Theor Appl Genet (1991) 83:137–140

THEORETICAL AND APPLIED GENETICS

Joint estimation of immigration and mating system parameters in gymnosperms using the *EM* algorithm

C. Y. Xie *, F. C. Yeh, B. P. Dancik and C. Strobeck [1]

Department of Forest Science, University of Alberta, Edmonton, Alberta T6G 2H1, Canada
[1] Department of Zoology, University of Alberta, Edmonton, Alberta T6G 2H1, Canada

Received February 5, 1991; Accepted May 16, 1991
Communicated by P. M. A. Tigerstedt

Summary. An *EM* algorithm procedure is presented for the maximum-likelihood joint estimation of immigration and mating system parameters of mixed-mating system models for gymnosperms. In addition to accommodating multiallelic and multilocus data in mature populations and pollen pools, the *EM* estimates are insensitive to allelic frequency changes in foreign population and may approach closer to global maximum-likelihood estimates with each iteration, regardless of initial starting values. Estimates of rates of selfing ($\hat{S}$), outcrossing ($\hat{O}$), and immigration ($\hat{I}$) derived from the model are bounded strictly within the natural biological range (i.e., $0 \leq \hat{O} + \hat{I} \leq 1$; $\hat{S} + \hat{O} + \hat{I} = 1$).

Key words: Maximum-likelihood estimation – Immigration – Mating systems – Gymnosperms – *EM* algorithm

Introduction

Distribution of allozyme variation within and among plants is closely associated with their mating systems and gene dispersal mechanisms (e.g., Loveless and Hamrick 1984). Models that enrich the study of plant mating systems include multilocus estimation (Shaw et al. 1981; Neale and Adams 1985; Yeh and Morgan 1987; Xie et al. 1991), measurement of effective selfing (Ritland 1984), and measurement of differential male fertility (Schoen and Clegg 1984; Schoen and Cheliak 1987). Methods developed to estimate gene flow include movement of dispersal units or dispersal vectors (Levin and Kester 1974; Thomson and Plowright 1980; Waser and Price 1982), allelic frequency distribution between and among populations (Slatkin 1981), and paternity analysis (Smith and Adams 1983; Ellstrand 1984; Friedman and Adams 1985). An extension of measurements of mating system and gene flow is to jointly estimate their levels in plant populations (Adams and Birkes 1989; Adams and Birkes 1990). In this paper, we extend the multilocus Expectation-Maximization (*EM*) algorithm for mating systems (Xie et al. 1991) to include immigration for gymnosperms. Explicit expressions are given to determine rates of selfing, outcrossing, and immigration, and frequencies of multilocus haplotypes in local and foreign pollen pools.

Description of the model

The *EM* algorithm is an iterative procedure wherein each cycle consists of an expectation step (*E*) followed by a maximization step (*M*). In gymnosperms, sources of pollen that contribute to the pollen pool of a given population can be divided into pollen from maternal plants themselves (selfing), from other plants in the same population (local outcrossing), and from plants in foreign populations (outcrossing due to immigration), but it is not possible to know which event has occurred. An embryo that carries a nonmaternal pollen at any locus is categorized as discernibly outcrossed, and otherwise as ambiguous. Starting by assuming arbitrary values of unknowns, rates of selfing (S), immigration (I), and outcrossing (O), and frequencies of multilocus pollen genotypes in local (P_j) and foreign (i_j) gymnosperm populations, the *E* step derives the expected number of gametes from selfed and outcrossed events and the *M* step provides maximum-likelihood estimates for S, I, O, P_j, and i_j. These estimates are used in another *E* and *M*

* Present address: Forest Science Research Branch, Ministry of Forests, 31 Bastion Square, Victoria, British Columbia, Canada V8W 3E7

step, and the *EM* algorithm continues until successive estimates converge to a specified criterion. Data used to estimate S, I, O, P_j, and i_j using the *EM* algorithm are: (1) known or inferred maternal genotypes in local populations, (2) genotypes of open-pollinated progeny arrays from known or inferred maternal genotypes, (3) known maternal contribution to heterozygous embryos from heterozygous maternal plants of the same genotype, and (4) known or inferred genotypes for a sample of trees from foreign sources.

