# 人类蛋白质表型标注预测算法研究展望

刘砺志

复旦大学 计算机科学技术学院

**摘要**

人类表型本体（Human Phenotype Ontology，HPO）是描述人类异常表型及其语义关系的标准化术语集，是对人类表型组进行大规模计算分析的有效工具。建立人类基因/蛋白质与异常表型之间的关联是研究遗传疾病及其表征的一项基本而又重要的工作，对遗传疾病和复杂疾病的诊断、治疗及新药的开发有着重要的帮助。当前预测人类蛋白质表型标注的研究工作主要集中于从蛋白质与 HPO 术语间关联的不同角度出发而产生的三种预测任务：（1）以蛋白质为中心的预测：确定新蛋白质的全部 HPO 注释；（2）逐对预测：识别缺失的蛋白质与 HPO 术语关系；（3）以 HPO 术语为中心的预测：对与某个 HPO 术语相关的候选蛋白质进行优选排序。除此之外，HPO 注释预测领域还有其它仍未探索的问题。本文，我们将介绍可变剪接异构体的标注预测、负样本的选取策略、噪声标注识别以及假基因的标注预测四个任务。这些问题在蛋白质功能预测领域已有研究，但在表型预测领域未有涉足。对于每个问题，我们都将首先介绍问题背景，并系统综述在功能预测领域的已有方法以及它们的优劣，最后给出在表型预测领域可能的研究方向。我们希望本文可以给后续的研究人员提供新的见解，促进表型标注预测领域不断向前发展。

## 1 背景

理解人类疾病背后的遗传机理是设计针对疾病的预防、诊断和治疗的重要一环 [1]。为了便于描述人类疾病中出现的表型异常，人类表型标准用语联盟（The Human Phenotype Ontology Project）在 2007 年提出了一个人类异常表型及其语义关系的标准化术语集，即人类表型本体（Human Phenotype Ontology，HPO）[2]。HPO 团队通过整合医学文献以及 OMIM [3]、Orphanet [4] 和 DECIPHER [5] 等数据库，构建起人类遗传疾病的 HPO 注释，并通过已知的基因—疾病关联，建立了人类基因的 HPO 注释。由于基因支配着蛋白质的合成并进而控制表型，破解人类基因/蛋白质与异常表型之间的联系十分重要。

* 本文撰写日期：2021 年 3 月

** 作者联系方式：liulizhi1996@gmail.com

截至 2020 年 10 月，只有 4484 个人类基因拥有 HPO 标注，仅占目前已知的蛋白质编码基因数的约四分之一。可是，在实际研究中，要确定基因/蛋白质和表型之间的关系，往往需要投入大量的人力、物力和财力。在后基因组时代，随着高通量测序技术的发展，海量的蛋白质序列被测定。在 2018 ～ 2020 年间，仅新增约 800 个蛋白质完成 HPO 标注。于是，在已标注和未标注蛋白质数量间形成了一条鲜明的鸿沟，单纯依靠手工标注蛋白质的表型注释显得力不从心。因此，开发一种准确高效的自动化人类蛋白质表型标注预测工具变得尤为迫切。

从机器学习的角度看，当前的人类蛋白质表型标注预测算法可以从对蛋白质—表型关联的不同视角分为三大类：（1）以蛋白质为中心（protein-centric）：确定新蛋白质（或完全未被标注的蛋白质）的全部 HPO 注释，可抽象为层次化多标签分类（hierarchical multi-label classification）问题，现有模型包括 [6-12]；（2）逐对预测（pairwise）：识别缺失的蛋白质-HPO 术语关系，可抽象为矩阵填充（matrix completion）或链接预测（link prediction）问题，现有算法包括 [13-18]；（3）以 HPO 术语为中心（term-centric）：对与某个 HPO 术语相关的候选蛋白质进行优选排序（candidate protein prioritization），可抽象为二元分类（binary classification）问题。该任务先前主要集中于功能注释预测领域，如 [19-26]。最近，刘砺志等人 [27] 基于深度图神经网络提出了第一个以 HPO 术语为中心的相关蛋白质预测算法。相关方法的综述参见文献 [28]。

尽管这些工作在各自领域都取得了当前最优的性能，但是目前仍存在不少尚未探索的空白领域。本文，我们选取了可变剪接异构体的标注预测、负样本的选取策略、噪声标注识别以及假基因的标注预测四个方面进行介绍。这些领域均已在蛋白质功能注释预测领域有所探索，但未见表型标注预测的相应研究。针对每个预测任务，我们都将首先介绍问题背景，并对在功能注释预测领域的已有算法进行系统综述，最后指出在表型预测领域可能的研究方向。本文旨在通过抛出未解问题、分析他山之石、点明探索之路，为感兴趣的研究者给予启迪，也为表型标注预测研究的不断发展添砖加瓦。

## 2 可变剪接异构体

蛋白质异构体（protein isoform）是起源于单个基因或基因家族的一组高度相似的蛋白质中的一员，是遗传差异的结果。尽管许多蛋白质异构体具有相同或相似的生物学作用，但某些异构体却具有独特的功能。一组蛋白质异构体可由单个基因的可变剪接（alternative splicing，AS）、可变启动子使用（variable promoter usage）或其他转录后修饰（post-transcriptional modification）形成，其中可变剪接是产生蛋白质异构体的主要机制。通过 RNA 剪接机制，mRNA 能够选择基因的不同蛋白质编码片段（外显子），甚至从 RNA 中选择外显子的不同部位以形成不同的 mRNA 序列。编码蛋白质的 mRNA 的选择性剪接是真核基因表达中控制蛋白质正常功能的重要调节机制，也与线粒体和各种离子通道的生理调节有关。在人类中，95% 的多外显子基因会经历选择性剪接。由于错接会通过改变或废除重要的生理蛋白质功能而导致各种人类疾病，微调的可变剪接平衡对于人体健康至关重要。大量的文献资料突显了剪接异构体在各种疾病

