基于拓扑势加权的动态 PPI 网络复合物挖掘方法

雷秀娟[1],高 银[1],郭 玲[2]
(1. 陕西师范大学计算机科学学院,陕西西安 710119; 2. 陕西师范大学生命科学学院,陕西西安 710119)

摘 要: 从动态蛋白质相互作用(PPI)网络中挖掘蛋白质复合物是当前复合物挖掘研究的一个热点,但是目前大都采用未加权网络进行聚类分析,由于不能准确地描述网络的拓扑特性,因此其正确率不高. 鉴于此,本文提出采用拓扑势场的方法来构造加权网络,网络中的每一个蛋白质都被视作一个物理粒子,在它周围存在一个虚拟的作用场,由此网络中所有蛋白质的相互作用联合形成一个拓扑势场,文中定义了结点间的拓扑势的概念,并以此来构造加权网络,之后采用马尔科夫聚类算法在 DIP 数据和 Krogan 数据上进行复合物挖掘. 与其它经典算法相比,该方法的 *precision* 和 *f*-measure 值较高,能更好地识别蛋白质复合物.

关键词: 蛋白质相互作用网络; 拓扑势; 马尔科夫聚类
中图分类号: TP301.6 **文献标识码:** A **文章编号:** 0372-2112 (2018)01-0145-07
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2018.01.020

Mining Protein Complexes Based on Topology Potential Weight in Dynamic Protein-Protein Interaction Networks

LEI Xiu-juan[1], GAO Yin[1], GUO Ling[2]
(1. *College of Computer Science, Shaanxi Normal University, Xi'an, Shaanxi* 710119, *China*;
2. *College of Life Science, Shaanxi Normal University, Xi'an, Shaanxi* 710119, *China*)

Abstract: At present, many researchers focus on identifying protein complexes in the dynamic protein-protein interaction (PPI) network. But most of the methods have been used in unweighted network, the accuracy is not high since they could not accurately describe the network topology characteristics. In this paper, we put forward the topology potential to construct the weighted network. A protein can be regarded as a physical particle in the network which has a virtual field around it, and the interaction of all proteins forms a topology potential field. By computing the value of topology potential between nodes, the weighted network is constructed, and Markov clustering algorithm is used to identify protein complexes. The experimental results compared with the classic algorithms on DIP data and Krogan data indicate that their precision and *f*-measure value are higher and the proposed algorithm is more suitable to identify the protein complexes.

Key words: Protein-protein interaction network; the topology potential; Markov clustering

1 引言

近年来,生物网络的研究已经成为生物信息学的热点. 将蛋白质结点及其相互作用采用计算机图论理论抽象成一个复杂网络,用网络的理论来研究蛋白质分子内部的相互作用是研究热点之一. 传统图理论方法有基于划分的方法、基于层次聚类的方法、基于密度的局部搜索方法等. 基于划分的聚类方法目的是寻找一个最优划分,通过计算相似性得到最终的聚类模块. King 等人[1]提出基于代价函数的划分算法 RNSC,该算法给每一个可能的模块定义相应的代价值代表模块聚类的好坏,然后寻找代价值较小的模块. 其聚类结果依赖于初始划分的质量,且划分后的每个蛋白质只能属于一个功能模块,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,因此基于图划分的方法并不适合 PPI 网络的聚类分析. 基于层次的聚类方法能以树状的形式呈现蛋白质网络的层次结构. Pržulj 等人[2]提出基于图连通性的 HCS 算法,该算法通过反复迭代移除图中的最小割集,进而将整个网络分割成若干个独立的 cluster. Wang 等[3]提出一种快速层次聚类方法 HC – PIN 用于识别蛋白质复合物. 这些方法对噪声非常敏感,而且很难挖掘交叠蛋白质复

收稿日期:2016-05-16;修回日期:2016-12-09;责任编辑:李勇锋
基金项目:国家自然科学基金(No. 61672334, No. 61502290, No. 61401263);陕西省工业科技攻关项目(No. 2015GY016);中国博士后科学基金(No. 2015M582606)

合物. 基于密度的局部搜索方法通常将蛋白质复合物看作稠密子图,是最常用的复合物识别方法. Palla 等人[4]提出的 CPM 算法首先定义一组模块统计特征向量,搜索得到 k 个结点的完全子图,再比较两个子图之间共有结点,若两个子图中有 $k-1$ 个公共结点,则在新的图中对应加边,最后把连通的子图作为挖掘到的复合物. Li Min 等人[5]在密度的基础上引入距离作为复合物识别的参数,提出基于距离测定的复合物识别算法 IPCA. 这些方法可在扩充过程中允许某个蛋白质重复出现,但无法识别 PPI 网络中非稠密的子图结构.

