中国科学: 信息科学 2018 年 第 48 卷 第 8 期: 1035–1050
SCIENTIA SINICA Informationis
《中国科学》杂志社
SCIENCE CHINA PRESS
论文

基于稀疏语义的蛋白质噪声功能标注识别

路畅[1], 陈霞[1], 王峻[1], 余国先[1*], 余志文[2]

1. 西南大学计算机与信息科学学院, 重庆 400715
2. 华南理工大学计算机科学与工程学院, 广州 510006
* 通信作者. E-mail: gxyu@swu.edu.cn

收稿日期: 2017–05–16; 接受日期: 2017–10–13; 网络出版日期: 2018–01–31

国家自然科学基金 (批准号: 61402378, 61572199, 61741217) 和重庆市基础与前沿研究计划项目 (批准号: cstc2014jcyjA40031, cstc2016jcyjA0351) 资助

摘要 蛋白质功能自动标注是生物信息学领域的关键问题之一. 蛋白质功能标注信息来源广泛, 噪声标注信息不可避免地被引入. 已有蛋白质功能预测研究更关注预测功能信息完全未知 (或部分已知) 蛋白质的功能, 极少关注识别蛋白质的噪声功能标注. 本文提出一种基于稀疏语义相似度的蛋白质噪声功能标注识别方法 (identifying noisy functional annotations of proteins using sparse semantic similarity, NFA). NFA 首先利用一个蛋白质 – 功能标签关联矩阵存储蛋白质功能标注信息, 对不同证据的功能标注信息分别加权, 再利用功能标签间层次结构关系向上传播这些权重到拓展的功能标注上; 其次, 在加权后的关联矩阵上利用 l_1-norm 约束的稀疏表示计算蛋白质之间的语义相似度; 最后基于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. 在酵母菌和拟南芥这两个模式生物上的实验结果表明, NFA 较现有算法能更准确识别蛋白质噪声功能标注, 剔除 NFA 识别出的噪声功能标注能够提升现有蛋白质功能预测算法的精度.

关键词 蛋白质功能, 噪声功能标注, 稀疏表示, 语义相似度, 标签结构

1 引言

各种高通量生物技术的广泛应用产生了海量的生物数据. 蛋白质作为一类主要的生物大分子, 是细胞生命活动的主要载体, 执行着生物体内各种重要的功能. 对蛋白质进行准确的功能标注能促进疾病分析与治疗、新药品研发、农作物促产和生物能源开发等领域的发展 [1~3]. 由于生物湿实验自身的限制和生物学家研究兴趣的偏向性, 当前基因本体 (gene ontology, GO) [4] 数据库中登记的经过实验验证的蛋白质功能标注信息具有较大的偏向性, 覆盖度有限, 且高通量技术测得的功能信息较浅层 [5,6]. 另一方面, 通过湿实验标注蛋白质功能的速度远小于新发现蛋白质的速度 [7]. 针对这些问题, 基于各种生物数据的大规模蛋白质功能预测方法被广泛研究, 并证明能够为湿实验验证提供具有较高置信度

引用格式: 路畅, 陈霞, 王峻, 等. 基于稀疏语义的蛋白质噪声功能标注识别. 中国科学: 信息科学, 2018, 48: 1035–1050, doi: 10.1360/N112017-00105
Lu C, Chen X, Wang J, et al. Identifying noisy functional annotations of proteins using sparse semantic similarity (in Chinese). Sci Sin Inform, 2018, 48: 1035–1050, doi: 10.1360/N112017-00105

表 1 基因本体的证据分类

Table 1 The categorization of GO evidence codes

Experimental	EXP	IDA	IPI	IMP	IGI	IEP					
Computational	ISS	ISO	ISA	ISM	IGC	IBA	IBD	IKR	IRD	RCA	IEA
Author	TAS	NAS									
Curatorial	IC	ND									

的功能参照信息 [1, 8, 9]. 实际上, UniProt 中 99% 的蛋白质功能标注, GO 数据库中 95% 的蛋白质功能标注都是基于蛋白质的特征信息通过计算学方法推断而来 [5~7].

随着生物学知识的不断更新完善和蛋白质功能信息的不断积累, GO 中存储功能标签间结构关系的 GO 文件和存储多个物种蛋白质的 GO 标注数据 (GO annotations, GOA) 的 GOA 文件也在不断更新 [4, 5]. 这些更新不仅包括功能标签和蛋白质功能标注信息的增补, 而且包括一些功能标签和功能标注信息的移除. 如 Gillis 等 [10] 统计分析了蛋白质功能标注信息的不稳定性, 发现 20% 的蛋白质在两年后一些功能信息不再保留. 尽管 GO 组委会采用多种质量控制方法来确保各物种 GOA 文件的一致性与质量 [4, 5], 但已有研究表明, 由于蛋白质功能信息收集来源广泛, GOA 文件中仍然包含一些错误的蛋白质功能标注信息 (即蛋白质并不具有某些功能却被标注了这些功能), 本文将这些错误标注的蛋白质功能信息定义为蛋白质的噪声功能标注. 现有蛋白质功能预测工作通常关注预测功能信息完全未知的蛋白质功能 [11, 12], 增量标注蛋白质的缺失功能 [13, 14] 和预测蛋白质的不相关功能 [15~18]. 这些工作均假设已知的蛋白质功能标注信息是准确无误的, 忽视了噪声标注信息对功能预测结果的影响. 而实际上, 这些噪声功能标注会误导后续蛋白质功能的研究与应用 [19, 20], 如药物靶标和药物设计 [2]、生物标志物发现、基因富集分析 [19, 20] 和生物网络分析 [3] 等. 因此, 有效地识别蛋白质噪声功能标注, 将有助于提高蛋白质已有功能标注信息的可靠性, 方便后续研究与应用.

GOA 文件中存储的蛋白质功能标注信息均附属了相应的 “证据” 属性, 表明该标注信息的来源或获取方式 [4]. GO 当前使用的证据类别有 21 种[1)] (如表 1 所示). 除 IEA (inferred from electronic annotations) 外, 其他所有证据类别的功能标注信息均经过 GO 组委会审核. Thomas 等 [21] 认为可以使用证据类别作为衡量蛋白质功能标注的可靠性指标, 通过统计分析指出属于 Experimental 和 Author 这两大类的功能标注信息可靠性更高. Gross 等 [22] 对不同证据的蛋白质功能标注信息进行统计分析, 发现它们在 GO 3 个分支 (biological process, BP; cellular component, CC; molecular function, MF) 可靠性不同. Jones 等 [23] 发现基于证据为 ISS (inferred from sequence or structural similarity) 的功能标注信息预测蛋白质功能会导致更高的错误率, 因此建议尽可能避免利用证据为 ISS 的功能标注. Rogers 等 [24] 利用功能标注的证据评估蛋白质功能预测方法的性能, 并声明在对比不同预测方法时结合证据属性可以避免偏向性. 此外, 蛋白质功能标注的证据属性还被用于计算蛋白质之间的语义相似度 [25, 26]. 蛋白质之间的语义相似度通常基于成对蛋白质已有功能标注信息和功能标签间的结构关系计算. 研究表明, 语义相似度与蛋白质之间的序列相似度和蛋白质之间的互作等正相关 [27, 28], 它已被用于蛋白质功能预测和蛋白质之间交互作用预测等领域 [29, 30]. Benabderrahmane 等 [25] 对不同证据的功能标注设置不同的权重, 再采用一种基于路径的方法在 GO 的功能标签节点构成的有向无环图上计算蛋白质之间的语义相似度, 分析表明结合证据属性可以更准确地描述蛋白质之间的语义关系. 尽管这些方法分析了不同证据功能标注的可靠性, 指出 GO 数据库中存储的蛋白质功能信息存在一定的噪声和剔除噪声功能标注的重要性, 但并没有对如何准确识别 GOA 中的噪声功能标注展开相关的研

1) http://www.geneontology.org/page/guide-go-evidence-codes.

究. 针对这一状态, Lu 等 [31] 提出一种蛋白质噪声功能标注识别方法 (NoisyGOA). NoisyGOA 首先计算蛋白质之间的语义相似度和 GO 功能标签之间的分类相似度, 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相似度, 最后将与这些近邻蛋白质具有最小分类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, 且没有考虑功能标注的证据属性, 预测精度有限. 如何更准确地识别蛋白质噪声功能标注亟待进一步研究.

