蛋白质相互作用网络的蜂群信息流聚类模型与算法

雷秀娟 田建芳
(陕西师范大学计算机科学学院 西安 710062)

摘 要 蛋白质相互作用网络的聚类算法研究是充分理解分子的结构、功能及识别蛋白质的功能模块的重要方法.很多传统聚类算法对于蛋白质相互作用网络聚类效果不佳.功能流模拟算法是一种新型聚类算法,但该算法没有考虑到距离的作用效果并且需要人为地设置合并阈值,带有主观性.文中提出了一种新颖的基于蜂群优化机理的信息流聚类模型与算法.该方法中,数据预处理采用结点网络综合特征值的排序来初始化聚类中心,将蜂群算法的蜜源位置对应于其聚类中心,蜜源的收益度大小对应于模块间的相似度,采蜜蜂结点的所有邻接点按照结点网络综合特征值的降序排列,作为侦察蜂的搜索邻域.采用正确率、查全率等指标对聚类效果做出客观评价,并对算法的一些关键参数进行仿真、对比与分析.结果表明新算法不仅克服了原功能流模拟算法的缺点,且其正确率和查全率的几何平均值最高,能够有效地识别蛋白质功能模块.

关键词 信息流;蜂群算法;聚类;蛋白质相互作用网络
中图法分类号 TP301 **DOI号**: 10.3724/SP.J.1016.2012.00134

The Information Flow Clustering Model and Algorithm Based on the Artificial Bee Colony Mechanism of PPI Network

LEI Xiu-Juan Tian Jian-Fang
(*College of Computer Science, Shaanxi Normal University, Xi'an* 710062)

Abstract The clustering algorithm of Protein-Protein Interaction (PPI) networks is an important method to fully understand the organizations and functions of molecules and identify the functional modules of protein. There are lots of traditional clustering algorithms which do not perform well in clustering PPI networks. Recently functional flow simulation algorithm is a novel clustering algorithm. However, it does not take the effect of distance into account and the merging threshold is set manually which is subjective. This paper proposes a novel information flow clustering model and algorithm based on the mechanism of Artificial Bee Colony (ABC) optimization. This method firstly sorts the network comprehensive feature value of nodes to initialize the cluster centers during the procedure of data pre-processing. The nectar source of ABC algorithm is corresponding to cluster center, the income level of nectar stands for the similarity between modules. Afterwards all the adjacent nodes of employed bee node are sorted in the descending order according to the network comprehensive feature value of nodes, which are regarded as the searching neighborhood of scouts. In the end, the algorithm adopts precision, recall and other criteria to evaluate the cluster effect in an objective way. In addition, some significant parameters of the algorithm is simulated, compared and analyzed. The experiment results show that the new algorithm not only overcomes the shortcomings of original algorithm, but also the harmonic mean value of precision and recall gets greatly improved, which can effectively identify the functional modules of protein.

收稿日期:2011-01-04;最终修改稿收到日期:2011-08-24.本课题得到国家自然科学基金(61100164)、陕西省自然科学基础研究计划项目(2010JQ8034)、中央高校基本科研业务费专项资金(GK200902016)资助.**雷秀娟**,1975年生,女,博士,副教授,中国计算机学会(CCF)会员,主要研究方向为智能计算、生物信息计算.E-mail: xjlei168@163.com. **田建芳**,1986年生,女,硕士研究生,主要研究方向为数据挖掘、生物信息计算.

Keywords　information flow；Artificial Bee Colony (ABC) algorithm；clustering；Protein-Protein Interaction (PPI) network

1　引　言

随着人类基因组计划的完成，生命科学研究步入后基因时代. 很长时间以来，研究基因的功能都是针对单个基因来进行的，其思路是“序列→结构→功能”，相对于功能基因组这一研究目标来说，这种“一次一个基因”的研究模式在效率上已经完全不能适应要求，生物的功能一般都是通过一批基因的相互作用形成网络而得以发挥的，因此采用“相互作用→网络→功能”的新思路，整合基因和蛋白质的不同方面、不同层次的信息进行基因功能分析，已经成为当前功能基因组研究的新方向.

研究蛋白质间相互作用网络(Protein-Protein Interaction Network，PPI)的一个主要目标是识别与分析细胞环境中生物分子的相互作用，以便深入地理解生物分子相互作用与执行功能的机制. 细胞中的各种生命活动与蛋白质间的相互作用密切相关，因此深入理解蛋白质相互作用是揭示生命活动奥秘的前提.

蛋白质相互作用网络与其它网络诸如万维网、人类社会网络一样，通常具有无尺度特性和小世界特性[1-5]，每种蛋白质有多个相互作用因子，且具有时间和空间差异性，因此蛋白质相互作用的研究更复杂、更具难度. 蛋白质相互作用网络可以被表示为一个无向图，其中每个结点表示一个蛋白质，每条边表示一对蛋白质结点之间的相互作用. 已有研究表明，蛋白质并不是功能分开而是以功能模块的形式相互作用的，进而形成蛋白质相互作用网络[6-8].

蛋白质相互作用网络的研究可以帮助我们预测功能未知的蛋白质功能，了解特定生物功能的分子机制，为探讨重大疾病的机理、疾病治疗、疾病预防和新药开发提供重要的理论基础. 从大量的蛋白质相互作用网络数据中提取出功能模块，即蛋白质相互作用网络的聚类研究，是生物体行为理解、蛋白质功能预测和药物设计的基础. 而且研究蛋白网络的方法具有一定的普适意义，在蛋白网络中被成功应用的算法可以应用到当前网络科学研究其它领域，如互联网、人际关系网和生物代谢网等，处理类似的问题，对其它领域的研究有一定的借鉴意义.

蛋白质相互作用网络属于复杂网络，其数据量十分庞大，而且蛋白质相互作用网络通常由大量高密度的蛋白质结点构成，其中也不缺乏稀疏连接结点，其形态各异，没有规则可循，具有小世界、无尺度特性，这就决定了聚类是在蛋白质相互作用网络中进行数据挖掘的首要工具.

根据适合蛋白质相互作用网络的聚类算法的特点，分为传统方法和新型方法. 传统方法有：划分的方法(partitioning method)、基于密度的方法(density-based method)、基于网格的方法(grid-based method)、基于模型的方法(model-based method)、层次的方法(hierarchical method)、模糊聚类方法(fuzzy clustering method)[9]等. 新型方法有：谱聚类方法、信息流模拟聚类方法、整体聚类方法等[10].

划分的方法是最基本的方法，但其最大缺点就是要事先知道要划分的目标类的确定个数，且大都根据对象之间的距离进行聚类，利于发现紧凑的簇，对于发现任意形状的簇遇到了困难. 基于密度的方法不能对存在大量稀疏结点的网络进行分类. 层次聚类的方法的缺陷是它对噪声数据的鲁棒性比较差. 谱聚类方法虽然能在任意形状的样本空间上聚类，并可以处理大规模稀疏数据的聚类问题，但邻域矩阵的选取始终是一难题. 整体聚类方法的目标就是把多个独立的不同聚类融合成为单一的全面聚类[11]，从而提高对无尺度网络聚类的质量. 但此方法缺乏全局目标函数，在每一步局部地确定需要合并的聚类；此外，该方法的时空复杂度较高. 综上所述，以上方法有的需要事先知道聚类的数目，有的适应性差，对大型蛋白质相互作用网络都不能取得良好的聚类效果.

