基于蛋白质网络的模块动态特性挖掘研究*

郭 阳，尚学群，李 晶
（西北工业大学 计算机学院，西安 710072）

摘 要：目前对于蛋白质网络的多数分析方法都是基于静态网络框架提出的，然而，事实上蛋白质之间的相互作用关系是随着时间变化的，具有明显的动态特性。从动态角度出发，基于蛋白质交互网络，结合时间序列的基因表达数据，揭示蛋白质功能模块变化的动态特性，提出了一种动态网络的构建方法及在动态快照网络中挖掘功能模块的算法，采用模糊匹配的方法揭示了蛋白质功能模块的变化过程。实验证明，该方法能够有效地揭示蛋白质功能模块的动态变化特性。

关键词：蛋白质网络；局部相关性；功能模块；动态特性
中图分类号：TP311 **文献标志码**：A **文章编号**：1001-3695(2012)12-4495-05
doi:10.3969/j.issn.1001-3695.2012.12.024

Analyzing dynamic properties of modules in PPI network

GUO Yang, SHANG Xue-qun, LI Jing
(*School of Computer Science, Northwestern Polytechnical University, Xi'an* 710072, *China*)

Abstract: Although many methods were proposed to analyze PPI network in past years, most of them are based on static network analysis framework. However, the PPI network is dynamic changing in living organisms. This paper analyzed the dynamic properties of modules in PPI network. Based on the time-series gene expression, it constructed the dynamic PPI snapshots and proposed a novel algorithm to mine functional modules in the snapshot networks. Thus, it adopted a fuzzy matching method to reveal the dynamic properties of modules. Experimental results illustrate the proposed method is helpful to analyze the dynamic properties of modules in PPI network.

Key words: PPI network; local similarity; functional module; dynamic property

0 引言

蛋白质交互(protein-protein interaction,PPI)网络分析是后基因时代蛋白质组学研究的一项重要内容。对蛋白质网络进行分析研究，挖掘蛋白质相互作用的潜在规则、结构特性等信息是蛋白质网络分析的一项重要任务，也是后基因时代蛋白质组学研究的一项重要难题。蛋白质结构和功能的多样性决定了蛋白质之间相互作用的复杂性，实验表明，实际中的蛋白质相互作用网络是具有小世界特性的复杂网络。提出简单、有效的分析、挖掘方法来揭示蕴藏在这些蛋白质相互作用背后的潜在规律或特征是摆在众多从事数据挖掘相关研究工作者面前的一项具有挑战性的任务。目前，对于蛋白质网络分析的主要方法大多集中在功能模块挖掘、模体识别、路径发现、网络特性分析等几个方面。所谓功能模块是指网络中蛋白质之间相互作用相对密切、共同协调完成某种具有特定意义的生物功能而形成的模块化结构。模体是指网络中出现的较为频繁的相互作用的模式。蛋白质之间相互作用是具有一定先后顺序的，路径发现的任务便是在蛋白质网络中寻找具有明显顺序特征的作用路径。在过去的一段时间里，很多方法已被提出并被用来分析蛋白质交互网络，其中大部分方法都是基于经典图论中的度量衡，如度分布、紧密度、最短路径、中介性等来对网络进行分析。随着研究的深入和测序技术的发展，从某种角度来说，人们更加倾向于对具体生物过程的揭示和理解，因为高效、准确地揭示具体生物过程对于人们了解生命活动的内在机制具有重要意义。虽然利用以前这些分析方法一定程度上能够揭示一定的生物现象，但要揭示具体的生物过程却存在其局限性。这种局限性主要是由于它们对于蛋白质网络的整体分析是基于静态的网络框架进行的。然而在具体的生物过程中，蛋白质之间的相互作用却是随着时间的进行而动态变化的。这就要求必须从动态的角度出发，依动态的思想来分析蛋白质交互网络。目前，已有一些基于动态思想的分析方法被陆续提出，其中，Jin 等人[1]尝试从动态的角度出发，挖掘动态功能模块；You 等人[2]试图以学习的方法在动态的网络中识别一定的作用模式；Rohian 等人[3]提出一种融合了时间信息的动态交互规则的发现方法。由于在动态条件下，原来在静态网络中分析得到的相关信息可能也是动态变化的，而不是一成不变的，这决定了动态网络分析的复杂性。据笔者所了解的情况而言，对于动态网络分析目前还没有形成一个较为完善的研究框架，许多研究都是基于尝试性的探索阶段。

本文从动态角度出发，基于蛋白质交互网络，结合时间序列的基因表达数据，试图揭示蛋白质网络中功能模块的动态特性。首先提出了一种基于时间序列基因表达数据的动态网络

收稿日期：2012-03-27；**修回日期**：2012-05-11 **基金项目**：国家“973”计划资助项目(2012CB316203)；国家自然科学基金资助项目(61033007)；西北工业大学基础研究资助项目(JC201042)

作者简介：郭阳(1986-)，男，硕士研究生，主要研究方向为数据挖掘(gyangnetsky@163.com)；尚学群(1973-)，女，教授，博士，主要研究方向为数据挖掘、数据库技术、生物信息学等；李晶(1987-)，女，硕士研究生，主要研究方向为数据挖掘.

