遗传 HEREDITAS (Beijing) 2008 年 5 月, 30(5): 642－648
ISSN 0253-9772 www.chinagene.cn

DOI: 10.3724/SP.J.1005.2008.00642

群体遗传学教学与研究辅助软件的设计与应用

高婧, 潘沈元, 曹静

江苏省徐州师范大学生命科学学院, 徐州 221116

摘要: 为群体遗传学教学与研究的需要, 采用 Visual Basic6.0 开发设计了群体遗传学教学与研究辅助软件。软件设计综合考虑了影响群体结构各方面因素, 通过各种条件参数的设置、理论推算、计算机模拟和显示状态的选择, 以图、表的方式精确、形象直观地输出群体基因频率和基因型频率在世代间的变化、分布规律和统计特征。该软件无论从功能上, 还是从操作、界面上都是目前较为完善的教学与研究辅助软件。

关键词: 群体遗传学; 教学与研究; 辅助软件; 基因频率; 基因型频率

Design and application of computer-assisted software for teaching and research of population genetics

GAO Jing, PAN Shen-Yuan, CAO Jing

College of Life Science, Xuzhou Normal University, Xuzhou 221116, *China*

Abstract: The computer-assisted software for population genetics' teaching and research is designed with Microsoft Visual Basic 6.0. It takes all factors affecting population genetic structure into account. By way of setting various conditional parameters, calculating theoretically, simulating with computer and choosing the states of show, the system displays the results in charts and tables accurately. It also presents distribution pattern, statistical characteristics and the change of population gene frequencies and genotype frequencies in the inter-generation visually. The software is a mature system for teaching and researches on the aspects of its functions, operations and interface.

Keywords: population genetics; teaching and research; computer-assisted software; gene frequency; genotype frequency

群体遗传学是用数学和统计学的方法研究群体的遗传结构及其变化规律的遗传学分支学科。群体的遗传结构是由群体的基因频率和基因型频率决定。影响群体遗传结构的因素主要是群体大小、交配方式、选择、突变、迁移等。研究群体遗传的方法主要是通过化石记录、调查现有的自然群体、设计实验群体、构造数学模型、计算机程序设计等。各种方法各有其优缺点, 需要相互补充和相互验证。由于群体遗传学研究要涉及的世代很多, 且影响群体遗传结构的因素也很多, 往往需要较大和较多的群体才能反映群体遗传的统计规律性, 这时化石记录、自然群体、实验群体很难全面反映这种规律。数学模型便于对群体做定性定量的分析, 研究基因频率变化和平衡的规律, 但若要综合考虑影响群体遗传结构变化的诸多因素似乎得不到一种综合的数学模型, 只能对影响因素做简化处理, 或研究几个限制的影响因素。此时计算机分析就成了群体遗传研究的重要辅助手段。

将计算机应用于群体遗传学研究, 主要归功于 20 世纪 80 年代微型计算机的普及和应用, 当时的研

收稿日期: 2007−10−27; **修回日期:** 2007−11−30
基金项目: 江苏省精品课程建设项目资助[Supported by Elite Course Construction Project of Jiangsu Province]
作者简介: 高婧(1986−), 女, 2004 级生物科学(师范)专业本科生。E-mail: xuanyunfeng@163.com
通讯作者: 潘沈元(1957−), 男, 教授, 硕士, 研究方向: 遗传学, 生物统计。E-mail: PanShenyuan@263.net

究主要针对影响群体的某一两个因素进行特定模拟。近年来，国外的群体模拟研究不断深入，主要涉及针对特殊群体设计模拟模型，提高群体遗传模拟性能和算法的研究，并将研究结果应用于分子进化研究，如研究 SNP 的起源与进化问题[1~3],同时也在设计有关自然选择和基因重组的群体遗传模拟软件[4]。国内关于群体遗传模拟的研究涉及很少，有关程序设计主要用于群体遗传教学[5]，且程序的模拟功能十分有限，与国外差距甚远。本研究设计的软件主要定位于辅助群体遗传学的教学与研究，从理论计算和计算机模拟两个方面，综合考虑各种因素对群体遗传结构改变的影响，以数据、曲线和图形等多种形式反映群体遗传结构的变化情况及统计规律，以达到对教科书和数学模型无法涉及的复杂条件的预测效果。

