中国科学: 信息科学 2016 年 第 46 卷 第 4 期: 461–475
SCIENTIA SINICA Informationis
《中国科学》杂志社
SCIENCE CHINA PRESS
论 文

基于有向混合图的蛋白质新功能预测

傅广垣[①], 余国先[①②*], 王峻[①], 张自力[①]

① 西南大学计算机与信息科学学院, 重庆 400715
② 吉林大学符号计算与知识工程教育部重点实验室, 长春 130012
* 通信作者. E-mail: gxyu@swu.edu.cn

收稿日期: 2015–05–20; 接受日期: 2015–06–11; 网络出版日期: 2016–04–13
国家自然科学基金 (批准号: 61402378, 61101234)、重庆市基础与前沿研究计划 (批准号: cstc2014jcyjA40031) 和中央高校基本科研业务费 (批准号: XDJK2014C044, 2362015XK07) 资助项目

摘要 蛋白质执行着生物体内各种重要生物活动, 对蛋白质功能的准确标注能极大地促进生命科学研究与应用. 传统的湿实验法通量低, 已无法测定高通量技术产生的海量蛋白质功能, 基于计算模型的大规模蛋白质功能预测是后基因时代生物信息学的核心任务之一. 当前基于机器学习的方法通常仅关注对完全未标记功能的蛋白质的功能预测, 而忽略了已标注功能的蛋白质可能存在的自身功能标记的不完整性, 预测精度有限. 本文结合基因本体层次结构关系和蛋白质互作网信息, 设计了一种有向混合图 (directed hybrid graph, dHG) 对上述信息进行描述, 并在此基础上提出一种基于有向混合图重启动随机游走的蛋白质功能预测方法 —— dHG. 本文提出的 dHG 方法不仅能补充已知部分功能标记的蛋白质新功能, 还能预测功能完全未知的蛋白质新功能. 在酵母菌和人类蛋白质上的实验结果表明, dHG 在多种评价度量上的预测性能均优于现有方法, 且效率更高.

关键词 蛋白质功能预测 机器学习 有向混合图 随机游走 基因本体 蛋白质互作网

1 引言

蛋白质是生命活动的物质基础, 它执行着生物体内各种重要的功能, 如生物组织的构造, 信号的识别和传导等. 蛋白质的功能信息不仅能为生命科学研究提供参照信息, 对疾病机理分析与调控、新药品研发、农作物促产、生物能源开发等研究领域都有着促进作用 [1]. 在过去 10 多年中, 蛋白质的功能不断从生物实验中测出, 并添加到蛋白质功能标记数据库中 (如 gene ontology) [2]. 然而随着海量高通量数据的产生, 待标注功能的蛋白质数量仍在飞速增长. 传统基于湿实验的功能测定方法成本高, 通量低且受限于实验条件, 导致已标注功能蛋白质的功能标记通常是不完整的, 存在缺失的功能标记. 当前已知的约 20000 个人类蛋白质中有 2/3 的蛋白质功能信息是未知或不完整的, 亟待进一步深入研究 [3]. 基于机器学习的计算方法能够大规模自动预测蛋白质功能, 为后续蛋白质功能测定实验提供指导, 减少实验验证的规模和成本, 近些年得到了广泛的研究 [1, 4].

引用格式: 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, doi: 10.1360/N112015-00109

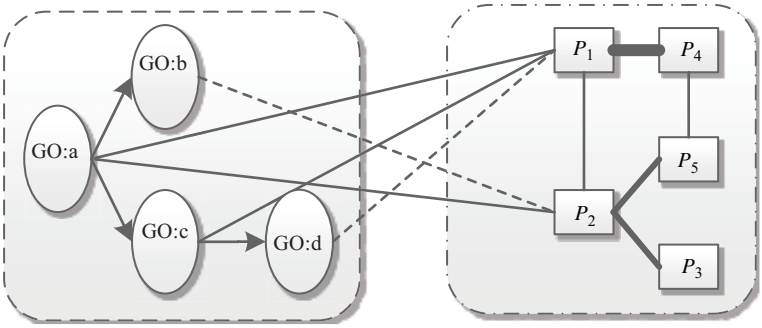

图 1　已知部分标记蛋白质示意图

Figure 1　An example of the partially labeled proteins

研究发现, 拥有相似氨基酸序列和三维结构的蛋白质通常共享相同的功能 [5, 6]. 蛋白质之间通过互作完成具体的生物功能, 这个互作的蛋白质之间通常形成一个功能模块或通路 [7, 8]. 基于蛋白质的上述生物学特性, 研究者提出多种基于蛋白质序列、三维结构、基因表达信息、互作网和通路数据的蛋白质功能预测方法 [4, 8]. 随着各种异构生物数据的不断获取, 一些研究者提出了多种数据集成方法对上述蛋白质数据进行整合利用, 在一定程度上克服了单种数据源描述蛋白质生物信息的不足, 进一步提高了蛋白质预测的精度 [9, 10]. 同一个蛋白质通常参与到不同的生命活动中, 具有多个不同的功能, 每个功能可以视为一个标记. 早期的蛋白质功能预测方法通常把功能预测问题转化为二分类问题, 对每个功能标记分别进行预测 [6], 这类方法忽略了功能标记间的关联关系和功能标记的不平衡特性, 预测精度不高. 近期, 研究者开始将功能预测问题转化为多标记学习问题进行研究 [10~13], 这类方法通过利用功能标记间的关联关系获得了较高的预测精度. 但这类方法通常仅利用了标记间的水平关系, 未充分利用标记间的层次结构关系.

已有研究发现, 功能标记间的层次结构关系在蛋白质功能预测中发挥着至关重要的作用 [14, 15]. 基因本体 (gene ontology, GO) 作为一种广泛使用的蛋白质 (或基因产物) 功能标记范式, 它用一个有向无环图描述功能之间的关系. 图 1 中左侧的子图即为一个 GO 有向无环图的简单示例, 子图中节点间的有向箭头表示标记之间的层次结构关系, 如 GO:a 为 GO:b 和 GO:c 的父节点, GO:a 为 GO:d 的祖先节点. 右边的子图为一个蛋白质互作网的示例, 节点的连线表示蛋白质的互作, 连线的粗度表示互作的强度或置信度. 蛋白质的功能标记信息通过功能标记和蛋白质之间的连线表示, 其中实线表示蛋白质已知的功能标记信息, 虚线表示蛋白质的缺失功能标记 (蛋白质实际具有该功能, 但该功能尚未被标注到该蛋白质上), 这些缺失的功能标记称为蛋白质的新功能. 为了简洁, 完全未标注蛋白质 (P_3, P_4 和 P_5) 的新功能标记并未在图中体现. 可以发现, 每个蛋白质的已知功能标记集合可以定义一个层次结构图, 该蛋白质的缺失功能标记只能是这个结构图中节点的子孙节点. 如 P_1 的层次结构图由 GO:a 和 GO:c 构成, 它的缺失标记通常是 GO:a (或 GO:c) 的子孙节点. 这些缺失功能是蛋白质已知功能的进一步细化, 包含了更多的生物学信息, 更具有指导意义. 本文将基于这一原理, 研究如何利用功能标记间的层次结构关系和蛋白质互作网, 预测已知部分功能蛋白质的新功能 (即将图中的虚线变为实线) 和完全未标注蛋白质的新功能. 国际大规模蛋白质功能预测评测组织近期也关注到这个问题, 并将预测已知部分功能蛋白质的新功能作为一个新的任务 [1, 16].

