一种基于代谢路径构建系统发生树的有效方法

赵建邦,高 琳,宋 佳
(西安电子科技大学计算机学院,陕西西安 710071)

摘 要: 构建系统发生树是研究物种起源和演化的重要手段.本文基于KEGG(Kyoto Encyclopedia of Genes and Genomes)代谢路径,引入图论的"核"概念,提出一种构建系统发生树的方法.首先解决在无数据丢失前提下,代谢路径数据的提取和表示问题,其次将不同代谢路径的相似度定义为图的核部分与非核部分各自匹配程度的加权之和,利用距离矩阵构建物种间的系统发生树.通过大量试验数据和NCBI(National Center for Biotechnology Information)分类法进行比较,验证了本文方法的有效性.

关键词: 系统发生树;代谢路径;核方法;路径比对
中图分类号: Q81, TP312 **文献标识码:** A **文章编号:** 0372-2112 (2009) 08-1633-06

An Efficient Method for Constructing Phylogenetic Trees Based on Metabolic Pathway

ZHAO Jian-bang, GAO Lin, SONG Jia
(*School of Computer Science and Technology, Xidian University, Xi'an, Shaanxi* 710071, *China*)

Abstract: Constructing the phylogenetic tree of life is an important resort of learning the origin and the evolution among species. By introducing the concept of "kernel", a method to achieve phylogenetic tree based on KEGG metabolic pathway is presented. We firstly solved the problem of pathway abstraction with no metabolic information lost, and secondly we defined the similarity between different metabolic pathways as the summation of weighted matching score of the kernel subgraph and the non-kernel one respectively. Based on the distance matrix obtained by the two steps above, we construct the phylogenetic tree of several species. The experiments show that it is an efficient method according to the comparison between the trees obtained and NCBI taxonomy.

Key words: phylogenetic tree; metabolic pathway; kernel-based method; pathway alignment

1 引言

系统发生(Phylogeny)是指一群有机体发生或进化的历史.系统发生树(Phylogenetic tree,又称Evolutionary tree进化树)是描述这一群有机体发生或进化顺序的拓扑结构,它可以用来研究不同物种间的进化关系,从达尔文时代起便是生物学的研究热点[1].

从上世纪70年代开始,分子测序技术的发展使得蛋白质、DNA数据大量出现,从而使系统发生学的研究进入了分子水平.许多基于序列的比对算法被应用在研究生命进化关系中,例如动态规划、HMM、语言学方法等.不同学者对于序列间距离的定义不尽相同,因而所建造的进化树也存在差别.目前国际上公认的系统发生树建树分子标准为小亚基核糖体核糖核酸(Small Subunit rRNA,SSUrRNA),因为这些碱基序列广泛存在于各个物种的基因中并保持着高度保守(Conserved)状态.基于序列进行系统发生树的构建存在的缺陷是:由于大多数生物的全基因组数据并没有得到测序,因此在所得构建系统发生树的规模和正确性上存在不足.

后基因组时代,生命科学从分子生物学时代开始进入系统生物学(Systems Biology)时代.生命体系实际上是一种由不同的生物化学反应通路模块组成的分子网络系统.生物分子网络作为一种描述生物分子间相互作用关系的方法,在揭示生物体的生长、发育、衰老和疾病等生命系统的基本分子过程和规律中受到越来越多的重视.运用计算生物学的工具,开展生物分子网络分析的理论研究已经成为计算生物学的一个非常重要的研究方向.蛋白质相互作用网络和代谢路径数据的出现为基于生物体内部的功能模块比对来研究系统发生学提供了条件.目前,已经存在一些利用系统发生分析来处理代谢路径信息的方法.Tohsato[2]等人提出一种基于动态规划来计算代谢路径相似性的算法,但其缺陷是只能计

收稿日期:2008-08-10;修回日期:2008-11-03
基金项目:国家自然基金(No.60574039);博士科学点基金(No.200807010013)陕西省自然科学研究计划项目(No.SJ08—ZT150);教育部留学回国人员基金

