ISSN 1000-9825, CODEN RUXUEW
*Journal of Software*, Vol.21, No.9, September 2010, pp.2089−2106
doi: 10.3724/SP.J.1001.2010.03860
© by *Institute of Software*, *the Chinese Academy of Sciences*. All rights reserved.
E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

# 生物网络比对的模型与算法*

郭杏莉[1,2+], 高 琳[1], 陈 新[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)

# Models and Algorithms for Alignment of Biological Networks

GUO Xing-Li[1,2+], GAO Lin[1], CHEN Xin[1]

[1](School of Computer Science and Technology, Xidian University, Xi'an 710071, China)
[2](School of Software Engineering, Xidian University, Xi'an 710071, China)
+ Corresponding author: E-mail: guoxingli.xd@gmail.com

**Guo XL, Gao L, Chen X. Models and algorithms for alignment of biological networks. *Journal of Software*, 2010,21(9):2089−2106.** http://www.jos.org.cn/1000-9825/3860.htm

**Abstract**: Biological network alignment is an important approach in the study of organisms' structure, function and evolution. In this review, the recent studies in the field of biological network alignment are surveyed. First, a definition of biological network alignment is defined formally. Secondly, the methods for alignment are reviewed and described into three categories according to their mathematical properties. Some models and algorithms in every category are analyzed comprehensively and comparably. Next, tools for alignment are listed and analyzed. In the end, some applications and key problems in the field are highlighted, as well as the future progress of biological network alignment.
**Key words**: biological network; alignment; graph matching; constrained optimization; divide and conquer

**摘 要**: 生物网络比对是生物体结构、功能和进化分析的重要研究手段.首先给出了生物网络比对问题的形式化定义;其次重点分析了现有的比对模型和算法,按照比对方法的数学特性对其进行了分类,并对典型算法结合应用进行了深入探讨,对3类比对方法的特点进行了总结与比较;再次,分析归纳了生物网络比对软件,阐述了生物网络比对研究的意义和应用;最后指出了生物网络比对研究中的关键问题及生物网络比对未来的研究方向.
**关键词**: 生物网络;比对;图的匹配;约束优化;分治策略
**中图法分类号**: TP391 **文献标识码**: A

随着实验测定技术(如酵母双杂交[1−3]、质谱分析[4]、染色体免疫共沉淀[5,6]、串联亲和纯化[7,8]、蛋白质芯片[9−11]、噬菌体显示技术[12,13])和文献挖掘技术[14]的发展,产生了大量的分子相互作用数据,也称为生物网络数据,例如蛋白质相互作用网络、代谢网络、基因表达网络、基因调控网络和信号传导网络,并且这些数据呈指

* Supported by the National Natural Science Foundation of China under Grant No.60933009 (国家自然科学基金); the Specialized Research Fund for the Doctoral Program of Higher Education of China under Grant No.200807010013 (高等学校博士学科点专项科研基金)

Received 2009-06-18; Revised 2009-11-26; Accepted 2010-03-30

© 中国科学院软件研究所 http://www.jos.org.cn

数级增长态势[15].目前,已开展了大量针对生物网络数据的研究工作,如采用复杂网络理论对生物网络的度分布、聚集系数、小世界特性的研究;采用子图搜索算法和子图比较算法挖掘生物网络模体;采用聚类方法挖掘生物网络模块等[16−23].其中,一类重要的研究工作是生物网络数据的比较分析,即生物网络比对.通过比对来认识和研究生物体[16,21,23],发现它们结构和功能的相关性,基于生物网络数据的比对结果研究生物的进化和演变,通过不同网络之间的比较进行知识迁移,从而借助已知生物研究未知生物.例如:通过生物网络比对进行蛋白质功能的预测[24,25]、蛋白质相互作用的预测[26]和不同物种之间蛋白质同源关系的预测[27];通过不同生物网络之间的局部比对,发掘保守的子网络结构,预测蛋白质复合物[28,29]和保守功能模块[25,30−34].

2000 年,Ogata 等人[35]通过生物网络比对研究了代谢网络中的酶及其编码基因在基因组中的位置相关性,这一研究工作被认为是生物网络比对研究的起点.此后,有很多学者从事这方面的研究,其中值得关注的是 Sharan 的研究小组,他们着重于蛋白质相互作用网络的比对,通过比对预测保守通路[36]、保守功能模块[25]和蛋白质复合物[28,29],在结构分析的基础上进一步预测蛋白质功能[25]和蛋白质相互作用[26].近期,该小组有关生物网络比对的研究工作主要包括两个方面:一是高效的多个生物网络的查询比对算法[37,38];二是比对方法与具体应用的结合,例如将比对应用到人类某些疾病或病毒的研究中.2006 年,Purdue 大学的 Koyutürk 等人[31]较早研究了两个网络的局部比对问题,提出了 MaWISH 比对方法.其核心在于构建两个蛋白质相互作用网络的比对图,并借助蛋白质相互作用网络的进化模型构造比对的相似度函数.同年,中国科技大学的 Liang Z 等人[26]提出的比对方法借助比对图中的团完成两个蛋白质相互作用网络的比对,挖掘保守的功能模块,随后又开展了多个生物网络比对的研究工作.同年,Stanford 大学的 Flannick 等人提出了 Græmlin 比对方法[32,33],可以进行多种类型生物网络的全局和局部比对.以上关于生物网络比对的研究主要是通过构建比对图将两个或者多个生物网络之间的比对问题转化为一个图中的问题,借助有关图论算法予以求解.生物网络比对问题也是一类组合优化问题,可以建立合理的优化模型借助优化方法对其进行求解.2006 年,德国 Koeln 大学的 Berg 等人[39]采用统计模型模拟生物网络中顶点和边的动力学演化过程,借助统计方法对比对问题进行求解;同一研究小组的 Kolar 等人将该方法用于疱疹病毒的研究中[24].2007 年,中国科学院章祥荪的研究小组将比对问题归约为优化问题,借助整数二次规划方法进行求解[40].2009 年,Klau[41]同样借助整数线性规划方法求解比对问题.2008 年,麻省理工学院的 Singh 等人[27,42]将比对问题归约为矩阵的特征值问题,利用幂法得到矩阵的特征值,求解比对问题.借助生物网络的模块化结构,将生物网络划分为模块,通过模块之间的比对完成整个网络的比对,在某些应用下不失为一个高效简捷的求解方法.2007 年,California 大学的 Narayanan 等人[34]采用这种模块化比对的方法完成了两个蛋白质相互作用网络的局部比对.在目标网络中查询给定结构的子网络,是生物网络比对中一个重要的研究子问题.2005 年,Pinter 等人[43]较早地研究了代谢网络中查询模式的比对问题,他们利用子树同胚算法对问题进行求解.针对查询模式的比对问题,研究者们开展了大量的研究工作,提出了很多有效的解决方法[44−50].

我们分析了国内外关于生物网络比对问题的研究成果,许多学者开展了大量的研究工作,针对生物网络比对问题中的各种情况提出了很多的算法.但是,生物网络比对是一个新兴的研究领域,还处在起步阶段,目前大多数研究工作只是针对某个特定问题或特定应用,算法的时间复杂度较高,多个网络比对算法的运行效率不高.生物网络比对模型和算法研究的目的是开发一个通用的生物网络比对软件,可以高效地进行多个生物网络及多种应用模式的比对,类似于序列比对软件 BLAST.此前,Sharan 等人[15]综述了生物网络比对的发展和应用,分别讨论了生物网络比对的 3 种应用模式,并将生物网络比对的发展与序列比对的发展相比较,指出了生物网络比对发展的方向,预言了生物网络比对在生物信息学研究中的重要地位.章祥荪的研究小组[51]对生物网络比对中的查询模式应用进行了分析讨论并总结了相关的比对工具,指出了这种应用模式在系统生物学中的重要性.本文从比对方法的模型和算法两方面对其进行综述.重点分析了三类比对方法的数学特性及其应用特点,并结合应用进行了具体的分析论述;从方法学上分析探讨了比对研究的成果及未来发展。

## 1 生物网络比对问题

生物网络可以抽象为图,网络中的节点对应图的顶点,网络中节点之间的相互作用对应图的边,生物网络的

© 中国科学院软件研究所 http://www.jos.org.cn

比对可以看作图的比较.图的顶点和边可以带有属性,边可以是有向边或无向边,不同类型的生物网络对应的图也是不同的.蛋白质相互作用网络 PIN(protein interaction network)可以抽象为顶点带标签的无向图,顶点上的标签用来标注不同的蛋白质,边上可以带权重用以描述 PIN 的更多特征.例如:可以用边上的权重说明相互作用存在的可信度;代谢网络(metabolic networks)可以抽象为顶点带标签的有向图.

### 1.1 生物网络比对问题的定义

生物网络就是一个图,生物网络之间的比对就是图的比对,下面给出其形式化定义.

**定义 1(生物网络比对)**. 给定两个生物网络,分别用图 $G(U,E)$和图 $H(V,F)$表示,$U$ 和 $E$ 是图 $G$ 的顶点集和边集,$V$ 和 $F$ 是图 $H$ 的顶点集和边集.$R$ 定义为 $U$ 到 $V$ 的映射,即 $R\subseteq U\times V$,图 $G$ 和图 $H$ 的比对对应 $U$ 到 $V$ 的映射 $R^*$,满足下面的要求:

$$sim(G,H)=\arg\max\sum_{\langle a,b\rangle\in R^*}sim(a,b) \tag{1}$$

使得图 $G$ 和图 $H$ 之间的相似性 $sim(G,H)$最大的映射为他们之间的比对结果.其中,$sim(a,b)$是图 $G$ 中的顶点 $a$ 与图 $H$ 中的顶点 $b$ 之间的相似性.

