一种用于生物网络数据的频繁模式挖掘算法

赵建邦[1],董安国[2],高 琳[1]
(1.西安电子科技大学计算机学院,陕西西安 710071; 2.长安大学理学院,陕西西安 710064)

摘 要: 频繁模式挖掘是生物网络数据分析中的一个核心问题,对于研究生物网络的组织结构和功能模块具有重要意义.本文提出了子图环分布的概念并构造了子图搜索算法,提高了搜索效率.其次设计了动态抽样算法计算子图频率,用于提高非穷举搜索的精度.利用 4 个真实生物网络数据进行仿真实验研究,验证了本文算法在效率和精度上相对于现有算法的优势.

关键词: 生物网络; 频繁模式; 子图搜索
中图分类号: TP311.12 **文献标识码:** A **文章编号:** 0372-2112 (2010) 08-1803-05

An Algorithm for Frequent Pattern Mining in Biological Networks

ZHAO Jian-bang[1], DONG An-guo[2], GAO Lin[1]
(1. *School of Computer Science and Technology, Xidian University, Xi'an, Shaanxi* 710071, *China*;
2. *School of Science, Chang'an University, Xi'an, Shaanxi* 710064, *China*)

Abstract: Frequent pattern mining has emerged as a key issue for analyzing the biological networks since it gives us insights into the organism and functional modules. A novel algorithm for this problem is proposed, which can efficiently obtain all these frequent subgraphs in networks based on the distribution of ring. To improve the accuracy of subgraph mining in non-exhaustive enumerate mode, additionally, we provide a dynamic sample algorithm. The experimental results in four real bio-networks show the superiority of our algorithm to existing algorithms.

Key words: biological network; frequent pattern; subgraph search

1 引言

网络模体的识别是研究结构组成和功能单元的重要手段.自 2002 年 R Milo 等人[1]在 Science 上提出网络模体这个概念以来,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,网络模体的发现和分析已经成为目前生物信息学的研究重点和热点之一.网络模体识别的核心计算问题是在生物网络构建的拓扑结构图中进行频繁模式的挖掘.

2000 年,A Inokuchi 等人[2]提出了一个基于 Apriori 思想的频繁子图模式挖掘算法 AGM,最初 AGM 只能发现频繁导出子图,后来该算法得到扩展可以发现所有的频繁子图.2002 年以后,各种不同的频繁模式挖掘的算法被提出来,比较有影响的有:X Yan 等[3]提出的 gSpan 算法、Han J 等[4]提出的 FFSM 算法,以及 M Kuramochi 等人[5]提出的 FSG 算法.2004 年 N Kashtan 等[6]在 R Milo 算法的基础上提出了一种以采样方法搜索子图的改进算法——ESA (Edge Sampling Algorithm),但是 ESA 算法存在偏差估计的缺陷[7].2006 年 S Wernicke[8]提出了一种枚举所有子图的新的子图搜索算法——ESU (Enumerate Subgraphs),此算法克服了 ESA 算法的种种缺陷,并且运行效率也得到了提高.除了上述这些图模式挖掘的算法外,研究人员还提出了运用于实际问题的带有约束的子图模式挖掘算法[9].尽管目前有很多关于频繁模式挖掘的算法,但是由于子图搜索本身的复杂性,导致现有的算法效率还不尽如人意.

频繁模式挖掘包括两个主要步骤:子图搜索和子图频率计算.本文首先引入了节点环分布的概念,提出了一种基于节点环分布的子图搜索算法.子图频率计算涉及到子图同构分类,该问题是 NP - 完全问题,当子图阶数较高时(大于 5),对子图进行同构分类的时间代价很高,甚至无法在有效时间内完成,成为频繁模式挖掘的一个瓶颈.为此,本文设计了动态抽样算法,将子图搜索算法和抽样过程结合在一起,以概率取子图样本,利用同构分类算法估计出子图频率,在确保计算精度的前提下提高了频繁子图挖掘的效率.

收稿日期:2008-12-09;修回日期:2010-03-24
基金项目:国家自然科学基金重点项目(No.60933009);高等学校博士学科点专项科研基金(No.200807010013)

2 符号说明及基本结论

文中涉及到的关于图的基本术语和概念与图论中一致,这里只对算法中用到的概念和符号进行说明.设一个给定的图 $G=(V,E)$,记 $n=|V|$,表示图 G 的阶数,V 中的节点依次用 $v_1,v_2,\cdots,v_n$ 来表示,E 表示图 G 中边的集合.