算没有分支的代谢路径，不能应用在有分支或者更复杂的代谢路径比对中. Liao[3]等人提出将代谢路径中的所有子路径定义为一个集合，特定物种(organism)便可以表示为一个关于代谢路径的二进制向量 Profile，物种间的相似性根据该向量给出，不过实验结果和根据 16s rRNA 所建立的进化树存在一定差距. Heymans[4]等人将 KEGG 代谢路径重组为不同的酶之间的相互作用图，把代谢路径相似性问题简化为酶图的相似性比对问题. 以 Heymans 的思路为基础，S June Oh[5]等人利用图的核"kernel"的概念，提出改进算法来计算代谢路径间的相似性，该算法局限性在于没有突出"非核"子图部分在计算代谢路径相似性中的作用. 本文首先对 KEGG 代谢路径的提取方法进行了改进，使实验数据更加精确，接着基于"核"概念提出一种新的计算子图间相似性的方法，实验结果证明，通过本方法构建的系统发生树更接近于 NCBI 分类法.

本文的组织结构如下：首先是数据的预处理，在确保原始数据信息 完整的前提下，对 KEGG 代谢路径进行提取和重组；第三节提出了一种基于核结构计算代谢路径之间相似性的方法，基于该相似性构建系统发生树；第四节是仿真试验，将本文提出的算法和现有的算法进行性能上的比较，并与 NCBI 分类标准的进化树进行比较；最后对本文进行总结，分析了系统发生树构建中所面临的问题和挑战.

2 数据预处理

基于 KEGG[6]代谢路径构建系统发生树，必须对代谢路径数据进行预处理：一方面提取构建系统发生树所需要的数据，另一方面将提取的数据通过新的数据结构来表示，使后续的建树工作更加方便.

本节首先对 KEGG 代谢路径进行简单介绍，并给出提取酶之间相互作用关系子图的算法.

2.1 KEGG 代谢路径简介

KEGG(Kyoto Encyclopedia of Genes and Genomes)生物数据库是由日本京都大学所提供的系统分析基因功能、联系基因组信息及功能信息的知识库. 主要提供四种生物数据支持，其中 KEGG PATHWAY 完成了代谢路径在生物分子级别上的可视化工作，将分子间的反应关系和相互作用关系用图形的形式表现出来.

KEGG PATHWAY 数据库根据代谢物的类型将其进行分类，如代谢类型主要包括碳水化合物代谢、类脂代谢、能量代谢等. 每种代谢类型包括了若干不同代谢物的新陈代谢信息，如碳水化合物代谢包括 17 种不同物质的网络，有柠檬酸盐循环(TCA cycle，以下简称 TCA)、糖酵解/糖质新生(Glycolysis/Gluconeogenesis，以下简称 G/G)等. KEGG 对每种代谢物的代谢信息实现了图解表示，即用一个图来表示该代谢路径中酶和化合物的反

