一种基于极大团的蛋白质相互作用预测方法①

王建新② 蔡 钊 李 敏

(中南大学信息科学与工程学院 长沙 410083)

摘 要 针对大规模、高通量方法识别出的蛋白质相互作用数据集包含噪声较高的问题,根据蛋白质相互作用网络中噪声数据的特点和网络拓扑属性,提出了一种基于极大团的相互作用预测方法以提高识别出的这种数据集的可靠性。该方法通过蛋白质与蛋白质团之间的关联程度来预测蛋白质间是否存在相互作用,同时结合噪声数据在网络中的特点将预测结果放大并将伪相互作用分离出来,最后可获得可靠性较高的预测数据集。实验结果显示,所提出的方法不仅可以预测更多相互作用而且具有更高的可靠性。

关键词 蛋白质相互作用网络, 相互作用预测, 假阳性, 极大团

0 引 言

人类基因组测序的完成标志着一个生物学研究新时代——后基因组时代的来临。生物学家们的研究热点又回到了蛋白质上,全基因组的序列信息并不足以解释及推测细胞的各种生命现象,因为蛋白质才是细胞活性及功能的最终执行者。蛋白质之间的相互作用对细胞中的所有生理过程都会产生重要影响[1],因此识别出各种蛋白质间的相互作用完全集对于理解生物细胞中的生理过程至关重要,这已成为现代生物学的研究热点之一[2,3]。迄今为止,已研究出了很多用于识别蛋白质相互作用的实验方法,这些方法大致可以分为两类:小规模(低通量)方法,如 co-IP[4-7];大规模(高通量)方法,如酵母双杂交[8]和串联亲和纯化[9]。

在蛋白质相互作用网络较大时,需要检测的蛋白质的数量会非常大,如果采用小规模的方法来检测相互作用,所耗费的时间和其它资源都会非常大,所以这时一般采用大规模的检测方法来识别蛋白质的相互作用。但是,与小规模方法相比大规模方法识别出的数据集更容易出现错误,其可靠性相对较低[10,11]。这些错误分为两种[12]:假阳性(false positive, FP)和假阴性(false negative, FN)。假阳性是指实验检测出的在真实的蛋白质网络中并不存在的两蛋白质间的相互作用;假阴性是指实验没有检测出来而真实的蛋白质网络中却存在的相互作用。文献[12]用实验证明:在用高通量实验检测出来的数据集中,这两种错误出现的概率相差很大,其中绝大部分(甚至高达 92.5%)错误是 FN。所以一般认为,通过大规模方法获得的数据集中,蛋白质间如果不存在相互作用,在一定程度上并不能说明该相互作用在真实的网络中不存在,而更可能的是实验没有检测出这个相互作用[12]。因此非常有必要设计有效的计算方法来纠正数据集中的大部分假阴性。

近年来已经提出了很多种蛋白质相互作用识别的计算方法。在早期,计算方法主要是通过识别物种的 motifs 残基来预测蛋白质相互作用[13]。之后,许多根据基因组序列的蛋白质相互作用预测算法不断被提出,例如:基于相互作用蛋白质间氨基酸序列中突变的相互关系分析[14],基于相邻基因和基因排列保守区域的查找[15],基于基因融合或“罗塞塔碑”[16,17]的方法。在 2001 年 Sprinzak 和 Margalit 首先提出基于功能域的预测方法[18]后,又有许多人提出了基于功能域的改进预测方法[19-24]。

最近越来越多的基于蛋白质相互作用网络自身特点的相互作用预测算法被提出,这些算法运算速度比较快,并且预测的可靠性也比较高,例如基于聚类的 RNSC 算法[25]和基于 k-core 的 MCODE 算法[26]。这两个算法本来是用于预测蛋白质复合物的,但是根据在同一个复合物内的任意两个蛋白质都存在相互作用的普遍观点,这些算法也可以用于蛋白质相互作用预测。2006 年,Haiyuan Yu 等人从蛋白质相互作用网络拓扑属性角度,提出了一种基

