STATISTICAL ASPECTS OF THE NON-DARWINIAN THEORY

W. J. EWENS
UNIVERSITY OF TEXAS AT AUSTIN

1. Introduction

Quite apart from its possible biological relevance, the non-Darwinian theory of evolution currently under discussion is of considerable interest to statisticians. This is so, of course, because any mathematical formulation of the proposition that a considerable proportion of observed allelic substitutions have no selective significance and have occurred purely by chance, and, hence, any quantitative testing of this theory, must be conducted by statistical methods. The ramifications of the statistical testing which will be required to discuss the half dozen or so major supporting arguments for the theory can hardly be supposed yet to have been analyzed, even superficially; in particular, this is is true for those arguments relying on protein sequence data, amino acid frequencies, and the genetic code. It may be that novel forms of statistical tests will be required for these analyses. The main aim of this paper, while going in this direction, though rather restricted, is to devise a statistical test of the non-Darwinian theory based on a form of data currently being obtained in large volume by biologists, namely, the number and frequencies of different alleles at a locus provided by a sample of individuals from one generation of a population. A subsidiary aim is to show that quite simple statistical arguments can cast some doubt on the usefulness of one support for the non-Darwinian theory, namely, the support arising from the principle of substitutional loads. It will be convenient to consider this subsidiary aim first.

2. The substitutional load

Our aim is not to question the validity of the concept of substitutional load itself, but rather its usefulness as a support for non-Darwinian evolution. To do this we trace the main outline of the way the substitutional load is calculated.

Consider a diploid population of fixed size N and suppose that at a certain locus two alleles, A_1 and A_2, are possible. The three genotypes are supposed subject to differential selection and it is assumed that this selection acts entirely

On leave from the Department of Mathematics, LaTrobe University, Bundoora, 3038, Victoria, Australia.
Written under support of USPHS Grant GM-15769.

through differential preadult mortality; specifically, that the probabilities that newborn individuals survive to reach the age of sexual reproduction are 1 for A_1A_1, $1 - sh$ for A_1A_2 and $1 - s$ for A_2A_2. Here s is assumed small and positive, $0 \leqq h \leqq 1$. Suppose in any generation that the frequency of A_1 is x. Then a proportion $1 - s(1 - x)\{1 + x(2h - 1)\}$ of all offspring survive and, hence, each parent generation is required to produce $N[1 - s(1 - x)\{1 + x(2h - 1)\}]^{-1}$ offspring so that, after the preadult selection has occurred, exactly N offspring survive to reach sexual maturity. The number of "selective deaths" is thus

$$(1) \quad N[1 - s(1 - x)\{1 + x(2h - 1)\}]^{-1} - N \cong Ns(1 - x)\{1 + x(2h - 1)\} = N\ell,$$

say. The substitutional load L is conventionally defined as the sum of the values of ℓ, whereby x increases from $x_0(\approx 0)$ to $x_1(\approx 1)$, and is

$$(2) \quad L = \int_0^\infty \ell \, dt = \int_{x_0}^{x_1} \ell \frac{dt}{dx} dx = \int_{x_0}^{x_1} [s(1 - x)\{1 + x(1 - 2h)\}][sx(1 - x)\{1 - h + x(2h - 1)\}]^{-1} dx.$$

For $h = \frac{1}{2}$, $x_0 = 10^{-4}$, we have $L = 18.4$, while for other values of h, L is sometimes greater than, sometimes less than, this value. A "representative" value for L is conventionally taken as $L = 30$.

We now turn to the biological interpretation of L. If the replacement process of A_1 for A_2 requires T generations, then clearly from (1) and (2) $NT + NL$ total offspring are required during this process; in other words, on average each individual must produce $1 + L/T$ offspring to face the possible forces of preadult selection. Alternatively, we may say that the optimal genotype leaves an average of $1 + L/T$ offspring who reach sexual maturity and that *all* genotypes must produce this number of offspring.

