文章编号: 1001- 2486(2008) 03- 0048- 05

大规模代谢网络分解的生物信息学研究*

王正华, 周婷婷
(国防科技大学 并行与分布处理国家重点实验室, 湖南 长沙 410073)

摘 要: 随着大规模分子相互作用数据的不断涌现, 生物学网络方面的研究正日益得到重视。代谢网络处于生物体的功能执行阶段, 其结构组成方式不仅反映了生物体的功能构成, 也直接影响代谢工程中的途径分析和研究。作为代谢网络研究的重要环节, 实现网络的合理分解不仅对于基因组范围内分子网络的结构和功能研究具有重要意义, 也是代谢工程的途径分析和优化得以顺利进行的前提之一。在回顾代谢网络宏观结构和拓扑特征研究成果的基础上, 通过对现有分解方法的深入分析, 指出缺乏合理且有针对性的模型评估准则是目前网络分解研究中亟待解决的问题之一。今后的研究趋势在于如何整合更多的信息和发展更先进的分析方法, 建立更合理的模型, 并进一步拓展网络分解的应用范围。

关键词: 生物信息学; 代谢网络; 网络分解; 评估准则
中图分类号: TP301. 6 文献标识码: A

A Review on Bioinformatics Analysis of Genome-scale Metabolic Network Decomposition

WANG Zheng-hua, ZHOU Ting-ting
(National Laboratory for Parallel & Distributed Processing, National Univ. of Defense Technology, Changsha 410073, China)

Abstract: As the large-scale data on molecular interaction available increases, the study on biological networks has ignited more and more attention. As the execution level of cellular functions, the structural composition of metabolic networks not only reflects the execution of cellular functions step by step, but also influences the pathway analysis of metabolic engineering. Hence, as a key step, a sound decomposition of global metabolic networks is not only significant for further exploring the structure and function of genome-scale biological networks, but also necessary for favorably developing the pathway analysis of metabolic engineering. Having reviewed the fruitful study of the macro-structure and topological characteristics, and analyzed some available decomposition approaches, this paper concludes that the lack of reasonable and purposeful evaluation criterions is one urgent problem in the present study of metabolic network decomposition. Besides, the future trend is to develop more effective decomposition models with more information and by more advanced approaches, as well as expanding the application range of network decomposition.

Key words: bioinformatics; metabolic network; network decomposition; evaluation criterion

随着大规模测量技术的广泛应用和大量分子相互作用数据的出现, 近年来各种分子相互作用网络的研究正逐渐成为生物信息学领域的研究热点。根据这些数据构建的生物学网络规模越来越大, 使得研究和分析愈加困难。因此, 如何在尽量不破坏某些网络特性(如结构与功能的对应关系)的前提下, 将大规模生物网络按照一定原则划分为结构和功能相对独立的组成部分, 正是当前生物学网络研究的难点之一。

代谢处于生命活动的功能执行层面, 是驱动生命过程的化学引擎。代谢网络把细胞内所有生化反应表示为网络形式, 反映了代谢活动中所有化合物及酶之间的相互作用。代谢工程是在代谢网络系统分析的基础上对细胞代谢的控制和改造过程, 其应用和发展关系着国计民生。作为代谢工程的重要组成部分, 以基元模式分析和极端途径分析为常用手段的途径分析一直是人们关注的重点。但由于存在组合爆炸的问题, 目前的途径分析方法仅能够应用在单途径分析或对几条简单途径构成的小规模代谢

* 收稿日期: 2007- 11- 20
基金项目: 国家自然科学基金资助项目(60603054); 国家重点实验室开放研究基金资助项目
作者简介: 王正华(1962—), 男, 教授, 博士生导师。

网络进行分析的情况。随着代谢网络规模的逐渐增大, 人们更希望能够从整个网络的角度实现代谢工程的分析和控制。因此, 对于大规模代谢网络而言, 需要首先根据其拓扑结构特征分解为结构和功能相对独立的子网络。然而目前的分解方法大都将整个代谢网络硬性分割, 造成途径分析中基元模式的丢失, 而直接影响了代谢工程的控制和优化。因此, 研究更加合理的代谢网络分解方法势在必行。

