



## 基于层次聚类的生物网络全局比对算法

田盼盼<sup>1</sup>, 陈璟<sup>1,2</sup>

(1. 江南大学 人工智能与计算机学院, 江苏 无锡 214122;

2. 江南大学 江苏省模式识别与计算智能工程实验室, 江苏 无锡 214122)

**摘要:** 生物网络比对是研究生物进化过程的重要手段, 不同物种间的比对不仅有助于理解物种的知识转移, 同时也有助于进行功能预测和检测保守功能成分。然而, 现有比对算法很难实现拓扑度量和生物度量同时最优。设计 JAlign 算法, 将拓扑相似性与归一化序列相似性相结合构成目标函数, 基于种子-扩展算法和模块检测进行全局比对。在种子筛选阶段, 利用 Jerarca 聚类算法划分功能模块, 借助目标函数计算模块间的相似性进行最优模块匹配, 并从匹配结果中提取部分节点对作为种子节点。在扩展阶段, 将比对从种子节点扩展至其邻居节点, 在选择节点对进行扩展比对时综合考虑节点之间的连接关系、度差值、节点相似性等因素。在此基础上, 为避免遗漏分散节点, 找到剩余未匹配的节点构建二分图, 以贪心方式进行最大加权二分图匹配, 并将匹配结果合并到比对集合中, 完成最终匹配。实验结果表明, JAlign 算法能够实现拓扑度量和生物度量的良好平衡, 其边正确性指标、诱导保守子结构得分、对称子结构得分和生物质量使用功能一致性指标均优于 L-GRAAL、SPINAL 和 ModuleAlign 算法, 在时间效率上也具有优势。  
**关键词:** 蛋白质相互作用网络; 网络比对; 层次聚类; 功能模块检测; 种子-扩展算法

开放科学(资源服务)标志码(OSID):



中文引用格式: 田盼盼, 陈璟. 基于层次聚类的生物网络全局比对算法[J]. 计算机工程, 2022, 48(2): 65-71, 78.

英文引用格式: TIAN P P, CHEN J. Global biological network alignment algorithm based on hierarchical clustering[J]. Computer Engineering, 2022, 48(2): 65-71, 78.

## Global Biological Network Alignment Algorithm Based on Hierarchical Clustering

TIAN Panpan<sup>1</sup>, CHEN Jing<sup>1,2</sup>

(1. School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, Jiangsu 214122, China; 2. Jiangsu Provincial Engineering Laboratory of Pattern Recognition and Computing Intelligence, Jiangnan University, Wuxi, Jiangsu 214122, China)

**[Abstract]** Biological network alignment is an important means to study the process of biological evolution. The comparison between different species assists not only in understanding the knowledge transfer between species, but also in functional prediction and conserved functional component detection. However, existing comparison algorithms are usually unable to achieve optimal topological measure and biological measure at the same time. This paper presents design of the JAlign algorithm, which combines topological similarity and normalized sequence similarity to form the objective function, and performs global alignment based on the seed-and-extend algorithm and module detection. In the seed selection stage, the Jerarca clustering algorithm is used to divide the functional modules. The similarity between the modules is calculated by using the objective function to perform optimal module matching, and some node pairs are extracted from the matching results as seed nodes. In the extension stage, the comparison extends from the seed node to its neighbor nodes. When selecting the node pair for extended comparison, the connection relationship between nodes, degree difference, node similarity and other factors are comprehensively considered. On this basis, in order to avoid missing scattered nodes, the remaining unmatched nodes are found to build a bipartite graph. The maximum weighted bipartite graph is matched in the greedy way, and the matching results are merged into the comparison set to complete the final matching. The experimental results show that the JAlign algorithm can achieve a good balance between topological measure and biological measure. It provides better results in Edge Correctness(EC) index, Induced Conserved-structure Score(ICS), Symmetric Sub-structure Score( $S^3$ ) and Functional Coherence(FC) index than L-GRAAL, SPINAL and ModuleAlign algorithms. The proposed algorithm also displays advantages in time efficiency.  
**[Key words]** Protein-Protein Interaction(PPI) network; network alignment; hierarchical clustering; functional module detection; seed-and-extend algorithm

DOI: 10.19678/j.issn.1000-3428.0060360

基金项目: 江苏省青年科学基金(BK20150159)。

作者简介: 田盼盼(1995—), 女, 硕士研究生, 主研方向为复杂网络; 陈璟(通信作者), 副教授、博士。

收稿日期: 2020-12-22 修回日期: 2021-02-08 E-mail: chenjing@jiangnan.edu.cn

## 0 概述

近年来,随着高通量技术的发展,通过实验方法检测到蛋白质相互作用(Protein-Protein Interaction, PPI)的数量大幅增加,形成了越来越多的PPI网络。发现并理解蛋白质之间的相互作用,是生物学领域内的重要课题之一<sup>[1]</sup>。对PPI网络的分析能够增进对生物学过程的理解,不同物种间相互作用组的比对在蛋白质功能预测、保守功能成分检测、物种间知识转移等方面具有重要意义<sup>[2]</sup>。

