Molecular Ecology Notes (2002), 10.1046/j.1471-8278 .2002.00228.x

PROGRAM NOTE

GIMLET: a computer program for analysing genetic individual identification data

NATHANIEL VALIÈRE
Laboratoire de Biométrie et Biologie Evolutive, UMR 5558, Université Claude Bernard, 43, boulevard du 11 novembre 1918, F-69622 Villeurbanne Cedex, France

Abstract

Growing interest in microsatellite genotyping, combined with noninvasive genetic sampling has led to the increased production of data. New tools to analyse these data are required. GIMLET is a user-friendly software package designed to perform several simple tasks: (i) construction of consensus genotypes from repeated genotyping; (ii) estimation of genotyping error rates; (iii) identification of identical genotypes; (iv) comparison of new genotypes to a set of reference genotypes; (v) determination of the kinship; and (vi) estimation of several population parameters such as allele frequencies, heterozygosity, probability of identity, and population size.

Keywords: data treatment, individual identification, microsatellites, population size, software

Received 19 November 2001; revision received 6 March 2002; accepted 19 March 2002

The revolution occurring in molecular ecology partially results from the development and the use of microsatellite genotypes (see Luikart & England 1999) as molecular marks (see for example Palsbøll 1999; Waits & Leberg 1999). The tremendous growth in size of new datasets asks for the development of new tools to organise and analyse the data. In the case of noninvasive genetic sampling (Taberlet *et al*. 1996), each sample has to be typed independently several times to correct for genotyping errors. After repeated genotyping, it is necessary to construct a consensus multilocus genotype [the most likely genotype based on all polymerase chain reaction (PCR) amplifications of a sample] and to estimate the genotyping error rates. In the case of individual identification based on microsatellite markers, errors could be mainly false homozygote (FH) and false alleles (FA). The first error is due to the detection of only one allele out of two in the case of heterozygous individuals (e.g. differential amplification after pipetting disequilibrium, gel misinterpretation). The second type of error is due to the 'creation' of a new allele (e.g. due to the slippage of the polymerase or a contamination), which is amplified and scored.

Correspondence: N. valière. Fax: +33 (0) 478 892 719; E-mail: valiere@biomserv.univ-lyon1.fr

Then, consensus genotypes have to be compared with each other to find unique genotypes and avoid redundancy. Moreover, previously identified individuals must be found among the recently genotyped samples. Additionally, potential parentage between individuals can be determined. Finally, population parameters such as allele frequencies, heterozygosity, probability of identity, or population size could be estimated.

GIMLET (Genetic Identification with MultiLocus Tags) is a user-friendly software package which performs all these tasks. GIMLET can be freely downloaded from http://pbil.univ-lyon1.fr/software/Gimlet/gimlet.htm. The software can be installed and run on any WINDOWS™. operating systems (WINDOWS95 and higher).

The input file format used in GIMLET is a text file (ASCII) with GENEPOP (Raymond & Rousset 1995) format. See tutorial document of GENEPOP program for a description of the required format. Although the GENEPOP program requires only 2-digits format for alleles, GIMLET accepts the 3-digits format for each allele.

Tasks are presented under two main menus. The 'Calculator' menu contains functions that construct consensus genotypes and estimates error rates for a set of genotypes from repeated PCR; estimates population genetic parameters such as allele frequencies, heterozygosity and probability of identity; and estimates demographic parameters such as population size or survival rate. The menu 'Identification'

contains functions that perform three tasks: the pooling of multilocus genotypes; the comparison of one or several multilocus genotypes with reference genotypes; and the determination of the parentage. Ranked according to their utilization order, the tasks can be listed as follows: (i) construction of consensus and estimation of error rates; (ii) identification of genotypes to obtain unique genotypes only and comparison of the new genotypes with previously collected samples; (iii) determination of the potential kinship between individuals; (iv) calculation of primary parameters such as allelic frequencies, heterozygosity, probability of identity; and finally (v) estimation of population size or other population parameters.

(i) GIMLET allows users to construct consensus genotypes from a set of PCR repetitions for each sample and to estimate the genotyping error rate. Both the construction of consensus genotypes and the estimation of error rates require a GENEPOP file containing the results of the repeated genotyping for each sample. Here, one 'pop' of GENEPOP corresponds to one sample; the different lines of a 'pop' are the different multilocus genotypes for this sample. The consensus genotype of a sample is constructed by retaining, for each locus, the allele(s) that are observed more than a given number of times (threshold number set by user). At each locus, when no allele can be retained (i.e. all allele scores are below the threshold) or when more than two alleles are retained, the genotype is considered missing data. As the construction of the consensus genotype is an automated procedure, the choice of the threshold value is very important: when the threshold is 1 (default value), any allele that occurs at least once is retained. In this case, false alleles are always retained. On the other hand, when the threshold is too high, no allele or only one of two could be retained in samples with poor amplification. The consensus genotypes can be saved into a GENEPOP format file.

(ii) The estimation of error rates at each locus is performed by comparing the repeated genotypes and the associated consensus genotype for each sample. GIMLET detects four categories of error:

1 False homozygote (or allelic dropout): when a heterozygote (from the consensus genotype) is typed as a homozygote (in one repeat).
2 False allele: when a homozygote (from the consensus genotype) is typed as a heterozygote (in one repeat).
3 Double error: when the error could be explained by an FH then an FA, or an FA then an FH.
4 Complete error: when the error is not explained by any of the types cited above.

