文章编号:1673-1689(2008)05-0015-06

复杂生物网络分析的图聚类方法研究进展

梅 娟[1,2]， 王正祥[1,2]， 石贵阳[1,2]， 李炜疆*[1,2]
(1. 江南大学 生物工程学院,江苏 无锡 214122; 2. 江南大学 工业生物技术教育部重点实验室,江苏 无锡 214122)

摘 要: 基因组学和高通量技术提供了大量生命系统组成元件(如蛋白质)之间相互关系的数据,由这些关系数据构成的复杂生物网络蕴含着丰富的生命系统运行机制的知识,挖掘这些隐蔽的知识成为当前系统生物学的主要任务之一。作为知识发现重要手段的图聚类方法,在复杂生物网络分析上受到了普遍关注,在远同源性探测、蛋白质功能预测、代谢途径发现等方面取得了令人瞩目的结果。同时也注意到,由于生命系统的高度复杂性,其他领域中卓有成效的方法往往在复杂生物网络分析中遇到困难。评述了近年来图聚类算法在复杂生物网络分析中的进展,简要分析了复杂生物网络研究的图聚类途径所面临的主要问题。

关键词: 复杂生物网络;图聚类;蛋白质相互作用网络;代谢网络;蛋白质相似性网络
中图分类号:Q 7 **文献标识码**: A

Progress of Graph Clustering on Analysis of Complex Biological Networks

MEI Juan[1,2], WANG Zheng-xiang[1,2], SHI Gui-yang[1,2], LI Wei-jiang*[1,2]
(1. School of Biotechnology, Jiangnan University, Wuxi 214122, China; 2. Key Laboratory of Industrial Biotechnology, Ministry of Education, Jiangnan University, Wuxi 214122, China)

Abstract: Genomics and high-throughout technologies have produced large amount of relational data about the components of living systems. Such data from various complex networks that carry rich knowledge about the systems. A current challenge of system biology is to mine the hidden knowledge stored in the networks. As an important means for knowledge discovery, graph clustering attracts much attention in analysis of the complex biological networks, attaining remarkable results in remote homology detection, protein function prediction, and metabolic pathway discovery. Meanwhile, due to the high complexity of living systems, methods what successful in other fields often encounter difficulties when applied to the complex biological networks. Here we briefly reviewed the main efforts dedicated to furthering clustering analysis of complex biological networks.

Key words: complex biological networks; graph clustering; protein interaction network; metabolic network; protein similarity network

收稿日期:2008-07-11.
基金项目:国家 863 计划项目(2006AA020204).
作者简介: 梅娟(1980-),女,江苏盐城人,发酵工程博士研究生.
*通讯作者: 李炜疆(1964-),男,陕西榆林人,理学博士,教授,博士生导师.主要从事生物信息学,计算分子生物学研究. Email: wjlee01@gmail.com

细胞是由大量蛋白质、核酸及化合物构成的复杂系统，为理解整个系统的运行机制，不仅需要了解其基本元件(如蛋白质)的个体特性，还需要掌握各元件之间的相互作用关系。随着基因组学研究和高通量技术的飞速进步，大量生命系统组成元件之间的相互关系数据迅速积累，由这些关系数据构成的生物网络成为系统生物学的主要研究对象[1]。

主要的生物网络有蛋白质相互作用网络、代谢网络、蛋白质相似性网络等，这些网络因节点众多、节点间连接极不规则而被称为复杂生物网络。复杂生物网络有小世界[2](节点对之间的平均距离较短)和高度不均—[3](即节点连接度接近幂率分布，少数节点与其他节点的联系众多，而大多数节点只有很少联系)的特征。同时，复杂生物网络常常含有隐蔽的集团结构，而集团往往与生物学功能直接相关[1]。例如，代谢网络中的集团可能对应着功能独立的代谢途径[4]。因此，挖掘复杂生物网络中的集团结构对人们了解生命系统意义重大。

