基于动态图的PPI网络构建和复合物挖掘算法研究

李 鹏[1,2,3],闵 慧[4],罗爱静[1,3]
(1.中南大学湘雅三医院,湖南长沙 410013;2.湖南中医药大学信息科学与工程学院,湖南长沙 410208;3.医学信息研究湖南省普通高等学校重点实验室(中南大学),湖南长沙 410006;4.湖南信息职业技术学院软件学院,湖南长沙 410200)

摘 要: 动态蛋白质网络的构建和复合物挖掘问题是目前研究的热点.针对现有的算法在解决前述问题上的不足,文中考虑了蛋白质的活性周期和连接强度,首先提出了一种基于动态图的蛋白质网络构建算法.然后基于密度聚类设计了一种在动态蛋白质网络上挖掘复合物的算法(PCMA).整个挖掘过程包含三个步骤:基于DBSCAN(Density-Based Spatial Clustering of Applications with Noise)算法的蛋白质复合物生成;基于合并增益的蛋白质复合物合并和基于归属度的复合物调整.在多个公开的生物数据集上进行了实验,实验结果表明,所提算法在查全率、查准率和F-measure方面的性能都要优于现有的算法,且对输入参数不敏感.在保证蛋白质复合物挖掘准确性的前提下,算法的时间复杂度处于一个合理的范围之内.

关键词: 动态蛋白质网络;蛋白质复合物;动态图;密度聚类;查全率;查准率;时间复杂度
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112(2021)08-1489-09
电子学报URL: http://www.ejournal.org.cn **DOI:** 10.12263/DZXB.20200357

Research on PPI Network Construction and Complex Mining Algorithm Based on Dynamic Graph

LI Peng[1,2,3], MIN Hui[4], LUO Ai-jing[1,3]
(1. *The Third Xiangya Hospital of Central South University*, *Changsha*, *Hunan* 410013, *China*;
2. *School of Informatics*, *Hunan University of Chinese Medicine*, *Changsha*, *Hunan* 410208, *China*;
3. *Key Laboratory of Medical Information Research (CSU)*, *College of Hunan Province*, *Changsha*, *Hunan* 410006, *China*;
4. *Software Department*, *Hunan College of Information*, *Changsha*, *Hunan* 410200, *China*)

Abstract: Dynamic protein network construction and complex mining problem is a hot topic. In view of the shortcomings of existing algorithms in solving the above problems, a protein network construction algorithm based on dynamic graph is firstly proposed by considering the active period and the connection strength of proteins in this paper. Then, a protein complex mining algorithm(PCMA) on dynamic protein network is designed based on the density clustering. The whole mining process consists of three steps: the generation of protein complex based on DBSCAN(density-based spatial clustering of applications with noise) algorithm; the combination of protein complex based on the combination gain and the adjustment of protein complex based on the degree of membership. Experiments are carried out on several open biological datasets. The experimental results show that the performance of the proposed algorithm is better than that of the existing algorithms in terms of recall, precision and F-measure, and it is not sensitive to the input parameters. On the premise of ensuring the accuracy of protein complex mining, the time complexity of the proposed algorithm is in a reasonable range.

Key words: dynamic protein network; protein complex; dynamic graph; density clustering; recall; precision; time complexity

1 引言

随着基因测序工程的完成和生命科学研究重点的转移[1,2],生物体的蛋白质相互作用网络(Protein-Protein Interaction Network, PPIN)[3]的相关研究已经成为生物信息学领域的重点之一,对该网络进行准确的建

收稿日期:2020-04-13;修回日期:2020-08-27;责任编辑:孙瑶
基金项目:国家社会科学基金重点项目(No.17AZD037);国家重点研发计划(No.2017YFC1703306);湖南省卫生健康委科研项目(No.202112072217);湖南自然科学基金青年项目(No.2019JJ50453);湖南自然科学基金面上项目(No.2018JJ2301);湖南省科技厅重点项目(No.2018JJ2301);湖南省教育厅一般项目 (No.19C1318)

模和分析成为一项十分重要的基础性工作. 蛋白质相互作用网络是指生物体内所有蛋白质之间在遗传上或物理上存在相互作用关系从而形成的一种拓扑结构,简称蛋白质网络. 其中,在相同时间和空间内由若干个蛋白质通过相互作用共同组成的一种多分子结构称为蛋白质复合物[4](Protein Complex). 大量的研究表明[5,6],蛋白质复合物是细胞行使其某些功能的关键载体,通过蛋白质的相互作用,几乎所有的生物过程都能被精确地执行,例如DNA复制、转录、翻译、物质代谢、信号传导以及细胞周期控制等一般生物过程. 采用计算机相关技术对蛋白质网络进行分析和研究是目前的一种主流手段,蛋白质复合物的挖掘可以看作一种典型的大图数据挖掘问题. 将蛋白质网络建模成一个图,然后利用各种图聚类算法来挖掘蛋白质复合物是目前的普遍做法.

然而,由于蛋白质的存在或降解较易受到环境的影响,在不同的生物学环境下,蛋白质的表达存在较大的差异,因此蛋白质之间的相互作用具有动态变化的特点,即蛋白质网络的结构对于环境的刺激具有较强的应激性. 但遗憾的是,目前收集到的蛋白质相互作用数据大多存在假阳性和假阴性等噪声,这给蛋白质网络的分析带来了极大的困难,使得如何准确地对蛋白质之间的动态相互作用进行衡量已经成为目前面临的一个关键挑战[7]. 由于该问题不能得到有效的解决,因此成为制约研究生物信息学领域中诸多问题(功能预测、关键蛋白识别等)的瓶颈. 为此,本文将对动态蛋白质网络的构建问题进行研究,并在此基础上提出新的蛋白质复合物挖掘方法,该方法具有重要的实际价值:它可以帮助生物学家从蛋白组学的观点出发更好地了解生物体的动态发展变化过程;有利于医药工作者高效地分析疾病的致病基因,并为药物靶点设计提供理论依据. 本文的研究方法还可以有效推广到社交网、语言网和航空网等复杂信息处理领域,具有普适意义.

