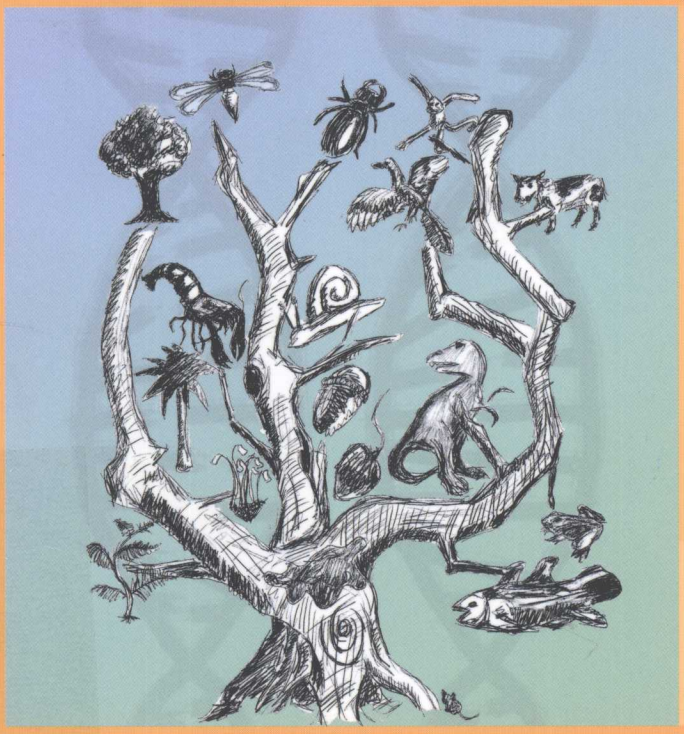
复旦大学进化生物学丛书

Computational Molecular Evolution

计算分子进化

〔英〕杨子恒（Ziheng Yang）/著
钟 扬 张文娟 梅 旖 王 莉 /译
杨 继 /校



復旦大學出版社
www.fudanpress.com.cn

复旦大学进化生物学丛书

Computational Molecular Evolution

计算分子进化

〔英〕杨子恒（Ziheng Yang）/著
钟 扬 张文娟 梅 旖 王 莉 /译
杨 继 /校

复旦大學出版社
www.fudanpress.com.cn

图书在版编目(CIP)数据

计算分子进化/〔英〕杨子恒著;钟扬等译. —上海:复旦大学出版社,2008.6
(复旦大学进化生物学丛书)
书名原文:Computational Molecular Evolution
ISBN 978-7-309-06042-3

Ⅰ. 计… Ⅱ. ①杨…②钟… Ⅲ. 分子进化 Ⅳ. Q75

中国版本图书馆 CIP 数据核字(2008)第 060310 号

图字:09 - 2007 - 106

计算分子进化

〔英〕杨子恒(Ziheng Yang) 著

钟 扬 张文娟 梅 旖 王 莉 译 杨 继 校

出版发行 复旦大学出版社 上海市国权路 579 号 邮编 200433
86-21-65642857(门市零售)
86-21-65100562(团体订购) 86-21-65109143(外埠邮购)
fupnet@fudanpress.com http://www.fudanpress.com

责任编辑 林 琳
出 品 人 贺圣遂

印 刷 上海复文印刷厂
开 本 787×960 1/16
印 张 24
字 数 405 千
版 次 2008 年 6 月第一版第一次印刷

书 号 ISBN 978-7-309-06042-3/Q·69
定 价 58.00 元

总 序

进化生物学在最近20年经历了一个快速发展和变革的时期，成为当今生命科学领域发展最为迅速的分支学科之一。这场变革一方面体现在我们对自然界生命起源和进化历史有了更深入的了解，在基因组数据大量积累和从分子水平对生物发育机制进行研究的基础上，将形态发生(morphogenesis)与发育调控基因结合起来，在一定程度上阐明了不同生物类群形态进化的分子基础和机制，从根本上改变了传统的研究思路和研究模式，促进了生命科学中遗传、发育和进化的统一；另一方面，进化生物学在近20年来极大地拓展了研究领域，向具有广泛的社会实用性的方向转变，尤其在揭示人类重大遗传疾病的分子基础、传染性疾病爆发与病原生物进化变异的关系，以及生物对环境变化的响应和适应机制方面显示出巨大的潜力，表现出显著的社会效应。