快照构建方法以及在网络快照中挖掘相对稀疏的蛋白质功能模块的新型算法;随后,试图采用模糊匹配的方法来揭示蛋白质功能模块在生物过程中的具体变化过程。实验表明,本文的分析方法能够较好、有效地揭示功能模块的动态变化特性。

1 问题定义

本文将整个 PPI 网络以图 $G(V,E)$ 的形式表示。其中:V 为 PPI 网络中所有蛋白的集合;E 为 PPI 网络中所有交互边的集合。由于在动态条件下,各蛋白质之间的相互作用都是随时间动态变化的,所以整个网络的拓扑结构也是动态变化的。为分析问题,有以下相关定义。

定义 1 时间区间对。对于图 G 中任意的一条边 $e=(v_1,v_2)\in E$,e 的一个时间区间对表示为 $r(e)=[i,i']:[j,j']$。其中:i、j 分别为 v_1、v_2 一次相互作用的开始时间;i'、j' 分别为 v_1、v_2 一次相互作用的终止时间。

由于蛋白质之间的相互作用具有一定的次序性,所以对于任意边的一个时间区间对来说,开始和终止的时间点可能各不相同。

定义 2 存在时间集合。对于图 G 中任意一条 $e=(v_1,v_2)\in E$,其所有时间区间对的集合为 $S(e)=\{[a_1,a_1']:[b_1,b_1'],[a_2,a_2']:[b_2,b_2'],[a_3,a_3']:[b_3,b_3'],\cdots\}$,则 e 的存在时间集 $ET(e)=\{[t_1,t_1'],[t_2,t_2'],[t_3,t_3'],\cdots\}$,其中,$[t_1,t_1']=[a_1,a_1']\cup[b_1,b_1']$,$[t_2,t_2']=[a_2,a_2']\cup[b_2,b_2']$,$[t_3,t_3']=[a_3,a_3']\cup[b_3,b_3']$,$\cdots$。

实际蛋白质网络中,蛋白质之间的相互作用是动态变化的,边的存在集合描述了任意相互作用边的存在时间区间。

2 分析方法及过程

2.1 局部相关性分析及时间区间追踪

2.1.1 局部相关性分析

局部相关性分析方法用来匹配计算两个向量之间最优的相关性。这种方法与以往的相关性分析(如皮尔逊相关系数等)的最大不同在于它从局部出发,能够从局部的变化趋势中发现最优的相关性。该方法由 Ruan 等人[4,5]提出,并将其成功应用于实际的问题分析中。为便于理解,以下对该方法主要步骤作简要阐述。

a)向量规整化。该方法采用正态变换[4]的方法对原始向量进行规整化。对于任意向量 $\boldsymbol{X}=(x_1,x_2,x_3,\cdots,x_n)$,假设对其进行排序后的向量为 $\boldsymbol{S}^x=(S_1^x,S_2^x,\cdots,S_N^x)$,则其正态变换后的向量为 $\boldsymbol{X}'=(x_1',x_2',x_3',\cdots,x_n')$,其中 $x'_i=\Phi_{-1}(S_i^x/(N+1))$,$i=1,\cdots,n$。

b)相关性计算。当对原始待比较向量进行规整化后,局部相关性通过构造两个向量的正、负相关矩阵来计算两个向量的相关系数。由于基于局部元素的比较,该方法将在预先定义的误差范围 D 内计算最大的相关系数。对于任意两个规整化的向量 $\boldsymbol{v}_1=(v_{11},v_{12},\cdots,v_{1n})$,$\boldsymbol{v}_2=(v_{21},v_{22},\cdots,v_{2n})$,则它们的正相关矩阵 $\boldsymbol{P}_{n\times n}$ 和负相关矩阵 $\boldsymbol{N}_{n\times n}$ 按如下步骤计算:

(a)对于任意 $i,j=1,\cdots,n$,$P_{0,j}=P_{j,0}=0$,$N_{0,j}=N_{j,0}=0$;