由于传统的图理论方法都有自身的缺点,所以近年来学者们又提出一些其它的蛋白质复合物和功能模块挖掘算法,其中,Gavin 等人[6]提出基于核心-附件(COACH)的算法,一个复合物由一个核心组成部分和附件构成. 核心蛋白质是高度共表达的,每个附件绑定到核,从而形成具有生物特性的复合物. Nepusz 等人[7]提出一种名为 ClusterONE 的算法,能从 PPI 网络中找到重叠复合物. Lei 等人[8]提出基于连通强度的 PPI 网络蚁群优化聚类算法.

以上算法均是在无权网络中识别蛋白质复合物,近年来通过给蛋白质网络加权,得到较真实的网络结构再进行聚类的方法也不断涌现. 比如 Bader 等人[9]提出一种基于种子-扩充的局部搜索算法 MCODE,该算法首先给结点赋权值,然后通过计算结点的邻居结点,将权重高的结点扩展进来形成簇. Cho 等人[10]利用 GO 注释信息赋予每条蛋白质相互作用边一个权重,提出一种基于流的模块化算法,用于加权 PPI 网络中交叠复合物的挖掘. Peña 等人[11]用语义相似度获得边权重然后用一种图熵的方法来识别蛋白质复合物.

马尔科夫聚类算法(Markov Clustering,MCL)[12]是一种快速的可扩展无监督聚类算法,对于网络的变化具有很好的鲁棒性,抗噪声能力强,且能迅速而精确地从大规模蛋白质网络中侦测复合物,是目前比较好的一种聚类方法. 接着,V Satuluri 等人[13]提出一种正则化的马尔科夫聚类 R-MCL 对基本的 MCL 进行改进,但 R-MCL 并未考虑生成的蛋白质复合物中重叠的部分,Y K Shih等人[14]提出一种 SR-MCL 算法,将识别到的重叠复合物剔除.

蛋白质网络构建过程中,大量的蛋白质可以构建成一个无向无权图. 但由于生物实验的局限和计算方法的错误等原因,使得 PPI 网络数据不可避免地存在假阳性和假阴性. 而且由于生物网络中每个蛋白质有着不同的功能,不同的边重要性也不同,所有本文引入结点间拓扑势的概念,对 PPI 网络进行加权,最后用 MCL 算法识别蛋白质复合物,算法简记为 MCL-TP. 该算法在 DIP 数据和 Krogan 数据上进行实验,结果表明,文章中加权网络的构建能更为真实地体现网络结构,同时可以更有效地挖掘蛋白质复合物.

2 加权网络构建

本文引入文献[15]中的 3-sigma 法则来区分一个基因在时间序列中哪些时间点上是具有活性的,因为蛋白质只有在活跃状态时才与其它蛋白质有共同作用并执行某些功能,然后利用 PPI 数据和基因表达数据为 12 个时间序列构建相互作用网络,从而形成动态蛋白质网络.

2.1 拓扑势场

从数据场思想出发,我们将网络 G 看作一个包含 n 个结点及其相互作用的抽象物理系统,每个结点周围存在一个物理场,位于场中的任何结点都将受到其它结点的联合作用,与数据场的定义不同,网络中结点间的作用只能通过边来传递. 结点间的相互作用与结点属性以及结点间的网络距离密切相关,根据真实网络的模块化与抱团特征,我们认为,结点间相互作用具有局域特性,每个结点的影响能力会随着网络距离的增长而快速衰减,因此,倾向于采用代表短程场且具有良好数学性质的高斯势函数描述结点间的相互作用,并称相应的场为拓扑势场[16,17].

2.2 结点的拓扑势

无权网络反映了结点间的简单连接方式和相互作用的基本信息. 如果给无权网络每条边都赋予相应的权值,该网络就成为加权网络. 加权网络能够描述结点间联系的紧密程度,更真实、详尽地表达复杂网络的结构,所以本文引入拓扑势的概念对蛋白质网络进行加权. 网络中的每一个蛋白质都能被视作一个物理粒子,在它周围存在一个虚拟的作用场,由此网络中所有蛋白质的相互作用联合形成一个拓扑势场. 网络结点拓扑势的定义如下[16]:

给定网络 $G=(V,E)$,其中,$V=(v_1,\cdots,v_n)$ 为结点的非空有限集,$E\in V*V$ 为结点之间边的集合. 结点 $v_i\in V$ 的拓扑势定义为

$$\varphi(v_i)=\sum_{j=1}^{n}\left(m_j\times e^{-\left(\frac{d_{ij}}{\sigma}\right)^2}\right) \tag{1}$$

其中,m_j 表示结点 $v_j(j=1,2,\cdots,n)$ 的质量,用来描述每个结点的固有属性;d_{ij} 表示网络中两结点间的最短距离;σ 用于控制每个结点的影响范围.

通常,复杂网络中最优影响因子是通过势熵的计算获得. 根据高斯函数的数学性质,对于给定的 σ 值,每个结点的影响范围近似为$\lfloor 3\sigma/\sqrt{2}\rfloor$跳的局部区域,根据研究得知,每个结点的影响范围仅为 2 跳[16]. 所以我们认为蛋白质相互作用的最佳影响因子 σ 为 0.9428.

