Heredity **78** (1997) 655–664 *Received 23 September 1996*

Estimating selective coefficients of allelic substitutions from patterns of interspecific allozymic mobility difference

D. J. COLGAN

*Ken and Yasuko Myer Molecular Evolutionary Biology Unit, The Australian Museum, 6 College Street, Sydney 2000, Australia**

The neutral hypothesis of molecular evolution predicts that there should be a 1:1 ratio between allozymes in a species which are faster or slower than allozymes at the corresponding loci in a second species. Methods are presented in this paper for estimating selective differentials from observed values of r, the deviation from a fraction of 1/2 in the proportion of substitutions conferring a mobility change in a specified direction. Although the methods were developed for allozymes, they are readily applied to cases where two taxa may differ in other properties, such as DNA sequence. r is related to the selection coefficient s under a 'genic selection' model. If a change in one direction is favoured in one species and a change in the other direction is equally favoured in the second species, both s and N are assumed constant and only one substitution has occurred at each locus, then the estimate of s is approximately $(\ln(1/2+r)-\ln(1/2-r))/4N$. If selective coefficients follow an unspecified distribution $f(s)$ and N is assumed constant, then the average value of s lies between $r/2N(1/2+r)$ and $r/2N(1/2-r)$. Specifying the distribution of s to enable point estimation of $E(s)$ is mathematically difficult for the usual probability density functions. A new family of distributions is suggested to overcome these difficulties. The analysis is extended to cover estimation of r where more than one substitution has occurred. The range of the various estimates of s provided by these methods from electrophoretic data is usually below $1/N$ and mostly substantially so.

Keywords: electrophoresis, isozymes, neutral theory, selective coefficients.

Introduction

The neutral hypothesis of molecular evolution, in its present most general forms (Kimura, 1986, 1990; Ohta, 1992) suggests that there is a range of bases which satisfy currently prevailing functional constraints at a DNA sequence site. Where more than one base is acceptable, the fate of mutations between them will generally be determined by genetic drift. Mutations to bases which cannot satisfy the constraints will usually be eliminated by selection, although a non-negligible fraction of such deleterious mutants will also be fixed by chance (Fisher, 1930; Ohta, 1973). Some fraction, held to be very small, of mutations is functionally superior to the incumbent base(s). Such mutations will be favoured by selection. The selectionist alternative to the neutral hypothesis proposes that a substantial fraction of allelic substitutions occurs because the successful form is selectively favoured.

*E-mail: donc@amsg.austmus.gov.au

One approach to testing the neutral hypothesis, using information on allozymic variation, was suggested by Ward & Skibinski (1982) and Colgan (1992). Under the hypothesis there should be no correlation between loci in the relative mobilities of successfully substituting allozymes. Which of the alternative acceptable bases occupies a given site at a given time is a matter of chance if the process has established a steady-state. If there is a charge difference between two amino acids specified by acceptable bases, then a new neutral mutation from one such base to another should have an equal probability of increasing or decreasing net negative charge. Thus, the overall frequency amongst successful new mutants of those which are faster than their antecedent is expected to be the same as the frequency of those which are slower.

Initial analyses of electrophoretic data from animals did not support the prediction of a 1:1 ratio of faster to slower mobilities (Colgan, 1992). Deviations from this expected ratio do not, however, provide an estimate of the selective advantage of

substituting alleles. Such an estimate is attempted here by mathematical modelling of the relationship of the deviations and selective coefficients. Although this modelling has been couched in terms of patterns of allozymic substitution, it is readily applicable to other situations. In particular, it can be extended to comparison of DNA sequences.

The genic selection model and parameter definition

The species are assumed diploid, with two alleles *a* and *b* at each locus. A change from allele *b* to allele *a* is assumed to confer a negative (−) shift in relative mobility and a change from *a* to *b* is assumed to give a positive (+) shift. In actual substitutions, there would be variation between loci in the differences of the relative mobilities of *a* and *b* alleles (owing to changes affecting conformation, variation in the charge difference between different types of substitutions, etc.). Such variations would not, however, affect the mathematics of the model, which is based solely on the direction of the changes. The relative fitness of the heterozygote *ab* is assumed to lie exactly halfway between the fitnesses of the two homozygotes. This standard 'genic selection' model is extensively discussed by Crow & Kimura (1970), Nei (1987) and Gale (1991).

f_l = the observed frequency of species pairs in which the substitution events produce relative mobility changes in the same direction in each of two loci — the frequency of 'like' changes.
f_u = the observed frequency of species pairs in which the substitution events produce relative mobility changes in opposite directions in each of two loci — the frequency of 'unlike' changes.
$Fr(b)$ = the frequency of substitutions of *b*-type alleles.
$Fr(a)$ = the frequency of substitutions of *a*-type alleles.
r = the deviation from a fraction of 1/2 in the proportion of substitutions conferring a mobility change in a specified direction. For instance, if the probability of substitution in the + direction equals $(1/2+r)$ then the probability in the − direction equals $(1/2-r)$.
z = the deviation from a fraction of 1/2 in the proportion of total substitutions which occur in a given species. Thus the probability of a substitution occurring in species A is $(1/2+z)$ and the probability of it occurring in species B is $(1/2-z)$.
n = the number of substitutions between two species which have occurred at a locus.