(b)对于 $i,j=0,\cdots,n$ 且 $|i-j|\leqslant D$,则

$$P_{i+1,j+1}=\max\{0,P_{i,j}+v_{1,i+1}\cdot v_{2,j+1}\}$$
$$N_{i+1,j+1}=\max\{0,N_{i,j}+v_{1,i+1}\cdot v_{2,j+1}\}$$

(c)$P(\boldsymbol{v}_1,\boldsymbol{v}_2)=\max\limits_{1\leqslant i,j\leqslant n}P_{i,j}$;$N(\boldsymbol{v}_1,\boldsymbol{v}_2)=\max\limits_{1\leqslant i,j\leqslant n}N_{i,j}$;

(d)$\mathrm{maxScore}(\boldsymbol{v}_1,\boldsymbol{v}_2)=\max(P(\boldsymbol{v}_1,\boldsymbol{v}_2),N(\boldsymbol{v}_1,\boldsymbol{v}_2))$;

$\mathrm{flag}(\boldsymbol{v}_1,\boldsymbol{v}_2)=\mathrm{sgn}(P(\boldsymbol{v}_1,\boldsymbol{v}_2)-N(\boldsymbol{v}_1,\boldsymbol{v}_2))$;

(e)$R(\boldsymbol{v}_1,\boldsymbol{v}_2)=\mathrm{maxScore}(\boldsymbol{v}_1,\boldsymbol{v}_2)/n$。

其中:$R(\boldsymbol{v}_1,\boldsymbol{v}_2)$ 为 $\boldsymbol{v}_1$、$\boldsymbol{v}_2$ 的局部最优相关系数,$\mathrm{flag}(\boldsymbol{v}_1,\boldsymbol{v}_2)$ 用来标示相关性的符号。

2.1.2 时间区间追踪

通过局部相关性分析,可以得到任意两个等长向量在误差 D 范围内的相关矩阵 $\boldsymbol{P}_{n\times n}$ 和 $\boldsymbol{N}_{n\times n}$,如何从这两个矩阵中获取它们在整个时间序列中体现相关性的具体时间信息是人们所关注的另一个重要问题。由相关矩阵的计算可以看出,如果两个向量的变化趋势相同,则随着它们在误差 D 范围内的对应匹配,将导致它们的正相关矩阵 $\boldsymbol{P}_{n\times n}$ 中的相应元素逐步递增;同理,如果两个向量变化趋势相反,则将导致它们的负相关矩阵 $\boldsymbol{N}_{n\times n}$ 中的相应元素逐步递增。由此可见,相关矩阵中较大元素的出现是由于在它们之前的相应变化趋势的匹配所导致的。基于这种思路,本文试图沿导致这一较大元素变大的匹配路径追踪到它们趋势开始匹配的时间点。本文的具体追踪步骤描述如下:

a)在正、负相关矩阵未标记区中找出最大元素 m_{ij}。若最大元素在 $\boldsymbol{P}_{n\times n}$ 中出现,则其相关性标记为“+”;若最大元素在 $\boldsymbol{N}_{n\times n}$ 中出现,则其相关性标记为“-”。

b)从当前最大元素 m_{ij} 开始,沿向前的方向追踪到比当前元素小的另一元素;重复此过程,直到此过程不能继续进行为止,此时假设到达元素为 m_{st}。

c)当前该最优匹配的时间区间为 $[s,i]:[t,j]$;在正、负相关矩阵中标记所有下标处于 $[s,i]$ 和 $[t,j]$ 之间的所有元素。

d)返回步骤 a),寻找次优匹配时间区间,直到正、负相关矩阵中所有的元素都被标记过为止。

步骤 b)中,所谓向前是指,对于任意元素 m_{ij},其向前的方向只有三个,即 $m_{i-1,j}$,$m_{i-1,j-1}$,$m_{i,j-1}$。

假设待比较向量 $\boldsymbol{a}$、$\boldsymbol{b}$,其一个相关性矩阵如图 1 所示,按照以上描述方法,则当前最大元素为 5,按照上面所介绍的时间追踪方法,当按照箭头所示方向到达 2 时,一次时间区间追踪结束。此时,获得的一个时间区间为 $[2,4]:[2,4]$。