自20世纪50年代起，著名遗传学家和进化生物学家谈家桢院士就开始在复旦大学从事生物遗传变异和进化研究，为复旦大学进化生物学学科发展奠定了坚实的基础，把进化的思想和视角渗透、融汇到生命科学各个领域的教学和研究中，培养了一批从事生物进化和生物多样性研究的杰出人才。2003年，在谈家桢院士的积极倡导下，复旦

大学成立了我国第一个生态与进化生物学系；2006 年，又组建了跨学科的“进化生物学研究中心”，旨在充分发挥复旦大学的学科特点和优势，进一步加强培养从事生物进化和生物多样性研究的高层次人才，提升我国进化生物学研究的整体水平和实力。

编辑和出版《复旦大学进化生物学丛书》，系统介绍进化生物学的理论体系、研究方法和最新研究进展，以满足专业人才培养的需要，同时也是针对我国目前生物进化理论教育相对薄弱的现状，秉承“通达民情，化育人心”的教育传统，普及现代进化生物学知识，培养“进化意识”，使在处理人与自然关系的过程中能自觉地、理性地调整我们的价值观念和行为，促进自然和人类文明的协同进化。

金力

2008 年 5 月

中译本序

分子进化学(即在分子水平上对生物进化过程的研究)始于20世纪60年代,近20年来由于分子遗传学资料的迅速积累,其重要性日益突出,成为计算生物学和生物信息学等新兴学科的重要组成部分,是生命科学中最引人注目的领域之一。进化遗传学、分类学、生态学、基因组学等对分子进化方法的需求自不必说,即便是病毒学、生物化学和发育生物学等实验生物学领域现在也离不开应用分子进化的方法和软件来分析解释自己的资料。然而,分子进化分析涉及相当复杂的现代统计学理论和计算方法,对很多生物学家来讲无疑是一种挑战。

本书旨在为生物学研究者介绍分子进化领域的统计分析工具,着眼点放在对资料分析有重要价值的方法,使读者系统而深入地了解其全貌,避免分析数据时的"黑箱作业",对生物学问题无关紧要的方法则不多费笔墨。我相信具有一定生物统计基础的生物学研究者有能力阅读本书,或许对某些章节会感到吃力,但书中的计算实例可帮助对方法的理解,并提高用这些方法处理现代遗传学与进化生物学问题的兴趣。近年来,国内在分子进化研究领域有了长足的进步,分子实验技术已很完善并在基因组测序等领域取得了举世瞩目的成就,在分析和解释分子数据方面也日趋成熟,而本书中文版可望对此能有所促进。

我也希望本书能吸引国内的统计学家和计算机学家来从事分子进化领域的研究,与生物学家合作开发统计工具与计算程序。实际

上，统计学的诞生和发展与遗传、进化的研究有着很深的渊源。例如，相关和回归的概念是 Francis Galton 和 Karl Pearson(伦敦大学学院，UCL)在分析人类遗传资料时提出的；现代统计学的奠基人 R. A. Fisher 曾两次担任遗传学教授(UCL 和剑桥大学)，但从未任过统计学教授。Jerzy Neyman 曾描述分子进化研究是一个提供崭新统计学问题的源泉，事实也是如此，分子进化研究中一些根本性的统计学问题在统计学教科书中没能找到现成答案。例如，进化树可靠性的评价(6.4 节)、似然法在比较进化树时的统计学特性(6.2.3 节)、用似然法估计物种分化年代(7.3.5.3 节)以及用贝斯法* 估计进化树(5.6.4 节)等，对这些问题的深入研究也许会成为对统计学的贡献也未可知。

感谢复旦大学钟扬教授和杨继教授及其同事们花费大量的时间和精力翻译本书，可喜的是英文版中的一些错误趁此机会也得以更正。如果本书的中文版对国内分子进化研究的发展有所促进，我将甚感欣慰。

