• 医学信息研究 •

基于 MATLAB 生物信息学工具箱构建分子系统发生树

张乐平 黄 非 闵 波 李东方

（第二军医大学计算机教研室 上海 200433）

〔摘要〕 分子水平的系统发生分析相对于经典系统发生方法更加科学、可靠。概述系统发生分析基础，详细介绍基于距离的系统发生分析方法，相关实验表明 MATLAB 具有强大的数据处理能力和方便实用的工具箱，便于系统发生分析的研究与应用。

〔关键词〕 生物信息学；分子进化；系统发生树

Construction of Molecular Phylogenetic Tree based on MATLAB Bioinformatics Toolbox *ZHANG Le – ping, HUANG Fei, MIN Bo, LI Dong – fang, Department of Computer, Second Military Medical University, Shanghai 200433, China*

〔**Abstract**〕 Comparing with classic phylogenetic methods, the molecular phylogenetic analysis can be more scientific and reliable. The paper outlines the basis of phylogenetic analysis, introduces the distance – based phylogenetic analysis method in detail. The related experimental results show that MATLAB possesses powerful data processing capability and provides many convenient and practical toolboxes, which are beneficial for the research and application to phylogenetic analysis.

〔**Keywords**〕 Bioinformatics; Molecular evolution; Phylogenetic tree

分子系统发生分析是生物信息学中一种研究进化的基本方法，一个可靠的系统发生推断将有助于对地球上不同物种进化关系的认识。随着后基因组时代的到来，从生物学领域到基因组学再到病毒学领域，进化树在解决生物学的很多重大问题上都有非常重要的意义[1]。在国际学术界，MATLAB 已经被公认为准确、可靠的科学计算标准软件。从 MATLAB 6.5.1 首次增加了引人注目的生物信息学工具箱以来，如今的 MATLABR2009a 版生物信息学工具箱的数据处理能力得到了极大的提升。本文以 MATLAB 生物信息学工具箱为基础，重点讨论利用距离法构建分子系统发生树。

1 系统发生分析基础

1.1 系统发生学相关概念

系统发生学通过比较物种的特征研究生物形成或进化的历史，其研究结果以系统发生树表示。系统发生树是由节点和分支组成的一种二叉树，节点代表分类单元（物种或序列），而分支则表示物种之间的进化关系。经典系统发生学主要通过形态学和生理学途径获取生物的特征，并成功构建了很多植物和动物的进化树，形成了大量有价值的生物学认识。但是，依靠这样的生物表型特征进行研究是有局限的，例如，有些关系很远的生物由于趋同进化也会造成相似的表型。

〔收稿日期〕 2010 – 01 – 11

〔作者简介〕 张乐平，副教授，主要研究方向为计算机应用、生物信息学。

随着分子测序技术的飞速发展，分子序列数据呈指数级增长，进化论的研究进入分子水平。分子序列数据常常可以用一个有限的字符集合来描述，例如，不论是细菌、植物还是动物，DNA 序列都是由 A、T、C、G 4 种碱基组成，这样任何生物基于分子序列都可以进行比较。而且，分子序列（例如 DNA）的进化具有统计规律性，从而可以用严格的数学模型描述其变化，容易形成关于进化过程的可验性假设。因此，分子水平的系统发生分析结果更加科学、可靠。

1.2 构建分子系统发生树的方法

用于在分子水平构建系统发生树的特征数据分成两类。距离数据：分子序列之间的距离是指一个序列变化到另一个序列所需的最小变化数目，常常用距离矩阵来描述；特征数据：表示能体现序列之间差异的分子水平的特征。基于距离的构树方法呈现的是序列之间的整体差异，而基于特征的方法强调的是那些特殊的信息位点。因为基于距离的方法和基于特征的方法所采用的分析有着本质上的区别，所以它们关于进化关系结论的一致性可以看作是对一棵系统发生树的正确性的认可[2]。

