计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2015.20140801
52(12):2824-2833,2015

一种基于随机游走的迭代加权子图查询算法

张小驰　于　华　宫秀军
(天津大学计算机科学与技术学院　天津　300072)
(天津市认知计算与应用重点实验室　天津　300072)
(1037903644@qq.com)

A Random Walk Based Iterative Weighted Algorithm for Sub-Graph Query

Zhang Xiaochi, Yu Hua, and Gong Xiujun
(*School of Computer Science and Technology, Tianjin University, Tianjin* 300072)
(*Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin* 300072)

Abstract　Nowadays, technological development in measuring molecular interactions has led to an increasing number of large-scale biological molecular networks. Identifying conserved and stable functional modules from such networks helps not only to disclose the function of inner components, but also to understand their relationships systematically in complex systems. As one of classical NP-complete problems, the sub-graph query problem is gaining research efforts in analyzing such behaviors from the fields of social networks, biological networks, and so on. Calculating node similarities and reducing the sizes of target graphs are two common means for improving query precisions and reducing computational complexity in the study of sub-graph algorithms. For the problem of querying sub-graphs among complex protein interaction networks, this paper presents a sub-graph query algorithm based on semi-Markov random walk model. A comprehensive similarity measurement based on semi-Markov random walk model is designed to integrate the similarities of nodes, structures and their neighbors. Meanwhile, an iterative procedure is applied to reduce the size of targeted graph by removing nodes in a cluster with lower similarities by calculating the global correspondence score. The experimental results on multiple real protein query networks demonstrate that the proposed algorithm improves its performance in both query precisions and computational complexity.

Key words　semi-Markov random walk model; iterative weighting; protein-protein interaction; biological molecule network; sub-graph query

摘　要　作为经典的NP完全问题之一,子图查询算法近年来在社交网络、生物分子网络等复杂系统分析中引起研究人员的极大关注.结点相似性计算和目标图约简是子图查询算法中提高查询准确率和降低计算复杂性的2种常用手段.针对复杂生物分子网络之间的子图查询问题,提出了一种基于半Markov随机游走的迭代加权子图查询算法.在结点相似性计算中,设计了基于半Markov游走模型的集成结点本身相似性、结构相似性及邻居结点相似性的综合度量方法;同时,在目标图约简过程中,通过迭代递减目标图中低相似性结点,以降低目标图的规模.对多个真实蛋白质网络查询的实验结果表明,算法在精度和时间复杂性方面都有明显提高.

收稿日期:2014-09-03;**修回日期**:2015-02-03
基金项目:国家自然科学基金项目(61170177);国家"九七三"重点基础研究发展计划基金项目(2013CB32930X)
通信作者:宫秀军(gongxj@tju.edu.cn)

关键词 半 Markov 游走模型;迭代加权;蛋白质相互作用;生物分子网络;子图查询

中图法分类号 TP18

从复杂的大规模网络结构中辨识保守、稳定的功能模块不仅可以揭示复杂系统内部元件的功能,而且可以系统化地理解元件内部的相互关系.例如在社交网络中,内聚的子网代表了内部各角色的共性行为,同时,普通用户也可以从了解他们邻居的行为中获得启发[1].在生物网络中,模块化的结构,如蛋白质复合物、蛋白质域,在生物功能的实现中占据着重要的地位[2].随着生物信息技术的发展,研究人员得以更多地运用数据挖掘的方法识别和分析重要的生物分子结构,如生物分子网络的聚类算法、重要的拓扑结构、化学组成分析以及生物分子图谱等[3-4].因而,子图挖掘正在成为机器学习领域的研究热点之一.

在生物信息领域,子图查询(query)或比对(alignment)的目标是从一个大规模生物网络中识别出一个与给定的查询网络高度相似的子网区域.以蛋白质为例,这些算法需要将 2 个网络中蛋白质本身的相似性(如序列相似性、结构相似性、功能相似性等)和它们在网络中的拓扑结构相似性融合起来,从而获得具有重要研究价值的生物结构.这种技术通常被用作从一些研究程度比较深入的物种或组织中去发现一些其他不易获得研究数据的物种或组织的类似的生物结构.然而,到现在为止,该问题仍然是一个类似于子图同构(subgraph isomorphism)问题的 NP 完全问题[5],巨大的计算开销成为制约其应用的一个重要瓶颈.

基于此,研究人员试图从不同角度来改进经典子图算法的性能.PathBLAST[6]是一种基于贪婪的"种子扩展"算法,它从 2 个网络中识别相似线性通路的网络查询,但是由于其高计算复杂度的缺陷,使得这个算法只适用于短通路的识别.为了降低计算复杂性,研究人员提出了基于颜色编码的 QPath[7]和 QNet[8]等算法,但只适用于一些规则的拓扑结构查询,类似地还有 PathMatch[9]和 GraphMatch[9]等众多算法.

2008 年,Singh 和 Xu 等人[10]提出了基于 pagerank 的 IsoRank 算法,将生物分子的序列相似性和其在生物网络中的拓扑相似性结合起来,不仅突破了查询网络的结构限制,而且提高了子网匹配的准确性;2012 年,Sahraeian 等人[11]提出了一种计算复杂性低的网络查询算法——RESQUE,该算法利用 Markov 随机游走模型迭代计算查询子图与目标图的全局匹配得分,进而递减目标图的规模,从而降低了子图的查询计算复杂性.

总结近几年在网络比对或网络查询领域中的优秀算法,RESQUE 和 IsoRank 算法均能适应于各种结构的子网查询或匹配,而没有只局限在线性或拥有特殊的图形结构(如树形结构)的子网上.然而,RESQUE 算法的计算复杂度较低,但查询准确率也较低;IsoRank 算法的查询准确率比较高,但计算复杂度也比较高.本文针对 IsoRank 和 RESQUE 算法的上述缺陷,提出一个在保证 RESQUE 算法的低计算复杂性的情况下,又同时具有 IsoRank 算法较高查询准确率的改进算法——RESCUEH 算法,算法在处理复杂数据、庞大生物分子网络查询任务时的运行效率有明显提高.