网络比对按映射关系可分成局部比对和全局比对。局部比对旨在找到高度保守的小型网络区域,并生成多对多节点映射,例如LePrimAlign<sup>[3]</sup>、LocalAli<sup>[4]</sup>、Pin-Align<sup>[5]</sup>。全局比对旨在找到较大的保守区域并生成一对一的节点映射,例如Isorank<sup>[6]</sup>、GRAAL<sup>[7]</sup>、H-GRAAL<sup>[8]</sup>、MI-GRAAL<sup>[9]</sup>、C-GRAAL<sup>[10]</sup>、L-GRAAL<sup>[11]</sup>、NETAL<sup>[12]</sup>、SPINAL<sup>[13]</sup>、PINALOG<sup>[14]</sup>、ModuleAlign<sup>[15]</sup>、PROPER<sup>[16]</sup>、AligNet<sup>[17]</sup>、IsoRankN<sup>[18]</sup>、BEAMS<sup>[19]</sup>、SMETANA<sup>[20]</sup>等,其中后3种算法属于多网络比对,其余为2个网络的比对。本文着重研究2个网络的全局比对。

在现有2个网络的全局比对算法中,IsoRank算法是先进的全局比对算法,其主要利用类似谷歌页面排序算法(PageRank)的方法计算节点相似性,并采用贪心算法进行比对。GRAAL系列算法利用图形度标签相似性作为节点的拓扑相似度,结合其他搜索算法比对PPI网络。NETAL算法基于拓扑和生物相似性构建相似性矩阵,利用贪心搜索方法比对网络。SPINAL算法首先基于二分图构建初始相似性矩阵,利用种子-扩展算法并基于迭代交换局部改进比对节点。PINALOG算法首先利用聚类算法提取密集模块,计算模块间的相似度并通过模块匹配提取比对的种子集合,并将比对扩展至种子节点的邻域。ModuleAlign算法基于输入网络的层次聚类计算蛋白质间的同源性得分,整合了序列信息以及局部和全局网络拓扑作为评分方案,通过迭代算法寻找一种在实现较好整体得分的同时最大化保守相互作用数量的比对方法。PROPER算法首先根据序列信息设置阈值筛选种子集合,然后利用渗透图匹配(Percolation Graph Matching, PGM)算法<sup>[21]</sup>扩展种子。AligNet算法是成对网络比对算法,其先将网络划分成多个重叠簇,再进行簇间的局部比对,最后将其组合扩展为全局比对。

上述算法大多从拓扑和生物2个角度考虑蛋白质的相似性,由于PPI网络存在噪声,仅考虑拓扑特征可能对比对产生误导,而序列信息的不完全性使得仅考虑生物信息会影响比对的准确性,序列上相似不一定代表功能相似。此外,生物网络被观察到是高度模块化的<sup>[22]</sup>,同一个功能模块中蛋白质相互

连接密集,不同模块间蛋白质相互连接稀疏<sup>[23]</sup>,部分算法利用层次聚类<sup>[24]</sup>、密度聚类<sup>[25]</sup>等模块检测技术从PPI网络中提取出具有相似功能的蛋白质,将功能模块检测融入网络比对中,更准确地预测蛋白质功能,但通常需要借助同源信息。因此,利用较少的额外信息实现良好的拓扑与生物一致性的平衡,是需要进一步研究的问题。

本文基于种子-扩展算法和功能模块检测,提出一种生物网络全局比对算法JAlign。结合节点及其邻居节点的拓扑特征和序列信息构建目标函数,利用Jerarca层次聚类算法<sup>[26]</sup>划分功能模块,并通过匈牙利算法从中提取种子,综合种子与邻居节点的连接关系、节点相似性和度,将比对从种子节点扩展至其邻居节点,同时对剩余节点再次进行加权二分图匹配,找到更多匹配节点,从而实现拓扑和生物一致性的平衡。

## 1 基于层次聚类的生物网络全局比对

### 1.1 问题描述

2个PPI网络分别用无向图 $G_1=(V_1, E_1)$ 和 $G_2=(V_2, E_2)$ 表示,其中: $V_1, V_2$ 表示网络中的节点集合; $E_1, E_2$ 表示网络中的边集合;节点表示蛋白质;边表示2个蛋白质间的相互作用; $N(i)$ 表示节点 $i$ 的邻居节点集合; $N(j)$ 表示节点 $j$ 的邻居节点集合。全局比对 $f: V_1 \rightarrow V_2$ ,是将 $G_1$ 中的 $V_1$ 节点映射到 $G_2$ 的 $V_2$ 节点上,形成一对一的映射关系,令 $f(V_1)=\{f(v) \in V_2 | v \in V_1\}$ , $f(E_1)=\{(f(u), f(v)) \in E_2 | (u, v) \in E_1\}$ 。全局网络比对的目的是找到比对节点对相似性得分之和最大的映射关系。

### 1.2 JAlign算法

本文提出JAlign算法,利用功能模块检测和种子-扩展算法实现全局比对。该算法主要分为3个阶段,分别是种子筛选阶段、扩展阶段和局部优化阶段。在种子筛选阶段,先基于拓扑和序列信息构建目标函数,再利用层次聚类算法划分功能模块并进行模块间的比对,从模块对中提取高相似度节点对作为种子节点。在扩展阶段,先根据种子节点计算其邻居节点的相似性,再迭代比对高得分的节点对,逐步将比对扩展至所有可比对的节点。在局部优化阶段,先对剩余节点构建二分图,再利用最大加权匹配比对剩余节点,将比对节点对合并到比对集合。