① 国家自然科学基金(60433020)、新世纪优秀人才支持计划(No. NCET-05-0683)和长江学者和创新团队发展计划(No. IRT0661)资助项目。
② 男,1969 年生,博士,教授;研究方向:计算机算法、网络优化理论、生物信息学;联系人,E-mail: jxwang@mail.csu.edu.cn
(收稿日期:2007-11-23)

于瑕团(defective clique, DC)的蛋白质相互作用预测算法(DC算法)。根据两两极大团的相互交叠关系,DC算法把两两关联比较紧密的极大团合并成更大的极大团,合并过程中要添加的边即为预测的相互作用。与以前其它算法相比,DC算法预测出的蛋白质相互作用数量和可靠性都有了很大提高。但是这种基于图中团与团之间相互交叠的相互作用预测方法对网络拓扑结构要求特别严格,致使能预测出的相互作用数量相对有限;而且由于高通量数据集本身含有较高的噪声,所以当对两个团之间的交叠比例要求降低时,其预测的可靠性就会急剧下降。

由于目前基于蛋白质相互作用网络的预测方法都没有分析FP在网络中的分布特点,这使得它们的预测性能在一定程度上受到了限制。本文从蛋白质相互作用网络的拓扑结构出发,结合高通量数据集中FP在网络中的分布特点以及蛋白质相互作用网络的相关特性,提出了一种在高通量数据集中基于极大团的FN识别方法,该方法在能预测出更多相互作用的同时,其预测结果集可以保持较高的可靠性。实验结果表明,本文的方法预测出的相互作用在数量和可靠性上,比DC算法都有了较大的提高。

1 预测方法

为了方便讨论,下面将给出本文中会涉及的相关概念和定义。

本文将蛋白质相互作用网络看作一个无向图 $G, G = (V, E)$: V 表示图中所有顶点的集合,E 表示图 G 中所有边的集合,$E \subseteq V \times V$。图 G 中每个顶点表示一个蛋白质,每条边表示一对蛋白质的相互作用,与顶点 v 相关联边的数量称作顶点 v 的度 (v)。

对于图 $G = (V, E)$, $\exists V' \subseteq V$, 如果顶点集 V 导出的子图 $G' = (V', E')$ 是完全图,则称子图 G' 为图 G 中的团;如果 $\neg \exists v \in V$ 且 $v \notin V'$, 顶点集 $V' \cup \{v\}$ 的导出子图 $G'' = (V' \cup \{v\}, E'')$ 是完全图,则称 G' 为图 G 中的极大团 T。极大团 T 所包含顶点的个数表示极大团 T 的大小 $|T|$,图 G 中顶点 v 与极大团 T 相关联边的数量表示顶点 v 到极大团 T 的关联度 $D(v_t)$。

1.1 蛋白质相互作用网络的无标度性质

作为复杂系统之一的蛋白质相互作用网络,它也属于一种无标度网络,网络中顶点度分布遵循“幂次定律”[27]。通过大规模、高通量的实验方法识别出的蛋白质相互作用数据集可靠性不高,假阳性所占比率比较大,比如在酵母中甚至高达50%[28]。但是我们通过统计分析发现,绝大部分的假阳性出现在网络中度较低的顶点上,在各顶点度上的分布大致也遵循“幂次定律”。图1中描述了来自Uetz(包含1387条相互作用,1314个蛋白质,234条假阳性)[29]和Ito(包含4363条相互作用,3255个蛋白质,673条假阳性)[30]的高通量蛋白质相互作用网络中假阳性在各种网络中各顶点度上的分布,纵坐标表示假阳性在各度值顶点上的关联度总数,横坐标表示网络中顶点度的范围。Uetz和Ito中顶点度的最大值分别是125和28,而且分别有88.2%和90.8%的假阳性与网络中度小于63和14的顶点相关联。从图1中可以发现,在这两个高通量相互作用网络中,如果忽略与度较高顶点相关联的那小部分假阳性,其余假阳性在各顶点度上的分布总体上逐步下降,近似地遵循“幂次定律”。

