代谢网络的蝴蝶结结构特征及其功能意义

赵 静[①②④] 陶 林[②] 俞 鸿[②] 骆建华[①] 曹志伟[②*] 李亦学[①②③*]

(① 上海交通大学生命科学技术学院, 上海 200240; ② 上海生物信息技术研究中心, 上海 200235; ③ 中国科学院上海生命科学研究院, 上海 200031; ④ 重庆后勤工程学院数学教研室, 重庆 400016. * 联系人, E-mail: zwcao@scbit.org; yxli@sibs.ac.cn)

摘要 研究全基因组代谢网络的结构拓扑对于了解结构与功能间的联系是必不可少的. 而可视化有助于获取网络组织结构的直观信息. 从网络拓扑的角度考察了 75 个物种的全基因组代谢网络. 提出了一个展开蝴蝶结模型, 实现了对代谢网络的蝴蝶结结构的清晰的可视化. 展开蝴蝶结所揭示的代谢网络的结构特征, 有助于我们设计高效的且更能反应代谢网络自身特点的网络算法. 同时, 这个粗粒化的网络模型还可实现对网络中脆弱连接的可视化, 因此对于疾病研究和药靶发现都有潜在意义. 通过对蝴蝶结中心部分的双向连接及主核的研究表明, 代谢网络的蝴蝶结结构有其内在的、有意义的拓扑特征, 而这些特征是随机网络所不具备的.

关键词 生物信息学 代谢网络 蝴蝶结结构 随机网络

近几年来, 随着数百种生物的全基因组测序的完成, 以及对基因功能不断深入和广泛的了解, 使得从基因组信息可靠地重建物种特异的代谢网络成为可能[1~6]. 因此, 通过代谢网络预测生物学功能, 成为后基因组时代生物信息学研究的一个基本问题[4,7~11]. 然而, 要探究代谢网络的结构与功能之间的可能联系, 我们必须首先对代谢网络的拓扑结构特征有足够的认识.

对代谢网络复杂的拓扑结构的研究, 已成为最近几年的热点问题之一[12~20]. 许多学者应用图论的方法将代谢网络表示成一个有向图, 其顶点对应代谢物, 有向弧则代表代谢物之间的生化反应. 人们发现, 代谢网络与其他真实世界的复杂网络一样, 具有与简单随机网络截然不同的拓扑特征[21], 说明它们这种非随机的结构可能蕴涵着代谢系统的具有生物学意义的组织原则.

通过对 65 个物种的全基因组代谢网络的计算分析, Ma等人[14]发现代谢网络的宏观结构呈现一种蝴蝶结的形态. 而Csete等人[22]从生物系统物质流和信息流的角度, 认为代谢系统的组织结构应是嵌套的蝴蝶结形式, 他们指出蝴蝶结结构有利于生物学系统的稳健性, 然而也具有基于其自身设计而带来的内在的、可预测的脆弱性; 而蝴蝶结结构这种大规模的组织构架, 是对复杂生物学过程进行更精确模拟的必要起点[22].

由于网络的蝴蝶结结构已引起很大关注[14,22~25], 而网络中大量顶点和边的存在, 使得对此结构的清晰的可视化成为必要. 同时蝴蝶结结构的中心部分GSC也需要更深入细致的研究. 将网络作k-核分解(k-cores)的方法[26], 使我们能同时按其连接度和在网络中的中心地位两个方面将网络结点分类. 此方法已成功用于分析蛋白相互作用网络[27], 它将有助于我们探究GSC部分的拓扑特征. 另一方面, 代谢网络同时包含可逆反应和不可逆反应, 而它们对代谢过程的调节起着不同的作用. 因此, 可逆反应在代谢网络中分布特征, 可能会有助于我们了解代谢调节, 而可逆性参数(reciprocity)指标即可量化有向网络中双向边的分布[28].

本研究重建了 75 个物种的代谢网络, 其中包括 8 种真核生物、56 种细菌和 11 种古生菌. 提出了一个基于粗粒化网络的展开蝴蝶结模型, 实现了对代谢网络的蝴蝶结结构的清晰的可视化. 接着研究了蝴蝶结 GSC 部分的可逆性参数特征和主核(main core). 为寻求代谢网络内在的结构特性, 将 *E.coli* 代谢网络与其恰当的随机化对照网络进行比较. 对照网络保留了 *E.coli* 代谢网络每个顶点的连接度及与其相连的有向边和无向边的总数, 而将顶点之间的连线进行了随机重连.

2006-08-01 收稿; 2006-11-06 接受

国家科技攻关计划(编号: 2003CB715900, 2004CB720103)、国家自然科学基金项目(批准号: 30500107, 30670953)和上海市科学技术委员会科研计划(批准号: 04DZ19850, 04DZ14005)项目资助

1 材料与方法

1.1 数据准备与网络重建

本研究的代谢相关数据来自于Ma等人[5]在KEGG代谢反应数据库[2,3]的基础上进行修正补充后得到一个新数据库. 此数据库明确给出了每个反应的可逆性信息. 另外, 在细胞内, ATP, ADP, NADH和NAD+等通用代谢物(currency metabolites)一般是电子转移或某些功能基团(磷酸基、氨基、一碳单位和甲基等)转移的携带者, 参与了大量的代谢反应, 但并没有参与反应产物的合成. 为了显示细胞中主要化合物的转化以及在代谢网络中确切地显示生化反应的步骤, 这些通用代谢物及一些小分子化合物(如H_2O, NH_3, O_2和CO_2等)不能作为反应之间的有效中间物, 否则将得出从葡萄糖到丙酮酸只需要两步反应的错误结论. 应注意的是, 简单地去掉所有反应中的某些通用代谢物和小分子化合物是不恰当的, 应针对具体反应进行考虑, 例如谷胺酸(GLU)和 2-酮戊二酸(AKG)在许多反应中用作转移氨基的通用代谢物, 但在以下反应中:

$$AKG + NH_3 + NADPH = GLU + NADP^+ + H_2O,$$

AK 参与了 GLU 的合成, 即它们是基本代谢物. 因此在此反应中它们应被保留. 通过此数据库可重建物种特异的代谢网络. 在这里用有向图表示一个代谢网络, 其中顶点代表代谢物, 弧代表代谢物间的反应. 不可逆反应用单向弧表示, 可逆反应用双向弧表示. 例如以下不可逆反应:

L-Glutamine + 2-Oxoglutarate→L-Glutamate, 对应两个单向弧, 即 L-Glutamine→L-Glutamate 和 2-Oxoglutarate→L-Glutamate.