The assumptions of the mixed-mating model are that: (1) there are only two types of mating events (outcrossing and selfing); (2) there is homogeneity in allelic frequencies among cross-fertilization pollens over array of maternal plants sampled; (3) all embryos, regardless of mating event, have equal fitness; (4) maternal plants sampled outside a given population are representative of the foreign pollen pool; and (5) there is no association between alleles at different loci, that is, the pollen pool is not in gametic phase disequilibrium. We use the term "gametic disequilibrium" instead of "linkage disequilibrium" to refer to associations between independent, as well as linked, loci. Likelihood equations for complete sample space using the *EM* algorithm are derived from multinomial sampling. Given the above assumptions, the probability of observing a pollen with the j^{th} multilocus pollen haplotype in offspring of a mother plant in the local population is given by Eqs. 1 and 2.

$$P[G_M(j,k)] = S\phi(k) + Ii_j + Op_j \quad \text{Ambiguous (self or outcross)} \tag{1}$$

$$P[G_O(j)] = Ii_j + Op_j \quad \text{Observed outcrosses} \tag{2}$$

where $G_M(j,k)$ is the j^{th} multilocus pollen haplotype that could be produced by the maternal plant, which is assumed to have k heterozygous loci; $G_O(j)$ is the j^{th} multilocus pollen haplotype that could not be produced by the maternal plant; $\phi(k)$ is the product of segregation parameters of k heterozygous loci in the maternal plant, which can be estimated using haploid megagametophytes of heterozygous trees; S is the proportion of selfing; I is the proportion of outcrossing due to foreign pollen; O is the proportion of outcrossing due to local pollen; p_j is the frequency of the j^{th} multilocus pollen haplotype in the outcrossed pollen pool of the local population; and i_j is frequency of the j^{th} multilocus pollen haplotype in the foreign pollen pool.

Let the observed number of pollens, $G_M(j,k)$ and $G_O(j)$, in the pollen pool of a given population be $N_M(j,k)$ and $N_O(j)$, respectively. The expected numbers of pollen from selfing ($\hat{N}_S$) and outcrossing [local ($\hat{N}_O$) and foreign ($\hat{N}_I$)] are:

$$\hat{N}_S = \sum_{j=1}^{N}\sum_{k=0}^{n} \frac{S\phi(k)}{S\phi(k) + Ii_j + Op_j} N_M(j,k) \tag{3}$$

$$\hat{N}_O = \sum_{j=1}^{N}\left\{\sum_{k=0}^{n} \frac{Op_j}{S\phi(k)+Ii_j+Op_j} N_M(j,k) + \frac{Op_j}{Ii_j+Op_j} N_O(j)\right\} \tag{4}$$

$$\hat{N}_I = \sum_{j=1}^{N}\left\{\sum_{k=0}^{n} \frac{Ii_j}{S\phi(k)+Ii_j+Op_j} N_M(j,k) + \frac{Ii_j}{Ii_j+Op_j} N_O(j)\right\}, \tag{5}$$

where N is the total number of multilocus pollen haplotypes observed and n is the number of loci scored. The maximum-likelihood estimators of S, I, and O are:

$$\hat{S} = \frac{\hat{N}_S}{\hat{N}_S+\hat{N}_I+\hat{N}_O},\quad \hat{I} = \frac{\hat{N}_I}{\hat{N}_S+\hat{N}_I+\hat{N}_O},\quad \hat{O} = \frac{\hat{N}_O}{\hat{N}_S+\hat{N}_I+\hat{N}_O} \tag{6}$$

The expected number of pollen "j" in the local outcrossed pollen pool is:

$$\hat{N}_O(j) = \sum_{k=0}^{n} \frac{Op_j}{S\phi(k)+Ii_j+Op_j} N_M(j,k) + \frac{Op_j}{Ii_j+Op_j} N_O(j). \tag{7}$$

Thus, the corresponding frequency of the j^{th} multilocus pollen haplotype in the outcrossed pollen pool of the local population, $\hat{p}_j$, is:

$$\hat{p}_j = \frac{\hat{N}_O(j)}{\hat{N}_O}. \tag{8}$$

The frequency of pollen "j" in the foreign pollen pool, i_j, is estimated from the product of frequencies of alleles composing multilocus genotypes of this pollen. Let $A(l,r)$ be the l^{th} allele at the r^{th} locus and the expected number of $A(l,r)$ in the foreign pollen pool be $\hat{n}A(l,r)$. Then,

$$\hat{n}A(l,r) = FA(l,r) + \sum_{j=1}^{N}\left\{\sum_{k=0}^{n} \frac{Ii_j}{S\phi(k)+Ii_j+Op_j} N'_M(j,k) + \frac{Ii_j}{Ii_j+Op_j} N'_O(j)\right\}, \tag{9}$$

where $N'_M(j,k)$ and $N'_O(j)$ are, respectively, the observed number of pollen, $G_M(j,k)$ and $G_O(j)$, with allele $A(l,r)$ in the local population; and $FA(l,r)$ is the observed number of alleles $A(l,r)$ in the foreign population. Let the frequency of allele $A(l,r)$ be $\hat{a}(l,r)$. Thus,

$$\hat{a}(l,r) = \frac{\hat{n}A(l,r)}{\hat{N}_I + F_T}, \tag{10}$$

where F_T is the total number of alleles observed in the foreign population. Therefore, the frequency of pollen "j" in the foreign pollen pool is:

$$\hat{i}_j = \prod_{r}^{n} \hat{a}(l,r), \tag{11}$$

where $\hat{a}(l,r)$'s are frequencies of alleles composing the multilocus genotype of pollen "j".