现有 GO 中功能标签的数量已超过 43000 个, 这些功能标签分布在 GO 3 个分支 (BP, CC, MF), 每个分支均通过一个有向无环图描述这些功能标签节点间的层次结构关系 [4], 子标签是父标签功能信息的进一步细化. 当蛋白质被标注有某个 GO 功能标签对应的功能时, 表明该蛋白质也拥有该标签所有祖先功能标签对应的功能; 而当蛋白质不具有某个标签对应的功能时, 该蛋白质也不会标注该标签及其子孙标签对应的功能; 这一规则称为 True Path Rule [4, 29]. 一个蛋白质通常仅标注有数十个甚至是几个功能标签对应的功能. 因此, 存储蛋白质功能标注信息的蛋白质 – 功能标签关联矩阵是一个非常稀疏且包含一定噪声元素的矩阵. 蛋白质之间的语义相似性度量容易被稀疏矩阵中的噪声干扰. 稀疏表示 [32, 33] 近些年被广泛应用于图像去噪、信号去噪和稀疏特征学习等任务中 [34~36]. 当输入信号是稀疏的且包含有噪声时, 稀疏表示能够较好地避免噪声特征的干扰, 准确地描述真实信号信息之间的关系.

基于上述理论和观察, 结合对已有研究成果的总结分析, 本文针对蛋白质的噪声功能标注识别这一亟待解决的新问题进行研究, 提出一种基于稀疏语义的蛋白质噪声功能标注识别方法 (NFA). NFA 首先根据功能标注的证据属性和 GO 功能标签的层次结构关系对蛋白质 – 功能标签关联矩阵进行加权; 其次在加权的关联矩阵上利用稀疏表示计算蛋白质之间的语义相似度; 最后利用近邻加权投票方法识别蛋白质的噪声功能标注. 在酵母菌和拟南芥这两个模式生物上的实验结果表明蛋白质噪声功能标注是可识别的, NFA 比现有相关算法 (如 NoisyGOA [31]) 能更准确地识别蛋白质噪声功能标注, 剔除 NFA 识别出的噪声功能标注能够显著地提高蛋白质功能预测精度. 本文研究还证实结合蛋白质功能标注信息的证据属性并设置相应权重和利用稀疏表示度量蛋白质之间的语义相似度均可提高噪声功能标注识别精度, 整合它们能进一步提高识别精度.

2 基于稀疏语义的蛋白质噪声功能标注识别

NFA 主要由基于证据的蛋白质 – 功能标签关联矩阵加权, 关联矩阵上基于稀疏表示的蛋白质之间语义相似性度量计算, 以及结合两者的蛋白质噪声功能标注识别 3 部分构成. 下文将首先描述基于证据的关联矩阵加权, 然后介绍加权关联矩阵上的稀疏表示, 最后在前 2 部分内容的基础上给出最终的噪声功能标注识别方法.

2.1 基于证据的蛋白质功能标注加权

令 $\boldsymbol{A} \in \mathbb{R}^{N\times|\mathcal{T}|}$ 为蛋白质 – 功能标签关联矩阵, N 为蛋白质的个数, $\mathcal{T}$ 为 GO 功能标签集合, $|\mathcal{T}|$ 为功能标签的个数. 矩阵 $\boldsymbol{A}$ 的定义如下:

$$\boldsymbol{A}(i,t)=\begin{cases}1, \text{ 如果蛋白质 } i \text{ 标注 } t \text{ 或者 } t \text{ 的子孙标签},\\ 0, \text{ 其他}.\end{cases} \tag{1}$$

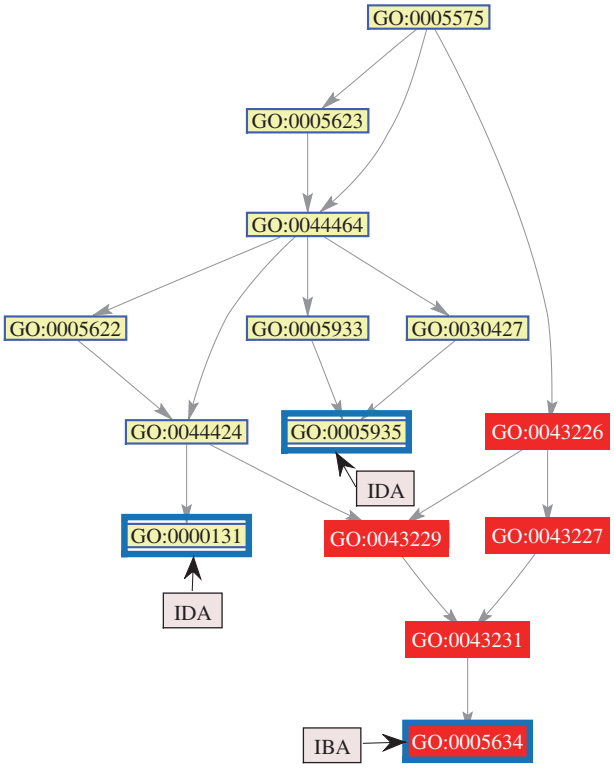

图 1 (网络版彩图) 酵母菌中蛋白质 “UBP5” 的 GO 功能标注 (红色功能标签为噪声标注)

Figure 1 (Color online) GO annotations of “UBP5” of S. cerevisiae (noisy annotations are in red rectangles)

蛋白质噪声功能标注识别的目标就是找出 $\boldsymbol{A}$ 中的噪声标注 (即错误的蛋白质 – 功能标记关联), 并将 $\boldsymbol{A}$ 中相应元素值从 1 更新为 0, 进而剔除噪声标注. 蛋白质噪声功能标注识别不同于蛋白质功能预测 [11~14], 后者将矩阵 $\boldsymbol{A}$ 中部分元素值从 0 更新为 1, 显式表明相应蛋白质标注了该功能标签对应的功能. 也不同于蛋白质负样例预测 [15~18], 后者将 $\boldsymbol{A}$ 中部分元素值从 0 更新为 −1, 显式表明相应蛋白质不具有该功能标签对应的功能.

已有研究表明不同证据属性的蛋白质功能标注可靠性不同. GOA 文件只记录每个蛋白质当前最详细且能准确描述该蛋白质功能的标注信息, 这些功能标注通常称为蛋白质的 “直接” 功能标注, GOA 文件记录了每个直接功能标注的 “证据” 类别. 例如, 在图 1 中, 酵母菌中的蛋白质 “UBP5” (日期: 2015-11-09) 有 3 个直接功能标注 (“GO:0005634”, nucleus; “GO:0000131”, incipient cellular bud site 和 “GO:0005935”, cellular bud neck), 它们的证据属性分别对应表 1 中的 IBA 和 IDA. 通过应用 GO 的 True Path Rule 可将这些直接标注对应功能标签的祖先标签也标注到该蛋白质上, 故图 1 中的蛋白质被共计 14 个 GO 功能标签标注. 而在更新的 GOA 文件 (日期: 2016-04-11), “UBP5” 的功能标注发生变化, 图 1 中红色矩形中的 5 个功能标签 (“GO:0043226”, “GO:0043229”, “GO:0043227”, “GO:0043231”, “GO:0005634”) 均不再标注在该蛋白质上, 为 “UBP5” 的噪声功能标注. 由于通过生物实验获取蛋白质的功能标注通常比基于计算方法推测的功能标注稳定且可信度更高 [21,37], 因此证据类别为 IDA 的功能标注可信度高于证据类别 IBA 的功能标注. 根据 True Path Rule 和图 1 中标签间结构, “GO:0005634” 及它的祖先标签比 “GO:0000131” 和 “GO:0005935” 更有可能为该蛋白质的噪声功能标注, 这与图 1 中所示的 5 个噪声标注完全相符. 上述观察表明结合蛋白质功能标注的证据属

表 2 GO 中 21 种证据的权重

Table 2 Weights assigned to 21 evidence codes of GO

EC	Experimental						Computational											Author		Curatorial	
	EXP	IDA	IEP	IGI	IMP	IPI	ISS	ISO	ISA	ISM	IGC	IBA	IBD	IKR	IRD	RCA	IEA	TAS	NAS	IC	ND
Weight	1	1	1	1	1	1	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	1	0.8	0.6	0.4

性并对它们区分加权有助于识别噪声标注.

受上述观察启发, 结合表 1 中证据分类和文献 [25] 中基于证据的语义相似性度量, 本文对 GO 中 21 个不同的证据类型设置了如表 2 所示的权重. 文献 [21] 指出属于 Experimental 和 Author 这两大类的蛋白质功能标注信息比其他的更可靠. 因此, 本文将 EXP, IDA, IEP, IPI, IMP, IGI, TAS 的权重设置为 1. 相比 TAS (traceable author statement), NAS (non-traceable author statement) 其可靠性要低一些, 所以其权重设置为 0.8, 但仍然大于其余所有类型证据. ND (no biological data available) 的权重设置为最低值 0.4, 剩余的所有类型的证据权重都设置为 0.6. 本文同样对文献 [38] 中建议的权重设置策略进行了实验研究, 实验结果与分析参见本文实验部分.

