文章编号:1000-5862(2013)03-0273-06

基于拓扑和生物特征的权重网络中络合物抽取

于凤英,杨志豪*,林鸿飞
(大连理工大学计算机科学与技术学院,辽宁 大连 116024)

摘要:蛋白质关系网络的权重设置对蛋白质络合物的抽取有着较大的影响. 综合考虑蛋白质关系网络的拓扑结构特征和生物信息特征,提出一种新的权重设置策略,并向原有蛋白质关系网络中添加高可信度关系. 基于修正后的蛋白质关系网络,用监督学习方法抽取蛋白质络合物,在 DIP 数据集上,F 值达到 0.570 5.

关键词:权重网络;权重设置;蛋白质关系;蛋白质络合物;监督学习
中图分类号:TP 391 **文献标志码**:A

0 引言

蛋白质相互关系在不同的生物进程中扮演着重要的角色. 近年来,随着酵母菌双杂交实验、串联亲和纯化实验、质朴分析实验等高通量实验的发展和广泛应用,产生了大量的蛋白质关系数据,也随之出现了多个公开的蛋白质关系数据集,从而使得在蛋白质层面上进行蛋白质相互关系的研究成为可能. 蛋白质关系数据集对于解密蛋白质分子机制,预测蛋白质功能,抽取蛋白质络合物等方面都有着重要的影响. 然而,通过高通量实验得到的蛋白质关系数据集中存在着假阳性关系,在这些数据集中仅有 30% ~50% 的关系被认为是生物相关的,而且各种蛋白质关系数据集之间的交集非常小[1]. 有研究[2]表明,在一些人类蛋白质关系数据集中,超过 50% 的关系被认为是假阳性的. 由此可以看到,在高通量实验获得的蛋白质关系数据集中,假阳性关系所占比例非常大,从这些蛋白质关系数据集中抽取的蛋白质络合物和预测的蛋白质功能,正确率不高. 因此,修正原始的蛋白质关系网络使其更加可靠就显得尤为重要.

近年来,各种各样修正蛋白质关系网络的方法相继被提出. C. Brun 等[3]提出了 CDdistance 方法,H. N. Chua 等[4]提出了 FSWeight 方法,Liu Guimei 等[5]提出了 AdjustCD 方法,这些方法是基于考虑 2 个蛋白质共同邻接点的数量来评价蛋白质关系的可靠性. 唐楠等[6]提出计算蛋白质对间的基因序列相似度,谢东等[7]提出先计算基因有向无环图中基因对间的相似度,再计算蛋白质对间的基因序列相似度. 这几种方法都基于考虑蛋白质对间的生物信息来评价蛋白质关系的可靠性. 本文针对蛋白质关系数据集中存在的假阳性和假阴性问题,提出一种蛋白质关系权重网络设置策略. 实验表明基于修正后的蛋白质关系数据集,利用监督学习方法抽取蛋白质络合物,能有效地提升抽取络合物的准确性.

1 方法描述

1.1 权重网络构建

DIP 酵母菌关系数据集[8]中的蛋白质作为结点,蛋白质关系作为边,构造一个无权重蛋白质关系网络. 该网络中的关系是通过高通量实验得到的,存在着大量的假阳性关系. 为了修正关系网络,需要一种衡量蛋白质关系可靠性的策略. 蛋白质与蛋白质之间通过相互作用来完成某一种生物功能. 当某个蛋白质参与某项生物功能时,该蛋白质的邻接点也趋向于参与这项生物功能. 因此,当 2 个蛋白质拥有更多的共同邻接点时,这 2 个蛋白质越趋向于可以

收稿日期:2012-12-15
基金项目:国家自然科学基金(61272373,61070098,60973068),国家“863”计划(2006AA01Z151)和中央高校基本科研业务费专项基金(DUT10JS09)资助项目.
作者简介:杨志豪(1973-),男,黑龙江大庆人,副教授,博士,博士生导师,主要从事文本挖掘和信息检索方面的研究.

