生物网络模体发现算法研究综述

覃桂敏[1,2],高 琳[1],呼加璐[1]
(1. 西安电子科技大学计算机学院,陕西西安 710071; 2. 西安电子科技大学软件学院,陕西西安 710071)

摘 要: 网络模体发现是生物网络数据分析中的一个核心问题. 首先分析了网络模体发现中相关的基本计算问题:随机网络建模,子图搜索和模体统计意义评价等. 其次对生物网络模体发现算法进行了综述和评价,从研究方法上将模体分为精确模体,概率模体和其它模体三类,并对识别每类模体的典型算法进行研究和分析. 为了对网络模体进行深入分析与研究,引入了与模体发现密切相关的生物网络模块发现问题. 最后讨论了网络模体发现算法的最新进展和下一步的研究方向.

关键词: 生物网络; 网络模体; 算法
中图分类号: TP391 **文献标识码**: A **文章编号**: 0372-2112 (2009) 10-2258-08

A Review on Algorithms for Network Motif Discovery in Biological Networks

QIN Gui-min[1,2], GAO Lin[1], HU Jia-lu[1]
(1. *School of Computer Science and Technology, Xidian University, Xi'an, Shaanxi* 710071, *China*;
2. *Software School, Xidian University, Xi'an, Shaanxi* 710071, *China*)

Abstract: Network motif discovery is a key problem in data analysis of biological networks. This survey first analyzes basic computational issues in motif discovery, which are modeling random networks, searching subgraphs, and evaluating motif significance. Then the motif discovery algorithms are reviewed and evaluated. The algorithms are divided into three major categories from the aspect of research methods, i. e. exact motif discovery, probability motif discovery, and other motif discovery. And several typical algorithms in each category are analyzed. In order to make a thorough analysis on network motifs, we introduce a related problem, modular discovery. Finally, we report the recent progresses from the algorithmic viewpoint and further discuss the future research trends.

Key words: biological networks; network motifs; algorithms

1 引言

细胞进化是由各种分子如蛋白质,DNA 和代谢物之间的相互作用调控的. 通过高通量技术,人们收集了大量的分子相互作用数据. 这些相互作用数据主要分为四类,即代谢路径,基因调控网络,蛋白质 - 蛋白质相互作用(PPI)网络和信号传导网络. 它们不仅提供了预测单个蛋白质生物功能的有用信息,还提供了理解细胞模块组织的实验基础,因此对它们的分析是后基因组时代的重要任务[1]. 最近对这些网络分析的结果表明[2],许多网络本质上具有共同的全局性质,如小世界性质和无标度性质. 而在 2002 年由 R. Milo 首次提出的网络模体[3]则是其中非常重要的一种局部性质,网络模体的发现和分析引起了生物信息学,复杂网络研究和社会统计学等领域的广泛关注,已经成为目前生物信息学的研究重点和热点之一.

短短几年对网络模体的分析已经取得了大量的研究成果,Shen-Orr 等人的研究表明在基因调控网络中的一些网络模体是基本的信息处理模块,它们互相配合完成基因调控网络中的信息处理工作[4]. Lee 等人在细菌和酵母的基因调控网络中也发现了相同的三种有重要功能的网络模体[5],这些模体在 Milo 和 Shen-Orr 的论文中也有相关阐述,这表明在同类型的网络中可能具有相同的模体. 2004 年 Milo 在信号传导网络和基因调控网络中也发现了这三种重要的网络模体[6,7]. 同时生物网络模体的研究中还出现了模体簇和超家族这样的概念[8],可以帮助人们进一步理解生物网络的功能. 文献[9]从实验研究的角度对网络模体及其功能进行了综述.

虽然网络模体对未来的生物学研究有着重要的意义,但是模体发现却是一个非常复杂的问题,其中有许

收稿日期:2008-09-10;修回日期:2009-05-31
基金项目:国家自然科学基金(No. 60574039);教育部博士点基金(No. 200807010013);陕西省自然基金(No. SJ08-ZT150)

多难题需要更深入地研究. 最近几年,人们提出了许多模体发现算法并在多个会议和期刊上发表. 为了将这些分散的文献和资料集中起来,本文对网络模体发现算法进行较全面地综述.

2　模体的定义及模体发现中的基本问题

为了更深入地理解模体发现问题,下面介绍模体的定义,以及模体发现过程中涉及到的基本计算问题,包括随机网络建模,子图搜索和网络模体统计意义评价.

2.1　模体定义

定义 1　精确网络模体(网络模体)是满足下列条件的子图[3]:

①该子图在与真实网络对应的随机网络中出现的次数(频率)大于它在真实网络中出现次数的概率是很小的,通常要求这个概率小于某个阈值 P,如 $P = 0.01$.

②该子图在真实网络中出现的次数 N_{real} 不小于某个下限 U,如 $U = 4$.

③该子图在真实网络中出现的次数 N_{real} 明显高于它在随机网络中出现的次数 N_{rand},一般要求 $(N_{\text{real}} - N_{\text{rand}}) > 0.1 N_{\text{rand}}$.

定义 2　网络模体发现问题:给定一个真实网络和参数 k,找出所有 k 规模的网络模体.

根据定义 1 可以得出,典型的网络模体发现算法包括三个步骤:①根据真实网络的性质生成一组对应的随机网络;②在真实网络和随机网络中搜索特定规模的子图,确定哪些子图是同构的,并将同构的子图归为一类;③通过比较每一类子图在真实网络和随机网络中出现的次数以确定其统计意义,从而确定网络模体.

定义 3　图 g 的概率矩阵 每个元素表示 g 中的两个顶点之间有边相连的概率.