功能流模拟聚类方法可以被称为信息传递[12-16]的聚类方法，每个结点(蛋白质)就像一个“蓄水池”，它容纳了每次迭代时由此结点流向它的邻接结点的流量，而每一条边的权值代表了它的流量限度. 在确定的时间内对功能流的流向分布进行仿真，从而得到与流量相对应的每个蛋白质的“功能得分”. 因而可以确定在这个网络中，一个蛋白质在功能上能对其它蛋白质产生多少影响. 因需要考虑全部结点的信息流传递，该方法的时间复杂度较高，但此方法的思想符合蛋白质的实际作用效果，易于理解实现. 但

如何自动获得最佳相似度阈值来最终确定聚类数仍没得到解决,目前文献中的参数是人工设定的[15-16].

2 基本概念和原理

2.1 蛋白质相互作用网络的特性

众所周知,蛋白质相互作用网络之所以复杂是因为每种蛋白质有多个相互作用因子,且具有时间、空间差异性,呈现出小世界、无尺度特性. Watts 和 Strogatz[17]最先提出了小世界网络模型,它是介于规则网络和完全随机网络之间的一类模型,也就是说小世界网络既有大的聚集系数又有小的平均距离,这种网路与现实网络更为接近. Barabasi 和 Albert[1]指出真实网络的一个重要统计特征是:网络结点的度服从幂分布 $P(t)=t^{-k}$,t 是结点的度,k 是一个与 t 无关的常数,一般取(2,3)之间的随机数,具有这种特性的网络称为无尺度网络. 而所谓无尺度是指所研究的对象与尺度无关,即无论测量的单位如何改变,研究的对象的性质如形态、复杂程度等都不会发生变化. 直观上看,无尺度网络具有这样的特征:只有少数结点的度较大,而其它大部分结点的度较小,而这些少数的度很大的结点在调节其它大量的结点之间的相互作用中起着重要作用.

2.2 蜂群算法

蜂群算法 ABC[18](Artificial Bees Colony algorithm)是一种基于蜜蜂行为的优化算法,是建立在蜜蜂自组织模型和群体智能基础上的一种非数值计算优化方法. Seely[19]最先提出了蜂群的自组织模拟模型,Teodorvic[20]提出了蜂群优化算法,Karaboga 等[21]提出了人工蜂群算法. ABC 算法实现采蜜的集体智能行为需要 4 部分:蜜源、采蜜蜂、跟随蜂和侦察蜂. 一个蜜源对应一只采蜜蜂,一个蜜源的位置代表优化问题的一个可能的解. 蜜源值取决于多种因素,诸如蜜源与蜂巢的接近程度,蜜源的集中程度等. 这里以"收益度"来衡量蜜源的特点. 采蜜蜂有保持优良蜜源的作用,具有经营特性,采蜜蜂与具体的蜜源联系在一起,这些蜜源是它们当前正在采集的蜜源,采蜜蜂携带了很多与收益度有关的蜜源信息. 跟随蜂搜索蜂巢附近的新蜜源,跟随蜂增加较好蜜源对应的蜜蜂数量,侦查蜂在全局范围内搜索新的蜜源.

首先,ABC 算法按照一定规则生成 M(种群规模)个代表蜜源的初始种群,每个解 $\boldsymbol{pop}(i)$ 是一个 clu_num(聚类个数)维的向量. 采蜜蜂先对其蜜源做一次邻域搜索,并选择花蜜量较多的也就是收益度高的蜜源. 所有的采蜜蜂完成搜索之后,然后跳摇摆舞与跟随蜂共享蜜源信息. 跟随蜂根据蜜源信息量的大小以一定选择机制选择蜜源,信息量大的采蜜蜂吸引跟随蜂的概率大于信息量小的采蜜蜂. 跟随蜂在蜜源附近进行邻域搜索,采用贪婪准则,比较跟随蜂搜索解与原采蜜蜂搜索解,当搜索到的新解优于原采蜜蜂的解时,替换原采蜜蜂的解,完成角色转换,反之,不做改变. 侦察蜂则在邻域范围内随机搜索蜜源,扩大解的多样性,有利于算法跳出局部最优.

2.3 基本功能流聚类模型与算法[22]

前已述及,蛋白质相互作用网络通常被表示成一个无向图,这样网络的初始输入简单、计算量小,但聚类结果不尽如人意. 功能流算法给每条边赋予权值,表示通过该边的流量上限. 这样可以真实模拟蛋白质的实际作用效果. 网络流的概念类似于图的切割,功能流模型利用网络流的思想,每一个已知的注释蛋白质被认为是功能流的"蓄水池",利用交互网络的边作为导管,访问没有被标注的结点. 通过简单的局部规则来访问结点.

Nabieva[14]等人首先提出了蛋白质相互作用网络的功能流模型,他们认为每一个结点就像一个"蓄水池",它容纳了每次迭代时由此结点流向它的邻接结点的流量,每一条边上的流量限度表明了在一次迭代过程中能够通过该边的流量的最大限度. 在确定的时间内对功能流的流向分布进行仿真,从而得到与流量相对应的每个蛋白质的"功能得分",并把它存储在"蓄水池"中. 但是这个模型并没有考虑到距离的作用效果,也就是说每一个已标记的蛋白质和跟它有着不同距离的其它蛋白质之间的相互作用被忽略掉了.

Cho[15-16]在蛋白质相互作用网络中提出了另外一种基于流的模块化方法来识别重叠的功能模块. 该方法要求输入一个带权重的相互作用网络,该功能流模型实现需要 3 个步骤:选择信息量大的结点;运用流模拟产生初始的模块;根据模块相似度合并改善模块结果.

第 1 步. 根据蛋白质的加权度 $dw(i)$ 选择信息量大的蛋白质,将这些蛋白质进行标记,成为注释蛋白质.

第 2 步. 流模拟方法开始于每一个注释的蛋白质. 流模拟是基于功能信息的概念的,即蛋白质 s 在这个带权重的相互作用网络中流过的所有可能的路径. 因此我们可以量化蛋白质 s 对网络中其它蛋白质的影响程度.

流 $f_s(x\rightarrow y)$代表了蛋白质 s 对流经蛋白质 x 到 y 所产生的影响程度，而 $inf_s(y)$代表蛋白质 s 对蛋白质 y 所产生的影响量. 算法首先对每一个注解蛋白质 s 分配权重 $dw(s)$，并把它作为最初的流量 $inf_s(s)$，同时给其它没有携带信息的蛋白质全都赋值为 0. 对于每一个携带信息的蛋白质 s 来说，它最初的流量是 $inf_s(s)$，在 s 向它的邻接结点 y 传递流量的过程中，$inf_s(s)$的值会随着对应边的权重的值的改变而逐渐减小. 因此 s 的初始流 f_s的流量定义如下：

$$f_s(x\rightarrow y)=w_e(s,y)\cdot inf_s(s) \tag{1}$$

对每一个蛋白质 $y\in N(s)$，$N(s)$表示蛋白质 s 的邻接结点并且满足条件 $0\leq w_e(s,y)\leq 1$ 的结点的集合，s 在蛋白质 y 上的流量 $inf_s(y)$，可以通过计算 y 的所有的邻接结点流入 y 的流量 f_s的和来进行更新：

$$inf_s(y)=\sum_{x\in N(y)} f_s(x\rightarrow y) \tag{2}$$

在初始流的条件下，$f_s(x\rightarrow y)$是等于 $inf_s(s)$的，蛋白质 s 对于 y 的影响将继续遍历图中所有与 y 连接的边，定义如下：

$$f_s(y\rightarrow z)=w_e(y,z)\cdot\frac{inf_s(y)}{|N(y)|} \tag{3}$$

其中 $y,z\in E$，并且满足条件 $0\leq w_e(y,z)\leq 1$. 算法通过式(2)反复地将所有流入 s 的结点的流量加起来，再通过式(3)把流量传递到所有与之相互连接的结点. 随着加权度 dw 的变化，从一个蛋白质到另一个蛋白质的作用值逐渐较小. 如果加权度等于 0，作用值快速减小，相反，如果蛋白质 i 和 j 完全可靠，例如 $w_e(i,j)=1$，则 $f_s(i)$完整地传递给蛋白质 j.

