动态加权蛋白质相互作用网络构建及其应用研究

胡 赛 [1] 熊慧军 [1] 赵碧海 [1] 李学勇 [1] 王 晶 [1]

摘 要 一个蛋白质可能在不同条件或不同时刻与不同的蛋白质发生相互作用, 这称为蛋白质的动态特性. 蛋白质在分子处理的不同阶段参与到不同的模块, 与其他的蛋白质共同完成某项功能. 因此, 动态蛋白质相互作用的研究有助于提高蛋白质功能预测的准确率. 结合蛋白质相互作用网络和时间序列基因表达数据, 构建动态蛋白质相互作用网络. 为降低 PPI 网络中假阴性对功能预测产生的负面影响, 结合结构域信息和复合物信息, 预测和产生新的相互作用, 并对相互作用加权. 基于构建的动态加权网络, 提出一种功能预测方法 D-PIN (Dynamic protein interaction networks). 基于三个不同的酵母相互作用网络实验结果表明, D-PIN 方法的综合性能比现有方法提高了 14 % 以上. 结果验证了构建的动态加权蛋白质相互网络的有效性.

关键词 动态加权网络, 功能预测, 蛋白质相互作用网络, 基因表达

引用格式 胡赛, 熊慧军, 赵碧海, 李学勇, 王晶. 动态加权蛋白质相互作用网络构建及其应用研究. 自动化学报, 2015, **41**(11): 1893−1900

DOI 10.16383/j.aas.2015.c150211

Construction of Dynamic-weighted Protein Interactome Network and Its Application

HU Sai[1] XIONG Hui-Jun[1] ZHAO Bi-Hai[1] LI Xue-Yong[1] WANG Jing[1]

Abstract A protein would interact with different proteins under different conditions or at different time instants, which is the dynamic attribute of interactions. Proteins participate in different functional modules in different stages of molecular processing to perform different functions with other proteins. So, research of dynamic protein-protein interaction would contribute to the accuracy improvement of protein functions prediction. We construct a dynamic protein interaction network (D-PIN) by integrating protein-protein interaction network and time course gene expression data. To reduce the negative effect of false "negative" on the protein function prediction, we predict and generate some new protein interactions which combine with proteins′ domain information and protein complexes information and weight all the interactions. Based on the weighted dynamic network, we propose a method for predicting protein function, named D-PIN. Experimental results compared with using three different yeast interactome networks indicate that the comprehensive performance of D-PIN is 14 % higher other competing methods. Results also verify the effectiveness of the constructed dynamic-weighted protein interactome network.

Key words Dynamic weighted network, functions prediction, protein interactome network, gene expression

Citation Hu Sai, Xiong Hui-Jun, Zhao Bi-Hai, Li Xue-Yong, Wang Jing. Construction of dynamic-weighted protein interactome network and its application. *Acta Automatica Sinica*, 2015, **41**(11): 1893−1900

由于蛋白质在不同的生物过程中扮演重要角色, 蛋白质功能注释已经成为后基因时代的一个重要挑战. 目前仍有大量已知的基因和蛋白质没有通过实验获取特征, 它们的功能是未知的. 生物功能不是由单个蛋白质实现, 而是通过复杂的相互作用与众多的蛋白质共同完成.

然而, 在新的条件或刺激下, 不仅蛋白质的数量和位置会发生变化, 蛋白质之间的相互作用也在发生变化. 一个蛋白质可能在不同条件或不同时刻与不同的蛋白质发生相互作用. 相互作用在细胞的每一个生物过程中都非常重要. 许多重要的分子过程都是由大型的多分子机器执行, 如 RNA 剪接、多聚腺苷酸化、蛋白质输出等, 这些都是由大量的相互作用的复合物形成的. 因此, 蛋白质之间可能会形成稳定的模块, 还有一些蛋白质会随着变化的相互作用而形成临时的、动态的模块. 蛋白质在不同的条件和不同的周期时刻具有不同的功能. 因此, 传统的通

收稿日期 2015-04-13 录用日期 2015-08-18
Manuscript received April 13, 2015; accepted August 18, 2015
国家自然科学基金 (11501054), 湖南省自然科学基金项目 (13JJ4106, 14JJ3138), 湖南省教育厅项目 (10C0408, 15C0124), 湖南省科技计划项目 (2010FJ3044, 2015GK3072) 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), and Science and Technology Plan Project of Hunan Province (2010FJ3044, 2015GK3072)
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
1. Department of Mathematics and Computer Science, Changsha University, Changsha 410022

过分析静态网络预测蛋白质功能的方式是不合适的.

