多关系蛋白质网络构建及其应用研究

胡 赛 [1] 熊慧军 [1] 李学勇 [1] 赵碧海 [1] 倪问尹 [1] 杨品红 [2] 刘 臻 [3]

摘 要 考虑到不同类型的相互作用对于功能预测的作用各不相同, 结合蛋白质相互作用网络和蛋白质结构域信息构建多关系蛋白质网络, 并为每种类型的相互作用赋予不同的遍历优先级. 基于多关系网络, 提出一种蛋白质功能预测方法 FPM (Functions prediction based on multi-relational networks). 对于未注释的蛋白质, 算法遍历与该蛋白质相连的, 具有最高优先级的所有相互作用, 形成一个候选邻居节点集合. 最后根据邻居节点集合形成预测的功能集合, 并为每一项功能评分、排序. 与其他算法对比结果表明, FPM 方法的性能优于其他的功能预测方法.

关键词 多关系网络, 蛋白质功能, 蛋白质相互作用, 优先级, 结构域

引用格式 胡赛, 熊慧军, 李学勇, 赵碧海, 倪问尹, 杨品红, 刘臻. 多关系蛋白质网络构建及其应用研究. 自动化学报, 2015, **41**(12): 2155−2163

DOI 10.16383/j.aas.2015.c150187

Construction of Multi-relation Protein Networks and Its Application

HU Sai[1] XIONG Hui-Jun[1] LI Xue-Yong[1] ZHAO Bi-Hai[1] NI Wen-Yin[1] YANG Pin-Hong[2] LIU Zhen[3]

Abstract Considering the different influences of different types of interactions in protein function prediction, we construct a multi-relation interaction network by integrating protein-protein interaction (PPI) network and proteins′ domain information and provide different priorities for different types of interactions. We propose a protein function prediction method, named functions prediction based on multi-relation networks (FPM) to infer protein functions. Given an unannotated protein, all interactions with highest priority associated with the protein are visited to form a candidate neighbor nodes set. Finally, we sort all functions of the neighbor nodes and select part of them to predict an unknown protein. Comparson with other methods show that our method outperforms other function prediction methods.

Key words Multi-relation network, protein functions, protein-protein interaction (PPI), priority, domain

Citation Hu Sai, Xiong Hui-Jun, Li Xue-Yong, Zhao Bi-Hai, Ni Wen-Yin, Yang Pin-Hong, Liu Zhen. Construction of multi-relation protein networks and its application. *Acta Automatica Sinica*, 2015, **41**(12): 2155−2163

蛋白质功能的准确注释是从分子水平理解生命活动的关键, 而且对生物医学和制药也有很大的影响. 由于内在难度和高昂费用, 实验功能注释方法已经无法满足日益增长的序列和结构数据. 因此, 蛋白质功能的计算方法注释已经成为计算生物学和分子生物学的前沿问题. 高通量的实验方法已经产生了大量高质量、大规模的蛋白质相互作用数据. 同时, 作为蛋白质的功能描述/定义的功能注释计划也得到了很好的发展. 这使得通过蛋白质相互作用数据预测未注释的蛋白质功能成为可能, 并且可行.

在过去的十年中, 从 PPI (Protein-protein interaction) 数据中预测蛋白质功能的计算方法发展迅速. Schwikowski 等[1] 提出邻居计数法, 以邻居中频繁出现的功能为功能未知的蛋白质注释. 但是这种方法忽视了不同功能注释的背景频率. Hishigaki 等[2] 改善了邻居计数法, 用卡方统计代替次数统计作为蛋白质功能评分方法. 除了直接邻居, Chua 等[3] 通过为直接邻居和间接邻居赋予不同的权值, 推断蛋白质的功能信息. 现有的方法一般近似以网络中的最短距离度量两个蛋白质间的距离, 但是大多数蛋白质彼此接近. 因此, Nabieva 等[4] 根据图的扩散性引入 DSD (Diffusion state distance) 距离, 旨在捕捉 PPI 网络中功能注释传输接近细粒度的区

收稿日期 2015-04-10 录用日期 2015-09-23
Manuscript received April 10, 2015; accepted September 23, 2015
国家自然科学基金 (11501054), 湖南省自然科学基金 (13JJ4106, 14JJ3138), 湖南省教育厅项目 (10C0408, 15C0124), 湖南省科技计划项目 (2010FJ3044, 2015GK3072), 水产高效健康生产湖南省协同创新中心资助
Supported by National Natural Science Foundation of China (11501054), Natural Science Foundation of Hunan Province (13JJ4106, 14JJ3138), National Scientific Research Foundation of Hunan Province (10C0408, 15C0124), Science and Technology Plan Project of Hunan Province (2010FJ3044, 2015GK3072), and Collaborative Innovation Center for Efficient and Health Production of Fisheries in Hunan Province
本文责任编委 张学工
Recommended by Associate Editor ZHANG Xue-Gong
1. 长沙学院数学与计算机科学系 长沙 410022 2. 湖南文理学院生命科学学院 常德 415000 3. 长沙学院生物与环境工程系 长沙 410022
1. Department of Mathematics and Computer Science, Changsha University, Changsha 410022 2. School of Life Science, Hunan University of Arts and Science, Changde 415000 3. Department of Biological and Environmental Engineering, Changsha University, Changsha 410022

别. Chi 等[5] 提出了一种新的基于迭代的蛋白质功能预测方法, 这种方法结合了蛋白质功能局部和全局的语义影响. 一些基于网络的方法将 PPI 网络分割为多个功能模块, 在同一模块中的蛋白质被分配相同的功能. Lee 等[6] 改变了现有的模块解决方案, 提出一种更好的使用模块信息预测蛋白质功能方法.

