中国科学: 信息科学 2017 年 第 47 卷 第 10 期: 1349–1368
SCIENTIA SINICA Informationis
论文

基于降维的蛋白质不相关功能预测

余国先[1*], 傅广垣[1], 王峻[1], 郭茂祖[2]

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
* 通信作者. E-mail: gxyu@swu.edu.cn

收稿日期: 2017–01–10; 接受日期: 2017–03–16; 网络出版日期: 2017–08–31
国家自然科学基金 (批准号: 61402378, 61571163, 61532014, 61671189)、重庆市研究生科研创新项目 (批准号: CYS16070)、重庆市基础与前沿研究计划项目 (批准号: cstc2014jcyjA40031, cstc2016jcyjA0351) 和中央高校基本科研业务费 (批准号: 2362015X-K07, XDJK2016B009, XDJK2017D061) 资助项目

摘要 蛋白质是生命活动的重要物质基础, 对其功能的准确标注可以极大地促进生命科学的研究与发展. 已有的蛋白质功能预测方法通常仅关注利用蛋白质具有某些功能的信息 (正样例), 并没有关注利用蛋白质不相关的功能信息 (负样例). 已有研究表明, 结合蛋白质负样例可以降低蛋白质功能预测的复杂度并提高预测精度. 本文提出一种基于降维的蛋白质不相关功能预测方法 (predicting irrelevant functions of proteins based on dimensionality reduction, IFDR). IFDR 通过在蛋白质互作网邻接矩阵和蛋白质 – 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 2 个矩阵投影降维为低维实数矩阵, 最后利用半监督回归预测负样例. 在酵母菌、人类和拟南芥的蛋白质数据集上的实验表明, IFDR 比已有相关算法能够更准确地预测负样例, 对互作网络和功能标记空间的降维均可以提高负样例预测精度.

关键词 蛋白质功能预测, 正负样例, 蛋白质互作网, 功能标记, 降维

1 引言

蛋白质是最主要的生命活动载体和功能执行者, 对蛋白质功能的准确标注可以帮助人类更好的理解生命机理, 在药物研发、疾病分析等方面都有着很多应用. 随着高通量测序技术和分析技术的广泛应用, 收集获取的基因/蛋白质序列和网络数据日益增多, 基于生物湿实验测定蛋白质功能的方法通量低、成本高, 难以满足对海量蛋白质数据进行快速功能标注的要求. 如何有效地利用海量数据进行蛋白质功能预测是后基因组时代生物信息学的核心问题之一 [1~3]. 计算学方法能够利用各种生物数据并融合生物学规律实现大规模蛋白质功能预测, 为后续湿实验提供较高置信度的蛋白质功能信息, 减少实验规模, 节约实验成本和时间. 蛋白质功能标注数据库 [4] 中约 95% 的蛋白质功能信息由计算学

引用格式: 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, doi: 10.1360/N112017-00009
Yu G X, Fu G Y, Wang J, et al. Predicting irrelevant functions of proteins based on dimensionality reduction (in Chinese). Sci Sin Inform, 2017, 47: 1349–1368, doi: 10.1360/N112017-00009

方法获得 [5].

研究发现蛋白质的氨基酸序列决定其结构, 进而决定其生物功能, 因此很多方法基于氨基酸序列和结构预测蛋白质功能 [6,7]. 蛋白质之间通过互作完成具体的生物功能, 互作的蛋白质通常共享一些相同的功能, 构成蛋白质互作网, 因而大量基于互作网的分类方法被应用到功能预测中 [8~10]. 单种蛋白质数据 (如氨基酸序列和蛋白质互作网) 描述蛋白质功能特性的能力有限, 一些学者尝试整合多种异构生物数据更全面地描述蛋白质, 提出多种基于数据融合的蛋白质功能预测方法 [11~14].

基因本体 (gene ontology, GO) [4] 是一种广泛使用的基因及其产物 (包括蛋白质和 RNA) 的功能标记范式, GO 利用有向无环图描述功能标记间层次结构关系, 图中每个节点描述并对应一种功能标记, 节点间有向边描述功能之间的关系 (is a part of 和 regulate), 子节点是父节点功能的进一步细化. GO 中存在称为 True Path Rule 的规则: 当一个蛋白质标注某个节点对应的功能时, 它也标注该节点的所有祖先节点对应的功能; 而当明确蛋白质不具有该节点对应的功能时, 它也不具有该节点的所有子节点对应的功能 [4,15]. 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, 可标注多个功能标记, 因此蛋白质可以看作多标记样本 [16~18]. 早期的蛋白质功能预测方法 [9,10] 通常把功能预测问题转化为二分类问题, 单独对每个功能标记进行预测分析, 这类方法忽略了标记之间的关联性, 取得的精度有限. 一些学者将功能预测问题转化为多标记分类问题进行研究 [16~19], 通过利用标记间的关系提高了功能预测的精度. 然而这些方法仅利用了标记间的水平关系, 并未考虑标记间层次结构关系 [20]. 近期一些学者结合标记间层次结构关系进行功能预测提高了预测精度 [12,15,20~22].

虽然多标记学习方法已被广泛用于蛋白质功能预测, 精度也在不断地提升, 但其中的假阴性问题仍未很好地解决, 原因是蛋白质功能标注数据库通常仅登记蛋白质具有某个功能的信息 (正样例), 极少登记与该蛋白质不相关的功能信息 (亦即蛋白质功能的负样例). 为保持与正样例的对称和便于行文, 下文简称蛋白质的不相关功能为蛋白质负样例. 数据库中并未登记的蛋白质与标记的关联性并不代表该标记为蛋白质的负样例, 仅表明该标记与蛋白质是否正关联尚需生物实验验证或目前缺乏相关证据 [5,23,24]. 由于已知的蛋白质功能信息并不完整, 这种未登记的关联性占据非常大的比例, 而现有很多蛋白质功能预测算法均把这些未登记的关联性假定为负样例 [17,18,25], 损失了预测精度.

从 GO 的 True Path Rule 可知, 若已知某个标记为蛋白质的负样例, 则该标记节点的所有子孙节点对应的标记也为该蛋白质的负样例. 研究发现, 通过蛋白质已知的功能信息, 预先选出一部分负样例, 可以显著地缩小功能预测问题的规模, 提高预测精度 [24,26,27]. Mostafavi 等 [11] 利用一种启发式方法选择负样例, 若存在兄弟关系的成对标记中仅已知一个标记为某个蛋白质正样例, 则另一个标记选为该蛋白质的负样例. 由于蛋白质的已知功能信息很不完善且受生物学家研究兴趣的影响 [5,23], 这种启发式方法通常容易错误地选择负样例 [26]. Youngs 等 [24] 提出一种参数化 Bayes 方法 (ALBNeg) 计算成对标记间的经验条件概率, 再结合每个蛋白质已标注的功能标记和上述条件概率预估其他标记也标注到该蛋白质的概率, 选择概率值最小的标记为该蛋白质的负样例. ALBNeg 在计算标记间条件概率时仅考虑了数据库中登记的正样例, 并没有考虑标记间的结构关系, 计算的条件概率存在偏差. 为此, Youngs 等 [26] 通过 GO 上的 True Path Rule 规则将一个蛋白质已标注的标记节点的祖先节点对应的标记也标注到该蛋白质, 再重新计算标记间经验条件概率, 提出一种负样例预测方法 SNOB, SNOB 利用与 ALBNeg 类似的方法预测负样例. 此外, 他们将每个蛋白质视为一个文档, 蛋白质已标注的标记为该文档的单词, 再利用 Lantent Dirichlet Allocation [28] 预测负样例. 上述方法预估的标记间条件概率完全依赖于不完整的蛋白质功能信息. 针对上述不足, Fu 等 [29] 提出 NegGOA 方法综合利用标记间的层次结构相似度和经验条件概率, 再结合蛋白质已知的功能标记在基因本体所在的有向无环图上进行重启动随机游走 [30] 预测蛋白质的负样例. 实验表明 NegGOA 选取的负样例的假阴性数更少,

选择的负样例显著地提升了功能预测精度. 上述方法仅关注对标记结构和已有正样例的利用, 忽略了对已有少量负样例和蛋白质其他特征信息的利用. 傅广垣等 [31] 提出一种基于正负样例的蛋白质功能预测方法 (ProPN), 该方法通过符号混合图描述蛋白质与标记的正关联 (正样例) 和负关联 (负样例)、蛋白质互作和标记间关联关系, 再通过混合图上的符号标记传播预测蛋白质负样例. 蛋白质互作网中存在一定量的假阳性互作, 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非常少, 在标记传播中这些稀疏标记容易被其他标记覆盖, 也会降低负样例预测的准确性.

在分析上述研究工作的基础上, 本文借助成分扩散分析 [32~34] 和奇异值分解 (single value decomposition, SVD) [35] 提出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 (predicting irrelevant functions of proteins using dimensionality reduction, IFDR). IFDR 首先在蛋白质互作网对应的邻接矩阵上进行重启动随机游走挖掘蛋白质潜在的互作关系, 再利用 SVD 获得互作网的低维实数特征向量矩阵, 矩阵中每行刻画对应蛋白质在互作网中的主要拓扑结构信息. 同时, IFDR 基于蛋白质已知的功能和基因本体结构利用重启动随机游走预估蛋白质的缺失正样例, 再利用 SVD 将蛋白质 – 功能标记关联矩阵投影转化为一个低维关联矩阵. 最后在降维后的蛋白质互作网和蛋白质 – 功能标记关联矩阵上利用半监督回归预测蛋白质负样例. 在酵母菌、人类和拟南芥 3 个模式生物上的负样例预测实验表明, IFDR 能够较已有相关方法更准确预测负样例, 通过 SVD 对互作网和标记空间进行降维可以挖掘并利用蛋白质互作关系和标记间关联关系, 提升负样例预测精度.

