

分子系统发生学

黄 原 编著



科学出版社



分子系统发生学

第二版



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 章对复杂数据集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>

北京通州皇家印刷厂印刷

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、卢慧薏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

前言

第 1 章 系统发生学概论	1
1.1 系统发生与系统发生学	1
1.2 系统发生关系的含义	2
1.2.1 表征关系	2
1.2.2 分支关系	3
1.2.3 遗传关系	4
1.2.4 系统发生关系	5
1.2.5 年代关系	6
1.2.6 地理分布关系	7
1.3 分子系统发生分析的原理和假设	8
1.3.1 分子系统发生分析的原理	8
1.3.2 分子系统发生分析的假设	13
1.3.3 分子数据的优点	14
1.4 分子系统发生学的方法论	15
1.5 分子系统发生学的发展历史	16
1.6 系统发生分析的策略与步骤	18
1.7 分子系统发生学的文献资源	20
1.7.1 分子系统发生学期刊	20
1.7.2 分子系统发生学领域主要专著和教科书	20
1.8 分子系统发生学的成就和问题	21
第 2 章 系统发生分析基础	23
2.1 分子进化基础	23
2.1.1 分子进化的动力	24
2.1.2 分子进化的中性理论	27
2.1.3 溯祖理论	29
2.2 系统发生分析的分类学基础	31
2.2.1 系统发生与分类学的关系	31
2.2.2 分类阶元的系统发生意义	32
2.3 性状和性状分析方法	35
2.3.1 性状的分类	36
2.3.2 关于性状的基本假设	36
2.3.3 性状进化分析方法	37
2.3.4 性状的加权	39

2.3.5	性状的同源	39
2.3.6	性状的同型	43
2.4	系统发生分析的数学基础	44
2.5	系统发生分析的统计学基础	45
2.5.1	概率分布	45
2.5.2	系统发生的统计学检验	45
2.5.3	零假设与零模型	46
2.5.4	常用检验方法	46
2.5.5	随机数据及其在系统发生中的应用	48
2.6	理论系统发生学	49
2.7	模拟系统发生研究	50
2.7.1	系统树的模拟	50
2.7.2	序列的模拟	51
2.7.3	系统发生模拟研究的优势	51
2.8	系统发生分析的算法	52
2.8.1	精确算法	52
2.8.2	启发式算法	53
第3章	系统树	58
3.1	系统树的概念和含义	58
3.2	系统树的要素	58
3.2.1	系统树的拓扑结构	59
3.2.2	系统树的节点	59
3.2.3	系统树的分枝和分枝长度	59
3.3	演化历史与系统树的完整性	60
3.4	系统树表达的信息	61
3.5	系统树概念和表达形式的发展	62
3.6	系统树的类型	67
3.6.1	树状图与网状图	67
3.6.2	有根树和无根树	68
3.6.3	标度树与未标度树	70
3.6.4	基因树和物种树	70
3.6.5	基础树和合一树、源树和超树	71
3.6.6	期望树与实际树	73
3.6.7	普适生命树与完全树	74
3.6.8	二歧树和多歧树	74
3.6.9	系统树的表示形式	75
3.7	系统树的数学描述	79
3.7.1	系统树各部位的名称	79
3.7.2	二分树及其表示方式	79
3.7.3	二歧树的性质	80

