一种改进的基于加权网络的蛋白质复合物识别算法

赵碧海[1,2] 熊慧军[1] 倪问尹[2,3] 刘志兵[1] 胡 赛[1]
(长沙学院信息与计算科学系 长沙 410003)[1] (中南大学信息科学与工程学院 长沙 410083)[2]
(长沙学院图书馆 长沙 410003)[3]

摘 要 不断增长的蛋白质相互作用数据使我们能够采用计算方法预测蛋白质复合物。然而,由于实验条件和技术的限制,现有的PPI网络中包含噪声。为了降低噪声对复合物识别所产生的负面影响,提出了一种改进的名为WPC的算法,用于从加权网络中识别蛋白质复合物。给定一个选定节点,所有邻居节点组成候选集,候选集中节点的邻居节点组成邻居集。对于候选集中的节点,若该节点在候选集与邻居集间的加权比低于设定阈值,则将该点剔除。处理后的候选集被标记为复合物。对于没有包含在任何复合物中的节点,如果节点在某一复合物内的平均加权度超过一个自适应的阈值,则将其补充到该复合物中。对WPC算法和现有的几种经典蛋白质复合物识别算法的性能进行了综合比较。实验结果表明,WPC算法的性能优于几种对比的复合物识别算法。

关键词 平均加权度,蛋白质复合物,蛋白质相互作用网络,加权比
中图法分类号 TP301 **文献标识码** A

Improved Weighted-network Based Algorithm for Predicting Protein Complexes

ZHAO Bi-hai[1,2] XIONG Hui-jun[1] NI Wen-yin[2,3] LIU Zhi-bing[1] HU Sai[1]
(Department of Information and Computing Science, Changsha University, Changsha 410003, China)[1]
(School of Information Science and Engineering, Central South University, Changsha 410083, China)[2]
(Library of Changsha University, Changsha 410003, China)[3]

Abstract The increasing amount of protein-protein interaction (PPI) data has enabled us to predict protein complexes. Due to the limitation of experimental conditions and techniques, there is a lot of noise in the PPI networks. To reduce the negative effects of noise on protein complex prediction, a new improved method named WPC (Weighted-network based method for Predicting protein Complexes) was proposed. Given a selected node, candidate set consists of all neighbors of the node and neighbor set consists of neighbors of all nodes in the candidate set. If the weighted ratio of a node between the candidate set and the neighbor set is lower than a threshold, the node is removed from the candidate set. After repeating the process for all nodes in the candidate set, the candidate set is represented as a complex. For a node not being included in any complexes, if its average weighted degree within a complex exceeds a self-adjustment threshold, WPC adds the node to the complex. A comprehensive comparison among the competitive algorithms and WPC was made. Experimental results show that WPC outperforms the state-of-the-art methods.

Keywords Average weighted degree, Protein complex, Protein-protein interaction network, Weighted ratio

1 引言

蛋白质是细胞内发生的生物过程的基本构建单位。它们通过相互作用发挥自己的功能,由此构成一个非常广泛的网络交互系统。高通量的方法[1]产生了大量蛋白质相互作用数据。蛋白质复合物由多种蛋白质相互作用组装的蛋白质分子聚合而成,是大分子组织的基本单元,并在单个基因产物进行有用的细胞功能中发挥关键作用。目前,蛋白质组装技术,如串联亲和纯化[2]等,可用于蛋白质复合物的预测。然而,这些实验方法都不能得到令人满意的结果。计算方法作为实验方法的补充,能够从可用的PPI数据中预测蛋白质复合物。已经提出了多种图聚类算法用于识别蛋白质相互作用图中高度连接的节点。基于团渗透方法,CFinder[3]算法试图找到所有的k团链,k团链是具有k个顶点并且完全连通的子集。另一种常见的聚类算法(MCL)是马尔可夫聚类算法[4]。MCL算法通过构建邻接矩阵,在蛋白质相互作用网络图中模拟流体流动,通过计算区分出高流动性区域和低流动性区域。Amin等提出了DPClus[5]算法,它是一种从PPI网络提取密集连接区域的聚类算法,通过扩展其在原图中的邻居节点来实现重叠复合物的挖掘。SPICi[6]算法原理类似于DPClus。不同之处在于,SPICi算法利用一个集群扩展方法,使用不同的选种标准。Liu等提出了一种基于最大团的聚类算法

