PPI网络聚类的评价方法的研究与应用

尤梦丽 雷秀娟
(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(Protein-Protein Interaction,PPI)聚类结果的评价方法的研究是检测PPI网络功能模块聚类结果正确与否的关键。介绍并分析了4种有代表性的PPI网络聚类的评价方法,即p-value、匹配统计量、基于准确率和查全率的综合评价以及基于层结构的hF-measure,在此基础上考虑了主错误划分类与该预测类的相似性,提出了新的罚分函数和新的Sf-measure评价方法。仿真结果表明了各评价方法的特点及Sf-measure评价方法的有效性及合理性。

关键词 蛋白质相互作用网络,评价方法,调和平均值,主错误划分类,Sf-measure
中图法分类号 TP391 **文献标识码** A

Study and Application of Evaluating Methods of PPI Network Clustering

YOU Meng-li LEI Xiu-juan
(School of Computer Science, Shaanxi Normal University, Xi'an 710062, China)

Abstract The research in evaluating clustering results for PPI (Protein-Protein Interaction) network is the key to detect clustering results of function module in PPI network. The four typical methods evaluating clusters of PPI (protein-protein interaction) network were introduced and analyzed in this paper, which are p-value, matching statistics, f-measure based on recall and precision and hF-measure based on hierarchical structure. Besides, considering the similarity between the main error classification and the cluster predicted, a new penalty function and the new Sf-measure evaluation method were put forward lately. The simulation results show the features of various evaluation methods and the rationality and effectivity of Sf-measure method.

Keywords Protein-protein interaction(PPI) network, Evaluation method, f-measure, Main error classification, Sf-measure

1 引言

近些年来随着生物技术尤其是高通量实验技术的发展,已经产生了大量的蛋白质与蛋白质相互作用的数据,而这些数据使得我们能够从系统水平上去理解细胞机制的组成[1]。众所周知,蛋白质并不是孤立地存在的,或者说很少单独地表达特定的功能,而是通过彼此间复杂的相互作用来进行特定的生命活动[2]。为了从蛋白质与蛋白质相互作用网络(PPI)中识别聚类,人们已经进行了很大的努力。通常人们期望被识别的聚类是蛋白质复合物或者有着特定作用的功能模块。

迄今为止,已经提出了许多进行PPI聚类的计算方法,包括基于密度的和局部搜索的方法[3-7]、图像分割法[8,9]、层次化的聚类方法[10,11]、监督图像聚簇法[12]。目前涌现了很多的聚类方法[11,13],我们课题组对基于群体智能优化机理的聚类方法进行了研究[14-17]。随着聚类算法的增多,聚出的新类的数目也越来越多,在生物学家将它们用于生化试验之前,合理地对聚类结果进行有效的生物验证即评价也很重要。

评价顾名思义就是一种有依据的合理判断。对PPI聚类分析方法的评价就是对它们分析结果的有效验证过程。由于不同的聚类方法得到的聚类个数和规模都存在较大的差异,因此采用哪种评价方法才能更合理、公平地评价它们就成了一个巨大的挑战。最直接的方法就是把它们与生化试验已经获得的蛋白质复合物或功能模型进行匹配。本文首先介绍了4种有代表性的评价方法,并在此基础上进行了一定的改进,将模块间的相似性对聚类结果的影响引入到实际的应用中;然后对同一算法的不同聚类结果用不同的评价方法进行仿真,并比较和分析这些评价方法的特点;最后讨论在这个领域有发展前景的研究热点。当然,要想找到一个统一的、可靠的PPI聚类的评价方法主要取决于两个方面的改善:a)实验技术,要求生物学家提供丰富而可靠的生物数据集以便评价时有一个"黄金"参考标准;b)数据评价技术,要求计算机科学家提供有效而健壮的计算方法,全面而合理地对聚出的结果进行评价。

2 PPI聚类分析的常见评价

对PPI网络进行聚类是为了通过已知蛋白质的功能预测

到稿日期:2013-02-01 返修日期:2013-04-30 本文受国家自然科学基金青年基金(61100164,61173190),教育部留学回国人员科研启动基金(教外司留[2012]1707号),陕西省2010年自然科学基础研究计划青年基金(2010JQ8034)资助。
尤梦丽(1985—),女,硕士生,主要研究领域为数据挖掘、生物信息计算;雷秀娟(1975—),女,博士,教授,硕士生导师,CCF会员,主要研究领域为智能计算与智能优化、生物信息计算等,E-mail:xjlei168@163.com(通信作者)。