结点的拓扑势 $\varphi(v_i)$ 表示的是一个结点受到其它所有结点作用之和,拓扑势值越大,说明该结点在网络

中越重要. 而本文是从另外一个全新的角度出发,只考虑单个结点与结点之间的作用关系,计算结点之间的拓扑势值,拓扑势值越大,说明该边在网络中越重要. 根据所有结点间的拓扑势值得到相应的关系矩阵,从而形成加权的 PPI 网络.

2.3　结点间的拓扑势定义及加权网络构建

给定网络 $G=(V,E)$,其中,$V=(v_1,\cdots,v_n)$为结点,$E\in V*V$ 为结点间边的集合. 假定结点在顺着网络中的边能散发出一个作用场,则位于场内的任何结点都将受到其它结点的作用. 结点在网络结构中的拓扑位置,相当于结点所处的位势,反映了它影响结点(也反映了被结点影响)的能力. 本文首次采用结点间的拓扑势对蛋白质间的相互作用进行加权. 结点 v_i 和 v_j 间的拓扑势定义为

$$\varphi(v_i,v_j)=m_j\times e^{-\left(\frac{d_{ij}}{\sigma}\right)^2} \tag{2}$$

其中,m_j 表示结点 $v_j(j=1,2,\cdots,n)$的固有属性;d_{ij}表示网络中结点 i 和 j 之间的最短距离;σ 为控制每个结点间相互作用的衰减速度与范围的参数,本文将 σ 取值为 0.9428[16].

拓扑势的计算过程中,m_j 和 d_{ij}的取值都会对实验结果有很大的影响. 在关于拓扑势的复杂网络研究中,d_{ij}一般取结点间最短距离,m_j 的值一般取为 1[17],然而,在生物网络中,结点的固有属性具有丰富的生物含义,网络中的每个蛋白质有着不同的功能和特性,所以本文对它们进行重新定义.

(1)计算网络中结点间的最短距离 d_{ij}

文章在计算任意两结点间的最短距离时,基本思想是:从任意结点 i 到任意结点 j 的最短路径有 2 种可能,1 是直接从 i 到 j,2 是从 i 经过若干个结点 k 到 j:

①从任意一条单边路径开始,所有两点之间的距离是边的权,若两点之间没有边相连,则进行步骤②.

②若不满足①,则说明两结点 i 和 j 不直接相连,此时对于每一对顶点 i 和 j,看是否存在一个顶点 m 使得从 i 到 m 再到 j 比已知路径更短. 如果是更新它.

(2)计算 m_j 的值

由于不同的蛋白质有着不同的属性和功能,所以我们可以采用不同的方法定义 m_j. 本文取 m_j 值为 1 和网络中心性 NC[18] 作为 m_j 的固有属性分别进行了实验.

网络中心性 NC 定义如下:

$$m_j=NC(v_j)=\sum_{v_i\in N_{v_j}}\frac{Z_{v_i,v_j}}{\min(d_{v_i}-1,d_{v_j}-1)} \tag{3}$$

其中,$Z_{vi,vj}$表示网络中包含边 $e(v_i,v_j)$的三角形的数量;d_{vi}和 d_{vj}分别表示结点 v_i,v_j 的度;N_{vj}表示结点 v_j 的所有邻居结点的集合.

文中采用 $m_j=1$ 作为固有属性计算拓扑势的方法记为 TP,NC 作为固有属性计算拓扑势的方法记为 TP-NC. 通过式(2)计算结点间的拓扑势值,我们可以获得蛋白质网络中相应边的权重值,为网络加权. 图 1 为蛋白质网络加权后的示意图.

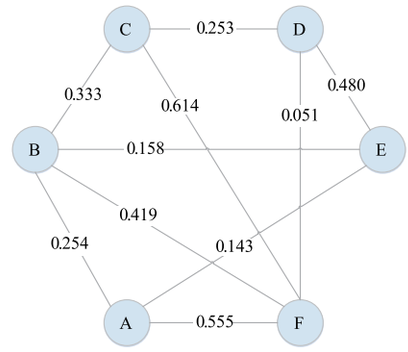

图1　蛋白质网络加权图

通常,我们使用的网络加权方法有:度中心性 DC、边聚集系数 ECC、GO 语义相似性、皮尔逊相关系数 PCC 等. DC 仅考虑各结点的邻居结点个数,而没有对网络进行全局性分析. ECC 计算任意两结点 i,j 的公共邻居个数,即网络中包含边 E_{ij}的三角形的个数,只考虑了图中的边关系. DC 和 ECC 是仅从某一拓扑特性出发考虑结点之间的连接关系,所以这两种方法对网络进行加权比较片面. GO 语义相似性是选取基因本体信息度量蛋白质结点间的功能相似性,构造功能相似性网络,该方法从功能角度出发,不再简单的使用拓扑特性,但是并未融合多种数据,同样不能很好地反应结点间的相互关系. PCC 加权是通过计算两列表达谱的皮尔逊相关系数得到表达谱的相似性,刻画出两个基因之间的相似关系,即从生物特性角度进行相似性描述. 由于 PPI 网络中结点之间关系和物理场中粒子之间的相互关系具有相似性质,本文从新的角度,定义了结点之间的拓扑势,使用拓扑势性质优化网络,能够较为真实地描述复杂网络的结构和蛋白质相互作用之间的可靠程度.

