生物多样性　第 8 卷,第 2 期,2000 年 5 月
CHINESE BIODIVERSITY　8 (2) :238 ~ 240, May, 2000

一个实用的群体遗传学分析软件包
——GENEPOP 3.1 版*

刘俊娥　乔传令*
(中国科学院动物研究所,　北京　100080)
侯　鑫
(中国科学院植物研究所系统与进化植物学开放研究实验室,　北京　100093)

摘　要　GENEPOP 是一个非常实用的群体遗传学分析软件包,适用于对大量的群体遗传学数据进行分析。它主要有以下 3 个方面的用途 :1)进行正合检验,如对哈迪_温伯格平衡、种群差异和位点间的连锁不平衡进行检验 ;2)估算经典的群体遗传学参数,如 *Fst* 和其它相关指数及基因频率等 ;3)可把 GENEPOP 的输入文件转换为其它常用的群体遗传学分析软件包(如 BIOSYS、FSTAT 和 LINKDOS)所要求的输入文件格式。与软件 BIOSYS 相比,它在所用的统计学检验方法、适用的群体遗传学研究数据类型及输入文件等方面具有一定的优点。

关键词　群体遗传学,软体包

A useful population genetics software package—GENEPOP (Version 3.1)/LIU Jun_Er[1)], QIAO Chuan_Ling[1)], HOU Xin[2)]

Abstract　GENEPOP is a software package used to analyse data of population genetics. It is able to perform three major tasks, i. e., exact test for Hardy_Weinberg equilibrium, population differentiation and genotypic disequilibrium among pairs of loci, estimation of classical population parameters, such as *Fst* and other correlations, allele frequencies, etc., convering the input GENEPOP files to formats used by other programs, like BIOSYS, FSTAT and LINKDOS. It has some advantages over other population genetics software such as BIOSYS in statistics test, data type and input files.

Key words　population genetics, software package

Author's address　1) Institute of Zoology, Chinese Academy of Sciences, Beijing　100080
2) Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing　100093

群体遗传学是遗传多样性研究的重要理论,对利用同工酶技术和各种 DNA 分析技术获得的大量群体遗传学数据的分析是群体遗传学研究的重要内容。GENEPOP 3.1 版是由法国学者 Raymond 和 Rousset 开发的 GENEPOP 1.2 版(Raymond & Rousset,1995)的升级版本,该软件是一个非常实用的软件。

GENEPOP 3.1 版可用于对单倍体或二倍体数据的群体遗传学分析。对于同工酶数据可将基因型用 2 位数字编号,对于微卫星序列等 DNA 数据,可按碱基数目将不同的序列编号进行分析,非常方便。GENEPOP 的主要功能如下 :1)进行正合检验(exact test),如对哈迪_温伯格平衡及杂合子过量或不足的检验,对种群差异和位点间的连锁不平衡进行检验,提供检验的 *P* 值 ;2)估算经典的群体遗传学参数,如 *Fis*、*Fst*、*Fit*、*Rho_is*、*Rho_st*、*Rho_it* 和其它相关指数及基因频率、杂合子及纯合子的观测值和期望值等 ;3)可把 GENEPOP 的输入文件转换为其它常用的群体遗传学分析软件包,如 BIOSYS(Goudet,1995)、FSTAT(Swofford & Selander,1981)和 LINKDOS(Garnier_Gere & Dilmann,1992)等所要求的输入文件格式。

GENEPOP 没有自带的文本编辑器,它的数据输入文件为纯文本文件,数据文件要与 GENEPOP 存于同一

通讯作者
收稿日期 :1999 - 03 - 15 ;修改稿收到日期 :1999 - 10 - 12
侯鑫 e_mail :biolog@ nmg2. imu. edu. cn

子目录下。所有的输出文件均可在 EDIT 等文字处理程序中进行查看和编辑。

GENEPOP 在 MS_DOS 环境下运行,主菜单主要有如下几个选项:

