基于最短路径的蛋白质相互作用网络拓扑分析①

李　敏②　陈建二　王建新③

(中南大学信息科学与工程学院　长沙 410083)

摘　要　为了进行对蛋白质相互作用网络的拓扑分析,应用最短路径技术对蛋白质相互作用数据库(DIP)中包括酵母在内的7个物种的8个蛋白质相互作用网络进行了研究,包括对网络直径、特征路径长度、连通效率、顶点介数与顶点度的相关性以及高介数边和长间隔边在网络连通中的作用的研究。分析发现,这些网络对随机移除一定数量的蛋白质顶点(或边)具有很好的健壮性,但对高介数顶点(或边)的确定性移除却相当脆弱,而且按顺序移除2%高介数顶点所引起的网络连通效率下降明显大于随机移除10%顶点所引起的网络连通效率变化;所研究的7个物种的网络都存在不同比例的边缺失替代路径,绝大多数网络在移除一定比例的长间隔边后网络连通效率下降。

关键词　生物信息学, 蛋白质相互作用网络, 最短路径, 特征路径长度, 介数, 间隔

0　引　言

在后基因组时代,一个重要的挑战就是系统地分析和全面理解蛋白质之间如何通过相互作用完成生命活动[1]。蛋白质相互作用在生命活动中起核心作用,不仅是正常生理过程如DNA复制、转录、翻译、物质代谢、信号传导以及细胞周期控制的基础,也在病理过程中起着重要的作用[2-5]。分析蛋白质相互作用网络,研究其拓扑特性,不仅有助于研究未知蛋白质的生物学功能,识别人类疾病的诊断标记和寻找治疗的新靶点,也为我们进一步了解蛋白质之间是如何通过相互作用来完成生命活动提供理论依据。

一个生物体内所有蛋白质相互作用被称为蛋白质相互作用网络(protein-protein interaction network, PIN)[2,6]。目前拓扑分析研究最多的是酵母(S. cerevisiae)蛋白质相互作用网络[7-11],已经发现了酵母蛋白质相互作用网络不同于随机图的一些特性,如小世界特性和度分布服从幂规律的特性。文献[7]指出,在酵母蛋白质相互作用网络中顶点度越高的蛋白质越重要,蛋白质顶点的度和重要性之间存在正相关性。但文献[10]在酵母蛋白质相互作用网络中发现了一些度很小但相当重要的蛋白质顶点,并引入介数来评估顶点的重要性,他们认为高介数的顶点在网络中更关键。文献[12]分析了大肠杆菌(*E. coli*)、线虫(*C. elegans*)、果蝇(*D. melanogaster*)和幽门螺杆菌(*H. pylori*)的蛋白质相互作用网络,发现这些网络的度分布有宽尾现象,并不服从严格的幂规律分布,而需要某种修改的幂规律形式来描述。文献[13]对人类蛋白质相互作用网络的分析结果否定了蛋白质顶点的度和重要性之间存在正相关性的推定。文献[14]引入介数和向心性度量方法来分析不同数据库来源的人类蛋白质相互作用网络,发现了大量低度高介数的蛋白质顶点。

目前,对各物种的蛋白质相互作用网络的拓扑分析还处于刚刚起步阶段,很多特性还有待于进一步的探索和证明。本文应用最短路径技术深入分析了包括酵母在内的7个物种的8个蛋白质相互作用网络,并对顶点介数和顶点度的相关性进行了分析,探讨了随机移除顶点或边、按序移除高介数的顶点或边以及按序移除长间隔的边对网络的影响。

1　相关定义

蛋白质相互作用网络可以被表示成为一个无向简单图 $G=(V,E)$,例如图1所示的酵母蛋白质相互作用网络[7],图中的每个顶点表示一个酵母蛋白质,每条边表示一对蛋白质之间的相互作用。在蛋白质相互作用网络图中任意两个顶点 i 和 j 之间的