1　软件设计思想和原理

由于遗传方式不同，分常染色体遗传和 X 连锁遗传两大模块设计。

1.1　条件参数设置

要求尽量全面反映影响群体遗传结构的因素，各类设置参数可以随意组合，既可以做单一因素的分析，又可以做各种综合因素的分析。

1.1.1　一般参数设置

包括模拟世代数、统计群体数、统计分组数、群体大小、初始群体基因型频率、各种基因型的适合度、基因的正反突变率、群体迁移率，迁移群体基因型频率。群体大小可以按雌雄群体分别设置，各个世代的群体大小可以相同，也可以不同，可以任意设置，也可以按某一函数关系设置。群体大小及以后的参数都可以按雌雄性别设置，可以相同、也可以不同。

1.1.2　特殊参数设置

包括：(1)是否存在选型交配？若是，显隐性关系？各种表现型的选型交配率？是正选型还是负选型？(2)是否存在家系大小问题？若存在，家系大小是由父方、母方还是父母双方决定？家系大小是固定还是随机分布(泊松分布)？各种基因型亲本的平均后代数？均可以做相同或不同的设置；(3)是否存在依频选择？若存在，被选择的基因型是完全显性还是无显性？选择系数与基因型频率的关系是线性函数、二次函数、对数函数还是平方根关系？与前面设定的适合度的关系是相加还是相乘？

1.2　理论计算原理

用以上设置的条件参数推导出一个数学公式，表示上下代基因型间的函数关系，将非常复杂甚至是不可能的，但它又是客观存在的关系纳入了该函数中。利用遗传学原理、各个参数的作用时期和计算机的逻辑判断和计算能力，逐步分解，可以实现上下代基因型间关系的推算。以二倍体常染色体遗传为例，介绍关键计算方法:

Step1: 计算双亲基因型产生 2 种类型配子的概率，若存在正反突变，矫正配子的概率;

Step2: 根据选型交配率和上代基因型频率，计算交配体系中，选型交配和随机交配的总概率，以及选型交配中各种基因型交配的概率和随机交配中各种基因型交配的概率，进而推算出整个交配体系中各种基因型相互交配的概率;

Step3: 若考虑家系大小，计算父母基因型组合的平均后代数，进而根据 Step2 中的结果计算交配体系中各种基因型相互交配并产生后代的概率;

Step4: 根据 Step3 和 Step1，计算双亲各种基因型交配并产生 3 种基因型的概率，进一步统计产生 3 种基因型后代的总概率;

Step5: 若存在依频选择，根据上一代的基因型频率按某种函数关系，调整 3 种基因型的适合度;

Step6: 将 Step4 中 3 种基因型后代的概率乘 Step5 中相应的适合度，并进行归一化矫正，得到选择后的基因型频率;

Step7: 若存在迁移，按雌雄群体迁移率和迁移群体基因型频率与当前群体基因型频率加权平均，得到迁移后的基因型频率。

循环迭代 Step2~Step7，可得到各个世代的基因型频率。

1.3　计算机模拟原理

计算机模拟同样是从上代基因型频率出发，按照设定的条件参数，以随机的方式模拟亲本基因型的组合-产生配子-突变-配子结合产生子代-子代接受选择等过程，仍以二倍体常染色体遗传为例，介绍关键计算方法:

Step1: 判断上代基因型是否固定，若是，直接确定以后世代的基因型，跳出循环;

Step2: 根据选型交配率和上代基因型频率，计算交配体系中，选型交配和随机交配的总概率，以及选型交配中各种基因型交配的概率和随机交配中

各种基因型交配的概率;