选项 C 可用于改变要进行分析的数据文件或修改文件中的数据。选项 1 提供种群内各位点的平衡状态及种群整体的平衡状态的检验。根据检验的条件和假设不同共有 5 个子选项:1)假设 H_1 = 杂合子不足的 U 检验;2)假设 H_1 = 杂合子过剩的 U 检验;3)基于 H_0 = 配子自由组合的概率检验;子选项 4 和 5 是对应于子选项 1 和 2 的多样本检验。每种检验都有两种算法可供选择:各种群中每个位点的等位基因总数目不大于 4 时,用完全枚举法(Louis & Dempster,1987)进行精确的 HW 检验,如果每个种群中每个位点的等位基因数都大于 4,GENEPOP 自动用马可夫链法(Markov chain method)(Guo & Thompson,1993)计算 HW 检验中无偏估计的 P 值。每次检验得出的参数有:1)P 值;2)标准误差(只有用马可夫链法时才有);3)*Fis* 的 W&C 估计值(Weir & Cockerham,1984)和 R&H 估计值(Robertson & Hill,1984);4)完全枚举法中所用的表格数。

选项 2 可用于处理单倍体或二倍体数据,对于单倍体数据,使用者先用子选项 3 为所有个体的所有位点加上一个相同的假想等位基因,使之变为二倍体数据,然后用主菜单中的 C 选项读取这个新文件,再用子选项 1 进行检验,得出的参数有 P 值和标准误差。

选项 3 是关于种群间遗传差异和基因型差异的检验。子选项 1 检验的假设是 H_0 = 各种群内的等位基因频率相同。子选项 3 检验的假设是 H_0 = 各种群内的等位基因型频率相同。子选项 1 和 3 对每个位点进行检验,计算无偏估计的 P 值和标准误差。子选项 2 和 4 对应于子选项 1 和 3,在每个位点上对所有种群对进行检验。

选项 4 用私有基因法给出有效迁移个体数的多位点估计值。共提供了 4 个 *Nm* 估计值,其中 3 个是用 B & S(Barton & Slatkii,1986)发表的文章中的回归曲线法计算出来的,另外一个是修正的估计值,用最近似的回归线计算而得。

选项 5 提供了数据文件的基本信息。对于各种群中的每个位点,输出文件提供了以下数据:基因型矩阵,纯合子和杂合子数目的期望值和观察值,基因频率表,根据 W & C (Weir & Cockerham,1984)计算的 *Fis* 估计值,以及根据 W & C (Weir & Cockerham,1984)和 R & H (Robertson & Hill,1984)计算得出的 2 个综合的 *Fis* 估计值。

选项 6 根据等位基因频率和等位基因大小分别用 F 统计和 *Rho* 统计(Rousset,1996)来估算群体间的相似度。子选项 1 和 2 用 F 统计估算有关的统计量。子选项 1 的输出文件中列出了每个位点上各种群的基因型矩阵和每个位点上所有种群的 *Fis*,*Fst* 和 *Fit* 估计值,最后给出了多位点的综合的 *Fis*,*Fst* 和 *Fit* 的估计值。子选项 2 列出了每对种群每个位点的 *Fst* 估计值及多位点的综合的 *Fst* 估计值。子选项 3 和 4 则根据等位基因大小用 *Rho* 统计估算有关的统计量。子选项 3 和 4 的输出文件的内容和子选项 1 和 2 相似,只是有关的统计量表示为 *Rho_is*,*Rho_st* 和 *Rho_it*。子选项 5 用于分析距离隔离。

选项 7 可将 GENEPOP 输入文件转换为 FSTAT(Goudet,1995)、BIOSYS(Swofford & Selander,1981)、ANOVA(Weir,1990)和 LINKDOS(Louis & Dempster,1993)文件格式。选项 8 的子选项 1 用 Dempster 等(Dempster et al,1977)的 EM 列举法计算一个哑等位基因存在时基因频率的最大似然估计值。输出文件中列出了哑等位基因存在时的基因频率和杂合子及纯合子的估计值,GENEPOP 认为数据组中编号最大的等位基因是哑等位基因。子选项 2 将一个单倍体数据"二倍体化"。子选项 3 用于删除所有可能的临时性文件。子选项 4 可将等位基因重新编号。