无论是距离构树法还是特征构树法，系统发生树的推断都是基于某种最优原则的假设。例如，在距离法中，紧邻法是基于最小进化原理，即寻找树的所有分支长度和最小的拓扑结构；在特征法中，最大简约法是基于最小替代数的思想，认为具有最小替代数的拓扑结构是最优树，而最大似然法是基于最大似然率的思想，也就是选出似然率最大的拓扑结构为最优树。

2 基于距离的分子系统发生分析

2.1 序列比对

MATLABR2009a 版生物信息学工具箱提供了分子系统发生树的距离法建树工具，因此重点介绍基于距离的分子系统发生分析的基本原理和过程[2-3]。在进化过程中，分子序列可能会发生插入、删除和置换等变化。通过序列最优比对，分子序列因间隔的插入变得等长，同源性得到更好的体现。因此序列比对是进行同源分析的一种基本手段，是分析序列之间差异的基础。

序列比对的加权和可以根据常用的打分矩阵计算，如果分子序列是氨基酸序列，则用 PAM 矩阵、BLOSUM 矩阵等；如果分子序列是 DNA 或者 RNA，则用单位矩阵、核苷酸转换——颠换矩阵或者 BLAST 矩阵等。

2.2 距离计算

对于两条长度为 N 的序列，统计差异数目 Nd，两条序列之间的距离常规地定义为 p = Nd /N，所有序列之间的两两距离则构成一个距离矩阵。由于序列比对过程中，常常用间隔（-）表示插入或删除，这些间隔增加了距离计算的复杂度。在距离估计中一般忽略间隔，具体有两种方法：完全删除和成对删除，见图 1。显然，成对删除利用了更多有用信息，因此经常被采用。

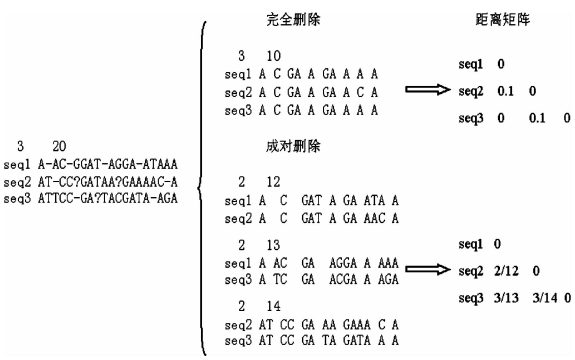

图 1 完全删除和成对删除

2.3 距离校正

如果序列之间差异很大，那么如上简单计数会严重低估序列在最近的祖先之后发生的替换数目。因此，距离必须通过数学模型加以校正。为了能正确无误地估计序列之间的分歧度，数学替换模型必须考虑到回复和平行突变等情况。例如，Jakes - Cantor 单参数模型假设 4 种核苷酸以相同的概率替换；而 Kimura 双参数模型考虑到真实序列之间转换（嘌呤转变成嘌呤，嘧啶转变成嘧啶）的频率远远高于颠换（嘌呤转变成嘧啶，嘧啶转变成嘌呤）频

率，允许转换和颠换以不同速率发生；如果是氨基酸序列，可采用泊松校正距离。在众多数学模型中如何选择需要具体问题具体分析。

2.4 基于距离的构树方法

利用物种或分类单元间的两两进化距离，依据一定的原则及算法构建系统进化树。这里介绍其中的两种：非加权配对算术平均法（Unweighted Pair Group Method with Arithmetic，UPGMA）和邻接法（Neighbor Joining，NJ）。

UPGMA 算法首先将两个距离最近的物种合成一个复合物种组。假设距离矩阵中的最小值是 Dab，物种 A 和 B 则合成一组（AB）。第一次聚类以后，更新距离矩阵，计算新组（AB）和其他物种之间的两两距离。然后，将新的距离矩阵中距离最小的两个物种再次合成一个复合物种组。如此反复，直到所有的物种都聚为一类。UPGMA 算法的最大优点是对于表型数据和分子数据，甚至是两者结合都很适用。但缺点是该算法假定树的所有分支的进化速率是相同的，因此，当不同分支的进化速率差异很大或有同源序列平行进化时常常得到错误的分子进化树。