简言之,生物网络比对就是要找到生物网络顶点之间的映射关系,使得生物网络之间的相似性得分最高.图 1 给出了两个生物网络 $G$ 和 $H$ 的局部映射关系,其顶点之间的相似性得分见图 1 左部的表格.基于表中给出的相似性得分,我们得到比对的映射关系如图 1 右部所示,虚线所关联的两个点在图 $G$ 和图 $H$ 的比对中被映射在一起.可以看出:$G$ 中的 $a_1$, $a_1'$ 和 $H$ 中的 $a_2$ 是多对一的映射;$f_1$ 和 $g_1$ 在 $H$ 中没有映射对象,就认为他们映射到一个 *gap*,在左部的表中用$\Delta$来表示一个 *gap*.同理,$H$ 中的 $f_2$ 和 $g_2$ 也映射到一个 *gap*,由此得到图 $G$ 和图 $H$ 的比对对应映射 $R^*=\{\langle a_1,a_2\rangle,\langle a_1',a_2\rangle,\langle b_1,b_2\rangle,\langle c_1,c_2\rangle,\langle d_1,d_2\rangle,\langle e_1,e_2\rangle\}$ ,在映射 $R^*$下 $G$ 和 $H$ 的相似性得分最高.

| | $a_2$ | $b_2$ | $c_2$ | $d_2$ | $e_2$ | $f_2$ | $g_2$ | $\Delta$ |
|---|---|---|---|---|---|---|---|---|
| $a_1$ | 1 | 0 | 0 | 0 | 0 | −1 | −1 | −1 |
| $a_1'$ | 1 | 0 | 0 | 0 | 0 | −1 | −1 | −1 |
| $b_1$ | 0 | 1 | 0 | 0 | 0 | −1 | −1 | −1 |
| $c_1$ | 0 | 0 | 1 | 0 | 0 | −1 | −1 | −1 |
| $d_1$ | 0 | 0 | 0 | 1 | 0 | −1 | −1 | −1 |
| $e_1$ | 0 | 0 | 0 | 0 | 1 | −1 | −1 | −1 |
| $f_1$ | −1 | −1 | −1 | −1 | −1 | −1 | −1 | −1 |
| $g_1$ | −1 | −1 | −1 | −1 | −1 | −1 | −1 | −1 |
| $\Delta$ | −1 | −1 | −1 | −1 | −1 | −1 | −1 | −1 |

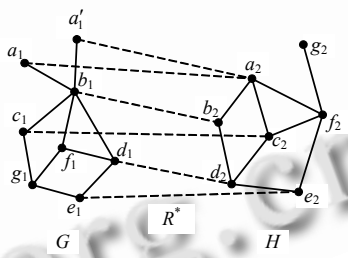



Fig.1 Mapping of biological network alignment

图 1 生物网络比对的映射关系

由公式(1)可以看出,相似性计算是生物网络比对的基础,它提供了生物网络匹配程度的精确的、量化的度量方式.两个网络的相似性得分通过求解其相似度函数得到.

**定义 2(相似度函数)**. 相似度函数是两个生物网络相似性得分的计算公式,其形式化定义如下:

$$\underset{\theta}{sim}(G,H)=\alpha\sum_{\substack{u\in U\\ \theta(u)\in V\cup\varnothing}}f(u,\theta(u))+\beta\sum_{\substack{e\in E\\ \theta(e)\in F\cup\varnothing}}g(e,\theta(e))+\gamma h(G,H) \tag{2}$$

其中:$\theta$是 $G$ 和 $H$ 之间的一个映射;$f$ 是定义在匹配顶点对上的相似度函数;$g$ 是定义在匹配边对上的相似度函数;$h$ 是 $G$ 和 $H$ 之间的进化相似度函数;$\varnothing$是空集合;$\alpha$,$\beta$和$\gamma$是用户自定义的参数,用来调节 3 类相似度函数所占的比重.

目前,多数生物网络比对的相似性计算只涉及相似度函数的前两项;进化相似性还是一个有待深入研究的内容,它的计算需要和生物网络的生物学规律、功能和进化关系等特性结合起来.在 PIN 比对中,顶点的相似性用蛋白质的氨基酸序列的比对结果来衡量,边的相似性通过网络拓扑结构的相似性来衡量.

© 中国科学院软件研究所 http://www.jos.org.cn

### 1.2 生物网络比对问题的特点

生物网络比对问题和图的匹配问题密切相关,在此,我们以 PIN 为例来说明它们之间的关系.PIN 可以看作顶点带标签的无向图,不同生物的 PIN 规模不同,目前的相互作用数据尚不完善,因此,PIN 之间的比对对应图的非精确匹配.生物网络比对问题又不同于一般图的匹配问题,原因在于生物网络数据的特殊性:一方面是生物网络数据的不完整性和噪声,大量的相互作用通过目前的实验手段还未被检测到,检测到的数据存在假阳性(实际上不存在,但实验结果呈阳性)和假阴性(实际上存在,但实验结果呈阴性)数据.研究发现,相互作用数据中的假阳性和假阴性数据的比例分别高达 70%和 90%[52].因此,我们在进行生物网络比对的时候要充分考虑到数据的这些特性,建立合理准确的图模型;另外一方面,生物网络的规模都比较大,以 DIP(database of interacting proteins)数据库中酵母的蛋白相互作用网络来说明.酵母的 PIN 中共有 4 943 个蛋白质,18 440 个相互作用,因此,一般的图的匹配算法很难直接应用到生物网络的比对中,生物网络比对需要利用生物网络数据的生物学特点和拓扑结构的特征来建立合适的图模型,针对具体应用借助有关图的匹配算法或其他的图论算法对问题进行求解.

## 2 生物网络比对的模型与算法

比对模型和算法是生物网络比对方法的两个核心:比对模型是对生物网络比对问题的抽象和数学建模,一般是基于生物网络的图模型和具体问题的特点来进行比对模型的构建;比对算法指的是比对模型上的可计算步骤,用来实现比对问题求解.本节从模型和算法两个方面对生物网络比对方法进行分类探讨.

### 2.1 生物网络比对方法的分类

我们基于比对模型和算法对其进行分类,见表 1.表中给出了每一类方法的模型、算法、特点以及典型方法.第 1 类是基于图模型的启发式搜索方法,该方法基于多个生物网络建立相应的比对图,它可以是积图(product graph)或者其他形式.比对图中的顶点对应一组分别来自不同生物的相容元素,相似度信息作为属性附加在比对图上.基于比对图模型设计启发式的搜索算法,完成比对问题求解.第 2 类是基于目标函数的约束优化方法,该方法将比对问题转化为某个已知求解方法的优化问题,借用已知的算法求解比对问题.第 3 类是基于分治策略的模块化方法,生物网络的规模较大并且具有模块化的结构,基于分治策略的思想将生物网络划分为模块,降低问题求解的难度,通过较小规模的模块比对来完成生物网络比对.

**Table 1** Classification of alignment based on models and algorithms

**表 1** 基于模型和算法的比对方法分类

| Alignment methods | Models & algorithms for methods | Characteristics of methods | Examples |
|---|---|---|---|
| Heuristic search based on alignment graph | Model based on alignment graph, heuristic search algorithm | The problem between two or more graphs is changed into one alignment graph; similarity is the feature of the alignment graph; heuristic search algorithm is designed for alignment graph; alignment mapping is reflected in alignment graph directly | MaWish[31] NetworkBLAST[25] Græmlin[32,33] NetAlign[26] |
| Constrained optimization based on objective function | Model based on another optimization problem, algorithm based on resolved optimization solution | Alignment is translated to an objective function; similarity measures are changed into the constraints; at the end, the change between the solution to the optimization and the alignment mapping is needed | IsoRank[27,42] MNAligner[40] G. W. Klau[41] Bayesian inferring[24,39] |
| Modular alignment based on divide and conquer strategy | Model based on many to many module alignment, graph partition algorithm cooperates with module alignment algorithm | The networks have been divided into modules; network alignment is changed into many to many module alignment; divide and conquer strategy is applied with similarity computation | “Match-and-Split”[34] BiNA[53] DivAfull[54] |

### 2.2 基于图模型的启发式搜索方法

基于图模型的启发式搜索方法将来自不同生物 PIN 的一组同源蛋白质作为比对图的顶点,将生物网络的相似性转化为比对图上的属性,将多个网络的比对问题转化为一个比对图上的问题,例如最大权重子图[31]、最

© 中国科学院软件研究所 http://www.jos.org.cn

大团[26]或者最大公共子图[30]等问题.针对比对图上的问题,设计启发式的搜索算法予以求解.启发式搜索算法的设计一般采用种子生长方式的贪心策略,选定满足某种要求的顶点集合作为种子,由种子初始化目标集合,开始自底向上的搜索.基于局部最优的原则选择一个点加入到目标集合中,类似种子的生长,每一个目标集合对应比对结果的一个子集.这类方法多应用于挖掘不同生物网络中的保守模块及蛋白质复合物预测[25,26,28,29,31]、蛋白质相互作用预测[25]、蛋白质功能预测[25,26]、同源蛋白质预测[27]等.下面分析几个典型比对方法来具体说明这一类比对方法的求解策略及其特点.

2.2.1 MaWISH 比对方法

Koyutürk 等人提出的 MaWISH 比对方法[31]基于 PIN 的复制/变异模型建立比对图,来挖掘两个 PIN 中的保守功能模块.比对图的顶点是一对分别来自两个 PIN 的同源蛋白质,文中以两个蛋白质序列比对的 *BLAST E_value* 值为参考来估计其同源性.比对图上的边代表生物进化中的复制和进化保守性带来的匹配以及变异导致的失配 3 类事件,边上的权重表示这 3 类事件对比对相似性的影响大小.

1. 比对图的顶点

给定 $PIN_1$ 和 $PIN_2$,它们的比对图为 $G(V,E)$,其顶点集合 $V$ 的定义见公式(3).$V$ 中的顶点是一对分别来自两个 PIN 的同源蛋白质,要求两个同源蛋白质的相似度函数 $S$ 的值大于 0,$S$ 的计算见公式(4):

$$V=\{n=\{u,v\}:u\in PIN_1,v\in PIN_2 \text{ and } S(u,v)>0\} \tag{3}$$

$$S(u,v)=P(E(u,v)<\tilde{E}\mid O_{uv})=\frac{|\{u'v'\in O:E(u',v')<\tilde{E}\}|}{|O|} \tag{4}$$

公式(4)以数据库 COG(cluster of orthologous groups of proteins)作参考,用两个蛋白质 $u$ 和 $v$ 序列比对的 *BLAST E_value* 值 $\tilde{E}$ 来估计蛋白质 $u$ 和 $v$ 同源的统计显著性,以此衡量 $u$ 和 $v$ 的相似性.