聚类是探索关系数据组织结构的重要手段，试图将数据对象划分为"自然的"集团，使得集团内部成员相似度(或关联度)高，而不同集团的成员之间的相似度低。这种划分方法不需要事先输入已知的分类信息，因而属于非监督学习。由于这个特点，聚类成为复杂现象中知识发现的极为重要的工具，在人工智能、图像处理、市场分析、社会学研究、生物信息学等众多领域有广泛应用[5]。

图聚类是一类非常重要的聚类方法，近年来引起了广泛关注。其原因有二：一是待研究对象间的相似度可以表示为图，一般的聚类问题常可转化为图聚类问题；二是复杂网络，特别是复杂生物网络研究兴趣的急速增长，对图聚类方法提出了新的要求。

目前，图聚类的方法在分析复杂生物网络方面取得了一些进展，如从单联聚类等局部算法到使用了全局信息的谱聚类等，都在蛋白质家族分类和功能预测等方面取得了较好的结果。本文总结几种常见的复杂生物网络及其聚类算法，探讨了其中存在的部分问题和发展趋势。

1 蛋白质相互作用网络

蛋白质分子之间的相互作用对蛋白质功能的实现至关重要。酵母双杂交技术[6]和质谱分析技术[7]等高通量技术使蛋白质相互作用数据得以迅速扩充，已经有多种数据库专门收集整理蛋白质相互作用数据，如MIPS[8]、DIP[9]。蛋白质相互作用网络(PIN，protein interaction network)的数学模型是无向图，图的节点为蛋白质，连接节点的边表示相应蛋白质间的相互作用。PIN是典型的复杂生物网络，其节点连接度高度不均匀。

在PIN中，连接紧密的子网络往往由功能相近的蛋白质构成，例如，Bu等[10]从MIPS数据库中高置信度相互作用组成的酵母PIN中找到48个准团(quasi-clique，大部分节点间都有连接的子网络)，其中功能不一致的蛋白质平均只占36%，他们通过计算邻接矩阵的本征向量寻找准团，这个方法的主要问题在于准团对连接稠密性的要求太高，大部分构成有生物学意义的集团的蛋白质之间的相互联系没有那么紧密，因而只能对少量最稠密集团做出推测。由于该方法对数据误差的敏感性[11]，如果降低准团连接稠密度的要求，则PIN中普遍存在的高误差率会严重影响聚类结果。

Bu等的直接基于邻接矩阵的谱方法并不常用，常用的是基于拉普拉斯矩阵的谱聚类算法，其物理模型可以解释为网络上的随机行走过程[12]，由于拉普拉斯矩阵通常非常稀疏，因此，谱聚类算法具有很高的效率，在模式识别等领域应用广泛[13]。但是应用于蛋白质相互作用网络分析时遇到困难，Sen等[14]用谱聚类算法对GRID数据库中相互作用组成的酵母PIN聚类分析，正确地预测了一些在更高版本GRID数据库中存在的蛋白质间的相互作用，但是无法找到统计意义显著的集团。这是因为蛋白质相互作用网络节点的度接近幂率分布，呈现高度非均一性，在对此类网络建模时，以往的方法不能奏效。

鉴于蛋白质相互作用数据库中有些功能类的蛋白质之间的连接稀疏以及拓扑形状任意，大部分的聚类方法只能找到较小的集团，丢弃一些连接度较低但带有重要信息的点，甚至产生许多孤立节点。Hwang等[15]将PIN模拟成一个信号转导系统来统计地分析任一蛋白质对这个网络拓扑结构的影响。通过计算每个节点到其他所有节点的信号转导值函数，找出最大的函数值对应的节点并将具有相等最大函数值的节点归为一类，这样就形成了一些初始的集团。在此基础上，定义两个集团的相似度函数并设定一个阈值，类似于凝聚式层次聚类的做法，将满足条件的群合并，到不能合并为止。他们选用DIP数据库中2 526个酵母蛋白间可靠的5 949个相互作用，以MIPS数据库中的描述蛋白质的致命性信息为基准，发现前555个具有较大函数值的蛋白质中有233个是致命的。对这个酵母PIN聚类得到60个集团，集团的平均密度只有0.214 5，