为了对蛋白质已有直接功能标注按照证据属性进行加权, 本文通过以下方式加权蛋白质 – 功能标签关联矩阵 $\boldsymbol{A}^d_{\rm ec}$:

$$\boldsymbol{A}^d_{\rm ec}(i,t) = \boldsymbol{A}^d(i,t) \times \boldsymbol{W}_{\rm ec}(i,t), \tag{2}$$

$\boldsymbol{A}^d \in \mathbb{R}^{N\times|\mathcal{T}|}$ 是仅仅由 GOA 文件中的蛋白质直接功能标注信息初始化的蛋白质 – 功能标签关联矩阵. $\boldsymbol{W}_{\rm ec} \in \mathbb{R}^{N\times|\mathcal{T}|}$ 为直接功能标注的证据权重 (参照表 2 设置). 如果同一个蛋白质 – 功能标签关联有多种证据, 本文选用这些证据中最大的权重初始化 $\boldsymbol{W}_{\rm ec}$.

基于 True Path Rule, 本文将直接功能标注的权重传播到它们的所有祖先功能标签, 方式如下:

$$\boldsymbol{A}_{\rm ec}(i,s) = \max\{\boldsymbol{A}^d_{\rm ec}(i,t) | s \in {\rm anc}(t)\}, \tag{3}$$

${\rm anc}(t)$ 表示 GO 中功能标签 t 所有的祖先功能标签集合. 如果功能标签 s 是多个直接功能标注的祖先, 则将 s 与蛋白质 i 的关联权重 $\boldsymbol{A}_{\rm ec}(i,s)$ 设置为这些权重中的最大值. 如图 1 中 "GO:0044464" 可分别由直接功能标注 "GO:0000131" 和 "GO:0005634" 基于 GO 的功能标签间结构关系推断获得. "GO:0000131" 的证据类别为 IDA, "GO:0005634" 的证据类别为 IBA. IDA 的权重大于 IBA, 因此, "GO:0044464" 与 "UBP5" 关联大小等于 IDA 的权重, "GO:0044424" 与 "UBP5" 的关联大小则等于 IBA 的权重, "GO:0044424" 比 "GO:0043326" 更可能为 "UBP5" 的噪声标注. 这种设置还可以确保蛋白质与祖先功能标签的关联权重大于或等于与所有子孙标签的关联权重. 相反, 如果蛋白质与祖先标签的关联权重小于其与子孙标签的关联权重, 则祖先标签将比子孙标签更有可能判定成为该蛋白质的噪声功能标注, 这种设置是不合理的.

2.2 基于稀疏表示的语义相似度计算

基于蛋白质功能标注的证据属性可以识别蛋白质噪声功能标注, 但这种识别方法仅考虑了该蛋白质自身的功能标注特性, 并没有考虑其他蛋白质的功能标注信息. 蛋白质之间通过协作完成具体的生物学功能, 蛋白质之间的互作和序列相似度等与蛋白质之间的语义相似度均正相关 [27, 28]. 为利用其他蛋白质已有的功能标注信息, 本小节在证据加权的蛋白质 – 功能标签关联矩阵 $\boldsymbol{A}_{\rm ec}$ 上计算蛋白质之间的语义的相似度, 方便后续基于稀疏语义的蛋白质噪声功能标注识别. 尽管研究者们已经提出了

多种语义相似度计算方法, 但这些方法普遍受到蛋白质的缺失功能标注和噪声标注的干扰 [27, 28]. 稀疏表示由于可以在一定程度上克服样本缺失特征和噪声特征的干扰, 已经广泛应用于图像和语音数据分析等领域 [34~36]. 由于绝大多数物种的蛋白质功能信息均存在不同程度的缺失 [4, 5], $\boldsymbol{A}_{\rm ec}$ 中的非 0 元素不超过 2%, 因此, $\boldsymbol{A}_{\rm ec}$ 是一个带有噪声的稀疏矩阵. 结合稀疏表示在高维噪声数据上的优势和 $\boldsymbol{A}_{\rm ec}$ 的特点, 本文利用 l_1-norm 约束的稀疏表示方法计算蛋白质之间的语义相似度, 方式如下:

$$\hat{\boldsymbol{\alpha}}_i = \arg\min_{\boldsymbol{\alpha}_i} ||\boldsymbol{A}_{\rm ec}(i,\cdot) - \boldsymbol{\alpha}_i \bar{\boldsymbol{A}}^i_{\rm ec}||_2 + \beta||\boldsymbol{\alpha}_i||_1, \quad \text{s.t. } \boldsymbol{\alpha}_i \geqslant 0, \tag{4}$$

$\bar{\boldsymbol{A}}^i_{\rm ec} \in \mathbb{R}^{(N-1)\times|\mathcal{T}|}$ 是 $\boldsymbol{A}_{\rm ec}$ 移除第 i 行后的子矩阵, $\beta > 0$ 是正则化参数, 用来平衡重构误差项与 l_1 范式稀疏项 ($||\boldsymbol{\alpha}_i||_1$) 之间的重要性. 式 (4) 的目的是对于蛋白质 i 的功能标记向量 $\boldsymbol{A}_{\rm ec}(i,\cdot)$, 期望用 $\boldsymbol{A}_{\rm ec}$ 中除 $\boldsymbol{A}_{\rm ec}(i,\cdot)$ 以外的其他尽可能少的向量线性重构 $\boldsymbol{A}_{\rm ec}(i,\cdot)$, $\boldsymbol{\alpha}_i \in \mathbb{R}^{1\times(N-1)}$ 为重构系数. $\boldsymbol{\alpha}_i(j)$ 为 $\boldsymbol{A}_{\rm ec}(j,\cdot)$ 对重构 $\boldsymbol{A}_{\rm ec}(i,\cdot)$ 的贡献大小系数, 其值越大代表向量 $\boldsymbol{A}_{\rm ec}(j,\cdot)$ 与向量 $\boldsymbol{A}_{\rm ec}(i,\cdot)$ 越相似, 亦即它们之间的语义相似度越大.

对于每一个 $\boldsymbol{A}_{\rm ec}(i,\cdot)$, 都可以得到对应 $\boldsymbol{\alpha}_i$, 本文根据 SLEP 软件包 [39] 求解式 (4), 再基于稀疏表示系数 $\{\boldsymbol{\alpha}_i\}_{i=1}^N$ 定义蛋白质之间的语义相似度. 蛋白质 i 和 j 之间的语义相似度计算如下:

$$\boldsymbol{S}(i,j) = \begin{cases} \boldsymbol{\alpha}_i(j), & j < i, \\ \boldsymbol{\alpha}_i(j-1), & j > i, \\ 0, & \text{其他}. \end{cases} \tag{5}$$

$\boldsymbol{S} \in \mathbb{R}^{N\times N}$ 是一个主对角线为 0 的语义相似度矩阵. 为保持距离度量的对称性, 令 $\boldsymbol{S} = (\boldsymbol{S} + \boldsymbol{S}^{\rm T})/2$ 为 N 个蛋白质之间的稀疏语义相似度矩阵. 实际上式 (4) 和 (5) 已经被广泛应用于样本之间相似性度量和近邻选取. 相关实验结果表明, 基于稀疏表示定义的样本之间相似度在高维且包含噪声特征的样本上很有效 [34~36].