共同参与某项生物功能. 基因本体已成为生物信息领域中一个极为重要的资源. 基因本体数据库 GO[9] 按照语义包含 3 个不同的领域: 分子功能、生物过程、细胞组件. 每个领域又由数万条表示基因属性的概念词 *term* 组成一个有向无环图. 在基因有向无环图中,子孙基因是由祖先基因派生出来的,子孙基因拥有的功能,祖先基因也都拥有. 为了方便比较,每个领域都被添加一个初始点作为该领域的类根点. 离类根点越远的基因 *term* 标注的蛋白质个数越少,基因 *term* 之间的差距越大. 由于每个蛋白质具有多种分子功能,参与多项生物过程,因而蛋白质通常由一个描述该蛋白质所具有的已知的所有功能、属性以及在细胞中存在位置的标注集合来标注. 当 2 个标注集合相同的基因 *term* 越多时,标注集合间的相似度越大.

本文综合考虑蛋白质对间的拓扑相似度和基因语义相似度,将这 2 种相似度有机地结合起来,得到一种更全面的评分策略($weight$),来评价蛋白质关系网络中关系的可信度.

考虑 A 蛋白质由标注集合 T_A 标注,B 蛋白质由标注集合 T_B 标注,$com(A,B)$ 为 A 和 B 蛋白质共同标注的 *term* 集合,$sim(i)$ 为被基因 $term_i$ 所标注过的蛋白质集合. $sim_{\max}$ 为所有参与标注蛋白质的基因 *term* 中所标注过的蛋白质个数最多的 *term* 所标注的蛋白质的集合,$com(A,B)$ 为 A 和 B 蛋白质共同邻接蛋白质集合,绝对值表示取集合中元素的个数. $TP(A,B)$ 为 A 和 B 蛋白质共同邻接蛋白质的个数.

$$weight(A,B) = \left(-|com(A,B)| \times \log\left(\min_{i \in com(A,B)} |sim(i)| / |sim_{\max}| \right) + TP(A,B) \right)^{1/2}, \quad (1)$$

由(1)式不难发现,当 A 和 B 蛋白质拥有的共同邻接蛋白质个数增多,共同标注基因 *term* 个数增多,或公共基因集合中标注过的蛋白质个数最少的 *term* 越远离有向无环图的类根点,A 和 B 蛋白质关系的权重值都会增大.

用(1)式对 DIP 酵母菌关系数据集中原有的蛋白质关系进行评分,然后向原有蛋白质关系网络中添加权重值大于某一阈值的蛋白质关系,从而构造一个可信度较高的蛋白质权重关系网络. 利用该网络使得在监督学习方法抽取蛋白质络合物中扩张完全子图时能够选择可信度更高的蛋白质添加进去,从而使得预测到的蛋白质络合物更加准确.

1.2 监督学习方法抽取络合物

1.2.1 *训练集构建* 构建 3 类别训练集: 正例、中间例、负例. 正例采用 Aloy、MIPS06、SGD、TAP06 共 4 个标准络合物集合的并集. 由于标准络合物集合和蛋白质关系网络来自不同的数据集,因而标准络合物中有些蛋白质可能并没有在 DIP 酵母菌关系数据集中出现过,将这样的蛋白质作为噪音过滤掉. 由于蛋白质络合物是指具有 2 个以上功能相关的多肽链通过二硫键或其他蛋白质相互作用所形成的蛋白质大分子结构. 故将正例中只含有 1 个蛋白质的络合物也作为噪音过滤掉. 最后剩余 732 个蛋白质络合物作为训练过程中的正例. 由于 COACH 方法[10] 是目前经典的络合物抽取算法中效果较好的,故中间例选用 COACH 方法生成的蛋白质络合物中除去与正例相匹配的,剩下的 425 个蛋白质络合物作为训练过程中的中间例. 由于标准络合物集合中络合物的大小成指数分布,故负例采用与正例络合物大小成比例、随机生成的方式,生成 2 000 个蛋白质络合物作为训练过程中的负例.

1.2.2 *特征构建及模型训练* 采用监督学习方法学习蛋白质络合物的正例、中间例、负例的 11 个拓扑结构特征. 采用回归模型 Regression[11] 对训练集的这 11 个拓扑结构特征进行训练,得到一个分类模型. 在 11 个拓扑结构特征中,有 4 个是非权重特征: 无权重网络图密度、无权重网络图度的平均值、无权重网络图度的中位数和无权重网络图度的相关系数; 7 个是权重特征: 权重网络图密度、权重网络图度的统计量的最大值、权重网络图度的统计量的平均值、权重网络图边权统计量的平均值、低权重差值比、高权重差值比和权重聚类系数.

