J. Math. Biol. (1991) 29: 693–713

Journal of
Mathematical
Biology

Selection at a diallelic autosomal locus in a dioecious population

Joseph M. Szucs

Department of General Academics, Texas A&M University at Galveston, P.O. Box 1675, Galveston, TX 77553, USA

Received April 14, 1990; received in revised form August 2, 1990

Abstract. The model used is that of an infinite dioecious population with nonoverlapping discrete generations and random mating. If the fitnesses are constant and heterozygotes are viable, it is proved that the allelic frequencies converge to equilibria as the number of generations tend to infinity. The results complement those of Karlin and Lessard [1] and Selgrade and Ziehe [5] in that hyperbolicity of equilibria is not assumed, use of index theory is avoided and it is determined how the number of equilibria and phase portraits depend on the fitnesses in the most general case. Lessard [2] gives, in the same situation, a condensed proof of convergence of allelic frequencies off the separatrix under the hypothesis that 1 is not an eigenvalue at any equilibrium. Our method of study is elementary.

Key words: Equilibrium – Global convergence – Diallelic autosomal locus – Selection

1 Introduction

In this paper we study the inheritance dynamics of a diallelic autosomal locus in an infinite dioecious population with nonoverlapping discrete generations and random mating (cf. [1, 2, 3, 5]). We will assume that the heterozygotic genotypes are viable and the fitnesses are constant. We will prove that the allelic frequencies converge to equilibria (as the number of generations increases to infinity) and we will determine how the number of equilibria and the phase portraits of the transformation T describing transition of allelic frequency ratios from a generation to the next generation depend on the fitnesses. Such results, with a different kind of analysis of the dependence of the phase portraits on the fitnesses, have been obtained by Karlin and Lessard [1] and Selgrade and Ziehe [5] under the hypothesis that the equilibria are hyperbolic. Furthermore, Lessard [2] has given a condensed proof of convergence of allelic frequencies off the separatrix under the hypothesis that 1 is not an eigenvalue of T' at any equilibrium. In this paper we do not make any explicit assumption on the eigenvalues of T'. Instead, we will prove through a multiplicity argument that neither eigenvalue can be equal to 1 at a polymorphic equilibrium that is unstable for neighboring initial allelic frequencies which are either both smaller or both larger than those of the equilibrium. It will follow from other arguments that the smaller eigenvalue is not smaller than -1 and not larger than 1. If it is -1, then it is easy to prove by observing how T transforms areas near the equilibrium that the domain of attraction of the equilibrium is a curve as described in case $(\leftarrow\rightarrow)$ of this Introduction (cf. Fig. 4). If the smaller eigenvalue is larger than -1, then a generalization of the stable manifold theorem [6] to maps that are not necessarily one-to-one [4, 5] implies easily that the domain of attraction of the equilibrium is a curve again as described in case $(\leftarrow\rightarrow)$ of this Introduction. (We could avoid the use of [4] at the expense of length.)

The allowance of equilibria with the eigenvalue 1 introduces two new types of phase portrait absent in [1] and [5]. In one of these types a polymorphic equilibrium is present that is stable (unstable) for neighboring initial allelic frequencies that are smaller (larger) than those of the equilibrium or vice versa (cf. cases $(\rightarrow\rightarrow)$ and $(\leftarrow\leftarrow)$ of this Introduction and Fig. 3). Such an equilibrium has multiplicity 2 and is not hyperbolic. The other new type of phase portrait occurs when there are infinitely many equilibria (cf. case $(-)$ of this Introduction and Fig. 5).

Our method of study is elementary. We avoid the use of index theory (cf. [5]) and even the stable manifold theorem (cf. [4, 6]) is needed only to prove that convergence to a hyperbolic unstable polymorphic equilibrium occurs on a curve (cf. case $(\leftarrow\rightarrow)$ of this Introduction) rather than in a domain between two curves.

The following is a list of the types of phase portraits, much of which is known (cf. [1, 2, 3, 5]). In the description below we use the "infinite" square $S = \{(x, y) : 0 \leqslant x, y \leqslant \infty\}$ as our state space, where $x = Q/(1-Q)$, $y = q/(1-q)$, Q being the frequency of one of the alleles at the locus under consideration in the female population and q the frequency of the same allele in the male population. We will use the notation $(\rightarrow)$, $(\rightarrow\leftarrow)$, etc. to indicate the nonexistence of polymorphic equilibria or the existence of a polymorphic equilibrium of a certain kind. This one-dimensional notation is justified by the fact that the phase portraits listed below are essentially one-dimensional. Of course, we will use the agreement that the equilibrium $(0, 0)$ precedes the equilibrium (∞, ∞).

At equilibrium, if an allele is fixed in one of the sexes, then it is also fixed in the other sex. In the list of types of phase portraits below, note that two unstable polymorphic equilibria never occur, and consequently, combination of $(\leftarrow\rightarrow)$ with $(\rightarrow\rightarrow)$ or $(\leftarrow\leftarrow)$ does not occur. Furthermore, if both fixation equilibria are stable, then there is a unique polymorphic equilibrium.

Cases $(\rightarrow)$ or $(\leftarrow)$. There is no polymorphic equilibrium. In this case there is a stable fixation equilibrium and possibly an unstable one. If the allelic frequencies are not associated with the unstable fixation equilibrium initially, then they converge to the allelic frequencies of the stable fixation equilibrium. In case $(\rightarrow)$ there is global convergence to (∞, ∞) except that $(0, 0)$ may be an equilibrium. In case $(\leftarrow)$ there is global convergence to $(0, 0)$ except that (∞, ∞) may be an equilibrium (Fig. 1).

$(\rightarrow\leftarrow)$ There is a unique polymorphic equilibrium G, which is locally stable. Unless the allelic frequencies are associated with a fixation equilibrium initially, they converge to the allelic frequencies of G (Fig. 2).

$(\rightarrow\rightarrow)$ There is a polymorphic equilibrium G that is locally stable (unstable) for initial allelic frequency ratios x, y that are smaller (larger) than those of G. There is a strictly decreasing continuous curve Γ in S, passing through G,

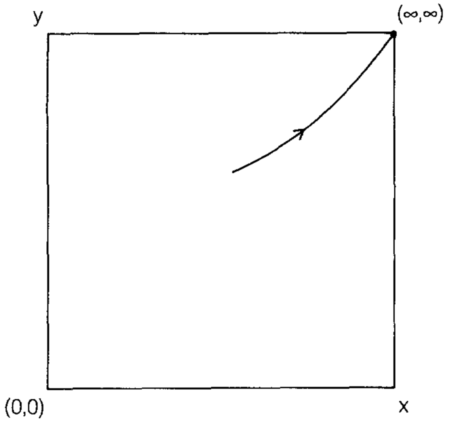

Fig. 1. No polymorphic equilibrium exists

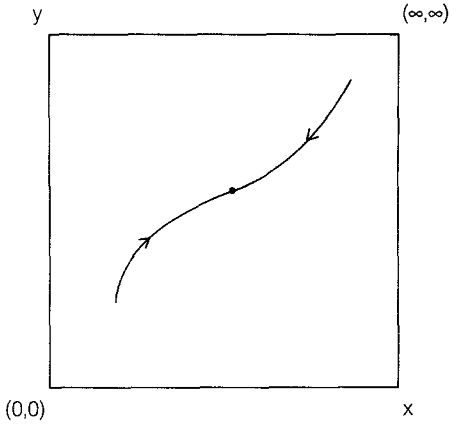

Fig. 2. The unique polymorphic equilibrium is locally stable

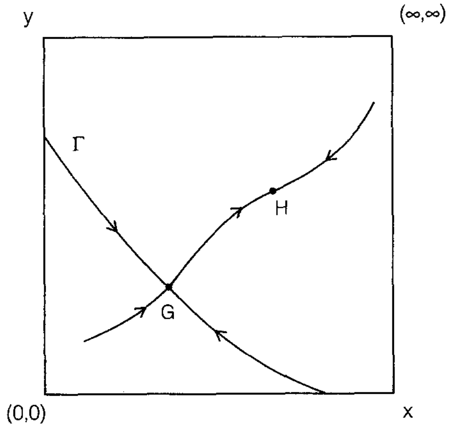

Fig. 3. A polymorphic equilibrium exists that is stable (unstable) from below (above)

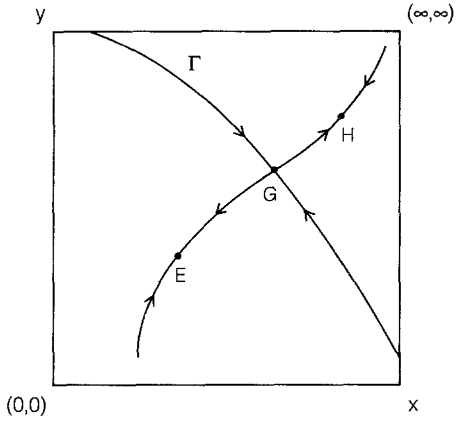

Fig. 4. A polymorphic equilibrium exists that is unstable from both above and below

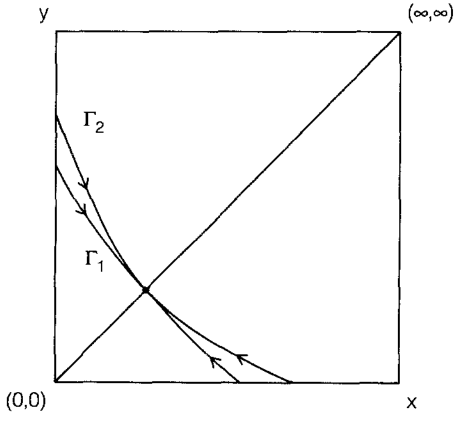

Fig. 5. There are infinitely many equilibria

separating S and having the following properties: if the initial gene frequencies correspond to a point on or below (above) Γ and the genetic state of the population is not in equilibrium initially, then the gene frequencies converge to those of G (to the gene frequencies of another equilibrium H for which $G < H$ coordinatewise) (Fig. 3).

($\leftarrow\leftarrow$) This case is the "mirror image" of ($\rightarrow\rightarrow$), where "below" is interchanged with "above" and "smaller" with "larger"; we will denote the equilibria by E and G, $E \leqslant G$.