Step3: 产生随机数, 按选型交配和随机交配的总概率确定亲本交配类型, 若为随机交配, 转 Step5; 若为选型交配, 继续;

Step4: 产生随机数, 按选型交配中各种基因型交配的概率, 确定交配亲本的基因型, 转 Step6;

Step5: 产生随机数, 按随机交配中各种基因型交配的概率, 确定交配亲本的基因型;

Step6: 确定双亲产生后代的个体数, 若后代数设为随机, 则按 Poisson 分布随机产生该家系的后代数(算法略);

Step7: 根据双亲的基因型, 产生随机数, 随机产生雌雄配子; 根据基因的正反突变率, 产生随机数, 判断基因是否发生突变, 最终确定雌雄配子基因型;

Step8: 配子结合产生下代个体, 确定其基因型, 产生随机数, 按该基因型的适合度确定该个体是保留还是淘汰(若存在依频选择, 先按要求调整适合度), 若淘汰, 转 Step7;

Step9: 产生随机数, 判断子代个体的性别, 累计雌雄各种基因型的个体数;

Step10: 若双亲产生的后代个体数小于规定的家系后代数, 转 Step7;

Step11: 若累计雌雄各种基因型个体数总和小于当代设置的群体大小, 转 Step2;

Step12: 若存在迁移, 根据当代设置的群体大小和雌雄群体迁移率, 计算迁移个体数;

Step13: 产生随机数, 根据雌雄迁移群体基因型频率, 随机确定迁移个体基因型, 累计雌雄各种基因型的个体数, 循环直到迁移个体数达到要求;

Step14: 根据 Step9 和 Step13 中的结果, 计算迁移后的群体基因型频率。

循环迭代 Step1~Step14, 可得各个世代的基因型频率。

1.4 结果显示与控制

根据 1.2 和 1.3 得到的理论和模拟的世代基因型频率, 不难得到各个世代雌雄群体的基因频率, 进而可以计算群体平均基因频率和基因型频率, 统计或计算基因频率的分布及平均数和方差。为便于教学、分析和研究, 程序设计除了要求能够显示、保存尽量多的遗传信息外, 还应该图文并茂, 以数据、曲线和图形等多种形式反映群体遗传结构的变化情况及统计规律, 同时应操控方便, 按不同目的要求, 随意组合显示内容。

2 功能实现

根据上节程序设计思想和原理, 设计了 8 个具有独立功能的窗体, 主要由 4 个窗体来完成上节要求的全部的功能。分常染色体遗传和 X 连锁遗传, 他们分别是子群体世代基因、基因型频率分化窗体和群体世代基因频率分布窗体。其余 2 个用于演示常染色体遗传和 X 连锁遗传漂变的动态过程, 2 个仅考虑突变、选择、迁移因素的模拟, 运算速度较快。其中常染色体遗传子群体世代基因、基因型频率计算与模拟窗体如图 1 所示。各窗体界面主要分为 3 个部分:

2.1 参数设置部分

分 2 页完成 1.1 中介绍的一般参数设置和特殊参数设置。

2.2 结果显示部分

两类窗体共分 10 个结果显示页, 分别是:

(1) 反映各世代子群体分化的基因频率和基因型频率圆饼图;

(2) 反映各世代子群体间基因频率和基因型频率离散过程的折线图;

(3) 反映各世代子群体内雌、雄群体和合并后的基因频率、基因型频率离散过程折线图;

(4) 反映各世代群体基因频率变化统计规律的频数直方图;

(5) 反映各世代群体分化过程的 3 维基因频率频数分布图;

(6) 反映各世代各群体平均基因频率变化的折线图;

(7) 各世代各子群体雌、雄及其合并后的基因型频率表;

(8) 各世代各子群体雌、雄及其合并后的基因频率表;

(9) 各世代群体基因频率频数分布表;