1.2 蝴蝶结结构

一个有向图具有蝴蝶结结构, 若它的顶点集可划分为四个不相交的子集: 庞大强连通体(giant strong component , GSC), 底物子集(substrate subset, S), 产物子集(product subset, P)和孤立子集(isolated subset, IS), 其中GSC是网络中最大的强连通体(所谓强连通体, 是指有向图的一部分子图, 其中任意两个顶点可通过有向路径彼此到达[29]), 它比网络中其他强连通体大很多, 是网络中顶点间连接最紧密和最复杂的部分. GSC中任两种代谢物可以通过一系列反应彼此生成; 底物子集中的代谢物均可以转化成为GSC中的代谢物, 但反之不能; 产物子集中代谢物均可以由GSC中的代谢物经一系列反应转化得到; 而孤立子集中的代谢物既不能转化成GSC 中的代谢物, 也不能由GSC 中的代谢物转化得到.

1.3 展开的蝴蝶结模型

利用展开的蝴蝶结模型, 可以实现对代谢网络的蝴蝶结结构的清晰的可视化. 该模型包括三个步骤:

(ⅰ) 通过一个粗粒化过程, 在保持网络信息流的前提下, 减少其顶点和弧的总数: 将每个强连通体当作一个整体而压缩成一个结点, 同时将从一个强连通体到另一个强连通体的所有弧压缩成一条弧.

(ⅱ) 进一步依据它们各自的连接模式和生物学特性, 将底物子集(S)、产物子集(P)和孤立子集(IS)再细分成特定的子集.

对粗粒化图的顶点集 S 和 P 中每个顶点 v, 由以下四个参数刻画其连接特性: $d^-(v)$ 为顶点 v 的入度, 即指向顶点 v 的弧总数; $d^+(v)$为顶点 v 的出度, 即从顶点 v 指出的弧总数; $d(v, G)$为从顶点 v 到 GSC 的有向距离, 即从v到 GSC 的最短有向路上弧的总数; $d(G, v)$为从 GSC 到顶点 v 的有向距离, 即从 GSC 到 v 的最短有向路上弧的总数.

根据这些参数, S和P分别被划分成以下公式描述的四个子集: S = S1∪S2∪S3∪S4; P= P1∪P2∪P3∪P4, 其中

$$S1=\{v \mid d^-(v)=0,\ d(v,G)>1\},$$

$$S2=\{v \mid d^-(v)>0,\ d(v,G)>1\},$$

$$S3=\{v \mid d^-(v)>0,\ d(v,G)=1\},$$

$$S4=\{v \mid d^-(v)=0,\ d(v,G)=1\},$$

$$P1=\{v \mid d^+(v)=0,\ d(G,v)=1\},$$

$$P2=\{v \mid d^+(v)>0,\ d(G,v)=1\},$$

$$P3=\{v \mid d^+(v)>0,\ d(G,v)>1\},$$

$$P4=\{v \mid d^+(v)=0,\ d(G,v)>1\}.$$

然后根据孤立子集中的点是与 S 集还是 P 集直接相连, 将其分别划分到 S-IS 和 P-IS 子集中.

(ⅲ) 用图论软件Pajek画出展开的蝴蝶结[30]. 首先, 用命令“draw-partition-vector”画出粗粒化的网络, 其中“partition”和“vector”分别代表各顶点所属子集和顶点大小的向量. 然后在图形窗口中, 用命令“layers-in y direction”, 将属同一子集的点排列在同

一水平线上. 最后, 手工调节个别顶点的位置, 以得到更清晰的布局.

1.4 可逆性参数

在代谢网络中, 不可逆和可逆反应分别用单向弧和双向弧(或称可逆弧)表示. 可逆性参数可量化一个有向网络的连接可逆性的程度[28], 它定义为有向网络的邻接矩阵的元素之间的相关系数:

$$\rho = \frac{\sum_{i\neq j}(a_{ij}-\bar{a})(a_{ji}-\bar{a})}{\sum_{i\neq j}(a_{ij}-\bar{a})^2},$$

其中 $\bar{a} = \frac{\sum_{i\neq j} a_{ij}}{N(N-1)} = \frac{L}{N(N-1)}$, $\bar{a}$ 代表网络中实际的单向弧总数与可能有的单向弧总数的比率, L 和 N 分别是网络中弧和顶点的总数. 一般而言, 一个有向网络的可逆性程度居于两个极端之间. 一个极端是其所有弧都是单向弧的纯粹单向网络(如论文引用网络, 其中后出版的论文可引用前期的, 而反之不行), 可逆性指数等于–1. 另一个极端是其所有弧都是双向弧的完美双向网络(如 Internet, 信息总是沿网络电缆在计算机间双向流动), 可逆性指数等于 1. ρ 的取值范围是[–1, 1]. 可逆性参数可用于比较不同大小的网络的可逆性程度.

1.5 图的 *k*-核

在图论中, 图G的k-核H是G的一个极大子图, 使得H中每个顶点的连接度至少为k, 如图 1[26]. k值最大的核是网络中连接最稠密的部分, 称为主核.

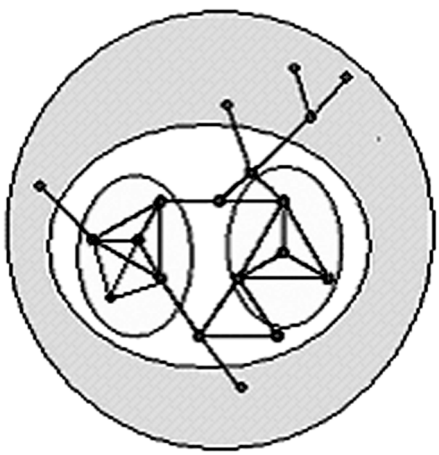

图 1 1-核, 2-核及 3-核

核具有两个基本特征: 核可能是不连通的子图; 核是嵌套的, 当 $i>j$ 时, i-核一定是该图 j-核的子图

1.6 集聚系数

在一个无向图中, 顶点 v 的集聚系数描述了其邻接点之间的连接程度, 即网络的局部集团化程度, 其数学定义为:

$$CC(v) = \frac{2|N(v)|}{d(v)(d(v)-1)},$$

其中$d(v)$是顶点v的度, $|N(v)|$代表顶点v的$d(v)$个邻接点之间的直接连线的数目[31]. $CC(v)$的值介于 0 与 1 之间. 网络的集聚系数定义为其所有顶点的集聚系数的平均值, 它在一定程度上反应了网络的集团化水平[31].