$F_s(b)$ = the probability of fixation of new *b* type alleles for a given value of s. This is given by:

$$\frac{1-e^{-2s}}{1-e^{-4Ns}}$$

(Fisher, 1930; Malécot, 1952; Kimura, 1962). For s small, this is approximately equal to $2se^{4Ns}/(e^{4Ns}-1)$.

$F_{-s}(a)$ = the probability of fixation of new *a*-type alleles. For $|s|$ small, this is approximately:

$$-2s/(1-e^{4Ns})$$

as the formula for $F(s)$ also holds for s small and negative (Crow & Kimura, 1970).

Estimation of *r*

When species pairs have undergone substitutions at two loci

Initially, it is assumed that mobility changes in a specified direction are favoured in one species, and are also at a disadvantage in the other. The expected frequencies of like and unlike pairs of changes are detailed in Table 1. Then r may be estimated by subtracting the observed frequency, f_u, of unlike pairs of changes from the frequency of like pairs, f_l. These frequencies are found by summation of the columns in Table 1 to be:

$$\begin{aligned} f_l &= (1/2+r)^2[(1/2-z)^2+(1/2+z)^2+2(1/2-z)(1/2+z)] \\ &\quad +(1/2-r)^2[(1/2-z)^2+(1/2+z)^2 \\ &\quad +2(1/2-z)(1/2+z)] \\ &= (1/2+r)^2[(1/2-z)+(1/2+z)]^2 \\ &\quad +(1/2-r)^2[(1/2-z)+(1/2+z)]^2 \\ &= (1/2+r)^2+(1/2-r)^2 = 1/2+2r^2. \end{aligned} \quad (1)$$

Similarly,

$$f_u = 1/2-2r^2. \quad (2)$$

Whence,

$$r = [(f_l-f_u)/4]^{1/2}. \quad (3)$$

If there is a differential probability of substitution of the two types of allozymes (faster or slower) in only one species, rather than in both, then the estimate of r is increased by a factor of $1/(1/2-z)$. So:

$$r = (f_l-f_u)^{1/2}/(1-2z). \quad (4)$$

The effect of supposing that differential probabilities occur in only one species depends on z, which may take values between $-1/2$ and $+1/2$. When

Table 1 The expected frequencies of the 16 possible combinations where there are two substitutions, which may alter mobility in either direction, in each of two species. The relative frequency of changes for species A are $(1/2-r)$ for the + direction and $(1/2+r)$ for the − direction. For species B, the frequencies are $(1/2+r)$ for the + direction and $(1/2-r)$ for the − direction. The proportion of substitutions in species A is $(1/2+z)$ and in species B is $(1/2-z)$. Combinations resulting in relative mobility changes in the same direction for the two loci are written on the left-hand side

	Similar mobilities			Opposite mobilities		
	Species			Species		
Locus	A	B	Frequency	A	B	Frequency
1		+	$(1/2+r)^2(1/2-z)^2$		+	$(1/2-r)(1/2+r)(1/2-z)^2$
2		+			−	
1		−	$(1/2-r)^2(1/2-z)^2$		−	$(1/2-r)(1/2+r)(1/2-z)^2$
2		−			+	
1	+		$(1/2-r)^2(1/2+z)^2$	+		$(1/2-r)(1/2+r)(1/2-z)^2$
2	+			−		
1	−		$(1/2+r)^2(1/2+z)^2$	−		$(1/2-r)(1/2+r)(1/2-z)^2$
2	−			+		
1	+		$(1/2-r)^2(1/2+z)(1/2-z)$	+		$(1/2-r)(1/2+r)(1/2-z)(1/2+z)$
2		−			+	
1		−	$(1/2-r)^2(1/2+z)(1/2-z)$		+	$(1/2-r)(1/2+r)(1/2-z)(1/2+z)$
2	+			+		
1	−		$(1/2+r)^2(1/2+z)(1/2-z)$	−		$(1/2-r)(1/2+r)(1/2-z)(1/2+z)$
2		+			−	
1		+	$(1/2+r)^2(1/2+z)(1/2-z)$		−	$(1/2-r)(1/2+r)(1/2-z)(1/2+z)$
2	−			−		

$z = -1/2$, all substitutions occur in the species with differential probabilities. If $z = 1/2$, they would all occur in the other species and (4) is undefined.

