融合人工鱼群机理的PPI网络聚类模型与算法

吴 爽 雷秀娟
(陕西师范大学计算机科学学院 西安 710062)

摘 要 预测蛋白质交互作用(Protein-Protein Interaction,PPI)网络中未知蛋白质的功能,是生物信息学的一个研究热点。目前基于功能流的方法能有效地解决PPI网络的聚类问题,但是其正确率偏低、时间复杂度较高。为此提出了一种融合人工鱼群机理的PPI网络聚类模型与算法:将人工鱼看作一组聚类中心,觅食行为是指从每个聚类中心开始向它的邻接结点搜索并添加结点到该聚类模块中;接下来将目标函数值最大的人工鱼对应的一组聚类模块看作初始的聚类结果,对应鱼群的追尾行为;剩下的人工鱼开始执行聚群行为,判断对应的聚类模块与初始的聚类结果之间的相似度。如果相似度低于给定的阈值,则将聚类模块添加到初始的聚类结果中。PPI数据集上的仿真实验表明,该算法可以自动确定聚类数目,而且聚类结果的正确率和算法的运行效率都优于功能流算法。

关键词 人工鱼群算法,蛋白质交互作用网络,加权聚集系数
中图法分类号 TP391.4 **文献标识码** J

PPI Networks Clustering Model and Algorithm Combining with the Principle of Artificial Fish School

WU Shuang LEI Xiu-juan
(School of Computer Science, Shaanxi Normal University, Xi'an 710062, China)

Abstract Predicting function of unknown proteins in the protein-protein interaction networks is a hot topic in the bioinformatics. Recently functional flow method has effectively solved the problem of clustering PPI networks. However, the accuracy is relatively low and the time complexity is high. So the PPI networks clustering algorithm combining with the principle of artificial fish school was proposed, which considered an artificial fish as a set of cluster centers. The foraging behavior was regarded as searching the neighbor nodes of initial cluster centers and adding the nodes into cluster module. Afterwards the set of cluster modules having the highest fitness value was selected as the initial cluster result, which was corresponding to the following behavior of artificial fish school. Then other artificial fish began to execute swarming behavior and judged the similarities between the corresponding cluster modules and initial cluster result. If the similarity was lower than given threshold, the cluster module was added into the initial cluster result. The simulation experiment on PPI datasets shows that this algorithm can automatically determine cluster number. In addition, both the accuracy of cluster result and efficiency of algorithm are superior to functional flow algorithm.

Keywords Artificial fish school algorithm, Protein-protein interaction networks, Weighted clustering coefficient

1 引言

蛋白质与蛋白质交互作用是生物体中众多生命活动过程的重要组成部分,在生物体中几乎无所不在。因此,研究生命活动过程中蛋白质交互作用,有助于揭示生命过程的许多本质问题[1]。随着蛋白质的实验数据的积累以及研究手段的快速发展,目前已产生了大量的蛋白质交互作用(Protein-Protein Interaction, PPI)网络数据库,如DIP、BIND、HPRD、MIPS[2]等。研究PPI网络中的聚类模块,能有效地预测出未知蛋白质的功能,为揭示一些重大疾病的机理和判断治疗以及药物的筛选、合成提供理论基础。

目前,PPI网络已经被证实具有小世界特性和无尺度特性。网络的小世界特性[3]主要表现为具有较高的聚集系数和较低的特征路径长度。网络的无尺度特性[4]则主要表现为增长和偏好连接两个特性。一般来说,在同一个模块内的蛋白质具有相似的表达谱、亚细胞定位[5]和基因显性,这说明了功能与模块组织的相关性。因此,针对PPI网络独特的拓扑特征,可以通过聚类分析来揭示它的内部结构,进一步预测聚类内功能未知的蛋白质。在解决PPI网络聚类的问题上,可以采用基于划分的方法、层次聚类方法、基于密度的方法以及基于模型的方法等。但是这些方法都存在一定的缺陷,有的算法对噪声点比较敏感,有的算法则会丢弃大量的拥有少数连接的节点。因此,近年来产生了一些新颖的算法来分析和研究PPI网络。2005年,Elena Nabieva等人[6]首次提出了一种

