神经网络和条件随机场来预测异构体功能[17]; iMILP[18] 和 IsoFun[19] 使用网络传播将基因标签传播到异构体, 得到异构体标签; Deep-IsoFun[20] 引入邻域自适应(domain adaptation, DA)将基因功能迁移到异构体, 得到异构体功能; IsoResolve[21] 结合偏最小二乘(partial least squares, PLS)回归和 DA 对齐基因域和异构体结构域来得到异构体功能; IsoFunGo[22] 使用 GO 嵌入方法保留 GO 术语的层次结构以准确预测异构体功能。

尽管如此, 异构体功能的预测准确率仍然有待提高。本文提出一种异构体功能预测方法 LossIsoFun, 首先使用 GO 嵌入技术得到基因的压缩 GO 注释, 保留了 GO 注释的层次结构和语义信息, 同时降低了预测负担。随后, 构建异构体功能网络, 以学习异构体的特征表示。最后, 使用基于注意力机制的 MIL 网络, 将异构体的特征表示聚合得到基因的特征表示, 并提出基于注意力权重的损失函数; 利用该损失函数, 最小化基因的压缩 GO 注释和基因的特征表示的差异来训练模型。在人类的基准数据集上的实验结果表明, LossIsoFun 的性能优于现有方法, 更具有可解释性并加快了运行速度。

1　LossIsoFun 的相关工作

1.1　数据收集

本文从 NCBI SRA 数据库收集了 384 个人类 RNA-seq 数据[22], 并对其组织类型进行统计分析, 结果如表 1 所示。此外, 分别从健康/疾病状态及实验处理方式两个层面对其生物学条件进行了分析。其中, 健康组样本 200 例, 疾病组样本 184 例; 150 例样本未经过任何处理, 90 例样本接受了药物处理, 72 例样本进行了基因敲除实验, 其余 72 例样本接受了其他类型的处理。上述统计分析表明, 该数据集在不同组织和条件下的分布较为均衡。通过数据处理[23], 得到人类 9 003 个基因和 32 769 个异构体。本文使用该数据集评估 LossIsoFun 模型并与其他方法进行性能对比。

表 1　组织类型数量分布

Table 1　Distribution of tissue type counts

组织	数量	组织	数量
大脑–皮层	12	肾脏–髓质	9
大脑–小脑	8	肾脏–肾小管	8
大脑–海马	7	胃–胃底	8
大脑–下丘脑	6	胃–幽门	8
大脑–杏仁核	4	胃–胃体	7
脊髓	3	胰腺–胰岛	7
肝脏–左叶	12	胰腺–外分泌腺泡	7
肝脏–右叶	12	胰腺–胰管	6
肝脏–胆管	6	肠道–小肠	9
肝脏–门静脉区	5	肠道–大肠	9
心脏–左心室	10	皮肤–表皮	8
心脏–右心室	8	皮肤–真皮	7
心脏–心房	7	血液–T细胞	12
心脏–冠状动脉	5	血液–B细胞	10
肺–肺泡	10	血液–单核细胞	10
肺–支气管	10	血液–中性粒细胞	9
肺–毛细血管	9	血液–红细胞	9
肾脏–皮质	10	骨骼	12
脾脏	10	甲状腺	9
脂肪	8	睾丸	7
卵巢	7	前列腺	6
食道	6	膀胱	6
肌肉	6	胎盘	5
淋巴结	5	子宫	4
眼睛	3	耳朵	3

1.2　LossIsoFun

本文对 GO 注释数据进行了以下处理，t 个 GO 项的关系矩阵 $\boldsymbol{A}$ 是由 GO 有向无环图直接得到的，如果项 b 是 a 的直接后代，则 $\boldsymbol{G}(a,b)=1$，否则 $\boldsymbol{G}(a,b)=0$。对于 m 个基因的 GO 注释 $\boldsymbol{Y}\in\mathbf{R}^{m\times t}$，如果 b 或 b 的后代被正向注释到基因 g，则 $\boldsymbol{Y}(g,t)=1$，否则 $\boldsymbol{Y}(g,t)=0$。对于基因和异构体的关联矩阵 $\boldsymbol{B}\in\mathbf{R}^{m\times n}$，如果异构体 iso 是由基因 g 剪切得到，则 $\boldsymbol{B}(g,\text{iso})=1$，否则 $\boldsymbol{B}(g,\text{iso})=0$

1.2.1　LossIsoFun 模型框架

LossIsoFun 包括 GO 嵌入、Isoform 数据融合、Loss-MIL 3 个模块。如图 1 所示。

Go 嵌入：LossIsoFun 对 GO 结构和文本信息进行处理，得到 GO 术语的低维表示，以保留 GO 的层次结构并且减少预测负载，然后将 GO 嵌入基因注释中得到压缩的基因 GO 注释。

Isoform 数据融合：LossIsoFun 利用 RNA-seq 数据集[24]、异构体序列数据[25] 和异构体互作网络融合得到异构体功能网络，并对异构体功能网络和异构体特征矩阵输入 GCN 中，得到异构体的 d 维表示。

Loss-MIL：LossIsoFun 对异构体的 d 维表示进行聚合，得到基因聚合矩阵，并提出了一个基于注意力机制的损失函数，通过最小化 m 个基因的压缩 GO 注释 $\hat{\boldsymbol{Y}}$ 和基因的聚合矩阵 $\bar{\boldsymbol{Y}}$ 的差异来训练模型。最终，通过训练完成的解码器来预测异构体功能。