该算法中所有结点的作用值都是在流模拟过程中在上一次的基础上累积的，蛋白质 s 的作用值的积累是一个主要因素，它决定蛋白质 i 和 s 属于同一个功能模块的可能性有多大. 因此流模拟通过路径访问所有结点，连接稠密的结点相对于连接稀疏的结点来讲，离注释结点比较近，能够得到较大的作用值. 当某条路径上结点的作用值小于最小阈值时停止，访问其他路径. 流模拟算法开始于注释结点 s 直到没有流流向其他结点时结束.

第 3 步，根据模块之间的相似度合并相似度大的模块得到最终的模块. 2 个模块 M_s和 M_t之间的相似度 $S(M_s,M_t)$可以通过加权相互连接性的计算得到，定义如下：

$$S(M_s,M_t)=\frac{\sum_{x\in M_s,y\in M_t} c(x,y)}{\max(|M_s|,|M_t|)} \tag{4}$$

其中，

$$c(x,y)=\begin{cases}1, & x=y\\ w_e(x,y), & x\neq y\ \text{且}\langle x,y\rangle\in E\\ 0, & \text{其它}\end{cases} \tag{5}$$

根据式(4)，重复合并具有最大相似度的 2 个模块，直到最大相似度低于某个人工设定的合并阈值. 文献[16]使用的是人为设定的阈值，显然计算结果带有主观性.

2.4　结点网络综合特征值

结点的度和加权度反映结点间的连接强度，而结点的网络综合特征值[23]既反映了结点间的连接强度、结点间距离的作用效果，又反映了结点局部范围内的相互连接密度与强度.

结点 v_i的度为与结点 v_i连接的边的总数，定义如下：

$$d_i=|\{v_j\mid(v_i,v_j)\in E,\quad v_i,v_j\in V\}| \tag{6}$$

结点 v_i的聚集度 dk_i为与结点 v_i连接的近邻结点之间的连接的边的个数，定义如下：

$$dk_i=|\{(v_j,v_k)|(v_i,v_j)\in E,(v_i,v_k)\in E,(v_j,v_k)\in E,\ v_i,v_j,v_k\in V\}| \tag{7}$$

结点 v_i的加权度 dw_i为与结点 v_i连接的边的权值之和，定义如下：

$$dw_i=\sum_{(v_i,v_j)\in E} w_{ij} \tag{8}$$

结点 v_i的加权聚集度为：

$$wk_i=\sum_{(v_i,v_j)\in E} w_{jk} \tag{9}$$

其中，k 取所有结点 v_i的聚集度的集合 dk_j中的点.

结点 v_i的聚集系数为：

$$wc_i=\frac{2\times wk_i}{d_i\times(d_i-1)} \tag{10}$$

结点的加权网络综合特征值 $com\text{-}value_i$ 定义如下：

$$com\text{-}value_i=\alpha\times wc_i+(1-\alpha)\times dw_i/N \tag{11}$$

其中，N 是网络中结点的总个数，$\alpha\in(0,1)$. 结点网络综合特征值 $com\text{-}value$ 既反映了结点与其它结点的连接强度，结点间距离的作用效果又反映了结点局部范围内的相互连接密度与强度，因此结点 $com\text{-}value$ 值高的结点含有较丰富的结点信息.

3　基于蜂群优化搜索的信息流聚类模型

基本功能流聚类模型中用结点的加权度 dw 确

定聚类中心并且将结点的 dw 作为注释蛋白质初始功能流，但是结点的 dw 只反映了结点与其它结点的连接强度，没有考虑结点之间距离的作用效果，含有的结点信息比较少. 因此用结点的 dw 确定聚类中心不够合理，而且每次对一个结点的邻接点分流时都是按照“先来先服务的”顺序进行的，这样有可能导致含有丰富结点信息的结点由于被访问得晚，而将其错误地划分到其它类中. 此外，基本功能流聚类模型还需要人为地设置合并阈值进行合并，带有主观因素，也会影响聚类效果. 根据人工蜂群算法的优良特性，能够有效地改善算法的性能. 因此本文提出了一种新的基于蜂群优化搜索的信息流聚类模型. 由于结点的加权网络综合特征 *com-value* 含有丰富的结点信息，因而该方法中用结点的 *com-value* 确定聚类中心，将 *com-value* 作为注释蛋白质的初始信息流，按照 *com-value* 的降序遍历结点的邻接点，降低了将结点错误分类的概率，此外该方法中的合并阈值是参照文献[24]得到的最佳阈值，而不是人为设定的，消除了主观因素，并且该方法中聚类的合并是在蜂群优化搜索过程中进行的，将模块相似度高的聚类合并，再由侦查蜂搜索新的模块相似度高的聚类.

3.1 聚类中心及蜜源的确定

基本功能流模型中，计算每个结点的 dw，并对 dw 从大到小排序，取前 *clu_num* 个 dw 对应的结点作为聚类中心. 本文根据结点的 *com-value* 确定聚类中心：对结点的 *com-value* 按照降序排列，取前 *clu_num* 个 *com-value* 对应的结点作为聚类中心. 并将聚类中心的结点按照任意的顺序排列作为一个蜜源.

3.2 解空间设计及蜜蜂的采蜜流程

初始时刻，所有蜜蜂都没有任何先验知识，根据结点的 *com-value* 初始化蜜源，$\boldsymbol{pop}(i,j)$，$i=1,\cdots,M$，$j=1,\cdots,clu\_num$，M 代表种群规模，*clu_num* 代表聚类个数，每个蜜源都代表一组聚类中心，计算适应度函数，将适应度函数按照降序排列. 前 $M/2$ 适应度所对应的蜜蜂为采蜜蜂，剩下 $M/2$ 适应度所对应的蜜蜂为跟随蜂. 由于蛋白质相互作用网络的特性，要尽可能地遍历结点信息丰富的结点，所以将采蜜蜂的邻接点的集合作为跟随蜂进行搜索的邻域，由于结点的邻接点是有限的，所以跟随蜂在每个蜜源附近的搜索也是有限的. 跟随蜂完成搜索之后，根据轮盘赌选择策略选择原采蜜蜂邻接点中的结点作为新的蜜源，进行邻域搜索. 当跟随蜂连续 *limit* 次搜索不到新解时，将相似度大的模块进行合并，再由侦查蜂进行全局搜索，更新蜜源.

为了更好地理解基于蜂群优化搜索的信息流聚类模型，表 1 中给出了该模型与蜂群算法中蜜蜂采蜜行为的对应关系.