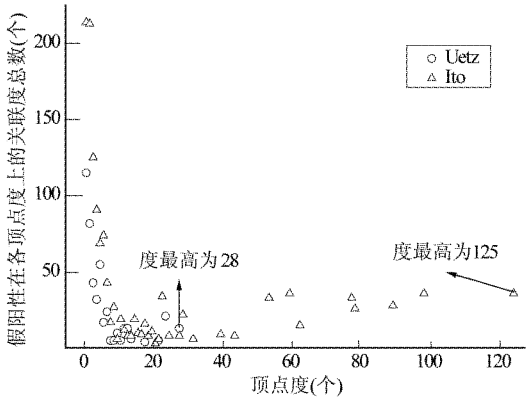

图1 假阳性在高通量网络中各顶点度上的分布

此外,蛋白质相互作用网络也遵循复杂网络的优先连接(preferential attachment)[31,32]原则,即拥有大的连接数、度(v)大的顶点拥有更大的概率连接到新的顶点。

1.2 算法 VTC(Vertex to Clique)

在蛋白质相互作用网络中,如果一个蛋白质跟不包含该蛋白质的复合物关联比较密切,那么这个蛋白质跟复合物中的蛋白质功能相似,它们之间存在相互作用。这种蛋白质与复合物的关系可以转换成图论中顶点与极大团的关系,即如果图中的某个极大团中大部分顶点与非该极大团的顶点相关联,则这个顶点与极大团中非关联的顶点之间也存在关联边,我们应该把这些丢失的边重新加入图中使这个极大团扩展成更大的极大团。例如,对于一个包含5个顶点

A、B、C、D、E 的极大团 T 和图中不包含在 T 中的顶点 v，v 与顶点 A、B、C 相关联，$D(v_t)/|T| = 3/5 = 0.6$，这时我们认为顶点 v 与极大团 T 的关联比较紧密，则应该把顶点 v 到 D 和 E 丢失的边重新添加到图中，使极大团 T 扩大成更大的极大团 $(T + v)$，这些添加的新边也就是预测出的新相互作用。

基于这种思想，VTC 算法首先比较图中任意极大团 T 和非极大团中的顶点 v。如果它们关联比较紧密，则把丢失的边保存到初始为空的相互作用预测结果集合中，并且立即把该边添加到图中，使得在下一次对与该边相关联顶点预测时，可以提高该顶点与极大团的关联度，从而放大预测结果，特别是放大了假阳性参与的预测。因为对于度比较高的顶点，边的密度比较高，增加的那部分关联度对于提高边的密度效果并不明显，只有对于度比较低的这部分顶点，其放大效果才比较突出。而前文提到网络中绝大部分的假阳性出现在顶点度较低范围内的顶点上，所以由这些大部分错误信息所预测出的伪相互作用(假阳性)数也得到了放大，应当删除这些伪相互作用。

根据前文提到的无标度网络优先连接原则可以删除预测结果集合中放大的假阳性。优先连接原则认为对于相互作用网络中度比较低的顶点，在预测结果集中与该顶点相关联的边也应该比较少；相反，如果比较多，则可以认为预测出错，即与该顶点相关联的相互作用是假阳性，应该丢弃预测结果集中与该顶点相关联的所有相互作用。

本算法的主要思想是基于图中顶点到极大团的关系来补全图中丢失的边，使小的极大团扩充成更大的极大团。所以首先我们可以利用现有的高效极大团枚举算法枚举出图中所有的极大团；然后预测任意顶点和极大团之间是否存在丢失的边，如果有则在预测过程中把预测的新边立即添加到图中，并把所有预测的新边保存到一个初始为空的预测结果集中；最后根据优先连接原则删除预测结果集中的假阳性。其详细描述如图 2 所示。

算法 VTC 中有 4 个阈值，即 δ、Q、ε 和 K，其中 δ 用于控制顶点与极大团的交叠比率，Q 用于删除预测结果集中的假阳性，ε 用于控制顶点与极大团的最小关联度，K 表示只对大于等于 K 的极大团进行预测。之所以设 ε 和 K 这两个阈值是因为在蛋白质相互作用网络中关联度一样时，相比于和大复合物的关联，顶点与小复合物的关联更有可能趋向于随机事件，而且小复合物的密度在现有蛋白质相互作用网络中比较低，用基于拓扑结构的方法预测相互作用容易产生假阳性[25]。