2. 比对图的边

比对图 $G(V,E)$是一个边赋权图,基于 PIN 网络的复制/变异模型定义匹配事件、失配事件和复制事件,分别见公式(5)~公式(7):

$$\begin{aligned}M=\{&u,u'\in PIN_1,v,v'\in PIN_2:S(u,v)>0,S(u',v')>0,\\&((uu'\in PIN_1\wedge \Delta_{PIN_2}(v,v')\leqslant\bar{\Delta})\vee(vv'\in PIN_2\wedge\Delta_{PIN_1}(u,u')\leqslant\bar{\Delta}))\}\end{aligned} \tag{5}$$

$$\begin{aligned}N=\{&u,u'\in PIN_1,v,v'\in PIN_2:S(u,v)>0,S(u',v')>0,\\&((uu'\in PIN_1\wedge \Delta_{PIN_2}(v,v')>\bar{\Delta})\vee(vv'\in PIN_2\wedge\Delta_{PIN_1}(u,u')>\bar{\Delta}))\}\end{aligned} \tag{6}$$

$$D=\{u,u'\in PIN_1:S(u,u')>0\}\cup\{v,v'\in PIN_2:S(v,v')>0\} \tag{7}$$

以比对图中的两个顶点$(u,v)$和$(u',v')$之间的边及其权重来说明如何用匹配、失配、复制这 3 种进化事件来度量比对的相似性:如果 $u$ 和 $u'$在 $PIN_1$ 中有相互作用,$v$ 和 $v'$在 $PIN_2$ 中也有相互作用,那么用匹配事件$\mu$来奖励,比对图中顶点$(u,v)$和$(u',v')$之间的边上的权重为$\mu$;如果 $u$ 和 $u'$在 $PIN_1$ 中有相互作用,$v$ 和 $v'$在 $PIN_2$ 中没有相互作用,那么用失配事件$\nu$来惩罚,比对图中顶点$(u,v)$和$(u',v')$之间的边上的权重为$-\nu$;如果 $u$ 和 $u'$之间或 $v$ 和 $v'$之间有复制事件,如果是,in-paralog 用$\sigma$来奖励,否则用$\sigma$来惩罚.如果不存在这 3 个事件中的任何一个,就认为这两个顶点之间没有边相连.其中,$\mu$,$\nu$和$\sigma$是基于顶点相似性的参数化函数.另外,也可以考虑 PIN 中蛋白质之间的间接相互作用来定义失配和匹配事件.

图 2 给出了一个比对图构建的例子.图 $G$ 和图 $H$ 是两个 PIN 的局部,其中,颜色相同的顶点代表它们是相似的.这里只关心直接相互作用,图 2 中右边的图是 $G$ 和 $H$ 的比对图.由上面关于比对图的构建方法得知,比对图中的顶点是分别来自两个 PIN 的相似蛋白质,因此,其顶点集合为$\{\{u_1,v_1\},\{u_2,v_1\},\{u_3,v_3\},\{u_4,v_2\},\{u_4,v_4\}\}$.比对图中的边就依据上面边权的定义来进行赋值.例如,图 $G$ 中 $u_1$ 和 $u_4$ 没有相互作用,图 $H$ 中 $v_1$ 和 $v_2$ 有相互作用,因此,比对图中顶点$\{u_1,v_1\}$和$\{u_4,v_2\}$之间的连接边用一个失配事件进行惩罚,其权重为$-\nu$;图 $G$ 中 $u_1$ 和 $u_2$ 有相互作用,$v_1$ 和自己相互作用,并且 $u_1$ 和 $u_2$ 之间是 out-paralog 的复制关系,因此,比对图中顶点$\{u_1,v_1\}$和$\{u_2,v_1\}$之间的连接边用一个匹配事件奖励同时用一个复制事件来惩罚,其权重为$\mu-\sigma$.

© 中国科学院软件研究所 http://www.jos.org.cn

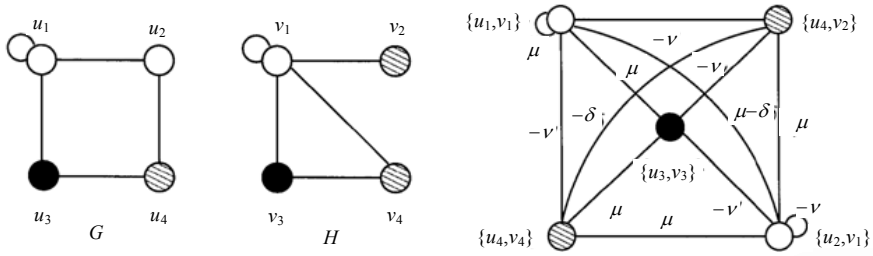



Fig.2 Construction of alignment graph in MaWISH[31]

图 2 MaWISH 方法中比对图的构建[31]

3. 比对算法

通过建立比对图,将两个 PIN 之间的比对问题转化为比对图中的最大权重子图问题,设计启发式搜索算法来完成问题的求解.算法的步骤如下:

(1) 选取一个关联匹配边个数最多的顶点,作为一个种子;

(2) 保守子图对应的顶点集合记为 $U$,用种子以及和种子通过匹配边连接的顶点初始化集合 $U$,并计算 $U$ 的导出子图的权重;

(3) 所有顶点组成的集合记为 $Q$,给 $Q$ 中的顶点赋一个权值,用来说明将该顶点从集合 $U$ 当中删除(如果该顶点已经在集合 $U$ 中)或者将该顶点加入到集合 $U$ 中(如果该顶点没有在集合 $U$ 中)对 $U$ 的导出子图所带来的权重变化;

(4) 从 $Q$ 中选取权值最大的那个顶点,考察它给 $U$ 的导出子图带来的权重变化.如果增加权重,那么将其加入到集合中或者从集合中删除,并修改当前 $Q$ 中顶点的权值以及 $U$ 的导出子图的权重.重复这个过程直到 $Q$ 为空;

(5) $U$ 对应的导出子图就是一个局部最优的比对结果.

MaWISH 方法提供了一个建立比对图的参考,顶点的匹配关系直接体现在比对图的顶点对中,相互作用的匹配借助复制/变异模型的 3 种进化事件用比对图边上的权重来描述.基于比对图模型将两个 PIN 的比对转化为比对图中权重最大子图的搜索问题,设计启发式搜索算法完成问题求解,权重最大子图即对应两个 PIN 的一个局部比对.

2.2.2 NetworkBLAST 比对方法

Sharan 等人[25]提出了多个 PIN 之间局部比对方法用以挖掘保守子模块,在此基础上进行蛋白质功能和相互作用的预测.该方法首先用相互作用的可信度给其赋权,基于多个赋权的 PIN 建立比对图.类似于 MaWISH 中的比对图,其顶点对应一组序列相似的蛋白质,它们分别来自不同的生物,边代表蛋白质之间存在的保守相互作用.该方法的一个重要特色在于建立了一个概率模型,如公式(8)所定义的 log likelihood ratio 模型.其中,$M_c$ 代表真实 PIN 的保守模型,$M_n$ 代表对应随机网络的空模型,$U$ 是保守模块对应的的顶点集合.

$$L(U)=\log\frac{\Pr(O_U\mid M_c)}{\Pr(O_U\mid M_n)} \tag{8}$$

基于这样一个模型分别定义比对图中顶点对和边对的相似度得分函数,得到一个顶点和边都带权重的比对图.将多个 PIN 之间的比对问题转化为在比对图中搜索权重最大子图问题,设计启发式搜索算法求解问题.

该方法的特点在于,借助公式(8)中的概率模型设计保守模块中顶点和边的相似性得分函数,借助该模型计算相似性得分,倾向于将较高的相似性得分赋给那些统计显著性较高的保守模块.这样做有利于提高比对结果的统计显著性.

2.2.3 Græmlin 比对方法

Flannick 等人提出了一个通用的多个生物网络的全局比对方法 Græmlin 1.0[32],类似 NetworkBLAST 中的

© 中国科学院软件研究所 http://www.jos.org.cn

相似度度量方法,基于 log likelihood ratio 概率模型定义顶点的得分函数和边的得分矩阵.搜索算法将匹配的蛋白质形成的集合作为一个等价类,在完成比对搜索时采用逐步扩张的方式,从一组相似性得分最高的蛋白质集合开始,逐步搜索和它们关联的相似蛋白质,最终形成一个以等价类为顶点的比对图,每个等价类映射一组相似的蛋白质.该方法中的等价类等同于前两个方法中比对图的顶点,不同之处在于,并不是一开始就建立一个完整的比对图,而是从一组匹配的蛋白质开始逐步搜索扩充完善比对图,直到算法执行结束时建立了一个完整的比对图.多个 PIN 的比对结果就隐含在这个完整的比对图中.

作者随后在 Græmlin1.0 的基础上又提出了 Græmlin2.0[33]比对方法,与 Græmlin1.0 的框架相同,但做出了如下改进:

(1) 在相似度函数的定义中引入多种特征参数;

(2) 针对不同类型的生物网络比对,建立自适应的相似性得分函数.通过参数训练算法为不同类型生物网络的比对选取合适的特征参数,确定化相似度函数;

(3) 提高了搜索算法运行效率;

(4) 建立了 benchmarks 用来对不同的生物网络比对方法进行测试比较.

2.2.4 基于图模型的启发式搜索方法的特点

从上面的分析可以看出,基于图模型的启发式搜索方法求解比对问题的特点如下:

(1) 比对图的建立,将多个生物网络融合成一个比对图,例如点积图[25,31]或边积图[26];

(2) 相似性得分函数的定义,根据生物网络比对的匹配准则设计相似性得分函数;

(3) 比对问题的转化,基于前两步的工作将多个生物网络之间的比对问题转化为一个比对图中的对应问题,例如最大权重子图问题[25,31]或最大团问题[26];

(4) 搜索算法的设计,设计启发式的搜索算法完成比对图中的问题;

(5) 比对结果直接反映在比对图中,搜索得到的每个子图对应一个局部映射.

### 2.3 基于目标函数的约束优化方法

生物网络比对问题是一类优化问题,具有一般优化问题的特征,基于目标函数的约束优化方法解决生物网络比对问题,通过将比对问题转化为某个常规的优化问题,采用已知的求解方法予以解决.此处的目标函数是一个更广义的概念,它是对生物网络比对映射关系的一种描述,可以呈现为传统优化问题中明确定义的目标函数.生物网络比对的匹配准则可以转化为约束条件,但不局限于此.通过将比对问题转化为常规优化问题进行求解,间接得到比对问题的解.不同于基于图模型的搜索方法:一是并不明确定义相似度函数,很多时候将其转换为约束条件;二是所得到的最优解并不直接反映比对结果,而是需要转换.

2.3.1 IsoRank 比对方法

Singh 等人[27,42]提出的 IsoRank 比对方法完成两个或者多个生物网络的全局比对.该方法是一种基于谱的方法,建模的工具是矩阵,矩阵有很多的特性可以巧妙地用以解决其他领域的很多难题.IsoRank 方法正是借助矩阵的特征值去计算相似性,基于相似性矩阵提取比对结果.