1 算法描述

随机游走是一种模拟物体运动过程的数学模型.该模型首次由 Pearson[12]于 1905 年引入,现已成功应用到经济学、物理学、计算机科学及生物学等多个领域.随机游走模型假定停留在某一位置的时间是相同的,根据物体行走过程中行走步长及相邻状态之间的关系,模型可以分为不同的类型.如果行走步长服从高斯分布、相邻状态满足 Markov 过程,则称之为高斯 Markov 随机游走模型(Gaussian Markov random walk model);如果停留时间也是随机的,则称之为半 Markov 随机游走模型(semi-Markov random walk model, SMRW).本文利用 SMRW 模型来计算子图的相似度.具体算法描述如下:

假设目标蛋白质相互作用网络可以抽象成一个有权无向图 $\boldsymbol{G}=(V,E,W)$,其中,V 是蛋白质相互作用网络中蛋白质节点的集合,总数为 N,E 是该蛋白质相互作用网络中相互作用的集合,总数为 m;而 $\omega:\varepsilon\to\mathbb{R}$,表示的是一个相互作用的可信度.同样,定义查询网络 $\boldsymbol{G}_Q=(V_Q,E_Q,W_Q)$ 是包含蛋白质结点的集合 V_Q 及可信度为 W_Q 的蛋白质相互作用集合 E_Q 的蛋白质相互作用网络.对于每一对蛋白质对 (q_i,v_j)(其中 q_i 属于查询网络,v_j 属于目标网络),其本身的相似性用 $s(q_i,v_j)$ 来表示,可以用 2 个蛋白质之间的序列相似性来度量,也可以选择用蛋白质之间的结构相似性或者功能相似性来表示该属性.

算法的基本流程图如图 1 所示：

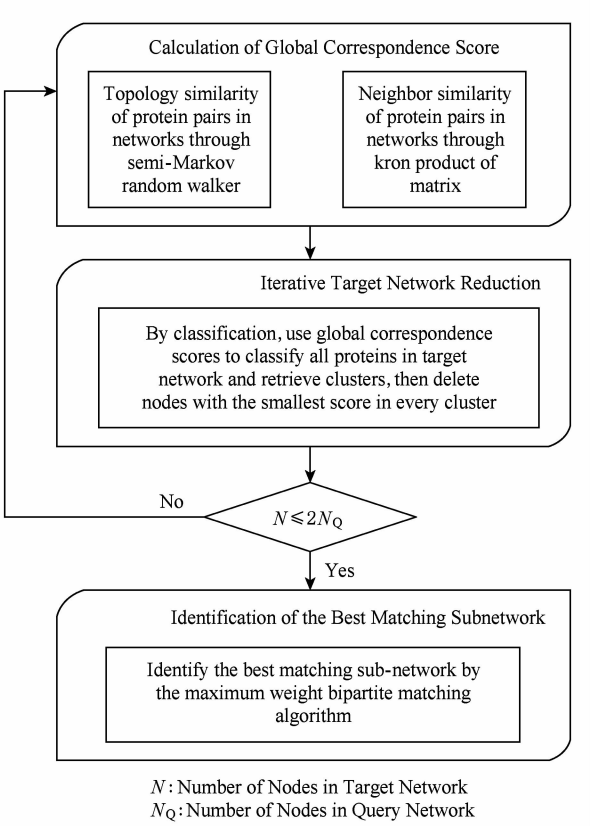


Fig. 1 The main stream of the algorithm.

图 1 算法的主要流程

1.1 计算蛋白质对的全局相似性得分

2 个网络之间的蛋白质对的全局相似性得分(global correspondence score, GCS)是进行最优子网匹配过程中的重要影响因素. IsoRank 算法考虑了以下 3 个影响因素：匹配蛋白对本身的相似性、匹配蛋白质对周围 interaction 结构的相似性以及匹配蛋白质对周围邻居蛋白质之间的相似性；而 RESQUE 算法忽略了第 3 个因素，仅仅考虑了前 2 个因素. 经研究发现，RESQUE 算法识别最优匹配子网的准确度比 IsoRank 算法低，因此，猜想这是由于匹配蛋白质对周围邻居蛋白质之间的相似性这个因素的缺失造成的，所以我们将匹配蛋白质对周围邻居蛋白质之间的相似性这个因素融合到全局相似性的计算当中. 利用 2 个网络中蛋白质之间的 blast score 表构成的矩阵 $\boldsymbol{S}$、查询网络的无权图的邻接矩阵 $\boldsymbol{Q}$，以及目标网络的无权图的邻接矩阵 $\boldsymbol{T}$，通过 kron 积的特性——式(1)(2)，计算出每对蛋白质之间的邻居节点的序列相似性平均值.

$$(\boldsymbol{T}^{\mathrm{T}} \otimes \boldsymbol{Q})vec(\boldsymbol{S}) = vec(\boldsymbol{QST}) = vec(\boldsymbol{H}), \quad (1)$$

$$h(q_i, v_j) = \frac{\sum_{q_{i'} \in N_i, v_{j'} \in N_j} s(q_{i'}, v_{j'})}{N_i \times N_j}, \quad (2)$$

其中，$\boldsymbol{H}$ 是由 $h(q_i, v_j)$ 构成的矩阵，q_i 是查询网络中的蛋白质节点，v_j 是目标网络中的蛋白质节点；N_i，N_j 分别是 q_i，v_j 的邻居节点的个数.