#### 1.2.1 种子筛选阶段

种子筛选阶段包括目标函数构建、功能模块检测及种子筛选2个部分。

##### 1) 目标函数构建

目标函数用于衡量节点间的相似性,是后续比对的重要依据,结合序列信息与拓扑特征可避免仅考虑生物信息或拓扑信息对比对结果产生的误导。节点 $i$

和节点 $j$ 的序列相似性根据 BLAST bit-score 值<sup>[27]</sup>计算, 如式(1)所示:

$$B(i, j) = \frac{b_{\text{blast}}(i, j) - \min_{u \in V_1, v \in V_2} b_{\text{blast}}(u, v)}{\min_{u \in V_1, v \in V_2} b_{\text{blast}}(u, v) - \min_{u \in V_1, v \in V_2} b_{\text{blast}}(u, v)} \quad (1)$$

其中:  $b_{\text{blast}}(i, j)$  表示节点  $i, j$  之间的 BLAST bit-score 值。

若 2 个节点的邻居节点拓扑相似, 则这 2 个节点拓扑相似的可能性更高, 通过同时考虑邻居节点的拓扑相似性和节点本身的相似度计算节点对的拓扑相似性, 能够更全面地衡量节点间的拓扑特征。计算节点  $i, j$  拓扑相似性得分  $T(i, j)$  的过程如下: 首先初始化  $T^0(i, j) = 1$ ; 然后构建二分图  $G_B = (V_B, E_B)$ , 其中  $V_B$  由  $N(i)$  节点和  $N(j)$  节点的 2 个不相交集组成,  $E_B$  中的边  $(i', j')$  由  $N(i), N(j)$  中节点所有可能的连接组成,  $i' \in N(i), j' \in N(j)$ , 边的权重  $w(i', j') = T^{i_i}(i', j')$ ; 最后利用贪心算法, 每次选中权重最大的边添加到匹配集合, 遍历完所有边后得到  $G_B$  的匹配集合  $M$ , 计算该匹配  $M$  对应的  $T^{i_i+1}(i, j)$  值。  $T^{i_i+1}$  计算公式如式(2)所示:

$$T^{i_i+1}(i, j) = \theta \times \frac{\sum_{(u, v) \in M} T^{i_i}(u, v)}{\max\{|N(i)|, |N(j)|\}} + (1 - \theta) \times \frac{\min\{d(i), d(j)\}}{\max_{u \in V_1 \cup V_2} \{d(u)\}} \quad (2)$$

其中:  $|N(i)|, |N(j)|$  表示节点  $i, j$  的邻居节点数;  $d(i), d(j)$  表示节点  $i, j$  的度;  $\max_{u \in V_1 \cup V_2} \{d(u)\}$  表示  $G_1, G_2$  中节点度的最大值;  $i_i$  是迭代次数;  $\theta$  是平衡邻居节点和节点本身拓扑相似性比重的参数,  $0 \leq \theta \leq 1$ 。经过多次迭代后, 矩阵  $T$  的最终值为节点的拓扑相似性。

在计算序列和拓扑相似性之后, 可得到节点的相似性得分, 如式(3)所示:

$$S(i, j) = \alpha \times B(i, j) + (1 - \alpha) \times T(i, j) \quad (3)$$

其中:  $\alpha$  是平衡拓扑和序列权重的参数,  $0 \leq \alpha \leq 1$ ;  $B(i, j)$  和  $T(i, j)$  分别是序列得分和拓扑得分。

## 2) 功能模块检测及种子筛选

生物网络的模块化特征使得具有相似生物功能的蛋白质连通密集, 功能模块检测将功能相似的节点划分为同一模块。从功能模块中筛选种子, 将功能信息融入到比对中, 相较于在整个网络内筛选种子节点, 缩小了种子节点的选择范围, 通过对高质量的种子进行扩展, 提高了比对结果质量。种子筛选阶段的算法流程如图 1 所示。首先利用 Jerarca 聚类算法划分输入网络中的功能模块, 构成模块集合, 结合匈牙利算法先进行模块内比对, 计算模块间相似性; 然后模块间比对, 得到模块间的最佳匹配结果; 最后从中筛选出高相似性的节点对作为种子节点, 其中 Jerarca 聚类算法通过节点间距离划分模块, 使得在聚类时不会遗漏一些常规的不完全连通的功能

模块, 也不需要 GO 文件辅助划分模块, 从而减少了输入信息。

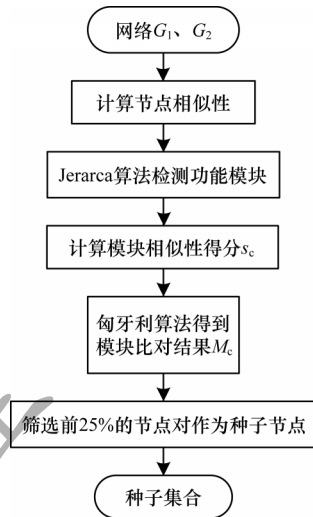


图 1 种子筛选流程

Fig.1 Procedure of seed selection

种子筛选阶段伪代码算法 1 所示。

## 算法 1 SeedSelect 算法

输入 2 个 PPI 网络  $G_1, G_2$ , 参数  $\theta, \alpha$

输出 种子筛选结果 seed

1. for all  $u \in V_1$  and  $v \in V_2$  do
2. 按式(3)计算节点相似性  $S(u, v)$
3.  $C_1, C_2 \leftarrow$  Jerarca 算法划分  $G_1, G_2$  的功能模块
4. for all  $c_i \in C_1$  and  $c_j \in C_2$  do
5. 按式(4)结合匈牙利算法计算模块相似性  $s_c(c_i, c_j)$
6.  $M_c \leftarrow$  匈牙利算法进行  $C_1, C_2$  内匹配
7. seed  $\leftarrow$  筛选前 25% 节点对
8. return seed