代谢网络所具有的宏观结构和拓扑特征保证了网络分解的合理性, 而机器学习中各种先进算法和数据挖掘中聚类分析手段的应用则使得代谢网络的分解更加有效。本文将在简要叙述代谢网络结构特征的基础上, 对大规模代谢网络分解模型的研究现状进行分析和综述, 并对目前存在的问题和今后的发展趋势提出一些看法。

1 代谢网络的宏观结构和拓扑特征

通常代谢网络以图的形式表示, 代谢物(或酶) 和反应分别表示成顶点和边。借助于复杂网络理论, 人们以顶点的度、度分布、平均聚集系数、平均路长等静态几何量来描述网络的拓扑性质。

大量研究成果表明[2], 代谢网络的度分布具有幂律形式, 说明细胞的代谢网络具有无尺度(Scale-free) 特征, 即反应中多数代谢物只参与少数的反应, 而少数代谢物则参与多数反应, 发挥着代谢中枢的作用; 代谢网络的平均路长很小且与网络规模成对数递增关系, 而平均聚集系数则远大于同等规模的随机网络, 说明代谢网络具有小世界(Small-world) 特征, 即大多数代谢物或酶之间只需极少的几步就可以完成相互转化, 且代谢物的浓度变化在很短的时间内就会传遍整个网络。代谢网络的平均聚集系数远大于同等规模的无尺度网络, 表明模块化是生物体内代谢网络高度固有的潜在特征。而顶点聚集系数服从尺度法则, 则表明连接度越高, 聚集系数越小, 即顶点的分布具有层次性。代谢网络的这些拓扑性质, 特别是分层模块化的特征, 正是代谢网络能够合理分解的理论依据。

另外, Ma 等人[3] 发现代谢网络宏观上呈现一种类似蝴蝶结(Bow-tie) 的结构(如图 1) 。全部顶点根据连接度划分为四部分, 庞大强连通体 GSC 是其中最重要的结构和功能构成。Zhao 等人[4] 则发现此蝴蝶结结构也具有分层嵌套特征, 即这四部分中也具有分层模块化的性质。代谢网络的这种宏观结构特征使得分解方法的设计更有针对性。

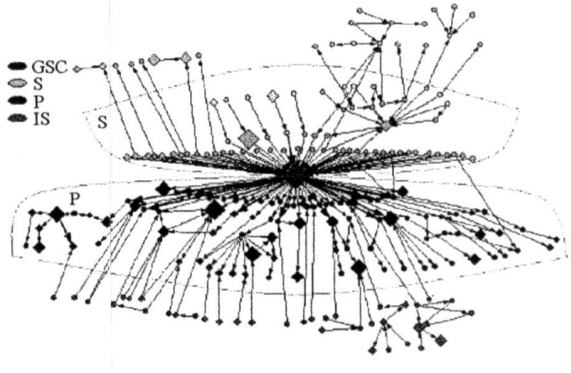

图 1 E. Coli 代谢网络的蝴蝶结特征, 中间最为集中的“ ◇” 部分为 GSC。此图取自文献[11]

Fig. 1 The bow-tie structure of E. Coli metabolic network

2 代谢网络的分解

代谢网络的分解原则是最大限度保持模块间的独立性和模块内代谢物及反应信息的完整性。体现在途径分析方面, 则要求对基元模式的破坏程度最小。

目前的分解方法大致可以分为两类: 一是基于生物学意义的划分, 按照代谢物及反应的生化功能(如糖代谢、氨基酸代谢、脂代谢等) , 直观地将代谢网络划分为子网[5]; 二是基于网络结构特征的划分, 即基于生物信息学的研究应用聚类算法实现网络分解。前者虽然能够最大限度地满足网络分解原则, 却需要以大量已知反应信息为基础, 无法适应网络规模日益增大的趋势; 后者则能够根据网络结构与功

能的对应关系,“自动”地将大规模代谢网络分解为“内紧外松”的子网络。因此对于基于全基因组信息重建的代谢网络分解,更倾向于应用基于拓扑结构特征的各种生物信息学方法。

2.1 非层次聚类法(Nonhierarchical Clustering)