2.3 噪声功能标注识别

投票法是一种简单直观的蛋白质噪声功能标注识别方法. 该方法基于每个蛋白质的语义近邻蛋白质的功能标注信息对标注到该蛋白质的每个功能标签分别进行投票, 若某个功能标签获得的票数少, 则说明该蛋白质的近邻标注该功能标签的个数也少, 该功能标签更有可能是噪声标注. 这种思想已经广泛应用于解决蛋白质功能标注的不一致性和整合多种来源的蛋白质功能标注信息 [40, 41], 近期还被应用到蛋白质噪声标注识别 [31] 中, 但这些方法并未较好地综合考虑不同投票者和不同功能标注的差异. 考虑到不同投票者之间和不同功能标注的差异性, NFA 利用证据加权的蛋白质 – 功能标签关联矩阵 $\boldsymbol{A}_{\rm ec}$ 和语义相似度矩阵 $\boldsymbol{S}$ 识别蛋白质噪声功能标注. 如果功能标签 t 已标注到蛋白质 t 上, 即 $\boldsymbol{A}_{\rm ec}(i,t) > 0$, t 最终得到的加权投票大小根据以下方式计算:

$$\boldsymbol{V}(i,t) = \sum_{j=1}^{N} \boldsymbol{S}(i,j) \times \boldsymbol{A}_{\rm ec}(j,t). \tag{6}$$

式 (6) 本质上是一个加权的 k 近邻分类器. 如果一个蛋白质标注 t, 但该蛋白质的语义近邻蛋白质均没有或者很少标注 t 或者对应的 $\boldsymbol{A}_{\rm ec}(j,t)$ 很低, 则 t 是该蛋白质的噪声功能标注的可能性极大, 对应 $\boldsymbol{V}(i,t)$ 的值很小. NFA 不仅利用稀疏表示来减少蛋白质噪声功能标注和缺失功能标注带来的负面影响, 还利用了直接功能标注的证据属性和功能标签间结构关系. NFA 将会比单独使用稀疏表示或者单独使用证据加权获得更高的蛋白质噪声功能标注识别精度, 下文实验也将证实 NFA 的上述优点.

表 3 拟南芥和酵母菌中蛋白质功能标注信息的统计

Table 3 Statistics of GO annotations of A. thaliana and S. cerevisiae

	Branch ($\|\mathcal{T}\|$)	Annotation	Noisy annotation
A. thaliana (24314)	BP (5390)	540073	10039
	CC (3853)	240184	1862
	MF (2773)	200008	2290
S. cerevisiae (5907)	BP (5161)	265224	3745
	CC (1017)	109934	683
	MF (2401)	70604	700

3 实验

3.1 数据集

由于没有现成的蛋白质噪声功能标注数据集可用于检验噪声功能标注识别算法的性能, 本文利用两个不同时期的 GO 文件和对应时期的模式生物 (拟南芥和酵母菌) GOA 文件来提取噪声功能标注数据, 以此来检验 NFA 的性能. 本文从 GO 官方网站[2)]分别下载 2015-11-09 和 2016-04-11 两个日期分别归档的 GO 文件和对应的 GOA 文件. 为了避免 GO 自身更新对噪声功能标注提取和预测结果的影响, 本文采取一种与 2nd CAFA (critical assessment of protein function annotation algorithms) [9] 相似的方法来处理数据: 对于 GO 功能标签数据, 本文仅考虑同时保留在 2015-11 和 2016-04 的功能标签, 排除仅出现在 2015-11 或 2016-04 中的功能标签, 即取历史的 GO 功能标签与现在的 GO 功能标签的交集. 然后利用上述两个时期归档的 GOA 文件基于式 (1) 中的方式分别初始化蛋白质 – 功能标签关联矩阵 $\boldsymbol{A}^h$ (2015-11) 和 $\boldsymbol{A}^r$ (2016-04). 本文假设 $\boldsymbol{A}^h$ 中出现但在 $\boldsymbol{A}^r$ 中消失的功能标注为噪声功能标注. 经过上述处理后, GO 在其 3 个分支上的功能标签总数分别为: 28111 (BP), 3853 (CC), 9966 (MF). 表 3 统计了拟南芥和酵母菌在 GO 3 个分支的功能标注信息. 例如, 拟南芥的 24314 个蛋白质在 BP 分支共计有 540073 个功能标注, 这些蛋白质与 5390 种功能标签存在关联, 其中有 10039 个噪声功能标注.

3.2 对比方法与评价度量

为对比分析 NFA 的性能, 本文采用 5 种相关的方法作为对比方法, 它们分别为:

(1) Least frequency (LF) 在每个蛋白质的已有功能标注中选取在 N 个蛋白质中出现频率最低的功能标签作为该蛋白质的噪声功能标注.

(2) Sparse representation (SR) 仅利用稀疏表示识别蛋白质噪声功能标注. 它首先在未加权的蛋白质 – 功能标签关联矩阵 $\boldsymbol{A}$ 上计算蛋白质之间的稀疏语义相似度, 然后再利用矩阵 $\boldsymbol{A}$ 和式 (6) 中 k 近邻分类器识别噪声功能标注.

(3) Evidence code (EC) 仅基于证据加权的蛋白质 – 功能标记关联矩阵 $\boldsymbol{A}_{\rm ec}$ 识别蛋白质噪声功能标注.

(4) NtN 是一种基于语义相似度的蛋白质功能预测方法 [42]. 该方法也可以用于识别蛋白质噪声功能标注. NtN 将每个蛋白质视作一个文档, 将标注到该蛋白质的功能标签视为单词. 它首先利用向量空间模型 [43] 中的词频 (功能标签频率) 和逆文档频率对蛋白质的功能标注进行加权, 再用 GO 功

2) http://geneontology.org/.

能标签间的结构关系调整这些权重, 接着在上述加权调整后的蛋白质 – 功能标签关联大小矩阵上进行奇异值分解重构, 将在原始关联矩阵中元素值大于 0 而在重构矩阵中对应元素值最小的关联判定为对应蛋白质的噪声功能标注.

(5) NoisyGOA [31] 是一种基于语义的噪声功能标注识别方法. 该方法在式 (1) 中未证据加权的蛋白质 – 功能标签关联矩阵 $\boldsymbol{A}$ 上采用余弦相似度计算蛋白质的语义相似度, 再通过一种基于路径的方法计算 GO 功能标签间的分类相似度, 最后聚合一个蛋白质的已有功能标签与其语义近邻蛋白质的已有功能标签间的分类相似度, 选定聚合的相似度最小的功能标注为噪声功能标注.

本文测试了式 (4) 中的参数 $\beta \in [0.1, 1]$ 的实验结果, 发现 NFA 在该范围内结果稳定, 因此本文将 β 设置为该范围的中间值 0.5. NtN 和 NoisyGOA 中的参数设置与原始论文中报告或建议的方式一致. 本文用 q 表示蛋白质 i 的噪声功能标注数目, 将式 (6) 中 $\boldsymbol{V}(i,\cdot) \in \mathbb{R}^{|\mathcal{T}|}$ 中最小的 q 个值所对应的功能标签预测为该蛋白质的噪声功能标注. 根据 True Path Rule, 当判定一个功能标签为某个蛋白质的噪声功能标注, 若该功能标签的子孙标签已标注到该蛋白质上, 则这些子孙标签也为被判定蛋白质的噪声功能标注.

为量化地分析上述蛋白质噪声功能标注识别算法的性能, 参照文献 [44], 本文采用了 6 种性能评价度量, 分别为 MacroP, MacroR, MacroF1, MicroP, MicroR, 和 MicroF1, 它们的定义如下:

$$p_i = \frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FP}_i}, \quad r_i = \frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FN}_i}, \tag{7}$$

$$\mathrm{MacroP} = \frac{1}{N}\sum_{i=1}^{N} p_i, \quad \mathrm{MacroR} = \frac{1}{N}\sum_{i=1}^{N} r_i, \tag{8}$$

$$\mathrm{MacroF1} = \frac{1}{N}\sum_{i=1}^{N} \frac{2 \times p_i \times r_i}{p_i + r_i}, \tag{9}$$

$$\begin{aligned} \mathrm{MicroP} &= \frac{\sum_{i=1}^{N} \mathrm{TP}_i}{\sum_{i=1}^{N} (\mathrm{TP}_i + \mathrm{FP}_i)}, \\ \mathrm{MicroR} &= \frac{\sum_{i=1}^{N} \mathrm{TP}_i}{\sum_{i=1}^{N} (\mathrm{TP}_i + \mathrm{FN}_i)}, \end{aligned} \tag{10}$$

$$\mathrm{MicroF1} = \frac{2 \times \mathrm{MicroP} \times \mathrm{MicroR}}{\mathrm{MicroP} + \mathrm{MicroR}}, \tag{11}$$

TP_i 为第 i 个蛋白质的噪声功能标注中被正确识别的标签数目, FP_i 为被错误识别的标签数目, FN_i 为未被识别测出的噪声标签数目. p_i 和 r_i 分别表示在第 i 个蛋白质上的准确率和召回率.

3.3 蛋白质噪声功能标注识别结果分析

本小节主要测试并对比分析 NFA 在蛋白质噪声功能标注识别方面的性能. 与 CAFA2 [9] 中的实验设置类似, 本文独立重复运行每个对比算法 500 次, 每次随机选取 85% 的蛋白质及其功能标注进行检验, 并计算各个对比算法在上述 6 个评价度量上的值. 500 次独立运行的实验结果均值和方差汇报在表 4 和 5 中. 表格中加粗的结果表明该结果在配对 t 检验 (95% 置信度) 中显著优于其他结果, 或该结果与最优结果之间无显著性差异.