2 基于网络和标记空间降维的蛋白质功能负样例预测

IFDR 主要由蛋白质互作网上的成分扩散分析及基于 SVD 的降维压缩表示, 蛋白质 – 功能标记关联矩阵中的缺失正样例预估及基于 SVD 的降维压缩表示, 基于压缩的蛋白质特征向量数据和蛋白质 – 功能标记向量数据的半监督线性回归 3 部分构成, 下文分别对上述内容进行详细分析介绍.

2.1 蛋白质互作网的降维表示

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, 蛋白质互作网是其中一种最常见和常用的蛋白质数据集, 互作网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信息. 互作网可以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中, 这些互作的蛋白质更有可能具有相同的功能. 很多学者对蛋白质互作网的生物学特性进行建模分析, 设计方法预测蛋白质功能 [8~10,22,31,32,36]. 然而高通量技术收集的蛋白质网络数据包含的互作信息并不全面, 还存在一定的噪声互作, 这些假阳性互作和缺失的 (假阴性) 互作会降低功能预测的精度.

近期 Cao 等 [32] 利用成分扩散分析刻画互作网中每个蛋白质细粒度拓扑结构并加以利用, 提升了预测精度. 但这类方法依然受假阳性和假阴性互作的影响. 为此, Cho 等 [33] 和 Wang 等 [34] 将降维与成分扩散分析进行结合, 首先在蛋白质互作网的邻接矩阵上进行基于重启动随机游走的成分扩散分析并更新邻接矩阵, 获得每个蛋白质节点的分布信息及其与其他蛋白质的关联信息, 再利用一种基于 Kullback [37] 的 Logistic 回归模型对邻接矩阵进行降维, 获得每个蛋白质的低维特征向量, 通过低维特征向量刻画该蛋白质与其他蛋白质之间的拓扑结构. 实验表明, 与在原始互作网上进行基于成分扩散分析的功能预测方法相比, 结合降维可以进一步提高预测精度.

为了克服假阳性和假阴性互作对蛋白质负样例预测的影响, 受上述工作启发, IFDR 在蛋白质互

作网的邻接矩阵上进行成分扩散分析, 再利用 SVD 将更新的邻接矩阵通过降维转化为低维向量形式. 令 $G_P \in \mathbb{R}^{n\times n}$ 为由 n 个蛋白质构成的互作网对应邻接矩阵, 当蛋白质 i 与 j 存在相互作用时, 则 $G_P(i,j)$ 存在一条权重不为 0 的边, 其边权重表示这两个蛋白质互作的强度或置信度. 为方便计算, 本文对 G_P 进行了归一化处理以保证一个蛋白质与其他蛋白质互作的强度之和为 1, 具体归一化方式如下:

$$G_{PP} = D_P^{-1} G_P, \tag{1}$$

D_P 是一个对角矩阵, $D_P(i,i) = \sum_{j=1}^{n} G_P(i,j)$. 令 $S_P^0 = G_{PP}$, $S_P^t \in \mathbb{R}^{n\times n}$ 表示第 t 步重启动随机游走后蛋白质互作网的扩散矩阵:

$$S_P^{t+1}(i,v) = (1-\gamma)\sum_{j=1}^{n} G_{PP}(i,j) S_P^t(j,v) + \gamma G_{PP}(i,v), \tag{2}$$

其中 $\gamma \in (0,1)$ 控制随机游走的重启动概率, 它调节随机游走对互作网全局和局部拓扑结构的影响, γ 越大表示局部结构权重越大. 由于 $0 \leqslant G_{PP}(i,j) \leqslant 1$, 式 (2) 必收敛. 令 S_P 表示重启动随机游走达到收敛时的平稳分布, 亦即最终的扩散状态. 互作网中两个蛋白质拥有相似的扩散状态, 意味着它们在互作网中相对其他节点有相似的位置, 表明它们的功能很可能相似 [32]. 但由于探知的蛋白质互作信息受限于高通量实验技术和生命分子活动的随机性 [3], 上述方法获得的最终扩散状态仍在一定程度上受噪声互作的影响. 此外, 若将 S_P 的每行当作一个蛋白质的特征向量, 再在其上训练分类器预测蛋白质功能, 则会面临高维数据上的巨大计算开销和维数灾难问题.

为克服上述问题, IFDR 在 S_P 上应用 SVD 将它降维压缩为实数向量特征矩阵, 具体方式如下:

$$S_P = U_P \Lambda_P V_P^{\mathrm{T}}, \tag{3}$$

其中 $U_P \in \mathbb{R}^{n\times n}$ 和 $V_P \in \mathbb{R}^{n\times n}$ 也称为 S_P 的左右特征向量矩阵, $\Lambda_P \in \mathbb{R}^{n\times n}$ 为对角奇异值矩阵, 对角线中的每个元素均不小于 0. U_P 中的每列由 $S_P {S_P}^{\mathrm{T}}$ 的特征向量构成, V_P 中的每列由 ${S_P}^{\mathrm{T}} S_P$ 的特征向量构成, 由于 S_P 为对称矩阵, 因此 $U_P = V_P$.

IFDR 选取 U_P 中前 d 个列向量构成矩阵 $U_P^d \in \mathbb{R}^{n\times d}$ 和 Λ_P 中前 d ($d \ll n$) 个最大对角元素构成对角矩阵 $\Lambda_P^d \in \mathbb{R}^{d\times d}$ 实现对 S_P 的低维向量表示. 令 $X_P = [x_1, \ldots, x_n] \in \mathbb{R}^{n\times d}$ 为 n 个蛋白质对应的低维向量特征矩阵, X_P 的计算方式如下:

$$X_P = U_P^d (\Lambda_P^d)^{\frac{1}{2}}. \tag{4}$$

通过将高维 S_P 压缩为 X_P 不仅避免了维数灾难问题, 也能在保持网络扩散模式的同时降低假阳性和假阴性互作的影响, 后续实验分析也将证明基于 SVD 对蛋白质互作网进行降维可提高负样例的预测精度.

2.2 蛋白质 – 功能标记关联矩阵的降维表示

蛋白质的功能标记非常不平衡 [15], 大量稀疏标记标注的蛋白质个数非常少, 只有少量标记标注的蛋白质个数较多. 例如在蛋白质功能标注数据库中 (截止 2016–08–31), 已知人类的约 20000 个蛋白质与 24429 个功能标记存在关联, 而这些标记中只有 9216 个标注的蛋白质数量大于 3, 有 5976 个标记标注的蛋白质数量超过 10, 仅有 1513 个标记标注的蛋白质数量大于 100, 其他 (约 62%) 的标记标注的蛋白质数量均小于 3. 这些稀疏标记的正样例非常少以致训练的分类器精度不高, 容易出现过

拟合问题. Yu 等 [20] 观察发现蛋白质新增的功能标记通常对应该蛋白质已有功能标记节点的子孙节点, 这些子孙节点对应的标记标注的蛋白质个数通常小于其祖先节点. 在负样例预测中, 若仅基于功能标记的频率信息预测负样例, 则很容易误判稀疏标记为蛋白质的负样例 [29]. 从 GO 结构和其上的 True Path Rule 规则可知, 这些稀疏标记通常描述了更详细的生物学功能. Pandey 等 [38] 发现利用标记间的关联关系可以提升功能预测算法在稀疏标记上的精度. 基于多标记学习的蛋白质功能预测方法通过不同的策略利用标记间的关联关系也提升了精度 [17,20,25,27,39]. 然而上述这些方法并没有充分考虑标记不平衡特性. 针对这一问题, Wang 等 [34] 提出 clusDCA 方法. clusDCA 在 GO 有向无环图对应的邻接矩阵上进行成分扩算分析, 利用 SVD 对邻接矩阵进行降维压缩, 再将蛋白质 – 功能标记关联矩阵 $G_F \in \mathbb{R}^{n\times m}(m$ 为标记个数) 投影到低维空间 $Y_F^c \in \mathbb{R}^{n\times c}(c \ll m)$, 最终通过优化一个矩阵 $R \in \mathbb{R}^{d\times n}(X_P R Y_F^c)$ 预测 n 蛋白质与 c 个压缩的标记之间的关联性. 然而 clusDCA 并没有考虑蛋白质功能信息的不完整性和蛋白质新增功能标记的模式规律.

与 clusDCA 相比, IFDR 不对 GO 对应的邻接矩阵进行类似蛋白质互作网的成分扩散分析和降维. IFDR 结合标记间的层次结构关系和蛋白质已有的功能信息, 在 GO 的有向无环图上进行有向的重启动随机游走, 对蛋白质的缺失正样例进行建模, 进而减少误判稀疏功能标记为蛋白质负样例的风险, 再利用 SVD 对噪声特征鲁棒的特点对蛋白质 – 功能标记关联矩阵 G_F 进行降维压缩表示, 降低错误预估正样例的破坏作用.

蛋白质 – 功能标记关联矩阵 G_F 通过如下方式进行初始化: 当标记 s 为蛋白质 i 的正样例时 $G_F(i,s) = 1$; 当标记 s 为蛋白质 i 的负样例时 $G_F(i,s) = -1$; 当它们之间的正负关联性未知时 $G_F(i,s) = 0$. Yu 等 [20] 统计发现, 对于蛋白质的缺失正样例, 来自直接父节点标记预估的置信度远大于其他祖先节点标记. 这是因为当已知该蛋白质标注了父节点标记时, 由 True Path Rule 可知该蛋白质也标注它的其他祖先节点标记, 反之则不一定成立. 基于上述发现, 本文初始仅考虑具有直接父子关系的标记节点间的信息传递, 并通过以下方式设置标记间的转移概率:

$$p(v|s) = \begin{cases} n_v/n_s, & n_v > 0, \\ 1/|\mathrm{ch}(s)|, & n_v = 0, \end{cases} \tag{5}$$

其中 n_s 和 n_v 为分别标注 s 和 v 的蛋白质数量, $\mathrm{ch}(s)$ 为 s 的直系孩子节点集合, $v \in \mathrm{ch}(s)$, $|\mathrm{ch}(s)|$ 为该集合的大小. 由 True Path Rule 可知 $n_s \geqslant n_v$. 当 $n_v = 0$ 时, 仅表明 n 个蛋白质中目前还没有蛋白质标注功能 v, 原因可能是缺少相关的实验验证, 针对这种情况, 本文设定 s 向 v 的转移概率为 s 的孩子节点数的倒数. 为使 s 向其所有孩子节点转移概率总和为 1, 对转移概率进行如下归一化:

$$\tilde{p}(v|s) = \frac{p(v|s)}{\sum_{u\in \mathrm{ch}(s)} p(u|s)}. \tag{6}$$

类似地, 对 G_F 进行如下归一化:

$$G_{FF} = D_F^{-1} G_F, \tag{7}$$

其中 $D_F \in \mathbb{R}^{n\times n}$ 是一个对角矩阵 $D_F(i,i) = \sum_{s=1}^{m} |G_F(i,s)|$, 取 $G_F(i,s)$ 的绝对值是因为 G_F 中存在正负样例.