到稿日期:2013-07-30 返修日期:2013-10-13 本文受国家自然科学基金(61232001),湖南省十二五规划课题(XJK011CXJ002)资助。
赵碧海(1980-),男,博士生,讲师,CCF会员,主要研究方向为生物信息学、数据挖掘,E-mail:bihaizhao@163.com;胡 赛(1978-),女,硕士,讲师,主要研究方向为概率统计、生物信息学,E-mail:husaiccsu@163.com(通信作者)。

(CMC)[7]用于从加权 PPI 网络中发现复合物。蛋白质相互作用间的可靠性使用一个迭代计分的权值表示。最近,Nepusz 等提出了一种名为 ClusterONE[8] 的蛋白质复合物识别算法。ClusterONE 算法能从 PPI 网络找到重叠的蛋白复合物。

前述的算法都是基于稠密子图挖掘的思想。为了进一步突破,我们应该考虑复合物固有的组织特性。Gavin[9] 等已对复合物组织结构做了深入研究。研究结果显示,一个复合物应该由一个核心组成部分和附件构成。核心蛋白质是高度共表达的,每个附件绑定到核,从而形成具有生物特性的复合物。受到这种发现的启发,提出了一些基于核-附件的算法,如 Core-Attachment[10],COACH[11]。

虽然基于计算方法的蛋白复合物预测已有了巨大的进步,但是如何准确地识别蛋白复合物仍然是一个瓶颈。一个尖锐的问题是如何衡量 PPI 网络的可靠性,减少和容忍噪声的负面影响。研究表明,只有 30%～50%的高通量相互作用是实际存在的[12]。因此,发现蛋白质复合物的关键之一是将噪声数据从蛋白质相互作用数据中分离出来。

基于这个问题,本文根据蛋白质相互作用网络的拓扑特性构造一个加权网络,权值的大小表示相互作用的可靠性。权值为 0 意味着测定的相互作用可能不存在,从而将这组相互作用从网络中移除。通过这种方式可以有效地减少和控制噪声对于蛋白复合物预测所产生的负面影响。

针对蛋白质复合物识别,本文提出了一种名为 WPC (Weighted network based method for Predicting protein Complexes)的算法,用于从加权网络中识别蛋白质复合物。现有的复合物识别算法一般将挖掘的稠密子图视为一个复合物,强调子图内部的连接,而忽视内部节点与外部节点的联系。结合软件工程中的模块化思想,我们认为,一个挖掘的子图被标识为复合物,需要具备两个条件:

(1)子图内部节点之间应该具有较强的凝聚力。

(2)子图内部节点与外部邻居子图的耦合度应该比较低。

为了验证 WPC 算法的性能,本文选取了 5 个经典的具有代表性的复合物识别算法进行对比分析,包括:ClusterONE[8]、MCL[4]、CFinder[3]、CMC[7]、SPICi[6]。这 5 种算法都可以从加权蛋白质相互作用网络中识别复合物。我们将 WPC 和 5 种对比算法在酵母 PPI 网络上运行。实验结果表明,WPC 算法的性能要优于其他算法。

2 WPC 算法

在描述 WPC 算法原理前,本节将介绍一些与算法有关的概念。

定义 1(加权网络) 设$G=(V,E)$表示 PPI 网络,其中 V 是一组顶点(蛋白质),E 是一组边(相互作用)。对应的加权网络定义为 $G=(G,W_E)$,其中 $W_E:E\to[0,1]$是一个表征边权值大小的函数。