到稿日期:2011-08-29 返修日期:2011-11-22 本文受2011年国家自然科学基金(61100164),陕西省2010年自然科学基础研究计划项目(2010JQ8034),2009年中央高校基本科研业务费专项资金项目(GK200902016),陕西师范大学研究生创新基金(2011CXS030)资助。
吴 爽(1987-),女,硕士生,主要研究方向为数据挖掘、生物信息计算,E-mail:persistencewu@126.com;雷秀娟(1975-),女,博士,副教授,硕士生导师,CCF会员,主要研究方向为智能计算、生物信息计算。

功能流模型，即在确定的时间内对结点的流向和流量进行仿真并产生聚类模块，但是该模型没有考虑到距离的作用效果。Cho Young-Rae[7]提出了一种模块化方法来识别重叠的功能模块，该方法可以确定一个蛋白质在功能上对其它蛋白质产生多少影响，能有效解决PPI网络聚类问题，但是查准率和查全率[8]偏低。

人工鱼群算法是国内学者李晓磊[9]模仿鱼类行为方式提出的一种基于动物自治体的优化方法，是一种群智能全局随机优化算法，具有跳出局部极值、取得全局极值的能力，而且对搜索空间具有一定的自适应能力，有较快的收敛速度。目前，人工鱼群算法已应用到许多领域。李晓磊最初将人工鱼群算法用于解决连续性函数优化问题和组合优化问题，它具有良好的寻优效果[10]。曲良东等人[11]结合混沌搜索的特点和人工鱼群算法的性能，提出了一种混沌人工鱼群优化算法CAFSA(Chaotic Artificial Fish School Algorithm)，并在测试函数上做了仿真实验。结果表明，该算法全局能力更强、搜索效率更高。随后，刘白等人[12]将人工鱼群算法和传统的k-means算法相结合，提出了一种基于人工鱼群的聚类分析算法。Iris数据集上的实验结果表明，该算法克服了需要聚类数目先验知识的缺陷，在时间效率和聚类结果的正确率上都取得了较好效果。

结合PPI网络的拓扑结构特点和人工鱼的觅食、追尾以及聚群行为特点，本文提出了一种融合人工鱼群机理的PPI网络聚类算法；第2节简单介绍了人工鱼群算法的基本原理以及算法在优化过程中需要采用的目标函数，提出了一种聚类结果的评价方法；第3节则将人工鱼群算法的机理融入到PPI网络聚类问题中，详细阐述了算法的编码、解空间设计以及具体的操作步骤和流程；第4节在PPI网络数据集上进行仿真，结果表明，本算法可以自动确定聚类数目，并且聚类结果的正确率和算法的运行效率都得到了提高；最后是结束语。

2 相关知识介绍

2.1 人工鱼群算法的基本原理

人工鱼群(Artificial Fish School，AFS)算法[13]是模仿鱼群的觅食等寻优行为而设计的一种优化算法，是集群智能算法的一个具体应用。它收敛速度较快，最初用于解决连续性函数优化问题和组合优化问题，取得了较好的寻优效果。随后，算法在神经网络训练以及聚类分析等问题上都表现出了鲁棒性强、全局收敛性好的特点。因此，本文采用人工鱼群算法的优良特性来解决PPI网络的聚类问题。首先简单阐述一下鱼群算法的基本原理。

在鱼群的活动过程中，觅食行为、追尾行为和聚群行为等与优化问题的解决有着密切的关系。觅食行为是指鱼群循着食物浓度比较丰富的方向游动。聚群行为则是为了避免过分拥挤，尽量向邻近伙伴的中心移动。追尾行为是一种向邻近的最优伙伴的方向移动的过程。假设人工鱼当前的状态可以表示为$X=(x_1,x_2,\cdots,x_n)$，其中$x_i(i=1,2,\cdots,n)$为寻优变量。人工鱼当前所在的位置的食物浓度表示为$Y=f(X)$，Y是要优化的目标函数值。人工鱼个体之间的距离一般采用欧式距离$d_{i,j}=\|X_i-X_j\|$，$Visual$表示人工鱼的可视距离，$Step$表示人工鱼移动的最大步长，δ为拥挤度因子。

