Heredity **74** (1995) 607–615 *Received 29 July 1994*

Genetic identity combining mutation and drift

JÜRGEN TOMIUK* & VOLKER LOESCHCKE†

Department of Population Genetics, University of Tübingen, Auf der Morgenstelle 28, D-72076 Tübingen, Germany and †Department of Ecology and Genetics, University of Aarhus, Ny Munkegade, DK-8000 Aarhus C, Denmark

A model for the genetic identity between diploid sexual populations is presented that considers simultaneously mutation and genetic drift as affecting gene frequencies. In contrast to other measures of genetic identity the proposed model allows the genetic identity to be estimated directly from a data set. The new model is integrated into the existing body of population genetics theory. For an infinite population size the model becomes identical to the pure mutation model and if mutation is neglected, it becomes equal to the well-known drift model. The proposed measure of genetic identity between a population and its ancestral population is independent of the population size and equal to the number of ancestral alleles found in the present population. Using data on protein variability from ten primate species, it is shown that the estimate of genetic identity proposed here correlates closely with other identity measures that do not consider genetic drift. The conclusions from hitherto existing studies on the genetic similarity of species, therefore, seem to be reliable. Finally, implications for estimates of the ancestral degree of homozygosity are discussed.

Keywords: ancestral gene, genetic drift, genetic identity, mutation model.

Introduction

In addition to the application of genetic identity or distance measures to phylogenetic studies, such measures are commonly used for comparisons between populations. Genetic identity measures are also incorporated into models in conservation biology for making phylogenetically based decisions on reserve choice and the characterization of evolutionarily unique species or species complexes (Vane-Wright *et al.*, 1991; Crozier & Kusmierski, 1994; Witting *et al.*, 1994). Most measures of genetic identity use the allelic similarity between populations which is averaged over a number of loci (e.g. Nei, 1972). Tomiuk & Loeschcke (1992) using results of protein electrophoresis of polyploid species demonstrated that such an approach is often biased and insufficient. They analysed an identity model that was based on genotypic distributions and considered only the similarity of electromorphs for defining an estimator of genetic identity.

Many different measures of genetic identity have been suggested in the past, but all of them have weaknesses. Some are theoretically unsound or have no evolutionary interpretation. Among those measures of genetic identity that are commonly used, the main limitation is that they either do not incorporate genetic drift, and thus assume population sizes to be infinite (e.g. Nei, 1972), or that they neglect mutation and are more suited for characterizing short-term population differentiation (Reynolds *et al.*, 1983). Thus they all confound effects of drift and mutation.

Here genetic drift is integrated into a population genetic model with random mating that otherwise only assumes mutation to change genotype frequencies. This is done by building upon a pure mutation model suggested previously for estimating genetic identity (Tomiuk & Loeschcke, 1991). It defines four discrete genotype classes according to homozygosity and heterozygosity and whether alleles are assumed to be ancestral or whether they arose by mutation. A simple direct procedure is given for estimating genetic identities and it is shown that the genetic identity can be estimated independently of population size, i.e. that varying population sizes in the evolutionary process do not affect the proposed estimator of genetic identity. The procedure allows us to analyse the effect of population size and the effect of the ancestral degree of heterozygosity on the relative frequencies of the genotype classes present.

The model

The model is based on Fisher–Wright's genetic drift model (Fisher, 1930; Wright, 1931) and the measure of genetic identity proposed by Tomiuk & Loeschcke (1991). Let us first exclude genetic drift and back mutations. Assume in an ancestral diploid population the proportion of homozygotes to be F_0 and of hetero-

*Correspondence.

zygotes to be H_0 with $F_0 + H_0 = 1$. If the mutation rate, α, is constant we expect the frequency of homozygous genotypes in generation t to be

$$F(t) = F_0 \cdot (1-\alpha)^{2t} \tag{1}$$

and the frequency of heterozygous genotypes that contain only alleles that are identical with those in the ancestral population to be

$$H(t) = H_0 \cdot (1-\alpha)^{2t}. \tag{2}$$

The heterozygous genotypes with combinations of ancestral and mutated alleles in generation t have a frequency of

$$NH(t) = 2 \cdot [(1-\alpha)^t - (1-\alpha)^{2t}], \tag{3}$$

and the class of genotypes with two mutated alleles has a frequency of

$$NG(t) = [1-(1-\alpha)^t]^2. \tag{4}$$

The grouping of genotypes into four discrete classes allows then the estimation of the genetic identity between two populations with random mating (Tomiuk & Loeschcke, 1991). The genetic identity I between two populations, 1 and 2, at time t is estimated by the product of the genetic identity I_1 between the common ancestral population ($t=0$) and population 1 and the genetic identity I_2 between the common ancestral population and population 2 with

$$I = I_1 \cdot I_2 = (1-\alpha_1)^t \cdot (1-\alpha_2)^t \tag{5}$$

which can be approximated by

$$I \approx e^{-(\alpha_1+\alpha_2)t}. \tag{6}$$

For $\alpha_1 = \alpha_2$, we have the well-known genetic identity function $I = e^{-2\alpha t}$ (Nei, 1972).