表 1 蛋白质相互作用网络聚类问题与蜜蜂采蜜行为的对应关系

蜂群采蜜行为	PPI 网络聚类问题
蜜源位置	一组聚类中心
蜜源的收益度大小	模块间的相似度
跟随蜂，侦察蜂职责	搜索还没有被访问的结点
高收益度食物源	模块之间相似度最低的一组聚类

3.3 基于蜂群优化搜索的信息流聚类模型

该算法中首先计算每个结点的 *com-value*，记录每个结点的所有邻接点并将其邻接点按照 *com-value* 的降序进行排列，从结点的 *com-value* 值高的结点中选取前 *clu_num* 个结点作为聚类中心即注解蛋白质，对每一个注解蛋白质 s 分配特征值 $com\text{-}value_e(s)$，并把它作为最初的流量 $inf_s(s)$，同时给其它没有携带信息的蛋白质全都赋值为 0. 对于每一个注解蛋白质 s 来说，它最初的流量是 $inf_s(s)$，在 s 向它的邻接结点 y 传递流量的过程中，$inf_s(s)$ 的值会随着对应边的权重的值的改变而逐渐减小. 因此 s 的初始流 f_s 的流量定义同式(1). 对每一个蛋白质 $y\in N(s)$，$N(s)$ 表示蛋白质 s 的邻接结点并且满足条件 $0\leq w_e(s,y)\leq 1$ 的结点的集合，s 在蛋白质 y 上的流量 $inf_s(y)$，可以通过计算 y 的所有的邻接结点流入 y 的流量 f_s 的和来进行更新，同式(2).

对聚类中心的这 *clu_num* 个结点按照任意的顺序排列作为 ABC 算法的一个蜜源 $\boldsymbol{pop}(i,j)$，$i=1,\cdots,M$，$j=1,\cdots,clu\_num$，每个蜜源代表一组聚类中心，计算适应度，根据适应度确定采蜜蜂和跟随蜂，采蜜蜂个数等于跟随蜂个数，跟随蜂按照轮盘赌选择策略将采蜜蜂结点的邻接点集合作为搜索邻域进行搜索产生新解 $onlooker\_pop(i,j)$，依据式(3)对跟随蜂对应的结点进行流量传递，根据最小流量阈值，判断该结点是否归入该采蜜蜂所对应的类中，并计算其适应度，更新最优解. 当跟随蜂搜索不到新的结点时，计算该跟随蜂所在的类与同一蜜源中其它类之间的相似度 *Sim*，若相似度最大值 $Sim\_max > merge\_thred$，则合并相似度高的模块，由侦查蜂进行全局搜索，产生新的蜜源即新的注释蛋白质，重复以上过程. 这里 *merge_thred* 是参照文献[25]得到的最佳合并阈值.

将适应度函数最小值即收益度最高的一组聚类

作为初始聚类，计算每个聚类的结点个数，将结点个数很少的一类认为是孤立结点，从聚类结果中删除，得到最终聚类结果.

3.4 蜂群算法的目标函数

我们用式(12)来描述基于蜂群优化搜索的信息流聚类算法的目标函数

$$fval=\frac{\sum_{s=1}^{clu\_num}\sum_{t=1}^{clu\_num}\boldsymbol{S}(M_s,M_t)}{clu\_num\times clu\_num} \tag{12}$$

其中，$\boldsymbol{S}(M_s,M_t)$为相似度矩阵，根据式(4)、(5)进行计算，clu_num 为聚类个数.

4 实验仿真及分析

4.1 蛋白质相互作用数据库

多年来人们发现了很多的蛋白质相互作用，并且创建了一些数据库来存储蛋白质之间的联系，例如慕尼黑蛋白质序列信息中心 MIPS(The Munich Information Center for Protein Sequence)[25]、蛋白质相互作用数据库 DIP(The Database of Interacting Proteins)[26]、生物分子相互作用网络数据库 BIND(The Biomolecular Interaction Network Database)[27]、IntAct[28]包含了诸如实验方法、实验环境和相互作用范围之类的数据和分子相互作用数据库 MINT(A Molecular INTeraction Database)[29].

4.2 聚类结果评价

实验中用到的蛋白质相互作用数据集来自 MIPS[30]数据库，该数据集提供了关于开放阅读框架、RNA 基因和其它遗传因素的信息. 该数据集包含两组数据：第 1 组数据给出了蛋白质结点以及结点之间的相互作用关系，作为实验数据；第 2 组数据中给出了属于同一类的功能模块相似的蛋白质，作为标准数据库. 我们将 MIPS 的标准数据库作为基准来评价我们采用的方法对蛋白质功能及复杂性的预测.

本文采用正确率、查全率及它们的几何平均值来进行算法聚类效果评价. 正确率($precision$)[12]是：实验结果与标准数据库中最大匹配的结点的个数与实验结果中结点个数的比值. 查全率($recall$)[12]是实验结果与标准数据库中最大匹配的结点的个数与标准数据库中结点个数的比值. 计算公式如下

$$precision(C|F)=\frac{MMS(C,F)}{|C|} \tag{13}$$

$$recall(C|F)=\frac{MMS(C,F)}{|F|} \tag{14}$$

其中：C 代表实验得到的聚类结果；F 标准数据库的结果；$|C|$代表测试得到的聚类的结点个数；$|F|$表示标准数据库中聚类结点的个数；MMS 表示实验结果与标准数据库中最大匹配的结点的个数.

通常情况下，较大的模块具有较高的查全率，因为一个较大的模块 C 很可能包含 F 中许多结点，极端的情况下，所有的结点都聚到一个模块中，查全率将达到最大，相反，较小的模块具有较高的正确率，因为较小模块的这些结点很可能具有同样的性质，极端的情况是每一个结点都是一个模块，这些模块具有最大的正确率，因此我们可以用正确率和查全率的几何平均值 $f\_measure$[12]来评价模块的准确性.

$$f\_measure=\frac{2\cdot precision\cdot recall}{precision+recall} \tag{15}$$

4.3 算法实现步骤

算法具体步骤如下：

1. 设置聚类个数 clu_num，种群规模 M，最大迭代次数 $maxiter$，是否丢弃解的参数 $limit$，跟随蜂搜索不到新解的次数 $Bas(i,j)$，$i=1,\cdots,M$，$j=1,\cdots,clu\_num$，流量阈值 $flow\_thred$，结点的流量 $flowamount(i)$，$i=1,2,\cdots,c$，结点个数 c，$size\_thred$ 模块结点是否为孤立点的阈值；

2. 将结点的 $com\text{-}value$ 值最大的前 clu_num 个结点作为聚类中心，对聚类中心的这 clu_num 个结点随机排序产生一个蜜源，$\boldsymbol{pop}(i,j)$，$i=1,2,\cdots,M$，$j=1,2,\cdots,clu\_num$，一个蜜源代表一组聚类中心，一个蜜源对应一只蜜蜂；

3. 分别将每个蜜源中的 clu_num 个结点作为注释蛋白质按照式(1)，(2)对其进行分流. 计算每个蜜源的适应度函数 $fval(i)$，$i=1,2,\cdots,M$，将适应度函数从小到大进行排序，前 $M/2$ 个适应度所对应的蜜蜂作为采蜜蜂 $employ\_pop$，后 $M/2$ 个适应度所对应的蜜蜂作为跟随蜂 $onlooker\_pop$；

4. $iter=1$，控制算法迭代次数；

5. 跟随蜂根据轮盘赌选择策略进行搜索，并对结点 $onlooker_pop(i,j)$按照式(3)进行分流，如果 $flowamount(onlooker\_pop(i,j))>flow\_thred$，则将该跟随蜂所对应的结点归入相应的类中；

6. 记录当前的全局最优值 $gbest\_fval$；

7. 由于蛋白质网络的特性，结点的邻接点个数是有限的，也就是跟随蜂能够搜索的邻域范围是有限的，当跟随蜂搜索不到新解时，$Bas(i,j)=Bas(i,j)+1$；

8. 如果 $Bas(i,j)\geq maxiter$，计算 $employ_pop(i,j)$所对应的类与其它类的相似度 Sim，若 $Sim\_max>merge\_thred$，合并这 2 个类. 再由侦查蜂进行全局搜索，搜索一个未被访问的结点作为新的蜜源，$Bas(i,j)=0$；

9. $iter=iter+1$，如果 $iter\leq maxiter$，返回步 5，否则聚类结束，生成初始聚类结果；

10. 当模块结点个数 $size\_node(i)<size\_thred$，认为是

孤立点,从聚类结果中剔除,然后输出聚类结果.

