SIMULATION OF THE GENETIC STRUCTURE AND REPRODUCTION IN PLANT POPULATIONS: SHORT NOTE

Dušan Gömöry

Faculty of Forestry, Technical University, SK–96053 Zvolen, Slovakia

Received February 14, 1995; accepted July 10, 1995

ABSTRACT

A set of programs which can be used to simulate a population with defined genetic structure and some reproduction processes is described. Distribution of the number of alleles at individual loci is simulated based on the modified Poisson distribution and serves for the generation of genotypes. Spatial location of genotypes can be assigned following 5 different distribution patterns. Progeny can be generated either under conditions of panmixia, or under spatial restrictions in mating, depending one of three pollen dispersion functions. Male and female gametic haplotypes are joined at random independently at each locus. Progeny individuals can be spatially distributed using the same dispersion functions as for pollen with different parameters. The programs are written in Turbo Basic and available on the Internet.

Key words: simulation, genetic structure, reproduction, pollen and seed dispersal

INTRODUCTION

The genetic structure of a population is a result of a complex of natural evolutionary factors as well as anthropic influences. An exact determination of factors, which have substantially contributed to forming of the present genetic architecture of a population, is frequently problematic in experimental research. In the case of short-lived organisms, the results of various evolutionary forces on population genetic structure can be assessed directly from experiments with living material in artificial conditions. This approach, however, is practically excluded in forest trees, where generation lengths are in the order of several tens of years. Therefore, with the exception of some types of selection, which can be studied under artificial conditions (BERGMANN & SCHOLZ 1989; GEBUREK & SCHOLZ 1985), investigating the effects of most management practices on genetic structure of forest stands, as well as the effects of natural processes are practically impossible. Mathematical modelling and simulation of these processes seems to be an appropriate substitution. Most of the published models describe changes in the population allelic structure in one locus (CUGUEN 1986). This paper describes a set of programs for creating a set of multilocus genotypes with selected characteristics representing adult population and for simulating of reproduction process to produce one or more offspring generations.

MATHEMATICAL MODELS AND SIMULATION METHODS

Simulation of maternal population

The program creates a population with Hardy-Weinberg genotype proportions at each locus, with a given mean effective number of alleles (CROW & KIMURA 1970; $\overline{n_e} = \sum_{j=1}^{n_l} (1 \,/\, \sum_{i=1}^{n_{a_j}} q_{ij}^2) \,/\, n_l)$ and/or mean expected heterozygosity ($\overline{H_e} = \sum_{j=1}^{n_l} (1 - \sum_{i=1}^{n_{a_j}} q_{ij}^2) \,/\, n_l$), and with a given mean number of alleles per locus, using the Monte Carlo method (q_{ij} is the frequency of the i-th allele at the j-th locus, n_{a_j} is the number of alleles at j-th locus, n_l is the number of loci).

The distribution of alleles at individual loci was used as a primary characteristic of the population, and as a basis of all further calculations. The modified Poisson distribution is used for its simulation. Published studies employing extensive sets of isozyme loci on several tree species (*Abies* sp. – JACOBS *et al.*, 1984; *Larix laricina* – CHELIAK *et al.* 1988; *P. brutia* – CONKLE *et al.* 1988) were used to prove the suitability of this distribution. The algorithm following DAVIES (1971) was applied for the simulation of the actual number of alleles at the j-th locus n_{a_j}: uniformly

© ARBORA PUBLISHERS

distributed random numbers u_i ($u_i \in \langle 0, 1 \rangle$) are generated as long as they match the inequality:

$$\prod_{i=1}^{n_{a_j}-1} u_i < e^{-(\overline{n_a}-1)} \quad [1]$$

$(n_{a_j} - 1)$ is then a random integer variable with Poisson distribution. Because the Poisson distribution starts from zero, and the number of alleles must be greater or equal to 1, this random integer must be increased by 1. $\overline{n_a}$ is the mean number of alleles per locus (input parameter).

Effective numbers of alleles at individual loci are calculated on the basis of the distribution simulated in the preceding step by multiplying the actual number of alleles by a random number u. The mean effective number of alleles ($\overline{n_e}$) is an input parameter. To obey the following conditions: $n_{e_j} \geq 1$; $n_{e_j} \leq n_{a_j}$, and $n_{e_j} = 1$ if and only if $n_{a_j} = 1$, effective number of alleles at the j-th locus n_{e_j} is simulated as

$$n_{e_j} = (n_{a_j} - 1) \,.\, u + 1 \quad [2]$$

Because the average of simulated effective numbers of alleles should approximate the mean effective number of alleles given as input ($\sum n_{e_j}/n_l \approx \overline{n_e}$; n_l is the number of loci), u should be a random number from the interval $\langle 0, 1 \rangle$ with an expectation k (the ratio $k = \dfrac{\overline{n_e} - 1}{\overline{n_a} - 1}$ is determined by the input data). Therefore, u is simulated as a random number uniformly distributed in the intervals $\langle 0, k \rangle$ and $\langle k, 1 \rangle$; selected at random from the first interval at frequency proportionate to $(1 - k)$ and from the second interval at frequency proportionate to k to ensure $E(u) = k$.