定义 1 $G=(V,E)$为一个给定的图,称 $G_s=(V_s,E_s)$为 G 的 k 阶子图,当且仅当 $V_s\subseteq V$,$E_s\subseteq E$ 且 $|E_s|=k$,特别地,如果 E_s 包含 E 中所有连接 V_s 中节点的边,则称 $G_s=(V_s,E_s)$为 G 的导出子图.

对一个给定的 k,图 G 中的所有 k 阶子图 $S_k(G)$可以按同构关系分为若干个类,称为等价类,第 i 个等价类中的子图集记为 $S_k^i(G)$,第 i 类子图的频率定义为 $f_i=\dfrac{|S_k^i(G)|}{\left(\sum\limits_j|S_k^j(G)|\right)}$.

定义 2 在图 $G=(V,E)$中,从节点 i 到 j 需要经过的最少的边数称为节点 i 到 j 的距离.

定义 3 设 G_k 是一个 k 阶子图,v 是 G_k 的一个节点,G_k 中与 v 的距离为 l 的节点的集合称为节点 v 的第 l 个环.

设 G_k 是一个 k 阶子图,v_0 是 G_k 的一个节点,则 v_0 最多有 $k-1$ 个环,各环中分布的节点个数$[x_1,x_2,\cdots,x_{k-1}]$称为 G_k 以 v_0 为中心的环分布,所有可能的环分布用 $\boldsymbol{P}^{(k)}$ 表示,即 $\boldsymbol{P}^{(k)}=(\boldsymbol{P}_{ij}^{(k)})_{N\times(k-1)}$,其中 $N=2^{k-2}$.$\boldsymbol{P}_{ij}^{(k)}$表示第 i 类环分布在节点 v_0 的第 j 个环上的节点个数.例如,3 阶连通图所有可能的分布类型有两种,即 $\boldsymbol{P}^{(3)}=\begin{bmatrix}2&0\\1&1\end{bmatrix}$.

在图 G 中搜索包含节点 v_0 的所有 k 阶子图时,只要遍历完 2^{k-2}种分布类型的子图即可.

3 基于环分布的子图搜索算法

3.1 算法描述

为了描述该算法,用 R 表示 G 中所有 k 阶子图集合,R_j 表示最小的节点序号为 j 的所有 k 阶子图集合,显然,

$$R=\bigcup_{j=1}^{n-k+1}R_j,\ R_i\cap R_j=\Phi,\ i\neq j \tag{1}$$

要得到 R,首先要搜索出 $R_j(j=1,2,\cdots,n-k+1)$.根据 R_j 的定义,如果 R_1 的搜索算法已经实现,只要将 G 中的节点 1 以及与它相连的边去掉,并将所有节点序号均减 1,重复 R_1 的搜索过程即可得到 $R_j(j=2,3,\cdots,n-k+1)$.

3.2 算法流程

定义 $\boldsymbol{A}$ 为图 G 的邻接矩阵,定义函数 ExtendSubgraph(*subgraph*, i, s, R)完成第 i 个分布类中第 s 个环的节点扩展(每步扩展 $\boldsymbol{P}_{is}$个节点),其中 R 为当前搜索的环,环的扩展过程由函数 Nexcl(w, *subgraph*)实现,它表示与 w 至少有一个节点有连边,而与 *subguaph* 没有连边的节点的集合,即 Nexcl(w, *subgraph*) = Neighbor(w) − Neighbor(*subgraph*).则关于图 G 的所有 k 阶子图搜索算法如图 1 所示.

```
Algorithm: ESRD(Enumerate Subgraphs based on Ring-Distribution)
Input: A graph G=(V,E) and an integer 1≤k≤|V|.
Output: All size-k subgraphs in G.
01: for  j=1: n-k+1 do:
02:      subguaph = w = {j};
03:      R = Neighbor(w);
04:      for i=1:2^(k-2) do
05:          s=1;
06:          subgraph = ExtendSubgraph(subgraph, i, s, R)
07:      end for
08:      delete the first row and column of A;
09: end for
ExtendSubgraph(subgraph, i, s, R)
E01: if |subgraph| = = k then output subgraph and return;
E02: while |subgraph| ≠0 and |R| ≥P_is
E03:    for all w(w 表示从 Nexcl(w, subgraph)中任取的 s 个节点)
E04:    R_1 = Neighbor(w, subgraph);
E05:    subgraph = subgraph ∪ w;
E06:    s = s+1;
E07:    subgraph = ExtendSubgraph(subgraph, i, s, R_1)
E08:    end for
E09: end while
```