以 MIPS 数据库为基准发现每个集团中大部分蛋白质有相同功能,丢弃的蛋白质只占 7.8%,而最大准团法[16]和上面提到的准团等算法的平均丢弃率达到 59%。这个方法解决了部分算法在分析 PIN 时遇到的由于同一集团中蛋白质间连接稀疏而引起的问题,但是不能避免数据中掺杂的假阳性和假阴性对聚类结果的影响。

Brun 等[17]提出了一种基于密度函数的方法来对蛋白质聚类分析,对于聚成的类采用"少数服从多数"的原则对未知功能进行注释。基于如果蛋白质的共同邻居在蛋白质的相互作用中有较大比重,则该两个蛋白质趋于具有相同的功能这样的假设,采用 Czekanowski-Dice 距离定义两个蛋白质在网络中的相似度。事先他们还对数据进行了预处理,去掉了连接度小于 3 的点,这就使预测的覆盖度下降了,但同时也使对未知蛋白质的预测准确度大大提高。为每个节点定义一个密度值,鉴于具有高密度值的节点聚成一类的可能性较大的思想,他们定义了核,即联通子网络,要求子网络中每个节点的密度局部最大并且大于平均密度值,核的数量就是要聚成的类的个数,最后将剩余的节点归类。作为算法的检验,作者将此算法与他们以前提出的 PRODISTIN 方法[18]进行了比较,采用同样的库,此算法聚成的 33%的类等于或包含于 PRODISTIN 方法所得到的类,55%的类有至少 70%的重叠部分,另外分析发现此方法找出的集团更具生物意义,并且成功地预测了 37 个酵母蛋白质的功能。

2　代谢网络

代谢网络把细胞内所有生化反应表示为一个网络,反映了所有参与代谢过程的反应物之间以及所有催化酶之间的相互作用。代谢网络可以用化合物－反应二部图来完整地描述,图中有两类节点,分别是化合物和反应。在此图的基础上可以导出 3 种类型的图:化合物导出图,酶导出图,反应导出图。在网络分析时常使用只包含化合物的简化图来表示代谢网络。如图 1 所示,(a)中的(1) (2) (3) 式是 3 个生化反应;(b)是反映了所有信息的化合物－反应二部图;(c)是化合物导出图,反映了化合物之间的关系。许多生化反应是不可逆的,所以代谢网络通常用有向图来表示。在研究代谢网络的某些特性时,可以忽略反应的方向将网络表示为无向图。

Jeong 等[19]率先提出了利用复杂网络理论研究代谢网络的拓扑特性,系统地研究了 43 种生物体的代谢网络,发现这些代谢网络有相同的组织结构和相同的拓扑特性,如高度不均一和小世界等特性。完整的代谢网络包含的数据量巨大,要挖掘其中蕴含的功能信息,对于网络节点聚类或将网络模块化都是重要的手段。