1.7 比较代谢网络与随机网络的方法

根据Maslov等人[32]提出的方法, 我们用Z-score来量化一个真实的代谢网络与它的随机化对照网络之间的差异:

$$Z = \frac{P-\bar{P}_r}{\Delta P_r},$$

其中 P 是真实网络的某个参数, $\bar{P}_r$ 和 ΔP_r 分别是随机化网络集合中该参数的均值和标准差.

2 结果与讨论

2.1 *E. coli* 代谢网络的展开蝴蝶结

本节中, 我们应用展开蝴蝶结模型来获得对 *E.coli* 代谢网络的蝴蝶结拓扑结构的可视化.

E.coli K-12 MG1655 的代谢网络包含 934 结点及 1437 条弧. 该网络最大的连通部分有 575 个结点并具有蝴蝶结拓扑结构, 其 GSC, S, P 和 IS 部分分别包含 234, 85, 177 和 79 个结点. 根据展开蝴蝶结模型, 该 575 个结点的最大的连通部分首先被压缩成只有 215 个结点的粗粒化的图. 然后, 该粗粒化图的顶点被划分成 11 个子集(表 1). 最后, 图 2 给出了 *E. coli* 代谢网络的粗粒化图的展开蝴蝶结结构图.

由展开的蝴蝶结结构, 可看出GSC部分是代谢物转换的枢纽. GSC以外的代谢流是高度分枝的, 且分枝之间的交流很少, 而S, P和IS内的大多数顶点间是单线连接. 在S→GSC及GSC→P的共 90 个连接中, 有 67 个单线连接, 共占 75%, 即大多数这类连接是单线的. 了解这一拓扑特征, 将有助于设计专门针对代谢网络的更高效的算法, 更有利于对代谢的分析. 例如, 我们曾据此设计了一个算法将代谢网络分解成子模块[33], 该算法先分解GSC部分, 再通

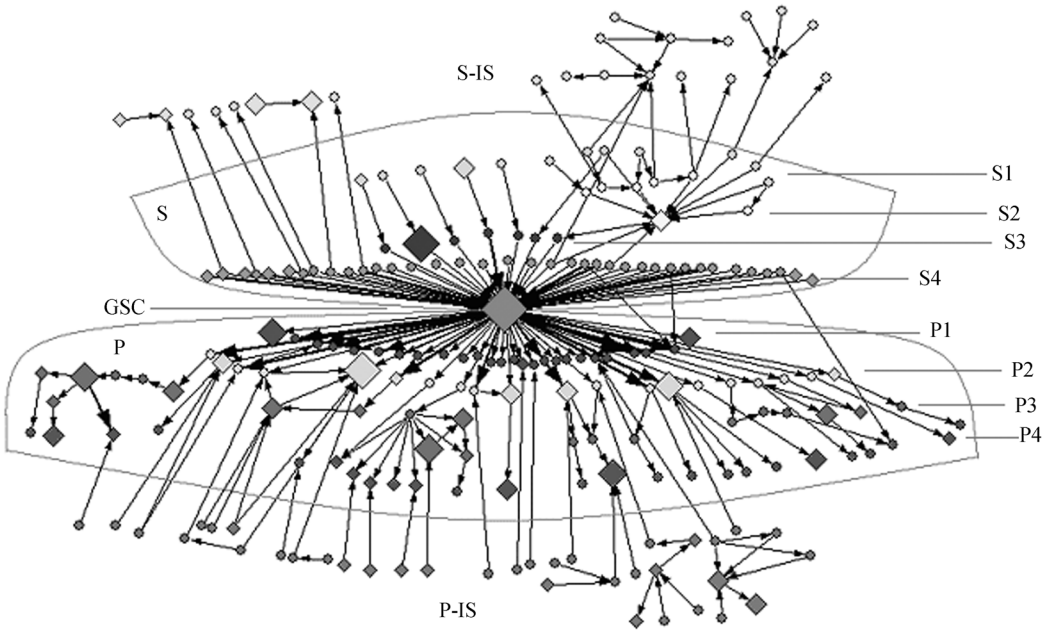

图 2 *E. coli* 代谢网络粗粒化图的展开蝴蝶结结构

属于不同子集的顶点位于不同的层次; 菱形代表至少包含 2 个结点的强连通体, 其大小与结点个数成比例(除 GSC 外); 圆形代表仅一个结点的强连通体; 弧的粗细与对应的 2 个强连通体之间的反应个数成比例, 该图用 Pajek 作成

表 1 不同蝴蝶结模型中每个子集包含的顶点数

		原网络的蝴蝶结结构	粗粒化网络的蝴蝶结结构	粗粒化网络的展开蝴蝶结结构
GSC		234	1	1
S	S1	85	63	12
	S2			7
	S3			7
	S4			37
P	P1	177	92	26
	P2			20
	P3			21
	P4			25
IS	S-IS	79	49	23
	P-IS			26
Total		575	215	215

过一个“多数原则”将GSC的子类扩展到全网络[23]. 分解所产生的每个子模块具有蝴蝶结的拓扑结构, 这些子蝴蝶结相互作用而构成全基因组代谢网络. 我们的分解更好地揭示了代谢网络特有的组织模式.

另外, 此图可显示每个结点在整个代谢网络中所起的作用. 我们定义的 11 个子集不仅代表了顶点在图中的连接特性, 同时也具有生物学意义. S1和S4中结点对应的代谢物是 *E.coli* 从外界摄取的营养物; P1 和 P4 中是代谢的终产物; 而 S2, S3, P2 和 P3 是中间代谢物. S3 和 S4 中结点对应的代谢物是 GSC 中代谢物的直接底物, 而 S1 和 S2 中的化合物则至少需要两步反应才能生成 GSC 中的物质. 类似地, P1 和 P2 是 GSC 中代谢物的直接产物, 而 P3 和 P4 的物质则需 GSC 至少通过两步反应才能获得.