杨 玉

2008 年 4 月于伦敦

* Bayesian(自 Bayes) method 的传统中译为“贝叶斯方法”，并不符合英语发音习俗，故建议本书改译为“贝斯方法”。——作者注

前 言

分子水平的进化研究致力于两大问题：重建物种间的进化关系以及了解进化过程的动力与机制。前者属于系统学领域，传统上是用形态性状和化石来开展研究的。分子数据的广泛应用和简便易得已使其成为重建大部分物种类群系统发育关系中最常用的数据类型。后一个关于分子进化机制的问题则通过估计核苷酸和氨基酸的置换速率以及采用序列数据来检测突变与选择模型来进行研究。

过去几十年来，由于遗传序列数据的爆炸性累积、计算机软件和硬件的改进以及适用于解决人们关注的生物学问题的复杂统计方法的发展，上述两个研究领域都有了惊人的进展。种种迹象表明，这种进展，特别是在数据生成方面的增长，一定还会持续。系统发育分析已经进入了基因组时代，对包含数百甚至上千的物种或序列的大数据集进行分析已经成为很常见的工作。形态学与分子之间的争论在很大程度上已宣告结束，大多数研究者都能充分意识到两者的价值。有关简约性与似然性的哲学争论仍在进行，但渐趋缓和。在发展和实现有用的统计方法与模型方面已取得许多振奋人心的进展，并且已被应用到实际数据集的常规分析中了。

现在到了总结这一领域中方法学进展的时候。本书正是这样的一种尝试。我并未勉强自己去概括求全，这在目前没有必要。感谢 Joseph Felsenstein 的近作(2004)，其中已经讨论了几乎所有与系统发育相关的内容。不过，我所采取的角度是将分子进化分析(包括系统发育重建和进化过程推断)作为一个统计推断问题来处理(Cavalli-

Sforza and Edwards, 1967)，因而公认的统计方法(如似然法和贝斯法)是本书遵循的标准。启发式和近似方法也采用统计学的角度来进行讨论，因为它们比较简单、直观，所以在介绍更严格的方法之前，一般用它们来引导核心概念。本书也包括了一些算法实现方面的讨论，可供发展数据分析方法的研究者参考。

本书是为高年级本科生、研究生以及进化生物学、分子系统学和群体遗传学领域的研究人员编写的。希望用程序分析过自己数据的生物学家会觉得本书特别有用，因为通过本书可以了解这些程序和方法的原理。本书除强调基本概念外，还包括详尽的数学推导，因而也可供有意涉足计算生物学这一奇妙领域的统计学家、数学家以及计算机科学家阅读。

本书假定读者具有遗传学基础知识，如掌握 Graur 和 Li(2000)的第 1 章中的内容。读者也应具有基本的统计学或生物统计学知识。在一些章节中还需要微积分和线性代数知识。似然率和贝斯统计会通过简单的例子来介绍，然后应用于更为复杂的分析。打算系统而全面了解这些方法的读者应该查阅更多更好的概率论和数理统计方面的教材。例如，初级水平的读者可阅读 DeGroot 和 Schervish(2002)的教材，水平较高的读者则可看看 Davison(2003)、Stuart 等(1999)以及 Leonard 和 Hsu(1999)的著作。

本书结构如下：第 1 部分包含两章，介绍序列进化的马尔可夫过程模型。第 1 章讨论核苷酸置换模型以及成对序列间距离的计算。这也许是最简单的系统发育分析，我还将借此机会介绍马尔可夫链理论和最大似然法，这些会在之后几章中广泛应用。所以，本章对生物学背景的读者而言可能是最富挑战性的一章。第 2 章描述了氨基酸和密码子置换的马尔可夫过程模型，以及它们在计算两个蛋白质序列间的距离、估计两个编码 DNA 序列间同义置换与非同义置换率的作用。第 2 部分涉及系统发育重建方法。第 3 章简要讨论了简约法和距离法，似然法和贝斯法则放在第 4、第 5 章中讨论。第 5 章是 Olivier Gascuel 主编的《系统发育与进化中的数学》(*Mathematics in*