在上述设置的基础上, IFDR 结合蛋白质已知的功能信息在有向无环图上进行重启动随机游走预估蛋白质的缺失正样例 ($G_F(i,v) = 0$), 具体预估方式如下:

$$S_F^{t+1}(i,v) = (1-\gamma) \sum_{s\in \mathrm{par}(v)}^{n} S_F^{t+1}(i,s)\tilde{p}(v|s) + \gamma G_{FF}(i,v), \tag{8}$$

其中 $\mathrm{par}(v)$ 为 v 的父母节点集合, $S_F^t(i,v)$ 为第 t 次随机游走时预估的 i 与 v 的关联大小, $S_F^0 = G_{FF}$. 由于 $\gamma \in (0,1)$, $\tilde{p}(v|s) \leqslant 1$, $S_F^t(i,v) \leqslant S_F^t(i,s)$, 亦即蛋白质 i 与 s (父节点标记) 的关联大小总是不小于与 v (子节点标记) 的关联大小.

这种预估方法较好地利用了蛋白质新增正样例的模式, 但容易引入较多的假阳性预估. 为了降低假阳性预估的影响, IFDR 利用 SVD 对噪声特征鲁棒的优点, 将 S_F (式 (8) 收敛时 n 个蛋白质与 m 个标记的关联大小矩阵) 分解为

$$S_F = U_F \Lambda_F V_F^{\mathrm{T}}, \tag{9}$$

其中 $U_F \in \mathbb{R}^{n\times n}$, $V_F \in \mathbb{R}^{m\times m}$, $\Lambda_F \in \mathbb{R}^{n\times m}$ 为对角奇异值矩阵. 一种常用的去噪方式是取 U_F 的前 c 列构成矩阵 $U_F^c \in \mathbb{R}^{n\times c}$, Λ_F 的前 c ($c \ll m$) 个最大奇异值构成对角矩阵 $\Lambda_F^c \in \mathbb{R}^{c\times c}$ 和 V_F 的前 c 列构成矩阵 $V_F^c \in \mathbb{R}^{m\times c}$, 通过 $U_F^c \Lambda_F^c V_F^c$ 近似重构矩阵 S_F. 在此基础上, 再利用 X_P 和去噪后的 S_F 进行蛋白质功能预测. 然而这种方法涉及的标记集合依然非常大, 甚至出现蛋白质数量小于标记数量的情况, 训练的分类器面临过拟合的风险.

针对上述问题, IFDR 利用 U_F^c 和 Λ_F^c 将 S_F 投影到低维空间 $Y_F^c = [y_1, y_2, \ldots, y_n]$ ($Y_F^c \in \mathbb{R}^{n\times c}$), 方式如下:

$$Y_F^c = U_F^c (\Lambda_F^c)^{\frac{1}{2}}, \tag{10}$$

IFDR 再基于 X_P 和 Y_F^c 进行蛋白质功能预测. 假定 $\tilde{Y}_F^c \in \mathbb{R}^{n\times c}$ 为 IFDR 预测的 n 个蛋白质与压缩的 c 个功能的关联性矩阵, IFDR 通过 $\tilde{Y}_F^c P_F^m$ ($P_F^m \in \mathbb{R}^{c\times m}$) 将 $\tilde{Y}_F^c$ 映射回原始的 m 个功能标记空间, P_F^m 的定义如下:

$$P_F^m = (\Lambda_F^c)^{\frac{1}{2}} V_F^c. \tag{11}$$

Y_F^c 中的每行可以看作对应蛋白质压缩的 c 维标记向量, P_F^m 中每列可以看作是原始标记的 c 维实数特征表示, 该列编码存储了对应标记与其他标记的关系 [35, 40]. 通过对 S_F 的压缩表示, 蛋白质的正负样例在 c 维标记空间可向它们相似的标记传递.

2.3 蛋白质负样例预测

IFDR 基于蛋白质互作网 G_P 的压缩向量矩阵 X_P 和蛋白质 – 功能标记关联矩阵 G_F 的压缩向量矩阵 Y_F^c 训练分类器进行负样例预测. 不同于一般的 0-1 (或 -1, 1) 形式的标记指示矩阵, Y_F^c 为实数向量矩阵, 且其中包含值为负的元素. 为此, 本文采用半监督线性回归 [41] 预测蛋白质负样例. 基础线性方程如下:

$$f(x) = W^{\mathrm{T}} x + b, \tag{12}$$

其中 $W \in \mathbb{R}^{d\times c}$ 为投影预测向量, $b \in \mathbb{R}^{c\times 1}$ 为偏移向量, $f(x) \in \mathbb{R}^c$ 为 x 在 c 个压缩的功能标记上的输出, 或 x 与 c 个标记的关联性大小.

类似流形正则化半监督分类框架 [42], 本文的半监督线性回归目标方程形式如下:

$$J(W,b) = \underset{W,b}{\operatorname{argmin}} \sum_{i=1}^{n} \Psi(x_i, y_i, f(x_i)) + \alpha \|f\|_I^2 + \beta \|f\|_H^2, \tag{13}$$

其中 $\Psi(x_i, y_i, f(x_i))$ 为预先定义的损失函数, $\|f\|_I^2$ 为基于 n 个蛋白质之间的特征相似度定义的平滑损失项, $\|f\|_H^2$ 为控制 $f(x)$ 复杂度避免其过度拟合的正则项. 在本文中, $\Psi(x_i, y_i, f(x_i))$ 选用平方误差

损失函数:

$$\Psi(x_i, y_i, f(x_i)) = \|f(x_i) - y_i\|_2^2 = \mathrm{tr}((W^{\mathrm{T}}x_i + b - y_i)(W^{\mathrm{T}}x_i + b - y_i)^{\mathrm{T}}), \tag{14}$$

其中 tr() 为求矩阵的迹.

$\|f\|_I^2$ 的定义与计算方式如下:

$$\begin{aligned}\|f\|_I^2 &= \frac{1}{2}\sum_{i,j=1}^{n}\|f(x_i) - f(x_j)\|_2^2 S_{ij} = \frac{1}{2}\sum_{i,j=1}^{n}\left\|W^{\mathrm{T}}x_i - W^{\mathrm{T}}x_j\right\|_2^2 S_{ij}\\ &= \mathrm{tr}\left(W^{\mathrm{T}}\sum_{i=1}^{n}\left(x_i S_{ii} x_i^{\mathrm{T}}\right)W - W^{\mathrm{T}}\sum_{i,j=1}^{n}(x_i S_{ij} x_j^{\mathrm{T}})W\right)\\ &= \mathrm{tr}\left(W^{\mathrm{T}}X_P^{\mathrm{T}}(D-S)X_P W\right) = \mathrm{tr}\left(W^{\mathrm{T}}X_P^{\mathrm{T}}LX_P W\right),\end{aligned} \tag{15}$$

其中 S_{ij} 为蛋白质 x_i 与 x_j 之间的相似度, 本文采用余弦相似性度量计算蛋白质之间的相似度. D 为对角矩阵, $D_{ii} = \sum_{j=1}^{N} S_{ij}$, $L = D - S$ 为图 Laplace 矩阵, 它对称半正定. 最小化 $\|f\|_I^2$ 的目的是使具有相似特征表示的蛋白质标注相似的功能标记集合, 因为这些特征相似的蛋白质在互作网中有相似的成分扩散模式, 通常构成一个功能模块, 协作完成生物学功能 [36, 43]. 控制分类器复杂度的正则项 $\|f\|_H^2$ 定义如下:

$$\|f\|_H^2 = \mathrm{tr}(W^{\mathrm{T}}W). \tag{16}$$

在上述定义的基础上, $J(W, b)$ 可以表示为

$$J(W,b) = \mathrm{tr}\left(\left(W^{\mathrm{T}}X_P^{\mathrm{T}} + b1_n^{\mathrm{T}} - Y_F^{c\mathrm{T}}\right)\left(W^{\mathrm{T}}X_P^{\mathrm{T}} + b1_n^{\mathrm{T}} - Y_F^{c\mathrm{T}}\right)^{\mathrm{T}} + \alpha W^{\mathrm{T}}X_P^{\mathrm{T}}LX_P W + \beta W^{\mathrm{T}}W\right), \tag{17}$$

其中 $1_n \in \mathbb{R}^{n\times 1}$ 为元素值全为 1 的列向量. 分别对 $J(W, b)$ 求 W 与 b 的导数, 并令导数为 0, 可得

$$W = \left(X_P^{\mathrm{T}}UX_P + \alpha X_P^{\mathrm{T}}LX_P + \beta I_d\right)^{-1}X_P^{\mathrm{T}}UY_F^c, \tag{18}$$

$$b = \frac{1}{n}\left(Y_F^c - W^{\mathrm{T}}X_P\right)^{\mathrm{T}}1_n, \tag{19}$$

其中 I_d 为 $d \times d$ 的单位矩阵, $U = I_n - 1/n$, I_n 为 n 维单位矩阵.

令 $\tilde{Y}_F^c = [f(x_1), f(x_2), \ldots, f(x_n)]$, IFDR 通过 P_F^m 将 $\tilde{Y}_F^c$ 映射回原始标记空间:

$$\tilde{Y}_F^m = \tilde{Y}_F^c P_F^m, \tag{20}$$

$\tilde{Y}_F^m \in \mathbb{R}^{n\times m}$ 存储 IFDR 预测的 n 个蛋白质与 m 个功能的关联度大小. $\tilde{Y}_F^m(i, v)$ 值越大表明 v 为蛋白质 i 的正样例的可能性越高, 值越小表明 v 为该蛋白质负样例的概率越高.

