系统发生树构建技术综述

李建伏, 郭茂祖
(哈尔滨工业大学计算机科学与技术学院, 黑龙江哈尔滨 150001)

摘 要: 随着不同的分子测序技术的飞速发展使得大量的 DNA 分子数据不断涌现, 这给生物学家提供了大量的数据使其实现重构地球上所有生命的进化树的梦想. 并且, 进化树的研究对于解决现代分子生物学中的许多问题都是非常关键的, 如多序列比对、蛋白质结构和功能预测以及药物设计等等. 但是构建进化树又是一个非常复杂的问题. 因此, 进化树的研究成了一个研究热点. 本文介绍了进化树研究的发展、研究现状, 最后在总结现有的进化树构建技术存在的问题的基础上探讨了该领域进一步的研究方向.

关键词: 系统发生树; 邻接法; 最大简约法; 最大似然法
中图分类号: Q811 文献标识码: A 文章编号: 0372-2112 (2006) 11-2047-06

A Review of Phylogenetic Tree Reconstruction Technology

LI Jian-fu, GUO Mao-zu
(*School of Computer Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang 150001, China*)

Abstract: With the rapid development of sequencing technologies, molecular data are accumulated with unprecedented pace which enable biologists to reconstruct the tree of life of all the organisms in the world. Moreover, the use of evolutionary trees is a fundamental step in many biological problems, such as multiple sequence alignments, protein structure and function prediction, and drug design. In this paper, we review the development and studying status of phylogeny reconstruction technology. The major limitations in phylogeny reconstruction technologies are analyzed. Then some vital aspects that may be conducted in the future investigations are discussed.

Key words: phylogenetic tree; neighbor joining; maximum parsimony; maximum likelihood

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)就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, 这一直是生物学的研究热点. 自 1859 年 Darwin 的《物种起源》(Origin of Species) 发表以来, 重构地球上所有生命的进化史并以系统树的形式描述这部历史一直是每一个生物学家的梦想. 由于化石保存的不完备性使得由化石记录推导出的系统发生树缺乏中间环节, 虽然利用现存物种的形态和生理学的研究大致填补了化石系统发生树树的空缺, 但由于形态和生理性状的进化非常复杂, 因此对分类单元何时与最近祖先分歧等细节性问题含糊不清.

自从上世纪 50 年代末期, 尤其是 1985 年产生的 PCR (Ploymerase Chain Reaction) 技术之后, 不同的分子测序技术的飞速发展使得大量的 DNA 分子数据不断涌现, 进化论的研究也进入了分子水平, 使得这种局面大大改观. 由于所有的生命蓝图都用 DNA(在某些病毒中则用 RNA)来书写, 因此人们可以通过比较 DNA 来研究它们的进化关系. 分子途径较经典的形态学和生理学途径有如下优点. 首先, DNA 仅由 4 种碱基组成, 即: 腺嘌呤(A)、胸腺嘧啶(T)、胞嘧啶(C)和鸟嘌呤(G). 所有生物, 不论是细菌、植物和动物中的 DNA 均由这 4 种碱基组成. 因而, 可用它们比较所有有机体的进化关系. 这在经典进化研究方法中是不可能做到的. 其次, DNA 的进化演变存在某种程度的规律性, 因而能用数学模型来描述其变化并可比较亲缘关系较远的生物间的 DNA. 而形态性状的进化演变, 即使在一段较短的进化时间, 也是极其复杂的. 因而, 基于形态的系统发生树的研究必然会有各种各样的假设, 但这些假设往往难以令人信服. 第三, 所有生物的基因组都是由长长的核酸序列组成, 比形态性状包含的进化信息要多得多.