($\leftarrow\rightarrow$) There is a polymorphic equilibrium G that is locally unstable for initial frequencies (x, y) that are either smaller or larger than those of G. Then there is a strictly decreasing continuous curve Γ in S, going through G, separating S and possessing the following properties: (a) if the point corresponding to the initial gene frequencies lies on Γ, then the gene frequencies converge to those of G; (b) if the point corresponding to the initial gene frequencies is below (above) Γ and the genetic state of the population is not in equilibrium initially, then the gene frequencies converge to an equilibrium E (H) for which $E < G$ ($G < H$) coordinatewise (Fig. 4).

($-$) The equilibria form an infinite line segment connecting $(0, 0)$ with (∞, ∞) (and including them). There is global convergence to equilibria. The domain of attraction of a polymorphic equilibrium G is bounded by two strictly decreasing continuous curves Γ_1 and Γ_2, going through G and separating S. The curves Γ_1 and Γ_2 are included in the domain of attraction of G and Γ_1 passes below Γ_2 in the wider sense. The domain of attraction of a fixation equilibrium is the singleton containing the equilibrium itself (Fig. 5).

In Sect. 2 we discuss the model, cases of extinction, and fixation equilibria. Section 3 contains the statement of our results. Theorem 1 concerns the case where the two sexes are alike in some sense, and Theorem 2 treats the remaining cases. The subsequent sections contain the proofs of Theorems 1 and 2. Section 4 is devoted to the study of the curves in the (x, y)-plane where the allelic frequencies do not change in the male or female sex in one generation. In this section we prove also equations for the gene frequency ratios at equilibrium. Section 5 contains a lemma involving convergence of gene frequencies on a strictly decreasing continuous curve in the (x, y)-plane that separates S, is symmetric with respect to the diagonal $x = y$, contains an equilibrium, and is invariant under the transformation T associated with transition from a generation to the next one. This lemma is used in the proofs of both Theorems 1 and 2 to prove convergence of allelic frequencies in the case ($\leftarrow\rightarrow$). Section 6 contains the proof of Theorem 1 and Sect. 7 the proof of Theorem 2.

For the convenience of the reader we have included two appendices containing a result of Selgrade and Ziehe [5] (Appendix A) and a result of Karlin and Lessard [1] (Appendix B), including proofs. Appendix A contains a result stating that T^2 does not have fixed points in some regions of S. (Of course, all fixed points of T^2 in S are also fixed points of T. However, we need Theorems 1 and 2 to see this.) Appendix A is crucial in the proof of the lemma in Sect. 5. Appendix B excludes the possibility of three polymorphic equilibria in case the fixation equilibria are locally stable and the two sexes are unlike in some sense. This result is used in Theorem 2 to exclude some a priori possible phase portraits and to prove that the larger eigenvalue is larger than 1 at a polymorphic equilibrium that is unstable for neighboring initial allelic frequencies which are either both smaller or larger than those of the equilibrium (case ($\leftarrow\rightarrow$)).

2 Model, extinction, and fixation equilibria

Following Nagylaki [3] or Karlin and Lessard [1] or Selgrade and Ziehe [5], we first construct a model for the inheritance dynamics of a diallelic autosomal locus in an infinite dioecious population with nonoverlapping discrete generations and random mating. Then we identify all cases of extinction and give a complete description of fixation equilibria. The results of this section are simple and mostly known.

Let $\boldsymbol{P}$ and $\boldsymbol{Q}$ be the two alleles of a diallelic autosomal locus in an infinite dioecious population with nonoverlapping discrete generations. Let P (Q) denote the frequency of $\boldsymbol{P}$ ($\boldsymbol{Q}$) in the gametic output of the female population and let p (q) have an analogous meaning for the male population. Suppose mating is random. Then the frequency of unordered genotype $\boldsymbol{PQ}$ ($\boldsymbol{PP}$, $\boldsymbol{QQ}$) in the female and in the male population immediately after fertilization is $Pq + Qp$ (Pp, Qq). Let W_{PQ}, W_{PP}, and W_{QQ} be the fitnesses of female genotypes $\boldsymbol{PQ}$, $\boldsymbol{PP}$, $\boldsymbol{QQ}$, respectively. Let w_{PQ}, w_{PP} and w_{QQ} denote the analogous fitnesses in the male population. We will assume that the fitnesses are constant. The quantities $W = W_{PP}Pp + W_{PQ}(Pq + Qp) + W_{QQ}Qq$ and $w = w_{PP}Pp + w_{PQ}(Pq + Qp) + w_{QQ}Qq$ are the mean female and mean male fitnesses, respectively. Suppose that the population does not become extinct, i.e., $Ww > 0$. Then the frequency P^* (Q^*) of $\boldsymbol{P}$ ($\boldsymbol{Q}$) in the gametic output of the next generation is

$$\begin{aligned} P^* &= (2W)^{-1}[2W_{PP}Pp + W_{PQ}(Pq + Qp)], \\ Q^* &= (2W)^{-1}[2W_{QQ}Qq + W_{PQ}(Pq + Qp)]. \end{aligned} \tag{1a}$$

The analogous equations for the male population are

$$\begin{aligned} p^* &= (2w)^{-1}[2w_{PP}Pp + w_{PQ}(Pq + Qp)], \\ q^* &= (2w)^{-1}[2w_{QQ}Qq + w_{PQ}(Pq + Qp)]. \end{aligned} \tag{1b}$$

Let us assume that $W_{PQ}w_{PQ} > 0$ and introduce the new parameters $a = 2W_{QQ}/W_{PQ}$, $b = 2W_{PP}/W_{PQ}$, $c = 2w_{QQ}/w_{PQ}$, $d = 2w_{PP}/w_{PQ}$ and the new variables $x = Q/P$, $y = q/p$. If $Ww > 0$ in all generations, i.e., if no generation dies out, then Eqs. (1) are equivalent to

$$x^* = \frac{axy + x + y}{b + x + y}, \qquad y^* = \frac{cxy + x + y}{d + x + y} \tag{2}$$

if we allow $x = \infty$ and $y = \infty$ with the agreement that $0 \cdot \infty = \infty \cdot 0 = 0$, and that x^* and y^* are calculated by taking double limits if they are not defined algebraically. (The double limits always exist in the absence of extinction.)

Extinction occurs if and only if $Ww = 0$ in a generation. Since it is assumed that $W_{PQ}w_{PQ} > 0$, the vanishing of W (w) implies that $Pq + Qp = 0$, i.e., either $P = p = 0$ or $Q = q = 0$. If $P = p = 0$, then $W = 0$ ($w = 0$) is equivalent to $W_{QQ} = 0$ ($w_{QQ} = 0$) and if $Q = q = 0$, then it is equivalent to $W_{PP} = 0$ ($w_{PP} = 0$). Therefore, extinction occurs if and only if $P = p = W_{QQ}w_{QQ} = 0$ or $Q = q = W_{PP}w_{PP} = 0$.

On the other hand, if $W^*w^* = 0$, then by what we have just proved, either $P^* = p^* = W_{QQ}w_{QQ} = 0$ or $Q^* = q^* = W_{PP}w_{PP} = 0$. Since $W_{PQ}w_{PQ} > 0$, from $P^* = p^* = W_{QQ}w_{QQ} = 0$ we obtain through (1) that $Pq + Qp = 0$, which now implies $P = p = 0$ because $Q = q = 0$ is impossible for then $Q^* = q^* = 0$ by (1), which contradicts $P^* = p^* = 0$. Summing up, $P^* = p^* = W_{QQ}w_{QQ} = 0$ is impossible since it implies the extinction of the preceding generation. So does

$Q^* = q^* = W_{PP}w_{PP} = 0$ for similar reasons. Therefore, we have proved the proposition below.

Proposition 1 *Extinction may occur only in the first generation. It occurs if and only if $P = p = W_{QQ}w_{QQ} = 0$ or $Q = q = W_{PP}w_{PP} = 0$. In terms of a, b, c, d, x and y, extinction occurs if and only if $x = y = \infty$, $ac = 0$ or $x = y = 0$, $bd = 0$.*

The right-hand sides of Eqs. (2) define functions that are jointly continuous in x and y in the "square" $\boldsymbol{S} = \{(x, y) : 0 \leqslant x, y \leqslant \infty\}$. Therefore, x^* and y^* are well-defined even in the presence of extinction. It is easy to verify that x^* and y^* are "strongly increasing" [5] functions of x and y in $\mathrm{int}\boldsymbol{S} = \{(x, y) : 0 < x, y < \infty\}$ provided that $(a + b)(c + d) > 0$. By this it is meant that if (x, y) and (X, Y) belong to $\mathrm{int}\boldsymbol{S}$ and $x \leqslant X$, $y \leqslant Y$ with at least one strict inequality, then $x^* < X^*$, $y^* < Y^*$.

Equations (2) have two types of symmetry. One concerns the interchange of $\boldsymbol{P}$ and $\boldsymbol{Q}$ and is given by the substitutions $a \leftrightarrow b$, $c \leftrightarrow d$, new $x = 1/(\text{old } x) = \text{old } P/\text{old } Q$, new $y = 1/(\text{old } y) = \text{old } p/\text{old } q$. The other concerns the interchange of sexes and is given by the substitutions $a \leftrightarrow c$, $b \leftrightarrow d$, $x \leftrightarrow y$.

If $b > 0$, then $x^* > 0$ or $x^* = 0$ according as $x + y > 0$ or $x + y = 0$. If $b = 0$, then $x^* = 1$ if $x = 0$ or $y = 0$. Consequently, $x + y > 0$ always implies $x^* > 0$. By symmetry, $x^* < \infty$ provided that $x < \infty$ or $y < \infty$. On the other hand, $a > 0$, $x = y = \infty$ implies $x^* = \infty$. If $a = 0$, then $x^* = 1$ provided that $x = \infty$ or $y = \infty$. Similar statements hold for y^*.

The observations of the preceding paragraph imply the following.

Proposition 2 *In a fixation equilibrium the same allele is fixed in both sexes. An allele participates in a fixation equilibrium if and only if both of its homozygotic fitnesses are different from zero. In terms of x, y, a, b, c, and d, the points $(0, 0)$ and (∞, ∞) are the only possible equilibria on the boundary of $\boldsymbol{S}$. Furthermore, $(0, 0)$ ((∞, ∞)) is an equilibrium if and only if $bd > 0$ ($ac > 0$).*

In this paragraph we are going to discuss how our notation translates into the notation of Karlin and Lessard [1] and of Selgrade and Ziehe [5]. Our symbols $\boldsymbol{P}$, $\boldsymbol{Q}$, P, Q, p, q, x, y, a, b, c, d correspond to A_2, A_1, $1 - p$, p, $1 - q$, q, x, y, $2f_1$, $2f_2$, $2m_1$, $2m_2$ in [1] and to A_2, A_1, $p_2^{♀}$, $p_1^{♀}$, $p_2^{♂}$, $p_1^{♂}$, x_1, x_2, a, b, c, d in [5].