预测已知部分功能蛋白质的新功能可以转化为多标记弱标记学习方法进行研究 [17~19]. 如 Sun 等 [17] 提出一种多标记弱标记学习方法, 该方法假定不同标记的样本之间存在较大的间隔, 每个标记仅包含少量的样例和标记之间的关系可通过一组基于低秩逼近的关联描述. 但该方法需通过耗时的

二次规划求取低秩逼近, 难以处理较大的数据和标记集合. Bucak 等 [18] 提出一种基于排序损失和组稀疏损失的弱标记学习方法 MLR-GL, 该方法可以利用不完整标记的样本预测标记完全未知样本的标记. Yu 等 [19] 提出一种基于弱标记学习的功能预测方法, 该方法基于功能标记之间的关联关系和蛋白质已知的功能, 预估蛋白质的缺失功能. 在此基础上, 结合蛋白质互作网进一步预测已知部分功能蛋白质的新功能和完全未标注蛋白质的功能. Yu 等 [20] 假定蛋白质的功能依赖于蛋白质的特征信息, 提出一种最大化依赖的功能预测方法 ProDM.

然而, 上述弱标记学习方法仅利用了功能标记之间的水平关系, 忽略了对标记间层次结构关系的利用. 李彦辉等 [21] 结合功能标记层次结构关系定义功能特异的蛋白质互作子网, 基于子网预测蛋白质的新功能. Tao 等 [22] 结合功能标记间的层次结构和蛋白质的已知功能标记, 定义蛋白质之间的语义相似性度量, 并基于此度量定义蛋白质之间的近邻关系, 再借助 k 近邻分类器预测蛋白质的新功能. Done 等 [23] 在蛋白质功能标记关联矩阵上结合向量空间模型, 潜语义分析和奇异值分解预测人类蛋白质的新功能. 但这些方法无法预测功能完全未知蛋白质的功能. Yu 等 [24] 提出一种基于层次弱标记的功能预测方法 PILL. PILL 基于标记间的层次结构关系和标记的分布信息定义了一种标记间关系度量, 基于此度量预估蛋白质的缺失功能标记, 再结合蛋白质互作网预测已知部分功能蛋白质的新功能和完全未标注蛋白质的新功能. 然而该方法在预估缺失功能标记过程中并没有考虑标记间的层次结构关系, 容易产生较多的假阳性预估.

针对上述问题, 本文提出一种基于有向混合图的蛋白质新功能预测方法 dHG (novel protein function prediction using directed hybrid graph). dHG 的主要流程如下: (1) 基于功能标记结构关系和蛋白质互作网构造一个有向混合图 (如图 1 所示), 图中每个节点对应一个功能标记或蛋白质; (2) 结合蛋白质已有的功能标记和蛋白质互作关系在该图上定义标记之间的转移概率, 及标记与蛋白质之间的转移概率; (3) 在该图上进行重启动随机游走预测蛋白质的功能. 在酵母菌和人类蛋白质上的实验结果表明: dHG 不仅能够预测已知部分功能蛋白质的新功能, 而且能够预测功能完全未知蛋白质的功能, 其性能显著优于现有方法. 下文将对该方法具体工作原理和过程进行详细描述.

2 基于有向混合图的弱标记预测

dHG 主要由功能标记之间的转移概率定义和有向混合图上的重启动随机游走两部分构成, 本节将首先对有向混合图上的重启动随机游走算法进行重点描述, 再对功能标记之间的转移概率定义进行介绍.

2.1 有向混合图上的重启动随机游走

蛋白质之间通过互作完成多种生物功能, 这些互作的蛋白质之间构成了一个蛋白质互作网, 它们可能共享相同的功能, 而蛋白质之间的互作可以通过图进行描述, 因此很多基于图的分类方法被应用到蛋白质功能预测中 [12, 25, 26]. 这类方法通常把每个蛋白质当作图上的一个节点, 再在图上设计算法 (如标签传播) 预测蛋白质的功能. 文献 [12], [26] 和 [27] 在标签传播过程中引入功能标记之间的关联关系, 提高了预测精度, 但这些方法在标签传播的过程中容易出现已有功能标记被覆盖的风险. 为了克服这个问题, 文献 [10] 提出了一种有向双关系图, 该图限制功能标记信息仅能向蛋白质互作子网传播, 该方法虽提高了一定的精度, 但与上述方法类似, 它没有很好地利用功能标记间的层次结构信息.

受有向双关系图启发, 本节基于功能标记之间的层次结构关系、蛋白质互作网和蛋白质与功能标记之间的关系构造一种如图 1 所示的有向混合图. 该混合图由两种节点和 3 种边构成, 两种节点分别是功能标记节点和蛋白质节点, 3 种边分别是蛋白质之间的边、功能标记之间的有向边、蛋白质与功能标记之间的边. 当两个蛋白质之间存在互作时, 它们之间存在一条边; 当两个功能节点存在父子关系时, 它们之间存在一条由父节点指向子节点的边; 当已知蛋白质的某个功能标记时, 该蛋白质与这个功能标记之间也存在一条边.

令 N 表示蛋白质个数, C 表示功能标记的个数. 有向混合图上的随机游走可通过图上节点间的概率转移矩阵 $W \in \mathbb{R}^{(C+N)\times(C+N)}$ 描述,

$$W = \begin{bmatrix} \alpha W_{\mathrm{FF}} & (1-\alpha)W_{\mathrm{FP}} \\ (1-\alpha)W_{\mathrm{PF}} & \alpha W_{\mathrm{PP}} \end{bmatrix}, \tag{1}$$

其中, $W_{\mathrm{FF}} \in \mathbb{R}^{C\times C}$ 和 $W_{\mathrm{PP}} \in \mathbb{R}^{N\times N}$ 分别描述功能标记之间和蛋白质之间的转移概率, $W_{\mathrm{FP}} \in \mathbb{R}^{C\times N}$ 描述功能标记与蛋白质之间的转移概率, W_{PF} 为 W_{FP} 的转置, $\alpha \in (0,1)$ 对功能标记子图和蛋白质互作子图的相对重要性进行调控, $1-\alpha$ 对随机游走子从一个子图跳到另一个子图的频率进行调控. W_{FF} 和 W_{PP} 的计算方式为

$$W_{\mathrm{FF}} = D_{\mathrm{R,FF}}^{-1/2} S_{\mathrm{FF}} D_{\mathrm{L,FF}}^{-1/2}, \quad W_{\mathrm{PP}} = D_{\mathrm{PP}}^{-1} S_{\mathrm{PP}}, \tag{2}$$

其中, S_{FF} 是一个 $C \times C$ 的矩阵, 表示功能标记之间的关联关系, 其具体设置将在后面的小节中详细介绍. S_{PP} 是一个 $N \times N$ 的对称矩阵, 表示蛋白质之间互作的强度, $D_{\mathrm{R,FF}}$ 和 $D_{\mathrm{L,FF}}$ 分别为 $C \times C$ 的对角矩阵, 其对角元素分别为 S_{FF} 的行和与列和, D_{PP} 是一个对角矩阵, 其对角元素为 S_{PP} 的行和(或列和).

类似地, W_{FP} 和 W_{PF} 的计算方式为

$$W_{\mathrm{FP}} = D_{\mathrm{FP}}^{-1/2} S_{\mathrm{FP}} D_{\mathrm{PF}}^{-1/2}, \quad W_{\mathrm{PF}} = D_{\mathrm{PF}}^{-1} S_{\mathrm{PP}} D_{\mathrm{FP}}^{-1/2}, \tag{3}$$

其中, $S_{\mathrm{PF}} \in \mathbb{R}^{N\times C}$ 是蛋白质与功能标记之间的关联矩阵, 若蛋白质 i 具有功能 c, 则 $S_{\mathrm{PF}}(i,c) = 1$; 否则 $S_{\mathrm{PF}}(i,c) = 0$. S_{FP} 是 S_{PF} 的转置. $D_{\mathrm{FP}} \in \mathbb{R}^{C\times C}$ 和 $D_{\mathrm{PF}} \in \mathbb{R}^{N\times N}$ 均为对角矩阵, 其对角元素分别为 S_{PF} 的列和与行和.