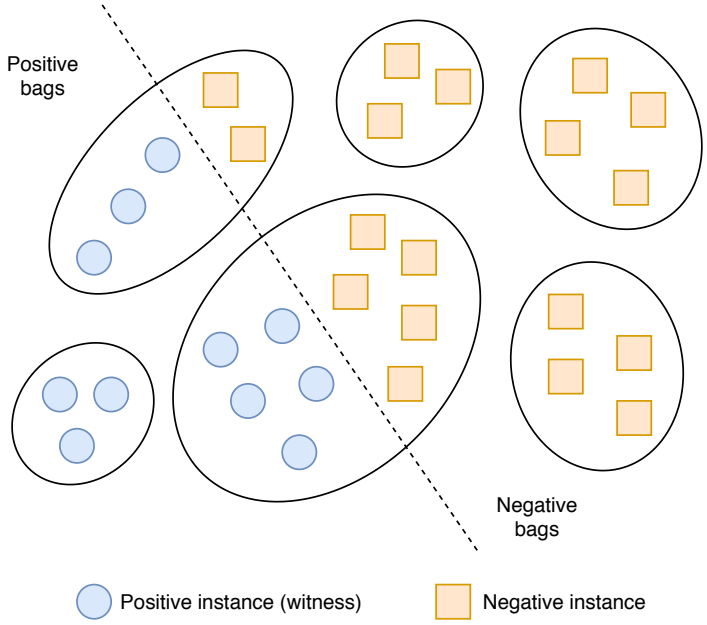


**图 1 多示例学习的框架。每个基因（椭圆框）被视为一个包，其中每个可变剪接异构体（椭圆内的圆圈或方块）视为包内的一个示例。关联于给定功能标签的基因称为正包，所有与该功能标签相关的示例（异构体）称为证人。**

中的重要性，这些疾病包括神经退行性疾病、癌症、免疫和传染性疾病、心血管疾病和代谢疾病等 [29]。

由于高通量测序技术的出现与快速发展，海量的剪接异构体水平的测序数据以及表达数据都已大量生成，但是对于异构体的功能标注和其疾病关联大多只能依靠“湿”实验人工确定，因而这些数据还极度匮乏。目前，成熟的功能注释数据库、疾病关系数据库、互作关系数据库等都主要停留于基因层面，或者说是权威异构体（canonical isoform）层面，更细化的标注信息库尚未健全。因而，不少研究人员提出了多种计算方法以实现异构体的自动化标注任务 [30]。目前，大量的算法都集中于可变剪接异构体的 GO 标注预测这一领域，近来亦出现对异构体-疾病关系的预测算法，但未见对异构体进行 HPO 标注预测的研究。

对于可变剪接异构体的 GO 标注预测，现有的算法可以归为四大类：（1）基于多示例学习（multiple instance learning，MIL）的方法 [31-35]；（2）基于深度学习的方法 [36-37]；（3）基于矩阵分解的方法 [38-39]；（4）基于迁移学习的方法 [36, 40-41]。下面我们一一简述这些算法。

图1描述了多示例学习的思想。在多示例学习框架中，基因被视为一个“包”（bag），异构体被视为包内的一个“示例”（instance），关联于给定功能标签的基因称为“正包”（positive bag），包内与该功能标签相关的所有示例称为“证人”（witness）。由此，可以假定若一个基因被某个功能所注释，则至少有一个该基因的异构体也被相应功能所注释；若一个基因没有被标注，则其所有的异构体都不被标注。算法的目标就是要识别出对基因的功能“负责”的那些剪接异构体。

Eksi 等人 [31] 提出了基于多示例支持向量机的利用 RNA-Seq 数据的算法 isoPred。作者在论文中提出了两种不同的 SVM 模型。第一种算法 mi-SVM 试图将所有正包中非证人示例

归为负样本，然后将该问题视为监督学习问题。第二种算法 MI-SVM 尝试从每个正包中一个为正标签负责的证人，然后仅基于这些证人建立分类器而将其他示例从分类过程中删除。两种算法的区别在于对间隔（margin）的定义不同：mi-SVM 寻找一个分离的超平面，使负基因的所有异构体均位于负半空间，而正半空间中的每个正基因均至少具有一个异构体，同时相对于所选标签最大化间隔；而在 MI-SVM 中，间隔的定义扩展到包层级。一个包相对于分离超平面的间隔可被定义为其示例的最大间隔。对于正包，包间隔由最正的示例定义，而对于负包，包间隔由最负的示例定义。因此，mi-SVM 中正包中的每个示例都会对间隔最大化过程产生影响，而在 MI-SVM 中，每个正包中仅考虑一个示例并仅此一个实例来确定包的间隔。

Panwar 等人 [32] 延续了 [31] 的假设，即在正基因的众多剪接异构体中，至少一个剪接异构体负责执行 GO 术语功能，而负基因的所有剪接异构体都不负责该特定功能。与 [31] 类似，所提出的基于 MIL-SVM 的算法 IsoFunc 的目标是从正基因中找出一个剪接异构体的子集，并使它们与负基因的异构体之间的差异最大化。作者使用基于“最大间隔分类”（maximum-margin-based classification）来最大化正负异构体之间差异。同样，作者选取 RNA-Seq 数据作为特征。

Luo 等人 [33] 提出了基于带权逻辑斯蒂回归（weighted logistic regression）的多示例学习模型 WLRM，将稀疏单纯形投影（即非凸稀疏感应正则器，nonconvex sparsity-inducing regularizer）与多示例学习框架集成在一起。通过在单纯形上的稀疏投影，不仅可以学习到原始基因水平特征在判别特征空间中的映射，还可以习得用以度量异构体对于给定功能的贡献大小的异构体权重向量以帮助检测关键异构体并减小正基因中负异构体带来的影响。该框架非常灵活，可以融入各种平滑和非平滑的损失函数，例如逻辑斯蒂损失和合页损失。为了解决由此产生的高度非平凡的非凸和非平滑优化问题，作者进一步开发了一种有效的块坐标下降（block coordinate descent）算法。作者在文章只使用 RNA-Seq 数据作为特征。

Li 等人 [34] 提出了基于多示例标签传播的算法 iMILP 并整合了多种从全基因组 RNA-Seq 数据构建的异构体共表达网络。整个算法流程包含两步。第一步是“网络的选择和组合”，即从所有输入的异构体共表达网络中选择与给定 GO 类别相关的网络的最佳子集，然后将它们整合为单个网络，并作为第二步的输入。第二步是“预测”，即作者所提出的多示例标签传播算法，它以组合网络为输入，并返回异构体的功能预测。不同于其它多示例标签传播算法，作者设计的标签传播规则是：在正包中，连接到更多正包内节点的节点（示例）获得更大的预测分数；而任何没有连接到其它正包内节点的节点都将被赋予零预测得分。在标签传播方法中反复应用此规则可以清楚地标识所有有资格继承包标签的示例。值得注意的是，作者构建了三标签分类问题，除了常规的正、负两种标签外，作者还为那些没有标注的基因的异构体赋予“未知”这一标签。

上述算法都将基因、异构体、GO 标注三者分开考虑，而未涉及它们间复杂的内在关联。为此，Yu 等人 [35] 提出了基于异质网络上双随机游走（bi-random walk）的 IsoFun 算法。该算法在构建的异质网络上运行定制的标签传播算法来预测异构体功能。IsoFun 首先根据从多个

RNA-Seq 数据集中收集的异构体表达谱构建一个异构体功能关系网络，然后将基因的所有注释分配给其异构体。接下来，它构建了一个由异构体、基因和 GO 术语组成的异质网络，以编码基因与异构体之间的关系、GO 术语之间的层次关系以及异构体之间的功能关联。这种异质网络可以在基因水平相互作用、可用的基因 GO 注释以及基因与异构体间的关系之间实现协同作用，从而减少不完整的单个数据源的影响。然后，IsoFun 在构造的异质网络上引入了基于双随机游走的标签传播，以预测异构体功能标注。另外，为确保基因的已知功能至少被其异构体之一所继承，IsoFun 在每次迭代中将已知功能“钳位到”（clamp）最“负责任”（responsible）的异构体上。