实验方法测定的相互作用可能不会在生物体内出现, 或者只存在于细胞周期的某一时刻或某一时间段内. 生物网络中蛋白质之间的相互作用随着时间、外部条件、刺激以及细胞的不同阶段而变化的特性, 称为蛋白质相互作用网络的动态特性. 我们认为, 构造动态的蛋白质相互作用网络, 在不同条件、不同时刻识别动态的生物模块, 进而实现蛋白质功能注释是提高功能预测性能的又一新的有效途径.

Yook 等[1] 得出的结论是大部分功能或位置以全蛋白质相互作用网络的隔离子网形式出现. 已有学者[2−4] 成功构建动态网络, 并用于提高功能模块预测, 实验结果证实了模块的动态性. Tang 等[2] 利用基因表达数据和蛋白质相互作用网络构造了一个名为 TC-PINS (Time course protein interaction networks) 的时序蛋白质相互作用网络, 并成功地应用于功能模块的识别. 在 PPI (Protein-protein interaction) 网络中存在相互作用的两个蛋白质, 若它们在某一时刻的基因表达值都超过某一固定阈值, 则认为它们在这一时刻共表达, 并在该时刻的 TC-PINS 网络中添加一条边. 然而, 不同蛋白质的表达量存在很大差异, 最小值接近 0, 最大值超过 150. 细胞中不同基因有不同的表达模式, 有的蛋白质在细胞里发挥重要作用, 却在整个细胞周期内具有很低的基因表达水平. 因此, 设置统一的阈值对所有蛋白质进行过滤的方式会使得构造的动态网络不准确. 考察一个蛋白质在某一时刻是否处于活跃状态应该考虑蛋白质自身的表达水平. Wu 等[3] 和 Wang 等[4] 等提出 3-sigma 准则构造动态网络, 阈值根据各自蛋白质基因表达的平均值和标准差计算得到.

本文的工作主要体现在以下几方面: 1) 在研究已有动态网络构建的基础之上, 考虑基因表达的周期性, 改进动态网络的构建方式; 2) 结合蛋白质结构域信息和蛋白质复合物信息, 补充缺失的相互作用, 并对动态网络加权, 降低假阴性对功能预测造成的影响; 3) 考虑模块和功能的动态特性, 基于构建的动态网络, 提出一种蛋白质功能预测方法 D-PIN (Dynamic protein interaction networks).

我们在三个不同来源的酵母蛋白质相互作用网络中运行 D-PIN 和其他几种对比的功能预测算法. 实验结果表明, D-PIN 算法的预测性能优于现有的蛋白质功能预测算法.

1 动态网络构建

动态网络已经引起了人们的广泛关注, 从信息交互网络到社交网络, 再到生物网络、疾病分子网络等. 从抽象的数学模型角度看, 动态网络其实是一个有序的图序列, 表示复杂系统在不同时刻的快照[5]. 动态网络的形式化定义如下所示:

定义 1. 动态网络[6] G 定义为一系列网络 $G=\{G_1, G_2, \cdots, G_i, \cdots G_k\}$, $G_i=(V_i, E_i)$ 是采样时刻 i 的网络, $V_i=\{v_{i1}, v_{i2}, \cdots, v_{in}\}$ 表示 i 时刻的蛋白质集合, $E_i=\{e_{i1}, e_{i2}, \cdots, e_{im}\}$ 表示 i 时刻的蛋白质相互作用的集合. 假定 $e^+=(u, v) \in (E_i \backslash E_{i-1})$ 表示 $i-1$ 时刻不存在, 而 i 时刻新增的相互作用; $e^-=(u, v) \in (E_{i-1} \backslash E_i)$ 表示 $i-1$ 时刻存在, 而 i 时刻消失的相互作用.

根据采样时机的不同, 动态网络可以划分为空间动态网络和时间动态网络. 所谓空间动态网络是指蛋白质之间在不同的空间状态或条件下具有不同的相互作用, 如不同的细胞定位信息等. 时间动态网络是指蛋白质之间在不同的采样时刻表现出不同的相互作用, 如基因或蛋白质在不同时刻表达水平的差异使得蛋白质之间的相互作用发生动态变化. 本文主要介绍基于时间动态网络的构建及其在蛋白质功能预测中的应用.