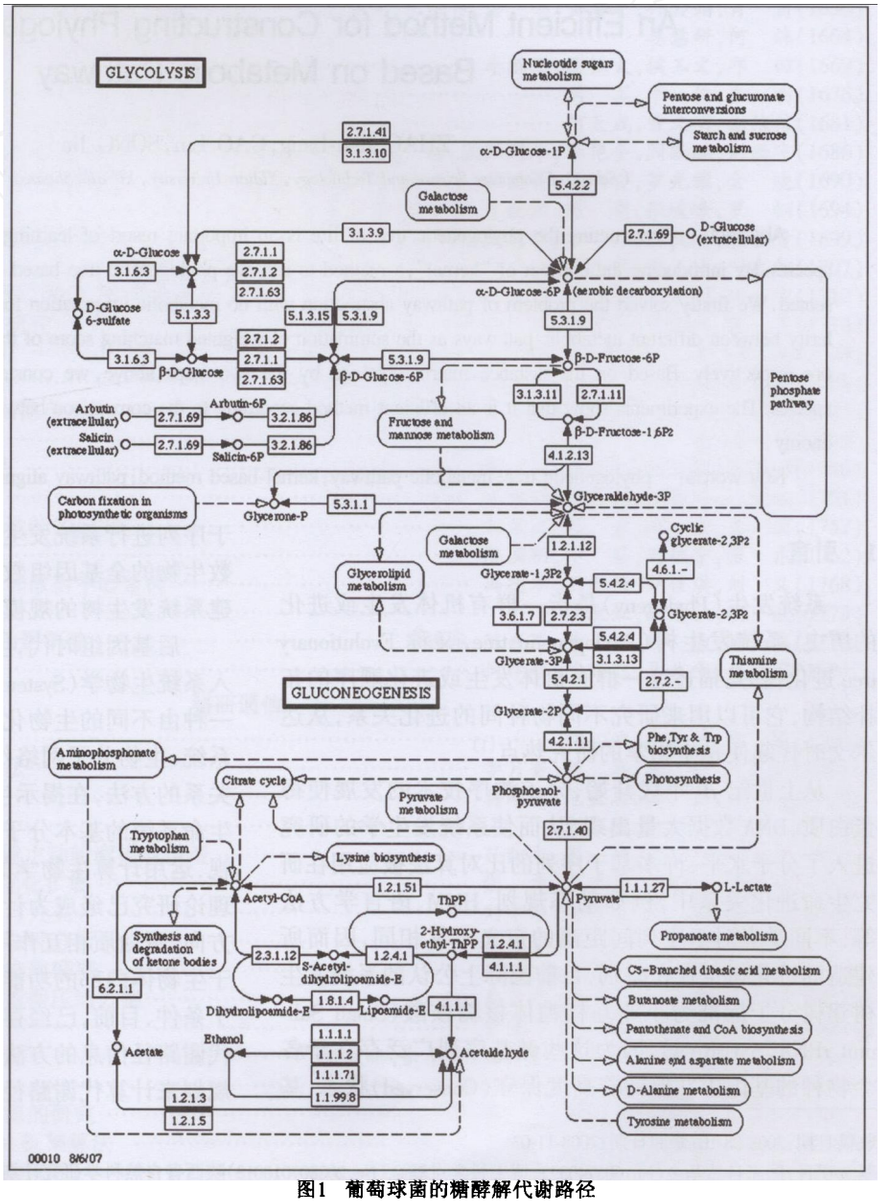


图1 葡萄球菌的糖酵解代谢路径

应关系,这里称该图为合成图.每个合成图都有一个详细的文本文件来对其进行解释,称其为解释文本(explication file).如 TCA 的解释文本链接为"http://www.genome.jp/dbget-bin/show_pathway? map00020".对于每种代谢物,在不同生物体内的代谢路径是不同的,主要是由于参与代谢的酶个数和种类不尽相同,因此所产生的化合物也不尽相同.如果酶 A 的产物(product)是酶 B 的底物(substrate),则 A 和 B 之间存在一条有向边,通过提取,特定代谢物质在某一物种内的代谢路径形成合成图的子图,如图 1 所示,为葡萄状球菌(缩写为 sau)的糖酵解代谢路径,显示为绿色的部分表示在该代谢过程中参与反应的酶,所有的酶(包括绿色和灰色)均属于合成图.

2.2 代谢路径提取方法

要对代谢路径进行分析,首先将各物种对应的子图从合成图中提取出来.现有的几种对代谢路径的提取核转换方法都是以酶的分类号(Enzyme Classification number)作为节点,但是这种提取方法会丢失有效数据,原因是:(1)有的酶在同一个代谢路径中会催化不同的反应,即子图中 EC Number 相同的酶会有不同的 ID,如图 2 中的酶"1.2.4.1"和"4.1.1.1",若将此类酶作为一个节点处理,会导致错误的反应关系;(2)合成图中由相同的酶所催化的反应不一定在子图中同时有效,如图 1 中的"2.7.1.69"和"3.2.1.86",若将此类酶作为一个节点处理,会增加本不存在的反应关系.这两种情况会影响原始数据的真实性.本文提出用代谢路径中酶的序号(ID)来表示唯一节点的观点,和现有的几种对代谢路径的提取和转换方法[4,5]相比,这种方法可以避免当子图中存在相同的酶时,提取过程对代谢路径所产生的数据损失.