图 1 子图搜索算法 ESRD

3.3 与现有算法的比较

根据节点是否带有标记,子图搜索问题可以应用于不同的背景.对于带有标记的 k 阶频繁模式挖掘,不需要搜索所有的 k 阶子图,只要逐步扩展满足“频繁”要求的子图就可以达到目的.典型的算法有前面提到的 AGM[2], FSG[5], gSpan[3] 和 FFSM[4] 等.其中,AGM 和 FSG 基于启发式的广度优先搜索,其它几种是基于启发式的深度优先搜索.

针对网络规模较大的单个图内的穷尽子图搜索以及子图的生物意义评估问题,现有的主要算法有 ESA[6] 和 ESU[8].

ESA 算法有两部分内容:首先对真实的生物网络进行穷尽搜索,找到所有的 k 阶子图,并对这些子图按照拓扑进行同构类统计;其次在大量的随机网络内(随机网络个数不少于 100)进行采样搜索,每次采样只得到一个 k 阶子图.ESA 穷尽搜索基于边扩展的方式,每步扩展时,从已有子图出发,将与子图邻接的边加入从而形成更大规模的子图.这种方式的缺陷是:由于边数一般大于节点数,因此遍历边会增加复杂度.该算法的穷

尽搜索效率在表1中给出的几种算法中较低,但是其采样速率不受网络规模的影响.其主要应用是:通过随机采样,利用很少一部分的采样子图估计出相对精确的子图浓度.

ESU是一种效率较高的子图穷举搜索算法,运行复杂度是 $O(n^{k-1})$,其中 n 为节点个数,k 为子图规模.其局限性在于:从规模为 i 的子图集进行第 $i+1$ 个节点的扩展时,每次只能从候选集中挑选一个节点,继续下次迭代运算.而ESRD在进行迭代的过程中可以选择多个候选节点,减少了搜索空间及大量由于递归造成的集合间的合并操作,节约了很大程度的时间,付出的代价是占用较大的内存空间.三种算法的性能比较参考表1.

表1　子图数量及搜索时间

node/Edge			size-4		size-5		size-6		size-7	
			number	Time	number	Time	number	Time	number	Time
E coli	418/519	ESRD	83893	0.031	1433502	0.516	22532584	8.516	319521581	301.56
		ESU	83893	0.344	1433502	6.28	22532584	155.6	319521581	2923
		ESA	83893	2.844	1433502	194.453	—	>3h	—	>3h
SeaUrchin	45/83	ESRD	2212	0.0	11043	0.016	49320	0.031	196082	0.125
		ESU	2212	0.0	11043	0.05	49320	0.28	196082	0.9
		ESA	2212	0.078	11043	0.984	49320	20.641	—	>3h
Yeast	688/1079	ESRD	183174	0.047	2508149	0.781	32883898	20.36	416284878	603.9
		ESU	183174	0.65	2508149	11.1	32883898	218.4	416284878	5148.5
		ESA	183174	7.063	2508149	357.516	—	>3h	—	>3h
Protein	270/716	ESRD	118129	0.062	1685010	0.641	22990600	8.656	297549099	157.06
		ESU	118129	0.375	1685010	4.172	22990600	69.3	297549099	1228.08
		ESA	118129	4.328	1685010	144.656	—	>3h	—	>3h

4　基于环分布的动态抽样算法

为了找出频繁子图,对从图 G 中搜索到的每一个子图都要按同构关系进行分类.而当子图节点数较大(大于5)时,搜索到的子图数量很大,则同构分类的时间代价会很高.事实上,为了得到频繁子图,没有必要精确地计算各等价类子图发生的频率,根据大数定律,只要按一定的规则从图 G 中随机抽样出部分子图(N 个),对这 N 个子图进行同构分类,就可以得到各等价类子图发生频率的近似值,并且误差通过关于 N 的函数来控制.下文将给出动态抽样的规则并从理论上进行论证.

4.1　动态抽样规则描述

为了确保频率估计的精度,必须保证两个条件,首先,样本量不能太小,其次,从图 G 中随机抽样出的样本要具有代表性.