基因表达有条件和时序地打开和关闭, 基因表达数据在生物过程的不同条件或不同阶段能反映蛋白质存在的动态性. 因此, 研究不同时间点和不同条件下基因表达数据将为研究蛋白质相互作用的动态变化提供途径. 直观地, 若两个蛋白质在 PPI 网络中存在相互作用, 并且在某一个时刻的基因表达水平都超过阈值, 则认为这两个蛋白质在该时刻共表达. 我们利用蛋白质在不同时刻的共表达特性, 构建动态的蛋白质相互作用网络. Wu 等[3] 和 Wang 等[4] 等分析指出, 不同蛋白质的表达水平差异较大, 因此不适合采用统一的阈值判定两个蛋白质是否共表达, 而应该根据各自的表达水平自适应确定. 不同于 3-sigma 准则[3−4], 本文中, 如果蛋白质在某时刻的表达水平超过自身的平均表达水平, 则认为蛋白质在该时刻表达. 也就是说本文取消了阈值, 这样还可以减少原方法中阈值对算法的影响, 提高算法的适应性.

此外, Tu 等[7] 通过研究发现, 酵母菌株经过一个简要的饥饿期后生长到高密度, 之后开始自发地周期性呼吸, 整个过程通过耗氧量测定. 进一步地, 他们进行了微阵列基因表达的分析, 评估证实了周期基因表达的存在. 如图 1[7] 所示, 蛋白质 36 个时刻的基因表达分为三个周期, 每 12 个时刻为一个周期.

Tang 等[2]、Wu 等[3] 和 Wang 等[4] 在建立动

态网络时, 都是分别在 36 个时刻构建一个蛋白质相互作用网络, 而本文则是为基因表达周期的 12 个时刻构建相互作用网络, 每一个时刻的基因表达值为三个周期的平均值, 即:

$$T^{'}(i)=\frac{T(i)+T(i+12)+T(i+24)}{3} \tag{1}$$

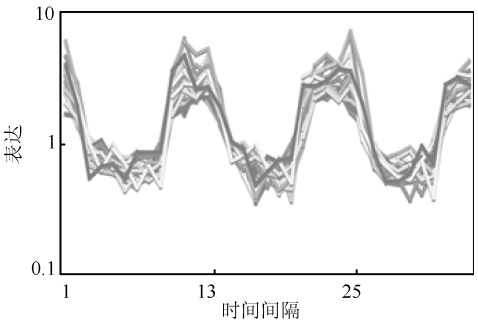

图 1　基因的周期表达

Fig. 1　Gene expression during the metabolic cycle

研究表明, 通过高通量的生物实验方法获取的蛋白质相互作用数据中包含较高比例的假阳性和假阴性. 尽量减少假阳性和假阴性造成的负面影响是提高蛋白质功能预测性能的关键和瓶颈. 已有多种功能预测算法利用生物数据添加新的蛋白质相互作用, 减少假阴性的影响. 然而不幸的是, 这些方法在降低假阴性的同时不可避免地会导致假阳性的提高. 也就是说, 利用多元生物信息预测的相互作用数据中包含假阳性. 由此可见, 假阳性和假阴性成为一对互斥的矛盾. 本文利用蛋白质结构域信息、蛋白质复合物信息、PPI 网络拓扑特性及其构建的动态蛋白质相互作用网络, 旨在降低假阳性和假阴性造成的负面影响. 针对假阴性, 我们利用蛋白质共享结构域信息和复合物信息的特性, 在原有的 PPI 网络中添加新的相互作用; 至于假阳性, 我们利用蛋白质的共表达特性, 构建动态蛋白质网络, 将不共表达的蛋白质相互作用作为假阳性予以去除.

图 2 描述了动态网络的构建过程, 分为三个步骤执行.