随着人类基因组计划的完成, 大量的分子数据不断涌现, 刺激了分子生物学的飞速发展. 这个崭新的时代又赋予了种系发生树的新的意义. 除了对地球上物种的进化史的研究, 进化树的研究还有更多更加重要的意义. 它有助于了解病毒传播的方式, 例如在非典时期, 对各种 SARS 病毒的研究, 通过构建系统发生树能确定各种病毒之间的关系, 得出病毒到底

收稿日期: 2005-10-15; 修回日期: 2006-05-15
基金项目: 国家自然科学基金(No. 60671011); 黑龙江省自然科学基金(No. F2004-16); 哈尔滨工业大学校基金(No. HIT. 2003.53)

是由人类传染给动物,还是由动物传染给人类的;有一些序列比对算法要依赖于进化树,所以它可以帮助科学家更好地比对蛋白质和 DNA 序列;有助于基因功能的研究,基因功能的预测往往是由于从该基因的进化史中提炼得到的——知道在一个机体中一个特定基因的功能对于在与该机体亲缘关系紧密的机体中的相似基因的意义的了解非常重要;进化树可以很好的研究进化,告诉我们一些进化机制以及不同的进化事件以及其产生的原因;假如用序列的进化史作为指导树,在数据库中搜索同源序列关系的时候将会变得更加方便等等.由此可见,进化树在在解决生物学的很多重大问题上都有非常重要意义,从生物学领域到基因组学再到病毒学领域.因此,系统发生树的研究成了一个研究热点.

2 推断系统发生树的研究现状

系统发生树推断就是一个根据某种标准,从给定的一组序列数据中推导出这些对象之间"最好"的系统发生树的过程.在生物领域内,待处理对象通常是生命机体、基因组和基因序列.用统计方法重建系统发生树分别独立地起始于形态学性状的数值分类法和分析基因频率数据的群体遗传学.在这些学科中发展起来的某些统计方法至今仍然用于分子数据的系统发生分析.近年来,随着计算机技术的飞速发展,又产生了许多新的方法[1].现在最常用的推断系统发生树的方法可以分为两大类[2]:基于算法(algorithm based)的方法和基于最优原则的方法(criteria based methods).但是两种方法之间没有一个严格的界限,某些算法可以看作是属于第一类也可以属于第二类的.基于算法的方法就是通过一系列的步骤来产生一个进化树,而基于最优原则的方法是首先定义评价每个进化树的好坏的标准,从而确定哪一棵进化树是最好的树.从方法的定义可以看出,基于最优原则的方法相对于基于算法的方法一个最大的优点就是基于最优原则的方法给每个可能的系统发生树一个评价值,那么我们可以根据这个分值对拓扑结构的好坏程度进行排队,这为我们提供了评价系统发生树的好坏的一个定量标准.另一方面,从计算复杂性上来讲,由于基于算法的方法不需要评价每一个可能的拓扑结构,因此基于算法的方法的计算速度要比基于标准的方法快得多.下面将对每种方法的原理以及优缺点进行简单的介绍.

2.1 基于算法的方法

目前最常见的基于算法的构建系统发生树的方法有距离法.在距离法[3]中,首先需要根据某种进化模型计算所有对象间的进化距离,然后根据不同的算法,从进化距离最短的开始依次聚类,利用距离方阵计算出最优树,或将分支长度之和最小化,获得最优树.可见距离法完全以距离矩阵为基础,对于核苷酸序列,两序列间的距离常规地定义为每位点发生的替代数.由于一个位点上可能发生多次替代,两序列间有差异位点的比例不能反映进化间的真实距离,需对多次替代事件进行校正.一般假定替代模型遵从马尔可夫过程,而且通常对核苷酸替代的模型也做进一步的假设,如 Jukes 和 Cantor[4] 的公式假设 4 种核苷酸以同等概率替换;Kimura[5] 的公式则容许转换(transition,一个嘌呤即腺嘌呤或鸟嘌呤被另一个不同的嘌呤所替代,或一个嘧啶即胸腺嘧啶或胞嘧啶被另一个不同嘧啶所替代)和颠换(tranversion,一个嘌呤被任意的一个嘧啶所替代,或者一个嘧啶被任意的嘌呤所替代)以不同速率发生.这些约束性假定就形成了很多替代模型及相应的距离计算公式.