表 1　本文所用到的符号及其解释

符号	解释
M	特定代谢物的代谢路径合成图
Z	M 中的酶集合
RA	M 中的反应集合
RL	M 中酶的相互关系集合
s	特定物种所对应的代谢路径
$Z(M,s)$	物种 s 在代谢路径 M 中的酶集合
rl	RL 中的元素变量
ra	RA 中的元素变量
e	酶-酶相互作用边
$Z(e)$	在 e 中出现的酶集合
$Z(rl)$	在 rl 中出现的酶集合
$T(ra)$	反应 ra 的产物集合
$S(ra)$	反应 ra 的底物集合
$Gs(M,s)$	物种 s 在代谢路径 M 中所对应的子图

表 1 列出了子图提取算法所涉及的一些符号,每个子图 $G_s(M,s)$ 由若干条有向边组成,每条边的格式为〈酶 1,酶 2,方向〉,其中"方向"的取值为 0 或 1,分别表示从前一个酶到后一个酶之间有一条双向边或单向边.

下面给出提取子图的算法:

INPUT: the explication file F and all tshe enzymes in a specific species $Z(M,s)$;

OUTPUT: Gs, the subgrpah of metabolic pathway in the specific species.

01 Read the explication file and store the Entry set, Reaction set, Relation set respectively;
02 For each Relation rl do:
03 Find Reaction ra_1: $\{ra_1 \mid RA(ra_1) = RA(rl.id1)\ \&\&\ T(ra_1) = M(rl)\}$;
04 Find Reaction ra_2: $\{ra_2 \mid RA(ra_2) = RA(rl.id2)\ \&\&\ S(ra_2) = M(rl)\}$;
05 Calculate the edge type;
06 Add the edge (rl.id1 rl.id2 type) to G;
07 End for;
08 Register all the enzymes in a specific species;
09 For each edge e in G do:
10 Add edge e to Gs if $Z(e) \subseteq Z(M,s)$;
11 End for;
12 Return G_s.

3 系统发生树的构建

基于代谢路径构建系统发生树分为三个步骤:

(1)对于某一特定的代谢路径,提取一组不同物种的酶-酶关系图,其中每一个酶图对应一个特定的物种.这一步骤已经在第一部分实现.

(2)对这一组图进行两两比对,得到这些物种的相似度矩阵,这一步骤的关键是如何给出两个图之间的相似性定义,也是本文的主要贡献所在.

(3)基于相似性矩阵,构建系统发生树,和进化树标准 NCBI 进行比较来检验所生成的进化树的可靠性.

本节首先将代谢路径的比对问题进行抽象,将其

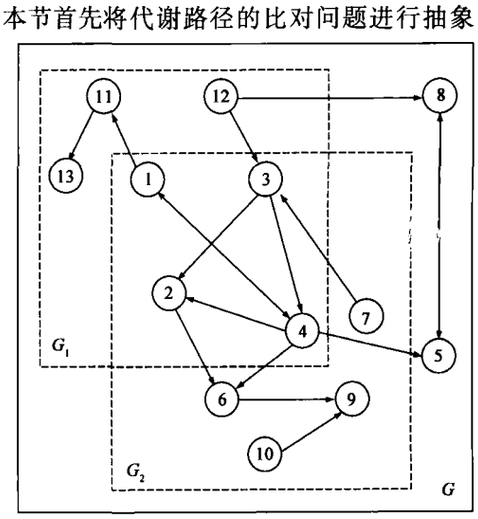


图2　代谢路径的抽象提取及其子网络

转化为图之间的相似性问题,然后提出基于"核"计算子图间的相似性算法,子图的相似性由"核"部分和非"核"部分相似性加权计算求得.

3.1 代谢路径的比较

代谢路径的相似性度量是构建系统发生树的关键.对于同一个代谢物在不同的两个物种内的代谢路径,提取出来的两个子图在比较时要同时考虑节点的相似性和连接关系之间的相似性.将两条路径的相似度定量表示为[0,1]之间的实数,相似度越接近1,说明二者越相似.

图2是酶-酶关系图 $G=(V,f,E)$ 的一种抽象表示,其中 V 表示酶的集合,f 表示酶及其标记的对应关系,E 表示边的集合.每个节点对应一个酶,每一条有向边对应不同酶之间的关系.设位于蓝色虚线框和红色虚线框的部分分别对应该代谢路径在两种不同的物种内的代谢子图 G_1 和 G_2,则在酶-酶关系图中计算其不同子图的相似性问题便转化为计算有标记的有向子图的相似性问题.