Allelic frequencies q_{mj} (m-th allele at the j-th locus) were simulated based on the effective number of alleles under conditions: $\sum_{i=1}^{n_{a_j}} q_{ij} = 1$ and $\sum_{i=1}^{n_{a_j}} q_{ij}^2 = 1/n_{e_j}$, as a number chosen at random within the limits given by:

$$q_{mj} = \frac{1 - \sum_{i=1}^{m-1} q_{ij}}{n_{a_j} - (m - 1)} \pm \sqrt{\frac{(n_{a_j} - m)\left[(n_{a_j} - m + 1)\left(\frac{1}{n_{e_j}} - \sum_{i=1}^{m-1} q_{ij}^2\right) - \left(1 - \sum_{i=1}^{m-1} q_{ij}\right)^2\right]}{n_{a_j} - (m - 1)}} \quad [3]$$

for $m = 1$ to $(n_{a_j} - 2)$. The last two allelic frequencies ($m = n_{a_j} - 1$, n_{a_j}) are calculated as follows:

$$q_{mj} = \frac{1 - \sum_{i=1}^{m-1} q_{ij}}{2} + \frac{\sqrt{2 \,.\, \left(\frac{1}{n_{e_j}} - \sum_{i=1}^{m-1} q_{ij}^2\right) - \left(1 - \sum_{i=1}^{m-1} q_{ij}\right)^2}}{2} \quad [4]$$

For more details on derivation of the formulae, see Appendix I.

As there is a functional relationship between the effective number of alleles and expected heterozygosity (h_{e_j}) at a single locus ($n_{e_j} = 1/(1 - h_{e_j})$), simulation is based on effective numbers of alleles even if the mean expected heterozygosity is used as input characteristic. However, mean heterozygosity corresponds to the harmonic mean of effective numbers of alleles, not to the arithmetic mean, so that the values of effective number of alleles at individual loci must be adjusted.

Subsequently, genotypes are generated from obtained allelic frequencies. Maternal and paternal gametes join at random, the model in the present version does not consider the linkage of loci. The alleles are chosen at random proportionately to allelic frequencies.

Simulation of the reproduction process

The program makes it possible to generate progeny either under the conditions of panmixia (probability of mating is the same for any pair of individuals) or dependent on the distance between parental individuals. All mating situations from full allogamy to full autogamy can be simulated. Mother individual is chosen at random or by the user. The distance to the pollen parent d is then generated based on one of three pollen dispersion functions depending on the parameter of pollen flow β (input parameter). Pollen density h and/or mating probability between two individuals ϕ are generated as random uniformly distributed numbers (u), the corresponding distance is calculated based on the dispersion function. The coordinates of the paternal individual are determined on the basis of the simulated distance d and a random chosen azimuth: $x_{pat} = x_{mat} + d \,.\, \sin(2\pi \,.\, u)$, $y_{pat} = y_{mat} + d \,.\, \cos(2\pi \,.\, u)$. The individual closest to the point $[x_{pat}, y_{pat}]$ is then chosen as the pollen parent. The pollen dispersion functions are following ones:

A. Gaussian curve; the distance d from the maternal individual to the paternal one is a normally distributed random variate with the mean 0 and standard deviation $\beta \,.\, \bar{d}$ ($\bar{d}$ is the average distance between individuals). A random variate v ($v \in N(0,1)$) is generated using the formula following DAVIES (1971):

$$v = \sum_{i=1}^{12} u_i - 6 \qquad [5]$$

(u_i is uniformly distributed random number from ⟨0, 1⟩). The distance between mates can be then generated as follows:

$$d = |v| \,.\, \beta \,.\, \bar{d} \qquad [6]$$

B. Dispersion function following ADAMS & BIRKES (1991):

$$\phi(i) = \frac{e^{-\beta . d_i}}{\sum_{k=1}^{n} e^{-\beta . d_k}} \qquad [7]$$

where $\phi(i)$ is the relative mating success of the i-th male partner at the distance d_i from the mother tree, n is the number of all possible pollen parents. An equal pollen fertility of all males is assumed. Term $e^{-\beta . d_i}$ is calculated for each individual (potential male) and the male is selected at random proportionately to its mating success. For seed dispersal, it is assumed that the seed density $h(d)$ decreases exponentially with the distance ($h(d) = e^{-\beta d}$), based on which the distance can be simulated as follows:

$$d = -\log u/\beta \qquad [8]$$

C. Pollen density $h(d)$ is indirectly proportional to distance d; $h(d) \in \langle 0, 1 \rangle$:

$$h(d) = \beta/(\beta + d) \qquad [9]$$

Thus, pollen dispersion can be simulated as follows:

$$d = \frac{1/u - 1}{\beta} \qquad [10]$$

(u is uniformly distributed random number from ⟨0, 1⟩).