从这两个表中的结果可以观察到, 无论在 Macro 还是 Micro 度量上, NFA 在绝大多数情况都能获得较其他对比算法更高的准确率和 F1 值. 虽然 SR 和 NFA 都利用了稀疏表示来计算蛋白质之间的

表 4 在拟南芥上预测噪声标注的表现

Table 4 Performance of predicting noisy annotations in A. thaliana on archived GOA files

		LF	SR	EC	NtN	NoisyGOA	NFA
BP	MacroP	21.18±0.39	20.89±0.39	19.48±0.34	16.75±0.29	17.31±0.34	**27.79±0.46**
	MacroR	21.34±0.39	21.25±0.40	36.27±0.58	**44.12±0.63**	24.62±0.45	28.56±0.47
	MacroF1	21.25±0.39	21.05±0.39	23.58±0.39	22.10±0.35	19.48±0.37	**28.11±0.46**
	MicroP	51.20±0.58	46.50±0.65	36.67±0.41	26.30±0.36	35.90±0.48	**60.72±0.51**
	MicroR	51.93±0.58	47.96±0.65	**79.75±0.42**	79.08±0.38	55.71±0.68	63.55±0.51
	MicroF1	51.57±0.58	47.22±0.65	50.24±0.45	39.47±0.45	43.66±0.55	**62.10±0.51**
CC	MacroP	34.11±1.03	41.83±1.14	41.63±1.17	30.25±0.92	38.33±1.09	**46.34±1.31**
	MacroR	34.47±1.04	42.81±1.16	**71.22±1.68**	51.92±1.35	58.73±1.52	46.66±1.31
	MacroF1	34.21±1.03	42.25±1.15	**46.32±1.24**	35.27±1.01	42.72±1.16	**46.48±1.31**
	MicroP	63.87±0.96	68.60±0.84	37.73±1.24	34.63±0.86	47.14±0.97	**74.48±0.81**
	MicroR	64.63±0.95	71.29±0.86	**93.41±0.37**	81.83±0.68	84.22±0.69	75.52±0.81
	MicroF1	64.25±0.95	69.92±0.84	53.75±1.03	48.66±0.95	60.45±0.91	**75.00±0.81**
MF	MacroP	26.53±0.67	**29.93±0.73**	24.17±0.58	26.54±0.59	25.18±0.66	**30.06±0.72**
	MacroR	26.57±0.67	30.37±0.74	50.23±1.03	**56.27±1.08**	31.47±0.78	30.47±0.73
	MacroF1	26.55±0.67	**30.12±0.74**	29.02±0.64	33.41±0.69	27.07±0.69	**30.24±0.73**
	MicroP	56.62±0.81	**60.19±0.74**	30.28±0.52	35.23±0.57	51.99±0.70	**59.93±0.76**
	MicroR	56.85±0.81	61.49±0.76	**83.02±0.51**	**83.10±0.51**	68.70±0.74	62.37±0.76
	MicroF1	56.74±0.81	**60.83±0.75**	44.37±0.60	49.48±0.62	59.19±0.69	**61.13±0.76**

语义相似度, 但是 SR 并未考虑功能标注的证据属性, 因而其识别效果较 NFA 低一些. EC 仅利用了基于证据加权的蛋白质 – 功能标签关联矩阵识别蛋白质噪声功能标注, 而没有考虑蛋白质已有的功能标注信息和蛋白质之间的稀疏语义相似度, 其准确率和 F1 值也在大多数情况下低于 NFA. 本文也测试了在未加权的蛋白质 – 功能标签关联矩阵 $\boldsymbol{A}$ 上利用稀疏表示计算蛋白质之间的语义相似度, 再在式 (6) 中利用该相似度和加权的蛋白质 – 功能标签关联矩阵 $\boldsymbol{A}_{\text{ec}}$ 进行噪声功能标注识别, 该方法的结果 (未报告在表中) 略优于 SR, 但不及 NFA. 本文采用 Wilcoxon 符号秩检验 [45] 来评估 NFA 与这些对比算法在多个数据集和多种评价度量上的差异, 对应 p 值均小于 0.04, NFA 显著性地优于其他对比算法, 上述实验结果对比分析证明了本文提出的 NFA 在蛋白质噪声功能标注预测中的有效性.

NFA 在 Micro 系列度量 (MicroP, MicroR, MicroF1) 上的结果往往优于在 Macro 系列度量上的结果. 其主要原因是 Macro 系列度量平等地对待每个蛋白质, 如果某个蛋白质只含有一个噪声功能, 且未被预测出, 此蛋白质的准确率、召回率和 F1 值均为 0, 则会拉低相应的 Macro 度量值. 然而, Micro 系列度量平等地对待每个功能标注, 它们受那些具有较少噪声功能标注的蛋白质影响较小. NFA 在 MacroR 和 MicroR 的指标中表现不如 EC 和 NtN, 原因是 NFA 选取的噪声功能标注往往为 (或者靠近) 叶子标签, 由 True Path Rule 可知 NFA 选择的噪声功能标注数远小于 EC 和 NtN, 从而导致了较低的 MacroR 和 MicroR 值. 这些对比方法在 GO 的 BP 分支上的结果不及其他两个分支, 原因是 BP 分支上的噪声标注相比其他分支更多.

与 NtN 和 NFA 类似, NoisyGOA 也利用了蛋白质之间的语义相似度, 还利用了功能标签之间的分类相似度. 在大多数情况下, NoisyGOA 比 NtN 和 LF 具有更高的准确率, 偶尔还能获得最高的 F1 值, 这表明分类相似度对蛋白质噪声功能标注识别有贡献. 然而, NoisyGOA 的结果通常不及 SR,

表 5 在酵母菌上预测噪声标注的表现

Table 5 Performance of predicting noisy annotations in S. cerevisiae on archived GOA files

		LF	SR	EC	NtN	NoisyGOA	NFA
BP	MacroP	9.25 ±0.31	9.29±0.34	12.37±0.40	6.93±0.22	9.76±0.36	**13.07±0.42**
	MacroR	9.32±0.31	9.70±0.35	20.75±0.58	**26.46±0.65**	13.15±0.46	13.49±0.43
	MacroF1	9.28±0.31	9.47±0.34	**14.45±0.44**	10.11±0.29	10.86±0.39	13.26±0.43
	MicroP	32.37±0.74	28.77±0.88	27.14±0.68	15.02±0.34	26.80±0.73	**41.43±0.99**
	MicroR	33.20±0.75	30.39±0.92	65.11±0.95	**69.75±0.75**	42.33±1.01	44.18±1.03
	MicroF1	32.78±0.74	29.55±0.90	38.31±0.80	24.72±0.51	32.82±0.84	**42.76±1.01**
CC	MacroP	30.53±1.50	37.91±1.73	34.19±1.57	20.72±1.05	37.3±1.57	**40.72±1.73**
	MacroR	30.53±1.50	38.59±1.77	**54.75±2.15**	**53.27±2.17**	52.07±1.99	41.47±1.76
	MacroF1	30.53±1.50	38.22±1.75	38.40±1.67	26.81±1.24	**41.86±1.68**	**42.02±1.74**
	MicroP	58.93±1.53	62.89±1.35	38.98±1.56	22.60±0.91	51.93±1.19	**69.56±1.40**
	MicroR	59.10±1.52	64.54±1.41	82.49±1.03	**79.38±1.11**	78.55±1.13	71.29±1.46
	MicroF1	59.01±1.52	63.71±1.37	52.93±1.56	35.18±1.19	62.52±1.14	**70.42±1.42**
MF	MacroP	17.09±0.75	17.82±0.81	19.19±0.81	12.46±0.51	13.02±0.67	**21.38±0.87**
	MacroR	17.18±0.75	18.08±0.82	30.69±1.11	**40.55±1.27**	14.13±0.72	22.18±0.89
	MacroF1	17.13±0.75	17.94±0.81	**21.23±0.86**	17.25±0.64	13.44±0.68	**21.68±0.87**
	MicroP	36.56±1.10	36.50±1.13	22.83±0.75	17.19±0.53	28.51±1.23	**38.70±1.31**
	MicroR	37.35±1.12	36.50±0.73	60.15±1.28	**62.22±1.17**	35.16±1.38	43.66±1.16
	MicroF1	36.95±1.11	37.35±1.14	33.10±0.96	26.93±0.73	31.49±1.28	**41.03±1.15**

这表明语义相似度比分类相似度在蛋白质噪声功能标注识别中发挥着更大的作用. NFA 显著性优于 NoisyGOA 的原因有两方面: (1) NFA 根据蛋白质之间的稀疏语义相似度来赋予每个近邻蛋白质不同的投票权重, 而 NoisyGOA 并未区分近邻蛋白质之间的差别; (2) NFA 利用了功能标注的证据属性对蛋白质 – 功能标签关联矩阵进行加权, 但 NoisyGOA 未利用.