1. 利用矩阵的特征值建模比对问题

如果两个顶点的邻居比较相似的话,那么这两个顶点就比较相似.基于这个匹配准则,通过邻居的相似度计算得到顶点本身的相似度,计算规则如公式(9)所描述.

$$R = AR,\ \text{where}\ A[i,j][u,v] = \begin{cases} \dfrac{1}{|N(u)||N(v)|}, & \text{if}\ (i,u) \in PIN_1, (j,v) \in PIN_2 \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

给定 $PIN_1$ 有 $m$ 个蛋白质,$PIN_2$ 有 $n$ 个蛋白质,$R$ 是一个$(m\times n)\times 1$ 的列向量,其元素值是对应蛋白质之间的相似性.

从公式(9)可以看出,$R$ 是 $A$ 矩阵特征值 1 所对应的特征向量.由 $A$ 矩阵的性质可知,1 是其模最大的特征值,那么 $R$ 是 $A$ 矩阵模最大的特征值对应的特征向量.因此,$R$ 可以通过幂法求解得到,具体计算如公式(10):

© 中国科学院软件研究所 http://www.jos.org.cn

$$R^{k+1} \leftarrow AR^k / |AR^k| \tag{10}$$

如果考虑蛋白质序列的相似性度量,其迭代计算如公式(11):

$$R = \alpha AR + (1-\alpha)E,\ 0 \leqslant \alpha \leqslant 1 \tag{11}$$

其中,$E$ 是两个 PIN 中蛋白质序列的相似度,通过 BLAST 工具比对得到,并对其进行归一化处理.

上述的步骤完成了比对问题的建模,借助矩阵将比对问题转化为矩阵的特征值问题,通过幂法求解该特征值问题得到两个 PIN 比对的相似性向量 $R$.

2. 映射关系提取

对于 $k$ 个 PIN 的比对,通过上述方法得到两两之间的相似性向量 $R$,建立一个 $k$ 部图,如图 3 所示.其中,每一层的点代表一个 PIN 中的所有蛋白质,不同层之间的蛋白质用一条边相连接,边上的权重为两个蛋白质之间的相似性,也就是这两个 PIN 之间的相似性向量 $R$ 中的元素值.例如,$PIN_1$ 中的蛋白质 $1_2$ 和 $PIN_2$ 中的蛋白质 $2_3$ 之间有一条权重为 $R_{23}^{12}$ 连接边.从图 3 的 $k$ 部图中提取 $k$ 个 PIN 比对的映射关系,采用启发式的搜索算法来完成.首先从 $k$ 部图中提取权重最大的两个边所关联的蛋白质初始化目标集合,在后续的过程采用局部最优的策略最大限度地扩充该集合,使其中的蛋白质尽可能的相似,直至没有满足要求的蛋白质可以加入到该集合中.此时,目标集合就对应一组匹配的蛋白质,从 $k$ 部图中删除这一组顶点,从余下的 $k$ 部图中继续这一过程,直至 $k$ 部图为空.最后,将每个匹配集合中的蛋白质还原到 $k$ 个 PIN 中,每个 PIN 中由这些被匹配的蛋白质形成的导出子图就对应保守子网络.

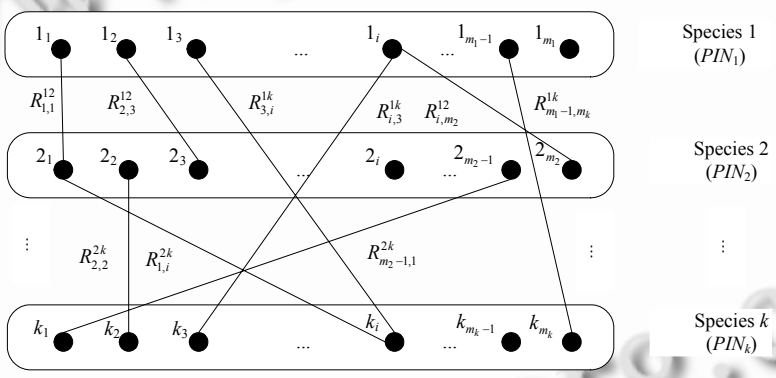



Fig.3 $K$-Partite graph of biological network alignment in IsoRank

图 3 IsoRank 中生物网络比对的 $k$ 部图

3. IsoRankN 比对方法

与 IsoRank 同一研究小组的 Liao 等人[55]采用同样的建模和相似性计算方法得到了多个 PIN 之间的相似性向量,建立了赋权的 $k$ 部图,通过在 $k$ 部图上运行一个谱聚类算法 PageRank Nibble 发掘保守的功能模块.

2.3.2 MNAligner 比对方法

Li 等人提出的 MNAligner 比对方法[40]将生物网络比对问题转化为一个整数二次规划问题予以求解.该方法描述如下:

(1) 确定生物网络的邻接矩阵、两个生物网络顶点之间的相似性矩阵,两个生物网络的匹配矩阵作为问题的待求解目标;

(2) 借助步骤 1 中定义的矩阵将两个生物网络比对的最大匹配归约为如下目标函数,约束条件为比对的匹配准则.

© 中国科学院软件研究所 http://www.jos.org.cn

$$\max_{X} f(G_1,G_2)=\lambda\sum_{i=1}^{m}\sum_{j=1}^{n}s_{ij}x_{ij}+(1-\lambda)\sum\sum\sum\sum a_{ik}b_{jl}x_{ij}x_{kl}\ \text{s.t.}\ \begin{cases}\sum_{j=1}^{n}x_{ij}\leqslant 1,\ i=1,2,\ldots,m\\ \sum_{i=1}^{m}x_{ij}\leqslant 1,\ j=1,2,\ldots,n\\ x_{ij}=0,1,\quad i=1,2,\ldots,m;j=1,2,\ldots,n\end{cases} \tag{12}$$

其中,目标函数为两个生物网络 $G_1$ 和 $G_2$ 比对的相似性得分.其中,$X$ 是待求解的匹配矩阵.目标函数分为两部分:一是顶点对之间的相似性得分,另外一部分是两个生物网络的结构匹配得分,这两部分所占比重通过可调参数 $\lambda$ 来控制.

通过上面的建模,将生物网络比对问题归约成一个整数二次规划问题,进一步松弛条件得到一个二次规划问题,也可以将此整数二次规划问题转化为一个整数线性规划问题,借助已有的方法对其进行求解.

2.3.3 基于目标函数的约束优化方法的特点

通过以上的分析,我们归纳基于目标函数的约束优化方法解决生物网络比对问题的特点:

(1) 比对问题的规约.生物网络比对问题也是一类优化问题,通过发掘比对问题的特性将其规约为一个已知的优化问题.例如,Klau[41]将比对问题转化为一个非线性整数规划问题;

(2) 优化问题求解.通过前一步的转化之后,一般采用已知的求解方法予以解决.例如,文献[41]将比对问题转化为非线性整数规划问题,采用拉格朗日松弛法对问题进行求解;

(3) 解的还原.不同于第 2.2 节中的方法,比对结果直接反映在比对图上;基于目标函数的约束优化方法由于将比对问题转化为其他问题进行求解,因此有时需要将比对结果进行还原.例如,IsoRank 中通过在 $k$ 部图中运行启发式搜索算法提取比对的映射关系.

### 2.4 基于分治策略的模块化比对方法

由于生物网络的规模较大并且具有模块化的结构[15−22,56],因此采用分治策略将原来的网络划分为小的模块,在模块比对的基础上完成网络比对.图 4 给出了这类比对方法的图示.原来的两个生物网络 $G$ 和 $H$,通过模块划分方法分别得到两组子网络 $G_1,G_2,G_3$ 和 $H_1,H_2$,然后在子网络之间两两进行比对.如果子网络的规模较大,那么对子网络继续进行模块划分,直至子网络规模满足模块比对算法的要求.

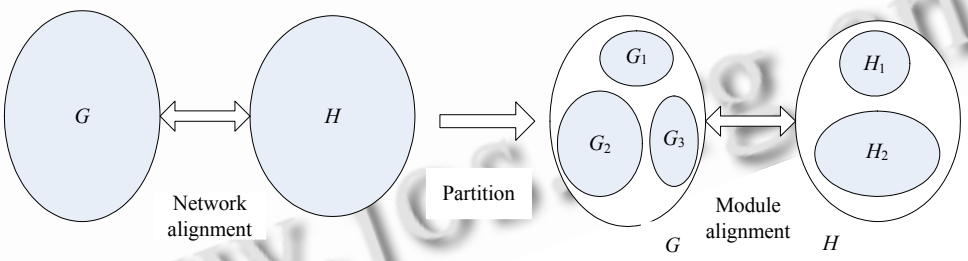



Fig.4 Modularized biological network alignment

图 4 模块化比对方法

2.4.1 Match-and-Split 比对方法

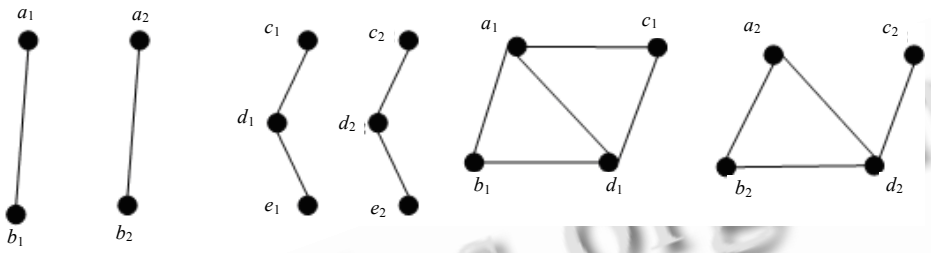
Narayanan 等人[34]提出的 Match-and-Split 比对方法基于分治策略采用模块化思想实现了两个 PIN 的比对.该方法中模块的划分通过匹配和分裂两个过程完成,匹配指网络中顶点的匹配,在匹配过程中没有匹配的点被删除,分裂指根据删除操作之后网络的连通性对图进行自然的划分,得到的每一个连通子网络即为一个模块.该方法的特点在于,其模块的比对是隐含在模块的划分过程中,模块的划分是原图在删除失配顶点之后导致的自然划分,不能划分的模块就意味着一对满足匹配规则的保守子网络.