考虑到 PPI 网络中的高噪声和已注释的蛋白质数量不足, 一些研究者尝试通过结合其他生物信息从而提高预测性能. Cozzetto 等[7] 提出一种整合方法迎接功能注释的挑战, 他将各种各样的生物信息结合成一个单一的框架, 涵盖了序列、基因表达和蛋白质相互作用数据. 在 FALCON 方法中, Kourmpetis 等[8] 提出一种新的离散优化算法, 基因本体被模型化为一个离散的贝叶斯网络. Zhang 等[9] 提出一种新的蛋白质功能预测方法, 该方法结合了蛋白质结构域组成信息和 PPI 网络. 最近, 一种通过整合蛋白质相互作用网络和蛋白质结构域信息的 DCS (Domain combination similarity)[10] 方法被用于蛋白质功能预测. 不同于 Zhang 的方法, DCS 改变了结构域相似性的计算方法, 结合了蛋白质自身和邻居的结构域组成. Liang 等[11] 采用结构域信息建立蛋白质重叠网络的人工网络模型 PON (Protein overlap network). 蛋白质的功能通过计算 GO (Gene ontology) 在 PON 网络中与其共结构域的邻居中出现的频率得到.

尽管从 PPI 网络中预测蛋白质功能或整合 PPI 网络和其他生物信息进行蛋白质功能预测的方法取得了很大进步, 但是基于计算方法的蛋白质功能预测依然面临很大的挑战. 其中一个重要的原因是, 从高通量生物实验中获得的 PPI 数据已发现含有显著比例的假阳性和假阴性. 假阳性的存在降低了预测的准确率, 而假阴性则使某些功能未知的蛋白质丧失被注释的机会. 一些方法试图通过整合 PPI 网络和其他生物信息重新构造更加可靠的网络, 从而减小随机噪声对预测性能造成的影响. 然而, 在生物学中, 蛋白质之间存在复杂和多样的联系. 比较相互作用数据是困难的, 因为它们常常源于不同条件, 扮演不同的角色. 现有的蛋白质功能预测方法忽略了这一点, 将不同情形下产生的相互作用合并, 从而导致相互作用网络中出现人为噪声.

本文利用 PPI 网络和多元的生物信息, 构造多关系蛋白质网络, 网络中的两个蛋白质之间可能存在一种以上的相互作用. 利用网络拓扑特性和蛋白质共结构特性, 我们对构造的多关系蛋白质网络加权. 基于加权多关系网络, 提出一种蛋白质功能预测方法 FPM (Function prediction based on multi-relational networks). 实验对比结果充分验证了多关系网络的有效性.

1 多关系网络构建

随着人类社会的日益网络化, 复杂网络已成为新的研究热点, 已渗透到数理学科、生命科学等领域. 复杂网络具有小世界特性和无标度特性. 小世界特性主要表现为具有高聚集系数和低特征路径长度. 无标度特性主要表现为成长性和优先连接, 节点度服从幂律分布. 现实生活中, 复杂网络中节点之间的联系呈现多样性的特点. 以人际关系网为例, 人与人之间可以通过邮件、电话、MSN 等多种方式联系, 这样人际关系网就构成一个多关系的复杂网络. 在复杂生物网络, 如 PPI 网络中, 蛋白质之间也存在着多样的联系, 如共基因表达、共功能注释等. 多关系网络具有不同于单一性质的简单网络的复杂特性.

定义 1 (多关系网络)[12]**.** 网络 M 由顶点的非空有限集合 V 和边的集合 E 组成, 其中, M 中每两个顶点通过有限数量 (可能为零) 的边相连. 如果两个或更多的边连接同一对 (不同) 的顶点, 那么这些边称为平行边. 在多关系网络中, 一条边被允许连接一个顶点到自身, 这样的边称为环.

图 1 是几个典型的多关系网络的实例, 网络 M_3 中有两个节点都有环. 在蛋白质相互作用网络中, 我们不考虑蛋白质的自相互作用. 因此本文中构建的多关系网络将不会出现环.

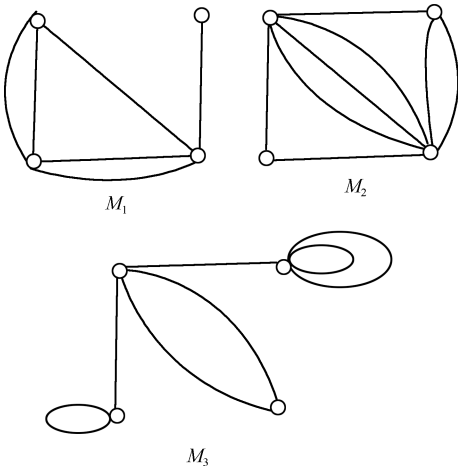


图 1 多关系网络实例

Fig. 1 An example of multi-relational network

为了提高预测的准确率, 本文整合了 PPI 网络和蛋白质的结构域 (Domain) 信息. 研究发现, 蛋白质通常由一个或多个具有独立功能的结构域 (Domain) 组成. 不同的蛋白质间的结构域组合可能存在差异, 识别这些结构域对揭示蛋白质功能具有重要意义. Zhang 方法和 DCS 方法已经成功地将 Domain 信息应用于蛋白质功能预测, 并有效地提高了预测的准确率. 两种方法都是通过计算蛋白质之间的结构域上下文相似性实现功能预测. 不同之

处在于, DCS 方法在 Zhang 方法的基础上, 扩充了邻居的结构域信息, 从而改变了相似性计算公式.

本文构造的多关系网络中, 两个蛋白质之间最多具有三种不同性质的联系: 共享结构域 (Domain)、共享邻居和在 PPI 网络中存在物理的相互作用. 多关系网络建立在原有的 PPI 网络基础之上, 如果两个蛋白质至少包含一个共同的结构域, 则为两个蛋白质之间添加一组相互作用; 如果两个蛋白质在 PPI 网络中有相互作用, 并且有相同的邻居 (在原始的 PPI 网络中存在相互作用的蛋白质), 则在这两个蛋白质之间也增加一组相互作用. 两个蛋白质即便共享邻居, 但是在 PPI 网络中不存在相互作用, 这两个蛋白质之间不会建立共享邻居的相互作用. 采取这种策略的目的在于防止 Hub 节点生成大量的新的相互作用, 降低人为假阳性的产生. 图 2 是本文将要构造的多关系网络示意图.