(10) 各世代雌、雄及其合并后的群体基因频率平均数和方差表。

计算分析结果的部分显示类型如图 2 所示。

2.3 状态显示与控制部分

用来控制程序的执行; 是否显示程序的执行过程以动态反映群体随世代的变化情况; 是否显示图

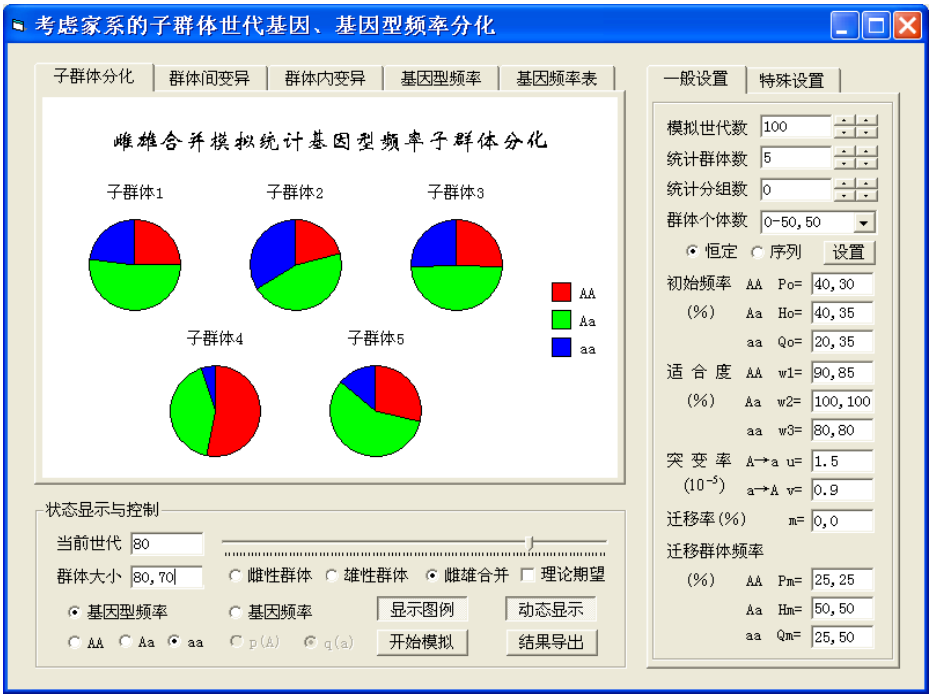
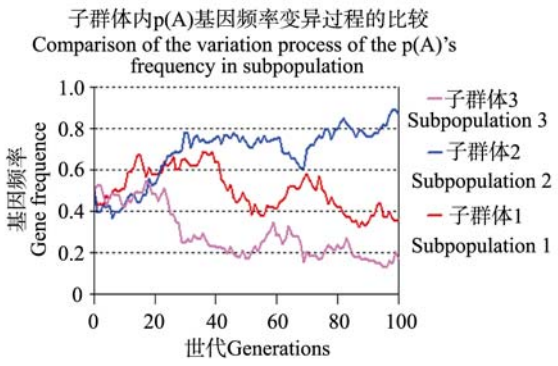
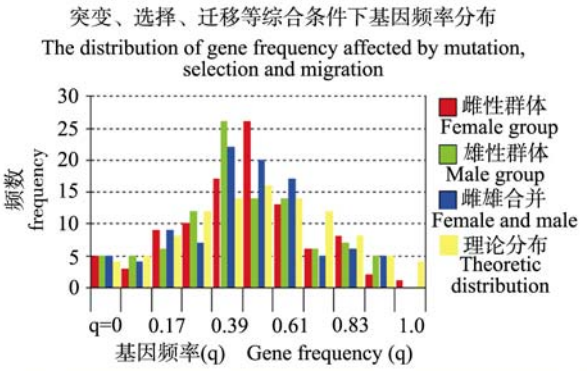
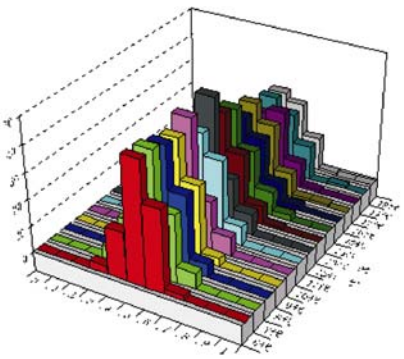