1. 局部匹配规则与模块划分

文中定义了两类局部匹配规则完成顶点之间的匹配:一类是 $p$-path 规则,一类是 $s$-similar 规则.如果顶点 $u$ 和 $v$ 基于 $p$-path 规则匹配,那么就有一个包含 $u$ 的长度为 $p$ 的路径和一个包含 $v$ 的长度为 $p$ 的路径匹配;如果 $u$

© 中国科学院软件研究所 http://www.jos.org.cn

和 $v$ 基于 $s$-similar 规则匹配,那么至少有 $u$ 的 $s$ 个邻居和 $v$ 的 $s$ 个邻居匹配.图 5 给出了 1-path 和 2-path 的匹配规则示例,图 5(a)中,$a_1$ 和 $b_1$ 是网络 1 中的结点,$a_2$ 和 $b_2$ 是网络 2 中的结点,且 $a_1$ 和 $a_2$ 以及 $b_1$ 和 $b_2$ 分别相似.由于 $a_1$ 和 $a_2$ 相似,且包含 $a_1$ 和 $a_2$ 的有一条长为 1 的匹配路径,则说明 $a_1$ 和 $a_2$ 基于 1-path 准则匹配.图 5(b)中 $d_1$ 和 $d_2$ 基于长为 2 的相似路径匹配.图 5(c)给出了基于长为 1 或者长为 2 的相似路径匹配的两个完全匹配图,其中,具有相同字母标号的点相似.




(a) Similar length—1-paths (b) Similar length—2-paths (c) Match based similar length—$p$-paths ($p$=1,2)

(a) 基于 1-path 的相似匹配 (b) 基于 2-path 的相似匹配 (c) 基于 $p$-paths 相似匹配($p$=1,2)

Fig.5 $p$-path local match rule[34]

图 5 $p$-path 局部匹配规则[34]

当选择一种规则来匹配时,对网络 1 中每一个结点遍历,找网络 2 中与其匹配的点,并删除没有匹配的点;同理,反过来可以删除网络 2 中没有匹配的点,这样就完成了一轮匹配和模块的划分.

2. 算法描述

输入:代表两个生物网络的图 $G$ 和 $H$,以及 $G$ 和 $H$ 中顶点之间的相似度用来描述其匹配关系;

输出:$G$ 和 $H$ 中匹配的保守功能模块.

(1) 计算 $G$ 和 $H$ 中基于 $p$-path 或者 $s$-similar 准则的匹配顶点集合;

(2) 求出匹配顶点集合的导出子图 $G'$和 $H'$;

(3) 计算 $G'$和 $H'$的连通性,基于连通性,分别将 $G'$和 $H'$划分为模块 $G_1,G_2,\ldots,G_c$ 和 $H_1,H_2,\ldots,H_d$;

(4) 对 $G'$和 $H'$中的模块应用上述匹配准则继续匹配和划分,直到完全匹配.

2.4.2 基于分治策略的其他模块化比对方法

Towfic 等人提出的 BiNA 方法[53]首先将原来的生物网络分解为多个子网络,原有网络之间的比对转化成多个子网络之间的比对,借助图的核函数完成子网络比对,并将该方法用于两个生物网络之间的比对.基于多个生物网络之间的两两比对结果构建系统发生树,其实验结果与已知结果一致,从而验证了该方法的有效性.Jancura 等人提出的 DivAfull 方法[54]基于不同生物中的同源蛋白质对网络进行划分,采用一种迭代地精确搜索的方式完成模块之间的比对,发掘不同生物中保守的蛋白质复合物.前者采用了聚类的方法对模块进行划分,后者基于不同网络中蛋白质之间的同源关系进行模块划分.Match-and-Split 方法则是基于顶点之间的相似性进行划分.不同的划分方法产生的模块具有不同的特性,针对这些特性设计模块比对算法,进而通过多个模块之间的比对完成生物网络之间的比对.

2.4.3 基于分治策略的模块化比对方法的特点

基于分治策略的模块化比对方法求解生物网络比对问题,有两个关键步骤:一是模块的划分,二是模块的比对.这也是该方法的两个重要特点.

1. 模块的划分.

有两种方法完成模块的划分:一种方法基于网络的拓扑结构借助已有的模块划分算法(例如 MCODE[57]、CFinder[58]、GN 算法[59]和 Kernighan-Lin 算法[60])对生物网络进行划分,例如 Towfic 等人的 BiNA 比对方法正是借助这一类的模块划分算法来完成模块的划分;另外一种方法结合比对的匹配规则、拓扑结构或者其他的网

© 中国科学院软件研究所 http://www.jos.org.cn

络特性进行划分,例如 Match-and-Split 方法和 Jancura 等人的 DivAfull 方法.两种方法各有特色,前者可以借鉴已有的聚类算法进行划分,后者可以根据比对的不同应用目的设计更有针对性的划分算法;

2. 模块的比对.

模块规模相比原来的生物网络会小很多,参考图论中的匹配算法设计模块比对算法,也可以根据比对的准则和模块的特性设计启发式的模块比对算法.

### 2.5 不同类型比对方法的发展与比较

基于图模型的启发式搜索方法仍然是比对方法中的主流,将图的相似性计算方法和具体比对问题的特点相结合定义相似度函数,根据不同应用对问题进行建模,借助图论中算法进行求解.根据生物网络数据的特性,借鉴图的相似性计算和匹配的研究成果研究生物网络比对问题将会有很广阔的研究空间,并且互有助益.生物网络比对问题的最初研究正是基于图模型来展开的,目前依然有很多的比对方法是借助图模型和启发式的搜索算法来完成[33,37,38,61−66].研究者在这一类比对方法的研究中引入了越来越多的其他理论和模型,例如隐马尔可夫模型[67]等.

基于目标函数的约束优化方法的特点在于对生物网络比对问题的归约和裁剪,将某些优化方法应用到问题的求解中.其关键之处在于,我们需要充分挖掘问题本身的特点对其进行建模,例如基于图的矩阵表示形式,借助矩阵的特征以及矩阵运算来求解问题;借助优化解的某些特性来归约问题;或者将问题归约到其他领域的某个问题,利用已有方法进行求解.

基于分治策略的模块化比对方法基于生物网络的模块化结构将网络划分为模块,借助多个模块之间的比对完成原网络的比对.这类方法一方面降低了原问题的难度,另外一方面可以针对模块比对的特点充分利用局部信息加速比对过程.该方法的局限性在于可能会忽略掉模块之间的连接关系,比较依赖所用的模块划分方法.但是在通过比对挖掘保守功能模块方面,该方法的高效性会得到很好的发挥.借鉴生物网络的结构分析,模块挖掘及聚类算法的研究成果可以促进模块化比对方法的研究.

## 3 生物网络比对软件

基于成熟的生物网络比对算法,研究者也设计开发了相应的比对软件.我们整理了一些主要的生物网络比对软件(见表 2),第 1 列给出了软件的名称,第 2 列给出了软件的网址并对软件的功能和具体应用进行概要说明,第 3 列给出了软件的相关参考文献.根据软件的应用特点,可以将这些软件分为路径查询匹配软件和图的查询匹配软件两大类.

路径查询匹配软件主要完成生物网络比对中通路的查询和比对,表的上半部分主要是对路径查询匹配软件的分析和总结.其主要应用在代谢网络的比对中,由于代谢网络反映的是生物体内霉的催化反应,一般情况下,通路结构可以很好地反映细胞内的这一生化过程,因此,代谢网络的比对分析多数是针对代谢网络中的通路结构.另外也有一些针对蛋白质相互作用网络中保守路径的比对研究,例 PathBLAST[68]是面向 PIN 的路径匹配软件,可以进行蛋白质相互作用网络中通路结构的查询和保守路径的挖掘.

图的查询匹配软件主要完成生物网络比对中图的查询和比对,表的下半部分主要是图的查询匹配软件.图的查询匹配软件主要应用于 PIN 的比对研究中,但是通过修改参数,也可以完成其他类型网络的比对.例如,MNAligner[40]可以完成多种类型生物网络的比对,针对不同类型的生物网络设置不同的邻接矩阵和相似性矩阵.Græmlin 也是面向多种类型网络的通用比对软件,Græmlin2.0 有专门的参数训练方法可以自适应到不同类型网络的比对中.SAGA[48]是一个图数据库查询软件,采用了图的索引技术,完成在图数据库中搜索与模式图匹配的图.

## 4 生物网络比对的关键问题

生物网络比对是生物学、数学、信息学和计算机科学等学科综合交叉领域的研究问题.开展生物网络比对

© 中国科学院软件研究所 http://www.jos.org.cn

研究要对生物学的问题和研究需求有比较清楚的认识,这是通过数学方法对问题进行建模的基础;数学的思想方法和工具为很多学科的研究提供理论支撑,同样,它也为生物网络比对的研究提供重要的理论支持,尤其是图论算法和组合优化算法的相关理论;生物网络比对又可以看作是对生物网络数据的信息分析和处理,因此需要借鉴信息学的理论和方法进行比对问题的研究;计算机是生物网络数据承载、处理和呈现的介质,计算机科学对于生物网络比对问题的研究更是至关重要.生物网络比对是一个多学科交叉领域的综合问题,需要借助多个学科的理论和工具,需要多个学科研究者的共同研究和探索.在此,我们对生物网络比对研究中几个亟待解决的关键问题进行了分析与归纳.