3 Main results

In this section we state the main results of our paper. Theorem 1 concerns the case where the two sexes are alike in some sense (either $a + b \leqslant ab$, $c + d \leqslant cd$ or $a + b \geqslant ab$, $c + d \geqslant cd$). This is by far the simpler case. Theorem 2 treats the remaining cases.

It is clear from the first equation in (2) that any polymorphic equilibrium (x, y) satisfies the equation

$$[(1 - a)x - 1]y = -x^2 - (b - 1)x \tag{3}$$

and we will prove in the next section that if $a + b > 0$, then the x-coordinate of a polymorphic equilibrium satisfies the equation

$$a_3x^3 + a_2x^2 + a_1x + a_0 = 0,$$

$$a_3 = a + c - ac,$$

$$\begin{aligned} a_2 &= (c-1)(a+b-ab)+(a-1)(d-a)+2(b-c), \\ a_1 &= -[(d-1)(a+b-ab)+(b-1)(c-b)+2(a-d)], \\ a_0 &= -[b+d-bd]. \end{aligned} \tag{4}$$

Theorem 1 *The following is a complete nonoverlapping list of types of phase portraits where $a+b \geqslant ab$, $c+d \geqslant cd$ or $a+b \leqslant ab$, $c+d \leqslant cd$.*

Case ($\rightarrow$) of Introduction occurs if $(a+b)(c+d)>0$ and either $1/a+1/b \geqslant 1$, $1/c+1/d \geqslant 1$, $1/b+1/d>1$, $1/a+1/c \leqslant 1$ or $a+b \leqslant 1$, $1/c+1/d \leqslant 1$, $1/b+1/d \geqslant 1$, $1/a+1/c<1$.

Case ($\leftarrow$) occurs if $(a+b)(c+d)>0$ and either $1/a+1/b \geqslant 1$, $1/c+1/d \geqslant 1$, $1/b+1/d \leqslant 1$, $1/a+1/c>1$ or $1/a+1/b \leqslant 1$, $1/c+1/d \leqslant 1$, $1/b+1/d<1$, $1/a+1/c \geqslant 1$.

Case ($\rightarrow\leftarrow$) occurs if either $(a+b)(c+d)=0$ or $(a+b)(c+d)>0$, $1/a+1/b \geqslant 1$, $1/c+1/d \geqslant 1$, $1/b+1/d>1$, $1/a+1/c>1$. If here $a+b>0$, then the x-coordinate of the polymorphic equilibrium is the only positive solution of (4) for which $1-b<x$ if $b<1$ and $x<1/(1-a)$ if $a<1$; the y-coordinate can be calculated from (3). If $a+b=0$, then

$$\left(1, \frac{c-d+\sqrt{(c-d)^2+4}}{2}\right)$$

is the polymorphic equilibrium.

Case ($\leftarrow\rightarrow$) with $E=(0,0)$, $H=(\infty,\infty)$ occurs if $1/a+1/b \leqslant 1$, $1/c+1/d \leqslant 1$, $1/b+1/d<1$, $1/a+1/c<1$. The x-coordinate of the polymorphic equilibrium is the unique positive solution of (4) and the y-coordinate can be calculated from (3).

Case ($-$) occurs if and only if $a=d>1$, $b=c=(a-1)/a$. The equilibria correspond to all points of the line segment $y=(b-1)x$, $0 \leqslant x \leqslant \infty$.

Remark. The relation $1/a+1/b=1$ (<1, >1) is equivalent to $a+b-ab=0$ (<0, >0) if $a+b>0$.

For the statement of the next theorem we need the discriminant

$$\delta = -4a_0a_2^3 + a_1^2a_2^2 + 18a_0a_1a_2a_3 - 27a_0^2a_3^2 - 4a_1^3a_3$$

of Eq. (4) in case $a_3 \neq 0$. If $a_3=0$, $a_2 \neq 0$, then the sign of δ is equal to the sign of the discriminant of the quadratic Eq. (4). Similarly, if $a_0=0$, $a_1 \neq 0$, then the discriminant of the quadratic equation for $1/x$, obtained from (4) by division by x^3 has the same sign as δ. We will also use the following conditions in Theorem 2: Condition ($+-$) will mean that the first nonvanishing coefficient in the sequence a_0, a_1, a_2, a_3 is positive and the last nonvanishing coefficient is negative. The symbols ($-+$), ($--$) and ($++$) will have similar meanings.

Theorem 2 *Let $a+b-ab<0$, $c+d-cd>0$ throughout this theorem. The following is a complete nonoverlapping list of types of phase portraits.*

Case ($\rightarrow$) of Introduction occurs when: (i) *all a_i's are nonpositive*; (ii) *in case ($--$) there is at least one positive a_i and $\delta<0$.*

Case ($\leftarrow$) occurs when: (i) *all a_i's are nonnegative*; (ii) *in case ($++$) there is at least one negative a_i and $\delta<0$.*

Case ($\rightarrow\leftarrow$) occurs when: (i) *in case ($-+$) we do not have $a_0<0$, $a_1>0$, $a_2<0$, $a_3>0$*; (ii) *$a_0<0$, $a_1>0$, $a_2<0$, $a_3>0$, $\delta<0$*; (iii) *$a_0<0$, $a_1>0$, $a_2<0$, $a_3>0$, $\delta=0$, $a_2^2-3a_1a_3=0$.*

Case $(\rightarrow\rightarrow)$ *occurs when*: (i) *in case* $(--)$ *there is a positive* a_i *and* $\delta = 0$ (*in this case* $H = (\infty, \infty)$); (ii) $a_0 < 0$, $a_1 > 0$, $a_2 < 0$, $a_3 > 0$, $\delta = 0$, $a_2^2 - 3a_1a_3 \neq 0$ *and the double root of* (4) *precedes the simple one* (*in this case* H *is the polymorphic equilibrium whose* x*-coordinate is the larger* (*simple*) *root of* (4) *and* $(\rightarrow\rightarrow)$ *can be refined to* $(\rightarrow\rightarrow\leftarrow)$).

Case $(\leftarrow\leftarrow)$ *occurs when*: (i) *in case* $(++)$ *there is negative* a_i *and* $\delta = 0$ (*in this case* $E = (0, 0)$); (ii) $a_0 < 0$, $a_1 > 0$, $a_2 < 0$, $a_3 > 0$, $\delta = 0$, $a_2^2 - 3a_1a_3 \neq 0$ *and the double root of* (4) *exceeds the simple one* (*in this case* E *is the polymorphic equilibrium whose* x*-coordinate is the smaller* (*simple*) *root of* (4) *and* $(\leftarrow\leftarrow)$ *can be refined to* $(\rightarrow\leftarrow\leftarrow)$).

Case $(\leftarrow\rightarrow)$ *occurs when*: (i) $a_0 < 0$, $a_1 > 0$, $a_2 < 0$, $a_3 > 0$, $\delta > 0$ (*in this case* E, G *and* H *are all polymorphic equilibria and* $(\leftarrow\rightarrow)$ *can be refined to* $(\rightarrow\leftarrow\rightarrow\leftarrow)$); (ii) *case* $(+-)$ *holds* (*then* $E = (0, 0)$ *and* $H = (\infty, \infty)$ *and therefore*, G *is the only polymorphic equilibrium*); (iii) *in case* $(--)$ *there is a positive* a_i *and* $\delta > 0$ (*then* E *and* G *are polymorphic*, $H = (\infty, \infty)$ *and* $(\leftarrow\rightarrow$ *can be refined to* $(\rightarrow\leftarrow\rightarrow)$); (iv) *in case* $(++)$ *there is a negative* a_i *and* $\delta > 0$ (*then* G *and* H *are polymorphic*, $E = (0, 0)$ *and* $(\leftarrow\rightarrow)$ *can be refined to* $(\leftarrow\rightarrow\leftarrow)$).

The positive solutions of (4) *are identical with the* x*-coordinates of all polymorphic equilibria and the* y*-coordinates can be calculated from* (3).

If $a + b > ab$, $c + d < cd$, *then results similar to those in this theorem can be obtained by interchanging the roles of* x *and* y, *which also entails the interchanges* $a \leftrightarrow c$, $b \leftrightarrow d$.

4 Equations for equilibria and study of the curves where x or y does not change in one generation

In this section we will prove that (4) in the preceding paragraph is indeed an equation for the x-coordinate of a polymorphic equilibrium in case $a + b > 0$ and study the curve (3) in int$\boldsymbol{S}$; this curve contains exactly these points (x, y) for which $x^* = x$.

Equation (3) restricted to $(x, y) \in \text{int}\boldsymbol{S}$ is indeed an equation for exactly those points for which $x^* = x$. This can be seen by substituting x for x^* in the first equation in (2) and then multiplying by $b + x + y > 0$ and rearranging. Let us perform the interchanges $x \leftrightarrow y$, $a \leftrightarrow c$, $b \leftrightarrow d$ in (3) and then substitute, in the equation $[(1 - c)y - 1]x = -y^2 - (d - 1)y$ thus obtained, the following expression of y obtained from (3):

$$y = \frac{-x^2 - (b - 1)x}{(1 - a)x - 1}. \tag{5}$$

Upon bringing to the common denominator $[(1 - a)x - 1]^2$ and rearranging, we obtain

$$[(1 - a)x - 1]^{-2}x\{x(b + x - 1)^2 - [(1 - c)x + d - 1](b + x - 1)[(1 - a)x - 1] \\ - [(1 - a)x - 1]^2\} = 0.$$

The at most cubic equation obtained by equating the expression in the curly brackets with zero becomes Eq. (4) after rearrangement.

Let $0 < x, y < \infty$. If $[(1 - a)x - 1] = 0$, then (3) is equivalent to $x = 1/(1 - a) = 1 - b$, which in turn is equivalent to $a = b = 0$ because we must have $a, b < 1$. Therefore, if $a + b > 0$, then in int$\boldsymbol{S}$, (3) is equivalent to (5), and consequently, the polymorphic equilibria are exactly those points $(x, y) \in \mathrm{int}\boldsymbol{S}$ for which x is a solution of (4) and the right-hand side of (5) calculated from this x is y (and, therefore, positive).