未知蛋白质的功能，进而指导对重大疾病的机理的研究和判断治疗。因此聚类结果的质量直接影响生物学家做出的临床实验方案和对实验结果的判断，而评价方法正是用来估测聚类结果质量的。现介绍几种常见的有代表性的评价方法。

2.1 基于功能富集分析的 p-value 评价分析法

GO(gene ontology，GO)是基因本体联合会所建立的数据库[18]，并且是一个存储相互作用和功能数据的组织结构。它采用分级结构，使用有控制的词汇表和严格定义的概念关系，以有向无环图形式，用 3 种描述符(分子功能、生物过程与细胞组分)对基因产品(大多数是蛋白质)进行描述。由于不同的功能注释之间的分布不是统一的，并且该数据集是有限的，很多研究者在评价聚类分析结果的时候直接采用能够获得的蛋白质功能注释信息进行功能富集分析。预测的 cluster 对于某一功能富集程度通常用 P 值来评估，其计算公式如下[6]：

$$p\text{-}value=1-\sum_{i=0}^{m-1}\frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}} \tag{1}$$

式中，N 表示整个 PPI 网络中包含的蛋白质总个数，而具有特定功能 f 的蛋白质个数为 M，n 表示预测的 cluster 中蛋白质个数，而含有特定功能 f 的蛋白质个数为 m。

具有超几何聚集分布的 p 值表明了预测的 clusters 能够随机出现某种功能的概率。p 值越小，越接近于 0，则预测的 cluster 中的蛋白质能够随机出现这种功能的概率就越低，可能越有生物意义。p 值本身对 cluster 的规模存有偏见，它更偏向于规模较大的类。通过每个预测的 cluster 取得最小 p 时所对应的功能可以预测未知蛋白质的功能。

为了评估聚类算法的整体性能，有研究者在 p 值的基础上定义了综合评价指标 clustering score，如式(2)所示[19]：

$$clustering\ score=\frac{\sum_{i=1}^{n_S}\min(p_i)+(n_I*cutoff)}{n_S+n_I} \tag{2}$$

或者如式(3)所示[20]：

$$clustering\ score=1-\frac{\sum_{i=1}^{n_S}\min(p_i)+(n_I\times cutoff)}{(n_S+n_I)\times cutoff} \tag{3}$$

式中，n_S 和 n_I 分别表示有显著意义和无显著意义的 clusters 的个数，$\min(p_i)$表示第 i 个 cluster 对应的所有 p 值中最小的一个。因为每个蛋白质可能有多个不同的功能，而每个 cluster 的功能为其所包含的所有蛋白质的功能的集合，因此每个 cluster 对应不同的功能可以计算出不同的 p 值。参数 $cutoff$ 是用来区分有意义和无意义的 cluster。如果一个预测的 cluster 的 p 值小于 $cutoff$，那么就称这个 cluster 是有意义的，反之，则称它为无意义的。通常可以把这个参数取作 0.05。

2.2 基于 f-measure 的评价方法

2.2.1 匹配统计量

为了对算法在 PPI 网络中分析的结果进行评估，将算法预测出的 cluster 与已知的标准数据集进行匹配是最直接、最有效的方法。最常用的标准数据集包括 MIPS 中已知蛋白质复合物数据、Nature 和 Science 公开发布的实验方法或者系统方法分析得到的蛋白质复合物数据，以及 GO 数据库中的功能注释信息等。重叠得分(overlap score，OS)具体描述复合物的匹配程度，OS 的计算公式如下[21]：

$$OS=\frac{i^2}{a\times b} \tag{4}$$

式中，i 是指被预测复合物与基准复合物重合的蛋白质数，a 是指被预测复合物中的蛋白质数，b 是指基准复合物中的蛋白质数。OS 的取值范围是[0，1]。OS 为 0 表示被预测复合物中没有蛋白质在基准复合物中出现，二者完全不匹配；OS 为 1 表示被预测复合物中的蛋白质与基准复合物中的蛋白质完全重合，是完美匹配。OS 越高表明预测的复合物越接近于基准复合物。通常情况下，如果它们的匹配程度 OS 超过给定的阈值，就称它们匹配成功。文献[22]指出，当 $0.3\geqslant OS\geqslant 0.2$ 时，大部分没有生物意义的预测复合物已经被过滤掉了。