非层次聚类法一般直接根据代谢网络的无尺度性质进行网络分解。无尺度网络具有顶点偏好依附的性质,网络中多数顶点的连接边较少,而少数顶点则拥有大量的连接边。因此,按照一定原则删除某些高连接度顶点就会断开大量的连接边,从而使网络被割裂成为相对独立的部分。这些连接度较高的代谢物被定义为外部代谢物① (External)。

Schuster等人[6]的方法较为简单直观。他们以保证子网络具有合适的规模为原则设定顶点连接度的阈值,将度值高于阈值的代谢物作为外部代谢物删除,实现网络的分解。Dandekar等人[7]则首先从数学角度罗列出所有可能的外部代谢物集合,然后以使基元模式数量最少为原则,利用基于Monte Carlo思想建立的Metropolis算法完成内外代谢物的判定及网络的分解。Huss等人[8]则根据Newman提出的 Q 函数,利用矩阵特征值运算来判定冗余(Abundant Substrates)代谢物及分解网络。若删除某个节点能够使得网络模块性 Q 增加,则该节点就被定义为冗余代谢物。这种网络分解的方法避开了由于代谢物功能复杂而易产生混淆的问题,从数学角度讲更为严谨。

通过删除外部代谢物实现网络分解的方法虽然直观,却并未完全体现出要求模块“内紧外松”的分解原则。于是人们基于优化思想和随机流理论提出了模拟退火算法(SACL)和马尔可夫算法(MCL)。模拟退火算法以Newman提出的模块性 M 的负值作为评价函数,在每个温度 T 通过随机移动和随机合并两类运动来改变网络模块结构;而马尔可夫聚类算法则无需预先设定聚类数目,以“直到一个紧密连接的模块中多数节点已被访问过随机流才会离开”为原则,通过概率改变和矩阵的反复修改来实现随机流模拟。Guimera等人[9]利用模拟退火算法对12种生物体的代谢网络进行了研究,以要求模块内部顶点连接尽可能紧密而模块间连接尽可能稀疏作为优化准则,当模块性 M 达到最大值时得到网络的最优分解,发现绝大部分模块具有相对独立的生物学功能。侯静[10]等人通过研究叶绿体(Chlo)及蓝细菌(Syw)的代谢网络对模拟退火和马尔可夫聚类两种方法的分解效果进行了对比,认为模拟退火算法比马尔可夫聚类算法得到的划分结果具有更好的模块性。从代谢网络功能分类及分析的角度来讲,模拟退火算法更胜一筹。

2.2 层次聚类法(Hierarchical Clustering)

层次聚类法大都以树状结构体现网络模块的层次性,定义任意两点间的相似指数(Similarity Index)或相异指数(Dissimilarity Index)来量化两点属于同一模块的概率[11]。层次聚类法一般分为凝聚法(Agglomerative Clustering Method)和分裂法(Divisive Method)两种。前者基于“bottom-up”思想,初始状态每个点都作为一个单独的子集,每次合并两个相似度最高的子集,直到所有相似顶点合并成一类为止;而后者则属于“top-down”方法,初始时把所有点视为一个集合,每次在相异度最高处进行分割,直到所有点分属不同类为止。

Ravasz等人[13]以E. Coli代谢网络为研究对象,首先在降低复杂度而不改变网络拓扑结构的基础上,将代谢网络表示成更为紧凑的形式;之后定义顶点间的拓扑重叠度(Topological Overlap)为相似指数,计算整个代谢网络的拓扑重叠矩阵。基于重叠越多的代谢物属于同一模块的可能性越大的假设,采用凝聚法构建网络分解模型。分解结果表明,同属一种生化分类的代谢物之间连接紧密,再一次验证了代谢网络结构和功能的对应关系。同样采用层次凝聚法,Ma等人[12]则从代谢网络的蝴蝶结结构特征入手,定义相异指数为顶点间最短的有向距离,首先对网络的GSC部分进行分解,然后按照与GSC分解模块中顶点连接关系的强弱将剩余顶点划分到相应模块中去,完成整个网络的分解。这种方法事先考虑到了GSC部分的重要性,相对而言生物学意义更强。Holme等人[14]则将代谢网络表示为化合物–反应