第二大类方法是基于深度学习技术的，这些方法大多以基因和异构体水平的特征为输入，训练一个深度神经网络，由此得到基因和异构体的功能标注预测。

之前介绍的算法都是使用多示例学习这一半监督学习技术。但是，有标签训练数据的匮乏阻碍了这些方法的性能。为此，Shaw 等人 [36] 提出了 DeepIsoFun 算法。该算法将多示例学习与域适配（domain adaptation，DA）思想结合在一起，使用 RNA-Seq 数据预测异构体的 GO 标注。在 DeepIsoFun 算法中存在两个域：基因域和异构体域。在异构体域中，每个基因看作是一个包，其所有异构体都是包中的示例。此外，单个基因具有表达信息，并有 GO 功能注释。因此，根据定义，一个基因同时是异构体域中的一个包，也是基因域中的一个示例。通过域适配技术，基因的功能和表达间的关系可以迁移到异构体域上，并由此提升了性能。DeepIsoFun 中的深度神经网络架构包含四个模块：（1）自编码器，以提取两个域的共同特征；（2）基因功能预测器，以标记每个基因的功能；（3）异构体功能预测器，以标记每个异构体的功能；（4）域标签预测，以确保知识从基因域转移到异构体域。这四个组件整体构成了一个深度前馈网络，并通过最小化基因功能预测损失和异构体功能预测以及最大化域分类预测损失来进行训练。尽管取得了一定的提升，但是作者也指出，DeepIsoFun 对于标签极度不平衡的情形处理的不是很好。

注意到之前的方法都只单单使用表达谱数据作为特征，而忽视了异构体序列本身这一可能的信息源。而异构体序列包含活性位点、结合位点、信号肽、模体和蛋白质结构域等多种潜在有用的信息。为此，Chen 等人 [37] 提出了基于深度学习的从序列和表达谱中预测异构体功能标注的算法 DIFFUSE。该算法分为两个阶段。第一阶段是一个深度神经网络，用以从异构体序列和保留域中抽取特征并获得初始预测结果。第二阶段是条件随机场，利用深度神经网络的输出和共表达信息，得到最终的预测结果。为了克服异构体水平训练标签的不足，作者提出了一种基于多示例学习框架的迭代半监督训练算法。

第三大类是基于推荐系统中矩阵分解的算法。在预测蛋白质异构体功能标注的语境下，推荐系统中的“用户”变为如基因或蛋白质等实体，而“商品”变为如功能注释等生物属性。此类方法通常将特征映射到隐空间内产生预测结果。

mFRecSys 算法 [38] 是由 Kandoi 提出的一种基于三因子分解（tri-factorization）的方法。区别于传统的矩阵分解算法直接将异构体-GO 术语关系矩阵直接进行分解，作者引入了第三

个因子矩阵以平衡异构体和 GO 术语数量上的差异以及缓解冷启动问题，并显式的导入异构体和 GO 术语特征，同时使用非线性映射。特别地，作者不仅使用了 mRNA 和蛋白质的序列信息，还首次使用了组织特异性表达数据。事实上，一些异构体是条件或组织特异的，因此，只有在特定条件下才具有功能活性。

Wang 等人 [39] 提出了另一个基于协同矩阵分解（collaborative matrix factorization）的预测算法 DisoFun。它假定基因的功能注释是从那些关键异构体的注释聚合而来。它协同将异构体表达数据矩阵和基因-GO 术语关系矩阵分解为低秩矩阵，以同时探索潜在的关键异构体，并通过将预测汇总到其导出基因来实现功能预测。此外，它利用 PPI 网络和基因本体结构进一步协调矩阵分解。

最后一类是基于迁移学习的算法。事实上，DeepIsoFun 中的域适配思想就是一种迁移学习思想。域适配思想还被 Li 等人 [40] 采用并提出了基于偏最小二乘法（partial least square）的 IsoResolve 算法。该算法将基因水平和异构体水平的特征分别视为源域和目标域。它采用域适配将两个域投影到隐变量空间中，使得来自两个域的隐变量具有相似的分布，这使得基因域信息可用于异构体功能预测。作者仅使用 RNA-Seq 数据作为特征，其中一个数据集还使用了组织特异性表达数据。

最近，Ferrer-Bonsoms 等人 [41] 提出了 ISOGO 算法。该算法采用了迁移学习的思想，可看作两个阶段。第一阶段以蛋白质水平的特征为输入，得到预测蛋白质 GO 标注的模型，该阶段又包含两个步骤。第一步是训练组件模型。其一组件模型是“相关法”（correlation method），其接收基因的共表达数据，并用 Spearman 相关系数计算基因的相似度矩阵，然后使用 Wilcoxon 检验将有给定功能标注的基因与未标注的基因进行比较，得到 Wilcoxon z-score 作为基因被给定功能所标注的可能性。其二组件模型是基于结构域的“弹性网络正则化逻辑斯蒂线性模型”（elastic-net regularized logistic linear model），其接收蛋白质结构域信息，得到蛋白质被给定功能所标注的 logit 值。第二步是通过贝叶斯逻辑斯蒂回归（Bayesian logistic regression）算法，以相关法的 z-score 和基于结构域方法的 logit 值以及它们的乘积和它们各自的平方值为输入，整合两个组件模型，得到最终的预测模型。算法的第二阶段则是将之前用蛋白质水平特征训练出的模型，改用异构体水平的特征为输入，得到异构体的 GO 注释预测。但是他们的方法没有将异构体信息融入进预测模型中，而只是将蛋白质水平特征训练出的模型直接套用到异构体水平的特征上，因而模型对于异构体功能标注的判别能力有限。

上面介绍的都是可变剪接异构体的 GO 标注预测。近来，一些研究人员开始关注异构体-疾病关系预测问题。Huang 等人 [42] 提出了一种称为 IDAPred 的基于矩阵分解的多示例学习方法，融合了基因组学和转录组学数据，来预测异构体的疾病本体（Disease Ontology，DO）标注。考虑到基因与其剪接异构体之间的包-示例关系，IDAPred 引入了分发和聚合（dispatch and aggregation）操作，以将基因-疾病关联分配给单个异构体，并将这些分配的关联反向聚集到关联的基因上。接下来，它融合了不同的基因组学（包括核酸序列和互作网络）和转录组学（包括从 RNA-Seq 数据构造的组织特异性异构体共表达网络）数据，以补充基因-疾病关联

并诱导线性分类器以连贯的方式预测异构体-疾病关联。另外，为了减轻对观察到的基因-疾病关联的偏差，作者增加了一个正则项，以将当前观察到的关联与未观察到的（潜在的）关联区分开。

最近，Yu 等人 [43] 基于深度神经网络提出了 DeepIDA 算法以预测异构体与疾病的关联。作者将该预测任务建模成二元分类问题。DeepIDA 整合了包括 mRNA 序列四元组频率和从 RNA-Seq 数据中提取的表达谱特征在内的异构体水平特征信息，以及功能注释和 miRNA-target 两种基因水平特征信息，并通过矩阵映射将基因水平的特征和疾病注释转换为异构体水平，通过前馈神经网络以及缓解标签不平衡的 Focal 损失函数 [44] 对模型进行训练，从而直接预测输入异构体关联疾病的概率。

表1对上面介绍的算法进行了总结。可以看到，现有的方法大量依赖于 RNA-Seq 表达谱数据，一些方法还使用了序列、PPI 网络、GO 的层次化结构、保守结构域等信息。它们中普遍存在以下几个问题：

第一，由于实验验证的异构体标注数据极为有限，研究人员在对模型进行评估时通常采用下面三个策略：（1）对于单异构体基因（single isoform gene，SIG），则直接将基因的标注作为异构体的标注，并对这些异构体的预测结果进行评估；（2）对于多异构体基因（multi-isoform gene，MIG）实施基因水平的评估，即将赋予其所有异构体的打分的最大值作为该基因的预测分数；（3）对于少数存在实验验证标注的异构体，通过案例分析的形式，计算在这个小数据集上的预测准确率。但是，这些折衷的评估策略不能完全替代直接在异构体水平上的全面评估，因此给出的结果都是片面的、低可依赖的，尤其是对于多异构体基因的实际预测性能目前还不得而知。