2 相关工作

目前,国内外众多学者对动态蛋白质网络的构建和复合物挖掘算法进行了研究. 例如,Tang等人[8]认为,如果PPI网络中两个蛋白质在某个时间点的基因表达值都超过某一阈值则它们之间存在基于时间依赖的相互作用,以此为基础提出了时序PPI网络(Time Course Protein Interaction Networks, TC-PINS),并成功地用TC-PINS进行了复合物的挖掘. 然而由于受到蛋白质的表达模式、环境、生物活性等因素的影响,不同蛋白质之间的表达量经常存在较大差异,这导致TC-PINS方法通过统一设置阈值过滤构建得到的网络在很多情况下并不准确,影响了复合物的挖掘质量.

另外,雷秀娟等人[9]借鉴物理学中拓扑势场的概念来构建动态蛋白质网络,文中将每个蛋白质看作一个物理粒子,通过分析它们之间的作用场来得到一个加权的蛋白质网络,并基于马尔科夫聚类算法进行复合物挖掘. 然而该方法对噪声非常敏感,而且很难挖掘交叠的蛋白质复合物. Shen等人[10]提出一种基于邻域亲和力和动态蛋白质相互作用网络的复合物发现算法(DPC-NADPIN). 该算法首先选择聚类系数高的蛋白质及其邻域中的每一个进行合并,形成初始聚类. 根据聚类内邻域间亲和力与聚类外邻域间亲和力的关系,将聚类内的邻域蛋白与聚类外的邻域蛋白相结合,形成蛋白质复合物. 然而该方法的结果严重依赖初始划分的质量,且划分后的每个蛋白质只能属于一个复合物,而实际的PPI网络中每个蛋白质可能具有多个功能,参与多个不同的生物进程,因此该方法的局限性较大. Lei等人[11]提出一种改进的花授粉算法以挖掘多关系重建的动态PPI网络中的蛋白质复合物. 文中首先考虑了蛋白质在寻找本质相互作用中的重要性,然后设计了多关系重构的动态PPI网络,并发现了网络中蛋白质复合物的潜在核心. 最后,提出了一种基于花授粉机制的IFPA(Improved Flower Pollination Algorithm)算法,通过模拟花粉过程生成蛋白质复合物. 该方法可在扩充过程中允许某个蛋白质重复出现,但无法挖掘PPI网络中非稠密的子图结构,挖掘精度还有待提高.

针对以上方法存在的不足,本文首先提出了一种基于动态图的蛋白质网络构建方法,并在此基础了设计了一种基于密度聚类的蛋白质复合物挖掘方法. 最后的仿真实验也验证了本文方法的有效性.

3 构建动态蛋白质网络

3.1 相关概念和定义

蛋白质之间的相互作用具有动态性,本文采用动态图对蛋白质网络进行建模,然后在构建得到的动态蛋白质网络上进行复合物的挖掘. 下面给出网络构建过程中所需用到的相关定义.

定义1(动态图) 给定图$G=(V,E)$和它的一系列子图$S_G=G_1,G_2,...,G_T$,其中$\bigcup_{i=1}^{T}G_i=G$. 设$TS_T=t_1,t_2,...,t_T$为子图相对应的时间序列,则称$\Omega=(G,S_G,TS_T)$为动态图. 就PPI网络而言,设$G_t=(V_t,E_t)$表示t时刻的蛋白质网络快照$(t=1,2,...,T)$;V_t表示t时刻蛋白质网络中节点的集合;E_t表示t时刻蛋白质网络中边的集合. 所有时刻的蛋白质网络快照构成动态蛋白质网络,即一个动态图.

定义2(蛋白质复合物) 设$CP_t=\{CP_{t,i}\}$. 其中,$CP_{t,i}$表示G_t上存在的第i个复合物;CP_t表示G_t中的蛋白质复合物.

定义3(蛋白质的活性周期) 对于任意的一个蛋白质P而言,如果在一个给定的时间周期T内,P的基因表达平均值$u(P)$都不低于阈值ε,则称$T(P)$为P的活性周期.

定义4(公共节点) $\forall P \in V_t$,如果在一个蛋白质网络中至少存在2个蛋白质复合物CP_1和CP_2都包含了P,则称P为公共节点.

定义5(增量边) 设$e^+ = (u,v) \in E_t, e^+ \notin E_{t-1}$表示在$t$时刻蛋白质网络中新增的边;$e^- = (u,v) \in E_{t-1}, e^- \notin E_t$表示在$t$时刻蛋白质网络中移除的边.

定义6(间接邻居节点$INB(v)$) 设$v \in V_t$,$NB^+(v)$表示蛋白质节点v的邻居节点集合(包含v本身);$NB(v)$表示不包含v本身的邻居节点集合. 间接邻居节点指节点邻居的邻居,可以定义为

$$INB(v) = \{t | t \in NB(u) \cap u \in NB(v)\} \tag{1}$$

定义7(节点关联性$Cor(u,v)$) 两个蛋白质节点u和v之间的关联性定义为

$$Cor(u,v) = \frac{|NB^+(u) \cap NB^+(v)|}{|NB^+(u) + NB^+(v)|} \tag{2}$$

其中,$|NB^+(u)|$表示蛋白质节点u的邻居蛋白质节点数目.

定义8(节点间距离$d(u,v)$) 两个蛋白质节点u和v之间的距离定义为

$$dis(u,v) = \sqrt{(1 - Cor(u,v))^2} \tag{3}$$

定义9(复合物合并增益Δg_{ij}) 对于任意的一个蛋白质网络G_k而言,当其中的两个蛋白质复合物CP_i和CP_j进行合并时,Δg_{ij}定义为

$$\Delta g_{ij} = 2(e_{ij} - q_i q_j), q_i = \frac{s\_d_i}{2m} \tag{4}$$

其中,e_{ij}表示CP_i和CP_j之间相连的边占总边数的比例$(i \neq j)$;q_i表示与CP_i中节点相连,但另一个节点不属于CP_i的边数占网络总边数的比例;s_d_i表示CP_i中节点的度数之和;m表示G_k的总边数.

