声明

本电子书由中国农业出版社数字出版，相关权利归中国农业出版社拥有。读者、著作权人和（或）依法可以行使著作权的权利人如有疑问，请与中国农业出版社联系：

地址：北京市朝阳区麦子店街１８号楼

邮编：100026

电话：010-64194921　010-65005894

E-mail：lishanzhao@sina.com

中国农业出版社

需要全本请访问 www.ertongbook.com

遗传学数据分析

——群体遗传学离散型数据分析方法

[美] Bruce S. Weir 著
徐云碧 王志宁 俞志华 译
朱 军 校

中国农业出版社

遗传学数据分析

——群体遗传学离散型数据分析方法

〔美〕 Bruce S. Weir 著
徐云碧 王志宁 俞志华 译
朱 军 校

中国农业出版社

需要全本请访问 www.ertongbook.com

据作者提供的修正稿以及我们的教学实践对原著作了多处修正。为了方便读者学习，我们将原作以 Fortran/77语言编写的程序全部列出。此外，在教学中，我们还用 Quick BASIC 语言为几乎全部的习题编写了计算程序，可以提供给有兴趣的读者和教师。

在本书出版之际，我们对浙江农业大学申宗坦和季道藩两位教授的关心和帮助，表示感谢。

由于译者水平所限，本书不免有翻译不当之处。我们热忱欢迎读者批评指正。

译校者

一九九四年六月于浙江农业大学

原　序

本书阐述了离散型群体遗传数据的统计分析方法。这些数据通常是一些计数，可以是基因型也可以是碱基类型。而由度量诸如株高或产量之类遗传性状所得的数量遗传数据，则不属本书所述范围。

群体遗传学家正面临着分析新型遗传数据的令人兴奋的任务，现在似乎是评述一些适合于分析限制性片段长度多态性和DNA序列的统计技术的时候了。本书可望表明，即使是新型的数据，也能用R. A. Fisher，J. B. S. Haldane和Sewall Wright在早期研究中发展起来的，并在此后为许多工作者所扩展的统计技术来处理。我希望以此填补20年前Regina Elandt—Johnson所著《遗传学中的概率模型和统计方法》一书中精彩的处理方法与当前科技文献之间存在的空缺。我力图拓宽涉及面，并在适当处参阅了一些更加专门的书籍。例如，读者也许要阅读Jurg Ott的《人类遗传连锁之分析》一书，以便详细了解连锁、尤其是人类连锁的估计，以及Masatoshi Nei的《分子进化遗传学》，以便全面合理地分析分子遗传数据。期望本书能成为一本用传统方法分析哈迪—温伯格及连锁不平衡，以及描述群体结构和估计遗传距离的指导书。本书还向读者介绍了一些构建进化树的方法，以及DNA或蛋白质序列排列和比较中所遇到的一些难题。

撰写有关分子遗传数据分析方法的书籍，好比是瞄准一个移动着的靶子。随着人类基因组计划中数据的积累，日趋复杂的分析方法正迅速涌现，并且这些方法也确实需要。如果能够收到有关未来版本中应包括的新论题的建议，以及告知本版中的错误或不明确之处，我将会很高兴由于许多人认真仔细的审阅，才使本书得以较为完善。我衷心感谢Bill Hill (Edinburgh)，Wyatt Anderson，

Jonathan Arnold, Jamie Cuticchia 和 Yunxin Fu(Athens, Georgia), Joe Felsenstein (Seattle), Mike Clegg (Riverside), 和 Chris Basten, Ken Dodds, Rebecca Doerge, Spencer Muse, 和朱军 (Raleigh)。在一些专业问题上我还得到了 Walfer Fitch (Irvine), Anthony Edwards (Cambridge) 和 John Monahan (Raleigh) 的指点。本书出自我在 Raleigh 讲课时所用的一些讲稿。几位北卡罗来纳大学和 Duke 大学的学生帮助证明了我所用的一些方法。

感谢许多作者和出版社允许我引用一些未发表过的数据和享有版权的资料。

在成书过程中，我与 Sinauer Associates 一起工作始终感到很愉快。要特别致谢 Andy Sinauer，正是在他的鼓励下，我才去啃 LATEX。我学得迟缓，他对此赋予极大的耐心，因而使得这本书价格适中，并避免了导致印刷错误的一个环节。在使用这一打字软件过程中，我得到了 Chris Basten 和 Hal Caswell 的帮助，尤其是在绘图上 Chris Basten 给予了帮助。

我战战兢兢地在书末列上了一些计算机程序的源编码清单，并冒着使应用者可能得出错误结论的危险。但是，由于不断有人向我索要有关不平衡和 F—统计量程序的拷贝，这鼓励我将这些程序附上。我很高兴能够得到有关在本书中所列程序的应用价值的评论。尽管这些程序并非特别有效，但它们至少是正确的。使用者应对照我所列出的样本输入文件对其进行检验。这些程序都是用 FORTRAN/77编写的，并在微机上运行过。不过对于据此所作的任何分析，我不承担责任。