在觅食行为中，人工鱼当前的状态是X_i，在其感知范围$Visual$内随机选择一个状态X_j，如果该状态的目标函数值优于状态X_i的目标函数值，则向该方向前进一步。移动之后，人工鱼的状态为

$$X_{i|next}=X_i+\text{Rand}()\cdot Step\cdot\frac{X_j-X_i}{\|X_j-X_i\|} \tag{1}$$

否则重新选择随机状态X_j，再次判断是否满足前进条件。这样反复尝试$trynumber$次后，如果仍不满足前进条件，则该人工鱼随机移动一步。

在聚群行为中，人工鱼首先探测当前邻域$Visual$内的伙伴数目n_f以及中心位置X_c。如果$Y_c/n_f>\delta Y_i$，表明伙伴中心的食物浓度比较高，而且并不拥挤，则该人工鱼朝伙伴的中心位置方向前进一步。否则，执行觅食行为。

在追尾行为中，人工鱼如果搜索到当前邻域内的伙伴X_j的目标函数值Y_j最优，并且满足条件$Y_j/n_f>\delta Y_i$，表明伙伴X_j具有较高的食物浓度，而且其周围不太拥挤，则朝伙伴X_j的方向前进一步。否则，执行觅食行为。

2.2 目标函数

PPI网络中结点的度和加权度反映了结点与其它结点的连接强度。结点的聚集系数[8]是指与该结点相连的邻接结点之间实际存在的边的数目与可能存在的边的数目的比例，反映了网络中结点的聚集程度。结点的加权聚集系数则反映了此结点局部范围内的相互连接密度和强度，表示为

$$\begin{aligned} w(i)&=\sum_{j=1}^{d_i-1}\sum_{k=j+1}^{d_i}weight(j,k)\\ coeff(i)&=2n_i/d_i(d_i-1)\\ wcoeff(i)&=2*w(i)/n_i(n_i-1)\end{aligned} \tag{2}$$

式中，i是PPI网络中的一个蛋白质结点，d_i表示与结点i相邻接的蛋白质结点的数目，n_i指的是与结点i相邻接的蛋白质结点之间存在的边的数目，$weight(j,k)$代表结点i的第j个和第k个邻接结点之间的边的权重，$w(i)$得到的是结点的加权度，$coeff(i)$代表了结点i的聚集系数。在加权的PPI网络中，根据式组(2)中的第3个公式可以得到结点的加权聚集系数$wcoeff(i)$。

在带权重的PPI网络中，结点的加权聚集系数表征了每一个蛋白质结点的特征。因此，一个聚类模块的聚集系数可以用来衡量该模块内所有蛋白质结点之间的拓扑结构特征，它定义为该聚类模块内的所有蛋白质结点的平均加权聚集系数[14]，计算公式为

$$C_B=\frac{\sum_{i=1}^{h}wcoeff(B(j))}{h} \tag{3}$$

式中，B是一个聚类模块，该模块包含了h个蛋白质结点。C_B得到的是聚类模块B的加权聚集系数，就是PPI网络聚类过程中要采用的目标函数。该值越大，模块内部蛋白质结点之间的连接强度和密度越高，则这些结点属于同一个功能模块的可能性也就越大，因此该聚类模块的正确率就越高。

2.3 聚类结果的评价方法

在PPI网络中，识别功能模块可以为未知蛋白质的功能预测提供指导作用。由于采用不同的方法会产生不同的效

果，而且同一种方法采用不同的参数也可能会产生不同的效果，因此选择聚类结果的评价方法具有很重要的作用。比较常用的评价方法就是查准率和查全率。假定一个模块 X 被映射到标准数据集的一个类 F_i 中，查准率 *precision* 表示模块 X 和 F_i 的交集的结点个数与模块 X 中的蛋白质结点个数的比值；查全率 *recall* 则表示模块 X 和 F_i 的交集的结点个数与类 F_i 的蛋白质结点个数的比值[8]。