3 实验

3.1 数据集

本文从 BioGrid 数据库 [44] (日期: 2016–08–01) 中分别下载 3 个典型模式物种 (人类 (H.sapiens), 酵母菌 (S.cerevisiae), 拟南芥 (A.thaliana)) 的蛋白质互作网数据, 同时还下载了对应的 GO 数据库和蛋白质 – 功能标记关联数据, 并对互作网中的蛋白质进行功能标注. GO 描述了功能标记之间的层次

表 1　实验数据集统计信息, Avg±Std 对应每个蛋白质的平均功能标记个数和对应的方差

Table 1　Dataset statistics, Avg±Std is the average number of annotations per protein and standard deviation

	Proteins (n)	Branch	Functions (m)	Positives (Negatives)	Avg±Std
H. sapiens	16082	BP	15373	790787 (16324)	49.17±63.14
		CC	2931	307635 (26963)	19.13±34.49
		MF	5990	158369 (12042)	9.84±18.55
S. cerevisiae	6017	BP	5256	222754 (1374)	37.02±31.65
		CC	2566	120392 (5456)	20.00±23.85
		MF	2501	47558 (799)	7.90±6.89
A. thaliana	9289	BP	5948	229193 (3132)	24.67±28.01
		CC	2397	179944 (45523)	19.37±31.44
		MF	2553	67695 (1846)	7.29±9.29

结构关系, 这些标记分布在 3 个不相交的分支上, 分别是生物过程 (biological process, BP)、分子功能 (molecular function, MF) 和细胞组分 (cellular component, CC). 与以往实验类似, 本文剔除 GO 中 'obsolete' 的功能标记; 为避免循环预测, 还剔除了证据属性为 IEA (inferred from electronic annotation) 的标注信息. 考虑到数据库仅登记蛋白质的直接功能标记, 这些标记过于稀疏, 本文利用 GO 上的 True Path Rule 对蛋白质功能信息进行增补, 将蛋白质正样例对应标记节点的祖先节点标记也标注到相应的蛋白质上, 蛋白质负样例对应标记节点的子孙节点也标注为相应蛋白质的负样例. 表 1 列出了上述 3 个物种的蛋白质已有标记统计信息, 可以看到 GO 数据库已经开始登记少量的蛋白质负样例, 但相对正样例来说其数量仍然很少. 最后一列 Avg±Std 对应每个蛋白质的平均功能标记个数和对应的方差, 较大的方差说明蛋白质的功能信息并不完整, 一些蛋白质的功能信息较完备, 另一些蛋白质的功能信息存在较多缺失, 还有一部分蛋白质的功能信息完全未知.

3.2　对比方法与评价度量

本文将通过负样例预测的假阴性个数和预测的负样例对蛋白质功能预测精度的提升情况综合检验负样例预测的有效性. 在负样例预测实验中, 本文以当前最新的 NegGOA [29], ProPN [31], SNOB [26] 和基准方法 Random 作为 IFDR 的对比方法. 为分析 IFDR 中缺失标记预估和半监督线性回归的影响, 本文引入 clusDCA [34] 和 IFDR-DCA 作为对比方法. NegGOA, ProPN, SNOB 和 clusDCA 在引言或 2.2 小节中做了详细介绍, 不做赘述. IFDR-DCA 是 IFDR 的变种, 它在蛋白质互作网和 GO 上分别进行成分扩散分析和 SVD, 再将蛋白质 – 功能标记关联矩阵映射到低维空间, 然后利用 IFDR 的半监督线性回归目标方程预测负样例. 基准方法 Random 从所有未标注到某个蛋白质的标记集合中随机选择标记为该蛋白质负样例, 为了减少随机偏差, 本文随机运行基准方法 100 次, 取 100 次运行结果的均值作为其负样例预测的最终结果. 所有对比方法的参数设置都按照原始论文提供的参数或方法进行设置. 如 NegGOA 中, $\alpha = \beta = 0.5$, 迭代次数为 4; ProPN 中 $\alpha = 0.1$, $\beta = 0.9$; clusDCA 中 $bp = 0.8$, 对 GO 降维的目标维度为 2500. IFDR 中蛋白质互作网空间投影维度 d 和标记空间投影维度 c 统一设置为 300, 下文实验将对这 2 个参数的敏感性进行分析.

在负样例预测实验中, 本文选用假阴性预测数 (false negatives, FNs) [26, 29] 为评测指标, 它统计预测结果为负样例但真实结果为正样例的错误情况. 本文下载了上述 3 个物种蛋白质早期的功能标注数据 (归档日期分别为 2015–07–01 和 2014–06–01), 并用 3.1 小节中同样方法和流程对蛋白质互作网中

的蛋白质进行功能标注. 所有方法均基于 2015 年的蛋白质功能标注数据进行蛋白质负样例预测, 再利用 2016 年更新的蛋白质功能标注数据检验负样例的预测质量, 若一个预测的负样例在更新的蛋白质功能标注数据中为正样例, 则产生了一个 FN.

本文另外选用 4 个常用的评价度量 MarcroF1, RAccuracy, AUC 和 Fmax 评估蛋白质功能预测的质量. MarcroF1 是一种以标记为中心的评价度量, 它先求取每个标记的 F1-score, 再取这些标记 F1-score 的均值, 这一评价度量受稀疏功能标记影响较大. RAccuracy 从全局上检验 n 个蛋白质中有多少缺失正样例被准确预测. AUC 首先对每个标记分别计算 ROC (receiver operating characteristic) 曲线下的面积, 再取 m 个标记对应面积的均值. 与 AUC 一样, Fmax 是国际大规模蛋白质功能预测评测组织 [2,3] 推荐的评价度量, 它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, 最后选择最大 F1 值为 Fmax 的值. 上述几个度量的形式化定义可参考文献 [3, 16, 21]. 对比算法在这 4 个度量上的值越高, 表示其预测质量越好. 从这些度量的定义可知它们从不同的角度评估蛋白质功能预测的质量, 一个算法很难在所有度量上均超过另一个算法.

3.3 负样例预测结果分析

本小节主要测试分析各个对比算法预测蛋白质负样例的假阴性数和各算法预测的负样例对蛋白质功能预测精度的提升效果. IFDR 分别对蛋白质互作网和蛋白质功能关联矩阵分别进行降维表示, 再基于半监督线性回归预测负样例, 最终得到 n 个蛋白质与 m 个功能标记的关联预测矩阵 $\tilde{Y}_F^m \in \mathbb{R}^{n\times m}$. IFDR 从 $\tilde{Y}_F^m$ 中选取最小的 l 个元素为预测的负样例, 再与 2016 年更新的蛋白质功能标注数据进行对比, 统计预测的假阴性个数 FNs. 其他对比算法也通过类似流程分别预测负样例再统计各自的 FNs. 表 2 给出了不同算法在人类数据集 (2015~2016) 上的验证结果. 限于篇幅, 这些算法在其他数据集上的实验结果报告在补充材料的表 S1 和 S2 中.

从这些表中的对比结果不难发现, 本文提出的 IFDR 方法在绝大部分实验设置 (不同的 l 和时间段) 下都取得相比其他对比算法更小的 FNs. 以人类数据集 (2015~2016) 中 BP 分支的结果为例, 在选取 80000 ($l = 80k$) 个预测的负样例做检验时 IFDR 无假阴性预测, NegGOA 产生了 2 个假阴性预测, ProPN 和 SNOB 分别产生了 51 个和 24 个假阴性预测, clusDCA 和 IFDR-DCA 分别产生了 189 和 26 个假阴性预测. 通过 Wilcoxon 符号秩检验 [45] 统计 IFDR 和其他对比算法在多个数据集上的负样例预测结果的差异显著性, 发现对应 p 值都小于 0.05. 从上述实验结果分析可知 IFDR 是一种有效的蛋白质不相关功能预测方法.

SNOB 仅基于蛋白质已知的正样例计算标记之间的经验条件概率, 再结合蛋白质已知的功能标记预估其他标记也标注到该蛋白质上的概率, 选择概率值最低的标记为该蛋白质的负样例. 这类经验条件概率在频率较高的浅层次功能标记间较为可靠, 但在频率较低的深层次功能标记间存在较大偏差, 而且这些深层 (稀疏) 标记很有可能是蛋白质的缺失标记 (正样例). 此外, 与 NegGOA 一样, 它忽略对蛋白质其他特征信息的利用. 因此, 它的 FNs 比 IFDR 多. Random 通过随机选取未标注到蛋白质的标记为该蛋白质的负样例, 它有时候获得了较 SNOB 更少的 FNs. 原因是 Random 的随机选择有一定的结构性, 稀疏标记通常对应 GO 中深层次节点, 它们被随机选择到的概率较大, 更有可能被预测为负样例, 而这些负样例在更新的蛋白质功能标注数据中较难被验证. 但由于稀疏标记也很可能为蛋白质的缺失正样例, 因而 Random 通常获得较其他对比算法更高的 FNs.