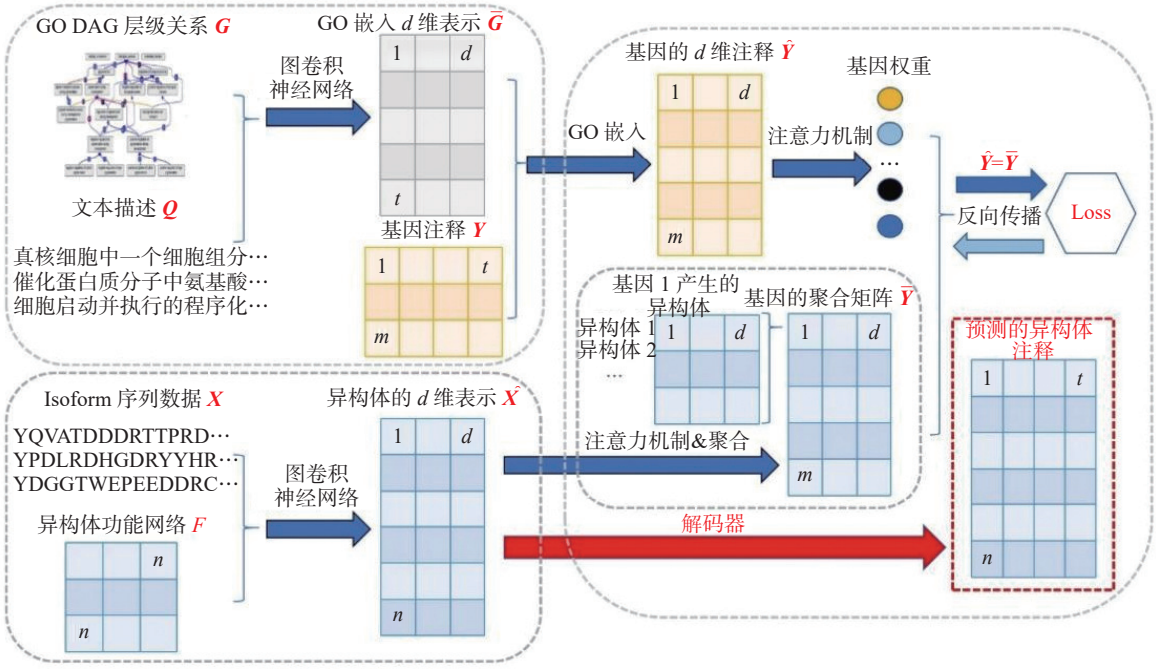


图 1　LossIsoFun 示意

Fig. 1　Schematic of LossIsoFun

1.2.2 GO 嵌入

基因和异构体可以看作 MIL 中的包和实例[18], 因此异构体功能预测问题被考虑为多个二进制 MIL 问题[17], 而忽略了 GO 术语的层次结构。随着技术的进步, 一些研究人员提出利用哈希[14] 及矩阵分解[15,26] 技术以考虑 GO 的层次结构, 取得了不错的效果。然而, GO 术语的大量文本语义并没有考虑在内。本文在保留 GO 层次结构的前提下, 引入 GCN(graph convolutional network)[27] 融合 GO 的文本语义和层次结构。

GCN 是一种有效的工具, 可用于融合节点属性和网络拓扑来学习节点表示[28]。本文输入无向的 GO DAG $\hat{\boldsymbol{G}}$和语义数字向量$\hat{\boldsymbol{Q}}$到 GCN 中, 得到低维的t个 GO 项嵌入表示$\bar{\boldsymbol{G}}$。首先, 使用 SimCSE (simple contrastive learning of sentence embeddings) 方法嵌入不同长度的文本数据[29], 该方法通过引入噪声数据增强和对比学习, 使得模型能够生成更高质量的句子嵌入, 并将文本语义$\boldsymbol{Q}$变成语义向量$\hat{\boldsymbol{Q}} \in \mathbf{R}^{t\times d}$, 公式为

$$\hat{\boldsymbol{Q}} = \mathrm{SimCSE}(\boldsymbol{Q}) \tag{1}$$

随后, 处理 GO DAG 得到无向版本的$\hat{\boldsymbol{G}}$, 将处理之后的$\hat{\boldsymbol{G}}$和$\hat{\boldsymbol{Q}}$输入到 GCN 中获得$\bar{\boldsymbol{G}}$:

$$\bar{\boldsymbol{G}} = \mathrm{GCN}(\theta_1, \hat{\boldsymbol{Q}}, \bar{\boldsymbol{G}}_1) \tag{2}$$

式中: θ_1是 GCN 的参数, $\bar{\boldsymbol{G}}_1$是对称归一化图拉普拉斯算子得到的矩阵$\hat{\boldsymbol{G}}$, $\bar{\boldsymbol{G}}_1 = \hat{\boldsymbol{D}}^{-1/2}\left(\boldsymbol{I}+\hat{\boldsymbol{G}}\right)\hat{\boldsymbol{D}}^{-1/2}$。为保留自信息, 本文将添加的自连接的邻接矩阵定义为$\hat{\boldsymbol{D}} = \boldsymbol{I}+\hat{\boldsymbol{G}}$, 其中$\boldsymbol{I}$是单位矩阵, 最终$\bar{\boldsymbol{G}}$是$t$个 GO 项的$d$维表示。

然而, 这种方式得到的$\bar{\boldsymbol{G}}$并不能很好地代表 GO 的层次结构[30], 而 GO 的层次结构在预测过程中起重要作用。本文引入了 Lin 相似度[31] 来衡量两个层次组织术语之间的相似度。例如: GO 术语t_1是t_2和t_3最近的共同祖先, 则t_2和t_3的层次相似度$H_{\mathrm{sim}}(t_2,t_3)$计算公式为

$$H_{\mathrm{sim}}(t_2,t_3) = \frac{2\times \mathrm{IC}(t_1)}{\mathrm{IC}(t_2)+\mathrm{IC}(t_3)} \tag{3}$$

$$\mathrm{IC}(t) = 1 - \frac{\log(1+|\mathrm{desc}(t)|)}{\log\tau} \tag{4}$$

式中: $\mathrm{IC}(t)$是t的层次信息含量, $|\mathrm{desc}(t)|$是t的后代 GO 术语的数量, τ是所有考虑术语的数量。t的后代越多, t的后代所涵盖的功能越广泛; t的信息量越少, $\mathrm{IC}(t)$也就越小。因此, 如果t_1靠近t_2和t_3但离根术语较远, 那么层级相似度就大, 否则相似度就小。因此, $H_{\mathrm{sim}}(t_2,t_3)$可以捕捉到$s$个 GO 术语之间的层次关系。然而, H_{sim}可能会错过祖先和后代 GO 项之间的方向信息。这里本文只考虑每个节点的后代, 通过式 (3) 构建一个非对称矩阵$\boldsymbol{H}^a_{\mathrm{sim}}$。即当且仅当术语$s$是术语$t$的后代时$\boldsymbol{H}^a_{\mathrm{sim}}(t,s) > 0$, 否则$\boldsymbol{H}^a_{\mathrm{sim}}(t,s) = 0$。

基于层次相似度$\boldsymbol{H}_{\mathrm{sim}}$, 本文引入一种三元组排序损失[22], 以更严格地保留 GO 层次结构。三元组排序损失衡量 3 个 GO 术语违反层次关系的程度, 定义为

$$L_{\mathrm{hp}}\left(\bar{\boldsymbol{G}},\boldsymbol{H}\right) = \sum_{t=1}^{\tau}\sum_{\boldsymbol{H}^a_{\mathrm{sim}}(t,i)>\boldsymbol{H}^a_{\mathrm{sim}}(t,j)} \max(\mathrm{dist}(t,i) - \mathrm{dist}(t,j),0) \tag{5}$$

式中: $\mathrm{dist}(t,v)$表示嵌入向量的余弦相似度$\tilde{g}_t$和$\tilde{g}_v$。通过最小化三元组损失, 可以获得优化后的$\bar{\boldsymbol{G}}$, 它保留了 GO 术语之间的多重关系, 并大大减少了标签的规模。

基于压缩 GO 注释$\bar{\boldsymbol{G}}$, 本文使用自动编码器θ_{enc}将基因的高维 GO 注释压缩为d维注释$\hat{\boldsymbol{Y}}$:

$$\hat{\boldsymbol{Y}} = \theta_{\mathrm{enc}}\left(\boldsymbol{Y},\bar{\boldsymbol{G}}\right) \tag{6}$$