① 国家自然科学基金(60433020)、新世纪优秀人才支持计划(NCET-05-0683)和长江学者和创新团队发展计划(IRT0661)资助项目。
② 女,1978年生,博士生,工程师;研究方向:生物信息学;E-mail: limin@mail.csu.edu.cn
③ 通讯作者,E-mail: jxwang@mail.csu.edu.cn
(收稿日期:2007-10-31)

最短路径长度 d_{ij} 是指从顶点 i 到顶点 j 至少经过的边数。为了描述方便，接下来给出本文所需的相关定义(定义中的图 G 即指蛋白质相互作用网络)：

图 1 酵母的蛋白质相互作用网络

为了描述方便，接下来给出本文所需的相关定义：

定义 1 网络直径(network diameter) D 是指图 G 中所有顶点对之间最短路径长度的最大值，如式

$$D = \max\{d_{ij} \mid i \neq j;\ i,j = 1,2,\cdots,|V|\} \quad (1)$$

所示。

由于能够获得的蛋白质相互作用数据具有不完整性，某些蛋白质相互作用网络可能是非连通图，整个网络分为若干个连通子图。严格来说，非连通图的网络直径是无穷大。本文将非连通图的网络直径定义为所有连通子图的网络直径的最大值。

定义 2 特征路径长度(characteristic path length) L 是指图 G 中所有最短路径长度的平均值，如式

$$L = \frac{\sum_d df(d)}{\sum_d f(d)} \quad (2)$$

所示，其中，$f(d)$ 表示长度为 d 的最短路径出现的频率。

特征路径长度描述了网络中顶点间的分离程度，即网络有多小。复杂网络研究中一个重要的发现是很多大规模真实网络的特征路径长度比想象的小得多，称之为“小世界效应”。

在某些蛋白质相互作用网络中可能存在顶点对之间的最短路径为无穷大的情况，定义 2 所描述的特征路径长度并没有考虑这些顶点对。为了体现最短路径为无穷大的顶点对对整个网络拓扑特性的影响，我们通过研究网络连通效率 E 来代替特征路径长度 L。E 的计算公式如下：

$$E = \frac{\sum_{i \neq j} \frac{1}{d_{ij}}}{|V|(|V|-1)} \quad (3)$$

由公式(3)可知，$0 \leqslant E \leqslant 1$，当网络为全连通图时 E 为 1，当网络中顶点完全离散时 E 为 0。E 越大表明网络的连通性越好。

定义 3 顶点介数(vertex betweenness) B_v[15]是指图 G 中所有的最短路径中经过某个特定顶点 v 的数量比例。

顶点介数反映了该顶点在整个网络中所处位置的重要程度。一般，比较关键的起枢纽作用的蛋白质都具有比较高的顶点介数。

定义 4 边介数(edge betweenness) B_e[16,17]是指图 G 中所有的最短路径中经过某条特定边 e 的数量比例。

边介数在蛋白质相互作用网络的模块化结构分析中具有重要作用，一般认为不同的模块之间相连的边具有比较高的边介数[16-19]。

定义 5 间隔(range)[20]是指图 G 的一条边上的两个顶点不通过该条边的最短路径大小，或者说是移除这条边后，这两个顶点之间的最短距离。

“间隔”是 Watts 和 Strogatz 在提出小世界(WS)模型[20]时一同提出来的，他们认为相对少量的几条长间隔的边比较重要。

定义 6 聚集系数(clustering coefficient)[20]刻画了某顶点的邻居顶点彼此之间联系的紧密程度。设某个顶点的度为 k_i，其对应的 k_i 个邻居顶点之间实际存在的边数为 E_i，则该顶点的聚集系数为 $2E_i/(k_i(k_i-1))$。