$$precision=\frac{|X\cap F_i|}{|X|} \tag{4}$$

$$recall=\frac{|X\cap F_i|}{|F_i|} \tag{5}$$

当所有的蛋白质结点都分到了一个类中，那么该类的查全率就达到最大值。反之，如果聚类模块比较小，那么查准率就会非常高。因此，我们折中了这两种评价方法，用 *f-measure* 值来评价聚类方法的优劣，它是查准率、查全率和算法的运行时间 *time* 的调和平均值：

$$f\text{-}measure=\frac{3}{\frac{1}{precision}+\frac{1}{recall}+time} \tag{6}$$

3 基于人工鱼群机理的 PPI 网络聚类模型

3.1 数据的预处理及聚类中心的选择

针对 PPI 网络数据的结构特点，为了提高程序的运行效率，算法首先将 PPI 网络中的蛋白质名称用数字编号标识。接下来初始化人工鱼，对应每条人工鱼，生成一组聚类中心。而聚类中心是根据每个蛋白质结点的度以及加权聚集系数来选择的，如果某个结点的度和加权聚集系数大于给定的阈值，则这个结点可以作为初始的聚类中心。为了避免聚类结果中出现较多的重叠的聚类模块，在聚类中心初始化的过程中，每条人工鱼内部所对应的聚类中心必须互异，若干条人工鱼之间所对应的聚类中心也是不同的。

3.2 聚类模型设计

在进行数据预处理之后，算法将人工鱼群算法的机理应用于 PPI 网络聚类问题中，提出了一种融入人工鱼群机理的 PPI 网络聚类算法，简记为 CPAF(Clustering Protein-protein interaction networks combing with Artificial Fish)算法。算法首先初始化若干条人工鱼，其中一条人工鱼对应一组聚类中心，而聚类中心是根据结点的度以及加权聚集系数来进行选择的。每条人工鱼在执行觅食行为的过程中，算法从初始的聚类中心开始一级一级地搜索其邻接结点。人工鱼群算法机理与 PPI 网络聚类问题的对应关系如表 1 所列。

表 1 人工鱼群算法机理和 PPI 网络聚类问题的对应关系

人工鱼群算法机理	PPI 网络聚类问题
初始的人工鱼	一组聚类中心
食物浓度	聚类模块的目标函数值
可视距离 Visual	结点的第 Visual 级邻接结点
觅食行为	从初始的聚类中心搜索得到聚类模块
觅食之后的人工鱼	一组聚类模块
随机行为	对目标函数值没有更新的聚类模块，随机添加一个结点
追尾行为	选择目标函数值最优的人工鱼所对应的聚类模块作为初始的聚类结果
聚群行为	添加其它与初始聚类结果相似度较低的聚类模块，得到最终的聚类结果

在表 1 中，算法将人工鱼群算法中的觅食行为、随机行为、追尾行为以及聚群行为与 PPI 网络聚类问题一一对应。假设一条人工鱼对应的聚类中心是 $C_1,C_2,\cdots,C_k$，对于第一个聚类中心 C_1，首先将它的第一级邻接结点 $C_{11},C_{12},\cdots$ 添加到该模块中，根据式(3)计算该模块 M_1 的目标函数值，将其作为该模块的初始适应度值 *fitness*。接下来搜索该聚类中心的第二级邻接结点 $C_{111},C_{112},\cdots$，并将其添加到模块 M_1 中。如果该模块的适应度值高于目标函数值 *fitness*，则更新模块 M_1，否则丢弃第二级邻接结点。执行上述过程 *trynumber* 次后，得到最终的聚类模块 M_1，并计算其对应的适应度值 *new_fitness*，如果高于 *fitness*，则模块 M_1 就是一个以 C_1 为中心的聚类模块。如果以 C_k 为聚类中心的模块在执行觅食行为 *trynumber* 次之后，目标函数值没有得到更新，则执行随机行为，从 C_k 的第二级邻接结点中随机选择一个结点，添加到模块 M_1 中，得到一个以 C_k 为中心的聚类模块。图 1 是一条人工鱼对应的聚类中心形成聚类模块的过程。