式中: $\hat{\boldsymbol{Y}}$是m个基因的压缩 GO 注释。另一方面, 使用解码器θ_{dec}对$\hat{\boldsymbol{Y}}$进行解压, 得到解压之后的Y_{dec}:

$$\boldsymbol{Y}_{\mathrm{dec}} = \theta_{\mathrm{dec}}\left(\hat{\boldsymbol{Y}}\right) \tag{7}$$

式中$\boldsymbol{Y}_{\mathrm{dec}}$是$m$个基因的解压 GO 注释。同样解码器$\theta_{\mathrm{dec}}$也可以用来解压异构体功能的低维表示。

1.2.3 异构体数据融合

目前已被证实异构体的功能注释十分稀少, 为了建立异构体关联网络, Luo 等[32] 在 RNA-seq 数据上建立模型, Yu 等[23] 利用异构体序列数据, 并发现序列数据包含有助于区分单个异构体功能的重要功能位点。然而这些方法得到的关联网络仍然未能准确反映异构体功能。

具有相似表达谱特征的异构体更可能具有相似的功能。本文基于人类基准数据集的 RNA-seq 数据, 使用皮尔逊相关系数构建异构体的共表达网络F_{exp}。基于异构体序列数据, 使用 BLAST 构建序列相似性网络F_{seq}[33]。随后, 从 STRING 数据库下载得到公开的蛋白质间互作网络, 使用互作关系强的网络(相互作用强度得分大于 900), 并将其映射到异构体上, 得到异构体互作网络F_{iso}。在融合网络方面, NEMO[34] 和相似性网络融合(similarity network fusion,SNF)[35] 取得了良好的结果, 本文引入 SNF 方法对F_{exp}、F_{seq}和F_{iso}进行融合, 得到功能网络F:

$$F = \mathrm{SNF}\left(F_{\mathrm{exp}},F_{\mathrm{seq}},F_{\mathrm{iso}}\right) \tag{8}$$

SNF 通过基于相似性的加权更新机制, 使不同网络的信息在迭代过程中不断优化, 同时对相似性矩阵进行归一化处理, 以避免某一网络对最终融合结果产生过大影响。对于冲突信息, SNF 采用基于局部相似性的 k 近邻策略, 使信息在不同网络之间进行动态传递, 从而减少单一网络异

常值的影响。此外，SNF 通过多轮迭代实现对各网络权重的自适应调整，确保最终融合的综合相似性矩阵能够兼顾不同来源的信息，形成更加稳健的异构体功能网络。因此，SNF 通过信息传递与权重调整等策略，在数据融合过程中有效降低了冗余信息的影响，并缓解了不同网络之间可能存在的冲突，使融合结果更具生物学意义和可靠性。

本文对异构体序列数据经过 K-mer 处理得到异构体的特征矩阵 $\boldsymbol{X}$（此处 K=3，d=8 000），然后将异构体的特征矩阵 $\boldsymbol{X}$ 和功能网络输入到 GCN 中，得到异构体的 d 维表示 $\hat{\boldsymbol{X}}$：

$$\hat{\boldsymbol{X}} = \mathrm{GCN}\left(\theta_2, \boldsymbol{X}, \hat{F}\right) \tag{9}$$

式中：θ_2 为异构体嵌入的参数，$\hat{F}$ 为 F 归一化后的网络，并且保持 $\bar{\boldsymbol{G}}$ 和 $\hat{\boldsymbol{X}}$ 的维度保持一致。

1.2.4　Loss-MIL

在多实例学习中，训练数据由多个包组成，每个包分成若干个实例，但只有一个标签。如果若干实例中至少有一个是正类，则这个包的标签为正样本；如果所有实例均为负类，则这个包的标签为负样本。传统多实例学习通过实例是否为正类来判断包的标签是否为正样本[36]。然而基因−异构体关联 $\boldsymbol{B}$ 是利用基因（包）的注释合理分配到单个异构体（实例）的重要桥梁。Wang 等[26] 在利用此关联时，采用最大池化和矩阵分解等 MIL 策略，对异构体的功能进行预测，但他们通常假设基因的功能仅与单一异构体有关[37]，这违背了生物学事实。例如，SR45 基因通过可变剪接产生的两种异构体 SR45.1 和 SR45.2 在拟南芥的发育和开花时间调节中共同发挥作用[38]。尽管它们在结构上有所不同，但它们在调节开花相关基因的剪接和表达方面具有相同的功能，确保植物的正常发育和生长。Shaw 等[20] 将基因的 GO 注释统一分发给所有的异构体，但是由一个基因剪切得到的异构体可能具有完全不同的功能，不符合基因−异构体关联。Qiu 等[22] 提出一种基于注意力机制的 MIL 方法来找出包的显著实例。该方法通过注意力机制找出显著实例，各个异构体的标签很大程度上取决于习得的权重，仅通过注意力机制习得的权重不具有可解释性。

针对以上不足，LossIsoFun 利用池化层来模拟生物学事实，Loss-MIL 利用池化层对局部特征进行聚合操作，不仅能有效降低模型的复杂性，还能增强基因和异构体之间的局部关联表达，同时减少噪声影响，从而提高模型的泛化能力与稳定性。同时，定义了一个基于注意力权重的损失函数，该损失函数通过权重 α 对每个样本的误差进行加权，使得某些样本对总损失的贡献更大或更小，从而使模型关注更重要的数据，让噪声数据对模型的影响更小，在处理不平衡数据时效果显著，并且可解释性更强。

根据上述得到的异构体特征矩阵 $\hat{\boldsymbol{X}}$ 和基因−异构体关联 $\boldsymbol{B}$，可以得到基因的聚合注释矩阵：

$$\bar{\boldsymbol{Y}} = \mu\left(\hat{\boldsymbol{X}}, \boldsymbol{B}\right) \tag{10}$$

本文通过池化层 μ 和 $\boldsymbol{B}$ 来聚合 n 个异构体的潜在压缩注释。同理，可以根据基因的聚合注释 $\bar{G}$ 解压缩分配给异构体。本文提出的基于注意力权重的损失函数，通过最小化 m 个基因的压缩 GO 注释 $\hat{\boldsymbol{Y}}$ 和基因的聚合矩阵 $\bar{\boldsymbol{Y}}$ 来训练模型。

$$\alpha = \mathrm{SoftMax}\left(\mathrm{Linear}\left(\mathrm{Tanh}\left(\mathrm{Linear}\left(\hat{\boldsymbol{Y}}\right)\right)\right)\right) \tag{11}$$

$$L_{\mathrm{mil}}\left(\hat{\boldsymbol{Y}}, \bar{\boldsymbol{Y}}\right) = \left\| \left(\hat{\boldsymbol{Y}} - \bar{\boldsymbol{Y}}\right) \times \alpha \right\|_2^2 \tag{12}$$

式中 α 为 m 个基因的注意力权重。为此，本文将 LossIsoFun 的损失函数定义为

$$L = L_{\mathrm{mil}}\left(\hat{\boldsymbol{Y}}, \bar{\boldsymbol{Y}}\right) + \omega L_{\mathrm{hp}}\left(\bar{\boldsymbol{G}}, \boldsymbol{H}\right) \tag{13}$$