NegGOA 首先计算标记间的经验条件概率和基于本体结构的标记间条件概率, 再结合蛋白质已有的功能标记和标记间条件概率利用重启动随机游预测蛋白质负样例, 它通常获得比 SNOB 和 Random 更小的 FNs. 但 NegGOA 仅利用了基因本体结构和已有的功能标注信息, 并没有利用蛋白质的其

表 2　人类数据集上不同负样例预测数下的假阴性个数 (2015∼2016)

Table 2 FNs of H. sapiens under different numbers of predicted negative examples. Negative examples are predicted by available annotations in 2015, and validated by updated annotations in 2016

	Data set	l							
		10k	20k	30k	40k	50k	60k	70k	80k
BP	IFDR	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
	IFDR-DCA	5	23	23	24	24	26	26	26
	clusDCA	44	75	99	121	139	157	174	189
	ProPN	3	24	33	35	43	44	51	51
	NegGOA	1	2	2	2	2	2	2	2
	SNOB	4	4	12	17	17	18	20	24
	Random	3.94	7.93	12.53	16.42	20.02	24.95	29.56	33.14
CC	IFDR	**0**	**0**	**1**	**1**	**2**	**2**	**2**	**2**
	IFDR-DCA	1	2	2	2	3	3	3	3
	clusDCA	81	159	227	276	332	373	423	455
	ProPN	1	2	5	6	14	15	15	19
	NegGOA	**0**	**0**	**0**	3	4	4	4	6
	SNOB	18	18	18	18	18	18	22	22
	Random	5.91	12.02	17.21	23.82	30.01	35.11	40.86	47.23
MF	IFDR	0	3	4	5	6	8	8	8
	IFDR-DCA	1	1	2	2	2	4	5	**6**
	clusDCA	12	15	22	29	35	40	44	51
	ProPN	11	46	53	69	70	74	76	76
	NegGOA	**0**	**0**	**0**	**0**	**0**	**0**	**2**	8
	SNOB	38	38	38	38	38	38	39	41
	Random	1.82	4.27	6.48	8.35	10.98	12.45	13.65	17.04

他特征信息 (如蛋白质互作网和氨基酸序列等), 同时它也没有对已知的少量负样例加以应用, 所以 NegGOA 通常获得较 IFDR 大的 FNs, 部分情况下也获得比 ProPN 大的 FNs. ProPN 根据已知的蛋白质互作信息和蛋白质正负样例信息构建一个符号混合图, 在该混合图上进行标记传播预测蛋白质负样例. ProPN 在拟南芥数据集上的 FNs 比 NegGOA, SNOB 和 Random 低, 但它在人类和酵母菌数据集上的 FNs 比其他对比算法高, 有时候比基准方法 Random 还高. 原因是 ProPN 在进行符号混合图上的标记传播时容易过度传播蛋白质的正负样例信息, 降低了负样例预测的精度. 另一个原因是 ProPN 并没有考虑标记间的层次结构关系, 容易误判蛋白质的缺失正样例为该蛋白质的负样例.

IFDR 和 clusDCA 均在蛋白质互作网上进行了成分扩散分析与降维, 不同的是 clusDCA 在 GO 对应的有向无环图上进行了无向的成分扩散分析和降维, 再求解两种低维向量之间的关联映射矩阵 R 进行负样例预测. IFDR 在蛋白质 – 功能标记关联矩阵上进行缺失标记预估再对该矩阵进行降维. clusDCA 的 FNs 总是比 IFDR 多, 也通常比其他对比方法多. 原因是它求取的关联映射矩阵并不一定适合负样例预测问题; 另一个原因是它假定未登记的蛋白质与标记间关联为蛋白质的负样例, 这种假定误导了后续的负样例预测. IFDR-DCA 与 clusDCA 一样在蛋白质互作网和 GO 上进行成分扩散分析再降维, 然后将高维蛋白质 – 功能标记关联矩阵降维到低维向量空间, 再利用半监督线性回归预测

蛋白质负样例. IFDR-DCA 的 FNs 远小于 clusDCA, 这表明本文选用的半监督线性回归可以有效地预测蛋白质负样例, 也进一步证实了结合已知的负样例进行负样例预测的有效性. IFDR-DCA 在大部分情况下的 FNs 比 IFDR 多, 主要原因是 IFDR 对蛋白质潜在的缺失功能标记进行了预估, 降低了判定稀疏标记为负样例的概率, 进而提高了负样例预测精度.

本文还基于 2014 年 6 月的蛋白质功能标注数据, 采用上述类似的方法预处理和设置后预测蛋白质负样例, 再用 2015 年 7 月更新的蛋白质功能标注数据检验负样例预测性能, 对应实验结果汇报在补充材料表 S3∼S5 中. 从这些表中的结果可以看到 IFDR 在大部分对比实验中均获得较其他对比算法更小的 FNs, 这些表中的结果和结论与时间段 (2015∼2016) 类似. 这些实验结果进一步证明了 IFDR 在负样例预测中的有效性.

本文还将 IFDR 拓展应用到上述 3 个物种的蛋白质序列数据, 并对比分析 IFDR 分别在蛋白质互作网, 序列数据构造的网络, 及其与互作网组成的混合网络上的负样例预测结果. 实验收集了 UniProt[1)] (日期: 2017–02–20) 中相应蛋白质的序列数据, 采用 BLAST 默认设置进行序列相似度计算, 保留 E 值小于 10 的相似度构造成对蛋白质之间的边连接, 再对网络进行归一化处理, 保证网络中每个蛋白质与其他蛋白质的边权重总和为 1. 对应的实验结果汇报在补充材料的表 S6∼S11 中, 表中 IFDR-Seq 仅在序列数据构造的网络上预测负样例, IFDR-Com 在混合网络上预测负样例, IFDR 仅在互作网上预测负样例. 从这些表中的结果可以看到 IFDR-Seq 和 IFDR-Com 与 IFDR 的 FNs 在少数情况下有可比的结果, 但它们的 FNs 通常大于 IFDR. 这是因为所有成对蛋白质进行序列比对后构造的网络能够比较好地描述成对蛋白质之间功能关联, 在该网络上进行成分扩散分析引入了成对蛋白质之间额外的关联, 这些额外的关联蛋白质之间原始序列比对的 E 值大于 10, 因而 IFDR 利用序列数据后降低了负样例预测效果. 蛋白质互作网中存在一定的缺失互作和噪声互作, 通过成分扩散分析可以挖掘缺失的互作, 再通过 SVD 则可以降低成分扩散分析引入的噪声互作和互作网中已有噪声互作的干扰, 进而提高负样例预测效果. 基于上述实验结果, 本文在后续实验中仅利用蛋白质互作信息进行负样例选择.

3.4 负样例可提高功能预测精度

由于已有蛋白质的功能信息并不完善, 统计 FNs 仅能在一定程度上反映各个负样例预测算法的性能. 研究发现, 结合蛋白质负样例进行功能预测可以提高预测精度 [24, 26, 27, 31]. 为进一步对比分析上述方法预测的负样例, 本文基于 Youngs 等 [24] 提出的改进 GeneMANIA [11] 的方法 SWSN 进行蛋白质功能预测. SWSN 可以同时利用正负样例整合多个蛋白质功能关联网络进行功能预测. 在此部分实验中本文采用 Mostafavi 等 [46] 收集整理的 Yeast 和 Human 多个功能关联网络数据为数据集, 并基于 2015 的蛋白质功能标记数据标注网络中的蛋白质功能. SWSN 将各负样例预测算法预测的负样例, 蛋白质已知的正负样例和多个网络作为输入, 优化这些网络对应的权重, 并将它们加权合并成一个复合网络, 再在复合网络上进行蛋白质功能预测, 再用 2016 年更新的数据检验预测的性能. 通常蛋白质的负样例数远大于正样例数, 在此部分实验中本文设置预测的负样例数量为已知正样例的十倍. Myers 等 [47] 指出特别稀疏的功能标记很难被生物湿实验检验, 参考 SNOB 和 NegGOA 中的实验设置, 实验中不考虑标注的蛋白质个数少于 3 的标记. 在上述实验设置下, 表 3 列出了各对比算法的预测结果. 由于 clusDCA 和 Random 的 FNs 通常远大于其他对比方法, 表 3 没有包含这 2 个方法对应的结果.

从表 3 中可以看到 IFDR 提供的负样例在大部分情况下都获得较这些对比算法更高的精度, 在 24 个 (2 物种 ×4 种评价度量 ×3 个 GO 分支) 对比实验中, IFDR 分别在 70.8%, 75%, 79.2% 和 91.7%

1) http://www.uniprot.org/.

表 3　酵母菌和人类多网络数据集上的蛋白质功能预测结果

Table 3　Results of protein function prediction on multiple networks of Yeast and Human datasets

		Yeast					Human				
		IFDR	ProPN	NegGOA	SNOB	IFDR-DCA	IFDR	ProPN	NegGOA	SNOB	IFDR-DCA
MacroF1	BP	**0.8518**	0.8435	0.7622	0.7626	0.7723	**0.8212**	0.8182	0.8182	0.8099	0.7766
	CC	**0.7459**	0.6406	0.5381	0.5087	0.5452	**0.8221**	0.8124	0.6892	0.6473	0.6341
	MF	**0.9118**	0.9104	0.8252	0.7928	0.8323	**0.8566**	0.8512	0.8408	0.8345	0.7940
RAccuracy	BP	**0.2292**	0.2118	0.2231	0.2219	0.2211	**0.2960**	0.2905	0.2905	0.2895	0.2922
	CC	**0.4217**	0.4026	0.4192	0.3800	0.4131	**0.4094**	0.4068	0.4082	0.4009	0.4041
	MF	**0.2777**	0.2500	0.2706	0.2691	0.2694	**0.4722**	0.4616	0.4676	0.4408	0.4267
AUC	BP	0.9593	0.9616	0.9653	**0.9697**	0.9116	0.9203	0.9305	**0.9381**	0.9329	0.8845
	CC	0.9789	0.9782	0.9797	**0.9808**	0.7728	0.9360	0.9436	**0.9468**	0.9442	0.8203
	MF	**0.9817**	0.9801	0.9809	**0.9817**	0.9505	0.9422	0.9489	**0.9528**	0.9496	0.9032
Fmax	BP	0.6902	0.7758	**0.7763**	0.6388	0.7036	**0.8227**	0.7915	0.8111	0.6547	0.7710
	CC	0.7602	0.7870	**0.8038**	0.7113	0.7717	**0.7888**	0.7779	0.7878	0.7191	0.7880
	MF	**0.8145**	0.7953	0.8018	0.7154	0.8028	**0.8497**	0.8271	0.8318	0.7840	0.8388