对于其中任何两个特定物种的代谢图来说,可能有一部分节点是二者共有的,那么由这些公共节点所诱导的子图也必定同构.将这二者共有的子图称为"核"结构,如 G_1 和 G_2 所共有的节点 N_1、N_2、N_3 和 N_4 所诱导的子图 $G_c(V_c,f_c,E_c)$,称为 G_1 和 G_2 的"核".一般来说,"核"结构在子图中所占规模越大,则两个子图就越相似;对于其它不属于"核"的节点,应用 Heymans M 所提出的方法[4],得到非"核"子图的相似性,整体的相似性应该综合考虑上述两部分的结果,通过参数调节来得到比较理想的相似性矩阵.

设对 n 个物种的集合 $O=\{O_1,\cdots,O_n\}$,在特定代谢路径 $G=(V,f,E)$ 中所提取的子图集合为 $G_S=\{G_1,\cdots,G_n\}$,则任意两个子图 $G_i(V_i,f_i,E_i)$ 和 $G_j(V_j,f_j,E_j)$ 的相似性由以下两部分组成:

(1)"核"子图部分:

$$Sim_1(i,j)=\frac{|V_c|^2}{|V_i|\cdot|V_j|}\cdot\frac{|E_c|}{\max\{|E_i|,|E_j|\}} \tag{1}$$

(2)非"核"子图部分:

对于非"核"部分,如图3中的 G_1 和 G_2,非"核"子图分别为由 N_{11}、N_{12} 和 N_{13} 所诱导的子图 G'_1,和由 N_6、N_7、N_9 和 N_{10} 所诱导的子图 G'_2,必须对二者进行节点的二部匹配,这里定义如下两个规模为 $|G'_1|\times|G'_2|$ 的矩阵:节点相似度矩阵 S 和节点匹配矩阵 M.

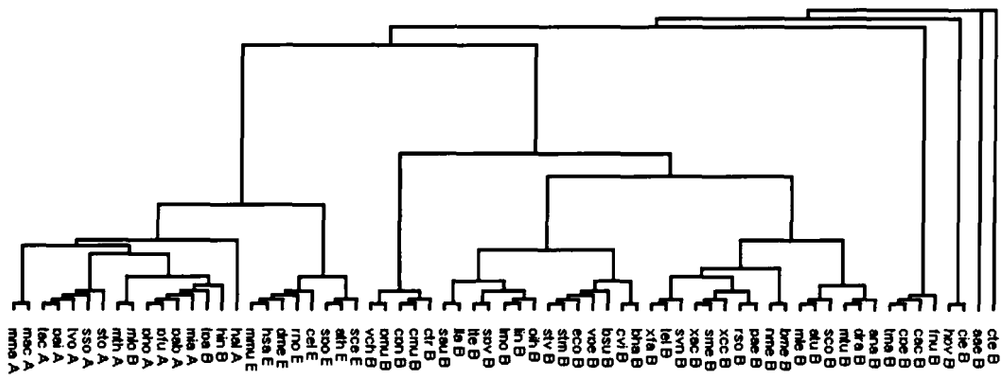


图3 通过67个物种的糖酵解代谢路径所构建的系统发生树

根据 G'_1 和 G'_2 的每个节点对的酶分类号和节点间的连接关系求出 $S(a,b)$[4],其中 $a\in G'_1$,$b\in G'_2$.接着利用匈牙利算法(Hungarian Algorithm[4])对其进行二部匹配,得到:

$$M(a,b)=\begin{cases}1 & \text{if}(match(a,b))\\ 0 & otherwise\end{cases} \tag{2}$$

则 G_1 和 G_2 的非"核"子图部分的相似性计算方法为:

$$Sim_2(1,2)=\frac{\sum_{a\in G'_1,b\in G'_2,M(a,b)=1}S(a,b)}{\sqrt{|G'_1|\cdot|G'_2|}} \tag{3}$$