In the following analysis genetic drift, i.e. the finite population size N, is explicitly taken into consideration. The approach used to calculate the frequencies of the genotype classes is based on a model that assumes discrete generations. The basic procedure is given in population genetics text books (e.g. Hartl & Clark, 1989) and allows one to calculate the degree of genetic heterozygosity at an equilibrium between mutation and random drift. A diploid population with N reproductive individuals consists of $2N$ hypothetical gamete pools. The probability of twice drawing a gamete from the same gamete pool is $1/2N$, and the probability of drawing from different gamete pools is $(1-1/2N)$. These probabilities can be combined easily with the probability that a genotype falls into one of the four genotype classes. Considering additionally the influence of mutation, the frequencies of the genotype classes in generation $t+1$ are

$$F(t+1) = [(F(t) + H(t) + NH(t)/2)/2N + (1-1/2N)\cdot F(t)]\cdot(1-\alpha)^2, \tag{7}$$

$$H(t+1) = H(t)\cdot(1-1/2N)\cdot(1-\alpha)^2, \tag{8}$$

$$NH(t+1) = NH(t)\cdot(1-1/2N)\cdot(1-\alpha) + 2\cdot[F(t)+H(t)]\cdot\alpha\cdot(1-\alpha) \text{ and} \tag{9}$$

$$NG(t+1) = 1 - F(t+1) - H(t+1) - NH(t+1). \tag{10}$$

The difference equations, (7)–(9), are approximated by a system of linear differential equations. Neglecting terms of the order of α^2 and α/N, it follows that

$$\frac{dF}{dt} = -2\cdot\alpha\cdot F(t) + [NH(t) + H(t)/2]/2N, \tag{11}$$

$$\frac{dH}{dt} = -[1/2N + 2\alpha]\cdot H(t) \qquad \text{and} \tag{12}$$

$$\frac{dNH}{dt} = 2\cdot\alpha\cdot F(t) - [1/2N+\alpha]\cdot NH(t) + 2\cdot\alpha\cdot H(t), \tag{13}$$

where $F(0) = F_0$, $H(0) = H_0$, $NH(0) = NH_0$ and $NG(0) = NG_0$ are the initial frequencies of the genotype classes at $t=0$ with F_0, H_0, NH_0, $NG_0 \geq 0$ and $F_0 + H_0 + NH_0 + NG_0 = 1$. The solution of equation (12) is given by

$$H(t) = H_0 \cdot e^{-(1/2N+2\alpha)t}. \tag{14}$$

Define now $y(t) := [F(t), NH(t)]^T$, then the linear differential equation is

$$\frac{dy}{dt} = \begin{bmatrix} -2\alpha & 1/4N \\ 2\alpha & -[1/2N+\alpha] \end{bmatrix} \cdot y + \begin{bmatrix} [H_0/2N]\cdot e^{-(1/2N+2\alpha)t} \\ 2\cdot\alpha\cdot H_0\cdot e^{-(1/2N+2\alpha)t} \end{bmatrix}. \tag{15}$$

The eigenvalues of the homogeneous system are $\tau_1 = -\alpha$ and $\tau_2 = -(1/2N+2\alpha)$, and the solution of the differential equation (15) is given by

$$F(t) = c_1 e^{-\alpha t}/4N\alpha - [c_2/2 + H_0/(1+2N\alpha)]\cdot e^{-(1/2N+2\alpha)t} \qquad \text{and} \tag{16}$$

$$NH(t) = c_1 e^{-\alpha t} + [c_2 - 4N\alpha\cdot H_0/(1+2N\alpha)]\cdot e^{-(1/2N+2\alpha)t}, \tag{17}$$

where $c_1 = 4N\alpha\cdot[F_0 + H_0 + NH_0/2]/(1+2N\alpha)$ and $c_2 = [NH_0 - 4N\alpha\cdot F_0]/(1+2N\alpha)$.

Characteristics of the model

(1) When the population size approaches infinity ($N \to \infty$), the mutation-drift model represented by equations (14), (16) and (17) converges to the mutation model proposed by Tomiuk & Loeschcke (1991). When $NH_0 = NG_0 = 0$ then

$$\lim_{N\to\infty} F(t) = F_0 \cdot e^{-2\alpha t},$$

$$\lim_{N\to\infty} H(t) = H_0 \cdot e^{-2\alpha t} \quad \text{and}$$

$$\lim_{N\to\infty} NH(t) = 2 \cdot e^{-\alpha t} \cdot (1 - e^{-\alpha t}).$$

(2) If mutation is neglected ($\alpha \to \infty$), equations (14), (16) and (17) converge to those of the random drift model. These equations are well-known from population genetics text books (e.g. Hartl & Clark, 1989, p. 79). When $NH_0 = NG_0 = 0$, then

$$\lim_{\alpha\to\infty} F(t) = 1 - (1 - F_0) \cdot e^{-t/2N} \text{ and } \lim_{\alpha\to\infty} H(t) = H_0 \cdot e^{-t/2N}.$$

The function $F(t)$ also corresponds to Reynolds's *et al.* (1983) pure drift model with $F_0 = 0$.