1.2.3 *蛋白质络合物抽取* 用 E. Tomita 等[12] 提出的 Cliques 算法抽取蛋白质关系网络中大小 $\geqslant 3$ 的完全子图. 由于蛋白质关系网络的密度非常大,抽取的完全子图的重合率也非常高,故需要对完全子图进行过滤. 先将完全子图使用训练集训练得到的分类模型进行评分,并按评分高低逐个比较. 若 2 个完全子图的重合率大于某一阈值($\geqslant 3$ 的完全子图过滤阈值为 2),则将评分低的完全子图过滤掉. 过滤后的完全子图作为初始集合,逐个进行迭代扩张. 对任意完全子图,每次逐个迭代加入使其评分得到最大提高的邻接点,直到没有邻接点的加入可以使该完全子图的评分再提高为止. 扩张后的完全子图作为候选络合物集合.

由于网络中的 2 个完全子图在扩张过程中,有可能加入过多相同的邻接点,因而候选络合物的重合率依然较高,这就需要对候选络合物进行进一步的过滤. 先将候选络合物使用训练集训练得到的分类模型进行评分,并按评分高低逐个比较. 若 2 个候选络合物的重合度 $overlap(C_i, C_j) >$ 阈值 0.8(阈值分别取 0.1,0.2,…,0.9,进行 9 组实验,选取使得 F 值达到最大的阈值为最终阈值,此处为 0.8) ,则合并这 2 个候选络合物得到新的络合物 C_k,比较 C_i、C_j、C_k 3 个络合物的评分,最终保留评分最高的那个络合物

$$overlap(C_i, C_j) = |C_i \cap C_j| / |C_i \cup C_j|. \quad (2)$$

2　实验结果与分析

2.1　评价指标

蛋白质络合物抽取的常用评价指标有: 准确率(*precision*) 、召回率(*recall*) 、F 值、敏感值(*sensitivity*) 、阳性预测值(*PPV*) 和精确率(*accuracy*) 等[13]. 在介绍这些指标之前,先介绍一下匹配值. 当预测络合物 p 与标准络合物 b 的匹配值 $NA(p, b) > 0.25$ 时,认为此预测络合物 p 预测正确,预测到 1 个与标准络合物 b 相匹配的络合物. V_p 表示 1 个预测络合物的蛋白质集合,V_b 表示 1 个标准络合物的蛋白质集合,绝对值表示集合的大小,预测络合物 p 与标准络合物 b 的匹配值计算方法为

$$NA(p, b) = |V_p \cap V_b|^2 / (|V_p| \times |V_b|).$$

准确率衡量预测到的络合物集合中有多少个络合物被预测正确. 召回率衡量标准络合物集合中有多少个络合物被召回. F 值为准确率与召回率的调和平均数. $|N_p|$ 表示预测络合物集合中至少与 1 个标准络合物相匹配的络合物个数,$|N_b|$ 表示标准络合物集合中被预测到的标准络合物的个数. $|P|$ 表示预测到的络合物的个数,$|B|$ 表示标准络合物的个数,且

$$N_p = \{p \mid p \in P, \exists b \in B, NA(p, b) \geqslant 0.25\},$$
$$N_b = \{b \mid b \in B, \exists p \in P, NA(p, b) \geqslant 0.25\},$$
$$precision = |N_p| / |P|, \quad recall = |N_b| / |B|,$$
$$f\text{-}score = (2 \times precision \times recall) / (precision + recall).$$

敏感值、阳性预测值、精确率是近几年提出的络合物预测的评价指标. 敏感值衡量预测络合物集合对标准络合物集合的覆盖程度,阳性预测值衡量预测络合物为真实络合物的可能性,精确率是敏感值和阳性预测值的几何平均数. 需要说明的是,这一系列的评价指标并不完全合理,只能作为参考. 例如络合物抽取算法发现一个覆盖了标准络合物里大多数蛋白质分子的络合物,敏感值会特别高. MIPS 数据集本身作为测试集,阳性预测值只能达到 0.772,而准确率和召回率可以达到 1. 因此,F 值系列的评价指标更合理.