更一般地,$G_i(V_i,f_i,E_i)$ 和 $G_j(V_j,f_j,E_j)$ 的非"核"子图部分的相似性计算方法为:

$$Sim_2(i,j)=\frac{\sum_{a\in G'_i,b\in G'_j,M(a,b)=1}S(a,b)}{\sqrt{|G'_i|\cdot|G'_j|}} \tag{4}$$

综上所述,$G_i(V_i,f_i,E_i)$ 和 $G_j(V_j,f_j,E_j)$ 的相似性定义为:

$$Sim(i,j)=Sim_1(i,j)+\eta\cdot Sim_2(i,j) \tag{5}$$

其中 η 为一个可以调节的参数,用来平衡核部分和非核部分在子图相似性计算上的重要性.为了确保 $Sim(i,j)$ 的值在范围[0,1]内,在本文中 η 取值为 $\frac{(|V_i|-|V_c|)\cdot(|V_j|-|V_c|)}{|V_i|\cdot|V_j|}$,并且分配子图的核部分边和非核部分边的重要性分别为 $\frac{|V_c|^2}{|V_i|\cdot|V_j|}$ 和 $\frac{(|V_i|-|V_c|)\cdot(|V_j|-|V_c|)}{|V_i|\cdot|V_j|}$.

3.2 系统发生树的构建

对于 N 个物种的集合 $O=\{O_1,\cdots,O_N\}$,通过计算

物种间代谢路径的相似性,得到一个 $N \times N$ 的相似度矩阵 *Sim*, *Sim* 为一个对角线元素为 1 的对称矩阵,每个元素的值在[0,1]区间内.利用如下公式得到该 N 个物种的距离矩阵:

$$Dist(i,j) = 1 - Sim(i,j) \quad (6)$$

在本文中我们使用 NJ(Neighbor Join)[7]算法对距离矩阵进行分析,生成两棵系统发生树,并用 TreeView[8] 将其表示出来.其中利用 67 个物种的 G/G 代谢路径所建立的系统发生树,(其中 Eubacteria(B) 46 种, Archaea (A) 13 种, Eukaryota(E) 8 种.)如图 3 所示;以及利用 71 个物种的 TCA 代谢路径所构建的系统发生树,(其中 Eubacteria(B) 49 种, Archaea(A) 14 种, Eukaryota(E) 8 种.)如图 4 所示.

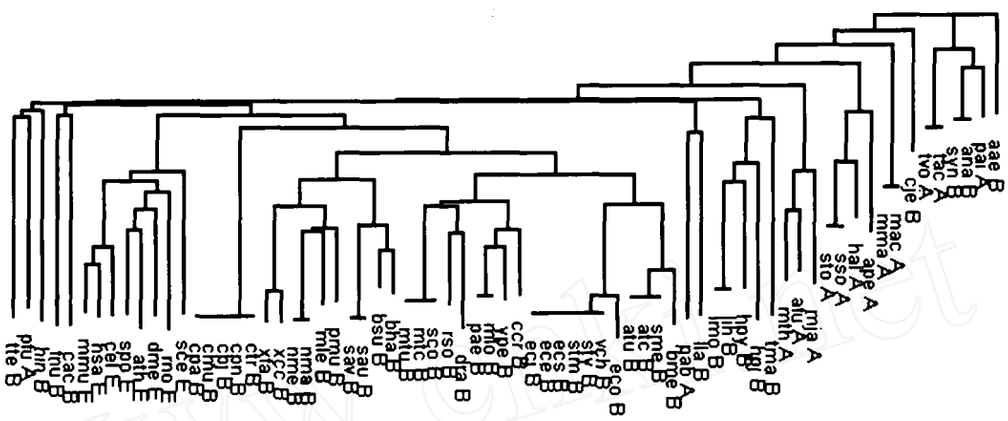
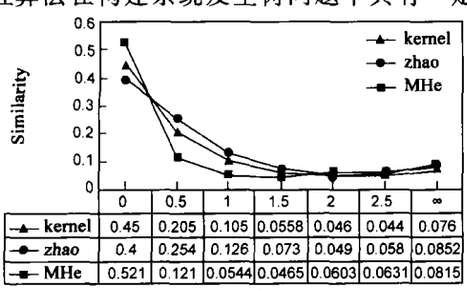


图4　通过71个物种的TCA代谢路径所构建的系统发生树

4　实验结果与讨论

对于构建的不同规模的系统发生树,我们对其正确性进行了评估.主要参考对象是 NCBI 分类(NCBI taxonomy)数据库[9].利用 Dennis Shasha 等人提出的树型数据间 cousin pair 概念[10],通过改变 cousin pair 的最大距离(max_distance of cousin),这里我们取最大距离的离散值范围为{0,0.5,1,1.5,2,2.5,∞},其中 ∞ 表示"无最大值限制".由此得到若干组本文所建树与 NCBI 标准树的相似度曲线.