式中 ω 用于平衡两个模块。通过最小化上述损失函数，训练异构体功能预测模型。

2　实验结果与验证

2.1　实验相关设置

本文使用人类基准数据集，以及收集的基因/异构体水平注释来验证 LossIsoFun 的性能。对数据的研究发现，大部分 GO 术语注释小于 50 个基因[17, 23, 32, 37]。然而，目前大部分已有方法只能处理注释量为几十个的 GO 术语，忽视了注释量较大的 GO 术语。为了测试 GO 术语的注释数量对 LossIsoFun 性能的影响，本文根据 GO 术语的注释数量将 GO 术语分为 [3,50), [50,100), [100,300) 3 个区间。

在异构体功能预测中，功能性异构体（正样本）的数量可能显著少于非功能性异构体（负样本）。这种数据不平衡会影响模型评估的准确性。在实际应用中，正确预测功能性异构体的能力通常更重要。例如，错误地将功能性异构体预测为非功能性可能导致关键生物学信息的遗漏。结合异构体功能预测问题的特点，受试者工作特征曲线下面积（area under the receiver operating characteristic curve, AUROC）和精确率−召回率曲线下面积（area under the precision-recall curve, AUPRC）可以从全局和局部两个层面综合评估模型性能：AUROC 提供全局性能评估，帮助了解模型在区分所有样本时的总体能力；AUPRC 更关注正样本预测能力，在功能性异构体稀少的情况下，AUPRC 更能反映模型实际应用场景下的效用。因此，本文使用 AUROC 及 AUPRC 两个常用指标来综合评估模型的预测准确率。两个指标的值越高，代表预测的准确度越高。

本文与目前已有的 6 种方法进行对比和验证：IsoFun[19]、DisoFun[26]、IsoResolve[21]、DIFFUSE[17]、DMIL-IsoFun[23]、IsoFunGo[22]。上述所有方法使用 LossIsofun 相同的数据集进行训练，并按照各个方法共享的参数配置。对于 LossIsoFun，本文的实验参数设置如表 2 所示。

表 2　实验参数设置

Table 2　Experimental parameter settings

参数	默认值
共表达网络的最近邻数设置k	5
异构体的嵌入维度embedded_d	256
训练批次大小batch_size	256
损失函数中参数ω	7
训练轮数epoch	50
学习率learning_rate	0.01

在神经网络的训练过程中，本文对共表达网络中的最近邻数 k 的选取进行优化，通过加权基因共表达网络分析（weighted gene co-expression network analysis, WGCNA）方法，调整 k 值并结合软阈值优化，计算异构体之间的拓扑重叠矩阵（topological overlap matrix, TOM），从而确定最佳 k 值[39]。结果发现当 k=5 时，网络的划分最为合理，且网络稳定性最好，所以选 k=5。根据本文使用数据集的规模大小，将异构体和 GO 术语的嵌入维度设置为 256，训练批次大小设置为 256，并将训练轮数（epoch）设置为 50。当 ω=7 时，模型训练过程中的损失值最小，因此本文将 ω 设置为 7，在初始训练阶段，学习率设为 0.001。随着训练的进行，本文通过标准的学习率步长衰减方法更新学习率[20]。如果发现学习发散（如观察到非常大的损失值），将初始学习率改变一个数量级，直至收敛为止，最终选定学习率为 0.01。

2.2　与现有方法比较

本文采用 80%/10%/10% 的比例随机分配训练集、测试集和验证集，进行 50 轮独立实验，并确保同一基因的所有异构体在每轮中都被划分为同一组。此外，从 eggNOG 数据库[40] 中获得了直系同源蛋白质组（COGs）的簇，并进一步确保属于同一 COG 的同源基因被划分为同一组。本文使用配备 16 核 32 线程的 Intel (R) Xeon(R) Gold 6135 CPU @ 3.40 GHz 和 4 块 NVIDIA Tesla V100S_PCIe _ 32 GB 显卡的 Linux 服务器进行训练。

本文使用 GO 的 3 个类别分别进行实验：生物学过程（biological process, BP）、细胞组分（cellular component, CC）和分子功能（molecular function, MF）。BP 涉及基因产物参与的生物学过程，CC 描述基因产物所在的细胞或亚细胞部分，MF 指基因产物在分子水平上的具体功能。通过分析上述 3 个类别的 GO 术语，能够更全面地评估 LossIsoFun 在不同层次上的性能，并为进一步优化模型提供依据。本文在上述 3 个类别进行实验，结果如表 3~5 所示。

表 3　人类数据集 BP 过程的异构体预测结果

Table 3　Isoform prediction results for human dataset in the BP process

方法	[3,50)		[50,100)		[100,300)	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
IsoFun[19]	0.5543	0.0030	0.5497	0.0089	0.5203	0.0204
DisoFun[26]	0.6161	0.0036	0.5931	0.0075	0.5377	0.0188
IsoResolve[21]	0.6108	0.0153	0.5975	0.0132	0.5585	0.0243
DIFFUSE[17]	0.5842	0.0201	0.5592	0.0308	0.5517	0.0401
DMIL-IsoFun[23]	0.6077	0.0238	0.6937	0.0202	0.6168	0.0328
IsoFunGo[22]	0.6171	0.0328	0.6573	0.0663	0.7080	**0.0936**
LossIsoFun	**0.6331**	**0.0414**	**0.7315**	**0.0786**	**0.7342**	0.0893

注：加粗为本列最优结果。

表 4　人类数据集 CC 过程的异构体预测结果

Table 4　Isoform prediction results for human dataset in the CC process

方法	[3,50)		[50,100)		[100,300)	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
IsoFun[19]	0.6693	0.0379	0.7098	0.0981	0.7060	0.1208
DisoFun[26]	0.5639	0.0038	0.5291	0.0056	0.5331	0.0191
IsoResolve[21]	0.5891	0.0041	0.5501	0.0097	0.4922	0.0193
DIFFUSE[17]	0.6013	0.0129	0.6177	0.0133	0.5614	0.0281
DMIL-IsoFun[23]	0.5943	0.0719	0.5859	0.0887	0.5987	0.1105
IsoFunGo[22]	0.6918	0.1635	0.7671	0.2948	**0.7285**	**0.2878**
LossIsoFun	**0.7371**	**0.1689**	**0.7862**	**0.2963**	0.7171	0.2533

注：加粗为本列最优结果。

表 5　人类数据集 MF 过程的异构体预测结果

Table 5　Isoform prediction results for human dataset in the MF process

方法	[3,50)		[50,100)		[100,300)	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
IsoFun[19]	0.728 1	0.133 1	0.551 7	0.008 0	0.531 0	0.020 1
DisoFun[26]	0.547 0	0.004 2	0.575 2	0.009 1	0.527 0	0.020 3
IsoResolve[21]	0.661 1	0.007 1	0.581 0	0.008 1	0.552 0	0.031 0
DIFFUSE[17]	0.672 1	0.039 0	0.591 1	0.035 3	0.578 0	0.047 5
DMIL-IsoFun[23]	0.692 3	0.019 1	0.695 0	0.021 0	0.632 0	0.038 7
IsoFunGo[22]	0.817 1	0.239 2	0.825 0	0.253 0	0.837 0	0.377 0
LossIsoFun	**0.823 7**	**0.261 4**	**0.845 8**	**0.357 6**	**0.887 8**	**0.492 2**