步骤 1. 网络加权, 利用结构域信息、复合物信息和网络拓扑特性增加和移除一些相互作用, 并对相互作用网络加权. 一般而言, 相互作用的存在概率可以通过挖掘相连的两个顶点的共同邻居数量确定. 对于相互作用的节点 u 和 v, 如果节点 u 或 v 的度等于 1, 则相互作用的存在概率等于 0; 否则, 我们采用 PN 计算相互作用的存在概率, 其定义如下:

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

其中, N_i 和 N_j 分别表示 v_i 和 v_j 的邻居节点集合. 如果 v_i 和 v_j 之间没有相互作用, 则 $\mathrm{PN}(v_i, v_j)=0$. 接下来, 利用结构域信息和蛋白质复合物信息补充 PPI 网络中的相互作用. 对于 PPI 网络中的任意两个蛋白质 v_i 和 v_j 且都至少包含一个结构域, D_i 和 D_j 分别表示 v_i 和 v_j 的结构域组成的集合, $\mathrm{PD}(v_i, v_j)$ 表示共享结构域的概率, 其计算公式如下所示:

$$\mathrm{PD}(v_i,v_j)=\frac{|D_i\cap D_j|^2}{|D_i||D_j|} \tag{3}$$

式 (3) 中, $D_i\cap D_j$ 表示的 v_i 和 v_j 共同结构域集合. 同理可得 v_i 和 v_j 共享复合物的概率 PC(v_i 和 v_j), 计算公式如下式 (4).

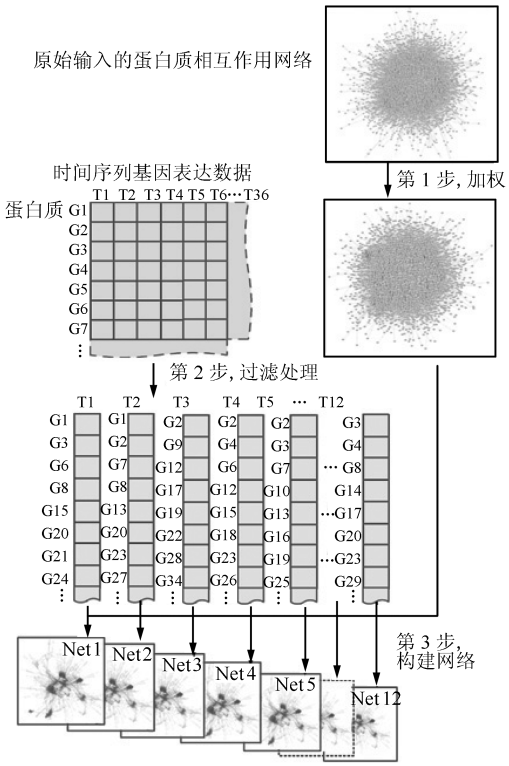

图 2　动态网络构建

Fig. 2　Construction of dynamic networks

$$\mathrm{PC}(v_i,v_j)=\frac{|C_i\cap C_j|^2}{|C_i||C_j|} \tag{4}$$

其中, C_i 和 C_j 分别表示包含 v_i 和 v_j 的复合物集合, $C_i\cap C_j$ 表示同时包含 v_i 和 v_j 的复合物集合. 如果 v_i 或 v_j 没有被任何复合物包含, 则 $\mathrm{PC}(v_i, v_j)=0$. 对于 PPI 网络 $G=(V, E)$ 中的任意两个蛋白质 v_i 和 v_j, 如果 $(v_i, v_j)\in E$, 则它们之间相互作用的权值计算方法如下所示:

$$W(v_i, v_j) = \mathrm{PN}(v_i, v_j) + \mathrm{PD}(v_i, v_j) + \mathrm{PC}(v_i, v_j) \tag{5}$$

否则, v_i 和 v_j 的之间权值计算为

$$W(v_i, v_j) = \mathrm{PD}(v_i, v_j) + \mathrm{PC}(v_i, v_j) \tag{6}$$

$E^{'} = \{(v_i,\ v_j) |\ v_i,\ v_j \in V,\ (v_i,\ v_j) \notin E,\ \mathrm{PC}(v_i,\ v_j) + \mathrm{PC}(v_i,\ v_j) > 0\}$ 表示通过结构域信息和复合物信息新增的相互作用集合. 加权后的网络可以描述为 $G^{'} = (V,\ E \cap E^{'},\ W)$, $W = \{w(e_1),\ w(e_2),\ \cdots,\ w(e_m)\}$, $w(e_i)$ 表示边 e_i 的权值.

步骤 2. 根据式 (1), 将基因表达的三个周期合并为一个周期, 然后在周期的 12 个时刻, 过滤基因表达值低于平均值的蛋白质.

步骤 3. 构建动态网络, 在时刻 i ($i \in [1, 12]$), 若两个蛋白质共表达并在 PPI 网络中存在相互作用, 则在该时刻的动态网络中增加一组相互作用.