LF 选择在 N 个蛋白质中出现频率最低的功能标签为蛋白质的噪声功能标注. LF 的准确率和 F1 值偶尔比 NtN 和 NoisyGOA 更高, 这表明功能标签的频率是一种对噪声功能标注识别有帮助作用的重要特征. 事实上 NFA, SR 和 NoisyGOA 也利用了这一特征, 它们在噪声功能标注识别中均统计一个蛋白质的功能标签也标注到近邻语义蛋白质的频率, 以此来判断该蛋白质是否应该标注这个标签. SR 在此基础上对不同相似度的近邻设置不同的权重, NFA 在 SR 的基础上对蛋白质 – 功能标签关联矩阵进行了基于证据属性的加权.

本文还从每个蛋白质已有功能标注中随机选择功能标注并判定为噪声, 该方法的实验结果不及 LF, 远低于 NFA, 因此未在表中报告. 上述对比结果证明了蛋白质噪声功能标注的可识别性, 也证明了本文采用的稀疏语义相似度在蛋白质噪声功能标注识别中的有效性.

3.4 剔除噪声功能标注对蛋白质功能预测的贡献分析

本文还进行了另一组实验来检验分析 NFA 剔除的噪声功能标注对蛋白质功能预测的贡献. 实验从 BioGrid[3)] 下载 (日期为 2016-06-01) 了酵母菌和拟南芥两个物种的蛋白质互作网数据, 将式 (6) 中

3) http://thebiogrid.org/.

表 6 在拟南芥上的蛋白质功能预测结果 a)

Table 6 Results of protein function prediction on A. thaliana with/without removing noisy annotations

	BP		CC		MF	
	Historical	Removed	Historical	Removed	Historical	Removed
MicroAvgF1	75.38	**77.11**	78.36	**80.08**	66.83	**69.03**
MacroAvgF1	**70.25**	68.77	71.18	**71.49**	54.86	**58.07**
1-HammLoss	99.39	**99.44**	98.85	**98.94**	99.42	**99.46**
1-RankLoss	**98.17**	98.06	99.12	**99.13**	**99.04**	98.99
AvgPrec	67.11	**69.20**	76.32	**78.15**	64.46	**66.67**
AvgAUC	**83.05**	82.25	82.53	**83.42**	75.59	**76.81**
Fmax	88.27	**88.47**	**93.82**	**93.82**	92.17	**92.26**

a) Historical 与 Removed 对比中更好的结果用粗体表示.

表 7 在酵母菌上的蛋白质功能预测结果 a)

Table 7 Results of protein function prediction on S. cerevisiae with/without removing noisy annotations

	BP		CC		MF	
	Historical	Removed	Historical	Removed	Historical	Removed
MicroAvgF1	97.99	**98.00**	97.13	**97.14**	96.01	**96.02**
MacroAvgF1	**95.05**	**95.05**	95.65	**95.66**	93.99	**94.00**
1-HammLoss	**99.92**	**99.92**	**99.97**	**99.97**	**99.93**	**99.93**
1-RankLoss	**99.51**	**99.51**	**99.08**	**99.08**	**99.30**	**99.30**
AvgPrec	93.07	**93.30**	94.18	**94.29**	92.45	**92.55**
AvgAUC	97.19	**97.20**	**98.04**	**98.04**	**97.30**	97.29
Fmax	95.91	**96.03**	96.66	**96.71**	95.51	**95.56**

a) Historical 与 Removed 对比中更好的结果用粗体表示.

$\boldsymbol{V}(i,t)$ 值为 0 的功能标注判定为噪声功能标注, 更新式 (1) 中的蛋白质 – 功能标签关联矩阵 $\boldsymbol{A}$, 再利用文献 [46] 中基于互作邻居的投票方法和更新的 $\boldsymbol{A}$ 进行蛋白质功能预测. 本文也在未去噪的原始蛋白质 – 功能标签关联矩阵 $\boldsymbol{A}$ 上用同样的投票方法来预测蛋白质功能标注. 对应的实验结果汇报在表 6 和 7 中, 表中 Historical 列为原始预测结果, Removed 列为剔除 NFA 识别出的噪声功能标注后的预测结果.

蛋白质功能预测可以看作多标记学习问题进行研究 [47, 48], 本文选用 MicroAvgF1, MacroAvgF1, AvgAUC, RankLoss, HammLoss, AvgPrec 和 Fmax 这 7 种评价度量评价蛋白质功能预测的质量, 这些评价度量常被用于评价多标记学习和蛋白质功能预测的性能, 其中前 6 个多标记学习度量的定义可参见文献 [47]. AvgAUC 针对每个功能标签分别计算 Area Under receiver operating Curve (AUC) 的值, 取这些标签 AUC 均值作为最终的评价结果. 与 AvgAUC 类似, Fmax 也是国际大规模蛋白质功能预测评测组织推荐的评价准度量, 它的具体定义可参见文献 [8, 9]. 不同于其他度量, RankLoss 和 HammLoss 的值越小表示预测的质量越高. 为保持一致性, 实验中以 1-RankLoss 代替 RankLoss, 1-HammLoss 代替 HammLoss. 这些度量从不同的方面评测蛋白质功能预测的质量, 一个算法很难在所有度量上面超过另一个算法.

从表 6 和 7 中的结果可以看出, 剔除 NFA 识别出的噪声功能标注后蛋白质功能预测的质量在多个评价度量上普遍有了提升. 本文使用 Wilcoxon 符号秩检验 [45] 来检验 (Historical 和 Removed 列)

表 8　不同的证据权重分配策略, NFA 使用 List2

Table 8　Different weight configurations of evidence codes. NFA sets the weights of evidence codes via List2

	Experimental						Computational											Author		Curatorial	
	EXP	IDA	IEP	IGI	IMP	IPI	ISS	ISO	ISA	ISM	IGC	IBA	IBD	IKR	IRD	RCA	IEA	TAS	NAS	IC	ND
List1	1	1	0.6	1	1	1	0.4	0.4	0.4	0.4	0.6	0.6	0.6	0.6	0.6	0.6	0.4	0.8	0.4	0.8	0
List2	1	1	1	1	1	1	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	1	0.8	0.6	0.4
List3	1	1	1	1	1	1	0.4	0.4	0.4	0.4	0.6	0.6	0.6	0.6	0.6	0.6	0.6	1	0.8	0.6	0.4

表 9　不同证据权重方式在拟南芥上的噪声标注识别结果 [a)]

Table 9　Results of noisy annotations prediction on A. thaliana under different weight configurations of evidence codes

		List1	List3	NFA
BP	MacroP	20.21±0.39	**27.92±0.46**	**27.79±0.46**
	MacroR	20.57±0.39	**28.69±0.47**	**28.56±0.47**
	MacroF1	20.36±0.39	**28.24±0.47**	**28.11±0.46**
	MicroP	45.19±0.76	**61.06±0.52**	**60.72±0.51**
	MicroR	47.19±0.77	**63.77±0.52**	**63.55±0.51**
	MicroF1	46.17±0.77	**62.38±0.52**	**62.10±0.51**
CC	MacroP	44.11±1.25	**45.85±1.31**	**46.34±1.31**
	MacroR	44.45±1.26	**46.18±1.32**	**46.66±1.31**
	MacroF1	44.26±1.25	**46.00±1.31**	**46.48±1.31**
	MicroP	72.40±0.84	**74.07±0.80**	**74.48±0.81**
	MicroR	73.64±0.84	**75.15±0.80**	**75.52±0.81**
	MicroF1	73.02±0.84	**74.61±0.80**	**75.00±0.81**
MF	MacroP	26.25±0.69	**30.04±0.69**	**30.30±0.73**
	MacroR	26.73±0.70	**30.74±0.71**	**30.94±0.75**
	MacroF1	26.43±0.70	**30.34±0.70**	**30.57±0.74**
	MicroP	49.81±0.96	59.29±0.75	**60.73±0.73**
	MicroR	55.35±0.92	62.89±0.75	**63.36±0.73**
	MicroF1	52.43±0.93	61.03±0.75	**62.02±0.73**

a) 成对 t-test 检验 (95% 的置信度) 下更好的结果用粗体表示.