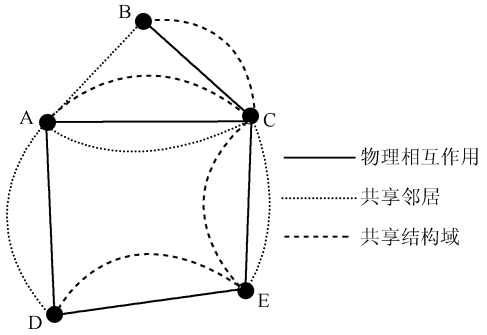


图 2　构造的多关系网络实例

Fig. 2　Reconstructed multi-relational network

为了降低噪声造成的负面影响, 提高预测算法的准确率, 现有的蛋白质功能预测算法开始整合 PPI 网络和多元的生物信息, 使得蛋白质之间可能存在多种联系. 在处理这样复杂的生物网络时, 现有算法都采取最简单方式, 把不同性质的边合并为一条边. 图 3 给出了现有方法整合 PPI 网络和其他生物信息重构网络的实例.

图 3 (a) 是原始的 PPI 网络, 图 3 (b) 是功能共注释网络, 图 3 (c) 是基于基因表达信息的共表达网络 (皮尔逊相关系数超过 0.8), 图 3 (d) 是基于 PPI 网络和生物信息重构的网络. 蛋白质 YPR018W 和 YJL115W 在 PPI 网络中之间不存在物理的相互作用, 由于它们共表达, 因此重构的网络中, 这两个蛋白质之间有边相连. 蛋白质 YDR181C 和 YJL115W 在 PPI 网络中存在相互作用, 共注释、共表达. 在重构的网络中两者之间三种不同性质的边进行了合并. 现实生活中, 复杂网络中两个节点间的多条边不能简单地合并, 因为它们源于不同的条件, 重要性也各不相同. 如图 3 的实例, 重构后的网络图 3 (d) 中, 相互作用 (YPR018W, YJL115W) 与相互作用 (YDR181C, YJL115W) 具有相同的地位和可靠性 (两条边的边聚集系数均为 1), 但是通过分析 PPI 网络图 3 (a)、共注释网络 v(b) 和共表达网络图 3 (c) 可以看出, 相互作用 (YDR181C, YJL115W) 应具有更强的可靠性, 因为蛋白质 YDR181C 和 YJL115W 在三个网络中都存在相互作用. 重构后的网络显然隐藏了这些信息. 具体到本文所涉及的功能预测的问题, 简单地合并不同类型的相互作用将会增加预测功能中假阳性功能的比例, 影响预测的性能. 因此, 在解决功能预测的问题时, 针对不同类型的相互作用采取不同的处理方式非常有必要, 而多关系蛋白质网络构建是解决这一问题的一个非常有效的途径.

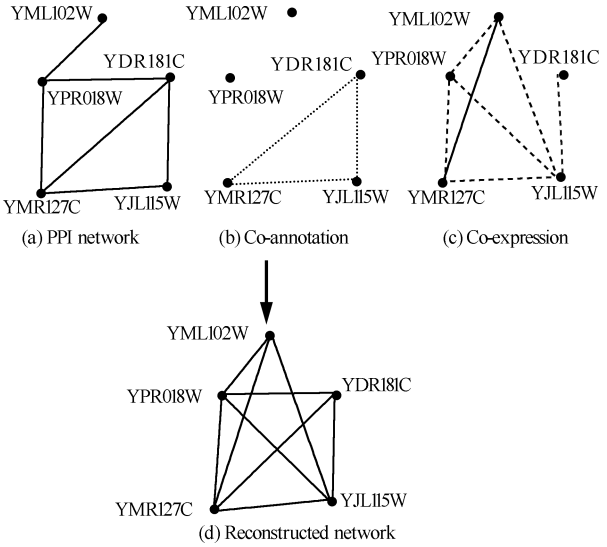


图 3　现有算法的网络重构实例

Fig. 3　Reconstructed network using current algorithms

构建的多关系蛋白质网络需要融合为简单网络才能进行功能预测. 在处理融合网络时, 一个最简单的方式是把不同性质的边合并为一条边进行处理. 正如前面的分析所言, 这种方式不适用于蛋白质功能预测. 本文考虑到蛋白质之间不同性质的相互作用对于功能预测的重要性各不相同, 采取对不同性质的边授予不同的优先级的处理方式. 优先级高的相互作用优先得到处理, 优先级相同的相互作用同一批次处理, 不同优先级的相互作用不能同时得到处理. 通过这种方式, 将多关系的复杂网络分解为多个具有不同优先级的简单网络. 换句话说, 本文首先将多关系网络分解为三个独立的单关系的简单网络, 并为每一个网络赋予一个优先级. 针对某一待预测功能的蛋白质, 本文设计的算法会在优先级高的网络遍历邻居节点, 只有在遍历不到邻居节点的情况下, 才会继续遍历优先级次之的网络. 这种遍历策略的好处在于: 1) 对于某一待预测的蛋白质, 算法可

能会遍历多个网络, 从而增大了其找到具有共同功能的邻居的可能性; 2) 只在同一类型 (优先级) 网络遍历, 能够降低假阳性功能的数量, 提高整体预测性能.

如何为三种不同性质的相互作用设置优先级对蛋白质功能预测的准确率有重要影响. 本文通过统计分析各自对于功能预测准确率的影响确定优先级的顺序, 图 4 是统计结果. 本次统计基于后续实验所使用的酵母物种的 DIP (Database of interacting proteins) 数据集[13]. 本次统计中, 我们将分别在三个单关系网络中运行 NC (Neighbor counting) 方法预测蛋白质功能, 通过计算三个网络各自的平均准确率进行对比分析. 其中共享结构域网络包含 2 972 个蛋白质和 23 749 组相互作用, 共享邻居网络包含 2 388 个蛋白质和 13 871 组相互作用, 物理相互作用网络包含 5 093 个蛋白质和 24 743 组相互作用. 所谓准确率是指, 某一蛋白质的已知功能中, 被正确预测的功能所占的比例.