第二，少有方法涉及组织或条件特异的预测。由于一些剪切异构体是组织或条件特异的，因而它们仅能在某些特定情形下才能发挥功能。尽管 mFRecSys 和 IsoResolve 的实验中提供了有限的组织特异性评估，但是这些评估还很不全面。此外，没有方法考虑到如年龄、性别和发展阶段等协变量的特异性对功能表现的影响。

第三，现有的方法主要关注于人类和小鼠物种，而未探索其他物种上的预测。众所周知，选择性剪接的 mRNA 异构体序列通常强烈依赖于剪接机制，而剪接机制可能因物种而异。因此，在人类 RNA-Seq 数据集上训练的预测模型可能会学习到特定于人类（和紧密相关物种）异构体的潜在标注规律，但可能无法准确预测拥有不同拼接机制的物种（如植物）的异构体注释。当前的计算方法可以通过使用来自多个物种的数据训练模型，使它们的预测更加准确、稳健和通用。

第四，机器学习方法中正负样本的选取对模型的训练至关重要。但是由于生物学数据中常常只有正样本标注，而鲜见负样本标注，这就给负样本的制作带来困难。现有的算法中，大多使用随机抽取的未标注基因作为负样本，并由此将问题建模为二分类问题。但事实上，未标注基因可能只是囿于现有的生物医学技术而尚未观测到与给定功能标签的关系，但不代表两者之间必然不存在关联。因此，这样的负样本选取可能会引入噪音和偏差。iMILP 和 IsoFun

表 1 可变剪接异构体标注预测算法总结

| Method | Task | Category | Technique | Data source |
|---|---|---|---|---|
| isoPred [31] | Function | Multi-instance learning | mi-SVM & MI-SVM | RNS-Seq |
| IsoFunc [32] | Function | Multi-instance learning | MIL-SVM | RNA-Seq |
| WLRM [33] | Function | Multi-instance learning | MIL with weighted logistic regression | RNA-Seq |
| iMILP [34] | Function | Multi-instance learning | Instance-oriented MI label propagation | RNA-Seq |
| IsoFun [35] | Function | Multi-instance learning | Bi-random walk on heterogeneous network | RNA-Seq, PPI, GO |
| DeepIsoFun [36] | Function | Deep learning, Transfer learning | MIL with domain adaptation | RNA-Seq |
| DIFFUSE [37] | Function | Deep learning | Deep neural network & Conditional random field | RNA-Seq, sequence, domain |
| mFRecSys [38] | Function | Matrix factorization | Tri-factorization | RNA-Seq, sequence, PPI, GO |
| DisoFun [39] | Function | Matrix factorization | Collaborative matrix factorization | RNA-Seq, PPI, GO |
| IsoResolve [40] | Function | Transfer learning | Partial least square | RNA-Seq |
| ISOGO [41] | Function | Transfer learning | Wilcoxon test, Elastic-net regularized logistic linear model, Bayesian logistic regression | RNA-Seq, domain |
| IDAPred [42] | Disease | Matrix factorization | MIL with graph-regularized matrix factorization | RNA-Seq, DisGeNET, PPI, sequence |
| DeepIDA [43] | Disease | Deep learning | Deep neural network | RNA-Seq, sequence, GO, miRNA-target |

除了常规的正、负两种标签外，还特别设置了“未知”标签于那些无标注基因，并由此将问题建模为三分类问题，这样对问题的设置更为合理。

第五，现有的方法多是以术语为中心进行预测的。由于严重的标签不平衡问题，模型对于低频术语的预测性能往往不如高频术语理想。而且，以术语为中心的预测大多没有考虑到标注的层次结构，使得预测结果不满足一致性。

尽管可变剪切异构体功能预测领域迅速发展，但是对其疾病关联的预测尚处于起步阶段，而异常表型的标注预测更是一片空白。在未来几年中，剪切异构体表型的精确识别可能会促进将异构体用作不同疾病的生物标志物，使其成为首选的治疗靶标 [30]。因此，对可变剪切异构体进行 HPO 标注预测研究将是一个充满前景的方向。此外，构建起收录完整、规模浩大、条目精准、不断更新的异构体标注数据库，提供多物种、多组织、多条件下的异构体注释信息，将会极大的促进该领域的研究发展。

## 3 负样本的选取

负样本，即与给定标注确定无关的基因或蛋白质，通常很少记录在基因组和蛋白质组注释数据库中。例如，在 2020 年 10 月发布的 HPO 标注数据集中，仅存储了 1500 余条疾病的 HPO 负注释，这与正标注的规模相去甚远。出现这种情况是由于实验的限制：实验分析通常应用于单个蛋白质，并且蛋白质的功能可能取决于上下文环境，从而使负陈述/标签非常不确定，并导致很少（或对于大多数蛋白质而言没有）被证实为负样本。但是，对于使用机器学习技术的蛋白质标注预测工具而言，负样本的选取对最终性能会带来显著影响：它们通常需要足够的正负样本示例以训练准确的预测器。大多数现有的计算方法都是从未标注蛋白质中随机选取负样本或将所有未标注蛋白质均视为负样本，因而限制了其预测精度的提高。所以，设计精巧的负样本选取策略对于提升计算工具的预测性能大有裨益。

事实上，类似的问题在文本分类领域中早有研究，并被命名为“PU 学习”（Positive-Unlabeled learning），典型的方法有信息检索领域相关反馈的 Rocchio 算法 [45] 和两阶段 PU 学习技术 1-DNF 算法 [46]。

当前提出的负样本选取策略都是在蛋白质功能预测语境下的。Mostafavi 等人 [47] 提出了称为“兄弟”（sibling）的启发性策略。在该策略中，所有标注为某术语的任意兄弟节点的蛋白质都作为该术语的负样本。背后的想法是，蛋白质很少被同一父节点的多个子节点同时标注。但是，正如 [47] 指出的那样，“兄弟”策略存在问题：因为许多兄弟节点不是互斥的，存在某些蛋白质被多个兄弟节点注释，使得某些功能类别没有满足这些要求的蛋白质以至于无法生成负样本。

Zhao 等人 [48] 基于两阶段 PU 学习技术提出了 AGPS（Annotating Genes with Positive Samples）技术。该算法的目标是在学习过程中从未标记的数据中自动生成负样本。具体地说，作者首先将蛋白质的互作数据、基因表达谱和蛋白质复合物数据等异构信息源整合到一张功

能联系图（functional linkage graph）中，然后采用奇异值分解技术来降低维数并从数据中消除噪声。最后，利用提出的 AGPS 算法来定义负样本并预测未知蛋白质的功能。尽管该方法整合了多种信息源，但是却未考虑 GO 的层次化结构。

Youngs 等人 [49] 基于可参数化的贝叶斯（parameterizable Bayesian）技术计算每个蛋白质的先验偏差（prior bias）并以此选择负样本的算法 ALBias。作者定义，对于给定的术语 $c$，蛋白质 $i$ 属于术语 $c$ 的条件先验概率是所有给定蛋白质 $i$ 的标注术语 $m$ 可见 $c$ 的经验条件概率之均值。于是，所有在与预测的术语相同的本体中带有注释的基因且其相对于该术语的先验值为 0 者均被视为负样本。直观地，如果在其他任何基因的注释 $c$ 旁边都没有出现蛋白质 $g$ 的最具体注释，则应将 $g$ 视为注释 $c$ 的负样本。作者将该策略与 GeneMANIA 算法集成，发现显著提高了性能。