定义 4　概率网络模体是由一组相似而不一定同构的子图构成,由概率矩阵表示[10]. 假设 p 个相似子图经过比对后(确定对应的节点序)得到的邻接矩阵为 $\{M^1, M^2, \cdots, M^p\}$,则概率模体 $\overline{M}$ 可由式(1)计算得到:

$$\overline{M} = \frac{1}{p} \sum_{\alpha=1}^{p} M^{\alpha} \tag{1}$$

2.2　随机网络模型

随机网络在许多领域都得到广泛的研究,本文仅讨论网络模体发现中的随机网络模型. 要建立适当的随机网络模型,首先要求随机网络应与真实网络具有相似的统计性质. 因为度分布是复杂网络中最重要的全局统计性质,而且许多研究人员用度分布来刻画复杂网络,因此通常通过建立与真实网络具有相同度分布的随机网络. 研究结果表明[11],大部分生物网络都是无标度网络,其度分布服从幂率分布,即 $P(k) \sim k^{-\gamma}$,其中 $2 < \gamma < 3$.

其次,在典型的网络模体发现算法中,随机网络模型进一步保持了真实网络的度序列,文献[12]分析了三个根据度序列生成随机网络的算法. 第一个是交换算法,该算法首先根据度序列构造一个网络,接着执行一系列的蒙特卡罗交换步,即随机选择一对边(如 $A \rightarrow B$, $C \rightarrow D$),然后进行交换($A \rightarrow D$, $C \rightarrow B$). 若该交换导致了多重边或自回路,则取消该交换. 第二个是匹配算法,该算法根据度序列给每个顶点设置入边和出边集合. 然后随机地选择一个顶点的入边和另一个顶点的出边,并建立它们之间的连边关系. 若该连边导致了多重边或自回路,则把该网络抛弃,上述过程重新开始. 最后一个算法是 Go with the Winner 算法,该算法同时考虑多个网络,对每一个网络的操作过程与匹配算法相似,过程中将导致多重边或自回路的网络抛弃,因此网络数不断减少. 为了弥补网络个数不断减少的情况,过一段时间就把剩余的网络复制. 重复这个过程直到所有的入边和出边都得到连接,然后随机地选择一个随机网络作为输出.

最近,有人提出用随机几何模型来建模 PPI 网络,且其度分布服从泊松分布[13,14].

2.3　子图搜索

子图搜索是网络模体发现的第一步. 它是计算复杂度很高的问题,因为子图的数量随着网络的规模和待搜索子图的规模呈指数级增长. Milo 等人采用穷尽递归搜索算法,搜索给定规模的所有子图[3,4]. 该算法用 $N * N$ 的邻接矩阵表示输入网络,通过枚举其中所有的 $n * n$ 的子矩阵得到对应的导出子图,对每个 n 搜索时间复杂度为 C_N^n —N 和 n 分别表示输入网络的规模和子图的规模,因此仅能有效地发现小规模的模体,如 3 - 模体和 4 - 模体. 为了提高子图搜索的效率,N. Kashtan 等提出了采样算法 ESA(Edge Sampling)来估计子图的相对出现次数[15]. ESA 算法随机地选择一条边,将其关联顶点加入到子图顶点集合 $V_{subgraph}$ 中. 然后不断地将 $V_{subgraph}$ 中顶点的邻接点增加到 $V_{subgraph}$ 中,直到子图达到指定规模为止,得到一个采样的子图. 重复该过程,直到采样的子图数达到预定义的个数. 由于该方法与网络规模是独立的,因此对分析规模非常大的网络很有效,而且能发现较大规模的模体,可以达到 8 - 模体. 然而,ESA 算法不能保证得到所有的子图而且同一个子图可能被多次采样.

另一个搜索算法是 ESU(Enumerating Subgraph)[16,17],该算法可以找出所有的子图. ESU 算法从输入网络的一个顶点 v 开始,增加具有下面两个性质的顶点到集合

$V_{extension}$中,一是它们的下标必须比 v 的下标大,二是它们不属于 $V_{subgraph}$中某个顶点的邻接点. 图 1[16,17]给出了 ESU 搜索子图的过程. ESU 从规模为 1 的子图出发逐步增大子图的规模,尽管也是枚举子图,但是在过程中,用了很多限制条件,保证了一个子图仅仅被搜索一次,且不会产生无意义的子图,因此本质上是一个回溯算法. 对于稀疏网络,ESU 是一个很有效的方法. 但由于问题本身的复杂性,搜索的子图规模最大也只能达到 6 个顶点. Rand-ESU 将 ESU 和概率方法相结合,在 ESU 枚举子图过程中,通过概率来判断是否需要继续扩展某个子图,可以大大提高算法的效率,而且可以把搜索规模扩展到 14 - 模体,是一种非常有效的方法.