The value of the pollen dispersion parameter β depends on the selected function and makes it possible to simulate also the infinite pollen flow (which is equivalent to full panmixia). In the first two variants, the appropriate value of β can be chosen based on the distance from maternal tree, in which 50 % of mating events occur. Because the last dispersion function has a divergent improper integral ($\int_{d=0}^{\infty} h(d) \to \infty$), pollen flow parameter β can be chosen based on the distance in which the actual pollen density decreases to 50 % of the maximum.

Seed dispersal can be simulated using the same functions changing the parameter β.

DESCRIPTION OF SIMULATION PROGRAMS

SIM1: The program simulates allelic frequencies of the maternal population. Input data comprise number of loci, required mean number of alleles, and required mean effective number of alleles and/or mean expected heterozygosity. Output contains the vector of number of alleles in individual loci and the matrix of allelic frequencies.

SIM2: The program simulates spatial coordinates for a given number of individuals. It allows to choose among 5 types of distribution patterns: regular square network, regular triangle network, irregular random distribution, irregular square network and irregular groupwise distribution. Number of individuals, distribution type and average distance between neighbors are required for all distribution patterns. In addition, minimum distance between neighboring individuals must be supplied in case of the irregular square network, and the number and relative compactness of groups on a scale from 0 (random distribution) to 10 (most compact groups) must be specified for the groupwise distribution. Output is represented by the set of coordinates in a two-dimensional space.

SIM3: The program creates a population with defined parameters. It requires allelic frequencies (created by the program SIM1, or from a real population). Genotypes are generated as random combinations of alleles at each locus. The current version does not allow to consider the linkage. Spatial coordinates (e.g., created by SIM2) can be assigned at random to the individuals.

SIM4: This program generates the progeny genotypes under the conditions of panmixia (probability of uniting of gametes is equal for any pair of individuals). Female and male gametes are generated by a random choice of one allele at each locus from the maternal and/or paternal diploid genotype and combined into progeny genotypes. The input file comprises the set of individuals (genotypes) of the parental population, while the output contains the set of progeny genotypes.

SIM5: The program simulates the creation of progeny in the case of limited pollen flow. Any of the dispersion functions presented in section 2.2 can be used for simulation. A graphic presentation of the dispersion diagram displayed on the monitor allows to choose the proper input parameters. Spatial coordinates can be assigned to the progeny genotypes using the same dispersion functions as in case of pollen flow.

© ARBORA PUBLISHERS

Otherwise, input and output data files are the same as with the program SIM4.

All the programs are written in the Turbo BASIC language. Executive as well as source (ASCII) files are available on anonymous ftp server vsld.tuzvo.sk in the directory /pub/incoming on Internet (compressed by pkzip; filename sim.zip), or can be requested by sending the diskette to the author. Source texts of all programs are supplied, so that further adaptations of these programs by the user are possible. All the programs can produce output files, which may be used directly for the BIOSYS-1 (SWOFFORD & SELANDER 1981), the most frequently used program for the analysis of population genetic data.

ACKNOWLEDGEMENT

Author is grateful to Dr. Vladimír Vacek, Department of Mathematics, Technical University in Zvolen, for helpful comments and discussions.

REFERENCES

ADAMS, W.T. & BIRKES, D.S., 1991: Estimating mating parameters in forest tree populations. *In:* Biochemical markers in the population genetics of forest trees. (eds. S. Fineschi, M.E. Malvolti, F. Cannata & H.H. Hattemer). p. 157–172. SPB Academic Publishing bv, The Hague.

BERGMANN, F. & SCHOLZ, F., 1989: Selection effects of air pollution in Norway spruce (*Picea abies*) populations. *In:* Genetic effects of air pollutants in forest tree populations. (eds. F. Scholz, H.-R. Gregorius & D. Rudin). p. 143–160. Springer-Verlag, Berlin – Heidelberg – New York.

CHELIAK, W.M., WANG, J. & PITEL, J.A., 1988: Population structure and genic diversity in tamarack, *Larix laricina* (Du Roi K. Koch). *Canadian Journal of Forest Research* **18**(10):1318–1324.

