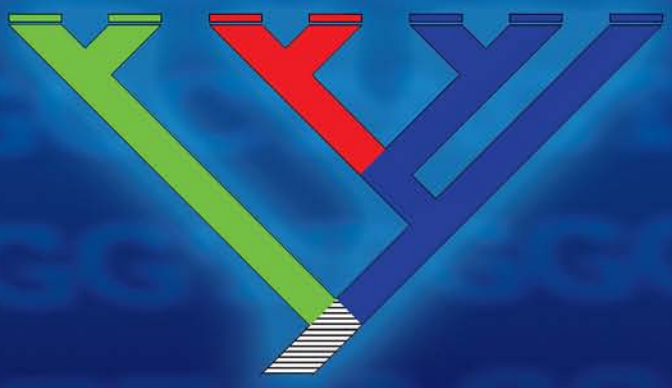
分子系统发生学

黄 原 编著



科学出版社

分子系统发生学

黄　原　编著

科学出版社

北　京

内 容 简 介

分子系统发生学是应用分子数据重建系统发生关系的学科。本书全面系统地论述了分子系统发生学的基础、原理、方法及应用。全书由 18 章组成，可以归纳为五大部分：第一部分包括第 1～3 章，分别介绍了系统发生和系统树的基本知识；第二部分包括第 4～7 章，是分子系统发生分析的基础，其中第 4 章和第 5 章是分子系统发生学的信息学基础，第 6 章是数据集系统发生信号评估，第 7 章讨论了分子进化模型及模型选择原理与方法；第三部分中的第 8～12 章是各种系统发生分析方法，分别就目前主要的系统发生分析方法（距离矩阵法、简约法、最大似然法、贝叶斯推论法和系统发生网络法等）从原理、软件操作、应用及局限性等方面进行了详细的介绍，第 13 章讨论了系统发生假设检验的原理和方法，第 14 章讨论了系统发生分析可靠性与影响因素；第四部分主要涉及各类数据集分析策略，其中第 15 章总结了不同类型数据的分析策略，第 16 章对复杂数据系统发生的分析策略与方法进行了详细地介绍，第 17 章是多基因数据分析策略和方法；最后一部分即第 18 章是系统树的可视化、注释与应用方面的内容。

本书可作为生物学、生物技术、生态学和生物信息学专业的本科生、研究生及科研人员学习分子系统发生学的教材或参考资料。

图书在版编目（CIP）数据

分子系统发生学/黄原编著. —北京：科学出版社，2012
ISBN 978-7-03-033026-0

Ⅰ. ①分… Ⅱ. ①黄… Ⅲ. ①分子进化—系统发育—研究 Ⅳ. ①Q75

中国版本图书馆 CIP 数据核字（2011）第 260101 号

责任编辑：王海光 矫天扬 刘 晶/ 责任校对：刘小梅
责任印制：钱玉芬/ 封面设计：耕者设计工作室

科学出版社 出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com

印刷
科学出版社发行 各地新华书店经销
*
2012 年 6 月第 一 版 开本：787×1092 1/16
2012 年 6 月第一次印刷 印张：34 1/2 插页：12
字数：790 000

定价：120.00 元
（如有印装质量问题，我社负责调换）

前　言

分子系统发生学是应用分子数据重建系统发生关系的学科。由于系统发生关系已经成为整合包括生物多样性在内的生物学知识的基本框架，所以构建生物类群之间的系统发生关系成为当代生物学的基本研究方法。本书全面系统地论述了分子系统发生学的基础、原理、方法及应用。

本书是作者20余年来在系统发生学领域研究和教学的总结，写作过程中注意兼顾基本概念的解释和最新进展的评述，使读者能够快速掌握本学科各个方面的基础知识。需要声明的是，作者本人主要从事分子进化和分子系统学领域的教学和研究工作，也就是分子系统发生学的应用方面，对系统发生学的理论、算法和软件并无任何创新，本书涉及的内容主要来自国外的期刊论文、课程网站和课件等，尤其需要指明的是，书中的软件列表主要来自 Joseph Felsenstein 维护的网站（http://evolution.genetics.washington.edu/phylip/software.html），部分图表的来源由于时间的关系未查明出处，在此特表歉意。如果不适当地使用了版权资料，还望作者或读者来信指明，以便将来有机会时更正。本书介绍的软件及其使用说明可以在作者实验室主页 http://www.molevobio.snnu.edu.cn 的“读者园地”中下载。

分子系统发生学涉及计算机、统计学、分子生物学、进化论、生物信息学等许多学科，限于本人的知识水平，书中疏漏谬误在所难免；另外，近年来分子系统发生学在理论和方法等方面发展迅猛，由于时间仓促，有些重要的进展未能加以详细地介绍和评述，敬请读者批评指正。

感谢为本书作出贡献的历届研究生，尤其感谢叶维萍硕士、卢慧甍博士和黄建华博士后在系统发生软件使用方面的贡献；感谢中国科学院动物研究所梁爱萍研究员、扬州大学杜予洲教授和湖北大学曾庆韬教授应邀参与研究生系统发生分析方法的讲座与讨论；感谢科学出版社王海光和矫天扬编辑在本书撰写、审稿、出版过程中所给予的帮助以及对本书提出的宝贵意见。

本书的研究工作得到了国家自然科学基金项目（39570110，30070114，30470238，30670279，30970346）的资助，本书的出版得到了陕西师范大学出版基金的资助，在此特表感谢。

黄　原

陕西师范大学生命科学学院教授

2011年8月

目　　录

第 1 章　系统发生学概论

1.1　系统发生与系统发生学

系统发生（phylogeny，由希腊词根 phylon＝stem、tribe、race 和 genesis＝origin 构成）是指任何生物实体（基因、个体、种群、物种和种上阶元）的起源和演化关系。

达尔文首次使用系统发生一词是在《物种起源》第 5 版提及 Haeckel 的著作 *Generelle Morphologie* 时，并将系统发生等同为所有生物的传代线（the lines of descent of all organic beings），这与 Haeckel 的原意不同，Haeckel 书中的系统发生是生命之树的传代线上生物在形态上发生的主要改变，而非传代线本身（Dayrat，2003）。但达尔文此处对系统发生概念的使用与我们现在的定义基本上一致。