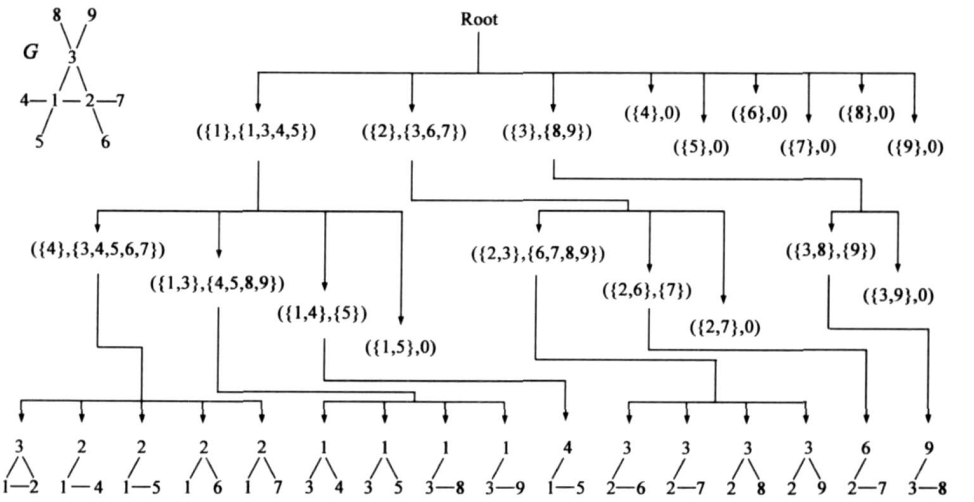

图1 ESU算法搜索子图过程

另一种子图搜索方法是枚举出所有特定规模的子图,然后在大图中对每个子图进行查询. J. A. Grochow 和 M. Kellis 提出了一个基于子图查询的模体发现算法[18]. 由于大多数真实网络都是稀疏的,因此我们提出了在稀疏网络中挖掘子图的算法[19],该算法可以高效地挖掘非树形子图. 另外,现在已经发现的模体中具有哈密尔顿环性质,因此我们基于矩阵论的方法挖掘哈密尔顿环频繁子图[20],该算法是一种有效的挖掘特定类型的子图算法.

2.4 网络模体统计意义评价

关于模体的统计意义评价,目前还没有统一的标准. 在关于网络模体的开创性文献中[3,4],模体的统计意义用 *Z-score* 来表示.

$$Z_i = (N_{reali} - \langle N_{randi}\rangle)/std(\sigma_{randi}) \quad (2)$$

其中 N_{reali}表示子图 g_i 在真实网络中的出现次数,而 $\langle N_{randi}\rangle$和 $std(\sigma_{randi})$ 分别表示子图 g_i 在随机网络集合中的平均出现次数和标准差. 很多模体发现算法都采用 Z - *score* 来判断一个子图是否为模体,文献[15]用子图 g_i 的浓度来代替次数计算 Z - *score*.

将子图 g_i 的 Z - *score* 规范化得到其 SP(significance profile)[6],式(3)给出了 SP 的计算方法. SP 强调子图的相对统计意义,而不是绝对意义. 它可以用来比较不同规模的网络,还可以用来了解一个给定网络的子图分布.

$$SP_i = Z_i / \sqrt{\sum_N Z_j^2} \quad (3)$$

3 网络模体发现算法

我们把现有网络模体发现算法识别的模体类型分为三类:第一类是精确网络模体,即遵循 Milo 的网络模体概念(定义 1);第二类是概率网络模体,这类模体是由一组相似的子图构成(定义 4);第三类是不属于上述两类模体的其它网络模体.

3.1 精确网络模体发现算法

精确网络模体发现算法遵循 Milo 提出的网络模体的定义来识别模体. 本文分析两个经典的精确网络模体发现算法,一个是 N. Kashtan 等人提出的算法[15],它提出了 ESA 子图采样算法,并提出子图浓度的概念和计算子图浓度的方法. 该算法首先采用穷尽递归搜索方法枚举出真实网络中的所有子图;然后产生一组与真实网络具有相同度序列的随机网络,并用 ESA 算法在这组随机网络中采样子图,计算每个子图的概率 P,并增加该子图的权重 $W = 1/P$ 到其同构类的得分 S 中. 对任意一个同构类 G_i,其浓度为

$$C_i = S_i / \sum_{k=1}^{L} S_k \quad (4)$$

由浓度计算 Z - *score*,

$$Z = (C_{real} - \langle C_{rand}\rangle)/\sigma_{rand} \quad (5)$$

从而确定网络模体. 他们还开发了软件 Mfinder 实现了该算法[15].

另一个是 S. Wernicke 的算法[16,17,21],提出了 ESU 和 Rand-ESU 两个子图搜索方法. 该算法首先产生一组与真实网络具有相同度序列的随机网络;然后用 ESU 和 Rand-ESU 算法来搜索子图,并基于这些子图计算模体的统计意义. 该算法也有对应的软件 FANMOD[21]. 此外,文献[16,17]还提出了一个不需要生成大量的随机网络,而是直接计算模体统计意义的理论方法,但是该方法的计算复杂性很高.

文献[3]首次提出了网络模体的概念,而文献[15]和文献[16,17]则是网络模体发现算法的两个开创性的成果,对后续网络模体发现算法的研究起了重要的指导意义. 文献[15]采用 ESA 算法搜索子图解决了子图数量随输入网络规模和子图规模呈指数级增长的问题,理论上可以在大规模的输入网络中搜索较大规模的子图,而且具有较高的效率. 但是由于在采样出子图后计算其概率,这需要较长的计算时间,因此该算法能发现的网络模体的规模也较小. 正如 S. Wernicke 的分析[16,17],ESA 算法有几个不足之处:计算子图概率 P 具有较高的计算复杂性,对相同的子图可能会多次采样而另外的子图被采样的概率很低,从而导致最后发现的模体具有不准确性等. 因此,S. Wernicke 提出了 ESU 算法,该算法是一种系统搜索方法,通过规定约束条件大大减少了搜索空间,是一个高效的方法. 但由于问题固有的复杂性,ESU 能搜索的子图并不能达到较大的规模. 因此,FANMOD[21] 软件采用 Rand-ESU 搜索子图. Rand-ESU 是一个很好的算法,但其参数的设置是一个难点,如何设置参数使得搜索的子图出现次数分布与真实网络中的分布尽可能一致是需要进一步研究的问题.