式(4)主要计算一个预测的 cluster 与已知蛋白质复合物的匹配程度，而聚类算法聚出的整个结果的好坏通常用特异性 Sp(Specificity，简称 Sp)和灵敏度 Sn(Sensitivity，简称 Sn)来表示。特异性是指成功匹配的复合物在所有预测复合物中所占的比重，其定义为 $Sp=[TP/(TP+FP)]$[23]，其中 TP(True Positive，简称 TP)表示匹配成功的蛋白质复合物数，即 OS 大于给定阈值的数量，FP(False Positive，简称 FP)是指预测复合物中没有匹配成功的数量。灵敏度是指匹配成功的复合物在基准复合物中所占的比重，其定义为 $Sn=[TP/(TP+FN)]$，其中 FN(False Negative，简称 FN)表示基准复合物中没有被成功匹配的复合物数量。

由于不同的聚类分析方法得到的聚类数目也不尽相同，聚类的数目越大，则可能与基准复合物匹配的复合物数量也越多，灵敏度越高，而特异性越低；反之亦然。为了避免灵敏度和特异性所带来的偏见，通常采用一个综合评价指标 f-measure 来评估整个算法的结果，其计算公式如下[24]：

$$f\text{-}measure=\frac{2\times Sn\times Sp}{Sn+Sp} \tag{5}$$

2.2.2 基于准确率和查全率的评价

蛋白质与蛋白质相互作用网络通常用一个无向图 $G(V,E)$表示，$v\in V$ 表示蛋白质，边 $(u,v)\in E$ 表示蛋白质 u 和蛋白质 v 之间的相互作用。通过某种特定的聚类分析方法可以获得一种聚类结果，则一个有 n 个蛋白质的 cluster 可表示为集合 $C=\{v_1,v_2,\cdots,v_n\}$。

查全率(recall)是最大匹配时预测 cluster 中匹配成功的蛋白质个数与标准库中具有相应功能的 cluster 包含的蛋白质个数的比值；正确率(precision)是最大匹配时预测 cluster 中匹配成功的蛋白质所占的比值。如，给定一个网络 $G(V,E)$、一个预测 cluster C 和功能 f，k 是 C 中具有功能 f 的蛋白质结点的个数，$Q\subseteq V$ 表示图 G 中具有功能 f 的结点的集合。那么关于 f 的 cluster C 的查全率和正确率被定义如下[25]：

$$recall(C)=\frac{k}{|Q|} \tag{6}$$

$$precision(C)=\frac{k}{|C|} \tag{7}$$

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 Q 中许多结点，极端的情况下，Q 中所有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相同的性质，极端的情况下，一个结点是一个模块，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调和平均值 f-measure[25]来评价预测 cluster 的准确性。

$$f\text{-}measure=2\times\frac{recall\times precision}{recall+precision}=\frac{2\times k}{|C|+|Q|} \tag{8}$$

2.2.3 HF-MEASURE[26]

在如GO这样的系统层次上，蛋白质功能注释结构是分层的，通常用有向无环图(DAG)表示，在该图中每个结点表示一个功能。$Ancestors(x_i)$表示一个结点x_i的祖先集，$L(x_i)$是指从根到x_i的最长路径长度；x_i和x_j的共同祖先被记为$C_A(x_i,x_j)=Ancestors(x_i)\cap Ancestors(x_j)$，且$C_A$中有着最长定位层的结点$x_k$被称为它们最近的祖先，简记为$x_k=\aleph^A(x_i,x_j)$。$x_i$和$x_j$的距离$d(x_i,x_j)$就可以被定义为：

$$d(x_i,x_j)=S_p(x_i,\aleph^A(x_i,x_j))+S_p(x_j,\aleph^A(x_i,x_j)) \tag{9}$$

式中，$S_p(a,b)$表示在DAG中从节点a到节点b最短路径的长度，当$a=b$时，$S_p(a,b)=0$。并且基于本体论结构和距离评估的分层组织给出了新的相似度$hS(x_i,x_j)$如下：

$$hS(x_i,x_j)=\frac{L(\aleph^A(x_i,x_j))}{L(\aleph^A(x_i,x_j))+\mu\times d(x_i,x_j)}=\frac{L^2(\aleph^A(x_i,x_j))}{L^2(\aleph^A(x_i,x_j))+L_A\times d(x_i,x_j)} \tag{10}$$