3　MCL-TP 算法

3.1　MCL 聚类

MCL 算法是基于马尔科夫链模型的一种图聚类方法. 它是通过构建邻接矩阵,通过扩张和膨胀这两种操作的交替,在蛋白质相互作用网络图中模拟随机游走. MCL 算法的主要过程为:

矩阵 $\boldsymbol{M}$ 定义为:

$$\boldsymbol{M}(i,j)=\frac{\boldsymbol{A}(i,j)}{\sum_{k=1}^{n}\boldsymbol{A}(k,j)} \tag{4}$$

其中 $\boldsymbol{A}$ 为添加自循环的邻接矩阵.

矩阵的扩张操作定义为:

$$Expand(\boldsymbol{M}) = \boldsymbol{M} \times \boldsymbol{M} \tag{5}$$

矩阵的膨胀操作定义为：

$$\boldsymbol{M}_{\inf}(i,j) = \frac{\boldsymbol{M}(i,j)^r}{\sum_{k=1}^{n} \boldsymbol{M}(k,j)^r} \tag{6}$$

其中 r 为非负实数. 对矩阵 $\boldsymbol{M}$ 的每列进行系数为 r 的幂运算和归一化计算.

3.2 MCL-TP 算法的基本思想

本文先利用基因表达数据和 PPI 数据分析出蛋白质的动态特性以及蛋白质之间相互作用的动态变化特性获得关系矩阵的表达形式得到动态子网络. 然后，根据数据场中势函数的定义，计算和分析蛋白质结点之间的拓扑势值，得到表示网络结点间拓扑势的邻接矩阵，从而形成加权的 PPI 网络.

在这个动态加权网络中，再对每个时间序列建立一个马尔科夫链模型，利用该模型的转移矩阵来描述该时间序列的动态特征，从而将对这些时间序列的聚类问题转换为对马尔科夫链模型的聚类问题. 利用马尔科夫聚类模型检索簇，获得基于时间片的动态加权 PPI 网络聚类结果.

3.3 伪代码描述

算法 MCL-TP 算法

输入：无权 PPI 网络 $G=(V,E)$，膨胀操作参数 r，prune 操作阈值 pru

输出：蛋白质复合物 PC

(1) $PC=\Phi$;

(2) FOR all edges E_{ij} in Matrix DO

(3) IF $v_i \in N_{vj}$//结点 v_i 是 v_j 的邻居结点

(4) $Node\_Inter = N(v_i \cap v_j)$//$v_i$ 和 v_j 共有邻居结点

(5) $Most\_tri = min(Degree_{vi}-1, Degree_{ij}-1)$;

//最多构成三角形的个数

(6) $Node\_Value\_DW = \sum_{vi \in Nij}(Node\_Inter/Most\_tri)$;

//计算结点加权度

(7) IF $M(i,j) < M(i,m) + M(m,j)$

(8) $M(i,j) = M(i,m) + M(m,j)$; //计算结点间的最短距离

(9) $M(i,j) = Node\_Value\_DW(i,j) * \exp(-(Network\_Distance(i,j)/sigma)\text{^}2)$;//计算 TP 的值，得到加权网络

(10) $\boldsymbol{M} = \boldsymbol{M} + I$; //$\boldsymbol{M}$ 为添加自循环的邻接矩阵

(11) $\boldsymbol{M} = Expand(\boldsymbol{M})$; //扩张操作

(12) $\boldsymbol{M} = Inflate(\boldsymbol{M}, r)$; //膨胀操作

(13) $\boldsymbol{M} = Prune(\boldsymbol{M}, pru)$;

(14) Repeat (11)(12)(13), until the Matrix to convergence;

(15) Return PC.

3.4 算法时间复杂度分析

该算法的时间复杂度由三部分组成. 计算结点加权度时，是把每个结点和邻居结点的交集统计一遍，所以对于 n 个结点的时间复杂度为 $O(n)$；然后通过一个图的权值矩阵求出它的每两点间的最短路径矩阵. 该算法必须计算 N 个矩阵 $\boldsymbol{D}_1, \boldsymbol{D}_2, \cdots, \boldsymbol{D}_n$，其中每一个矩阵包含 N^2 个元素，总共必须计算 N^3 个元素，所以时间复杂度为 $O(n^3)$；MCL 算法在进行扩张操作时，将矩阵和矩阵相乘，其时间复杂度为 $O(n^2)$. 所以该算法总的时间复杂度为 $O(n+n^3+n^2)$，其中，n 表示蛋白质结点个数.