种子筛选阶段的具体过程如下:

- 1) 根据式(3)计算节点间的相似性得分  $S(u, v)$ 。
- 2) 利用 Jerarca 聚类算法检测  $G_1, G_2$  中的功能模块, 并将模块信息分别存入集合  $C_1, C_2$  中。

3) 计算模块间的相似性得分(算法 1 第 4 行和第 5 行)。利用匈牙利算法进行模块内比对, 得到模块内节点间的最佳映射关系  $M_H$ , 计算比对模块的相似性得分  $s_c$ , 如式(4)所示:

$$s_c(c_i, c_j) = \sum_{(u, v) \in M_H} S(u, v) \quad (4)$$

其中:  $c_i, c_j$  分别表示  $C_1, C_2$  中第  $i, j$  个模块;  $M_H$  是通过匈牙利算法得到的  $c_i, c_j$  内节点的比对集合。

4) 根据模块间的相似性得分  $s_c$ , 采用匈牙利算法得到模块间的匹配结果  $M_c$ , 依据相似性得分值分布的第 3、4 分位数, 筛选出前 25% 的节点对作为种子继续扩展(算法 1 第 6 行和第 7 行)。

## 1.2.2 扩展阶段

种子筛选阶段能够将部分重要节点比对上, 但仍存在多数节点未涉及。为了将比对从种子节点扩展至整个网络, 本文提出一种新的扩展方法, 基本思



路是某一对节点的邻居节点相似性越高,则这对节点比对上的概率越高,即比对上的邻居节点数越多,该节点对越可能被比对上。以种子节点为中心,遍历其邻居节点,将比对逐步从种子节点扩展至周围节点,但仅依靠节点间的连接关系确定比对节点也不全面,会存在多对节点满足扩展条件,而每步节点对的选择都会对后续节点的比对产生影响,从多角度考虑可以得出较为全面的结果。本阶段通过综合考虑节点间的连接关系、度特征、节点相似性,更全面、谨慎地确定扩展节点,产生更优的比对结果。

扩展阶段伪代码如算法2所示。

#### 算法2 Extension 算法

输入  $G_1, G_2$ , 种子集合  $seed$ , 节点相似性  $S$

输出 扩展比对结果  $M$

```

1.  $M \leftarrow seed$ 
2. for all  $(u, v) \in seed$  do
3. 按式(5)计算邻居节点  $score_{stru}(u', v')$ 
4.  $seed = \emptyset$ 
5. while 存在  $sim_{stru}(u', v') > 0$  do
6.  $N \leftarrow sim_{stru}$  得分最高的节点对
7. if  $|N| > 1$  then
8.  $N \leftarrow \{(i, j) | (i, j) \in N \text{ and } |d_i - d_j| = \min \{ |d_i - d_j| \} \}$ 
9. if  $|N| > 1$  then
10.  $seed = seed \cup \{(i, j) | (i, j) \in N \text{ and } S(i, j) = \max \{ S(i, j) \} \}$ 
11. else
12.  $seed = seed \cup N$ 
13. else
14.  $seed = seed \cup N$ 
15.  $M = M \cup seed$ 
16. return  $M$ 

```

扩展阶段的具体过程如下:

1) 将种子节点对添加到比对集合  $M$  (算法2第1行)。

2) 计算种子的邻居节点间的结构相似性(算法2第2行和第3行)。遍历种子节点的所有邻居节点,计算每一对邻居节点在已比对节点集合中的公共邻居节点数作为邻居节点的结构相似性得分  $s_{stru}$ , 如式(5)所示:

$$s_{stru}(u, v) = |\{(u', v') | (u', v') \in M, (u, u') \in E_1, (v, v') \in E_2\}| \quad (5)$$

3) 综合邻居节点结构相似性  $s_{stru}$ 、度、节点相似性  $S$  选择扩展节点对(算法2第5~15行)。首先对所有结构相似性从高到低排序,将得分最高的节点对添加到集合  $N$ ,若  $N$  中存在多个得分最高的节点对,计算节点间的度差值,保留集合  $N$  中度差值最小的节点对,若  $N$  中存在多个最小度差值节点对,则根据式(3)中相似性得分  $S$  选择节点对,将相似性得分最高的节点对添加到比对集合中。

4) 每有一对新的节点比对成功,更新其邻居节点的相似性得分(算法2第2行),重复上述过程,直至没有可选择的节点对。

### 1.2.3 局部优化阶段

扩展阶段只是将比对从种子节点扩展到节点的邻域,这对于种子节点的选择非常重要,而模块检测是从网络中提取出连通密集模块并从中筛选出种子,因此,选择的种子节点多数是在连通密集的子图中的,对于较为离散或连通度不高的子图,比对上的概率很低,而这一部分子图中可能有部分节点是重要节点。为了解决这一问题,局部优化阶段对剩余节点进行局部优化比对,通过二分图的加权匹配比对剩余节点,使得孤立节点也能参与比对,将比对覆盖至整个网络。具体过程如下:

- 1) 查找出  $G_1, G_2$  网络中未比对上的节点,构建二分图  $G_b$ ,所有边的权重为该节点对的相似性得分  $S$ 。
- 2) 选择权重最大的边合并到比对集合  $M$  中。
- 3) 删除步骤2中选中的节点对及其相关的边。
- 4) 重复步骤2、3,直至图中没有边存在。