式中，L_A表示DAG图中从根到每个顶点的最大长度，也即DAG的高度；而$\mu=\frac{1}{\alpha}$是用来调整相似度$hS(x_i,x_j)$的，且依据功能分化的特性，设$\alpha=\frac{L(\aleph^A(x_i,x_j))}{L_A}$。由式(10)可看出任意的两结点$x_i$和$x_j$的$hS(x_i,x_j)$都是0和1之间的正数。

根据相似度越大，功能越相近，定义错误罚分e_p如下：

$$e_p=\begin{cases}0, & \text{if } x_j=x_k\\ -\delta, & \text{if } x_j\neq x_k \text{ and } \aleph^A(x_j,x_k)=x_k\\ \delta, & \text{otherwise}\end{cases} \tag{11}$$

式中，$\delta=\frac{d(x_j,x_k)}{L(\aleph^A(x_j,x_k))+d(x_j,x_k)+L(x_k)}$，$x_j$与$x_k$越接近，罚分越小。对于特殊的情况$x_j=x_k$，$e_p$为0。如果$x_j$是$x_k$的儿子或者后代，就可以获得这个蛋白质更详细的信息。因此，应该给一些奖励。

根据错误罚分的定义，一个蛋白质$v_i\in C$对于具有功能x_k的聚类C组成的贡献定义如下：

$$w_{x_k}(v_i)=\max_{x_j\in F_i}\{hS(x_j,x_k)-e_p\} \tag{12}$$

式中，F_i是蛋白质v_i的注释功能的集合。假设一个聚类C包含n个蛋白质，例如$C=\{v_1,v_2,\cdots,v_n\}$，并且每个蛋白质v_i有一个功能注释集F_i，那么，分配到聚类C的可能的功能被描述为$F_C^*=\bigcup_{i=1}^n F_i$，对每一个功能$x_i\in F_C^*$，可以计算一个相应$hF\text{-}measure(C,x_i)$。

基于上面的定义，$hF\text{-}measure(C,x_i)$的公式被定义如下：

$$hF\text{-}measure(C,x_i)=\frac{\sum_{v_j\in C}w_{x_i}(v_j)+k}{|C|+|Q|} \tag{13}$$

式中，Q是V的子集，其中每一个蛋白质都有功能x_i，且k是C和Q相交的顶点数。

这种评价方法对PPI聚类结果的先验知识要求比较高，不仅要求对各功能模块之间的关系明确，而且对各个蛋白质的所有功能也要知道，目前已有的数据还不能给予其充分的支持。因此该法适于评价的类很有限，有研究者用其评价从基准类中替代蛋白质得到的预测类。

2.3 其他评价分析方法

基于GO识别的聚类的评价方法，除上述外还有一些，例如建立列联表将聚类分析结果与已知蛋白质复合物进行匹配。给定n个已知蛋白质复合物以及m个预测的clusters，该列联表是一个$n*m$的矩阵，其中行i表示第i个已知蛋白质复合物，列j表示第j个预测的cluster。矩阵中每个元素T_{ij}表示复合物i和cluster j的交集的规模。此外，Co-annotation和Co-localization[27]也通常用来分析预测的cluster中蛋白质的同源性和集中性。其值越大表明预测的cluster中的蛋白质的功能相似性越高，或者位于相同组分的可能性越高。给定一个预测的cluster C，其Co-annotation值的计算公式如下：

$$Co\text{-}annotation=\frac{2\sum_{u,v\in C}\mathrm{Sim}(u,v)}{|C|\times(|C|-1)} \tag{14}$$