When species pairs have undergone substitutions at three loci

The observed distributions of relative mobilities where three loci differ between species pairs may be used to estimate r. Here f_1 is defined as the frequency of comparisons where one or other species has the faster allozyme(s) at two loci and f_3 as the frequency of comparisons where one or other species has the faster allozyme(s) at all three loci. If there are different probabilities for substitution in the two directions in both species, with the favoured direction in one species being at an equal and opposite disadvantage in the other species, then r is estimated by:

$$r = [(3f_3 - f_1)/3]^{1/2}/2. \quad (5)$$

As before, if there are differential probabilities of substitution of faster and slower allozymes in only one of the species, then the estimate of r can be shown to be:

$$r = [(3f_3 - f_1)/3]^{1/2}/(1-2z). \quad (6)$$

An alternative method of estimating r

As detailed in Colgan (1992), comparing mobilities of the allozymes of two species can be considered a series of Bernoulli trials with probability of success p. The number of loci for which a species has faster allozymes ranges from 0 to k (k even) and from 0 to $(k-1)$ (k odd). The distribution of these is a random variable. The entire set of observations can be treated as a sample from the random variable with density:

$$g(x) = \Sigma K_i g_i(x)/N, \quad (7)$$

where N is the total number of comparisons, K_i is the number of species pairs which have i substitutions and $g_i(x)$ is the density of the random variable for i substitutions. The $g_i(x)$ are found by summing probabilities from a binomial table with the appropriate value of p. For instance, if p equals 0.45 and four loci have substitutions, then the probability of having one of the species with three allozymes

faster than those in the other species is 0.2995 plus 0.2005 (= 0.5).

$g(x)$ provides an alternative method of estimating r — by finding (through iteration) the value of p that produces concordance between the means of the observed and expected distributions. $g(x)$ can also be used to test for randomness in the number of loci for which a species has faster allozymes than those in the species with which it is compared. The observed means of the overall distribution are tested against the neutral expectation using the standardized normal statistic:

$$z = (\bar{x}-\mu)/(\sigma/\sqrt{N}). \tag{8}$$

The effect of multiple substitutions

The comparison of relative mobilities between species does not discriminate between cases where there is only one substitution and where there are multiple substitutions. Under the neutral hypothesis multiple substitution does not alter the expected 1:1 ratio of mobility classes, but if r is nonzero for each substitution then successive instances would increase observed deviations from expectation. The magnitude of this effect, assuming that there are n substitutions all in the one species, can be calculated as follows. The distribution of genotypes with a specified number of substitutions conferring a faster mobility and a specified number conferring a slower mobility is found by expansion of:

$$(Fr(b)+Fr(a))^n.$$

The excess of apparently faster allozymes over slower is determined by the series:

$$[(Fr(b))^n-(Fr(a))^n]$$

$$+\left[\binom{n}{n-1}Fr(b)Fr(a)((Fr(b))^{n-2}-(Fr(a))^{n-2})\right]$$

$$+\left[\binom{n}{n-2}(Fr(b))^2(Fr(a))^2((Fr(b))^{n-4}-(Fr(a))^{n-4})\right]\ldots$$

Replacing $Fr(b)$ by $(1/2+r)$ and $Fr(a)$ by $(1/2-r)$, and ignoring terms in higher powers of r in the expansion of successive terms within square brackets, this is:

$$2\frac{nr}{2^{n-1}}+2\binom{n}{n-1}(n-2)\frac{r}{2^{n-1}}+2\binom{n}{n-2}(n-4)\frac{r}{2^{n-1}}\ldots,$$

which equals:

$$\frac{2r}{2^{n-1}}\sum_{i=0}^{t}\binom{n}{n-i}(n-2i), \tag{9}$$

where t equals $(n/2-1)$ for n even and $(n-1)/2$ for n odd. For $n = 3$ or 4, the excess of faster allozymes is $3r$, for $n = 5$ or 6, it is $60r/16$, and so on. For $n = 1$, the excess is $2r$, so that a doubling of the observed excess with a constant value of r would require at least five substitutions.

Relating *s* to *r* in the genic selection model

s *assumed constant*

s and N are assumed constant. It is assumed that the probability of mutation to b-type alleles in an a background (with relative fitnesses of 1 for aa, $1+s$ for ab and $1+2s$ for bb) is the same as that for mutation to an a-type allele in a b background. The relative fitnesses for this latter scenario will be assumed to be $1-2s$ for aa, $1-s$ for ab and 1 for bb.