而蛋白质对在网络中关联的 interaction 的结构相似性，我们参考 RESQUE 算法中的方法，利用半 Markov 随机游走模型. 首先将查询网络与目标网络的积抽象成一个 Markov 链，然后计算该网络中每个蛋白质节点的转移概率 $a_q(i,k)$，如式(3)所示，得到整个转移概率矩阵 $\boldsymbol{A}_Q$：

$$a_q(i,k) = \frac{\omega_q(i,k)}{\sum_{k' \in N_i} \omega_q(i,k')}, \quad (3)$$

其中，$\omega_q(i,k)$ 是网络中的蛋白质相互作用的可信度. 同样的方法，我们计算出目标网络的状态转移矩阵 $\boldsymbol{A}$. 然后，利用状态转移矩阵，通过寻找转移矩阵的特征向量，计算出每个节点的稳态值 $\pi_Q(q_i)$(目标网络 $\pi(v_j)$)，它表示了在随机游走过程中每个节点上停留的时间长短.

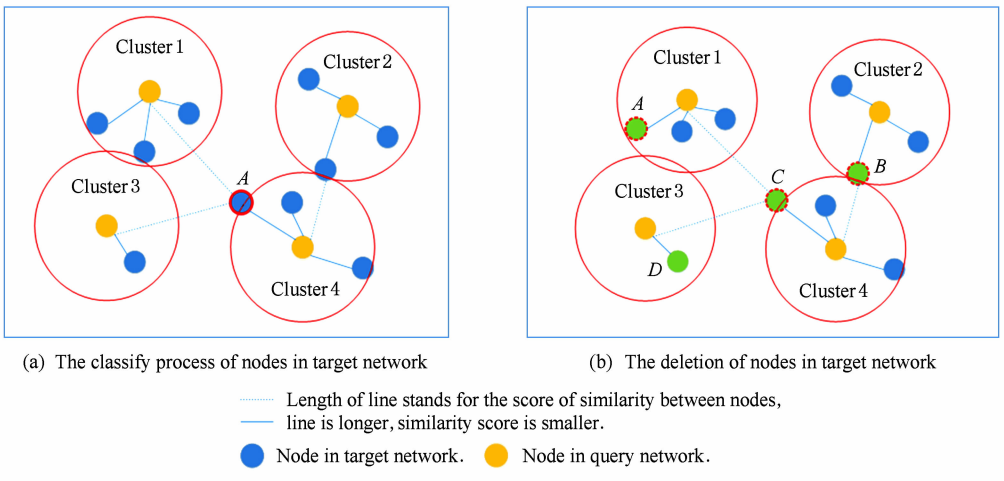
在计算每对蛋白质对(一个蛋白质属于查询网络，一个蛋白质属于目标网络)的 GCS 时，按照式(4)融合上述 3 种影响因素的计算得分，继而得到 GCS 矩阵 $\boldsymbol{C}$(一般 $\alpha=0.5$).

$$c(q_i, v_j) = (1-\alpha)\frac{h(q_i, v_j) \times s(q_i, v_j)}{\sum_{i',j' \in S} h(q_{i'}, v_{j'}) \times s(q_{i'}, v_{j'})} + \alpha \frac{\pi_Q(q_i) \times \pi(v_j) \times s(q_i, v_j)}{\sum_{i',j' \in S} \pi_Q(q_{i'}) \times \pi(v_{j'}) \times s(q_{i'}, v_{j'})}. \quad (4)$$

1.2 迭代递减目标网络

在进行最优匹配子网的查询过程中，为了提高匹配的可靠性，RESQUE 算法提出了一种使目标网络迭代递减的思想，当目标网络中存在大量的与查询网络中蛋白质节点非相关的节点时，矩阵 $\boldsymbol{C}$ 中的节点对应的 GCS 会倾向于较低可靠性. 为解决这一问题，RESQUE 算法通过更新 GCS 矩阵，计算 2 个网络中所有蛋白质对 GCS 极小值，最终选取目标网络与查询网络中蛋白质的 GCS 都最小的蛋白质进行删除. 但是在查询网络和目标网络的规模都比较大的情况下，目标网络中与查询网络中节点相似的节点可能会比较多，导致迭代的次数会迅速增长，计算量呈指数级增大. 所以，在保证匹配准确率的前提下，改进这一缺陷，考虑将 GCS 矩阵按与查询网络中蛋白质的相似性大小分类，如图 2(a)所示，给出了目标网络节点的分类过程. 节点 A 和 3 个查询网络中的节点相似，但是 Cluster4 中的节点和节点 A 最为相似，因此节点 A 被归类到 Cluster4 中. 删除每个类别中距离查询网络蛋白质最远(即相似性最

小)的目标网络中的蛋白质,除非它们是该分类中唯一的蛋白质节点,以此来快速减少迭代的次数,如图 2 所示. 图 2(b)中,节点 A,B,C,D 分别与其对应查询网络中的蛋白质节点最远,但由于节点 D 是 Cluster3 中唯一的节点,因此节点 D 被保留下来,删除节点 A,B,C. 算法通过迭代递减过程,分类过滤出与查询网络中蛋白质相似性最小的目标网络蛋白质节点集合,迅速缩小目标网络规模;同时提高待匹配蛋白质节点之间 GCS 的可靠性,为下一步的最优子网匹配提供一个规模相对较小的目标网络. 为了与 RESQUE 算法的查询结果进行比较,本文参考 RESQUE 算法,当目标网络中的节点数目小于查询网络中节点数目的 2 倍时,停止迭代,进入下一个阶段——查询最优匹配子图[11].



(a) The classify process of nodes in target network (b) The deletion of nodes in target network

Fig. 2 The iterative reduction of the target network.