定义 2(加权度) 给定加权网络 $G=(V,E,W)$和节点 $v_a\in V$,其中 $V=\{v_1,v_2,\cdots,v_n\}$,$E=\{e_1,e_2,\cdots,e_m\}$,$W=\{w(e_1),w(e_2),\cdots,w(e_m)\}$。节点 v_a 在加权网络 G 内的加权度定义为:

$$WD(v_a,G)=\sum_{i=1}^{n}w(v_a,v_i),(v_a,v_i)\in E$$

定义 3(平均加权度) 给定加权网络 $G=(V,E,W)$及节点 $v_a\in V$,其中 $V=\{v_1,v_2,\cdots,v_n\}$,$E=\{e_1,e_2,\cdots,e_m\}$,$W=\{w(e_1),w(e_2),\cdots,w(e_m)\}$。节点 v_a 在加权网络 G 内的平均加权度定义为:

$$AWD(v_a,G)=\sum_{i=1}^{n}w(v_a,v_i)/|V|,(v_a,v_i)\in E$$

定义 4(加权比) 给定加权网络 $G=(V,E,W)$和节点 $v_a\in V$,其中 $V=\{v_1,v_2,\cdots,v_n\}$,$E=\{e_1,e_2,\cdots,e_m\}$,$W=\{w(e_1),w(e_2),\cdots,w(e_m)\}$。$G_N=(V_N,E_N,W_N)$,其中,

$$V_N=\{v_i\,|\,dis(v_i,v_a)=2\}\cup\{v_a\}$$

$$E_N=\{(v_i,v_a)\,|\,dis(v_i,v_a)=2\}$$

$$W_N=\{w(v_i,v_a)\,|\,dis(v_i,v_a)=2\}$$

其中,$dis(v_i,v_a)$表示 v_i 和 v_a 之间的距离。

节点 v_a 的加权比定义为:

$$WR(v_a,G)=\frac{WD(v_a,G)}{WD(v_a,G)+WD(v_a,G_N)}$$

定义 5(加权稠密度) 给定加权网络 $G=(V,E,W)$,其中 $V=\{v_1,v_2,\cdots,v_n\}$,$E=\{e_1,e_2,\cdots,e_m\}$,$W=\{w(e_1),w(e_2),\cdots,w(e_m)\}$,加权网络 G 的加权稠密度定义为:

$$WDensity(G)=2\times\sum_{i=1}^{m}w(e_i)/(|V|\times(|V|-1))$$

WPC 算法主要分为 4 个阶段实施:

(1)输入加权 PPI 网络,对于加权网络中的任意节点,计算该节点在其邻居图(由邻居节点及其与之相连的边构成的子图)中的平均加权度。加权网络中平均加权度为 0 的节点和权值为 0 的边被作为噪声移除。处理之后的加权以邻接矩阵的形式输出。在形成邻接矩阵的过程中,节点按照平均加权度降序排列。若两节点平均加权度相同,再按度的大小降序排列。邻接矩阵中,所有节点平均加权度的平均值 *Self_Thres* 作为自适应参数输出,用于后续的蛋白质复合物识别。

(2)从邻接矩阵中按照顺序选择蛋白质节点作为种子节点,种子节点的一级邻居节点形成候选集。对于候选集中的节点,若节点的加权比低于设定的阈值,则从候选集中移除该节点,同时该节点失去成为种子节点的机会。这一过程对所有的节点不断重复,从而形成蛋白质复合物集合。

(3)对于每一个非种子节点,依次与第(2)阶段形成的每一个复合物对比,若节点在复合物内的平均加权度高于第(1)阶段形成的自适应阈值 *Self_Thres*,则将节点补充至该复合物中。

(4)计算经过第(3)阶段处理后形成的复合物相互间的重叠率,当重叠率超过指定的阈值时,移除那些具有更低的加权稠密度或更少蛋白质的复合物。

以下是 WPC 算法的伪代码描述:

算法 1 构造邻接矩阵算法

```
输入:加权 PPI 网络 G=(V,E)
输出:矩阵 Matrix,自适应阈值 Self_Thres
(1)FOR all v_i∈V DO
    AWD_i= AWD(v_i,G_Ni);// G_Ni为 v_i 的邻居图
(2)Noise_process(WG);
(3)Self_Thres=AWD_i,i∈[1,|V|]
(4)Matrix= Convert(WG);
(5)Sort(Matrix);
(6)Return Matrix,Self_Thres
```