图 1　子群体世代基因、基因型频率分化计算与模拟窗体

Fig. 1　The calculation and simulation form of subpopulation differentiation of gene and genotype frequencies in each generation







世代	雌性基因型频率			雄性基因型频率			合并基因型频率		
	P(AA)	H(Aa)	Q(aa)	P(AA)	H(Aa)	Q(aa)	P(AA)	H(Aa)	Q(aa)
0	0.25	0.50	0.25	0.25	0.50	0.25	0.25	0.50	0.25
1	0.21	0.52	0.27	0.35	0.35	0.29	0.28	0.44	0.28
2	0.27	0.45	0.27	0.20	0.53	0.27	0.24	0.49	0.27
3	0.20	0.61	0.20	0.26	0.46	0.28	0.23	0.54	0.24
4	0.30	0.34	0.36	0.13	0.53	0.34	0.21	0.43	0.35
5	0.12	0.52	0.37	0.20	0.58	0.23	0.16	0.55	0.30
6	0.17	0.52	0.30	0.13	0.46	0.41	0.15	0.49	0.36
7	0.13	0.46	0.41	0.22	0.52	0.26	0.17	0.49	0.33
8	0.09	0.65	0.26	0.09	0.54	0.37	0.09	0.59	0.32
9	0.18	0.53	0.30	0.17	0.45	0.38	0.17	0.49	0.34
10	0.09	0.55	0.36	0.19	0.55	0.26	0.14	0.55	0.31
11	0.09	0.54	0.37	0.15	0.48	0.37	0.12	0.51	0.37
12	0.16	0.29	0.55	0.06	0.47	0.47	0.11	0.38	0.51
13	0.09	0.32	0.59	0.04	0.41	0.55	0.06	0.36	0.57
14	0.07	0.39	0.55	0.04	0.59	0.38	0.05	0.49	0.46

图 2　计算和模拟结果的主要显示方式

Fig. 2　The main way to show the results of calculation and simulation

例；显示基因频率还是基因型频率；显示哪种基因型、哪种基因的频率变化曲线；显示雌性群体、雄性群体还是两者的平均或同时显示；显示理论计算结果还是计算机模拟结果或同时显示；显示哪个世代中的结果以及当前世代的主要统计特征；是否将参数和计算结果输出到 Excel 中，以便编辑与保存等等。

3 主要特色

3.1 综合功能

本软件的参数设置基本上涵盖了群体遗传学教科书中影响群体遗传结构改变的主要因素，并有所拓展，多种参数可以根据研究的需要，随意组合，从而可以综合地分析群体遗传结构的变化特点与规律。国内已有的群体遗传模拟报道多涉及遗传漂变的计算机模拟，即便是有选型交配、选择的单项模拟，内容不但简单，如只涉及到完全显性情况下的完全正选型交配，而且也不是真正意义上的随机模拟，只是通过理论计算得到结果[6]。

将理论计算和计算机模拟集成在同一软件中，对相同影响因素的参数组合同时采用理论计算和计算机模拟方法，可对分析结果相互支撑和比较。目前群体遗传学研究尚没有似乎也不可能有综合各种影响因素反映群体基因频率和基因型频率变化的理论模型，但根据遗传规律，通过计算机的逻辑判断和计算能力，可以实现各个世代基因频率和基因型频率变化期望值的精确预测。根据计算机计算和模拟的结果，也可以对理论上推导的某些简化条件下的近似公式进行验证。