图 2 目标网络迭代递减

1.3 查询最优匹配子图

当完成目标网络的递减过程后,获得了一个有着与查询网络较高相似性的规模较小的子目标网络 $\boldsymbol{G}^{\infty}=(V^{\infty},E^{\infty},W^{\infty})$(节点规模不超过 $2N_Q$ 节点). 我们的目标是在这个子目标网络 $\boldsymbol{G}^{\infty}$ 中匹配出与查询网络 $\boldsymbol{G}_Q$ 最相似的子网. 为了识别最优匹配子网,采用比较常用的最大权值偶图匹配算法——KM(Kuhn-Munkres)算法. 在这个算法中,偶图 $\boldsymbol{G}_B$ 中的节点集合是 $V_Q \cup V^{\infty}$,边的集合是由一组由权值的边组成,每条边中的 2 个节点分别是 $q_i \in V_Q, v_j \in V^{\infty}$,权值是 $\omega(i,j)=\boldsymbol{C}^{\infty}[i,j]$. 然后将权值转化成可行顶标,假设 M 为匹配集合,寻找 M 中的可增广轨 P,P 包含 $2k+1$ 条边,其中 k 属于 M,$k+1$ 不属于 M. 更新 M 为 $M \cup P$,直至找不到可增广轨,若 V_Q 中所有的节点都在 M 中有匹配节点,则获得一组完备匹配,若无法求出完备匹配,则修改可行顶标,重复执行上述过程,直至找到完备匹配为止,此时的完备匹配就是最优匹配,即我们最终查询得到的最优匹配子网.

1.4 评价方法

评价算法的识别准确率方面,我们分别使用 2 种结构指标和 2 种生物指标. 结构指标采用实际匹配数($hits$)和有效匹配数(meanning hits, MH),生物指标采用功能内聚性(functional coherence, FC)和特异性(specificity, SP). 用运行时间和迭代次数来评价算法的计算性能.

2 实验及结果分析

2.1 蛋白质复合物作为查询网络

分别选取 H. sapiens,S. cerevisiae,D. melanogaster,M. muscle 四个物种的蛋白质相互作用(PPI)网络作为目标网络,数据来自于 ISObase①. 其中,H. sapiens 的 PPI 网络中包含 10 403 个蛋白质和 55 168 条 PPI;S. cerevisiae 的 PPI 网络中包含 5 524 个蛋白质和 82 932 条 PPI;D. melanogaster 的 PPI 网络中包含7 396 个蛋白质和 25 054 条 PPI;M. muscle 的 PPI 网络中包含 623 个蛋白质和 592 条 PPI. 查询网络则选用 4 个物种中蛋白质个数大于 4 的蛋白质复合物. 4 个物种的查询网络个数分别为 693,27,263,176.

① http://groups.csail.mit.edu/cb/mna/isobase

将改进算法分别与 RESUCE 算法及 IsoRank 算法比较,分别测试所有匹配子网的拓扑结构有效性(*MH*)、生物意义(*FC* 及 *SP*)以及一次子网匹配的运行时间和迭代次数. 3 个算法中,均选择输出一个最大权值匹配图,实验结果如图 3 所示. 图 3 中的 *MH* 是指匹配出来的子图需要满足其包含的节点个数至少是查询子图包含节点数的 1/2,至多为其节点数的 3/2. 因为 RESQUEH 算法的目的主要是为了识别具有生物意义的子图,匹配子图中包含节点过多或过少都容易使最终的统计结果产生偏差,