注：加粗为本列最优结果。

分析表 3~5 可知，LossIsoFun 在上述 3 组区间内准确率普遍优于目前已有的方法。这是因为 LossIsoFun 在保留 GO 的层次结构和语义信息的同时，通过构建异构体互作网络，完善异构体之间的关联信息，为预测异构体功能提供了有利的条件和其在生物学研究中的意义。此外，基于注意力权重的的损失函数在训练模型时也为模型参数的更新做出贡献。IsoFun[19] 和 DisoFun[26] 对 GO DAG 进行建模以处理稀疏术语，但它们的模型只能挖掘异构体和 GO 项之间的线性关系，对非线性关系并不敏感。IsoResolve[21]、DIFFUSE[17] 和 DMIL-IsoFun[23] 执行多个二元分类任务来预测异构体的功能，它们的准确率都低于 LossIsofun，并且后两者也进行了网络融合。这表明了 LossIsofun 引入 GO 层次结构和文本语义的有效性。IsoFun[19] 和 DisoFun[26] 完全依赖于 GO 层次结构，忽略了文本语义，并且主要捕获 GO 术语和异构型数据之间的线性关系，证明了图卷积神经网络融合 GO 层次结构和文本语义的有效性。IsoFunGO[22] 虽然也对 GO 进行了与 LossIsoFun 相同的处理，但 LossIsoFun 使用了异构体互作网络和基于注意力机制的损失函数，更自然地模拟基因−异构体关联，从而在大部分实验结果上优于 IsoFunGo[22]。在区间 [100,300) 上，LossIsoFun 和 IsoFunGo[22] 准确率接近，这是由于 GO 术语注释量过大，导致这些 GO 术语所含信息量较少，并且注释量在该区间上的 GO 术语数量较少。因此，在区间 [100,300) 上准确率接近并不能否定 LossIsoFun 的优势。最后，LossIsoFun 只需要执行 d 项任务，相较于其他需要执行众多二元 MIL 任务预测异构体功能的方法，LossIsoFun 运行时间大幅减少，表明 LossIsoFun 在大规模异构体功能预测方面的效率。

为了验证 LossIsoFun 模型的泛化性，本研究从 NCBI SRA 数据库收集 40 个玉米 RNA-seq 数据[25]，进行与上述相同的处理和实验，结果如表 6~8 所示。分析表 3~5 与表 6~8 可得，LossIsoFun 在玉米数据集上的表现优于在人类数据集上的表现。这是因为人类数据集比玉米数据集更为复杂，人类基因通常比玉米基因具有更多的可变剪接的异构体，这使得预测人类异构体的功能比预测玉米异构体功能更加困难。

表 6　玉米数据集 BP 过程的异构体预测结果

Table 6　Isoform prediction results for maize data in the BP process

方法	[3,50)		[50,100)		[100,300)	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
IsoFun[19]	0.572 6	0.010 1	0.580 2	0.010 0	0.554 4	0.022 8
DisoFun[26]	0.638 2	0.011 7	0.593 1	0.011 2	0.562 7	0.021 1
IsoResolve[21]	0.640 8	0.020 3	0.597 5	0.010 8	0.578 5	0.030 8
DIFFUSE[17]	0.631 0	0.028 7	0.603 7	0.033 1	0.543 8	0.044 7
DMIL-IsoFun[23]	0.654 4	0.035 1	0.701 2	0.025 7	0.637 6	0.037 8
IsoFunGo[22]	0.673 1	0.037 7	0.690 7	0.070 1	0.739 6	0.092 1
LossIsoFun	**0.706 3**	**0.047 1**	**0.785 6**	**0.081 2**	**0.791 0**	**0.095 7**

注：加粗为本列最优结果。

表 7　玉米数据集 CC 过程的异构体预测结果

Table 7　Isoform prediction results for maize data in the CC process

方法	[3,50)		[50,100)		[100,300)	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
IsoFun[19]	0.687 1	0.040 1	0.703 4	0.090 1	0.713 2	0.130 1
DisoFun[26]	0.580 0	0.005 5	0.555 1	0.007 5	0.590 3	0.028 5
IsoResolve[21]	0.603 1	0.007 3	0.572 2	0.011 7	0.501 4	0.033 6
DIFFUSE[17]	0.624 7	0.020 8	0.633 6	0.021 4	0.563 1	0.030 9
DMIL-IsoFun[23]	0.603 8	0.073 8	0.602 4	0.090	0.600 7	0.174 6
IsoFunGo[22]	0.714 6	0.177 9	0.787 1	0.301 2	**0.735 0**	0.283 3
LossIsoFun	**0.757 7**	**0.182 6**	**0.802 4**	**0.312 9**	0.729 6	**0.293 6**

注：加粗为本列最优结果。

表 8　玉米数据集 MF 过程的异构体预测结果

Table 8　Isoform prediction results for maize data in the MF process

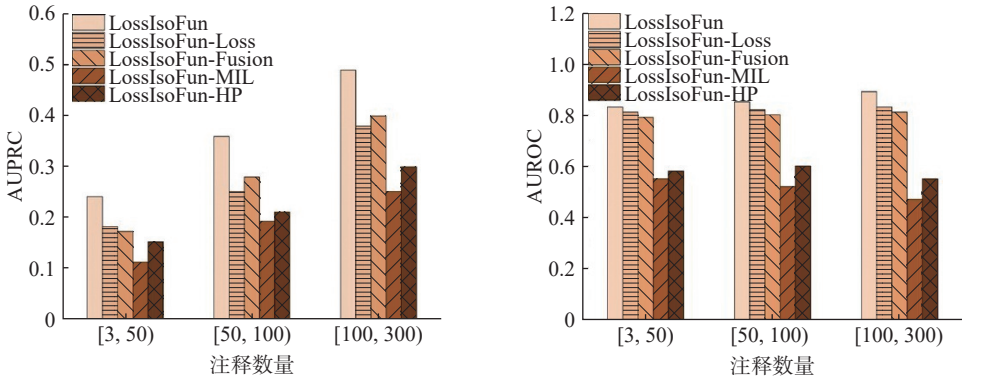
方法	[3,50)		[50,100)		[100,300)	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
IsoFun[19]	0.721 0	0.155 4	0.574 2	0.010 0	0.541 5	0.022 3
DisoFun[26]	0.583 4	0.007 8	0.569 3	0.011 6	0.530 8	0.027 6
IsoResolve[21]	0.686 9	0.009 8	0.594 4	0.009 8	0.567 9	0.037 5
DIFFUSE[17]	0.681 2	0.043 9	0.603 8	0.049 0	0.588 9	0.055 3
DMIL-IsoFun[23]	0.694 6	0.020 4	0.716 9	0.030 5	0.655 7	0.049 6
IsoFunGo[22]	0.824 1	0.239 3	0.830 5	0.273 5	0.840 7	0.391 2
LossIsoFun	**0.833 2**	**0.273 6**	**0.864 5**	**0.377 1**	**0.900 1**	**0.503 8**