Let $1/(1-a) = 1-b$, i.e., $a+b = ab$. Then (3) is equivalent to the pair of equations $x = 1-b$, $y = (b-1)x$. If here $ab = 0$, then $a = b = 0$. Let Γ_x denote the line $x = 1$ in this case (Fig. 6). If $ab > 0$ (and $a+b = ab$), then $a, b > 1$. Let Γ_x denote the line $y = (b-1)x$ in this case (Fig. 6). If $a = 1$, then (3) is equivalent to the equation $y = x^2 + (b-1)x$, which is the equation of a concave-up parabola passing through the points $(0, 0)$ and $(1-b, 0)$. Let us denote this parabola by Γ_x. In this case $\Gamma_x \cap \mathrm{int}\boldsymbol{S}$ is an increasing portion of a concave-up parabola starting from $(0, 0)$ if $b \geqslant 1$ and from $(1-b, 0)$ if $b < 1$ (Fig. 7). If $a \neq 1$ and $a+b \neq ab$, then (3) is equivalent to (5), which in turn is equivalent to

$$y = \frac{1}{1-a}x - \frac{a+b-ab}{(1-a)^2} - \frac{a+b-ab}{(1-a)^2[(1-a)x-1]},$$

which is the equation of a hyperbola with the asymptotes

$$x = \frac{1}{1-a}, \qquad y = -\frac{1}{(1-a)}x - \frac{a+b-ab}{(1-a)^2}.$$

If $a+b < ab$, then $a, b > 1$, the vertical asymptote misses the half-plane $x \geqslant 0$, the slope of the nonvertical asymptote is positive and its y-intercept is also positive. Let Γ_x denote the right-hand side branch of this hyperbola. Since $(0, 0)$ obviously satisfies (5), in this case $\Gamma_x \cap \mathrm{int}\boldsymbol{S}$ is a portion of an increasing concave-down hyperbola branch approaching $(0, 0)$ (Fig. 8). If $a+b > ab$, $a > 1$, then the vertical asymptote misses the half-plane $x \geqslant 0$, the slope of the nonvertical asymptote is positive and its y-intercept is negative. Let Γ_x denote the right-hand side hyperbola branch again. Since $(0, 0)$ satisfies (5), in this case $\Gamma_x \cap \mathrm{int}\boldsymbol{S}$ is an increasing portion of a concave-up hyperbola branch approaching $(0, 0)$ if $b \geqslant 1$ and $(1-b, 0)$ if $b < 1$ (Fig. 9). If $a+b > ab$, $a < 1$, then the vertical asymptote of the hyperbola (5) passes through the half-plane $x > 0$, the slope of the nonvertical asymptote is negative and its y-intercept is also negative. Let Γ_x be the left-hand side hyperbola branch. Since $(0, 0)$ obviously satisfies (5),

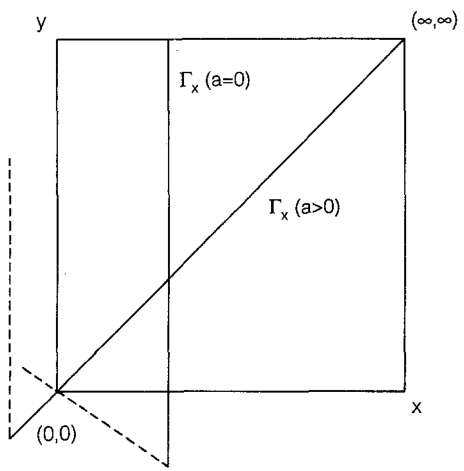

Fig. 6. Γ_x for $a+b=ab$

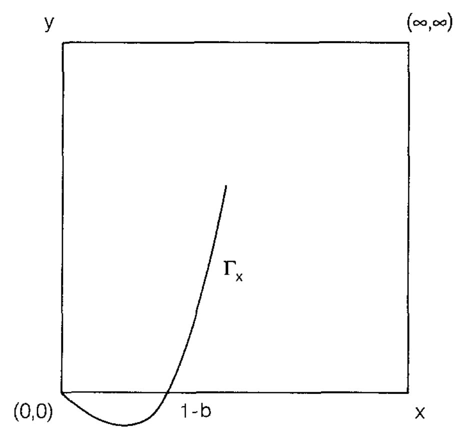

Fig. 7. Γ_x for $a=1$

in this case $\Gamma_x \cap \text{int}\boldsymbol{S}$ is an increasing portion of a concave-up hyperbola branch approaching $(0, 0)$ if $b \geqslant 1$ and $(1-b, 0)$ if $b < 1$ (Fig. 10).

It is clear that the curve Γ_x is defined in such a way that $\Gamma_x \cap \text{int}\boldsymbol{S}$ coincides with the curve $x^* = x$ in $\text{int}\boldsymbol{S}$.

The curve Γ_y can be calculated from Γ_x by interchanging the roles of x and y.

The following simple observation is crucial in the proofs of both Theorems 1 and 2. Since $x^* \geqslant x$ on the line $x = 0$ (otherwise $\boldsymbol{S}$ would not be invariant under T), in $\text{int}\boldsymbol{S}$ we have $x^* > x$ to the left of $\Gamma_x \cap \text{int}\boldsymbol{S}$. Similarly, $\text{int}\boldsymbol{S}$, $y^* > y$ below $\Gamma_y \cap \text{int}\boldsymbol{S}$. Since in $\text{int}\boldsymbol{S}$, both Γ_x and Γ_y are graphs of strictly increasing continuous functions (the only anomaly occurring for $a = b = 0$ $(c = d = 0)$, when Γ_x (Γ_y) is a vertical (horizontal) line), a combination of the observations above can be worded to yield the following:

Lemma 1 *Let* $(a+b)(c+d) > 0$. *For* $z \in \text{int}\boldsymbol{S}$ *we have* $z^* > z$ $(z^* < z)$ *if* z *is above* (*below*) $\Gamma_x \cap \text{int}\boldsymbol{S}$ *and below* (*above*) $\Gamma_y \cap \text{int}\boldsymbol{S}$

The following lemma is a direct consequence of Lemma 1.

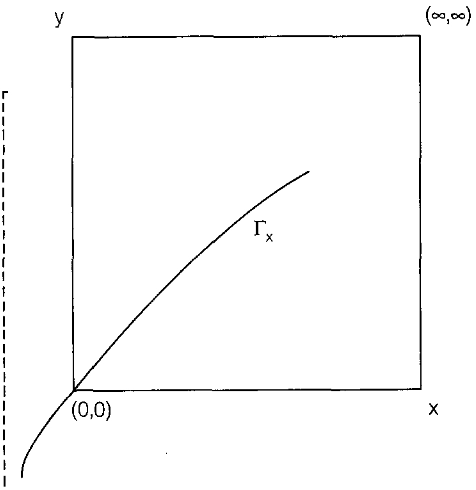

Fig. 8. Γ_x for $a+b<ab$

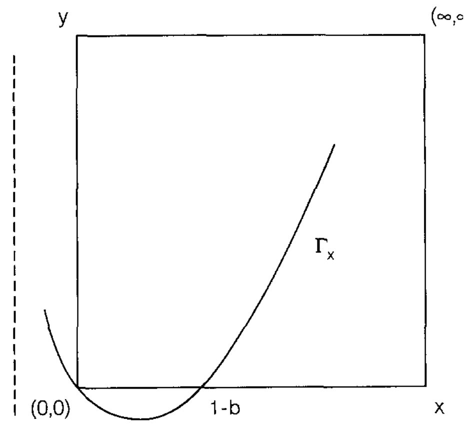

Fig. 9. Γ_x for $a+b>ab$, $a>1$

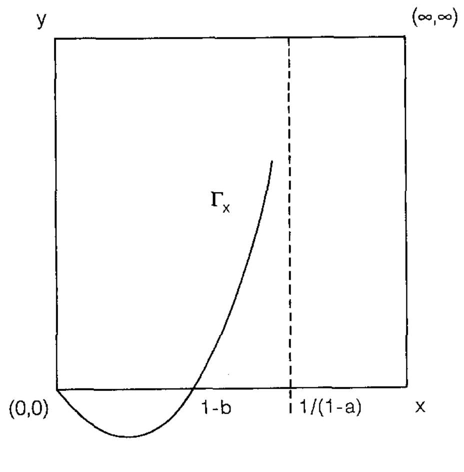

Fig. 10. Γ_x for $a+b>ab$, $a<1$

Lemma 2 *Let $(a+b)(c+d)>0$ and let E (G) be an equilibrium such that $E<G$ coordinatewise or let $E=(0,0)$ $(G=(\infty,\infty))$. If Γ_x passes strictly below (strictly above) Γ_y in the open rectangle $(E,G)=\{z: E<x,y<G$ coordinatewise$\}$, then $T^n z\to G(E)$ as $n\to\infty$ for $z\neq E$ (G) in the closed rectangle $[E,G]=\{(x,y): E\leqslant x,y\leqslant G\}$.*

Proof. We only prove the statement concerning convergence to G. The other statement follows by symmetry. Let $E<z\leqslant G$ coordinatewise. Then the hypotheses of the lemma ensure the existence of $z_0\in(E,G)$ above Γ_x and below Γ_y such that $z_0<z$. By Lemma 1, $Tz_0\geqslant z_0$ and then by the monotonicity of T and the invariance of G under T, $z_0\leqslant T^n z_0\leqslant T^{n+1}z_0\leqslant G$. Therefore, by the continuity of T, $T^n z_0$ converges to an equilibrium between z_0 and G, which must be G by the hypotheses of the lemma (cf. Lemma 1). On the other hand, $T^n z_0\leqslant T^n z\leqslant G$. Then $T^n z\to G$ by the sandwich theorem.

If for $z=(x,y)$ we have $E\leqslant z\leqslant G$, $z\neq E$ and $x=x_E$ or y_E, where $E=(x_E,y_E)$, then the hypothesis $(a+b)(c+d)>0$ ensures that $E<Tz\leqslant G$ and the first part of this proof applies to Tz. This completes the proof of Lemma 2.