The ratio of the frequencies of substitutions of b alleles and a alleles, which is estimated by $(1/2+r)/(1/2-r)$, theoretically equals the ratio of the probabilities of fixation of new mutations. Approximately:

$$\frac{\frac{1}{2}+r}{\frac{1}{2}-r}=\frac{2s}{1-e^{-4Ns}}\cdot\frac{1-e^{4Ns}}{-2s}$$

$$=e^{4Ns}.$$

This implies that:

$$4Ns = \ln(\tfrac{1}{2}+r)-\ln(\tfrac{1}{2}-r). \tag{10}$$

It could be argued that the relative fitnesses for mutation to a in a b-type background are better represented as $1-2s/(1+2s)$ for aa, $1-s/(1+2s)$ for ab and 1 for bb. For s small, the quantitative difference between the two estimates of Ns can be seen to be small. With the alternative fitnesses, it can be shown that:

$$4Ns \approx \ln(1+2r)-\ln(1-2r)+\ln(1+2s).$$

s *assumed variable with an unspecified distribution*

The assumption of constant s is relaxed by assuming it follows the unspecified density function $f(s)$. N is assumed constant. The relative genotypic fitnesses for new b-type alleles in an a background are 1 for aa, $1+s$ for ab and $1+2s$ for bb for any given value

of s. The relative fitnesses for new a-type alleles in a b background will be assumed to be $1-2s$ for aa, $1-s$ for ab and 1 for bb. It is assumed that the density $f(s)$ is the same for both types of substitutions. This is clearly not the case for new mutations in general (Mukai *et al.*, 1965) but is more realistic for allozymic mutations which are usually detected as functional products. That proportion of allozymic variants which are highly deleterious mutations is ignored as being rapidly eliminated from the population. Then, if x and y are the extremes of the range of s, the ratio of the frequency of substitutions by b-type alleles to the frequency of substitutions by a-type alleles is given by:

$$\frac{\int_x^y f(s)e^{4Ns}\frac{2s}{e^{4Ns}-1}ds}{\int_x^y f(s)\frac{2s}{e^{4Ns}-1}ds} = \frac{\int_x^y 2sf(s)+\int_x^y \frac{2sf(s)}{e^{4Ns}-1}ds}{\int_x^y \frac{2sf(s)}{e^{4Ns}-1}ds} = \frac{1/2+r}{1/2-r}. \tag{11}$$

Then,

$$E(s) = \frac{r}{1/2-r}\int_x^y \frac{2sf(s)}{e^{4Ns}-1}ds < \frac{r}{2N(1/2-r)}\int_x^y \frac{e^{4Ns}-1}{e^{4Ns}-1}f(s)ds = \frac{r}{2N(1/2-r)}. \tag{12}$$

Alternatively,

$$\frac{\int_x^y f(s)e^{4Ns}\frac{2s}{e^{4Ns}-1}ds}{\int_x^y f(s)\frac{2s}{e^{4Ns}-1}ds} = \frac{\int_x^y e^{4Ns}f(s)\frac{s}{e^{4Ns}-1}ds}{\int_x^y \left(\frac{1-e^{4Ns}}{e^{4Ns}-1}+\frac{e^{4Ns}}{e^{4Ns}-1}\right)sf(s)ds} = \frac{\int_x^y e^{4Ns}f(s)\frac{s}{e^{4Ns}-1}ds}{-\mathrm{E}(s)+\int_x^y e^{4Ns}f(s)\frac{s}{e^{4Ns}-1}ds} = \frac{1/2+r}{1/2-r}. \tag{13}$$

Rearranging this equation and noting that:

$$\frac{4Ns}{e^{4Ns}-1} = \frac{1}{\sum_{i=1}^{\infty}\frac{(4Ns)^{i-1}}{i!}} > \frac{1}{\sum_{i=0}^{\infty}\frac{(4Ns)^{i}}{i!}} = \frac{1}{e^{4Ns}}, \tag{14}$$

we have:

$$E(s) = \frac{2r}{1/2+r}\frac{1}{4N}\int_x^y \frac{4Ns}{e^{4Ns}-1}e^{4Ns}f(s)ds > \frac{r}{2N(1/2+r)}\int_x^y \frac{e^{4Ns}}{e^{4Ns}}f(s)ds = \frac{r}{2N(1/2+r)}, \tag{15}$$

so that:

$$\frac{r}{2N(1/2+r)} < E(s) < \frac{r}{2N(1/2-r)}. \tag{16}$$

A similar range can be established when N is variable between taxa but constant within them (during the substitution) if the distributions of s and N (density $l(N)$) are statistically independent. Then, following iterative integration over ds and subsequently over dN, a generalized form of the Integral Mean Value theorem guarantees (Olmsted, 1961) that there is a value of N, say N^*, such that:

$$\int_{n_1}^{n_2}\frac{1}{N}l(N)dN = \frac{1}{N^*}\int_{n_1}^{n_2}l(N)dN = \frac{1}{N^*}. \tag{17}$$

N^* can be substituted for N in eqn (16), although it should be noted that N^* is not the average value of N. Any attempt to calculate N^* would require specification of the distribution of N.