如图2所示,先选中权重最大的边  $(a, d)$ ,若  $a, d$  节点均未在比对集中出现过,则将节点对  $(a, d)$  添加到比对集合中,删除  $a, d$  节点及其相关的边,再从剩余边中选择权重最大的边,选择的边是  $(b, e)$ ,确定  $b, e$  节点未出现在现有比对集合中,将  $(b, e)$  添加到比对集合,删除  $b, e$  节点及其相关的边,此时,  $G_b$  中无可匹配节点对存在,比对结束,将节点对  $(a, d)$ 、 $(b, e)$  合并至比对集合,得到最终比对结果。

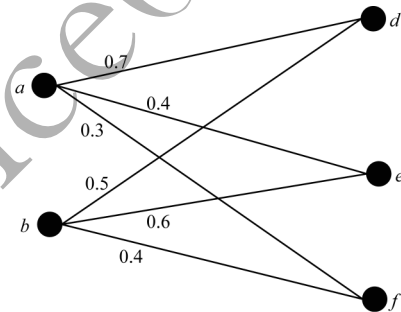


图2 剩余节点匹配示例

Fig.2 Example of the matching for remaining nodes

### 1.3 时间复杂度分析

令  $n = \max \{|V_1|, |V_2|\}$ ,  $m = \max \{|E_1|, |E_2|\}$ 。第一阶段是种子筛选,计算节点相似性的时间复杂度为  $O(n^2)$ ,聚类算法划分模块的时间复杂度为  $O(n^2 \log_a n)$ ,匈牙利算法的时间复杂度是  $O(n^2)$ ,因此,第一阶段的时间复杂度为  $O(n^2 \log_a n)$ ;第二阶段是扩展比对,种子节点对最多有  $n$  对,每个种子节点的邻居节点最多有  $m$  个,因此,复杂度为  $O(nm^2)$ ;第三阶段是剩余节点比对,剩余节点最多有  $n$  个,因此,第三阶段的时间复杂度为  $O(n^2)$ 。至此,本文算法总的时间复杂度为  $O(n^2 \log_a n + nm^2)$ 。在生物网络普遍性假设中,  $m \approx n \log_a n$ ,所以,本文算法的时间复杂度为  $O(nm^2) = O(n^3 \log_a n)$ 。

2 实验与结果分析

2.1 实验数据

实验所用数据来自 Isobase<sup>[28]</sup>数据库的真实网络数据和 NAPAbench<sup>[29]</sup>合成网络数据,真实网络数据包括 Caenorhabditis Elegans(CE)、Saccharomyces Cerevisiae(SC)、Drosophila Melanogaster(DM)、Homo Sapiens(HS)等 4 个物种,合成网络包括 Crystal Growth(CG)、Duplication Mutation Completion(DMC)、Duplication with Random Mutation(DMR)。生物网络的节点和边信息见表 1,其中真实网络的数据均经过预处理,排除了自循环、重复的边和节点。

表 1 网络数据信息

Table 1 Network data information

网络	物种	节点数	边数
真实网络	CE	2 974	4 827
	SC	5 523	82 656
	DM	7 387	24 937
	HS	10 296	54 654
CG	AA	3 000	11 987
	BB	4 000	15 987
DMC	AA	3 000	6 090
	BB	4 000	8 112
DMR	AA	3 000	6 017
	BB	4 000	8 238

2.2 评价指标

对于网络比对算法的性能,从拓扑质量和生物质量 2 个方面进行评估。

1) 拓扑质量度量

边正确性(Edge Correctness, EC)<sup>[7]</sup>是衡量网络比对拓扑质量最常用的指标,通过计算  $f$  映射下保守边在源网络中的比例来评估比对的质量,不能惩罚将稀疏网络映射到密集网络的比对。EC 计算公式如式(6)所示:

$$E_c(f)=\frac{|f(E_1)|}{|E_1|}$$
 (6)

其中:  $|E_1|$  表示  $G_1$  网络的边数;  $|f(E_1)|$  表示以  $f$  映射方式覆盖到  $G_2$  中的边的边数。

诱导保守子结构得分(Induced Con-served-structure Score, ICS)<sup>[30]</sup>计算保守边与诱导边的比例,克服了 EC 的问题,但不能惩罚将密集网络映射到稀疏网络的比对。ICS 计算公式如式(7)所示:

$$I_{cs}(f)=\frac{|f(E_1)|}{|E_{G_2[f(V_1)]}|}$$
 (7)

其中:  $|E_{G_2[f(V_1)]}|$  表示  $G_2$  的诱导子网络的边数,子网络包含  $G_2$  中所有比对上的节点;  $|f(E_1)|$  表示以  $f$  映射方式覆盖  $G_2$  中的边的边数。

对称子结构得分(Symmetric Sub-structure Score,  $S^3$ )<sup>[31]</sup>针对源网络和目标网络,既能惩罚稀疏到密集的比对又能惩罚密集到稀疏的比对。 $S^3$  的计算公式如式(8)所示:

$$S^3(f)=\frac{|f(E_1)|}{|E_1|+|E_{G_2[f(V_1)]}|-|f(E_1)|}$$
 (8)

其中:分母表示根据比对  $f$  将  $G_1$  和  $G_2$  诱导子图重叠得到复合图计算图中唯一边的数目;  $|E_1|$  表示  $G_1$  网络中的边数。