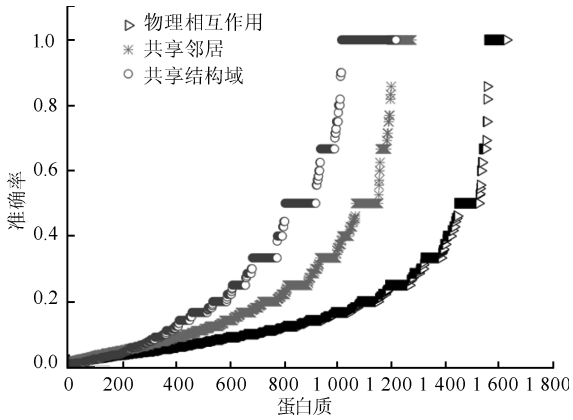


图 4 不同类型相互作用对预测的影响

Fig. 4 The impact of different interactions

统计结果中, 共享邻居是指在 PPI 网络中存在相互作用的两个蛋白质间具有共同的邻居, 即边聚集系数 (Edge clustering coefficient, ECC) 大于零. 从上述结果可以看出, 三种类型的准确率的区别非常明显. 为此, 我们为共享结构域的相互作用授予最高优先级, 共享邻居方式的优先级次之, PPI 网络中没有共同邻居的物理相互作用的优先级最低.

与此同时, 我们前期的蛋白质复合物预测和关键蛋白质识别的研究结果表明, 基于加权网络的预测算法的性能要优于基于非加权网络的同种算法的性能. 原因非常简单: 权值代表了相互作用的可靠性, 加权网络比非加权网络能更加代表真实的相互作用网络. 因此, 本文将进一步对多关系网络中的相互作用加权.

蛋白质分子中, 肽链主键形成的二级结构相互靠近可能形成二级结构的聚集体, 称为蛋白质结构域. 对于较大的蛋白质分子, 其多肽链往往包含一个以上的结构域, 而对于小分子蛋白质, 则是单结构域的. 同一种结构域可能在多个蛋白质中出现, 分子进化过程中常常通过结构域重组形成新的功能. 蛋白质结构域是重要的蛋白质功能标志物, 与蛋白质功能存在一定的联系. 因此我们认为, 两个蛋白质共享结构域的概率越大, 他们共享功能的可能性越高, 两者间相互作用的权值也就越高. 针对共享结构域的相互作用, 我们利用共享结构域的数量与各自结构域数量的比值描述该类型的相互作用的权值. 加权计算方式如下所示:

$$P(v_i, v_j) = \frac{|D_i \cap D_j|^2}{|D_i||D_j|} \tag{1}$$

其中, $P(v_i, v_j)$ 表示蛋白质 v_i 和 v_j 共享结构域的概率, D_i 和 D_j 分别表示蛋白质 v_i 和 v_j 的结构域集合, $D_i \cap D_j$ 表示两个蛋白质共同的结构域构集合. 若 D_i 或 D_j 等于 0, 则权值简单地设置为 0. 针对共享邻居的相互作用, 我们采用类似的、以共同邻居数量刻画相互作用的可靠性, 其权值计算公式如下:

$$P(v_i, v_j) = \frac{|N_i \cap N_j|^2}{(|N_i| - 1)(|N_j| - 1|)} \tag{2}$$

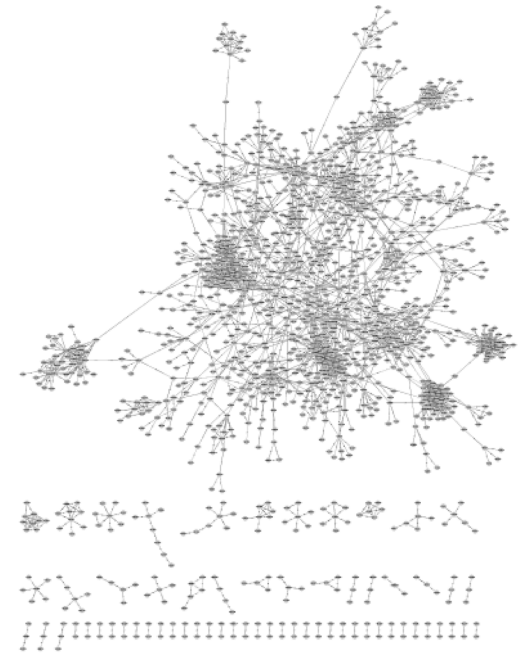
$P(v_i, v_j)$ 表示蛋白质 v_i 和 v_j 共享邻居的概率, 该概率计算公式是边聚集系数 (ECC) 的变换. 其中, N_i 和 N_j 分别表示 v_i 和 v_j 的邻居节点集合. 至于第三种情形, 没有共同邻居的物理相互作用, 本文只是简单地将这种类型的相互作用的权值设置为 1.

本文采用虚拟化直观地展示了构造的多关系的酵母相互作用网络, 如图 5 所示.

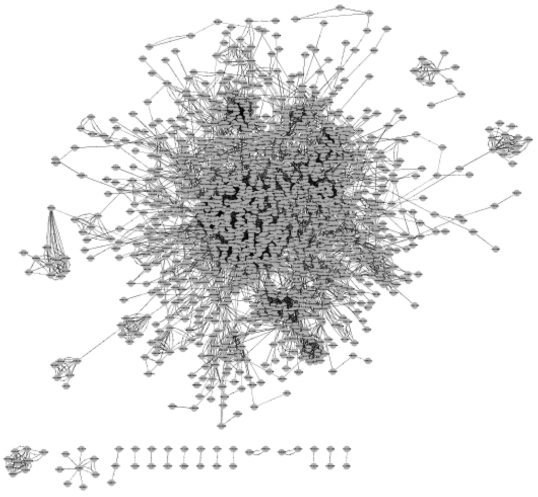
图 5 (a) 是输入的酵母相互作用网络, 由 1 412 个蛋白质和 3 004 组相互作用组成; 图 5 (b) 是构造的多关系网络, 构建的多关系相互作用包括 1 412 个蛋白质和 8 093 组相互作用. 对比图 5 (a) 和图 5 (b) 不难看出, 构建的多关系网络明显比原始 PPI 网络稠密, 这是因为通过共结构域和共享邻居关系新增了一部分相互作用. 图 5 (c) 是多关系网络的一个子网的放大展示, 该图清晰地显示节点间的多元关系.