4.4 实验结果分析

4.4.1 各算法性能比较

结点的度和加权度只反映了结点间的连接强度,没有考虑结点间距离的影响,基本功能流算法中,用结点的 dw 作为结点的初始流量,进行聚类.而结点的网络综合特征值含有比度和加权度更丰富的结点信息.因此在流算法中将结点的网络综合特征值作为结点的初始流量进行聚类,使聚类效果得到显著改善,而且用网络综合特征值确定聚类中心的聚类效果与用加权度和度确定聚类中心的聚类效果相比,前者极大地改善了算法的性能.

表 2 中将各算法的缩写所代表的算法的实际意义进行了具体描述.

表 2 算法缩写与其实际意义的对应关系

算法缩写	算法表示的实际意义
degree_degree_flow	用结点的度确定聚类中心,且度作为注释蛋白质初始功能流的基本功能流聚类算法
com-value_degree_flow	用结点的网络综合特征值确定聚类中心,度作为注释蛋白质初始功能流的功能流聚类算法
dw_dw_flow	用结点的加权度确定聚类中心,加权度作为注释蛋白质初始功能流的功能流聚类算法
com-value_dw_flow	用结点的网络综合特征值确定聚类中心,加权度作为注释蛋白质初始功能流的功能流聚类算法
com-value_com-value_flow	用结点的网络综合特征值确定聚类中心,且网络综合特征值作为注释蛋白质初始功能流的功能流聚类算法
basic_degree_flow	结点的度作为注释蛋白质初始功能流的基本功能流聚类算法
ABC_degree_flow	结点的度作为注释蛋白质初始信息流的基于蜂群优化搜索的信息流聚类算法
basic_dw_flow	结点的加权度作为注释蛋白质初始功能流的基本功能流聚类算法
ABC_dw_flow	结点的加权度作为注释蛋白质初始信息流的基于蜂群优化搜索的信息流聚类算法
basic_com-value_flow	结点的网络综合特征值作为注释蛋白质初始功能流的基本功能流聚类算法
ABC_com-value_flow	结点的网络综合特征值作为注释蛋白质初始信息流的基于蜂群优化搜索的信息流聚类算法

表 3 初始聚类个数为 20,该表中的结果是运行 10 次的平均值.该表中可以看出,用结点的度和加权度确定聚类中心的基本功能流聚类算法的聚类效果很差,这是因为结点的度和加权度只反映结点与其它结点的连接强度,含有的结点信息比较少,所以用结点的度和加权度确定聚类中心是不合理的.而结点的网络综合特征值既反映了结点与其它结点的连接强度、结点间距离的作用效果,又反映了结点局部范围内的连接密度与强度,结点信息非常丰富,因此用结点的网络综合特征值确定聚类中心的基本功能流聚类算法聚类效果得到显著提高.

表 3 基本功能流算法聚类结果比较

算 法	正确率	查全率	几何平均	聚类个数	运行时间/s
degree_degree_flow	0.0626	0.6501	0.1142	16.5	89.2
com-value_degree_flow	0.4524	0.5489	0.4960	16.0	52.6
dw_dw_flow	0.0688	0.5740	0.1229	15.6	87.2
com-value_dw_flow	0.5137	0.6174	0.5608	14.7	60.2
com-value_com-value_flow	0.4005	0.6706	0.5015	13.4	60.8

表 4 中结果是在 $merge\_thred=0.3$, $limit=10$, $\alpha=0.5$, $size\_thred=2$, $maxiter=70$ 的情况下运行 10 次的平均值.表 4 中对基本功能流聚类算法和基于蜂群优化的信息流聚类算法分别在将结点的度、加权度和网络综合特征值作为注释蛋白质初始信息流的在不同的初始聚类个数情况下的聚类结果做了比较.表中可以看出将结点的度和加权度作为注释蛋白初始功能流的基本功能流聚类算法聚类效果很不理想,当聚类个数逐渐增大的时候,聚类效果逐渐变好,但是聚类效果依然很差.这是因为由于度和加权度含有的结点信息比较少,用其确定的聚类中心不够合理,并且合并阈值是人为设定的,带有一定的主观性,对聚类结果的影响很大.用结点的网络综合特征值作为注释蛋白质初始功能流的基本功能流聚类算法效果比较好,并且随着聚类个数的增加,聚类效果持续的得到改进,当聚类个数为 100 时聚类效果最好,因为结点网络综合特征值结点信息比较丰富,用其确定的聚类中心比较合理,并且虽然人为设定的合并阈值具有一定的主观性,但是根据其将相似度大的聚类进行合并,也在一定程度上改善了聚类结果.分别将结点度、加权度和网络综合特征值作为注释蛋白质初始信息流的基于蜂群优化搜索的信息流聚类算法聚类效果非常好,因为该方法中,遍历结点的邻接点是按照结点信息量大小的顺序进行的,并且在蜂群算法搜索过程中就将模块相似度高的模块进行了合并,提高了算法的聚类效果.对于度和加权度的基于蜂群优化搜索的信息流聚类算法聚类个数为 60 时聚类效果最好,而对于用结点网络综合特征值确定注释蛋白质初始信息流的基于蜂群优化搜索的信息流聚类算法聚类个数为 100 时聚类效果最好,但是由于聚类个数的增大,增大了问题的规模,具有高的模块相似度的聚类个数增大,需要合并的聚类个数增加,而每合并一个聚类,就需要侦查

蜂进行一次全局搜索，侦查蜂进行全局搜索的次数也增大，因此时间复杂度变得很大.