随后，Youngs 等人 [50] 又对 ALBias 进行改进，提出了 SNOB（Selection of Negatives through Observed Bias，通过观测偏差挑选负样本）和 NETL（Negative Examples from Topic Likelihood，从主题似然性挑选负样本）。与 ALBias 不同，SNOB 将所有术语（包括这些直接注释的祖先注释）进行平均，而非最具体的术语，从而考虑本体的层次结构；并且不同于将所有先验分数为 0 的蛋白质选为负样本，SNOB 提供给用户一个参数 $n$，表示所需的负样本个数，并选择分数最低的 $n$ 个蛋白质作为负样本。NETL 则从文本挖掘的角度出发，将蛋白质类似于“文档”对待，每个蛋白质的 GO 术语均类似于文档的“单词”，但此外作者还认为蛋白质具有潜在的“主题”。这些隐藏的主题代表了蛋白质的“真实”功能，既考虑了新功能（因须验证/测试而未注释功能），也包括错误（error）和错义注释（missannotation）（因标注中存在潜在错误，具有 GO 注释不能保证蛋白质实际发挥作用）。于是，作者应用多主题推断算法，隐含狄利克雷分布（Latent Dirichlet Allocation，LDA），来学习这些潜在主题或“真实”功能的分布，也可以学习“单词”或带注释的 GO 术语的条件分布。一旦知道了这些分布，NETL 就会将那些隐含主题分布与正样本尽可能不同的蛋白质选择为负样本，并允许用户指定所需的负样本数量。通过实验发现，SNOB 相比于那些 PU 学习借鉴而来的方法以及启发式策略取得了更低的预测错误率。这表明，尽管缺乏大量的否定注释，但 GO 数据库通过其肯定注释对可能的负样本的隐式信息进行了编码。

基于当前可用的蛋白质注释选择负样本存在很大的偏差。由于生物学研究人员研究兴趣的偏倚，与感兴趣领域相关的术语预计会比其他术语更频繁地被注释，因而导致蛋白质的功能注释不完整。虽然 [50] 通过真路径规则在直接注释上进行传播，但仍然不能完全表征蛋白质的功能角色。由于缺乏实验支持或其他原因，一个原本应使用更具体的术语进行注释的蛋白质，目前仅被该术语的祖先术语所注释。众所周知，一个具体的（或稀疏的）术语所标注的蛋白质要少于其祖先术语，而稀疏的术语占据了整个本体的绝大部分。[49-50] 等方法更倾向于选择稀疏术语作为蛋白质的负样本，因为它们仅利用不完整的注释来近似术语之间的经验条件概率，而忽视了本体结构本身。此外，蛋白质的 GO 注释常常会随着时间流逝而改变，越来越多当前术语的后代节点会被加入进来。为此，Fu 等人 [51] 通过整合本体结构和传播后

的蛋白质注释提出了 NegGOA 算法。该算法利用 GO 术语之间的层次化语义相似度融入 GO 层次信息，并应用带重启动的向下随机游走（downward random walks with restart）来考虑 GO 注释的演变，同时根据两个术语共同注释蛋白质的经验条件概率对缺失的注释进行建模。由此，算法整合了蛋白质的可用注释和缺失注释的潜在可能性来挑选负样本。

上述方法仅关注对 GO 的层次结构和已有正样本的利用, 而忽略了对已有少量负样本和蛋白质其他特征信息（如蛋白质互作关系和氨基酸序列信息）的利用。为此，傅广垣等人 [52] 提出了基于正负样例的蛋白质功能预测方法 ProPN。该算法首先通过构造一个有向符号混合图描述已知的蛋白质与功能标记的正负关联信息、蛋白质之间的互作信息和功能标记间的关联关系，再通过符号混合图上的标签传播算法预测蛋白质功能，包括已标注部分功能标记蛋白质的负样例。但是，由于蛋白质互作网络中存在一定量的假阳性互作关系，这些假阳性互作会引起正负样例的过度传播。此外, 大部分功能标记是非常稀疏的，它们标注的蛋白质个数极少，在标记传播中这些稀疏标记容易被其他标记覆盖，因而也会降低负样例预测的准确性。

之后，余国先等人 [53] 借助成分扩散分析和奇异值分解提出了一种基于网络和标记空间降维的蛋白质不相关功能预测方法 IFDR。IFDR 首先通过在蛋白质互作网络邻接矩阵和蛋白质-功能标记关联矩阵上分别进行随机游走，挖掘蛋白质之间的内在关系和预估蛋白质的缺失功能标记，再分别利用奇异值分解将上述两个矩阵投影降维为低维实数矩阵以降低噪声的破坏作用，最后利用半监督回归在降维后的两个低维矩阵上预测负样例。

最近，Vesztrocy 等人 [54] 介绍了一个全新的、与开放世界假设（Open World Assumption，OWA）兼容的基准测试框架，其中提出了一种从系统发育树上蛋白质家族的专家标注注释中获取负标注的算法。事实上，专家标注人员已使用 PANTHER 系列的系统发育注释和推测工具使用 GO 术语对系统发育树上基因的祖先状态进行了正负注释。然后，这些祖先注释沿系统发育树向下传播到现存基因。如果有证据表明某功能在特定的子树中不存在，则对该子树的根节点及其后代节点分配否定注释，这意味着基因在该分支上失去了特定功能。于是，通过扫描 PAINT 注释中的此类实例，可以导出大量的负样本。

由于开放世界假设，确定负样本预测方法的准确性本身并不是一件容易的事。在有限的论文中，研究人员多是采用以下两个策略对结果进行评估：（1）负样本预测准确率：对于在 $t_0$ 时刻做出的预测，利用 $t_0$ 至 $t_1$ 时刻积累的少量负标注数据作为标准答案，统计预测结果上的命中率；（2）对蛋白质功能预测性能的帮助：通过提出的负样本采集方法选取负样本，结合已有的正样本，送入蛋白质功能预测模型中，对比用随机抽取负样本或将全部未标注蛋白质当作负样本所训练得到的模型的性能，计算性能的提升比例。

表2总结了本小节综述的若干蛋白质功能标注负样本选取方法。可以看到，现有的方法要么基于统计学手段，从 GO 的层次结构中挖掘与给定功能可能相关的负样本；要么利用机器学习技术，根据蛋白质功能预测模型的预测需要生成负样本以帮助提升预测精度。尽管 GO 的层次结构已经隐含了不少负样本信息，但是必然存在一些情形，仅 GO 本身不足以预测一组好的负样本。例如，兼职蛋白（moonlighting proteins）在行使主要功能之外还兼有其他次要

**表 2 蛋白质功能标注的负样本选取算法总结**

| Method | Keyword | Data source |
|---|---|---|
| SibNeg [47] | Sibling negatives | GO hierarchy |
| AGPS [48] | Two-stage PU learning, Single value decomposition | Protein interaction network, Gene expression profiles, Protein complex data |
| ALBias [49] | Parametric Bayesian priors | GO hierarchy |
| SNOB [50] | Parametric Bayesian priors | GO hierarchy |
| NETL [50] | Lantent Dirichlet Allocation | GO hierarchy |
| NegGOA [51] | Random walks with restart, Probability | GO hierarchy, GO annotation |
| ProPN [52] | Direct signed hybrid graph, Label propagation | GO hierarchy, GO annotation, Protein interaction network |
| IFDR [53] | Random walk, Single value decomposition, Semi-supervised regression | GO hierarchy, GO annotation, Protein interaction network |
| Vesztrocy's method [54] | Phylogeny | Curated gene phylogenies |