① 此处的内外部代谢物仅是网络构成意义上的区分,位于子网外的为外部代谢物,子网内连接紧密的为内部代谢物。

二部有向图, 定义化合物顶点间的相异指数为与其相连的反应顶点的介数①, 依照介数由大到小的顺序依次去掉相应的反应顶点和边, 以层次分裂法逐步将网络分解为子网。他们的分解结果再次印证了代谢网络具有蝴蝶结结构特征。

传统的聚类分析[15] 认为, 非层次聚类只是简单地将网络划分为不重叠的子集, 属于划分聚类(Partitional Clustering); 而层次聚类则兼顾到网络的层次关系, 其分解结果可以看作子网络的嵌套。因此, 基于层次聚类建立的各种分解模型更多地考虑了代谢网络的层次模块性, 设计和使用都更加合理。

3 网络分解中存在的问题

代谢网络的分解主要有两大目的, 一是加深对网络结构和功能的理解, 二是缓解途径分析研究中基元模式计算的组合爆炸趋势。虽然人们借助于机器学习、数据挖掘等领域的分析方法已经建立了许多有效的分解模型, 然而由于代谢网络本身的复杂性, 研究中仍存在着许多问题亟待解决。

3.1 合理定义外部代谢物

非层次聚类中通过定义和删除外部代谢物实现网络分解的方法, 外部代谢物定义的合适与否对于分解模型的优劣有很大的影响。对于以分析结构和功能为目的的网络分解, 若内外代谢物区分不当, 将会影响分解后子网功能的判断, 从而影响整个网络结构和功能的对应关系; 对于以缓解基元模式组合爆炸趋势为目的的网络分解, 若内外代谢物区分不当, 则会造成子网途径中基元模式丢失的情况 ——由于某些重要代谢物被强行定义为外部代谢物而无法满足代谢通量平衡的条件[16], 由此而降低网络分解的意义。目前外部代谢物的判断主要依据静态统计量或从数学角度提出, 判断方法的生物学意义并不强, 而对于上述问题的解决也只能在判断后根据其相应的生化功能做二次判定, 所以对于删除外部代谢物分解网络的模型, 目前需要发展一种既兼顾代谢物生物学意义又易于实现的外部代谢物判定方法, 以提高分解模型的有效性。

3.2 建立合适的模型评估标准

无论是为了何种研究目的分解网络, 分解模型选取不当都会对后续研究造成影响。因此, 在目前的网络分解研究阶段, 亟须提出合适的模型评估标准, 针对不同的应用目的来评估模型的有效性。而就目前研究来看, 缺少合适的模型评估标准, 是目前代谢网络分解研究中存在的最大问题。

我们认为, 作为基于聚类分析建立的分解模型, 其基本的评估标准可以借鉴聚类分析中的簇评估技术, 如簇的凝聚性 CC(Cluster Cohesion) 和分离性 CS(Cluster Separation) 等[15]。然而考虑到研究对象和研究目的的特殊性, 应对此基本评估标准以某种合适的方式加以改进, 如定义分解模型的有效性 V 为两部分评估指标的线性组合: $V = C_1 + C_2$。其中, C_1 为基本的簇评估指标, 可以为 CC、CS 或其他; C_2 为根据代谢网络特征和研究目的提出的附加评估指标。若网络分解的目的为分析代谢网络结构和功能的对应关系, 则 C_2 取值为结构与功能对应程度的评估指标; 若网络分解的目的为缓解基元模式计算的组合爆炸趋势, 则 C_2 则应取值为分解模型对基元模式的保留程度。

3.3 网络分解方法对于途径分析的影响

从途径分析角度考虑, 虽然代谢网络的分解原则已经最大限度地照顾了代谢通路的完整性, 但无论是从生物学角度分解还是从拓扑学角度分解, 仍然不可避免某些代谢途径特别是横跨几个子网之间的途径会被切断, 从而造成基元模式的缺失。