Specifying the distribution of s

N is assumed constant. As substitutions of a-type alleles are also being considered, the absolute value of $2s$ should be one or less. The probability density should be highest for small values of s and decline rapidly for larger values. The mathematics seems intractable for previously suggested exponential (Ohta, 1977) and gamma (Kimura, 1983) distributions (reviewed by Gillespie, 1991) and for other commonly used distributions (Normal, Poisson, Cauchy, etc.). $f(s)$ might be chosen to be c/s (for c a constant) but this is too sensitive to the lower limit of integration, which must be changed from 0 as the density would be infinite at that point. More promising is the family of distributions specified by:

$$f(s) = c(e^{-4N(k-1)s} - e^{-4Nks}) = \frac{c(e^{4Ns}-1)}{e^{4Nks}} \quad (18)$$

where k (>1) is a constant chosen by the investigator or fitted to the data, and c is a constant depending on k and the upper limit of integration. $f(s)$ is defined at 0 (as 0), but is an increasing function of s until attaining a maximum at $s = (\ln(k) - \ln(k-1))/4N$. The vanishing of the density at 0 probably has some effect on the density's biological accuracy. The function certainly decreases at values of s above $1/N$ and the total probability density below this value is greater than $(1/N)f(1/N)$ for moderate values of k (>1.16, Appendix 1). This family of densities may also be varied, once k is set to a value, by specifying a constant indicating what proportion of new mutants have negative values of s if it is assumed that more are disadvantageous than are advantageous. Other properties of the family of densities are given in Appendix 1.

The ratio of the frequency of substitutions by b-type alleles to the frequency of substitution by a-type alleles is given by:

$$\frac{\int_0^{1/2} f(s)e^{4Ns}\frac{2s}{e^{4Ns}-1}ds}{\int_0^{1/2} f(s)\frac{2s}{e^{4Ns}-1}ds} = \frac{\int_0^{1/2} 2se^{-4N(k-1)s}ds}{\int_0^{1/2} 2se^{-4Nks}ds}$$

$$= \frac{\left[-e^{-4N(k-1)s}\frac{4N(k-1)s+1}{16N^2(k-1)^2}\right]_0^{1/2}}{\left[-e^{-4Nks}\frac{4Nks+1}{16N^2k^2}\right]_0^{1/2}}$$

$$= \frac{\frac{1}{(k-1)^2}(1-e^{-2N(k-1)}(2N(k-1)+1))}{\frac{1}{k^2}(1-e^{-2Nk}(2Nk+1))} \quad (19)$$

$$= \frac{k^2}{(k-1)^2}$$

$$= \frac{1/2+r}{1/2-r}.$$

Using the last two expressions on the right-hand side of (19) to find r in terms of k:

$$r = \frac{2k-1}{4k^2-4k+2}.$$

Using this expression it can be shown from equation (A2) that:

$$E(s) \simeq \frac{(2k-1)r}{N((2k-1)-2r)}. \quad (20)$$

Alternatively, if it is wished to estimate k from the data (with limits of integration from 0 to 1/2), rearranging the last two expressions on the right-hand side of eqn (19) and solving the resulting quadratic gives:

$$k = \frac{1+2r\pm\sqrt{1-4r^2}}{4r}. \quad (21)$$

Estimation of *r* from observed substitutions

Colgan (1992) found significant deviations from expectation in three large taxonomic groups, the insects, reptiles and fish, and in the combined data set. Most of the 'Results' section of Colgan (1992) was devoted to minimizing the statistical effects of outliers, as it was recognized that these might be caused by nomenclatural problems or mis-scoring. Subsequently, Slatkin (M. Slatkin, personal communication) has identified nomenclatural problems in three references which included outlying observations. For this paper, all references have been re-examined and only those which explicitly state that allozymic designations (either alphabetical, ordinal or other numerical) are in order of electrophoretic mobility have been used in analyses. These references have been augmented by data which have become available since 1992 and which have been assessed with the more stringent admission criteria. The data listing has been lodged with *Heredity* and is available from the author or on the World Wide Web at http://www.austmus.gov.au.

Despite the more stringent criteria, statistical testing continues to suggest that the patterns of allozymic substitution are nonrandom. In the following, 'locus-level' indicates the set of comparisons for species which differ at a particular number of loci. Including all the data in Table 2, and using eqn (8), the probability of obtaining the observed mean by chance is less than 0.03. The probability remains less than 0.05 if the singleton outlier at locus-level 17 with a difference of 11 is excluded. The probability of obtaining the observed sample sum of squares for all data is less than 0.01 and remains less than 0.05 if the 11-difference outlier is ignored. For χ^2-tests of

Table 2 Observed numbers of types of relative mobility differences between species. The data considered in this paper comprise those used in Colgan (1992) with additions from subsequent literature searches and emendations due to the use of the more stringent admission criteria explained in the text. Methods of comparing taxa are explained in Colgan (1992). Rows indicate the number of loci where substitutions have occurred between the members of a given pair. Columns headed by numerals show the absolute difference between the numbers of loci for which each member of the pair has the allozyme(s) with greater anodal mobility. The second-last column is the observed mean and the last the ratio of the observed and expected means for each row. The last two rows give the total expected and observed numbers for each column