或隐藏功能，仅通过 GO 的层次结构很难分析此类蛋白质的负注释。此外，这一类方法在某种程度上局限于那些已有一定研究基础的蛋白质，而这样的蛋白质在很多物种的基因组中只占很小的比例。因此，整合如基因表达、蛋白质互作、结构域等多种信息源对于提升预测效果会有帮助。近来的一些机器学习方法虽然考虑了多种信息源，但是其受信息源中噪音的影响较大，易产生“假阳性”预测；而且，由于一些标签极度不平衡，训练的分类器面临过拟合的风险。

虽然在蛋白质功能预测领域对负样本选取方法的研究取得了一定进展，但是对表型预测问题中的负样本选取研究还是一片处女地。目前 HPO 标注数据集中实验验证的负标注条数很少，而且现有数据库中的标注还很不完善，预测工具选取所有未标注蛋白质作为负样本将会导入大量噪音降低预测性能。因而此项研究工作对于帮助完善负标注验证、提升表型预测工具性能有着可预期的贡献。

## 4 噪声标注识别

HPO 团队在构建基因的 HPO 标注时，首先会通过四种方式完成疾病的 HPO 标注，再从疾病数据库中下载疾病与基因的关联，由此建立起基因与 HPO 术语间的关系。标注人员对疾病进行注释的四种来源分别是：(1) 从电子注释推断 (IEA)；(2) 已发表的临床研究 (PCS)；(3) 个人临床经验（ICE）和（4）可追溯的作者陈述（TAS)。在这其中，IEA 是通过文本挖掘程

序自动的从 OMIM 数据库内“临床特征”（Clinical Features）一章抽取该疾病的异常表型关联。于是一些被程序错误识别的信息引入了数据库中，导致了错误的标注。例如，在 2020 年 8 月发布的基因的 HPO 标注文件中，与 HP:0000246 鼻窦炎（*Sinusitis*）相关的基因包括 ARMC4、BLM、CCDC114、CCDC151 和 TTC25，但是这些关联在随后 10 月发布的标注文件中被删除。事实上，类似的情况还有不少。本文中，我们将这些错义标注定义为噪声标注。现有的 HPO 标注预测工作均假设已知的蛋白质表型标注信息是准确无误的，忽视了噪声标注信息对预测结果的影响。而且，这些噪声标注会误导后续基因组学、蛋白质组学、表观遗传学、临床诊断与治疗、疾病药物靶标和药物设计等多个领域的研究与应用。因此，有效地识别 HPO 噪声标注，将有助于提高已有标注信息的可靠性，方便后续研究与应用。

事实上，蛋白质的 GO 标注数据中也存在不少噪声标注。研究人员发现，证据属性为 IEA、ISS 等的标注中存在不少错误。尽管已有一些研究工作，对不同证据的 GO 标注的可靠性展开了分析。但是，对于如何准确识别蛋白质功能标注中的噪声标注，相关的研究工作还十分匮乏。现有的几个方法可以根据其思想划分为两大类：（1）根据某些功能的物种特异性判断功能标注的合理性；（2）基于机器学习方法推断可能的噪声标注。

对于第一类方法，其基于这样一个事实：某些蛋白质功能是具有物种特异性的。例如，“泌乳”（lactation）过程仅出现于哺乳动物中，“线粒体”（mitochondrion）只会在真核生物细胞内。于是，通过比较不同物种间同一功能术语标注的一致性，可以发现可能的错误标注。Deegan 等人 [55] 提出了若干基于物种的标注约束规则，并通过程序进行一致性检查，发现已有标注文件中违反约束的注释。最近，Wei 等人 [56] 也基于物种差异，提出比较不同物种间同一 GO 术语的“标注比率”（ratio of annotation rates，RAR），将 RAR 相对较低的标注视为潜在的不正确注释。尽管这些方法能够成功的发现一些错误的功能注释，但是它们所基于的物种特异性假设并非永远成立。例如，一些宿主-病原体相互作用关系容易被忽略，误导了程序的判断。一个例子是，GO:0061630 泛素蛋白连接酶活性（*Ubiquitin protein ligase activity*）是细菌中罕见的 GO 术语，因为泛素依赖性蛋白降解是真核生物特异性蛋白分解代谢途径，故该术语与动物大量相关，但只标注了一种细菌蛋白（SspH2，UniProt ID：P0CE12）。但事实上，细菌的功能注释是正确的，因为 SspH2 是一种 E3 泛素连接酶，在沙门氏菌感染后会干扰真核宿主中的泛素化途径。类似的情形降低了算法的准确率，使得标注人员还必须对自动识别的所谓“错误标注”逐一进行人工检查。

路畅等人 [57] 使用分类学（taxonomic）和语义相似度提出了 NoisyGOA 算法。该方法首先使用 GO 层次结构和基因之间的语义相似度来度量本体术语之间的分类学相似度，然后计算汇总一个蛋白质的每个功能标注与它语义近邻蛋白质的功能标注的最大分类学相似度，最后将与这些近邻蛋白质具有最小分类学相似度的功能标注判定为该蛋白质的噪声功能标注。然而，NoisyGOA 在计算语义相似度的时候易受蛋白质已有噪声功能标注的影响，也忽视了功能标注的证据属性。为此，余国先等人 [58] 基于证据属性和稀疏表示提出了 NoGOA 算法。NoGOA 首先在基因-术语关联矩阵上应用稀疏表示来减少噪音注释的影响，并利用稀疏表示

**表 3 蛋白质噪声功能标注识别算法总结**

| Method | Category | Keyword |
|---|---|---|
| Deegan's method [55] | Taxon-based | Taxon constraints |
| RAR [56] | Taxon-based | Ratio of annotation rates |
| NoisyGOA [57] | ML-based | Semantic similarity, Taxonomic similarity |
| NoGOA [58] | ML-based | Evidence codes, Sparse representation |
| NFA [59] | ML-based | Evidence codes, Sparse representation |

系数来度量基因之间的语义相似度。然后，它基于来自该基因语义近邻基因"投票"的总票数，初步预测该基因的噪音注释。接下来，NoGOA 基于在不同版本的基因功能标注文件估计每个证据代码的噪音注释的比率，并通过估计的比率和 GO 的层次结构对关联矩阵中的元素进行加权。最后，算法通过加权基因-术语关联矩阵和来自近邻基因的汇总投票给出噪音注释的最终预测结果。路畅等人 [59] 后又对 NoGOA 进行修改，不再进行基于稀疏表示的初步预测，而是将稀疏表示与基于证据属性的加权关联矩阵整合起来，提出了 NFA 算法。NFA 同样先根据功能标注的证据属性和 GO 的层次结构对蛋白质-功能标签关联矩阵进行加权，然后在加权的关联矩阵上利用 $\ell_1$-norm 约束的稀疏表示计算蛋白质之间的语义相似度。区别是最后一步，算法利用一个蛋白质的语义近邻蛋白质的功能标注信息投票识别该蛋白质的噪声功能标注。不过这两个方法都没有考虑功能的物种特异性信息。