2 FPM 算法

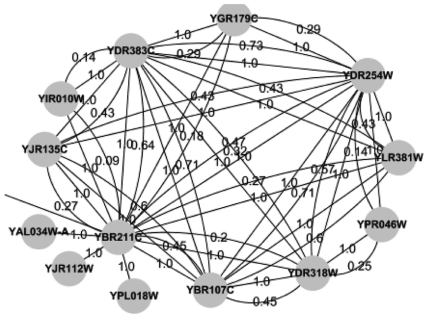
为了验证构建的多关系蛋白质网络的性能, 本文设计了一种基于多关系网络的功能预测算法 FPM. FPM 算法设计的目的在于: 1) 提高预测的敏感性 (Sensitivity), 能够为更多的蛋白质预测功能; 2) 提高预测的特异性 (Specificity), 减少预测的功能中假阳性功能的数量. 算法的核心思想是, 遍历某一蛋白质的邻居时, 只在优先级相同的边上游走



(a) 原始的PPI网络
(a) Original PPI network



(b) 构造的多关系蛋白质网络
(b) Constructed multi-relational network



(c) 多关系子网实例
(c) Subnetwok of multi-relational network

图 5　构造的多关系蛋白质网络

Fig. 5　Constructed multi-relational interaction networks

访问, 而不是把与该蛋白质存在相互作用的所有蛋白质作为候选蛋白质. 并且在遍历时, 永远从优先级最高的边开始访问. 算法分为三个步骤:

1) 根据 PPI 网络的拓扑属性和蛋白质的结构域信息, 构造多关系的相互作用网络.

2) 对于功能未知的某一蛋白质 u, 算法先在分解得到的优先级最高的单关系网络中遍历功能已知的邻居节点. 如果没有发现邻居节点, 则进入优先级次之的网络中继续遍历, 直至形成候选邻居节点集合. 如果三个网络遍历完毕都不能形成邻居节点集合, 则表示该蛋白质无法通过算法进行功能注释.

3) 根据候选邻居节点集合, 产生候选功能集合, 并为功能评分、排序, 选取排名靠前的一定数量功能作为未知功能的蛋白质功能. 如何从候选功能集合中选取功能注释蛋白质成为影响功能预测算法性能的关键之一. 针对功能选取, 现有算法主要采取两种方式实现: 1) 以 Zhang 方法和 DCS 方法为代表, 计算蛋白质之间的相似性, 然后将相似性值最高的蛋白质的所有功能赋予未知功能的蛋白质; 2) 以 NC 方法为代表, 由邻居节点的所有功能形成候选功能集, 然后根据一定策略对集合中的功能评分、排序. 本文对已注释的蛋白质之间的功能重叠情况进行了统计分析, 以便确定功能选取方案. 统计结果表明, 有超过一半的蛋白质对的功能重叠率介于 0.4 和 0.6 之间, 重叠率超过 0.6 的蛋白质对仅占 11.99 %. 基于统计分析结果, FPM 算法采取第二种功能选取方式, 即选择排名靠前的若干项功能注释蛋白质. 接下来就需要确定候选功能的评分函数.

设 $P = \{p_1, p_2, \cdots, p_n\}$ 是功能未知的蛋白质 u 的候选邻居节点集合, $F = \{f_1, f_2, \cdots, f_m\}$ 是由 P 中所有蛋白质的功能形成的候选功能集合. 对于 F 中某一候选功能 f_j, 其排名得分计算方法如下所示:

$$Score(f_j) = \sum_{i=1}^{n} P(u, p_i) t_{ij}, \quad t \in [1, n] \qquad (3)$$

其中, $P(u, p_i)$ 表示待预测功能的蛋白质 u 与蛋白质 p_i 间相互作用的权值; 若蛋白质 p_i 包含功能 f_j, 则 $t_{ij} = 1$; 否则, $t_{ij} = 0$. 算法将候选功能按照式 (3) 计算得到的分值降序排列, 然后从候选功能列表中选取 S 项功能作为预测功能, 对蛋白质 u 进行功能注释. S 的值将根据网络的拓扑结构信息[13] 确定, 具体来讲就是根据 u 的邻居节点的功能数量确定. 设 $Num = \{N_1, N_2, \cdots, N_n\}$ 是候选邻居节点集合 P 中所有节点的功能数量构成的集合, 并且集合中的数值已经按照从小到大的顺序排列. S 取值

Num 集合的中位数, 具体计算方法如下所示:

$$S=\begin{cases} N_{\lfloor \frac{n}{2} \rfloor+1}, & N \bmod 2=1 \\ \dfrac{N_{\frac{n}{2}}+N_{\frac{n}{2}+1}}{2}, & N \bmod 2=0 \end{cases} \tag{4}$$

3 实验结果与分析

酿酒酵母相互作用网络作为一种标准数据, 被广泛应用于基于网络的蛋白质功能预测算法, 源于它的蛋白质相互作用数据和功能注释数据是比较完善和可靠的. 本文实验也将基于酿酒酵母相互作用数据. PPI 网络数据源于 DIP[14] 数据库的 2010 年 10 月 10 日的版本, 去除自相互作用和重复的相互作用后, 该 DIP 数据包含 5 093 个蛋白质和 24 743 组相互作用. 为充分验证算法的有效性, 我们还基于 Krogan[15] 数据集简要对比分析了 FPM 算法和其他竞争算法的性能. Krogan 数据集由 3 672 个蛋白质和 14 317 组相互作用组成. 用于验证算法性能的蛋白质功能注释数据是从 GO 官方网站下载的最新版本[16]. 为了避免太特殊或者太一般化, 仅仅使用那些至少注释了 10 个或者最多注释了 200 个蛋白质的 GO Term 来进行实验验证, 处理后的 GO Term 数量为 267 个. 此外, GO 注释数据中, 本文利用 Uniprot[17] 网站将蛋白质格式从 UniProtKB[18] 转换为 Ensemble Genomes Protein[19], 以便与 PPI 网络中蛋白质的格式匹配. 本文所用的结构域 (Domain) 数据从 Pfam 数据库下载得到[20], 包含 1 107 个不同的结构域, 涉及 PPI 网络中的 3 056 个蛋白质.