(3) An estimate for the genetic identity exists which is independent of the population size. From equations (14), (16) and (17), the frequency of ancestral alleles at time t is

$$F(t) + H(t) + NH(t)/2 = e^{-\alpha t}. \tag{18}$$

Therefore, the observed frequency of ancestral alleles in a diploid population provides an estimate for the genetic identity, I_{new}, between a population at time t and its ancestral population at time zero when the population size N and the mutation rate α are constant during the time period t.

Assume now that the population size is N_1 for $0 \le t \le t_1$ and that $F_0 + H_0 = 1$. Then we have

$$F(t_1) = F_0 \cdot e^{-(1/2N_1 + 2\alpha)t_1} + [e^{-\alpha t_1} - e^{-(1/2N_1 + 2\alpha)t_1}]/[1 + 2N_1\alpha], \tag{19}$$

$$H(t_1) = [1 - F_0] \cdot e^{-(1/2N_1 + 2\alpha)t_1} \quad \text{and} \tag{20}$$

$$NH(t_1) = 4N_1\alpha \cdot [e^{-\alpha t_1} - e^{-(1/2N_1 + 2\alpha)t_1}]/[1 + 2N_1\alpha]. \tag{21}$$

If we further assume that the population size is N_2 for $t_1 < t \le t_2$, then it follows from equations (14), (16) and (17) that

$$F(t_2) = c_1 e^{-\alpha(t_2 - t_1)}/4N_2\alpha - [c_2/2 + H(t_1)/(1 + 2N_2\alpha)] \cdot e^{-(1/2N_2 + 2\alpha)(t_2 - t_1)}, \tag{22}$$

$$H(t_2) = H(t_1) \cdot e^{-(1/2N_2 + 2\alpha)(t_2 - t_1)} \quad \text{and} \tag{23}$$

$$NH(t_2) = c_1 e^{-\alpha(t_2 - t_1)} + [c_2 - 4N_2\alpha \cdot H(t_1)/(1 + 2N_2\alpha)] \cdot e^{-(1/2N_2 + 2\alpha)(t_2 - t_1)}, \tag{24}$$

where

$$c_1 = 4N_2\alpha \cdot [F(t_1) + H(t_1) + NH(t_1)/2]/[1 + 2N_2\alpha] = 4N_2\alpha \cdot e^{-\alpha t_1}/[1 + 2N_2\alpha] \quad \text{and}$$

$$c_2 = [NH(t_1) - 4N_2\alpha \cdot F(t_1)]/[1 + 2N_2\alpha].$$

Calculating the frequency of ancestral alleles at time t_2, we get from equations (22)–(24)

$$F(t_2) + H(t_2) + NH(t_2)/2 = e^{-\alpha t_2}.$$

The frequency of ancestral genes in a present population therefore yields an estimate for the genetic identity between the present population and its ancestral population. This estimate is not a function of the population size and thus is not affected by genetic drift, even if the population size differs in succeeding time periods, e.g. if $N = N_1$ for $0 \le t \le t_1$ and $N = N_2$ for $t_1 < t \le t_2$. By mathematical induction this proof can be generalized for any number of finite time periods.

The model assumes that genotypes can unequivocally be grouped into the four genotype classes. In practice, however, we cannot distinguish between differences that are caused by drift and mutation. From formula (18) the proportion of identical alleles present in two populations, $\hat{I}$, is estimated that pools drift and mutation and ignores the number of ancestral alleles that are present in one of the considered populations and absent in the other, $I \cdot (1 - I)$. Taking this fact into consideration the genetic identity value, $\hat{I}$, estimated by formula (18) can be corrected easily through $I = \hat{I} + I \cdot (1 - I)$. An estimate of the genetic identity, I, between a population and the common ancestral population then is the square root of the estimate given by formula (18).