很明显，只要你打开这本书，就会感觉到 Clark Cockerman 的见解对我的影响有多大。Clark 指导了我25年，我深感荣幸。

最后，我还要感谢 Beth, Claudia 和 Henry 所给予的支持与耐心。

布鲁斯·威尔 (Bruce S. Weir)

1990年3月于北卡罗来纳

目　　录

译者的话

原序

1.引言 …… 1

1.1　遗传数据的属性 …… 1

1.2　遗传数据实例 …… 1

1.2.1　表现型数据 …… 1

1.2.2　共显性的表现型数据 …… 2

1.2.3　等位基因酶数据 …… 5

1.2.4　蛋白质序列数据 …… 8

1.2.5　限制性片段数据 …… 10

1.2.6　DNA 序列数据 …… 12

1.3　遗传抽样与统计抽样 …… 13

1.4　符号和术语 …… 16

1.5　孟德尔与费歇尔 …… 17

1.6　小结 …… 26

1.7　练习 …… 27

2.频率估计 …… 28

2.1　多项基因型计数 …… 28

2.1.1　多项矩量 …… 30

2.1.2　群体内基因频率的方差 …… 33

2.1.3　指示变量 …… 35

2.1.4　群体内基因频率的协方差 …… 37

2.1.5　实例 …… 38

2.1.6　基因频率的总方差 …… 41

2.1.7　Fisher 的方差近似公式 …… 43

2.2 似然估计 …… 45
2.2.1 最大似然估计量的特性 …… 48
2.2.2 求 MLE 的 Bailey 法 …… 52
2.2.3 似然方程的迭代求解法 …… 55
2.2.4 EM 算法 …… 56
2.2.5 实例 …… 59
2.2.6 配子频率 …… 63
2.3 矩量方法 …… 66
2.4 小结 …… 67
2.5 练习 …… 68
3. 不平衡 …… 70
3.1 哈迪-温伯格不平衡 …… 70
3.1.1 不平衡系数 D_A 的估计 …… 72
3.1.2 用 D_A 测验哈迪-温伯格不平衡 …… 73
3.1.3 HWE 的正合测验 …… 75
3.1.4 HWE 的似然比测验 …… 79
3.1.5 对数-线性模型（log-linear models） …… 81
3.1.6 复等位基因 …… 83
3.1.7 HWE 测验的功效 …… 87
3.2 连锁不平衡 …… 89
3.2.1 两个基因座的配子不平衡 …… 89
3.2.2 配子不平衡的正合测验 …… 90
3.2.3 复等位基因的配子不平衡 …… 93
3.2.4 配子连锁不平衡的方差和协方差 …… 93
3.2.5 三个或四个基因座的配子不平衡 …… 94
3.2.6 正规化的配子不平衡（normalized gametic disequilibria） …… 96
3.2.7 两基因座基因型不平衡 …… 96
3.2.8 复合的基因型不平衡 …… 101
3.2.9 连锁不平衡的对数-线性测验 …… 103
3.2.10 实例 …… 107
3.3 多重测验 …… 109
3.4 同质性测验 …… 110
3.5 小结 …… 113

3.6 练习 …… 113
4. 多样性 …… 115
4.1 引言 …… 115
4.2 杂合性 …… 115
4.2.1 杂合性的群体内方差 …… 116
4.2.2 杂合性的总方差 …… 118
4.3 基因多样性 …… 124
4.3.1 基因多样性群体内方差 …… 125
4.3.2 基因多样性的总方差 …… 127
4.3.3 估计多样性方差 …… 128
4.3.4 实例 …… 130
4.4 小结 …… 133
4.5 练习 …… 133
5. 群体结构 …… 135
5.1 引言 …… 135
5.2 固定群体 …… 136
5.2.1 基因频率 …… 136
5.2.2 Jackknife 法 …… 137
5.2.3 Bootstrap 法 …… 140
5.2.4 方差分析 …… 143
5.3 随机群体 …… 145
5.3.1 单倍体数据 …… 145
5.3.2 二倍体数据 …… 152
5.4 群体亚分 …… 155
5.4.1 三级谱系（three-level hierarchy） …… 156
5.4.2 四级谱系（four-level hierarchy） …… 159
5.5 遗传距离 …… 162
5.5.1 几何距离 …… 163
5.5.2 共祖距离（coancestry distance） …… 166
5.5.3 内氏遗传距离（Nei's genetic distance） …… 168
5.5.4 距离估计值的方差 …… 170
5.6 小结 …… 171
5.7 练习 …… 171

