中国科学: 信息科学 2019 年 第 49 卷 第 9 期: 1159–1174
SCIENTIA SINICA Informationis
论文

基于 0-1 矩阵分解的蛋白质功能预测

赵颖闻[1], 王峻[1], 郭茂祖[2,3], 张自力[1], 余国先[1*]

1. 西南大学计算机与信息科学学院, 重庆 400715
2. 北京建筑大学电气与信息工程学院, 北京 100044
3. 北京建筑大学建筑大数据智能处理方法研究北京市重点实验室, 北京 100044
* 通信作者. E-mail: gxyu@swu.edu.cn

收稿日期: 2018–12–12; 修回日期: 2019–02–11; 接受日期: 2019–02–28; 网络出版日期: 2019–09–04

国家自然科学基金 (批准号: 61872300, 61741217, 61873214, 61871020)、国家重点研发计划 (批准号: 2016YFC0901902)、中央高校基本科研业务 (批准号: XDJK2019B024) 和重庆市基础与前沿研究 (批准号: cstc2018jcyjAX0228, cstc2016jcyjA0351) 资助项目

摘要 准确地标注蛋白质功能是功能基因组学的核心任务之一. 蛋白质功能标注信息存在大量缺失且功能标签空间巨大. 近期一些标签压缩方法被提出并应用于蛋白质功能预测, 但是这些方法获取的压缩标签可解释性差, 且面临着多标记学习中的阈值划分难题. 为解决这些问题, 本文提出一种基于 0-1 矩阵分解的蛋白质功能预测方法 (zero-one matrix factorization, ZOMF). ZOMF 首先将蛋白质 – 功能标签关联矩阵分解成两个低秩 0-1 矩阵, 挖掘蛋白质和功能标签间的内在关联. 其次它利用蛋白质互作网和基因本体结构信息分别针对上述两个低秩矩阵定义了平滑正则项, 约束指导低秩矩阵的优化. 最后它利用优化获取的低秩矩阵重构关联矩阵, 进而实现蛋白质功能预测. 通过在酵母菌、拟南芥、老鼠和人类数据集上的实验表明, ZOMF 比已有的相关算法能够更准确地预测蛋白质功能, 它无需对重构的关联矩阵进行阈值划分, 压缩的 0-1 标签可解释性更直观.

关键词 蛋白质功能预测, 矩阵分解, 蛋白质互作网, 基因本体, 阈值划分

1 引言

蛋白质是生命活动的主要执行者, 一切生命活动都依赖于蛋白质功能的正确发挥, 如生物组织的构造、新陈代谢所需生物化学反应的催化、细胞环境的维护、生物信号的识别和传导等 [1~3]. 随着高通量生物技术的广泛应用, 收集到的蛋白质数据迅速增长, 如氨基酸序列、互作网和基因表达数据等. 蛋白质的功能信息也不断地从生物实验中检测出, 并添加到蛋白质功能标签数据库 (如 gene ontology [4], GO) 中. 蛋白质功能信息的精准标注对蛋白质机制的解析、疾病机理分析与调控、新药品研发、农作物促产和生物能源开发等诸多领域的研发都有着极大的促进作用 [3,5,6]. 然而由于高通量

引用格式: 赵颖闻, 王峻, 郭茂祖, 等. 基于 0-1 矩阵分解的蛋白质功能预测. 中国科学: 信息科学, 2019, 49: 1159–1174, doi: 10.1360/N112018-00331
Zhao Y W, Wang J, Guo M Z, et al. Protein function prediction based on zero-one matrix factorization (in Chinese). Sci Sin Inform, 2019, 49: 1159–1174, doi: 10.1360/N112018-00331

生物技术自身的不足和生物学家研究兴趣的偏向性, 蛋白质 (基因) 现有的功能标注信息还存在大量缺失, 且覆盖度有限, 也较浅层 [5]. 新发现的蛋白质速度远超过湿实验标注蛋白质功能的速度. 针对这些问题, 基于计算模型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的功能标注信息, 显著降低实验规模和成本 [3]. 然而这些方法普遍仅能给出蛋白质 – 功能标签间关联的概率值, 难以准确地判定蛋白质的相关功能标注和不相关功能标注, 且难以处理较大规模的功能标签集合.

已有研究发现, 蛋白质功能标签间的结构关系在蛋白质功能预测中发挥着至关重要的作用 [7,8]. 基因本体 [4] (GO) 作为一种广泛使用的蛋白质功能标注范式, 它通过一个有向无环图来刻画功能标签间层次结构关系. 标签之间存在一种叫做 True Path Rule 的规则 [4,7]: 子节点是父节点功能的进一步细化, 当一个蛋白质被标注为某个节点对应的功能时, 该蛋白质同时标注该节点的所有祖先节点对应的功能. 当一个蛋白质不具有某个功能时, 则该蛋白质不会拥有该功能节点及其所有子节点对应的功能. GO 功能标签划分在 3 个分支: 生物过程功能 (biological process, BP), 细胞成分功能 (cellular component, CC) 和分子功能 (molecular function, MF). GO 的功能标签数量巨大, 目前已经超过 45000 个.

一个蛋白质通常参与到不同的生命过程中, 发挥多个不同的生物学功能, 可以同时标注多个功能标签, 因此蛋白质功能预测问题可以看作是多标记学习问题 [9~13]. 然而由于蛋白质功能标注信息存在大量缺失和标签空间巨大, 现有基于多标记学习的预测方法面临着标注缺失和标签空间巨大等难题. 早期的一些方法首先应用功能标签筛选策略 [9,10], 选取至少被标注到 30 个蛋白质的标签作为研究对象, 忽略了量大且富含精细功能信息的稀疏标签, 而这些稀疏标签对应的功能信息更具有生命医学指导意义. 近期一些标签压缩方法被提出并被应用于蛋白质功能预测 [14~16], 提升了预测精度. 最近的研究表明有效地利用基因本体结构和蛋白质特征信息进行标签压缩可进一步提高预测精度 [17]. 然而现有的标签压缩方法获取的压缩标签为低维实数向量, 可解释性差, 难以从预测的蛋白质 – 功能标签关联概率矩阵中准确判定蛋白质的相关标注和不相关标注.

在总结分析已有研究工作的基础上, 本文提出一种可以融合蛋白质互作网和基因本体等生物数据的 0-1 矩阵分解方法 (zero-one matrix factorization, ZOMF) 预测蛋白质功能. ZOMF 首先将蛋白质 – 功能标签关联矩阵分解为 2 个低秩的 0-1 矩阵, 再在低秩矩阵上结合利用蛋白质互作信息和标签间的内在关联等生物学知识指导低秩矩阵的分解优化, 最后基于优化获取的低秩矩阵重构蛋白质 – 功能标签关联矩阵, 实现蛋白质功能预测. 本文工作的主要创新点如下:

• 提出的 ZOMF 能有效地融合蛋白质互作和基因本体数据挖掘蛋白质和标签间的潜在关联关系, 并在低秩空间有较好的可解释性. 需指出本文的方法还可以融合其他类型蛋白质特征数据.

• ZOMF 重构的蛋白质 – 功能标签矩阵为整数矩阵, 避免了在关联矩阵上进行阈值划分的难题, 而已有方法普遍仅能获得概率关联矩阵.

• 相比其他相关的蛋白质功能预测方法 [13~16], ZOMF 不仅拥有更高的预测精度和效率, 而且对参数鲁棒.

2 相关工作

针对功能标签间的层次结构关系, 一些研究者把蛋白质功能预测问题转化为层次多标记分类方法进行研究. Valentini [7] 提出了一种基于 True Path Rule 的蛋白质功能预测方法, 该方法针对每个功能标签分别训练二分类器, 再利用功能标签之间的层次结构关系整合和优化这些二分类器的预测结果,

取得了较好的预测精度. 但是该方法假定已标注的蛋白质功能信息完整, 再对完全未标注的蛋白质进行功能预测, 忽略了作为训练数据的蛋白质自身功能标注信息的不完整性, 预测精度有限. Tao 等 [18] 提出一种基于语义的蛋白质功能预测方法, 该方法首先利用功能标签间的层次结构关系计算功能标签间的分类相似度, 然后基于标注到两个蛋白质的成对标签之间的最大分类相似度衡量蛋白质之间的语义相似度, 最后利用 k 近邻分类器进行蛋白质功能预测. Done 等 [19] 结合 GO 结构和向量空间模型对蛋白质功能标注进行加权, 再利用奇异值分解预测蛋白质功能. Yu 等 [20] 提出一种基于层次弱标注的蛋白质功能预测方法, 该方法综合考虑功能标签间的水平关系和层次结构语义相似度, 再基于蛋白质已有的功能标注预估该蛋白质缺失的功能标注, 最后结合蛋白质互作网数据进行蛋白质功能预测. 然而该方法在预测缺失功能标注过程中没有充分考虑功能标签的层次结构信息, 易引入较多的假阳性预测.