2 D-PIN 算法

为了验证本文构建的动态加权蛋白质相互作用网络对于蛋白质功能预测的有效性, 本文设计一种基于动态加权网络的功能预测算法 D-PIN. 算法首先根据待预测功能的蛋白质在 12 个动态网络中的邻居节点形成候选蛋白质集合, 并计算每一个候选蛋白质的得分. 候选蛋白质的得分是 12 个时刻动态网络中该蛋白质与待预测蛋白质之间相互作用权值的总和. 候选蛋白质得分的形式化表达如下:

动态网络集合 $G = \{G_1, G_2,\ \cdots,\ G_{12}\}$ 是根据前述方法得到的 12 个网络, $G_i = (V_i, E_i, W_i)$ ($i \in [1,\ 12]$), u 为功能未知等待预测的蛋白质, v 为网络中一个功能已知的蛋白质, v 在预测 u 的功能时的得分计算公式如下所示:

$$S\_Protein(v) = \sum_{i=1}^{12} W(v, u) \times t_i \tag{7}$$

式中, 如果 $(v, u) \in E_i$, 则 $t_i = 1$; 否则 $t_i = 0$. 设 $P = \{p_1, p_2,\ \cdots,\ p_n\}$ 是根据上述方法预测 u 的功能时形成的候选蛋白质集合, $F = \{f_1, f_2,\ \cdots,\ f_m\}$ ($m \geq n$) 是 P 集合中所有蛋白质的已知功能构成的集合. 对于 F 中某一候选功能 f_j, 其得分计算方法如下所示:

$$S\_Function(f_j) = \sum_{i=1}^{n} S\_Protein(p_i) \times t_{ij} \tag{8}$$

式中, $S\_Protein(p_i)$ 是候选蛋白质 p_i 的得分. 若蛋白质 p_i 包含功能 f_j, 则 $t_{ij} = 1$; 否则, $t_{ij} = 0$. 所有候选功能根据得分降序排列, 算法从中选取前 N 项功能作为功能未知的蛋白质的预测功能列表. D-PIN 算法统计每一个候选蛋白质包含候选功能的数量, 包含候选功能最多的蛋白质的功能数量将作为 N 的取值. 即 N 的计算如下所示:

$$N = GoNum(p_i), \quad \text{其中 } p_i = \max\left(\sum_{i=1}^{m} t_{ij}\right) \tag{9}$$

其中, $GoNum(p_i)$ 表示邻居蛋白质 p_i 的功能数量. 算法 D-PIN 说明了基于动态网络的蛋白质功能预测方法的整体框架.

D-PIN 算法

Input: A set of dynamic networks $G = \{G_1,\ G_2,\ \cdots,\ G_{12}\}$

Output: The set of predicted functions PF

1) For each un-annotated protein u Do
2) Get a proteins set $P = \{p_1, p_2,\ \cdots,\ p_n\}$
3) Get a functions set $F = \{f_1, f_2,\ \cdots,\ f_m\}$
4) For each function fi in F Do
5) $S\_Function(f_j) = \sum_{i=1}^{n} S\_Protein(p_i) \times t_{ij}$
6) End For
7) Order functions of F descendant by scores
8) $PF = \{f_1, f_2,\ \cdots,\ f_N\}$; //$N$ is computed using Equation (9)
9) Output PF
10) End For

3 实验结果和分析

我们用于实验分析的 PPI 网络来源于酿酒酵母, 因为它已经通过基因敲除实验被很好地特征化, 并被广泛应用于功能预测的评估. 酿酒酵母相互作用网络的功能注释数据、基因表达数据和复合物数据是比较完善和可靠的. PPI 网络数据源于 DIP[8] 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. 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[9]. 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 网站将蛋白质格式从 UniProtKB 转换为 Ensemble Genomes Protein, 以便与 PPI 网络中蛋白质的格式匹配.

本文所用的结构域 (Domain) 数据是从 Pfam 数据库下载得到[10], 包含 1 107 个不同的结构域, 涉及 PPI 网络中的 3 056 个蛋白质. 蛋白质复合物数据采用 CYC2008[11] 数据集, CYC2008 包含

408 个通过生物方法预测得到的复合物, 并被作为标准的已知复合物集合, 广泛应用于蛋白质复合物预测方法评价. 酵母的基因表达数据[7] 共包含 6 776 个基因产品 (蛋白质) 在 36 个不同时刻的采样数据. 6 776 个蛋白质中, 有 4 902 个蛋白质包含在 DIP 数据集中. 关键蛋白质的覆盖率超过 95 %(4 902/5 023 = 97.59 %). 对于没有基因表达数据的蛋白质, 我们只是简单地将基因表达值设为 0.