从图 1 可以观察到每个功能标记节点与一组同它有关联的蛋白质之间存在边连接, 本文把第 c 个功能节点和已知具有该功能的蛋白质共同作为一个集合 G_c, 其定义如下:

$$G_c = v_c^{\mathrm{F}} \cup \{v_i^{\mathrm{P}} | S_{\mathrm{PF}}(i,c) = 1\}, \tag{4}$$

其中, v_c^{F} 为混合图中第 c 个功能标记对应的节点, v_i^{P} 是混合图中第 i 个蛋白质节点. 当要判定蛋白质 j 是否具有功能标记 c 时, 可以基于蛋白质 j 与 G_c 中多个蛋白质之间的关系判定, 而不是基于单个蛋白质判定. 这种设置符合蛋白质互作网的生物学特性和拓扑结构特征: 具有相同功能的蛋白质之间通常存在互作, 这些蛋白质通常会形成一个蛋白质功能模块或功能通路 [7, 8].

令 $Y \in \mathbb{R}^{(C+N)\times C}$ 表示 C 个功能标记在混合图上 $C+N$ 个节点上的分布信息, $Y_c \in \mathbb{R}^{C+N}$ 表示功能标记 c 在这些节点上的分布信息, 其定义如下:

$$Y_c = \begin{bmatrix} \beta Y_c^{\mathrm{F}} \\ (1-\beta) Y_c^{\mathrm{P}} \end{bmatrix}, \tag{5}$$

其中, Y_c^{F} 的第 c 个元素为 1, 其他均为 0; $Y_c^{\mathrm{P}} \in \mathbb{R}^N$ 是 c 在 N 个蛋白质上的分布信息, 若 $S_{\mathrm{PF}}(i,c)=1$, 则 $Y_c^{\mathrm{P}}(i)=1/\sum_{i=1}^{N} S_{\mathrm{PF}}(i,c)$, 否则 $Y_c^{\mathrm{P}}(i)=0$. $\beta \in (0,1)$ 调整标记节点和蛋白质节点之间的权重. Y_c 可以看作一个随机游走子 c 在 $C+N$ 个节点上的初始分布.

互作的蛋白质之间很可能共享功能, 因此一个蛋白质的功能可借助与其互作蛋白质的功能进行预测 [25,27]. 在有向混合图定义的基础上, 重启动随机游走目标方程为 [28]

$$\boldsymbol{F}^{(t+1)}(i) = (1-\gamma)\sum_{j=1}^{C+N} W(i,j)\boldsymbol{F}^{(t)}(j) + \gamma \boldsymbol{Y}(i), \tag{6}$$

其中, $\boldsymbol{F}^{(t)}(i) \in \mathbb{R}^C$ 表示在第 t 次迭代中第 i 个节点上预测的功能标记分布概率向量, $\boldsymbol{Y}(i) \in \mathbb{R}^C$ 表示第 i 个节点初始的标记分布概率. $W(i,j)$ 表示节点 i 和 j 之间边的权重, 它的作用是传递蛋白质 (或功能标记) 节点 j 上的标记信息到节点 i, 进而预测节点 i 上的功能标记分布概率. $\gamma \in (0,1)$ 是随机游走重启动的概率. 令 $\boldsymbol{F}^{(0)} = \boldsymbol{Y}$, 可得

$$\boldsymbol{F}^{(t+1)} = ((1-\gamma)W)^{t+1}\boldsymbol{Y} + \gamma\sum_{k=0}^{t}((1-\gamma)W)^k\boldsymbol{Y}. \tag{7}$$

因 $\gamma \in (0,1)$, $0 \leqslant W(i,j) \leqslant 1$, 上式中的第一项在多次迭代后将趋向 0. 上式中的第二项 $\sum_{k=0}^{t}((1-\gamma)W)^k$ 是一个等比数列, 其极限为

$$\lim_{t\to\infty}\sum_{k=0}^{t}((1-\gamma)W)^k = (I-(1-\gamma)W)^{-1}, \tag{8}$$

其中, $I \in \mathbb{R}^{(C+N)\times(C+N)}$ 是一个单位矩阵. 在此基础上, 可得式 (7) 的显式解

$$\boldsymbol{F} = \gamma\boldsymbol{Y}(I-(1-\gamma)W)^{-1}, \tag{9}$$

从式 (9) 可以看出, 最终的预测结果 $\boldsymbol{F}$ 由 W 和 $\boldsymbol{Y}$ 共同决定, 而 $\boldsymbol{Y}$ 和 W 中的子图 W_{PP}, W_{PF}, W_{FP} 取决于已知的蛋白质功能标记和蛋白质之间互作信息, 均由数据库直接提供. 因此, 本文针对相关方法未充分利用标记之间层次结构关系的不足 [10,12,19,20], 在功能标记子网 W_{FF} 中引入功能标记节点间的层次结构关系, 利用这种关系提高蛋白质功能预测的精度.

2.2 功能标记间的转移概率

大量研究发现结合功能标记之间的相关性可以提高蛋白质功能预测的精度 [14,15,24]. 很多现有算法利用余弦相似度和 Jaccard Index 相似度等描述功能标记之间的关联关系, 这些方法通常仅关注了标记之间的水平关系, 并未考虑功能标记之间的层次结构关系 [10,12,19,20,27]. PILL 综合考虑了功能标记间的水平关系和层次关系, 但在预测的过程中并未利用标记之间的层次结构关系 [24]. 而蛋白质的新功能通常是已知功能的进一步细化, 这些细化的功能带有更多的生物学信息 [15,22,24]. 因此, 功能标记之间的层次结构关系在蛋白质的新功能预测中至关重要. 在结合已知的蛋白质功能标记分布信息和基因本体结构关系的基础上, 本文设计一种初始化功能标记子网 W_{FF} 的方法.

假定 c 为功能标记节点 d 的父节点, N_c 和 N_d 为 N 个蛋白质中分别标注功能 c 和 d 的数量, 由功能标记之间的 True Path Rule 规则可知 $N_c \geqslant N_d$. 令 $P(d|c)$ 表示已知一个蛋白质具有功能 c 时该蛋白质具有功能 d 的概率, $P(d|\bar{c})$ 表示已知一个蛋白质具有功能 $\bar{c}$ (除 c 以外的其他祖先节点) 时该

蛋白质具有功能 d 的概率. Yu 等统计研究发现 $P(d|c) > P(d|\bar{c})$, 因为当已知一个蛋白质标注有功能 c 时, 该蛋白质也标注了 c 的祖先节点对应的功能, 但当一个蛋白质标注有 c 的祖先节点对应的功能时, 该蛋白质不一定标注有节点 c 对应的功能 [24]. 由于蛋白质的功能标记信息不完整, 不宜直接利用 N_c 和 N_d 的比值定义 $P(d|c)$, 而通过以下方式计算,

$$P(d|c) = \begin{cases} \dfrac{N_d \bar{N}_d}{N_c^2}, & N_d > 0, \\ 1/h, & N_d = 0, \end{cases} \tag{10}$$