2) 生物质量度量

生物质量使用功能一致性(Functional Coherence, FC)<sup>[32]</sup>来衡量,FC 利用 GO 术语测量比对蛋白质的功能一致性。GO 术语描述蛋白质的生物特性,包括分子功能(Molecular Function, MF)、细胞成分(Cellular Component, CC)、生物过程(Biological Process, BP)<sup>[33]</sup>。通常认为具有相似 GO 术语的蛋白质其生物功能相似。FC 计算公式如式(9)和式(10)所示:

$$F_s(u,f(u))=\frac{|G_o(u)\cap G_o(f(u))|}{|G_o(u)\cup G_o(f(u))|}$$
 (9)

$$F_c(f)=\frac{\sum_{u\in V_1}F_s(u,f(u))}{|V_1|}$$
 (10)

其中:  $G_o(u)$  和  $G_o(f(u))$  表示节点  $u$  和  $f(u)$  被注释的 GO 集合。

2.3 结果分析

本文将 JAlign 算法与 L-GRAAL、SPINAL、ModuleAlign 算法进行比较,在参数设置上,拓扑相似性计算中  $\theta$  设置为 0.5,迭代 2 次,总相似性中  $\alpha$  设置为 0.4。Jerarca 聚类中选择 RCluster<sup>[26]</sup>迭代算法,迭代 3 次,层次树构建使用 UPGMA<sup>[26]</sup>算法。其他算法的参数设置参照原文设置,具体参数设置如表 2 所示。

表 2 3 种对比算法的参数设置

Table 2 Parameters setting of three alignment algorithms

算法	命令行	参数
L-GRAAL	无	无
SPINAL	-mode II -alpha $\alpha$	$\alpha=0.7$
ModuleAlign	-alpha $\alpha$	$\alpha=0.4$

2.3.1 合成网络比对结果分析

4 种算法在合成网络上的比对结果如表 3 所示,其中,加粗的数据为最佳结果,加下划线的数据为次优结果,下同。在拓扑指标上, JAlign 算法的结果明显优于其他 3 种算法;在生物指标 FC 上, JAlign 算法仅次于 SPINAL 算法, L-GRAAL 算法表现最差。总体而言, JAlign 算法在合成网络上表现最好,在保证了较好的生物特性的基础上实现了最好的拓扑质量。

表 3 合成网络比对结果

Table 3 Alignment result for synthesis networks

网络	算法	EC	ICS	S <sup>3</sup>	FC
CG	L-GRAAL	0.530	0.710	0.530	0.470
	SPINAL	<u>0.810</u>	<u>0.820</u>	<u>0.690</u>	<b>0.730</b>
	ModuleAlign	0.680	0.680	0.520	0.630
	JAlign	<b>0.820</b>	<b>0.840</b>	<b>0.710</b>	<u>0.710</u>
DMC	L-GRAAL	0.534	0.533	0.364	0.125
	SPINAL	<u>0.592</u>	<u>0.548</u>	<u>0.397</u>	<b>0.610</b>
	ModuleAlign	0.426	0.466	0.286	0.134
	JAlign	<b>0.607</b>	<b>0.609</b>	<b>0.437</b>	<u>0.298</u>
DMR	L-GRAAL	0.681	0.647	0.497	0.259
	SPINAL	<b>0.821</b>	<u>0.719</u>	<b>0.622</b>	<b>0.732</b>
	ModuleAlign	0.640	0.620	0.460	0.355
	JAlign	<u>0.776</u>	<b>0.736</b>	<u>0.607</u>	<u>0.407</u>

2.3.2 真实网络比对结果分析

在真实网络上,分别从EC、ICS、S<sup>3</sup>、FC这4个方面对比4种算法,4种算法在EC指标上的结果如表4所示。实验结果表明:在CE-SC、CE-HS、DM-HS中,ModuleAlign算法表现最好;在SC-HS、SC-DM中,JAlign算法表现最好。

表 4 不同算法在不同物种上的 EC 指标

Table 4 EC index of different algorithms for different species

实验组	L-GRAAL	SPINAL	ModuleAlign	JAlign
CE-SC	0.545	0.643	<b>0.765</b>	<u>0.671</u>
CE-DM	<b>0.609</b>	0.489	<u>0.583</u>	0.535
CE-HS	0.563	<u>0.612</u>	<b>0.657</b>	0.562
SC-HS	<u>0.119</u>	0.105	0.117	<b>0.152</b>
SC-DM	0.082	0.081	<u>0.094</u>	<b>0.100</b>
DM-HS	0.319	0.286	<b>0.368</b>	<u>0.340</u>

表5是不同算法在ICS指标上的对比结果。实验结果表明:L-GRAAL算法在CE-SC、CE-DM、CE-HS、DM-HS实验中结果最好;JAlign算法在这4组实验中表现仅次于L-GRAAL算法,在SC-HS、SC-DM实验中结果最好;SPINAL在各组实验中结果最差。

表 5 不同算法在不同物种上的 ICS 指标

Table 5 ICS index of different algorithms for different species

实验组	L-GRAAL	SPINAL	ModuleAlign	JAlign
CE-SC	<b>0.091</b>	0.052	0.066	<u>0.089</u>
CE-DM	<b>0.342</b>	0.163	0.202	<u>0.278</u>
CE-HS	<b>0.250</b>	0.110	0.133	<u>0.229</u>
SC-HS	0.227	0.201	<u>0.236</u>	<b>0.287</b>
SC-DM	0.305	0.298	<u>0.350</u>	<b>0.371</b>
DM-HS	<b>0.205</b>	0.147	0.188	<u>0.197</u>