为处理较大的功能标签集合, Wang 等 [14] 提出 ClusDCA 方法. 该方法分别在蛋白质互作网络和功能标签所在的层次结构网络上进行成分扩散分析, 补全蛋白质互作信息和功能标签间结构信息, 再通过奇异值分解 SVD (singular value decomposition) 分别求取蛋白质互作网和功能标签压缩的实数特征表示, 最后在压缩的标签空间通过 Logistic 回归预测蛋白质的功能. 近期 Yu 等 [15] 将图哈希 (Hash) 学习引入到大规模结构标签压缩中, 提出 HashGO 方法预测蛋白质功能. HashGO 利用图结构保持准则优化哈希函数, 进而将大量的功能标签二进制编码, 与此同时在低维空间中保持并利用标签间内在关联; 再将蛋白质 – 功能标签关联矩阵投影到低维哈希空间, 并计算蛋白质之间的语义相似度, 最后基于语义 k 近邻分类器进行蛋白质功能预测. 针对 HashGO 无法较好地保持标签间的层次结构关系, Zhao 等 [16] 提出一种基于 GO 结构保持的蛋白质功能预测方法 HPhash. 该方法首先计算功能标签间的分类相似度, 然后基于排序损失的层次保持哈希方法将大量的功能标签进行二进制编码, 并且在低维哈希空间保持功能标签间的层次结构关系; 再将蛋白质 – 功能标签关联矩阵映射到低维空间并计算蛋白质之间的语义关系; 最后通过基于语义的 k 近邻分类器实现蛋白质功能预测.

上述基于标签压缩的方法虽能较好地处理大量标签, 提升预测精度, 但是对应的压缩标签集合的可解释性差. 如 ClusDCA 方法在 SVD 分解后的压缩标记存在负值, 而且还需要设置合适的阈值去准确判定蛋白质的相关标注和不相关标注. 为避免上述问题, 本文考虑将蛋白质 – 功能标签重构为一个整数矩阵, 将该矩阵中非 0 元素对应的关联判定为蛋白质的相关标注, 其他元素为不相关元素. 为此, 本文提出一种基于 0-1 矩阵分解的蛋白质功能预测方法 ZOMF. ZOMF 获取的蛋白质功能标签关联矩阵为整数矩阵, 它分解的低秩矩阵不仅融入了生物特征信息和基因本体结构知识, 还避免了阈值划分的难题, 具有较强的低维可解释性. 它能够更精准地挖掘蛋白质和标签间的潜在关联关系, 提升预测效果. 实验证明, 相对于传统的实数矩阵分解, 本文引入的 0-1 矩阵分解不仅没有降低预测能力, 反而提升了预测效果.

3 基于 0-1 矩阵分解预测蛋白质功能

已知有 n 个蛋白质共计被 c 个不同的功能标签标注, $\boldsymbol{Y} \in \mathbb{R}^{n\times c}$ 存储这些蛋白质已知的功能标注信息. $\boldsymbol{Y}$ 基于 GO 结构初始化, 具体地当一个蛋白质标注有标签 t 对应的功能时, 该蛋白质也标注有 t 的祖先节点对应的功能, 反之则不一定. 基于这一规则, 本文对蛋白质 – 功能标签关联矩阵 $\boldsymbol{Y}$ 进行如下初始化:

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

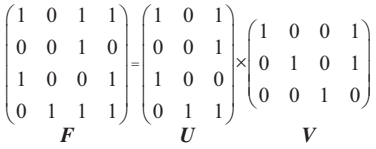

图 1　0-1 矩阵分解示例

Figure 1　An example of zero-one matrix factorization

需要指出的是 $\boldsymbol{Y}(i,t)=0$ 并不表示蛋白质 i 不应该标注标签 t, 而是目前还没有证据表明该蛋白质具有标签 t 对应的功能. 这一设置受蛋白质功能标注信息的不完整性和开放世界假设 (open world assumption) [5] 的影响. GO 数据库中通常仅登记蛋白质具有某个功能的信息, 极少登记该蛋白质不具有的功能信息, 原因是准确测定蛋白质所具有的全部功能非常困难, 生物学家通常更关注蛋白质具有的功能信息.

3.1　0-1 矩阵分解

一个蛋白质通常仅标注 GO 中 45000 多个功能标签中的几 (或几十) 个标签. 为处理大量的功能标签集合, 研究者提出了一些基于标签压缩的蛋白质功能预测方法. 如 Done 等 [19] 受 SVD 能够挖掘文本与单词间潜在关联的启发, 将 SVD 运用到蛋白质功能预测中. 具体地, 他们首先在 $\boldsymbol{Y}$ 上应用 SVD 分别挖掘蛋白质与标签间的潜在关联, 再基于 SVD 的低秩近似矩阵重构新的功能关联矩阵, 实现蛋白质功能预测. Wang 等 [14] 结合基因本体结构和蛋白质互作信息, 针对这两种信息分别进行 SVD 分解, 最后映射到统一空间进行蛋白质功能预测, 再将压缩的蛋白质 – 功能标签关联矩阵映射回原始空间实现最终的蛋白质功能预测, 但该方法仍然存在阈值难划分的问题 [21]. 原因是这些方法预测出来的结果是蛋白质与功能标签关联的概率值, 需采用合适的阈值划分技术将概率值转化为 0-1 值, 其中 1 表示蛋白质应标注该功能, 0 表示不标注该功能. 选择合适的阈值是公开的难题 [21], 不合适的阈值很可能引入较多的错误预测. 此外, 基于 SVD 分解的低秩矩阵是实数矩阵, 可解释性较差. 基于哈希学习的蛋白质功能预测方法 [15,16] 仅能在低维哈希空间获取整数型的关联矩阵, 在原始标签空间的关联矩阵仍为实数矩阵.

图 1 为 0-1 矩阵分解的示例, 其中 "×" 表示矩阵乘法运算. 假设矩阵 $\boldsymbol{F}$ 中每行对应一个蛋白质的功能标注信息, 每列代表一种功能标签, $\boldsymbol{F}(i,t)=1$ 表示已知第 i 个蛋白质标注第 t 个标签对应的生物功能. 本文希望通过 0-1 矩阵分解将 $\boldsymbol{F}$ 分解为两个低秩 0-1 矩阵 ($\boldsymbol{U}$, $\boldsymbol{V}$), 并且挖掘 3 个功能集合. 如 $\boldsymbol{U}$ 表示蛋白质 – 功能集合关联矩阵, $\boldsymbol{U}(i,k)=1$ 表示第 i 个蛋白质具有第 k 个功能集合的所有功能; 而矩阵 $\boldsymbol{V}$ 表示功能集合 – 功能标签隶属矩阵, $\boldsymbol{V}(k,t)=1$ 表示第 k 个功能集合包含第 t 个功能标签. 低秩 0-1 矩阵 $\boldsymbol{U}$ 和 $\boldsymbol{V}$ 通过乘法运算重构蛋白质 – 功能标签关联矩阵 $\boldsymbol{F}$. 因此, 利用 0-1 矩阵分解能将大量功能标签压缩到低维空间且具有较好的可解释性. 不同于已有基于矩阵分解的蛋白质功能预测研究 [14,17], 本文的方法从集合的角度考察不同功能标签集合之间的关联和集合内部多个标签间的内在关联. 需指出的是近期广泛研究的布尔矩阵分解 [22,23] 与本文的 0-1 矩阵分解均可获得元素值为 0-1 的因子矩阵, 但现有布尔矩阵分解方法在融合生物特征信息时普遍依赖于模块种子选取 [24,25], 而最优的模块种子通常很难选取, 导致无法有效地整合并利用不同的生物数据.

考虑到 $\boldsymbol{Y}$ 的稀疏高维非负特性和 0-1 矩阵分解在文本分析领域的成功应用 [26], 本文首先在蛋白质 – 功能标签关联矩阵 $\boldsymbol{Y}$ 上应用 0-1 矩阵分解, 以期挖掘蛋白质与大量标签间内在关联, 具体最小化

的目标方程为

$$\begin{aligned} &\min\ J_0(\boldsymbol{A},\boldsymbol{B}) = \sum_{i=1}^{n}\sum_{s=1}^{c}(\boldsymbol{Y}_{is} - (\boldsymbol{AB})_{is})^2 \\ &\text{s.t.}\ \ \boldsymbol{A}_{ik}^2 - \boldsymbol{A}_{ik} = 0;\ \boldsymbol{B}_{ks}^2 - \boldsymbol{B}_{ks} = 0, \end{aligned} \tag{2}$$

其中 $\boldsymbol{A} = (\boldsymbol{A}_{1\cdot}; \boldsymbol{A}_{2\cdot}; \ldots; \boldsymbol{A}_{n\cdot}) \in \mathbb{R}^{n\times k}$ 和 $\boldsymbol{B} = (\boldsymbol{B}_{\cdot 1}, \boldsymbol{B}_{\cdot 2}, \ldots, \boldsymbol{B}_{\cdot c}) \in \mathbb{R}^{k\times c}$ 为 2 个低秩 0-1 矩阵, 它们分别在压缩的 k ($k < \min(n, c)$) 维空间描述 n 个蛋白质的语义特征信息和 c 个功能标签的特征信息. 约束 $\boldsymbol{A}_{ik}^2 - \boldsymbol{A}_{ik} = 0$ 的引入使得 $\boldsymbol{A}$ 中元素的取值仅为 0 或 1, 进而实现 0-1 矩阵分解. 式 (2) 通过低秩 0-1 矩阵分解挖掘隐藏在 $\boldsymbol{Y}$ 中蛋白质之间的语义关联和标签间内在关联, 进而发现蛋白质与功能标签间的潜在关联. 然而式 (2) 没有显式考虑标签间的关联关系, 也没有考虑蛋白质的其他特征数据 (如蛋白质互作网和氨基酸序列数据等), 预测精度有限.