4.1.1　抽样规则的确定

为了确保样本的代表性,对每一类分布结构的子图按概率 p 进行独立抽样,只要将ESRD算法中函数ExtendSubgraph(*subgraph*, i, s, R)的E03部分改为:

for all w(w 表示在 R 的所有 $\boldsymbol{P}_{is}$ 个节点的组合中以概率 p_s 取出的样本),则子图搜索算法便成为动态抽样算法.通过抽样,不需要找出图 G 中的所有子图,这样就提高了算法的空间效率和时间效率.下文的分析以及实验均表明,通过抽样估计出的子图频率具有较高的精度.

4.1.2　子图数量的估计

为确定抽样概率 p,首先要估计子图总数 M,为此先给 p 赋一个很小的值($p=0.001$).假设按4.1.1的抽样规则得到的样本总数为 $\overline{M}$ 个,则子图总数 $M\approx\overline{M}/p$.

4.1.3　抽样概率确定

设子图总数为 M,当 $M<10^6$ 时,利用第2节和第3节的算法,对频率进行精确计算,时间代价也不高(见实验结果).不失一般性,假定 $M\geqslant 10^6$.由于子图数量比较庞大,为了确保计算精度,抽取样本数量控制在 10^6 左右,即每个子图被选到的概率 $p=10^6/M$,为了确定 p,可以利用4.1.2的方法估计 M.

4.2　动态抽样算法的理论分析

4.2.1　各层抽样概率 p_s 的确定

根据上文,由 h_1^r 的定义,h_1^r 服从二项分布 $B(p_1, h_1)$,所以,其数学期望 $E(h_1^r)=p_1h_1$.同理,$E(h_{2i}^r)=p_1h_{2i}$,$E(h_{3j}^r)=p_1h_{3j}$,显然,第一层选到的节点数的期望值 $E(h_1^r)$ 是原有节点数 h_1 的 p_1 倍,选到的节点在第二层的分支中又以概率 p_1 进行抽样,所以第二层节点数的期望值 $E(h_2^r)$ 是 h_2 的 p_1^2 倍.同理,$E(h_3^r)=p_1^3h_3$,这样该类分布结构的子图被选到的概率为 p_1^3,由于 $p_1^3=p$,所以 $p_1=\sqrt[3]{p}$.一般的,若环分布类的环数为 u,则

各环上的抽样概率为 $p_1=\sqrt[u]{p}$.

4.2.2 各等价类子图频率估计及误差分析

设 $S_k(G)$表示图 G 的所有 k 阶子图，S_i 表示图 G 的具有第 i 种分布结构的 k 阶子图，其中，属于第 j 个等价类的集合为 S_i^j，A_j 表示第 j 个等价类的子图，对于 k 阶子图 $G_k\in S_k$，定义随机变量 $X=j$(如果 $G_k\in A_j$)，并设 X 的概率分布为 $P(X=j)=p_j$，这样，只要估计出 p_j 的值，就完成了频繁模式的挖掘.为此，引入随机变量 $Y=i$(如果 $G_k\in S_i$)，根据全概率公式，$p_j=P(X=j)=\sum_i P(X=j/Y=i)P(Y=i)$，显然 $P(Y=i)=|S_i|/|S_k|$.设抽样到的第 i 个分布结构的子图数量为 m_i，其中，属于第 j 个等价类的个数为 m_i^j，由最大似然估计，$P(X=j/Y=i)$的估计值是 m_i^j/m_i，所以 p_j 的估计值为：

$$\bar{p}_j=\sum_i \frac{m_i^j}{m_i}|S_i|/|S_k|=\sum_i \frac{m_i^j}{p|S_k|}=\frac{1}{N}\sum_i m_i^j \quad (2)$$

其中 N 表示样本总数.显然，m_i^j 表示涤 S_i^j 中以概率 p 抽到的子图个数，所以，$m_i^j\sim B(|S_i^j|,p)$，从而，$E(m_i^j)=p|S_i^j|$，$D(m_i^j)=p(1-p)|S_i^j|$，由中心极限定理，$\bar{p}_j=\frac{1}{N}\sum_i m_i^j\sim N(\mu,\sigma^2)$，其中，

$$\mu=E(\frac{1}{N}\sum_i m_i^j)=\frac{1}{N}\sum_i p|S_i^j|=\frac{p}{N}|A_j|=\frac{|A_j|}{|S_k|}=p_j \quad (3)$$

所以 $\bar{p}_j-p_j\sim N(0,\frac{(1-p)}{N}p_j)$，根据 3σ 原则，$|\bar{p}_j-p_j|<3\sqrt{\frac{(1-p)p_j}{N}}<3\sqrt{\frac{p_j}{N}}$，相对误差 $\frac{|\bar{p}_j-p_j|}{p_j}<3\sqrt{\frac{1}{Np_j}}$，对于频繁子图对应的同构类，其频率不会太小，按保守估计，频繁子图的频率 $p_j>0.01$，由 4.1.1 的说明，$N\approx 10^6$，所以相对误差 $\frac{|\bar{p}_j-p_j|}{p_j}<0.03$.