整个网络的聚集系数为所有顶点聚集系数的平均值，它体现了相互作用的顶点聚集成簇的整体趋势。

2 蛋白质相互作用网络拓扑特性分析

2.1 数据样本

本文的蛋白质相互作用数据选自相互作用的蛋白质数据库(database of interacting proteins, DIP)[21]。我们从 2007 年 7 月的 DIP 数据库中下载得到相互作用多于 200 的 7 个物种的 8 个最新数据集(其中酵母分为全集和核心集两个数据集)，去掉这些数据集中的自相互作用和冗余相互作用，得到蛋白质相互作用网络分析的样本数据，其基本信息如表 1 所

示。为了描述方便,我们在表 1 中给出了各个物种的标记符号(marker),例如酵母的数据全集被记作 YS、酵母的核心数据集被记作 YSC、线虫被记作 CE。

表 1　各物种 PIN 的基本信息

物种名称	物种标记	蛋白质	相互作用
S.cerevisiae (酵母, 全集)	YS	5024	17535
S.cerevisiae Core (酵母, 核心集)	YSC	2633	5969
E. coli 大肠杆菌	EC	2503	6971
D.melanogaster (果蝇)	DM	7476	22828
C.elegans 线虫	CE	2666	4037
H.pylori (幽门螺杆菌)	HP	731	1420
M.musculus (家鼠)	MM	404	346
H.sapiens (人类)	HS	1258	1534

2.2　网络直径和特征路径长度分析

在本节实验中,我们首先采用广度优先算法计算网络图中所有顶点之间的最短路径,并根据公式(1)、(2)和(3)分别计算网络直径 D、特征路径长度 L 和网络连通效率 E,计算结果如表 2 所示。

表 2　各物种 PIN 的网络直径(D)、特征路径长度(L)和网络连通效率(E)的计算结果

物种标记	平均度	D	L	E
YS	6.95	11	4.15	0.198
YSC	4.53	14	5.23	0.177
EC	7.72	11	3.80	0.252
DM	6.08	12	4.39	0.124
CE	3.02	14	4.81	0.184
HP	3.87	9	4.14	0.251
MM	1.68	9	3.47	0.011
HS	2.42	21	6.63	0.048

从表 2 可以看出,人类的蛋白质相互作用网络直径最长,为 21;幽门螺杆菌和家鼠的蛋白质相互作用网络直径最短,为 9。对于平均度大于 4 的几个蛋白质相互作用网络,平均度越大(即网络越稠密),网络直径越小,但对顶点平均度小于 4 的几个蛋白质相互作用网络不存在这种规律。从表 2 可以看出,除了人类以外的其他几个物种的蛋白质相互作用网络的特征路径长度都小于 6,符合复杂网络的小世界特性,即“六度原理”。但对于家鼠和人类这样非常稀疏的蛋白质相互作用网络(目前通过实验获得的蛋白质相互作用还很有限)来说,特征路径长度还不足以判断其是否具有小世界网络特性。各物种存在相互作用路径的蛋白质顶点之间的最短路径长度的分布情况如图 2 所示。