其中, $\bar{N}_d = N_c - N_d$, h 为 c 的直系孩子节点个数. 本文对功能标记的不平衡情况进行调和, 当 d 对应的蛋白质数量 N_d 较多且靠近 N_c 时, 此时 $\bar{N}_d$ 较小, 表明缺失功能标记 d 的蛋白质数量可能较少, 通过 $\bar{N}_d$ 可以将条件概率 $P(d|c)$ 调小; 当已知 N_d 较小而 N_c 较大时, 此时 $\bar{N}_d$ 较大, 表明缺失功能标记 d 的蛋白质数量可能较多, 通过 $\bar{N}_d$ 可以将 $P(d|c)$ 调大; 当 $N_d = 0$ 时, 表明可能还没有在这 N 个蛋白质上进行相关的生物学功能实验检测, 因此设置 $P(d|c)$ 为 $1/h$. 同理, 若 $N_c = 0$ 且 c 有 h 个直系孩子节点, 也设置 $P(d|c)$ 为 $1/h$. 由于蛋白质的新功能通常是已知功能的进一步细化, 这些细化的功能通常是已知功能节点的子孙节点. 因此, 以蛋白质当前已知的功能标记为随机游走子, 在功能标记所在的有向无环图上进行有方向的重启动随机游走, 能够预测蛋白质的新功能. 需指出的是 GO 中一个节点可能存在多个子节点, 也可能存在多个父节点, 可以同时是某一个节点的父节点和该节点的其他祖先节点, 图 1 和式 (10) 可以对上述节点间的关系进行描述. 如当 d 有另一个直系父亲节点 c' 时, 可以利用式 (10) 计算 $P(d|c')$. 当 d 既是 c 的直系孩子节点, 同时也是 c 的孙子节点时, 考虑 c 为 d 的父节点, 再基于式 (10) 计算 $P(d|c)$.

在 $P(d|c)$ 的基础上, 本文通过下式初始化 S_{FF},

$$S_{\mathrm{FF}}(d, c) = \frac{P(d|c)}{\sum_{d' \in \mathrm{ch}(c)} P(d'|c)}, \tag{11}$$

其中, $\mathrm{ch}(c)$ 为 c 所有直系孩子节点的集合. 上述归一化操作保证了一个节点向其直系孩子节点随机游走的概率总和为 1. $\mathrm{ch}(c)$ 不为 c 所有子孙节点的集合原因有两方面. 一方面是 c 包含了 c 的其他祖先节点的信息 [29], 也就是 $P(d|c) > P(d|\bar{c})$, 即当功能标记 d 缺失时, 基于 c 可以得到较其他祖先节点更准确的预测. 另一方面是在重启动随机游走过程中, d 的祖先节点 (除 c 以外) 对缺失功能标记 d 的预测会随着功能标记所在的层次结构图逐层向下推进, 其预测概率小于父节点 c 的预测概率. 由上述定义可以看出, 所提出的 dHG 方法不仅能弥补以往方法未能充分利用功能标记间层次结构关系的不足, 还可以借助有向混合图预测蛋白质的新功能.

3 实验

3.1 数据集

本文在 3 个不同的蛋白质互作网上检验提出的 dHG 性能, 并将其与其他相关算法进行比较. 这 3 个网络分别是 KroganPPI [30], ScPPI 和 HumanPPI, 其中前 2 个来源于酵母菌, 第三个来源于人类. KroganPPI 网络中蛋白质之间的互作可靠且权重是实数值. ScPPI 和 HumanPPI 均为不加权的网络, 若两个蛋白质之间存在物理交互, 则它们之间的权重为 1, 否则为 0. 这两个网络分别从 BioGrid[1)] 下

1) http://thebiogrid.org/.

表 1 实验数据集统计信息

Table 1 Dataset statistics

Dataset	N(number of proteins)	C(number of functions)	Avg±Std[a)]
KroganPPI	2670	564	15.21±10.51
ScPPI	5700	794	12.93±10.02
HumanPPI	19703	2649	44.68±58.16

a) Avg±Std 对应每个蛋白质的平均功能个数和对应的方差 (Avg±Std is the average number of functions per protein and the standard deviation)

载获得, 下载日期为 2015-01-15, 其中 ScPPI 由酵母菌互作网的最大连通子网构成. 本文从 GO 官方网站[2)]分别下载 (日期 2015-01-15) 酵母菌和人类蛋白质功能标记数据库, 并选用其中的生物过程 (biological process) 功能. 功能标记数据库中存在一些证据属性为 IEA (inferred from electronic annotation) 的功能标记, 为避免循环预测, 剔除这些功能标记. 功能标记数据库仅提供了蛋白质当前可知的最细粒度功能标记, 利用 True Path Rule 规则, 把这些功能标记的父母及祖先节点对应的功能也标注到该蛋白质上. 蛋白质的功能标记分布不均衡, 当某个功能标记仅存在于几个蛋白质上时该功能并不能在湿实验中测出, 对生物学研究意义并不大[31]. 因此本文过滤掉附属的蛋白质数量少于 10 的功能标记, 由于 HumanPPI 中的功能集合很大, 实验中过滤掉附属的蛋白质数量少于 30 的功能标记. 类似地, 去掉根节点 GO:0008150 (生物过程) 对应的功能标记. 最终使用的数据集统计情况如表 1 所示, 以 ScPPI 为例, 其包含 5700 个蛋白质, 这些蛋白质共计被 794 个不同的功能标记标注, 每一个蛋白质平均拥有 12.93 个功能标记. 从表 1 中可以发现, 这些数据集中蛋白质的功能标记并不均匀, 原因是一些蛋白质的功能标记信息较详细因而标注了较多的功能, 而另外一些蛋白质的功能标记信息较粗略, 仅被标注了少量的功能, 这也反映了蛋白质功能标记信息还待进一步补充完善.

3.2 对比算法及评价准则

为了分析比较 dHG 算法的性能, 本文把 dHG 与 TPR[14], MLR-GL[18], FCML[12], ProDM[20], PILL[24] 和国际大规模蛋白质功能预测评测组织推荐的基准方法 Naive[1] 进行比较. 其中, TPR 针对每个功能标记训练一个二分类器, 再利用标记间的层次结构关系调整和整合这些二分类器的结果, 从而预测蛋白质的功能. FCML 在 Green 函数中引入功能标记之间的关联关系预测蛋白质功能, 本质是一种基于多标记学习的蛋白质功能预测方法, 和 TPR 一样没有特别考虑蛋白质自身功能标记的不完整性. Naive 基于功能标记在蛋白质集合上的频率预测蛋白质的功能[1]. 其他对比算法已在相关工作中做了详细介绍. 这些对比算法的参数设置参照原始论文中提供的值或建议的范围进行设置. 本文通过在训练数据集上的 5 重交叉验证优化 dHG 的参数 α, β 和 γ (范围为 0.1 至 0.9, 步长为 0.1), 最终设置为 $\alpha = 0.1$, $\beta = 0.9$ 和 $\gamma = 0.1$.

目前已存在多种蛋白质功能预测评价准则[14,15], 这些准则从不同方面衡量预测算法的性能, 不同的预测算法在不同的评价准则下性能也不尽相同. 为了综合评价性能, 选用的评价度量分别为 *MacroF*1, *AvgROC*, *RankLoss*, *Coverage*, *AUC* 和 *Fmax*. 这 6 个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 4 个多标记学习度量的定义可参见文献 [11]. *AvgROC* 针对每类功能标记分别计算在不同阈值下的真阳性率和假阳性率并绘制对应的 ROC(receiver operating characteristic curve) 曲线, 然后计算每类标记对应 ROC 曲线下的面积, 最后计算这些曲线下面积的均值, 将该均值作为评

2) http://geneontology.org/.