The crux of the argument arises when many loci are considered simultaneously. If we assume the same fitnesses at all loci as those given above, and if the substitutions at the different loci start, on average, n generations apart, there will be T/n substitutions in progress at any one time. The mean number of offspring of the optimal genotype (that is, the genotype having the configuration "A_1A_1" at each locus) is

$$(3) \quad \left(1 + \frac{L}{T}\right)^{T/n} \cong \exp\left\{\frac{L}{n}\right\},$$

and using the value $L = 30$ and the estimate $n = \frac{1}{2}$ quoted as deriving from protein sequence data, this gives $e^{60} \cong 10^{26}$ offspring. In other words, the argument suggests that in order for the substitutions at all the loci to proceed at the required rate, each individual must leave 10^{26} offspring and that these survive differentially in such a way that N offspring eventually reach sexual maturity. The substitutional load argument centers on the impossibility that each individual can leave such a large number of offspring.

It is at this point that elementary statistical arguments suggest that this reasoning be reviewed. We first consider the probability that an individual chosen at random is of the optimal genotype. If $h = \frac{1}{2}$, $x_0 = 10^{-4}$, $x_1 = 1 - 10^{-4}$, $s = 0.01$, $n = \frac{1}{2}$, there are 7,360 substitutions in progress at any one time and the probability that any randomly chosen individual has the optimal genotype A_1A_1 at all 7,360 substituting loci is about $10^{-15,000}$. It seems of dubious value to base our considerations on such individuals. Rather, statistical considerations suggest that in any real population, no individual will have a genetic configuration which is too "extreme" (note that this in no way slows down the substitution rates), and, in particular, the mean number of surviving offspring required of the optimal genotype *we can expect to appear* in the population will not differ much from unity. This requirement can be calculated using the statistics of extreme values (of a sample of N individuals) and effectively assuming "independence" of loci. (This assumption, which, for example, ignores linkage between loci, will provide an upper bound to the offspring requirement, which should in fact be rather lower if linkage were taken into account.) For $N = 10^5$, $s = 0.01$, $h = \frac{1}{2}$, we find that this offspring requirement is about 1.5. In other words, substitutions at the required rate can occur if all individuals leave 1.5 offspring and these survive differentially according to their genotype (with the optimal *existing* genotype all surviving) in such a way that N offspring altogether survive to sexual maturity. When the effects of linkage, extensive linkage disequilibrium, and the possibility that considerable selection actually occurs through fertility difference, are all taken into account, the value 1.5 is probably reduced to about 1.2. There appears little difficulty for a population actually to achieve this value: that is to say, the statistical form of reasoning we have adopted suggests that there is little difficulty in ascribing the observed substitutions to selective forces with selective differentials of order one per cent. (Of course, this does not mean selection must be the responsible agency; there appears to be no reason, however, for us to say selection *cannot* be responsible.)

A further statistical point relating to the above argument is as follows. In a typical size population of, say, 10^6, no individual can leave more than 10^6 offspring who survive to maturity. The calculation that 10^{26} offspring of certain genotype do so survive must then result from inexact modelling; in this case, from the implicit assumption that mean numbers are "multiplicative over loci." In other words, not only have the load arguments supporting the non-Darwinian theory been carried out entirely with reference to individuals whose probability of occurrence is of order $10^{-15,000}$, they have also ascribed mean numbers of viable offspring to such (essentially nonexistent) individuals about 10^{20} times in excess of the maximum possible value.

3. The sampling theory of neutral alleles

A considerable amount of data is being obtained currently of the following form: a sample of n individuals ($2n$ genes) is taken from a population of un-

known size N. In the sample it is observed that k different alleles occur, with numbers $n_1, n_2, \cdots, n_k$ ($\sum n_i = 2n$). We now ask, can such data be used to test the hypothesis that the alleles are selectively neutral with respect to each other?

To do this, is it assumed that new alleles are formed in the following fashion: each gene will mutate with fixed (but unknown) probability u, to an allelic type not currently existing, nor previously existing, in the population. (Note that this assumption is inspired by protein sequence data, where our detailed knowledge of the sequence of amino acids determined by any gene makes this assumption reasonable. The assumption is perhaps less valid if our mode of differentiating alleles is by electrophoresis, since the theoretical effects of the nonidentification seemingly unavoidable in this procedure are not yet known. It is, therefore, possible that application of the following techniques to electrophoretically obtained data should be viewed with extreme caution.) More explicitly, we assume a multinomial mode of sampling the genes of a daughter generation from the genes of the parent generation. This implies that if we fix attention on some allele A_k, and suppose that i genes of this allele exist in any generation, then the probability $p_{i,j}$ that there exist j genes of this allele in the next generation is

$$p_{i,j} = \binom{2N}{j}\left\{\frac{i}{2N}(1-u)\right\}^j\left\{1-\frac{i}{2N}(1-u)\right\}^{2N-j}. \tag{4}$$

Note that we are assuming that no selective differences exist between alleles; thus, our aim is to develop distribution theory under the neutral hypothesis, with a view to subsequent testing of it using real data.