而且, 由于同一强连通体中的顶点能彼此到达, 与此强连通体相连的任何顶点也就能以同样方式与其中的任何顶点相连. 因此, 压缩强连通体的这种简化方式, 是在保持全网络的代谢流信息基础上的合理简化, 而所得的粗粒化图可看作是代谢网络的一个宏观结构框架. 这样, 原来交错而复杂的代谢网络被简化成一个清晰的分枝状的展开蝴蝶结模型. 这将帮助我们从全局上了解生物代谢流.

2.2 全网络和 GSC 的可逆性

为了研究代谢网络中可逆反应的分布特征, 我们对 Ma 等人数据库中 75 个物种的代谢网络, 分别计算了全网络和 GSC 部分可逆性指数. 详细结果参见网络版附录A. 图 3 分别画出了这些物种的全代谢网络和 GSC 部分的可逆性指数的正态概率图. 两个图中的

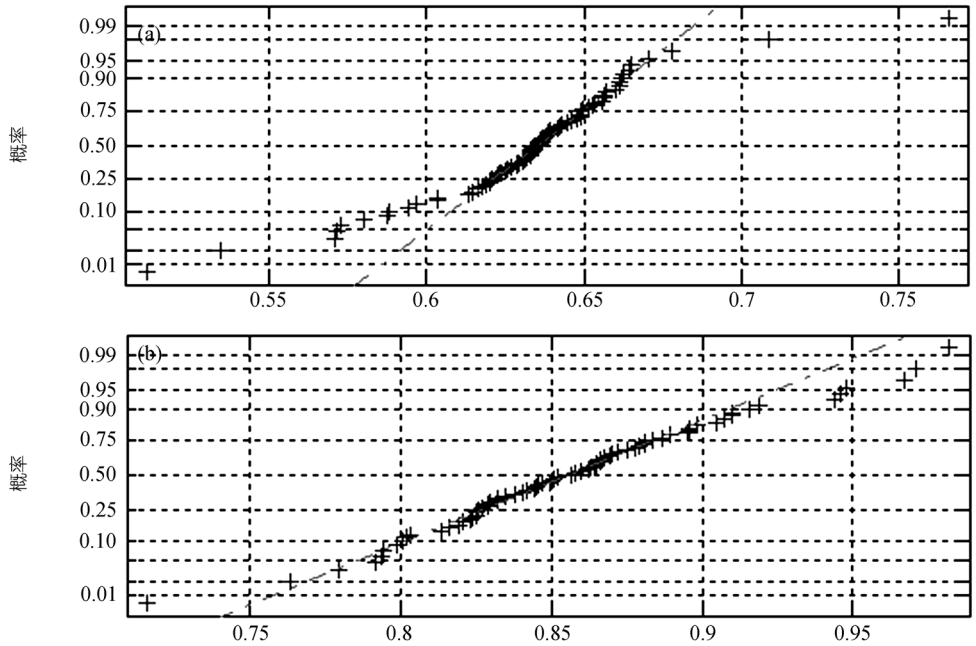

图 3 Ma 和 Zeng 数据库中 75 个物种的全代谢网络的可逆性指数(a)和 GSC 部分的可逆性指数的正态概率图(b)

点都近似于线性分布，说明 75 个物种的全代谢网络和 GSC 部分的可逆性指数都服从正态分布. 由最小方差无偏估计，得出全网络可逆性指数的均值和方差分别为 0.6324 和 0.0344，在 GSC 部分则分别为 0.8575 和 0.0482. 因此, GSC 部分的平均可逆性指数显著高于全网络. 这说明代谢网络的 GSC 部分富集可逆反应. 全网络和 GSC 部分的正态分布均值的显著差异也说明，可逆反应的这种分布特性是代谢网络内在的拓扑特性.

2.3 GSC 的主核

作为网络中连接最稠密的部分，主核可能在网络功能方面有重要意义. 通过研究 GSC 部分的连接密度，我们发现 *E. coli* 代谢网络的 GSC 部分有 2 个主核. 如图 4，这 2 个主核是 3-核(3-core)，由 2 个三角形状的结点(对应代谢物 2-脱水-3-脱氧-6-磷酸-D-葡萄糖酸和甲基乙二醛)连接它们.

当对应到 KEGG 代谢途径数据库[2,3](http://www.genome.jp/kegg/pathway.html)时，发现主核中大多数反应属于碳代谢，而仅有 6 个代谢物参与氨基酸代谢. 引人注意的是，三个重要代谢途径—糖酵解，三羧酸循环(TCA)和戊糖磷酸循环占主核中碳代谢的大多数. 而且，共同用于生物合成的 12 个前体代谢物中[22]，有 6 个包含在主核中. 这些发现说明，从代谢网络拓扑结构上看，作为生命基本代谢的碳代谢，尤其是糖酵解，三羧酸循环(TCA)和戊糖磷酸循环对随机突变是有高度容错性的.

2.4 *E. coli* 代谢网络与随机网络集的比较

Maslov等人[32,34]提出，真实网络应与其保留了一些基本拓扑特征(如顶点度分布等)的恰当随机化的网络比较. 而真实网络的某个拓扑特征，只有当与其零模型的对应拓扑特征具有统计显著性的偏差时，才是有意义的. 因此，通过比较代谢网络及其随机化的对照网络，能揭示它们之间内在的差异.

在此研究中，我们了解到双向弧的分布特征是代谢网络内在的特性. 因此，在构造代谢网络的随机化零模型时，我们不但保留每个顶点的连接度[32,34]，还保留每个顶点所连的单向弧和双向弧总数. 这里用我们在文献[33]中提出的算法构造*E.coli*网络的 60 个随机化对照网络. 这 60 个随机网络的拓扑参数详见附录B，与*E.coli*代谢网络的比较见表 2. 拓扑分析显示，这些随机网络的宏观结构仍然是蝴蝶结，但与*E.coli*代谢网络相比，它们在蝴蝶结各部分的大小以及网络的集聚系数方面都有显著差异.