2.2 动态快照网络构建

通过时间追踪步骤,可以获得网络中任意一条边的存在时间集合,进而可以据此构建任意时间点上的快照网络,具体构建方法由定义 3 给出。

定义 3 对于任意时刻 $t=t_0$,网络 G 中任意一条边 e 在该时刻存在,当且仅当 t_0 被边 e 的存在时间集合中任意存在时间区间所包含。

根据定义 3,可以在原始蛋白质网络的基础上构建出任意时间点上的快照网络。

2.3 功能模块挖掘

在静态条件下,对于功能模块的挖掘,多数算法都是基于紧密度的概念提出来的。假设网络中的点之间交互越紧密,它们就越有可能是协调完成某项功能的相似功能蛋白族,从而组成一个相对紧密的团。大量实验已证明这种假设具有其合理性,并且这种模块化的结构普遍存在于复杂网络中,如社交网络、蛋白质网络等。然而在动态条件下,原来静态网络中的一

些交互边可能已经不存在,这就可能导致一个问题,即在动态条件下可能网络整体就相对比较稀疏,如果仍使用基于紧密度的方法来挖掘功能模块,则可能漏掉一些实际相对稀疏但确实具有功能相似性的功能模块。这种情况在蛋白质网络中尤为常见,且被大量实验所证明。例如,Jin 等人[1]证明在酵母蛋白质网络中存在大量相对稀疏的功能模块。

为能在动态条件下挖掘出相对稀疏的功能模块,本文提出了一种基于扩展修饰的模块挖掘方法——SFMM 算法,相关定义如下:

定义 4　种子点。对于网络 G 中的任意点 P,假设 P 有 n 个直接邻居节点,P 及其所有直接邻居节点之间有 m 条边,若 $\gamma_{(P)}=2m/n(n+1)\geqslant\sigma$,则 P 为一个种子节点。其中 σ 是一个参数。

定义 5　紧密度。对于模块 M,若 M 包含 n 个顶点,m 条边,则 M 的紧密度定义为 $D_{(M)}=2m/n(n-1)$。

定义 6　基本模块。对于网络 G 中任意种子节点 P,则由 P 及其直接邻居和它们之间的所有边组成的模块称为由 P 决定的基本模块。

定义 7　初级模块。对于一个基本模块 M,对其进行扩展后(定义 8),M 将成为一个初级模块。

定义 8　扩展规则。对于基本模块 M,对与其直接相连的点 P,若 P 的度为 n,其与 M 中的点相连接的边数为 m。如果 $r=m/n\geqslant\varepsilon$,则可以将 P 及其与 M 相关的边加入到 M。

由于需要计算待考察点的与当前模块相连的边和其自身全部边的比例,这就要求当前模块中的点必须在一定范围内保持相对稳定。所以对于基本模块的扩展,本文采用层次结构的扩展方法,即考察完所有的与当前模块直接相连的点后再将该层次所有满足扩展条件的点及相关边加入待扩展模块。当 $\sigma=0.5$,$\varepsilon=0.5$ 时,一个简单例子如图 2 所示。

a \ b	1	2	3	4
1	3	4	0	0
2	4	2	4	0
3	0	5	3	5
4	0	4	4	5

图 1　时间区间追踪实例

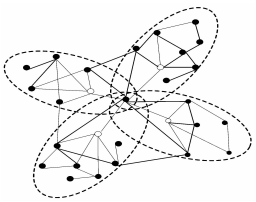

图 2　初级模块实例

因 r 一定程度上反映了一个待考察点属于当前模块的可能性,一般情况下取 ε 不小于 0.5 较为合适。

通过上面的定义,按照定义 8 的扩充规则,如果 ε 的取值不是很大,则可以尽可能大地将那些有可能与当前待扩展模块中的点较为相似的考察点包含到当前模块中来。然而可以预料到,当 ε 不是很大时,这样扩展得到的初级模块将有可能非常地稀疏。虽然实验已证明在网络中存在大量的稀疏模块,然而,交互的紧密程度确实能在一定程度上反映点之间交互的倾向性和相似性。为使挖掘所得到的模块不至于太稀疏(太稀疏则其之间功能相似的可能性相对较小),需要对经过扩展得到的初级模块进行修饰,以保证其相对紧密的特性。对初级模块进行修饰的过程可以通过删除当前具有最小度的点及其相关边开始。具体的挖掘功能模块的方法可以用算法 1 来描述。

算法 1　SFMM(sparse functional modules mining)

输入:$G(V, E)$,ε,σ,ω;
输出:功能模块集合 M。

1. for G 中任意点 P_i
2. if $\gamma(P_i)\geqslant\sigma$ then 将 P_i 放入优先队列 Q;
3. end for
4. while Q 不为空 do
5. 从 Q 中取出一个种子节点 P_j,对由 P_j 所决定的基本模块 M_j 按层次结构进行扩展,直到没有满足条件的点可以加入 M_j;
6. while $D_{(M_j)}\leqslant\omega$ do
7. 从 M_j 中删除具有最小度的点及其相关边;
8. end while
9. 输出 M_j 到集合 M;
10. end while
11. 对 M 中的重叠模块进行合并;
12. end