Phylogeny and Evolution)(牛津大学出版社 2005 年版) 中一章的扩展版本。第 6 章给出了比较不同系统发育重建方法的研究评论,并介绍了如何对系统发育树进行统计检验。第 3 部分综合评述了一些用系统发育方法来研究进化过程的应用,如检验分子钟和用分子钟估计物种分歧时间(第 7 章),以及用密码子置换模型检测对蛋白质进化产生影响的自然选择(第 8 章)。第 9 章讨论计算机模拟的基本技术。第 10 章则讨论该领域所面临的挑战和未来的前景。附录 C 对主要的系统发育分析软件包进行了简要评述。用星号(*)标注的部分是技术性的,可以略过不读。

本书中所用的数据集范例以及用于实现算法的一些 C 语言小程序均置于本书网页(http://abacus. gene. ucl. as. uk/CME/)。网页上还包括一个勘误表。恳请你将新发现的错误告知作者(z. yang@ucl. ac. uk)。

我要向一批同行们致谢,他们阅读了本书章节的早期版本并提出了建设性的意见和批评,他们是:Hiroshi Akashi (第 2、8 章)、Adam Eyre-Walker(第 2 章)、Jim Mallet(第 4、5 章)、Konrad Scheffler(第 1、5、8 章)、Elliott Sober(第 6 章)、Mike Steel(第 6 章)、Jeff Thorne(第 6 章)、Simon Whelan(第 1 章)以及 Anne Yoder(第 6 章)。特别感谢 Karen Cranston、Ligia Mateiu 和 Fengrong Ren,他们阅读了全书并提供了很多详细的建议。Jessica Vamathevan 和 Richard Emes 是我在试验一些晦涩段落时的"受害者"。当然,所有遗留下来的错误都归结于我。我还要感谢牛津大学出版社的 Ian Sherman 和 Stefanie Gehrig 启动该项目并给予宝贵的支持和自始至终的耐心。

杨 2006

2006 年 3 月于伦敦

目 录

第 1 部分 分子进化建模

第 1 章 核苷酸置换模型 …… 3
1.1 引言 …… 3
1.2 核苷酸置换和距离估计的马尔可夫模型 …… 6
1.2.1 JC69 模型 …… 6
1.2.2 K80 模型 …… 9
1.2.3 HKY85、F84、TN93 等模型 …… 11
1.2.4 转换/颠换比率 …… 17
1.3 位点间可变的置换率 …… 18
1.4 最大似然估计 …… 21
1.4.1 JC69 模型 …… 22
1.4.2 K80 模型 …… 25
*1.4.3 概形与积分似然方法 …… 28
1.5 马尔可夫链和广义模型下的距离估计 …… 30
1.5.1 广义理论 …… 30
1.5.2 广义时间可逆(GTR)模型 …… 33
1.6 讨论 …… 37
1.6.1 不同置换模型下的距离估计 …… 37
1.6.2 成对比较的局限 …… 37
1.7 练习 …… 38

第 2 章 氨基酸和密码子置换模型 …… 40
2.1 引言 …… 40
2.2 氨基酸替代模型 …… 40
2.2.1 经验模型 …… 40
2.2.2 机理模型 …… 44

2.2.3 位点间的异质性 …… 44
2.3 两条蛋白质序列间距离的估计 …… 45
2.3.1 泊松模型 …… 45
2.3.2 经验模型 …… 46
2.3.3 伽马距离 …… 47
2.3.4 例子:猫和兔的 p53 基因间的距离 …… 47
2.4 密码子置换模型 …… 48
2.5 同义和非同义置换率的估计 …… 49
2.5.1 计数法 …… 50
2.5.2 最大似然法 …… 58
2.5.3 方法比较 …… 61
* 2.5.4 对距离的诠释及滥用 …… 62
* 2.6 转换概率矩阵的数值计算 …… 68
2.7 练习 …… 70

第 2 部分 系统发育重建

第 3 章 系统发育重建:概述 …… 75
3.1 树的概念 …… 75
3.1.1 术语 …… 75
3.1.2 树之间的拓扑距离 …… 79
3.1.3 一致树 …… 81
3.1.4 基因树和物种树 …… 82
3.1.5 树重建方法的分类 …… 83
3.2 穷举式或启发式树搜索 …… 84
3.2.1 穷举式树搜索 …… 84
3.2.2 启发式树搜索 …… 85
3.2.3 分枝交换 …… 86
3.2.4 树空间的局部峰 …… 89
3.2.5 随机树搜索 …… 90
3.3 距离方法 …… 91
3.3.1 最小二乘法 …… 92
3.3.2 邻接法 …… 94
3.4 最大简约法 …… 95