3.2 结合蛋白质互作信息和功能标签关联信息

3.2.1 结合蛋白质互作信息

矩阵 $\boldsymbol{A}$ 中每行可以看作是相应的蛋白质在矩阵 $\boldsymbol{B}$ 刻画的 k 个功能集合中的 0-1 向量表示, 但这种向量表示并没有结合蛋白质的其他特征信息. 高通量技术的广泛应用产生了海量的多源异构蛋白质数据, 如蛋白质互作网数据、结构域数据和氨基酸序列数据等, 其中蛋白质互作网是一种最常见和常用的蛋白质特征数据. 蛋白质互作网描述蛋白质如何通过互作来完成特定生物功能和参与到具体的生命过程, 其中每个节点对应一个蛋白质, 节点间的边描述蛋白质之间的互作. 研究表明互作的蛋白质更有可能具有相同的功能 [12], 为此本文拟在 $\boldsymbol{A}$ 上引入蛋白质互作网络的约束, 使得存在较强关联的蛋白质 i 和 j, 它们的低维向量表示 $\boldsymbol{A}_{i\cdot}$ 和 $\boldsymbol{A}_{j\cdot}$ 也相似. 为实现上述目标, 受平滑性假设启发 [27], 本文引入蛋白质互作网络上的平滑性约束项:

$$J_1(\boldsymbol{A}) = \frac{1}{2}\sum_{i,j=1}^{n} \|\boldsymbol{A}_{i\cdot} - \boldsymbol{A}_{j\cdot}\|^2 \boldsymbol{W}_{ij} = \mathrm{tr}(\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{D} - \boldsymbol{W})\boldsymbol{A}) = \mathrm{tr}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{A}), \tag{3}$$

其中 $\boldsymbol{W} \in \mathbb{R}^{n\times n}$ 为由 n 个蛋白质构成的互作网对应的邻接权重矩阵, 当蛋白质 i 与 j 存在互作时, $\boldsymbol{W}_{ij} > 0$, $\boldsymbol{W}_{ij}$ 的大小表示这两个蛋白质互作的强度或置信度. $\boldsymbol{D} \in \mathbb{R}^{n\times n}$ 为对角矩阵, $\boldsymbol{D}_{i,i} = \sum_{j=1}^{n} \boldsymbol{W}_{i,j}$, $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$. 最小化式 (3) 可以使得互作的成对蛋白质在低维语义空间彼此靠近, 这一目标也遵循了蛋白质之间的语义相似度与蛋白质之间的特征相似度正相关的特点. 因此, $J_1(\boldsymbol{A})$ 可以融合蛋白质互作网络 (或氨基酸序列等数据) 约束指导 $\boldsymbol{A}$ 的分解.

3.2.2 结合功能标签关联信息

一个蛋白质通常标注多个功能标签, 这些标签存在不同程度的关联和共现概率. 蛋白质功能预测问题可以转化为多标记学习问题进行研究, 面向蛋白质功能预测的多标记学习方法能够利用标签间的关联关系指导蛋白质功能预测, 显著提升了蛋白质功能预测精度 [9, 11, 20, 28], 式 (2) 仅通过矩阵分解隐式地挖掘蛋白质与标签间的关联关系, 稀疏标签容易由于标注的蛋白质个数较少而被忽略. Yu 等 [20] 统计发现, 对于蛋白质的缺失功能标注信息, 来自其直接父节点标注预估的置信度远大于其他祖先节点. 这是因为当已知该蛋白质标注了父亲节点标记时, 由 True Path Rule 可知该蛋白质也标注它的其他祖先节点对应的生物功能, 反之则不一定. 基于上述发现, 本文考虑利用功能标签节点间的父子关系, 这也体现了 GO 的有向无环图性质.

假设矩阵 $\boldsymbol{G} \in \mathbb{R}^{c\times c}$ 表示由 c 个结构功能标签构成的有向无环图, 它的每一列表示功能标签 t 和它的直接父标签节点构成的功能标签集合, $\boldsymbol{G}$ 的设置如下:

$$\boldsymbol{G}(t,s)=\begin{cases}1, \text{ 如果 } t \text{ 是 } s \text{ 的直接父节点},\\ 0, \text{ 其他}.\end{cases} \tag{4}$$

类似式 (3), 本文利用 $\boldsymbol{G}$ 对矩阵 $\boldsymbol{B}$ 的优化进行如下约束:

$$J_2(\boldsymbol{B})=\frac{1}{2}\sum_{s,t=1}^{c}\|\boldsymbol{B}_{\cdot s}-\boldsymbol{B}_{\cdot t}\|^2\boldsymbol{G}_{st}=\mathrm{tr}(\boldsymbol{B}(\boldsymbol{D}^c-\boldsymbol{G})\boldsymbol{B}^{\mathrm{T}})=\mathrm{tr}(\boldsymbol{B}\boldsymbol{L}^c\boldsymbol{B}^{\mathrm{T}}), \tag{5}$$

其中 $\boldsymbol{D}^c\in\mathbb{R}^{c\times c}$ 为对角矩阵, $\boldsymbol{D}^c_{s,s}=\sum_{t=1}^{c}\boldsymbol{G}_{t,s}$, $\boldsymbol{L}^c=\boldsymbol{D}^c-\boldsymbol{G}$. 矩阵 $\boldsymbol{B}$ 中的每列可以看作是对应标签隶属 k 个功能集合的 0-1 向量表示, 在 GO 中存在较强关联的标签 s 和 t, 它们的低维向量表示 $\boldsymbol{B}_{\cdot s}$ 和 $\boldsymbol{B}_{\cdot t}$ 应该类似. 因此最小化式 (5) 可以使得存在较强关联的标签拥有相似的低维向量表示, 进而使得存在较强关联的功能标签更可能标注到同一个蛋白质上.

3.3 统一的目标方程与优化求解

在 3.2 小节分析设计的基础上, 为处理较大的标签集合, 利用标签间关联性和蛋白质特征信息, 本文整合 $J_0(\boldsymbol{A},\boldsymbol{B})$, $J_1(\boldsymbol{A})$ 和 $J_2(\boldsymbol{B})$ 定义 ZOMF 的目标方程如下:

$$\begin{aligned}&\min\ J(\boldsymbol{A},\boldsymbol{B})=\sum_{i}^{n}\sum_{s}^{c}(\boldsymbol{Y}_{is}-(\boldsymbol{A}\boldsymbol{B})_{is})^2+\alpha\sum_{i,j}^{n}\|\boldsymbol{A}_{i\cdot}-\boldsymbol{A}_{j\cdot}\|^2\boldsymbol{W}_{ij}+\beta\sum_{s,t}^{c}\|\boldsymbol{B}_{\cdot s}-\boldsymbol{B}_{\cdot t}\|^2\boldsymbol{G}_{st}\\ &\text{s.t.}\ \ \boldsymbol{A}_{ik}^2-\boldsymbol{A}_{ik}=0,\ \boldsymbol{B}_{ks}^2-\boldsymbol{B}_{ks}=0,\end{aligned} \tag{6}$$

其中 $\alpha>0$ 和 $\beta>0$ 用于调控蛋白质互作网和标签关联性对低秩矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 的影响. 在获取优化后的低秩矩阵 $\boldsymbol{A}^*$ 和 $\boldsymbol{B}^*$ 之后, 本文通过

$$\boldsymbol{Y}^*=\boldsymbol{A}^*\boldsymbol{B}^* \tag{7}$$

重构蛋白质 – 功能标签之间的关联矩阵, 实现蛋白质功能预测. 由于 $\boldsymbol{A}^*$ 和 $\boldsymbol{B}^*$ 均为 0-1 矩阵, $\boldsymbol{Y}^*$ 是一个整数矩阵, 本文可直接判定 $\boldsymbol{Y}^*$ 中元素值为 0 的关联对应蛋白质的不相关标注, 其他为相关标注.