价准则. 与 *AvgROC* 不同, 本文采用的 AUC 首先对每个蛋白质功能标记预测向量中的元素从大至小排序; 然后在控制每个蛋白质上预测的功能个数从 1 增至 C 的同时, 计算这些蛋白质上对应的真阳性率和假阳性率; 最后, 再计算真阳性率和假阳性率对应曲线下的面积, 以此面积大小评估多标记分类的性能. AUC 的具体定义可参见文献 [18]. $Fmax$ 是国际大规模蛋白质功能预测评测组织推荐的评价准则 [1], 其定义为

$$Fmax = \max_t \frac{2 \times p(t) \times r(t)}{p(t) + r(t)},$$

其中, $t \in [0,1]$ 是阈值, $m(t)$ 是阈值为 t 时单个蛋白质的预测功能标记概率向量中至少有一个元素大于 t 的蛋白质总数, $p(t) = \sum_{i=1}^{m(t)} p_i^t / m(t)$ 是在 $m(t)$ 个蛋白质上的平均预测准确率, $r(t) = \sum_{i=1}^{N} r_i(t)/N$ 是在 N 个蛋白质上的平均召回率, $p_i(t)$ 为第 i 个蛋白质上功能标记预测的准确率, $r_i(t)$ 为对应的召回率. *RankLoss* 和 *Coverage* 的值越小表示预测的精度越高. 为保持一致性, 实验中以 1-*RankLoss* 代替 *RankLoss*, *Coverage* 的值通常大于 1, 故不作类似处理. 这些准则从不同的方面评测蛋白质功能预测的性能, 一个算法很难在所有度量上面超过另一个算法.

3.3 预测已知部分功能蛋白质的新功能

本小节主要测试算法预测已知部分功能蛋白质新功能的性能. 由于 GO 和功能标记数据库的持续更新, 并没有现成的数据集可用于检验算法预测已知部分功能蛋白质新功能的性能. 本文假定当前蛋白质的功能标记信息是完整的, 针对每个蛋白质已知功能标记的层次结构图, 随机隐藏图中叶子节点对应的功能, 当图中非叶子节点的子节点对应的功能全部被隐藏后, 该节点也变为一个叶子节点, 对应的功能也可被隐藏, 这些隐藏的功能为蛋白质的新功能, 被用来评价算法补全新功能标记的性能. 为方便表示, 实验中用 m 表示一个蛋白质缺失标记的数量, 例如 $m = 3$ 表明该蛋白质有 3 个功能被隐藏 (或缺失). 如果一个蛋白质的功能标记数不足 m, 不会将其所有的标记隐藏, 而是保证至少有一个功能标记. 数据集中有一小部分蛋白质的功能标记完全未知, 为保持蛋白质互作网的结构, 这些蛋白质保留于实验中, 但不在这些蛋白质上隐藏标记和测试算法性能.

基于上述实验设置, 在实验中将 N 个蛋白质同时作为训练集和测试集, 针对每个蛋白质随机隐藏 m 个功能标记. 为了减少随机因素的影响, 在每个数据集上针对每个算法重复 15 次独立随机实验, 并记录每个对比算法在给定 m 下的 15 次平均结果. 表 2~4 分别给出这些算法在 KroganPPI, ScPPI 和 HumanPPI 上的实验结果, 结果以均值 ± 方差 (avg±std) 的形式表示. 表中 ↓ 表示值越小, 算法的性能越高. 表中加粗的结果表明其在配对检验 (95% 置信度) 中显著优于其他结果, 或与最优结果之间无显著性差异. HumanPPI 数据集较大, MLR-GL 和 Naive 的预测概率非常离散, 导致评价度量 *Fmax* 和 AUC 在该数据集上非常耗时, 因此表 4 中未包括 MLR-GL 和 Naive 在 HumanPPI 上的结果.

从 3 个表可以观察到 dHG 在绝大多数情况都能获得较其他对比算法更好的结果. 在 3 个数据集上的 54 种对比实验中, dHG 的结果一直优于 PILL, ProDM, MLR-GL, TPR 和 Naive. dHG 和 FCML 在两种对比实验中获得了相似的最优结果, 在其他对比实验中 dHG 均超过后者. 由于 ProDM 和 FCML 利用了功能标记之间的水平关系, PILL 综合利用了功能标记之间的层次结构关系和水平关系, 它们均获得了较基线方法 Naive 更好的结果, 这表明仅依赖于功能标记的频率信息无法准确预测蛋白质的新功能. 在这 3 个方法当中, PILL 获得了较 ProDM 和 FCML 更好的结果, 这说明功能标记之间的层次结构关系在蛋白质功能预测中至关重要. 但 PILL 在很多评价度量上的结果均被 dHG 显著性超过, 特别是评价度量 *MacroF*1. 主要原因是 *MacroF*1 受算法在细粒度功能上的性能影响较大, PILL 在预估蛋白质新功能的过程中并没有考虑功能标记之间的层次结构关系, 容易把一些粗粒度的

表 2 KroganPPI 上已知部分功能蛋白质的新功能预测结果

Table 2 Results of novel function prediction for partially annotated proteins on KroganPPI

Metric	m	dHG	PILL	ProDM	FCML	MLR-GL	TPR	Naive
	1	**95.16 ± 0.10**	93.27±0.03	83.28±0.10	92.79±0.15	17.53±0.11	26.70±0.14	2.97±0.05
*MacroF*1	3	**86.53 ± 0.15**	83.44±0.18	74.46±0.46	81.99±0.32	17.26±0.24	24.81±0.13	2.95±0.01
	5	**77.75 ± 0.25**	74.81±0.34	65.06±0.27	71.92±0.57	16.80±0.32	22.70±0.24	2.94±0.01
	1	**99.61 ± 0.03**	99.45±0.04	97.10±0.01	99.42±0.05	50.27±0.20	63.59±0.02	49.08±0.00
AvgROC	3	**98.38 ± 0.07**	98.12±0.07	95.65±0.04	98.12±0.07	50.83±0.20	62.25±0.06	49.08±0.00
	5	**96.72 ± 0.08**	96.43±0.06	93.78±0.09	96.48±0.07	51.31±0.51	60.93±0.15	49.08±0.00
	1	**99.74 ± 0.02**	99.48±0.04	96.28±0.04	98.39±0.05	32.39±0.27	38.39±0.05	84.04±0.02
1-*RankLoss*	3	**98.92 ± 0.06**	98.00±0.11	95.95±0.02	95.64±0.07	31.99±0.37	35.86±0.04	83.51±0.04
	5	**97.90 ± 0.10**	96.32±0.10	93.07±0.08	93.39±0.15	30.70±0.42	33.06±0.05	82.90±0.04
	1	**99.28 ± 0.00**	99.16±0.01	98.10±0.00	98.66±0.03	51.18±0.25	73.76±0.02	83.75±0.01
AUC	3	**98.79 ± 0.03**	98.39±0.04	97.36±0.02	97.09±0.04	51.33±0.35	72.87±0.04	83.52±0.01
	5	**98.05 ± 0.05**	97.33±0.04	96.39±0.03	95.42±0.07	51.16±0.19	71.81±0.03	83.19±0.02
	1	**95.01 ± 0.04**	93.56±0.04	82.62±0.01	90.41±1.30	21.68±3.65	39.15±0.04	39.03±0.00
Fmax	3	**86.54 ± 0.10**	85.06±0.04	73.87±0.06	79.87±3.83	24.40±2.43	37.94±0.08	39.03±0.00
	5	**78.61 ± 0.18**	76.94±0.02	66.28±0.05	63.40±14.86	22.98±1.40	36.44±0.08	38.76±0.15
	1	**27.49 ± 0.30**	36.95±1.14	60.41±0.24	77.57±2.67	532.21±1.22	426.35±0.81	353.70±0.82
Coverage↓	3	**58.61 ± 2.34**	79.70±2.74	100.65±1.44	157.88±3.04	540.90±1.33	454.46±0.79	379.70±0.35
	5	**95.28 ± 3.58**	125.69±2.60	143.28±3.01	214.93±3.11	543.44±1.61	479.97±1.26	384.96±2.14

功能标记预估为蛋白质新功能. 而 dHG 在蛋白质新功能预测的过程中考虑了功能标记之间的层次结构关系, 预估的新功能通常是该蛋白质已有功能标记节点的子节点, 子节点描述的功能粒度更细. 从评价度量 *Coverage* 的结果可知, 为了覆盖蛋白质的完整功能标记集合, dHG 在功能标记预测向量上的平均查找长度小于其他对比算法. 上述对比结果证明了本文提出的有向混合图在蛋白质新功能预测中的有效性.