3.5 算法的后处理

假设算法识别的复合物是 A，基准复合物是 B，若它俩的重叠率得分 $OS(A,B) \geqslant t$，(其中 t 为一个预先设定的阈值，根据文献[4]得出，OS 取 0.2 ~ 0.3 时可以过滤掉更多不重要的重叠蛋白质复合物，所以通常 t 取值为 $0.2^{[4,6]}$)，则表示预测蛋白质和已知蛋白质匹配，同时将不符合的蛋白质复合物剔除. OS 定义如下：

$$OS(A,B) = \frac{|V_A \cap V_B|^2}{|V_A| * |V_B|} \tag{7}$$

其中，V_A 和 V_B 是指复合物 A 和 B 中的数量.

采用 MCL-TP 算法聚类得到复合物后，根据式(7)计算预测到的复合物和已知基准类的匹配程度，当重叠率低于给定阈值时，则将该复合物剔除，所有预测到的蛋白质复合物检测完后，得到最终的蛋白质复合物结果.

4 实验仿真及分析

4.1 实验数据集

本文实验数据采用 DIP 数据集[19] 和 Krogan 数据集[20]，DIP 数据中列出了已知的成对蛋白质，成对蛋白质间的相互作用是指通过生物实验发现两个氨基酸链之间的绑定关系，有助于人们研究蛋白质网络的组织结构和复杂性. 本文使用的 DIP 数据是 DIP20140427 版本. 去除自相互作用和重复相互作用后，该网络中有 4995 个蛋白质和 21554 条边. Krogan 数据是用串联亲和纯化来处理 4,562 不同标签酵母蛋白质. 去除自相互作用和重复相互作用后，该网络中包含 2674 个蛋白质和 7075 条可靠的相互作用. 由于这两个原始网络数据集均没有权重，本文使用拓扑势对其进行加权. 为了评估预测的蛋白质复合物，本文将 CYC2008[21] 作为基准数据集，该数据集中包含 408 个通过生物方法预测到的蛋白质复合物，每个复合物包含两个或两个以上的蛋白质.

4.2 评价方法

本文采用正确率、查全率及它们的几何平均值 *f-measure* 来进行算法聚类效果的评价. 正确率(*precision*)是实验结果和标准数据库中最大匹配的结点的个数和实验结果中结点个数的比值. 查全率(*recall*)是实验结果与标准数据库中最大匹配的结点的个数与标准数据库中结点个数的比值. 计算公式如下：

$$precision(C|F)=\frac{MMS(C,F)}{|C|} \tag{8}$$

$$recall(C|F)=\frac{MMS(C,F)}{|F|} \tag{9}$$

其中,C 表示实验得到的聚类结果;F 表示标准数据库的结果;$|C|$表示测试得到的聚类的结点个数;$|F|$表示标准数据库中聚类结点的个数;*MMS* 表示实验结果与标准数据库中最大匹配的结点个数.

综合考虑正确率和查全率两方面,提出了 *f-measure* 指标来评价模块的准确性. 公式如下:

$$f-measure=\frac{2\times precision\times recall}{precision+recall} \tag{10}$$

4.3　实验结果分析

4.3.1　DIP 数据实验结果

图 2 是使用不同加权方法实验挖掘到的蛋白质复合物与标准库的对比结果. 从图中可以看出, ECC 加权后的方法未能识别 YDR359C 和 YGR002C 两个蛋白质,是因为其只考虑了结点间边的关系;DC 加权后的方法有 3 个结点未能预测到,而且 YEL018W 结点中错误挖掘出了几个结点,该方法对蛋白质网络数据的分析较为单一,所以得到不完全匹配的结果;而使用 TP-NC 加权考虑了蛋白质结点的拓扑特性,同时结合了蛋白质的固有属性和网络中结点间的距离对该网络进行分析,减小了实验数据与真实网络间的差距,所以 TP-NC 加权后得到较好的聚类结果.

为了进一步衡量不同加权方法的性能,本文还用 *precision-recall* 曲线评价方法对不同的查全率水平绘制查准率的曲线图. 如图 3 所示,当 *recall* 值增大时,*precision* 的值都在变小. 当 *recall* 值为 0.45 ~ 0.6 时,MCL-ECC 的 *precision* 值更高,而 *recall* 取不同值时,MCL-TP-NC 方法都有较高的质量较好,该算法挖掘蛋白质复合物性能较优.