注：加粗为本列最优结果。

2.3　消融实验

为了验证 LossIsoFun 模型中的各个模块及损失函数中每个部分对该模型的影响。本文引入 LossIsoFun 的 4 种变体：LossIsoFun-Loss、LossIsoFun-Fusion、LossIsoFun-MIL 和 LossIsoFun-HP。

LossIsoFun-Loss 使用均方误差代替本文中采用的基于注意力权重的损失函数；LossIsoFun-Fusion 在网络融合中不使用异构体互作网络，仅使用异构体的共表达网络和序列相似性网络进行网络融合；LossIsoFun-MIL 和 LossIsoFun-HP 是对损失函数中的权重参数进行消融实验，前者仅使用 L_{hp} 作为损失函数，后者使用 L_{mil} 作为损失函数，以验证不同部分损失项对实验结果的影响。消融实验结果如图 2 所示。



(a) LossIsoFun 及其变体在 AUPRC 上的对比　　(b) LossIsoFun 及其变体在 AUROC 上的对比

图 2　消融实验对比

Fig. 2　Ablation experiment comparison chart

图 2(a) 给出了 LossIsoFun 与其变体在 AUPRC 上的对比；图 2(b) 给出了 LossIsoFun 与其变体在 AUROC 上的对比。从图 2 中可以看出 LossIsoFun 在 AUROC 和 AUPRC 上都优于其变体，表明异构体间互作网络可以使异构体的信息更加丰富，对预测异构体的功能更加有效。此外，基于注意力权重的损失函数优于传统的损失函数，能更自然地模拟基因−异构体关联，并符合同一基因的两个或多个异构体在相同功能上合作的生物学事实。最后，损失函数两个部分的消融实验也表明了损失函数 L_{hp} 和 L_{mil} 两部分的重要性。消融实验证明了 LossIsoFun 各个部分对异构体功能预测都起到了正向的影响。

2.4　对预测结果进行验证

为进一步研究实验预测的人类异构体功能，本文收集了从 6 个基因剪接而来的 15 种人类异构体及其功能注释，与各方法的预测结果进行比较，结果如表 9 所示。这些 GO 术语描述了多种酶和蛋白质的具体功能，包括甲基转移酶活性（GO:0008170）：催化甲基基团的转移，影响基因表达和蛋白质功能；作用于蛋白质的催化活性（GO:0140101）：催化涉及蛋白质的化学反应，如蛋白激酶和磷酸酶；翻译调节因子活性（GO:0140359）：调节蛋白质合成过程，包括促进或抑制翻译的因子；氨酰 tRNA 合成酶活性（GO:0101005）：将氨基酸附着到其对应的 tRNA 上，是蛋白质合成的关键步骤；ATP 酶活性（GO:0016887）：催化 ATP 水解，释放能量用于细胞功能，如主动运输和信号传导；以及结构构成核糖体的成分（GO:0003735）：作为核糖体的一部分参与蛋白质合成[41]。

表 9　LossIsoFun 对异构体注释的预测结果（√/×）

Table 9　Prediction positive/nnegative (√/×) annotations for each isoform by each comparison method

GO术语	基因	异构体	注释	LossIsoFun	IsoFunGO[22]	DMIL-IsoFun[26]	DIFFUSE[17]	IsoResolve[21]	DisoFun[25]	IsoFun[19]
0008170	DNMT1	P26358	√	√	×	×	×	×	×	×
		K7ENW7	×	×	×	√	×	√	×	×
0140101	ELAC2	G5E9D5	√	√	√	×	×	×	×	×
		V9GZ72	√	√	√	×	×	×	×	×
		E7ES68	×	×	×	×	√	×	×	×
		H7C2I4	×	×	×	×	√	×	×	×
014359	ABCB11	A0A3B3IS78	√	√	√	√	√	×	√	×
		A0A3B3ISD4	×	×	√	√	×	×	×	×
0101005	USP19	O94966	√	√	√	√	√	√	×	√
		A0A0A0MR08	×	×	×	×	×	×	×	×
0016887	MCM3	A0A499FHX9	√	×	√	×	√	×	×	×
		J3KQ69	√	×	×	×	√	×	×	×
		Q7Z6P5	×	×	×	×	×	×	×	×
0003735	RPL13	J3KS98	√	√	√	√	×	×	√	×
		J3QSB4	√	√	√	×	√	√	×	×
准确率/%				86.67	80.00	46.67	60.00	46.67	53.33	46.67

注：“√”代表基因剪切得到的异构体具有该功能，“×”则代表无，下划线标记代表该方法的预测结果与实际收集得到的结果不同。

表 9 表明，LossIsoFun 正确区分了 15 个 GO 注释中的 13 个，准确率最高，可以更准确地区分从同一基因剪接的不同异构体的功能，还表明了同一基因的两种或多种异构体在相同的功能上的协同作用。此外，LossIsoFun 预测的异构体功能具有明确的生物学意义。例如，实验预测的异构体“P26358”主要负责维持脱氧核糖核酸（deoxyribonucleic acid, DNA）甲基化模式，确保在 DNA 复制过程中将原有的甲基化标记传递给新合成的 DNA 链，该异构体的功能仅通过 LossIsoFun 预测得到；异构体“A0A3B3IS78”参与肝胆酸稳态和脂质稳态；异构体“O94966”生成一种去泛素化酶，可调控多种蛋白质的降解；异构体“G5E9D5”和“V9GZ72”编码的蛋白具有线粒体 tRNA 3′末端加

工的核酸内切酶活性, 并参与 tRNA 的成熟过程[41]。

相比之下, DisoFun[26] 基于矩阵分解的解决方案大多忽略了这一事实, 并且遗漏了许多正面注释(即认定异构体具有特定功能的注释)。此外, 正面注释的基因和负面注释的基因之间在大多数 GO 术语方面存在巨大不平衡。因此, IsoResolve[21]、DisoFun[26] 和 IsoFun[19] 都倾向于预测负面注释。DMIL-IsoFun[23] 使用两个不同的网络预测异构体函数功能, 忽略了 GO 层次结构, 并且精度也较 LossIsoFun 低得多。IsoFunGO[22] 虽然也采用了 GO 层次结构, 但是其采用的注意力机制并不能合理地将基因注释分配给各个异构体。

2.5　LossIsoFun 的可解释性

为更加严谨地分析 LossIsoFun 可解释性, 本文分别从多实例学习和注意力机制两个方面展开讨论。

多实例学习在异构体功能预测中的可解释性主要体现在其对基因及其异构体的建模方式上。在多实例学习框架下, 基因可以被视为一个“包”(bag), 而异构体作为包中的“实例”(instances)。这种方式能够有效地反映基因与异构体之间的关系, 并提供生物学上的可解释性。由于基因的整体功能是已知的, 而各个异构体的具体功能可能尚不明确, 多实例学习可以基于基因的整体功能推测其异构体的潜在功能。此外, 多实例学习能够评估不同异构体对基因功能的影响, 结合注意力权重, 推测哪些异构体在特定功能中起关键作用。