情况下优于 NegGOA, ProPN, SNOB 和 IFDR-DCA; 在 29.2%, 25%, 20.8% 和 8.3% 的情况下被这 4 个对比方法超过; 在 0%, 0%, 4.16% 和 0% 的情况下获得跟它们一样的结果. 本文再次通过 Wilcoxon 符号秩检验评估 IFDR 与其他对比算法预测结果的差异性, 对应的 p 值分别为 0.089, 0.034, 0.0008 和 0.0002, 可见 IFDR 显著性优于 ProPN, SNOB 和 IFDR-DCA. IFDR 虽与 NegGOA 的结果相比较显著性并不明显, 但也存在一定优势. 从评价度量上可以发现 IFDR 在 MacroF1 和 RAccuarcy 上总能比其他对比算法获得更高的精度. 这一原因是 IFDR 通过成分扩散分析和 SVD 降低了假阳性互作的破坏作用, 并通过蛋白质 – 功能标记关联矩阵上的缺失正样例预估和 SVD 挖掘了蛋白质与标记间潜在的关联, 从而针对稀疏标记的负样例预测更加准确, 进而帮助 SWSN 更准确地预测蛋白质功能, 特别是蛋白质缺失的功能标记和稀疏标记. 蛋白质的缺失功能标记通常对应蛋白质已知功能标记的子孙节点, 是对这些已知功能的进一步细化. RAccuarcy 评估多少缺失的标记被准确预测, MacroF1 受稀疏标记的影响更大, 因此 IFDR 选择的负样例能够帮助 SWSN 获得较其他算法更高的 RAccuarcy 和 MacroF1. 这种实验结果进一步证明了对蛋白质缺失标记进行建模和对蛋白质网络与标记空间进行降维学习与压缩的必要性和有效性.

IFDR 在大部分情况下可以取得比其他算法更高的 Fmax, 但获得的 AUC 低于其他对比算法. 原因是 IFDR 通过对蛋白质 – 功能标记关联矩阵先进行了降维再进行升维, $\tilde{Y}_F^m$ 存在较多的非零元素, AUC 汇总每个标记在不同阈值下的预测结果, 较多的非零元素影响不同阈值下的预测结果, 所以 IFDR 获得的 AUC 低于其他算法; 而 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 评价功能预测性能, 因此它较少受 $\tilde{Y}_F^m$ 中非零元素过多的干扰.

综上所述, IFDR 不仅能够较其他对比算法更准确地预测蛋白质负样例, IFDR 预测的负样例对蛋白质功能预测精度的提升也通常优于这些对比方法.

CAFA (critical assessment of protein function annotation algorithms) 是蛋白质功能预测领域的一个专业比赛 [2, 3], 本文收集了 CAFA2 的基准数据集 (2013.09~2014.09), 并整理出了人类、酵母菌和拟南芥 3 个物种的相关数据 (包括序列数据、基因表达数据、结构功能域和互作网), 进一步验证本文算

表 4 INGA 结合负样例后蛋白质功能预测结果

Table 4 Results of protein function prediction by INGA without/with using negative examples predicted by IFDR

		S. cerevisiae		H. sapiens		A. thaliana	
		INGA	INGA-Neg	INGA	INGA-Neg	INGA	INGA-Neg
AUC	BP	**0.1509**	0.1507	**0.3819**	0.3815	0.2210	**0.2212**
	CC	0.2050	**0.2051**	0.5030	**0.5053**	**0.2754**	0.2711
	MF	**0.1858**	0.1851	**0.6634**	0.6624	0.2360	**0.2393**
Fmax	BP	0.5383	**0.5385**	0.4558	0.4558	**0.4706**	0.4703
	CC	0.7020	**0.7086**	0.5519	**0.5558**	0.7529	**0.7551**
	MF	0.6462	**0.6470**	0.5702	**0.5729**	**0.5840**	0.5827
RankingLoss[a)]	BP	0.1163	**0.1127**	0.0867	**0.0818**	0.0842	**0.0766**
	CC	0.0949	**0.0902**	0.1108	**0.0912**	0.0333	**0.0280**
	MF	0.0547	**0.0512**	0.0937	**0.0857**	0.1763	**0.1512**
Smin[a)]	BP	12.8237	**12.8212**	26.4586	**26.4523**	14.043	**14.0204**
	CC	5.1867	**5.1677**	6.8148	**6.8088**	5.0567	**4.9987**
	MF	4.2951	**4.2387**	8.1087	**8.0944**	5.2668	**5.1093**

a) The lower value means the better performance.

法的有效性. 本文选用 CAFA2 中 2 种排名前 10 的算法 INGA [48] 和 MSkNN [49] 检验 IFDR 选择的负样例对蛋白质功能预测的贡献. 其中 INGA 分别在蛋白质互作网和结构功能域网络上进行 GO 富集分析, 以及序列同源性进行初步功能预测, 再整合这些异构数据源上的预测结果. MSkNN 通过在每种数据上训练一个 k 近邻分类器, 再整合这些分类器的预测结果. 此处, 本文参照 CAFA2 中采用的基准评价度量 Smin, Fmax 和 AUC, 同时还引入 RankingLoss [16] 作为第 4 种评价度量. Fmax 和 AUC 在前一节中已经有所介绍, Smin 取不同阈值下预测错误标记的结构信息损失和未预测到标记的结构信息损失的最小值, Smin 的值越小表示在结构信息上的损失量越小; RankingLoss 表示在对每个蛋白质的预测结果进行排序后其不相关标记排在相关标记前的比值. 与 Smin 类似, RankingLoss 值越小表明预测质量越好. 考虑到蛋白质功能的负样例数远大于其正样例数, 在此部分实验中本文设置预测的负样例数量为已知正样例的 2 倍. 表 4 和补充材料中的表 S12 分别报告 INGA 和 MSkNN 的预测结果和结合 IFDR 选择的负样例后的预测结果, 其中 INGA-Neg 和 MSkNN-Neg 分别对应它们结合 IFDR 选择的负样例后的结果, INGA 和 MSkNN 对应各自原始结果.

从表 4 和 S12 中的结果可以发现, INGA-Neg 和 MSkNN-Neg 较 INGA 和 MSkNN 在 RankingLoss 和 Smin 上都有了较明显的下降, 在 Fmax 上有微量提升或保持近似. 通过 Wilcoxon 符号秩检验分别统计 MSkNN-Neg 与 MSkNN, INGA-Neg 与 INGA 在这些数据集上的差异性, 对应 p 值分别为 0.048 和 0.00035. 上述实验和统计结果表明 IFDR 预测的负样例能有效地缩小蛋白质功能预测问题规模, 提升已有蛋白质功能预测算法的精度. INGA-Neg (MSkNN-Neg) 利用负样例后在评价度量 AUC 上的提升并不明显甚至有所下降, 这是因为 INGA-Neg (MSkNN-Neg) 预测的蛋白质 – 功能标记关联概率矩阵中存在较多的非零元素, AUC 是一种以标记为中心的度量, 它汇总每个标记在不同阈值下的预测结果, 较多的非零元素影响其不同阈值下的预测结果.

3.5 降维贡献分析

IFDR 通过在蛋白质互作网的邻接矩阵和蛋白质 – 功能标记关联矩阵上分别进行基于 SVD 的维

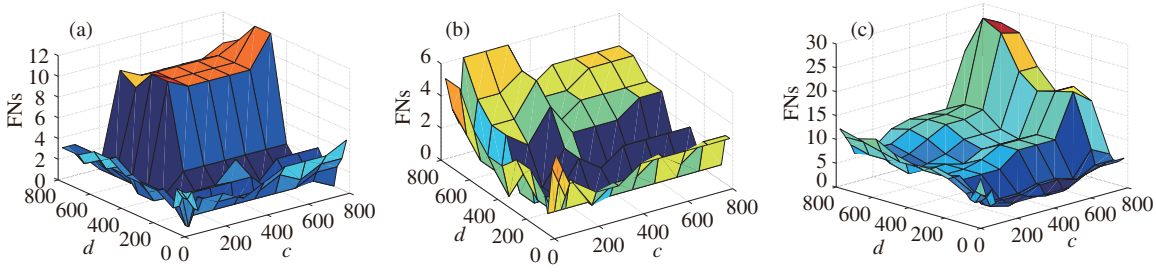

图 1 (网络版彩图) 人类数据集上降维目标维度影响 (d 和 c 分别为蛋白质互作网和标记空间降维的目标维度)

Figure 1 (Color online) Influence of the size of target dimensionality on H.sapiens. d and c represent the target dimensionality of PPI network and function label space, respectively. (a) BP; (b) CC; (c) MF

数约减, 再利用半监督回归进行负样例预测. 本小节探索降维的目标维度对 IFDR 的影响, 令 d 和 c 分别表示蛋白质互作网和蛋白质 – 功能标记关联矩阵降维后的特征维度大小, 在本部分实验中将 d 和 c 分别从 20 增加到 800, 并登记不同 d 和 c 组合下 IFDR 的 FNs. 在人类蛋白质数据集上的实验结果 (2015~2016) 如图 1 所示.

从图 1 中可以发现, 总体上随着投影维度的下降, FNs 越变越小, 但在 $d = 20$ 或 $c = 20$ 时, FNs 并不是最小的. 原因是这 2 类数据均需要一定量的特征描述原始蛋白质之间的互作和蛋白质与功能标记间的关联. 在 $d > 500$ 或 $c > 500$ 时, FNs 不断增大, 原因是较大的目标维度易引入较多的噪声特征, 降低半监督回归的性能从而增大 FNs. 在对蛋白质互作网进行降维后, FNs 下降比较明显, 这是因为蛋白质互作网本身存在一部分噪声互作, 虽然成分扩散分析能够在一定程度上降低噪声互作的干扰, 但是成分扩散分析本身也会传播噪声互作. 上述观察表明在成分扩散分析更新的邻接矩阵上进行 SVD 是合理也且必要的. 在蛋白质 – 功能标记关联矩阵上的降维效果情况与蛋白质互作网相似, FNs 随着维度下降而增大. 这一观察表明 SVD 可以在一定程度上降低预估的假阳性缺失标记的破坏作用. 从图中还可以发现对蛋白质互作网和蛋白质 – 功能标记关联矩阵分别进行降维压缩表示是有效的. 本文统计了 IFDR 在上述实验中取得最小 FNs 时, d 和 c 各应取所有奇异值总和中约前 25% 和 45% 的奇异值对应的特征向量实现对网络空间和标记空间的降维.