NJ 法是目前应用最广泛的距离法，基于最小进化原理构建进化树，即树的所有分支长度和最小的拓扑结构为最优树。在每一轮聚类过程中，考虑所有可能的物种树，把树的整个分支长度和最小的物种对聚为一组，并产生新的距离矩阵。该方法的关键步骤一是计算发散系数，二是生成一个速率校正距离矩阵。但是，当物种数较大时，主要采用启发式搜索，可能会遗漏一些拓扑结构更合理的树。

2.5 自举检验

所有的系统进化树都是关于物种或序列的进化历史的假设，因此，需要评价系统进化树及其分支的置信度水平。自举检验是一种重采样技术，通过统计分析，量化整棵树及其不同部分的置信度水平。基本方法：从原数据集中随机抽取（同时替换）部分数据组成新的数据集，即自举数据集，然后用这个新的数据集构建系统发生树。重复该过程，产生成百上千的自举数据集，并同时生成对应的自举树。产生相同分组的自举树的数目常常标注在系统发生树相应节点的旁边，表示树中每个部分的相对置信度。例如，如果某个分支在 100 个自举树中相同的次数为 95，则该分支的可靠程度为 95%。尽管自举检验过程非常耗时，但自举法已经成为系统发生分析中很受欢迎的检验算法。

3 分子系统发生分析实验

3.1 基于 MATLAB 生物信息学工具箱构建分子系统发生树的基本过程

获取序列数据，主要有两种方式：网络数据库（getgenbank 等命令）和本地文件（load，fastaread 等命令）。计算序列间的距离，主要的命令：seqpdist。构建分子进化树，主要的命令：seqlinkage 和 seqneighjoin。自举检验，由于 MATLAB 工具箱没有直接可用的命令，需要编程实现，分 4 步：生成自举数据集；构建自举树；重复第一、二步若干次；计算每个分支的置信度。显示分支带有置信度的系统发生树。

3.2 分别用 UPGMA 和 NJ 法构建 12 种灵长类动物的系统发生树

数据来源于 MATLAB 提供的 FASTA 格式的文件“primatesaligned. fa”，文件中包含有 12 条预先序列比对好的 DNA 序列。在测试程序中，首先计算距离矩阵，然后分别用 UPGMA（seqlinkage）和 NJ（seqneighjoin）法对其构建进化树，最后自举检验，次数为 100，结果见图 2、图 3。