为评估本文提出的 FPM 方法的性能, 我们从多方面对比分析 FPM 方法与其他的经典功能预测方法: Zhang[9]、DCS[10]、NC[1]、PON[11] 的预测结果.

交叉验证是广泛采用的, 被用于评估蛋白质功能预测算法性能的方法. 该方法人为地将网络中部分蛋白质的功能注释去除, 并将这部分蛋白质作为测试集, 而剩余的具有功能注释的蛋白质作为训练集. 利用训练集中的功能信息来预测测试集中蛋白质的功能, 并将预测得到结果与真实的功能进行比较, 从而确定功能预测算法的性能. 本次实验中, 每一轮仅有一个功能未知的蛋白质在测试集, 而剩余的所有蛋白质在训练集中. 依次对所有的蛋白质进行测试, 所有蛋白质预测结果的平均值将用于评估预测算法的性能. 每一个在测试集中的蛋白质根据在训练集中的蛋白质的功能, 被赋予具有各个功能的可能性. 然后根据预测算法设定的功能选取策略选取一定数量的功能. 预测结果的性能取决于预测的功能与实际功能之间的匹配程度. 有两种广泛使用的评价指标来衡量预测结果的好坏. 其中一个是特异性 (Specificity, Sp), 它衡量的是在预测的功能中有多少比例的功能与已知的功能之间能够匹配. 另外一个是敏感性 (Sensitivity, Sn), 它衡量的是已知的蛋白质功能有多少能够与被预测的功能所匹配. Specificity 和 Sensitivity 的形式化定义如下:

$$\mathrm{Sp}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{5}$$

$$\mathrm{Sn}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{6}$$

其中, TP (True positive) 表示的是预测的蛋白质功能有多少能被已知的功能所匹配. FP (False positive) 表示的是预测的功能有多少不能被已知的功能所匹配. FN (False negative) 表示的有多少已知的蛋白质功能没有被预测的功能所匹配. 在预测过程中, 选取的功能数量越多, 待预测蛋白质的功能匹配数量越高, 即 Sp 越高, 但同时会导致 Sn 降低. F_measure 则能较好地反映算法的综合性能, 它是 Sp 和 Sn 的调和平均值, 计算公式如下所示:

$$\mathrm{F\_measure}=\frac{2\times \mathrm{Sp}\times \mathrm{Sn}}{\mathrm{Sp}+\mathrm{Sn}} \tag{7}$$

在本次实验选定的 PPI 网络中, 共有 5 093 个蛋白质, 其中有 2 894 个蛋白质具有功能注释. 我们首先分析 FPM 和其他 4 种对比方法对这 2 894 个蛋白质实施功能预测的整体性能, 图 6 是实验对比结果. 图 6 显示了各种方法对 2 894 个蛋白质进行功能预测的平均特异性、敏感性和 F_measure 值. 2 894 个蛋白质中被 FPM、Zhang、DCS、NC 和 PON 5 种方法至少正确预测一个功能的蛋白质数量分别为 1 407、801、1 118、1 626 和 566 个. FPM 的覆盖蛋白质数量比 Zhang、DCS 和 PON 分别提高了 75.66 %, 25.85 % 和 148.59 %.

从图 6 中可以看出, FPM 具有最高的特异性 (Specificity), 这意味着 FPM 算法预测的功能中错误 (噪声) 功能所占比例最少. 敏感性 (Sensitivity) 方面, FPM 比 Zhang、DCS 和 PON 分别提高了 73.65 %, 22.54 % 和 173.46 %. 这说明, 相比这三种经典的功能预测算法, FPM 算法在不增加噪声功能比例的前提下能够注释更多的蛋白质. FPM 算法的敏感性低于 NC. 这是因为, FPM 只选择了排名靠前的部分功能注释功能未知的蛋白质, 而 NC 方法是将邻居的所有功能全部赋予待预测的蛋白质. 但是, 这种策略导致 NC 方法预测的功能中包含大量的噪声功能, 使得特异性急剧下降. 本次实验中, 虽然 NC 方法的敏感性比 FPM 提高了 29.49 %, 但是特异性却比 FPM 下降了 192.92 %. 就综合性能而言, FPM 方法的 F-

measure 值分别比 Zhang、DCS、NC 和 PON 分别提高了 61.27%、16.04%、91.34% 和 148%. 由此可见, FPM 方法具有最高的综合性能.

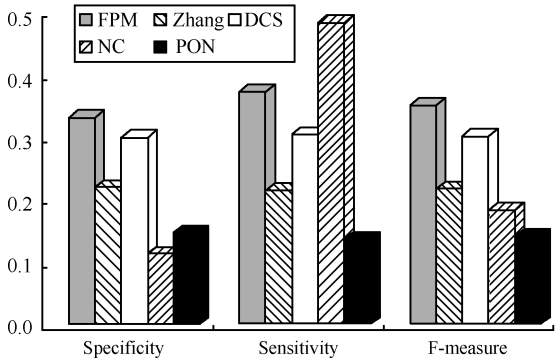


图 6　各种算法综合性能对比

Fig. 6　Overall performance comparison of various algorithms

由于 5 种预测方法分别采取了不同的功能数量选取策略, 为了更加全面、客观地对比分析各种方法的性能, 我们将尽可能地为各种方法选择相同的功能数量选取策略, 对每一个蛋白质, 分别选取各种预测方法的前 K 项功能进行预测. 针对 Zhang 方法和 DCS 方法, 选取前 M ($M \leq K$) 个最相似的蛋白质, 从这 M 个蛋白质的功能列表中选取前 K 项功能作为预测的功能. 功能根据蛋白质的相似值的最大值降序排列 (例如, 有多个蛋白质具有某项功能 F_i, 则取这些蛋白质中与待预测的蛋白质最相似的蛋白质的相似值作为功能 F_i 的排序得分); 对于 FPM、NC 和 PON 方法, 我们分别选取各自方法预测的前 K 个 GO Term 对功能未知的蛋白质进行功能注释. K 的取值从 1 到 50, 对于不同 K 值, 分别计算各种方法的平均 F_measure 值, 对比结果如图 7 所示.