If $(a+b)(c+d)>0$, then it follows from (5) that the slope of Γ_x at $(0,0)$ is $b-1$ and it follows from this by symmetry that the slope of Γ_y at $(0,0)$ is $1/(d-1)$ (with the agreement that $1/0=\infty$). Therefore, in int$\boldsymbol{S}$, Γ_x passes below Γ_y near $(0,0)$ if $d\leqslant 1$ or $d>1$, $b-1<1/(d-1)$ and above it if $d>1$, $b-1>1/(d-1)$. These observations lead to the following:

Lemma 3 *Suppose $(a+b)(c+d)>0$. If $1/b+1/d>1$ (<1), then in* int$\boldsymbol{S}$, *Γ_x passes below (above) Γ_y near $(0,0)$. By symmetry, if $1/a+1/c>1$ (<1), then in* int$\boldsymbol{S}$, *Γ_x passes above (below) Γ_y near (∞,∞).*

The curve Γ_x (Γ_y) is strictly concave-up or concave-down or straight according as $a+b$ $(c+d)$ is larger (smaller) than or smaller (larger) than or equal to ab (cd). If we combine this observation with the expression for the slope of Γ_x (Γ_y), we obtain the following:

Lemma 4 *Suppose $(a+b)(c+d)>0$. If $a+b\geqslant ab$, $c+d\geqslant cd$ $(a+b\leqslant ab$, $c+d\leqslant cd)$ and strict inequality holds in at least one of the two inequalities, then $1/b+1/d=1$ implies that in* int$\boldsymbol{S}$, *Γ_x passes above (below) Γ_y near $(0,0)$ and $1/a+1/c=1$ implies that in* int$\boldsymbol{S}$, *Γ_x passes below (above) Γ_y near (∞,∞).*

If $a+b\geqslant ab$, $c+d\geqslant cd$ or $a+b\leqslant ab$, $c+d\leqslant cd$, then Lemmas 1–4 above are sufficient to determine the dynamics of (2) near $(0,0)$ and near (∞,∞). However, if $a+b$ is smaller (larger) than ab and $c+d$ is larger (smaller) than cd, then we need a refinement of Lemmas 3 and 4. For the sake of definiteness, let us assume that $a+b<ab$ and $c+d>cd$. In this case both Γ_x and Γ_y are concave-down in int$\boldsymbol{S}$ and if their slopes at $(0,0)$ are equal, then it is not clear which curve passes below the other near $(0,0)$ in int$\boldsymbol{S}$. Lemma 5 below enables us to determine this and also to determine the relative position of Γ_x and Γ_y near a polymorphic equilibrium.

Lemma 5 *Let $a+b<ab$, $c+d>cd$. If the first nonvanishing coefficient in the sequence a_0,a_1,a_2,a_3 is negative (positive), then in* int$\boldsymbol{S}$, *Γ_x passes below (above) Γ_y near $(0,0)$. If the last nonvanishing coefficient in the sequence a_0,a_1,a_2,a_3 is negative (positive), then in* int$\boldsymbol{S}$, *Γ_x passes below (above) Γ_y near (∞,∞). If a polymorphic equilibrium G corresponds to a simple or triple (double) root of* (4), *then Γ_x and Γ_y are on opposite sides (on the same side) of each other to the left and to the right of G near G.*

Proof. Out of the statements concerning $(0, 0)$ and (∞, ∞) we will only prove the one concerning $(0, 0)$ because the other statement follows from this one by symmetry. The condition $a_0 < 0$ $(a_0 > 0)$ is equivalent to the condition $b + d > bd$ $(b + d < bd)$. If $bd > 0$, then this condition in turn is equivalent to the condition $1/b + 1/d > 1$ (<1) and the statement of Lemma 5 concerning $(0, 0)$ follows from Lemma 3. If $a_0 \neq 0$ and $bd = 0$, then $a_0 < 0$ and $1/b + 1/d > 1$ and the statement concerning $(0, 0)$ follows from Lemma 3 again.

In proving the statement concerning $(0, 0)$ we can therefore assume that $a_0 = 0$, i.e., $(b - 1)(d - 1) = 1$. Then the slopes of Γ_x and Γ_y at $(0, 0)$ are equal to the same positive number (the slope of Γ_x is positive at $(0, 0)$ because $a + b < ab$). The function ϕ defined by the right-hand side of (5) after the interchanges $x \leftrightarrow y$, $a \leftrightarrow c$, $b \leftrightarrow d$ is therefore strictly increasing for $y \geqslant 0$ (and is zero for $y = 0$). The function Φ defined by the right-hand side of (5) is strictly increasing for $x \geqslant 0$ (and is zero for $x = 0$). It follows from the definition of Φ and ϕ that

$$[(1 - a)x - 1]^2[(1 - c)\Phi(x) - 1][x - \phi(\Phi(x))] = a_3x^4 + a_2x^3 + a_1x^2 + a_0x.$$

Let ϕ^{-1} denote the inverse function of ϕ in int$\boldsymbol{S}$. By the definitions of Φ and ϕ, the curve Γ_x passes below (above) Γ_y near $(0, 0)$ if and only if $\Phi(x)$ is smaller (larger) than $\phi^{-1}(x)$ for small positive x or if and only if $x - \phi(\Phi(x))$ is larger (smaller) than 0 for small positive x. Since $\Phi(0) = 0$, the signs of $x - \phi(\Phi(x))$ and of $a_3x^4 + a_2x^3 + a_1x^2 + a_0x$ are opposite for small positive x. The latter is negative (positive) if the first nonvanishing coefficient in the sequence a_0, a_1, a_2, a_3 is negative (positive). This proves the assertion of Lemma 4 concerning $(0, 0)$.

The assertion of Lemma 5 concerning a polymorphic equilibrium G can be proved similarly to the assertion concerning $(0, 0)$. We only have to note that if $c + d > 0$, then Γ_y is the graph of a strictly increasing continuous function in the neighborhood of G and neither $(1 - a)x - 1$ nor $(1 - c)y - 1$ vanishes at an equilibrium (x, y) unless $a = b = 0$ or $c = d = 0$.

5 Convergence of allelic frequencies on decreasing invariant curves containing an equilibrium

As before, let T denote the transformation $T(x, y) = (x^*, y^*)$. The sole purpose of this section is the proof of the lemma below. This lemma will be used in proving convergence of $T^n z$ to unstable equilibria. Before stating and proving Lemma 6 below, we study the inversion of T. Let $(a + b)(c + d) > 0$ and let us introduce the variables $r = xy$, $s = x + y$. Equations (2) are then equivalent to

$$ar + (1 - x^*)s = bx^*, \qquad cr + (1 - y^*)s = dy^*.$$

This is a system of linear equations for r and s. Consequently, the set of solutions (r, s) is either the empty set or a singleton or a line or the entire (r, s)-space. The points (r, s) for which there are $x, y \in$ int$\boldsymbol{S}$ such that $r = xy$, $s = x + y$ are the points (r, s) that are on or above the parabola $s^2 = 4r$ in the strictly positive quadrant of the (r, s)-plane. The points on this parabola correspond to a single point on the diagonal $x = y$ and the points strictly above the parabola correspond to two distinct points (x, y) and (y, x). Therefore, if the image Tz_0 of a point $z_0 \in \mathrm{int}\boldsymbol{S}$ has more than two inverse images under T in $\mathrm{int}\boldsymbol{S}$, then the set of points (r, s) corresponding to these inverse images is the intersection of a line with the domain $D = \{(r, s) : 0 < s,\ 0 < 4r \leqslant s^2\}$ or it is the entire domain D. In either case, there is a line segment l stretching between two boundary points of D and having the property that $Tz = Tz_0$ for points $z \in \mathrm{int}\boldsymbol{S}$ on the curve l' corresponding to l in $\mathrm{int}\boldsymbol{S}$. The curve l' is the nonempty intersection of a hyperbola or a pair of lines or a line with $\mathrm{int}\boldsymbol{S}$. Consequently, we have proved the following.

If the image Tz_0 of a point $z_0 \in \mathrm{int}\boldsymbol{S}$ has more than two inverse images, then there are two distinct points on the boundary of $\boldsymbol{S}$ that are connected with a curve γ in $\mathrm{int}\boldsymbol{S}$ on which T assumes the constant value Tz_0.

For if l has an endpoint W_1 on the parabola $s^2 = 4r$, then the other endpoint W_2 of l is on the boundary of the positive quadrant of the (r, s)-plane. The curve l' will then contain a broken line segment γ or a hyperbola branch γ, either of them symmetric with respect to the line $x = y$ and connecting the boundary points of $\boldsymbol{S}$ corresponding to W_2. If both endpoints of l are on the boundary of the positive quadrant of the (r, s)-plane, then the italicized statement is obvious.

Lemma 6 *Suppose that $(a + b)(c + d) > 0$. Let Γ be a strictly decreasing continuous curve in $\boldsymbol{S}$ that* (1) *separates $\boldsymbol{S}$,* (2) *is symmetric with respect to the diagonal $x = y$,* (3) *is invariant under T,* (4) *contains a polymorphic equilibrium G. Then $T^n z \to G$ as $n \to \infty$ provided that $z \in \Gamma$.*

Proof. First let us assume that Γ contains an interior point z_0 whose image Tz_0 has more than two inverse images under T. If γ in the italicized statement at the beginning of this section has a point Z not on Γ, then, since Γ separates $\boldsymbol{S}$, there is a point z_1 on Γ which is comparable to Z. Then due to the strong monotonicity of T, the points $TZ = Tz_0$ and Tz_1 are comparable distinct points on Γ, which contradicts the strictly decreasing property of Γ. Therefore, $\gamma \subseteq \Gamma$. Since γ connects two distinct boundary points of $\boldsymbol{S}$, we must have $\gamma = \Gamma$. Then $G = Tz_0$ and $Tz = G$ for $z \in G$ and the lemma is proved. (This case actually occurs when, for example, $a = b > 0$, $c = d > 0$, $G = (1, 1)$ and $\Gamma = \gamma$ is the concave-up branch of the hyperbola $xy = 1$.)

In the rest of this proof we can therefore assume that (5) *no point Tz_0 of Γ, where z_0 is an interior point of Γ, has more than two inverse image under T.*

Since it follows from the increasing property of Γ_x in $\mathrm{int}\boldsymbol{S}$, proved in the preceding section, that the fixed points of T are linearly ordered, Γ does not contain any fixed points of T other than G due to the decreasing property of Γ. Let H, $H \neq G$, be a fixed point of T^2 on Γ. Then TH is also a fixed point of T^2 on Γ. We have $TH \neq H$ because Γ does not contain fixed points of T other than G. Therefore, we have two distinct points, H and TH, on Γ, transformed into each other by T. The intermediate value theorem implies that there is a fixed point of T on Γ strictly between H and TH, which then has to be G. This contradicts Appendix A. Summing up, G is the only fixed point of T^2 (and then of T) on Γ.