由表 2 可见，*E.coli* 代谢网络的 GSC 部分及 2-核都分别比随机网络相应小很多. 更为重要的是，随机网络中没有 3-核. 由 *Z*-score 的值说明这种差异不是

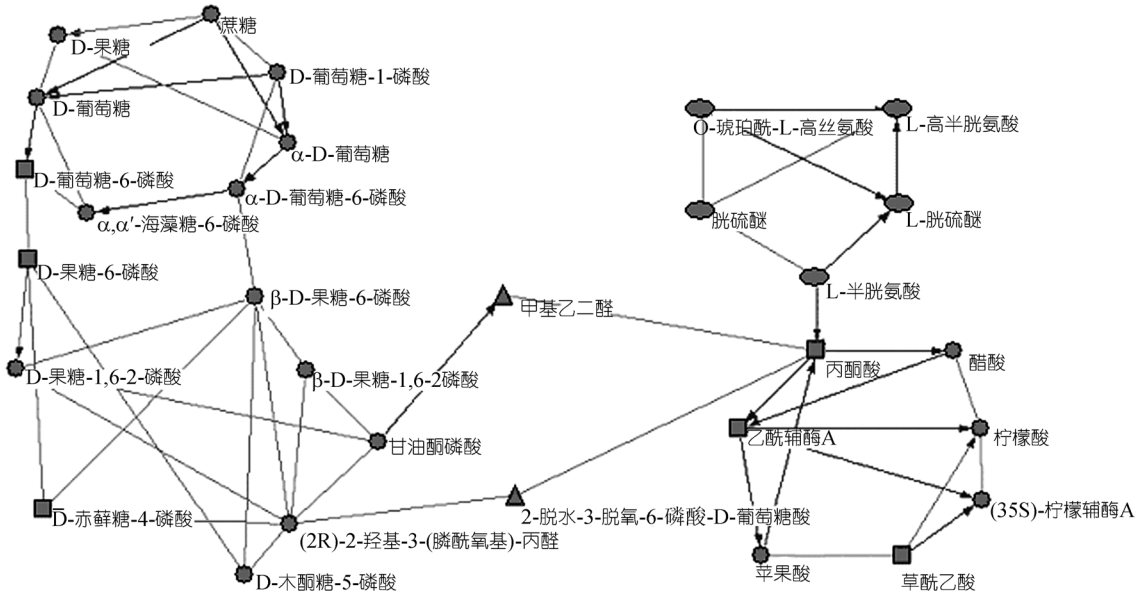

图 4 GSC 的两个主核以及直接连接它们的两个代谢物(三角形结点)

椭圆形结点表示的代谢物参与氨基酸代谢, 其他所有的代谢物参与碳代谢; 主核中包含 6 个前体代谢物(方形结点); 为简化图形, 可逆反应用无向边表示

表 2 *E.coli* 代谢网络与 60 个随机网络间的比较

		GSC	S	P	IS	2-核	3-核	C[a)]
60 个随机网络	均值	287	90	126	71	205	0	0.0027
	标准差	15.86	10.23	14.37	13.72	12.43	0	0.0019
E. coli 代谢网络		234	85	177	79	163	28	0.0646
Z-score		−3.40	−0.52	3.53	0.61	−3.37	∞	31.91

a) C 为网络的平均集聚系数

随机发生的. 表 2 还说明, 与*E.coli*代谢网络截然不同, 随机网络的集聚系数接近于 0. 这些显著的差异说明, 随机网络整体连接比较松散平均, 而代谢网络有一些更加紧致的区域. 从图 5 中可明显看出它们 2-核的不同连接模式. *E. coli*代谢网络的 2-核有一些明显的紧密连接区域, 这些区域之间的连接较为稀疏. 而随机网络的整个 2-核都以大致相同的密度连接, 其中没有明显的“团”状结构, 尤其是没有 3-核. 这些比较结果说明代谢网络的蝴蝶结结构是它自身具有的有意义的拓扑特征, 而不是随机产生的. 这种结构模式可能是代谢途径内的局部相互作用的结果, 这种特征也与代谢网络的模块化组织模式相符合[33,35~37].

2.5 蝴蝶结结构对代谢的意义

代谢网络的拓扑结构可以是其形成与进化的动力学反映, 而蝴蝶结结构也许是构造代谢网络的简单而巧妙的方式. 从标准的生物化学的观点看, 代谢系统的组织方式可看成一个蝴蝶结, 其中心由少量的活性载体(即一些通用代谢物)和12种前体代谢物构成, 有大量的营养物通过分解代谢输入此中心, 而中心内的物质通过生物合成途径输出多种产物[22,38]. 研究发现许多不同的生物学系统如信号转导系统、转录与翻译过程以及免疫系统, 也具有这种组织模式[22,25,38~41]. 而本文研究的蝴蝶结模型, 从拓扑结构的角度对生物信息流给出了另一种诠释, 这个蝴蝶结的中心比上面提到的蝴蝶结的中心包含更多的物质. 我们将用小蝴蝶结和大蝴蝶结来区分这两个模型. 这两个蝴蝶结模型的相似之处是它们都特别地处理通用代谢物. 要注意的是, 除通用代谢物外, 小蝴蝶结模型的中心仅包含 12 种前体代谢物, 而大蝴蝶结模型也包含这 12 种物质, 此外还有生成它们的 3 个重要代谢通路——三羧酸循环, 戊糖磷酸循环及糖酵解通路, 以及更多的其他代谢物和反应. 尽管这两种蝴蝶结模型在细节上有不同之处, 它们相似的组织模式都有利于构建一个复杂的系统, 使之能同时满足调控的稳健性、基因组规模及生物化学上可能的酶的数量等方面的复杂要求[22].