CONKLE, M.T., SCHILLER, G. & GRUNWALD, C., 1988: Electrophoretic analysis of diversity and phylogeny of *Pinus brutia* and closely related taxa. *Systematic Botany* **13**(3):411–424.

CROW, J.F. & KIMURA, M., 1970: An Introduction to Population Genetics Theory. Harper and Row, New York.

CUGUEN, J., 1986: Differentiation génétique inter- et intra-population d'un arbre forestier anémophile: le cas du hêtre. Thèse, Université des Sciences et Techniques du Lan-guedoc, Montpellier, 83+77 pp.

DAVIESR.G. 1971: Computer Programming in Quantitative Biology. Academic Press, London – New York, 492 pp.

GEBUREKT. & SCHOLZ, F., 1985: Über Selektionswirkungen bei Forstpflanzenpopulationen infolge von Luftverunreinigungen. *Forstarchiv* **56**(6):234–238.

JACOBS, B.F., WERTH, C.R. & GUTTMANN, S.I., 1984: Genetic relationships in *Abies* (fir) of eastern United States: an electrophoretic study. *Canadian Journal of Botany* **62**:609–616.

SWOFFORD, D.L. & SELANDER, R.B., 1981: BIOSYS-1: a Fortran program for the comprehensive analysis of electrophoretic data in population genetics and systematics. *Journal of Heredity* **72**:281–283.

Appendix I

Allele frequencies at each locus depend on the effective number of alleles n_{e_j} and the actual number of alleles n_{a_j}. If there are two alleles at the j-th locus, then the allele frequencies are function of the effective number of alleles. If $n_{e_j} = 1 / \sum_{i=1}^{2} q_{ij}^2$, then for the frequency of the first allele holds $q_{1j}^2 + (1 - q_{1j})^2 = 1 / n_{e_j}$, from which the frequency q_{1j} can be derived as:

$$q_{1j} = \frac{1}{2} + \sqrt{\frac{2 \cdot \frac{1}{n_{e_j}} - 1^2}{4}} \qquad [11]$$

whereas the frequency of the second allele is the complement to 1.

If the number of alleles n_{a_j} is higher than 2, then the frequencies of the first $(n_{a_j} - 2)$ alleles are not function of n_{e_j}, nevertheless, they are limited by this value. The frequency of the first allele is maximum, when the remaining $(n_{a_j} - 1)$ alleles are equally represented, i.e. their frequency is $\frac{1 - q_{1j\,\max}}{n_{a_j} - 1}$. Therefore, the upper limit of the frequency of the first simulated $q_{1j\,\max}$ allele can be derived from:

$$q_{1j\max}^2 + (n_{a_j} - 1) \cdot \left(\frac{1 - q_{1j\max}}{n_{a_j} - 1} \right)^2 = \frac{1}{n_{e_j}} \qquad [12]$$

The same holds for the lower limit of the frequency of the first allele; the frequency reaches its minimum, when the remaining alleles are equally represented. Based on equation [12], the limits for the frequency of the first allele are given by:

$$q_{1j} = \frac{1}{n_{a_j}} \pm \sqrt{\frac{(n_{a_j} - 1) \cdot \left(n_{a_j} \cdot \frac{1}{n_{e_j}} - 1^2 \right)}{n_{a_j}^2}} \qquad [13]$$

For the second, third *etc.* alleles, the frequency is limited also by frequencies of alleles which were simulated before. After simulating $(m - 1)$ alleles, the number of alleles n_{a_j} for the formula [13] is decreased to $[n_{a_j} - (m - 1)]$, remaining summary allele frequency 1 becomes $\left(1 - \sum_{i=1}^{m-1} q_{ij} \right)$, and the inverse value of the effective number of alleles must be subtracted of the sum of squared frequencies of the first $(m - 1)$ alleles: $1 / n_{e_j} - \sum_{i=1}^{m-1} q_{ij}^2$. The generalized formula then becomes:

$$q_{mj} = \frac{1 - \sum_{i=1}^{m-1} q_{ij}}{n_{a_j} - (m - 1)} \pm$$

$$\pm \sqrt{\frac{(n_{a_j} - m) \left[(n_{a_j} - m + 1) \cdot \left(\frac{1}{n_{e_j}} - \sum_{i=1}^{m-1} q_{ij}^2 \right) - \left(1 - \sum_{i=1}^{m-1} q_{ij} \right)^2 \right]}{[n_{a_j} - (m - 1)]^2}} \qquad [14]$$

Formula [4] can be derived from formula [11] using the same procedure.

© ARBORA PUBLISHERS