对上述方法的总结见表3。为了评估算法的识别精度，研究人员通常采用以下两种评估策略：（1）在 $t_0$ 时刻基于当时的标注数据给出预测结果，然后以那些在 $t_0$ 时刻出现但在之后 $t_1$ 时刻消失的标注为基准，评估预测结果的精度；（2）将预测出的噪音标注从发布的标注文件中剔除，并用更新后的标注数据训练蛋白质功能预测模型，并与使用原始标注训练的模型对比，评估移除预测的噪音标注后会否提升功能预测模型的性能。不过，由于错误的标注在整个数据库中所占的比例是比较低的，因而得出的评估指标的置信程度还有待进一步讨论。

尽管已有一些识别 GO 标注中错误注释的算法，但遗憾的是，目前，识别 HPO 标注中的噪声标注这一课题尚无人触及。但是，现有 HPO 标注中存在的一些错误标注会给蛋白质表型预测工具的性能产生不利影响。因此，我们认为，探索准确高效的 HPO 噪声标注识别算法将是一个值得深入研究的方向。

## 5 假基因的 HPO 标注

假基因（Pseudogene），又称伪基因，是基因家族在进化过程中形成的无功能的残留物，是展现了人类基因组进化编年史的"基因组化石"。长期以来，生物学家们认为假基因是没有功能的"垃圾 DNA"（junk DNA），但近年来的研究显示，假基因和其他非编码片段一样，拥有调控基因表达的功能 [60]。假基因的调控作用对维持生物体的正常生理活动有着重要意义，一些研究发现一部分假基因在某些疾病的发展中扮演着重要角色 [61]。一个著名的例子是假基因

PTENP1 在几种癌症中对抑癌基因 PTEN 的转录调控 [62]。文献 [63] 从不同方面对假基因与人类遗传疾病的相关性进行了详细的综述。显然，用 HPO 术语对假基因进行注释可以帮助研究人员更好的理解人类疾病背后的遗传学因素。然而，目前还无人涉足此领域，更没有收录实验验证的假基因 HPO 标注的数据库。

最近，Fan 等人 [64] 基于图卷积神经网络提出了用于假基因功能预测的半监督学习模型 Pseudo2GO。假基因功能预测的最大挑战是缺少足够的特征和功能注释，这使得训练预测模型变得困难。作者考虑到假基因与其共享大量 DNA 序列的亲本编码基因之间的功能相似性，以及编码基因具有丰富的 GO 注释，指出通过基于图的方式借鉴编码基因的信息来预测假基因功能。作者首先构建序列相似性图以连接假基因和编码基因。包括表达谱、与 microRNA 的相互作用、蛋白质相互作用和遗传相互作用等在内的多种特征信息作为节点属性融入图中使其成为一个属性图。最后，应用图卷积网络在图上传播节点属性，以对假基因进行分类。模型在由编码基因组成的训练集上进行训练，对由假基因组成的测试集进行预测。作者使用 GENCODE [65] 发布的假基因和编码基因注释进行实验。就我们所知，这是第一个也是唯一一个在基因本体上直接预测假基因功能的算法。

我们认为，构建假基因的 HPO 标注数据库，探索假基因 HPO 标注自动化预测算法，对揭开疾病的致病机理、理解假基因与人类疾病的关联有潜在的帮助，是未来值得研究的方向。

## 6 总结

本文介绍了可变剪接异构体的标注预测、负样本的选取策略、噪声标注识别以及假基因的标注预测四个在功能注释预测领域已有进展，但在表型注释预测领域未有探索的任务。通过对问题背景进行清晰描绘，对已有工作进行系统总结，对未来方向进行展望，作者希望可以由此促进该领域的研究发展不断前进，期待着有越来越多的研究人员加入到这项工作中来。

## 参考文献

[1] OPAP K, MULDER N. Recent advances in predicting gene–disease associations [J]. F1000Res., 2017, 6.

[2] ROBINSON P, et al. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease [J]. Am. J. Hum. Genet., 2008, 83(5): 610-615.

[3] HAMOSH A, et al. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders [J]. Nucleic Acids Res., 2002, 30(1): 52-55.

[4] PAVAN S, et al. Clinical Practice Guidelines for Rare Diseases: The Orphanet Database [J]. PLoS One, 2017, 12(1): e0170365.

[5] FIRTH H, et al. DECIPHER: database of chromosomal imbalance and phenotype in humans using ensembl resources [J]. Am. J. Hum. Genet., 2009, 84(4): 524-533.

[6] WANG P, et al. Inference of gene-phenotype associations via protein-protein interaction and orthology [J]. PLoS One, 2013, 8(10): e77478.

[7] DOĞAN T. HPO2GO: prediction of human phenotype ontology term associations for proteins using cross ontology annotation co-occurrences [J]. PeerJ, 2018, 6: e5298.

[8] KAHANDA I, et al. PHENOstruct: Prediction of human phenotype ontology terms using heterogeneous data sources [version 1; peer review: 2 approved] [J]. F1000Research, 2015, 4(259).

[9] NOTARO M, et al. Prediction of Human Phenotype Ontology terms by means of hierarchical ensemble methods [J]. BMC Bioinformatics, 2017, 18(1): 1-18.

[10] NOTARO M, et al. Ensembling descendant term classifiers to improve gene-abnormal phenotype predictions [C]//International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics. [S.l.]: Springer, 2017: 70-80.

[11] KULMANOV M, HOEHNDORF R. DeepPheno: Predicting single gene loss-of-function phenotypes using an ontology-aware hierarchical classifier [J]. PLoS Comput. Biol., 2020, 16(11): e1008453.

[12] LIU L, et al. HPOLabeler: improving prediction of human protein-phenotype associations by learning to rank [J]. Bioinform., 2020, 36(14): 4180-4188.

[13] PETEGROSSO R, et al. Transfer learning across ontologies for phenome-genome association prediction [J]. Bioinform., 2017, 33(4): 529-536.

[14] HAN S, et al. Metrical Consistency NMF for Predicting Gene–Phenotype Associations [J]. Interdiscip. Sci., 2018, 10(1): 189-194.

[15] ZHANG Y, et al. GC$^2$NMF: A Novel Matrix Factorization Framework for Gene–Phenotype Association Prediction [J]. Interdiscip. Sci., 2018, 10(3): 572-582.

[16] GAO J, et al. AiProAnnotator: Low-rank Approximation with network side information for high-performance, large-scale human Protein abnormality Annotator [C]//IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2018, Madrid, Spain, December 3-6, 2018. [S.l.]: IEEE Computer Society, 2018: 13-20.

[17] GAO J, et al. HPOAnnotator: improving large-scale prediction of HPO annotations by low-rank approximation with HPO semantic similarities and multiple PPI networks [J]. BMC Med. Genomics, 2019, 12 (10): 187.

[18] LIU L, et al. HPOFiller: identifying missing protein-phenotype associations by graph convolutional network [J]. 2021.

[19] GLIGORIJEVIC V, et al. deepNF: deep network fusion for protein function prediction [J]. Bioinform., 2018, 34(22): 3873-3881.

[20] XUE H, et al. Towards Gene Function Prediction via Multi-Networks Representation Learning [C]//The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019. [S.l.]: AAAI Press, 2019: 10069-10070.