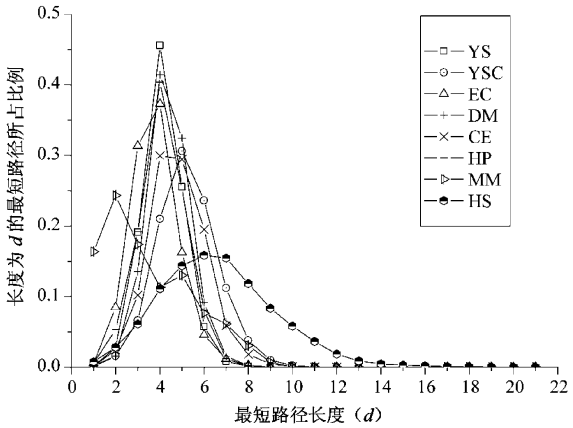

图 2　各物种的蛋白质相互作用最短路径分布图

从图 2 可以看出,除家鼠以外其他几个物种的蛋白质相互作用的最短路径都近似地服从正态分布。

网络连通效率从另外一个角度描述了网络连通性的好坏。从直觉来看,网络越稠密,其连通性应该越好。从表 2 可以看出,两个最稀疏的家鼠和人类的蛋白质相互作用网络的网络连通效率最低。从同一物种酵母的两个不同数据集来看,其核心数据集的网络连通效率低于数据全集的网络连通效率。但从表 2 中的其他几个物种来看,网络连通效率与网络顶点平均度之间没有直接关系,大肠杆菌和幽门螺杆菌的顶点平均度分别为 7.72 和 3.87,但都具有较高的网络连通效率。线虫比果蝇的顶点平均度低,却较果蝇具有更高的网络连通效率。

2.3　介数分析

顶点介数描述了网络图中经过该顶点的最短路径数。直观上感觉,顶点的度越高,其对应的介数也应该越高。8 个蛋白质相互作用网络的顶点度与顶点介数的关系如图 3 所示。从图 3 可以看出,酵母(包括核心集和数据全集)、大肠杆菌、果蝇和幽门螺杆菌的蛋白质相互作用网络的顶点介数基本上随着顶点度的增加而增加,说明这些物种网络的顶点介数和顶点度之间存在较高的正相关性。但在线虫、家鼠和人类的蛋白质相互作用网络中却没有这种明显的正相关性,特别是在家鼠和人类的蛋白质相互作用网络中存在大量低度高介数的蛋白质顶点。

本文通过分析部分顶点被移除后整个网络的拓扑变化来进一步分析这些顶点在整个网络中的重要程度。对 8 个蛋白质相互作用网络分别实施随机移除和有选择地移除一定比例顶点(从 0%到 10%,按

2 个百分点间隔)的操作。实施随机移除顶点的操作后各物种的网络连通效率的变化如图 4(a)所示，实施按序移除高介数顶点后各物种的网络连通效率的变化如图 4(b)所示。