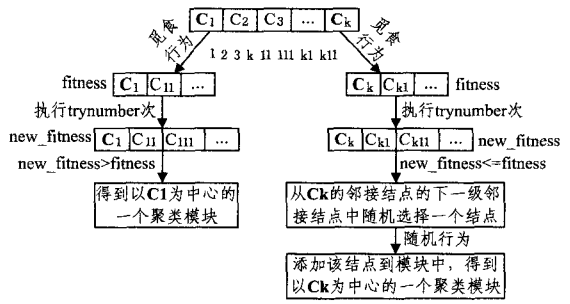

图 1 人工鱼的觅食过程

假设算法中有 n 条人工鱼，每条人工鱼初始时对应 k 个聚类中心，则最终产生了 $n\times k$ 个聚类模块。接下来采用人工鱼群算法中的追尾行为和聚群行为优化，得到聚类模块。首先分别计算每条人工鱼所对应的 k 个聚类模块的目标函数值之和，并将 n 条人工鱼按照目标函数值之和从大到小排序，根据人工鱼群算法中的追尾行为，选择目标函数值之和最大的人工鱼所对应的 k 个聚类模块作为初始的聚类结果 *Fcluster*。剩下的 $(n-1)\times k$ 个聚类模块按照其对应的人工鱼的排列顺序依次存入数组 *Rcluster* 中。最后计算 *Rcluster* 中第一个聚类模块与 *Fcluster* 中的聚类模块之间的相似度，如果小于给定的阈值，则将该模块添加到 *Fcluster* 中。持续该过程，直到 *Rcluster* 中所有的聚类模块都进行了判断，得到最终的聚类结果。

3.3 算法描述

根据第 3.2 节提出的模型得到相应的算法，具体步骤如下：

步骤 1 数据预处理并初始化 n 条人工鱼，每条人工鱼对应 k 个聚类中心，并设置人工鱼的可视距离 $Visual=1$，觅食行为的最大尝试次数 $trynumber=10$，相似度阈值 $sim\_threshold=0.1$ 以及访问标志 $i=1$；

步骤 2 从第 i 条人工鱼所对应的聚类中心开始执行觅食行为和随机行为，生成一组聚类模块，其中包含 k 个聚类模块；

步骤 3 令 $i=i+1$，如果 $i\leqslant n$，则返回步骤 2，否则执行步骤 4；

步骤 4　对所有的人工鱼执行追尾行为，就是选择适应度值最高的人工鱼所对应的一组聚类模块作为初始的聚类结果 *Fcluster*，剩下的聚类模块按照所对应的人工鱼的适应度值从大到小的顺序依次存入数组 *Rcluster* 中，并设置访问标志 $j=1$；

步骤 5　分别计算 *Rcluster* 中的第 j 个聚类模块与 *Fcluster* 中的每一个聚类模块之间的相似度，如果低于给定的阈值 *sim_threshold*，则将该模块添加到 *Fcluster* 中；