本文蝴蝶结模型的中心是网络中连接较紧密的

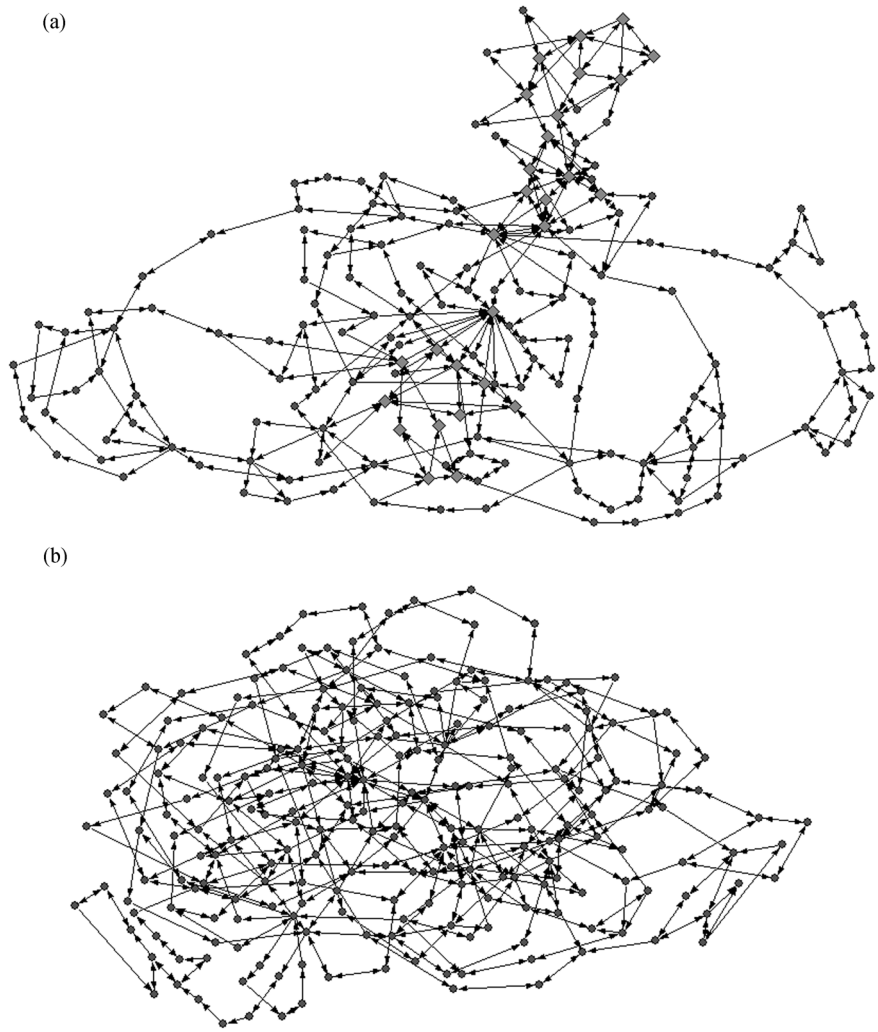

图 5 *E.coli* 代谢网络与随机网络的 2-核的比较

(a) *E.coli* 代谢网络的 2-核，共 163 个顶点，其 3-核用菱形顶点表示; (b) 随机网络的 2-核，该随机网络的 GSC, S, P 与 IS 部分分别包含 302, 84, 134 和 55 个顶点, 2-核包含 227 个顶点，它没有 3-核

部分，由大量中间代谢物构成. 这个较大的中心可能使其能对网络进行更灵活的控制，并提供更多的输入和输出接口，以应对紧急情况或生成新的产物. 另一方面，这个更大的中心可能揭示生物在营养物和产物之间相互转换的灵活性. *E. coli* 及大多数别的生物都大量地利用其他物种的代谢作为营养，同时也能仅依赖极少量的营养物而成活. 这种大的中心也许能反应这种灵活性，但需要进一步的研究来发现它们之间的联系.

蝴蝶结结构的另一个意义在于网络的稳健性. 代谢网络蝴蝶结的GSC部分对突变而言是最稳健的，因为GSC中任两个结点间都有多条通路[14,25]; 更重要的是，所有的 12 种前体代谢物都包含在*E.coli*代谢网络的GSC中，其中一半包含在网络最紧密的连接区域 3-核中，这显示了代谢网络的保护机制. 反之，GSC外低连接度的区域应该是整个系统功能最薄弱的地方. 这与Palumbo等人[10]及Samal等人[11]研究*S.cerevisiae*和*E.coli*代谢网络中致命酶(essential enzyme)分布特点所得出的结论一致. 由于展开蝴蝶结模型能可视化网络中的薄弱连接，即去除单线弧通常导致网络破碎，它可能会有助于寻找代谢网络的薄弱连接或药靶筛选.

3 结论

作为在探索复杂代谢网络的结构与功能方面的新的尝试，本文在合理简化代谢网络的基础上提出

了展开蝴蝶结模型. 此模型对错综复杂的代谢通路中的生物信息流提供了清晰的可视化, 可用于研究代谢网络的薄弱环节. 通过对蝴蝶结的 GSC 部分的可逆性参数和主核的研究, 发现 GSC 部分富集可逆反应且呈现成团化特征(即模块化), 而在统计可比的随机网络中却没有这些特征. 本研究的结果使我们对代谢网络的设计机理有了更深入的了解.

致谢 感谢 Ma H W 和 Zeng A P 提供的代谢网络数据库.