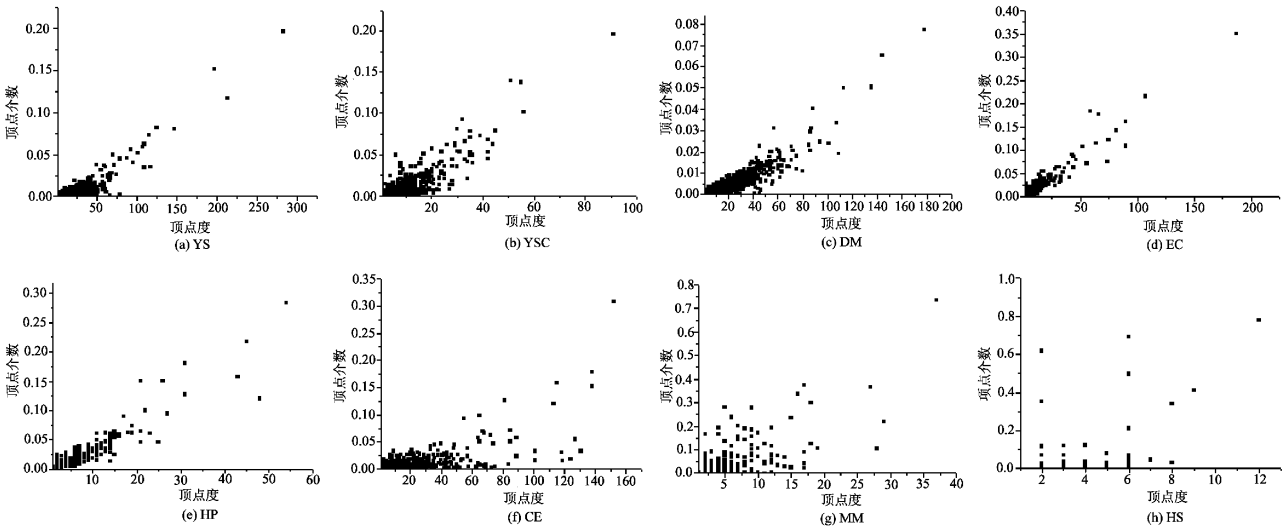

图 3 各物种的顶点度与顶点介数之间的关系图

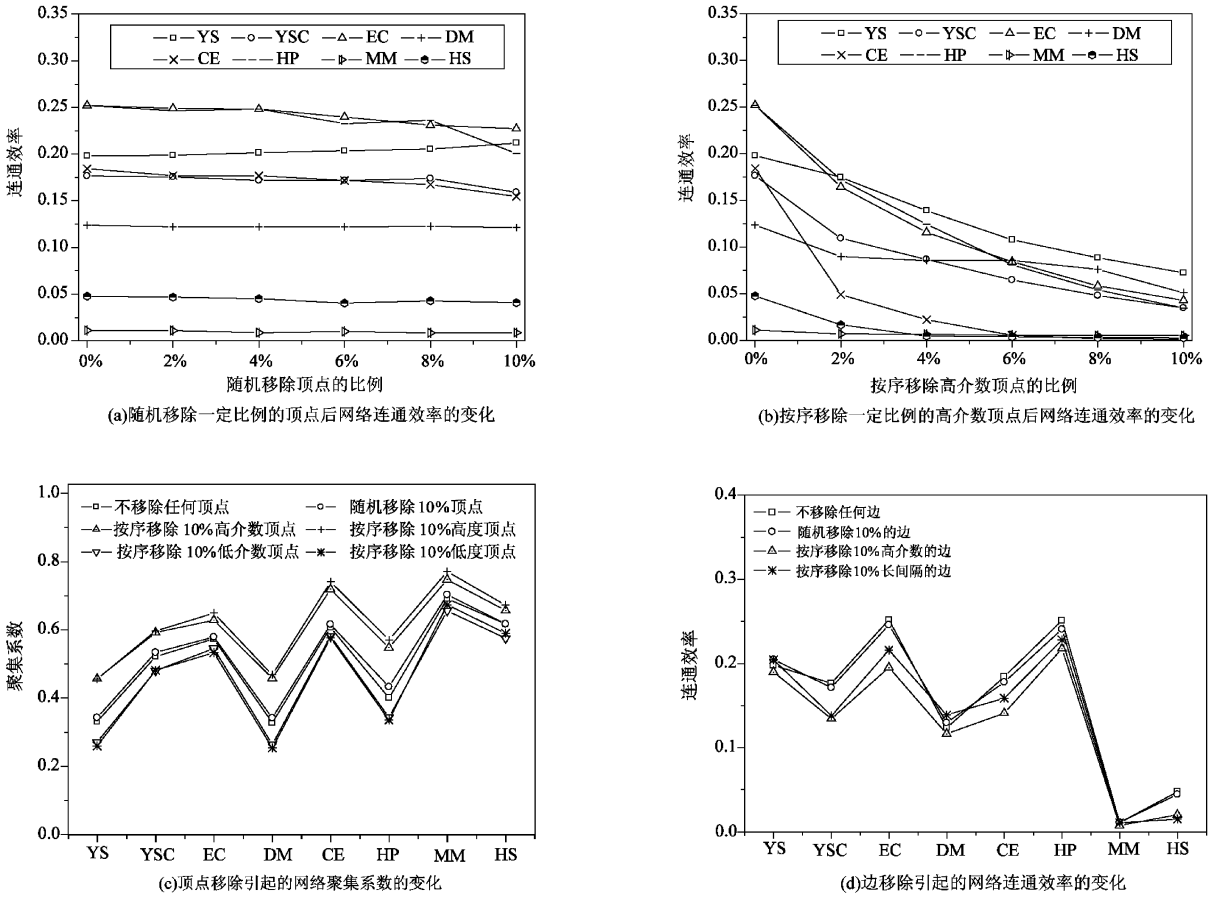

图 4 按不同规则移除部分顶点(或边)后网络连通效率和网络聚集系数的变化曲线图

从图 4(a)可以看出，随机移除 10%的顶点对各物种的网络连通效率几乎没有影响。从图 4(b)可以看出按序移除 2%的高介数顶点后绝大多数物种的网络连通效率已有明显下降；按序移除 10%的高介数顶点后各物种的网络连通效率大幅下降。其中线虫和人类的蛋白质相互作用网络在按序移除

10%的高介数顶点后网络连通效率较随机移除相同数量顶点的网络连通效率下降达 90%以上;果蝇和家鼠的蛋白质相互作用网络在两个不同移除操作下移除相同数量(10%)的顶点的差别最小,但按序移除后网络连通效率较随机移除后网络连通效率下降仍达 20%以上;酵母(核心集)、大肠杆菌和幽门螺杆菌的蛋白质相互作用网络在按序移除 10%高介数顶点后网络连通效率较随机移除相同数量的顶点后网络连通效率下降都达 59%以上。由此可见,蛋白质相互作用网络对随机移除具有较强的健壮性,但对高介数顶点的确定性移除却相当脆弱。这说明,高介数的顶点在各物种的蛋白质相互作用网络中具有重要作用。