算法 VTC
输入：图 G
输出：所有预测的相互作用
1. 枚举图中所有的极大团
2. 对于图中的第一个顶点 v
   对于图中的第一个大于等于 K 的极大团 T
   If $D(v_t) < \varepsilon$ AND $D(v_t)/|T| >= \delta$ then
   预测 v 与 T 中无关联的顶点之间有一条边并把新边添加到原始图和预测结果集中；
3. 对于预测结果集中的第一个顶点 v
   If $\frac{v\text{在预测结果集中的度}}{v\text{在原始数据集中的度}} > Q$ then
   把顶点 v 以及与 v 相关联的边从预测结果集中删除；
4. 输出预测结果集中的相互作用

图 2 算法 VTC 的描述

算法 VTC 第一步枚举图中所有的极大团是一个 NP 完全问题，解决这个问题的算法目前只能在非多项式时间内完成，它的时间复杂度上界是 $O(nm\mu)$[33]，其中 n 是顶点的个数，m 是边的条数，μ 是极大团的个数；第二步的时间复杂度是 $O(n\mu)$；第三步的时间复杂度为 $O(n)$。所以本算法的时间复杂度为 $O(nm\mu + n\mu + n)$。DC 算法的时间复杂度为 $O(nm\mu + \mu^2)$[12]，当图比较稠密时，枚举图中所有极大团的时间复杂度将会以非多项式增长，即 μ 以非多项式方式增长。所以当图比较稠密时，算法 VTC 在时间复杂度方面比 DC 算法有很大的优势。为了进一步提高算法的运行速度，在算法执行的第一步前可以对网络数据集进行预处理，对网络中的顶点度小于 δ 的顶点和其关联的边循环删除，这样可以大幅度地降低枚举所有极大团的时间。

2 实验结果

在 S. cerevisiae 蛋白质相互作用网络的高通量数据集上，本文分别对 DC、VTC 算法进行了实验分析，并把预测结果与 gold standard 相互作用集作了比较。gold standard 相互作用集具有较高的阳性(MIPS 数据库复合物记录中确定的处于同一个复合物中的蛋白质相互作用对[4])置信度或阴性(不同亚细胞位置无相互作用的蛋白质对[34])置信度，比如文献[2]

中的数据集。本文使用的 gold standard 数据集[2]包含 8250 对真阳性和 2708622 对真阴性相互作用。

预测算法的性能好坏一般是通过预测精度公式 Precision = $P_+/(P_+ + P_-)$ 来评价,然而在大多数生物实验中,由于阳性和阴性之间样本的偏差,精度值往往不能正确评价预测方法的性能,但相互作用预测似然比不会出现这种情况[12]。本文将使用相互作用预测似然比[2]作为评价算法是否有效的标准,具体定义为:

$$L = \frac{(P_+/G_+)}{(P_-/G_-)} \tag{1}$$

其中 P_+ 是指真阳性的数量,即预测结果集与 gold standard 中真阳性集重叠的相互作用数; P_- 是指假阴性的数量,即预测结果集与 gold standard 中真阴性集重叠的相互作用数; G_+ 是指在 gold standard 中所有真阳性的总数; G_- 是指在 gold standard 中所有真阴性的总数。

L 值越大表示预测的结果集与 gold standard 数据集中真阳性重叠越多或者说与真阴性重叠越少,这也意味着预测的结果更好。

2.1 DC 和 VTC 在 S.cerevisiae 数据集上的性能

本文使用来自文献[35]的 S. cerevisiae 蛋白质相互作用网络高通量数据集测试算法 DC 和 VTC 的性能,该数据集包含 6645 对相互作用,2283 个蛋白质。

在算法 DC 中有两个参数,两两极大团的交叠规模 k 和非交叠规模 l[12],这两个参数可以在一定范围内取值,在这里我们取其在预测出的相互作用数量和可靠性都较高的值,即 $k = 4, l = 3$ (文献[12]中使用的也是这两个值)。运行 DC 算法后,共预测出 388 对相互作用,其中 61 对与 gold standard 的真阳性数据集重叠,24 对与 gold standard 的真阴性数据集重叠,于是可以得出 $L = 834.5$。