依据不同的聚类算法,距离法又有以下几种:UPGMA[6](Unweighted pair group method with arithmetic mean,使用算术平均的不加权的组对法)、Fitch-Margoliash 法[7]、邻接法(Neighbor Joining,NJ)[8],其中每种方法都有自己的优缺点.UPGMA 法是建立在沿着树的所有分支的突变率相等的假设之上,因此在不同分支间进化速率有较大差异或有同源序列的平行进化时常得出错误的拓扑结构[9],而且当进化树的状态空间较大,UPGMA 法的可操作性极差,因而该建树方法的使用极为有限.Fitch-Margoliash 法则去除了 UPGMA 关于所有分支的突变率相同的假设.Fitch-Margoliash 法承认利用该方法得到的系统发生树的拓扑结构可能是不正确的,并建议考察其它的拓扑结构,然后利用百分标准差(centesimal standard deviation)来比较不同的拓扑结构,最好的进化树应该具有最小的百分标准差.因此,当各个分支的突变率不同的时候,Fitch-Margoliash 法一般能够得到比 UPGMA 法更加准确的结果,是距离法建树中一种相对较好的方法.NJ 法的运算速度最快,是目前应用最广泛的一种距离法.但该算法每迭代运算一次均只搜索最近邻居配对,对其他可能的配对不加考虑,最终只生成单一的最优树,可能会遗漏一些拓扑结构更合理的次优树;并且对于某些数据[10],NJ 得到的系统发生树的拓扑结构不够准确,因此目前出现了一些对于 NJ 的改进算法,例如文献[11~14].

2.2 基于最优原则的方法

基于最优原则的方法从数学角度讲就是在评价树的最优标准的基础之上,找到使得目标函数最优的树.目前最常用两种基于最优原则的方法有最大简约标准(maximum parsimony,简称 MP)和最大似然(maximum likelihood,简称 ML)标准.虽然每种最优原则都有不尽相同之处,但是不论是基于最大简约标准的方法还是基于最大似然标准的方法,它们构建系统发生树的最根本的思路是相同的,即首先找出由对象形成的所有的可能的树,然后利用最优标准对每一个树进行评价,给每个树赋予一个分值,最后将具有最好的分值的树作为最优树.但是 n 个对象可以有 $\Pi_{i=3}^{n}(2i-5)$ 种可能的进化树,所以当对象数目超过 10 时,在目前计算机所能够容忍的时间和空间内不能够对每一棵进化树都一一做评估.但是在分子数据大量积累的今天,往往所处理的数据量又是非常大的.并且,现在已经证明了构建 n 个对象的最大简约树[15]和最大似然树[16]都是 NP-难的.因此针对两种方法,目前出现了许多改进算法.

2.2.1 基于最优标准的方法

目前基于最优标准的方法有两种:最大简约法和最大似然法.分别介绍如下:

(1)最大简约法 最大简约方法 MP 源于形态性状研究[17],后来又出现了许多不同版本.MP 方法考虑 4 个或 4 个以上的核苷酸(或氨基酸)序列($m>4$),假设 4 种核苷酸(或

20 种氨基酸) 可突变为与自身不同的任何一种(即所有方向). 这样对于任一结定的拓扑结构, 可以推断每个位点的祖先状态. 对这一拓扑结构, 可以计算出用来解释整个进化过程所需核苷酸(或氨基酸) 的最小替代数目 s, 对所有可能正确的拓扑结构进行这种计算并挑选出所需替代数的最小拓扑结构作为最优系统发生树. 该方法的理论基础是奥卡姆哲学原则, 即解释一个过程的最好的理论是所需假设数目最少的那一个. Sober[18] 指出, 如果对系统发育推断所需要知道的进化过程愈少, 得到的结论就愈可信.