Difference Substitutions	0	1	2	3	4	5	6	7	8	9	10	11	Observed mean	O/E mean ratio
2	30		33										1.05	1.05
3		34		21									1.73	1.15
4	25		23		4								1.19	0.79
5		21		23		2							2.17	1.16
6	11		19		7		3						2.10	1.12
7		24		10		10							2.36	1.08
8	6		12		9		1						2.34	1.07
9		13		9		5		1					2.57	1.04
10	5		11		3								1.79	0.73
11		6		5		4				1			3.13	1.16
12	2		5		4		1				1		3.23	1.20
13		4		2		6		1					3.62	1.23
14	2		3		2		1				1		3.33	1.13
15		3				5				1			4.11	1.31
16			2		1				1				4.00	1.27
17		1				1		1				1	6.00	1.78
Observed total	81	106	108	70	30	33	6	3	1	2	2	1		
Expected total	81	126	106	65	31	19	8	5	2	1				

goodness of fit, $g(x)$ (eqn 8) was used to obtain expected values. Pooling all cells with differences of seven or more gives a statistic (7 d.f.) with $P<0.05$.

Fourteen of 16 locus-levels have observed means greater than expected. This is the same as for the original data set and is significant at the 0.01 level using a sign test. χ^2-tests of goodness of fit are significant (at the 0.05 level) for locus-levels 3 and 5 and nearly significant for locus-levels 7 and 11. These results suggest that the differences from the neutral expectation are not caused by any remaining mis-scoring/outliers, or to particular numbers of loci in the comparisons.

In the original data set, there was a significant regression of the ratio of observed and expected means for each locus level (Y) on the number of loci at which species differed (X). The regression equation for the amended data set is:

$$Y = 0.89 + 0.027X.$$

The linear coefficient is no longer significant.

The estimate of r from eqn (3) using the data for locus-level 2 in Table 2 is 0.109; from eqn (4), with $z = 0$, 0.218; from eqn (5) using the data for locus-level 3, $r = 0.209$; from eqn (6), with $z = 0$, 0.419. For the present overall data, the value of p which produces concordance between observed and expected means in eqn (8) is ≈ 0.43, giving an estimated r of about 0.07.

Discussion

Estimates of r can be applied to eqns (10) or (16) using a range of assumed values of N or N^* to give estimates of s or its range. For instance, with $r = 0.10$, eqn (10) gives s as $0.101/N$, and eqn 16, gives a range for the average value of s of ($0.083/N$, $0.125/N$). If $r = 0.07$, eqn 10 gives $0.07/N$ as an estimate of s. The maximum estimated value of r (0.419 from eqn 6) in the data was for comparisons of species pairs which have 'fixed' differences at three loci, with substitutions all occurring in only one species. This gives an estimate of s from eqn (10) of $0.607/N$ and range for the average value from eqn (16) of $0.228/N$, $2.586/N$. The unusually high upper limit here can be put in perspective by noting the

major effects of differences between estimates of r when this parameter is large. For instance if r is 0.3, the upper limit of eqn (16) is only $0.75/N$.

The estimates of s in the previous paragraph are small and, indeed, they are generally less than $1/N$ which is sometimes regarded as the criterion for effective selective neutrality. Yet the effect of selective differentials of these magnitudes may be notable over evolutionary time. $2r$ may be considered as an estimate of the proportion of electrophoretically detectable alleles which were successful because of the selective advantages directly conferred by the charge difference through which they are detected. The actual direction of change favoured by selection will vary between different geographical areas, different taxa and different epochs. r is, however, used as a magnitude, not a directional quantity, so that the conceptualization of $2r$ as the proportion of selected alleles will be heuristically useful as long as the selective 'pressure' continues (intermittently or with fluctuations in intensity) long enough for a few substitutions between species. It is assumed here that the relationship between s and r developed above is a reasonable indicator of the relationship of r and whatever average best measures the variation which occurs in the selective coefficient of an allele during the course of its substitution. Assume also that electrophoretically detectable variants have similar relative fitness distributions to other mutations. Given a substitution rate of about 4.65×10^{-9} per nucleotide per year (the average for synonymous substitution rates in 42 coding region genes surveyed by Li *et al.*, 1985; as augmented by Nei, 1987), a value of $r = 0.1$ ($2r = 0.2$) implies that, over one billion years, the number of selective substitutions within a genomic lineage is ≈ 0.93 ($= 4.65 \times 0.2$) times the number of nucleotide sites. Many of these would involve multiple changes at the same site. However, a very substantial fraction of nucleotide sites would have undergone selective substitution under currently observed values of r.