式 (6) 可以等价为

$$\begin{aligned}J(\boldsymbol{A},\boldsymbol{B})=&\frac{1}{2}\sum_{i}^{n}\sum_{s}^{c}(\boldsymbol{Y}_{is}-(\boldsymbol{A}\boldsymbol{B})_{is})^2+\frac{\alpha}{2}\sum_{i,j}^{n}\|\boldsymbol{A}_{i\cdot}-\boldsymbol{A}_{j\cdot}\|^2\boldsymbol{W}_{ij}+\frac{\beta}{2}\sum_{s,t}^{c}\|\boldsymbol{B}_{\cdot s}-\boldsymbol{B}_{\cdot t}\|^2\boldsymbol{G}_{st}\\ &+\frac{\lambda}{2}\sum_{i}^{n}\sum_{l}^{k}(\boldsymbol{A}_{il}^2-\boldsymbol{A}_{il})^2+\frac{\lambda}{2}\sum_{l}^{k}\sum_{s}^{c}(\boldsymbol{B}_{ls}^2-\boldsymbol{B}_{ls})^2\\ =&\frac{1}{2}\sum_{i}^{n}\sum_{s}^{c}(\boldsymbol{Y}_{is}-(\boldsymbol{A}\boldsymbol{B})_{is})^2+\alpha\mathrm{tr}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{A})+\beta\mathrm{tr}(\boldsymbol{B}\boldsymbol{L}^c\boldsymbol{B}^{\mathrm{T}})\\ &+\frac{\lambda}{2}\sum_{i}^{n}\sum_{l}^{k}(\boldsymbol{A}_{il}^2-\boldsymbol{A}_{il})^2+\frac{\lambda}{2}\sum_{l}^{k}\sum_{s}^{c}(\boldsymbol{B}_{ls}^2-\boldsymbol{B}_{ls})^2,\end{aligned} \tag{8}$$

其中 $\lambda\geqslant 0$ 为引入的 Lagrange 乘子.

本文引入梯度下降方法求解式 (8), 具体的更新公式为

$$\boldsymbol{A}_{ik}\leftarrow\boldsymbol{A}_{ik}+\eta_{ik}\frac{\partial J(\boldsymbol{A},\boldsymbol{B})}{\partial\boldsymbol{A}_{ik}},\ \ \boldsymbol{B}_{ks}\leftarrow\boldsymbol{B}_{ks}+\delta_{ks}\frac{\partial J(\boldsymbol{A},\boldsymbol{B})}{\partial\boldsymbol{B}_{ks}}, \tag{9}$$

Algorithm 1 ZOMF algorithm

Input: Protein-function association matrix $\boldsymbol{Y}$, protein-protein interaction matrix $\boldsymbol{W}$, GO adjacency matrix $\boldsymbol{G}$, low-rank parameter k, weight parameter α and β;

Output: Predicted protein-function association matrix $\boldsymbol{Y}^*$;

Initialize $\lambda = 10^{-16}$, $\varepsilon = 0.01$;

Randomly initialize matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ in the range of $(0, 1)$;

Normalize matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ according to Eq. (12);

DO

  Update matrix $\boldsymbol{A}$ according to Eq. (10);

  Update matrix $\boldsymbol{B}$ according to Eq. (11);

  $\lambda = 10\lambda$;

  Normalize matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ according to Eq. (12);

WHILE $(\boldsymbol{A}_{ik}^2 - \boldsymbol{A}_{ik})^2 + (\boldsymbol{B}_{ks}^2 - \boldsymbol{B}_{ks})^2 \geqslant \varepsilon$

Predict protein function using Eq. (7) and return matrix $\boldsymbol{Y}^*$.

其中 η_{ik} 和 δ_{ks} 是梯度下降中的步长参数. 本文利用偏导的特殊形式设置自适应步长 [29], 从而保证 $\boldsymbol{A}_{ik}$ 和 $\boldsymbol{B}_{ks}$ 的非负特性. 令 $\eta_{ik} = -\boldsymbol{A}_{ik}/((\boldsymbol{A}\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}})_{ik} + 2\lambda\boldsymbol{A}_{ik}^3 + \lambda\boldsymbol{A}_{ik} + 2\alpha(\boldsymbol{D}\boldsymbol{A})_{ik})$, 则 $\boldsymbol{A}_{ik}$ 的更新公式为

$$\begin{aligned}
\frac{\partial J(\boldsymbol{A},\boldsymbol{B})}{\partial \boldsymbol{A}_{ik}} &= -\sum_{s=1}^{c}(\boldsymbol{Y}_{is} - (\boldsymbol{A}\boldsymbol{B})_{is})\boldsymbol{B}_{ks} + 2\alpha(\boldsymbol{L}\boldsymbol{A})_{ik} + \lambda(2\boldsymbol{A}_{ik} - 1)(\boldsymbol{A}_{ik}^2 - \boldsymbol{A}_{ik}),\\
\boldsymbol{A}_{ik} &= \boldsymbol{A}_{ik} + \eta_{ik}\frac{\partial J(\boldsymbol{A},\boldsymbol{B})}{\partial \boldsymbol{A}_{ik}}\\
&= \boldsymbol{A}_{ik}\frac{(\boldsymbol{Y}\boldsymbol{B}^{\mathrm{T}})_{ik} + 3\lambda\boldsymbol{A}_{ik}^2 + 2\alpha(\boldsymbol{W}\boldsymbol{A})_{ik}}{(\boldsymbol{A}\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}})_{ik} + 2\lambda\boldsymbol{A}_{ik}^3 + \lambda\boldsymbol{A}_{ik} + 2\alpha(\boldsymbol{D}\boldsymbol{A})_{ik}}.
\end{aligned} \tag{10}$$

同理, 令 $\delta_{ks} = -\boldsymbol{B}_{ks}/((\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{B})_{ks} + 2\lambda\boldsymbol{B}_{ks}^3 + \lambda\boldsymbol{B}_{ks} + 2\beta(\boldsymbol{B}\boldsymbol{D}^c)_{ks})$, 则 $\boldsymbol{B}_{ks}$ 的更新公式为

$$\begin{aligned}
\frac{\partial J(\boldsymbol{A},\boldsymbol{B})}{\partial \boldsymbol{B}_{ks}} &= -\sum_{i=1}^{n}(\boldsymbol{Y}_{is} - (\boldsymbol{A}\boldsymbol{B})_{is})\boldsymbol{A}_{ik} + 2\beta(\boldsymbol{B}\boldsymbol{L}^c)_{ks} + \lambda(2\boldsymbol{B}_{ks} - 1)(\boldsymbol{B}_{ks}^2 - \boldsymbol{B}_{ks}),\\
\boldsymbol{B}_{ks} &= \boldsymbol{B}_{ks} + \delta_{ks}\frac{\partial J(\boldsymbol{A},\boldsymbol{B})}{\partial \boldsymbol{B}_{ks}}\\
&= \boldsymbol{B}_{ks}\frac{(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{Y})_{ks} + 3\lambda\boldsymbol{B}_{ks}^2 + 2\beta(\boldsymbol{B}\boldsymbol{G})_{ks}}{(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{B})_{ks} + 2\lambda\boldsymbol{B}_{ks}^3 + \lambda\boldsymbol{B}_{ks} + 2\beta(\boldsymbol{B}\boldsymbol{D}^c)_{ks}}.
\end{aligned} \tag{11}$$

通过迭代地计算式 (10) 和 (11), 最终求得优化的低秩 0-1 矩阵 $\boldsymbol{A}^*$ 和 $\boldsymbol{B}^*$.

为了防止 0-1 矩阵 $\boldsymbol{A}^*$ 和 $\boldsymbol{B}^*$ 中的某一方过于稀疏, 在迭代的过程中本文利用有界性定理 [26] 对矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 进行标准化, 具体的公式如下:

$$\boldsymbol{A}^* = \boldsymbol{A}\boldsymbol{D}_A^{-\frac{1}{2}}\boldsymbol{D}_B^{\frac{1}{2}},\ \boldsymbol{B}^* = \boldsymbol{D}_B^{-\frac{1}{2}}\boldsymbol{D}_A^{\frac{1}{2}}\boldsymbol{B}, \tag{12}$$

其中 $\boldsymbol{D}_A \in \mathbb{R}^{k\times k}$ 和 $\boldsymbol{D}_B \in \mathbb{R}^{k\times k}$ 都是对角矩阵, $\boldsymbol{D}_A = \mathrm{diag}(\max(\boldsymbol{A}_{\cdot 1}), \max(\boldsymbol{A}_{\cdot 2}), \ldots, \max(\boldsymbol{A}_{\cdot k}))$, $\boldsymbol{D}_B = \mathrm{diag}(\max(\boldsymbol{B}_{1\cdot}), \max(\boldsymbol{B}_{2\cdot}), \ldots, \max(\boldsymbol{B}_{k\cdot}))$.

ZOMF 算法的流程如算法 1 所示.