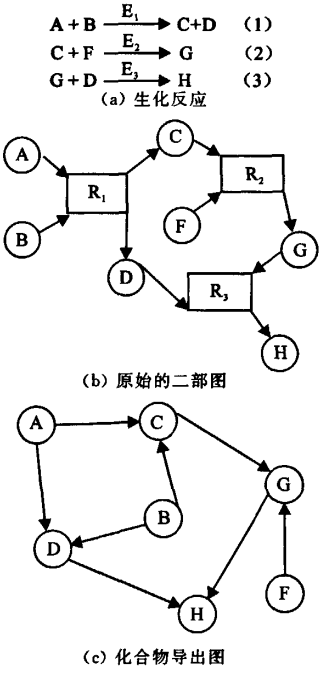

图 1　代谢网络的图表示

Fig. 1　Graph description of metabolic network

基于从网络的拓扑结构出发,将代谢网络划分成结构独立的模块,模块内部节点连接紧密,模块之间连接稀疏,Guimera 等[20]用模拟退火算法来优化这个问题。根据 Newman 等[21]提出的模块性定义,将模块性函数的负值作为模拟退火算法的目标函数,通过不断地迭代寻找最大的模块性函数值对应的模块化结果。算法不需要事先知道模块的个数,而是将其作为一个输出。他们重构了 12 种细菌和真核生物的代谢网络,以 KEGG 数据库中代谢途径的信息为基准,化合物导出图为对象评价了此算法。每个代谢网络平均被划分成 15 个模块,其中最多的 19 个,最少的 11 个,模块内的连接密度是模块间连接密度的 100 到 1 000 倍。通过模块度和参与系数这两个指标,分析了节点在网络中所处的地位。与 KEGG 数据库中的注释比较,发现大部分处于相同模块中的节点属于相同代谢途径,印证了代谢网络中功能模块与拓扑集团结构的对应关系。这个方法能够得到比较精确的结果,但由于模拟退火算法本身有接受次于当前解的机制,所以需要较大的计算量。

同样从分解代谢网络的观点出发，Schuster等[4]提出了基于代谢物的局部连接性的分解算法，能够自动地将代谢网络分解成一些子系统。他们在算法中设定一个阈值，将连接度大于阈值的节点从网络中移除，这样整个网络就分解成许多不连通子网络，这些子网络对应着聚类结果。作者用此算法分解了肺炎支原体(Mycoplasma pneumoniae)的全基因组代谢网络，根据KEGG数据库中的注释，将其分解成精氨酸降解、四茎叶酸系统、核酸代谢等19个子网络，证明了此方法能够得到一些具有相对独立生物功能的模块。这个算法存在的主要问题是阈值难以选取，由于代谢网络有高度不均一的特性，较小的阈值会产生过小的统计意义不显著的集团，甚至产生许多孤立节点；较大的阈值只能将hub节点去掉，产生许多零碎的子网络，不具有明显的生物学意义。

Holme等[22]提出了用分割式的层次聚类算法将代谢网络分解成一些子网络。作者根据WIT数据库中的信息重构了43种生物体(包含了被子植物，细菌，真核细胞)的代谢网络。将代谢网络表示为化合物－反应二部图，计算其中反应节点的介度，并依从大到小的顺序以分割式的层次聚类算法思想去掉相应的节点和边，逐步将整个代谢网络分解成子网络，反映了子网络间的层次关系。同样，Ravasz等[23]也用类似方法研究发现，代谢网络中存在一些小的高度连接的模块，这些小的模块组合成一些稍大的模块，稍大的模块又组合成一些更大的模块，也就是说代谢网络中的模块是按层次化的方式组织起来的。

为了降低分析代谢网络的复杂度，Ma等[24]通过对代谢网络的巨强分支(GSC)的分析来代替对整个网络的分析。作者构建了E. coli的全基因组反应导出图，将两个反应的相异度定义为它们之间较短的有向距离，采用凝聚式的层次聚类方法分析GSC，成功地将其分解为11个有明确生物功能的模块。

3 蛋白质相似性网络

与PIN及代谢网络所表示的物理化学相互作用不同，蛋白质相似性网络所表示的则是人为定义的蛋白质间的某种相互关系。这种网络通常用无向加权图来表示，图的节点为蛋白质，连接两个节点的边的权重是相应蛋白质之间的相似程度，称之为相似度。常见的计算相似度的工具有BLAST[25]、FASTA[26]和PSI-BLAST[27]等。其他的一些度量标准，如蛋白质序列间的Czekanowski-Dice距离，蛋白质序列间氨基酸组分的某种关系也可作为蛋白质序列的相似度。相似度的定义有较大任意性，不同的方法往往导致差异较大的结果，根据具体问题选择合适的度量标准是分析这类网络的一个关键。