定义10(增量节点ΔINt) 设NNt、VNt和$NCNt$分别表示在蛋白质网络G_t中t时刻相对于$t-1$时刻的新增节点、消失节点和邻居变化节点,则ΔINt定义为

$$\begin{aligned}
&NN_t = \{v | v \in V_t \cap v \notin V_{t-1}\},\\
&VN_t = \{v | v \in V_{t-1} \cap v \notin V_t\},\\
&NCN_t = \{v | (e^+ = (u,v) \cap ((u \in CP_{t-1,i} \cap v \in CP_{t-1,j}, i \neq j)\\
&\qquad \cup (u \in NN_t \cap v \in CP_{t-1,i})))\\
&\qquad \cup (e^- = (u,v) \cap ((u \in CP_{t-1,i} \cap v \in CP_{t-1,j}, i \neq j)\\
&\qquad \cup (u \in VN_t \cap v \in CP_{t-1,i}))\},\\
&\Delta IN_t = \{NN_t \cup VN_t \cup NCN_t\}
\end{aligned} \tag{5}$$

定义11(边变化率) 任意的蛋白质节点u的边变化率$Ecr(u)$为

$$Ecr(u) = \frac{|e_t^+(u)| + |e_t^-(u)| + 1}{|e_{t-1}(u)| + 1} \tag{6}$$

其中,$|e_t^+(u)|$表示蛋白质u在时刻t新增的边的数目;$|e_t^-(u)|$表示蛋白质u在时刻t丢失的边的数目;$|e_{t-1}(u)|$表示蛋白质u在时刻$t-1$拥有的边的数目.

定义12(归属度$MD_{v,i}$) 对于t时刻的蛋白质节点v而言,v对复合物$CP_{t,i}$的归属度定义为

$$\begin{aligned}
MD_{v,i} = &\frac{\sum\limits_{u \in CP_{t,i} \cap u \in NB(u,v)} Cor(u,v)}{\sum\limits_{u \in NB(v)} Cor(u,v)}\\
&+ \frac{1}{(\max\{deg(x)\})^2} \sum_{u \in CP_{t,i} \cap u \in INB(v)} \left(\frac{deg(u) \cdot deg(v)}{dis(u,v) + 1} \right)
\end{aligned} \tag{7}$$

其中,$deg(u)$表示蛋白质节点u的度;$\forall x \in G_t$,$\max\{deg(x)\}$表示G_t中蛋白质节点的度的最大值.

3.2 网络构建的主要思想

紧接着3.1节的定义,文中通过如下的三个步骤来构建动态蛋白质网络:

首先,我们根据蛋白质的基因表达值计算它的活性周期.

然后,基于蛋白质的不同活性周期将一个生物体内的所有蛋白质划分成不同的集合,属于同一集合的蛋白质具有相同的活性周期,拥有同一个时间片. 再根据定义13所示的连接强度对同一时间片内的所有蛋白质构建蛋白质子网.

最后,对上一阶段得到的各个蛋白质子网采用动态图进行建模,得到最终的动态蛋白质网络.

3.2.1 活性周期的计算

设$GEV_t(P_i)$表示蛋白质P_i在t时刻的基因表达值,$1 \leqslant t \leqslant n$. $u(P_i)$表示蛋白质P_i在一个时间片内(1到n)基因表达值的均值,$\sigma(P_i)$表示蛋白质P_i在一个时间片内的基因表达值的标准差. 则有

$$u(P_i) = \frac{\sum_{t=1}^{n} GEV_t(P_i)}{n} \tag{8}$$

$$\sigma(P_i) = \frac{\sum_{t=1}^{n} (GEV_t(P_i) - u(P_i))^2}{n} \tag{9}$$

进一步地,文中采用$F(P_i)$反映蛋白质P_i的基因表达曲线的波动性:

$$F(P_i) = \frac{1}{1 + \sigma^2(P_i)} \tag{10}$$

根据式(10)可知,$\sigma(P_i)$越大,$F(P_i)$越小,$0 \leqslant F(P_i) \leqslant 1$. 此外,我们采用如下的3-sigma准则[8]来定义活性阈值ε:

$$\varepsilon = u(P_i) \times F(P_i) + (u(P_i) + 3\sigma(P_i)) \times (1 - F(P_i)) \quad (11)$$

对于任意给定的一个时间片 T_α 而言，如果有 $u(P_i) \geqslant \varepsilon, i = 1, 2, ..., k$，则表明这 k 个蛋白质有相同的活性周期 T_α，可以在 T_α 内将这 k 个蛋白质构建成一个子网. 以此类推，我们通过计算所有蛋白质的活性周期可以划分出多个具有不同时间片大小的蛋白质集合，对每个集合中的所有蛋白质构建一个子网，从而构建得到多个蛋白质子网.

3.2.2 构建蛋白质子网

设 $P\_S = \{P_1, P_2, ..., P_n\}$ 表示具有同一活性周期的 n 个蛋白质的集合，文中通过分析这 n 个蛋白质之间的相互作用来构建蛋白质子网. 其中的关键是要确定在同一时间片内的不同蛋白质之间是否存在连接关系，为此，我们从多个角度出发定义了连接强度来衡量不同蛋白质之间的相互作用，如果连接强度超过某一固定阈值，则认为它们之间存在连接关系.

具体而言，文中从两个方面来计算蛋白质之间的连接强度：公共邻居数和连接差异数. 公共邻居数是不同蛋白质之间存在的共同邻居个数，如果它们之间的公共邻居越多，则表明相互作用越强；连接差异数是蛋白质节点之间的邻接边数和各个节点的度的最小值的比值. 它也可以间接地衡量蛋白质之间的相互作用关系. 例如，假设有蛋白质 $\{p1, p2, p3\}$，各自的度分别为 $\{18, 12, 6\}$，$p1 - p2$ 之间的邻接边数为 5，$p1 - p3$ 之间的邻接边数为 4. 则相对于 $p2$ 而言，$p3$ 与 $p1$ 的相互作用关系更强，这是因为 $p2$ 中还有更多的边可能与其他的蛋白质发生相互作用. 综上可得，连接强度的定义如下.