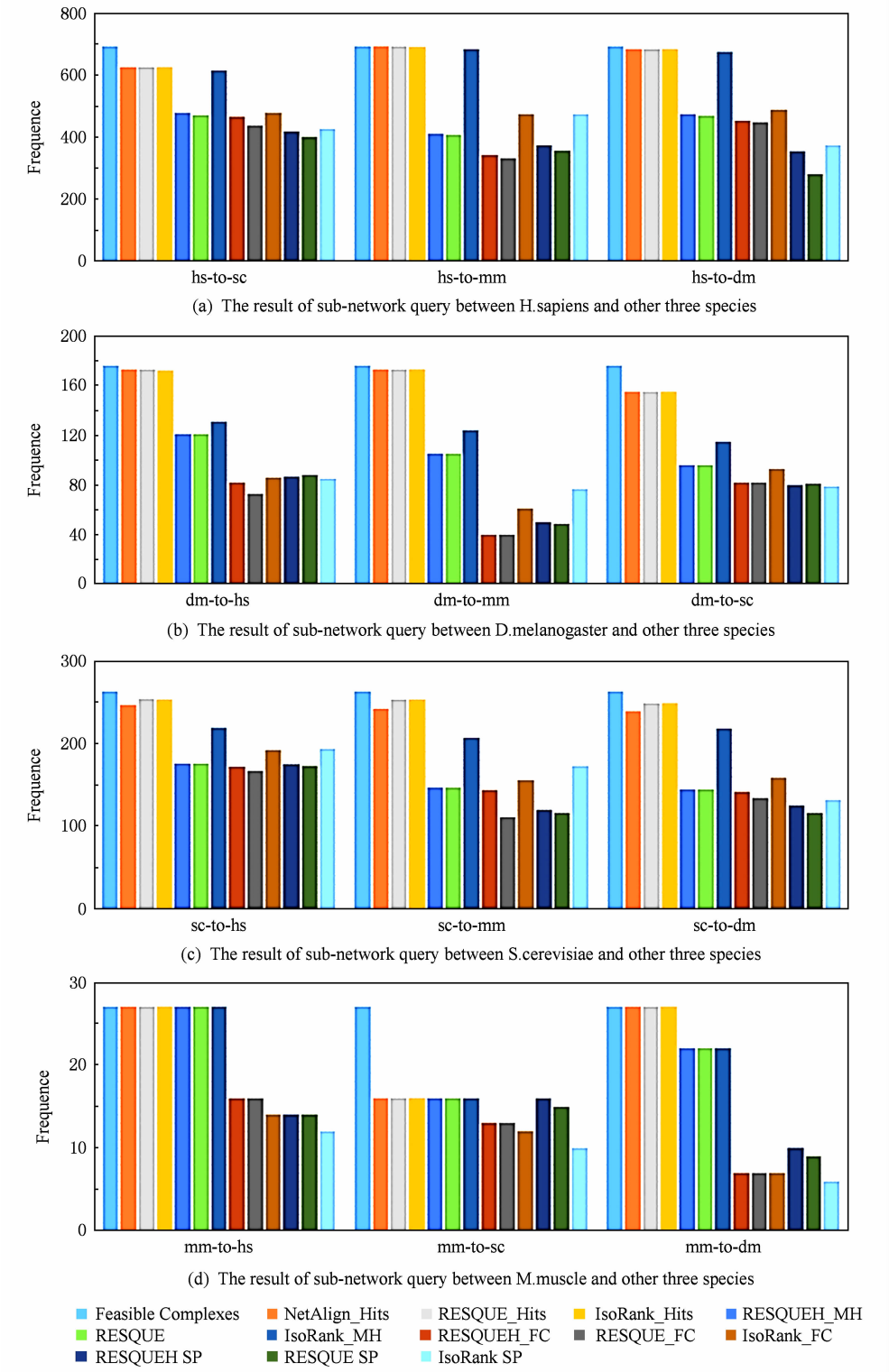


Fig. 3 The result of sub-network query among four species with different sub-network query algorithms.

图 3 4 个物种之间的不同算法的子网查询结果

所以对 *MH* 的子图进行生物意义的探究,比直接对所有的 hit 子图要更有价值. feasible hits 指的是经过数据预处理之后的所有蛋白质复合物的个数,实验中我们选取了大于等于 4 个蛋白质的蛋白质复合物作为用来测试的查询网络. 根据图 3 可知,我们的算法在识别有效子图方面和 RESQUE 算法差不多,分别是63.8%和63.4%. 为了分析算法的识别准确性,分别测试了不同算法得到的有效子图的 *FC* 和 *SP*. *FC* 主要检测那些有重要生物功能关联的子图数目,利用子图在基因本体(GO)上的富集度的 p-value 值. 而 *SP* 则是用来检测子图与一个已知的蛋白质复合物重合比例,如表 1 所示:

Table 1 Comparison of the Specificity and Functional Coherence of Protein Complexes

表 1 蛋白质复合体功能内聚性与特异性比较

Network	Complex	*FC*((*FC*/*MH*)/%)			*SP*((*SP*/*MH*)/%)		
		RESQUEH	RESQUE	IsoRank	RESQUEH	RESQUE	IsoRank
H. sapiens	M. muscle	343(83.6)	333(81.4)	475(69.4)	375(91.2)	358(87.5)	475(69.4)
	S. cerevisiae	467(97.7)	439(93.0)	479(77.9)	418(87.4)	402(85.1)	427(69.4)
	D. melanogaster	454(95.6)	448(95.3)	490(72.5)	356(74.9)	281(59.8)	375(55.5)
M. muscle	H. sapiens	16(59.3)	16(59.3)	14(51.9)	14(51.9)	14(51.9)	12(44.4)
	S. cerevisiae	13(81.3)	13(81.3)	12(75.0)	16(100)	15(93.8)	10(62.5)
	D. melanogaster	7(31.8)	7(31.8)	7(31.8)	10(45.5)	9(40.9)	6(27.3)
S. cerevisiae	M. muscle	144(98.0)	111(75.5)	156(75.4)	120(81.6)	116(78.9)	173(83.6)
	H. sapiens	172(97.7)	167(94.9)	192(87.7)	175(99.4)	173(98.3)	194(88.5)
	D. melanogaster	134(92.4)	125(86.2)	159(72.9)	125(86.2)	116(80.0)	132(60.6)
D. melanogaster	M. muscle	40(38.1)	40(38.1)	61(49.2)	50(47.6)	49(46.7)	77(62.1)
	S. cerevisiae	82(85.4)	82(85.4)	93(80.9)	87(71.9)	88(72.7)	79(68.7)
	H. sapiens	82(67.8)	73(60.3)	86(65.6)	80(83.3)	81(84.4)	85(64.9)

结合图 3 和表 1 可知,在识别效果上,RESQUE 算法和 RESQUEH 算法的 *FC*/*MH* 和 *SP*/*MH* 的值都大于 80%,IsoRank 的值大都小于 80%;单从识别的 *FC* 和 *SP* 的值可以看出,IsoRank 算法的效果是最好的,其次是我们的改进算法——RESQUEH,最后是 RESQUE 算法. 那些没有到达 *SP* 指标的有效匹配子图有可能成为以前未被发现的新蛋白质复合物. 我们的算法发现的与已知复合物不重合的 *MH* 有 393 个,RESQUE 算法的有 504 个,IsoRank 的则更多. 通过上述结果可以看出,利用 IsoRank 算法中邻居节点相似性的思想对 RESQUE 算法进行的改进,确实使得算法在匹配子网的精确性上有所提升. 另一方面,我们发现在 *MH* 的数量上,RESQUE 算法与我们的算法相似的. 原因可能是算法并未改变 RESQUE 算法在迭代终止时的条件. 相较于 IsoRank 算法,RESQUE 算法中的迭代终止条件更加满足目标网络迭代递减的思想,因此从有效子网的识别数目上来说,RESQUE 算法和 RESQUEH 算法的子网匹配效果是相似的.

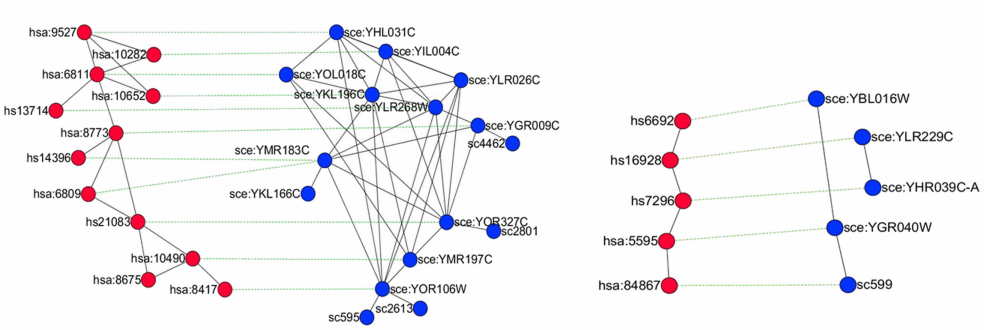
2.2 已知通路(pathway)作为查询网络

选取 H. sapeins 和 S. cerevisiae 二个物种,按 KEGG 中的 KO 获取 S. cerevisiae 的蛋白质子集作为查询网络(29 个)来测试 2 个物种在 KEGG 数据库上的子网匹配效果. 图 4(a)~4(d)分别是 S. cerevisiae 四个重要的生物过程(pathway)在 H. sapeins 目标网络中的匹配结果.