Methods and applied data

The frequencies of the genotype classes were calculated as described by Tomiuk & Loeschcke (1991). The average over all studied loci yields the expected frequency distribution of genotype classes which was used for the calculation of the genetic identity proposed here. Comparing two species, 1 and 2, the genetic identity between each species and their common ancestral population, $I_1 = e^{-\alpha_1 t}$ and $I_2 = e^{-\alpha_2 t}$, was estimated from the frequency of the genotype classes (the square root of the estimate that is given by equation 18). The product, $I_1 \cdot I_2$, is an estimate of the genetic identity between species 1 and species 2. The measure proposed here for the genetic identity, I_{new}, was estimated from the mutation-drift model and

compared with the genetic identity, I_{TL}, of the corresponding pure mutation model (Tomiuk & Loeschcke, 1991). The least square estimate of the latter identity is closely correlated with Nei's identity (Nei, 1972; see Tomiuk & Loeschcke, 1991). The logarithmic transformation of both measures gives the genetic distance functions, D_{TL} and D_{new}, which are linearly dependent on evolutionary time and mutation rate.

The electrophoretic data of 10 primate species, *Homo sapiens*, *Pan troglodytes*, *Pan paniscus*, *Gorilla gorilla*, *Macaca fascicularis*, *Macaca mulatta*, *Cercopithecus aethiops*, *Saimiri sciureus*, *Cebus apella* and *Aotus trivirgatus*, were used to study the attributes of the new measure of genetic identity proposed here. Allele frequencies at 49 protein loci of 10 Catarrhine and Platyrrhine species were taken from a phylogenetic study by Schmitt *et al.* (1990), and allele frequencies at

Table 1 List of proteins studied electrophoretically in at least two species of hominoids, cercopithecoids (old world monkeys) and ceboids (new world monkeys)

		Hominoids				Cercopithecoids			Ceboids		
Protein	Locus	Hs	Pt	Pp	Gg	Mf	Mm	Cae	Ss	Ca	At
Alcohol dehydrogenase(2)	*ADH1*	−	+	−	−	+	+	−	−	−	−
β-Hydroxybutyrate dehydrogenase(2)	*HBDH*	+	−	−	−	+	−	+	−	−	−
Glycerol-3-phosphate dehydrogenase	*GPD*	+	+	+	+	+	−	+	+	−	−
Sorbitol dehydrogenase	*SDH*	+	+	−	+	+	+	+	−	+	−
Lactate dehydrogenase	*LDHA*	+	+	+	+	+	+	+	+	+	+
	LDHB	+	+	+	+	+	+	+	+	+	+
Malate dehydrogenase	*MDH1*	+	+	+	+	+	+	+	+	+	+
	MDH2	+	+	+	+	+	+	+	−	−	+
Malic enzyme	*MOD1*	+	+	+	+	+	+	+	−	−	−
	MOD2	+	+	+	+	+	+	+	−	−	−
Isocitrate dehydrogenase	*ICD1*	+	+	+	+	+	+	+	+	+	+
	ICD2	+	+	+	+	+	+	+	+	+	+
Phosphogluconate dehydrogenase	*PGD*	+	+	+	+	+	+	+	+	+	+
Glucose-6-phosphate dehydrogenase	*G6PD*	+	+	+	+	+	+	+	−	−	−
α-Keto acid reductase	*KAR*	+	−	−	−	+	−	−	−	+	+
Glutamate dehydrogenase(2)	*GLUDH*	+	−	−	−	+	−	+	−	−	−
Diaphorase	*DIA1*	+	+	+	+	+	+	+	+	+	+
	DIA2	+	+	+	+	+	+	+	+	+	+
Catalase	*CAT*	+	−	−	−	+	+	+	−	−	−
Glutathione peroxidase(2)	*GPX*	+	−	−	−	+	−	−	−	−	−
Superoxide dismutase	*SOD1*	+	+	+	+	+	+	+	+	+	+
	SOD2	+	+	+	+	+	+	+	−	−	−
Nucleoside phosphorylase(2)	*NP*	+	+	−	−	+	−	+	−	−	−
Glutamate oxaloacetate transaminase	*GOT1*	+	+	+	+	+	+	+	+	+	+
	GOT2	+	+	+	+	+	+	+	+	+	+
Glutamate pyruvate transaminase	*GPT*	+	+	−	−	+	+	+	+	+	+
Hexokinase(2)	*HK*	+	−	−	−	+	−	+	−	−	−
Pyruvate kinase(2)	*PK-M1*	+	+	−	−	+	−	+	−	−	−
	PK-M2	+	+	−	−	+	−	+	−	−	−
Creatine kinase(2)	*CK-A*	+	−	+	−	+	−	+	−	−	−
	CK-B	+	−	+	−	+	−	+	−	−	−
Adenylate kinase(2)	*AK1*	+	+	+	+	+	−	+	+	+	+
	AK2	+	−	−	−	+	−	+	−	−	−