If it is supposed that sufficient time has elapsed for a stationary situation to be reached, then passing to the diffusion approximation to (4), standard theory (Ewens [2], Chapters 5 and 6) shows that the probability that, in any randomly chosen generation, there exists an allele in the population with frequency in the range $(p, p + \delta p)$ is

$$f(p)\,\delta p = \theta p^{-1}(1-p)^{\theta-1}\,\delta p, \tag{5}$$

where $\theta = 4Nu$. It will turn out that all of our subsequent theory can be carried out by using this "frequency spectrum" $f(p)$. In particular, we note that if $p_1, p_2, \cdots$ are the (unknown) frequencies of the various alleles present in any generation, and if $\phi(p)$ is any function of p that is $O(p)$ near $p = 0$, then

$$E\sum\phi(p_i) = \theta\int_0^1 \phi(p)p^{-1}(1-p)^{\theta-1}\,dp, \tag{6}$$

to a sufficiently close approximation. For example, if $\phi(p) = p$, use of (6) yields the trivial identity $E\sum p_i = 1$, while if $\phi(p) = p^2$, we find

$$E\sum p_i^2 = \frac{1}{1+\theta}. \tag{7}$$

Note that the left side in (7) is the probability that two genes drawn at random are of the same allelic type. It is, thus, a measure of genetic variability in the

population and the quantity $1/(1+\theta)$ is, consequently, of some interest to geneticists. We shall return later to the question of estimating this quantity from experimental data.

Now suppose a sample of $2n$ genes is drawn one by one from the population. We suppose $N \gg n$ so that binomial approximations are adequate. Then if we know the frequencies $p_1, p_2, \cdots$ of the various alleles in the population in the generation from which the sample was drawn, the probability that a previously unseen allele appears for the first time on the $(j+1)$th draw is

$$\sum_i (1-p_i)^j p_i. \tag{8}$$

Also the probability that the first j draws all yield the same allele is

$$\sum_i p_i^j. \tag{9}$$

Equation (6) shows that the unconditional probabilities to be attached to these events are

$$\theta \int_0^1 (1-p)^j p[p^{-1}(1-p)^{\theta-1}]\, dp = \frac{\theta}{\theta+j}, \tag{10}$$

and

$$\theta \int_0^1 p^j[p^{-1}(p-1)^{\theta-1}]\, dp = \theta(j-1)!\,\frac{\Gamma(\theta)}{\Gamma(\theta+j)}, \tag{11}$$

respectively. It follows from (11) that the probability that the first $(j+1)$ draws all yield one allelic type, given that this is true of the first j draws, is

$$\frac{\theta j!\Gamma(\theta)\Gamma(\theta+j)}{\theta(j-1)!\Gamma(\theta)\Gamma(\theta+j+1)} = \frac{j}{j+\theta}. \tag{12}$$

From this it follows that the probability that a new allele appears on the $(j+1)$th draw, given that the first j draws yield only one allelic type, is $\theta/(\theta+j)$. Note that this is identical to (10). We now argue more generally that the probability that a new allele appears on the $(j+1)$th draw is $\theta/(\theta+j)$ *whatever* the allelic composition was of the first j draws. A formal proof of this proposition has been given by Karlin and McGregor [6]. Intuitively, this (unusual) result can be seen as follows. If we label the new allele seen on the $(j+1)$th draw as A_k, then this gene is descended from some original mutant A_k allele. The stochastic behavior of the line of descendants of this mutant is independent of the allelic composition of the rest of the population, and so far as A_k is concerned the allelic forms of the non-A_k genes are just irrelevant labels. In particular, the probability that A_k appears for the first time at the $(j+1)$th draw is independent of this irrelevant labelling, that is, of the numbers and frequencies of the alleles which appeared on the first j draws.