定义 13（连接强度）

$$JS(P_i, P_j) = \frac{1}{1 + e^{-\left(\frac{dc_{ij}}{\min(deg_i, deg_j)} + \frac{|NB(P_i) \cap NB(P_j)|}{|NB(P_i)| + |NB(P_j)|}\right)}} \quad (12)$$

其中，$JS(P_i, P_j)$ 是蛋白质 P_i 和蛋白质 P_j 之间的连接强度；dc_{ij} 是蛋白质 P_i 和蛋白质 P_j 之间直接相连的边的数量；$NB(P_i)$ 是蛋白质 P_i 的邻居节点集合；deg_i 是蛋白质 P_i 的度；式中的 $\frac{1}{1 + e^{-(\cdot)}}$ 是一个 Sigmoid 函数[12].

综上所示，本文提出的动态蛋白质网络构建算法如算法 1 所示.

算法 1 动态蛋白质网络构建算法

输入：蛋白质相互作用数据，阈值 λ，基因表达数据

输出：动态蛋白质网络模型 $\Omega = (G, S_C, TS_T)$

Step1 以基因表达数据作为输入联立公式(7)~(9)来计算所有蛋白质的活性周期 $T(P)$，然后根据 $T(P)$ 的大小对其进行降序排序后采用列表存储，有

$T(P) = [T_1(P), T_2(P), ..., T_k(P)]$;

Step2 紧接着采用 $T(P)$ 构建动态蛋白质子网：

For $T_i(P), i = 1, 2, ..., k$ in $T(P)$:

在 $T_i(P)$ 中计算 $JS(P_i, P_j)$;

如果 $JS(P_i, P_j) \geqslant \lambda$，则认为蛋白质 P_i 和蛋白质 P_j 之间存在相互作用，在两者之间添加一条边，并标记该条边所对应的时间序列 TS_i;

Step3 重复执行 Step2，直到列表 $T(p)$ 为空，则结束.

4 蛋白质复合物的挖掘

4.1 蛋白质复合物挖掘算法总体设计

下面在上述已经构建好的动态蛋白质网络基础上，提出了基于密度聚类的蛋白质复合物挖掘算法. 不妨假设整个蛋白质网络由 T 个时刻的子网 S_G 组成，则挖掘算法主要包含以下几个步骤：

(1) t_1 时刻进行蛋白质间的距离计算；

(2) t_1 时刻生成蛋白质复合物；

(3) t_1 时刻的蛋白质复合物进行合并；

(4) 输出 t_1 时刻的蛋白质复合物；

(5) $t_i (2 \leqslant i \leqslant T)$ 时刻进行增量节点的计算；

(6) $t_i (2 \leqslant i \leqslant T)$ 时刻蛋白质网络新增节点间的距离计算；

(7) $t_i (2 \leqslant i \leqslant T)$ 时刻生成蛋白质复合物；

(8) 根据节点的复合物归属度对 $t_i (2 \leqslant i \leqslant T)$ 时刻的蛋白质节点的归属进行调整；

(9) $t_i (2 \leqslant i \leqslant T)$ 时刻的蛋白质复合物进行合并；

(10) 输出 $t_i (2 \leqslant i \leqslant T)$ 时刻的蛋白质复合物.

4.2 算法具体设计

根据上文给出的蛋白质复合物挖掘的总体思路，下面对其具体过程进行详细阐述.

4.2.1 蛋白质间的距离计算和复合物生成

为了进行蛋白质复合物的挖掘，首先需要计算蛋白质之间的距离，然后再采用聚类算法生成蛋白质复合物. 对于 t_1 时刻的蛋白质子网而言，考虑到在真实蛋白质网络中节点与邻居、间接邻居的关系较密切，而与其他节点较疏远，同时为了减少计算量，在计算蛋白质之间的距离时，只考虑蛋白质节点与邻居、间接邻居间的距离. 因此，蛋白质间的距离计算过程如下：首先通过式(2)和式(3)计算节蛋白质节点与各自邻居节点间的距离；然后计算蛋白质节点与间接邻居节点间的距离.

在获得蛋白质间的距离之后，本文采用基于密度的聚类方法(DBSCAN[13])来生成蛋白质复合物. 一般来说，真实蛋白质网络中的很多蛋白质可以同时属于多个复合物中的分子元件，但目前大多数基于密度的聚类算法没有考虑公共节点问题，使得部分蛋白质在聚类的时候只能归属于一个蛋白质复合物，这极大地降低了蛋白质复合物的挖掘质量. 为此，本文对这一挖

掘过程做了优化:

首先,根据预设的距离阈值d和核心点阈值th,采用DBSCAN算法对蛋白质网络进行聚类,一个聚类看作一个蛋白质复合物.

然后,在聚类过程中通过考查d-领域内蛋白质节点集合的个数与th的关系,获取公共节点及其所属的复合物集合,并将聚类间的公共节点同时划分到多个蛋白质复合物中.

4.2.2　蛋白质复合物的合并

在t_i时刻的蛋白质子网中,为了降低蛋白质相互作用数据中的假阳性或假阴性对于蛋白质复合物挖掘结果的影响,需进行蛋白质复合物的合并. 对于t_i时刻的蛋白质子网中蛋白质复合物的合并,有如下的性质.

性质1[14]　在任意给定的蛋白质网络中,对于任意的两个蛋白质复合物CP_i和$CP_j(i \neq j)$,设CP_k是CP_i和CP_j合并后生成的新复合物,CP_k、CP_i和CP_j内部的边数分别为ne_k、ne_i和ne_j,CP_i和CP_j之间相连的边数为ne_{ij},则有

$$ne_k = ne_i + ne_j + ne_{ij} \tag{13}$$

蛋白质复合物合并的主要过程如下.

首先,初始化一个合并增益矩阵$\boldsymbol{GM}_t = 0$,并根据式(4)计算t_i时刻的蛋白质复合物合并增益Δg_{ij},可得$\boldsymbol{GM}_t = [\Delta g_{ij}]$.