从生物意义的角度分析,这些高介数的顶点在物种的进化过程中具有重要作用,例如人类蛋白质相互作用网络中的高介数顶点(DIP 数据库中的编号) DIP:109N、DIP:232N、DIP:252N、DIP:1045N、DIP:1048N 和 DIP:1078N 等在家鼠的蛋白质相互作用网络中存在而且也具有较高的顶点介数。

图 4(c)描述了随机移除或按序移除 10%的顶点对网络聚集系数的影响。从图 4(c)可以看出,随机移除 10%的顶点对各物种的网络聚集系数几乎没有影响;按序移除 10%的低度或低介数的顶点后各物种的网络聚集系数都减小;而按序移除 10%的高度或高介数顶点后各物种的网络聚集系数都明显增加。这说明顶点的度和介数越高,其聚集系数越小。反之,顶点的度和介数越低,其聚集系数越大,越容易集聚成簇。

与顶点介数类似,边介数研究的是经过每条边的最短路径数。本文分别对 8 个蛋白质相互作用网络进行随机移除 10%的边和按序移除 10%高介数边的操作,这两种移除操作对网络连通效率的影响如图 4(d)所示。

从图 4(d)可以看出,随机移除网络中 10%的边对网络连通效率几乎没有影响;而按序移除 10%高介数的边后网络连通效率明显降低。这说明,高介数的边在网络中具有重要的连接作用,它是蛋白质相互作用网络中不同模块之间联系的桥梁。

2.4　间隔分析

如果一条边的间隔为无穷大,则该边在网络中不存在替代路径。8 个蛋白质相互作用网络中不存在替代路径的边所占的比例如图 5 所示,存在替代路径的边的间隔分布如图 5 所示。