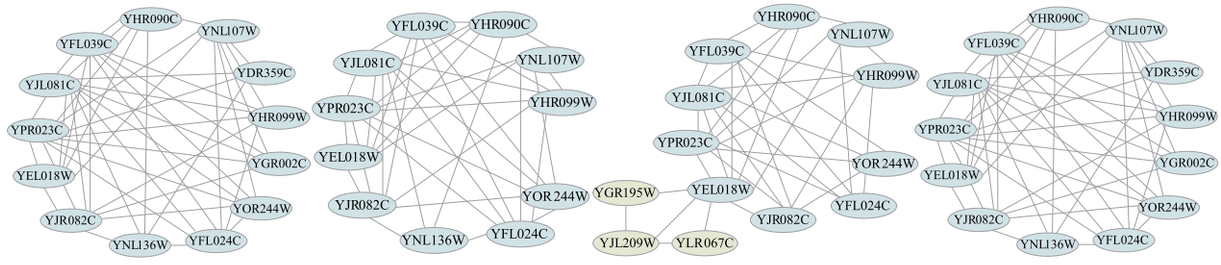

图2　不同加权方法识别NuA4 Histone Acetyltransferase Complex复合物与基准类的对比

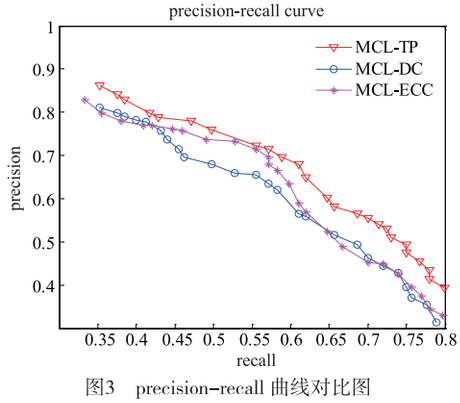

图3　precision–recall 曲线对比图

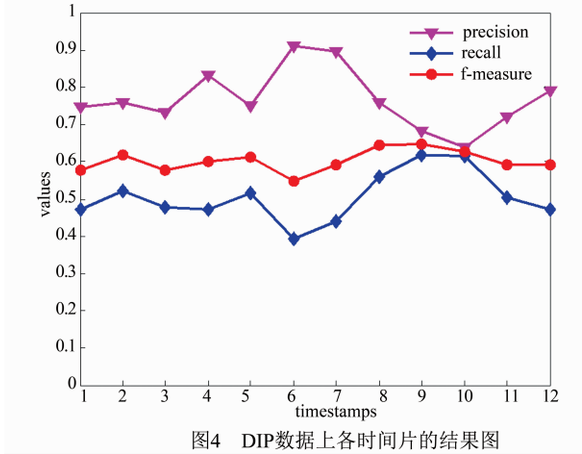

图4　DIP数据上各时间片的结果图

图 4 是在 DIP 数据上 12 个时间片中的 *precision*、*recall* *f – measure* 实验结果. 表 1 是不同聚类算法在 DIP 数据上的性能比较. 表中算法上标是指算法首次提出的文献序号,用于对比的实验结果参考文献如下(表 2 同):MCODE, DPClus, IPCA, Weighted Graph Entropy 来源于文献[11];MCL, R-MCL, SR-MCL 来源于文献[14];CPM,COACH 源于文献[22].

从表 1 可看出,部分算法(如 CPM、COACH、DPClus 等)中,并未对网络数据进行预处理,由于数据本身存在很大误差,不能很好的反映真实网络结构,同时由于算法自身的缺点而导致预测结果并不理想. Weighted Graph Entropy 使用语义度加权后用图熵的方法识别复合物,其只考虑了蛋白质功能,且使用图熵的方法聚类性能本身并没有 MCL 算法好,所以得到的正确率不高,使用 MCL 算法在原始数据中进行聚类,由于数据本身噪声太多,不

可避免的存在较大误差,导致算法性能不能很好体现. 而在动态加权的网络中,使用 MCL 算法聚类效果明显较优. MCL-DC、MCL-ECC 两种加权方法能有效优化原有数据结构,使数据更贴合真实数据,从而得到较高的正确率,但 DC 和 ECC 加权考虑片面,而 TP 加权考虑了结点间距离和固有属性,描述了网络中结点间的影响能力,能有效提高蛋白质相互作用的可靠性,得到的 *f-measure* 较高,尤其是 *precision* 的值比较理想. 但 TP 加权时 m_j 的不同取法对实验结果也有一定的影响,当 $m_j=1$ 时,聚类结果并不理想,而使用 NC 作为固有属性既考虑了结点之间的关系又考虑了边关系,所以挖掘到的复合物正确率更高,获得更好的聚类结果.

4.3.2 Krogan 数据实验结果

为评价新方法的性能,本文又在 Krogan 数据集上进行了实验,并与其它方法进行了比较,如表 2,其中 MCODE、ClusterONE、HUNTER、METHOD 来源于文献[23];COACH、COAN、CMC 来源于文献[24].