当然，我们所实现的所谓“综合”只是相对的综合，引起群体遗传结构改变的因素错综复杂，难以面面俱到，我们将努力对软件的综合性不断完善。

3.2 统一的操作界面，灵活的参数设置

为方便用户使用，本软件的主要界面在设计风格上力求统一，用户只要掌握一个界面的操作，就可以类推其他。在操作方面，对每一个输入和选择控件，当鼠标移到其上时，都会有相应的提示；运算前做参数检查，发现错误给以提示；对于群体大小、初始基因型频率、突变率、迁移率、迁移群体基因型频率的参数设置，若仅输入 1 个数，则表示雌雄群体相同，若输入 2 个数以逗号分割，前者表示雌性群体，后者表示雄性群体，无论是全角半角逗号都不会发生错误；即使输入的 3 种基因型频率总和不为 1，适合度不是相对适合度，程序都能自动给以矫正；若各世代群体大小不固定，将打开子窗口，其中嵌入了 Excel 图表对象，可以输入函数，自动生成具有某种变化规律的雌性群体大小序列，也可以手动按要求逐个设置，如需模拟瓶颈效应、奠基者效应等，最后还能以曲线方式反映雌雄性群体的变化曲线。总之，程序设计尽量考虑了参数设置的方便性和安全性。

3.3 多种结果显示功能

事实上，计算机计算或模拟的结果只是一大堆数据，虽然是第一手资料，但不直观，不容易发现其中的规律，需要统计归纳处理，以表、线、图等多种方式显示。正像图 2 所示，我们设计了 10 个子窗口来显示各种类型的分析结果，再通过显示状态控制，可以组合显示、比较不同世代、不同子群体、雌雄性群体、不同基因和基因型的理论的和模拟的分析结果。

4 应用举例

4.1 理论计算和计算机模拟相互验证

计算机模拟和理论计算的主要差别是前者带有随机性，随着群体变小而使随机漂变加大，但从理论上讲模拟结果的数学期望应该是理论计算值。利用这一特点可以验证程序设计的正确与否，因为计算机模拟和理论计算是于两个不同的体系，同时可以用来检验推导的理论公式的正确性。

理论计算值和模拟结果在世代中变化的比较，设定群体中雌雄个体数均为 10 000，初始基因型频率雌雄不同，分别是 P_{0f}=0.5，P_{0m}=0.65，H_{0f}=0.495，H_{0m}=0.34，Q_{0f}=0.005，Q_{0m}=0.01，基因的正反突变率分别是 1.3×10^{-5}，和 0.9×10^{-5}，3 种基因型 *AA*，*Aa*，*aa* 雌雄适合度也不相同，分别是 w_{1f}=0.9，w_{1m}=0.85，w_{2f}=1，w_{2m}=1，w_{3f}=0.7，w_{3m}=0.7，杂合子具有选择优势，不存在迁移，存在正选型婚配，完全显性，显性个体的同型交配率和隐性个体的同型交配率均为 50%；家系大小服从 Poisson 分布，由父方基因型决定，基因型 *AA*, *Aa*, *aa* 的平均后代数分别是 2, 2, 3，增加了 *aa* 在选择上的优势；不存在依频选择。从图 3 可以看到，两种算法基因型频率随世代的变化基本一致，同时也揭示了一个重要的遗传问题，我们将另做报道。