Identification of genotypes is especially useful to pool identical genotypes in the case of noninvasive sampling where several samples from the same individual may occur within a single session. The comparison task is useful to 'follow' a genotype through several capture sessions. For a multilocus genotype, when there is no match with any other genotype, the program compares genotypes locus by locus, and tries to identify the closest genotype(s). The closest genotype is selected by the highest score, which corresponds to the number of loci that are identical between two genotypes divided by the total number of loci used. For one genotype, when two or more genotypes have the same score, then the program identifies an 'ambiguity' in the output file. Additionally, GIMLET detects the pairs of samples for which only one allele is identical at one or two loci, or only two alleles are different at just one locus [e.g. pairs of samples from a single individual where one or two error(s) occur(s) in the genotype at one or two loci even using multitubes approach]. In this case, the user will judge whether to regroup or identify these genotypes. The development of a process to treat missing data is relatively difficult to assess, being a case-by-case process. Thus, when a multilocus genotype contains a missing locus, it will be considered as a distinctive genotype. In the locus-by-locus comparison of genotypes, the missing loci are not used in the comparison or identification procedures. After pooling genotypes, all distinctive genotypes can be saved into a GENEPOP format file. Moreover, after the identification process, a GENEPOP format file can be constructed with the reference genotypes and the unambiguous unique genotypes, for which no reference multilocus genotypes can be identified and for which the score of the closest genotype is below a threshold set by user (this threshold, calibrated as a number of loci, has to be set low enough to ensure that the genotype is unique).

(iii) Parentage between individuals is assessed simply by comparing genotypes. An individual will be a potential parent of an offspring if the multilocus genotypes share at least one allele of two for all loci. A threshold error can be introduced to obtain imperfect parentage. Additional information about individuals (such as age or sex) or population (such as gestation/incubation time, age of sexual maturation) can also be used to improve the determination of the parents.

(iv) The allele frequencies, the heterozygosity rate as well as the probability of identity (PI) for each locus and overall loci can be all computed. Allele frequencies at each locus are simply the observed frequencies. For heterozygosity, both the observed (number of heterozygous genotypes divided by the total number of genotypes analysed) and the expected heterozygosity (computed using equation $H_{exp} = 1 - \Sigma p_i^2$ where p_i is the frequency of the i^{th} allele) are calculated. The mean overall loci is calculated assuming independence of loci. PI (the probability that two individuals in the population share the same genotype) is computed using the equations of theoretical PI, unbiased

PI (with sample size correction) and PI for sibs given in Waits *et al*. (2001).

(v) GIMLET can be also used to estimate demographic parameters. I consider two methods for estimating population size. When several sampling sessions are conducted, population size can be estimated using capture-mark-recapture methods. In this case, a marked individual corresponds to an individual sampled during a session and is identified by its multilocus genotype. In this case, GIMLET produces an output file to use as input data in the program CAPTURE (Otis *et al*. 1978). This output file allows the estimation of population size using various models for capture probability available in CAPTURE program. The second method is the rarefaction curve method used by Kohn *et al*. (1999). These authors estimated the population size as the asymptote of the curve between the cumulative number of unique genotypes and the number of typed samples. GIMLET creates a file giving the number of occurrences for each unique genotype and can be employed in the R package (Ihaka & Gentleman 1996) using a script file generated by GIMLET. As the order in which samples are typed affects the estimation of asymptote (see Kohn *et al*. 1999), a random sampling of the genotypes was repeated 1000 times. Then *nls* function in the R package is used to estimate the asymptote value using nonlinear model approximation and two different equations. The first equation corresponds to that used by Kohn *et al*. (1999). The second equation corresponds to the expectancy of the number y of occupied boxes (unique genotypes) when we randomly distribute x balls (typed samples) in a boxes (number of individuals): $y = a - a[1 - (1/a)]x$. This equation is expected to give a more accurate estimation but further study is required to compare estimation of population size using these two equations. Other population parameters such as survival rate or recapture rate can be estimated by using a file constructed by GIMLET that can then be used in the MARK program (White & Burnham 1999) and SURGE program (Pradel and Lebreton 1993).

Acknowledgements

I thank D. Chessel for the equation and R code development for the estimation of population size using the rarefaction curve method. I thank C. Maudet, I. Till-Bottraud, G. Luikart, S. Regnaut, P. England and J.-M. Gaillard for useful comments and suggestions on earlier versions of the software and manuscript. The kinship module of GIMLET was inspired from the work of Alain Cercueil, Eva Bellemain and Stéphanie Manel and their program PARENTE (submitted).

References

Ihaka R, Gentleman R (1996) R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, **5**, 299–314.

Kohn MH, York E, Kamradt DA *et al*. (1999) Estimating population size by genotyping faeces. *Proceedings of the Royal Society of London, Serie B*, **266**, 657–663.

Luikart G, England PR (1999) Statistical analysis of microsatellite DNA data. *Trends in Ecology and Evolution*, **14**, 253–256.

Otis DL, Burnham KP, White GC, Anderson DR (1978) Statistical Inference from Capture Data on Closed Animal Populations. *Wildlife Monographs*, **62**.

Pradel R, Lebreton (1993) User manual for program SURGE, version 4.3. CEFE/CNRS, Montpellier, France.

Palsbøll P (1999) Genetic tagging: contemporary molecular ecology. *Biological Journal of the Linnean Society*, **68**, 3–22.

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

Taberlet P, Griffin S, Goossens B *et al*. (1996) Reliable genotyping of samples with very low DNA quantities using PCR. *Nucleic Acids Research*, **24**, 3189–3194.

Waits JL, Leberg PL (1999) Advances in the use of molecular markers for studies of population size and movement. *Transaction of the North American Wildlife Society*, 191–201.

Waits LP, Luikart G, Taberlet P (2001) Estimating the probability of identity among genotypes in natural populations: cautions and guidelines. *Molecular Ecology*, **10**, 249–256.

White GC, Burnham KP (1999) Program MARK: Survival estimation from populations of marked animals. *Bird Study*, **46**, 120–138.