MLR-GL 的结果显示其预测性能始终低于 dHG, PILL, ProDM 和 FCML, 可能的原因是 MLR-GL 采用的组稀疏和减少排序损失的策略并不适宜这种标记间存在层次结构关系的蛋白质功能预测问题. 另一个原因是 MLR-GL 基于已知部分标记的训练样本来预测完全未标注样本的标记, 并未关注对已知部分标记样本的新标记预测. TPR 利用了标记之间的层次结构关系, 但在部分评价度量上低于 Naive 方法, 原因主要是该方法隐式地假设用于训练的蛋白质的功能标记是完整的, 并基于这些实际上功能不完整标注的蛋白质, 为每种功能标记训练一个二分类器, 再利用标记间的层次结构关系调整和整合这些二分类器的结果. 因此 TPR 在针对二分类的评价度量 *AvgROC* 上获得了较 Naive 好的结果. 但 TPR 在评价度量 1-*RankLoss* 上总是低于 Naive, 原因是 1-*RankLoss* 偏好能够准确排序成对功能标记的分类器, 而这种偏好正好同 Naive 一致.

3.4 预测功能完全未标注蛋白质的新功能

本文还进行了另一组实验来检验 dHG 和其他对比方法预测完全未标注蛋白质新功能的性能. 实验中随机选择 80% 蛋白质作为训练集, 剩下 20% 作为测试集. 与前面的实验设置类似, 对训练集的每个蛋白质随机隐藏 $m = 3$ 个功能标记, 再基于这些被隐藏部分功能标记的蛋白质预测测试集中蛋白

表 3 ScPPI 上已知部分功能蛋白质的新功能预测结果

Table 3 Results of novel function prediction for partially annotated proteins on ScPPI

Metric	m	dHG	PILL	ProDM	FCML	MLR-GL	TPR	Naive
	1	**94.51 ± 0.04**	92.72±0.12	82.85±0.09	91.02±0.11	8.36±0.31	26.39±0.17	1.86±0.05
$MacroF1$	3	**85.10 ± 0.14**	81.25±0.18	72.68±0.14	80.20±0.15	8.54±0.15	24.26±0.19	1.87±0.07
	5	**75.69 ± 0.29**	70.68±0.34	62.37±0.21	69.03±0.26	8.72±0.32	21.70±0.19	1.86±0.03
	1	**99.66 ± 0.01**	99.44±0.03	98.40±0.03	99.56±0.02	50.84±0.39	69.65±0.04	43.05±0.00
$AvgROC$	3	**98.60 ± 0.04**	98.08±0.08	97.03±0.04	98.57 ± 0.04	50.90±0.45	67.46±0.13	43.05±0.00
	5	**97.03 ± 0.12**	96.07±0.10	95.12±0.15	97.09 ± 0.09	50.33±0.63	65.31±0.09	43.05±0.00
	1	**99.78 ± 0.01**	99.48±0.04	98.52±0.01	98.25±0.08	44.77±0.76	40.17±0.03	85.81±0.02
1-$RankLoss$	3	**98.73 ± 0.25**	97.84±0.08	96.79±0.10	95.20±0.05	43.26±0.43	37.70±0.08	84.86±0.02
	5	**97.79 ± 0.28**	96.14±0.12	95.60±0.03	93.01±0.08	39.50±0.86	35.17±0.04	84.37±0.02
	1	**99.50 ± 0.00**	99.37±0.01	99.37±0.00	98.83±0.03	46.95±0.84	77.72±0.01	85.41±0.00
AUC	3	**99.02 ± 0.06**	98.57±0.03	98.36±0.01	97.22±0.03	45.13±0.40	76.79±0.03	85.11±0.01
	5	**98.39 ± 0.07**	97.48±0.05	97.82±0.02	95.59±0.02	42.33±0.55	75.73±0.03	84.80±0.02
	1	**94.55 ± 0.04**	93.76±0.02	83.51±0.01	90.24±0.08	15.36±1.01	38.39±0.06	37.03±0.00
$Fmax$	3	**85.08 ± 0.10**	83.75±0.03	73.10±0.02	80.41±0.06	14.07±1.47	37.30±0.10	37.03±0.00
	5	**77.36 ± 0.09**	74.85±0.06	64.82±0.14	72.47±0.07	12.57±0.76	35.76±0.06	36.74±0.00
	1	**26.19 ± 0.43**	40.39±1.14	67.45±1.34	93.19±2.75	744.75±3.61	588.31±1.55	448.04±0.94
$Coverage\downarrow$	3	**67.85 ± 7.32**	100.04±1.37	105.04±0.96	209.63±3.11	758.26±2.59	626.16±1.20	480.93±0.98
	5	**113.13 ± 8.15**	166.36±3.02	163.94±3.03	291.82±1.37	767.56±3.31	652.24±1.31	497.70±1.06

质的新功能, 并把预测结果与测试集中蛋白质已有的功能标记作比较, 评价算法性能. 表 5~7 列出了每个对比算法在 KroganPPI, ScPPI 和 HumanPPI 上 15 次独立重复实验的平均结果. 与前一小节的原因类似, 表 7 中未包括 MLR-GL 和 Naive 在 HumanPPI 上的结果.

从这些结果可以看出 dHG 和 PILL 总能获得较其他对比算法更好的结果. dHG 和 PILL 在一些评价度量上能够获得相似的性能评估结果, 但在一些度量上具有不同的结果. 主要原因在于 PILL 预估的新功能标记粒度较粗, 带有一定的假阳性预估, 但这些粗粒度功能附属的蛋白质数目较多, 在蛋白质互作网上的标签传播过程中逐渐被其他蛋白质的同类功能标记影响, 从而降低了假阳性预估对预测性能的影响. 与其相比, dHG 预估的功能标记粒度较细, 这些细粒度功能标记附属的蛋白质较少, 在标签传播的过程中被大量的粗粒度功能影响, 细粒度功能的预测概率因此降低, 进而影响最终的预测精度. dHG 在 HumanPPI 上的结果绝大部分优于 PILL, 原因是该数据集中的蛋白质数量较多, 对式 (10) 中的预估相对较准确.

TPR 在预测已知部分功能蛋白质的新功能的实验中, 在评价度量 $AvgROC$ 和 $Fmax$ 上, 有时候获得较 ProDM 和 MLR-GL 更好的结果, 原因在于 TPR 在对完全未标注蛋白质的新功能预测过程中, 利用了功能标记间的层次结构关系调整了最终的预测概率, 而 ProWL 和 MLR-GL 均未显式利用功能标记间的层次结构关系. 这些实验结果表明在蛋白质功能预测中必须考虑功能标记间的层次结构关系, dHG 和 PILL 在蛋白质功能预测的不同阶段利用了这种层次结构关系, 因而获得较其他算法更好的结果.