3.2　概率网络模体发现算法

J. Berg 和 M. Lassig 认为若生物网络进化是一个随机过程,那么网络模体就不一定需要由同构的子图构成,因此提出了概率网络模体的概念[10]. 他们建立了一个概率模体出现次数的统计模型,从该模型得到得分函数,并根据该得分函数计算模体的统计意义. 他们认为现有的网络模体具有较高的内部连接性,因此仅考虑非树型子图. 该方法主要分为三个步骤:首先枚举所有给定规模的非树型子图;然后计算两两子图之间的失配值;最后根据得分函数用模拟退火算法求出最相似(即得分最高)的一组子图$\{G^1, G^2, \cdots, G^p\}$,对这组子图进行多图比对得到网络模体. 图 2 给出了三个相似子图比对得到概率模体的例子. 得分函数如式(6),

$$S(G^1, G^2, \cdots, G^p, A) = (\sigma - \sigma_0)\sum_{\alpha=1}^{p} L(c^\alpha) - \frac{\mu}{2p}\sum_{\alpha,\beta=1}^{p} M^{\alpha\beta} - \log(Z_{\sigma,\mu}/Z_{\sigma_0}) \quad (6)$$

其中 A,$L(c^\alpha)$ 和 $M^{\alpha\beta}$ 分别表示子图集合$\{G^1, G^2, \cdots, G^p\}$的比对,子图 G^α 的内部连边数和子图 G^α 和 G^β 之间的失配值. σ, σ_0 和 μ 为参数,$Z_{\sigma,\mu}$ 和 Z_{σ_0} 为规范化因子. 从式(6)可以看出,$(\sigma - \sigma_0)\sum_{\alpha=1}^{p} L(c^\alpha)$ 部分表示子图内部连边数越多,则得分越高,连接奖励因子 $\sigma - \sigma_0$ 加强了内部连边数的影响;$\frac{\mu}{2p}\sum_{\alpha,\beta=1}^{p} M^{\alpha\beta}$ 部分表示子图之间的相异度越小,则得分越高,失配惩罚因子 μ 加强了相异度的影响;$\log(Z_{\sigma,\mu}/Z_{\sigma_0})$ 可以视为规范化项,当子图之间有很大的相异性或子图的内部连边数很少时,赋一个负值[10]. 最大化式(6)的值意味着要找出一组最相似的内部连边数较多的子图.

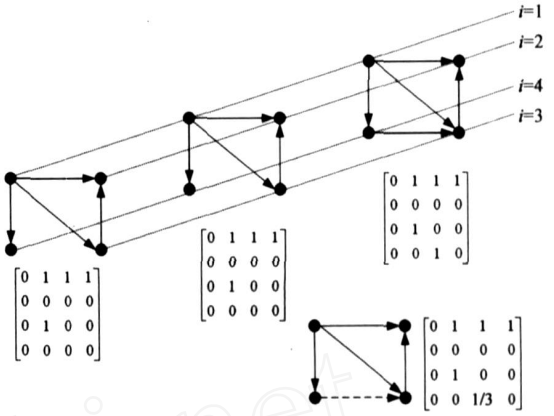

图2　三个相似子图的比对及其对应的网络概率模体

该算法第一次提出了概率网络模体的概念,且通过统计方法来求解网络模体,为网络模体发现的研究做出了重要贡献. 该方法不需要像传统方法那样产生大量的随机网络并在这些随机网络中搜索子图,因此在一定程度上具有较高的效率. 他们的方法侧重于理论的证明和推导,从算法的角度来看,还有许多值得讨论的地方. 得分函数的第三部分规范化因子的计算是非常复杂的,而且在优化的过程中一直需要计算该部分的值,因此算法总的来说是复杂的. 得分函数中参数的设置也是较难的问题,且与数据有关. 另外,用模拟退火来求最优化问题,并不能保证得到全局最优解,因此得到的结果不一定是最优的. 算法需要多次执行模拟退火算法来优化参数.

Rui Jiang 等人进一步认为输入网络也是概率网络 $P = (\pi_{ij})_{N*N}, 0 \le \pi_{ij} \le 1$,其中 π_{ij} 表示节点 i 和节点 j 之间连接的概率[22]. 他们建立了一个随机网络的有限混合模型,并采用 EM 算法来识别网络模体. 在他们的方法中,将输入随机网络视为以概率 λ 将一组相似的子图嵌入到背景随机网络集合中. 子图的集合定义了前景即随机网络模体,并用概率矩阵 $\Theta_1 = (\theta_{ij})_{n*n}, 0 \le \theta_{ij} \le 1$ 表示,θ_{ij} 为节点 i 和节点 j 连接的概率. 该算法的主要步骤如下:第一步,根据 P 产生 K 个邻接矩阵 $\{A^k\}_{k=1}^{K}$,$a_{ij}^k = 0, 1$,其中 $\Pr(a_{ij} = 1) = \pi_{ij}$ 和 $\Pr(a_{ij} = 0) = 1 - \pi_{ij}$. 在这 K 个邻接矩阵中,枚举所有的 n 规模子图,

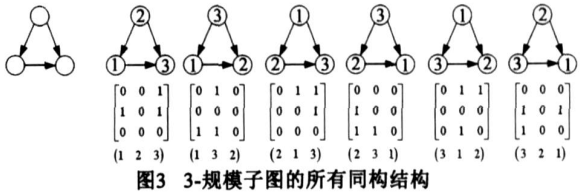

图3　3-规模子图的所有同构结构

并计算出这些子图的所有同构结构. 图 3 列出了一个 3 规模子图的所有同构结构的例子. 第二步,对每一个邻接矩阵 A^k 产生 L 个与 A^k 具有相同度分布的随机网络 $\{A_l^k\}_{l=1}^{L}$,此时得到 $K*L$ 个随机网络. 从这些随机网络中采样出 n 规模子图,并计算出这些子图的所有同构结构. 最后,根据式(7)的伪似然函数,采用 EM 算法估计参数 λ 和 Θ_1

$$L(\Theta|X,Y,Z)=\prod_{w=1}^{W}\prod_{h=0}^{1}[\lambda_h\prod_{p=1}^{P_w}\Pr(X_p^w|\Theta_h)^{Y_{hp}^w}] \quad (7)$$

其中 X 为观察值,是已知的,Y 和 Z 是未知的,可以视为缺失数据,$\Theta=\{\lambda,\Theta_0,\Theta_1\}$ 是待估计的参数[22].