简约法是一种不依赖任何进化模型的无噪声统计方法[19], 能快速地分析出大量序列之间的系统发生关系, 所构建的树中的短分支更接近真实. 但简约树的分值完全决定于所有重建祖先序列中的最小突变数, 而突变是否按照事先约定的核苷酸最少替代的途径进行是不得而知的, 单一的突变图谱可能会得出似是而非的结论. 再者, 所有分支的突变数不可能相同, 由于没有考虑核苷酸的突变过程, 使得长分支末端的序列由于趋同进化而显示较好的相似性趋同现象违背了简约法则, 导致的结果是对“长枝吸引”的敏感[20]. 因此, 当序列单位位点上核苷酸替代数相对较大时, MP 法则极可能得出错误拓扑结构的树[21].

(2) 最大似然法　最早将最大似然法 ML 用于系统发生树推断工作的是 Cavalli-Sforza 和 Edwards[22] 对基因频率数据的分析. 但他们在应用该方法时遇到了一系列问题. 其后, 基于核苷酸序列数据的分析, Felsenstein[23] 提出了一种用最大似然法构建系统发育树的算法. 在 ML 法中, 以一个特定的替代模型分析既定的一组序列数据, 使所获得的每一个拓扑结构的似然率均为最大, 挑出其最大似然值最大的拓扑结构选为最终树. 所考虑的参数不是拓扑结构而是每个拓扑结构的枝长, 并对似然值求最大来估计枝长. 单个位点的似然值是指在核苷酸替代模型中该位点每个可能被取代或再现的概率之和, 进化树的似然值就是所有位点似然值的乘积. 其分析的核心在于替代模型——根据碱基频率的相等或不等、转换和颠换速率的相等或不等、位点间替代速率已执行异质性的有无以及不变位点比例的高低等若干特征可以有 56 种之多. 而模型的正确选择也就成为最大似然法的关键之所在[24]. 另外, 最大似然法是一种建立在进化模型基础上的统计方法, 具有很多优越性, 如一致性、健壮性、能够在一个统计框架内比较不同的树以及能够充分利用原始数据等等, 但是由于计算每个树的似然值算法的复杂性以及基于最优标准算法的复杂性使得最大似然法实现起来非常困难, 因此目前出现了许多对于最大似然算法的改进算法.

2.2.2　降低算法复杂性的方法

目前经常用的减少计算复杂性的方法主要是从以下两个角度对系统发生树的构建算法进行改进:

(1) 减少计算过程中序列的位点的数目. 这样在评价每一棵树的时候, 就减少了对该树中每个内部节点的计算量, 从而降低了整体的计算量, 如文献[25]通过使用衡量子树质量的向量来加速最大似然法.

(2) 使用一些启发式信息来较少算法搜索空间的大小. 使用启发式搜索算法不能确保能够找到最优解, 但是能够使得计算速度大大提高. 目前, 关于这种使用启发式信息来引导构建进化树算法的方法非常多, 主要有在 (a) PAUP * [26], PHYLIP[27] 和 fastDNAML[28] 软件包中使用最多的分支交换法 (Branch-Swapping) 以及后来提出的对于分支交换技术的改进策略, 例如文献[29~31]. (b) Disk-Covering[32] 方法和基于 Quart 的方法[33~36] 为代表的分治算法 (Divide-and-Conquer Techniques). (c) 也有一些人将模拟退火技术来指导构建进化树, 并且取得了不同程度的成功, 例如文献[37~40]. (d) 将遗传算法用来指导构建进化树[41,42]. (e) 使用 Markov Chain Monte Carlo 方法来指导建树, 例如文献[43~49]等等.