算法 1 中,Noise_process()用于噪声处理,即去除平均加权度为 0 的节点(蛋白质)和权值为 0 的边(相互作用)。Sort()根据

节点的平均加权度和度的大小排序。排序结果对最终识别结果产生影响，平均加权度高的节点优先遍历有助于提高复合物识别的准确率。

算法 2　复合物识别算法

输入：邻接矩阵 Matrix，加权比阈值 Thres_WR

输出：候选复合物 PC

(1)PC=Φ;
(2)FOR all v_i in Matrix DO
(3)　IF TAG(v_i)=REJECTED CONTINUE;
(4)　CS={v_i|dis(v_i,v_a)=1}∪{v_i};
(5)　FOR all v_j∈CS DO
(6)　　IF WR(v_j,CS)<Thres_WR THEN BEGIN
(7)　　　Remove(v_j);
(8)　　TAG(v_i)=REJECTED;END
(9)　IF Size(CS)>1 THEN
(10)　　PC= PC∪CS;
(11)FOR all TAG(v_k)=REJECTED DO
(12)　FOR all pc_i∈PC DO
(13)　　IF AWD(v_k,pc_i)> Self_Thres THEN
(14)　　pc_i= pc_i∪{v_k};
(15)Return PC

初始时，邻接矩阵中所有节点都享有成为种子节点的机会，算法按照顺序选取种子节点。对于选定的种子节点，节点的所有一级邻居及与该节点的相互作用形成邻居图。对于邻居图中的节点，如果加权比低于设定的阈值，则将节点从邻居图中剔除，并标记为 REJECTED。被标记为 REJECTED 且尚未遍历的节点将失去成为种子节点的机会。处理后的邻居图成为候选复合物加入集合 *PC* 中。

算法中(11)－(14)步实现 WPC 算法的第(3)个阶段：对于所有的非种子节点，若平均加权度超过阈值 *Self_Thres*，则将其补充到复合物中。阈值 *Self_Thres* 根据输入的网络自适应调整。

WPC 算法的最后阶段是重叠处理。有些重叠模块可能有重要的生物特性，但对于重叠程度非常高的群体应当处理。对于一对重叠率超过阈值的复合物，算法将丢弃具有较小稠密度或较小尺寸的复合物。本算法中，复合物的加权稠密度根据定义 5 计算，这有别于其他的复合物识别算法。本文中重叠率的阈值设为 0.8[8]，重叠率的计算公式如下[10]：

$$NA(A,B)=|A\cap B|^2/|A||B| \tag{1}$$

3　实验结果与分析

本文实验数据采用酵母 PPI 网络，因为酵母 PPI 网络是所有物种中最完整和可靠的。本文在 Krogan[12] 数据集上运行了 WPC 算法和其他 5 种算法：CFinder、ClusterONE、MCL、CMC、SPICi。Krogan 数据集包含 3672 个蛋白质和 14317 组蛋白质间的相互作用。数据集中都去掉了自我相互作用和重复的相互作用。为了评估预测得到的蛋白质复合物，本文采用 CYC2008[13] 作为基准复合物集，CYC2008 包含 408 个通过生物方法预测得到的复合物，每个复合物包含 2 个或 2 个以上的蛋白质。

本文从 F-measure 和 P-Value 两个方面对结果进行详细分析。为了公平对比，其他 5 种算法的参数均按照作者的建议设定为最优值。

作为评估算法性能的方式之一，本文将各个算法识别的复合物与已知的基准复合物集进行匹配，并且计算相应的 Precision、Recall 和 F-measure 值。

假设算法识别的复合物集为 *PC*，基准复合物集为 *BC*，对于一个识别的复合物 $pc\in PC$ 和一个基准复合物 $bc\in BC$。根据式(1)，如果 $NA(pc, bc)\geqslant t$，则认为 *pc* 与 *bc* 匹配。其中，*t* 为一个预先设定的阈值，一般地，该阈值设置为 0.2[10]。

算法的准确率(Precision)和召回率(Recall)是用来评估复合物识别算法的两个重要指标。准确率是指算法识别的复合物中被匹配的部分所占比重；召回率是指已知复合物中被匹配的部分所占比重，如式(2)所示：

$$Precision=\frac{TP}{|PC|},Recall=\frac{TN}{|BC|} \tag{2}$$

其中，*TP*(True Positive)表示算法识别的复合物中与已知复合物匹配的数量，*TN*(True Negative)表示已知复合物中被匹配的数量。综合考虑准确率和召回率两个方面，提出了综合评价指标 F-measure，它是准确率和召回率的调和平均值，如式(3)所示：

$$F\text{-}measure=\frac{2\times Precision\times Recall}{Precision+Recall} \tag{3}$$