表 4　不同的初始聚类个数对聚类结果的影响

算　法	初始聚类个数	正确率	查全率	几何平均	聚类个数	运行时间/s
Basic_degree_flow	20	0.0626	0.6501	0.1142	16.5	89.2
ABC_degree_flow	20	**0.4888**	**0.7221**	**0.5830**	**15.3**	**261.4**
Basic_dw_flow	20	0.0688	0.5740	0.1229	15.6	87.2
ABC_dw_flow	20	**0.7414**	**0.5956**	**0.6604**	**16.0**	**226.2**
Basic_com-value_flow	20	0.4005	0.6706	0.5015	13.4	60.8
ABC_com-value_flow	20	**0.7295**	**0.7745**	**0.7513**	**11.7**	**268.5**
Basic_degree_flow	40	0.1160	0.7340	0.2003	15.7	210.2
ABC_degree_flow	40	**0.4712**	**0.7625**	**0.5825**	**16.7**	**969.5**
Basic_dw_flow	40	0.1012	0.7498	0.1783	15.3	208.7
ABC_dw_flow	40	**0.6784**	**0.7066**	**0.6922**	**16.0**	**984.2**
Basic_com-value_flow	40	0.4275	0.7075	0.5330	14.7	156.8
ABC_com-value_flow	40	**0.6726**	**0.7703**	**0.7181**	**27.0**	**996.8**
Basic_degree_flow	60	0.2069	0.9198	0.3378	35.7	330.3
ABC_degree_flow	60	**0.4594**	**0.7600**	**0.5726**	**18.0**	**1018.9**
Basic_dw_flow	60	0.1285	0.8091	0.2218	45.0	332.1
ABC_dw_flow	60	**0.7481**	**0.7333**	**0.7406**	**16.0**	**1036.7**
Basic_com-value_flow	60	0.5153	0.7686	0.6170	15.0	275.2
ABC_com-value_flow	60	**0.7191**	**0.8028**	**0.7585**	**19.0**	**2011.0**
Basic_degree_flow	80	0.1703	0.8748	0.2851	66.3	465.8
ABC_degree_flow	80	**0.6814**	**0.7133**	**0.6970**	**24.0**	**1302.1**
Basic_dw_flow	80	0.1802	0.8426	0.2969	61.7	471.5
ABC_dw_flow	80	**0.6375**	**0.8390**	**0.7245**	**24.0**	**1359.9**
Basic_com-value_flow	80	0.5494	0.8249	0.6595	16.3	409.7
ABC_com-value_flow	80	**0.6951**	**0.8352**	**0.7588**	**31.5**	**11390.8**
Basic_degree_flow	100	0.1886	0.8627	0.3095	73.0	623.6
ABC_degree_flow	100	**0.5726**	**0.6560**	**0.6115**	**31.0**	**1642.3**
Basic_dw_flow	100	0.1982	0.8613	0.3222	63.7	627.3
ABC_dw_flow	100	**0.6981**	**0.6925**	**0.6953**	**35.7**	**1603.0**
Basic_com-value_flow	100	0.5919	0.8057	0.6824	21.0	536.5
ABC_com-value_flow	100	**0.7270**	**0.8387**	**0.7789**	**36.7**	**14073.3**

4.4.2　参数分析

图 1 是基本功能流聚类算法中将结点的度、加权度和网络综合特征值分别作为注释蛋白质初始功能流的聚类结果比较. 图 1 中可以看出，在结点的度，加权度和网络综合特征值的基本功能流聚类算法中，因为结点的度只反映结点间的连接强度，含有的结点信息最少，因此结点的度对应的基本功能流聚类算法聚类效果最差，结点的加权度既反映了结点的连接强度又在一定程度上反映了结点间距离的作用效果，因此结点的加权度对应的基本功能流聚类算法的聚类效果比结点的度对应的聚类效果好. 结点的网络综合特征值反映了结点之间的连接强度，结点之间距离的作用效果还反映了结点局部范围内的连接密度，结点信息非常丰富，如图 1 所示，结点网络综合特征值对应的基本功能流聚类算法聚类效果最好.

图 2 是基于蜂群优化搜索的信息流聚类算法分别将结点的度、加权度和网络综合特征值作为注释蛋白质初始信息流的聚类结果比较. 图 2 中用结点的度作为注释蛋白质初始信息流的基于蜂群优化搜索的信息流聚类算法聚类效果最差，结点的加权度对应的算法聚类效果次之，结点网络综合特征值对应的算法聚类效果最好. 由于人工蜂群算法优良的寻优特性，能够快速有效地找到最优解，因此基于蜂群优化搜索的信息流聚类算法的性能得到了极大改善. 从图 1 与图 2 可以看出，与基本功能流聚类算法相比，基于蜂群优化搜索的信息流聚类算法在结点的度、加权度和网络综合特征值上聚类效果都得到了显著的提高.

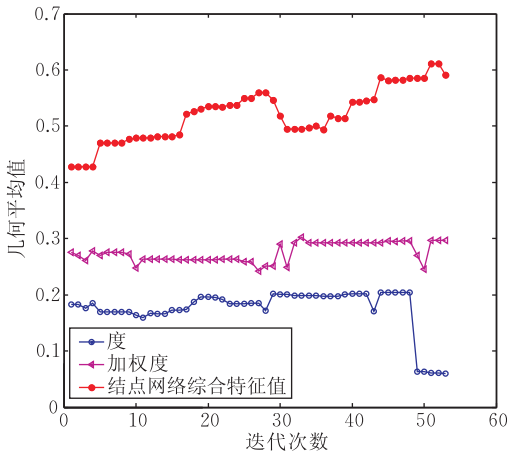

图 1　基本功能流聚类算法的聚类结果比较

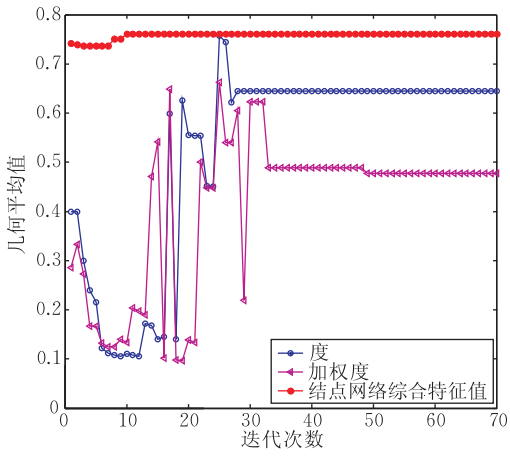

图 2　基于蜂群优化搜索的信息流聚类算法的聚类结果比较

图 3 中参数 *merge_thred* 是算法在迭代过程中，合并模块相似度高的阈值便于侦查蜂搜索新蜜源的一个合并阈值，可以看出 *merge_thred* 对于结点的度和加权度的基于蜂群优化搜索的信息流聚类算法的聚类效果影响很大，对于结点的度，合并阈值等于 0.4 时聚类效果最好；对于结点的加权度，合并阈值等于 0.2 时聚类效果最好，当合并阈值逐渐增

大，聚类效果逐渐变差.结点网络综合特征值的基于蜂群优化搜索的信息流聚类算法对于合并阈值的变化，聚类效果影响不大，当合并阈值等于 0.4 时，聚类效果最好，因此在聚类算法中要选取合适的合并阈值.

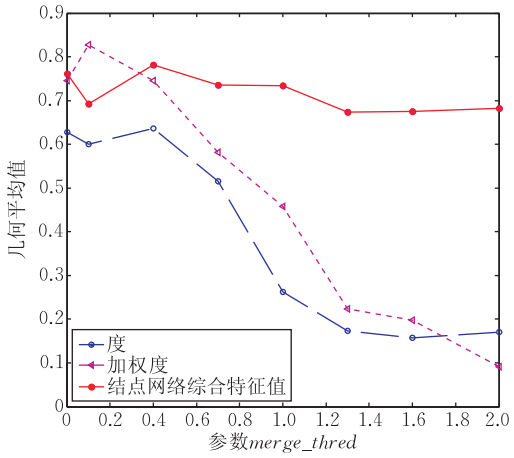

图 3 参数 *merge_thred* 对聚类结果的影响

图 4 中 α 是计算结点网络综合特征值的参数，当 $\alpha=0$ 时结点的网络综合特征值只反映了结点局部范围内的连接密度并没有考虑结点与其它结点的连接强度，当 $\alpha=1$ 时结点网络综合特征值只反映了结点与其他结点的连接强度和结点间距离的作用效果，没有考虑结点局部范围内的连接密度.$0<\alpha<1$ 时结点网络综合特征值既考虑了结点与其它结点的连接强度，结点间距离的作用效果，又考虑结点局部范围内的连接密度，充分表现了蛋白质交互网络的特性，结点信息非常丰富，因此聚类效果非常好，当 $\alpha=0.5$ 时，结点网络综合特征值中，将结点与其它结点的连接强度，结点间距离的作用效果以及结点局部范围内的连接密度均衡考虑，因此聚类效果最好.