The maximum-likelihood estimates for S, I, O, P_j, and i_j can be obtained by iterating the above estimation procedure until successive estimates stabilize. Initially, arbitrary values of S, I, O ($S+I+O=1$), p_j, and i_j are provided to begin the iterative procedure.

Analysis of data using the *EM* algorithm

We have reanalyzed the 1983 flowering year data from a clone bank of *Pseudotsuga menziesii* (Mirb.) Franco presented by Fast et al. (1986). In that study, rates of immigration (0.82) and outcrossing (0.56) were separately estimated from the models given by Smith and Adams (1983) and Neale and Adams (1985), respectively. The *EM* algorithm procedure gives a comparable selfing estimate of 0.50. However, the rate of immigration is significantly lower at 0.34. A large number of initial values of the set $\{S, I, O, p_j, i_j\}$ was used to test convergence of the iterative procedure to a stationary value or local or global maximum (Wu 1981). Regardless of the initial set, the *EM* algorithm procedure always converged to $\{\hat{S}, \hat{I}, \hat{O}, \hat{p}_j, \hat{i}_j\}$ for the estimate set. These results suggest that the present procedure is robust with respect to initial starting values, and that the values of $\hat{S}$, $\hat{I}$, $\hat{O}$, $\hat{p}_j$, and $\hat{i}_j$ are likely to be the global maximum-likelihood estimates.

In the Smith and Adams (1983) and Friedman and Adams (1985) models, the immigration rate, m, is estimated by dividing the proportion of observed immigrants, b, by the probability that a foreign (background) pollen grain has a distinguishable multilocus marker, d (i.e., $m=b/d$). Obviously, an underestimation of d will result in an overestimation of m. Fast et al. (1986) estimated d based on samples of trees from natural populations surrounding the seed orchard. However, pollen contributions from adjacent clone banks were ignored. Most likely, the sample of foreign genotypes in their study did not accurately reflect allele frequencies of all outside pollen sources.

We computed the seven most frequent foreign pollen multilocus haplotypes and their frequencies from background genotypes (trees from natural populations surrounding the seed orchard and from adjacent clone banks) for this clone bank. In theory, these seven foreign pollen multilocus haplotypes accounted for 25% of the total possible foreign multilocus haplotypes produced by background pollens, but none was observed in the 1983 data of Fast et al. (1986). The probability of not observing any of the seven most frequent foreign pollen multilocus haplotypes is very small, at 5.28×10^{-17}. This suggests that a part of the contribution of the foreign pollen pool was ignored and, consequently, that d must be an underestimation in Fast et al. (1986). Another way to show that d is underestimated is to compare the immigration rate (0.82) with the theoretical maximum immigration rate, which is the outcrossing rate (0.56). Obviously, the former is much larger than the latter. This indicates that the rate of immigration in this clone bank was overestimated by Fast et al. (1986).

In the *EM* algorithm procedure, the rate of pollen immigration is given by the expected number of immigrants, $\hat{N}_I$, divided by the expected total pollen array (i.e., $\hat{N}_S+\hat{N}_I+\hat{N}_O$). In contrast to sensitivity of the Smith and Adams model to d, the *EM* algorithm procedure is relatively stable with respect to allelic frequency changes in the foreign population. The expected number of the l^{th} allele at the r^{th} locus in the foreign pollen pool has two components (see Eq. 9). It is the sum of the observed number of alleles in the foreign population, $[FA(l, r)]$, and the expected number of pollen carrying that allele from outcrossing with foreign sources. In testing the sensitivity of the *EM* estimates to changes in allelic frequencies in the foreign population, we constrained $FA(l, r)$ in Eq. 9 to a wide range of values. The iterative procedure always converged close to the stationary value or local or presumed (likely) global maximum. Even in the extreme case when we set $FA(l, r)$ in Eq. 9 to zero (i.e., no information on the foreign population), the resulting estimate of 0.36 for proportion of outcrossing due to foreign pollen and 0.15 for proportion of outcrossing due to local pollen approximated the global maximum $\hat{I}$ (0.34) and $\hat{O}$ (0.16) values. It is always time-consuming and difficult to collect allelic frequency data from foreign (background) sources, especially for those nondisjunct populations. Thus, we believe that the apparent stability of $\hat{I}$ and $\hat{O}$ to changes in foreign allelic data is an appealing property of the *EM* algorithm procedure, for the joint estimation of mating system parameters and rate of immigration in gymnosperms.