在算法 VTC 第二步中有 3 个阈值,即 ε、K 和 δ,其大小是根据 DC 算法中相关参数取值,由(2)、(3)、(4)推导得出:

$$K = k + 1 = 5 \tag{2}$$

$$\varepsilon \geqslant k = 4 \tag{3}$$

$$\delta = \frac{k}{k+l} = \frac{1}{1 + l/k} \geqslant \frac{1}{1 + 3/4} = \frac{4}{7} \tag{4}$$

第三步中阈值 Q 根据图的密度来设定,即 $Q = |E|/|V|$。运行算法 VTC 后,共预测出 558 条边,把这些边添加到原图中后,使得原图中的极大团个数减少了 377 个,这符合我们把小的极大团扩展成更大的极大团使得极大团个数减少的出发点。在这 558 对相互作用中有 133 对与 gold standard 的真阳性数据集重叠,20 对与 gold standard 真阴性数据集重叠,从而可计算出似然比 $L = 2183.3$。这比 DC 算法($L = 834.5$)和贝叶斯方法[10]($L < 400$[2])得出的 L 值都要高很多;此外在这种预测可靠性高的情况下,VTC 所预测出的相互作用数量为 558,也要比 DC 的 388 高得多(详见表 1)。

表 1 DC 算法和 VTC 算法在来自文献[35]中 S.cerevisiae 高通量数据集上的性能

算法	新边总数	真阳性	真阴性	L
DC	388	61	24	834.5
VTC	558	133	20	2183.3

由(2)可知 $K = k + 1$,为了方便 VTC 与 DC 在其他参数值上性能的比较,本文把 K 和 k 分别作为 VTC 和 DC 算法的变量,固定其它参数,即设 DC 中 $4 \leqslant k \leqslant 7$、$l = 3$,VTC 中 $5 \leqslant K \leqslant 8$、$\varepsilon = 4$、$\delta = 4/7$,其结果描述如图 3。从图 3 中可以看出,在各参数上,VTC 的性能都要优于 DC。当 $K > 5, k > 4$ 时,VTC 和 DC 的性能比较接近,但当 $K = 5, k = 4$ 时,VTC 的预测性能比 DC 有了极大的提高,这说明 VTC 算法在 $K > 5$ 时,对假阳性数据放大的效果并不明显,使得这时预测性能跟 DC 相近。因为高通量蛋白质相互作用网络中,在度较低的范围内,假阳性在各顶点度上的分布总体逐步下降,所以对这个度范围内的假阳性的放大也是逐步下降的;在度较高的范围内,由于假阳性相对非常少,VTC 对假阳性几乎没有放大效用,这符合本文 1.1 节的分析。

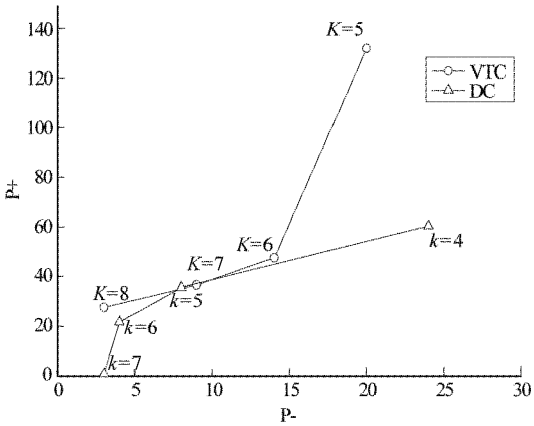

图 3 算法 VTC 和 DC 在其它参数上的预测结果

评价预测质量的另一个重要指标是功能富集,功能富集越高表示算法的性能越好。本文采用与文献[12]相同的方法来计算功能富集,即预测结果集

中的蛋白质相互作用对在功能上相似的频率与在蛋白质基因组中所期望频率的比值,其中功能分类信息来自 MIPS 数据库。

算法 VTC 和 DC 预测出来的结果集的似然率和功能富集的分布统计如表 2 和表 3 所示。从表 2 和表 3 中可以看出,即使是对于很小的交叠规模(overlap size),DC 算法和 VTC 算法预测出的边都大部分存在于 gold standard 的真阳性数据集中,并且它们的功能富集要比相互作用对的平均值高很多。由于 VTC 算法主要是针对度不高的顶点消除噪声数据,所以在交叠规模等于某些值时,DC 算法预测结果集的功能富集要比 VTC 算法稍高些,但是这部分数据只占结果集中很少的一部分。对于整个数据集来说,VTC 算法中删除的真阳性比假阳性要少得多,因此对于整个结果集上的功能富集算法 VTC 要优于算法 DC。