该算法进一步扩展了概率的范畴,认为不仅模体表现为概率的形式,输入网络也应该是概率随机网络,这也是符合实际情况的. 他们把生物网络视为一组相似的子图(模体)以一定的概率嵌入到背景随机网络中,然后用 EM 算法来估计相关的参数,具有创新性. 然而,该算法需要根据输入概率网络生成一组确定网络,并根据这组确定网络产生大量的随机网络,然后在确定网络中枚举所有特定规模的子图,在随机网络中采样特定规模的子图,并把所有子图的所有同构结构列举出来,仅仅这样的步骤就需要很大的计算量,因此效率是较低的.

上述两个算法只能发现规模很小的模体(3 - 模体和 4 - 模体),它们的实验部分并没有给出性能分析. 另外,对一个规模只能产生一个模体,而对特定规模的图而言,具有多个不同的拓扑结构,而且这些拓扑结构之间可能具有较大的差异性,这两个算法并没有讨论该问题.

3.3 其它网络模体发现算法

Jin Chen 等人提出的 NeMoFinder[23] 中的网络模体定义为重复出现的,独特的子图. 一个子图是重复出现的,若其在真实网络中出现的次数不少于 t_f. 一个子图是独特的,若其在真实网络中出现的次数至少比它在 n 个随机网络的 t_u 个中出现的次数要多,其中 t_f, t_u 和 n 是给定参数. NeMoFinder 采用了 SPIN 算法[24] 搜索重复出现子树的思想,并将其应用到子图的搜索中. 为了确定重复出现的子图,需要首先找到规模为 2 到 k 的重复出现子树,然后将网络划分为图集,执行图的合并操作来找到重复出现子图.

F. Schreiber 和 H. Schwobermeyer 基于子图之间的重叠提出了确定子图出现次数的三种情况,即 F1, F2 和 F3,分别对应顶点与边可以任意重叠,顶点可以重叠而边不重叠和顶点与边均不重叠三种情况[25]. 并提出了一个通用的算法框架 FPF,计算在这三种定义下的子图出现次数. 该算法框架首先计算 F1 定义下的子图出现次数及其实例集合,然后根据实例之间的重叠关系构造一个重叠图,图中的每个顶点对应一个实例,若两个实例间有重叠关系,则顶点之间有边相连. 计算该重叠图的最大独立集,得到 F2 和 F3 概念下的子图出现次数.

Chia-Ying Cheng 等人也提出了不同于 Milo 的模体定义,他们提出了发现 bridge 模体和 brick 模体的算法[26]. 在该算法中,为每个从节点 u 到节点 v 的边赋一个权重 $link(u,v)$,其值为对应的超几何系数 $C_{u,v}$. 若 $link(u,v)$ 在随机网络或真实网络中的值小于某个阈值 $link_{AVG}$ 减去对应的两个标准差,则认为该边是弱连接,否则认为是强连接. Bridge 模体仅由弱连接构成,且包含和其它模体既不交互也不重叠的模体. Brick 模体仅由强连接构成,且在定义全局拓扑结构中起重要作用. 为了识别 bridge 和 brick 模体,在真实网络和一组对应的随机网络上执行下列操作:首先,计算每条边的权重,根据定义标识每条边为强或弱连接. 然后,搜索 n 规模的 bridge 和 brick 子图. 最后,根据其 Z_{score} 值和 SP 值识别出 bridge 和 brick 模体.

Noga Alon 等人提出了用着色编码技术计算树和有界树宽子图的出现次数[27]. 他们的方法侧重于得到各种不同树的出现次数的分布,从而可以比较各种不同的生物网络. 该方法首先用 k 种颜色给输入网络 G 的每个顶点独立且均匀地着色,然后用动态规划技术来计算每个顶点唯一着色的树 T 的非导出子图的出现次数. 重复上面两步若干次,将得到的出现次数增加到树 T 的出现次数以估计其在网络 G 中的出现次数.

上述算法识别其它网络模体. NeMoFinder 算法能发现从小规模(规模为 2)到较大规模的子图,而且效率较高. 但该算法并不能把所有的模体都找出来,可能存在遗漏[28]. 后来 Jin Chen 等人对该算法进行了扩展,提出 LaMoFinder 算法[29],将模体发现应用到有标记的网络中. FPF 引入了重叠子图的概念,计算在不重叠,顶点重叠和边重叠三种情况下的子图的出现次数,严格来说并不是模体发现算法,只是频繁子图挖掘算法,它能发现从小规模到大规模的所有频繁子图. Chia-Ying Cheng 等人提出了 bridge 模体和 brick 模体的概念,并在生物网络中识别这两类模体,具有较大的创新性. Noga Alon 等人的算法更侧重于特殊子图出现次数的分布上,主要应用于比较各种不同的网络,与一般的模体发现算法有较大的区别.