One other advantage of the *EM* algorithm is its facility in coping with a high number of alleles at marker loci. It yields identical estimates for outcrossing (i.e., $\hat{I}+\hat{O}$) as with the strictly comparable maximum-likelihood analysis, provided the latter is less than unity (Brown et al. 1984). However, the *EM* estimate of outcrossing is bounded strictly within the natural biological range, between 0 and 1. Thus, when the *EM* estimate of outcrossing is equal to unity, its simultaneous estimates of local and foreign pollen allele frequencies will likely be biased.

Acknowledgements. We thank W. T. Adams, W. M. Cheliak, and H. R. Gregorius for their helpful comments on an earlier draft of this paper. This research was supported by grants from the Natural Sciences and Engineering Research Council of Canada, A0342 to B.P.D., A0502 to C.S., A2282 and IRC 8607 to F.C.Y., and F0001 to the Forest Genetics Group at the University of Alberta.

References

Adams WT, Birkes DS (1989) Mating patterns in seed orchards. Proc South For Tree Improv Conf 20:75–86

Adams WT, Birkes DS (1990) Estimating mating patterns in forest tree populations. In: Hattemer HH, Fineschi S (eds) Biochemical markers in the population genetics of forest trees. SPB Academic Publishing. The Hague, pp 157–172

Brown AHD, Barrett SCH, Moran GF (1984) Mating system estimation in forest tree: models, methods and meanings. In: Gregorius HR (ed) Population genetics in forestry. Springer, Berlin, pp 32–49

Ellstrand NC (1984) Multiple paternity within the fruits of the wild radish, *Raphanus sativus*. Am Nat 123:819–828

Fast W, Dancik BP, Bower RC (1986) Pollen contamination and mating system in a Douglas-fir clone bank. Can J For Res 16:1314–1319

Friedman ST, Adams WT (1985) Estimation of gene flow into two seed orchards of loblolly pine (*Pinus taeda* L). Theor Appl Genet 69:609–615

Levin DA, Kester HW (1974) Gene flow in seed plants. Evol Biol 7:139–220

Loveless MD, Hamrick JL (1984) Ecological determinants of genetic structure in plant populations. Annu Rev Ecol Syst 15:65–95

Neale DB, Adams WT (1985) The mating system in natural and shelterwood stands of Douglas-fir. Theor Appl Genet 71:201–207

Ritland K (1984) The effective proportion of self-fertilization with consanguineous matings in inbred populations. Genetics 106:139–152

Schoen DJ, Cheliak WM (1987) Male fertility variation in a polycross with Norway spruce, *Picea abies* (L.) Karst. Theor Appl Genet 74:554–559

Schoen DJ, Clegg MT (1984) Estimation of mating system parameters when outcrossing events are correlated. Proc Natl Acad Sci USA 81:5258–5262

Shaw DV, Kahler AL, Allard RW (1981) A multilocus estimator of mating system parameters in plant populations. Proc Natl Acad Sci USA 78:1298–1302

Slatkin M (1981) Estimating levels of gene flow in natural populations. Genetics 99:323–335

Smith DB, Adams WT (1983) Measuring pollen contamination in clonal seed orchards with the aid of genetic markers. Proc South For Tree Improv Conf 17:64–73

Thomson JD, Plowright RC (1980) Pollen carryover, vector rewards, and pollinator behavior with special reference to *Diervilla lonicera*. Oecologia 46:68–74

Waser NM, Price MV (1982) A comparison of pollen and fluorescent dye carryover by natural pollinators of *Ipomopsis aggregata*. Ecology 63:1168–1172

Wu C-H (1981) On the convergence of the EM algorithm. Tech Rep No 642. University of Wisconsin, Dept of Statistics, Madison/WI

Yeh FC, Morgan K (1987) Mating system and multilocus associations in a natural population of *Pseudotsuga menziesii* (Mirb.) Franco. Theor Appl Genet 73:799–808

Xie C-Y, Dancik BP, Yeh FC (1991) The mating system in natural populations of *Thuja orientalis* Linn. Can J For Res 21:333–339