式中，$|C|$表示预测cluster C中的规模，$\mathrm{Sim}(u,v)$表示的是cluster C中蛋白质u和v的语义相似性。而值的计算式与公式(14)一样，只是前者是基于GO数据库中的生物进程本体而后者则是基于组成成分本体计算的。两个蛋白质之间的语义相似性通常定义为它们对应包含的功能项之间语义相似性的最大值。另外将基于PPI网络的分析结果与基于随机网络的分析结果进行比较也是一个有效的评估方法。随机网络的生成方法有很多种，其中最常见的就是随机洗牌法[28,7]。再者，通过在原始PPI网络中随机地增加或者删除一定数量的相互作用，模拟大规模PPI网络中存在的假阳性和假阴性，可以评价一个算法的鲁棒性；预测的cluster中蛋白质的共表达性、子细胞本地化以及基因显性等也常用来评价预测的cluster的生物意义。

3 改进的Sf-measure

2.2.2节所述的评价方法中，每个模块的f-measure都是独立考虑的，错分的蛋白质无论分到哪个模块都被视为是一样的，这与生物体中各部分组件都有着不同程度的联系是不相符的。而2.2.3节所述评价方法在分析时采用的是蛋白质剔除法，限于GO中数据的有限性以及聚类结果的不可控性，若想将其使用到一般蛋白质模块的预测类的评价上目前还很有难度。再结合蛋白质聚类分析方法的特点：相似性强的蛋白质容易分到一个模块，因此，这里做了一些改进，即在用f-measure评价一个聚类模块的同时，也考虑其与错分模块的关系。一般相似性越大的模块，其关系越密切，但不作分层考虑，而是用另外一种形式来定义错误罚分。

在预测类与标准类进行匹配时，达到最大匹配的标准类和预测类被认为是同一类，这里我们把与该标准类达到次最大匹配的预测类作为主错误划分类。为了避免受到聚类过程中蛋白质间相似度选取的限制，用次最大匹配的蛋白质个数与次最大匹配预测类的规模之比作为与该标准类对应预测类和主错误划分类的相似性，相似性越大，罚分越小，因此用该相似性的倒数来定义罚分ep。如某标准类A与预测类C达到最大匹配，匹配蛋白质个数为k，且与预测类CC满足次最大匹配，匹配蛋白质个数为kk，则其相似性Similar表示为$Similar=\frac{kk}{|CC|}$，当蛋白质被错分到相似性低的模块中时给其以适当惩罚，相似性越小惩罚越重；反之，当蛋白质被错分到相似性较大的模块中时给其以适当奖励，相似性越高，奖励

越大。故其罚分可定义如下：

$$ep=\begin{cases}-\dfrac{kk*precision}{5*|CC|}, & Similar\geqslant 0.25\\ \dfrac{(|CC|-kk)*(1-recall)*(1-precision)}{5*|CC|}, & \text{others}\end{cases} \tag{15}$$

从式(15)可以看出，Similar 若大于阈值则在评价该类时给其以奖励，反之，给以惩罚。该惩罚函数与 cluster 的相似性、查全率和准确率相关，受惩罚时相似性越小，查全率和准确率越低，惩罚越重；受奖励时，一般其查全率不高，此时相似性和准确率越高，奖励值越大。这里根据经验 Similar 的阈值取 0.25。故可进一步定义其 Sf-measure 如下：

$$Sf\text{-}measure=\frac{2}{\frac{1}{recall}+\frac{1}{precision}}-ep \tag{16}$$

实际上 p-value 也可以作类似的改变，这里不再赘述。

4 实验结果

本文所采用的 PPI 数据集来自 MIPS 数据库[18]，评价的聚类结果来自文献[29]。MIPS 数据库提供了一些关于 RNA 基因和其他遗传因素的信息，可以将其作为评价预测蛋白质复合物、蛋白质功能模块的基准。标准比对库的详情见表 1。