表 1 给出了各种算法预测得到的复合物的基本信息。其中，#*PM* 表示算法识别的复合物总数，*Full* 是指识别的复合物中被完全匹配的个数。

表 1　各算法预测的复合物的基本信息

Algorithms	#PM	Full	TN	TP
WPC	**366**	**14**	**149**	**240**
ClusterONE	240	8	118	126
MCL	724	14	173	151
CMC	168	7	95	92
SPICi	378	6	138	118
CFinder	121	9	61	55

从表 1 不难看出，WPC 算法识别的复合物中有 240 个被匹配，在所有算法中匹配数量最多。完全匹配数目(14)也是所有算法中最多的。基准复合物集中被 WPC 算法匹配的复合物数量是 149，仅次于 MCL 算法，但是 WPC 识别的复合物总数(366)几乎只有 MCL 算法识别的复合物(724)数量的一半。

图 1 显示了各算法的运行结果，包括 Precision、Recall 和 F-measure(从左至右依次为 WPC、ClusterONE、MCL、CMC、SPICi、CFinder)。

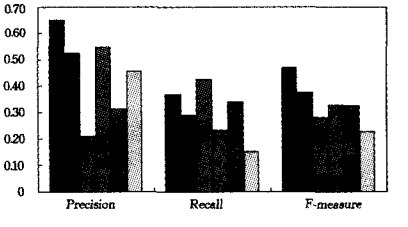

图 1　各算法的 Precision、Recall 和 F-measure

图 1 显示，WPC 算法具有最高的 F-measure 和 Precision 值，Recall 值仅次于 MCL 算法。WPC 算法的 F-measure 分别比 ClusterONE、MCL、CMC、SPICi 和 CFinder 提高了 26.02%，68.1%，43.84%，44.76%和 108.88%。

为了测试 WPC 算法识别的复合物的生物意义，我们采用功能富集分析(P-value)。P-value 被认为是衡量预测的复

合物是一个真正的蛋白质复合物的可能性。预测的蛋白复合物的低 P-value 值表明该复合物具有很高的统计学意义。可以设置一个阈值区分显著的复合物和非显著的复合物。本文中，该阈值设为 0.01[14]。表 2 给出了不同算法 P-value 对比结果。

表 2 各算法预测的复合物的显著性统计信息

Algorithms	#PM	#SC	Proportion	P-score
WPC	**366**	**305**	**83.33%**	**11.5**
ClusterONE	240	160	80%	9.72
MCL	724	263	36.33%	5.56
CMC	168	141	83.93%	8.52
SPICi	378	180	47.62%	7.12
CFinder	121	77	63.63%	8.1

在表 2 中，#*PC* 是预测的复合物的数目，#*SC* 是显著的复合物的数目，P-score 是所有显著性复合物的 $-\lg$(P-value) 的平均值。一般而言，较高的显著性复合物比例和 P-score 值表明，在相同的蛋白质复合物的蛋白质往往有着较高的功能相似，所以它们可以被用来评估预测的蛋白复合物的整体质量。而 P-score 能更加真实、全面地反映识别的复合物的整体生物统计特性。

其中，#*PM* 表示算法识别的功能模块总数，#*SC* 表示显著的蛋白质复合物数量，即 P-value$<$0.01 的复合物数量。表 2 显示，WPC 算法识别的复合物中显著性复合物的比例接近 CMC 算法，高于其他 4 种算法。WPC 算法的 P-score 值相比 ClusterONE、MCL、CMC、SPICi 和 CFinder 分别提高了 18.31%、106.83%、34.98%、61.52%和 41.98%。对比 WPC 算法和 CMC 算法，虽然 CMC 算法的显著性复合物比例略高于 WPC 算法，但是 CMC 算法的 P-score 明显低于 WPC 算法，原因在于，WPC 算法识别的复合物具有更小的 P-value 值。