利用 SFMM 算法,可以挖掘网络中的功能模块。由于动态条件下紧密的功能模块相对较少,而相对稀疏的功能模块可能大量存在。为能够挖掘出相对稀疏的功能模块,在修饰初级模块时紧密度参数 ω 设置得不宜过高。

2.4　匹配模块动态变化过程

利用 SFMM 算法可以从已知网络中挖掘出相对较稀疏的功能模块。将该算法用于任意时刻的动态快照网络中,可以得到在该时刻存在的模块集合。如果想要揭示某个模块在整个时间序列上的动态变化特性,则需要在整个时间序列上的快照网络的模块集合中来匹配。基于该思想,有如下有关模块动态变化关系的相关定义。

定义 9　增长关系。对于 t 时刻快照网络模块集合 S_t 中的任意模块 M,若在 $t+1$ 时刻快照网络的模块集合 S_{t+1} 中存在一个包含 M 的模块 M',则将 M 从 t 时刻到 $t+1$ 时刻变化为 M'的演化关系标记为增长关系。

定义 10　衰减关系。对于 t 时刻快照网络模块集合 S_t 中的任意模块 M,若在 $t+1$ 时刻快照网络的模块集合 S_{t+1} 中存在一个被 M 所包含的模块 M',则将 M 从 t 时刻到 $t+1$ 时刻变化为 M'的演化关系标记为衰减关系。

定义 11　稳定关系。对于 t 时刻快照网络模块集合 S_t 中的任意模块 M,若在 $t+1$ 时刻快照网络的模块集合 S_{t+1} 中存在一个与 M 完全相同的模块 M',则将 M 从 t 时刻到 $t+1$ 时刻变化为 M'的演化关系标记为稳定关系。

定义 12　改变关系。对于 t 时刻快照网络模块集合 S_t 中的任意模块 M,若在 $t+1$ 时刻快照网络的模块集合 S_{t+1} 中存在一个模块 M'满足条件

$$\delta=e/\min\{n,n'\}\geqslant c \text{ 且 } \delta\neq 1$$

则将 M 从 t 时刻到 $t+1$ 时刻变化为 M'的演化关系标记为改变关系。其中,e 为模块 M 和 M'中相同的边数(所谓边相同是指边的两个顶点及边上的标志符号都相同);n 为模块 M 所包含的边数,n'为模块 M'所包含的边数,$\min\{n, n'\}$ 表示 n 和 n'中的较小者;c 为预先定义的参数($0<c<1$)。

可以看到,对于以上四种变化关系的定义都是基于模块拓扑关系的(将边及边上作用标志符号纳入考察范围),其中对于改变关系的定义采用了模糊匹配的方式,这使得对于模块的动态演化匹配更具有针对性。基于模块拓扑而不是基于模块点的集合是因为在模块的动态变化过程中可能存在虽然模块的点集合并未改变但点之间的作用方式及关系却已经改变了的情况。基于上面四种关系的定义,可以在整个时间序列上对任意时刻快照网络的任意模块的动态变化过程进行揭示。具体步骤可以由算法 2 来描述。

算法 2　MDC(matching the dynamic changes)

输入:整个时间序列上快照网络的模块集合;
输出:任意模块的动态变化链。
1. 初始化模块变化链队列 Q,用于存放已经匹配的模块链;
2. 以 $t=1$ 时刻的快照网络模块集合 M_1 中每个模块为头模块初始化一条新的模块链放入 Q 中,并标记每个模块的时间信息为 1;
3. for $t=2$ to n
4. for t 时刻的快照网络的模块集合 M_t 的每个模块 M'
5. 标记时间信息为 t;
6. for Q 中的每个模块链的尾模块 M
7. if M 和 M'的关系满足 2.4 节所定义的四种关系之一
8. then 将 M'作为 M 的后续模块链入 M 的变化链;
9. end for
10. if M'不能被链入 Q 中任何模块链 then
11. 以 M'为头模块初始化一个新模块变化链放入 Q;
12. end for
13. end for
14. 输出所有的模块变化链;
15. end