表 1 标准比对库信息

聚类个数	聚类蛋白质数	总蛋白质数	最大类	最小类	类平均大小
89	516	1376	35	1	5.76

不同的评价方法基于的评价参数不尽相同，不能进行直接的比较，但这些评价方法都有一定的不足和偏见。图 1 显示的是将同一种聚类结果分别用基于正确率和查全率的 f-measure 及 p-value 进行评价。从图中可以看出预测的聚类结果与标准结果还有很大的差距。f-measure 越大，预测类越好，p-value 越小越好，而图 1 中有些类的 f-measure 和 p-value 都大，有些都小，这就说明这两种方法在评价效果上有时是不一致的，下面分析导致这种不一致的因素。由于生物学家主要关注计算方法所产生的处理结果的生物学意义，因此不对这些评价方法的时间复杂度和空间复杂度做比较分析。

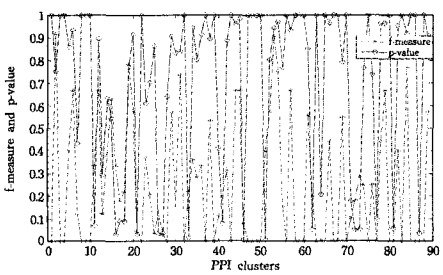

图 1 PPI 聚类的评价结果

表 2 给出了一种聚类结果中最大类、最小类、次大类和次小类的正确率、查全率、f-measure 和 p-value。从较有代表性的这 4 组数据中可以看出较大规模的 cluster 其查全率较高，反之正确率较高。最大类的正确率较低，查全率偏高，说明这个 cluster 预测不准确，p-value 偏大(通常认为 p-value 小于 0.05 的类是我们要找的有生物意义的显著性的类)；次大类的正确率和查全率都偏高，这个 cluster 预测较准确，p-value 偏小；最小类正确率极高，查全率极小，这个预测 cluster 不准确，但 p-value 偏小；次小类的正确率极高、查全率偏低，这个预测 cluster 也不理想，p-value 偏高。这就说明 p-value 在评价大类时较准确，评价小类时具有很大的偶然性。

表 2 基于正确率和查全率以及 p 值的评价结果

预测类	正确率	查全率	f-measure	p-vale
最大类	0.1111	0.7692	0.1942	0.4134
最小类	1	0.0286	0.0557	0.0254
次大类	0.5000	1	0.6667	0.0015
次小类	1	0.2308	0.3750	0.9719

表 3 给出具有不同数目的 clusters 结果的特异性、灵敏度和 f-measure 值。此时的阈值 x 取 0.2。从中可以看出聚类的数目越大，灵敏度越高，特异性越低，反之亦然。因为聚类数目越大，可能与基准 clusters 匹配的数目也就越多。而 f-measure 是一个综合评价指标，从 1、4 组数据，2、3 组数据可以看出其对聚类数目不太敏感。

表 3 基于匹配统计量的评价结果

聚类数目	特异性(Sp)	灵敏度(Sn)	f-measure
21	0.4268	0.1011	0.1642
43	0.4063	0.1461	0.2149
101	0.2277	0.2584	0.2421
168	0.1488	0.2809	0.1946

由于 hf-measure 是采用蛋白质剔除法来进行分析，对于由特定聚类分析方法得出的聚类结果其适用性受限，这里不再仿真，详解可见文献[22]。除此之外上面分析的其他 3 种评价方法均未考虑模块间相似度对聚类结果的影响，实际上蛋白质模块间并没有绝对清楚的界限，它们之间都有着千丝万缕的联系。

表 4 f-measure 与 Sf-measure 的比较

聚类 NO	recall	precision	f-measure	Similar	ep	Sf-measure
2	1	0.8462	0.9167	0	0	0.9167
5	0.2667	0.8000	0.4000	0.0435	0.0281	0.3719
7	0.2308	0.0811	0.1200	0.6000	−0.0097	0.1297
13	0.2286	0.0851	0.1240	0.0694	0.1314	−0.0073
17	0.3333	0.1250	0.1818	0.0270	0.1135	0.0683
19	0.6923	0.9000	0.7826	0.0217	0.0060	0.7766
20	0.4444	0.8000	0.5714	0.7500	−0.1200	0.6914
21	0.3333	0.0256	0.0476	0	0.1299	−0.0823
24	0.3333	0.1500	0.2069	0.0435	0.1084	0.0985
31	0.6364	0.8750	0.7368	0.3750	−0.0656	0.8025
40	0.2308	0.0732	0.1111	0.0139	0.1406	−0.0295
45	0.6667	0.6667	0.6667	0.0417	0.0213	0.6454
53	1	0.8333	0.9091	0	0	0.9091
57	0.5714	0.8000	0.6667	0.2500	−0.0400	0.7067
61	0.5556	0.5556	0.5556	0.6000	−0.0667	0.6222
66	0.6667	0.3333	0.4444	0.1429	0.0318	0.4063
71	0.0909	0.1250	0.1053	0.0488	0.1513	−0.0461
76	0.1579	0.6000	0.2500	0.7500	−0.0900	0.3400
84	0.7143	0.8333	0.7692	0.4000	−0.0677	0.8359
89	0.5000	0.6667	0.5714	0.0278	0.0324	0.5390