**Table 2** The Softwares for biological network alignment

**表 2** 生物网络比对软件

| Softwares | Website & software description | References |
|---|---|---|
| PathBLAST | http://www.pathblast.org/<br>蛋白质相互作用网络中的路径查询匹配和保守路径的比对 | [68] |
| MetaPathwayHunter | http://www.cs.technion.ac.il/~olegro/metapathwayhunter/<br>在一组路径中查找与给定路径相似的子路径,软件提供图形化界面,<br>有分别运行于 Windows,Linux 和 MAC 操作系统上的版本 | [43] |
| MetaPAT | http://theinf1.informatik.uni-jena.de/metapat/<br>代谢路径的查询匹配工具 | [47] |
| PathMath | http://faculty.cs.tamu.edu/shsze/pathmatch<br>完成生物网络中路径的匹配查询 | [49] |
| PathAligner | http://bibiserv.techfak.uni-bielefeld.de/pathaligner/<br>用来重构和恢复代谢路径并可进行它们之间的比较,<br>并提供可视化的图形显示界面 | [69] |
| MetaRoute | http://www-bs.informatik.uni-tuebingen.de/Services/MetaRoute<br>提供代谢网络中源点(source)和目的点(product)间路径搜索匹配 | [45] |
| NetworkBLAST | http://www.cs.tau.ac.il/~bnet/networkblast.htm<br>两个或多个蛋白质相互作用网络的比对,发掘保守的功能模块.<br>两个网络比对有 Web 服务,多个网络比对有 NetworkBLASTM 软件下载,<br>目前只提供 Linuxx86 平台上的软件版本 | [70] |
| MaWISH | http://vorlon.case.edu/~mxk331/software/index.html<br>完成两个蛋白质相互作用网络的局部比对 | [31] |
| SAGA | http://www.eecs.umich.edu/saga<br>用于近似子图匹配,用户可以在图数据库中匹配查询图,<br>其核心是一个可变形的图距离模型,它合并了点和结构的近似匹配,<br>并在里面加入了一个索引算法来加速搜索过程 | [48] |
| Græmlin | http://graemlin.stanford.edu<br>多个生物网络之间的全局和局部比对,可以实现比对模式和查询模式的生物网络比较,2.0 版本在 1.0 版本的基础上实现了多种类型生物网络的参数自适应比对,<br>并且建立统一的 benchmarks | [32,33] |
| NetAlign | http://www1.ustc.edu.cn/lab/pcrystal/NetAlign<br>提供了简单直观的用户界面,输入查询网络以及目标网络进行比较,<br>设定 BLAST E-value 阈值.结果页分别显示网络比较所得到保守子网络结构,<br>同时提供指向外部数据库的链接 | [71] |
| MNAligner | http://intelligent.eic.osaka-sandai.ac.jp/chenen/MNAligner.htm<br>应用于多种类型分子网络以及网络上有权重和方向的多种情况的比对.<br>既可以完成线性路径、树结构的比对也可以进行一般图结构的比对.<br>目前该软件只是发行了 beta 版本以供初始的测试 | [40] |
| GraphMatch | http://faculty.cs.tamu.edu/shsze/graphmatch<br>完成生物网络中图的匹配查询 | [49] |
| IsoRank | http://people.csail.mit.edu/kennyluck/biology/isorank/<br>进行两个蛋白质相互作用网络的全局比对,发掘两个物种之间的同源蛋白质信息 | [27,42] |
| IsoRankN | http://isorank.csail.mit.edu/<br>将谱方法运用到多个网络的全局比对,具有很强的容错性以及计算的高效性 | [55] |
| TORQUE | http://www.cs.tau.ac.il/~bnet/torque.html<br>给定蛋白质复合物或者通路,在目标网络中查询与之匹配的结构.<br>基于查询蛋白质的序列信息,不借助查询蛋白质之间的拓扑结构完成匹配,<br>查询规模可以达到 25 个蛋白质 | [38] |

© 中国科学院软件研究所 http://www.jos.org.cn

### 4.1 生物网络数据的预处理

通过整理汇总各种实验数据得到的各种生物网络数据目前还不完善,进行生物网络比对首先要对数据进行预处理.

(1) 提取关键数据.生物网络数据库(例如 DIP[72],KEGG[73])中针对每一项数据包含数据的来源、实验方法等一些细节信息,因此需要从这些数据中提取关键项,并对相关数据信息做一些简单的抽象和处理,目的是将实验数据映射到某种可计算的数据类型.例如,我们可以将生物网络数据抽象为图,将实验数据中的生物学信息转换成图中顶点和边上的属性;

(2) 可信数据建模.实验数据不完整并且有噪声,因此需要对实验数据的可信度进行评估,这样更有利于建立科学的符合客观实际的数学模型.例如,文献[25,32,34,40,52]中通过给生物网络中的相互作用赋一个权重来说明它的可信度.可信数据的建模也是生物信息学中的一个研究热点.新加坡国立大学 Limsoon Wong 的研究小组[74,75]针对蛋白质相互作用数据中大量的假阳性和假阴性数据通过计算的方法对酵母 PIN 中的相互作用赋权值,用以衡量其可靠性.Sharan 的研究小组采用 Bader 的方法对 PIN 中相互作用的可靠性进行评估,并将可靠性信息用于 PIN 的比对[25].贺福初的研究小组[76]采用多种数据源建模人的 PIN 可信数据.Nataša 的研究小组也有关于可信 PPI 数据建模的研究[77].

### 4.2 相似性计算

相似性计算是生物网络比对的基础.在图模型比对方法中,相似性是启发式搜索算法中的搜索导向;在目标约束优化方法中,相似性是定义目标函数的基础;在模块化方法中,相似性同样指导着模块的划分和比对.目前,实际应用中的相似度函数的定义包括顶点的相似性和拓扑关系的相似性,利用这两项来计算整个网络的相似性.值得进一步研究的问题在于,如何将网络结构的相似性与生物学进化规律和生物学现象结合,使得生物网络的相似性计算更具有生物学意义.即如何引入更多的更合理的有生物学意义的比对特征到相似度函数的定义中.值得借鉴的方法是 Græmlin2.0[33],该方法在蛋白质序列和拓扑结构相似的基础上将功能相似性引入到相似度函数定义中.

### 4.3 比对方法的评价

比对方法的评价包含两个方面的内容:一是基于算法分析理论对比对算法的执行效率进行评估,二是基于算法运行结果的生物学意义进行评估.目前,有关比对方法的评价分为 3 个层面:第 1 层面是结果评价,贴合比对的应用目的,将实验结果和已有的功能数据库进行比较说明比对方法的敏感性(sensitivity)和特定性(specifity).我们将实验数据和功能数据库中的参考数据相一致的部分称为匹配数据,所谓敏感性是指匹配数据所占参考数据的比例,所谓特定性是指匹配数据所占实验数据的比例[34];第 2 层面的评价是统计显著性评价,这个评价是为了说明算法的生物学意义.首先建立一组和原有网络度序列一致的随机网络,在随机网络上运行比对算法,基于随机网络上的比对结果和真实生物网络上的比对结果设计评价函数,通过评价函数的值来说明算法的统计显著性;第 3 层面的评价是不同算法的横向比较,通过建立统一的标准和测试用例,对不同的比对方法进行量化分析比较.目前,关于比对方法的评价大多是前两个层面上的比较分析.Græmlin2.0 中首次建立了比对方法评价的 benchmarks,将自身和 NetworkBLAST,MaWISH,IsoRank 以及 Græmlin1.0 进行了量化评测.

## 5 生物网络比对的应用

生物网络数据是由生物体内分子之间的相互作用形成的一类重要的生物数据,研究网络数据更能从系统层面揭示生物学的规律.生物网络比对的应用主要集中在以下 4 个方面:

(1) 结构预测.通过不同生物的网络数据的比对研究,发现其在结构上的异同,进而借助模式生物去研究其他生物的网络数据[15,18].例如,对于蛋白质相互作用数据的比对研究可以预测新的蛋白质相互作用[25,31−33],发掘保守的功能模块[26,28,78−80],借助局部相似性预测蛋白质网络在进化过程中的复制和变异事件[26,31],进而研究其网络结构的演化模型[17,39,81];

© 中国科学院软件研究所 http://www.jos.org.cn

(2) 功能预测.类似于结构预测,通过生物网络比对借助已知生物网络中蛋白质的功能预测其他生物体中同源蛋白质的功能[25,26],抑或是基于功能的相似性预测蛋白质之间的同源性[27,82];

(3) 系统发生分析.结构和功能的预测是基于比对结果的局部信息进行的,系统发生分析则借助比对结果的全局信息.通过比对得到两个物种生物网络的相似性,基于此进行系统发生分析的研究或者在网络划分的基础上基于多个模块比对的结果进行系统发生分析的研究[83];

(4) 与疾病等相关的特定研究.将生物网络比对的思想和方法直接用于特定的网络,针对某个具体问题开展研究.例如,结合疾病相关的表型数据和蛋白质相互作用数据,搜索和某种疾病相关的子结构,以期从基因组学的层面解释研究疾病的发病机制,进而为疾病的诊断治疗以及药物发现提供有价值的信息[24,84,85].

## 6 结 论

随着生物网络数据的出现,生物网络比对成为生物信息学领域内的一个重要研究问题,网络数据本身固有的复杂性以及生物网络比对问题的组合优化特性给该问题的求解带来了挑战.生物网络比对和图的匹配、图的相似性计算等理论相关,该问题的研究不仅对生物网络数据分析和信息挖掘具有重要的意义,而且对基于图的算法理论研究及应用都有积极的意义.因此,近年来很多的研究者都致力于该问题的求解,提出了很多有价值的算法.

本文首先给出了生物网络比对的形式化定义,并从模型和算法的角度对现有的生物网络比对方法进行了研究和分析,对 3 种类型的比对方法结合具体的应用进行了分类探讨,总结了每类方法的特点及其应用.基于图模型的启发式搜索方法主要借助图论算法求解问题.随着各类结构化数据的日益增多,图的理论研究和各类应用问题吸引了很多的研究者,并提出了很多新的算法,将图论算法和生物网络数据的结构特征相结合,设计合适的图模型有效解决比对问题,这将仍然是解决比对问题的较好途径.基于目标函数的约束优化方法将比对问题转化为已知的优化问题求解,生物网络比对本质上也是一类组合优化问题,设计合适的目标函数对比对问题进行归约,借助已有的优化算法对问题进行求解.组合优化领域有大量的可借鉴的模型和算法,通过将生物网络比对问题转化为某个优化问题求解,是值得研究者继续探索的方向之一.基于分治策略的模块化方法将模块划分和比对相结合完成问题求解,由于生物网络的规模较大并且具有模块化的结构特征,因此,借助已有的模块划分算法或者设计新的模块划分算法对生物网络进行分解,在模块的基础上进行比对将给比对问题的研究带来新的契机.目前,有关这方面的工作还较少,我们的研究小组已经开展了这方面的研究工作.

生物网络比对通过比较的方法研究生物体结构、功能和进化等方面的相关性,借此研究可以发掘网络数据中有价值的功能模块,进行各方面的预测分析.除此之外,将生物网络比对的研究和实际应用相结合,例如疾病的诊断[24]等,结合具体应用的特点开展有针对性的研究工作,将使得生物网络比对的研究有更实际的应用价值,也将是生物网络比对研究的一个新方向.

**致谢** 在此我们向对本文写作给予建议和帮助的其他小组成员表示感谢.

**References**:

[1] Uetz P, Giot L, Cangney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang MJ, Johnston M, Fields S, Rothberg JM. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature, 2000,403:623−627.

[2] Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, Timm J, Mintzlaff S, Abraham C, Bock N, Kietzmann S, Goedde A, Toksöz E, Droege A, Krobitsch S, Korn B, Birchmeier W, Lehrach H, Wanker EE. A human protein-protein interaction network: A resource for annotating the proteome. Cell, 2005,122:957−968.