3.8	系统树的赋根方法	82
3.9	系统树的生物学描述和解释	86
3.9.1	描述系统树的基本术语	86
3.9.2	系统树的分类学解释	87
3.9.3	系统树的进化解释	89
第4章	系统发生信息学	91
4.1	系统发生信息学概述	91
4.2	系统发生信息学研究内容	92
4.3	系统发生数据文件格式	92
4.3.1	数据文件格式	92
4.3.2	格式转换软件	99
4.3.3	系统树文件格式	101
4.4	系统发生分析软件	103
4.4.1	系统发生分析软件概述	103
4.4.2	系统发生分析软件的编程语言	104
4.4.3	系统发生分析软件的使用	104
4.5	PAUP* 软件及使用	109
4.5.1	PAUP* 软件的历史和版本	109
4.5.2	PAUP* 的安装	110
4.5.3	PAUP* 的功能	110
4.5.4	PAUP* 命令及操作	111
4.5.5	PAUP* 使用的一般步骤	113
4.5.6	ClustalX 和 PAUP* 连用	114
4.5.7	PAUP* 4 辅助软件	114
4.6	MEGA 5 软件包简介	115
4.7	DAMBE 软件包简介	116
4.8	SeaView 4 软件包简介	117
4.9	PHYLIP 软件包简介	118
4.10	系统发生的自动化分析工具	121
4.11	系统发生网络资源	121
4.11.1	系统发生软件目录	122
4.11.2	CIPRES	123
4.11.3	分子进化和系统发生专题研讨会	124
4.12	系统发生数据库介绍	125
4.12.1	系统发生知识数据库	125
4.12.2	生命之树数据库	126
4.12.3	Species 2000 数据库	127
4.12.4	NCBI 分类数据库	129
4.13	系统发生信息学展望	130

第 5 章 数据集准备与序列比对	131
5.1 分子数据的获得	131
5.1.1 自测数据	131
5.1.2 序列拼接	134
5.2 来源于公共数据库的分子数据	135
5.2.1 查看分类单元中已知基因序列分布的方法	135
5.2.2 查看一个分类单元被提交到 GenBank 中序列数量的方法	136
5.2.3 查看一个分类单元有序列记录物种数量的方法	137
5.2.4 数据库序列获取方法	137
5.2.5 批量下载序列的方法	139
5.2.6 比对序列数据库	140
5.3 序列比对	140
5.3.1 比对的概念和分类	140
5.3.2 序列比对的原理	141
5.3.3 序列比对算法	143
5.3.4 比对方法的分类	144
5.4 常用比对软件	144
5.4.1 ClustalX	145
5.4.2 T-Coffee	151
5.4.3 DIALIGN	152
5.4.4 MUSCLE 和 MAFFT	152
5.4.5 ProAlign	155
5.4.6 POA 和 ABA	157
5.5 比对软件的选择	157
5.6 不同类型的序列比对方法和策略	158
5.6.1 DNA 序列比对方法和策略	158
5.6.2 RNA 基因序列的比对方法与策略	159
5.6.3 蛋白质序列比对	162
5.7 比对结果的美化显示与格式转化	164
5.7.1 比对结果的美化和位点信息显示	164
5.7.2 比对结果的格式转化	165
5.8 比对与系统发生分析	166
5.9 数据集中空位、模糊区、多态位点和丢失数据的处理	167
5.9.1 数据集中空位的处理	167
5.9.2 模糊比对序列的处理	169
5.9.3 多态性状的处理	170
5.9.4 丢失数据的处理	171
5.10 多源数据集组装	171
5.10.1 公共数据库数据的组装	171
5.10.2 多基因数据的连接	172

5.11	序列管理与数据提交	173
5.11.1	序列管理	173
5.11.2	系统发生数据提交	174
第 6 章	数据集系统发生信号评估	176
6.1	系统发生数据信号描述	176
6.2	数据集质量的评价	177
6.2.1	数据集组成特征分析	178
6.2.2	替换型式分析	182
6.2.3	分子进化参数计算	187
6.2.4	替换饱和作图	192
6.3	系统发生信号与结构分析	200
6.3.1	序列数据系统发生信号强弱的评价	200
6.3.2	系统发生信号评估软件与方法	200
6.3.3	系统发生信号组成结构分析	205
6.4	系统发生数据探索与实验性分析	209
6.4.1	数据特征的探索	209
6.4.2	系统发生数据的实验性分析	209
第 7 章	进化模型及其选择	211
7.1	进化模型及其在系统发生分析中的作用	211
7.2	系统发生模型	211
7.3	形态性状进化模型	212
7.4	DNA 序列进化模型	213
7.4.1	DNA 序列上发生的进化改变	213
7.4.2	同质性模型	216
7.4.3	碱基组成异质性模型	222
7.4.4	Indel 模型	222
7.5	RNA 进化模型	223
7.5.1	结构 RNA 序列的进化特征	223
7.5.2	RNA 替换模型	224
7.6	蛋白质序列进化模型	225
7.6.1	蛋白质序列进化及建模	225
7.6.2	经验模型	226
7.6.3	机理模型	227
7.6.4	氨基酸频率变异和位点之间速率变异模型	228
7.6.5	混合模型	228
7.7	进化模型的选择	229
7.7.1	进化模型选择原理	229
7.7.2	LRT 检验法	229
7.7.3	AIC 信息标准法	231
7.7.4	贝叶斯信息标准法	232