Yona等提出的ProtoMap[28]和Pipenbacher等提出的ProClust[29]都是基于单联聚类的思想，属于局部算法。他们根据蛋白质序列的相似度将蛋白质构造成树形结构，选择适当的阈值将整个树划分成许多子树，即对应着聚成的集团。这种方法简单易行，层次结构一目了然。但是，不同的阈值对应着不同的聚类结果，放松的阈值易将本不属于一类的蛋白质聚到一起，保守的阈值能够将相似度较大的蛋白质聚成一类，对于相似度较小而实际上属于同一类的蛋白质聚类效果不佳，即难以发现比较松散的集团结构，这是这种算法的一个不足。

基于对单联聚类算法的改进，Kawaji等[30]将蛋白质序列的聚类问题转化为带权联接图的分割问题。算法的基本思想是去除了平衡参数的FM算法[31]。他们根据单联聚类生成单联树，由相似度给定一个初始划分，在此基础上采用"one-pass improvement"策略，找出第一次切断的最小权重，这对应着当前所寻找的一个划分，通过更新单联树，不断地迭代上述步骤。作者分别用此算法和单联聚类算法来对SWISS-PROT数据库中所有小鼠蛋白质序列进行家族分类，并以InterPro数据库中的家族分类作为基准比较了聚类结果。与理论分析相一致，对于不同的阈值，单联聚类的结果都不佳，产生了与实际不符的群。而本文的算法所聚成的蛋白质家族中有接近一半的家族与InterPro数据库中家族的符合率超过90%。虽然这种算法对蛋白质家族分类有一定效果，然而如何给定一个恰当的初始分类是一个难点，单单根据两两相似度这个局部信息来确定初始划分是不准确的。

随着蛋白质序列的急剧增长，多域蛋白质以及混杂域的出现(混杂域一般具有多种不同的功能)，已有的一些算法，如GeneRAGE[32]对蛋白质数量相对较少的原核生物进行家族分类取得了一定效果，但是将它应用到蛋白质数量较大的真核生物时则聚类效果不佳。针对这些问题Enright等提出了TRIBE-MCL算法[33]，是一种基于概率和图论的全局算法。用BLAST E-values的单调函数作为相似度建立相似度矩阵，对每列归一化产生随机矩阵(马尔可夫矩阵)，对此矩阵不断交替使用expansion操作和inflation操作，直到矩阵不再变化为

止,即为幂等矩阵,对应最后的聚类结果。为了评价算法在蛋白质家族分类方面的性能,他们以 InterPro 数据库和 SCOP 数据库中的描述作为基准,来对 SwissProt 数据库中80 000个蛋白质进行家族分类分析,结果显示 78%的家族有与注释完全一致的域结构,说明此方法对多域蛋白质家族分类有较好的效果。对 SCOP 数据库中的18 248个蛋白质进行家族分类,准确率可以达到 87%。

Paccanaro 等使用的谱聚类算法[34]也是全局算法,聚类的结果决定于所有蛋白质间相互关系的共同作用。同样用 BLAST E-values 的单调函数作为序列的相似度来构建相似度矩阵,对其拉普拉斯变换后求特征值和特征向量,对前 K 个最大的特征值对应的特征向量组成的矩阵处理后用 K-means 对行聚类。作者选用 SCOP 数据库中超家族和家族的分类作为基准来评价这个算法,并将聚类结果与 GeneRAGE 和层次聚类[35]进行比较。GeneRAGE 和层次聚类在家族层次上的聚类上取得一定的结果,而在超家族层次上聚类效果不佳,无法将分散的家族组成超家族,即这些局部算法难以发现比较松散的集团结构,因为同一家族中的蛋白质相似度高于同一超家族中蛋白质的相似度,这与笔者前面的分析相吻合。谱聚类算法在家族和超家族层次上都取得了比较好的结果,能够找出一些远同源性的蛋白质序列。但是这个方法存在两方面的问题,首先采用 BLAST E-values 的单调函数作为相似度在分析远同源性蛋白质序列时是不够准确的,其次是谱聚类算法本身对数据噪声相当敏感。