分子系统发生（molecular phylogeny）是利用各种分子性状构建的生物实体之间起源和演化关系，采用的分子数据主要是 DNA 和蛋白质序列，也包括其他类型的分子数据。

系统发生学（phylogenetics）是研究利用各种性状构建基因、个体、种群、物种和种上单元之间系统树或网络的原理和方法的学科。系统发生学重建进化历史依赖于对取样物种的性状分布进行数学推论，这种重建涉及不同类群共享的同源性状，并通过这些性状推断系统树。这种数学推断的准确性完全依赖于对性状进化的假设和模型。

20 世纪 50 年代以来，蛋白质和 DNA 测序技术为系统发生重建带来了曙光。DNA 和蛋白质序列数据作为生物信息分子具有线性数字编码特征，并且能够建立位点之间的同源关系，逐渐成为系统发生分析的主要数据来源。分子系统发生学（molecular phylogenetics）就是在这种背景下诞生的，是研究利用各种分子性状构建基因、个体、种群和物种之间系统树或进化网络的原理和方法的学科。

分子数据的使用导致了系统发生研究的革命。在 20 世纪 80 年代后期，由于保守引物的 PCR 扩增和 DNA 测序技术的应用，使得系统发生分析可利用的同源位点（即性状）数量达到 500 个，有的甚至超过数千个，与此前几十个、最多上百个形态特征相比，大大地增强了解决系统发生推论的数据力度。此时，一些线粒体基因和 rDNA 成为最广泛应用的标记，其中编码 SSU rRNA 的基因识别出了作为生命树的第三分支的古细菌（Archaea）。随着更多基因标记，尤其是大量单拷贝核基因的使用，基于单个基因推论的系统发生关系之间的冲突逐渐显露。而且，来自单个基因的信息经常不足以对系统发生的节点提供坚实的统计学支持。所以，自 20 世纪 90 年代以来，多基因数据逐渐成为分子系统发生研究的主流。

目前，成千上万个物种的全基因组序列信息已经通过新一代的高通量测序技术产生，并由此产生了一个新的分支学科——谱系基因组学（phylogenomics），就是在基因

组水平上进行系统发生研究。谱系基因组学将基因座位的进化作为一种随机过程看待，将分子水平的基因座位和序列位点进化模型及群体历史过程整合在一起，分析基因树和物种树之间的关系，引发分子系统发生学思想的又一次革命。基因组学数据增加了用于系统发生学分析的性状数量和类型，期望能够减少先前由于序列或基因取样偏差造成的系统发生推论误差。

分子系统发生学已经成为当前生物学研究的核心领域。根据 SCI Web of Science 引文数据库统计，到 2009 年底已经有 30 000 多篇关于系统发生分析的论文，并且每年以 3000 篇的速度增加（Pagel and Meade，2008）。Rokas 和 Carroll（2006）估计世界范围平均每天发表 15 棵系统树。最近发起的重建生命之树计划和 DNA 条形码计划是生物学历史上能够与基因组计划媲美的生物学大科学项目，加之廉价而快速的新一代高通量测序技术引发的全基因组测序的普及，如人类千人基因组计划、宏基因组学（metagenomics）、脊椎动物基因组 10K 计划和昆虫基因组 5K 计划等，将极大地推进分子系统发生学的研究。

分子系统发生学数据的增加速度很快，目前 NCBI 核苷酸数据库有序列记录的物种数超过 30 万种。过去 5 年 GenBank 的物种数以每年约 1.7 万种的速度增加，也就是 170 万种已描述物种中，每年约有 1%的物种被进行至少一个基因的测序。即便如此，至少含有一条分子序列的生物体只占全部已知物种的 17%左右。而在系统发生信息数据库 TreeBASE 中，目前只录入了 2000 多项研究的 5000 多棵系统树，包括100 000个类群（http：//www.treebase.org/）。因此，实现重建生命之树的宏伟计划还有漫长的路要走。

1.2 系统发生关系的含义

不同生物学家对系统发生概念的认识和理解有所不同。生物之间在系统发生学上的相关性称为系统发生关系（phylogenetic relationship）。生物之间存在着各种各样的相互关系，系统发生关系只是其中最重要的关系之一，其他的关系还包括表征的（phenetic）、分支的（cladistic）、时序的（chronistic）、遗传的或亲缘的（patristic）和相互作用（interaction）关系等，这些复杂的关系从不同的角度反映了生物之间的相关性。

1.2.1 表征关系

表征关系是不考虑进化关系，仅以所有可利用性状为基础的全面相似性程度排列的关系。Sneath 和 Sokal（1972）将表征关系定义为“在所研究的机体表型特征基础上的相似性”。以表征关系为基础的分类学研究称为表征分类学（phenetics），根据生物表征总体相似性为依据获得的有机体之间的关系图解称为表征图（phenogram）。表征分类学认为有机体之间的演化关系是无法弄清楚的，因而表征图不需要代表机体之间的演化关系。根据表征图显示的类群之间的聚类关系就可以直接转化为分类体系。

1.2.2　分支关系

分支关系指物种或类群之间与共同祖先相对近度（relative recency）的关系。以分支关系为基础的系统学研究称为支序系统学（cladistics）或系统发生系统学（phylogenetic systematics）（Hennig，1966）。支序系统学派认为，判别系统发生关系远近的唯一标准是共同祖先的近度（recency of common ancestry），共同祖先关系可以通过性状的分布分析来发现，支序系统学派将性状分为祖征（plesiomorphy）、共享祖征（symplesiomorphy）、衍征（apomorphy）、共享衍征（synapomorphy）和自裔衍征（autapomorphy），认为只有共享衍征才是共同祖先的证据，共享祖征及由趋同进化和平行进化形成的相似性（同型性状）均不能作为共同祖先的证据。通过共享衍征推论的有机体分支关系的树状图称为支序图（cladogram）。支序图的纵轴仅表示分支发生的相对时间，图上的二叉分支节点代表一次物种形成事件。

支序图只是关于共享衍征分布的陈述（图 1-1），而不是系统发生关系的陈述，要将支序图转化成系统发生关系还需要进一步对进化过程作出假设。支序图上的分类单元（无论是现存种还是化石种）总是在末端分枝，而在系统树上必须明确分类单元的祖裔关系。图 1-2 中左框的分支图（（A，B），C）就可以解释为右框中 6 种不同的系统树。