4 生物网络中的模块发现

前面讨论了生物网络中的模体发现算法. 从生物学角度来看,模体是指生物网络中的基本功能模块,而从计算角度来看,模体是相对于随机网络而言在真实网络中频繁出现的子图. 在生物网络分析中,还有一类

重要的问题即生物网络中的模块发现问题.同样,从生物学角度来看,模块也是指生物网络中的功能模块,而从计算角度来看,模块是模体簇或者是生物网络中稠密的子图.作为生物网络中的功能单元,模体和模块具有紧密的联系.图4[30]说明了基因调控网络的四个不同粒度的结构,构成基因调控网络最基本的节点是调控因子和被调控的基因,边对应它们之间的调控关系;小规模的节点和边的集合构成了模体,模体是基因调控网络中频繁出现的小规模的子图;模体进一步组合得到模体簇或模块,模块表示更大规模的生物学功能;多个模块相互连接构成了整个基因调控网络.在其它的生物网络中也有类似的层次结构.最近的研究表明功能上具有紧密联系的基因和蛋白质可以由网络的拓扑结构—生物网络模块推导出来.由于目前许多简单的生物如酵母的大部分基因和蛋白质的功能是未知的,如果能够设计可靠的方法通过已知的基因和蛋白质功能来预测未知元素的功能,则对生物学的发展将起着非常重要的作用[31].由于模体发现问题固有的复杂性,并不能用相似的算法来求解大规模的模块发现问题.目前关于模块发现也有许多成果[32~37].也有一些相关的文献将模体发现与模块发现结合起来研究[8,18,30,31].

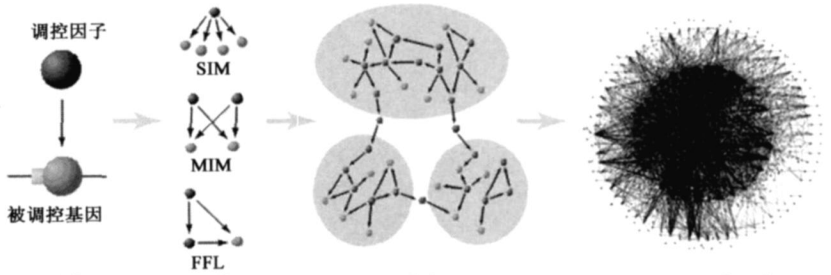

图4　基因调控网络的四个不同粒度的结构

5　讨论与结束语

本文对生物网络模体发现算法进行了综述和评价,对典型的模体发现算法进行了研究和分析,指出了每种算法的优势及其存在的问题.其次分析网络模体发现中相关的计算问题,如随机网络建模,子图搜索及模体评价等.除了本文分析的算法之外,目前还有许多模体发现算法,包括 Esti Yeger - Logem 等人提出的在集成网络中发现模体的算法[1],Kim Basherville 等人提出的一种新的图同构计算方法和反 - 模体的概念[38],Laxmi Parida 等人针对生物网络的特点提出的压缩网络来更有效地发现模体[39],Royi Itzhack 等人根据生物网络的无标度性质提出的对网络进行分解的思想以及并行计算的方法[40],Joshua A. Grochow 等人提出的子图查询方法,并引入多个优化技术有效地改进算法性能[18],C. Matias 等人从模体数量的均值,方差以及服从的统计分布来分析网络模体[41],F. Picard 等人扩展了[41]的思想更深入地分析网络模体[42].由于篇幅有限,本文不可能涵盖所有的算法,希望这篇综述能对网络模体发现算法的研究起到一定的参考作用.另外,尽管目前已经有较多的模体发现算法,但模体发现中还有几个方面需要进一步研究.

①有效的子图搜索算法

子图搜索是模体发现的关键步骤,它具有很高的计算复杂性.如何设计一个有效的子图搜索算法至关重要.对于枚举方法,系统搜索算法如回溯法和分支限界法可能是较好的选择,关键是如何设计适当的限界函数.对于采样方法,如何设置相关的参数使得找出子图的出现次数分布可以很好地代表真实网络和随机网络中的子图分布也是很难的问题.还有一种方法是进行子图查询.然而在这种情况下,首先必须枚举出某个规模的所有子图,然后需要设计一个有效的查询算法.另外,由于 3 到 6 规模的网络模体还是较容易发现的,可以考虑通过充分利用已发现的小规模模体来发现更大规模的模体,即进行子图的扩展.关于子图扩展策略,可以参考频繁子图挖掘算法[43~48].

②适当的随机网络模型

模体是与随机网络相比在真实网络中频繁出现的子图,因此随机网络建模是模体发现中的一个必要步骤.如何生成合适的随机网络是首先要考虑的问题.是否与真实网络具有相同度分布的随机网络就可以很好地用于解决模体发现问题?除了度分布或者度序列之外,还有其它的全局统计性质和局部性质用于刻画网络.在生成随机网络时,是否应该把这些因素考虑在内?而且,在建立随机网络时如何处理数据的不完整性问题?典型的算法中往往需要生成大量的随机网络,能否考虑仅生成少量的随机网络或者通过统计方法不用生成随机网络就能识别网络模体,从而提高模体发现的效率?

③模体统计意义评价标准

目前一般通过与随机网络比较计算子图的 *Z* - *score* 和 SP 值来评价模体统计意义,或者通过建立统计模型来求出网络模体.如何更有效地评价模体统计意义需要深入地研究.

④特定类型的网络模体

目前模体发现算法中,往往认为考虑子图之间的重叠是无意义的,因为一个蛋白质可能会同时起多个生物作用,但在特定的应用下,可能需要考虑子图重叠

的问题.另外,根据具体问题找出具有各种约束条件的模体也有重要的实际意义.