表 1 蛋白质功能标注信息统计

Table 1 Statistics of functional annotations of proteins

Species	Proteins (n)	Branch	2016 (Avg ± std)	2017 (Avg ± std)	Labels (c)
Yeast	6017	BP	55849 (9.28 ± 12.57)	56971 (9.47 ± 13.14)	2036
		MF	15783 (2.62 ± 3.56)	15899 (2.64 ± 3.58)	777
		CC	17872 (2.97 ± 4.30)	19765 (3.28 ± 4.58)	543
Arabidopsis	9228	BP	41486 (4.50 ± 10.12)	47159 (5.11 ± 11.53)	1649
		MF	11517 (1.25 ± 3.18)	14634 (1.59 ± 3.84)	600
		CC	13009 (1.41 ± 3.86)	14012 (1.52 ± 4.07)	321
Mouse	5585	BP	125100 (22.40 ± 35.41)	148721 (26.63 ± 42.03)	5077
		MF	23014 (4.12 ± 5.68)	28746 (5.15 ± 6.84)	1098
		CC	20842 (3.73 ± 5.36)	28118 (5.03 ± 7.04)	731
Human	16073	BP	153772 (9.57 ± 18.72)	170727 (10.62 ± 20.57)	5408
		MF	35524 (2.21 ± 3.42)	39028 (2.43 ± 3.63)	1626
		CC	23228 (1.45 ± 3.01)	27305 (1.70 ± 3.28)	769

4 实验与结果

4.1 数据集

不同于以往的实验评价方法, 本文采用一种历史到现在的方式对比 ZOMF 和其他相关方法的性能. 首先收集 4 种模式物种 (Yeast, Arabidopsis, Mouse 和 Human) 历史的 (归档日期: 2016–05–07) 功能标注数据作为训练集进行功能预测, 再利用新近 (归档日期: 2017–11–09) 的功能标注数据作为验证集. 本文还下载了同期归档的 GO 文件[1)], 并在 GO 3 个分支上 (BP, MF, CC) 分别对蛋白质进行功能标注. 为避免循环预测问题, 筛除证据属性为 IEA (inferred by electronic annotation), NR (not recorded), ND (no biological data available) 和 IC (inferred by curator) 的功能标注. 蛋白质互作网数据从 BioGrid 数据库收集获取 (Version 3.4.137). 参照 clusDCA [14] 的实验设置, 本文将标注的蛋白质个数不小于 3 且不大于 300 的标签均予保留进行实验分析. 表 1 中统计了 2016–05 和 2017–11 两个时间节点每个物种的蛋白质在 3 个分支上的功能标注数和相应的标签个数.

从表 1 可以看出蛋白质功能标注信息在不断地增多完善, 如 Human 的 16073 个蛋白质在 BP 分支的功能标注数从 153772 个增加到 170727 个, 这些蛋白质共计被 5408 个不同的功能标签标注, 从如此大的标签空间中准确预测蛋白质的功能非常困难. 此外, 表中 Avg ± std 对应数据集中每个蛋白质的平均功能标签个数和相应的方差, Avg < std 说明蛋白质的功能标注信息不平衡, 进一步增加了蛋白质功能预测的难度.

4.2 对比方法和评价度量

本文选取新近且具有代表性的蛋白质功能预测方法 ClusDCA [14], NewGOA [13], HPhash [16] 和经典的 MV [12] 作为对比方法. 其中, ClusDCA 在相关工作中已介绍, 不赘述. MV 是在蛋白质互作网上利用 'Guilt by Association' 规则, 基于蛋白质互作邻居投票的方式进行蛋白质功能预测. HPhash [16] 首先利用层次结构保持哈希将大量结构标签哈希到低维空间, 并将蛋白质 – 功能标签关联矩阵投影到低维哈希空间, 再结合蛋白质互作网和类似 MV 的方式在低维空间进行蛋白质功能预测, 最后将

1) http://geneontology.org/.

预测映射回原始空间. NewGOA 基于功能标签间的层次结构、蛋白质间的互作以及蛋白质 – 功能标签的关联关系构造混合图, 再在这个混合图上采用双随机游走策略预测蛋白质功能. 此外本文还引入 ZOMF(Y), ZOMF(PPI) 和 ZOMF(GO) 这 3 个 ZOMF 的变种作为对比方法. 其中, ZOMF(Y) 仅利用 $\boldsymbol{Y}$ 的分解重构 (即 $\alpha = 0$, $\beta = 0$) 进行蛋白质功能预测; ZOMF(PPI) 只利用蛋白质互作数据指导 0-1 矩阵分解 ($\alpha > 0, \beta = 0$), 再重构蛋白质 – 功能标签关联矩阵实现功能预测; 而 ZOMF(GO) 只用基因本体数据指导 0-1 矩阵分解 ($\alpha = 0, \beta > 0$). 这些对比算法的参数参照原文建议的方式优化后设置.

为综合评价上述功能预测算法的性能, 本文采用国际蛋白质功能预测评价组织 CAFA (community critical assessment of protein function annotation) [1] 推荐的评估度量: Fmax 和 Smin. 由于蛋白质功能预测问题可转化为多标记学习问题进行研究, 本文也采用了多标记学习度量 [30]: MacroF1 和 MicroF1. 其中 Fmax 和 Smin 是以蛋白质 (样本) 为中心的评价准则. Fmax 首先计算不同阈值下的准确率 (precision) 和查全率 (recall) 并计算该阈值对应的 F1 值, 最后选取最大 F1 值作为 Fmax 的值; Smin 结合基因本体结构首先计算不同阈值下未被预测到的功能标签和过度预测的错误标签之间的语义距离, 最后选择最小的距离作为 Smin 的值. MicroF1 和 MacroF1 均是以功能标签为中心的评价度量. MicroF1 计算不同功能标签 F1-Score 的和, 这一评价度量受频繁功能标签影响较大; MacroF1 先求取每个标签的 F1-Score, 再取这些标签 F1-Score 的均值, 这一评价度量受稀疏功能标签影响较大. 所有的对比方法在 Fmax, MicroF1 和 MacroF1 度量上的值越高, 表示其预测质量越好, 而在 Smin 度量上的值越小表示其预测质量越好.

MacroF1 和 MicroF1 需将蛋白质 – 功能标签关联概率矩阵转化为 0-1 矩阵. ZOMF 的预测结果是整数矩阵, 无需转化, 但对比方法预测的结果是 0 至 1 之间的概率值, 需进行转换. 参照 HPhash 和 NewGOA 的实验设置, 本文将预测的蛋白质 – 功能标签关联概率值先进行降序排序, 然后基于每个蛋白质更新的功能标注数量选择相应数量且置信值最高的标签作为该蛋白质相关标注.

4.3 实验结果

本文通过实验对比分析 ZOMF 在蛋白质功能预测中的有效性, 此部分实验中, 降维 (或低秩矩阵分解) 的目标维度统一设置为 80, 对应的实验结果见表 2∼5. 表 2∼5 中每种度量下最好的结果用粗体表示, 需指出表中 ↓ 表示值越小结果越好.

从表 2∼5 中可以看到 ZOMF 在整体上要优于其他对比算法以及自身变种. 由于表 2∼5 中结果是基于历史的蛋白质功能标注数据预测并用新近的功能标注数据检验, 因此结果中不存在方差, 为此本文利用 Wilcoxon 符号秩检验 [31,32] 分析对比 ZOMF 与对比方法在这些数据集和评价度量下的结果, 对应 p 值均小于 10^{-6}. 从上述对比结果可知, ZOMF 显著性优于已有基于随机游走、SVD 标签压缩和哈希标签压缩的蛋白质功能预测算法. 尽管 ZOMF 及其变种在判定蛋白质的相关标注和计算 MacroF1 及 MicroF1 度量时没有参照蛋白质已有的功能标注数量, 而其他对比算法在上述过程中均参照了蛋白质已有的功能标注信息, ZOMF 及其变种依然获得了较这些对比算法更好的预测效果, 这一对比表明 ZOMF 既可以避免阈值划分难题, 还能保持较好的性能. 上述实验结果证明了利用 0-1 矩阵分解进行蛋白质功能预测的有效性.

ClusDCA, NewGOA 和 HPhash 都利用了蛋白质互作信息和功能标签之间的层次结构关系, 它们均获得了较基线方法 MV 更好的结果, 这表明功能标签之间的层次结构关系在蛋白质功能预测中的重要性. 在这 3 个方法中, 本文将 ClusDCA, HPhash 与 NewGOA 进行符号秩检验, p 值分别是 0.195% 和 0.01%. ClusDCA 和 HPhash 获得了较 NewGOA 更好的实验结果, 这说明标签压缩方法可以较好地克服标注稀疏和缺失的问题, 原因是这些方法在标签压缩的过程中挖掘并利用了标签间结构关系.