根据算法 2,可以在整个时间序列范围内得到各个模块的变化链,这些模块的变化链反映了模块的动态变化特性。

3 实验及结果

为验证本文分析方法的正确性和有效性,考察分析过程中各算法的性能,使用 MIPS 数据库中啤酒酵母菌的蛋白质网络作为原始的静态 PPI 网络。该 PPI 网络包括 4 362 个蛋白和 13 317 条交互边。对于时间序列的基因表达数据,使用酵母菌细胞周期的基因表达数据作为本文的实验源数据。该数据由 Spellman 等人在 1998 年发布,可以从 Stanford 大学的基因数据库中下载得到(也可以从其他公用数据库中下载),其主要通过在啤酒酵母菌的整个细胞周期中每隔 7 min 提取、观察相应基因所转录的 mRNA 的表达量来得到[7]。通过对原始数据进行适当清理,该数据总共包含了 6 066 个基因从细胞生命周期的 G1 期到 M 期的 18 个时间点上的基因表达水平。

基于上面两个数据集,本文主要从两方面来验证该分析方法的有效性。一方面,基于 GO 数据库对本文的功能模块挖掘算法挖掘结果的准确性进行评价;另一方面,基于 KEGG 数据库中酵母菌细胞周期的蛋白质交互路径数据和 GO 数据库的基因注释对本文所揭示的模块的动态变化特性进行生物意义的可解释性考察。

3.1 SFMM 算法具有较高的准确性和良好的稳定性

为考察 SFMM 算法的性能,本文将其与 MINE[8] 算法(基于紧密度的一种模块挖掘算法)分别在三个不同时间点上的快照网络($t=4,7,15$)中的平均准确性进行比较,观察其在不同紧密度水平上的准确性及稳定性。在实验中,对于所挖模块的效果好坏按照基于 GO 平台的相似性打分机制[9] 进行评价。一般情况下,该相似性打分机制对一个模块中各蛋白进行两两比较,如果所有或大部分的比较结果在 0.45 以上,则说明该模块中的蛋白质整体相似性较好。定义算法准确率为经过 GO 相似性打分后结果较好的模块数与利用该算法所挖掘的所有模块的总数量的比值。在实验中,设置基本参数 $\sigma=0.5,\varepsilon=0.6$,对于 GO 最低相似性打分设置为 0.52。在模块评价中要求相似性大于 0.52 的比较结果比例达到 85% 以上后才认为该模块为整体相似度相对较好的模块。实验结果如图 3 所示。

从图 3 可以看出,本文的算法——SFMM 的最低准确率约为 61%,最高准确率约为 68%,平均准确率约为 64%,这一结果对于一般模块挖掘算法来说已相对较好。在紧密度低于 0.7 时,本文算法的准确率明显高于 MINE 算法的准确率,这说明本文算法在挖掘稀疏模块时具有很强的优势,随着紧密度的提高,虽然本文算法的准确率略低于 MINE 算法,但相差不是很大。总体看来,随着紧密度提高,本文算法的稳定性明显高于 MINE 算法,这充分体现了本文算法在对相对稀疏的模块进行挖掘时的有效性和优越性。

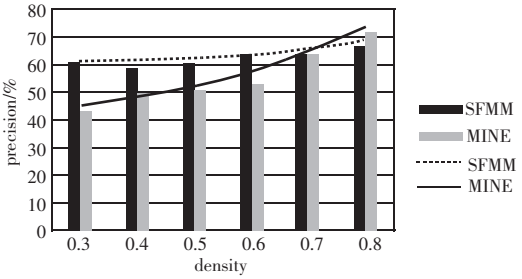

图 3 SFMM 与 MINE 算法精准率比较

3.2 揭示的模块的动态特性具有很强的生物意义

为说明本文所揭示的模块动态特性具有一定的实际意义(具有事实上的合理性),本文基于 KEGG 数据库中酵母细胞周期的蛋白质交互路径数据,结合 GO 数据库中的基因注释对本文的实验结果进行可解释性验证。在生物研究中,细胞的一个生命周期一般被分为 G1、S、G2、M 四个重要时期,其中 G1、S、G2 期组成细胞的分裂间期(约占细胞生命周期的 90% ~ 95%),M 期为细胞的有丝分裂期(约占细胞生命周期的 5% ~ 10%)。由于在本文的实验中采用了酵母的细胞周期的基因表达数据,所以将尝试从酵母细胞周期不同阶段的生物特征的常识性知识角度出发,对本文的实验结果的可解释性进行验证。在本文的具体实验中,对于 SFMM 算法中的基本参数设置为 $\sigma=0.5,\varepsilon=0.6,\omega=0.5$,在模块动态特性匹配算法 MDC 中设置参数 $c=0.8$。为了说明本文实验结果一定程度上能够反映具体的生物过程,本文列出了两个实验结果的例子,如图 4 所示。