⑤与生物知识结合

生物信息学的目标是为生物问题提供答案.对网络模体发现的研究应该与生物知识密切联系到一起.在大多数模体发现相关的文献中,生物网络通常建模为简单的无向图或有向图.而实际生物网络数据是非常复杂的,如网络是动态的,带标记的,有权的或集成的等等,如何在这样的生物网络中发现模体是一个挑战.

总而言之,作为一个新兴的研究领域,网络模体发现已经取得了长足的发展并得到了广泛的应用.但是,在几个关键计算问题上还需要进一步研究,以使得该领域可以长期且更深入地促进系统生物学和生物信息学的发展.

参考文献:

[1] Yeger-Lotem E,Sattath S,Kashtan N,Itzkovitz S,Milo R,Pinter R Y,Alon U,Margalit H. Network motifs in integrated cellular networks of transcription regulation and protein-protein interaction[J]. Proc Natl Acad Sci,2004,101(16):5934 - 5939.

[2] Newman M E J. The structure and function of complex networks[J]. SIAM Rev. 2003,45(2):167 - 256.

[3] R Milo,S S Shen-Orr,S Itzkovitz,et al. Network motifs:Simple building blocks of complex networks[J]. Science,2002,298(5594):824 - 827.

[4] S S Shen-Orr,R Milo,S Mangan,U Alon. Network motifs in the transcriptional regulation network of Escherichian Coli[J]. Nature Genetics,2002,31(1):64 - 68.

[5] Lee T I,et al. Transcriptional regulatory networks in saccharomyces cerevisiae[J]. Science,2002,298,799 - 804.

[6] Milo R,Itzkovitz S,Kashtan N,et al. Superfamilies of evolved and designed networks[J]. Science,2004,303,1538 - 1542.

[7] Lahav G,Rosenfeld N,Sigal A,et al. Dynamics of the p53-Mdm2 feedback loop in individual cells[J]. Nature Genetics,2004,36,147 - 150.

[8] Barabasi AL,Oltvai ZN. Network biology:Understanding the cell's functional organization[J]. Nature Reviews Genetics,2004,5(2):101 - 114.

[9] Uri Alon. Network motifs:theory and experimental approaches[J]. Nature,2007,8,450 - 461.

[10] Berg J,Lässig M. Local graph alignment and motif search in biological networks[J]. Proc Natl Acad Sci,2004,101(41):14689 - 14694.

[11] Albert-László Barabási, Zoltán N. Oltvai. Network biology:Understanding the cell's fuctional organization[J]. Nature,2004,5:101 - 113.

[12] Milo R,Kashtan N,Itzkovitz S,Neuman M E J,Alon,U. On the uniform generation of random graphs with prescribed degree sequences[OL]. cond-mat/0312028,2003.

[13] Przulj N,Corneil DG,Jurisica I. Modeling interactome:scale-free or geometric? [J] Bioinformatics,2004,20(18):3508 - 3515.

[14] Przulj N,Corneil DG,Jurisica I. Efficient estimation of graphlet frequency distribution in protein-protein interaction networks[J]. Bioinformatics,2006,22(8):974 - 980.

[15] N Kashtan,S Itzkovitz,R Milo,U Alon. Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs[J]. Bioinformatics,2004,20(11):1746 - 1758.

[16] Wernicke S. A faster algorithm for detecting network motifs[A]. In Proceedings of the 5th Workshop on Algorithms in Bioinformatics (WABI'05). Lecture Notes in Bioinformatics[C]. Mallorca,Spain:Springer Verlag,2005,3692:165 - 177.

[17] Sebastian Wernicke. Efficient detection of network motifs[J]. IEEE Trans on Computational Biology and Bioinformatics,2006,3(4):347 - 359,

[18] Grochow JA,Kellis M. Network motif discovery using subgraph enumeration and symmetry-breaking[A]. In RECOMB 2007,Lecture Notes in Computer Science[C]. Oakland,CA:Springer. 92 - 106.

[19] Xiaofeng Zhou,Lin Gao,AnGuo Dong. An algorithm for finding frequent patterns in a large sparse graph[A]. In IAENG International Conference on Bioinformatics (ICB2007)[C]. Hong Kong,China,2007. 290 - 294.

[20] An guo Dong,Lin Gao,Xiaofeng Zhou,HongYu Su. An algebra approach for finding frequent subgraphs with Hamilton cycle[A]. The 4th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD07)[C]. Haikou,China:IEEE,2007. 288 - 292.

[21] Sebastian Wernicke ,Florian Rasche. FANMOD:a tool for fast network motif detection[J]. Bioinformatics,2006,22,1152 - 1153.

[22] Jiang Rui,Tu Zhidong,Chen Ting,et. al. Network motif identification in stochastic networks[J]. Proc Natl Acad Sci,2006,103(25):9404 - 9409.

[23] Jin Chen,Wynne Hsu,Mong Li Lee,See-Kiong Ng. NeMoFinder:Dissecting genome-wide protein-protein interactions with meso-scale network motifs[A]. In KDD'06[C]. Philadelphia,Pennsylvania,USA:ACM,2006. 106 - 115.

[24] Huan J,Wang W,Prins J,et al. Spin:mining maximal frequent subgraphs from graph database[A]. In KDD 2004[C]. Seattle,Washington,USA:ACM,2004. 581 - 586.

[25] Schreiber F,Schwobbermeyer H. Frequency concepts and pattern detection for the analysis of motifs in networks[J]. Transactions on Computational Systems Biology,2005,3:89 - 104.