Table 1 *Continued*

Protein	Locus	Hominoids				Cercopithecoids			Ceboids		
		Hs	Pt	Pp	Gg	Mf	Mm	Cae	Ss	Ca	At
Uridine monophosphate kinase(2)	*UMPK*	+	+	−	−	+	−	+	−	−	−
Phosphoglucomutase	*PGM1*	+	+	+	+	+	+	+	+	+	+
	PGM2	+	+	+	+	+	+	+	+	+	+
	PGM3	+	+	+	+	+	−	+	−	−	−
Phosphoglyceromutase(2)	*PGAM1*	+	−	−	−	+	−	+	−	−	−
	PGAM2	+	−	−	−	+	−	+	−	−	−
Galactose-1-p-uridyl transferase	*GALT*	+	+	+	+	+	+	+	+	+	+
Carboxylic ester hydrolase	*ESD*	+	+	+	+	+	+	+	+	+	+
Acylcholine acyl hydrolase(1)	*CHES*	−	−	−	−	+	+	−	−	−	−
Glyoxalase	*GLO2*	+	+	+	+	+	−	+	−	−	−
Alkaline phosphatase	*ALP*	+	−	−	−	+	+	+	−	−	−
Acid phosphatase	*ACP1*	+	+	+	+	+	+	+	+	+	+
	ACP2	+	+	+	+	+	−	+	−	−	−
Fructose-1,6-di phosphatase	*FDP*	+	+	+	+	+	−	+	−	−	−
β-Glucosidase	*βGLU*	−	−	−	−	+	+	+	−	−	−
β-Galactosidase(2)	*βGAL*	−	−	−	−	+	−	+	−	−	−
Leucine amino peptidase(1)	*LAP*	−	−	−	−	+	+	−	−	−	−
Arginase	*ARG*	−	−	−	−	+	+	+	−	−	−
Adenosine deaminase	*ADA*	+	−	−	−	+	+	+	−	−	−
Cytidine deaminase(2)	*CDA*	−	−	−	−	+	−	+	−	−	−
Aldolase	*ALD*	+	−	−	−	+	+	+	+	+	+
Carbonic anhydrase	*CA1*	+	+	+	+	+	+	+	+	+	−
	CA2	+	+	+	+	+	+	+	+	+	+
Aconitate hydratase	*ACO1*	+	+	+	+	+	−	+	−	−	−
	ACO2	+	+	+	+	+	−	+	−	−	−
Phosphopyruvate hydratase(2)	*ENO*	+	+	−	−	+	−	+	−	−	−
Ribulose-5-phosphate-3-epimerase	*RPE*	+	−	−	−	+	−	+	−	+	+
Ribose phosphate isomerase	*RPI*	+	−	−	−	+	−	+	−	+	+
Mannose phosphate isomerase	*MPI*	+	+	+	+	+	−	+	+	+	+
Glucose phosphate isomerase	*GPI*	+	+	+	+	+	+	+	+	+	+
Haptoglobin	*HP*	+	−	−	−	+	+	+	−	−	−
Haemoglobin	*HB*	+	−	−	−	+	+	+	+	+	+
Protease inhibitor	*PI*	+	+	+	+	+	+	−	−	−	−
Albumin	*ALB*	+	+	+	+	+	+	+	+	+	+
Prealbumin(1)	*PA*	−	−	−	−	+	+	−	−	−	−
Thyroxin binding prealbumin(1)	*TBPA*	−	−	−	−	+	+	−	−	−	−
Ceruloplasmin(1)	*CP*	−	−	−	−	+	+	−	−	−	−
Transferrin	*TF*	+	+	+	+	+	+	+	−	+	+

Data from Schmitt *et al.* (1990) and supplemented by (1) data from Nozawa *et al.* (1977) and (2) Schmitt & Tomiuk (unpublished results). Loci are marked by + when the allele frequencies of the protein loci are given. Hs, *Homo sapiens*; Pt, *Pan troglodytes*; Pp, *Pan paniscus*; Gg, *Gorilla gorilla*; Mf, *Macaca fascicularis*; Mm, *Macaca mulatta*; Cae, *Cercopithecus aethiops*; Ss, *Saimiri sciureus*; Ca, *Cebus apella*; At, *Aotus trivirgatus*.

22 additional protein loci studied by Nozawa *et al.* (1977) and Schmitt & Tomiuk (unpublished results) were added to the data set. In Table 1, the protein loci are listed for which the allele frequencies of at least two of the 10 primate species are known.

Results

The analysis of genotype frequencies in 10 primate species resulted in 45 combinations of species pairs. The two measures of genetic identity were applied to each species pair using a minimum of 22 loci up to a maximum of 61 loci. The distance values were within a large range ($0.02 \leq D \leq 1.20$). First the correlation between the estimates of the two distance measures, D_{TL} and D_{new}, was calculated (Fig. 1). The correlation coefficient was extremely high ($r = 0.98$, $n = 45$) and the new estimate of the genetic distance correlated linearly with the estimate obtained from the pure mutation model. Furthermore, the estimates of genetic distance that were obtained by four different procedures (Nei's measure, Nei (1972); Nei's modified measure, Hillis (1984) and Tomiuk & Graur (1988); the pure mutation model, Tomiuk & Loeschcke (1991); the mutation-drift model, this paper) were closely linearly correlated ($r > 0.95$).