表 4 **HumanPPI 上已知部分功能蛋白质的新功能预测结果**

Table 4 Results of novel function prediction for partially annotated proteins on HumanPPI

Metric	m	dHG	PILL	ProDM	FCML	TPR
	1	**97.42 ± 0.03**	97.12±0.01	86.39±0.02	68.09±0.01	15.66±0.00
*MacroF*1	3	**93.64 ± 0.05**	92.32±0.01	81.74±0.02	65.17±0.04	15.47±0.01
	5	**90.17 ± 0.05**	88.07±0.02	77.52±0.07	62.58±0.03	15.20±0.01
	1	**99.84 ± 0.00**	99.24±0.00	98.56±0.01	96.08±0.02	68.40±0.01
AvgROC	3	**99.41 ± 0.01**	98.03±0.01	97.89±0.01	93.08±0.01	67.70±0.01
	5	**98.87 ± 0.01**	96.99±0.01	97.21±0.02	91.58±0.01	67.03±0.01
	1	**99.87 ± 0.00**	99.68±0.00	99.37±0.00	84.79±0.01	44.75±0.00
1-*RankLoss*	3	**99.55 ± 0.00**	99.02±0.01	98.60±0.01	82.91±0.02	44.27±0.01
	5	**99.17 ± 0.01**	98.34±0.02	97.84±0.01	81.12±0.06	43.81±0.02
	1	**98.85 ± 0.00**	98.75±0.00	97.50±0.00	73.14±0.00	72.79±0.00
AUC	3	**98.66 ± 0.00**	98.41±0.00	97.13±0.00	72.80±0.03	72.65±0.00
	5	**98.45 ± 0.00**	98.05±0.00	96.74±0.02	72.33±0.07	72.50±0.01
	1	**98.22 ± 0.00**	97.71±0.02	86.36±0.01	72.90±0.03	34.86±0.01
Fmax	3	**94.73 ± 0.01**	94.17±0.01	82.83±0.00	69.31±0.67	34.71±0.01
	5	**91.26 ± 0.03**	90.70±0.02	79.40±0.01	66.09±0.37	34.55±0.02
	1	**191.20 ± 0.56**	335.56±0.54	425.05±1.61	1580.9±0.70	1958.9±0.46
Coverage ↓	3	**354.36 ± 3.67**	574.78±1.19	645.86±3.74	1813.4±1.88	2048.6±1.73
	5	**483.96 ± 3.31**	728.72±6.80	809.31±6.00	1946.8±0.20	2105.1±2.29

表 5 **KroganPPI 上完全未标注蛋白质的新功能预测结果**

Table 5 Results of novel function predictions for complete unlabeled proteins of KroganPPI

Metric	dHG	PILL	ProDM	FCML	MLR-GL	TPR	Naive
*MacroF*1	32.70±0.48	**34.31 ± 1.18**	24.79±1.09	27.51±1.10	3.14±0.20	21.83±1.62	2.90±0.23
AvgROC	67.92±0.45	**70.42 ± 1.82**	57.54±1.37	66.32±1.82	47.01±1.67	60.31±0.80	48.90±1.06
1-*RankLoss*	**88.04 ± 0.92**	76.18±1.16	73.33±0.53	57.64±1.16	25.03±1.02	32.56±1.78	83.30±0.59
AUC	**89.30 ± 0.64**	78.77±0.55	73.77±0.45	63.53±0.95	55.90±1.02	71.89±0.77	83.43±0.38
Fmax	42.12±1.04	**46.31 ± 0.87**	27.30±0.88	38.46±0.55	16.99±1.30	36.07±1.11	39.74±0.99
Coverage↓	227.58±9.75	**137.45 ± 2.97**	312.02±12.07	415.13±6.31	529.46±3.44	465.81±6.13	376.69±6.12

3.5 算法运行时间分析

为了统计分析各个对比算法的效率, 与 3.3 小节的实验设置类似, 本文统计了每个算法在不同数据集上的运行时间, 并将每个算法 5 次独立运行 (不包括评价度量) 的平均时间报告在表 8 中. 实验运行平台为: Linux OS 2.6.32, Intel Xeon E7-4820, 64GB RAM.

由表 8 可知 dHG 的运行时间总是远小于其他相关对比算法. MLR-GL 利用组稀疏和最小化排序损失预测蛋白质功能, 因其依赖于支持向量机预先对每个标记进行预测, 所以其时间耗费大于其他对比算法, 特别是在样本数和标记规模较大的数据集上. FCML 需要针对蛋白质互作网对应的关联矩阵计算本征值分解问题, 本征值分解的时间复杂度为 $O(N^3)$, 因此其时间耗费也比较大. ProDM 需要计算样本之间的水平关系和标记与蛋白质属性之间的依赖性, 所以其时间耗费较大, 但远小于 MLR-GL

表 6 ScPPI 上完全未标注蛋白质的新功能预测结果

Table 6 Results of novel function predictions for complete unlabeled proteins of ScPPI

Metric	dHG	PILL	ProDM	FCML	MLR-GL	TPR	Naive
*MacroF*1	**30.11 ± 0.68**	**30.25 ± 1.07**	21.24±0.92	26.01±1.50	2.16±0.43	21.73±1.24	1.92±0.07
AvgROC	**76.79 ± 1.06**	**76.08 ± 0.31**	64.43±0.23	76.05±0.64	45.55±0.83	64.57±1.28	48.87±1.43
1-*RankLoss*	**91.42 ± 0.28**	78.16±1.03	59.28±0.18	62.00±1.16	32.49±2.31	35.43±0.94	84.52±0.23
AUC	**92.31 ± 0.26**	81.55±0.59	65.64±0.26	69.40±0.97	36.82±1.46	75.96±0.27	84.71±0.25
Fmax	36.95±0.69	**43.89 ± 0.91**	32.93±0.82	37.92±1.11	8.98±1.38	36.26±0.86	36.68±0.73
Coverage↓	268.09±6.70	**196.70 ± 3.87**	482.00±0.53	552.65±3.50	767.96±2.99	638.79±8.55	487.69±5.91

表 7 HumanPPI 上完全未标注蛋白质的新功能预测结果

Table 7 Results of novel function predictions for complete unlabeled proteins of HumanPPI

Metric	dHG	PILL	ProDM	FCML	TPR
*MacroF*1	**18.68 ± 0.68**	16.03±0.56	12.24±1.18	9.20±1.50	12.85±0.34
AvgROC	**65.77 ± 1.06**	62.57±0.61	60.32±0.42	49.45±0.64	61.22±0.62
1-*RankLoss*	**86.61 ± 0.28**	70.10±0.17	70.03±0.73	53.58±1.16	39.86±0.98
AUC	**85.24 ± 0.26**	73.05±0.21	66.48±0.43	50.96±0.97	69.91±0.42
Fmax	26.95±0.69	**35.55 ± 0.11**	10.86±3.46	28.82±0.36	22.28±0.26
Coverage↓	**1441.5 ± 27.95**	1938.2±8.47	1805.8±14.63	2435.3±13.94	2153.2±13.14

表 8 对比算法运行时间 (s)

Table 8 Runtime cost of comparing methods (s)

Metric	dHG	PILL	ProDM	FCML	MLR-GL	TPR
KroganPPI	5.06	63.82	107.63	322.25	396.27	17.29
ScPPI	14.32	162.63	271.91	1086.24	1446.24	54.43
HumanPPI	278.24	2131.45	8186.77	61065.61	241483.19	16451.44

和 FCML. PILL 需要预先计算功能标记之间的水平关系和层次结构关系, 预估蛋白质的新功能, 再进行基于图的蛋白质功能预测, 所以其时间耗费较 dHG 大. TPR 先在蛋白质互作网上进行基于图的蛋白质功能预测, 再基于功能标记之间的层次结构关系调整这些预测, 所以其时间耗费也大于 dHG, 特别是在样本和标记规模较大的数据集上. 本文提出的 dHG 通过把功能标记层次结构关系图和蛋白质互作网整合成一个混合有向图, 再在该图上进行有向的重启动随机游走预测蛋白质的新功能, 该图对应的关联矩阵为稀疏矩阵, 因此其时间耗费较其他对比算法要小. 上述实验结果表明本文提出的 dHG 算法不仅能获得较其他相关算法更高的预测精度, 也获得更高的效率.