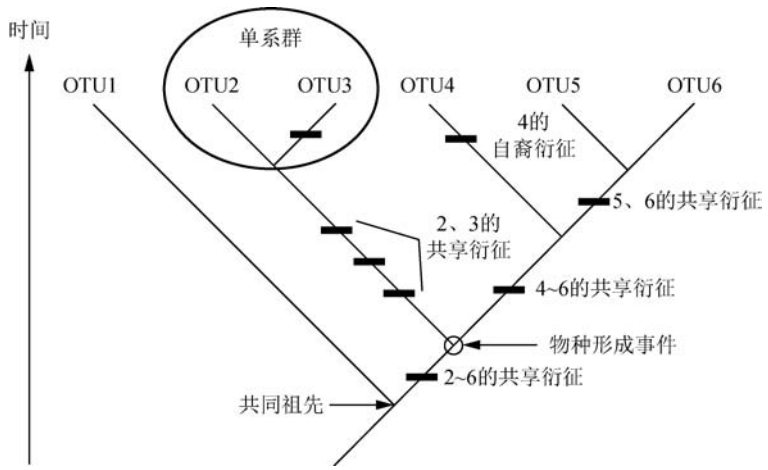


图 1-1　支序图上的性状类型

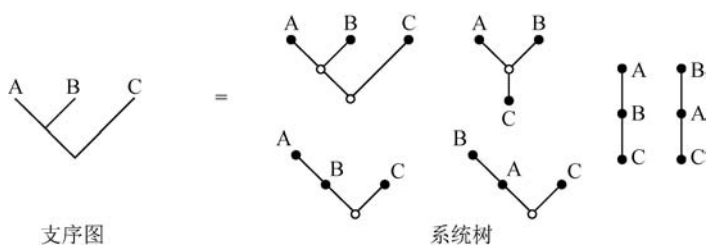


图 1-2　支序图与系统树的区别

支序系统学认为系统发生关系是生物之间最核心的关系，所有的分类学体系都必须建立在系统发生关系的基础上，也就是要求所有的分类单元必须是单系性的。已经建立了基于系统发生的分类学命名法规——Phylocode（www.ohio.edu/phylocode/）。

1.2.3 遗传关系

遗传关系是生物在遗传组成方面的关系，在群体遗传学中采用遗传相关性系数（coefficient of genetic relatedness）来度量，在种上阶元之间采用亲缘距离（patristic distance）来度量。亲缘距离是指在传代线内发生的遗传变异数量，表现在标度系统树上两个物种经过其共同祖先节点的所有通径的分枝长度之和。在分子系统树上，亲缘距离实际上等价于它们从共同祖先分歧以来在两个支系上发生的遗传改变，如果以基因组序列来度量的话，就等价于分支之间的遗传组成差异。

遗传关系起源于遗传物质的继承与传递，包括两种不同的遗传方式：垂直遗传和水平遗传。垂直传递是通过繁殖方式进行的，在有性生殖群体内个体之间的遗传关系是一种网状关系（特称为 tokogeny）。垂直遗传包括双亲遗传（如常染色体遗传）、父系遗传（如 Y 染色体遗传）和母系遗传（如线粒体基因组遗传）三种不同的方式。双亲遗传标记是生物之间的主要遗传标记，可以用于研究生物主要遗传组成的演化历史；父系遗传的标记可以推论父本谱系的历史；母系遗传标记可以推论母本谱系的历史。三类垂直遗传标记都可用于类群系统发生关系的重建。

水平遗传的主要方式是基因水平转移（horizontal gene transfer，HGT），也称为侧向转移（lateral gene transfer，LGT），是指在不同物种之间进行的遗传物质的交流。LGT 类似于物种内部的重组，但种内不同染色体/DNA 分子的重组是共享基因库分子之间的混合，虽然也产生了不同进化历史的 DNA 分子的嵌合体，但重组分子对推论物种之间的系统发生关系影响不大，因为它们的遗传传递方式与分支发生方式一致。而 LGT 是跨越生殖隔离的 DNA 分子之间的混合，是与分支发生关系毫无关联的遗传传递，因而会对系统发生关系产生误导。水平基因转移事件作为推动物种进化的重要动力，在生命起源和进化的早期发挥了十分重要的作用，后来也对原核生物基因组的进化产生了深刻的影响（图 1-3），但相对来说在真核生物中发生的规模不大。

垂直遗传和水平遗传的概念在细胞形态的生物之间是很容易区分的，因为垂直遗传是通过细胞膜体系和遗传系统的双重复制及分裂过程完成的，而水平遗传仅仅是部分遗传物质的整合。因此，也有人将这种以细胞传承为基础的垂直遗传系统发生称为细胞之树（tree of cell）。

以遗传关系为基础的系统学研究即分子系统学，从带遗传信息的分子数据建立的树状图称为分子树（molecular tree）或基因树（gene tree）。基因树可以是群体内部取样的等位基因之间的系统发生关系，特称为基因谱系（gene genealogy），反映的是等位基因的起源和演化关系；也可以是基因组内部一个基因家族成员之间的系统发生关系，反映的是基因重复事件；还可以是不同物种的直系或并系同源基因之间的系统发