4.2 基因频率的统计分布

以往的群体模拟软件注重于表现基因频率的随

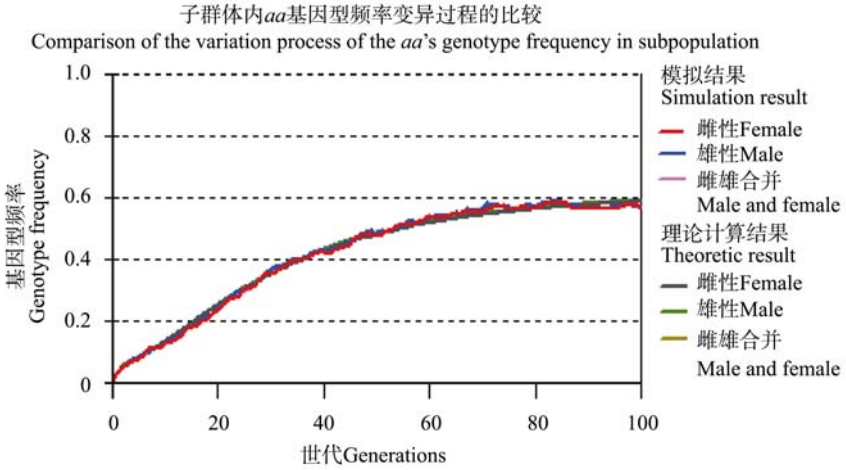


图 3　理论计算和计算机模拟结果比较
Fig. 3　Comparison of theoretic calculation and computer simulation

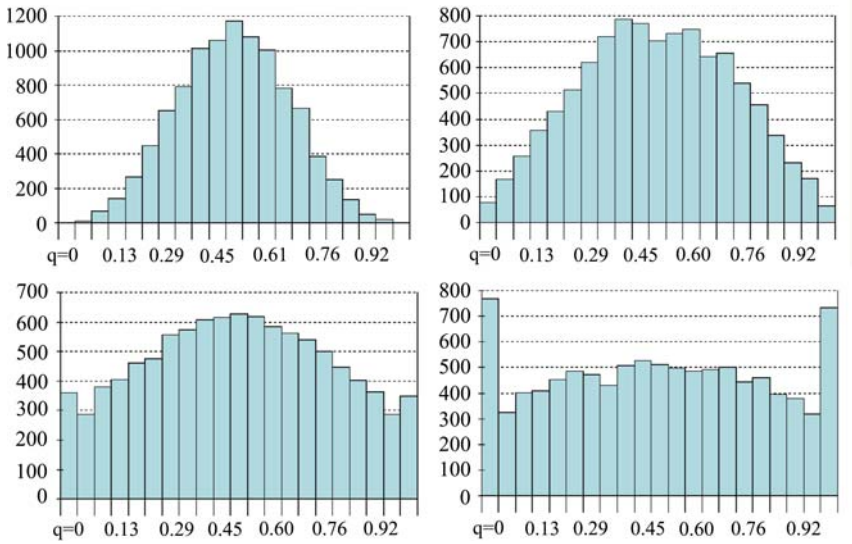


图 4　基因频率在某些世代的频数分布
Fig. 4　The frequency distribution of gene frequency in some generations

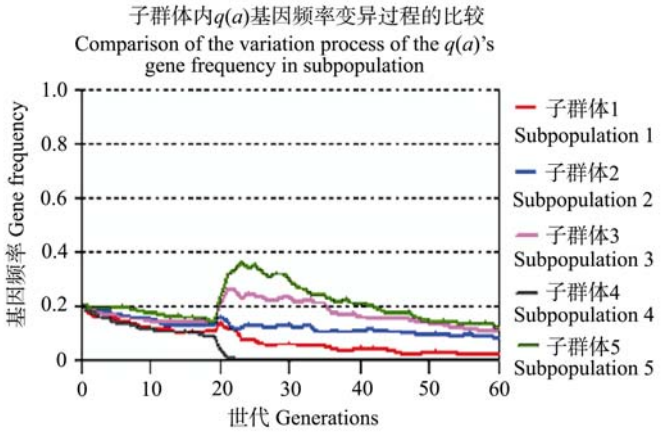


图 5　瓶颈效应模拟结果
Fig. 5　Simulation result of bottleneck effect

机变化，好像无规律可言，但是在大量的实验中，还是有统计规律的，描述随机过程最完整的方法就是基因频率在各个世代的统计分布，简要的方法就是分布的数字特征平均数和方差，本软件可以统计几百、几千甚至上万个子群体，绘制各个世代基因频率分布的直方图，给出平均数和标准差。图 4 描述了初始基因频率为 0.5，群体大小为 100 时，由于遗传漂变，在 25、50、75 和 100 世代基因频率统计分布状况。