需要全本请访问 www.ertongbook.com

6. 世代间分析 …… 174
6.1 异交估计 …… 174
6.1.1 根据纯合母本进行估计 …… 174
6.1.2 根据平衡群体估计 …… 176
6.1.3 根据任意母本的子代进行估计 …… 177
6.1.4 多基因座估计值 …… 181
6.1.5 估计父本数 …… 183
6.2 父性推断 …… 185
6.2.1 利用共显性基因座排除父亲身份 …… 186
6.2.2 利用显性基因座排除父亲身份 …… 188
6.2.3 父性指数 …… 189
6.2.4 DNA 指纹分析 …… 191
6.2.5 最可能的父本植株 …… 193
6.3 选择的估计 …… 194
6.3.1 选择的拟合优度检验 …… 194
6.3.2 一个世代内的估计 …… 196
6.3.3 选择的分量 …… 197
6.3.4 生存选择的最大似然估计 …… 200
6.3.5 部分自交生物的选择 …… 201
6.3.6 母—子数据的利用 …… 204
6.4 连锁的估计 …… 208
6.4.1 基因间距离 …… 209
6.4.2 双回交法 …… 210
6.4.3 F_2群体法 …… 213
6.4.4 直接法 …… 216
6.4.5 已知连锁相的 LOD 值 …… 217
6.4.6 未知连锁相的 LOD 值 …… 218
6.5 小结 …… 219
6.6 练习 …… 219
7. 分子数据 …… 221
7.1 引言 …… 221
7.2 限制性位点资料 …… 221
7.2.1 估计片段长度 …… 221

7.2.2 顺序排列片段 …… 224
7.2.3 估计限制性位点的位置 …… 225
7.2.4 核苷酸变异的推断 …… 225
7.2.5 疾病关联 …… 227
7.3 DNA 序列资料 …… 231
7.3.1 碱基组成 …… 231
7.3.2 二核苷酸频率 …… 232
7.3.3 Markov 链分析 …… 235
7.3.4 单一序列中的模式 …… 238
7.3.5 Shuffling 测验 …… 241
7.4 序列比较 …… 242
7.4.1 点标图 …… 242
7.4.2 序列间的完全匹配 …… 242
7.4.3 Queen 和 Korn 法则 …… 244
7.4.4 Needleman—Wunsch 法则 …… 246
7.5 蛋白质序列资料 …… 250
7.6 DNA 序列间的距离 …… 254
7.7 小结 …… 257
7.8 练习 …… 259
8. 种系发生结构 …… 260
8.1 引言 …… 260
8.2 距离矩阵法 …… 262
8.2.1 平均连接聚类法（Average Linkage Clustering）（UPGMA） …… 263
8.2.2 Fitch—Margoliash 法则 …… 265
8.3 约减（Parsimony）法 …… 270
8.4 似然法 …… 274
8.4.1 DNA 序列的似然模型 …… 274
8.4.2 两个序列 …… 275
8.4.3 三个序列 …… 278
8.4.4 相对频率检验 …… 281
8.4.5 多个序列 …… 283
8.4.6 其它突变模型 …… 284
8.4.7 对系统发生树 Bootstrap 抽样 …… 285

8.5 小结 …… 285
8.6 练习 …… 286
附录 A 统计表 …… 288
A.1 正态分布 …… 288
A.2 卡方分布 …… 290
A.3 非中心卡方分布 …… 292
附录 B 随机数 …… 295
B.1 随机数的产生 …… 295
B.2 Bootstrap 抽样（bootstrapping） …… 296
B.3 随机 DNA 序列 …… 296
B.3.1 有约束的随机 DNA 序列 …… 298
附录 C 电算程序 …… 299
C.1 连锁不平衡 …… 299
C.1.1 配子不平衡 …… 299
C.1.2 假定 HWE …… 318
C.1.3 复合不平衡系数 …… 345
C.2 F-统计量 …… 359
C.2.1 单倍体数据 …… 359
C.2.2 二倍体数据 …… 376
C.3 DNA 序列处理 …… 397
C.3.1 Shuffling 法重排序列 …… 397
C.3.2 Karlin 算法查找顺接重复序列 …… 400
名词索引 …… 409
参考文献 …… 413

1. 引　　言

1.1　遗传数据的属性

本书对离散型性状数据作了论述。这些性状包括从花朵颜色之类的形态性状到 DNA 序列中某个碱基的类型。不论数据的属性怎样，本书所提供的分析方法适用于分析离散型遗传单位（它们具有孟德尔单位（Mendelizing units）的特征）的观察结果。换言之，所观察的单位可由亲代向子代传递，并且该传递过程是随机的。这样的单位曾被称作基因。但是，现在已不再需要将注意力局限于那些编码蛋白质的 DNA 区段。可以发展一些统计理论，用以分析那些确实是基因的遗传单位，或者那些取决于基因产物的物理性质的遗传单位以及那些与编码序列毫无关系的遗传单位。