2.2　实验结果

蛋白质关系是蛋白质分子与蛋白质分子之间为共同完成某项生物功能而发生的相互联系. 这种相互联系,不仅仅体现在它拥有很强的生物意义,而且在拓扑结构上也有它自身的特点. 为此设置了仅利用蛋白质关系的拓扑结构特征,仅利用蛋白质关系的生物信息特征和将蛋白质关系的拓扑结构特征、生物信息特征按(1) 式有机结合起来的 3 组对比实验(对比实验结果如表 1 所示) . 对比实验的结果显示,仅利用蛋白质关系的生物信息特征要比仅利用蛋白质关系的拓扑结构特征的效果要好,这就说明,衡量蛋白质关系的可靠性,主要还是依靠蛋白质对间的生物信息特征. 但当将蛋白质关系对间的生物信息特征和拓扑结构特征按照(1) 式有机结合起来的时候,实验性能较好(正如本文 2.1 节中评价指标中提到的,*accuracy* 系列指标不太合理,只能作为辅助指标,*f-score* 才是当前最主要的评价指标,因此,这里及以下各实验提到的性能比较的都是 *f-score*) ,这就说明蛋白质对间的拓扑结构特征对修正蛋白质关系网络也有影响. 可靠的蛋白质关系应当是在拓扑结构和生物功能上联系均很紧密的关系. 在修正蛋白质关系网络时,合理地结合这 2 种特征,才能使得修正后的蛋白质关系网络更加可靠.

表 1　拓扑相似度、语义相似度、混合相似度的性能

方法	*num*	*precision*	*recall*	*f-score*	*sensitivity*	*PPV*	*accuracy*
拓扑相似度	789	0.471 5	0.422 1	0.445 4	0.402 4	0.677 2	0.522 1
语义相似度	861	0.623 7	0.476 8	0.540 4	0.472 9	0.714 7	0.581 4
混合相似度	848	0.641 5	0.485 0	0.552 4	0.437 7	0.740 1	0.569 1

由于 DIP 酵母菌关系数据集是通过高通量实验获得的,这就导致了有很多蛋白质对间本应该是有关系的,但是通过高通量实验,并没有发现这部分关系,那么向原有蛋白质关系网络中添加部分可信度高的关系就显得尤为重要. 表 2 为保持原有 DIP 酵母菌关系网络不变和向原有 DIP 酵母菌关系网络中添加权重值大于某一阈值的关系后的网络用于监督学习方法抽取蛋白质络合物的实验性能对比. 实验结果显示,向原有关系网络中添加关系虽然能使实验效果有所提升,但关系并不是添加的越多越好. 当向原有关系网络中添加的关系过多时,可能会引入噪音,把不可靠的关系也添加进网络,这样就会影响蛋白质关系网络的可靠性. 当向原有关系网络中添加的关系过少时,仍有较多可靠性很高的蛋白质关系未被添加进网络. 由此可见,向原有蛋白质关系网络中添加关系的数量是有限制的. 实验结果显示,将权重值 >7 的 2 282 组新的蛋白质关系添加进原有网络,得到修正后的蛋白质关系网络较可靠,实验性能较好.

表 2 添加不同权重值关系的性能

方法	*num*	*precision*	*recall*	*f-score*	*sensitivity*	*PPV*	*accuracy*
原网络	848	0.641 5	0.485 0	0.552 4	0.437 7	0.740 1	0.569 1
>4	845	0.616 6	0.501 4	0.553 0	0.490 2	0.720 6	0.594 3
>5	822	0.633 8	0.508 2	0.564 1	0.484 7	0.730 2	0.594 9
>6	844	0.654 0	0.502 7	0.568 5	0.456 7	0.729 8	0.577 3
>7	872	0.654 8	0.505 5	0.570 5	0.460 9	0.729 3	0.579 8
>8	863	0.651 2	0.480 9	0.553 2	0.453 1	0.730 7	0.575 4
>9	838	0.643 2	0.486 3	0.553 9	0.459 0	0.728 0	0.578 1

近年来随着高通量实验的发展和广泛应用,产生了大量的蛋白质关系数据,随之出现了多个公开的蛋白质关系数据集和多种修正蛋白质关系数据的方法. 表 3 是本文提出的修正网络策略与目前效果很好的修正网络策略在 DIP 数据集、Krogan 数据集[14]、Gavin 数据集[15]上性能的对比. 这 3 个数据集中训练集均按照本文前边提到的训练集构造方法构造. 表 4 是 3 个数据集的详细信息. 为了体现实验对比的公平性,各种修正网络策略均没有向原网络中添加蛋白质关系. 实验结果显示,该方法在 3 个数据集上修正的蛋白质关系网络都是比较可靠的,实验性能都是较好的. 在修正关系网络策略中,CDdistance、FSWeight、AdjustCD 方法是目前流行的基于拓扑结构特征修正蛋白质关系网络的方法,而唐楠等[6]的方法则是基于生物信息特征修正蛋白质关系网络的方法.