图 7 清晰地显示, 当 K 从 1 增长到 50 时, FPM 始终具有最高的平均 F-measure 值. 随着 K 值的增长, FPM 方法的 F-measure 值虽然略微下降, 但基本能维持在 0.32 左右, Zhang 方法和 DCS 方法的 F-measure 值则下降非常明显. 综合分析图 7 和图 5, 我们还可以发现一个有趣的现象: FPM 方法和 PON 方法采用相同的功能数量选取策略, 但是这种策略对各自的影响却不尽相同, FPM 方法能从图 7 的 0.32 提升至图 5 的 0.35, 而 PON 方法的 F-measure 值则从 0.21 下降到 0.14. 换句话说, 这种功能数量的选取策略反而抑制了 PON 方法的功能预测性能.

我们在 Krogan 数据集运行本文提出的 FPM 算法和其他竞争算法. 表 1 是各种算法在 Krogan 数据集的运行结果.

从表 1 不难看出, 采用 Krogan 数据集时, FPM 算法依然具有最高的综合性能 (F_measure).

综合上述分析可知, 相比其他几种功能预测算法, FPM 算法具有最高的预测准确率, 这有力地说明了本文构建的多关系蛋白质网络的可靠性. 本文提出的算法中, 我们根据网络结构信息设置功能选取策略, 为了验证该方式的有效性, 我们对比分析了该策略与已有方法的功能选取策略. 本次实验中, 我们用 NC 方法的功能选取策略替换现有的基于网络结构信息的功能选取策略, 表 2 是对比结果.

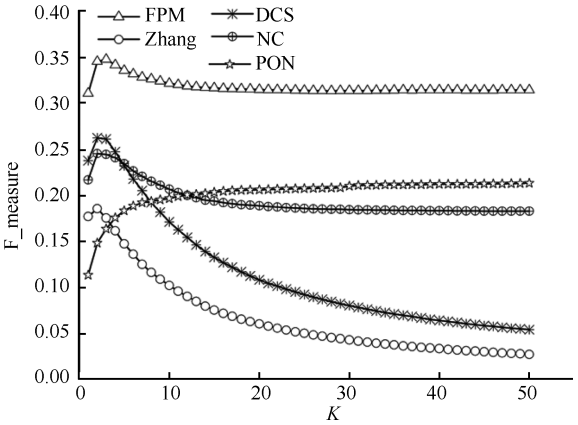


图 7　不同 K 值时各种算法的 F_measure 对比

Fig. 7　Comparison of average F_measure of various algorithms under different K values

表 1　Krogan 数据集运行结果

Table 1　Results on Krogan dataset

算法	特异性	敏感性	F_measure
FPM	0.317	0.342	0.329
Zhang	0.193	0.188	0.191
DCS	0.290	0.288	0.289
NC	0.120	0.421	0.186
PON	0.122	0.113	0.117

表 2　功能选取策略对预测结果的影响

Table 2　The influence of function selection strategy

功能选取策略	Sp	Sn	F_measure
基于网络结构	0.331	0.373	0.350
基于 NC 方法	0.193	0.546	0.315

从表 2 可以看出, 采用原有的功能选取策略, 虽然敏感性比较高, 但是特异性和综合评估指标 F_measure 明显低于基于网络结构信息的功能选取策略. 实验结果验证了该方式的有效性.

此外, FPM 算法通过为不同类型的相互作用设置不同的优先级的方式, 将多关系网络分解为多个

简单的网络. 为了验证 FPM 设置的优先级的有效性, 我们将尝试更改不同类型相互作用的优先级顺序, 分别进行功能预测. 实验结果如表 3 所示.

表 3 优先级对预测结果的影响

Table 3 The influence of priority

数据集	优先方式	Sp	Sn	F_measure
DIP	共享结构域	0.331	0.373	0.350
	共享邻居	0.274	0.308	0.290
	物理相互作用	0.120	0.493	0.193
Krogan	共享结构域	0.317	0.342	0.329
	共享邻居	0.280	0.290	0.285
	物理相互作用	0.128	0.432	0.197

从表 3 中可以看出, 在算法其他策略相同的情况下, DIP 数据集中, 采用共享结构域优先方式预测时的 F_measure 分别比共享邻居优先和物理相互作用优先两种方式提高了 20.69 % 和 81.35 %; Krogan 数据集相应的提高比例为 15.4 % 和 67 %. 从而验证了前述设置的相互作用优先级的有效性.

4 结论

存在大量未注释的蛋白质, 且相互作用数据包含大量假阳性和假阴性. 为解决这个难题, 现有功能预测方法整合 PPI 网络和多元的生物信息数据, 提高功能预测的准确率和性能. 融入多元生物信息后, 蛋白质间的相互作用变得多样化. 不同类型的相互作用在功能预测中的作用各不相同. 简单合并多种相互作用, 虽能增加预测的功能数量, 但同时增加了假阳性功能的数量. 本文根据 PPI 网络的拓扑特性和蛋白质的结构域信息构造具有多关系的蛋白质相互作用网络, 并为不同类型的相互作用赋予不同的优先级. 基于构造的多关系蛋白质网络, 本文提出一种名为 FPM 的蛋白质功能预测方法. 通过对比 FPM 方法和其他经典的蛋白质功能预测算法的性能, 验证了多关系蛋白质网络的有效性.