在开始讨论这一内容以前，先来看看群体遗传学中离散型数据的类型。本章列举了几种不同类型的数据。

1.2　遗传数据实例

1.2.1　表现型数据

最为人们熟知的表现型数据是孟德尔的豌豆（*Pisum sativum*）试验数据。孟德尔在一系列杂交组合中观察了 7 对性状，并报道了杂种自交后代的试验结果（表 1.1）。根据这些结果孟德尔提出了显性与隐性表现型之比为 3∶1，尽管有人对其结果与这一比例

的过分符合提出了讨论。关于这一点，在本章稍后一节中将会提到。

表 1.1　孟德尔对豌豆（*Pisum sativum*）7 个显性性状的观察结果

（孟德尔，1866）

性　状	显性表现	隐性表现
	籽粒性状	
A　种子外形	5474　圆	1850　皱
B　子叶颜色	6022　黄	2001　绿
	植株性状	
C　种皮颜色	705　灰棕色	224　白
D　豆荚形状	882　饱满	299　皱缩
E　未熟豆荚颜色	428　绿色	152　黄色
F　花着生位置	651　腋生	207　顶生
G　植株高度	787　高秆	277　矮秆

1.2.2　共显性的表现型数据

正如 Fisher（1936）在讨论孟德尔的研究工作时所指出的，显性性状有个问题，即无法区分杂合体与显性纯合体，故不能肯定亲本的基因型。而象血型这样的共显性性状，就不会引起这类问题。表 1.2 列举了 Race 等（1949）所报道的一组相当完整的共显性表现型数据。

表 1.2　Race 等（1949）报道的几个家系的 MNS 血型

序	双亲		子　女					
号	父亲	母亲	1	2	3	4	5	6
1	MSMs	MsMs	MsMs	MSMs	MSMs			
2	MM. S	MsMs	MSMs	MSMs				
3	MM. S	MM. S	MM. S	MM. S	MM. S			
4	MM. S	MM. S	MM. S					
5	MM. S	MM. S	MM. S					
6	MM. S	MM. S	MM. S	MM. S	MM. S			
7	MsMs	MsNs	MsNs	MsNs				
8	MsNs	MsMs	MsMs					
9	MsNs	MsMs	MsNs					
10	MsMs	MsNs	MsMs=	=MsMs				

（续）

序号	双亲		子女					
	父亲	母亲	1	2	3	4	5	6
11	MsMs	MsNs	MsNs	MsMs	MsNs	MsMs		
12	MsNs	MSMs	MSNs—	—MsMs				
13	MSMs	MsNs	MsNs	MSNs	MsMs			
14	MsNs	MSMs	MsMs	MSNs	MSMs			
15	MM. S	MsNs	MSMs	MSMs	MSNs	MSNs	MSNs	MSNs
16	MsNs	MM. S	MSMs	MSMs=	=MSMs			
17	MM. S	MsNs	MSMs	MSMs				
18	MsNs	MM. S	MSMs=	=MSMs				
19	MsNs	MM. S	MSNs					
20	MsNs	MM. S	MSMs	MSMs				
21	MsMs	MSNs	MSMs	MsNs	MSMs	MsNs		
22	MsMs	MSNs	MSMs	MSMs	MSMs—	—MSMs		
23	MSNs	MSMs	MM. S	MsNs				
24	MSNs	MSMs	MM. S	MsNs				
25	MsNS	MSMs	MsMs					
26	MM. S	MN. S	MM. S	MN. S	MN. S			
27	MM. S	MN. S	MM. S	MN. S				
28	MM. S	MN. S	MM. S	MM. S				
29	MN. S	MM. S	MN. S	MM. S				
30	MM. S	MN. S	MN. S					
31	MM. S	MN. S	MM. S	MN. S				
32	MM. S	MN. S	MN. S	MM. S				
33	MN. S	MM. S	MN. S—	—MM. S				
34	MM. S	MN. S	MM. S	MM. S				
35	MM. S	MN. S	MM. S	MM. S	MN. S			
36	NsNs	MsMs	MsNs	MsNs				
37	NsNs	MsMs	MsNs	MsNs	MsNs	MsNs	MsNs	MsNs
38	NsNs	MsMs	MsNs	MsNs	MsNs			
39	MsMs	NsNs	MsNs					
40	MSMs	NsNs	MsNs	MsNs				
41	MSMs	NsNs	MSNs	MsNs				
42	NsNs	MSMs	MSNs	MsNs	MSNs			
43	NsNs	MM. S	MSNs					
44	NN. S	MsMs	MsNS					

需要全本请访问 www.ertongbook.com