LossIsoFun 在注意力机制中采用 Tanh 激活函数来计算注意力权重。Tanh 的非线性特性有助于捕捉复杂的特征交互关系, 而其对称性和归一化范围则进一步增强了注意力权重的可解释性。对称性使得注意力权重能够区分不同异构体的影响程度, 使其作用方向更加直观。与此同时, Tanh 的归一化范围 (−1,1) 有效避免了极端值的出现, 确保权重在合理区间内分布, 使得不同异构体的权重具有可比性, 避免因数值过大或过小导致注意力分配失衡。

本文对注意力权重进行可视化, 结果如图 3 所示。图中, 纵坐标代表所有基因, 横坐标代表每个基因的所有可变剪接异构体, 每一行为该基因所有可变剪切异构体的注意力权重可视化。右侧图例表示颜色从紫色(下)到黄色(上)注意力权重逐渐增大, 表示基因对异构体的影响越大。

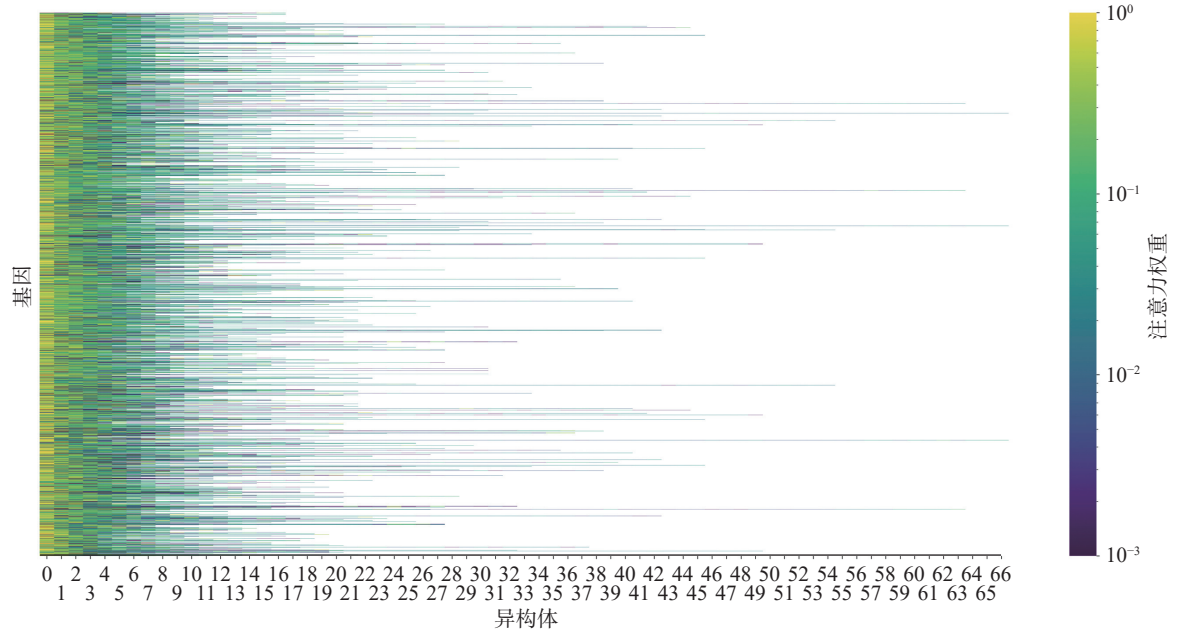


图 3　注意力权重可视化

Fig. 3　Visualization of attention weights

3　结束语

准确预测基因选择性剪接产生的异构体的功能, 有助于解析复杂疾病的机制, 并提升对功能基因组学的深入了解。本文提出了一种基于损失优化的异构体功能预测方法 LossIsoFun, 该方法首先生成 GO 术语的低维表示并将大量 GO 注释压缩为紧凑的注释, 随后将异构体表达和序列数据与异构体间互作网络进行融合, 并提出基于注意力权重的损失函数来训练模型。使用人类基准数据集验证 LossIsoFun 的有效性, 结果表明该方法提高了异构体功能预测的可解释性。

在未来工作中，异构体功能预测将更广泛地整合多种组学数据，如蛋白质组以及表观遗传组等，有助于更全面地理解异构体的功能调控机制。此外，目前许多 GO 术语和功能注释不够准确。未来将通过更精准的实验数据（如蛋白质晶体结构、功能性区域的标记等），提供更详细的异构体功能注释，提高异构体功能预测模型的准确性和实用性。

参考文献：

[1] PAN Qun, SHAI O, LEE L J, et al. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing[J]. Nature genetics, 2008, 40(12): 1413−1415.

[2] WANG E T, SANDBERG R, LUO Shujun, et al. Alternative isoform regulation in human tissue transcriptomes [J]. Nature, 2008, 456(7221): 470−476.

[3] CROWL S, COLEMAN M B, CHAPIV A, et al. Systematic analysis of the effects of splicing on the diversity of post-translational modifications in protein isoforms using PTM-POSE[EB/OL]. (2024−01−11)[2025−09−15]. https://doi.org/10.1101/2024.01.10.575062.

[4] SMITH L M, KELLEHER N L. Proteoforms as the next proteomics currency[J]. Science, 2018, 359(6380): 1106−1107.

[5] 曾杰. 基于深度多示例学习的可变剪接异构体相互作用预测研究[D]. 重庆: 西南大学, 2021.
ZENG Jie. Study on interaction prediction of alternative splicing isomers based on deep multi-instance learning [D]. Chongqing: Southwest University, 2021.

[6] HOWES A, ROGERSON C, BELYAEV N, et al. The FAM13A long isoform regulates cilia movement and coordination in airway mucociliary transport[J]. American journal of respiratory cell and molecular biology, 2024, 71(3): 282−293.

[7] MITTENDORF K F, DEATHERAGE C L, OHI M D, et al. Tailoring of membrane proteins by alternative splicing of pre-mRNA[J]. Biochemistry, 2012, 51(28): 5541−5556.

[8] GUO Miao, LIU Wei, SERRA S, et al. FGFR2 isoforms support epithelial-stromal interactions in thyroid cancer progression[J]. Cancer research, 2012, 72(8): 2017−2027.

[9] WANG Shiying, SUN Boyun, YUAN Jianye, et al. The different effects of VEGFA121 and VEGFA165 on regulating angiogenesis depend on phosphorylation sites of VEGFR2[J]. Inflammatory bowel diseases, 2017, 23(4): 603−616.

[10] HASSN MESRATI M, SYAFRUDDIN S E, MOHTAR M A, et al. CD44: a multifunctional mediator of cancer progression[J]. Biomolecules, 2021, 11(12): 1850.

[11] REVIL T, TOUTANT J, SHKRETA L, et al. Protein kinase C-dependent control of Bcl-x alternative splicing [J]. Molecular and cellular biology, 2007, 27(24): 8431−8441.