7.7.5	贝叶斯因子法	233
7.7.6	决策论法	233
7.7.7	进化模型选择注意事项	234
7.8	DNA 进化模型选择	235
7.8.1	用 PAUP* 选择模型的 LRT 检验	235
7.8.2	DNA 模型选择软件	236
7.8.3	jModelTest 的使用	236
7.9	蛋白质进化模型的选择和使用	240
7.9.1	蛋白质进化模型选择概述	240
7.9.2	蛋白质进化模型选择软件 ProtTest3.0	241
7.10	进化模型参数的准确估计	244
7.11	混合模型和平均模型	245
第 8 章	距离矩阵方法	247
8.1	遗传距离的概念	247
8.2	距离数据的数学特征和生物学意义	247
8.3	将序列数据转化为距离的方法	250
8.3.1	未校正的遗传距离	250
8.3.2	校正距离的计算方法	253
8.3.3	最大似然法估计的校正距离	254
8.3.4	LogDet 距离	255
8.3.5	基因组距离	255
8.3.6	蛋白质遗传距离	256
8.3.7	计算遗传距离的软件	257
8.3.8	校正距离的选择和使用注意事项	259
8.4	距离矩阵方法概述	260
8.5	聚类分析方法	261
8.6	邻接法	262
8.6.1	邻接法原理	262
8.6.2	邻接法的算法	263
8.7	最小进化法	265
8.8	叠加树法	266
8.8.1	原理	266
8.8.2	平均距离法	267
8.8.3	转换距离法	268
8.8.4	最小平方法	268
8.8.5	其他叠加树方法	269
8.9	距离树可靠性评价	270
8.10	距离矩阵建树方法的比较及应用	270
8.11	距离矩阵法建树软件	271

8.11.1	PAUP* 4 距离法建树	272
8.11.2	MEGA5 的距离法	275
8.11.3	TREECON 使用	276
8.11.4	T-REX 软件使用	278
8.11.5	ProfDist 使用方法	280
第 9 章	简约法	283
9.1	简约性方法原理	283
9.2	简约法的分析过程	284
9.2.1	性状分布模式	284
9.2.2	性状优化	285
9.2.3	多态性内部节点祖先状态的重建方法	291
9.2.4	性状加权	292
9.2.5	最简约树搜索	293
9.2.6	简约树分枝长度和树长的计算	295
9.2.7	最简约树的选择	295
9.2.8	MP 树分支支持度计算	296
9.3	数据集中同型性状水平的分析和评价	297
9.4	简约法分析结果	299
9.5	简约性方法的优缺点	299
9.6	简约法分析软件	300
9.7	用 PAUP* 进行 MP 法分析	301
9.7.1	利用 PAUP* 进行简单简约法分析	301
9.7.2	加权简约法分析	306
9.7.3	PAUP* 限制树搜索	308
9.7.4	PAUP* 4 简约法的脚本命令运行	309
9.8	TNT 软件	310
9.9	WinClada 和 NOVA	311
第 10 章	最大似然法	313
10.1	最大似然法原理及其在系统发生分析上的应用	313
10.2	最大似然法建树原理	314
10.3	最大似然法建树过程	314
10.3.1	进化模型的选择及参数计算	315
10.3.2	系统树搜索方法	316
10.3.3	分枝长度的优化	318
10.3.4	似然值的计算	319
10.3.5	分支支持度计算	322
10.4	最大似然法建树结果的表示	323
10.5	最大似然法的优缺点	323
10.5.1	最大似然法的优点	323