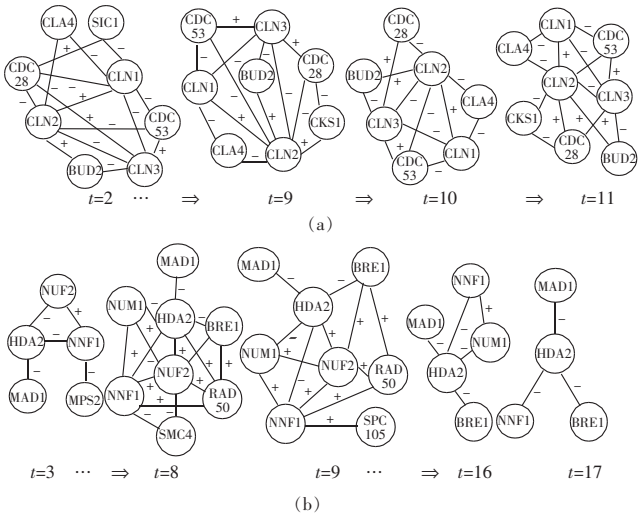

图 4 实验结果实例

图 4(a)描述了一个功能模块从细胞周期的 G1 期到 M 期的变换过程。根据 KEGG 数据库中细胞周期的蛋白质交互路径信息可以知道,CDC28 蛋白是细胞周期蛋白酶(CDK)的一个亚单元蛋白;CLN3/CDC28 之间相互作用是细胞周期开始的必要行为和重要标志,当它们作用一段时间后,其产物将进一

步促进 CLN1 和 CLN2 等其他 S 期蛋白的转录。而 CLN1、CLN2 又进一步与 CDC28 蛋白相互作用促进与周期蛋白酶(CDK)相关的 B 型周期蛋白的合成,从而促进 DNA 的复制和转录,促使细胞周期进入有丝分裂期(M 期)。从图 4(a)可以看到,当 $t=2$ 时,该模块包含 8 个蛋白,且这些蛋白都是与细胞 G1 和 S 期内的活动有关的蛋白。其中,CLN1、CLN2 和 CLN3 都是细胞 G1 期的周期蛋白,起调控细胞周期的作用。它们通过与 CDC28 相互作用来促进细胞周期由 G1 向 S 期的转变。SIC1 是细胞由 G1 期向 S 期转变的抑制蛋白,其与 CDC28 交互用于防止由 G1 期到 S 期的过早转变。CDC53 是 SCF 复合物(促进 G1/S 的转变)的一个遍在蛋白;CLA4 是 PAK 复合物的一个信号传导酶;BUD2 是 GTpase(鸟嘌呤磷酸酶)的激活蛋白。在本文模块中,CDC28/CLN3 呈现正向交互,同时,CDC28 与 CLN1 和 CLN2 之间呈现负向交互,这充分说明了 CDC28/CLN3 之间的交互是细胞周期开始的重要行为的生物事实。SIC1/CDC28 之间呈现正向的交互,SIC1/CLN1 之间呈现负向交互也充分说明了 SIC1 是 CDC28-CLB 酶合物的抑制蛋白的生物事实。同时,CLN1/CLN2 呈现正向的交互,也进一步说明了 CLN1 与 CLN2 具有功能相似性的生物事实。当 $t=9$ 时,细胞周期开始由 G1 期向 S 期转变,与 $t=2$ 时的模块相比,SIC1 蛋白已经消失,同时与 DNA 转录调节有关的蛋白 CKS1 开始出现,这说明在 S 期可能已经开始了与 M 期的有丝分裂过程相关的生物活动。当 $t=10$ 时,CKS1 蛋白却又消失了,这说明对于 M 期活动的预先转录调节可能只是短暂性的过程。当 $t=11$ 时,细胞周期进入 S 期,CDC28/CLN2 呈现正向交互,促进与周期蛋白酶相关的 B 型蛋白的合成,进而加强 DNA 的复制和转录,为细胞进入 M 期做好前期的准备。由此可见,本文的实验结果确实在一定程度上反映了具体的生物过程。