[3] Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proceedings of the National Academy of Sciences, 2001,98:4569−4574. [doi: 10.1073/pnas.061034498]

[4] Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, Yang L, Wolting C, Donaldson I, Schandorff S, Shewnarane J, Vo M, Taggart J, Goudreault M, Muskat B, Alfarano C, Dewar D, Lin Z, Michalickova

© 中国科学院软件研究所 http://www.jos.org.cn

K, Willems AR, Sassi H, Nielsen PA, Rasmussen KJ, Andersen JR, Johansen LE, Hansen LH, Jespersen H, Podtelejnikov A, Nielsen E, Crawford J, Poulsen V, Sørensen BD, Matthiesen J, Hendrickson RC, Gleeson F, Pawson T, Moran MF, Durocher D, Mann M, Hogue CW, Figeys D, Tyers M. System identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature, 2002,415:180−183.

[5] Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. Nature, 2001,409:533−538. [doi: 10.1038/35054095]

[6] Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert TL, Wilson CJ, Bell SP, Young RA. Genome-Wide location and function of DNA binding proteins. Science, 2000,290:2306−2309. [doi: 10.1126/science.290.5500.2306]

[7] Cheng YS, Liu JY. Tandem affinity purification technique and its application in proteomics. Progress in Biochemistry and Biophysics, 2004,31(4):379−383 (in Chinese with English abstract).

[8] Gavin AC, BoÈsche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, Remor M, HoÈfert C, Schelder M, Brajenovic M, Ruffner H, Merino A, Klein K, Hudak M, Dickson D, Rudi T, Gnau V, Bauch A, Sonja B, Huhse B, Leutwein C, Heurtier MA, Copley RR, Edelmann A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G, Furga GS. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature, 2002,415:141−147. [doi: 10.1038/415141a]

[9] Liu KD, Zhao JL. Progress of protein chip technology. China Biotechnology, 2004,24(12):48−52,58 (in Chinese with English abstract).

[10] Li M, Zhou ZC. Protein chip. Chemistry of Life, 2001,21(2):156−157 (in Chinese).

[11] Zhong CY, Peng R, Peng JX, Hong HZ. Protein chip technology. Biotechnology Bulletin, 2004,2:34−37 (in Chinese with English abstract).

[12] Yang HY, Ying WT, Qian XH. Current progress in proteome. Progress in Natural Science, 2002,12(1):13−17 (in Chinese).

[13] Ma XJ, Hu R, Lü H, Wei KK, Zhang LL, Xue SX, Hou YD. Transformation of human interferon a1c/86D with phage display. Science in China (Series C), 1999,29(2):209−216 (in Chinese with English abstract).

[14] Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, Niranjan V, Muthusamy B, Gandhi TKB, Gronborg M, Ibarrola N, Deshpande N, Shanker K, Shivashankar HN, Rashmi BP, Ramya MA, Zhao Z, Chandrika KN, Padma N, Harsha HC, Yatish AJ, Kavitha MP, Menezes M, Choudhury DR, Suresh S, Ghosh N, Saravana R, Chandran S, Krishna S, Joy M, Anand SK, Madavan V, Joseph A, Wong GW, Schiemann WP, Constantinescu SN, Huang L, Khosravi-Far R, Steen H, Tewari M, Ghaffari S, Blobe GC, Dang CV, Garcia JGN, Pevsner J, Jensen ON, Roepstorff P, Deshpande KS, Chinnaiyan AM, Hamosh A, Chakravarti A, Pandey A. Development of human protein reference database as an initial platform for approaching systems biology in humans. Genome Research, 2003,13:2363−2371. [doi: 10.1101/gr.1680803]

[15] Sharan R, Ideker T. Modeling cellular machinery through biological network comparison. Nature Biotechnology, 2006,24(4): 427−433. [doi: 10.1038/nbt1196]

[16] Almaas E. Biological impacts and context of network theory. The Journal of Experimental Biology, 2007,210:1548−1558. [doi: 10.1242/jeb.003731]

[17] Silva E, Stumpf MPH. Complex networks and simple models in biology. Journal of The Royal Society Interface, 2005,2:419−430. [doi: 10.1098/rsif.2005.0067]

[18] Srinivasan BS, Shah NH, Flannick JA, Abeliuk E, Novak AF, Batzoglou S. Current progress in network research: Toward reference networks for key model organisms. Brief in Bioinform 2007,8(5):318−332. http://bib.oxfordjournals.org/cgi/content/short/bbm038v1

[19] Sun JC, Xu JL, Li YX, Shi TL. Analysis and application of large-scale protein-protein interaction data. Chinese Science Bulletin, 2005,50(19):2055−2060 (in Chinese).

[20] Guan W, Wang J, He FC. The advance in research methods for large-scale protein-protein interactions. Chinese Bulletin of Life Sciences, 2006,18(5):507−512 (in Chinese with English abstract).

[21] Liu ZY, Li D, Zhu YP, He FC. Progress in the evolutionary analysis of protein interaction networks. Progress in Biochemistry and Biophysics, 2009,36(1):13−24 (in Chinese with English abstract).

[22] Liu W, Li D, Zhu YP, He FC. Bioinformatics analysis in signal transduction network. Science in China (Series C), 2008,38(11): 999−1006 (in Chinese with English abstract).

[23] Zhang S, Jin GX, Zhang XS, Chen LN. Discovering functions and revealing mechanisms at molecular level from biological networks. Proteomics, 2007,7:2856−2869. [doi: 10.1002/pmic.200700095]

© 中国科学院软件研究所 http://www.jos.org.cn

[24] Kolar M, Lassig M, Berg J. From protein interactions to functional annotation: Graph alignment in Herpes. BMC Systems Biology, 2008,2:90−99. [doi: 10.1186/1752-0509-2-90]

[25] Sharan R, Suthranm S, Kelley RM, Kuhn T, McCuine S, Uetz P, Sittler T, Karp RM, Ideker T. Conserved patterns of protein interaction in multiple species. PNAS, 2005,102(6):1974−1979. [doi: 10.1073/pnas.0409522102]

[26] Liang Z, Xu M, Teng MK, Niu LW. Comparison of protein interaction networks reveals species conservation and divergence. BMC Bioinformatics, 2006,7:457−472. [doi: 10.1186/1471-2105-7-457]

[27] Singh R, Xu JB, Berger B. Global alignment of multiple protein interaction networks with application to functional orthology detection. PNAS, 2008,105(35):12763−12768. [doi: 10.1073/pnas.0806627105]

[28] Sharan R, Ideker T, Kelley B, Shamir R, Karp RM. Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data. Journal of Computational Biology, 2005,12(6):835−846. [doi: 10.1089/cmb.2005.12.835]

[29] Hirsh E, Sharan R. Identification of conserved protein complexes based on a model of protein network evolution. Bioinformatics, 2006,23:e170−e176. [doi: 10.1093/bioinformatics/btl295]

[30] Tian WH, Samatova NF. Pairwise alignment of interaction networks by fast identification of maximal conserved patterns. Pacific Symposium on Biocomputing, 2009,14:99−110.

[31] Koyuturk M, Kim Y, Topkara U, Subramaniam S, Szpankowski W, Grama A. Pairwise alignment of protein interaction networks. Journal of Computational Biology, 2006,13(2):182−199. [doi: 10.1089/cmb.2006.13.182]

[32] Flannick J, Novak A, Srinivasan BS, McAdams HH, Batzoglou S. Graemlin: General and robust alignment of multiple large interaction networks. Genome Research, 2006,16:1169−1181. [doi: 10.1101/gr.5235706]

[33] Flannick J, Novak A, Do ChB, Srinivasan BS, Batzoglou S. Automatic parameter learning for multiple network alignment. In: Proc. of the RECOMB 2008. LNBI 4955, 2008. 21−231. http://portal.acm.org/citation.cfm?id=1804334 [doi: 10.1089/cmb.2009.0099]

[34] Narayanan M, Karp RM. Comparing protein interaction networks via a graph match-and-split algorithm. Journal of Computational Biology, 2007,14(7):892−907. [doi: 10.1089/cmb.2007.0025]

[35] Ogata H, Fujibuchi W, Goto S, Kanehisa M. A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters. Nucleic Acids Research, 2000,28(20):4021−4028. [doi: 10.1093/nar/28.20.4021]

[36] Kelley BP, Sharan R, Karp RM, Sittler T, Root DE, Stockwell BR. Conserved pathways within bacteria and yeast as revealed by global protein network alignment. PNAS, 2003,100(20):11394−11399. [doi: 10.1073/pnas.1534710100]

[37] Kalaev M, Bafna V, Sharan R. Fast and accurate alignment of multiple protein interaction networks. Journal of computational biology, 2009, 16(8):989−999. [doi: 10.1089/cmb.2009.0136]

[38] Bruckner S, Hüffner F, Karp RM, Shamir R, Sharan R. Topology-free querying of protein interaction networks. Journal of computational biology, 2010, 17(3):237−252. [doi: 10.1093/nar/gkp474]

[39] Berg J, Lässig M. Cross-Species analysis of biological networks by Bayesian alignment. PNAS, 2006,103(29):10967−10972. [doi: 10.1073/pnas.0602294103]

[40] Li ZP, Zhang SH, Wang Y, Zhang XS, Chen LN. Alignment of molecular networks by integer quadratic programming. Bioinformatics, 2007,23(13):1631−1639. [doi: 10.1093/bioinformatics/btm156]

[41] Klau GW. A new graph-based method for pairwise global network alignment. BMC Bioinformatics, 2009,10(Supp.):59−67. [doi: 10.1186/1471-2105-10-S1-S59]

[42] Singh R, Xu JB, Berger B. Pairwise global alignment of protein interaction networks by matching neighborhood topology. In: Proc. of the RECOMB 2007. LNBI 4453, 2007. 16−31. http://portal.acm.org/citation.cfm?id=1758224

[43] Pinter RY, Rokhlenko O, Lotem EY, Ukelson MZ. Alignment of metabolic pathways. Bioinformatics, 2005,21(162005):3401−3408. [doi: 10.1093/bioinformatics/bti554]

[44] Shlomi T, Segal D, Ruppin E, Sharan R. QPath: A method for query pathways in a protein-protein interaction network. BMC Bioinformatics, 2006,7:199. [doi: 10.1186/1471-2105-7-199]

[45] Blum T, Kohlbacher O. MetaRoute: Fast search for relevant metabolic routes for interactive network navigation and visualization. Bioinformatics, 2008,24(18):2108−2109. [doi: 10.1093/bioinformatics/btn360]

[46] Li YL, Ridder D, Groot MJL, Reinders MJT. Metabolic pathway alignment between species using a comprehensive and flexible similarity measure. BMC Systems Biology, 2008,2:111. [doi: 10.1186/1752-0509-2-111]

[47] Wernicke S, Rasche F. Simple and fast alignment of metabolic pathways by exploiting local diversity. Bioinformatics, 2007,23(25): 1978−1985. [doi: 10.1093/bioinformatics/btm279]