对目前存在的两种不同的通过比对 KEGG 代谢路径来建立系统发生树的算法[4,5](分别代称为"MHe"、"kernel"),我们将其与本文算法做了比对,实验结果如图 5、6、7 所示.图中横轴表示最大 cousin 距离,纵轴表示不同规模的系统发生树与 NCBI 标准树的相似性.通过比较可以看出,本文所提出的基于"核提取"的子图相似性算法在构建系统发生树问题中具有一定实用性.

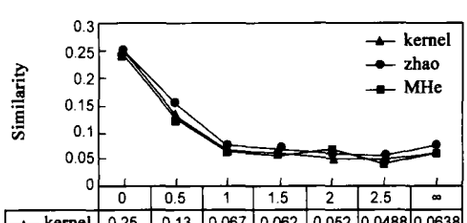


	0	0.5	1	1.5	2	2.5	∞
kernel	0.45	0.205	0.105	0.0558	0.046	0.044	0.076
zhao	0.4	0.254	0.126	0.073	0.049	0.058	0.0852
MHe	0.521	0.121	0.0544	0.0465	0.0603	0.0631	0.0815

图5　利用64个物种的TCA代谢路径所构建的三棵系统发生树性能比较

在实验中,我们发现如下规律:

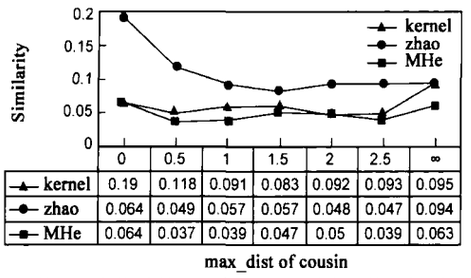
(1)物种的选择会对所构建的系统发生树的相似性产生影响.这是因为通过 NCBI 所得到的标准树不是严格的二叉树,会有多个孩子节点共享一个父亲节点的情况,而通过实验所得到的进化树为严格的二叉树.因此,如果标准树中非二叉子结构频繁出现,则本算法会丧失相对于其它参考算法的优势.如图 5 所示,cousin 的最大距离为 0 时,参考算法的结果表现较优,以及图 6 中所示各算法的有效性非常接近的情况.



	0	0.5	1	1.5	2	2.5	∞
kernel	0.25	0.13	0.067	0.062	0.052	0.0488	0.0638
zhao	0.25	0.1555	0.0772	0.0675	0.0612	0.058	0.0752
MHe	0.241	0.123	0.063	0.058	0.065	0.042	0.061

图6　利用64个物种的糖酵解代谢路径所构建的三棵系统发生树性能比较

(2)物种个数的多少会对所构建的系统发生树的相似性产生影响.一般情况下,物种越少,和对应的标准树的相似性越高.物种的个数较多时,本算法的有效性非常接近已有算法,甚至会低于二者,如图 5 中最大 cousin 距离取 0 或者 2.0 的情况.当物种数较少(小于 50)时,本算法有效性较高,如图 7.