表 4 给出了从一个聚类结果中抽取的 20 个 clusters 的 recall、precision、f-measure、Similar、ep、Sf-measure 值。从中可以看出当 Similar 大于等于 0.25 时，聚类的 ep 就为负值，其绝对值表示奖励值，Similar 越大，准确率越高，其奖励值越大；反之，ep 为正，表示惩罚值，Similar 越小，查全率和准确率越小，其惩罚就越重。这样最终得出的 Sf-measure 与原来的 f-measure 相比不再是孤立的，而是受其他相近类的影响。得到奖励的类其 Sf-measure 比原 f-measure 大，受到的惩罚的比其小，这样更能客观地评价蛋白质聚类算法聚出的结果的优劣。

结束语 随着高通量技术的发展和大量的蛋白质相互作用、数据信息的不断累积，PPI 网络的聚类分析方法越来越多，如何更确切地评价这些聚类方法成为生物信息学和系统生物学领域内的一个重要研究问题。本文介绍的评价方法是目前比较有代表性的几种，并针对这几种方法给出了分析与比较。实验结果表明这几种方法确实都有各自的偏见及缺陷，但在一定程度上也能反映聚类分析算法所得结果的好坏。近年来也有不少研究者开始关注评价这个重要的问题，试图寻找一种综合而又全面的方法。

未来该领域的研究可从以下两方面取得突破：

a)改善实验技术，获得更全面、更精确的蛋白质相互作用数据。评价方法中目前最直接的操作就是比对，比对就需要"黄金"标准；而作为被评价的预测类不仅需要知道其模块或复合物间的关系，而且蛋白质本身的信息也是越多越好。

b)数据评价技术本身，目前的评价方法都有对聚类的规模或数量有偏见这种缺陷。研究者应该寻找一种评价方法来把预测类的规模和数量甚至预测类之间的相似性都考虑进去。由于评价在聚类算法研究中的重要地位，评价方法本身将成为该领域研究的一个热点。

参 考 文 献

[1] Barabasi A L, Oltvai Z N. Network biology: understanding the cell's functional organization[J]. Nal. Res. ,2004,5:101-114

[2] Von Mering C, Krause R, Sne B, et al. Comparative Assessment of Large-Scale Data Sets of Protein-Protein Interactions [J]. Nature, 2002, 417(6887): 399-403

[3] Bader G D, Hogue C W. An automated method for finding molecular complexes in large protein interaction networks[J]. BMC Bioinformatics, 2003, 4: 2

[4] Li X L, Tan S, Foo C, et al. Interaction Graph Mining for Protein Complexes Using Local Clique Merging[J]. Genome Informatics, 2005, 16(2): 260-269

[5] Altaf-UI-Amin M, Shinbo Y, Mihara K, et al. Development and implementation of an algorithm for detection of protein complexes in large interaction networks[J]. BMC Bioinformatics, 2006, 7: 207

[6] Li M, Chen J, et al. Modifying the DPClus Algorithm for Identifying Protein Complexes Based on New Topological Structures [J]. BMC Bioinformatics, 2008, 9: 398

[7] Li M, Wang J, Chen J, et al. Identifying the Overlapping Complexes in Protein Interaction Networks[J]. Int. J. DataMing and Bioinformatics, 2010, 4(1): 91-108

[8] King A D, Przulj N, Jurisical I. Protein complexes prediction via cost-based clustering[J]. Bioinformatics, 2004, 20(17): 3013-3020