The effect of the population size and the ancestral degree of heterozygosity on the genotype class distribution was analysed. We assumed that the mutation rate and the population size were constant during the total evolutionary period. Fig. 2 shows the frequencies of the genotype classes, *F*, *H* and *NH*, as a function of the genetic identity where the ancestral degree of heterozygosity $H_0 = 0.1$ and 0.5, and the parameter $M = 2N\alpha$ varied between 0.01 and infinity (for example when $\alpha = 10^{-6}$ this corresponded to a population size of *N* between 50 000 and infinity). Obviously, the expected frequencies of the genotype classes for finite population sizes strongly deviated from that expected by the mutation model ($M = \infty$), even if the population size was large. With decreasing population size, the effect on the frequency of genotype classes decreased drastically. The ancestral degree of heterozygosity greatly influenced the frequencies of genotype classes (Fig. 2) when closely related species (*I* near 1) were considered, but was minimized for distantly related species (*I* near 0).

Discussion

Most of the hitherto existing long-term measures of genetic identity between species are exclusively based on mutational changes in infinite populations. Besides mutation, however, genetic drift is known to influence the genetic structure of populations. Drift has been assumed to affect estimates of evolutionary time. Our proposed procedure for estimating the genetic identity also considers the influence of finite population size on the genetic structure of populations. Surprisingly, we found that even varying population sizes and the

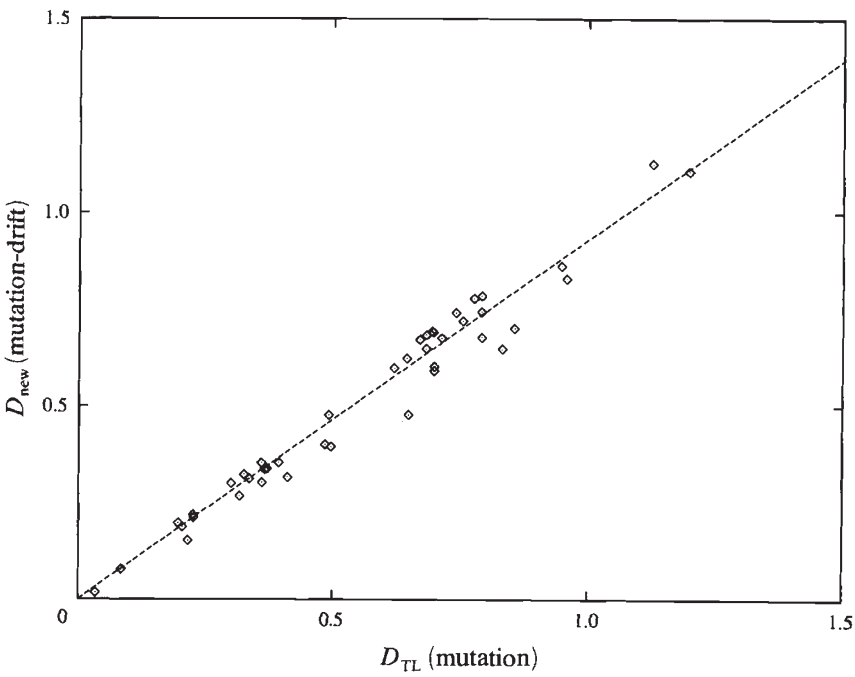

Fig. 1 Relationship between genetic distances based on the pure mutation model (D_{TL}) and on the mutation-drift model (D_{new}) based on electrophoretic data from 45 combinations of species-pairs of ten primate species. The regression $D_{new} = mD_{TL} + b$ has the estimated parameters $m = 0.94$ and $b = -0.02$ with $r = 0.98$.