2.3　检验标准

由于历史不会再重现, 因此无法直接考察有机体之间的系统发生关系, 也就不存在评估系统发生树的绝对标准. 但是, 无论使用何种方法重建系统发生树, 都必须对其拓扑结构以及整个树的可靠性的统计置信度进行检验. 评价系统发生树种每一个分支的可靠性, 统计学上用重复取样来排除随机误差的影响, 常用的方法有两种自举法(bootsrap method) 和刀切法(jackknife method)[50]. 在分子系统学中, 一般不可能去真正地重复取样, 只能是由原有数据产生假重复数据. 自举法是原有数据中的性状进行复置重复取样, 即随机抽取一个性状后, 再将该性状放回原数据, 继续随机取样, 直到新产生的一组数据大小与原是数据相同为止. 结果各个性状被抽取的次数可能不同. 刀切法的取样是不复置取样. Muller 和 Ayala 的刀切法是每次从原始数据中去掉一定数目的性状, 然后再对剩余的所有性状进行系统发育分析. Layon 的刀切法则是每次去掉一个分类群对象, 然后对剩余的所有分类群进行分析. 可见刀切法产生的数据小于原有数据. 由此可见, 自举法和刀切法的差别在于重复取样的方式有所不同. 前者从原始数据集中以相同的概率抽取每个位点(由于随机性, 有些位点重复了, 有些则缺失了), 直到新建数据集同原始数据集的位点总数相等, 然后对于产生的新数据集建树, 重复若干次, 得到特定分支格局的出现频率——自举值; 后者同前者的差别仅在于新建的数据集合要比原始数据集小, 而且不包含重复位点. 如果一个内部分支在原始系统发生树中产生的序列分割与在用多次抽样产生的新数据结构构建系统树中产生的序列分割相同, 那么就被赋值 1, 否则赋值 0. 此过程重复 100 次, 该原始树的每个内部分支得到的值 1 的次数百分比被计算, 用以表示每个内部分支的可靠程度. 例如, 如果某个分支在 100 个系统树中得到值 1 的次数为 95 次, 则该分支的可靠程度为 95%. 这个数值越高, 则该分支越可靠.

2.4　其他相关研究

为了进一步避免基因转移和缺少合适的多物种共有的保守序列的影响, 基于基因组信息的增加, 有不少学者提出了采用全基因组建立系统发生树的方法, 例如文献[51]. 其中有的采用对不同基因组中包含和缺少的基因及其拷贝数进行统计分析并建立相应的数学模型, 然后再对不同物种进行进化分析, 如 Bork 等及 Fitz-Gibbon 和 House 等[52] 基于基因组中相邻蛋白(or2thologous protein) 信息建立数据矩阵和进化树, 其结

果与先前的进化树基本相符. 国内的程淼等[53]有相类似的研究. 不过也有人如 Doolittle[54]提出这样的方法可能会对数据的选择产生偏爱, 由于只对不同基因组中含有或缺少的基因进行分类处理, 那些多物种中都存在的高保守性的基因就有可能被忽略. 不过也有研究表明这样的基因相对来说比较少. 还有人如郝柏林等[55]基于分析原核物种基因组中寡核苷酸片段(6-8 bps)的各种组合的出现频率建立系统发生树的全基因组建树方法, 也得到了与公认的生命之树基本相符的结果. 每一个物种都包含很多的编码蛋白, 它们都部分包含了该物种的进化信息.

虽然基于全基因组的进化树构建方法相对于基于序列数据的进化树构建方法具有很多优点, 如避免了基因树和物种树的问题, 不需要进行序列比对, 基因重组和复制事件比起核苷酸突变要少得多等, 这样使得我们能够跟踪的历史比序列数据更加久远. 但是, 基于全基因组数据的进化树构建方法也有缺点, 最重要的一点就是数据的缺乏, 真核细胞选择的都是基因组计划中的模型物种: 人类, 老鼠, 果蝇, 蠕虫, 芥菜, 酵母等, 尽管数目在迅速的增长, 但是在短时间内将不会超过所描述的基因组的一小部分. 数据的缺少反过来会产生另一个问题, 即没有针对于基因序列数据的好的进化模型. 缺少模型和第三个问题即基因序列的极端的复杂性, 产生了复杂的计算问题. 因此, 全基因组建树引入了更多的基因表型特征, 其结果的稳定性和客观性更强, 如能方便计算处理, 其应用必将会进一步广泛.