本文对算法 VTC 和 DC 在其它不同规模的高通量数据集上的性能也进行了测试,其实验结果如表 4所示。从表 4 可以看出,VTC 预测的相互作用明显多于 DC,并且包含了 DC 预测结果集中大约 80% ~ 90%的相互作用。这说明存在大量的相互作用能同时被算法 VTC 和 DC 预测出。为了分析这些同时被两个算法预测出来的相互作用是否具有统计意义,本文使用由超几何聚集分布(hyper-geometric cumulative)计算的 P 值来进行统计分析。

表 2 算法 DC 构造完全团的预测边集似然率和功能富集分布

交叠规模	新边总数	真阳性	真阴性	L	观测频率	功能富集
4	231	25	16	513.0	88	7.937
5	94	14	4	1149.1	45	9.973
6	50	21	1	6894.8	36	15.000
7	9	1	2	164.2	4	9.259
8	4	0	1	0	2	10.417
total	388	61	24	834.5	175	9.396

表 3 算法 VTC 构造完全团的预测边集似然率和功能富集的分布

交叠规模	新边总数	真阳性	真阴性	L	观测频率	功能富集
4	320	96	10	3151.9	167	10.873
5	176	11	3	1203.8	72	8.523
6	83	25	4	2052.0	54	13.554
7	8	1	0	N/A	5	13.021
8	7	0	3	0	3	8.929
total	558	133	20	2183.3	301	11.238

表 4 DC 算法和 VTC 算法在几种 S.cerevisiae 数据集上的预测性能比较 data1 来自文献[35]; data2 来自 DIP 数据库;data3 和 data4 来自文献[36],其原数据来自文献[10]

数据集	蛋白质数目	检测的相互作用数	DC 预测的相互作用数	VTC 预测的相互作用数	预测交叠	P 值
data1	2283	6645	388	558	311	$<10^{-130}$
data2	1361	3222	260	288	206	$<10^{-130}$
data3	4683	14455	638	1181	577	$<10^{-130}$
data4	1093	7440	2416	4119	2322	$<10^{-130}$

对于一个包含蛋白质数为 n 的相互作用网络,N 是样本空间,$N = n \times (n-1)/2$,$S1$ 是指 VTC 预测的相互作用数,$S2$ 是指 DC 预测的相互作用数,X 是指这两个算法预测的相互作用相同的个数。P 值的计算模型定义为:

$$P = 1 - \sum_{i=0}^{X-1} \frac{\begin{bmatrix} S1 \\ i \end{bmatrix} \begin{bmatrix} N - S1 \\ S2 - i \end{bmatrix}}{\begin{bmatrix} N \\ S2 \end{bmatrix}} \tag{5}$$

P 值越小说明其越具有统计意义。从表 4 可以看出,算法 VTC 和 DC 的 P 值在 10^{-130}以下,统计意义显著。

2.2 生物实例

由 2.1 可知,DC 和 VTC 算法应用在文献[35]数据集上,分别识别了 388 和 558 个相互作用。把这些相互作用添加到初始的网络中,可以识别很多在初始网络中没有的蛋白质复合物。例如:用于调控细胞生长和增殖的 Casein kinase II 复合物[37],它包含两个催化活性亚基(CKA1 和 CKA2)和两个调控亚基(CKB1 和 CKB2)。在原始的高通量数据集中,这 4 个蛋白质只构成了 2 个大小为 3 的团{CKA1, CKA2, CKB2}和{CKA1, CKB1, CKB2},CKA2 和 CKB1 之间的相互作用被丢失。通过 DC 和 VTC 算法,可以识别出这个丢失的相互作用,把该相互作用重新添加到网络中得到 1 个大小为 4 的团{CKA1, CKA2, CKB1,CKB2}的完整复合物。