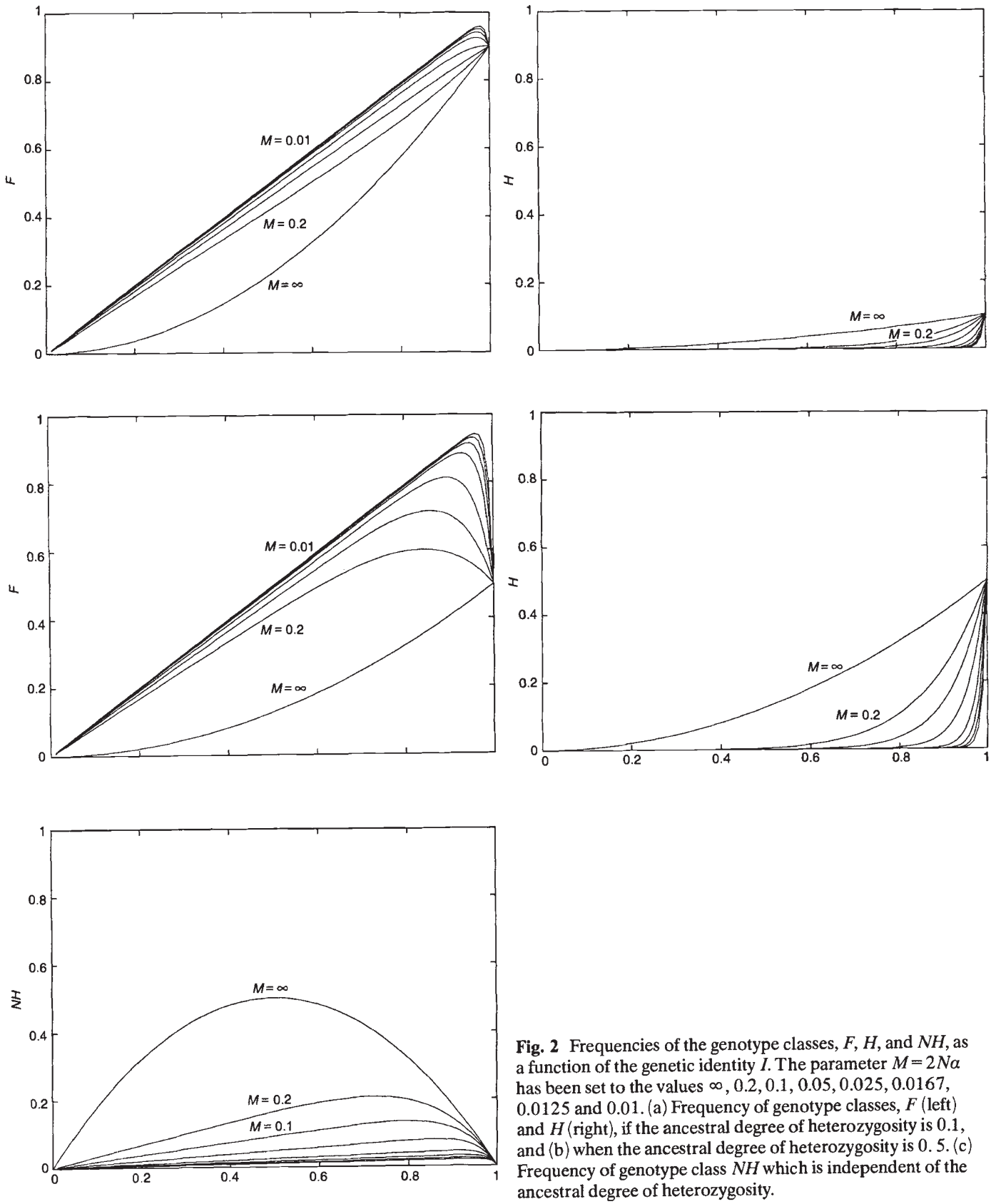

Fig. 2 Frequencies of the genotype classes, F, H, and NH, as a function of the genetic identity I. The parameter $M = 2N\alpha$ has been set to the values ∞, 0.2, 0.1, 0.05, 0.025, 0.0167, 0.0125 and 0.01. (a) Frequency of genotype classes, F (left) and H (right), if the ancestral degree of heterozygosity is 0.1, and (b) when the ancestral degree of heterozygosity is 0. 5. (c) Frequency of genotype class NH which is independent of the ancestral degree of heterozygosity.

heterozygosity of the ancestral population did not affect the new measure of genetic identity between species but the frequencies of the genotype classes in finite populations always strongly deviated from those of the pure mutation model, even if relatively large population sizes were considered.

Most of the known procedures for estimating the genetic identity between species consider the probability of finding identical alleles at single loci, and the average over all loci is taken as an approximation for the genetic identity (e.g. Cavalli-Sforza & Edwards, 1967; Nei, 1972). Our approach is the reverse; first the mean frequencies of identical alleles over all loci are calculated and their product then gives the genetic identity between the respective species. The new estimate of the genetic identity can easily be calculated directly. It is not necessary for it to be approximated by such functions as the cosine of the angle between two vectors of allele frequencies (Nei, 1972) or by a least square estimate (Tomiuk & Loeschcke, 1991). Furthermore, the estimates of the genetic identity which are obtained by some of these indirect approaches, are not reliable in certain cases. For example, even if the alleles found at one locus were identical in two species, the genetic identity is not equal to 1 when the allele frequencies are different between the two species (if we use Nei's genetic identity (1972) between two species with alleles a and b, where $a=0.1$ and $b=0.9$ for species 1 and $a=0.9$ und $b=0.1$ for species 2, we get $I=0.22$). Such kinds of differences, however, are obviously caused by genetic drift. This problem has been recognized by Reynolds *et al.* (1983) which led them to construct a pure drift model of genetic identities for short-term population differentiation. An advantage of our approach to measuring genetic identity is that it incorporates mutation, but at the same time allows us to consider the temporal influence of genetic drift in closely related species. In this case the genotype class distribution provides the most information on the order of magnitude of random drift.