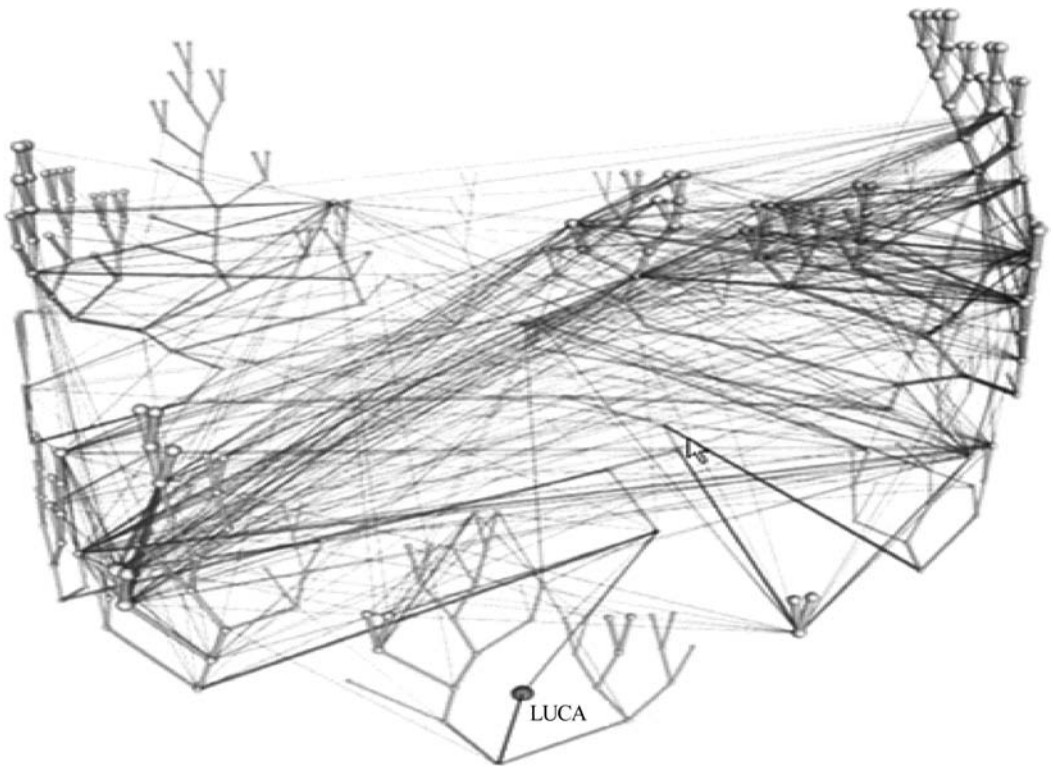


图 1-3　显示垂直遗传和水平遗传关系的生命之树鸟瞰图（Kunin et al.，2005）（见彩图）
圆点表示最近共同祖先 LUCA，垂直方向的粗分支表示系统发生关系，垂直分支之间的细线条表示基因水平转移，下方中间的分支为古细菌，两侧分支为真细菌

生关系，反映的是基因重复与物种形成双重进化事件。

基于垂直遗传分子标记构建的基因树可以转化为物种树，而根据水平遗传分子标记构建的基因树就不能直接转化为物种树。所以，只有垂直遗传关系的基因才能表达物种之间的系统发生关系。

1.2.4　系统发生关系

广义的系统发生（phylogenetic）或种系发生（phyletic）关系包括任何生物实体（基因、个体、群体、物种和种上阶元）的起源和演化关系，而狭义的系统发生关系仅指物种和种上单元的起源和演化关系。对狭义的系统发生关系的含义有不同的看法，有些人认为系统发生关系应是包括以上三种关系的总和，有些人则将系统发生关系仅看成是分支关系或遗传关系。在此，我们将狭义的系统发生关系定义为由分支发生（cladogenesis）产生的存在于任何支系（lineage）之间的祖裔关系和姐妹群关系。因此，系统发生关系实际上是由垂直遗传构成的分支关系，而分支发生实际上是连续的物种形成过程。在这个连续的过程中，永恒存在的仅仅是作为复制模板的细胞膜系统

和遗传系统，个体只是作为这两个复制模板的中间载体而短暂存在。除这种分支关系外，系统发生关系还包括少量的由物种杂交形成产生的网络关系。

自达尔文以来，系统发生关系被认为是树状分支的，表示有机体之间系统发生关系的树状图解称为系统树（phylogenetic tree 或 phylogram）或进化树（evolutionary tree）。传统上，系统发生关系通过寻找共同祖先及祖裔关系来重建。

由于现存的所有物种都是由共同祖先进化形成的，因此，系统发生关系的远近程度是一个相对概念。如果两个谱系享有比其他谱系更近的共同祖先，则这两个谱系相互之间的系统发生关系更接近且远离其他谱系。

由于灭绝和分类单元取样不完整等原因，系统发生关系一般无法重建完整的生物进化历史，而是简化的进化历史，是一种对进化历史的假设，但随着研究的深入可无限逼近进化历史。

以上四种关系是生物之间最核心的关系，它们之间在概念上的区别可以通过熟悉的爬行类和鸟类进化关系的树状图解来说明（图 1-4）。图 1-4 中显示出表征关系、分支关系和亲缘距离三种关系度量方法之间的矛盾。蜥蜴和鳄鱼的形态特征最接近，表征关系最近；鳄鱼和恐龙之间在系统树上的通径长度最小，故亲缘距离最近；恐龙和鸟类是拥有最近共同祖先的类群，它们之间的分支关系最近。

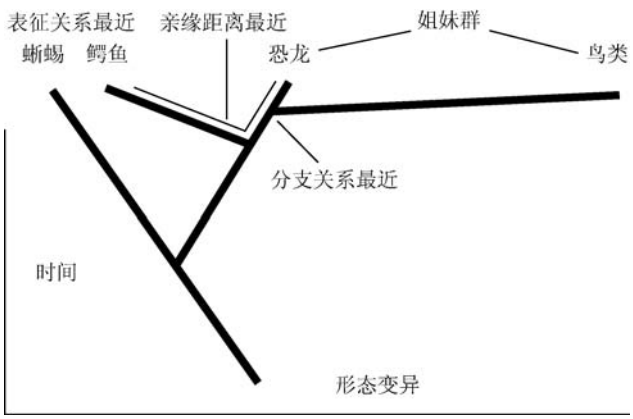


图 1-4 爬行类和鸟类进化关系示意图

（http://www.mun.ca/biology/scarr/Phenetic_Patristic_Cladistic.html）

1.2.5 年代关系

年代或时序关系在进化时间标度上标示的有机体之间的关系，亦即系统发生树的垂直轴（纵轴）上有机体之间的关系，这样的图也称为时序图（chronogram）（图 1-5），这种关系对研究进化速率和进化趋势有一定的意义。