差异性, 对应 p 值小于 0.002, 即剔除 NFA 识别出的噪声功能标注能显著地提升后续蛋白质功能预测的质量.

3.5　证据加权策略分析

为了分析不同证据加权策略对蛋白质噪声标注识别效果的影响, 参照 3.3 小节的实验设置本文测试了在 3 种不同的加权策略 (见表 8) 下的对应 NFA 的性能. 其中, List1 参考了 Buza 等 [38] 对不同类型证据的加权方式, List2 对应 2.1 小节中表 2 的加权策略. 根据 Jones 等 [23] 的研究, 以证据属性为 ISS 的功能标注为基础预测蛋白质功能不可靠, 因此 List3 在 List2 的基础上, 将 ISS 及其 3 个子类 ISO, ISA, ISM 的权重调低. 这 3 种不同的证据加权方式在酵母菌和拟南芥上的结果汇报在表 9 和 10 中.

从这些表中可以发现, 不同的证据权重分配策略影响蛋白质噪声标注识别的性能. 在 3 种证据分

表 10 不同证据权重方式在酵母菌上的噪声标注识别结果 a)

Table 10 Results of noisy annotations prediction on S. cerevisiae under different weight configurations of evidence codes

		List1	List3	NFA
BP	MacroP	**12.91±0.41**	**12.81±0.43**	**13.07±0.42**
	MacroR	**13.31±0.42**	**13.30±0.44**	**13.49±0.43**
	MacroF1	**13.09±0.41**	**13.02±0.43**	**13.26±0.43**
	MicroP	40.91±1.01	**41.35±1.03**	**41.43±1.16**
	MicroR	43.77±1.04	**44.22±1.08**	**44.18±1.45**
	MicroF1	42.29±1.02	**42.74±1.06**	**42.76±1.42**
CC	MacroP	38.56±1.79	**40.16±1.69**	**39.89±1.73**
	MacroR	39.18±1.82	**41.13±1.73**	**40.86±1.76**
	MacroF1	38.81±1.80	**40.53±1.70**	**40.26±1.74**
	MicroP	65.50±1.50	**66.80±1.37**	**66.74±0.99**
	MicroR	67.81±1.55	**70.35±1.41**	**70.24±1.03**
	MicroF1	66.64±1.51	**68.53±1.35**	**68.44±1.01**
MF	MacroP	17.51±0.79	20.91±0.83	**21.38±0.87**
	MacroR	18.25±0.83	21.54±0.86	**22.18±0.89**
	MacroF1	17.80±0.81	21.15±0.84	**21.68±0.87**
	MicroP	34.28±1.28	37.57±1.19	**38.54±1.16**
	MicroR	39.07±1.34	42.50±1.19	**44.43±1.17**
	MicroF1	36.52±1.30	39.88±1.19	**41.27±1.16**

a) 成对 t-test 检验 (95% 的置信度) 下更好的结果用粗体表示.

配策略中, List2 表现相对最好, 而 List1 的表现最差. 这说明 Buza 等 [38] 提出的证据加权策略并不适用于蛋白质噪声标注识别. 另外, 本文还发现相比于 List2 和 List3, IEA 在 List1 中的权重较低, 而证据为 IEA 的功能标注占据某一物种蛋白质所有已知功能标注比例是最高的, 因此, 证据属性 IEA 的功能标注权重设置通常影响到整个物种上的蛋白质噪声功能标注识别的精度. 基于 List1 的结果不及 List2 和 List3 是由于其 IEA 权重设置较低导致预测结果的下降, 这是因为证据类型为 IEA 的蛋白质功能标注的可靠性并不比其他基于计算方法获得的功能标注的可靠性差 [37, 49]. List3 的性能略低于 List2 说明了赋予 ISS 低的权重并不能提升基于稀疏语义的蛋白质噪声功能标注识别的效果.

4 结束语

当前的蛋白质功能预测方法研究主要关注于预测完全未标注蛋白质的功能或对已部分标注功能的蛋白质进行缺失功能标注补全. 随着蛋白质功能标注信息在多种领域的广泛应用, 亟待用计算方法对 GO 数据库中的蛋白质噪声功能标注进行准确识别, 降低噪声功能标注对后续研究与应用的不利影响. 本文主要研究了蛋白质噪声功能标注的可识别性与如何准确识别噪声功能标注. 为实现这一目标, 本文提出了一种基于稀疏语义的蛋白质噪声功能标注识别方法 (NFA). 实验结果证明了蛋白质噪声功能标注的可识别性, 相比其他对比方法, NFA 可以更为准确地识别噪声功能标注. 实验分析还证明剔除 NFA 识别的蛋白质噪声功能标注能够提升蛋白质功能预测的精度. 稀疏语义相似度和功能标注的证据属性均有助于蛋白质噪声功能标注识别.

虽然本文通过实验证实了对蛋白质已有功能标注信息进行基于证据的加权可以提高噪声功能标注的识别效果, 剔除识别出的噪声标注信息可以提高后续蛋白质功能预测精度, 但如何设置这些证据最佳的权重和如何更准确地描述蛋白质之间的语义相似度都有待进一步深入研究.

参考文献

1 Amarda S, Barbará D, Molloy K. A survey of computational methods for protein function prediction. In: Big Data Analytics in Genomics. Berlin: Springer, 2016. 225–298

2 Barabási A L, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet, 2011, 12: 56–68

3 Wang P, Chen Y, Lv J H, et al. Graphical features of functional genes in human protein interaction network. IEEE Trans Biomed Circ Syst, 2016, 10: 707–720

4 Gene Ontology Consortium. Expansion of the gene ontology knowledgebase and resources. Nucleic Acids Res, 2017, 45: D331–D338

5 Huntley R P, Sawford T, Martin M J, et al. Understanding how and why the gene ontology and its annotations evolve: the GO within UniProt. GigaScience, 2014, 3: 4

6 Schones A M, Ream D C, Thorman A W, et al. Biases in the experimental annotations of protein function and their effect on our understanding of protein function space. PLoS Comput Biol, 2013, 9: e1003063

7 UniProt Consortium. Ongoing and future developments at the universal protein resource. Nucleic Acids Res, 2011, 39: D214–D219

8 Radivojac P, Clark W T, Oron T R, et al. A large-scale evaluation of computational protein function prediction. Nat Methods, 2013, 10: 221–227

9 Jiang Y X, Oron T R, Clark W T, et al. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. Genome Biol, 2016, 17: 184

10 Gillis J, Pavlidis P. Assessing identity, redundancy and confounds in gene ontology annotations over time. Bioinformatics, 2013, 29: 476–482

11 Gao L, Li X, Guo Z, et al. Broadly predicting specific protein functions with protein-protein interactions and gene expression profiles. Sci China Ser C Life Sci, 2006, 36: 441–450 [高磊, 李霞, 郭政, 等. 结合蛋白质互作与基因表达谱信息大范围预测蛋白质的精细功能. 中国科学 C 辑 生命科学, 2006, 36: 441–450]

12 Yu G X, Fu G Y, Wang J, et al. Predicting protein function via semantic integration of multiple networks. IEEE/ACM Trans Comp Biol Bioinform, 2016, 13: 220–232

13 Li Y H, Guo Z, Ma W C, et al. Predicting specific functions of protein with partial functions by protein-protein interactions network. Chin Sci Bull, 2007, 52: 2367–2373 [李彦辉, 郭政, 马文财, 等. 通过蛋白质互作网络预测已知部分功能的蛋白质的精细功能. 科学通报, 2007, 52: 2367–2373]

14 Fu G Y, Yu G X, Wang J, et al. Novel protein-function prediction using a direct hybrid graph. Sci Sin Inform, 2016, 46: 461–475 [傅广垣, 余国先, 王峻, 等. 基于有向混合图的蛋白质新功能预测. 中国科学: 信息科学, 2016, 46: 461–475]

15 Youngs N, Penfold-Brown D, Drew K, et al. Parametric bayesian priors and better choice of negative examples improve protein function prediction. Bioinformatics, 2013, 29: 1190–1198

16 Fu G Y, Wang J, Yang B, et al. NegGOA: negative GO annotations selection using ontology structure. Bioinformatics, 2016, 32: 2996–3004

17 Fu G Y, Yu G X, Wang J, et al. Protein function prediction using positive and negative examples. J Comp Res Dev, 2016, 53: 1753–1765 [傅广垣, 余国先, 王峻, 等. 基于正负样例的蛋白质功能预测. 计算机研究与发展, 2016, 53: 1753–1765]