4 结束语

为蛋白质提供准确且详细的功能标记是生物信息学的核心问题之一, 针对先前的蛋白质功能预测方法未能较好处理功能标记间的层次结构关系和预先假定已标注蛋白质的功能标记信息完整的不足, 本文设计了一种有向混合图用于描述功能标记之间的层次结构关系及功能标记与蛋白质的关系, 提出了一种基于有向混合图的蛋白质新功能预测方法 dHG. 实验结果表明, dHG 能够获得较其他相关对

比算法更好的预测结果, 功能标记之间的层次关系能够提高蛋白质的新功能预测性能.

在将来的工作中, 将针对蛋白质功能标记的不平衡特性, 对同一个蛋白质上不同的功能标记设置不同的权重, 提高在细粒度功能标记上的精度. 此外, 蛋白质互作子网含有一定量的假阳性互作, 剔除这些噪声互作并融合其他类型的蛋白质数据提高混合图的质量, 也值得在将来的工作中进一步探索.

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

3 Legrain P, Aebersold R, Archakov A, et al. The human proteome project: current state and future direction. Mol Cell Proteomics, 2011, 10: 3309–3309

4 Pandey G, Kumar V, Steinbach M. Computational approach for protein function prediction. Technical Report TR06-028. Twin Cities: Department of Computer Science and Engineering, University of Minnesota, 2006

5 Lee D, Redfern O, Orengo C. Predicting protein function from sequence and structure. Nat Rev Mol Cell Biol, 2007, 8: 995–1005

6 Leslie C S, Eskin E, Cohen A, et al. Mismatch string kernels for discriminative protein classification. Bioinformatics, 2004, 20: 467–476

7 Spirin V, Mirny L A. Protein complexes and functional modules in molecular networks. Proc Natl Acad Sci, 2003, 100: 12123–12128

8 Cao M, Pietras C, Feng K, et al. New directions for diffusion-based network prediction of protein function: incorporating pathways with confidence. Bioinformatics, 2014, 30: i219–i227

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

10 Yu G X, Domeniconi C, Rangwala H, et al. Transductive multi-label ensemble classification for protein function prediction. In: Proceedings of 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, 2012. 1077–1085

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

12 Wang H, Huang H, Ding C. Function-function correlated multi-label protein function prediction over interaction networks. J Comput Biol, 2013, 20: 322–343

13 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 Bioinform, 2014, 11: 891–902

14 Valentini G. True Path Rule hierarchical ensembles for genome-wide gene function prediction. IEEE ACM Trans Comput Biol Bioinform, 2011, 8: 832–547

15 Valentini G. Hierarchical ensemble methods for protein function prediction. ISRN Bioinform, 2014: 901419

16 Dessimoz C, Skunca N, Thomas P D. CAFA and the open world of protein function predictions. Trends Genet, 2014, 29: 609–610

17 Sun Y Y, Zhang Y, Zhou Z H. Multi-label learning with weak label. In: Proceedings of the 25th AAAI Conference on Artificial Intelligence. USA: AAAI Press, 2011. 293–298

18 Bucak S S, Jin R, Jain A K. Multi-label learning with incomplete class assignments. In: Proceedings of the 24th International Conference on Computer Vision and Pattern Recognition, Columbus, 2011. 2801–2808

19 Yu G X, Rangwala H, Domeniconi C, et al. Protein function prediction with incomplete annotations. IEEE ACM Trans Comput Biol Bioinform, 2014, 11: 579–591

20 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

21 Li Y H, Guo Z, Ma W C, et al. Predicting specific functions of protein with partial functions by protein-protein interactions network. Chinese 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]

22 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: i529–i538
23 Dong B, Khatri P, Done A, et al. Predicting novel human gene ontology annotations using semantic analysis. IEEE ACM Trans Comput Biol Bioinform, 2010, 7: 91–99
24 Yu G X, Zhu H L, Domeniconi C. Predicting protein functions using incomplete hierarchical labels. BMC Bioinformatics, 2015, 16: 1–12
25 Sharan R, Ulitsky I, Shamir R. Network-based prediction of protein function. Mol Syst Biol, 2007, 3: 1–15
26 Zhang X F, Dai D Q. A framework for incorporating functional interrelationships into protein function prediction algorithms. IEEE ACM Trans Comput Biol Bioinform, 2012, 9: 740–753
27 Jiang J Q. Learning protein functions from bi-relational graph of proteins and function annotations. In: Proceedings of the 11th International Conference on Algorithms in Bioinformatics. Berlin: Springer, 2011. 128–138
28 Tong H H, Faloutsos C, Pan J Y. Random walk with restart: fast solutions and applications. Knowl Informa Syst, 2008, 14: 327–346
29 Teng Z X, Guo M Z, Liu X Y, et al. Measuring gene functional similarity based on group-wise comparison of GO terms. Bioinformatics, 2013, 29: 1424–1432
30 Krogan N J, Cagney G, Yu H Y, et al. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature, 2006, 440: 637–643
31 Myers C, Barrett D, Hibbs M, et al. Finding function: evaluation methods for functional genomic data. BMC Genomics, 2006, 7: 187

Novel protein-function prediction using a directed hybrid graph

Guangyuan FU[1], Guoxian YU[1,2*], Jun WANG[1] & Zili ZHANG[1]

1 *College of Computer and Information Science, Southwest University, Chongqing* 400715*, China*;
2 *Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun* 130012*, China*
*E-mail: gxyu@swu.edu.cn

Abstract Proteins carry out various important activities in an organism. Accurately annotating their functions can boost the advance of life-science research and application. High-throughput techniques generate such a large volume of proteomic and genomic data that it is beyond the capability of low-throughput wet-lab based techniques. Thus, computational model-based large-scale protein-function prediction is one of the key tasks in the post-genomic era. Current machine-learning based methods often focus on predicting the functions of completely unlabeled proteins. These methods ignore the incomplete labels of the labeled proteins, and hence have low accuracy. In this paper, we design a directed Hybrid Graph (dHG) based on the gene ontology hierarchy and the protein-protein interaction network. Next, we use the dHG to predict novel functions by performing a random walk with restart on it. The proposed dHG can predict not only new functions for partially labeled proteins, but also new functions for completely unlabeled proteins. Experimental results on proteins of yeast and humans show that dHG, across various evaluation metrics, achieves better results than other related methods, and costs less time than these methods.

Keywords protein function prediction, machine learning, directed hybrid graph, random walk, gene ontology

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

Jun WANG was born in 1983. She received a Ph.D. degree in Artificial Intelligence from 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. She is a member of the China Computer Federation (CCF).

Guoxian YU was born in 1985. He received a Ph.D. degree in Computer Science from South China University of Technology, Guangzhou in 2013. Currently, he is an Associate Professor at the College of Computer and Information Science, Southwest University, Chongqing. His research interests include data mining and bioinformatics. He serves as a PC member of several premier data-mining conferences. He is a member of the China Computer Federation (CCF) and IEEE DMTC.

Zili ZHANG was born in 1964. He received a Ph.D. degree in Computer Science from Deakin University, Australia in 2002. Currently, he is a Professor at the College of Computer and Information Science, Southwest University, Chongqing, and a Senior Lecturer at Deakin University, Australia. His research interests include bio-inspired artificial intelligence, agent-based computing, big data analysis, and agent-data mining interaction and integration. He is a fellow of the China Computer Federation (CCF).