10.5.2	最大似然法的缺点	324
10.6	最大似然法分析软件	324
10.6.1	PAUP* 4 的 ML 分析方法	325
10.6.2	PAUP* 与 ModelTest 联合运行选择进化模型	333
10.6.3	TREEFINDER 软件使用方法	334
10.6.4	TREE-PUZZLE 软件使用方法	336
10.6.5	RAxML	338
10.6.6	PhyML	339
10.6.7	MetaPIGA	340
10.6.8	IQPNNI	341
10.6.9	GARLI	342
第 11 章	贝叶斯系统发生推论法	343
11.1	贝叶斯系统发生分析原理	343
11.1.1	贝叶斯统计原理	343
11.1.2	贝叶斯系统发生推论法历史和现状	344
11.1.3	贝叶斯系统发生推论原理	345
11.2	贝叶斯分析过程	347
11.2.1	贝叶斯方法选择模型	347
11.2.2	先验概率的设置	348
11.2.3	马尔可夫链运行设置	349
11.2.4	提议、混合与接受	350
11.2.5	贝叶斯推论法克服局部优化的方法	351
11.2.6	评估和促进后验概率分布收敛的方法	351
11.2.7	影响系统树后验概率计算的因素	352
11.3	贝叶斯法运行结果汇总	353
11.4	贝叶斯推论法结果的分析、判断与表示	354
11.5	贝叶斯系统发生软件及使用	356
11.5.1	贝叶斯系统发生软件	356
11.5.2	MrBayes 3.2 使用方法	357
11.6	贝叶斯系统发生推论法优缺点	364
11.7	贝叶斯法与最大似然法的联系及区别	365
11.8	贝叶斯后验概率与自举支持度的关系	366
第 12 章	系统发生网络、超树和无比对方法	368
12.1	系统发生网络	368
12.1.1	网状进化型式与机制	368
12.1.2	系统发生网络的构建方法	368
12.1.3	网状图的构建软件	370
12.1.4	系统发生网络的应用	371
12.2	系统树的整合方法——超树	375
12.2.1	超树的概念	375

12.2.2	超树构建方法	375
12.2.3	超树方法的优缺点	376
12.3	无比对方法	377
12.3.1	比对和系统发生的联合估计方法	377
12.3.2	完全无比对方法	379
第 13 章	系统发生假设检验	381
13.1	系统发生假设检验概述	381
13.2	似然比检验	382
13.3	数据随机化检验	382
13.3.1	比较双树检验	383
13.3.2	PTP 检验和限制树 T-PTP 检验	383
13.4	配对位点检验	384
13.4.1	Templeton 检验	385
13.4.2	KH 检验	386
13.5	非参数自举法	387
13.5.1	SH 检验	388
13.5.2	AU 检验	389
13.6	参数自举法	389
13.7	贝叶斯统计检验法	391
13.8	PAUP* 执行的系统发生假设检验方法	391
13.9	CONSEL 软件使用	392
第 14 章	系统发生分析的可靠性与影响因素	394
14.1	系统发生分析方法的可靠性	394
14.1.1	方法可靠性的评价标准	394
14.1.2	系统发生分析方法的比较研究	395
14.1.3	不同构树方法的优缺点	397
14.2	系统树的可靠性	400
14.2.1	系统树的两类误差	400
14.2.2	系统误差和随机误差	400
14.2.3	检验系统树可靠性的统计学方法	401
14.3	随机误差及统计分析	402
14.3.1	评估分支支持度的方法	402
14.3.2	自举法	404
14.3.3	自减法	407
14.3.4	贝叶斯后验概率法	407
14.3.5	计算分支支持度的软件	408
14.4	系统误差的消除方法	409
14.4.1	系统误差的来源	409
14.4.2	导致系统误差的条件	410