4　其它复杂生物网络

对于其他一些复杂生物网络,通常也采用图聚类的分析方法。如 Yildirim 等[36]通过构建药靶网络(DT, drug-target)来挖掘药物和药物作用的靶标之间的关系。在 DT 网络的基础上,又导出了两个在生物上相关的网络,分别是药物网络(DN, drug network)和靶蛋白网络(TPN, target protein network)。在 DN 中,节点表示药物,当两种药物作用于至少同一个靶蛋白时,两者之间就有边相连;在 TPN 中,节点表示靶蛋白,当两个靶蛋白被至少同一个药物作用时,两者之间就有边相连。对这两个网络分别采用 DrugBank 提供的解剖学治疗学化学分类方法和 Gene Ontology 中提供的细胞的构成将节点进行分类。分类结果显示,新药趋向作用于已知的连接度较高的靶蛋白,这为药物的设计提供了一条途径。

5　结　语

图聚类的目标是直观、自然地隔离松散联系的密集子图。一般来说,好的聚类方法应使类内相似度尽可能高而类间相似度尽可能低。事实上,不存在独立于聚类最终目的的绝对最佳标准,往往通过与实际问题的含义来评判。现有的聚类算法在不同方面取得了一些进展,但仍有许多重要问题尚未解决,其中包括:

1) 蛋白质相互作用网络、代谢网络等典型复杂生物网络有高度不均一的特征,使得通常行之有效的图聚类算法遇到了较大困难,无法将整个网络划分为统计意义显著的子网络,因此针对网络的这种特征,必须设计新的算法。

2) 由于实验设备和技术的限制,数据中普遍存在着误差,如蛋白质相互作用网络中存在着假阳性和假阴性问题,严重影响了聚类的效果。

3) 复杂生物网络的节点数目众多,节点间关系错综复杂,网络聚类计算量通常很大,许多算法的严格求解是 NP-hard 难题。因此,实际应用中算法的执行效率也是需要特别关注的问题。

参考文献(References):

[1]Barabasi A L, Oltvai Z N. Network biology: understanding the cell's functional organization[J]. **Nat Rev Genet**, 2004(5): 101－113.

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

[3]Albert R. Scale－free networks in cell biology[J]. **J Cell Sci**, 2005, 118: 4947－4957.

[4]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: 351－361.

[5]Berkhin P. A survey of clustering data mining techniques[J]. **Grouping Multidimensional Data**, 2006: 25－71.

[6]Ito T, Chiba T, Ozawa R, et al. A comprehensive two-hybrid analysis to explore the yeast protein interactome[J]. **Proc Natl Acad Sci U S A**, 2001, 98: 4569－4574.

[7]Ho Y, Gruhler A, Heilbut A, et al. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry[J]. **Nature**, 2002, 415: 180－183.

[8]Pagel P, Kovac S, Oesterheld M, et al. The MIPS mammalian protein-protein interaction database[J]. **Bioinformatics**,

2005, 21: 832—834.

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

[10]Bu D, Zhao Y, Cai L, et al. Topological structure analysis of the protein—protein interaction network in budding yeast [J]. **Nucleic Acids Res**, 2003, 31: 2443—2450.

[11]Chang H, Yeung D Y. Robust path based spectral clustering[J]. **Pattern Recognition**, 2008, 41: 191—203.

[12]Meila M, Shi J. A random walks view of spectral segmentation[C]. Florida: Proceedings of the International Workshop on AI and Statistics(AISTATS). 2001.