18 Yu G X, Fu G Y, Wang J, et al. Predicting irrelevant functions of proteins based on dimensionality reduction. Sci Sin Inform, 2017, 47: 1349–1368 [余国先, 傅广垣, 王峻, 等. 基于降维的蛋白质不相关功能功能预测. 中国科学: 信息科学, 2017, 47: 1349–1368]

19 Mi H Y, Muruganujan A, Casagrande J T, et al. Large-scale gene function analysis with the PANTHER classification system. Nat Protoc, 2013, 8: 1551–1566

20 Kissa M, Tsatsaronis G, Schroeder M. Prediction of drug gene associations via ontological profile similarity with application to drug repositioning. Methods, 2015, 74: 71–82

21 Thomas P D, Mi H Y, Lewis S E. Ontology annotation: mapping genomic regions to biological function. Curr Opin Chem Biol, 2007, 11: 4–11

22 Gross A, Hartung M, Fer K, et al. Impact of ontology evolution on functional analyses. Bioinformatics, 2012, 28: 2671–2677

23 Jones C E, Brown A L, Baumann A U. Estimating the annotation error rate of curated GO database sequence annotations. BMC Bioinformatics, 2007, 8: 1–9

24 Rogers M, Ben-Hur A. The use of gene ontology evidence code in preventing classifier assessment bias. Bioinformatics, 2009, 25: 1173–1177

25 Benabderrahmane S, Smail-Tabbone M, Poch O, et al. IntelliGO: a new vector-based semantic similarity measure including annotation origin. BMC Bioinform, 2010, 11: 588

26 Caniza H, Romero A E, Heron S, et al. GOssTo: a stand-alone application and a web tool for calculating semantic similarities on the gene ontology. Bioinformatics, 2014, 30: 2235–2236

27 Mazandu G K, Chimusa E R, Mulder N J. Gene ontology semantic similarity tools: survey on features and challenges for biological knowledge discovery. Brief Bioinform, 2016, 18: 886–901

28 Guzzi P H, Mina M, Guerra C, et al. Semantic similarity analysis of protein data: assessment with biological features and issues. Brief Bioinform, 2011, 13: 569–585

29 Yu G X, Zhu H L, Domeniconi C, et al. Predicting protein function via downward random walks on a gene ontology. BMC Bioinform, 2015, 16: 271

30 Wu X M, Zhu L, Guo J, et al. Prediction of yeast protein-protein interaction network: insights from the gene ontology and annotations. Nucleic Acids Res, 2006, 34: 2137–2150

31 Lu C, Wang J, Zhang Z L, et al. NoisyGOA: noisy GO annotations prediction using taxonomic and semantic similarity. Comput Biol Chem, 2016, 65: 203–211

32 Donoho D L, Elad M, Temlyakov V N. Stable recovery of sparse overcomplete representations in the presence of noise. IEEE Trans Inf Theory, 2006, 52: 6–18

33 Wright J, Ma Y, Mairal J, et al. Sparse representation for computer vision and pattern recognition. Proc IEEE, 2010, 98: 1031–1044

34 Ma X, Zhuang W J, Feng J F. Loose sparse representation based undersampled face recognition with auxiliary dictionaries. Int J Pattern Recogn Artif Intell, 2016, 29: 439–446 [马晓, 庄雯璟, 封举富. 基于带补偿字典的松弛稀疏表示的小样本人脸识别. 模式识别与人工智能, 2016, 29: 439–446]

35 Wang J J, Bensmail H, Gao X. Feature selection and multi-kernel learning for sparse representation on a manifold. Neural Netw, 2014, 51: 9–16

36 Yu G X, Zhang G J, Zhang Z L, et al. Semi-supervised classification based on subspace sparse representation. Knowl Inf Syst, 2015, 43: 80–101

37 Rhee S Y, Wood V, Dolinski K, et al. Use and misuse of the gene ontology annotations. Nat Rev Genet, 2008, 9: 509–515

38 Buza T J, Mccarthy F M, Wang N, et al. Gene ontology annotation quality analysis in model eukaryotes. Nucleic Acids Res, 2008, 36: e12

39 Liu J, Ji S W, Ye J P. SLEP: Sparse Learning with Efficient Projections Version 4.1. 2013. http://www.yelab.net/publications/2009_slep.pdf

40 Good B M, Su A I. Crowdsourcing for bioinformatics. Bioinformatics, 2013, 29: 1925–1933

41 Good B M, Clarke E L, Alfaro L D, et al. The gene wiki in 2011: community intelligence applied to human gene annotation. Nucleic Acids Res, 2012, 40: D1255–D1261

42 Done B, Khatri P, Done A, et al. Predicting novel human gene ontology annotations using semantic analysis. IEEE/ACM Trans Comp Biol Bioinform, 2010, 7: 91–99

43 Salton G, Wong A, Yang C S. A vector space model for automatic indexing. Commun ACM, 1975, 18: 613–620

44 Zhou Z H. Machine Learning. Beijing: Tsinghua University Press, 2016. 24–28 [周志华. 机器学习. 北京: 清华大学出版社, 2016. 24–28]

45 Wilcoxon F. Individual comparisons by ranking methods. Biom Bull, 1945, 1: 80–83

46 Schwikowski B, Al E. A network of protein-protein interactions in yeast. Nat Biotechnol, 2000, 18: 1257–1261

47 Zhang M L, Zhou Z H. A review on multi-label learning algorithms. IEEE Trans Knowl Data Eng, 2014, 26: 1819–1837

48 Yu G X, Domeniconi C, Rangwala H, et al. Transductive multi-label ensemble classification for protein function

prediction. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, 2012. 1077–1085

49 Škunca N, Altenhoff A, Dessimoz C. Quality of computationally inferred gene ontology annotations. PLoS Comput Biol, 2012, 8: e1002533

Identifying noisy functional annotations of proteins using sparse semantic similarity

Chang LU[1], Xia CHEN[1], Jun WANG[1], Guoxian YU[1*] & Zhiwen YU[2]

1. *College of Computer and Information Sciences, Southwest University, Chongqing* 400715*, China*;
2. *College of Computer Science and Technology, South China University of Technology, Guangzhou* 510006*, China*
* Corresponding author. E-mail: gxyu@swu.edu.cn

Abstract Automatically annotating functions of proteins is a key task in bioinformatics. Functional annotations of proteins are collected from multiple sources; thus, noisy annotations are inevitably introduced. However, the current research in protein function prediction almost always focuses on predicting functions for completely unannotated (or incompletely annotated) proteins, and seldom identifies the noisy annotations of proteins. In this paper, we propose a method called identifying noisy functional annotations (NFAs) of proteins using sparse semantic similarity. NFA first utilizes a protein-function association matrix to store the functional annotations of proteins, differentially weighs the annotations using the evidence codes attached with these annotations, and subsequently upward propagates the weights to the expanded annotations via the hierarchical structure among the functional labels. Next, NFA measures the semantic similarity between proteins by the l_1-norm regularized sparse representation on the weighted protein-function association matrix. Finally, it identifies the noisy functions of a protein based on the functions annotated to its semantic neighborhood proteins. The experimental results on two model species (A. thaliana and S. cerevisiae) show that the NFA more accurately identifies noisy annotations than other related methods. Additionally, removing the identified noisy annotations improves the accuracy of the current function prediction model.

Keywords protein function, noisy functional annotations, sparse representation, semantic similarity, label structure

Chang LU was born in 1992. She received her B.Sc. degree in computer science from Jiangxi Normal University, Nanchang, in 2014. She is a master's student at the College of Computer and Information Sciences, Southwest University. Her research interests include machine learning and bioinformatics.

Xia CHEN was born in 1994. She received her B.Sc. degree in computer science from Chongqing Normal University, Chongqing, in 2016. Currently, she is a master's student at the College of Computer and Information Sciences, Southwest University. Her research interests include machine learning and bioinformatics.

Jun WANG was born in 1983. She received her Ph.D. degree in artificial intelligence from the Harbin Institute of Technology, Harbin, in 2010. Currently, she is an associate professor at the College of Computer and Information Science, Southwest University, Chongqing. Her research interests include machine learning, data mining, and their applications in bioinformatics.

Guoxian YU was born in 1985. He received his Ph.D. degree in computer science from the South China University of Technology, Guangzhou, in 2013. He is an associate professor at the College of Computer and Information Science, Southwest University, Chongqing, China. His research interests include data mining and bioinformatics.