[12] ASHBURNER M, BALL C A, BLAKE J A, et al. Gene ontology: tool for the unification of biology[J]. Nature genetics, 2000, 25(1): 25−29.

[13] ZHAO Yingwen, WANG Jun, GUO Maozu, et al. Cross-species protein function prediction with asynchronous-random walk[J]. IEEE/ACM transactions on computational biology and bioinformatics, 2019, 18(4): 1439−1450.

[14] ZHAO Yingwen, FU Guangyuan, WANG Jun, et al. Gene function prediction based on gene ontology hierarchy preserving hashing[J]. Genomics, 2019, 111(3): 334−342.

[15] YU Guoxian, WANG Keyao, FU Guangyuan, et al. NMFGO: gene function prediction via nonnegative matrix factorization with gene ontology[J]. IEEE/ACM transactions on computational biology and bioinformatics, 2020, 17(1): 238−249.

[16] CARBONNEAU M A, CHEPLYGINA V, GRANGER E, et al. Multiple instance learning: a survey of problem characteristics and applications[J]. Pattern recognition, 2018, 77: 329−353.

[17] CHEN Hao, SHAW D, ZENG Jianyang, et al. DIFFUSE: predicting isoform functions from sequences and expression profiles via deep learning[J]. Bioinformatics, 2019, 35(14): i284−i294.

[18] LI Wenyuan, KANG Shuli, LIU Chunchi, et al. High-resolution functional annotation of human transcriptome: predicting isoform functions by a novel multiple instance-based label propagation method[J]. Nucleic acids research, 2014, 42(6): e39.

[19] YU Guoxian, WANG Keyao, DOMENICONI C, et al. Isoform function prediction based on bi-random walks on a heterogeneous networkFree[J]. Bioinformatics, 2020, 36(1): 303−310.

[20] SHAW D, CHEN Hao, JIANG Tao. DeepIsoFun: a deep domain adaptation approach to predict isoform functionsFree[J]. Bioinformatics, 2018, 35(15): 2535−2544.

[21] LI Hongdong, YANG Changhuo, ZHANG Zhimin, et al. IsoResolve: predicting splice isoform functions by integrating gene and isoform-level features with domain adaptation[J]. Bioinformatics, 2021, 37(4): 522−530.

[22] QIU Sichao, YU Guoxian, LU Xudong, et al. Isoform function prediction by gene ontology embedding[J]. Bioinformatics, 2022, 38(19): 4581−4588.

[23] YU Guoxian, ZHOU Guangjie, ZHANG Xiangliang, et al. DMIL-IsoFun: predicting isoform function using deep

multi-instance learning[J]. Bioinformatics, 2021, 37(24): 4818−4825.

[24] 王可尧. 基于 RNA-seq 数据的可变剪接异构体功能预测方法研究[D]. 重庆: 西南大学, 2019.
WANG Keyao. Study on function prediction method of alternative splicing isomers based on RNA-seq data[D]. Chongqing: Southwest University, 2019.

[25] SU Yaqi, YU Zhejian, JIN Siqian, et al. Comprehensive assessment of mRNA isoform detection methods for long-read sequencing data[J]. Nature communications, 2024, 15(1): 3972.

[26] WANG Keyao, WANG Jun, DOMENICONI C, et al. Differentiating isoform functions with collaborative matrix factorization[J]. Bioinformatics, 2020, 36(6): 1864−1871.

[27] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[EB/OL]. (2016−09−09)[2025−09−15]. https://arxiv.org/abs/1609.02907.

[28] 张硕. 基于图神经网络的剪接异构体功能预测方法研究[D]. 长沙: 中南大学, 2022.
ZHANG Shuo. Study on function prediction method of splicing isomers based on graph neural network[D]. Changsha: Central South University, 2022.

[29] GAO Tianyu, YAO Xingcheng, CHEN Danqi. SimCSE: simple contrastive learning of sentence embeddings [EB/OL]. (2021−04−18)[2025−09−15]. https://arxiv.org/abs/2104.08821.

[30] ZHAO Yingwen, WANG Jun, CHEN Jian, et al. A literature review of gene function prediction by modeling gene ontology[J]. Frontiers in genetics, 2020, 11: 400.

[31] LIN Dekang. An information-theoretic definition of similarity [C]//Proceedings of the Fifteenth International Conference on Machine Learning. Madison: Morgan Kaufmann Publishers Inc. , 1998: 296−304.

[32] LUO Tingjin, ZHANG Weizhong, QIU Shuang, et al. Functional annotation of human protein coding isoforms via non-convex multi-instance learning[C]//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Halifax: ACM, 2017: 345−354.

[33] ALTSCHUL S F, GISH W, MILLER W, et al. Basic local alignment search tool[J]. Journal of molecular biology, 1990, 215(3): 403−410.

[34] RAPPOPORT N, SHAMIR R. NEMO: cancer subtyping by integration of partial multi-omic data[J]. Bioinformatics, 2019, 35(18): 3348−3356.

[35] WANG Bo, MEZLINI A M, DEMIR F, et al. Similarity network fusion for aggregating data types on a genomic scale[J]. Nature methods, 2014, 11(3): 333−337.

[36] 赵璐, 袁立明, 郝琨. 多示例学习算法综述[J]. 计算机科学, 2022, 49(S1): 93−99.
ZHAO Lu, YUAN Liming, HAO Kun. A survey of multi-instance learning algorithms[J]. Computer science, 2022, 49(S1): 93−99.

[37] EKSI R, LI Hongdong, MENON R, et al. Systematically differentiating functions for alternatively spliced isoforms through integrating RNA-seq data[J]. PLoS comput biol, 2013, 9(11): e1003314.

[38] ZHANG Shijia, LIU Huili, YUAN Li, et al. Recognition of CCA1 alternative protein isoforms during temperature acclimation[J]. Plant cell reports, 2021, 40(2): 421−432.

[39] LANGFELDER P, HORVATH S. WGCNA: an R package for weighted correlation network analysis[J]. BMC bioinformatics, 2008, 9: 559.

[40] HUERTA-CEPAS J, SZKLARCZYK D, HELLER D, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses[J]. Nucleic acids research, 2019, 47(D1): D309−D314.

[41] CONSORTIUM U. UniProt: the universal protein knowledgebase in 2021[J]. Nucleic acids research, 2021, 49(D1): D480−D489.

作者简介：

郭茂祖，教授，博士生导师，北京建筑大学智能科学与技术学院院长，中国人工智能学会机器学习专委会常委、中国建筑学会计算性设计学术委员会常委，主要研究方向为机器学习、计算生物学。获吴文俊人工智能自然科学奖二等奖。发表学术论文 100 余篇。E-mail：guomaozu@bucea.edu.cn。

周遨宇，硕士研究生，主要研究方向为深度学习和生物信息学。E-mail：18336331205@163.com。

段然，讲师，主要研究方向为生物信息学、网络科学、数据挖掘、机器学习。主持国家自然科学基金青年项目 1 项。发表学术论文 8 篇。E-mail：duanran@bucea.edu.cn。