此外,VTC 算法除了可以识别复合物中丢失的相互作用,还可以识别出复合物中蛋白质与复合物的某些附着蛋白质之间丢失的相互作用。例如:复合物 20S proteasome[38]{PRE10, PRE2, PRE3, PRE6, PRE8, PUP3, SCL1}及其附着蛋白质 RPN8,在原始网络中,丢失了 RPN8 与复合物 20S proteasome 中的蛋白质{PRE2, PRE8, PUP3, SCL1}之间的相互作用,应用 VTC 算法,可以识别出这些丢失的相互作用,但是 DC 算法却没有识别出这些相互作用。

3 结 论

根据蛋白质相互作用网络的拓扑结构特点,本文提出了一种基于极大团的蛋白质相互作用预测算法,该算法能够有效提高蛋白质相互作用预测数量,同时根据高通量的蛋白质相互作用网络中假阳性的分布特点,放大预测结果集中的伪相互作用,然后利用优先连接原则将其去除,从而可提高预测结果的可靠性。实验结果表明 VTC 算法不仅预测的可靠性和预测的相互作用数量优于 DC 算法,而且在功能富集方面也要好于 DC 算法。此外,如果把 VTC 中极大团从小到大的扩展看作成蛋白质网络复合物的聚类过程,那么 VTC 也可作为一个简单的蛋白质复合物预测算法。

参考文献

[1] Drewes G, Bouwmeester T. Global approaches to protein-protein interactions. *Curr Opin Cell Biol*, 2003, 15(2): 199-205

[2] Jansen R, Yu H Y, Greenbaum D, et al. A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, 2003, 302(5644):449-453

[3] Goldberg D S, Roth F P. Assessing experimentally derived interactions in a small world. *Proc Natl Acad Sci USA*, 2003, 4372-4376

[4] Mewes H W, Heumann K, Kaps A, et al. MIPS: a database for genomes and protein sequences. *Nucleic Acids Res*, 2002, 27(1):31-34

[5] Bader G D, Hogue C W. Analyzing yeast protein protein interaction data obtained from different sources. *Nat Biotech*, 2002, 20(10): 991-997

[6] Xia Y, Yu H Y, Jansen R, et al. Analyzing cellular biochemistry in terms of molecular networks. *Annu Rev Biochem*, 2004, 73:1051-1087

[7] Xenarios I, Salwínski L, Duan X J, et al. DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res*, 2002, 30(1):303-305

[8] Kumar A, Snyder M. Protein complexes take the bait. *Nature*, 2002, 415(6868): 123-124

[9] Rigaut G, Shevchenko A, Rutz B, et al. A generic protein purification method for protein complex characterization and proteome exploration. *Nature Biotechnol*, 1999, 17(10): 1030-1032

[10] Von Mering C, Krause R, Snel B, et al. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 2002, 417(6887): 399-403

[11] Jansen R, Lan N, Qian J, et al. Integration of genomic datasets to predict protein complexes in yeast. *J Struct Funct Genomics*, 2002, 2:71-81

[12] Yu H Y, Paccanaro A, Trifonov V, et al. Predicting interactions in protein networks by completing defective cliques. *Bioinformatics*, 2006, 22(7): 823-829

[13] Kini R M, Evans J H. Prediction of potential protein-protein interaction sites from amino acid sequence. Identification of a fibrin polymerization site. *FEBS Lett*, 1996, 385(1-2):81-86

[14] Pazos F, Helmer-Citterich M, Ausiello G, et al. Correlated mutation contain information about protein-protein interaction. *J Mol Biol*, 1997, 271(4):511-523

[15] Dandekar T, Snel B, Huynen M, et al. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci*, 1998, 23(9):324-328

[16] Enright A J, Iliopoulos I, Kyrpides N C, et al. Protein interaction maps for complete genomes based on gene fusion events. *Nature*, 1999, 402(6757):86-90

[17] Marcotte E M, Pellegrini M, Ng H L, et al. Detecting protein function and protein-protein interactions from genome sequences. *Science*, 1999, 285(5428):751-753