表 1 各算法在 DIP 数据上的性能比较

Algorithm	*precision*	*recall*	*f-measure*
CPM[4]	0.3429	0.2593	0.2953
COACH[6]	0.3829	0.5818	0.4618
MCODE[9]	0.2050	0.2980	0.2430
DPClus[11]	0.3130	0.3030	0.3080
IPCA[5]	0.2180	0.2880	0.2480
Weighted Graph Entropy[11]	0.4710	0.2430	0.3210
MCL[12]	0.3569	0.3879	0.3717
R-MCL[13]	0.3814	0.4866	0.4276
SR-MCL[14]	0.4443	0.5709	0.4997
MCL-DC 加权	0.6289	0.5312	0.5759
MCL-ECC 加权	0.6471	0.5268	0.5786
MCL-TP 加权	0.6818	0.5183	0.5889
MCL-TP-NC 加权	0.7380	0.5118	0.6045

表 2 各算法在 Krogan 数据上的性能比较

Algorithm	*precision*	*recall*	*f-measure*
MCODE[9]	0.72	0.16	0.26
ClusterONE[7]	0.58	0.33	0.42
HUNTER[23]	0.87	0.20	0.32
METHOD[23]	0.46	0.59	0.52
COACH[6]	0.62	0.34	0.44
COAN[24]	0.71	0.33	0.45
CMC[24]	0.75	0.24	0.36
MCL-DC 加权	0.6498	0.4954	0.5622
MCL-ECC 加权	0.6744	0.4933	0.5698
MCL-TP 加权	0.7386	0.4780	0.5804
MCL-TP-NC 加权	0.8458	0.4699	0.6041

图 5 反映的是在 Krogan 数据中 12 个时间片中的 *precision*、*recall*、*f-measure* 实验结果. 从表 2 可以看出,MCL-TP-NC 方法的结果优于其它所有方法,从不同的加权方法中可以看出,TP-NC 加权可以有效识别蛋白质复合物,聚类结果最优,再次证明 m_j 取值的重要性和 NC 作为固有属性的合理性,但是由于算法识别的复合物总数比较少,所以 *recall* 值并不理想.

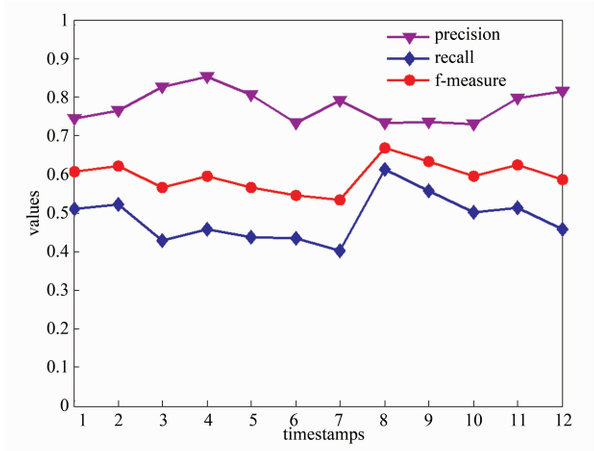

图5 Krogan数据上各时间片的结果图

5 总结

文章使用 3-sigma 准则构建动态网络,然后用结点间拓扑势的方法给 PPI 网络加权,对网络进行优化,得到动态加权网络,较好地体现了网络生物特性和结构特性,之后采用 MCL 算法挖掘蛋白质复合物,分别在 DIP 数据和 Krogan 数据上进行仿真实验,并将预测到的复合物与基准类进行对比分析. 从以上分析得出,在动态加权网络中识别复合物优势明显,但不同的加权方法会对结果产生影响,文中采用拓扑势的思想使网络中蛋白质之间的相互关系更准确,降低了数据假阳性和假阴性带来的负面影响,再结合快速高效的 MCL 算法,使得 MCL – TP 算法挖掘到的复合物与标准数据库匹配度较高,所以得到较高的 *precision* 值,相应的 *f-measure* 的值也较优. 但是由于算法对 *recall* 的值提高不明显,且 TP 中固有属性对结果影响明显,所以下一步将对该算法中距离参数和结点的固有属性进行改进,使算法具有更高准确率,且将动态网络加权思想运用于关键蛋白识别和疾病预测等相关研究中.

参考文献

[1] A D King, N Pržulj, I Jurisica. Protein complex prediction via cost-based clustering [J]. Bioinformatics, 2004, 20(17):3013 – 3020.

[2] Pržulj N, Wigle D A, Jurisica I. Functional topology in a network of protein interactions [J]. Bioinformatics, 2004, 20(3):340 – 348.