为了更加深入地分析在蛋白质互作网空间和功能标记空间上降维的贡献, 本文基于 IFDR 衍生出 4 个变种 IFDR-F, IFDR-P, IFDR-FSVD 和 IFDR-N. IFDR-F 只对蛋白质 – 功能关联矩阵进行潜在功能标记预估和降维, 直接利用原始蛋白质互作网进行负样例预测; IFDR-P 只对蛋白质互作网进行成分扩散分析和降维, 再在原始标记空间进行负样例预测; IFDR-FSVD 在蛋白质互作网上进行成分扩散分析和降维, 但仅对蛋白质 – 功能关联矩阵进行 SVD 降维. IFDR-N 直接在原始标记空间和蛋白质互作网上进行负样例预测. 在本小节实验中蛋白质互作网和标记空间压缩的目标维度 (d 和 c) 均设为 300. 限于篇幅, 表 5 仅报告 IFDR 和 4 个变种在人类数据集上的实验结果, 其他数据集上的实验结果在补充材料的表 S13 和 S14 中.

从这些表中数据可以看出, IFDR-N 的 FNs 通常是最高的, 这是因为 IFDR-N 没有对蛋白质的缺失功能标记进行预估, 也没有考虑蛋白质之间的噪声互作的破坏作用. IFDR-P 虽然通过成分扩散分析和基于 SVD 的降维降低了假阴性互作和假阳性互作的影响, 但并没有很好地解决蛋白质功能标记的不平衡性和稀疏性, 所以它的 FNs 也相对较高. IFDR-F 的 FNs 比 IFDR-N 和 IFDR-P 更小, 说明利用蛋白质已知的功能标记结合基因本体结构对潜在的正样例进行预估, 再通过 SVD 降低错误预估的正样例的破坏作用是有效的. 从 IFDR-F 与 IFDR 之间 FNs 的差值和 IFDR-P 与 IFDR 之间

表 5 人类数据集上不同 IFDR 变种的负样例预测结果

Table 5 Results of negative examples prediction on H. sapiens by different variants of IFDR

	Data set	l							
		10k	20k	30k	40k	50k	60k	70k	80k
BP	IFDR	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
	IFDR-F	1	1	3	5	13	13	17	18
	IFDR-P	30	63	101	134	160	192	228	264
	IFDR-N	28	52	120	174	211	259	300	329
	IFDR-FSVD	7	7	12	14	15	15	20	21
CC	IFDR	**0**	**0**	**1**	**1**	**2**	**2**	**2**	**2**
	IFDR-F	**0**	1	2	2	5	6	7	7
	IFDR-P	34	88	118	158	228	273	349	440
	IFDR-N	23	83	124	173	229	293	363	449
	IFDR-FSVD	**0**	3	8	12	16	16	18	18
MF	IFDR	**0**	3	4	5	6	8	**8**	**8**
	IFDR-F	7	11	15	15	17	20	21	22
	IFDR-P	10	23	49	66	83	105	124	132
	IFDR-N	12	26	30	36	46	53	58	63
	IFDR-FSVD	3	**3**	**3**	**4**	**4**	**5**	**8**	**8**

FNs 的差值可知, 相对于对蛋白质互作网进行降维, 对蛋白质的标记空间进行降维对负样例的预测精度提升作用更大. IFDR-P 的 FNs 通常小于 IFDR-N 进一步证实对蛋白质互作网数据进行成分扩散和降维的必要性. 由于实验中统一设置 $d = 300$, 并没有根据不同物种的蛋白质互作网进行优化, 因此 IFDR-P 有时候获得较 IFDR-N 高一些的 FNs. IFDR-FSVD 实际上是在 IFDR-P 基础上继续对蛋白质 – 功能标记关联矩阵降维, 它的 FNs 要少于 IFDR-P, 与 IFDR-F 在部分设置下能获得类似的 FNs, 这说明 SVD 能在保证正样例标记的特征下, 减少噪声对预测结果的影响. 但 IFDR-FSVD 的 FNs 通常大于 IFDR, 这进一步证明对潜在正样例预估可以提升负样例预测的准确性. 上述实验分析证明对蛋白质互作网的邻接矩阵和蛋白质 – 功能标记关联矩阵进行降维压缩是必要而且有效的, 这种处理可以降低负样例预测的假阴性数.

此外, 为了分析不同来源的蛋白质互作网数据对 IFDR 的影响, 本文还下载了 STRING 数据库 [50] 中的人类和酵母菌的蛋白质互作网络数据, 对网络中的蛋白质采用了类似 3.1 小节的方法标注, 再进行实验. 补充材料的表 S15 和 S16 登记了 IFDR 和上述 4 变种在 STRING 数据库上的结果. 这些方法在人类和酵母菌上的结果与 BioGrid 上结果类似, 均表明对网络和标记空间降维可提升负样例预测精度, 不同数据库对 IFDR 的影响很小.

3.6 运行时间分析

SNOB 需要计算成对标记间的经验条件概率, 对应时间复杂度为 $O(m^2)$, 它评估 n 个蛋白质与 m 个标记的正负关联性对应的时间复杂度为 $O(nm^2)$, 所以 SNOB 总的时间复杂度为 $O(nm^2)$. NegGOA 不仅需要计算成对标记间的条件概率, 还需计算标记之间的结构相似度, 对应的时间复杂度为 $O(m^2 + m^3)$; NegGOA 再进行随机游走拓展功能标记之间的关联关系, 进而预测蛋白质负样例, 此部分的时

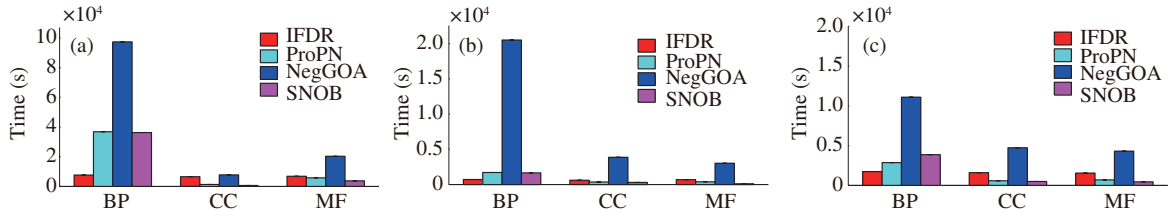

图 2 (网络版彩图) 对比算法在不同数据集上的运行时间统计

Figure 2 (Color online) Statistics of runtime cost of four comparing methods on different datasets. (a) H.sapiens; (b) S.cerevisiae; (c) A.thaliana

间复杂度为 $O(nm^2)$; 因此 NegGOA 总的时间复杂度为 $O(m^3+nm^2)$. ProPN 构建和初始化一个包含 $n+m$ 节点的符号混合图的时间复杂度为 $O(n^2+m^2+nm)$, 混合图上标记传播的时间复杂度为 $O((n+m)^2m)$, 因此 ProPN 总的时间复杂度为 $O((n+m)^2m)$. IFDR 分别在蛋白质互作网上进行成分扩散分析和在 GO 对应的有向无环图上进行 n 个蛋白质的缺失正样例预估, 对应的时间复杂度分别为 $O(n^3)$ 和 $O(nm^2)$, IFDR 再分别通过 SVD 进行奇异值分解降维, 这部分的时间复杂度分别为 $O(n^3)$ 和 $O(\min\{nm^2+n^2m\})$; 在 d 维特征空间和 c 维标记空间上的半监督线性回归的时间复杂度为 $O(nd^2+n^2c)$, 由于 $d\ll n$ 和 $c\ll m$, 因此 IFDR 总的时间复杂度为 $O(n^3+\min\{nm^2+n^2m\})$. 需指出的是蛋白质互作网和基因本体结构对应的邻接矩阵均为稀疏矩阵, 故这些算法的实际运行复杂度要低于上述分析结果. 此外 IFDR 仅需求解前 d (或 c) 个奇异值及其对应的左右特征向量, 故 IFDR 在 SVD 这一部分的时间运行复杂度也低于理论分析结果.

与 3.3 小节的实验设置一样, 本文记录了除基准方法 Random 以外其他算法在不同数据集上的运行时间 (5 次平均), 并报告在图 2 中. 所有算法均基于 Matlab 2012b (64 bit) 实现, 实验运行平台配置为: Linux OS 2.6.32, Intel Xeon E5-2678v3 和 256 GB RAM. 从运行时间上来看, SNOB 在 CC 和 MF 分支的运行时间最小, 但在 BP 上总是高于 IFDR. 这是因为 SNOB 需要计算 m 个标记之间的条件概率, 其时间复杂度为 m 的平方, BP 中的功能标记个数远大于 CC 和 MF 分支. NegGOA 不仅需要计算标间的条件概率, 还利用 GO 结构计算标记间的转移概率, 由于标记集合 m 很大, 所以其时间耗费远大于其他对比方法. 虽然 ProPN 采用一个混合图进行负样例预测, 混合图对应的邻接矩阵为 $n+m$ 的方阵, 远大于其他对比方法, 但由于对应的邻接矩阵为稀疏矩阵, 所以其运行时间总是小于 NegGOA, 有时跟 SNOB 的相近. ProPN 利用余弦相似性度量计算标记间的关联关系, 再针对 m 个标记进行负样例预测, 所以其运行时间在标记个数多的 BP 分支总是大于 IFDR, 在标记个数小的 CC 和 MF 分支小于 IFDR. 本文提出的 IFDR 由于利用 SVD 对蛋白质互作网络和蛋白质功能标记空间分别进行了降维压缩, 在功能标记最多的 BP 分支上的运行时间小于 ProPN 和 SNOB, 但在功能标记较少的 CC 和 MF 分支上大于这两个方法. 这是因为蛋白质个数远大于 MF 和 CC 分支的标记个数, 蛋白质互作网上的成分扩散分析和 SVD 的耗时远大于蛋白质 – 功能标记关联矩阵上 SVD 的耗时. 从图 2 和前 2 节的实验结果可以发现 IFDR 算法不仅能较相关算法更准确地预测蛋白质的负样例, 还能保持较高的效率, 特别在功能标记集合比较大的物种上.