这一问题可考虑利用模糊聚类的思想来解决, 即不把网络硬性割裂成完全独立的子网, 而是根据一定的约束条件, 直接以概率形式来表示交叉部分属于某个子网的"程度", 并让这些元素以概率形式参加基元模式的计算。这种方法避免了因途径被硬性断开而使基元模式缺失的情况, 但缺点在于不仅增加了计算复杂度, 而且需要更多的代谢物和反应的生化特征作为约束。

---

① 顶点的介数(Betweenness) 定义为网络中经过该顶点的所有最短路径数, 反映该顶点在网络中的影响力。

4 研究趋势

代谢网络的分解研究目前还处于起步阶段,分解模型大多根据网络结构功能分析、途径分析等具体研究需要而建立,缺少较为合理和统一的评估准则。提出合理的评估准则并针对不同研究目的建立更有效的分解模型,是今后的研究趋势之一。另外,随着系统生物学研究的不断深入和生物学网络整合程度的不断加深,代谢网络的研究将更多地考虑基因调控、信号转导等其他网络的信息及网络间的联系,因此其分解模型的建立也将要融合更多学科、更多领域的研究方法。

包括代谢网络在内的各种生物学网络共同构成了"网络的网络"。随着研究的深入和认识的加深,生物网络呈现出的状态越来越复杂,需要考虑和整合的信息也越来越多。因此,生物网络分解研究不应该仅是一种研究工具,更应该看做是对于复杂生物网络的细化过程,它也将同网络整合分析一样,成为今后系统生物学中生物网络研究的一个重要方向。

参 考 文 献:

[1] 孙之荣. 后基因信息学[M]. 北京: 清华大学出版社, 2002.

[2] 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[J]. 科学通报, 2006, 51(11): 1241- 1248.

[3] Ma H W, Zeng A P. The Connectivity Structure, Giant Strong Component and Centrality of Metabolic Networks[J]. Bioinformatics, 2003, 19(11): 1423- 1430.

[4] Zhao J, Yu H, Luo J H, et al. Hierarchical Modularity of Nested Bow-ties in Metabolic Networks[J]. BMC Bioinformatics, 2006, 7(386).

[5] Schilling C H, Palsson B O. Assessment of the Metabolic Capabilities of Haemophilus Influenzae Rd through a Genome-scale Pathway Analysis[J]. Journal of Theoretical Biology, 2000, 203(3): 249- 283.

[6] Schuster S, Pfeiffer T, Moldenhauer F, et al. Exploring the Pathway Structure of Metabolism: Decomposition into Subnetworks and Application to Mycoplasma Pneumoniae[J]. Bioinformatics, 2002, 18(2): 351- 361.

[7] Dandekar T, Moldenhauer F, Bulik S, et al. A Method for Classifying Metabolites in Topological Pathway Analyses Based on Minimization of Pathway Number[J]. Biosystems, 2003, 70(3): 255- 270.

[8] Huss M, Holme P. Currency and Commodity Metabolites: Their Identification and Relation to the Modularity of Metabolic Networks[J]. IET System Biology, 2007, 1(5): 280- 285.

[9] Guimera R, Amaral L A N. Functional Cartography of Complex Metabolic Networks[J]. Nature, 2005, 433(7028): 895- 900.

[10] 侯静, 宋安平, 王卓, 等. 图形聚类算法的代谢网络模块化分析[J]. 应用科学学报, 2006, 24(6): 588- 592.

[11] Zhao J, Yu H, Luo J H, et al. Complex Networks Theory for Analyzing Metabolic Networks[J]. Chinese Science Bulletin, 2006, 51(13): 1529 - 1537.

[12] 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[J]. Bioinformatics, 2004, 20(12): 1870- 1876.

[13] Ravasz E, Somera A L, Mongru D A, et al. Hierarchical Organization of Modularity in Metabolic Networks[J]. Science, 2002, 297(5586): 1551 - 1555.

[14] Holme P, Huss M, Jeong H. Subnetwork Hierarchies of Biochemical Pathways[J]. Bioinformatics, 2003, 19(4): 532- 538.

[15] 范明, 范宏建. 数据挖掘导论[M]. 北京: 人民邮电出版社, 2006.

[16] 何锋, 马红武, 赵学明, 等. 生物信息学用于代谢网络研究的进展与展望[J]. 化工学报, 2004, 55(010): 1593- 1601.