[21] PENG J, et al. Integrating multi-network topology for gene function prediction using deep neural networks [J]. Brief. Bioinform., 2020.

[22] FORSTER D, et al. BIONIC: Biological Network Integration using Convolutions [J]. bioRxiv, 2021.

[23] VALENTINI G, et al. RANKS: a flexible tool for node label ranking and classification in biological networks [J]. Bioinform., 2016, 32(18): 2872-2874.

[24] MOSTAFAVI S, et al. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function [J]. Genome Biol., 2008, 9(S1): S4.

[25] MOSTAFAVI S, MORRIS Q. Fast integration of heterogeneous data sources for predicting gene function with limited annotation [J]. Bioinform., 2010, 26(14): 1759-1765.

[26] CHO H, et al. Compact integration of multi-network topology for functional analysis of genes [J]. Cell Syst., 2016, 3(6): 540-548.

[27] LIU L, et al. HPODNets: deep graph convolutional networks for predicting human protein-phenotype associations [J]. 2021.

[28] LIU L, ZHU S. Computational methods for prediction of human protein-phenotype associations: a review [J]. 2021.

[29] KIM H, et al. Alternative splicing isoforms in health and disease [J]. Pflügers Arch. - Eur. J. Physiol., 2018, 470(7): 995-1016.

[30] MISHRA S, et al. Computational methods for predicting functions at the mRNA isoform level [J]. Int. J. Mol. Sci., 2020, 21(16): 5686.

[31] EKSI R, et al. Systematically differentiating functions for alternatively spliced isoforms through integrating RNA-seq data [J]. PLoS Comput. Biol., 2013, 9(11): e1003314.

[32] PANWAR B, et al. Genome-wide functional annotation of human protein-coding splice variants using multiple instance learning [J]. J. Proteome Res., 2016, 15(6): 1747-1753.

[33] LUO T, 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, NS, Canada, August 13 - 17, 2017. [S.l.]: ACM, 2017: 345-354.

[34] LI W, et al. High-resolution functional annotation of human transcriptome: predicting isoform functions by a novel multiple instance-based label propagation method [J]. Nucleic Acids Res., 2014, 42(6): e39-e39.

[35] YU G, et al. Isoform function prediction based on bi-random walks on a heterogeneous network [J]. Bioinform., 2020, 36(1): 303-310.

[36] SHAW D, et al. DeepIsoFun: a deep domain adaptation approach to predict isoform functions [J]. Bioinform., 2019, 35(15): 2535-2544.

[37] CHEN H, et al. DIFFUSE: predicting isoform functions from sequences and expression profiles via deep learning [J]. Bioinform., 2019, 35(14): i284-i294.

[38] KANDOI G. Machine learning tools for mRNA isoform function prediction [D]. [S.l.]: Iowa State University, 2019.

[39] WANG K, et al. Differentiating isoform functions with collaborative matrix factorization [J]. Bioinform., 2020, 36(6): 1864-1871.

[40] LI H, et al. IsoResolve: Predicting Splice Isoform Functions by Integrating Gene and Isoform-level Features with Domain Adaptation [J]. Bioinform., 2020.

[41] FERRER-BONSOMS J, et al. ISOGO: Functional annotation of protein-coding splice variants [J]. Sci. Rep., 2020, 10(1): 1-11.

[42] HUANG Q, et al. Isoform-Disease Association Prediction by Data Fusion [C]//Lecture Notes in Computer Science: volume 12304 Bioinformatics Research and Applications - 16th International Symposium, ISBRA 2020, Moscow, Russia, December 1-4, 2020, Proceedings. [S.l.]: Springer, 2020: 44-55.

[43] YU G, et al. DeepIDA: predicting isoform-disease associations by data fusion and deep neural networks [J]. IEEE/ACM Trans. Comput. Biol. Bioinform., 2021.

[44] LIN T, et al. Focal Loss for Dense Object Detection [C]//IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017. [S.l.]: IEEE Computer Society, 2017: 2999-3007.

[45] ROCCHIO J. Relevance Feedback in Information Retrieval [M]. Englewood Cliffs: Prentice Hall, 1971.

[46] YU H, et al. PEBL: positive example based learning for Web page classification using SVM [C]// Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July 23-26, 2002, Edmonton, Alberta, Canada. [S.l.]: ACM, 2002: 239-248.

[47] MOSTAFAVI S, MORRIS Q. Using the Gene Ontology Hierarchy when Predicting Gene Function [C]// UAI 2009, Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, Montreal, QC, Canada, June 18-21, 2009. [S.l.]: AUAI Press, 2009: 419-427.

[48] ZHAO X, et al. Gene function prediction using labeled and unlabeled data [J]. BMC Bioinform., 2008, 9: 57.

[49] YOUNGS N, et al. Parametric Bayesian priors and better choice of negative examples improve protein function prediction [J]. Bioinform., 2013, 29(9): 1190-1198.

[50] YOUNGS N, et al. Negative Example Selection for Protein Function Prediction: The NoGO Database [J]. PLoS Comput. Biol., 2014, 10(6).

[51] FU G, et al. NegGOA: negative GO annotations selection using ontology structure [J]. Bioinform., 2016, 32(19): 2996-3004.

[52] FU G, et al. Protein function prediction using positive and negative examples [J]. J. Comput. Res. Dev., 2016, 53(8): 1753.

[53] YU G, et al. Predicting irrelevant functions of proteins based on dimensionality reduction [J]. Sci. Sin. Inform., 2017, 47(10): 1349-1368.

[54] VESZTROCY A, DESSIMOZ C. Benchmarking gene ontology function predictions using negative annotations [J]. Bioinform., 2020, 36(Supplement-1): i210-i218.

[55] DEEGAN J, et al. Formalization of taxon-based constraints to detect inconsistencies in annotation and ontology development [J]. BMC Bioinform., 2010, 11: 530.

[56] WEI X, et al. Detecting Gene Ontology misannotations using taxon-specific rate ratio comparisons [J]. Bioinform., 2020, 36(16): 4383-4388.

[57] LU C, et al. NoisyGOA: Noisy GO annotations prediction using taxonomic and semantic similarity [J]. Comput. Biol. Chem., 2016, 65: 203-211.

[58] YU G, et al. NoGOA: predicting noisy GO annotations using evidences and sparse representation [J]. BMC Bioinform., 2017, 18(1): 350.

[59] LU C, et al. Identifying noisy functional annotations of proteins using sparse semantic similarity [J]. Sci. Sin. Inform., 2018, 48(8): 1035-1050.

[60] BALAKIREV E, AYALA F. Pseudogenes: are they "junk" or functional DNA? [J]. Annu. Rev. Genet., 2003, 37(1): 123-151.

[61] LU X, et al. Pseudogene in cancer: real functions and promising signature [J]. J. Med. Genet., 2015, 52 (1): 17-24.

[62] POLISENO L, et al. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology [J]. Nature, 2010, 465(7301): 1033-1038.

[63] SEN K. Relevance of Pseudogenes to Human Genetic Disease [J]. eLS, 2013.

[64] FAN K, ZHANG Y. Pseudo2GO: A Graph-Based Deep Learning Method for Pseudogene Function Prediction by Borrowing Information From Coding Genes [J]. Front. Genet., 2020, 11: 807.

[65] FRANKISH A, et al. GENCODE reference annotation for the human and mouse genomes [J]. Nucleic Acids Res., 2019, 47(D1): D766-D773.