Without loss of generality we may assume that G is not below the diagonal $x = y$ (otherwise we interchange the roles of x and y). Due to (1) and (2) in Lemma 6, Γ intersects the diagonal $x = y$. This point of intersection M is unique because Γ is decreasing. Let us parametrize Γ continuously by a parameter t that increases from 0 corresponding to the leftmost point of Γ to 1 corresponding to the rightmost point of Γ. Suppose $t = \frac{1}{2}$ corresponds to M and the parametrization follows the symmetry of Γ, i.e., if a certain value $t^\wedge$ of t corresponds to (x, y) on Γ, then $1 - t^\wedge$ corresponds to (y, x). Denote the value of t corresponding to G by t_0. The transformation T carries over to a continuous function $\tau : [0, 1] \to [0, 1]$. We note that $T(x, y) = T(y, x)$. Therefore, the graph of τ is symmetric with respect to the line $t = \frac{1}{2}$ and (5) implies that τ is one-to-one on $(0, \frac{1}{2})$ and on $(\frac{1}{2}, 1)$. Consequently, τ is strictly increasing or strictly decreasing on $(0, \frac{1}{2})$ and then on $[0, \frac{1}{2}]$ by continuity.

First suppose that τ is strictly increasing on $[0, \frac{1}{2}]$. Then, since $t_0 \leqslant \frac{1}{2}$ and t_0 is the only fixed point of τ according to the third paragraph of this proof, we have $t < \tau(t) < t_0$ if $0 \leqslant t < t_0$ and $\tau(t) < t$ if $t_0 < t \leqslant 1$ (because $\tau(\frac{1}{2}) < \frac{1}{2}$ if $t_0 \neq \frac{1}{2}$). It is a simple excercise to prove that the sequence obtained by applying the iterates of

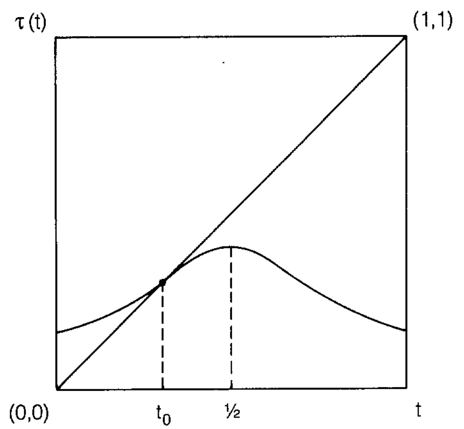

Fig. 11. τ is increasing

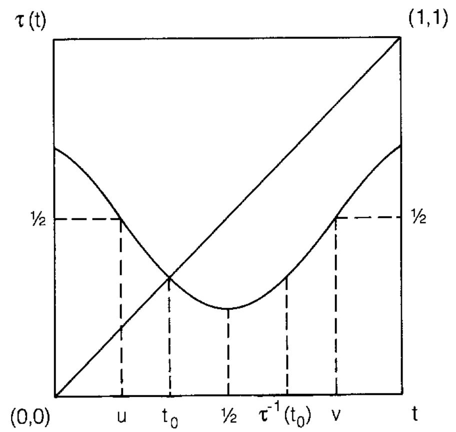

Fig. 12. τ is decreasing

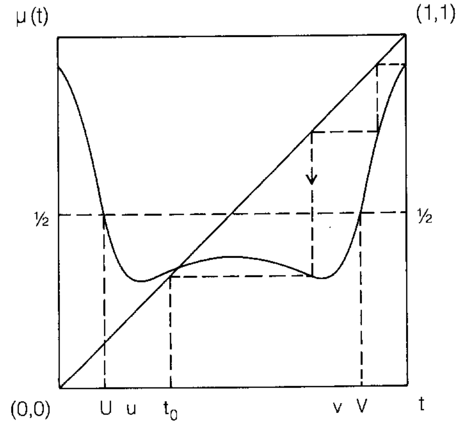

Fig. 13. Graph of μ when τ is decreasing

τ to a point in $[0, 1]$ converges to t_0, the convergence being monotone from a certain iterate (Fig. 11).

Let us now assume that τ is strictly decreasing on $[0, \frac{1}{2}]$. If G is on the diagonal, i.e., $G = M$, then $t_0 = \frac{1}{2}$ and introduction of the new parameter $t' = 1 - t$ reduces this case to the case discussed above. We can therefore assume that $t_0 < \frac{1}{2}$. Then $\tau(\frac{1}{2}) < \frac{1}{2}$, and since t_0 is the only fixed point of τ according to the third paragraph of this proof, $\tau(t) < t$ for $t > t_0$ (Fig. 12). If there is a point t for which $\tau(t) = \frac{1}{2}$, then $t \neq \frac{1}{2}$ because $\frac{1}{2}$ is not a fixed point of τ. Therefore, there are exactly two such points, t and $1 - t$, if there are any at all. Let u (v) be the smaller (larger) of these points if the points exist and let $u = 0$ $(v = 1)$ if they do not exist (Fig. 12). Then $\frac{1}{2} < \tau(t)$ if $t < u$ and $\tau(t) < \frac{1}{2}$ if $u < t < \frac{1}{2}$. Let us consider $\mu = \tau \circ \tau$, the composition of τ with itself. Then μ decreases for $t < u$, increases for $u < t < \frac{1}{2}$ and is symmetric with respect to the line $t = \frac{1}{2}$. Since according to the third paragraph of this proof, μ cannot have a fixed point other than t_0, $\mu(u) > u$ and $\mu(\frac{1}{2}) < \frac{1}{2}$ (Fig. 13).

If $m \leqslant t \leqslant \frac{1}{2}$, then $\mu^{(n)}(t) \to t_0$ as $n \to \infty$ for reasons similar to those in paragraph 5 of this proof. Let $\mu(U) = \mu(V) = \frac{1}{2}$, $U < V$ if such U, V exist. Otherwise let $U = 0$, $V = 1$. Then $U \leqslant u < \frac{1}{2} < v \leqslant V$. If $U \leqslant t \leqslant V$, then $u \leqslant \mu(t) \leqslant \frac{1}{2}$ and $\mu^{(n)}(t) \to t_0$. If $t < U$, then $\mu(t) > \frac{1}{2}$. If we also have $\mu(t) \leqslant V$, then $\mu^{(n)}(t) \to t_0$. We can therefore assume that $t > V$. Then $\frac{1}{2} < \mu(t) < t$. If still $\mu(t) > V$, then we apply μ to $\mu(t)$ again. Ultimately we find an N such that $\mu(t) > \cdots > \mu^{(N-1)}(t) > \mu^{(N)}(t) > V$ but $\mu^{(N+1)}(t) \leqslant V$ (because μ does not have fixed points to the right of V). Then $U < \frac{1}{2} < \mu^{(N+1)}(t) \leqslant V$, and therefore, $\mu^{(n)}(t) \to t_0$. This completes the proof of the lemma.

6 Proof of Theorem 1

First we note that the apparently missing case where $a + b \geqslant ab$, $c + d \geqslant cd$, $1/b + 1/d \leqslant 1$, $1/a + 1/c \leqslant 1$ or $a + b \leqslant ab$, $c + d \leqslant cd$, $1/b + 1/d \geqslant 1$, $1/a + 1/c \geqslant 1$ is also included in Theorem 1 since if the first set of inequalities holds, then $abcd > 0$ and the condition in case $(-)$ is satisfied and if the second set of inequalities holds, then $abcd = 0$ implies $a = b = 0$ or $c = d = 0$, which is included in case $(\to\leftarrow)$ and $abcd > 0$ implies the condition for case $(-)$.

In Theorem 1 it is assumed that $a + b \leqslant ab$, $c + d \leqslant cd$ or $a + b \geqslant ab$, $c + d \geqslant cd$. Section 4 informs us that barring the anomalies occurring when $a = b = 0$ or $c = d = 0$, in this case Γ_x and Γ_y are lines passing through $(0, 0)$ or hyperbola branches passing through $(0, 0)$ that are concave in opposite senses in int$\boldsymbol{S}$ (Γ_x is concave-up and Γ_y is concave-down or vice versa). Consequently, Γ_x and Γ_y can have at most one point of intersection (counting multiplicity) in int$\boldsymbol{S}$ unless they coincide. Coincidence can only happen when Γ_x and Γ_y are lines with the same finite positive slope, i.e., $a, b, c, d > 1$, $a + b = ab$, $c + d = cd$, $1/(a - 1) = c - 1$; these conditions are equivalent to the conditions listed in case $(-)$ of Theorem 1. If $\Gamma_x \neq \Gamma_y$, then in int$\boldsymbol{S}$ the curves Γ_x and Γ_y have 0 (1) point of intersection if Γ_x and Γ_y are on the same side (on opposite sides) of each other near $(0, 0)$ and near (∞, ∞).

To prove the statement of Theorem 1 concerning case $(\to)$ we first note that if $1/b + 1/d > 1$, $1/a + 1/c < 1$, then combination of Lemmas 2 and 3 and the second paragraph of this section shows that case $(\to)$ of Introduction occurs. (We do not even have to use the other inequality $1/a + 1/c < 1$.) To complete the proof of case $(\to)$ we only have to show that $a + b \geqslant ab$, $c + d \geqslant cd$, $1/b + 1/d > 1$, $1/a + 1/c = 1$ or $a + b \leqslant ab$, $c + d \leqslant cd$, $1/b + 1/d = 1$, $1/a + 1/c \leqslant 1$

implies that case $(\rightarrow)$ of Introduction occurs. To prove this we can apply Lemmas 2 and 4 combined with the second paragraph of this section.

Case $(\leftarrow)$ of Theorem 1 can be proved similarly to case $(\rightarrow)$.

Let us now assume that the conditions in case $(\rightarrow\leftarrow)$ of Theorem 1 are satisfied and $(a+b)(c+d)>0$. Then Lemma 3 combined with the second paragraph of this section shows that there is a unique polymorphic equilibrium G and Lemma 2 implies that $T^n z \rightarrow G$ for $z \leqslant G$, $z \neq (0,0)$ and for $z \geqslant G$, $z \neq (\infty, \infty)$. Then we can complete the proof by appealing to the sandwich theorem.