3 系统发生树研究的发展方向

随着分子生物学研究的深入, 越来越多序列的进化意义最终会被揭示, 对各种构树方法也提出了新的挑战, 如何提高算法对复杂数据的处理能力将成为今后系统发生树研究的重点. 具体包括以下几个方面:

(1) 从现在的分析来看, 最大似然法是目前最准确的一种方法, 但是一个最大的缺点就是计算复杂性高, 这使得它不能够处理大规模数据, 这与目前不断出现的大量数据之间的矛盾越来越明显, 因此需要引入新的技术来加速最大似然法的计算.

(2) 生物数据不同于普通的数据, 它有一定的生物学意义, 因此要充分挖掘其生物学意义, 结合到目前最常用的距离法和最优原则法当中, 得到真正具有生物学意义的系统发生树.

(3) 进化统计模型为一些方法提高了更加坚实的统计基础. 但是, 进化模型的好坏直接影响到构建的系统发生树质量的高低. 而现在所用的都是一些非常简单的模型, 因此深入研究分子的进化机制, 开发出更加有现实意义的进化模型, 这样才能使得建立在进化模型之上的系统发生树的研究更具有现实意义.

(4) 任何一种方法都不可能完全模拟出进化的真正历史, 每种方法都有有缺点, 因此应该将多种方法进行综合, 以开发出更好地更加全面的算法.

总之, 进化树的研究一直在不断的发展完善, 但与真实的物种进化关系还有一定距离. 在以上四点进一步引入分子进化分析后, 未来的系统发生树将在生物学研究中发挥更大的作用. 随着后基因组时代的到来, 各种生物信息的增多, 系统发生树的概念也会更加充实, 它不仅仅是某个特殊物种的可能进化反映, 而会成为整个进化历史的中心趋势的参照. 可以预见的是, 未来的进化树将会更加细致深入地反映生命的历史, 为我们对生命进化的预测和生物学相关问题的分析提供参考.

参考文献:

[1] Felsentein J. Phylogenies from molecular sequences: inference and reliability[J]. Annual Review of Genetic, 1988, 22: 521-565.

[2] [美]根井正利, 库马著(吕宝忠, 钟扬, 高莉萍译, 赵寿元, 张建之校). 分子进化与系统发生[M]. 北京: 高等教育出版社, 2002.

[3] Pauplin Y. Direct calculation of a tree length using a distance matrix[J]. Journal of Molecular Evolution, 2000, 51(1): 41-47.

[4] Jukes T H, Cantor C R. Mammalian Protein Metabolism[M]. New York: Academic Press, 1969. 21- 132.

[5] Kimura M. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences[J]. Journal of Molecular Evolution, 1980, 16(2): 111-120.

[6] Dave Thomas. Example calculation of phylogenies: the UPGMA method[EB/OL]. http://www.nmsr.org/upgma.htm, 2002-10-31/2005-10-10.

[7] Chris. Fitch-Margoliash algorithm for calculating the branch lengths[EB/OL]. http://www.bioinfo.rpi.edu/~bystrc/courses/biol4540/lecture12/sld002.htm, 2004-12-10/2005-10-10.

[8] 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.

[9] 李靖炎. 消偶合今祖法的提出[J]. 动物学研究, 1992, 13(4): 387- 396.
Li Jing Yan. The pdaric method for constructing molecular evolutionary trees from sequences data[J]. Zoological Research, 1992, 13(4): 387-396. (in Chinese)

[10] B M Moret, U Roshan, T Warnow. Sequence length requirements for phylogenetic methods[A]. Proc 2nd Int'l Workshop on Algorithms in Bioinformatics, Lecture Notes in Computer Science[C]. Heidelberg, Berlin: Springer, 2002, 2452: 343-356.

[11] Pearson WR, Robins G, Zhang T. Generalized neighbor-joining: more reliable phylogenetic tree reconstruction[J]. Molecular Biology and Evolution, 1999, 16(6): 806- 816.

[12] William J Bruno, Nicholas D Socci, Aaron L Halpern. Weight-

ed neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction[J]. Molecular Biology and Evolution, 2000, 17(1): 189-197.

[13] Vincent Ranwez, Olivier Gascuel. Improvement of distance-based phylogenetic methods by a local maximum likelihood approach using triplets[J]. Molecular Biology and Evolution, 2002, 19(11): 1952-1963.