参 考 文 献

1 Goto S, Nishioka T, Kanehisa M. LIGAND: Chemical database of enzyme reactions. Nucl Acids Res, 2000, 28: 380—382[DOI]
2 Goto S, Okuno Y, Hattori M, et al. LIGAND: Database of chemical compounds and reactions in biological pathways. Nucl Acids Res, 2002, 30: 402—404[DOI]
3 Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucl Acids Res, 2000, 28: 27—30[DOI]
4 Karp P D, Krummenacker M, Paley S, et al. Integrated pathway-genome databases and their role in drug discovery. Trends Biotechnol, 1999, 17: 275—281[DOI]
5 Ma H, Zeng A P. Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. Bioinformatics, 2003, 19: 270—277[DOI]
6 Overbeek R, Larsen N, Pusch G D, et al. WIT: Integrated system for high-throughput genome sequence analysis and metabolic reconstruction. Nucl Acids Res, 2000, 28: 123—125[DOI]
7 Lemke N, Heredia F, Barcellos C K, et al. Essentiality and damage in metabolic networks. Bioinformatics, 2004, 20: 115—119[DOI]
8 Patil K R, Nielsen J. Uncovering transcriptional regulation of metabolism by using metabolic network topology. Proc Natl Acad Sci USA, 2005, 102: 2685—2689[DOI]
9 Stelling J, Klamt S, Bettenbrock K, et al. Metabolic network structure determines key aspects of functionality and regulation. Nature, 2002, 420: 190—193[DOI]
10 Palumbo C, Giuliani F. Functional essentiality from topology features in metabolic networks: A case study in yeast. FEBS Lett, 2005, 579: 4642—4646[DOI]
11 Samal A, Singh S, Giri V, et al. Low degree metabolites explain essential reactions and enhance modularity in biological networks. BMC Bioinformatics, 2006, 7: 118[DOI]
12 Albert R, Barabasi A L. Statistical mechanics of complex networks. Rev Mod Phys, 2002, 74: 47—97[DOI]
13 Jeong H, Mason S P, Barabasi A L, et al. Lethality and centrality in protein networks. Nature, 2001, 411: 41—42[DOI]
14 Ma H W, Zeng A P. The connectivity structure, giant strong component and centrality of metabolic networks. Bioinformatics, 2003, 19: 1423—1430[DOI]
15 Wagner A, Fell D A. The small world inside large metabolic networks. Proc R Soc Lond B, 2001, 268: 1803—1810[DOI]
16 Arita M. The metabolic world of Escherichia coli is not small. Proc Natl Acad Sci USA, 2004, 101: 1543—1547[DOI]
17 Arita M. Scale-freeness and biological networks. J Biochem, 2005, 138: 1—4[DOI]
18 Tanaka R. Scale-rich metabolic networks. Phys Rev Lett, 2005, 94: 168101[DOI]
19 Tanaka R, Yi T M, Doyle J. Some protein interaction data to not exhibit power law statistics. EBS Lett, 2005, 579: 5140—5144
20 赵静, 俞鸿, 骆建华, 等. 应用复杂网络理论研究代谢网络的进展. 科学通报, 2006, 51: 1241—1248
21 Erdos P, Renyi A. On the evolution of random graphs. Publ Math Inst Hungar Acad Sci, 1960, 5: 17—61
22 Csete M, Doyle J, Bow ties. Metabolism and disease. Trends Biotechnol, 2004, 22: 446—450[DOI]
23 Ma H W, Zhao X M, Yuan Y J, et al. Decomposition of metabolic network into functional modules based on the global connectivity structure of reaction graph. Bioinformatics, 2004, 20: 1870—1876[DOI]
24 Broder A, Kumar R, Maghoul F, et al. Graph structure in the web. Compu Networks, 2000, 33: 309—320[DOI]
25 Kitano H. Biological robustness. Nat Rev Genet, 2004, 5: 826—837[DOI]
26 Seidman S B. Network structure and minimum degree. Soc Networks, 1983, 5: 269—287[DOI]
27 Wuchty S, Almaas E. Peeling the yeast protein network. Proteomics, 2005, 5: 444—449[DOI]
28 Garlaschelli D, Loffredo M I. Patterns of link reciprocity in directed networks. Phys Rev Lett, 2004, 93: 268701[DOI]
29 Bondy J A, Murty U S R. Graph Theory With Applications. London: Macmillan, 1976
30 Batagelj V, Mrvar A. Pajek-program for large network analysis. Connections, 1998, 21: 47—57
31 Watts D J, Strogatz S H. Collective dynamics of 'small-world' networks. Nature, 1998, 393: 440—442[DOI]
32 Maslov S, Sneppen K, Zaliznyak A. Detection of topological patterns in complex networks: Correlation profile of the internet. Physica A, 2004, 333: 529—540[DOI]
33 Zhao J, Yu H, Luo J, et al. Hierarchical modularity of nested bow-ties in metabolic networks. BMC Bioinf, 2006, 7: 386[DOI]
34 Maslov S, Sneppen K. Specificity and stability in topology of protein networks. Science, 2002, 296: 910—913[DOI]
35 Ravasz E, Somera A L, Mongru D A, et al. Hierarchical organization of modularity in metabolic networks. Science, 2002, 297: 1551—1555[DOI]
36 Spirin V, Gelfand M S, Mironov A A, et al. A metabolic network in the evolutionary context: Multiscale structure and modularity. Proc Natl Acad Sci USA, 2006, 103: 8774—8779[DOI]
37 Holme P, Huss M, Jeong H, et al. Subnetwork hierarchies of biochemical pathways. Bioinformatics, 2003, 19: 532—538[DOI]
38 Tanaka R, Csete M, Doyle J. Highly optimised global organisation of metabolic networks. IEE Proc Syst Biol, 2005, 152(4): 179—184[DOI]
39 Kitano H, Oda K. Robustness trade-offs and host-microbial symbiosis in the immune system. Mol Syst Biol, 2006, 2: E1—E10
40 Kitano H, Oda K, Kimura T, et al. Metabolic Syndrome and Robustness Tradeoffs. Diabetes, 2004, 53: S6—S15
41 Marhl M, Perc M, Schuster S. Selective regulation of cellular processes via protein cascades acting as band-pass filters for time-limited oscillations. FEBS Lett, 2005, 579: 5461—5465

附录

A 75 个物种的代谢网络及 GSC 部分的顶点分布及可逆性参数

表 A1 8 个真核生物

物种名称	全网络		GSC		
	顶点	ρ	顶点	在全网络中所占比例	ρ
hsa	651	0.5945	296	45.47%	0.7914
mmu	352	0.5345	44	12.50%	0.8891
rno	346	0.6327	115	33.24%	0.8716
dme	474	0.6036	183	38.61%	0.8296
cel	428	0.6378	155	36.21%	0.8778
ath	404	0.5718	120	29.70%	0.8239
sce	470	0.5971	163	34.68%	0.7634
spo	371	0.5799	64	17.25%	0.7154