表 2　在 Yeast 上的实验结果

Table 2　The results on Yeast

		MV	ClusDCA	NewGOA	HPhash	ZOMF(Y)	ZOMF(GO)	ZOMF(PPI)	ZOMF
MicroF1	BP	0.9368	0.9475	0.9455	0.9401	0.9351	0.9351	0.9491	**0.9510**
	MF	0.9378	0.9470	0.9491	0.9397	0.9363	0.9376	0.9502	**0.9554**
	CC	0.8911	0.8995	0.8965	0.8731	0.9129	0.9138	0.9193	**0.9200**
MacroF1	BP	0.9352	0.9397	0.9154	0.9252	0.9342	0.9353	**0.9435**	**0.9435**
	MF	0.9347	0.9464	0.9275	0.9236	0.9387	0.9391	**0.9474**	**0.9474**
	CC	0.9192	0.9252	0.8952	0.8956	0.9366	0.9376	0.9449	**0.9452**
Fmax	BP	0.8861	0.9508	0.9552	0.9716	0.9458	0.9497	0.9652	**0.9756**
	MF	0.8229	0.8706	0.8647	0.8814	0.8753	0.8759	0.8852	**0.8872**
	CC	0.7250	0.7684	0.7765	0.8162	0.8070	0.8070	0.8185	**0.8196**
Smin ↓	BP	1.5707	0.5481	0.3948	0.3986	0.4673	0.4677	0.3689	**0.3603**
	MF	0.4110	0.2011	0.2012	0.1740	0.1945	0.1980	0.1545	**0.1543**
	CC	0.3677	0.1625	0.1675	0.1357	0.1317	0.1232	**0.1093**	**0.1093**

表 3　在 Arabidopsis 上的实验结果

Table 3　The results on Arabidopsis

		MV	ClusDCA	NewGOA	HPhash	ZOMF(Y)	ZOMF(GO)	ZOMF(PPI)	ZOMF
MicroF1	BP	0.7977	0.8511	0.8479	0.8325	0.8818	0.8822	0.8850	**0.8851**
	MF	0.7344	0.7724	0.7709	0.7452	0.8250	**0.8260**	0.8224	0.8259
	CC	0.8551	0.8863	0.8877	0.8651	0.9099	0.9078	0.9170	**0.9171**
MacroF1	BP	0.8162	0.8593	0.8016	0.8337	0.8855	0.8856	0.8868	**0.8869**
	MF	0.7955	0.8044	0.7372	0.7771	0.8424	0.8432	0.8472	**0.8508**
	CC	0.8184	0.8370	0.8096	0.7893	0.8556	0.8610	0.8561	**0.8639**
Fmax	BP	0.8337	0.8928	0.9039	**0.9146**	0.9054	0.9057	0.9068	0.9068
	MF	0.7319	0.7643	0.7605	**0.8093**	0.8087	0.8087	0.7910	0.7910
	CC	0.6341	0.5882	0.6039	**0.7144**	0.7069	0.7101	0.7057	**0.7144**
Smin ↓	BP	2.1709	1.0860	1.0391	1.0097	1.0065	1.0056	0.9968	**0.9964**
	MF	0.9126	0.7410	0.7707	0.6449	0.6130	**0.5930**	0.6005	0.6003
	CC	0.4977	0.5761	0.5077	0.2576	0.2540	0.2546	0.2600	**0.2476**

ZOMF 的预测性能优于 NewGOA, 原因是 NewGOA 在进行重启随机游走时引入了一些噪声信息, 在一定程度上影响了预测结果. HPhash 在 Fmax 度量上显著优于 ZOMF, 原因是 Fmax 在 $[0,1]$ 的阈值范围内计算每个阈值对应的 F1-Score, 选择最大的 F1-Score 作为评价结果; 而 ZOMF 通过对蛋白质 – 功能标签关联矩阵 $\boldsymbol{Y}$ 先进行 0-1 矩阵分解再重构, 重构的矩阵 $\boldsymbol{Y}^*$ 为非负整数矩阵, 在不同阈值下的 F1-Score 不变.

ZOMF 比 ZOMF(PPI) 和 ZOMF(GO) 获得了更好的预测结果, 特别是在 MacroF1 上, 主要原因是 MacroF1 受算法在稀疏功能标签上的性能影响较大, 这说明有效地利用基因本体结构和蛋白质特征信息进行 0-1 矩阵分解可进一步提高预测精度. 实际上本文还将 ZOMF(PPI), ZOMF(GO) 与 ZOMF(Y) 进行符号秩检验, p 值分别是 10^{-6} 和 10^{-2}. 由此可见 ZOMF(Y) 对应的实验结果显著低于 ZOMF(PPI) 和 ZOMF(GO), 这进一步说明引入蛋白质互作信息和基因本体信息指导 0-1 矩阵的分解可以显著提高蛋白质的预测精度. 在 Arabidopsis 和 Human 数据集上, ZOMF 的部分评价度量结果不

表 4 在 Mouse 上的实验结果

Table 4 The results on Mouse

		MV	ClusDCA	NewGOA	HPhash	ZOMF(Y)	ZOMF(GO)	ZOMF(PPI)	ZOMF
MicroF1	BP	0.7646	0.8229	0.8211	0.8131	0.8527	0.8538	0.8682	**0.8703**
	MF	0.7482	0.7962	0.7942	0.7827	0.8575	0.8575	0.8580	**0.8581**
	CC	0.7061	0.7541	0.7542	0.7263	0.8138	0.8137	0.8197	**0.8202**
MacroF1	BP	0.7689	0.8284	0.7558	0.8015	0.8569	0.8570	0.8572	**0.8576**
	MF	0.7651	0.8098	0.7371	0.7833	0.8283	0.8342	0.8352	**0.8390**
	CC	0.7464	0.7706	0.7077	0.7498	0.8108	0.8137	0.8195	**0.8196**
Fmax	BP	0.7890	0.8582	0.8537	**0.8916**	0.8775	0.8776	0.8806	0.8806
	MF	0.7091	0.7862	0.7432	0.7983	0.7997	0.7997	**0.8001**	**0.8001**
	CC	0.6334	0.6697	0.6207	0.7062	0.7038	0.7037	0.7092	**0.7093**
Smin ↓	BP	7.2180	6.1819	5.2861	5.4010	2.5646	2.5648	2.5440	**2.5345**
	MF	1.1973	0.7469	0.8482	0.8490	0.6953	0.6953	0.6881	**0.6852**
	CC	0.9895	0.7845	0.9694	0.7990	0.6225	0.6126	0.6093	**0.6071**

表 5 在 Human 上的实验结果

Table 5 The results on Human

		MV	ClusDCA	NewGOA	HPhash	ZOMF(Y)	ZOMF(GO)	ZOMF(PPI)	ZOMF
MicroF1	BP	0.8538	0.8862	0.8876	0.8819	0.9051	0.9051	0.9131	**0.9139**
	MF	0.8638	0.8942	0.8993	0.8883	0.9130	0.9134	0.9219	**0.9228**
	CC	0.8356	0.8623	0.8608	0.8431	0.8752	0.8751	0.8854	**0.8951**
MacroF1	BP	0.8699	0.9015	0.8480	0.8865	0.9120	0.9121	0.9139	**0.9148**
	MF	0.8792	0.9153	0.8759	0.8932	0.9201	0.9202	**0.9225**	**0.9225**
	CC	0.8478	0.8776	0.8301	0.8520	0.8833	0.8834	0.8906	**0.8935**
Fmax	BP	0.7538	0.8637	0.8428	**0.8959**	0.8812	0.8812	0.8862	0.8863
	MF	0.6493	0.7408	0.6902	**0.7587**	0.7494	0.7499	0.7527	0.7559
	CC	0.4598	0.5502	0.4692	0.5643	0.5524	0.5623	0.5649	**0.5688**
Smin ↓	BP	3.1853	1.6946	1.4567	1.3245	0.7708	0.7701	**0.7504**	0.7510
	MF	0.5476	0.2589	0.3674	0.2541	0.2067	0.2060	0.1995	**0.1975**
	CC	0.4465	0.2037	0.3725	0.2298	0.1697	0.1698	0.1573	**0.1474**

及变种 ZOMF(GO) 和 ZOMF(PPI), 原因可能是 Arabidopsis 和 Human 数据集的功能标注信息过于稀疏. 需要指明的是, 这部分实验涉及的参数 α 和 β 分别取 0.1 和 0.01, 后面将对这两个参数的鲁棒性进行讨论.

4.4 参数敏感性分析

4.4.1 低秩大小 k 敏感性分析

ZOMF 结合基因本体结构和蛋白质互作网指导蛋白质 – 功能标签矩阵 $\boldsymbol{Y}$ 分解为低秩 0-1 矩阵: $\boldsymbol{A}$ 和 $\boldsymbol{B}$. 为分析不同秩大小 k 对预测结果的影响, 本文固定 $\alpha = 0.1$ 和 $\beta = 0.01$, 对 k 进行了敏感性分析, 不同 k 值 (5 至 200) 下的 Fmax 和 Smin 的结果值如图 2 所示.