然后,对于t_1时刻的蛋白质子网中生成的所有蛋白质复合物进行迭代合并:对于每一个$CP_{t,i} \in CP_t$,if$(CP_{t,i} \notin CP_{t-1})$,从矩阵$\boldsymbol{GM}_{t-1}$中删除与$CP_{t,j}$有边相连的蛋白质复合物与$CP_{t,j}$的合并增益行向量$\Delta g_j$;对于每一个$CP_{t,i} \notin CP_{t-1}$,if$(CP_{t,i} \in CP_t)$,向矩阵$\boldsymbol{GM}_{t-1}$中添加与$CP_{t,j}$有边相连的蛋白质复合物与$CP_{t,j}$的合并增益行向量$\Delta g_i$. 我们统计每个蛋白质复合物内部及复合物间的边数目,并基于性质1和式(4)迭代更新$\boldsymbol{GM}_t$.

最后,当$\boldsymbol{GM}_t$中的元素值全为0时,蛋白质复合物合并过程结束,生成最终的蛋白质复合物.

4.2.3　蛋白质的复合物归属调整

由于蛋白质网络的动态性,在进行$t_i(i > 1)$时刻的蛋白质复合物挖掘时需要以t_{i-1}时刻的网络快照为基础. 此时,我们需要求取两个时间片上蛋白质网络快照的增量节点(可根据式(4)计算得到). 相对于t_{i-1}时刻而言,在t_i时刻的蛋白质网络中可能发生蛋白质节点的新增、消失和邻居发生变动的情形,因此需要对这些蛋白质节点的复合物归属进行调整. 需要调整的蛋白质节点是NCN_t中的节点及其邻居节点. 在进行蛋白质复合物调整时,先通过式(6)计算NCN_t中节点的边变化率,决定蛋白质复合物归属调整顺序;再通过式(7)计算蛋白质节点对各蛋白质复合物的归属度大小,得到最大归属复合物,并把节点划分到归属度最大的复合物中;最终判断节点的归属复合物是否发生变化,并通过式(2)计算复合物归属发生变化的节点与其邻居节点的关联性,按关联性大小降序排序,对邻居节点进行蛋白质复合物归属调整.

4.3　算法流程

综上所述,本文提出的蛋白质复合物挖掘算法的流程如算法2所示.

算法2　蛋白质复合物挖掘算法(PCMA)

输入:动态蛋白质网络$G = G_1, G_2, ..., G_T$,距离阈值d;核心点阈值th

输出:各个时刻的蛋白质复合物$CP_t = \{CP_{t,i}\}$

Step1　初始化合并增益矩阵$\boldsymbol{GM}_t = 0$

Step2　对于t_i时刻的蛋白质子网,采用DBSACN算法进行复合物生成,具体执行如下操作:

(1) 任选一个蛋白质节点P_i,根据式(2)和式(3)计算P_i与其他蛋白质节点之间的距离.

(2) 找到距离小于等于d的所有蛋白质节点. 如果距P_i的距离在d之内的蛋白质个数小于th,那么P_i被标记为噪声. 如果距离在d之内的蛋白质个数大于th,则P_i被标记为核心样本,并被分配一个新的复合物标签.

(3) 访问P_i的所有邻居(在距离th以内),如果它们还没有被分配给一个复合物,那么就将刚刚创建的新的复合物标签分配给它们.

(4) 在t_i时刻的蛋白质子网中重新选取一个尚未被访问过的蛋白质节点,重复上述过程直到t_i时刻的蛋白质子网中所有蛋白质节点访问完毕.

Step3　根据式(4)计算合并增益,并进行蛋白质复合物的合并

Step4　$i = i + 1$,根据式(5)计算增量节点,并根据式(6)调整蛋白质节点的复合物归属情况,直到相邻两次调整结果不变,则算法结束

5　实验

我们用Python语言实现了PCMA算法,在一台8核16线程的计算机上进行了实验. 其中,CPU型号为Intel Core i9-9960X @ 3.10GHz,内存为16GB,操作系统为Ubuntu 16.04 LTS 64位. 为了验证PCMA的有效性,我们在多个数据集上将PCMA算法与目前较为典型的蛋白质复合物挖掘算法进行了性能比较:TC-PINS(Time Course Protein Interaction Networks)[8]、MCL-TP[9]、DPC-NADPIN[10]和Improved Flower Pollination Algorithm(IFPA)[11].

5.1　实验数据集

本文采用DIP数据集、MIPS数据集[15]和CYC2008[16]数据集作为研究对象. 具体而言,DIP数据集采用官方提供的20170205版本的数据(http://dip.doe-mbi.ucla.edu/),它涵盖了通过噬菌体、串联亲和纯化等生物实验测得的可靠的蛋白质相互作

用,通过去掉冗余和自相互作用等预处理后,该数据集还有4995个蛋白质和21554对相互作用. MIPS数据集是指慕尼黑蛋白质序列信息中心,对其采用和DIP数据集相同的预处理后,该数据集还有4546个蛋白质和12319对相互作用. 而CYC2008数据集中则提供了408个通过生物实验方法确定的蛋白质复合物,它们将作为文中实验对比的对象.

5.2 评价指标

为了综合衡量PCMA算法的优越性,采用如下的几种指标来评价多种不同的蛋白质复合物挖掘算法的性能.