[48] Tian Y, McEachin RC, Santos C, States DJ, Patel JM. SAGA: A subgraph matching tool for biological graphs. Bioinformatics, 2007,23(2):232−239. [doi: 10.1093/bioinformatics/btl571]

© 中国科学院软件研究所 http://www.jos.org.cn

[49] Yang QW, SZE SH. Path matching and graph matching in biological networks. Journal of Computational Biology, 2007,14(1): 56−57. [doi: 10.1089/cmb.2006.0076]

[50] Brevier G, Rizzi R, Vialette S. Pattern matching in protein-protein interaction graphs. In: Proc. of the FCT 2007. LNCS 4639, 2007. 137−148. http://www.springerlink.com/content/q6v6j0g852757016/

[51] Zhang SH, Zhang XS, Chen LN. Biomolecular network querying: A promising approach in systems biology. BMC Systems Biology, 2008,2:5. [doi: 10.1186/1752-0509-2-5]

[52] Ali W, Deane CM. Functionally guided alignment of protein interaction networks for module detection. Bioinformatics, 2009, 25(23):3166−3173. [doi: 10.1093/bioinformatics/btp569]

[53] Towfic F, Greenlee MHW, Honavar V. Aligning biomolecular networks using modular graph kernels. In: Proc. of the WABI 2009. LNBI 5724, 2009. 345−361. http://portal.acm.org/citation.cfm?id=1812935

[54] Jancura P, Heringa J, Marchiori E. Divide, align and full-search for discovering conserved protein complexes. In: Proc. of the LNCS EvoBIO. 2008. 71−82. http://portal.acm.org/citation.cfm?id=1792681

[55] Liao CS, Lu KH, Baym M, Singh R, Berger B. IsoRankN: Spectral methods for global alignment of multiple protein networks. Bioinformatics, 2009,25:i253−i258. [doi: 10.1093/bioinformatics/btp203]

[56] Hartwell LH, Hopfield JJ, Leibler S, Murray AW. From molecular to modular cell biology. Nature, 1999,402:C47−C52. [doi: 10.1038/35011540]

[57] Bader GD, Hogue CW. An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics, 2003,4:2−28. [doi: 10.1186/1471-2105-4-2]

[58] Adamcsek B, Palla G, Farkas IJ, Dereny I. CFinder: Locating cliques and overlapping modules in biological networks. Bioinformatics, 2006,22(8):1021−1023. [doi: 10.1093/bioinformatics/btl039]

[59] Girvan M, Newman MEJ. Community structure in social and biological networks. PNAS, 2002,99(12):7821−7826. [doi: 10.1093/bioinformatics/btl039]

[60] Kernighan BW, Lin S. A efficient heuristic procedure for partitioning graphs. Bell System Technical Journal, 1970,49(2):291−307.

[61] Deniélou YP, Boyer F, Viari A, Sagot MF. Multiple alignment of biological networks: A flexible approach. In: Proc. of the CPM 2009. LNCS 5577, 2009. 263−273. http://portal.acm.org/citation.cfm?id=1574253

[62] Guo X, Hartemink AJ. Domain-Oriented edge-based alignment of protein interaction networks. Bioinformatics, 2009,25:i240−i246. [doi: 10.1093/bioinformatics/btp202]

[63] Fertin G, Rizzi R, Vialette S. Finding exact and maximum occurrences of protein complexes in protein-protein interaction graphs. Journal of Discrete Algorithms, 2009,7:90−101. [doi: 10.1016/j.jda.2008.11.003]

[64] Zhang SJ, Li SR, Yang J. GADDI: Distance index based subgraph matching in biological networks. In: Proc. of the 12th Int'l Conf. on Extending Database Technology (EDBT 2009). 2009. 192−203. http://portal.acm.org/citation.cfm?id=1516384 [doi:10.1145/1516360.1516384]

[65] Zaslavskiy M, Bach F, Vert JP. Global alignment of protein-protein interaction networks by graph matching methods. Bioinformatics, 2009,25:i259−i267. [doi: 10.1093/bioinformatics/btp196]

[66] Cootes AP, Muggleton SH, Sternberg MJE. The identification of similarities between biological networks: Application to the metabolome and interactome. Journal of Molecular Biology, 2007,369:1126−1139. [doi: 10.1016/j.jmb.2007.03.013]

[67] Qian XN, Sze SH, Yoon BJ. Querying pathways in protein interaction networks based on hidden markov models. Journal of Computational Biology, 2009,16(2): 145−157. [doi: 10.1089/cmb.2008.02TT]

[68] Kelley BP, Yuan B, Lewitter F, Sharan R, Stockwell BR, Ideker T. PathBLAST: A tool for alignment of protein interaction networks. Nucleic Acids Research, 2004,32:W83−W88. [doi: 10.1093/nar/gkh411]

[69] Chen M, Hofestadt R. An algorithm for linear metabolic pathway alignment. Silico Biology, 2005,5:111−128.

[70] Kalaev M, Smoot M, Ideker T, Sharan R. NetworkBLAST: Comparative analysis of protein networks. Bioinformatics, 2008,24(4): 594−596. [doi: 10.1093/bioinformatics/btm630]

[71] Liang Z, Xu M, Teng MK, Niu LW. NetAlign: A Web-based tool for comparison of protein interaction networks. Bioinformatics, 2006,22(17):2175−2177. [doi: 10.1093/bioinformatics/btl287]

[72] Xenarios I, Salwínski L, Duan XJ, Higney P, Kim SM, Eisenberg D. DIP, the database of interacting proteins: A research tool for studying cellular networks of protein interactions. Nucleic Acids Research, 2002,30(1):303−305. [doi: 10.1093/nar/30.1.303]

[73] Ogata H, Goto S, Sato F, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Research, 1999,27(1):29−34. [doi: 10.1093/nar/27.1.29]

© 中国科学院软件研究所 http://www.jos.org.cn

[74] Liu GM, Li JY, Wong L. Assessing and predicting protein interactions using both local and global network topological metrics. In: Proc. of the 19th Int'l Conf. on Genome Informatics (GIW 2008). 2008. 138−149. http://eproceedings.worldscinet.com/9781848163324/9781848163324_0012.html

[75] Chua HN, Wong L. Increasing the reliability of protein interactomes. Drug Discovery Today, 2008,13(15/16):652−658. [doi: 10.1016/j.drudis.2008.05.004]

[76] Li D, Liu WL, Liu ZY, Wang J, Liu QJ, Zhu YP, He FC. PRINCESS, a protein interaction confidence evaluation system with multiple data sources. Molecular & Cellular Proteomics, 2008,7(6):1043−1052. [doi: 10.1074/mcp.M700287-MCP200]

[77] Kuchaiev O, Rasajski M, Highm DJ, Przulj N. Geometric de-noising of protein-protein interaction networks. PLoS Computational Biology, 2009,5(8):e1000454. [doi: 10.1371/journal.pcbi.1000454]

[78] Dutkowski J, Tiurn J. Identification of functional modules from conserved ancestral protein-protein interactions. Bioinformatics, 2007,23:i149−i158. [doi: 10.1093/bioinformatics/btm194]

[79] Dittrich MT, Klau GW, Rosenwald A. Identifying functional modules in protein-protein interaction networks: An integrated exact approach. Bioinformatics, 2008,24:223−231. [doi: 10.1093/bioinformatics/btn161]

[80] Rivera CG, Murali TM. Identifying evolutionarily conserved protein interaction modules using GraphHopper. In: Proc. of the BICoB 2009. LNBI 5462, 2009. 67−68. http://portal.acm.org/citation.cfm?id=1537782. [doi: 10.1007/978-3-642-00727-9_9]

[81] Milenkovic T, Pržulj N. Uncovering biological network function via graphlet degree signatures. Cacer Informatics, 2008,6:257−273.

[82] Yosef N, Sharan N, Noble WS. Improved network-based identification of protein orthologs. Bioinformatics, 2008,24:i200−i206. [doi: 10.1093/bioinformatics/btn277]

[83] Erten S, Li X, Bebek G, Li J, Koyuturk. Phylogenetic analysis of modularity in protein interaction networks. BMC Bioinformatics, 2009,10:333. [doi: 10.1186/1471-2105-10-333]

[84] Karni S, Soreq H, Sharan R. A network-based method for predicting disease-causing genes. Journal of Computational Biology, 2009,16(2):181−189.

[85] Ideker T, Sharan R. Protein networks in disease. Genome Research, 2008,18:644−652. [doi: 10.1101/gr.071852.107]

**附中文参考文献:**

[7] 程永升,刘进元.串联亲和纯化(TAP)技术在蛋白质组学中的应用.生物化学与生物物理进展,2004,31(4):379−383.

[9] 刘康栋,赵建龙.蛋白质芯片技术进展.中国生物工程杂志,2004,24(12):48−52,58.

[10] 李民,周宗灿.蛋白质芯片.生命的化学,2001,21(2):156−157.

[11] 钟春英,彭蓉,彭建新,洪华珠.蛋白质芯片技术.生物技术通报,2004,2:34−37.

[12] 杨何义,应万涛,钱小红.蛋白质组技术的研究进展.自然科学进展,2002,12(1):13−17.

[13] 马学军,胡荣,吕海,魏开坤,张丽兰,薛水星,侯云德.噬菌体显示技术改造人干扰素$\alpha$1c/86D 的研究.中国科学(C 辑:生命科学),1999,29(2):209−216.

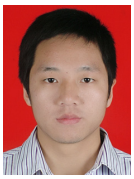
[19] 孙景春,徐晋麟,李亦学,石铁流.大规模蛋白质相互作用数据的分析与应用.科学通报,2005,50(19):2055−2060.

[20] 关薇,王建,贺福初.大规模蛋白质相互作用研究方法进展.生命科学,2006,18(5):507−512.

[21] 刘中扬,李栋,朱云平,贺福初.蛋白质相互作用网络进化分析研究进展.生物化学与生物物理进展,2009,36(1):13−24.

[22] 刘伟,李栋,朱云平,贺福初.信号传导网络的生物信息学分析.中国科学(C 辑:生命科学),2008,38(11):999−1006.

**郭杏莉**(1979－),女,陕西扶风人,博士生,讲师,主要研究领域为生物信息学,生物网络比对,图论算法的研究与应用.



**陈新**(1986－),男,硕士生,主要研究领域为生物网络比对.

**高琳**(1964－),女,教授,博士,博士生导师,主要研究领域为计算生物信息学,生物数据挖掘,图论与组合优化算法及其应用.

© 中国科学院软件研究所 http://www.jos.org.cn