It follows from this that if we write $\pi_{j,i}$ for the probability that the first j draws yield exactly i different alleles, we have

$$\begin{aligned} \pi_{j,1} &= (j-1)![(\theta+1)(\theta+2)\cdots(\theta+j-1)]^{-1}, \\ \pi_{j,j} &= \theta^{j-1}[(\theta+1)(\theta+2)\cdots(\theta+j-1)]^{-1}, \end{aligned} \tag{13}$$

and the recurrence relation

$$\pi_{j+1,i} = \pi_{j,i}\left\{\frac{j}{\theta + j}\right\} + \pi_{j,i-1}\frac{\theta}{\theta + j}. \tag{14}$$

We are particularly interested in the values $\pi_k = \pi_{2n,k}$. Solution of (14) shows that

$$\pi_k = \frac{\ell_k \theta^k}{L(\theta)}, \tag{15}$$

where

$$\begin{aligned} L(\theta) &= \theta(\theta + 1)\cdots(\theta + 2n - 1) \\ &= \ell_1\theta + \ell_2\theta^2 + \cdots + \ell_{2n}\theta^{2n} \end{aligned} \tag{16}$$

and the ℓ_k are moduli of Stirling numbers of the first kind. Equation (15) gives the probability distribution of the number k of different alleles in the sample. We have in particular

$$\begin{aligned} E(k) &= \frac{\theta}{\theta} + \frac{\theta}{\theta + 1} + \cdots + \frac{\theta}{\theta + 2n - 1}, \\ \mathrm{Var}(k) &= \frac{\theta}{\theta} + \frac{\theta}{\theta + 1} + \cdots + \frac{\theta}{\theta + 2n - 1} \\ &\quad - \left[\frac{\theta^2}{\theta^2} + \frac{\theta^2}{(\theta + 1)^2} + \cdots + \frac{\theta^2}{(\theta + 2n - 1)^2}\right], \end{aligned} \tag{17}$$

and, further, that for $2n$ large, k had an approximate normal distribution with this mean and this variance. Note also that the distribution (15) is complete.

We turn now to a more complex problem, namely, the distribution of the vector $\{k; n_1, \cdots, n_k\}$, where $n_1 \cdots n_k$ are the numbers of genes of the k alleles in the sample. We find (Ewens [3]) that this distribution is of the form

$$\frac{g(n_1, \cdots, n_k)\theta^k}{L(\theta)}, \tag{18}$$

where $g(n_1, \cdots, n_k)$ does not depend on θ. Comparison of (15) and (18) shows that k is sufficient for θ. Hence, any estimable function of θ is best estimated (in the sense of minimum variance unbiased) by some (unique) function of k. In particular, this is true of the (estimable) function $1/(1 + \theta)$, whose optimal estimator is

$$c(k) = \frac{\text{coeff } \theta^k \text{ in } \theta(\theta + 2)(\theta + 3)\cdots(\theta + 2n - 1)}{\text{coeff } \theta^k \text{ in } \theta(\theta + 1)(\theta + 2)\cdots(\theta + 2n - 1)}. \tag{19}$$

Curiously, because of the genetical interpretation of $1/(1 + \theta)$, (see the discussion following equation (7)), this function has traditionally been estimated by

$$\frac{n_1^2 + \cdots + n_k^2}{(2n)^2}. \tag{20}$$

This estimator has been found in Monte Carlo simulations to have very large variance (see, for example, Bodmer and Cavalli-Sforza [1]) and the present

theory indicates why this is the case, namely, that (20) uses precisely the inappropriate part of the vector $\{k; n_1, \cdots, n_k\}$. Further, Monte Carlo runs (Guess and Ewens [4]) suggest that the variance of $c(k)$ is only about 35 per cent that of (20). It is clear that some standard statistical theory is of some aid here in estimating an important genetical parameter.