3.4.1 简史 …… 95
3.4.2 对给定树计算最小变化数目 …… 95
3.4.3 加权简约法和颠换简约法 …… 97
3.4.4 长枝吸引 …… 100
3.4.5 简约法的假定 …… 101

第4章 最大似然法 …… 102
4.1 引言 …… 102
4.2 树的似然计算 …… 102
4.2.1 数据、模型、树及似然函数 …… 102
4.2.2 修剪算法 …… 103
4.2.3 时间可逆性、树根及分子钟 …… 107
4.2.4 缺失数据与对位排列间隔 …… 109
4.2.5 一个数值例子:猿类系统发育关系 …… 110
4.3 复杂模型下的似然计算 …… 111
4.3.1 位点间可变置换率模型 …… 111
4.3.2 多个数据集联合分析的模型 …… 118
4.3.3 非齐次与非稳定模型 …… 120
4.3.4 氨基酸与密码子模型 …… 121
4.4 祖先状态重建 …… 121
4.4.1 概述 …… 121
4.4.2 经验和等级贝斯重建 …… 123
*4.4.3 离散形态性状 …… 126
4.4.4 祖先重建中的系统偏差 …… 127
*4.5 最大似然估计的数值算法 …… 130
4.5.1 单变量最优化 …… 130
4.5.2 多变量优化 …… 132
4.5.3 树形固定时的优化 …… 135
4.5.4 树形固定时似然表面上的多重局部峰 …… 136
4.5.5 最大似然树搜索 …… 137
4.6 似然法的近似 …… 138
4.7 模型选择与稳健性 …… 139
4.7.1 LRT、AIC 和 BIC …… 139
4.7.2 模型的充分性与稳健性 …… 143

4.8 练习 …… 145

第5章 贝斯方法 …… 147
5.1 贝斯范式 …… 147
5.1.1 概述 …… 147
5.1.2 贝斯定理 …… 149
5.1.3 经典统计学与贝斯统计学的比较 …… 153
5.2 先验分布 …… 159
5.3 马尔可夫链蒙特卡罗算法 …… 161
5.3.1 蒙特卡罗积分 …… 161
5.3.2 Metropolis-Hastings 算法 …… 162
5.3.3 单一成分 Metropolis-Hastings 算法 …… 167
5.3.4 Gibbs 取样法 …… 167
5.3.5 Metropolis-偶联 MCMC(MCMCMC 或 MC^3) …… 167
5.4 简单建议及其建议比 …… 169
5.4.1 均匀分布的滑动窗口 …… 169
5.4.2 正态分布的滑动窗口 …… 169
5.4.3 多元正态分布的滑动窗口 …… 170
5.4.4 比例收缩和膨胀 …… 171
5.5 监测马尔可夫链及处理输出 …… 172
5.5.1 验证和诊断 MCMC 算法 …… 172
5.5.2 潜在尺度减约统计 …… 174
5.5.3 处理输出 …… 175
5.6 贝斯系统发育分析 …… 176
5.6.1 简史 …… 176
5.6.2 总体框架 …… 176
5.6.3 汇总 MCMC 输出 …… 177
5.6.4 贝斯法与似然法的比较 …… 179
5.6.5 一个数值例子:猿类系统发育关系 …… 181
5.7 溯祖模型下的 MCMC 算法 …… 182
5.7.1 概述 …… 182
5.7.2 θ 的估计 …… 182
5.8 练习 …… 184