Let us now assume that $(a+b)(c+d)=0$, for example, $a+b=0$. Then Γ_x is the line $x=1$. Since $\Gamma_x \cap \operatorname{int} S$ is always the graph of a function of x for $x>0$, there is a unique polymorphic equilibrium G. The rest of the proof is essentially the same as in the case $(a+b)(c+d)>0$ through Lemmas 2 and 3 and the sandwich theorem. In fact, Lemmas 2 and 3 (and the other lemmas in Sect. 4) could be extended to cover the case $(a+b)(c+d)=0$, as well. We have chosen not to do this because then the statements of our lemmas would become awkward.

Case $(\leftarrow\rightarrow)$ is the most complicated. Lemma 3 combined with the second paragraph of this section again implies the existence of a unique polymorphic equilibrium G and Lemma 2 shows that if $z \neq G$, then $T^n z \rightarrow (0,0)$ or (∞, ∞) according as $z \leqslant G$ or $z \geqslant G$. The proof of convergence of $T^n z$ for the remaining z's follows a line of reasoning used in [7]. Let D_1 be the set of points $z \in S$ such that $T^n z \rightarrow (0,0)$ as $n \rightarrow \infty$. Clearly, the rectangle $[(0,0), G]$ minus $\{G\}$ is contained in D_1, and $TD_1 \subseteq D_1$. Let Γ_1 be the boundary of D_1 in S. Then $G \in \Gamma_1$ and $T\Gamma_1 \subseteq \Gamma_1$. On the other hand, $z \leqslant z_0$, $z_0 \in D_1$ implies $z \in D_1$ by the monotonicity of T. Consequently, Γ_1 is the graph of a nonincreasing function where the gaps at the points of discontinuity are filled with line segments. Let l be a horizontal or vertical line segment of positive length that is part of Γ_1. Since $a, b, c, d > 0$ in case $(\leftarrow\rightarrow)$, the strong monotonicity of T implies that Tl is the graph of a strictly increasing function. Since $Tl \subseteq \Gamma_1$, this contradicts the nonincreasing property of Γ_1. Therefore, Γ_1 is the graph of a strictly decreasing continuous function. On the other hand, it is clear that the endpoints of Γ_1 belong to the boundary of S and also to Γ_1. Since $T(x,y)=T(y,x)$, D_1 and then Γ_1 are symmetric with respect to the diagonal $x=y$, Γ_1 satisfies the hypotheses of Lemma 6 and $T^n z \rightarrow G$ for $z \in \Gamma_1$. On the other hand, since every point of S strictly below Γ_1 belongs to D_1, we have $T^n z \rightarrow (0,0)$ for z below Γ_1.

Similarly to Γ_1, we can construct another strictly decreasing continuous curve Γ_2 separating S, passing through G and possessing the property that $T^n z$ converges to G or (∞, ∞) according as z is on or strictly above Γ_2. It is then clear that Γ_2 is above Γ_1 in the wider sense. If z is between the two curves, then the sandwich theorem ensures that $T^n z \rightarrow G$ due to the monotonicity of T.

In case $(\leftarrow\rightarrow)$, Γ_x (Γ_y) is strictly concave-down (concave-up) or a straight line and Γ_x and Γ_y both go through $(0,0)$. Therefore, Γ_x and Γ_y are not tangent to each other at G (since they do not coincide). Consequently, 1 cannot be an eigenvalue of the derivative matrix of T at G due to the following observation in [7]: The number 1 is an eigenvalue of the Jacobian of a dynamical system $x' = \xi(x,y)$, $y' = \eta(x,y)$ $(\partial x/\partial y > 0,\ \partial\eta/\partial x > 0)$ at an internal fixed point G if and only if the curves $x' = x$ and $y' = y$ are tangent to each other at G. (This statement can be proved by direct verification.)

On the other hand, the convergence pattern of $T^n z$ for z below (on) Γ_1 shows that the maximal (minimal) eigenvalue λ_1 (λ_2) at G must be $\geqslant 1$ ($\geqslant -1$ and $\leqslant 1$).

Therefore, $\lambda_1 > 1$ and $-1 \leqslant \lambda_2 < 1$. Suppose $\lambda_2 = -1$ and Γ_1 has points arbitrarily close to G that are not on Γ_2. Let ϱ be a nondegenerate rectangle between Γ_1 and Γ_2 and let z_u (z_l) be the upper right (lower left) vertex of ϱ. By what we have proved so far, $T^n z_u \to G$ and $T^n z_l \to G$. Consequently, due to the monotonicity of T, the area of $T^n \varrho$ converges to zero as $n \to \infty$. On the other hand, if ϱ is sufficiently close to G, then $T^n \varrho$ is so close to G that for all n the area of $T^{n+1}\varrho$ is larger than $(\lambda_1 + 1)/2$ times the area of $T^n \varrho$. (This follows from the integral transformation theorem.) Since $(\lambda_1 + 1)/2 > 1$, this implies that the area of $T^n \varrho$ goes to ∞ as $n \to \infty$, a contradiction. Consequently, either $\Gamma_1 = \Gamma_2$ in a neighborhood of G or $-1 < \lambda_2 < 1$.

Suppose $-1 < \lambda_2 < 1$. Then the extension of the stable manifold theorem to maps that are not necessarily one-to-one [4], combined with the monotonicity of T [5, Proof of Proposition 3.5] implies the existence of a strictly decreasing continuous curve near G (the local stable manifold of G) with the property that near G, $T^n z \to G$ exactly when z is on this curve. Consequently, $\Gamma_1 = \Gamma_2$ near G in this case, as well, and then always in case $(\leftarrow\rightarrow)$.

This contradicts the existence of points of int$\boldsymbol{S}$ strictly between Γ_1 and Γ_2 since then there would be distinct points z_1 and z_2 between Γ_1 and Γ_2 such that $z_1 \leqslant z_2$ (coordinatewise); then $T^n z_1$ and $T^n z_2$ would be distinct and we would have $T^n z_1 \leqslant T^n z_2$ (coordinatewise), which provides a contradiction if n is sufficiently large because $T^n z_1 \in \Gamma_1$, $T^n z_1 \to G$ and $T^n z_2 \in \Gamma_2$, $T^n z_2 \to G$ and Γ_1 and Γ_2 are strictly decreasing and coincide near G. Therefore, $\Gamma_1 = \Gamma_2$ can serve as the curve Γ occurring in case $(\leftarrow\rightarrow)$. This completes the proof of case $(\leftarrow\rightarrow)$.

Remark. It is easy to prove that in case $(\leftarrow\rightarrow)$, the derivative matrix of T is degenerate at only those internal unstable fixed points that are either on the diagonal $x = y$ or on a curve where T is constant. In the latter case, the assertion in $(\leftarrow\rightarrow)$ is obvious. In the former case, a little change in an appropriate parameter moves the unstable fixed point to a place where the derivative matrix of T is nondegenerate. Then the standard stable manifold theorem is sufficient to draw the phase portrait of T. The phase portrait of the original system can then be obtained by moving the new unstable internal fixed point back to its original position and exploit the monotonicity of T in the parameters a, b, c and d. This observation holds also for the case $(\leftarrow\rightarrow)$ of Theorem 2. Therefore, we do not really need the stable manifold theorem for maps that are not one-to-one, although the application of this theorem makes our proofs somewhat simpler.

Let us now assume that the condition in case $(-)$ is satisfied. Then all the points in $\boldsymbol{S}$ on the line $y = (b-1)x$ are fixed points of T and there are no more fixed points. Let z_0 be such a fixed point in int$\boldsymbol{S}$ and let D_1 (D_2) be the set of points z in $\boldsymbol{S}$ for which all cluster points of the sequence $\{T^n z\}$ are $\geqslant z_0$ $(\leqslant z_0)$. Arguments similar to those above show that the boundary Γ_1 of D_1 is never above the boundary Γ_2 of D_2 and by Lemma 6, $T^n z \to z_0$ if z is between Γ_1 and Γ_2 in the wider sense. We now study the dependence of Γ_1 and Γ_2 on the fixed point for which they are constructed. It is clear that if z on the line $y = (b-1)x$ converges increasingly to z_0, then $\Gamma_2(z)$ converges increasingly to a curve Γ_- that never passes above $\Gamma_1(z_0)$, contains z_0 and is invariant under T. It is also clear that Γ_- is nonincreasing and then arguments similar to those above show that it is strictly decreasing. Consequently, Lemma 6 implies that $T^n z \to z_0$ for $z \in \Gamma_-$. Then we must have $\Gamma_- = \Gamma_1(z_0)$. We can prove similarly that $\Gamma_1(z)$ converges decreasingly to $\Gamma_2(z_0)$ as the fixed point z converges decreasingly to z_0. Therefore, the domains between the curves Γ_1 and Γ_2 constructed for all equilibria in int$S$ fill int$S$. Since $Tz \in \text{int}S$ for all $z \in S$ other than $(0, 0)$ or (∞, ∞), the proof is complete.

7 Proof of Theorem 2

Since it is assumed that $a + b < ab$, the curve Γ_x contains no points with a positive x-coordinate and a negative y-coordinate. Therefore, all positive solutions of (4) furnish polymorphic equilibria. Consequently, relying on Descartes' rule concerning the number of sign changes in the sequence of the coefficients of an algebraic equation and using the theory of at most cubic equations, we can prove easily most of the statement of Theorem 2 concerning the number and multiplicity of polymorphic equilibria. The only exception is that to exclude the occurrence of two unstable polymorphic equilibria we have to appeal to Appendix B. The behavior of T^n between two consecutive equilibria and before (after) the smallest (largest) polymorphic equilibrium can be determined by using Lemmas 2 and 5. Since it follows easily from Appendix B that in case $(\leftarrow\rightarrow)$ two distinct unstable polymorphic equilibria or a multiple polymorphic equilibrium cannot exist, the behavior of T^n on the rest of S can be determined by using the same technique as in the proof of Theorem 1. (The only possible difference is that we may have to use the sandwich theorem in proving convergence to stable polymorphic equilibria if both initial gene frequency ratios are smaller (larger) than those of the unstable equilibrium. Furthermore, the strict decrease of the Γ curves follows from the fact that in Theorem 2 we have $(a + b)(c + d) > 0$ and $(a + c)(b + d) > 0$, and therefore, if l is a horizontal or vertical line segment of positive length of the Γ curves, then $T^2 l$ is a strictly increasing curve of positive length.)