步骤 6　令 $j=j+1$，如果 j 的值大于数组 *Rcluster* 的行数，则输出最终的聚类结果 *Fcluster*，否则继续执行步骤 5。

算法流程图如图 2 所示。

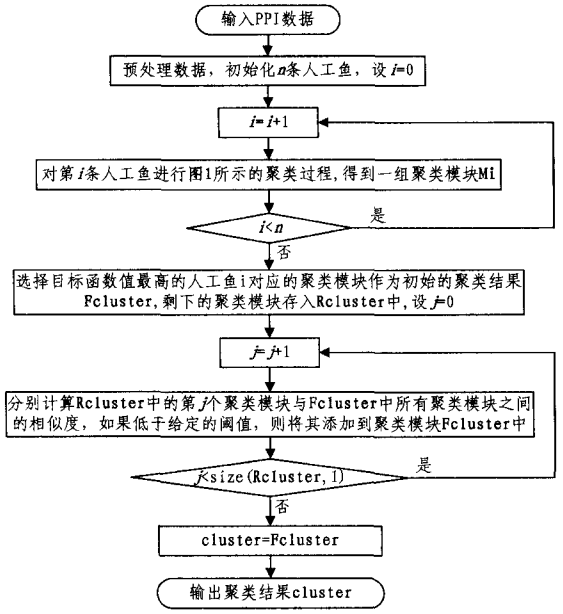

图 2　CPAF 算法流程图

4　仿真结果

为了测试算法的性能，本文采用 MIPS 数据库中的 PPI 数据集[15]进行实验。PPI 数据集由蛋白质以及蛋白质结点之间的相互作用的权重大小构成，因此首先要对数据进行预处理。除此之外，算法融入了人工鱼群算法的机理，其中涉及到人工鱼的可视距离 *Visual*、拥挤度因子、觅食行为的尝试次数 trynumber 以及人工鱼的数量等参数。但在融入人工鱼群机理的 PPI 网络聚类算法中，拥挤度因子的概念没有被引入模型设计中，因此接下来要讨论其它一些参数对算法聚类结果的影响。

4.1　参数分析

在本算法中，设置人工鱼的可视距离 $Visual=1$，觅食行为的最大尝试次数 $trynumber=10$，相似度阈值 $sim\_threshold=0.1$，这些参数的选取对聚类结果的影响不大。接下来分析种群数目以及每一条人工鱼所对应的一组聚类中心的数目对聚类结果产生的影响。

图 3 显示了种群数目对聚类结果的正确率的影响。从图 3 中可以看出，当种群数目取值为 3 时，聚类结果的查准率和 *f-measure* 值相对较高。图 4 则是聚类结果的聚类数目随种群数目的变化图。随着聚类数目的增加，得到的聚类结果的数目以下降的趋势不断减少。

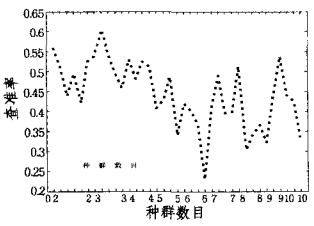

图 3　种群数目对聚类结果的查准率的影响

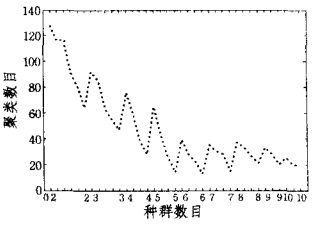

图 4　种群数目对聚类结果的聚类数目的影响

表 2 显示了随着种群数目和聚类中心数目的变化，聚类结果的查准率、查全率、*f-measure* 值以及最终得到的聚类数目的变化情况。由于 PPI 网络具有无尺度属性，拥有大量连接的结点的数目有限，因此当种群数目发生变化时，每一条人工鱼所对应的聚类中心的数目必须在一定的范围内进行选择。从表中可以看出，如果种群的聚类数目设置偏低，则每条人工鱼对应的一组聚类中心的数目相应地偏高，虽然聚类结果的查准率在一定程度上不会受到影响，但是算法在执行追尾和聚群行为时的全局搜索性能得不到很好的发挥，因此聚类结果的 *f-measure* 值偏低。如果种群数目设置偏高，则每条人工鱼对应的聚类中心的数目取值比较低，因此算法执行追尾行为之后得到的初始聚类结果的规模相对较小，导致最终得到的聚类结果的聚类数目比较少，而且查准率也偏低。结合表 2 所列以及以上分析得出结论：当种群数目设置为 3、每条人工鱼对应的一组聚类中心的数目取值为 60 时，聚类效果比较好。