表6是不同算法在S<sup>3</sup>指标上的对比结果,S<sup>3</sup>是从源网络和目标网络2个方面度量拓扑特征,能够更全面地分析算法的拓扑质量。实验结果表明:JAlign

算法在网络规模较大的物种上,相较于其他算法表现最好,也较为稳定;L-GRAAL算法仅次于JAlign,且在小网络上结果较好。总结上述结果,JAlign算法在拓扑质量上可以得到稳定、最好的实验结果,L-GRAAL算法整体表现仅次于JAlign算法。

表 6 不同算法在不同物种上的 S<sup>3</sup> 指标

Table 6 S<sup>3</sup> index of different algorithms for different species

实验组	L-GRAAL	SPINAL	ModuleAlign	JAlign
CE-SC	0.847	0.050	0.064	<b>0.850</b>
CE-DM	<b>0.280</b>	0.139	0.176	<u>0.224</u>
CE-HS	<b>0.210</b>	0.103	0.124	<u>0.194</u>
SC-HS	0.085	0.074	0.085	<b>0.110</b>
SC-DM	<u>0.069</u>	0.068	0.080	<b>0.085</b>
DM-HS	0.142	0.108	0.142	<b>0.143</b>

表7是不同算法在FC指标上的对比结果。从生物指标角度来看,SPINAL算法和JAlign算法结果相近,L-GRAAL算法表现最差。

表 7 不同算法在不同物种上的 FC 指标

Table 7 FC index of different algorithms for different species

实验组	L-GRAAL	SPINAL	ModuleAlign	JAlign
CE-SC	0.011	<u>0.031</u>	0.022	<b>0.033</b>
CE-DM	0.005	<b>0.028</b>	0.008	<u>0.027</u>
CE-HS	0.012	<u>0.028</u>	0.016	<b>0.029</b>
SC-HS	0.012	<u>0.039</u>	<b>0.045</b>	0.038
SC-DM	0.011	0.026	<b>0.033</b>	<u>0.026</u>
DM-HS	0.008	<b>0.025</b>	0.012	<u>0.023</u>

表8给出了不同算法运行时间的比较,实验计算机使用Inter Core i7-10510U 2.30 GHz处理器,内存16 GB,算法在Linux Ubuntu环境下运行。实验结果表明,在时间效率上JAlign算法也具有一定优势,SPINAL、ModuleAlign算法运行时间较长。

表 8 不同算法的运行时间

Table 8 Running time of different algorithms

算法	运行时间
L-GRAAL	60 min 50.336 s
SPINAL	121 min 50.840 s
ModuleAlign	207 min 46.732 s
JAlign	52 min 28.320 s

结合拓扑质量度量结果分析,虽然SPINAL算法在生物指标上表现很好,但在拓扑指标上表现较差,而L-GRAAL算法则与之相反,其在牺牲一定生物质量的基础上实现了较好的拓扑质量,4种算法中只有JAlign算法同时实现了最佳的拓扑度量和生物度量。进一步分析实验过程可知,JAlign算法在目标函数中添加了拓扑特征并降低了序列特征的比重,在聚类算法中也仅依靠节点的结构信息来划分模块,但随着序列信息所占比例的降低,其在生物指



标上表现仍优于大部分算法,这说明 JAlign 算法可以实现生物一致性和拓扑特性的良好平衡。

### 3 结束语

本文提出的生物网络全局比对算法 JAlign。基于种子-扩展算法,利用层次聚类算法检测功能模块并进行模块匹配,根据目标函数从模块中筛选种子节点。在此基础上,结合邻居节点与种子节点间的连接关系将比对扩展至种子节点的邻居节点,并对剩余节点构建二分图进行最大加权匹配,使得孤立节点也有机会被比对上。实验结果表明,该算法能够全面地考虑比对过程,实现拓扑质量与生物质量的良好平衡,相较于其他3种对比算法,其在拓扑和生物角度同时达到了最优的比对结果。为优化 JAlign 算法的效率,进一步提高算法在生物功能方面的性能,下一步将对功能模块检测部分做并行化处理,并将模块检测应用到目标函数计算和种子的扩展阶段。