第 6 章 方法比较与树的检验 …… 186
6.1 树重建方法的统计性能 …… 186
6.1.1 标准 …… 186
6.1.2 性能 …… 189
6.2 似然法 …… 191
6.2.1 与常规参数估计的不同之处 …… 191
6.2.2 一致性 …… 192
6.2.3 有效性 …… 193
6.2.4 稳健性 …… 197
6.3 简约法 …… 198
6.3.1 与病态似然模型的等价性 …… 199
6.3.2 与正常行为的似然模型的等价性 …… 202
6.3.3 假定与证明 …… 205
6.4 检验关于树的假设 …… 207
6.4.1 自展 …… 208
6.4.2 内枝检验 …… 211
6.4.3 Kishino-Hasegawa 检验及改进 …… 212
6.4.4 简约法分析中所用的指数 …… 214
6.4.5 例子:猿类的系统发育 …… 215
* 6.5 附录:Tuffley 和 Steel 的单性状似然分析 …… 216

第 3 部分 高 级 专 题

第 7 章 分子钟与物种分歧时间估计 …… 225
7.1 概述 …… 225
7.2 分子钟检验 …… 227
7.2.1 相对速率检验 …… 227
7.2.2 似然比检验 …… 229
7.2.3 分子钟检验的局限性 …… 230
7.2.4 离散指数 …… 230
7.3 分歧时间的似然估计 …… 231
7.3.1 全局分子钟模型 …… 231
7.3.2 局部分子钟模型 …… 232
7.3.3 试探式速率平滑方法 …… 233

7.3.4 确定灵长类分歧时间 …… 235
* 7.3.5 化石的不确定性 …… 237
7.4 分歧时间的贝斯估计 …… 247
7.4.1 总体框架 …… 247
7.4.2 似然计算 …… 248
7.4.3 速率的先验分布 …… 249
7.4.4 化石的不确定性与分歧时间的先验分布 …… 250
7.4.5 在灵长类和哺乳类分歧中的应用 …… 253
7.5 展望 …… 258

第 8 章 蛋白质的中性与适应性进化 …… 260
8.1 引言 …… 260
8.2 中性理论和中性检验 …… 261
8.2.1 中性与近中性理论 …… 261
8.2.2 Tajima 的 D 检验 …… 264
8.2.3 Fu 和 Li 的 D 检验与 Fay 和 Wu 的 H 检验 …… 265
8.2.4 McDonald-Kreitman 检验和选择强度估计 …… 266
8.2.5 Hudson-Kreitman-Aquade 检验 …… 268
8.3 经历适应性进化的谱系 …… 269
8.3.1 启发式方法 …… 269
8.3.2 似然法 …… 270
8.4 经历适应性进化的氨基酸位点 …… 272
8.4.1 三种策略 …… 272
8.4.2 随机位点模型下正选择的似然比检验 …… 274
8.4.3 处于正选择的位点鉴定 …… 276
8.4.4 人类主要组织相容性(MHC)基因的正选择 …… 277
8.5 影响特定位点和谱系的适应性进化 …… 280
8.5.1 正选择的分枝-位点检验 …… 280
8.5.2 其他类似模型 …… 282
8.5.3 被子植物光敏色素的适应性进化 …… 283
8.6 假定、局限与比较 …… 284
8.6.1 当前方法的局限 …… 284
8.6.2 中性检验与基于 d_N 和 d_S 的检验间的比较 …… 286
8.7 适应性进化的基因 …… 287

第 9 章 分子进化的计算机模拟 …… 293
9.1 简介 …… 293
9.2 随机数发生器 …… 294
9.3 连续随机变量的产生 …… 295
9.4 离散随机变量的产生 …… 296
9.4.1 离散均匀分布 …… 296
9.4.2 二项式分布 …… 297
9.4.3 广义离散分布 …… 297
9.4.4 多项式分布 …… 298
9.4.5 针对混合分布的组分法 …… 298
* 9.4.6 从离散分布中抽样的重影法 …… 299
9.5 分子进化的计算机模拟 …… 301
9.5.1 树形固定时序列的模拟 …… 301
9.5.2 生成随机树 …… 305
9.6 练习 …… 305

第 10 章 展望 …… 308
10.1 系统发育重建中的理论问题 …… 308
10.2 大规模和异质数据集分析中的计算问题 …… 309
10.3 基因组重排数据 …… 309
10.4 比较基因组学 …… 310

附录 …… 311
附录 A:随机变量函数 …… 311
附录 B: Δ 技术 …… 313
附录 C:系统发育分析软件 …… 316

参考文献 …… 318