[14] Satoshi Ota, Wen-Hsiung Li. NJML: a hybrid algorithm for the neighbor-joining and maximum-likelihood methods[J]. Molecular Biology and Evolution, 2000, 17(9): 1401-1409.

[15] W Day, D Johnson, D Sankoff. The computational complexity of inferring rooted phylogenies by parsimony[J]. Mathematical Biosciences, 1986, 81(1): 33-42.

[16] Sebastien Roch. A short proof that phylogenetic tree reconstruction by maximum likelihood is Hard[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2006, 3(1): 92-94.

[17] Henning, W. Phylogenetic systematics[M]. Urbana: University of Illinois Press. 1966.

[18] E Sober. Reconstructing the past: parsimony, evolution and inference[M]. Cambridge: MIT press. 1988.

[19] Sourdis J, Nei M. Relative efficiencies of the maximum parsimony and distance matrix methods in obtaining the correct phylogenetic tree[J]. Molecular Biology and Evolution, 1988, 5(3): 298-311.

[20] Holder M, Lewis P O. Phylogeny estimation: traditional and Bayesian approaches[J]. Nature Reviews Genetics, 2003, 4(3): 275-284.

[21] Li W H. Evolutionary change of restriction cleavage sites and phylogenetic inference[J]. Genetics, 1986, 113(1): 187-213.

[22] L L Cavalli Sforza, A W Edwards. Phylogenetic analysis: models and estimation procedures[J]. American Journal of Human Genetics, 1967, 19(3): 233-257.

[23] J Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach[J]. Journal of Molecular Evolution, 1981, 17(6): 368-376.

[24] Mike S, David P. Parsimony, likelihood, and the role of models in molecular phylogenetics[J]. Molecular Biology and Evolution, 2000, 17(6): 839-850.

[25] Alexandros P Stamatakis, Thomas Ludwig, Harald Meier. Accelerating parallel maximum likelihood-based phylogenetic tree calculations using subtree equality vectors[A]. In Proceedings of 15th IEEE/ACM Supercomputing Conference (SC2002)[C]. NW Washington: IEEE computer Society, 2002. 1-16.

[26] D Swofford. PAUP* —Phylogenetic Analysis Using Parsimony (and other methods)[CP/OL]. http://www.PAUP.CSIT.FSU.EDU, 2002-03-04/2005-10-10.

[27] Joseph Felsenstein. PHYLIP (Phylogeny Inference Package)[CP/OL]. http://www.phylip.com, 2005-8-30/2005-10-10.

[28] Olsen G J, H Matsuda, R Hagstrom, R Overbeek. FastDNAml: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood[J]. Computer Application in the Biosciences, 1994, 10(1): 41-48.

[29] S Guindon O, Gascuel. A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood[J]. Systematic Biology, 2003, 52(5): 696-704.

[30] Wim Hordijk, Olivier Gascuel. Improving the efficiency of SPR moves in phylogenetic tree search methods based on maximum likelihood[J]. Bioinformatics, 2005, 21(24): 4338-4337.

[31] L S Vinh, AV Haeseler. IQPNNI: Moving fast through tree space and stopping in time[J]. Molecular Biology and Evolution, 2004, 21(8): 1565-1571.

[32] D Huson, S Nettles, T Warnow. Disk-covering, a fast converging method for phylogenetic tree reconstruction[J]. Journal of Computational Biology, 1999, 6(3): 369-386.

[33] P Erdos, M Steel, L Szekely, T Warnow. Local quart splits of a binary tree infer all quartet splits via one dyadic inference rule[J]. Computers and Artificial Intelligent, 1997, 16(2): 217-227.

[34] K Strimmer, A Von Haeseler. Quarter puzzling: a quartet maximum-likelihood method for reconstruction tree topologies[J]. Molecular Biology and Evolution, 1996, 13(7): 964-969.