图 4(b)中,当 $t=3$ 时,该模块包含 NUF2、HDA2、NNF1、MAD1、MPS2 五个蛋白。其中,HDA2 是一种染色体端粒维持蛋白;NUF2 是相对保守的、与着丝粒相关的 NDC80 复合物的组成蛋白,其在纺锤体活动和着丝点分离等活动中发挥着重要作用;NNF1 是与纺锤丝微管有关的蛋白;MAD1 是一种与纺锤体结合有关的蛋白;MPS2 是大量存在于纺锤体和细胞核核膜上的一种重要的膜蛋白。可以看出,这五个蛋白都与有丝分裂过程密切相关。当 $t=8$ 时,该模块在原来基础上新加入了四个新的蛋白(BRE1、NUM1、RAD50、SMC4),但 MPS2 蛋白却消失了。BRE1 是 DNA 复制所需的一种重要蛋白;NUM1 是核仁迁移所必需的一种重要蛋白;RAD50 是一种端粒维护蛋白;SMC4 是在细胞分裂时染色体改编的凝合物。可以看出,这四个蛋白都与有丝分裂过程有关,且它们之间相互作用,为有丝分裂(M 期)期的到来做前期准备。当 $t=9$ 时(细胞周期开始进入 G2 期),可以看到,SPC105 蛋白加入了模块,与此同时,SMC4 却消失了,而 SPC105 是组成染色体的一种重要蛋白。这充分说明了在细胞周期中,染色体的大规模合成是从 G2 期开始的生物事实。随着该模块的不断变化,当 $t=16$ 时,模块规模衰减至五个蛋白,且其中心蛋白(HDA2)与其他蛋白都呈现负向交互,这充分说明了随着细胞周期进入有丝分裂后期,与染色体形成有关的蛋白活动将受到抑制的生物事实。这种抑制过程一直持续,直到 $t=18$ 该模块消失为止。可见,对于该模块变化的整个过程,都与一定的生物事实相对应。

综上所述,可以看出本文的分析结果确实是与一定的生物事实相对应的,且反映了具体的生物活动的变化过程。由此可见,本文的分析方法可以有效地揭示模块变化的动态特性,且具有很强的可解释性。

4　结束语

本文提出了一种在动态条件下揭示模块动态变化特性的新方法。基于原始静态的 PPI 网络,利用时间序列的基因表达数据对静态 PPI 网络中的相应蛋白交互活动添加时间信息,进而构造出动态的蛋白质交互的快照网络;随后,又提出了一种在动态快照网络中挖掘相对稀疏模块的新方法。最后,试图基于模糊匹配的方法在不同的时间快照网络中匹配模块的动态变化状态,从而揭示特定模块在整个时间序列上的动态变化特性。实验结果表明,本文的分析方法可以有效地揭示模块变化的动态特性,且具有很强的可解释性。虽然本文所有工作都是基于生物数据进行的,但其分析方法也可以应用到其他类型复杂网络中。对该方法的进一步完善和充实将是笔者下一步工作的一项重要内容。

参考文献:

[1] JIN Ruo-ming, McCALLEN S, LIU Chun-chi, *et al*. Identifying dynamic network module with temporal and spatial constraints [C]//Proc of Pacific Symposium on Biocomputing. 2009:203-214.

[2] YOU Chang-hun, HOLDER L B, COOK D J. Learning patterns in the dynamics of biological networks [C]//Proc of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2009:977-986.

[3] ROHIAN H, AN Ai-jun, ZHAO Jia-shu, *et al*. Discovering temporal associations among significant changes in gene expression [C]//Proc of IEEE International Conference on Bioinformatics and Biomedicine. Washington DC: IEEE Computer Society, 2009:419-423.

[4] RUAN Quan-song, DUTTA D, SCHWALBACH M S, *et al*. Local similarity analysis reveals unique associations among marine bacterioplankton species and environmental factors [J]. Bioinformatics, 2006, 22(20):2532-2538.

[5] RUAN Quan-song, STEELE J A, SCHWALBACH M S, *et al*. A dynamic programming algorithm for binning microbial community profiles [J]. Bioinformatics, 2006, 22(12):1508-1514.

[6] HAN Jing-dong, BERTIN N, HAO T, *et al*. Evidence for dynamically organized modularity in the yeast protein-protein interaction network [J]. Nature, 2004, 430(6995):88-93.

[7] BAR-JOSEPH Z. Analyzing time series gene expression data [J]. Bioinformatics, 2004, 20(16):2493-2503.

[8] RHRISSORRAKRAI K, GUNSALUS K C. MINE: module identification in networks [J]. BMC Bioinformatics, 2011, 12:192.

[9] 张少华,尚学群,王淼. 一种衡量基因语义相似度的新方法 [J]. 计算机应用研究, 2011, 28(3):957-960.

[10] AACH J, CHURCH G M. Aligning gene expression time series with time warping algorithms[J]. Bioinformatics, 2001, 17(6):495-508.

[11] KWOH C K, NG P Y. Network analysis approach for biology[J]. Cellular and Molecular Life Sciences, 2007, 64(14):1739-1751.

[12] ROBARDET C. Constraint-based pattern mining in dynamic graphs [C]//Proc of the 9th IEEE International Conference on Data Mining. Washington DC: IEEE Computer Society, 2009:950-955.