应用该软件进行分析的步骤如下:先将等位基因用 2 位或 3 位数字编号,建立数据文件,运行 GENEPOP 后即可选择各选项进行检验。建立数据文件时需注意以下几点:1)可以把几个位点的名称写在同一行内,中间用逗号隔开,但最后一个位点名称后没有标点符号;2)每个等位基因可以用 2 位或 3 位数表示,不需要连续编号,但每个位点等位基因的总数不可超过 99,种群数则没有限制;3)缺失的数据用 00 或 000 表示,文件的中间和末尾不能有空行;4)输出文件中种群的名称用的是输入文件中相应种群最后一个个体的名称;5)GENEPOP 不接受输入文件的扩展名。

GENEPOP 是一个功能强大的免费软件,它与常见的群体遗传学分析软件 BIOSYS 相比,具有如下优点:1)所用的统计学检验方法比 BIOSYS 更有功效。GENEPOP 运用的主要是正合检验,而 BIOSYS 运用的检验方

法主要是卡平方检验；2）数据文件比 BIOSYS 简单，且可用 2 位或 3 位数对数据进行编号；3）GENEPOP 输入文件中缺失的数据可用 00 或 000 表示，因此适用于某居群或某个体在某位点上基因型数据缺失的情况，而 BIOSYS 不适于这种情况；4）它不仅适用于二倍体数据，还适用于单倍体数据。但 GENEPOP 不能直接计算出居群间的实际遗传距离，只能根据 *Fst* 值进行估计。GENEPOP 可通过匿名 FTP 在服务器 ftp. cefe. cnrs_mop. fr 的 pub/pc/msdos/genepop 子目录下下载（需用二进制传输模式），或寄两张软盘到本文作者处获得。

参 考 文 献

Barton N H, Slatkin M, 1986. A quasi_equilibrium theory of the distribution of rare alleles in a subdivided population. *Journal of Heredity*, **56**: 409 ~ 415

Dempster A P, Laird N M, Rubin D B J, 1977. Maximum likelihood from incomplete data via the Emalgorithm. *Journal of Royal Statistics Society B*, **39**: 1 ~ 38

Garnier_Gere P, Dillmann C J, 1992. A computer program for testing pairwise linkage disequilibria in subdivided populations. *Journal of Heredity*, **83**: 239

Goudet J, 1995. Fstat version 1.2：a computer program to calculate *F*_statistics. *Journal of Heredity*, **86**: 485 ~ 486

Guo S W, Thompson E A, 1993. Performing the exact test of Hardy_Weinberg proportions for multiple alleles. *Biometrics*, **48**: 361 ~ 372

Louis E J, Dempster E R, 1987. An exact test for Hardy_Weinberg and muitiple alleles. *Biometrics*, **43**: 805 ~ 811

Raymond M, Rousset F, 1995. Genepop (Version 1.2)：population genetics software for exact tests and ecumenicism. *Journal of Heredity*, **86**: 248 ~ 249

Robertson A, Hill W G, 1984. Deviations from Hardy_Weinberg proportions：sampling variances and use in estimation of inbreeding coeffcients. *Genetics*, **107**: 713 ~ 718

Rousset F, 1996. Equilibrium values of measure of population subdivision for stepwise mutation processes. *Genetics*, **142**: 1357 ~ 1362

Swofford D L, Selander R B, 1981. Biosys_1：a FORTRAN program for the comprehensive analysis for electrophoretic data in population genetics and systematics. *Journal of Heredity*, **72**: 281 ~ 283

Weir B S, Cockerham C C, 1984. Estimating *F*_statistics for the analysis of population structure. *Evolution*, **38**: 1358 ~ 1370

Weir B J, 1990. Genetic data analysis. Sinauer Publ., Sunderland, MA

（本文责任编辑：孙大川）