(1)查全率、查准率和F-measure值. 其中,查全率(Recall),又称特异性,是指本文算法挖掘的蛋白质复合物与实验数据集中真实存在的蛋白质复合物的最大匹配数目除以实验数据集中真实存在的蛋白质复合物总数的比值;查准率(Precision),又称敏感度,是指本文算法挖掘的蛋白质复合物与实验数据集中真实存在的蛋白质复合物的最大匹配数目除以实验测得的蛋白质复合物总数的比值. 计算公式如下

$$Recall = \frac{MNM(ER, RR)}{ER} \tag{14}$$

$$Precision = \frac{MNM(ER, RR)}{RR} \tag{15}$$

其中,ER表示本文算法挖掘的蛋白质复合物;RR表示实验数据集中真实存在的蛋白质复合物;$MNM(ER, RR)$表示ER和RR之间的最大匹配数目. 综合考虑查全率和查准率两方面,可得*F-measure*的计算公式为

$$F\text{-}measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{16}$$

(2)功能富集程度. 该指标主要用于进一步分析挖掘的蛋白质复合物是否具有生物意义. 它对于预测蛋白质复合物的功能也具有重要作用. 我们通过计算蛋白质复合物的*P-value*值来衡量它对某个功能的富集程度,其计算公式[17]为

$$P\text{-}value = 1 - \sum_{t=0}^{\lambda-1} \frac{\binom{F}{t}\binom{N-F}{PN-t}}{\binom{N}{PN}} \tag{17}$$

其中,N表示蛋白质网络的规模;PN表示蛋白质复合物中的蛋白质数量;λ表示蛋白质复合物中具有某项功能的蛋白质数量;F表示PPIN中具有该功能的蛋白质数量. 一般而言,如果*P-value*值越小,则表明挖掘的蛋白质复合物是随机具有这种功能的概率越低,即蛋白质复合物更具有生物学意义.

(3)时间开销. 在多个数据集上衡量不同算法挖掘蛋白质复合物所耗费的时间,验证不同算法的运行效率.

5.3 实验结果与分析

5.3.1 参数d对于PCMA算法的性能影响

下面首先分析参数d对于蛋白质复合物挖掘结果的影响. 以CYC2008数据集作为测试对象,测试了本文算法的性能,结果如图1所示. 从图中可以看到,d在[0.4,0.7]之间取不同值时,本文算法挖掘的蛋白质复合物的*F-measure*值在[0.82, 0.86]之间波动,变化幅度小于0.04,波动具有随机性. 当d的取值超过0.7后,*F-measure*值基本不变. 因此,本文认为d的取值对于本文算法的蛋白质复合物挖掘精度并无显著影响. 仔细分析其原因可知,文中在采用基于密度的聚类来挖掘蛋白质复合物的过程中创新性地将公共节点划分到多个蛋白质复合物中,减少了部分蛋白质节点及其d-领域内的其他蛋白质节点成为噪声点的概率,降低了算法对于d的敏感性,提高了挖掘质量. 另外,本文还通过采用基于合并增益的蛋白质复合物合并过程来有效降低DBSCAN聚类生成蛋白质复合物后剩余的噪声点对于蛋白质复合物质量的影响,提高了挖掘蛋白质复合物的准确率.

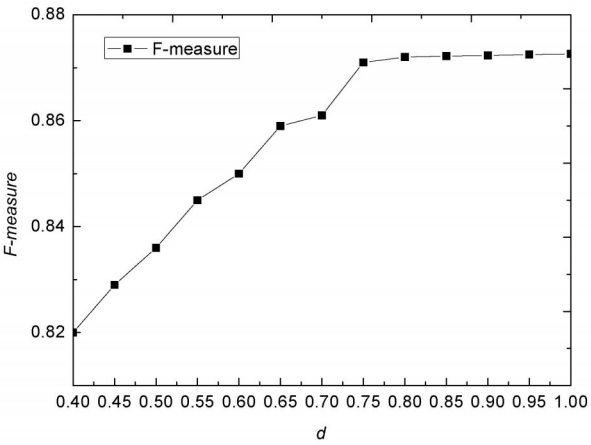

图1 d取不同值时本文方法的*F-measure*值

5.3.2 参数th对于PCMA算法的性能影响

接下来分析参数th对于蛋白质复合物挖掘结果的影响. 结果如图2所示,从图2可以看到,随着th的增大,本文算法的*F-measure*值先增大后减小. 当th在[2,6]内取不同值时,*F-measure*值在[0.82, 0.86]之间波动,变化幅度小于0.04,波动具有随机性. 这是因为当th取值较小时,采用DBSCAN算法进行蛋白质复合物挖掘后,噪声节点减少,核心节点增多,生成的蛋白质复合物增多. 而在蛋白质复合物合并的后处理步骤中,合并复合物会使生成的复合物数量减少,降低复合物挖掘的精度. 当th取值较大时,噪声节点增多,核心节点减少,生成的蛋白质复合物数量减少,同样会造成

部分蛋白质节点被误划分从而导致复合物挖掘的精度下降. 但是当 *th* 的取值超过7后,*F-measure* 值基本不变. 因此,本文认为 *th* 的取值对于本文算法的蛋白质复合物挖掘精度并无显著影响.

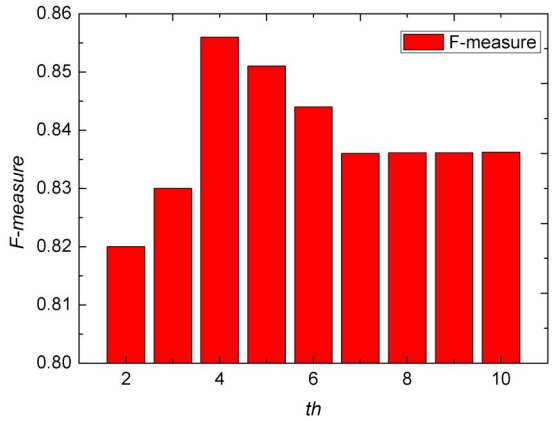

图2　*th* 取不同值时本文方法的 *F-measure* 值

5.3.3　PCMA算法与其他算法的比较

为了准确地分析PCMA算法在文中构建的动态蛋白质网络上挖掘复合物的性能,将PCMA算法与TC-PINS[12]、MCL-TP[13]、DPC-NADPIN[14]和IFPA[15]在DIP数据集和MIPS数据集上进行了比较. 文中对每一种算法都在所构建的每个动态子网上进行复合物挖掘,对于最终挖掘得到的复合物都采用文献[18]的后处理方式进行过滤,去掉重复的、包含蛋白质数量少于2的复合物. 下面的表1和表2分别列出了各种算法在DIP数据集和MIPS数据集上的实验结果.