[35] Haoyong Zhang. Design, implementation, and analysis of a novel quartet-based phylogenetic reconstruction method[D]. Canada: the University of Waterloo, 2000.

[36] Michael Hu. A collapsing Method for efficient recovery of optimal edges in phylogenetic trees[D]. Canada: the University of Waterloo, 2002.

[37] D Barker. LVB: Parsimony and simulated annealing in the search for phylogenetic trees[J]. Bioinformatics, 2004, 20(2): 274-275.

[38] L Salter. Simulated-based estimation of phylogenetic trees[D]. Columbus: The Ohio State University, 1999.

[39] L Salter, D Peal. A stochastic search strategy for estimation of maximum likelihood phylogenetic trees[J]. Systematic Biology, 2000, 50(1): 7-17.

[40] A Stamatakis. An efficient program for phylogenetic inference using simulated annealing[A]. Proceeding of 19th IEEE/ACM International Parallel and Distributed Proceedings Symposium[C]. NW Washington: IEEE computer Society, 2005. 198-200.

[41] P Lewis. A genetic algorithm for maximum-likelihood phylogeny inference using nucleotide sequence data[J]. Molecular Biology Evolution, 1998, 15(3): 277-283.

[42] Lemmon A, Milinkovitch M. The metapopulation genetic algorithm: an efficient solution for the problem of large phylogeny

estimation[J]. Proceedings of the National Academy of Sciences of the United State of America, 2002, 99(16): 10516-10521.

[43] S Li, D Pearl, H Doss. Phylogenetic tree construction using Markov chain Monte Carlo[J]. Journal of the American Statistical Association, 2000, 95(451): 493-508.

[44] B Mau, M Newton, B Larget. Bayesian phylogenetic inference via Markov chain Monte Carlo methods[J]. Biometrics, 1999, 55(1): 1-12.

[45] ZH Yang, B Rannala. Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method[J]. Molecular Biology Evolution, 1997, 14(7): 717-724.

[46] Rambaut A, Grassly NC. Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees[J]. Computer Application in Biosciences, 1997, 13(3): 235-238.

[47] Larget B, D Simon. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees[J]. Molecular Biology of Evolution, 1999, 16(6): 750-759.

[48] D Simon, B Larget. Bayesian analysis in molecular biology and evolution(BAMBE), version 2.03 beta[CP/OL]. http://www.mathcs.duq.edu/larget/bambe.html, 2001-04-18/2005-10-10.

[49] J P Heulsenbeck, F Ronquist. Mrbayes: Bayesian inference of phylogeny[J]. Bioinformatics, 2001, 17(8): 754-755.

[50] Jun Shao, Dongsheng Tu. The Jacknife and Bootstrap[M]. USA: Springer, 1996.

[51] Youri I Wolf, Igor B Rogozin. Genome trees and the tree of life[J]. Trends Genetic, 2002, 18(9): 472-479.

[52] Snel B, Bork P. Genome phylogeny based on gene content[J]. Nat Genet, 1999, 21(1): 108-110.

[53] 程淼, 阮戈, 饶子和. 利用物种的蛋白质系统发生图谱研究物种进化[J]. 自然科学进展, 2002, 12(12): 1309-1313.

[54] W F Doolittle. Some thoughts on the tree of life[A]. The Harvey Lectures series 99 2003-2004[C]. United states: University of Minnesota press, 2005. 111-128.

[55] Hao BL, Qi J, Wang B. Prokaryotic phylogeny based on complete genomes without sequence alignment[J]. Modern Physics Letter, 2003, 17(3): 91-94.

作者简介:

李建伏 女, 1979 年生于河北, 现为哈尔滨工业大学计算机科学与技术系博士研究生, 主要研究方向为生物信息学.
E-mail: jianfu_lili@163.com

郭茂祖 男, 1966 年生于山东德州, 哈尔滨工业大学教授, 博士生导师, 中国人工智能学会机器学习专业委员会委员, 主要研究方向为机器学习与专家系统、生物信息学与 DNA 计算、多主体系统、色彩匹配技术、随机算法与近似算法.