The close correlation of identity values between former measures (Nei, 1972; Hillis, 1984; Tomiuk & Graur, 1988; Tomiuk & Loeschcke, 1991) and the measure proposed here appears now to be valuable. The conclusions from hitherto existing studies on the genetic similarity of species seem to be reliable. However, even if the correlation between the mutation and the mutation-drift model is close, as shown for the large range of the analysed data, the functional dependence seems not to be simple and has some consequences for evolutionary analyses. Fig. 1 suggests that the bias is highest for identity values around 0.5.

Beyond the studies on protein polymorphisms, our findings can be extrapolated to the evolutionary analysis of any genetic variability differentiating populations or species. The use of the frequencies of single alleles does not result in unbiased estimators for the genetic identity, and as has been demonstrated even the frequencies of the pooled genotype classes are considerably influenced by genetic drift. The distribution of genotype classes depends on the order of magnitude of $1/N$ and α, both of which influence the genetic identity. The mean frequency of observed ancestral alleles is the basis for a good estimate of the genetic identity between species which counterbalances the bias at single loci caused by genetic drift. However, the analysis of the genotypic population structure additionally can provide information on the influence of the order of magnitude of the population size. But care must be used when the parameter $2N\alpha$ is estimated because the genotype distributions converge rapidly to threshold functions. This also implies that the estimation of the ancestral degree of homozygosity from equations (14), (16) and (17) can be biased strongly when population sizes are small.

Acknowledgments

We are grateful to Bob Krebs, Jotun Hein, Claus Hedegaard and an anonymous reviewer for comments on the manuscript, to Bernt Gulbrandtsen and Claus Hedegaard for help in preparing the figures and to the German Research Council (DFG-Ga 342/3-1) and the Danish Research Council for Natural Sciences (grant 11-9639-1) for financial support.

References

CAVALLI-SFORZA, L. L. AND EDWARDS, A. W. F. 1967. Phylogenetic analysis: models and estimation procedures. *Am. J. Hum. Genet.*, **19**, 233–257.

CROZIER, R. H. AND KUSMIERSKI, R. M. 1994. Genetic distances and the setting of conservation priorities. In: Loeschcke, V., Tomiuk, J. and Jain, S. K. (eds) *Conservation Genetics*, pp. 227–238. Birkhäuser Verlag, Basel.

FISHER, R. A. 1930. *The Genetical Theory of Natural Selection.* Clarendon Press, Oxford.

HARTL, D. L. AND CLARK, A. G. 1989. *Principles of Population Genetics.* Sinauer, Sunderland, MA.

HILLIS, D. M. 1984. Misuse and modification of Nei's genetic distance. *Syst. Zool.*, **33**, 238–240.

NEI, M. 1972. Genetic distance between populations. *Am. Nat.*, **106**, 283–292.

NOZAWA, K., SHOTAKE, T., OHKURA, Y. AND TANABE, Y. 1977. Genetic variations within and between species of Asian macaques. *Jap. J. Genet.*, **52**, 15–30.

REYNOLDS, J., WEIR, B. S. AND COCKERHAM, C. C. 1983. Estimation of the coancestry coefficient: basis for a short-term genetic distance. *Genetics*, **105**, 767–779.

SCHMITT, J., GRAUR, D. AND TOMIUK, J. 1990. Phylogenetic relationships and rates of evolution in primates: allozymic data from Catarrhine and Platyrrhine species. *Primates*, **31**, 95–108.

TOMIUK, J. AND GRAUR, D. 1988. Nei's modified identity and distance measure and their sampling variances. *Syst. Zool.*, **37**, 156–162.

TOMIUK, J. AND LOESCHCKE, V. 1991. A new measure of genetic identity between populations of sexual and asexual species. *Evolution*, **45**, 1685–1694.

TOMIUK, J. AND LOESCHCKE, V. 1992. Evolution of parthenogenesis in the *Otiorhynchus scaber* complex. *Heredity*, **68**, 391–397.

VANE-WRIGHT, R. I., HUMPHRIES, C. J. AND WILLIAMS, P. H. 1991. What to protect? — Systematics and the agony of choice. *Biol. Conserv.*, **55**, 235–254.

WITTING, L., McCARTHY, M. A. AND LOESCHCKE, V. 1994. Multispecies risk analysis, species evaluation, and biodiversity conservation. In: Loeschcke, V., Tomiuk, J. and Jain, S. K. (eds) *Conservation Genetics*, pp. 239–251. Birkhäuser Verlag, Basel.

WRIGHT, S. 1931. Evolution in Mendelian populations. *Genetics*, **16**, 97–159.