从表1和表2的结果可以看到,PCMA算法在两种数据集上的查全率和查准率都要优于另外的四种算法. 在DIP数据集上,PCMA算法的 *F-measure* 值要比TC-PINS、MCL-TP、DPC-NADPIN和IFPA分别高41%、31%、27%和17%. 在MIPS数据集上,PCMA算法的 *F-measure* 值要比TC-PINS、MCL-TP、DPC-NADPIN和IFPA分别高36%、23%、21%和7%. 主要原因是:(1)本文通过考虑蛋白质的活性周期和连接强度来构建分时的动态蛋白质网络,较为准确地预测了蛋白质之间的相互作用;(2)PCMA算法通过计算邻居节点发生变化的蛋白质的边变化率来确定蛋白质复合物归属调整,使得增量节点尽可能地被划分到正确的复合物中,并对复合物归属发生变化的蛋白质节点的邻居也进行了复合物归属判断,有效地缩小了蛋白质网络中增量节点对于网络中其他蛋白质的复合物归属造成的影响,降低了复合物划分的累计误差和网络动态变化带来的影响. 因此取得了比其他算法更好的结果.

表1　各个算法在DIP数据集上的性能比较

算法	*Precision*	*Recall*	*F-measure*
TC-PINS	0.5417	0.4702	0.5034
MCL-TP	0.6818	0.5683	0.5889
DPC-NADPIN	0.7919	0.5122	0.6221
IFPA	0.8051	0.6322	0.7081
PCMA	0.8605	0.8456	0.8530

表2　各个算法在MIPS数据集上的性能比较

算法	*Precision*	*Recall*	*F-measure*
TC-PINS	0.4998	0.5115	0.5056
MCL-TP	0.6639	0.5714	0.6142
DPC-NADPIN	0.7781	0.5236	0.6260
IFPA	0.7967	0.6880	0.7384
PCMA	0.8513	0.7451	0.7947

5.3.4　PCMA算法的功能富集分析

为了进一步分析PCMA算法挖掘的蛋白质复合物所具有的生物学意义,我们计算所有包含蛋白质数量大于2的复合物的 *P-value* 值. 再从中筛选出 *P-value* 值小于0.005的复合物进行功能富集分析,我们认为这一部分蛋白质复合物具有较强的生物学意义,可以在两个方面为生物学家的工作提供技术支持:

(1) 预测未知蛋白质的功能. 以MIPS数据集为例,功能未知的蛋白质CSNK2B包含在一个规模为6的复合物中,该复合物内其他5个蛋白质(NOLC1,NOS2,HDAC4, HDAC3, RBBP4) 都具有transposition功能. 因此,我们可以预测CSNK2B也具有transposition功能.

(2) 估计已知蛋白质的新功能. 我们还可以通过分析蛋白质复合物中具有最小 *P-value* 值时所对应的功能来注释已知蛋白质的新功能. 表3给出了PCMA算法挖掘的一个规模为7的蛋白质复合物,其中的每个蛋白质都具有多种功能信息,当该蛋白质复合物对应的功能为 Acts as component of the MCM2-7 complex 时,其 *P-value* 值最小. 具体分析该复合物可以发现,其中有5个蛋白质都具有 Acts as component of the MCM2-7 complex 功能,其他2个蛋白质(CDC6 protein, DBF4-related protein)在现有的功能注释信息中没有发现 Acts as component of the MCM2-7 complex 功能,但这2个蛋白质都参与了mRNAprocessing. 因此,我们可以估计这2个蛋白质具有 Acts as component of the MCM2-7 complex 功能.

5.3.5　PCMA算法的效率分析

蛋白质网络是一种典型的复杂网络,算法运行时间是动态蛋白质网络中复合物挖掘算法评价的主要标准之一. 下面分别给出了不同算法在DIP数据集和MIPS数据集上进行复合物挖掘的时间开销比较,结果如表4所示.

从表4可以看到,PCMA算法在两种数据集上的运行时间不超过10s,低于MCL-TP、DPC-NADPIN和IFPA

表3 蛋白质复合物中各个蛋白的功能注释信息

蛋白质	功能
Origin recognition complex subunit 2	(1) stabilizes ORC3; (2) Binds histone H3 and H4 trimethylation marks H3K9me3, H3K20me3 and H4K27me3; (3) Stabilizes LRWD1; (4) Acts as component of the MCM2-7 complex.
DNA replication licensing factor MCM2	(1) Acts as component of the MCM2-7 complex; (2) Plays a role in terminally differentiated hair cells development of the cochlea and induces cells apoptosis.
CDC6 protein	(1) Involved in the initiation of DNA replication; (2) participates in checkpoint controls that ensure DNA replication is completed before mitosis is initiated.
Replication protein A3, 14kDa homo-log RPA14	(1) Acts as component of the MCM2-7 complex; (2) it plays an essential role both in DNA replication and the cellular response to DNA damage.
DNA replication licensing factor MCM5	(1) Acts as component of the MCM2-7 complex; (2) Interacts with MCMBP.
DNA replication licensing factor MCM7	(1) Acts as component of the MCM2-7 complex; (2) Once loaded onto DNA, double hexamers can slide on dsDNA in the absence of ATPase activity.
DBF4-related protein	(1) Regulatory subunit for CDC7; (2) involved in regulating the initiation of DNA replication during cell cycle.

表4 各个算法的运行时间比较(单位:s)

算法	DIP数据集	MIPS数据集
TC-PINS	8.5	7.75
MCL-TP	9.7	9.33
DPC-NADPIN	11.2	10.28
IFPA	12.8	10.93
PCMA	9.1	8.2

算法,要略高于TC-PINS算法的运行时间.这是因为PCMA算法在采用DBSCAN算法进行聚类后,还增加了复合物合并的后处理步骤,这会增加一定的时间开销.此外,在蛋白质复合物归属调整时,PCMA算法不仅需要对邻居发生变化的蛋白质节点进行复合物调整,还需要对其邻居节点进行蛋白质复合物归属判断.然而,从性能折中的角度来看,我们认为在保证蛋白质复合物挖掘准确性的前提下,牺牲算法的部分效率完全是可以接受的.总的来说,本文提出的PCMA算法还是具有不错的运行效率,完全可以应用到大规模蛋白质网络中.