分析表明，利用抽样得到部分子图对子图的频率进行估计，其相对误差在 3% 以内，与实验结果相一致，优于文献[8]计算结果.

5 实验结果与分析

四种实验数据分别是是 E.coli、Yeast、SeaUrchin 的基因调控网络[10]和规模较小的一种蛋白质相互作用网络 Protein.

5.1 子图搜索速度比较

在实验中，对 4 个真实的网络，用本文的算法和 ESA[6]以及 ESU[8]，搜索了 4 到 7 个节点的子图，对搜索时间进行了比较，具体结果见表 1，从表中可以看出，本算法找出的子图数量与参考文献的结果完全一样，而运算效率明显高于参考文献.

说明：ESRD 是本文算法，ESA 和 ESU 分别是文献[6,8]的算法，时间单位是秒；‘>3h’表示运行时间超过 3 小时.

5.2 随机搜索算法的精度比较

根据 4.2.2 的分析，估计子图频率的误差可以严格控制，本文在估计 7 阶子图的频率时，利用动态抽样算法，抽取样本约 100 万个(抽样概率约为 0.003)，统计出频率，并与精确计算进行比较，如图 2、图 3 所示，实线和虚线分别表示基于抽样计算(sample)和搜索(search)在不同生物网络(E.coli 和 Protein)中得到的子图频率曲线.实验结果表明，本文的抽样方案优于文献[8]的方案，频率估计的精度高于文献[8]，所以，在精度要求不高的情况下，可以进一步减少样本数量，提高频繁模式的挖掘效率.

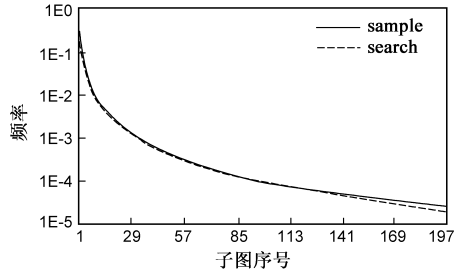


图2 Ecoli网络子图频率精确值与估计值

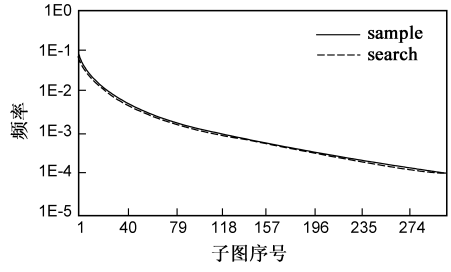


图3 Protein网络子图频率精确值与估计

5.3 频繁模式的挖掘结果

对 4 种不同的生物网络，利用基于环分布的动态抽样算法，对 4-7 阶子图的频率进行了估计，挖掘出各个网络的频繁模式，表 2 列出了前三个频繁模式结果.