In cases $(\rightarrow\rightarrow)$ and $(\leftarrow\leftarrow)$ a new phenomenon appears. We only discuss case $(\rightarrow\rightarrow)$. In this case let D be the set of $z \in S$ for which $T^n z \to H$ and let Γ be the boundary of D in S. Lemmas 2 and 5 (combined with the sandwich theorem in case $H \neq (\infty, \infty)$) imply that $z \in D$ for $G \leqslant z \leqslant (\infty, \infty)$, $z \neq G$, $z \neq (\infty, \infty)$. Furthermore, if $z_0 \in D$ and $z \geqslant z_0$, $z \neq (\infty, \infty)$, then $T^n z_0 \leqslant T^n z$, $T^n z_0 \to H$ implies that $G < T^n z < (\infty, \infty)$ for some n and then $T^n z \to H$ by the above. Consequently, Γ is a continuous nonincreasing curve stretching between two boundary points of S. Then Γ is strictly decreasing according to the last sentence of the preceding paragraph because Γ inherits invariance under T from D. It is also clear from the definitions of D and Γ that $T^n z \to H$ if z is strictly above Γ.

Now let $z \neq (0, 0)$ be on or below Γ. Lemmas 2 and 5 imply that $T^n z \to G$ if $(0, 0) \neq z \leqslant G$ and we can always choose z_0 such that $(0, 0) \neq z_0 \leqslant G$ and $z_0 \leqslant z$. Then $T^n z_0 \leqslant T^n z$, $T^n z_0 \to G$, combined with the fact that $T^n z$ is on or below Γ (due to the T-invariance of Γ), implies that $T^n z \to G$. Note that in the cases $(\rightarrow\rightarrow)$ and $(\leftarrow\leftarrow)$ we have not used Lemma 6. This completes the proof of Theorem 2.

Appendix A: The absence of points of minimal period two (after Selgrade and Ziehe [5, Theorem 4.2])

Suppose $(a + b)(c + d) > 0$. *Let* $Z = (X, Y)$ *be a fixed point of* T *in* int$S$.

(i) *If* $c(X - 1) \leqslant a(Y - 1)$, *then* $0 \leqslant x < X$, $\infty \geqslant y > Y$ *implies* $T^2(x, y) \neq (x, y)$.

(ii) *If* $c(X-1) \geqslant a(Y-1)$, *then* $\infty \geqslant x > X$, $0 \leqslant y < Y$ *implies* $T^2(x, y) \neq (x, y)$.

Proof. By interchanging the roles of $\boldsymbol{P}$ and $\boldsymbol{Q}$ if necessary, we can assume that $c > 0$. If, in addition, $a > 0$, then it is sufficient to prove, say (i), because then (ii) will follow by symmetry. If $c > 0$ and $a = 0$, then $b > 0$ (because the interchange mentioned above preserves the inequality $(a + b)(c + d) > 0$) and $x^* < 1$ for $x + y > 0$. In this case the left-hand side of the inequality that is the condition in (i) is negative and the right-hand side is zero. Therefore, the condition in (ii) is not satisfied. Summing up, to prove the statement in Appendix A it is sufficient to prove (i) under the hypothesis that $c > 0$ (and, of course, $(a + b) > 0$). Furthermore, since $0 < x^*$ for $(x, y) \neq (0, 0)$ and $y^* < \infty$ for $(x, y) \neq (\infty, \infty)$, it is sufficient to prove (i) for $0 < x < X$, $\infty > y > Y$. Consequently, in the rest of this proof we will assume this.

Consider the function

$$H(z) = H(x, y) = [a(y-1) + c(1-x)](x^* + y^*) + ady - bcx.$$

We will show that $H(x, y) > 0$. It follows from $c(X-1) \leqslant a(Y-1)$ that $a(y-1) + c(1-x) > 0$. This observation and the inequalities $\partial x^*/\partial y > 0$, $\partial y^*/\partial y > 0$ imply that $\partial H/\partial y > 0$. Therefore, H is strictly decreasing in y and thus $H(x, Y) < H(x, y)$. Let

$$h(x) = [a(Y-1) + c(1-x)]\left[\frac{axY + x + Y}{b + X + Y} + \frac{cxY + x + Y}{d + X + Y}\right] + adY - bcx.$$

Then $h(x) < H(x, Y)$. Since (X, Y) is a fixed point of T,

$$aXY + (1-X)(X+Y) - bX = 0 = cXY + (1-Y)(X+Y) - dY.$$

Hence $h(X) = 0$. Since $h(0) > 0$ and $y = h(x)$ represents a concave-down parabola, $h(x) > 0$. From this and previous inequalities,

$$H(x, y) > H(x, Y) > h(x) > 0.$$

Rearrangement of the inequality $H(x, y) > 0$ yields

$$acx^*y^* - a(y-1)(x^* + y^*) - ady < acx^*y^* + c(1-x)(x^* + y^*) - bcx.$$

If $a = 0$, then the left-hand side of this inequality is zero. Therefore, the right-hand side is positive, which is equivalent to $x^{**} > x$ since $c > 0$. This proves (i) if $a = 0$. On the other hand, $x^{**} = x$ implies the vanishing of the right-hand side, which in turn implies that the left-hand side is negative, which is equivalent to $y^{**} < y$ if $a > 0$. This completes the proof of the statement of Appendix A.

Appendix B: There is only one polymorphic equilibrium, counting multiplicity, if the fixation equilibria are stable (after Karlin and Lessard [1, pp. 214, 215 and 242])

The sole purpose of this appendix is the proof of the statement in the title. The multiplicity in the title refers to the multiplicity of the x-coordinate of the polymorphic equilibrium as a solution of (4). Let us introduce the variable $t = y/x$. For $0 < x, y < \infty$, (4) is equivalent to the equations

$$x = \frac{t-(b-1)}{1-(a-1)t}, \qquad y = \frac{(d-1)t-1}{(c-1)-t},$$
$$f(t) = t^3 - \alpha t^2 + \beta t - 1 = 0$$

with

$$\alpha = (a-1)(d-1) + (b-1) + (c-1),$$
$$\beta = (b-1)(c-1) + (a-1) + (d-1).$$

We have $f(0) = -1 < 0$. We have to prove the assertion in the title only if $a + b < ab$ and $c + d > cd$ or vice verse, since Theorem 1 covers the remaining cases and its proof is independent of this appendix. We may therefore assume that $a + b < ab$ and $c + d > cd$. Since (∞, ∞) is a stable equilibrium, $a + c < ac$ according to Lemmas 2 and 5. The inequality $a + b < ab$ implies that $a > 1$. We have

$$f((a-1)^{-1}) = (a+c-ac)(a+b-ab)(a-1)^{-3} \geqslant 0.$$

Therefore, $f(t_0) = 0$ for some t_0, $0 < t_0 \leqslant (a-1)^{-1}$. If $t_0 = (a-1)^{-1}$, then x is not defined. If $t_0 < (a-1)^{-1}$, then $t_0 - (b-1) < (a-1)^{-1} - (b-1) = (a+b-ab)(a-1)^{-1} < 0$ and $1-(a-1)t_0 > 1-(a-1)(a-1)^{-1} = 0$. Therefore, $x > 0$. Consequently, no equilibrium corresponds to the root t_0 of f and there are at most two distinct polymorphic equilibria.

If there are two distinct polymorphic equilibria, then the stability of the fixation equilibria implies through Lemmas 2 and 5 that the x-coordinate of one of the polymorphic equilibria is a double root of (4) and that of the other is a simple root. Since both Γ_x and Γ_y are strictly concave-down in int$\boldsymbol{S}$ and the derivative of the right-hand side of (5) with respect to b is positive for $x > 0$ (because $a > 1$), a slight change of b in the appropriate direction produces three different polymorphic equilibria (and stable fixation equilibria), a contradiction.

Therefore, there is exactly one polymorphic equilibrium. Because this equilibrium is unstable for neighboring initial allelic frequencies that are either both smaller or both larger than those of the equilibrium, the x-coordinate of the equilibrium has to be a simple or triple solution of (4) according to Lemma 5. Let us assume that it is a triple solution. According to the proof of Lemma 5, Γ_x and Γ_y are tangent to each other of order 2 at the polymorphic equilibrium in question. Let (X, Y) denote this polymorphic equilibrium. Then $Y((1-a)X-1) = -X^2 + (1-b)X$. If we change a and b in such a way that this equation is satisfied but we do not change c and d, then the new system will have an equilibrium at (X, Y). Let us express b from the above equation: $b = 1 + (1/X)[-X^2 - Y((1-a)X-1)]$. Substituting this expression of b and $x = X$, $y = Y$ into the formula for the derivative dy/dx along the curve Γ_x (the derivative with respect to x of $\Phi(x)$), we obtain that the slope of Γ_x at (X, Y) is $(X^3 + Y)/\{X[1+(a-1)X]\}$. This is increasing as a, $a > 1$, decreases. Consequently, if a is decreased by a sufficiently small amount (and b is calculated from the equation given for it), then Γ_x and Γ_y will have three distinct points of intersection in int$\boldsymbol{S}$ (one of which is (X, Y)) because Γ_x was passing below Γ_y for $y > Y$, near (X, Y) and Γ_x and Γ_y were tangent to each other at (X, Y) originally. Since the fixation equilibria have remained stable, we have arrived at a contradiction. Consequently, under the hypotheses of Appendix B, there is only one polymorphic equilibrium and if $a + b < ab$, $c + d > cd$, then the x-coordinate of the polymorphic equilibrium is a simple solution of (4).

References

1. Karlin S., Lessard, S.: Theoretical studies on sex ratio evolution. Princeton, N.J.: Princteon University Press 1986
2. Lessard, S.: Évolution du rapport numérique des sexes et modèles dynamiques connexes. In: Lessard, S. (ed.) Mathematical and Statistical Developments of Evolutionary Theory, pp. 269–325, Norwell, MA: Kluwer Academic Publishers 1990
3. Nagylaki, T.: Selection in one- and two-locus systems. Berlin Heidelberg New York: Springer 1977
4. Ruelle, D., Shub, M.: Stable manifolds for maps. In: Nitecki, Z., Robinson, C. (eds.) Global theory of dynamical systems (Lect. Notes Math., 819, pp. 389–392), Berlin Heidelberg New York: Springer 1980
5. Selgrade, J. F., Ziehe, M.: Convergence to equilibrium in a genetic model with differential viability between the sexes. J. Math. Biol. **25**, 477–490 (1987)
6. Smale, S.: Differentiable dynamical systems. Bull. Am. Math. Soc. **73**, 747–817 (1967)
7. Szucs, J. M.: Selection and mutation at a diallelic X-linked locus. J. Math. Biol. **29**, 587–627