表 3 不同权重方法的性能比较

方法	数据集	*num*	*precision*	*recall*	*f-score*	*sensitivity*	*PPV*	*accuracy*
Ours	DIP	848	0.641 5	0.485 0	0.552 4	0.437 7	0.740 1	0.569 1
	Krogan	293	0.716 7	0.453 3	0.555 3	0.496 5	0.739 8	0.606 1
	Gavin	336	0.761 9	0.513 7	0.613 7	0.531 7	0.714 6	0.616 4
CDdistance	DIP	806	0.416 9	0.390 7	0.403 4	0.447 6	0.642 1	0.536 1
	Krogan	306	0.542 5	0.379 2	0.446 4	0.464 8	0.708 4	0.573 8
	Gavin	332	0.572 3	0.416 1	0.481 9	0.576 4	0.675 1	0.623 8
FSWeight	DIP	816	0.414 2	0.389 3	0.401 4	0.447 4	0.641 7	0.535 8
	Krogan	309	0.530 7	0.382 7	0.444 7	0.466 3	0.708 4	0.574 7
	Gavin	337	0.560 8	0.414 4	0.476 6	0.582 7	0.680 8	0.629 8
AdjustCD	DIP	810	0.434 6	0.382 5	0.406 9	0.449 1	0.640 9	0.536 5
	Krogan	297	0.532 0	0.377 4	0.441 6	0.467 0	0.705 0	0.573 8
	Gavin	334	0.571 9	0.416 1	0.481 7	0.575 9	0.676 7	0.624 3
谢东等	DIP	784	0.390 3	0.340 2	0.363 5	0.555 7	0.655 0	0.603 3
	Krogan	288	0.534 7	0.373 9	0.440 1	0.521 9	0.696 4	0.602 9
	Gavin	329	0.522 8	0.422 9	0.467 6	0.657 5	0.703 0	0.679 9
唐楠等	DIP	815	0.471 2	0.405 7	0.436 0	0.500 3	0.659 8	0.574 5
	Krogan	303	0.580 9	0.409 2	0.480 1	0.492 2	0.716 8	0.594 0
	Gavin	323	0.569 7	0.436 6	0.494 4	0.613 8	0.691 7	0.651 6

表 4　数据集信息

数据集	蛋白质数量	蛋白质关系数量	训练集		
			正例数量	中间例数量	负例数量
DIP	1 428	17 201	732	425	2 000
Krogan	2 675	7 080	567	141	1 600
Gavin	1 430	6 531	584	86	1 600

随着高通量实验的发展和广泛应用,产生了大量的蛋白质关系数据,也随之出现了多种蛋白质络合物的抽取方法. 目前 G. Bader 等[16] 提出的 MCODE 方法(基于局部密度的聚类方法),G. Liu 等[5] 提出的 CMC 方法(基于最大完全子图的方法),Xiaoli 等提出的 COACH(基于蛋白质络合物的"核—附属"结构方法)方法,都是目前络合物抽取方法中的经典算法,实验效果也较好. 表 5 是本文提出的抽取络合物的方法与 MCODE、CMC、COACH 方法结果的对比. 由于本文采用的是监督学习方法抽取蛋白质络合物,为了体现与各种方法对比的公平性,这里采用了 5 倍交叉验证实验获得最终的络合物预测结果. 首先,将 732 个正例平均分成 5 个集合 $\{C_1, C_2, C_3, C_4, C_5\}$. 然后分别进行 5 组交叉验证实验,在每组交叉验证实验中,采用该集合中的 4 份作为训练集,一份作为测试集,并将预测结果中与训练集络合物匹配值(NA) >0.9 的络合物去掉. 最后将 5 组预测结果进行合并,由于络合物重叠率过高,进行了自去重,去掉重合度($overlap$) >0.6 的络合物. 经过 5 倍交叉验证实验获得的实验结果就避免了预测集与训练集重合的问题. 将 5 倍交叉验证实验结果与 MCODE、CMC、COACH 方法结果进行对比,结果显示,该方法性能优于前述的若干方法.