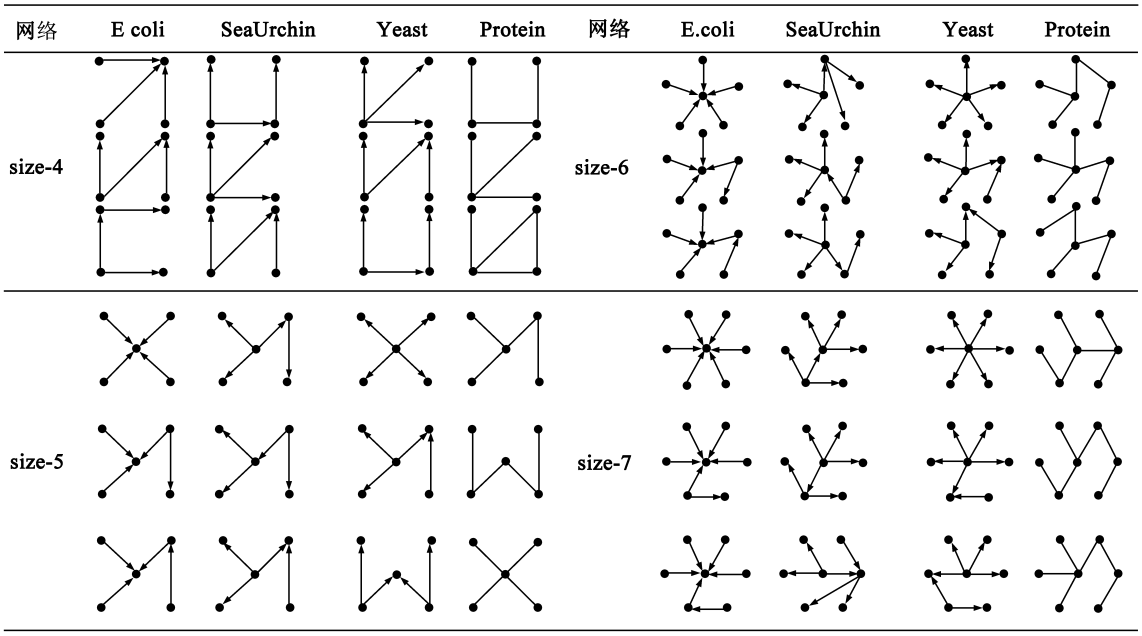
从表 2 可以看出，同一个网络中低阶频繁子图在多数情况下是高阶频繁子图的子图，例如 E.coli 网络的 4 个节点的频繁子图正好是 5 个节点的频繁子图的子图，这与预期的结果是一致的.

6 结论

围绕频繁模式挖掘问题，本文提出了环分布的概念，在此基础上构造了两个算法：第一个是基于环分布的子图搜索算法，它包括环分布类型的确定和子图搜索；第二个是动态抽样算法，将随机抽样过程嵌入到子

图搜索的每一个环节中,减少了搜索结果集和搜索时间,有效地减少了统计子图频繁的时间.通过对 4 个真实网络数据的仿真实验研究,表明本文提出频繁模式挖掘算法优于 ESU 算法.此外,该算法还可以应用到大型网络(生物网络,社会网络等)的模体发现问题中.

表 2　频率模式列表



参考文献:

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

[2] A Inokuchi, T Washio, H Motoda. An apriori-based algorithm for mining frequent substructures from graph data[A]. Proc of European Conf on Principles of Data Mining and Knowledge Discovery(PKDD 2000)[C]. London, UK: Springer-Verlag, 2000. 13 – 23.

[3] X Yan, J Han. gSpan: graph-based substructure patterns mining [A]. Proceedings of IEEE the 2002 International Conference on Data Mining(ICDM 2002)[C]. Washington DC, USA: IEEE Computer Society, 2002. 721 – 724.

[4] Huan J, Wang W, Prins J. Efficient mining of frequent subgraphs in the presence of isomorphism[A]. Proc of the IEEE International Conference on Data Mining (ICDM 2003)[C]. Washington DC, USA: IEEE Computer Society, 2003. 549 – 552.

[5] M Kuramochi, G Karypis. An efficient algorithm for discovering frequent subgraphs[J]. IEEE Transactions on Knowledge and Data Engineering, 2004, 16(9): 1038 – 1051.

[6] 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.

[7] 覃桂敏,高琳,呼加璐.生物网络模体发现算法研究综述[J].电子学报,2009,37(10):2258 – 2265.
QIN Gui-min, Gao Lin, Hu Jia-lu. A Review on algorithms for network motif discovery in biological networks[J]. Acta Electronica Sinica, 2009, 37(10): 2258 – 2265. (in Chinese)

[8] S Wernicke. Efficient detection of network motifs[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2006, 3(4): 347 – 359.

[9] H Hu, X Yan, Y Huang, et al. Mining coherent dense subgraphs across massive biological networks for functional discovery[J]. Bioinformatics, 2005, 21(1): 213 – 221.

[10] Uri Alon Lab. Weizmann Networks[DB/OL]. www.weizmann.ac.il/mcb/UriAlon/groupNetworksData.html, 2008.

作者简介:

赵建邦　男,1983 年生于陕西渭南,现在西安电子科技大学计算机学院攻读计算机应用技术专业博士学位,研究方向为生物信息数据挖掘.
E-mail: zjb9797@foxmail.com

董安国　男,1964 年生于浙江象山,西安长安大学教授,硕士生导师,于 1987 年毕业于西安交通大学数学系.主要从事生物信息学、图论与矩阵论算法及其应用的研究工作.