3 讨 论

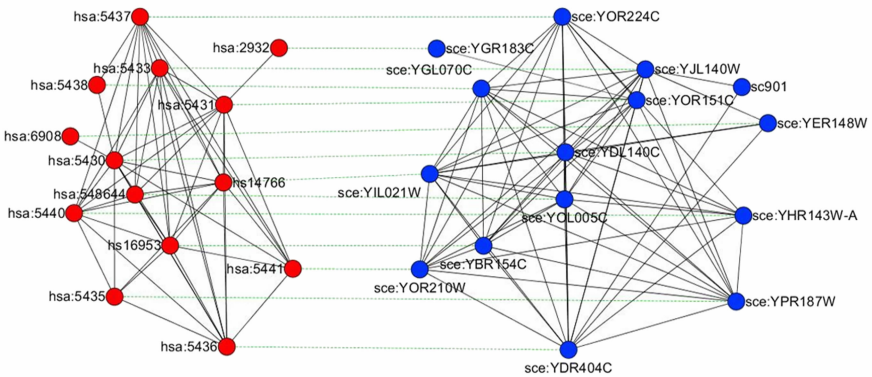
3.1 全局相似性得分中的选取

在计算全局相似性得分中,主要包括了匹配蛋白质对的自身相似性、匹配蛋白质对周围 interaction 结构的相似性以及匹配蛋白质对周围邻居蛋白质之间的相似性 3 个重要影响因素. 其中,后 2 个相似性的计算得分均与蛋白质对在其相应网络中的拓扑结构有关,它们分别体现了拓扑结构中的不同方面对全局相似性得分的影响. 为确定 α 的最优值,我们分别选取 0,0.1,0.3,0.5,0.7,0.9,1 值,对 RESQUEH 算法进行了测试. 测试选择的 2 个物种分别是 D. melanogaster 和 H. sapiens. 以 D. melanogaster 的 173 个蛋白质复合物为查询网络,H. sapiens 为目标网络. 实验结果如图 5 所示.

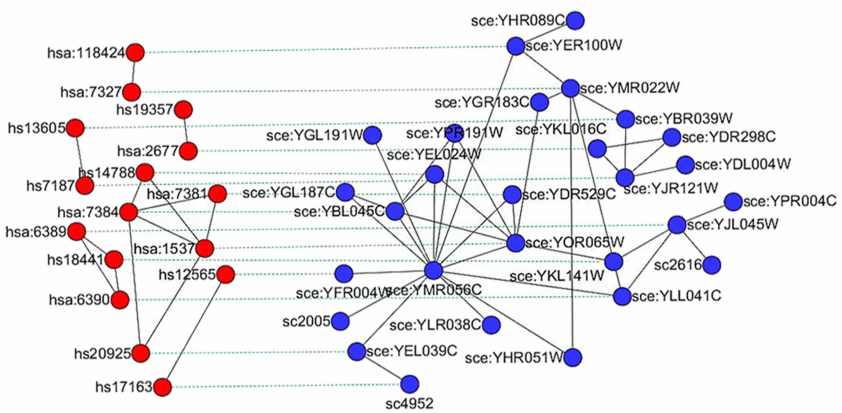


(a) SNARE interactions in vesicular transport

(b) MAPK signaling pathway



(c) RNA polymerase



(d) Oxidative phosphorylation

Fig. 4 Instances of sub-network query.

图 4 子网查询实例

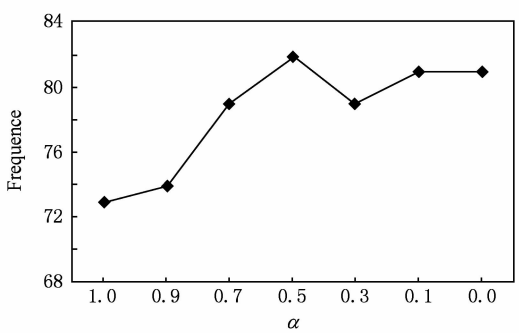


Fig. 5 Result of sub-network dm-to-hs query with different.

图 5 不同值得到的子网 dm-to-hs 查询结果

当 $\alpha=0$ 时,算法中仅考虑蛋白质对的邻居节点序列相似性得分和蛋白质对本身的序列相似性得分;当 $\alpha=1$ 时,算法基本上就是 RESQUE 算法.根据图 5 我们可以看出,当 $\alpha=0.5$ 时 FC 的数目最多,这说明此时算法的识别准确率最高.因此,当蛋白质对的邻居节点序列相似性得分和蛋白质对的关联相互作用的结构相似性得分平衡的时候,算法的识别准确率最高.

3.2 计算复杂性分析

RESQUE 算法的计算复杂性是 $O(mN+zN+N^2\lg(N_Q))$,由于我们的算法添加了对于邻居节点相似性的计算,将会导致计算复杂性有所增加,这一部分的复杂性是 $O((NN_Q)^2)$;同时,通过改进目标网络迭代递减时删除节点的规则,将目标网络中的蛋白质按照查询网络中的蛋白质分类,并删除各个分类中的最远距离的蛋白质,这一方法提高了迭代递减的速度,降低了计算复杂性.因此,最终改进算法的计算复杂性是 $O(mN/N_Q+zN/N_Q+N^2\lg N_Q+(NN_Q)^2)$,其中 N_Q 是查询网络中包含节点的个数,N 是目标网络中包含节点的个数,m 是目标网络中边的数目,z 是 2 个网络中同源或相似的节点数目(实际上 $z\ll N\times N_Q$).我们的算法不仅可以应用在无权的生物分子网络上,同样可以应用于有权生物分子网络,主要从以下 2 个方面具体说明.