### 参考文献

- [1] 祝家焯. 基于局部优化与二分图匹配的PPI网络比对算法[J]. 计算机应用与软件, 2018, 35(1): 281-287.  
ZHU J Y. PPI network comparison algorithm based on local optimization and bipartite graph matching[J]. Computer Applications and Software, 2018, 35(1): 281-287. (in Chinese)
- [2] MENG L, STRIEGEL A, MILENKOVIĆ T. Local versus global biological network alignment[J]. Bioinformatics, 2016, 32(20): 3155-3164.
- [3] MASKEY S, CHO Y R. LePrimAlign: local entropy-based alignment of PPI networks to predict conserved modules[J]. BMC Genomics, 2019, 20(9): 964-976.
- [4] HU J, REINERT K. LocalAli: an evolutionary-based local alignment approach to identify functionally conserved modules in multiple networks[J]. Bioinformatics, 2015, 31(3): 363-372.
- [5] AMIR-GHIASVAND F, NOWZARI-DALINIA, MOMENZADEH V. Pin-Align: a new dynamic programming approach to align protein-protein interaction networks[J]. Computational and Mathematical Methods in Medicine, 2014, 2014: 1-5.
- [6] SINGH R, XU J, BERGER B. Global alignment of multiple protein interaction networks with application to functional orthology detection[J]. Proceedings of the National Academy of Sciences, 2008, 105(35): 12763-12768.
- [7] KUCHAIEV O, MILENKOVIĆ T, MEMIŠEVIĆ V, et al. Topological network alignment uncovers biological function and phylogeny[J]. Journal of the Royal Society Interface, 2010, 7: 1341-1354.
- [8] MILENKOVIĆ T, NG W L, HAYES W, et al. Optimal network alignment with graphlet degree vectors[J]. Cancer Informatics, 2010, 9(8): 121-137.
- [9] KUCHAIEV O, PRŽULJ N. Integrative network alignment reveals large regions of global network similarity in yeast and human[J]. Bioinformatics, 2011, 27(10): 1390-1396.
- [10] MEMIŠEVIĆ V, PRŽULJ N. C-GRAAL: common-neighbors-based global graph alignment of biological networks[J]. Integrative Biology, 2012, 4(7): 734-743.
- [11] MALOD-DOGNIN N, PRŽULJ N. L-GRAAL: lagrangian graphlet-based network aligner[J]. Bioinformatics, 2015, 31(13): 2182-2189.
- [12] NEYSHABUR B, KHADEM A, HASHEMIFAR S, et al. NETAL: a new graph-based method for global alignment of protein-protein interaction networks[J]. Bioinformatics, 2013, 29(13): 1654-1662.
- [13] ALADAČ A E, ERTEN C. SPINAL: scalable protein interaction network alignment[J]. Bioinformatics, 2013, 29(7): 917-924.
- [14] PHAN H T T, STERNBERG M J E. PINALOG: a novel approach to align protein interaction networks—implications for complex detection and function prediction[J]. Bioinformatics, 2012, 28(9): 1239-1245.
- [15] HASHEMIFAR S, MA J, NAVEED H, et al. ModuleAlign: module-based global alignment of protein-protein interaction networks[J]. Bioinformatics, 2016, 32(17): 658-664.
- [16] KAZEMI E, HASSANI H, GROSSGLAUSER M, et al. PROPER: global protein interaction network alignment through percolation matching[J]. BMC Bioinformatics, 2016, 17(1): 527-543.
- [17] ALCALÁ A, ALBERICH R, LLABRÉS M, et al. AligNet: alignment of protein-protein interaction networks[J]. BMC Bioinformatics, 2020, 21(6): 1-22.
- [18] LIAO C S, LU K, BAYM M, et al. IsoRankN: spectral methods for global alignment of multiple protein networks[J]. Bioinformatics, 2009, 25(12): 253-258.
- [19] ALKAN F, ERTEN C. BEAMS: backbone extraction and merge strategy for the global many-to-many alignment of multiple PPI networks[J]. Bioinformatics, 2014, 30(4): 531-539.
- [20] SAHRAEIAN S M E, YOON B J. SMETANA: accurate and scalable algorithm for probabilistic alignment of large-scale biological networks[J]. PloS One, 2013, 8(7): 1-5.
- [21] KAZEMI E, HASSANI S H, GROSSGLAUSER M. Growing a graph matching from a handful of seeds[J]. Proceedings of the VLDB Endowment, 2015, 8(10): 1010-1021.
- [22] TRIPATHI B, PARTHASARATHY S, SINHA H, et al. Adapting community detection algorithms for disease module identification in heterogeneous biological networks[J]. Frontiers in Genetics, 2019, 10: 164-181.
- [23] 冀俊忠, 刘志军, 刘红欣, 等. 蛋白质相互作用网络功能模块检测的研究综述[J]. 自动化学报, 2014, 40(4): 577-593.
- [24] JI J Z, LIU Z J, LIU H X, et al. Research review on functional module detection of protein interaction network[J]. Acta Automata Sinica, 2014, 40(4): 577-593. (in Chinese)
- [25] LI C, BAI J, ZHAO W, et al. Community detection using hierarchical clustering based on edge-weighted similarity in cloud environment[J]. Information Processing & Management, 2019, 56(1): 91-109.
- [26] BRYANT A, CIOK K. RNN-DBSCAN: A density-based clustering algorithm using reverse nearest neighbor density estimates[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(6): 1109-1121.

(上接第 71 页)

- [26] ALDECOA R, MARÍN I. Jerarca: efficient analysis of complex networks using hierarchical clustering[J]. PloS One, 2010, 5(7): 11585-11592.
- [27] ALTSCHUL S F, GISH W, MILLER W, et al. Basic local alignment search tool[J]. Journal of Molecular Biology, 1990, 215(3): 403-410.
- [28] PARK D, SINGH R, BAYM M, et al. IsoBase: a database of functionally related proteins across PPI networks[J]. Nucleic Acids Research, 2010, 39(s1): 295-300.
- [29] SAHRAEIAN S M E, YOON B J. A network synthesis model for generating protein interaction network families[J]. PloS One, 2012, 7(8): 1-5.
- [30] PATRO R, KINGSFORD C. Global network alignment using multiscale spectral signatures[J]. Bioinformatics, 2012, 28(23): 3105-3114.
- [31] SARAPH V, MILENKOVIĆ T. MAGNA: maximizing accuracy in global network alignment[J]. Bioinformatics, 2014, 30(20): 2931-2940.
- [32] HUANG J, GONG M, MA L. A global network alignment method using discrete particle swarm optimization[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2016, 15(3): 705-718.
- [33] ZHU Y, LI Y, LIU J, et al. Discovering large conserved functional components in global network alignment by graph matching[J]. BMC Genomics, 2018, 19(7): 41-58.

编辑 金胡考