从图 5 可以看出,3 个最稀疏的网络(线虫、家鼠和人类)不存在替代路径的边所占比例最高,达 40%以上,其余网络不存在替代路径的边所占比例都不超过 20%。

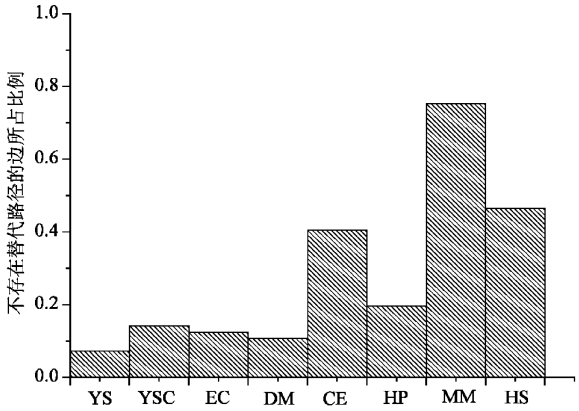

图 5　各物种 PIN 中不存在替代路径的边所占比例图

从图 6 可以看出,各物种的蛋白质相互作用网络中存在替代路径的边的间隔主要集中在 2 和 3,间隔大于 6 的边的数量很少。Watts 和 Strogatz 认为相对少量的几条长间隔的边对整个网络来说非常重要。将 8 个蛋白质相互作用网络移除 10%长间隔的边后网络连通效率的变化如图 4(d)所示。

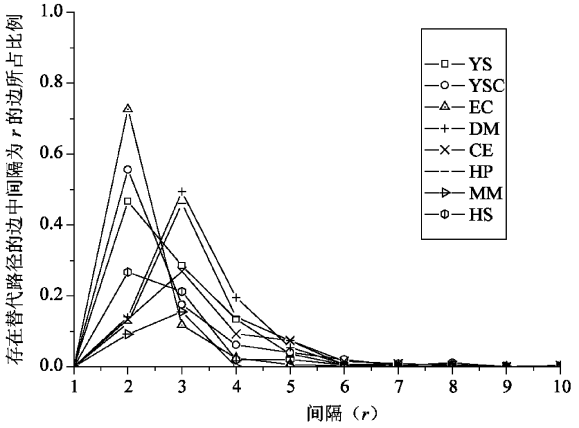

图 6　各物种的蛋白质相互作用间隔分布图

从图 4(d)可以看出,酵母数据全集的网络移除 10%长间隔的边后网络连通效率基本没有变化,而酵母核心数据集的网络在移除 10%长间隔的边后网络连通效率明显降低;线虫、大肠杆菌、幽门螺杆菌和人类的蛋白质相互作用网络在移除 10%长间隔的边后网络连通效率也都明显降低;而果蝇和家鼠的变化很小。

3　结　论

本文应用最短路径对 8 个不同规模的 7 个物种的蛋白质相互作用网络的拓扑进行了深入分析。分

析的结果表明蛋白质相互作用网络具有如下性质：

(1) 网络中任意两个蛋白质顶点之间的最短路径长度近似服从正态分布，这些最短路径长度的平均值一般不超过 6，具有典型的小世界特性。

(2) 酵母(包括核心集和数据全集)、大肠杆菌、果蝇和幽门螺杆菌的蛋白质相互作用网络的顶点介数基本上随着顶点度的增加而增加，即顶点介数和顶点度之间存在较高的正相关性。但在线虫、家鼠和人类稀疏的蛋白质相互作用网络中却没有这种明显的正相关性，特别是在家鼠和人类的蛋白质相互作用网络中存在大量低度高介数的蛋白质顶点。

(3) 各物种的蛋白质相互作用网络对随机移除一定数量的蛋白质顶点或边都具有很好的健壮性。

(4) 高介数的顶点和边在网络连接中起重要作用，移除一定比例的高介数顶点或边后各物种的网络连通效率明显降低。

(5) 平均度较小的线虫、家鼠和人类这 3 个物种的网络超过 40% 的边无替代路径，其他 5 个物种的网络中无替代路径的边都不超过 20%，所有网络中存在替代路径的边的间隔主要集中在 2 和 3，绝大多数网络在移除一定比例的长间隔的边后网络连通效率下降。

进一步的研究主要包括分析网络的模块化结构特性和动态特性，识别网络功能模块并进行蛋白质功能预测，特别是为那些通过直系同源方法无法进行功能预测的蛋白质进行功能注释。

参考文献

[1] Garrels J I. Yeast genomic databases and the challenge of the post-genomic era. *Funct Integr Genomics*, 2002, 2(4-5): 212-237