An alternative approach to estimating the proportion of 'selective' substitutions would be to estimate k in eqn (21) from observed values of r and evaluating the integral of the density function eqn (18) between whichever value of s is supposed to imply 'selection' and the maximum value of one-half. If r is 0.1, then the upper value (implying less selective substitution) of k from eqn (21) is 5.45. If r is 0.3, then the upper value for k is 4.00. For populations of effective size greater than a few individuals, the numerical effects on the definite integral of eqn (18) are overwhelmingly caused by the lower bound. If this is $1/N$, the estimated proportion of alleles with larger selective values is 1.00×10^{-7} for $r = 0.1$ and 2.42×10^{-5} for $r = 0.3$. If the lower bound is $0.1/N$, the estimated proportions of 'selected' alleles become much larger. The proportion for $r = 0.1$ is 0.42 and that for $r = 0.3$ is 0.60.

Values of r have implications for questions about the proportion of extant variation which is in the process of selective substitution. According to classical theory, the rate of substitution of neutral alleles is equal to the mutation rate μ. The rate for advantageous genes, under genic selection with s small, is $\approx 4Ns\mu$. If $2r$ can be regarded, as suggested above, as the fraction of selective substitutions, then $(1-2r)$ may be regarded as the fraction of neutral substitutions. The total (relative) time for this level of neutral substitution to be accomplished would be $(1-2r)\mu$ and the total (relative) time for $2r$ selective substitutions would be $(2r)(4Ns\mu)$. The ratio of the total times required for the substitution of advantageous alleles to the time required for neutral alleles is:

$$\frac{2r}{1-2r} 4Ns .$$

Noting that, under genic selection with constant N and s, $4Ns$ is estimated by $\ln(1/2+r) - \ln(1/2-r)$, this ratio can be represented as:

$$\frac{2r}{1-2r} (\ln(\tfrac{1}{2}+r) - \ln(\tfrac{1}{2}-r)) . \qquad \text{a22)}$$

For the typical values of r found here (say, 0.1), this ratio is about 0.1, suggesting that only a minor fraction of the variation which is observed at any given time would be in the process of selectively induced substitution. Recent studies (Gillespie, 1994) have suggested that the classical estimates of substitution rate may be significant overestimates for non-neutral variants. The 'TIM' case is the closest among those situations simulated by Gillespie (1994) to the standard genic selection model. The classical substitution rate is greater than the simulated rate for this model by a factor of ≈ 2.4. If the substitution rate for selectively advantageous variants is slowed to such an extent, then a greater proportion (albeit less than one-quarter) of extant variation may be expected to be in the process of selective fixation.

The analyses in this paper can be extended to cover other data for which two species may exhibit differences. In particular, the analyses are applicable to comparisons of DNA sequences. In this situation, the triplet codon is the unit of comparison and tests are carried out on the pattern of changes at different codons within a nominated sequence.

Other types of difference than charge changes could be tested. Sufficient data for this purpose are presently available for comparing the genes of rat and mouse (Wolfe & Sharp, 1993) and for investigating cytochrome b in many species. Data for comparing humans and chimpanzees may be investigated but may not yet be extensive enough to give accurate estimates of r. It will be interesting to compare the values of s generated from such sequence data with those derived from the methods of Sawyer & Hartl (1992). Should these two quite disparate approaches lead to similar conclusions regarding the size of selective coefficients, as appears possible from their respective applications to date, then we may begin to be confident of our estimates of the average values of s.

Acknowledgements

I would like to thank Drs O. Mayo and A. H. D. Brown and Professor J. S. F. Barker for commenting on the manuscript.

References

COLGAN, D. J. 1992. Interspecific isoenzymic substitution is not random. *Heredity*, **69**, 150–159.

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

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

GALE, J. S. 1991. *Theoretical Population Genetics*. Unwin-Hyman, London.

GILLESPIE, J. H. 1991. *The Causes of Molecular Evolution*. Oxford University Press, New York.

GILLESPIE, J. H. 1994. Substitution processes in molecular evolution. II. Exchangeable models from population genetics. *Evolution*, **48**, 1101–1113.

KIMURA, M. 1962. On the probability of fixation of mutant genes in a population. *Genetics*, **47**, 713–719.

KIMURA, M. 1983. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.

KIMURA, M. 1986. DNA and the neutral theory. *Phil. Trans. R. Soc. B*, **312**, 344–354.

KIMURA, M. 1990. Some models of neutral evolution, compensatory evolution and the shifting balance process. *Theor. Pop. Biol.*, **37**, 150–158.

LI, W.-H., WU, C.-I. AND LUO, C.-C. 1985. A new method for estimating synonymous and non-synonymous rates of substitution considering the relative likelihood of nucleotide and codon changes. *Mol. Biol. Evol*, **2**, 150–174.

MALÉCOT, G. 1952. Les processus stochastiques et la méthode des fonctions génératrices ou caractéristiques. *Extr. Publ l'ISUP*, **3**, 1–16.

MUKAI, T., CHIGUSA, S. AND YOSHIKAWA, I. 1965. The genetic structure of natural populations of *Drosophila melanogaster*. III. Dominance effect of spontaneous mutant polygenes controlling viability in heterogeneous genetic backgrounds. *Genetics*, **52**, 493–501.

NEI, M. 1987. *Molecular Evolutionary Genetics*. Columbia University Press, New York.