表 2　种群数目对聚类结果的影响

种群数目	聚类中心数目	查准率	查全率	运行时间(s)	f-measure 值	聚类数目
2	100	0.5585	0.2892	17.094	0.4801	128
2	90	0.5122	0.2911	15.531	0.4697	127
2	80	0.4373	0.3278	10.922	0.4734	117
2	70	0.4944	0.2592	9.4375	0.4360	91
3	**60**	**0.6042**	0.3052	14.391	**0.5058**	85
3	50	0.5455	0.3044	12.484	0.4903	63
3	40	0.4953	0.3202	10.219	0.4881	54
4	50	0.5299	0.2689	17.906	0.4541	76
4	40	0.4814	0.3323	21.266	0.4929	59
4	30	0.5239	0.2875	14.031	0.4697	38
4	20	0.5117	0.3061	10.787	0.4882	28
5	40	0.4073	0.3081	23	0.4477	65
5	30	0.4276	0.2631	17.781	0.4202	42
5	20	0.4858	0.3067	11.563	0.4748	26
6	30	0.4197	0.3415	16.906	0.4754	39
7	20	0.4892	0.3448	14.641	0.5047	30
8	20	0.3028	0.3008	17.344	0.3938	33
8	10	0.36493	0.38701	16.813	0.4744	21
9	20	0.3213	0.3339	19.672	0.4221	33
9	10	0.5381	0.32508	13.891	0.5055	20
10	20	0.4406	0.3249	21.25	0.4726	25
10	10	0.3256	0.3292	11.828	0.4220	19

4.2　算法性能测试

图 5 显示了 CPAF 算法在执行追尾和聚群行为前后的聚类结果的查准率对比情况。人工鱼在执行觅食行为和随机行为之后产生若干组聚类模块，每条人工鱼仅仅找到了个体的局部极值，因此聚类效果一般。在执行追尾行为时，选择适应度值最大的人工鱼对应的聚类模块作为初始的聚类结果，也就是搜索到种群的全局极值，在此基础上通过其它人工鱼的聚群行为来优化初始的聚类结果，得到最终的聚类结果。结果表明，采用人工鱼群算法的优良的全局搜索性能可以提高 PPI 网络聚类结果的正确率。为了更充分地体现 CPAF 算法

的性能优势，本文引言中提到了功能流 Flow 算法，它是目前解决 PPI 网络聚类问题的一种相对有效的方法，因此本文将 CPAF 算法与 Flow 算法进行了对比，效果如图 6 所示。

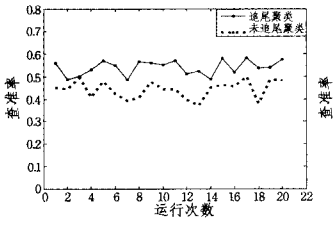

图 5 CPAF 算法执行追尾聚群前后的聚类结果的对比图

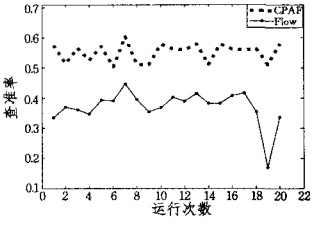

图 6 CPAF 算法与 Flow 算法在聚类结果的查准率上的比较

图 6 给出了 CPAF 算法和 Flow 算法在 PPI 网络数据集上各自运行 20 次的聚类结果的查准率的比较。从图中可以看出，CPAF 算法的聚类结果的查准率值在 0.5 和 0.6 之间波动，比较稳定；而 Flow 算法的聚类结果的查准率在 0.15 到 0.45 之间上下起伏，波动比较大。为了综合评价两种算法在查准率、查全率以及运行时间上的性能对比情况，图 7 给出了两种算法各自运行 20 次的聚类结果的 *f-measure* 值比较。从图中可以看出，CPAF 算法的聚类结果的 *f-measure* 值远远高于 Flow 算法，而且变化相对稳定。这是由于 CPAF 算法融入了人工鱼群机理，优化了聚类模块的生成过程，同时根据鱼群的追尾和聚群行为特点，对产生的初始聚类模块进行选择与添加操作，因此能够更有效、更快速地识别出具有相同功能的聚类模块。