4.3 瓶颈效应模拟

模拟东卡罗林群岛先天性失明症的遗传漂变。假设初始群体处于平衡状态，人数为 2 000，患者频率为 0.04，相对适合度为 0.8，基因突变率设置与 4.1 中相同，模拟 5 个子群体。经过 19 代的选择和漂变，5 个子群体的 *a* 基因频率分别为 0.01、0.02、0.02、0.01、0.02，*aa* 基因型频率分别为 0.01、0、0.07、0、0.10。在 20 世代时受到飓风的袭击，仅留下 9 个男人和 21 个女人，经过 10 代以 1.5 的增长率恢复到 1 500 人，由于瓶颈效应，在 23 世代时 5 个子群体的 *a* 基因频率分别为 0.08、0.12、0.23、0、0.37，*aa* 基因型频率分别为 0.01、0、0.07、0、0.10，子群体基因频率最高达 0.37，患者频率最高达 10%，同时也有出现基因丢失。模拟结果如图 5 所示。

在教学中要让学生明确瓶颈效应既可以使隐性不利基因在群体中以较高的频率保存，也可以使其丢失，小群体只是加大了基因频率变异的幅度，但方向无法预测。

参考文献(References):

[1] Fearnhead P. Perfect simulation from population genetic models with selection. *Theor Popul Biol*, 2001, 59: 263–279. [DOI]

[2] Maksymowicz AZ. Simulation of population growth and structure of the population. *Computer Physics Communications*, 2002, 147(1-2): 577–581. [DOI]

[3] Paul Fearnhead. Perfect Simulation from non-neutral population genetic models: variable population size and population sub-division. *Genetics*, 2006, 174: 1397–1406. [DOI]

[4] Spencer GCA, Coop G. SelSim: a program to simulate population genetic data with natural selection and recombination. *Bioinformatics*, 2004, 20: 3673–3675. [DOI]

[5] HUANG Yuan-Zhang. The application of computer simulation in teaching population genetics. *Hereditas*(*Beijing*), 1998, 20(4): 26–27.
黄远樟. 计算机模拟在群体遗传教学中的应用. 遗传, 1998, 20 (4): 26–27.

[6] MAO Sheng-Xian, HUANG Yuan-Zhang. Population Genetics and Programming. Beijing: Beijing Normal University Press, 1991, 215–225.
毛盛贤，黄远樟. 群体遗传及其程序设计. 北京: 北京师范大学出版社, 1991, 215–225

中国遗传学会第八届全国代表大会公司参展邀请

中国遗传学会第八届全国代表大会将于 2008 年 10 月 28 日-31 日在重庆市召开。为了促进科研用户与公司的交流，本次会议将举办展览，欢迎各大公司与我们联系参展。

本次大会议程丰富，规模盛大。大会的主题是：“新世纪的遗传学与社会和谐发展”。其函盖的领域为植物遗传学、医学遗传学、动物遗传学、微生物遗传学、群体与进化遗传学、基因组学、蛋白组学、发育遗传学、表观遗传学和药物基因组学等相关专业学科。各地报名参会十分踊跃，预计大会规模近千人。

出席本次大会的专家阵容强大。已邀请的院士报告为：李家洋院士，张启发院士，张亚平院士，贺福初院士，贺林院士等。另外李载平、吴常信、曾溢滔、夏家辉、沈岩、陈晓亚、杨焕明、孟安明等院士均参加会议。

大会报告的演讲人除邀请外，还将从提交的优秀论文中遴选报告人，会议将出版论文集和中国遗传学会成立三十周年大型纪念画册。

本次大会的招展工作全权委托上海市遗传学会和复旦大学遗传工程国家重点实验室共同负责。

展位内容及价格：详情请查看大会网页：www.congress-gsc.cn

联系方法：中国遗传学会　电话：01064889611；传真：01064853199

电子信箱：ccwang@genetics.ac.cn　联系人：王长城

上海市遗传学会　联系人：万波 / 张元元

电话/传真：86-21-65643404 电子信箱：sklge@fudan.edu.cn

中国遗传学会

2008 年 4 月 22 日