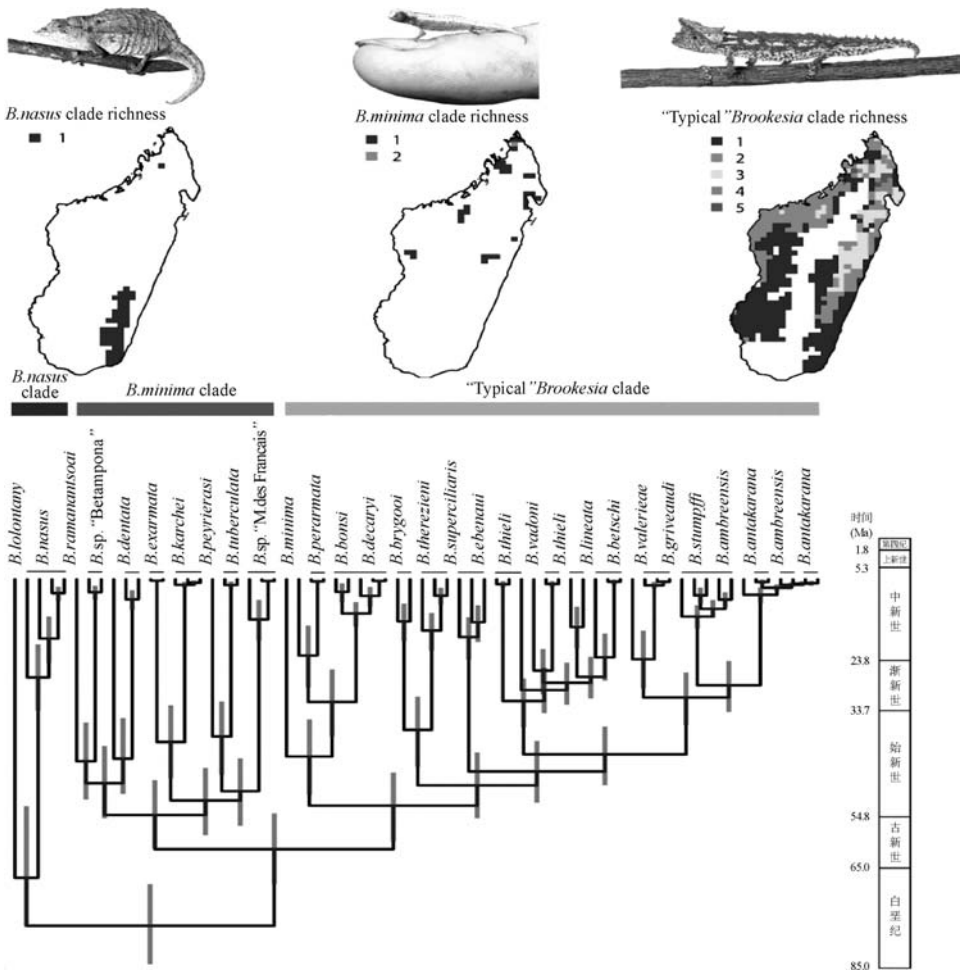


图 1-5　从三个线粒体基因和两个核基因估计的马达加斯加的变色龙三个单系群的时序图（Townsend et al.，2009）（见彩图）

节点上的线段表示估计时间的 95% 置信限范围，该图同时显示了马达加斯加岛上变色龙的物种丰富度的分布

1.2.6　地理分布关系

地理分布（chorologic）关系是指生物在进化过程中形成的空间分布上的关系。现存生物的地理分布式样，是在祖先区系被障碍分割或祖先区系扩散迁移以后形成的，这些过程与物种形成有密切关系，因此，生物的分布与分支发生紧密相关。替代生物地理学（vicariance biogeography）正是由此出发，强调生物的演化是与地球地理学的进化同步的，通过研究单系类群的分支关系（支序图）与其地理扩散轨迹（track）之间的关系而构建区域支序图（area cladogram）。表明不同单系类群的区域分支关系的相合性（一致性）称地理支序图（geographic cladogram）。分子生物地理学或谱系生物

地理学（phylogeography）主要利用分子标记研究近缘种及种内不同种群形成现有分布格局的历史原因和进化过程。谱系地理学不局限于解释现有物种或种群的分布状况，而是进一步探究其分布的起因，阐述其进化历程，分析地理种群在时间上和空间上的发展变化，从而重建生物区系的进化过程。

综上所述，有机体之间上述 6 种关系都可用一种分支图解来表示，这种图统称为树状图（dendrogram），表征图、支序图、系统树等都是树状图的一种，但表征图和支序图仅是有机体之间各种关系的一种，而系统树则是有机体多种关系的综合图解。系统发生研究、推论或重建（phylogenetic study，infernce，reconstruction）都是指通过对现存或化石生物性状的比较分析来建立生物类群的系统发生关系（系统树）。

1.3 分子系统发生分析的原理和假设

1.3.1 分子系统发生分析的原理

分子系统发生学是利用分子数据重建生物实体进化历史的一门学科，所以分子系统发生学的原理包括分子标记、生物进化和系统发生重建方法等方面的基本原理。分子标记是分子系统发生分析的依据，关于分子标记的性质、进化特点和表征方法将在《分子系统学》一书中详细介绍；系统发生分析方法（距离法、简约法、最大似然法、贝叶斯法、网络法等）主要基于数学与统计学的原理，是本书的核心内容；进化论是系统发生学的理论基础，这里就分子系统发生学上涉及的进化理论简述如下。

1. 进化的过程和型式

分子系统发生涉及分子和机体（organismal）两个水平的进化过程。机体水平的进化是新种形成过程，物种形成的模式可以区分为线系（anagensis）进化和趋异进化（divergent evolution）两种基本型式。线系进化是指在时间上世代延续的种，在进化过程中一个物种可以逐渐地演变为另一个种，即在一段地质时间中第二个种替代了第一个种，这样的种被称为时间种（chronospecies）。趋异进化也叫分支发生（cladogenesis），是指由同一祖先分支出两个和多个线系的进化型式（二歧和多歧树），是物种水平的基本进化型式；杂交是另外一种新种形成的方式，可以用系统发生网络表示。

机体水平的进化是以个体为单位承载和实施的，相互交配繁育的个体组成了群体，有基因流的群体构成了物种，而具有共同祖先的在不同时间形成的物种组成了属及其他种上单元。仔细考察机体进化过程（图 1-6），可以看出小进化和大进化是一个连续的过程。所有种上单元都是通过物种形成产生的，也就是说，系统树上的任何部分，都代表着那个时间的繁育群体。