从图 2 中曲线可以发现, ZOMF 的性能随着 k 的增大而不断提升, 而当 $k \geqslant 80$ 时预测结果趋于稳定, 这说明 ZOMF 对 k 是比较鲁棒的. ZOMF 在 $k \leqslant 80$ 时有一定的波动, 原因可能是蛋白质数目

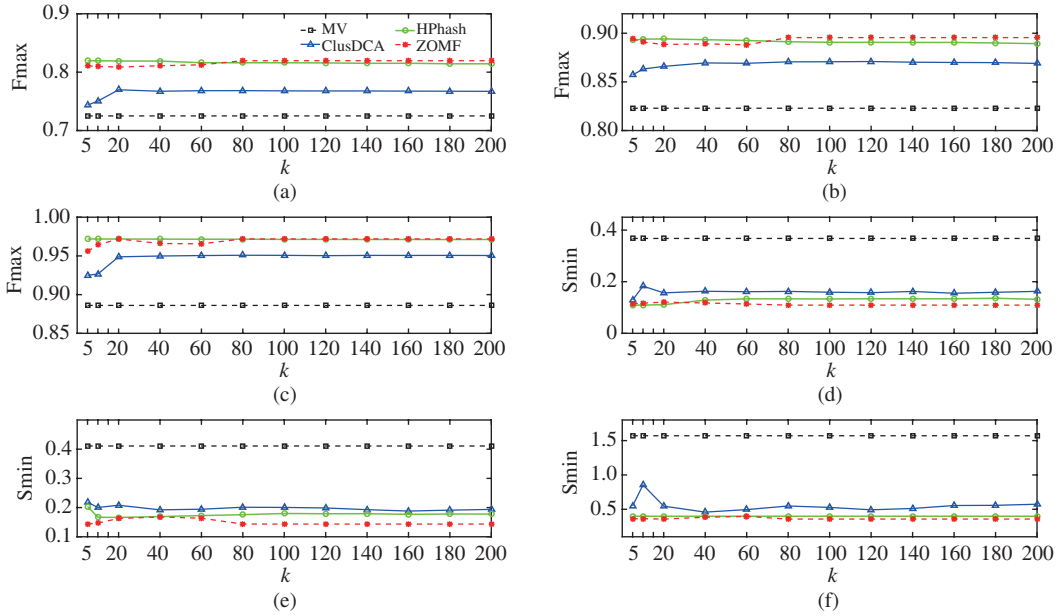

图 2 (网络版彩图) 在 Yeast 上低秩矩阵大小 k 的敏感性分析

Figure 2 (Color online) Sensitivity analysis of low-rank parameter k on Yeast. (a) CC (Fmax); (b) MF (Fmax); (c) BP (Fmax); (d) CC (Smin); (e) MF (Smin); (f) BP (Smin)

较多和功能标签空间巨大, 过小的 k 不足以充分刻画蛋白质和标签间内在关联, 从而对预测结果产生一定的影响. MV 的结果最稳定, 原因是 MV 不依赖于 k 的设置. ClusDCA 最不稳定, 它需选择一个合适的 k 进行奇异值分解, 但是选择一个合适的 k 是非常难的. 随着 k 的增大, HPhash 的结果逐渐趋于稳定, 这是因为当 k 较小时, 只能用有限的二进制位数将大量的功能标签进行编码, 会造成不同功能标签编码的重叠, 从而对预测结果造成一定的影响. 此外, 本文对参数 k 做了更高维度的敏感性分析, 预测精度保持不变. 原因是 $k \in [80, 200]$ 已经能较好地刻画蛋白质的语义特征信息和功能标签特征信息, 以及它们内在关联. 增大 k 会引入更多的特征并提高计算复杂度, 但不会提升精度.

4.4.2 参数 α 和 β 敏感性分析

ZOMF 整合基因本体结构和蛋白质互作网指导蛋白质 – 功能标签关联矩阵的 0-1 矩阵分解. 其中 α 和 β 分别为蛋白质互作网和基因本体约束所对应的权重参数 (见式 (6)). 为分析这两个参数对预测结果的影响, 本文在 3 个物种的 CC 分支上对 α 和 β 进行了敏感性分析. 本文固定维度参数 $k = 80$, 并将不同 α 和 β 值 (10^{-2} 至 10^{8}) 下的 Fmax 和 Smin 结果值汇报在图 3 中. 分析图 3 中结果可以发现, 当 $10 \leqslant \alpha \leqslant 10000$ 而 β 在给定的区间范围内任意取值时, 预测的结果几乎不变; 但是当 α 取值过小或过大时, 均会影响蛋白质功能预测的结果, 这表明蛋白质互作数据对蛋白质功能预测的影响较大. α 和 β 在较大组合范围内的取值变化对 ZOMF 的性能并没有明显的影响, 这说明 ZOMF 对 α 和 β 也是比较鲁棒的.

4.5 运行时对比

为了分析 ZOMF 和其他对比算法的运行效率, 本文登记了这些对比算法的实际运行时间, 如表 6 所示. 本部分实验设置与 4.1 小节实验设置类似, 各算法均基于 Matlab2014a (64 位) 编码实现, 实验运行平台配置为: Intel(R) Xeon(R) CPU E5-2678v3, Ubuntu OS 16.04.2, 256 GB RAM.

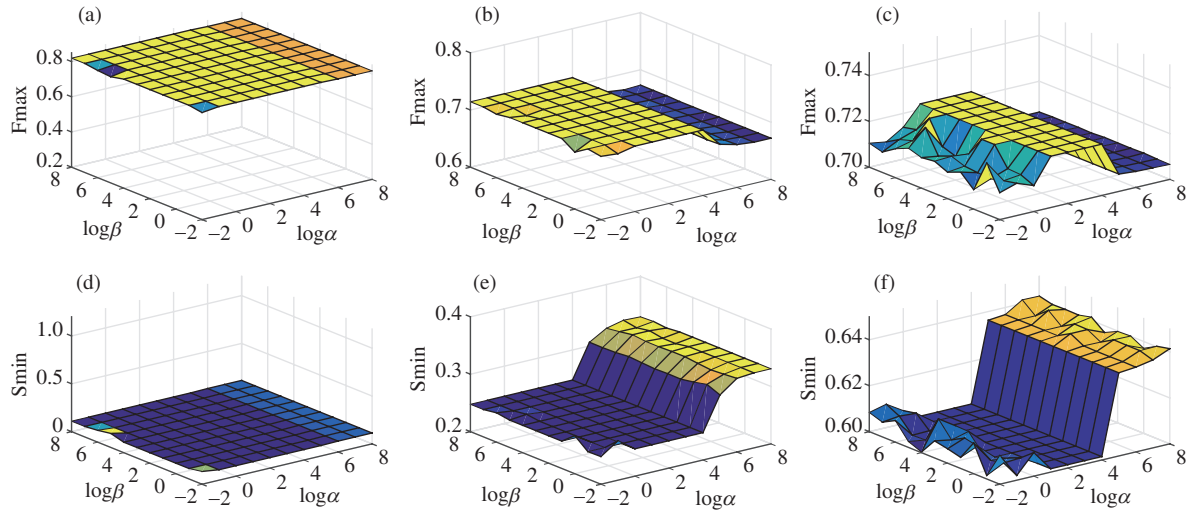

图 3 (网络版彩图) 权重参数 α 和 β 的敏感性分析

Figure 3 (Color online) Sensitivity analysis of weight parameter α and β. (a) Yeast CC (Fmax); (b) Arabidopsis CC (Fmax); (c) Mouse CC (Fmax); (d) Yeast CC (Smin); (e) Arabidopsis CC (Smin); (f) Mouse CC (Smin)

表 6 5 种对比算法在不同数据集上的运行时间统计

Table 6 Statistics of runtime cost of five comparing methods on different datasets

Species	Branch	MV	ClusDCA	NewGOA	HPhash	ZOMF
Yeast	BP	0.71	91.83	605.14	2548.95	88.56
	MF	1.28	67.61	224.92	267.09	34.86
	CC	0.92	64.04	89.00	131.76	40.23
Arabidopsis	BP	0.59	54.47	537.48	1909.87	46.78
	MF	0.33	32.27	207.91	163.81	26.52
	CC	0.21	22.35	83.88	58.36	12.13
Mouse	BP	1.76	208.15	725.90	55146.48	228.33
	MF	1.18	85.92	266.70	690.48	51.41
	CC	1.32	85.18	114.63	300.29	36.31
Human	BP	9.17	540.57	1292.68	64863.61	968.58
	MF	10.31	526.34	411.06	1308.74	470.72
	CC	10.26	591.77	183.00	257.22	310.38
Total		38.04	2370.50	4742.30	127646.66	2314.81

从表 6 中的运行时间统计数据可以看出, 除基准方法 MV 外, ZOMF 的运行时间最小. 原因是 MV 直接利用蛋白质互作网基于近邻投票的方式进行蛋白质功能预测, 没有利用基因本体结构信息, 更不涉及迭代优化问题. 此外, 可以观察到 ZOMF 在 Human 和 Mouse 的 BP 分支上比 ClusDCA 慢, 原因是 ZOMF 分解获取的低秩矩阵为整数矩阵, 且矩阵求解过程中使用了梯度下降算法进行迭代优化, 受蛋白质数量和功能标签数量的影响较大. NewGOA 需要在蛋白质互作网和大量功能标签组成的有向无环图上分别进行随机游走, 所以其时间消耗大于 ZOMF. HPhash 首先基于基因本体层次结构利用一系列偏序保持哈希函数将各个功能标签进行二进制编码, 之后在低维哈希空间中进行蛋白质功能预测, 但是其偏序保持哈希函数的优化需要耗费大量的时间, 所以其运行时间最大. 综合上述实验结

果可以发现 ZOMF 不仅比现有基于标签压缩的蛋白质功能方法预测结果更好, 还能保持较高的效率.