[13]Filippone M, Camastra F, Masulli F, et al. A survey of kernel and spectral methods for clustering[J]. **Pattern Recognition**, 2008, 41: 176—190.

[14]Sen T Z, Kloczkowski A, Jernigan R L. Functional clustering of yeast proteins from the protein—protein interaction network[J]. **BMC Bioinformatics**, 2006(7): 355.

[15]Hwang W, Cho Y R, Zhang A, et al. A novel functional module detection algorithm for protein—protein interaction networks[J]. **Algorithms Mol Biol**, 2006(1): 24.

[16]Spirin V, Mirny L A. Protein complexes and functional modules in molecular networks[J]. **Proc Natl Acad Sci U S A**, 2003, 100: 12123—12128.

[17]Brun C, Herrmann C, Guenoche A. Clustering proteins from interaction networks for the prediction of cellular functions [J]. **BMC Bioinformatics**, 2004(5): 95.

[18]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 Biol**, 2003(5):6.

[19]Jeong H, Tombor B, Albert R, et al. The large-scale organization of metabolic networks[J]. **Nature**, 2000, 407: 651—654.

[20]Guimera R, Nunes Amaral L A. Functional cartography of complex metabolic networks[J]. **Nature**, 2005, 433: 895—900.

[21]Newman M E, Girvan M. Finding and evaluating community structure in networks[J]. **Phys Rev E Stat Nonlin Soft Matter Phys**, 2004, 69: 26—113.

[22]Holme P, Huss M, Jeong H. Subnetwork hierarchies of biochemical pathways[J]. **Bioinformatics**, 2003, 19: 532—538.

[23]Ravasz E, Somera A L, Mongru D A, et al. Hierarchical organization of modularity in metabolic networks[J]. **Science**, 2002, 297: 1551—1555.

[24]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: 1870—1876.

[25]Altschul S F, Gish W, Miller W, et al. Basic local alignment search tool[J]. **J Mol Biol**, 1990, 215: 403—410.

[26]Pearson W R, Lipman D J. Improved tools for biological sequence comparison[J]. **Proc Natl Acad Sci U S A**, 1988, 85: 2444—2448.

[27]Altschul S F, Madden T L, Schaffer A A, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs[J]. **Nucleic Acids Res**, 1997, 25: 3389—3402.

[28]Yona G, Linial N, Linial M. ProtoMap: automatic classification of protein sequences and hierarchy of protein families[J]. **Nucleic Acids Res**, 2000, 28: 49—55.

[29]Pipenbacher P, Schliep A, Schneckener S, et al. ProClust: improved clustering of protein sequences with an extended graph—based approach[J]. **Bioinformatics**, 2002, 18(2): 182—191.

[30]Kawaji H, Yamaguchi Y, Matsuda H, et al. A graph-based clustering method for a large set of sequences using a graph partitioning algorithm[J]. **Genome Inform**, 2001(12): 93—102.

[31]Fiduccia C M, Mattheyses R M. A Linear-time heuristic for improving network partitions[C]. New York: 19th Design Automation Conf, 1982.

[32]Enright A J, Ouzounis C A. GeneRAGE: a robust algorithm for sequence clustering and domain detection[J]. **Bioinformatics**, 2000, 16: 451—457.

[33]Enright A J, Van Dongen S, Ouzounis Ç A. An efficient algorithm for large—scale detection of protein families[J]. **Nucleic Acids Res**, 2002, 30: 1575—1584.

[34]Paccanaro A, Casbon J A, Saqi M A. Spectral clustering of protein sequences[J]. **Nucleic Acids Res**, 2006, 34: 1571—1580.

[35]Corpet F. Multiple sequence alignment with hierarchical clustering[J]. **Nucleic Acids Res**, 1988, 16: 10881—10890.

[36]Yildirim M A, Goh K I, Cusick M E, et al. Drug-target network[J]. **Nat Biotechnol**, 2007, 25: 1119—1126.

(责任编辑:秦和平)