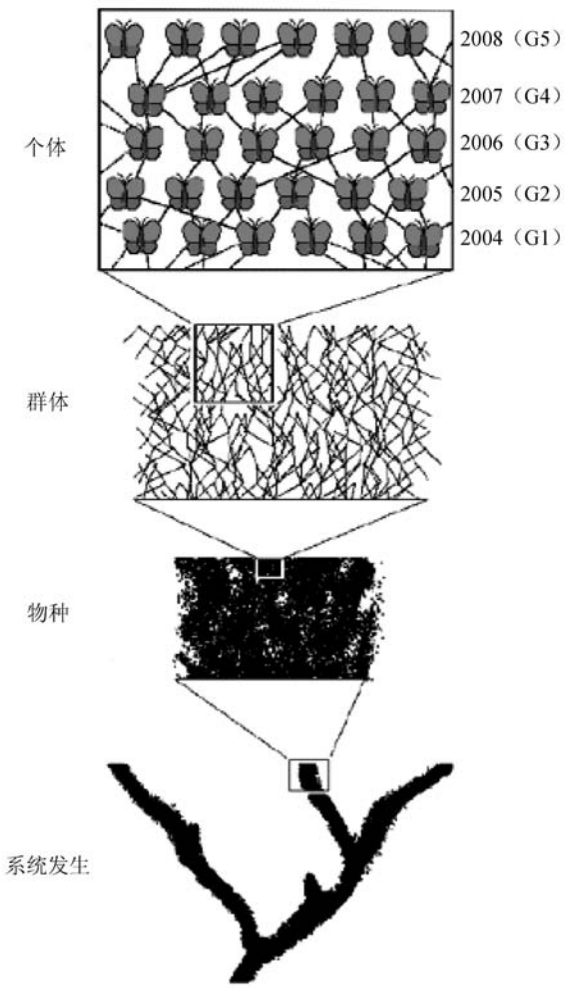


图 1-6　从个体到系统树分枝的详细遗传历史（Baum，2008）

2. 生物共祖原理

虽然是否存在地球上所有生命的最近共同祖先（last universal common ancestor，LUCA）仍然有激烈的争论，但毋庸置疑的是，所有现存生物类群都有远近不同的共同祖先，这种共同祖先不是一个个体，而是由一群能够繁育的个体组成的种群，这样的祖先称为一个生物类群的最近共同祖先（most recent common ancestor，MRCA 或 last common ancestor，LCA）。

地球上现在存在的所有物种，如果在时间坐标上溯本求源，就会发现它们都曾一个接一个地与其他物种互相汇合。以人类为例，在大约 700 万年前，人类与黑猩猩是共祖的，而几乎在同时，大猩猩与黑猩猩也是共祖的。再倒退数百万年，非洲古猿与猩猩也是共祖的。在比这更早的年代，人类的祖先是与长臂猿是共祖的。在更久远的过去，人类的祖先曾经与另外一些主要的哺乳动物类群，如啮齿动物、猫、蝙蝠、大

象等共祖。再往前，人类的祖先又与各种爬行动物、鸟类、两栖动物、鱼类、无脊椎动物、单细胞动物、真菌、植物、古细菌和细菌共祖。这样，以目前的系统发生知识，可以识别出人类在生命之树上与其他生物共有 39 个共祖节点（图 1-7）。

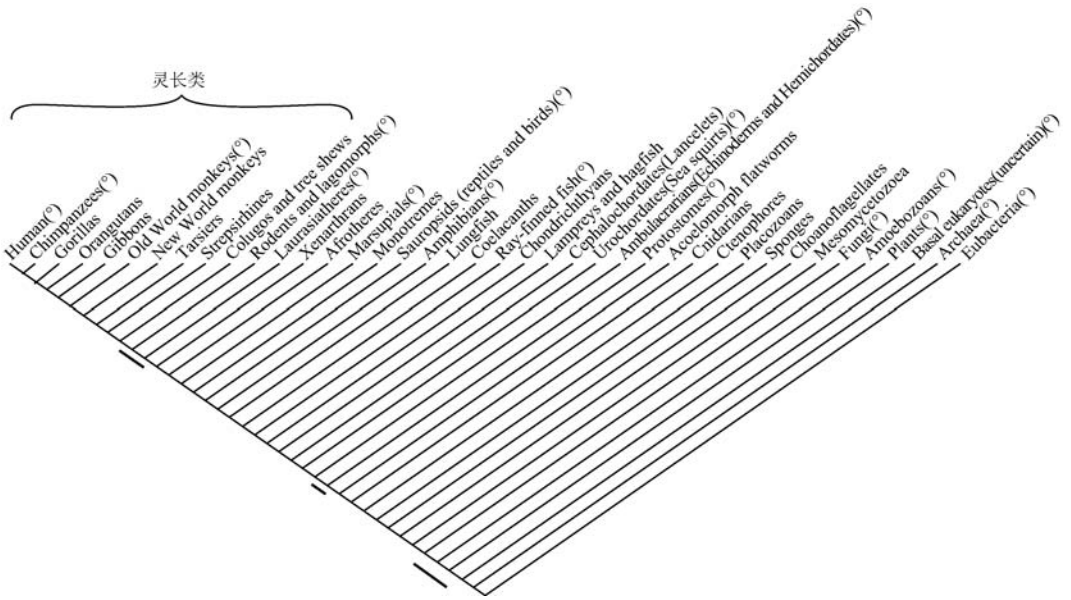


图 1-7　导致人类起源的所有分支发生过程的系统树（Castresana，2007）

“◦”表示该类群中至少有一个物种的基因组序列被测出

生物有共同祖先的特性称为单系性，由具有共同祖先的所有生物组成的一个类群称为单系群，系统发生分析要求分类单元必须是单系群。

3. 分子进化原理

遗传物质的复制是最基本、最核心的生物学过程，保持在一定阈值下的复制错误是突变的主要源泉，是遗传漂变和自然选择驱动的进化之所以发生的前提条件。生物的变异是与 DNA 复制（生物繁殖的前提）关联的，有繁殖必须有 DNA 复制，而有复制就有误差。复制误差导致 DNA 突变，这种突变如果发生在生殖细胞并遗传到下一代，就是变异。如果这个变异在群体中固定下来，并且它的频率逐渐增加，就是替换。

个体内部同源基因拷贝之间存在的遗传变异称为异质性（heteroplasmy）；种内个体之间存在的遗传变异称为多态性（polymorphism）；物种之间表现的遗传差异称为遗传分歧（genetic divergence）；不同物种继承它们共享祖先的多态性的不同形式称为支系分选（lineage sorting），或深度溯祖（deep coalescence），或跨种多态性（transpecies polymorphism）。