References

1 Schwikowski B, Uetz P, Fields S. A network of protein-protein interactions in yeast. *Nature Biotechnology*, 2000, **18**(12): 1257−1261

2 Hishigaki H, Nakai K, Ono T, Tanigami A, Takagi T. Assessment of prediction accuracy of protein function from protein-protein interaction data. *Yeast*, 2001, **18**(6): 523−531

3 Chua H N, Sung W K, Wong L. Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. *Bioinformatics*, 2006, **22**(13): 1623−1630

4 Nabieva E, Jim K, Agarwal A, Chazelle B, Singh M. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics*, 2005, **21**(S1): i302−i310

5 Chi X X, Hou J Y. An iterative approach of protein function prediction. *BMC Bioinformatics*, 2011, **12**: 437

6 Lee J, Gross S P, Lee J. Improved network community structure improves function prediction. *Scientific Reports*, 2013, **3**: 2197

7 Cozzetto D, Buchan D W A, Bryson K, Jones D T. Protein function prediction by massive integration of evolutionary analyses and multiple data sources. *BMC Bioinformatics*, 2013, **14**(S3): article S1

8 Kourmpetis Y A I, van Dijk A D J, ter Braak C J F. Gene Ontology consistent protein function prediction: the FALCON algorithm applied to six eukaryotic genomes. *Algorithms for Molecular Biology*, 2013, **8**(1): 10

9 Zhang S, Chen H, Liu K, Sun Z R. Inferring protein function by domain context similarities in protein-protein interaction networks. *BMC Bioinformatics*, 2009, **10**(1): 395

10 Peng W, Wang J X, Cai J, Lu C, Li M, Wu F X. Improving protein function prediction using domain and protein complexes in PPI networks. *BMC Systems Biology*, 2014, **8**(1): 35

11 Liang S D, Zheng D D, Standley D M, Guo H R, Zhang C. A novel function prediction approach using protein overlap networks. *BMC Systems Biology*, 2013, **7**(1): 61

12 Ding Zhao-Yun, Zhou Bin, Jia Yan, Zhang Lu-Min. Topical influence analysis based on the multi-relational network in microblogs. *Journal of Computer Research and Development*, 2013, **50**(10): 2155−2175
(丁兆云, 周斌, 贾焰, 张鲁民. 微博中基于多关系网络的话题层次影响力分析. 计算机研究与发展, 2013, **50**(10): 2155−2175)

13 Zhou T, Wang Y L. Causal relationship inference for a large-scale cellular network. *Bioinformatics*, 2010, **26**(16): 2020−2028

14 Zhao B H, Wang J X, Li M, Wu F X, Pan Y. Detecting protein complexes based on uncertain graph model. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2014, **11**(3): 486−497

15 Zhao B H, Wang J X, Li M, Wu F X, Pan Y. Prediction of essential proteins based on overlapping essential modules. *IEEE Transactions on NanoBioscience*, 2014, **13**(4): 415−424

16 Ji Jun-Zhong, Liu Zhi-Jun, Liu Hong-Xin, Liu Chun-Nian. An overview of research on functional module detection for protein-protein interaction networks. *Acta Automatica Sinica*, 2014, **40**(4): 577−593
(冀俊忠, 刘志军, 刘红欣, 刘椿年. 蛋白质相互作用网络功能模块检测的研究综述. 自动化学报, 2014, **40**(4): 577−593)

17 The UniProt Consortium. Update on activities at the universal protein resource (UniProt) in 2013. *Nucleic Acids Research*, 2013, **41**(D1): D43−D47

18 Lima T, Auchincloss A H, Coudert E, Keller G, Michoud K, Rivoire C, Bulliard V, de Castro E, Lachaize C, Baratin D, Phan I, Bougueleret L, Bairoch A. HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot. *Nucleic Acids Research*, 2009, **37**(S1): D471−D478

19 Wasson T, Hartemink A J. An ensemble model of competitive multi-factor binding of the genome. *Genome Research*, 2009, **19**(11): 2101−2112

20 Hawkins T, Chitale M, Luban S, Kihara D. PFP: automated prediction of gene ontology functional annotations with confidence scores using protein sequence data. *Proteins: Structure, Function, and Bioinformatics*, 2009, **74**(3): 566−582

胡 赛 长沙学院数学与计算机科学系副教授. 2003 年获得湖南大学数学与计量经济学院硕士学位. 主要研究方向为数理统计. E-mail: husaiccsu@163.com
(**HU Sai** Associate professor in the Department of Mathematics and Computer Science, Changsha University. She received her master degree from Hunan University in 2003. Her main research interest covers mathematical statistics.)

熊慧军 长沙学院数学与计算机科学系副教授. 2003 年获得湖南师范大学理学院硕士学位. 主要研究方向为生物信息学. E-mail: xionghuijunccsu@163.com
(**XIONG Hui-Jun** Associate professor in the Department of Mathematics and Computer Science, Changsha University. She received her master degree from Hunan Normal University in 2003. Her main research interest covers bioinformatics.)

李学勇 长沙学院数学与计算机科学系教授. 2003 年获得湖南大学计算机学院硕士学位. 主要研究方向为数据挖掘.
E-mail: xueyongli@163.com
(**LI Xue-Yong** Professor in the Department of Mathematics and Computer Science, Changsha University. He received his master degree from Hunan University in 2003. His main research interest is data mining.)

赵碧海 博士, 长沙学院数学与计算机科学系副教授. 2014 年获得中南大学信息学院博士学位. 主要研究方向为生物信息学, 数据挖掘. 本文通信作者.
E-mail: bihaizhao@163.com
(**ZHAO Bi-Hai** Associate professor in the Department of Mathematics and Computer Science, Changsha University. He received his Ph. D. degree from Central South University in 2014. His research interest covers bioinformatics and data mining. Corresponding author of this paper.)

倪问尹 博士，长沙学院数学与计算机科学系副教授. 2013 年获得中南大学信息学院博士学位. 主要研究方向为生物信息学. E-mail: masonni@163.com
(**NI Wen-Yin** Associate professor in the Department of Mathematics and Computer Science, Changsha University. He received his Ph. D. degree from Central South University in 2013. His research covers bioinformatics.)

杨品红 湖南文理学院生命科学学院教授, 博士. 主要研究方向为水生生物资源与利用. E-mail: yph098@163.com
(**YANG Pin-Hong** Professor at the School of Life Science, Hunan University of Arts and Science. He received his Ph. D. degree in 1999. His research interest covers aquatic biological resources and utilization.)

刘 臻 长沙学院生物与环境工程系副教授, 博士. 主要研究方向为分子营养与调控研究.
E-mail: Zhenliuccsu@163.com
(**LIU Zhen** Associate professor in the Department of Biological and Environmental Engineering, Changsha University. He received his Ph. D. degree in 2010. His research covers molecular nutrition and regulation.)