3.2.1 无权分子网络子网查询复杂性分析

我们统计了前面提到的 4 个物种之间的无权蛋白质相互作用网络之间在不同算法下的每一次子网查询的运行时间,如图 6 和表 2 所示.由表 2 可知,RESQUEH 算法和 RESQUE 算法的 1 次子网查询时间 99%集中在 200 s 以内,92%集中在 0.5 s 以内;而 IsoRank 算法 80%主要集中在 0.55 s 以内,只有 10%则是分布在 0.5 s 以内.由图 6 可知,RESQUEH 算法减少了 RESQUE 算法中那些非常耗时的子网查询,但是 RESQUEH 算法中 0.1~0.5 s 的实验数目较 RESQUE 算法有所增加.

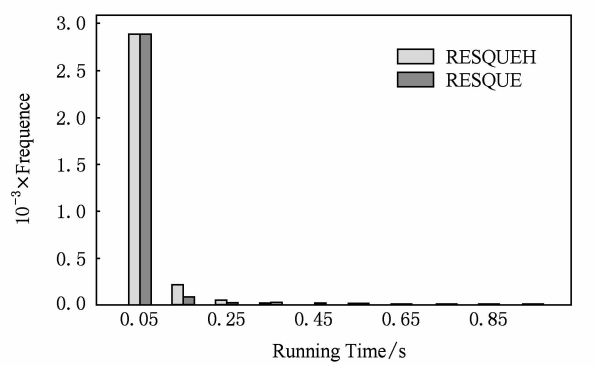


Fig. 6 Comparison between running time of RESQUEH and RESQUE algorithms.

图 6 2 种算法运行 3 477 个 cliques 运行时间比较

Table 2 Running Time of Every Sub-network Query Process in Four Species

表 2 4 个物种中所有子网匹配的一次运行时间

Running Time/s	RESQUEH	RESQUE	IsoRank
>200	1	4	4
50~200	6	8	201
5~50	21	31	1 341
1~5	34	51	915
0.5~1	41	34	532
0.1~0.5	337	140	72
<0.1	2 889	3 065	266

经分析,造成这个现象的主要原因有以下 2 个方面:

1) 我们对 RESQUEH 和 RESQUE 算法每次运行时迭代的次数进行了统计,结果发现 80%左右的一次子网查询的迭代次数为 1~2,这说明,我们尝试通过降低迭代次数来降低计算复杂度的方法在这种情况下基本失效,如图 7 所示:

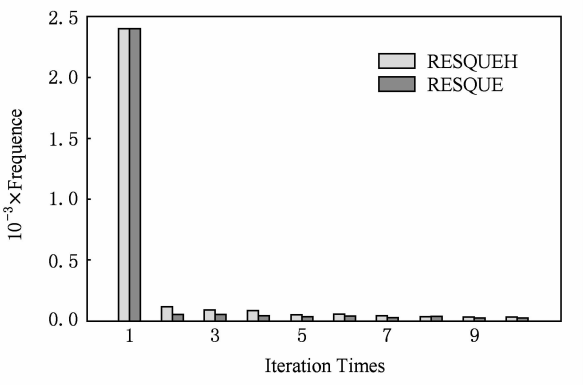


Fig. 7 Comparison between iteration times of RESQUEH and RESQUE algorithms.

图 7 2 种算法的每次子网查询的迭代次数比较

2) 为了提高 RESQUE 算法的识别精度而添加了计算邻居节点相似性的步骤,其时间复杂性是 $O((N\times N_Q)^2)$,因此在迭代次数较少(1~5)的情况下,该部分的运行时间占主要部分.我们还分析了 RESQUE 算法和 IsoRank 算法中运行时间超过 200 s 而改进后的算法中运行时间都有明显变化的 5 个子网查询的实例,其结果的详细信息如表 3 所示.

由表 3 可以看出运行时间长的子网查询实例分

以下 2 种:①查询网络结构简单(即包含节点少)但是迭代次数很多;②查询网络本身就很复杂,但是迭代次数不一定多.像情况 1,主要是因为查询网络和目标网络的节点本身的序列相似性信息比较完整,查询网络中的蛋白质初次迭代的匹配目标网络中的蛋白质就多,需要迭代的次数也多,计算邻居节点相似性在总运行时间的比例就会变小,RESQUEH 算法就比 RESQUE 算法更为高效.

Table 3 Instances of RESQUEH, RESQUE and IsoRank with Running Time Longer than 200 s (NA: No Result)

表 3 RESQUEH, RESQUE 和 IsoRank 中运行时间大于 200 s 的实例结果(NA:没计算出来)

Algorithm	Running Time/s(Iter)				
	55[6/11/DM_HS]①	96[9/32/DM_HS]①	175[4/6/DM_HS]①	2759[1153/7753/SC_MM]②	3022[1153/7753/SC_DM]③
RESQUEH	67.921 2(24)	1 275.3(26)	92.746(25)	2 571(1)	2 516(1)
RESQUE	209.5(681)	2 416.6(1415)	332.5(1439)	NA	6 499.1(1)
IsoRank	NA	NA	214.37(141)	NA	NA

① No of Experiment [Number of Nodes in Query Network/Number of Edges in Query Network/D. melanogaster_H. sapiens]

② No of Experiment [Number of Nodes in Query Network/Number of Edges in Query Network/S. cerevisiae_M. muscle]

③ No of Experiment [Number of Nodes in Query Network/Number of Edges in Query Network/S. cerevisiae_ D. melanogaster]

3.2.2 有权分子网络子网查询复杂性分析

我们用同样方法测试了有权生物分子网络子网查询的计算复杂性.我们利用文献[13]中的蛋白质网络构建方法构建了人类的脑和肾脏 2 个组织的蛋白质相互作用网络,并获取其中肾脏网络中的 30 个 cluster 作查询网络.由表 4 可以看出,我们的算法在有权生物分子网络中子网查询的计算复杂度也很低.