地球上的每一个生物个体都有自己的直接祖先，而其所携带的遗传物质则来自父母。由于 DNA 复制过程都伴随着突变，所以，所有生物都是父母遗传信息经过变异的后代。总的来说，遗传变异是一个小概率事件，可以期望大多数的替换发生在不同的位点上并能够在进化过程中保留下来。因此，一个支系上不同时间发生的遗传变异有不同的分布模式。越是早期发生的替换在所有后代中分布越广泛；越是近

期发生的替换，分布越局限于近期分歧的类群中。这种遗传变异的后代可以构成一个谱系（genealogy），在这个谱系上，不同世代的个体都有能够识别的印记，包括个体的、世代的和家族的，可以通过各自的特征把各个世代识别出来。因此，来自共同祖先的生物中不朽遗传物质的差异记录着这些生物分歧的历史。如果把一个物种的谱系一直回溯到单细胞时期或 LUCA 的话，就可以以同样的原理构建出地球上生物的生命之树。

4. 依据分子标记重建演化历史

按照生物学物种的定义，任何一个物种的所有成员共享基因库，因而，同一物种的所有基因都有潜在的机会相互构成一个基因组。当一个物种一分为二的时候，一个新的物种就诞生了，从此一个基因组被分为相互隔离的两个基因组。两个基因组在各自的种内通过性的重组互相混合。但是，一个物种分裂为两个物种之后，两套基因就不再互为伙伴，它们一般不会再在同一个躯体中相遇（杂交除外），两套基因积累各自支系上的突变，这就是物种之间的分歧进化（divergent evolution）现象。

图 1-8 表示的是物种形成期间单倍体基因的基因谱系，当物种形成刚开始时，两个亚群体中的个体之间关系密切，基因谱系呈现复系性或并系性，当物种形成完成后，它们之间呈现相互的单系性，支系分选也结束了，这个过程需要 4Ne 个世代（Ne 为有效群体大小）。

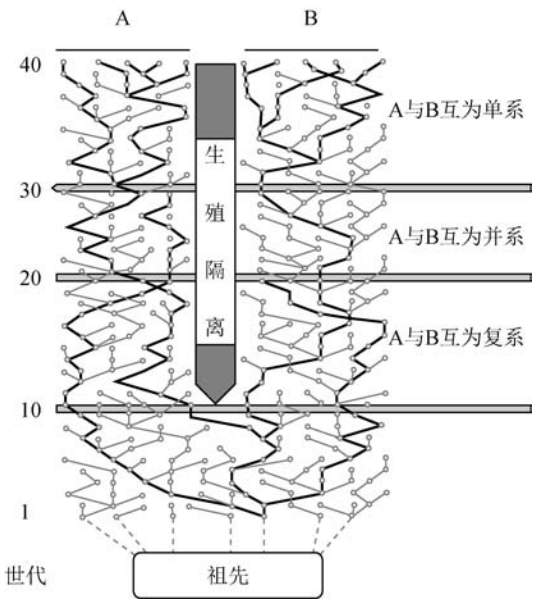


图 1-8　种群 A 和种群 B 在第 11 代开始建立生殖隔离后的支系分选过程（Avise，2000）
在开始时两个群体是复系性的，接着为并系性，之后成为相互单系性

在假定无自然选择、群体无限大和随机交配等理想状态下，DNA 和蛋白质序列的进化仅由替换速率和时间两个因素决定，这就是 DNA 和蛋白质序列进化分析的原理。

5. 从基因树到物种树

从序列数据中构建物种树有两种可能的误差：其一是，如果序列的数据结构杂乱或系统误差原因将导致不正确的基因树；其二是，即使基因树能正确地从序列中构建，基因的深度溯祖、基因重复、基因水平转移等现象也可导致基因树与物种树不一致（图 1-9）。

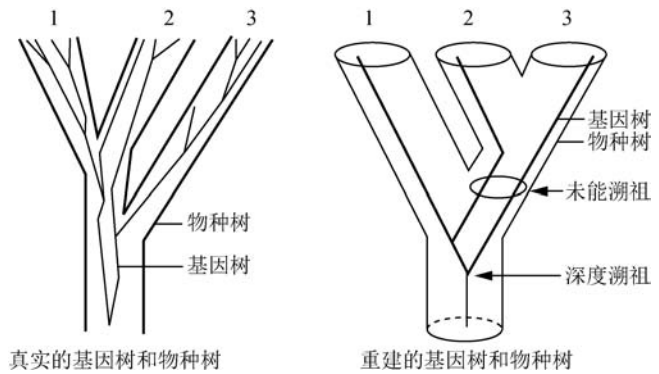


图 1-9　真实的基因树和物种树（物种 1 与物种 2 为姐妹群），
重建的基因树与物种树（物种 2 与物种 3 为姐妹群）

基因组水平上不同基因树之间的关系研究显示，在各种不同分类阶元水平，基因树之间都存在广泛的不相合性。最近分化的亚种/物种之间的不相合性主要是由于祖先多态性的支系分选；早期分歧的类群之间的不相合性主要是由于性状同型、基因重复/基因丢失和 LGT 等。

不论基因树的不相合性程度如何，对于一个特定的类群，其进化历史是唯一的，即物种树是唯一的。当所构建的多个基因树之间出现不相合性时，实质上是体现了部分（或者全部）的基因树没能正确反映物种进化关系。

Maddison（1997）认为每个基因树都是物种树的一部分，所有的基因树形成基因历史的“云团”（cloud），即物种树应该看成是所有基因树的统计分布。他提出了区分基因树之间不相合的可能的生物学过程的理论模型，包括支系分选（lineage sorting）、基因重复/基因丢失、基因水平转移/杂交。

目前，引起基因树与物种树不相合性的原因可以总结为以下 5 种（图 1-10）。

（1）并源基因引起的基因树的不相合性：并源基因的形成依赖于进化过程中的基因重复/基因丢失的动态发展。基因重复引起不一致的原因是并源基因抽样造成的，如果所研究的类群中有多拷贝时，常会发生这种情况。基因重复后又丢失拷贝造成直系同源基因和并系同源基因混淆这一因素不可忽视（图 1-10b）。