[2] 孙景春，徐晋麟，李亦学等. 大规模蛋白质相互作用数据的分析与应用. 科学通报，2005，50(19)：2055-2060

[3] Eisenberg D, Marcotte E M, Xenarios I, et al. Protein function in the post-genomic era. *Nature*, 2000, 405(6788): 823-826

[4] Wang J. Protein recognition by cell surface receptors: physiological receptors versus virus interactions. *Trends Biochem Sci*, 2002, 27(3): 122-126

[5] Loregian A, Marsden H S, Palu G. Protein-protein interactions as targets for antiviral chemotherapy. *Rev Med Virol*, 2002, 12(4): 239-262

[6] Legrain P, Wojcik J, Gauthier J M. Protein-protein interaction maps: a lead towards cellular functions. *Trends Genet*, 2001, 17(6): 46-352

[7] Jeong H, Mason S, Barabási A, et al. Lethality and centrality in protein networks. *Nature*, 2001, 411:41-42

[8] Yook S, Oltvai Z, Barabási A. Functional and topological characterization of protein interaction networks. *Proteomics*, 2004, 4:928-942

[9] Pržulj N, Wigle D A, Jurisica I. Functional topology in a network of protein interactions. *Bioinformatics*, 2004, 20(3):340-348

[10] Joy M, Brock A, Ingber D, et al. High-betweenness proteins in the yeast protein interaction network. *Journal of Biomedicine and Biotechnology*, 2005, 2:96-103

[11] Wuchty S, Almaas E. Peeling the yeast protein network. *Proteomics*, 2005, 5(2):444-449

[12] Goh K, Kahng B, Kim D. Graph theoretic analysis of protein interaction networks of eukaryotes. *Physica A*, 2005, 357: 501-512

[13] Gandhi T, Zhong J, Mathivanan S, et al. Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nature Genetics*, 2006, 38:285-293

[14] Bader D A, Madduri K. A graph-theoretic analysis of the human protein-interaction network using multi-core parallel algorithms. In: Proceedings of the 6th IEEE International Workshop on High Performance Computational Biology, Long Beach, CA, 2007. 26-30

[15] Freeman L. A set of measures of centrality based on betweenness. *Sociometry*, 1977, 40:35-41

[16] Girvan M, Newman M. Community structure in social and biological networks. *Proc Natl Acad Sci USA*, 2002: 7821-7826

[17] Ruth D, Frank D, Christopher M. The use of edge-betweenness clustering to investigate biological function in protein interaction networks. *BMC Bioinformatics*, 2005, 6:39

[18] Chen J C, Yuan B. Detecting functional modules in the yeast protein-protein interaction network. *Bioinformatics*, 2006, 22(18): 2283-2290

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

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

[21] Xenarios I, Rice D W, Salwinski L, et al. DIP: the database of interacting proteins. *Nucleic Acids Res*, 2000, 28:289-291

Shortest path-based analysis of protein-protein interaction networks

Li Min, Chen Jianer, Wang Jianxin

(School of Information Science and Engineering, Central South University, Changsha 410083)

Abstract

Eight protein-protein interaction networks of 7 species from the database of interacting proteins (DIP) were studied based on the shortest path technique for topological analysis of the protein-protein interaction networks of all species, including the studies of the networks' diameter, characteristic path length, connection efficiency, relation between vertex betweenness and degree, and the effect of edges with high betweenness and long range on network connection. The analyses show that all the 8 protein-protein interaction networks are robust against the arbitrary elimination of vertices (or edges), but vulnerable to the sequential deletion of vertices (or edges) by betweenness. The connection efficiency of a network decreases faster when 2% of vertices are sequentially removed from the highest betweenness than that when 10% of vertices are removed at random. There are a certain number of edges without substitute paths in all the 8 protein-protein interaction networks of 7 species. Most networks are vulnerable to sequential removal of edges with long range.

Key words: bioinformatics, protein-protein interaction network, shortest path, characteristic path length, betweenness, range