由此可见，WPC 算法预测的复合物具有最强的生物统计意义。

在算法 2 中，为根据加权比对候选集中节点进行筛选，本文引入了一个自定义的参数 *Thres_WR*。*Thres_WR* 用于描述一个节点在子图中的加权比，根据定义 4 可知，*Thres_WR*$\in$[0,1]。图 2 显示了当 *Thres_WR* 取不同值时，F-measure 的变化情况。

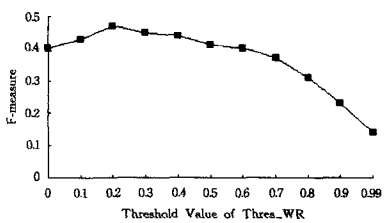

图 2 参数 *Thres_WR* 的影响

从图 2 不难看出，当 *Thres_WR*$\in$[0,0.2)时，对应的 F-measure不断升高，*Thres_WR*$\in$[0.2,1)，F-measure 逐渐降低。*Thres_WR* 在区间[0.19,0.23]内，获得最高的 F-measure值 0.47。当 *Thres_WR* 分别取值 0.19,0.2,0.21 和 0.22 时，完全匹配的复合物数量为 13 个，而 *Thres_WR* 取值 0.23 时，完全匹配的复合物数量为 14 个，因此，本文将 *Thres_WR* 设为 0.23。

结束语 考虑到蛋白质间相互网络中存在噪声，本文提出了一种改进的基于加权网络的复合物识别算法 WPC。不同于目前的识别算法，WPC 通过加权比判定一个稠密子图是否可以表达为高内聚、低耦合的蛋白质复合物。为了评估算法的性能，本文将 WPC 算法与其他 5 种经典的复合物识别算法进行了对比，包括 F-measure 和 P-value 值。实验结果表明，WPC 具有更高的识别准确率，识别的复合物具有更强的生物统计特性。

参 考 文 献

[1] Gavin A C, Bosche M, Krause R, et al. Functional organization of the yeast proteome by systematic analysis of protein complexes [J]. Nature, 2002, 415(6868): 141-147

[2] Puig O, Caspary F, Rigaut G, et al. The tandem affinity purification (TAP) method: a general procedure of protein complex purification [J]. Methods, 2001, 24: 218-229

[3] Adamcsek B, et al. CFinder: locating cliques and overlapping modules in biological networks [J]. Bioinformatics, 2006, 22(8): 1021-1023

[4] Enright A, Dongen S, Ouzounis C. An efficient algorithm for large-scale detection of protein families [J]. Nucleic Acids Research, 2002, 30(7): 1575-1584

[5] 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] Jiang P, Singh M. SPICi: a fast clustering algorithm for large biological networks [J]. Bioinformatics, 2010, 26(8): 1105-1111

[7] Liu G, Wong L, Chua H N. Complex discovery from weighted PPI networks[J]. Bioinformatics, 2009, 25: 1891-1897

[8] Nepusz T, Yu H, Paccanaro A. Detecting overlapping protein complexes in protein-protein interaction networks [J]. Nature Methods, 2012, 9(5): 471-475

[9] Gavin A, Aloy P, Grandi P, et al. Proteome survey reveals modularity of the yeast cell machinery [J]. Nature, 2006, 440(7084): 631-636

[10] Leung H, Xiang Q, Yiu S, et al. Predicting protein complexes from PPI data: a core-attachment approach [J]. Journal of Computational Biology, 2009, 16(2): 133-144

[11] Wu M, Li X, Kwoh C, et al. A core-attachment based method to detect protein complexes in ppi networks [J]. BMC Bioinformatics, 2009, 10: 169

[12] Krogan N, Cagney G, Yu H, et al. Global landscape of protein complexes in the yeast Saccharomyces Cerevisiae [J]. Nature, 2006, 440: 637-643

[13] Pu S, Wong J, Turner B, et al. Up-to-date catalogues of yeast protein complexes [J]. Nucleic Acids Research, 2009, 37(3): 825-831

[14] Hu H, Yan X, Huang Y, et al. Mining coherent dense subgraphs across massive biological networks for functional discovery [J]. Bioinformatics, 2005, 21: 213-221