我们将对比 D-PIN 方法与其他的功能预测方法: NC (Neighbor counting)[12]、Zhang[13]、DCS (Domain combination similarity)[14]、PON (Protein overlap network)[15] 和 DSCP (Domain combination similarity in context of protein complexes)[14] 的预测结果. 其中, DSCP 方法是 DCS 方法的改进, 结合了复合物信息. 为了评测 D-PIN 方法的性能, 我们采用交叉验证法, 这是一种广泛用于蛋白质功能预测算法评估方法. 该方法的基本思想是人为去除网络中部分蛋白质的功能注释, 并将这部分蛋白质作为测试集, 剩余的具有功能注释的蛋白质作为训练集, 并用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, 我们分别采用留一法和七倍交叉验证法测试各种算法的性能.

3.1 留一法验证

每一轮仅有一个蛋白质的功能被移除并放入测试集, 而剩余的所有蛋白质在训练集中用于对测试集中的蛋白质进行功能注释. 对每一个蛋白质进行预测, 并分别计算准确率 (Precision)、召回率 (Recall)、F-measure 和覆盖率. 准确率是指预测的功能中有多大比例与已知的功能之间能够匹配, 召回率是指已知的蛋白质功能有多大比例与被预测的功能匹配. F-measure 则能较好地反映算法的综合性能, 它是 Precision 和 Recall 的调和平均值. 覆盖率是指至少匹配一项功能的蛋白质在所有预测的蛋白质中所占比重. 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, 将依次作为测试蛋白质. 表 1 显示了 D-PIN、Zhang、DCS、NC、PON 和 DSCP 6 种方法预测的平均准确率、召回率、F-measure 值和覆盖率.

从表 1 可以看出, D-PIN 方法的覆盖率分别比 Zhang、DCS、PON 和 DSCP 方法提高了 88.02 %、34.7 %、166.08 % 和 17.66 %. 其中, 覆盖率提高百分比计算公式为

$$\frac{\text{D}-\text{PIN 覆盖率} - \text{其他算法覆盖率}}{\text{其他算法覆盖率}} \times 100\,\% \quad (10)$$

表 1 留一法验证结果

Table 1 Results using leave-one-out cross validation

算法	准确率 (%)	召回率 (%)	F-measure	覆盖率 (%)
D-PIN	39.1	44.4	41.6	52.0
Zhang	21.9	21.5	21.7	27.7
DCS	29.9	30.4	30.2	38.6
NC	11.3	48.3	18.3	56.2
PON	14.6	13.6	14.1	19.6
DSCP	34.9	35.4	35.1	44.2

D-PIN 方法的召回率和覆盖率仅次于 NC 方法, 这是因为, D-PIN 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, 从而能够匹配更多的正确功能. 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, 使得准确率急剧下降. 本次实验中, 虽然 NC 方法的召回率比 D-PIN 提高了 8.78 %, 但是准确率却比 D-PIN 下降了 246.02 %, 综合衡量指标 F-measure 比 D-PIN 下降了 127.32 %.

我们的 D-PIN 方法获得了最高的性能, F-measure 分别比 Zhang、DCS、NC、PON 和 DSCP 提高了 91.64 %、37.89 %、127.32 %、194.69 % 和 18.46 %. 而 D-PIN 方法的准确率比其他 5 种方法最少提高了 10 % 以上, 最多提高了 2 倍以上. 由此可见, D-PIN 方法具有最优的综合性能. D-PIN 算法具有仅次于 NC 方法的召回率, 说明构建的动态网络包含较少的假阴性, 而最高的准确率则表明动态网络中降低了假阳性的比例.

由于 6 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, 对每一个待预测的蛋白质, 分别选取各种方法预测的前 K 项功能进行预测. 针对 Zhang、DCS 和 DSCP 方法, 选取前 $M\,(M \le K)$ 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 D-PIN、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, 对于不同的 K 值, 分别计算各种方法的平均 F-measure 值, 对比结果如图 3 所示.