[18] Sprinzak E, Margalit H. Correlated sequence signatures as markers of protein-protein interactions. *J Mol Biol*, 2001, 311(4):681-692

[19] Kim W K, Park J, Suh J K. Large scale statistical prediction of protein protein interaction by potentially interacting domain (PID) pair. *Genome Inform Ser Workshop Genome Inform*, 2002, 13:42-50

[20] Ng S, Zhang Z, Tan S H. Integrative approach for computationally inferring protein domain interactions. *Bioinformatics*, 2003, 19(8):923-929
[21] Han D S, Kim H S, Seo J M, et al. A domain combination based probabilistic framework for protein-protein interaction prediction. *Genome Inform Ser Workshop Genome Inform*, 2003, 14:250-259
[22] Han D S, Kim H S, Jang W H, et al. PreSPI: design and implementation of protein-protein interaction prediction service system. *Genome Inform*, 2004, 15(2):171-80
[23] Chen X W, Liu M. Prediction of protein-protein interactions using random decision forest framework. *Bioinformatics*, 2005, 21(24): 4394-4400
[24] Singhal M, Resat H. A domain-based approach to predict protein-protein interactions. *BMC Bioinformatics*, 2007, 8 (13):199
[25] King A D, Przulj N, Jurisica I. Protein complex prediction via cost-based clustering. *Bioinformatics*, 2004, 20(17): 3013-3020
[26] Bader G D, Hogue C W. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, 2003, 4: 2
[27] Han J D, Dupuy D, Bertin N, et al. Effect of sampling on topology predictions of protein-protein interaction networks. *Nat Biotech*, 2005, 23(7):839-844
[28] Patil A, Nakamura H. Filtering high-throughput protein-protein interaction data using a combination of genomic features. *BMC Bioinformatics*, 2005, 6:100
[29] Uetz P, Giot L, Cagney G, et al. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, 403, 2000, 623-627
[30] Ito T, Chiba T, Ozawa R, et al. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci USA*, 2001,98: 4569-4576
[31] Albert R, Albert L, Barabási A L. Statistical mechanics of complex networks. *Review of Modern Physics*, 2002,74 :47-97
[32] Newman M E J. The structure and function of complex networks. *SIAM Review*, 2003, 45:167-256
[33] Tsukiyama S, Ide M, Ariyoshi H, et al. A new algorithm for generating all the maximal independent sets. *SIAM J Comput*, 1977, 6(3):505-517
[34] Kumar A, Agarwal S, Heyman J A, et al. Subcellular localization of the yeast proteome. *Genes Dev*, 2002, 16: 707-719
[35] Bader G D, Donaldson I, Wolting C, et al. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res*, 2003,31:248-250
[36] Deng M H, Sun F Z, Chen T. Assessment of the reliability of protein-protein interactions and protein function prediction. In: Proceedings of the 8th Pacific Symposium on Biocomputing, Hawaii, USA, 2003. 140-151
[37] Ackermann K, Waxmann A, Pyerin W, et al. Genes targeted by protein kinase CK2: A genome-wide expression array analysis in yeast. *Mol Cell Biochem*, 2001,227:59-66
[38] Gavin A C, Aloy P, Grandi P, et al. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 2005, 440: 631-636

A maximal clique-based method for predicting interactions in protein networks

Wang Jianxin, Cai Zhao, Li Min
(School of Information Science and Engineering, Central South University, Changsha 410083)

Abstract

This paper analyzes the problem that the protein-protein interaction datasets identified by present large-scale, high-throughput methods contain a relatively high level of noise, and then proposes a maximal clique-based method for predicting interactions in protein networks according to the distribution characteristics of noise data and topological properties in the networks, to overcome the drop of the reliability in the dataset identification due to the noise. The method predicts the interactions among proteins based on the degree of correlation between a protein and a protein clique, and then deletes the noise data after amplifying them in the predicted dataset, which improves the prediction reliability. The experimental results show that this method can predict not only more interactions but also higher reliable interaction datasets.

Key words: protein-protein interaction network, interaction prediction, false positive, maximal clique