5 结束语

针对蛋白质功能标签巨大, 现有基于标签压缩的蛋白质功能预测方法的可解释性差和阈值划分困难等问题, 本文提出了一种基于 0-1 矩阵分解的蛋白质功能预测方法. 该方法利用蛋白质互作数据和基因本体信息共同指导蛋白质 – 功能标签关联矩阵上的低秩 0-1 矩阵分解, 再利用低秩矩阵重构非负整数型蛋白质 – 功能标签关联矩阵, 在实现蛋白质功能预测的同时避免阈值划分问题. 在多个模式物种上的实验结果表明本文提出的方法获得了较其他相关算法更好的预测结果, 验证了其合理性和有效性. 此外, 该方法运行效率更高且参数鲁棒. 后续研究将探索鲁棒的离散矩阵分解方法和优化策略, 进一步提高蛋白质功能预测精度.

参考文献

1 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

2 Vazquez A, Flammini A, Maritan A, et al. Global protein function prediction from protein-protein interaction networks. Nat Biotechnol, 2003, 21: 697–700

3 Shehu A, Barbara D, Molloy K. A survey of computational methods for protein function prediction. In: Big Data Analytics in Genomics. Berlin: Springer, 2016. 225–298

4 Berardini T Z, Khodiyar V K, Lovering R C, et al. The Gene Ontology in 2010: extensions and refinements. Nucleic Acids Res, 2010, 38: 331–335

5 Schnoes 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

6 Legrain P, Aebersold R, Archakov A, et al. The human proteome project: current state and future direction. Mol Cellular Proteomics, 2011, 10: M111.009993

7 Valentini G. True path rule hierarchical ensembles for genome-wide gene function prediction. IEEE/ACM Trans Comput Biol Bioinf, 2011, 8: 832–847

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

9 Wu J S, Huang S J, Zhou Z H. Genome-wide protein function prediction through multi-instance multi-label learning. IEEE/ACM Trans Comput Biol Bioinf, 2014, 11: 891–902

10 Wang H, Huang H, Ding C. Function-function correlated multi-label protein function prediction over interaction networks. In: Proceedings of International Conference on Research in Computational Molecular Biology, 2012. 302–313

11 Yu G X, Domeniconi C, Rangwala H, et al. Transductive multi-label ensemble classification for protein function prediction. In: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2012. 1077–1085

12 Schwikowski B, Uetz P, Fields S. A network of protein-protein interactions in yeast. Nat Biotechnol, 2000, 18: 1257–1261

13 Yu G X, Fu G Y, Wang J, et al. NewGOA: predicting new GO annotations of proteins by bi-random walks on a hybrid graph. IEEE/ACM Trans Comput Biol Bioinf, 2018, 15: 1390–1402

14 Wang S, Cho H, Zhai C X, et al. Exploiting ontology graph for predicting sparsely annotated gene function. Bioinformatics, 2015, 31: 357–364

15 Yu G X, Zhao Y W, Lu C, et al. HashGO: hashing gene ontology for protein function prediction. Comput Biol Chem, 2017, 71: 264–273

16 Zhao Y W, Fu G Y, Wang J, et al. Gene function prediction based on Gene Ontology Hierarchy Preserving Hashing. Genomics, 2018. doi.org/10.1016/j.ygeno.2018.02.008
17 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预测. 中国科学: 信息科学, 2017, 47: 1349–1368]
18 Tao Y, Sam L, Li J R, et al. Information theory applied to the sparse gene ontology annotation network to predict novel gene function. Bioinformatics, 2007, 23: 529–538
19 Done B, Khatri P, Done A, et al. Predicting novel human gene ontology annotations using semantic analysis. IEEE/ACM Trans Comput Biol Bioinf, 2010, 7: 91–99
20 Yu G X, Zhu H L, Domeniconi C. Predicting protein functions using incomplete hierarchical labels. BMC Bioinf, 2015, 16: 1
21 Pillai I, Fumera G, Roli F. Threshold optimisation for multi-label classifiers. Pattern Recogn, 2013, 46: 2055–2065
22 Lu H L, Vaidya J, Atluri V. Optimal boolean matrix decomposition: application to role engineering. In: Proceedings of IEEE International Conference on Data Engineering, 2008. 297–306
23 Miettinen P, Vreeken J. Model order selection for boolean matrix factorization. In: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data mining, 2011. 51–59
24 Miettinen P, Mielikainen T, Gionis A, et al. The discrete basis problem. IEEE Trans Knowl Data Eng, 2008, 20: 1348–1362
25 Karaev S, Miettinen P, Vreeken J. Getting to know the unknown unknowns: destructive-noise resistant boolean matrix factorization. In: Proceedings of SIAM International Conference on Data Mining, 2015. 325–333
26 Zhang Z, Li T, Ding C, et al. Binary matrix factorization with applications. In: Proceedings of IEEE International Conference on Data Mining, 2007. 391–400
27 Mikhail B, Niyogi P, Sindhwani V. Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. J Mach Learn Res, 2006, 7: 2399–2434
28 Fu G Y, Yu G X, Wang J, et al. Novel protein-function prediction using a directed 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]
29 Cai D, He X F, Han J W, et al. Graph regularized nonnegative matrix factorization for data representation. IEEE Trans Pattern Anal Mach Intell, 2011, 33: 1548–1560
30 Zhang M L, Zhou Z H. A review on multi-label learning algorithms. IEEE Trans Knowl Data Eng, 2014, 26: 1819–1837
31 Wilcoxon F. Individual comparisons by ranking methods. Biometrics Bull, 1945, 1: 80–83
32 Demsar J. Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res, 2006, 7: 1–30

Protein function prediction based on zero-one matrix factorization

Yingwen ZHAO[1], Jun WANG[1], Maozu GUO[2,3], Zili ZHANG[1] & Guoxian YU[1*]

1. *College of Computer and Information Sciences, Southwest University, Chongqing* 400715*, China*;
2. *College of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing* 100044*, China*;
3. *Beijing Key Laboratory of Intelligent Processing for Building Big Data, Beijing University of Civil Engineering and Architecture, Beijing* 100044*, China*
* Corresponding author. E-mail: gxyu@swu.edu.cn

Abstract Accurately annotating the functions of proteins is one of the key tasks of functional genomics. A large portion of functional annotations of proteins is missing, and the functional label space is expansive. Moreover, label compression methods have been proposed and applied to predict protein function; however, such methods lack the interpretability of compressed labels and suffer from the inherent problem of thresholding labels in multi-label learning. To solve these problems, this paper proposes a protein function prediction method based on zero-one matrix factorization (ZOMF). ZOMF first factorizes the protein-function association matrix into two low-rank zero-one matrices and explores the inner latent relationship between proteins and labels. Subsequently, it defines two smoothness terms on these two low-rank matrices with respect to protein–protein interactions and the structural relationships between labels to guide the optimization of low-rank matrices. Finally, to predict protein function, it reconstructs the association matrix using the optimized two low-rank matrices. Experimental results on four model species (yeast, Arabidopsis, mouse, and human) show that ZOMF can predict protein functions more accurately than existing algorithms. However, it does not need to threshold the reconstructed matrix, and the compressed zero-one labels have more than one intuitive explanation.

Keywords protein function prediction, matrix factorization, protein-protein interaction network, gene ontology, thresholding segmentation

Yingwen ZHAO was born in 1994. She received her B.S. degree in Internet of Things Engineering from Jiangsu University, Jingsu, in 2016. Currently, she is a master's student at 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 Harbin Institute of Technology, Harbin, China, in 2010. Currently, she is an Associate Professor at College of Computer and Information Science, Southwest University, Chongqing. Her research interests include machine learning, data mining, as well as their applications in bioinformatics.

Maozu GUO was born in 1966. He received his Ph.D. degree in Computer Science from Harbin Institute of Technology, Harbin, China, in 1997. He is a Professor at Beijing University of Civil Engineering and Architecture, Beijing. His research interests include bioinformatics, machine learning, and data mining.

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