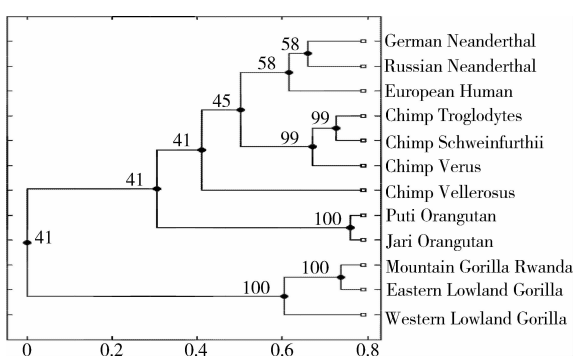

图 2 用 UPGMA 法构建 12 种灵长类动物的系统发生树

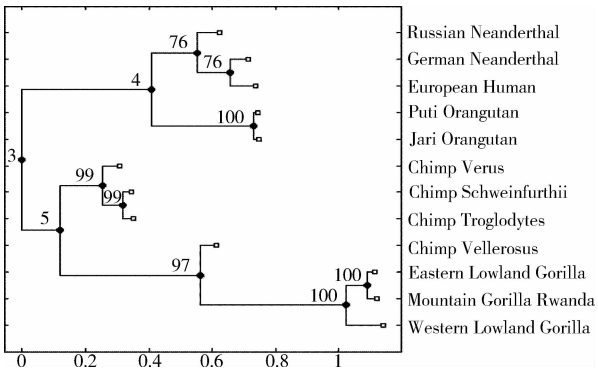

图 3 用 NJ 法构建 12 种灵长类动物的系统发生树

4 结论

从上面的实验结果可以看到，对于同一个数据集，无论是拓扑结构还是置信度水平，UPGMA 和 NJ 法产生的系统发生树差异较大。进一步试验发现，即使用同一种方法构树，如果参数设置不同，也会影响系统发生树。面对不同系统发生树的算法，应深入理解这些算法的逻辑基础，并结合具体的分子序列数据，选择合适的算法。生物数据不同于工程领域中的数据，需要充分挖掘数据的生物学意义，从而使所建的系统发生树传递可靠的信息。

MATLAB 具有强大的数据处理能力和方便实用的各种工具箱，可以节省传统计算机编程在算法细节实现中花费的大量精力，而研究人员可将注意力集中到需要解决的具体问题上，便于研发新的系统发生分析方法。例如，董安国等[4]在 MATLAB7.1 中基于最大似然法模型构建系统进化树，并计算了进化时间。

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.

2 Dan E. Krane，Michael L. Raymer 著，孙啸等译. 生物信息学概论〔M〕. 北京：清华大学出版社，2005.

3 Masatoshi Nei，Sudhir Kumur 著，吕宝忠等译. 分子进化与系统发生〔M〕. 北京：高等教育出版社，2002.

4 董安国，高琳，赵建邦，等. 基于 DNA 序列的系统进化树构建〔J〕. 西北农林科技大学学报（自然科学版），2008，36（10）：221－226.

（上接第 29 页）

3.2 引入 AIP

应用 AIP 具有重要的临床意义[8]。AIP 作为动脉硬化的血浆标记物，反映了冠状动脉硬化性心脏病危险性；将其编入 LIS 并自动计算，对于患者的医疗诊治具有重要作用。

3.3 一小时报告

检测快速，结果及时；以病人为中心，方便病人。福州总医院从肝功 5 项第 2 天报告到门诊 100 项一小时报告，说明优化组合 LIS 的配套技术是最重要的。

局域网、公式编制及优化 LIS 提高了工作与研究效率，在招标检验仪器时应讲究速度与效率，各级卫生部门应积极推广优秀的检验信息系统，以改进公共医疗卫生服务质量。

参考文献

1 Deshpande SD. ILIS－an integrated laboratory information system. i. biochemistry and hematology〔J〕. Clinical Chemistry，1982，28（2）：271－276.

2 邵松，王嘉. 检验科信息化管理〔J〕. 现代检验医学杂志，2005，20（5）：65－66.

3 胡望平，于萍，鲜荣华，等. 病例、论文与考试是实验诊断学教学和实习的三项重点〔J〕. 现代检验医学杂志，2005，20（2）：19－20.

4 安宁，吕顺超，华琛. 医院实验室信息系统的开发和应用〔J〕. 生物医学工程与临床，2006，10（3）：188－190.

5 石玉玲，李林海，徐德兴，等. 包含条形码的全信息彩色标签技术在检验科信息管理中的应用〔J〕. 中华检验医学杂志，2005，28（6）：652－653.

6 Henricus J. Automated Processing of Serum Indices Used for Interference Detection by the Laboratory Information System〔J〕. Clin Chem，2005，51（1）：244－245.

7 Harrison JP，McDowell GM. The Role of Laboratory Information Systems in Healthcare Quality Improvement〔J〕. Int J Health Care Qual Assur，2008，21（7）：679－691.

8 叶桂云，胡望平，张忠源，等. 2 型糖尿病血浆致动脉硬化指数及几种指数的比较〔J〕. 检验医学，2009，24（9）：667－670.