4 结束语

本文针对蛋白质负样例预测中的正样例信息不完整和标记空间过大等问题, 提出一种基于蛋白质

互作网和功能标记降维的负样例预测方法 (IFDR). IFDR 通过对蛋白质互作网络进行成分扩散分析和对蛋白质缺失正样例进行预估后, 再利用 SVD 对噪声特征鲁棒的特点分别对网络空间和标记空间进行降维压缩, 最后采用一个半监督回归方法预测蛋白质负样例. 与现有负样例预测方法相比, IFDR 不仅同时利用了蛋白质已知的正样例和少量负样例, 还结合成分扩散分析和降维降低了噪声的破坏作用, 提高了负样例预测效率和质量. 实验表明, 对蛋白质互作网络和功能标记空间进行降维是合理且有效的. 针对降维后数据的特性, 本文选用了一个简单的半监督线性回归方法预测负样例, 后续研究将探索新的回归方法, 进一步提高负样例预测精度. 本文实验中简单设置网络和标记空间的目标维度 (d 和 c) 均为 300, 后续工作将调研如何根据具体的数据集分别设置合适的 d 和 c, 探索其他降维方法对 IFDR 的影响也是一个值得研究的方向.

参考文献

1 Roberts R J. Identifying protein function—a call for community action. PLoS Biology, 2004, 2: e42

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

3 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 Biology, 2016, 17: 1–19

4 Ashburner M, Ball C A, Blake J A, et al. Gene ontology: tool for the unification of biology. Nature Genetics, 2000, 25: 25–29

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: 1

6 Lee D, Redfern O, Orengo C. Predicting protein function from sequence and structure. Nature Reviews Molecular Cell Biology, 2007, 8: 995–1005

7 Deng L, Chen Z. An integrated framework for functional annotation of protein structural domains. IEEE/ACM Trans Comput Biology Bioinform, 2015, 12: 902–913

8 Sharan R, Ulitsky I, Shamir R. Network-based prediction of protein function. Molecular Syst Biology, 2007, 3: 1–15

9 Schwikowski B, Uetz P, Fields S, et al. A network of protein-protein interactions in Yeast. Nature Biotech, 2000, 18: 1257–1261

10 Vazquez A, Flammini A, Maritan A, et al. Global protein function prediction from protein-protein interaction networks. Nature Biotech, 2003, 21: 697–700

11 Mostafavi S, Ray D, Warde-Farley D, et al. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. Genome Biology, 2008, 9: S4

12 Cesa-Bianchi N, Re M, Valentini G. Synergy of multi-label hierarchical ensembles, data fusion, and cost-sensitive methods for gene functional inference. Mach Learn, 2012, 88: 209–241

13 Yu G X, Zhu H L, Domeniconi C, et al. Integrating multiple networks for protein function prediction. BMC Syst Biology, 2015, 9: S3

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

15 Valentini G. True path rule hierarchical ensembles for genome-wide gene function prediction. IEEE/ACM Trans Comput Biology Bioinform, 2011, 8: 832–547

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

17 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

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

19 Yu G X, Domeniconi C, Rangwala H, et al. Protein function prediction using dependence maximization. In: Proceedings of the 24th European Conference on Machine Learning. Berlin: Springer, 2013. 574–589
20 Yu G X, Zhu H L, Domeniconi C. Predicting protein function using incomplete hierarchical labels. BMC Bioinform, 2015, 16: 1
21 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: 273
22 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]
23 Schnoes M, Ream D, Thorman A, et al. Biases in the experimental annotations of protein function and their effect on our understanding of protein function space. PLoS Comput Biology, 2013, 9: e1003063
24 Youngs N, Duncan P, Kevin D, et al. Parametric Bayesian priors and better choice of negative examples improve protein function prediction. Bioinformatics, 2013, 29: 1190–1198
25 Wang H, Huang H, Ding C. Function-function correlated multi-label protein function prediction over interaction networks. J Comput Biology, 2013, 20: 322–343
26 Youngs N, Penfold-Brown D, Bonneau R, et al. Negative example selection for protein function prediction: the NoGO database. PLoS Comput Biology, 2014, 10: e1003644
27 Yu G X, Rangwala H, Domeniconi C, et al. Protein function prediction with incomplete annotations. IEEE ACM Trans Comput Biology Bioinform, 2014, 11: 579–591
28 Blei D, Ng A, Jordan M. Latent dirichlet allocation. J Mach Learn Res, 2003, 3: 993–1022
29 Fu G Y, Wang J, Yang B, et al. NegGOA: negative GO annotations selection using ontology structure. Bioinformatics, 2016, 32: 2996–3004
30 Tong H H, Faloutsos C, Pan J Y. Random walk with restart: fast solutions and applications. Knowl Inf Syst, 2008, 14: 327–346
31 Fu G Y, Yu G X, Wang J, et al. Protein function prediction using positive and negative examples. J Comput Sci 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]
32 Cao M, Pietras C M, Feng X, et al. New directions for diffusion-based network prediction of protein function: incorporating pathways with confidence. Bioinformatics, 2014, 30: 219–227
33 Cho H, Berger B, Peng J. Diffusion component analysis: unraveling functional topology in biological networks. In: Proceedings of the 19th Annual International Conference on Research in Computational Molecular Biology. Berlin: Springer, 2015. 62–64
34 Wang S, Cho H, Zhai C X, et al. Exploiting ontology graph for predicting sparsely annotated gene function. Bioinformatics, 2015, 31: 357–364
35 Banerjee S, Roy A. Linear Algebra and Matrix Analysis for Statistics. BocaRaton: CRC Press, 2014
36 Guo M Z, Dai Q G, Xu L Q, et al. On protein complexes identifying algorithm based on the novel modularity function. J Comput Res Dev, 2014, 51: 2178–2186 [郭茂祖, 代启国, 徐立秋, 等. 一种蛋白质复合体模块度函数及其识别算法. 计算机研究与发展, 2014, 51: 2178–2186]
37 Kullback S, Leibler R A. On information and sufficiency. Ann Math Stat, 1951, 22: 79–86
38 Pandey G, Myers C L, Kumar V. Incorporating functional inter-relationships into protein function prediction algorithms. BMC Bioinform, 2009, 10: 1
39 Zhang X F, Dai D Q. A framework for incorporating functional interrelationships into protein function prediction algorithms. IEEE/ACM Trans Comput Biology Bioinform, 2012, 9: 740–753
40 Alter O, Brown P O, Botstein D. Singular value decomposition for genome-wide expression data processing and modeling. Proc National Academy Sci, 2000, 97: 10101–10106
41 Zhu X J. Semi-supervised learning literature survey. Comput Sci, 2008, 37: 63–77
42 Belkin M, Niyogi P, Sindhwani V. Manifold regularization: a geometric framework for learning from labeled and

unlabeled examples. J Mach Learn Res, 2006, 7: 2399–2434
43 Spirin V, Mirny L A. Protein complexes and functional modules in molecular networks. In: Proceedings of the National Academy of Sciences, 2003, 100: 12123–12128
44 Chatr-Aryamontri A, Breitkreutz B J, Oughtred R, et al. The BioGRID interaction database: 2015 update. Nucleic Acids Res, 2015, 43: 470–478
45 Wilcoxon F. Individual comparisons by ranking methods. Biometrics Bulletin, 1945, 1: 80–83
46 Mostafavi S, Morris Q. Fast integration of heterogeneous data sources for predicting gene function with limited annotation. Bioinformatics, 2010, 26: 1759–1765
47 Myers C L, Barrett D R, Hibbs M A, et al. Finding function: evaluation methods for functional genomic data. BMC Genomics, 2006, 7: 1
48 Piovesan D, Giollo M, Leonardi E, et al. INGA: protein function prediction combining interaction networks, domain assignments and sequence similarity. Nucleic Acids Res, 2015, 43: 134–140
49 Lan L, Djuric N, Guo Y, et al. MS-kNN: protein function prediction by integrating multiple data sources. BMC Bioinformatics, 2013, 14: S8
50 Szklarczyk D, Franceschini A, Wyder S, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res, 2015, 43: 447–452

Predicting irrelevant functions of proteins based on dimensionality reduction

Guoxian YU[1*], Guangyuan FU[1], Jun WANG[1] & Maozu GUO[2]

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*
* Corresponding author. E-mail: gxyu@swu.edu.cn

Abstract Proteins are the foundation for many life processes and accurately annotating their biological functions can significantly boost the development of life sciences. Current function prediction models focus on employing the knowledge that proteins perform specific functions (positive examples), but ignore the knowledge that some functions are irrelevant for a protein (negative examples). Recent research indicates that incorporating negative examples can reduce the complexity and improve the accuracy of protein function prediction. In this paper, we propose an approach for predicting irrelevant functions of proteins based on dimensionality reduction (IFDR). Initially, IFDR performs random walks through matrices in a protein-protein interactions (PPI) network, as well as the corresponding protein-function association matrices, in order to explore the underlying relationships between proteins and model the missing functional annotations of proteins. Next, IFDR uses single value decomposition to project these matrices into low-dimensional numerical matrices. Finally, IFDR uses semi-supervised regression to predict negative examples of proteins. Experiments on S. cerevisiae, H. sapiens, and A. thaliana data demonstrate that IFDR can more accurately predict negative examples when compared to related methods. Dimensionality reduction in the network space and label space can both improve the accuracy of negative example prediction.

Keywords protein function prediction, positive and negative examples, PPI network, function label, dimensionality reduction

Guoxian YU was born in 1985. He received a 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.

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

Guangyuan FU was born in 1993. He received a B.S. degree in computer science from Southwest University, Chongqing, in 2015. He is currently a Master's student at the College of Computer and Information Sciences, Southwest University. His research interests include machine learning and bioinformatics.

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