Our main use, however, of the sufficiency of k for θ will be to provide a test for the hypothesis of selective neutrality, for under this theory the distribution of $n_1, \cdots, n_k$, given k, must be independent of θ and, hence, the same for all models of the form (4) (whatever the values of N and u are). It is found that this conditional distribution is [3]

$$f(n_1 \cdots n_k | k, 2n, \text{neutrality}) = \frac{(2n)!}{k! \ell_k n_1 n_2 \cdots n_k}. \tag{21}$$

More precisely, we assume the k alleles in the sample have been labelled in some conventional fashion $A_1 \cdots A_k$: equation (21) gives the probability that there are n_1 genes of the allele labelled A_1, $\cdots$, and n_k genes of the allele labelled A_k. Note that the fact that (21) sums to unity is established by an identity for Stirling numbers going back at least to Cauchy (see, for example, Jordan [5], p. 146, equation (5)).

The distribution (under the hypothesis of selective neutrality) of any test statistic can be found from (21). The actual choice of test statistic is not easy since the alternative hypothesis (that selection exists) does not seem sufficiently precise to yield an unambiguous statistic. Here we shall be content with using (more or less arbitrarily) the (information) statistic

$$B = -\sum x_i \log x_i, \tag{22}$$

where $x_i = n_i/2n$. The mean and variance of B can readily be calculated from (21) and by noting that marginal distributions from (21) are calculated almost immediately. Hence, if we write $E(B)$ and $\sigma(B)$ for the mean and standard deviation of B under the hypothesis of selective neutrality, given the appropriate value of k, it is possible for any set of data to evaluate

$$L = \frac{B - E(B)}{\sigma(B)}, \tag{23}$$

which under selective neutrality is a random variable having mean zero and variance one. Evaluation of L seems particularly useful when sets of interrelated data (that is, from each of a number of species in each of a number of locations) are available, since the patterns of the values for L for such data are often revealing.

If, on the other hand, it is desired to carry out a test of hypothesis, an approximate but reasonable procedure is to suppose $B/\log k$ has a beta distribution (whose parameters are known from $E(B)$ and $\sigma(B)$). A standard transformation yields a variable having an F distribution under the neutrality hypothesis. A FORTRAN program carrying out all the required computations is provided in [3].

TABLE I

EXAMPLE USING $n = 350$ AND $k = 4$

Sample	Allele frequencies				L	F	($d.f.$)
1	.35	.30	.20	.15	2.42	34.44	(3.5, 4.7)
2	.83	.11	.04	.02	0.02	1.02	(3.5, 4.7)
3	.99	.005	.0025	.0025	−1.74	0.07	(3.5, 4.7)

In Table I we indicate the sort of result obtained by an example. Suppose $n = 350$, $k = 4$, and consider three different sets of values of $x_1, \cdots, x_4$. The first sample yields significant evidence of selection (of some form) holding all alleles at high frequency, while the third sample yields evidence of selection favoring one allele. The second sample has almost a "perfect" set of "neutral" frequencies.

It is proper to conclude on a note of caution. The present lack of theoretical knowledge on the effect of nonidentification may be sufficiently strong to vitiate application of the above to electrophoretically derived data. (The above theory assumes total ability to differentiate different alleles.) Our model also ignores the effects of linkage, possible fluctuation in population size, and so forth. (On the other hand, it is conjectured that the distribution (21) applies for a wide range of "neutral" models, not just the model (4).) Finally, the test does not appear to be particularly powerful (in a statistical sense) and using it one may often maintain the hypothesis of neutrality when, in fact, mild selection does occur. Altogether, it appears that the test of hypothesis, and use of the index function L, may best serve as a cautiously used adjunct to other and independent methods of testing the non-Darwinian theory.

REFERENCES

[1] W. F. BODMER and L. L. CAVALLI-SFORZA, "Variation in fitness and molecular evolution," *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley and Los Angeles, University of California Press, 1972, Vol. 5, pp. 255–275.

[2] W. J. EWENS, *Population Genetics*, London, Methuen, 1969.

[3] ———, "The sampling theory of selectively neutral alleles," *Theor. Pop. Biol.*, Vol. 3 (1972), pp. 87-112.

[4] H. GUESS and W. J. EWENS, "Theoretical and simulation results relating to the neutral allele theory," *Theor. Pop. Biol.*, to appear.

[5] C. JORDAN, *Calculus of Finite Differences*, New York, Chelsea, 1950.

[6] S. KARLIN and J. L. MCGREGOR, *Addendum* to "The sampling theory of selectively neutral alleles" [3], *Theor. Pop. Biol.*, Vol. 3 (1972), pp. 113–116.