14.4.3	系统误差的识别	410
14.4.4	系统误差的消除方法	411
14.5	系统发生分析疑难解答	411
14.5.1	有异常分支的系统发生	411
14.5.2	随机误差	412
14.5.3	分类单元抽样	413
14.5.4	序列长度与类型	414
14.5.5	序列比对问题	416
14.5.6	进化模型选择问题	417
14.5.7	建树方法的选择	418
14.5.8	搜索算法选择	418
14.5.9	分子进化速率对系统发生的影响	418
14.5.10	替换速率变异	419
14.5.11	碱基组成偏向性的影响	421
14.5.12	碱基组成异质性的影响	421
14.5.13	外群选择与系统树的赋根问题	422
14.5.14	谱系缺失的影响	423
14.5.15	数据缺失对系统发生分析的影响	423
14.5.16	基因水平转移	424
14.5.17	序列和位点同源关系	424
14.5.18	选择作用的影响	424
14.5.19	重组的影响	425
14.5.20	分支支持度低的问题	426
14.5.21	计算时间太长的的问题	427
14.5.22	总结	428
第 15 章	不同类型数据的分析策略	429
15.1	不同类型数据的特点	429
15.2	DNA 序列分析策略和方法	429
15.2.1	用 DNA 序列还是蛋白质序列	429
15.2.2	编码蛋白质 DNA 序列的分析	430
15.2.3	DNA 序列的加权简约法分析	431
15.2.4	DNA 序列的 ML 和贝叶斯法分析	434
15.3	蛋白质序列分析策略和方法	435
15.3.1	蛋白质序列数据的获得	435
15.3.2	必须使用蛋白质序列的情况	435
15.3.3	蛋白质序列的分析策略	435
15.3.4	蛋白质立体结构分析	439
15.4	RNA 序列分析策略和方法	440
15.4.1	RNA 序列数据的特点	440
15.4.2	rRNA 基因序列系统发生分析策略	440

15.4.3	rRNA 基因序列分析软件	442
第 16 章	复杂数据和困难系统发生的分析策略与方法	444
16.1	早期适应辐射的系统发生	444
16.2	近期发生过适应辐射的系统发生	448
16.3	存在长枝吸引问题的系统发生	450
16.3.1	长枝吸引现象	450
16.3.2	产生长枝吸引现象的可能原因	451
16.3.3	识别长枝吸引的方法	453
16.3.4	消除长枝吸引现象的方法	453
16.4	大数据集的系统发生	455
16.4.1	大数据集系统发生及其面临的问题	455
16.4.2	大数据集系统发生分析策略	455
16.4.3	大数据集的系统发生分析需要的计算机和软件	457
16.4.4	大数据集分析实例	458
16.5	碱基组成异质性数据集的分析	458
16.5.1	序列组成偏向性及其对系统发生分析的影响	458
16.5.2	碱基组成异质性数据分析方法	460
16.5.4	氨基酸组成异质性数据分析方法	461
16.6	种上与种下数据的联合分析	461
第 17 章	多源数据集分析策略和方法	465
17.1	多源数据集概述	465
17.2	数据集之间的不相合性及检验方法	466
17.2.1	不相合性的类型	466
17.2.2	数据集之间不相合性的原因	467
17.2.3	数据集之间不相合性的检验方法	469
17.3	多源数据集的分析策略	473
17.3.1	联合方法	473
17.3.2	分类学相合性分析	475
17.3.3	数据划分方法	476
17.4	多源数据集的划分分析实例	482
17.5	谱系基因组学方法	485
17.5.1	谱系基因组学	485
17.5.2	谱系基因组学分析策略	486
17.5.3	谱系基因组学分析方法	487
第 18 章	系统树的可视化、注释与应用	489
18.1	系统树的可视化	489
18.1.1	TreeView	491
18.1.2	Dendroscope	492
18.1.3	Mesquite	493
18.1.4	FigTree	494

18.1.5	MrEnt	494
18.1.6	2D 和 3D 曲面表示方法	495
18.1.7	iTOL	496
18.2	系统树的注释	497
18.2.1	分类学命名标注	497
18.2.2	分歧年代和地质时代的标注	499
18.2.3	重建祖先状态	502
18.2.4	性状进化	503
18.2.5	协同系统发生	504
18.3	系统树表达的信息及其应用	507
18.3.1	拓扑结构和分支长度	507
18.3.2	系统树的树形及应用	507
18.3.3	系统发生的不平衡性	509
18.3.4	系统树用于分析分歧速度	510
18.4	系统发生的应用	510
参考文献		511