（2）基因水平转移：基因水平转移。引起的基因树与物种树的不一致性是直观的，可以将亲缘关系很疏远的类群聚合在一起（图 1-10c）。

（3）重组：重组导致基因序列偏离严格的二分歧系统发生模型（图 1-10d）。发生重组的序列片段之间可以拥有不同的进化历史。

（4）杂交：杂交在系统树上形成回环（图 1-10e），只能用网络表示。

（5）支系分选/深度溯祖：Zou 等（2008）进行的一系列统计分析表明，随机误差、系统误差及生物学因素中的基因水平转移、杂交/渐渗、基因重复后的拷贝丢失等均无法解释基因树冲突，而谱系分选则可有效地解释基因树冲突。更为重要的是，理论研究表明，当连续物种形成事件间隔很短（如辐射进化）时，支系分选的后果将十分严重，尤其是在祖先有效群体很大时。

支系分选现象的产生取决于物种形成过程中，祖先物种等位基因多态性的程度及持续时间。因而支系分选时常出现短时间内（世代数很少）形成的新种，或有大范围（有效群体很大，分布很广）祖先物种的新种中。由此推测，支系分选应该是发生在系统树的短而密的分支上的（图 1-10f）。

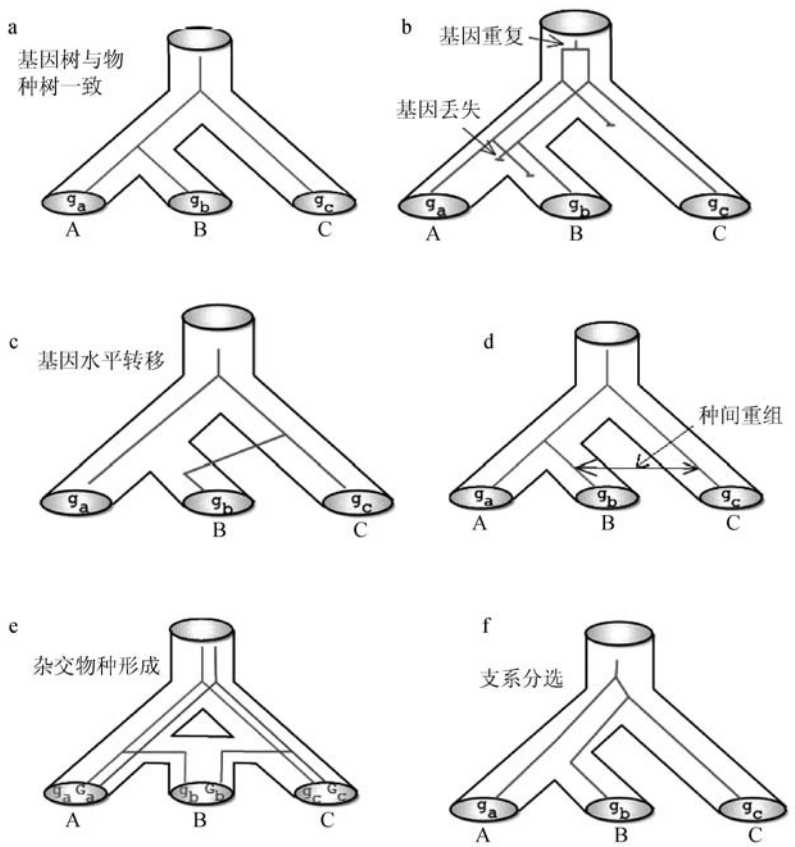


图 1-10　引起基因树与物种树不相合性的 5 种进化情景（见彩图）

a. 基因树与物种树一致；b. 由基因重复/基因丢失引起的基因树与物种树的不相合性；c. 由基因水平转移引起的不相合性；d. 由物种之间重组引起的不相合性；e. 由杂交物种形成引起的不相合性；f. 由支系分选/深度溯祖引起的不相合性

1.3.2　分子系统发生分析的假设

系统发生分析方法及其结果的准确性是与进化过程的模型密切相关的。如果一种

系统发生推论方法能够建立在一种完全已知进化过程的基础上，则这种方法可以避免分析中的系统误差，亦即如果有足够多的数据，依靠这种方法可获得正确的系统树。即使没有完全的进化知识，但如果建立的进化模型正确的话，该模型也能用于系统发生分析的基础而不受系统误差的影响。目前应用的各类分析方法都是建立在明确或隐含的关于进化的假设基础上，这些假设或模型能足够保证这些方法的有效性，但不同的方法有不同的假设。有的方法虽声称无需进化假设，但实际上仍隐含着进化过程的假设。

分子系统发生学对分子数据的假设通常包括以下几个方面。

（1）进化过程表现为二分枝式的树状结构，排除了多分枝式和网状进化型式，而实际上这两种型式是客观存在的。目前有些方法也能用于建立多分枝式进化树和网络。

（2）一个物种内部的所有性状具有相同的进化历史。

（3）所有分析的分类单元都是单系性的。

（4）物种鉴定和序列测定正确无误。

（5）基因序列是同源的。

（6）序列比对正确，即不同序列的同一个位点都是同源的。

（7）分类单元取样充足或有代表性。

（8）基因或序列数据有足够的信息可以解决感兴趣的问题。

（9）样本序列是随机进化的。

（10）序列中的所有位点的进化都是随机的。

（11）序列中的每一个位点的进化都是独立的。

1.3.3 分子数据的优点

自 20 世纪 80 年代以来，分子数据迅速取代形态学特征，成为重建系统发生关系的基本数据。在分子系统发生研究的早期，对于分子和形态数据的相对优劣问题及不相合性有相当多的争论，现在看来这些争论大多是无意义的。大量的比较研究表明，形态变化和分子变化是各自独立的，是对不同进化压力反应的结果，遵循着不同的进化规律。在系统发生学实践上关心的问题是所研究对象表现出的变异是否适合于所要解决的问题，而不是谁优谁劣的问题。30 年来，分子系统学的大量研究结果并没有广泛地排斥由形态学数据建立的系统发生关系，而且将分子与形态结合起来比用单一方法可对生物的多样性作出更好的描述和解释。

与传统的形态学特征相比，分子数据具有以下特点。

（1）可以识别的同源性状有宽广范围，核心代谢途径上的基因，如核糖体蛋白质和 rRNA 基因在所有细胞形式的生命体中都存在，可以进行最宽广范围的系统发生重建。

（2）可以提供大量的性状，如人类基因组可以提供约 32 亿个核苷酸位点，约 3 万个基因，以及大量的稀有结构变异等。

（3）提供有用的特性，无论何时，只要得到新的序列数据，它就能提供一些有用的系统发生的信息。