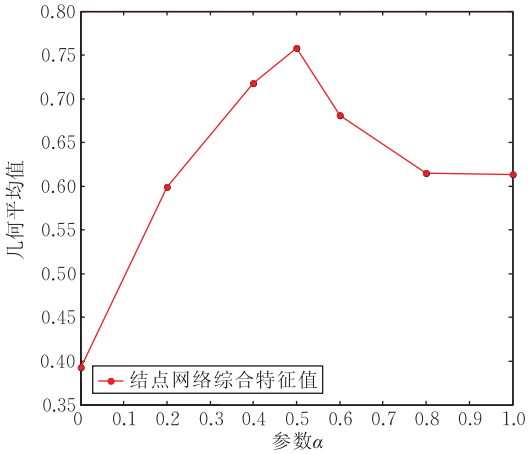

图 4 参数 α 对聚类结果的影响

图 5 中 *limit* 是蜂群算法中一个很重要的参数，用来决定是否合并模块相似度高的类，由侦察蜂进行搜索产生新解的一个参数.图 5 可以看出 *limit* 取得太小，侦察蜂就会不断地搜索新解来代替当前解，使算法丢弃最优解；取得太大，侦察蜂搜索新解的次数大大减少，又会使算法陷入局部最优，都会严重影响聚类效果.对于用结点的度和结点网络综合特征值确定注释蛋白质初始信息流的基于蜂群优化搜索的信息流聚类算法，*limit*＝10 时聚类效果最好，然后当 *limit* 逐渐增大时，聚类效果会逐渐变差，对于结点加权度确定注释蛋白质初始信息流的基于蜂群搜索的信息流聚类算法，当 *limit*＝15 时聚类效果最好，同样，当 *limit* 逐渐增大时，聚类效果会逐渐变得很差.

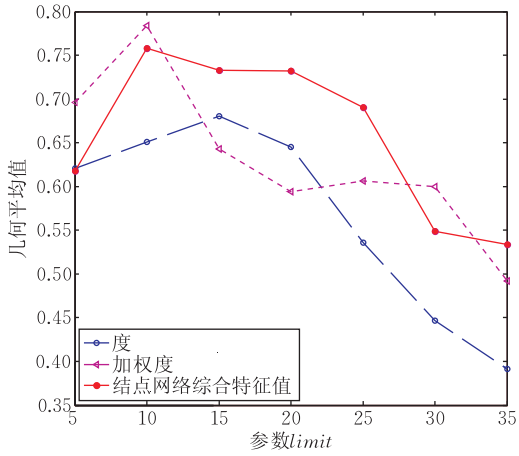

图 5 参数 *limit* 对聚类结果的影响

图 6 中 *size_thred* 是从初始聚类结果中剔除孤立点的聚类个数阈值.从图 6 可以看出，对于用结点的度和加权度确定注释蛋白质初始信息流的基于蜂群优化搜索的信息流聚类算法，当 *size_thred*＝1 时聚类效果最好，但是对聚类效果改进不是很大，对于

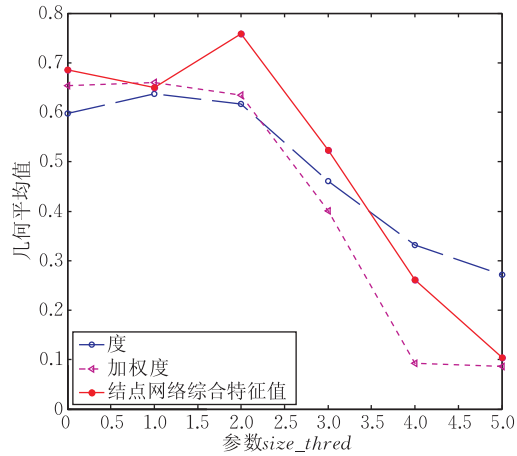

图 6 参数 *size_thred* 对聚类结果的影响

用结点网络综合特征值确定注释蛋白质初始信息流的基于蜂群优化搜索的信息流聚类算法，当 *size_thred*=2 时聚类效果得到显著改善. 对于这三种算法，当 *size_thred* 逐渐增大时，会将一些虽然结点个数少，但是信息丰富，连接紧密，相互作用强的模块作为孤立点被剔除，聚类结果的几何平均值急剧地变小. 因此在聚类算法中要选取合适的 *size_thred* 值.

5　结论及展望

本文中将结点的度 *degree*，加权度 *dw* 和网络综合特征值 *com_value* 分别作为注释蛋白质结点初始信息流的基本功能流算法的聚类结果做了比较，仿真结果表明用 *com_value* 作为初始信息流的基本功能流聚类算法聚类效果最好. 在此基础上，本文提出了一种新颖的基于蜂群优化搜索的信息流聚类模型与算法. 该方法中用结点的 *com_value* 来初始化聚类中心，将结点的 *com_value* 值作为注释蛋白质的初始信息流，将蜂群算法的蜜源位置对应于蛋白质相互作用网络的聚类中心，将蜜源的收益度大小对应于模块间的相似度，用跟随蜂和侦察蜂遍历结点. 采用正确率、查全率等指标对聚类效果做出客观评价，仿真结果表明该算法不仅克服了原算法的缺点，且其正确率和查全率的几何平均值最高，能够有效地识别蛋白质功能模块. 虽然本算法是针对蛋白质相互作用网络提出的，但是对于其它的具有类似结构的网络如人际关系网、因特网和 Web 网等复杂网络都具有借鉴意义.

参　考　文　献

[1] Barabasi A L, Albert R. Emergence of scaling in random networks. Science, 1999, 286(5439): 509-512

[2] Barabasi A L, Oltvai Z N. Network biology: understanding the cell's functional organization. Nature Review Genetics, 2004, 5(1): 101-113

[3] Schwikowski B, Uetz P, Fields S. A network of protein-protein interactions in yeast. Nat Biotechnol, 2000, 18(12): 1257-1261

[4] Yook S H, Oltvai Z N, Barabási A L. Functional and topological characterization of protein interaction networks. Proteomics, 2004, 4(4): 928-942

[5] Satuluri V, Parthasarathy S, Ucar D. Markov clustering of protein interaction networks with improved balance and scalability//Proceedings of the 1st ACM International Conference on Bioinformatics and Computational Biology. New York, USA, 2010: 247-256

[6] Wilkins M R, Sanchez J C, Gooley A A et al. Progress with proteome projects: Why all proteins expressed by a genome should be identified and how to do it. Biotechnology and Genetic Engineering Reviews, 1996, 13: 19-50

[7] Graves P R, Haystead T A. Molecular biologist's guide to proteomics. Microbiology and Molecular Biology Reviews, 2002, 66(1): 39-63

[8] Legrain P, Wojcik J, Gauthier J M. Protein-protein interaction maps: A lead towards cellular functions. Trends in Genetics, 2001, 17(6): 346-352

[9] Sun Peng-Gang, Gao Lin, Han Shan-Shan. Identification of overlapping and non-overlapping community structure by fuzzy clustering in complex networks. Information Sciences, 2011, 181(6): 1060-1071

[10] Wang Zheng-Hua, Dong Yun-Yuan, Wang Yong-Xian. Review on several clustering methods in protein-protein interaction network. Journal of National University of Defense Technology, 2009, 31(4): 81-86(in Chinese)
(王正华，董蕴源，王勇献. 蛋白质相互作用网络的几种聚类方法综述. 国防科技大学学报，2009，31(4)：81-86)

[11] Asur S, Ucar D, Parthasarathy S. An ensemble framework for clustering protein-protein interaction networks. Bioinformatics, 2007, 23(13): i29-i40