从图 3 可以清晰地看出, 当 K 从 1 增长到 50 时, D-PIN 方法的 F-measure 曲线始终位于最上方, 这也就意味着 PDN 具有最优的整体性能.

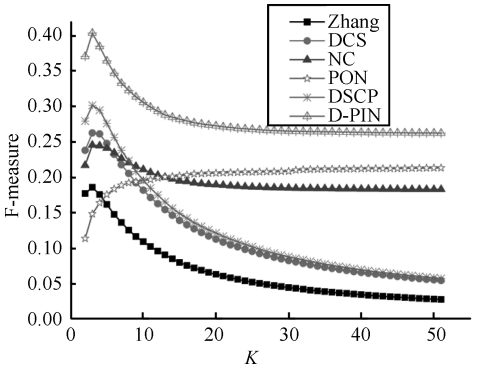

图 3 F-measure 变化曲线图

Fig. 3 F-measure curves as K values varies

3.2 七倍交叉验证

上述实验中采用留一法验证了 D-PIN 方法性能, 为防止留一法验证可能造成的偏差, 本节采用七倍交叉验证 (7-fold cross validation) 法进一步评估 D-PIN 方法的预测性能. 七倍交叉验证法属于留部分法, 它随机地将数据集分成 7 份, 轮流将其中 6 份做训练, 剩余的一份做测试, 7 次的结果的均值作为对算法精度的估计. 为了降低实验误差, 这一过程重复 100 次, 平均结果作为最终值. 表 2 列出了各种方法多次重复实验结果的平均偏差.

表 2 交叉验证误差统计结果

Table 2 Results of error statistics using cross validation

算法	准确率 (%)	召回率 (%)	F-measure	覆盖率 (%)
D-PIN	1.53	1.58	1.51	1.67
Zhang	1.44	1.33	1.84	1.57
DCS	1.57	1.81	1.65	1.95
NC	0.69	1.64	0.95	1.60
PON	1.50	1.42	1.45	1.91
DSCP	1.57	1.84	1.67	1.88

表 2 结果显示, 各种方法实验结果的平均偏差都比较小, 说明实验结果真实有效. 表 3 给出各种方法采用七倍交叉验证的实验结果. 从实验结果不难看出, 采用七倍交叉验证法, D-PIN 依然取得最高的准确率和 F-measure 值, 召回率和覆盖率略低于于 NC 方法.

表 3 七倍交叉验证结果

Table 3 Results using 7-fold cross validation

算法	准确率 (%)	召回率 (%)	F-measure	覆盖率 (%)
D-PIN	38.3	43.0	40.5	50.9
Zhang	20.8	20.6	20.7	26.6
DCS	29.0	29.6	29.3	37.8
NC	11.8	45.8	18.8	53.7
PON	14.3	13.3	13.8	19.1
DSCP	34.0	34.4	34.2	43.3

3.3 其他数据集结果分析

为了全面对比各种功能预测算法, 我们还采用留一法在其他两个不同的酵母 PPI 网络 (Krogan[16] 数据库和 Collins[17] 数据库) 测试了 D-PIN 方法和其他 5 种对比方法. 去除重复的相互作用和自相互作用后, Krogan 数据库由 3 672 个蛋白质和 14 317 组相互作用组成, Collins 包括 1 622 个蛋白质和 9 074 组相互作用. Krogan 和 Collins 有功能注释的蛋白质数量分别是 2 268 和 1 274. 表 4 和表 5 分别列出了 Krogan 和 Collins 网络上各种方法预测功能的实验结果.

采用留一法在 Krogan 和 Collins 两个 PPI 网络进行功能预测时, D-PIN 依然能够取得最高的准确率和 F-measure 值. 在不同数据上的测试结果也证明了 D-PIN 算法的可靠性.

表 4 Krogan 数据集结果

Table 4 Results on the Krogan data

算法	准确率 (%)	召回率 (%)	F-measure	覆盖率 (%)
D-PIN	36.7	40.5	38.5	49.3
Zhang	19.3	18.8	19.1	25.3
DCS	29.0	28.8	28.9	37.2
NC	12.0	42.1	18.6	49.3
PON	12.2	11.3	11.7	17.5
DSCP	33.2	33.2	33.2	42.2