[3] J Wang, M Li, J Chen, Y Pan. A fast hierarchical clustering algorithm for functional modules discovery in protein interaction networks [J]. IEEE/ACM Trans Comput Biol Bioin-

format,2011,8(3):607 – 620.
[4] Palla G,Derenyi I,Farkas I,et al. Uncovering the overlapping community structure of complex networks in nature and society[J]. Nature,2005,435(7043):814 – 818.
[5] Min Li,Jian-er Chen,Jian-xin Wang,Bin Hu,Gang Chen. Modifying the DPClus algorithm for identifying protein complexes based on new topological structures[J]. BMC Bioinformatics,2008,9(1):398.
[6] M Wu,X L Li,C K Kwoh,S K Ng. A core-attachment based method to detect protein complexes in PPI networks [J]. BMC Bioinformatics,2009,10(1):1 – 16.
[7] Nepusz T,Yu H,Paccanaro A. Detecting overlapping protein complexes in protein-protein interaction networks [J]. Nature Methods,2012,9(5):471 – 475.
[8] 雷秀娟,黄旭,吴爽,郭玲. 基于连通强度的 PPI 网络蚁群优化聚类算法[J]. 电子学报,2012,40(4):695 – 702.
Lei xiujuan,Huang Xu,Wu Shuang,Guo Lin. Joint strength based ant colony optimization clustering algorithm for PPI networks[J]. Acta Electronica Sinica,2012,40(4):695 – 702. (in Chinese)
[9] G D Bader,C W Hogue. An automated method for finding molecular complexes in large protein interaction networks [J]. BMC Bioinformatics,2003,4(2):1471 – 2105.
[10] Cho Y R,Hwang W,Ramanmathan M,et al. Semantic integration to identify overlapping functional modules in protein interaction networks [J]. BMC Bioinformatics, 2007,8(1):265.
[11] F I Peña,Y R Cho. Improvements of graph entropy approach to detect protein complexes by ontological analysis of PPIs[A]. IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW)[C]. IEEE,2012. 211 – 217.
[12] Van Dongen S M. Graph clustering by flow simulation [D]. Utrecht:University of Utrecht,2000.
[13] V Satuluri,S Parthasarathy. Scalable graph clustering using stochastic flows:Applications to community discovery[A]. KDD'09[C]. New York,2009. 737 – 746.
[14] Y K Shih,S Parthasarathy. Identifying functional modules in interaction networks through overlapping Markov clustering[J]. Bioinformatics,2012,28:473 – 479.
[15] J Wang,X Peng,M Li,Y Luo,Y Pan. Construction and application of dynamic protein interaction network based on time course gene expression data [J]. Proteomics, 2013,13(2):301 – 312.
[16] Min Li,Yu Lu,Jianxin Wang,Fang-Xiang Wu,Yi Pan. A topology potential-based method for identifying essential proteins[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics,2015,12(2):372 – 383.
[17] 张冰,杨静,张健沛,谢静. 面向聚类分析的邻域拓扑势熵数据扰动方法[J]. 哈尔滨工程大学学报,2014,35(9):1149 – 1155.
Zhang Bing,Yang Jing,Zhang jianpei,Xie Jing. A neighborhood topological potential entropy data perturbation method for clustering analysis[J]. Journal of Harbin Engineering University,2014,35(9):1149 – 1155. (in Chinese)
[18] J Wang,M Li,H Wang,et al. Identification of essential proteins based on edge clustering coefficient[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics,2012,9(4):1070 – 1080.
[19] L Xenarios,L Salwinski,X Duan,P Higney,S Kim,et al. DIP:the database of interaction proteins:a research tool for studying cellular networks of protein interactions[J]. Nucleic Acids Research,2002,30(1):303 – 305.
[20] Krogan N J,et al. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae[J]. Nature,2006, 440(7084):637 – 643.
[21] Pu S,Wong J,Turner B,et al. Up to date catalogues of yeast protein complexes [J]. Nucleic Acids Research, 2009,37(3):825 – 831.
[22] Shuliang Wang,Fang Wu. Detecting overlapping protein complexes in PPI networks based on robustness[J]. Proteome Science,2013,11(1):S18.
[23] YJ Zhang,HF Lin,ZH Yang,et al. A method for predicting protein complex in dynamic PPI networks[J]. BMC Bioinformatics,2016,17(Suppl 7):229.
[24] Yijia Zhang,Hongfei Lin,Zhihao Yang,Jian Wang. Construction of dynamic probabilistic protein interaction networks for protein complex identification[J]. BMC Bioinformatics,2016,17(1):1 – 13.

作者简介

雷秀娟　女,1975 年出生于陕西西安,教授,博士生导师,研究方向为智能计算、生物信息计算等.
E-mail:xjlei168@163.com

高　银　女,1991 年出生于陕西榆林. 陕西师范大学计算机科学学院 2014 级硕士研究生,研究方向为群智能优化算法、生物信息计算.
E-mail:gy_sophia@snnu.edu.cn

郭　玲　女,1980 年出生于陕西略阳. 陕西师范大学生命科学学院 2011 级博士研究生. 主要研究方向为生物信息学.
E-mail:glalguo2@163.com