OHTA, T. 1973. Slightly deleterious mutant substitutions in evolution. *Nature*, **246**, 96–98.

OHTA, T. 1977. Extension to the neutral mutation drift hypothesis. In: Kimura, M (ed.) *Molecular Evolution and Polymorphism*, pp. 148–167. National Institute of Genetics, Mishima.

OHTA, T. 1992. The nearly neutral theory of molecular evolution. *Ann. Rev. Ecol. Syst.*, **23**, 263–286.

OLMSTED, J. M. H. 1961. *Advanced Calculus*. Appleton-Century-Crofts, New York.

SAWYER, S. A. AND HARTL, D. L. 1992. Population genetics of polymorphism and divergence. *Genetics*, **132**, 1161–1176.

WARD, R. L. AND SKIBINSKI, D. O. F. 1982. Inter-locus allozyme mobility correlations and species divergence. *Experientia*, **38**, 654–655.

WOLFE, K. H. AND SHARP, P. M. 1993. Mammalian gene evolution: nucleotide sequence divergence between mouse and rat. *J. Mol. Evol.*, **37**, 441–456.

Appendix 1

This appendix details various properties of density functions of the form $f(s) = c(e^{-4N(k-1)s} - e^{-4Nks})$. The constant c (which depends on the limits of integration, being the assumed range of values of s) is calculated as follows:

$$1 = c\int_0^{1/2} \frac{e^{4Ns}-1}{e^{4Nks}}\,ds$$

$$= c\left[\frac{-1}{4N(k-1)}e^{-4N(k-1)s} + \frac{1}{4Nk}e^{-4Nks}\right]_0^{1/2} \qquad \text{(A1)}$$

$$= c\frac{1}{4Nk(k-1)}\left[1-ke^{-2N(k-1)}+(k-1)e^{-2Nk}\right]$$

$$\Rightarrow \quad c \simeq 4Nk(k-1).$$

The mean of $f(s)$ is approximated:

$$\begin{aligned}
\mathrm{E}(s) &= c\int_0^{1/2} s(e^{-4N(k-1)s}-e^{-4Nks})ds \\
&= \frac{c}{16N^2k^2(k-1)^2}\left[-e^{-4N(k-1)s}k^2(4N(k-1)s+1)+e^{-4Nks}(k-1)^2(4Nks+1)\right]_0^{1/2} \\
&\simeq \frac{c(2k-1)}{16N^2k^2(k-1)^2} \\
&\simeq \frac{2k-1}{4Nk(k-1)}.
\end{aligned} \tag{A2}$$

The second moment of $f(s)$ is:

$$\begin{aligned}
\mathrm{E}(s)^2 &= c\int_0^{1/2} s^2(e^{-4N(k-1)s}-e^{-4Nks})ds \\
&= c\left[\frac{-s^2e^{-4N(k-1)s}}{4N(k-1)}-\frac{2(4N(k-1)s+1)e^{-4N(k-1)s}}{64N^3(k-1)^3}+\frac{s^2e^{-4Nks}}{4Nk}+\frac{2(4Nks+1)e^{-4Nks}}{64N^3k^3}\right]^{1/2}
\end{aligned} \tag{A3}$$

Ignoring terms in $e^{-2N(k-1)}$ and e^{-2Nk}, the variance of $f(s)$ is approximately:

$$\begin{aligned}
\sigma^2(s) &= \mathrm{E}(s)^2-(\mathrm{E}(s))^2 \\
&\simeq c\left(\frac{2}{64N^3(k-1)^3}-\frac{2}{64N^3k^3}\right)-\frac{(2k-1)^2}{16N^2k^2(k-1)^2} \\
&\simeq \frac{1}{16N^2k^2(k-1)^2}(2k^3-2(k-1)^3-(2k-1)^2) \\
&= \frac{2k^2-2k+1}{16N^2k^2(k-1)^2}.
\end{aligned} \tag{A4}$$

The total probability density for values of s below $1/N$ can be shown to be greater than $(1/N)f(1/N)$ for moderate values of k:

$$\begin{aligned}
&c\int_0^{1/N}(e^{-4N(k-1)s}-e^{-4Nks})ds-\frac{f(1/N)}{N} \\
&= \frac{c}{4N}\left(\left[\frac{-1}{k-1}e^{-4N(k-1)s}+\frac{1}{k}e^{-4Nks}\right]_0^{1/N}-4e^{-4(k-1)}+4e^{-4k}\right) \\
&= \frac{ce^{-4k}}{4Nk(k-1)}(k(e^{4k}-e^4)-(k-1)(e^{4k}-1)-4k(k-1)e^4+4k(k-1)).
\end{aligned} \tag{A5}$$

The right-hand side of eqn (A5) is greater than 0 when:

$$e^{4k}-e^4(4k^2-3k)+4k^2-3k-1>0. \tag{A6}$$

This is so for k above 1.16.