表 5 Collins 数据集结果

Table 5 Results on the Collins data

算法	准确率 (%)	召回率 (%)	F-measure	覆盖率 (%)
D-PIN	44.3	48.6	46.3	58.9
Zhang	19.7	19.0	19.4	25.2
DCS	38.1	39.3	38.7	49.3
NC	21.0	60.3	31.1	69.1
PON	15.5	14.5	15.0	20.1
DSCP	39.9	41.1	40.5	51.6

4 结论

目前仍有大量已知的基因和蛋白质没有通过实验获取特征, 它们的功能未知. 功能预测的计算方法一般都是基于蛋白质相互作用网络, 然而, 生物网络中蛋白质之间的相互作用可能会随着时间、外部条件、刺激以及细胞的不同阶段而变化. 本文利用基因的周期表达特性, 结合蛋白质相互作用网络和多元的生物信息, 建立动态加权蛋白质相互作用网络, 并提出一种名为 D-PIN 的功能预测方法, 旨在降低相互作用网络中的假阳性和假阴性对功能预测造成的负面影响. 实验结果验证了动态加权网络的有效性.

References

1 Yook S H, Oltvai Z N, Barabási A L. Functional and topological characterization of protein interaction networks. *Proteomics*, 2004, **4**(4): 928−942

2 Tang X W, Wang J X, Liu B B, Li M, Chen G, Pan Y. A comparison of the functional modules identified from time course and static PPI network data. *BMC Bioinformatics*, 2011, **12**(1): 339

3 Wu F X, Ressom H, Dunn M J, Wang J X, Peng X Q, Li M, Pan Y. Construction and application of dynamic protein interaction network based on time course gene expression data. *Proteomics*, 2013, **13**(2): 301−312

4 Wang J X, Peng X Q, Peng W, Wu F X. Dynamic protein interaction network construction and applications. *Proteomics*, 2014, **14**(4−5): 338−352

5 Mantzaris A V, Bassett D S, Wymbs N F, Estrada E, Porter M A, Mucha P J, Grafton S T, Higham D J. Dynamic network centrality summarizes learning in the human brain. *Journal of Complex Networks*, 2013, **1**(1): 83−92

6 Chen Guan-Rong. Problems and challenges in control theory under complex dynamical network environments. *Acta Automatica Sinica*, 2013, **39**(4): 312−321
(陈关荣. 复杂动态网络环境下控制理论遇到的问题与挑战. 自动化学报, 2013, **39**(4): 312−321)

7 Tu B P, Kudlicki A, Rowicka M, McKnight S L. Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes. *Science*, 2005, **310**(5751): 1152−1158

8 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

9 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)

10 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

11 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

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

13 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

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

15 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

16 Krogan N J, Cagney G, Yu H Y, Zhong G Q, Guo X H, Ignatchenko A, Li J, Pu S Y, Datta N, Tikuisis A P, Punna T, Peregrín-Alvarez J M, Shales M, Zhang X, Davey M, Robinson M D, Paccanaro A, Bray J E, Sheung A, Beattie B, Richards D P, Canadien V, Lalev A, Mena F, Wong P, Starostine A, Canete M M, Vlasblom J, Wu S, Orsi C, Collins S R, Chandran S, Haw R, Rilstone J J, Gandi K, Thompson N J, Musso G, St Onge P, Ghanny S, Lam M H Y, Butland G, Altaf-Ul A M, Kanaya S, Shilatifard A, O'Shea E, Weissman J S, Ingles C J, Hughes T R, Parkinson J, Gerstein M, Wodak S J, Emili A, Greenblatt J F. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature*, 2006, **440**(7084): 637−643

17 Collins M O, Yu L, Choudhary J S. Analysis of protein phosphorylation on a proteome-scale. *Proteomics*, 2007, **7**(16): 2751−2768

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. 主要研究方向为生物信息学, 概率论与数理统计, 数据挖掘. 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 research interest covers bioinformatics, probability and mathematical statistics, data mining.)

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** Ph. D., associate professor in the Department of Mathematics and Computer Science, Changsha University. He received his Ph. D. degree from Central Sourth University in 2014. His research interest covers bioinformatics and data mining. Corresponding author of this paper.)

王 晶 博士, 长沙学院数学与计算机科学系副教授. 2009 年获得湖南师范大学理学院博士学位. 主要研究方向为图论.
E-mail: wangjing1001@hotmail.com
(**WANG Jing** Ph. D., associate professor in the Department of Mathematics and Computer Science, Changsha University. She received her Ph. D. degree from Hunan Normal University in 2009. Her main research interest is graph theory.)