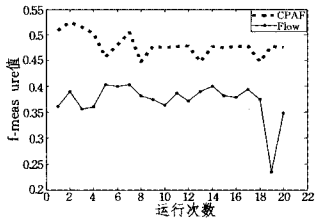

图 7 CPAF 算法与 Flow 算法在聚类结果的 *f-measure* 值上的比较

表 3 聚类结果中的前 10 个聚类模块

聚类模块	聚类正确的蛋白质	聚类错误的蛋白质
1	YGL240w, YFR036w, YHR166c, YDR118w, YBL048c, YKL022c, YOR249c, YNL172w, YDL008w, YLR102c, YLR127c	——
2	YML077w, YBR254c, YOR115c, YMR218c, YDR472w, YDR246w, YKR068c, YDR108w, YDR407c	——
3	YDR176w, YLR055c, YPL254w, YMR236w, YGR252w, YBR198c, YDR392w, YDR167w	YDR145w, YBR081c
4	YGR078c, YNL153c, YML094w, YEL003w, YLR200w	YER016w
5	YLR277c, YJR093c, YMR061w, YAL043c, YLR115w, YDR301w	YPR107c, YKR002w, YNL317w
6	YGL025c, YOL135c, YDL005c, YOR174w, YBR193c	——
7	YCR002c, YJR076c, YHR107c, YLR314c	——
8	YMR109w, YLR337c, YKL129c	YER084w
9	YDR318w, YPL018w	YPL008w
10	YKR054c, YMR294w	YHR191c, YDR150w, YOR265w, YMR299c, YLR381w

表 3 列出了采用 CPAF 算法得到的聚类模块。由于篇幅限制，表 3 只列出了前 10 个聚类模块中聚类正确的蛋白质和聚类错误的蛋白质。从表中可以看出，在聚类模块 1，2，6，7 中，每一个模块内的蛋白质结点都存在于标准数据库对应的聚类模块中，因此这些聚类模块的查准率都是 1。从表 3 的聚类结果中可以有效地识别出一些具有相同功能的蛋白质，从而预测未知蛋白质的功能。

结束语 针对 PPI 网络数据集的拓扑结构特点以及功能流算法在解决 PPI 网络聚类问题中的缺陷，本文将人工鱼群算法应用于 PPI 网络聚类问题中，提出了一种融入人工鱼群机理的 PPI 网络聚类算法。算法将人工鱼看作是一组聚类中心，通过觅食行为和随机行为产生若干组聚类模块，并利用鱼群的追尾和聚群行为优化聚类模块，产生最终的聚类结果。实验表明，算法可以自动确定聚类数目，而且聚类结果的查准率和查全率优于功能流算法。除此之外，算法的运行效率也得到了很大程度的提高。

参 考 文 献

[1] 刘昊. 基于聚类算法和相互作用网络的蛋白质功能预测研究[D]. 长沙：湖南大学，2009

[2] Guldener U, Munsterkotter M, Oesterheld M, et al. The MIPS protein interaction resource on yeast[J]. Nucleic Acids Research, 2006, 34: D436-D441

[3] Watts D J, Strogatz S H. Collective dynamics of 'small-world' networks[J]. Nature, 1998, 393: 440-442

[4] Barabási A L, Oltvai Z N. Network biology: understanding the cell's functional organization[J]. Nature Reviews: Genetics, 2004, 5: 101-113

[5] 张树波，赖剑煌. 蛋白质亚细胞定位预测的机器学习方法[J]. 计算机科学，2009，36(4)：29-33

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

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

[8] Zhang Ai-dong. Protein Interaction Networks[M]. New York, USA: Cambridge University Press, 2009

[9] 李晓磊，邵之江，钱积新. 一种基于动物自治体的寻优模式：鱼群算法[J]. 系统工程理论与实践，2002，22(11)：32-38

[10] 李晓磊，路飞，田国会，等. 组合优化问题的人工鱼群算法应用[J]. 山东大学学报，2004，34(5)：64-67

[11] 曲良东，何登旭. 一种混沌人工鱼群优化算法[J]. 计算机工程与应用，2010，46(22)：40-42

[12] 刘白，周永权. 一种基于人工鱼群的混合聚类算法[J]. 计算机工程与应用，2008，44(18)：136-138

[13] 李晓磊. 一种新兴的智能优化算法——人工鱼群算法[D]. 杭州：浙江大学，2003

[14] Li Xiao-li. Biological data mining in protein interaction networks [C]// IGI publishing. 2009

[15] Güldener U, Münsterkötter M, Kastenmüller G, et al. CYGD: the comprehensive yeast genome database[J]. Nucleic Acids Research, 2005, 33: D364-D368