[12] Zhang Ai-Dong. Protein Interaction Networks. New York, USA: Cambridge University Press, 2009

[13] Mézard M. Where are the exemplars. Computer Science, 2007, 315(5814): 949-951

[14] Nabieva E, Jim K, Agarwal A et al. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. Bioinformatics, 2005, 21(1): i302-i310

[15] Cho Y R, Hwang W, Ramanathan M et al. Semantic integration to identify overlapping functional modules in protein interaction networks. BMC Bioinformatics, 2007, 8(265): 1-13

[16] Cho Y R, Hwang W, Zhang Ai-Dong. Identification of overlapping functional modules in protein interaction networks: Information flow-based approach//Proceedings of the 6th IEEE International Conference on Data Mining-Workshops, 2006: 147-152

[17] Watts D J, Strogatz S H. Collective dynamics of 'small-world' network. Nature, 1998, 393: 409-410

[18] Karaboga D, Basturk B. On the performance of artificial bee colony (ABC) algorithm. Applied Soft Computing, 2008, 8(1): 687-697

[19] Seeley T D. The wisdom of the hive: The Social Physiology of Honey Bee Colonies. Boston, Massachusetts, USA: Harvard University Press, 1995

[20] Teodorovic D, Lucic P, Markovic G et al. Bee colony optimization: principles and applications. Neural Network Applications in Electrical Engineering, 2006, 8: 151-156

[21] Karaboga D, Akay B. A comparative study of artificial bee colony algorithm. Applied Mathematics and Computation, 2009, 214(1): 108-132

[22] Cho Y R, Hwang W, Zhang Ai-Dong. Optimizing flow-based modularization by iterative centroid search in protein

interaction networks//Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineer. Boston, MA, 2007: 342-349

[23] Ma Rui-Fu. Clustering analyses algorithm based on the complexity of the networks performance evaluation for the research and application [D]. Beijing: Beijing University of Technology, 2009(in Chinese)
(马瑞复. 基于聚类分析算法的复杂网络绩效评估算法的研究与应用[D]. 北京：北京工业大学，2009)

[24] Lei Xiu-Juan, Huang Xu, Zhang Ai-Dong. Improved artificial bee colony algorithm and its application in gene and PPI data clustering//The IEEE 5th International Conference on Bio-Inspired Computing: Theories and Applications (BICTA2010). Changsha, China, 2010: 514-521

[25] Mewes H W, Frishman D et al. MIPS: A database for genomes and protein sequences. Nucleic Acids Research, 2002, 30(1): 31-34

[26] Salwinski L, Miller C S et al. The database of interacting proteins: 2004 update. Nucleic Acids Research, 2004, 32 (Suppl. 1): D449-D451

[27] Bader G D, Betel D, Hogue C W. BIND: The bimolecular interaction network database. Nucleic Acids Research, 2003, 31(1): 248-250

[28] Kerrien S, Alam-Faruque Y et al. IntAct-open source resource for molecular interaction data. Nucleic Acids Research, 2007, 35(Suppl. 1): D561-D565

[29] Chatr-Aryamontri A, Ceol A et al. MINT: The molecular interaction database. Nucleic Acids Research, 2007, 35(Suppl. 1): D572-D574

[30] Güldener U et al. CYGD: The comprehensive yeast genome database. Nucleic Acids Research, 2005, 33: D364-D368

LEI Xiu-Juan, born in 1975, Ph.D., associate professor. Her current research interests include intelligent computing and bioinformatics and so on.

TIAN Jian-Fang, born in 1986, M. S. candidate. Her current research interests include data mining and bio-information computing.

Background

A major goal of studying the Protein-Protein Interaction (PPI) networks is to identify and analyze the interactions of biological molecules in the cellular environment so as to profoundly understand the interactions among biological molecules and the mechanism of performing functions.

Although there are a large number of clustering methods, partition based method, hierarchical approach, model based method, density based method, graph theory based method, spectral clustering method and so on. However, owing to the small-world and scale-free properties of PPI networks, proteins (nodes) contain large amount of information and their interactions are very complicated, these traditional methods do not perform well in solving this problem. Function flow simulation algorithm is a novel clustering method which is proposed internationally in recent years. This method is consistent with the actual effect of interactions among proteins in terms of principle, meanwhile easy to understand and implement. Whereas, this method has the essential to manually predefine the merge threshold which is a little subjective. In addition, this approach doesn't take the effect of distance among protein nodes into account. Therefore, the clustering results are not satisfactory.

With regard to the former problems, the artificial bee algorithm was first introduced in the clustering problem of PPI networks, and then a novel algorithm named after information flow clustering model and algorithm based on Artificial Bee Colony (ABC) optimization searching was proposed in this paper. We took full advantage of the network comprehensive feature value of node which contained abundant information of node to initialize the cluster centers, the nectar source of ABC algorithm was corresponding to the cluster center of PPI networks. The income degree of nectar was regarded as the similarity between cluster modules. And then this algorithm sorted all the adjacent nodes of employed bee according to the descending order of network comprehensive feature values of nodes. In the end, the ordered neighbor nodes were considered to be the searching neighborhood of scouts. The experiment result showed that this algorithm could not only improve the clustering effect, but also effectively identify the functional module of protein so as to further predict the function of unknown protein.

This paper is supported by the National Natural Science

Foundation of China (No. 61100164), the Natural Science Foundation of Shaanxi Province of China in 2010 (No. 2010JQ8034), the Fundamental Research Funds for the Central Universities (No. GK200902016).

Significance: The research on the novel intelligent clustering algorithm with regard to the complicated PPI networks has a wide range of applications in various fields, such as exploring the potential information containing biological significance, annotating the biological function of unknown proteins, understanding the mechanism of biological activities, providing theoretic basis for the target spot discovery and design of new drug. This algorithm can predict functions of unknown protein, understand the molecular mechanism of the specific biological function, and provide important theoretic basis for exploring the mechanism of major diseases, disease treatment, disease prevention and the design of new drugs.

Previous researches:

[1] Xiujuan Lei, Xu Huang, Lei Shi, Aidong Zhang. Clustering PPI Data Based on Improved Functional-Flow Model through Quantum-behaved PSO. International Journal of Data Mining and Bioinformatics, 2010, Adopted.

This article used the quantum particle swarm optimization algorithm to automatically obtain the clustering threshold of PPI networks which overcame the drawback of artificially setting the merge threshold of the original flow algorithm.

[2] Xiujuan Lei, Xu Huang, Aidong Zhang. Improved Artificial Bee Colony Algorithm and Its Application in Gene and PPI Data Clustering. The IEEE Fifth International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA2010), Changsha, China, Sep. 23-26, 2010, 514-521.

This article proposed an improved colony algorithm which was applied in the clustering problem of gene and PPI data to automatically optimize the threshhold.

[3] Qun Zhang, Xiujuan Lei, Xu Huang, Aidong Zhang. An Improved Projection Pursuit Clustering Model and its Application Based on Quantum-behaved PSO. 2010 Sixth International Conference on Natural Computation (ICNC'10), Yantai, China, Aug. 10-12, 2010, Vol. 5: 2581-2585.

This article presented a project pursuit clustering model to reduce the dimension and took advantage of the quantum particle swarm optimization algorithm to optimize the projection direction.

Papers [1] and [2] adopted the intelligent optimization algorithms to automatically optimize the threshold during the procedure of clustering the similarity modules. However, in this paper, the principle of artificial bee colony algorithm was directly integrated into the information flow model which could solve the novel modeling and optimizing problems based on integrated swarm intelligent algorithm of information flow in the projects.