[26] Chia-Ying Cheng, Chung-Yuan Huang, Chuen-Tsai Sun. Mining bridge and brick motifs from complex biological networks for functionally and statistically significant discovery[J]. IEEE Transactions on Systems, Man, and Cybernetics, 2008, 38(1): 17 - 24.

[27] Noga Alon, Phuong Dao, Iman Hajirasouliha, et al. Biomolecular network motif counting and discovery by color coding [J]. Bioinformatics, 2008, 24: i241 - i249.

[28] Giovanni Ciriello, Concettina Guerra. A review on models and algorithms for motif discovery in protein-protein interaction networks [J]. Briefings in Functional Genomics and Proteomics, 2008, 2: 147 - 156.

[29] Chen J, Hsu W, Lee ML, et al. Labeling network motifs in protein interactomes for protein function prediction [A]. In IEEE 23rd International Conference on Data Engineering (ICDE07) [C]. Istanbul, Turkey: IEEE Computer Society, 2007. 546 - 555.

[30] Babu M M, Luscombe N M, Aravind L, et al. Structure and evolution of transcriptional regulatory networks [J]. Curr Opin Struct Biol, 2004, 14: 283 - 291.

[31] Mason, O Verwoerd, M. Graph theory and networks in biology [J]. System Biology IET, 2007, 1(2): 89 - 119.

[32] Bader GD, Hogue CW. An automated method for finding molecular complexes in large protein interaction networks [J]. BMC Bioinformatics, 2003, 4: 2.

[33] King AD, Przulj N, Jurisica I. Protein complex prediction via cost-based clustering [J]. Bioinformatics, 2004, 20(17): 3013 - 3020.

[34] Van Dongen S. Graph Clustering by Flow Simulation [D]. Utrecht, Netherlands: University of Utrecht, 2000.

[35] Blatt M, Wiseman S, Domany E. Superparamagnetic clustering of data [J]. Phys Rev Lett, 1996, 76(18): 3251 - 3254.

[36] Arnau V, Mars S, Marin I. Iterative cluster analysis of protein interaction data [J]. Bioinformatics, 2005, 21(3): 364 - 378.

[37] Palla G, Dere'nyi I, Farkas IJ, et al. Uncovering the overlapping community structure of complex networks in nature and society [J]. Nature, 2005, 435 (7043): 814-818.

[38] Baskerville K, Paczuski M. Subgraph ensembles and motif discovery using a new heuristic for graph isomorphism [J]. Phys Rev 2006, E 74: 051903.

[39] Parida L. Discoverying topological motifs using a compact notation [J]. J Comput Biol, 2007, 4(3): 300 - 323.

[40] Royi Itzhack, Yelena Mogilevski, Yoram Louzoun. An optimal algorithm for counting network motifs [J]. Physica A, 2007, 381: 482 - 490.

[41] C Matias, S Schbath, E Birmele, et al. Network motifs: mean and variance for the count [J]. REVSTAT-Statistical Journal, 2006, 4(1): 31 - 51.

[42] F Picard, J.-J. Daudin, M. Koskas, et al. Assessing the exceptionality of network motifs [J]. Journal of Computational Biology, 2008, 15(1): 1-20.

[43] Inokuchi, T Washio, H Motoda. An apriori-based algorithm for mining frequent substructures from graph data [A]. Proc. Fourth European Conf. Principles and Practice of Knowledge Discovery in Databases (PKDD '00) [C]. Lyon, France: Springer-Verlag, 2000. 13 - 23.

[44] A Inokuchi, T Washio, H Motoda. Complete mining of frequent patterns from graphs: Mining graph data [J]. Machine Learning, 2003, 50(3): 321 - 354.

[45] M Kuramochi, G Karypis. An efficient algorithm for discovering frequent subgraphs [J]. IEEE Trans Knowledge and Data Eng, 2004, 16(9): 1038 - 1051.

[46] X Yan, J Han. g. Span: graph-based substructure pattern mining [A]. In Proc of 2002 IEEE Int'l Conf Data Mining (ICDM) [C]. Maebashi City, Japan: IEEE, 2002. 721 - 724.

[47] J Huan, W Wang, J Prins. Efficient mining of frequent subgraph in the presence of isomorphism [A]. In Proc of 2003 IEEE International Conference on Data Mining (ICDM '03) [C]. Melbourne, Florida USA: IEEE, 2003. 549 - 552.

[48] E Gudes, S E Shimony, N Vanetik. Discovering frequent graph patterns using disjoint paths [J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(11): 1441 - 1456.

[49] 高琳,覃桂敏,周晓峰. 图数据中频繁模式挖掘算法研究综述[J]. 电子学报, 2008, 36(8): 1603 - 1609.
Gao Lin, Qin Gui-min, Zhou Xiao-feng. An overview of algorithms for mining frequent patterns in graph data [J]. Acta Electronica Sinica, 2008, 36(8): 1603 - 1609. (in Chinese)

作者简介:

覃桂敏　女,1977 年生于广西象州. 西安电子科技大学软件学院讲师,现为西安电子科技大学计算机学院在职博士,目前感兴趣的问题为频繁子图的挖掘,生物网络数据的模式发现算法研究.
E-mail: gmqin@mail.xidian.edu.cn

高　琳　女,1964 年生于陕西乾县. 西安电子科技大学计算机学院教授,博士生导师,学术带头人. 2004 年 6 月至 2005 年 6 月被国家留学基金委批准选派赴加拿大 University of Guelph 做访问学者,主要研究方向包括计算生物信息学,生物数据挖掘,图论与组合优化算法及其应用等,在国内外核心期刊和国际会议发表学术论文 40 余篇.