表 5　不同方法在 DIP 语料上的性能

方法	*num*	*precision*	*recall*	*f-score*	*sensitivity*	*PPV*	*accuracy*
Ours	572	0.629 4	0.489 1	0.550 4	0.436 5	0.736 4	0.567 0
MCODE	59	0.542 4	0.088 8	0.152 6	0.235 1	0.637 2	0.387 1
CMC	173	0.630 1	0.312 8	0.418 1	0.349 9	0.722 3	0.502 7
COACH	747	0.431 1	0.468 6	0.449 0	0.432 7	0.692 2	0.547 3

3　总结和展望

本论文针对监督学习方法抽取蛋白质络合物提出了一种更好的修正关系网络策略,该策略将蛋白质关系网络的拓扑结构特征和生物信息特征有效的结合起来,不仅刻画了关系网络结构上的联系,还刻画了关系网络在生物意义上的联系.

本论文提出的修正蛋白质关系网络的策略虽然使得实验性能得到明显提高,但是,这种修正策略也还是传统意义上基于规则的方法. 下一步将尝试引用各种特征,用监督学习方法来自动设置蛋白质关系网络的权重.

4　参考文献

[1] Deane C M, Salwiński Ł, Xenarios I, et al. Protein interactions: two methods for assessment of the reliability of high throughput observations [J]. Molecular & Cellular Proteomics, 2002, 1(5): 349-356.

[2] Srihari S, Ning K, Leong H W. Refining Markov clustering for protein complex prediction by incorporating core-attachment structure [J]. Genome Informatics Series, 2009, 23(1): 159-168.

[3] Brun C, Chevenet F, Martin D, et al. Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network [J]. Genome biology, 2004, 5(1): 6.

[4] Chua Honnian, Sung Wingkin, Wong Limsoon. Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions [J]. Bioinformatics, 2006, 22(13): 1623-1630.

[5] Liu Guimei, Wong Limsoon, Chua Honnian. Complex discovery from weighted PPI networks [J]. Bioinformatics, 2009, 25(15): 1891-1897.

[6] 唐楠,杨志豪,吴佳金,等. 基于监督学习的蛋白质络合物抽取方法 [J]. 广西师范大学学报: 自然科学版, 2011, 29(2): 174-179.

[7] Wang Jian, Xie Dong, Lin Hongfei, et al. Identifying protein complexes from PPI networks using go semantic similarity [EB/OL]. [2012-10-13]. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6120506&tag=1.

[8] Xenarios I, Salwinski L, Duan X J, et al. DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions [J]. Nucleic acids re-

search,2002,30(1):303-305.

[9] Harris M,Clark J,Ireland A,et al. The gene ontology (GO) database and informatics resource [J]. Nucleic acids research,2004,32: D258-D261.

[10] Wu Min,Li Xiaoli,Kwoh C K,et al. A core-attachment based method to detect protein complexes in PPI networks [J]. BMC bioinformatics,2009,10(1):169.

[11] Cossock D,Zhang Tong. Subset ranking using regression [EB/OL]. [2012-10-16]. http://link.springer.com/chapter/10.1007%2F11776420_44#page-1.

[12] Tomita E,Tanaka A,Takahashi H. The worst-case time complexity for generating all maximal cliques and computational experiments [J]. Theoretical Computer Science,2006,363(1):28-42.

[13] Brohee S,Van Helden J. Evaluation of clustering algorithms for protein-protein interaction networks [J]. BMC bioinformatics,2006,7(1):488.

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

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

[16] Bader G,Hogue C. An automated method for finding molecular complexes in large protein interaction networks [J]. BMC bioinformatics,2003,4(1):2.

[17] 何文译,林鸿飞,杨亮. 基于群体智慧的电影排序模型[J]. 江西师范大学学报: 自然科学版,2013,37(2):136-141.

[18] 任巨伟,杨亮,林鸿飞. 情感图式构造及其在文体情感计算中的应用 [J]. 江西师范大学学报: 自然科学版,2013,37(2):130-135.

Complex Extraction from the Weighted Network Based on Topological and Biological Characteristics

YU Feng-ying,YANG Zhi-hao*,LIN Hong-fei

(School of Computer Science and Technology,Dalian University of Technology,Dalian Liaoning 116024,China)

Abstract: The weight setting of protein interaction network has a great effect on protein complex identification. A better strategy which considers both topological characteristics and biological characteristics of protein interaction network has been provided. Furthermore,some credible interactions have been added into the original network. Then,the updated protein interaction network has been used for complex identification based on a supervised method. *F*-score reached 0.570 5 in the DIP dataset.

Key words: weighted network; weight setting; protein interaction; protein complex; supervised learning

(责任编辑: 冉小晓)