Table 4 The Result of Running Time and Iteration of Sub-network Query Between Brain and Kidney of Human

表 4 人脑和肾脏组织子网查询运行时间和迭代次数结果

Algorithm	Running Time/s(Iter)			
	1[29/43]①	12[7/9]①	17[20/24]①	22[47/76]①
RESQUEH	0.075 3(20)	0.014 3(1)	0.025 5(5)	0.043 5(3)
RESQUE	0.207 4(62)	0.002 7(1)	0.043 9(14)	0.105 5(25)

① No of Experiment [Number of Nodes in Query Network/ Number of Edges in Query Network]

4 结论与展望

本文主要针对子网查询算法中的 RESQUE 算法在子网识别准确程度上的缺陷进行了改进,提高了 RESQUE 算法在复杂网络之间查询的识别准确率.具体来说,改进算法主要是参考 IsoRank 算法中计算蛋白质拓扑结构方面相似性的优点(不仅考虑了 2 个蛋白质周围 interaction 的权值对两者相似度的影响,还考虑 2 个蛋白质周围蛋白质相似性对其的影响),通过添加蛋白质对的邻居节点相似性,提高了 RESQUE 算法的准确度;同时,还改进目标网络迭代递减时的方法,大大缩减了迭代次数.同时,算法中还存在着一些不足,这主要是由于要计算蛋白质对的邻居节点相似性,导致计算量增加,当查询网络在匹配时本身需要较少迭代次数的时候,会导致算法运行时间的增加.根据上述的实验结果和讨论分析,我们的算法更适合同物种之间的子网查询.主要原因是这种蛋白质之间的同源比对数据较完全,因此蛋白质对的邻居节点相似性的效果会更为明显,同时查询网络匹配时的选择性较多,迭代次数也会较多,蛋白质对的邻居节点相似性计算消耗的时间对整体运行时间造成的影响就会降低.

在未来的工作中,我们将尝试进一步将算法改进成可以用于多网络比对或查询的算法,并且考虑将其应用于同一物种的不同组织或疾病的蛋白质相互作用网络,利用其发现不同组织或疾病之间的生物功能联系以及组织特异性的功能结构,发现疾病的关键诱因或各个组织中的重要生物结构之间的联系.

参 考 文 献

[1] Tung A K H. Large scale cohesive subgraphs discovery for social network visual analysis [J]. Proceedings of the VLDB Endowment, 2012, 6(2): 85-96

[2] Andorf C M, Honavar V, Sen T Z. Predicting the binding patterns of hub proteins: A study using yeast protein interaction networks [J]. PLoS One, 2013, 8(2): Article No e56833

[3] Krian M, Nagarajaram H A. Global versus local hubs in human protein-protein interaction network [J]. Proteome Research, 2013, 12(12): 5436-5446

[4] Glaab E, Baudot A, Krasnogor N, et al. EnrichNet: Network-based gene set enrichment analysis [J]. Bioinformatics, 2012, 28(18): i451-i457

[5] Huan Jun, Wang Wei, Prins J. Efficient mining of frequent subgraphs in the presence of isomorphism[C] //Proc of the 3rd IEEE Int Conf on Data Mining. Los Alamitos, CA: IEEE Computer Society, 2003: 549-552

[6] Kelley B P, Yuan Bingbing, Lewitter F, et al. PathBLAST: A tool for alignment of protein interaction networks [J]. Nucleic Acids Research, 2004, 32(Web Server issue): W83-W88

[7] Shlomi T, Segal D, Ruppin E, et al. QPath: A method for querying pathways in a protein-protein interaction network [J]. BMC Bioinformatics, 2006(7): Article No 199

[8] Dost B, Shlomi T, Gupta N, et al. QNet: A tool for querying protein interaction networks [J]. Computional Biology, 2008, 15(7): 913-925

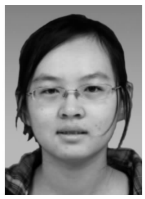
[9] Yang Q, Sze S H. Path matching and graph matching in biological networks [J]. Computional Biology, 2007, 14(1): 56-67

[10] Singh R, Xu Jinbo, Berger B. Global alignment of multiple protein interaction networks with application to functional orthology detection [J]. Proceedings of the National Academy of Sciences of the United States of America, 2008, 105(35): 12763-12768

[11] Sahraeian S M E, Yoon B J. RESQUE: Network reduction using semi-Markov random walk scores for efficient querying of biological networks [J]. Bioinformatics, 2012, 28(16): 2129-2136

[12] Pearson K. The problem of the random walk [J]. Nature, 1905, 294(72): 318-342

[13] Zhang Xiaochi, Gong Xiujun. Interaction network construction and enrichment analysis for human kidney tissue specific proteins [J]. Journal of Guangxi University: Natural Science Edition, 2013, 38(5): 1008-1116 (in Chinese)
(张小驰, 宫秀军. 人类肾脏组织特异性蛋白网络构建及分析[J]. 广西大学学报: 自然科学版, 2013, 38(5): 1008-1116)



Zhang Xiaochi, born in 1990. Master candidate in Tianjin University. Her main research interests include bioinformatics and data mining.

Yu hua, born in 1973. PhD candidate and engineer in Tianjin University. Her research interests include data mining and semantic Web.

Gong Xiujun, born in 1972. PhD and associate professor in Tianjin University. His main research interests include bioinformatics and data mining.