	0	0.5	1	1.5	2	2.5	∞
kernel	0.19	0.118	0.091	0.083	0.092	0.093	0.095
zhao	0.064	0.049	0.057	0.057	0.048	0.047	0.094
MHe	0.064	0.037	0.039	0.047	0.05	0.039	0.063

图7 利用48个物种的TCA代谢路径所构建的三棵系统发生树性能比较

5 总结与展望

现有的构建系统发生树的方法主要基于 DNA 序列和蛋白质序列数据,本文基于 KEGG 代谢路径构建系统发生树,首先解决在无数据丢失前提下,代谢路径数据的提取和表示问题,将物种间相似性的计算问题转化为图的相似性的计算.基于'核'概念提出一种新的计算代谢路径间相似性的方法,将两个子图的相似性分为"核"和非"核"二部分,两部分通过加权求和来得到子图的相似性.最后构建物种间的系统发生树,通过大量试验数据和 NCBI 分类法进行比较,验证了本文方法的有效性.本文将来的研究重点在于如何给出更合理的相似性定义.希望可以借助于 KEGG 代谢路径中模块(module)的提取来定义物种间的相似性.

基于代谢路径构建系统发生树还有大量具有挑战的研究工作,首先是需要改善树型结构比对算法,因为通过现有方法构建的系统发生树是严格的二叉树,而通过 NCBI 分类法所构建的树是普通树,二者本身就存在较大差异,而且现有的基于 cousin pair 来判断树的相似性只是关注二叉树的局部分支,并没有在全局层次上关注子树的相似性;其次,由实验所得到的系统发生树需要考虑非叶子节点的意义,使系统发生树的构建更多地为物种的演化提供理论依据.

参考文献:

[1] 李建伏,郭茂祖.系统发生树构建技术综述[J].电子学报,2006,34(11):2047 - 2052.
Li Jian-fu, Guo Mao-zu. A review of phylogenetic tree reconstruction technology[J]. Acta Electronica Sinica, 2006, 34(11):2047 - 2052. (in Chinese)

[2] Tohsato Y, Matsuda H, Hashimoto A. A multiple alignment algorithm for metabolic pathway analysis using enzyme hierarchy [A]. Proc Int Conf Intell Syst Mol Biol. 2000[C]. Menlo Park, CA: AAAI Press, 2000. 376 - 383.

[3] Liao L, Kim S, Tomb J F. Genome comparisons based on profiles of metabolic pathways[A]. In Sixth International Conference on Knowledge-Based Intelligent Information and Engineering Systems (KES 2002)[C]. Crema, Italy, 2002. 469 - 472.

[4] Heymans M, Singh A K: Deriving phylogenetic trees from the similarity analysis of metabolic pathways. [J] Bioinformatics, 2003, 19(1): 138 - 146.

[5] Oh S J, Joung J G, Chang J H, Zhang B T. Construction of phylogenetic trees by kernel-based comparative analysis of metabolic networks[J]. BMC Bioinformatics, 2006, 7: 284 - 295.

[6] Ogata H, et al. KEGG: Kyoto encyclopedia of genes and genomes[J]. Nucleic Acids Res. 1999, 27(1): 27 - 30.

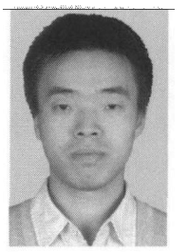
[7] Saitou N, Nei M. The neighbor-joining method: A new method for reconstructing phylogenetic trees[J]. Molecular Biology and Evolution, 1987, 4(4): 406 - 425, .

[8] Page RDM: TREEVIEW: An application to display phylogenetic trees on personal computers[J]. CABIOS 1996, 12(4): 357 - 358

[9] NCBI taxonomy[DB/OL]. http://www.ncbi.nlm.nih.gov/Taxonomy/, 2008.

[10] Shasha D, et al. Unordered Tree Mining with Applications to Phylogeny[A]. 20th International Conference on Data Engineering 2004[C]. Washington, DC, USA: IEEE Computer Society, 2004. 708 - 719.

作者简介:



赵建邦 男,1983 年生于陕西渭南,现在西安电子科技大学计算机学院攻读计算机应用技术专业博士学位,研究方向为生物信息数据挖掘.
E-mail: zjb9797@foxmail.com

高 琳 女,1964 年 11 月生,西安电子科技大学计算机学院教授,博士生导师.于 1987 年、1990 年和 2003 年获得西安交通大学学士、西北大学硕士和西安电子科技大学电子博士学位.2004 年 6 月至 2005 年 6 月在加拿大 University of Guelph 做访问学者,从事计算生物信息学交叉学科的研究工作.主要研究方向包括计算生物信息学,生物数据挖掘,图论与组合优化算法及其应用等.