6 结束语

在动态蛋白质网络中如何准确且高效地挖掘蛋白质复合物是目前生物信息学研究中的热点之一.文中首先提出一种基于动态图的动态蛋白质网络构建方法,然后在此基础上考虑蛋白质复合物的合并增益、增量节点和复合物归属增益等因素,进而设计了一种基于密度聚类的方法来挖掘蛋白质复合物.最后在多个生物数据集上验证了本文算法的有效性.在下一步工作中,我们将对未知蛋白质的功能进行预测研究,考虑到图卷积理论在处理大图数据上的优势,拟提出一种基于图卷积的蛋白质功能预测算法.

参考文献

[1] Larance M, Lamond A I. Multidimensional proteomics for cell biology[J]. Nature Reviews Molecular Cell Biology, 2015, 16(5): 269 – 280.

[2] Yang X, Coulombe-Huntington J, Kang S, et al. Widespread expansion of protein interaction capabilities by alternative splicing[J]. Cell, 2016, 164(4): 805 – 817.

[3] Lei X, Wang F, Wu F X, et al. Protein complex identification through Markov clustering with firefly algorithm on dynamic protein-protein interaction networks[J]. Information Sciences, 2016, 329: 303 – 316.

[4] Nepusz T, Yu H, Paccanaro A. Detecting overlapping protein complexes in protein-protein interaction networks[J]. Nature Methods, 2012, 9(5): 471 – 472.

[5] 李敏, 王晓桐, 罗慧敏, 等. 随机游走技术在网络生物学中的研究进展[J]. 电子学报, 2018, 46(8): 2035 – 2048.
Li M, Wang X T, Luo H M, et al. Progress on random walk and its application in network biology[J]. Acta Electronica Sinica, 2018, 46(8): 2035 – 2048. (in Chinese)

[6] 张媛, 贾克斌, 张爱东. 基于多视图融合的蛋白质功能

模块检测方法[J]. 电子学报, 2014, 42(12): 2337 – 2344.
Zhang Y, Jia K B, Zhang A D. Consistent protein functional module detection from multi-view of biological data [J]. Acta Electronica Sinica, 2014, 42 (12): 2337 – 2344. (in Chinese)

[7] Hegele A, Kamburov A, Grossmann A, et al. Dynamic protein-protein interaction wiring of the human spliceosome[J]. Molecular Cell, 2012, 45(4): 567 – 580.

[8] Tang X W, Wang J X, Liu B B, et al. A comparison of the functional modules identified from time course and static PPI network data[J]. BMC Bioinformatics, 2011, 12(1): 339.1 – 339.15.

[9] 雷秀娟, 高银, 郭玲. 基于拓扑势加权的动态PPI网络复合物挖掘方法[J]. 电子学报, 2018, 46(1): 145 – 151.
Lei X J, Gao Y, Guo L. Mining protein complexes based on topology potential weight in dynamic protein-protein interaction networks[J]. Acta Electronica Sinica, 2018, 46 (1): 145 – 151. (in Chinese)

[10] Shen X, Yi L, Jiang X, et al. Neighbor affinity based algorithm for discovering temporal protein complex from dynamic PPI network[J]. Methods, 2016, 110: 90 – 96.

[11] Lei X J, Fang M, Guo L, et al. Protein complex detection based on flower pollination mechanism in multi-relation reconstructed dynamic protein networks [J]. BMC Bioinformatics, 2019, 20(3): 63 – 74.

[12] Elfwing S, Uchibe E, Dova K. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning [J]. Neural Networks, 2018, 107: 3 – 11.

[13] Shen J, Hao X, Liang Z, et al. Real-time superpixel segmentation by DBSCAN clustering algorithm [J]. IEEE Transactions on Image Processing, 2016, 25 (12): 5933 – 5942.

[14] 乔少杰, 郭俊, 韩楠, 等. 大规模复杂网络社区并行发现算法[J]. 计算机学报, 2017, 40(3): 687 – 700.
Qiao S J, Guo J, Han N, et al. Parallel algorithm for discovering communities in large-scale complex networks [J]. Chinese Journal of Computers, 2017, 40(3): 687 – 700.(in Chinese)

[15] Lei H, Wen Y, You Z, et al. Protein-protein interactions prediction via multimodal deep polynomial network and regularized extreme learning machine [J]. IEEE Journal of Biomedical and Health Informatics, 2019, 23 (3): 1290 – 1303.

[16] Ruan P Y, Hayashida M, Akutsu T, et al. Improving prediction of heterodimeric protein complexes using combination with pairwise kernel [J]. BMC Bioinformatics, 2018, 19(1): 73 – 84.

[17] Pellegrini M, Baglioni M, Geraci F. Protein complex prediction for large protein interaction networks with the Core&Peel method [J]. BMC Bioinformatics, 2016, 17 (12): 372.1 – 372.30.

[18] Lei H, Wen Y, You Z, et al. Protein-protein interactions prediction via multimodal deep polynomial network and regularized extreme learning machine [J]. IEEE Journal of Biomedical and Health Informatics, 2018, 23 (3): 1290 – 1303.

作者简介

李　鹏　男,1983年11月出生,湖南泸溪人.博士、讲师,中南大学公共卫生与预防医学博士后流动站在站博士后.主要研究方向为生物信息学、机器学习、中医药大数据.
E-mail:lpchs617@csu.edu.cn

闵　慧　女,1986年12月出生,湖南湘潭人.硕士、讲师,主要研究方向为生物信息学、网络优化.
E-mail:mh1220@126.com

罗爱静(通信作者)　女,1962年出生,湖南安乡人.博士、教授、博士生导师,主要研究方向为医药信息管理、卫生信息管理、医药信息检索.
E-mail:805372510@qq.com