表 A2 56 种细菌

物种名称	全网络		GSC		
	顶点	ρ	顶点	在全网络中所占比例	ρ
eco	575	0.6475	234	40.70%	0.801
ecc	499	0.6181	170	34.07%	0.823
sty	534	0.6413	210	39.33%	0.7944
stm	594	0.6523	217	36.53%	0.8018
ype	529	0.6253	175	33.08%	0.8294
ypk	503	0.6339	165	32.80%	0.8452
sfl	530	0.6382	211	39.81%	0.7936
hin	324	0.6556	101	31.17%	0.8412
pmu	374	0.6281	112	29.95%	0.8161
xfa	331	0.5709	72	21.75%	0.8488
xcc	476	0.6326	137	28.78%	0.84
xac	487	0.6325	140	28.75%	0.844
vch	427	0.6286	150	35.13%	0.8273
pae	505	0.6347	186	36.83%	0.7986
son	417	0.6565	151	36.21%	0.8322
nme	348	0.6457	90	25.86%	0.8592
rso	519	0.6484	211	40.66%	0.7795
cje	265	0.6229	68	25.66%	0.8288
mlo	562	0.6359	174	30.96%	0.8033
sme	593	0.6364	184	31.03%	0.8133
atu	510	0.6488	173	33.92%	0.8196
bme	549	0.6642	186	33.88%	0.8324
bms	468	0.6564	134	28.63%	0.834
ccr	460	0.6703	132	28.70%	0.8517
bsu	507	0.6207	164	32.35%	0.8251
bha	478	0.6361	166	34.73%	0.8207
oih	471	0.678	180	38.22%	0.8468
sav	409	0.6314	115	28.12%	0.8648
sam	419	0.6266	115	27.45%	0.8648
lmo	397	0.6529	125	31.49%	0.8376
lin	366	0.6487	116	31.69%	0.8444
lla	363	0.6248	96	26.45%	0.8623
spy	254	0.6132	64	25.20%	0.8955
spn	345	0.5876	73	21.16%	0.867
san	268	0.6328	71	26.49%	0.8836
smu	336	0.6165	79	23.51%	0.8697

表 A1(续)

物种名称	全网络		GSC		
	顶点	ρ	顶点	在全网络中所占比例	ρ
cac	361	0.6491	115	31.86%	0.8662
tte	348	0.6202	69	19.83%	0.8806
mpu	100	0.6335	34	34.00%	0.9462
uur	57	0.5109	18	31.58%	0.9188
mtu	452	0.6418	147	32.52%	0.8256
mle	381	0.656	123	32.28%	0.8252
cgl	322	0.6424	86	26.71%	0.864
sco	482	0.6145	138	28.63%	0.8564
fnu	308	0.6306	79	25.65%	0.9155
ctr	124	0.6621	32	25.81%	0.9708
bbu	58	0.7661	27	46.55%	0.9477
tpa	111	0.7088	42	37.84%	0.9672
lil	434	0.6232	97	22.35%	0.8957
syn	339	0.6615	96	28.32%	0.8789
tel	299	0.6648	79	26.42%	0.8981
ana	343	0.6513	113	32.94%	0.8576
cte	295	0.6193	68	23.05%	0.8951
dra	349	0.6387	112	32.09%	0.8496
aae	279	0.6443	65	23.30%	0.9071
tma	354	0.5728	69	19.49%	0.9047

表 A3 11 种古细菌

物种名称	全网络		GSC		
	顶点	ρ	顶点	在全网络中所占比例	ρ
mja	183	0.5879	28	15.30%	0.9822
mac	235	0.6421	71	30.21%	0.866
mma	252	0.6302	45	17.86%	0.8808
afu	228	0.6382	36	15.79%	0.944
hal	178	0.6368	65	36.52%	0.8501
tac	245	0.6289	69	28.16%	0.863
tvo	210	0.6601	70	33.33%	0.8694
ape	256	0.6613	73	28.52%	0.9098
sso	249	0.6039	62	24.90%	0.8748
sto	217	0.62	58	26.73%	0.9099
pai	249	0.6612	64	25.70%	0.8867

B *E.coli* 代谢网络的随机对照网络

表 B1 *E.coli* 代谢网络的 60 个随机对照网络的拓扑参数

Network	GSC	S	P	IS	2-核	C[a)]
1	296	111	107	61	212	0.0002
2	270	86	138	81	189	0.0007
3	300	90	133	52	221	0.0036
4	237	116	105	117	161	0.0092
5	313	75	134	64	220	0.0052
6	305	84	122	64	223	0.0034
7	284	85	150	56	202	0.004
8	304	94	132	45	219	0.0011
9	296	88	120	71	213	0.0027
10	291	81	151	52	212	0.0016
11	291	71	118	95	199	0.0005

表 B1(续)

Network	GSC	S	P	IS	2-核	C[a)]
12	297	81	129	68	205	0.0031
13	295	100	126	54	209	0.0007
14	300	71	133	71	205	0.0061
15	310	92	118	55	220	0.0022
16	289	86	132	68	207	0.0019
17	283	99	120	73	203	0.003
18	282	89	126	78	192	0.0063
19	289	93	108	85	203	0.0014
20	297	85	114	79	202	0.0024
21	275	97	130	73	195	0.0018
22	293	74	125	83	204	0.0023
23	311	84	134	46	223	0.0039
24	271	83	132	89	194	0.0012
25	259	89	147	80	187	0.002
26	272	94	146	63	195	0.0018
27	300	89	110	76	219	0.0031
28	274	99	110	92	197	0.0062
29	307	78	122	68	218	0.0019
30	264	108	145	58	190	0.0014
31	276	107	127	65	209	0.0041
32	291	86	125	73	207	0.0021
33	284	97	136	58	207	0.0046
34	297	90	120	68	212	0.0027
35	296	78	112	89	211	0.0026
36	313	84	115	63	227	0.0029
37	291	86	123	75	202	0.0031
38	270	102	128	75	198	0.0009
39	290	105	112	68	208	0.0009
40	279	113	102	81	207	0.0035
41	278	86	137	74	200	0
42	289	89	112	85	198	0.0035
43	286	81	137	71	200	0.0019
44	270	88	164	53	185	0.0022
45	303	79	102	91	205	0.004
46	254	93	155	73	187	0.0076
47	296	102	124	53	205	0.0007
48	284	88	109	94	204	0.0006
49	273	106	121	75	199	0.0049
50	307	87	102	79	226	0.0008
51	302	84	134	55	227	0.0049
52	260	107	121	87	194	0.0048
53	294	85	125	71	214	0.0014
54	284	92	142	57	197	0.0008
55	286	97	120	72	206	0.0018
56	266	80	158	71	183	0.0059
57	305	83	131	56	219	0.0005
58	310	94	107	64	216	0.0016
59	293	89	132	61	194	0.0027
60	290	90	130	65	210	0.0012
eco	234	85	177	79	163	0.0646

a) C 为网络的平均聚集系数