[9] Ruan J H, Zhang W X. An efficient spectral algorithm for network community discovery and its applications to biological and social network[C]// Perner P, ed. Proceedings of the 7th IEEE International Conference on Data Mining. 2007, 72: 643-648

[10] Luo F, Yang Y, Chen C F, et al. Modular organization of protein interaction networks[J]. Bioinformatics, 2007, 23(2): 207-214

[11] Wang J, Li M, Deng Y, et al. Recent advances in clustering methods for protein interaction networks[J]. BMC Genomics 2010, 11(Suppl 3)

[12] Qi Y, Balem F, Faloutsos C, et al. Protein complex identification by supervised graph local clustering[J]. Bioinformatics, 2008, 24 (13): 250-258

[13] Li X, Wu M, Kwoh C, et al. Computational approaches for detecting protein complexes from protein interaction networks: a survey[J]. BMC Genomics, 2010, 11(suppl 1)

[14] Lei X J, Huang X, Zhang A. Improved artificial bee colony algorithm and its application in Data clustering[C]// The IEEE Fifth International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA2010). Changsha, China, Sep. 2010: 514-521

[15] Lei X J, Huang X, Shi L, et al. Clustering PPI Data based on improved Functional-Flow model through Quantum-behaved PSO [J]. International Journal of Data Mining and Bioinformatics, 2012, 6(1): 42-60

[16] Lei X J, Wu S, Ge L, et al. Clustering and Overlapping Modules Detection in PPI Network Based on IBFO[J]. Proteomics, 2013, 13(2): 278-290

[17] Lei X J, Wu S, Ge L, et al. The Clustering of PPI Data Based on Ant Colony Algorithm[J]. Chinese Journal of Electronics, 2013, 22(1): 118-123

[18] Ashburner M, et al. Gene ontology: tool for the unification of biology[J]. The gene ontology consortium. Nat Genet. , 2000, 25 (1): 25-29

[19] Ucar D, Asur S, Catalyurek U, et al. Improving Functional Modularity in Protein-Protein Interactions Graphs using Hub-Induced Subgraphs [J]. PKDD, 2006, 4213: 371-382

[20] Zhang Y, Zeng E, Li T, et al. Weighted Consensus Clustering for Identifying Functional Modules in Protein-Protein Interaction Networks [C]// Proc of the 2009 Int'I Conf on Machine Learning and Applications. 2009: 539-544

[21] Bader G, Hogue C. An automated method for finding molecular complexes in large protein interaction networks [J]. BMC Bioinformatics, 2003, 4(02)

[22] Krogan N, Cagney G, Yu Hai-yuan, et al. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae [J]. Nature, 2006, 440(7084): 637-643

[23] Song J, How S M. When should Interactome-derived Clusters be Used to Predict Functional Modules and Protein Function [J]. Bioinformatics, 2009, 25(29): 3143-3150

[24] LI Xiao-li, Wu Min, Kwoh C K, et al. Computational approaches for detecting protein complexes from protein interaction networks: a survey[J]. BMC Genomics, 2010, 11(suppl 1)

[25] Zhang A D. Protein Interaction Networks [M]. New York, USA: Cambridge University Press, 2009: 44-46

[26] Li Min, Wu Xue-hong, Wang Jian-xin, et al. A New Measurement for Evaluating Clusters in Protein Interaction Networks [C]// BIMBM'11 Proceedingsof the 2001 IEEE International Conference on Bioinformations and Biomedicine. 2011: 63-68

[27] Wu M, Li X, Kwoh C. Algorithms for Detecting Protein Complexes in PPI Networks: An Evaluation Study[C]// Proceedings of Third IAPR International Conference of Pattern Recognition in Bioinformatics (PRIB 2008). Australia, 2008: 135-146

[28] Altaf-UI-Amin M, Shinbo Y, Mihara K, et al. Development and Implementation of an Algorithm for Detection of Protein Complexes in Large Interaction Networks[J]. BMC Bioinformatics, 2006, 7: 207

[29] 雷秀娟，田建芳. 蛋白质相互作用网络的蜂群信息流聚类模型与算法[J]. 计算机学报，2012，35(1)：134-145