An Improved Method for Estimating the Rate of Fixation of Favorable Mutations Based on DNA Polymorphism Data

Wolfgang Stephan
Department of Zoology, University of Maryland

The hitchhiking effect has frequently been used as a method to detect the action of past directional selection at the DNA sequence level in natural populations of several Drosophila species (Aguade et al. [1989]; Stephan and Langley [1989]; Miyashita [1990]; Begun and Aquadro [1991]; Berry et al. [1991]; references to later papers can be found in Golding's [1994] volume) and in natural isolates of Escherichia coli (Guttman and Dykhuizen 1994). Advantageous mutations going to fixation leave footprints of reduced DNA sequence variation which can most easily be identified in genomic regions of very low recombination. Mathematical models of the hitchhiking effect have explored the observed correlation between DNA sequence variation and levels of crossing-over (Maynard Smith and Haigh 1974; Kaplan et al. 1989; Stephan et al. 1992). Following this work, a method for estimating the rate of fixation of favorable mutations from DNA polymorphism data has been proposed (Wiehe and Stephan 1993). A formula describing the functional relationship between expected nucleotide diversity, \( \pi \), and recombination rate \( r \) per nucleotide site, has been derived for an equilibrium hitchhiking model with one selected locus:

\[
\pi = H_{\text{neu}} \frac{\rho}{\rho + kav}.
\]  

In a diploid model, \( H_{\text{neu}} = 4N\mu \) is the neutral level of equilibrium nucleotide diversity, with \( N \) being the effective population size and \( \mu \) the neutral mutation rate; \( \alpha = 2Ns \), where \( s \) is the selection coefficient of the favorable mutation, \( \nu \) is the rate of the selected substitutions per nucleotide site, and \( k \) is a constant which is approximately equal to 0.075.

The value of the parameter \( av \) can be estimated by fitting formula (1) to DNA polymorphism data collected from a number of different loci with varying levels of crossing-over. For carrying out the fitting procedure, a Lineweaver-Burk transformation of model (1) was proposed. This procedure is generally used in biochemistry to fit models of the Michaelis-Menten type such as for the Lineweaver-Burk transformation. The values of these parameters can then be estimated using the geometric mean (GM) method, a Model II regression procedure, which takes into account that both random variables \( \pi \) and \( \rho \) are subject to error (Sokal and Rohlf 1981, chap. 14.13).

This procedure, however, generates two problems. First, Colquhoun (1969) has demonstrated that the estimates of the parameters \( V \) and \( K \) of the Michaelis-Menten hyperbola \( Y = \nu x/(K + x) \) have large variances when the Lineweaver-Burk transformation is used. He generated normally distributed observations by computer simulations under various specified standard deviations. In all cases, the Lineweaver-Burk transformation yielded the largest variances of the estimates of \( V \) and \( K \) among the three possible transformations (see below) that produce linear plots from the hyperbola \( Y = Vx/(K + x) \). Second, since the parameters of the Lineweaver-Burk transformed model (1) are combinations of \( av \) and \( H_{\text{neu}} \), it was not possible to provide confidence limits for the estimate of \( av \). This note presents an improved method for estimating \( av \) which circumvents these problems associated with the Lineweaver-Burk transformation.

Among the three possible transformations which produce linear regression equations from formula (1) (i.e., \( 1/\pi \) against \( 1/\rho \), \( \pi \) against \( \pi/\rho \), and \( \rho/\pi \) against \( \rho \)), we use

\[
\pi = H_{\text{neu}} - kav \frac{\rho}{\rho}.
\]  

This model never gives huge errors in the estimates of the parameters, but it may lead to a bias toward low
values if the experimental errors are large (Colquhoun 1969). Another advantage of this transformation is that the parameters \( H_{\text{rej}} \) and \( \alpha v \) are no longer combinations of each other, as for the Lineweaver-Burk transformation and also for the plot of \( p/\pi \) against \( p \). Therefore, the GM method applied to model (2) can provide confidence limits for \( \alpha v \), which also solves the second problem.

In the following, we will discuss the use of this procedure by analyzing the DNA polymorphism data of E. C. Kindahl and C. F. Aquadro (personal communication; see also Aquadro et al. 1994). These authors presented a four-cutter survey from a single \textit{Drosophila melanogaster} population for 15 third-chromosome loci (summarized in table 1). The study by Kindahl and Aquadro was conducted in a systematic way after a first analysis by Begun and Aquadro (1992) had shown a positive correlation between levels of DNA sequence variation and crossing-over for gene regions scattered throughout the \textit{D. melanogaster} genome and population samples from many different geographic areas. However, the reader should be advised that it is at present unclear to what extent the observed correlation between levels of genetic diversity and recombination rate is due to hitchhiking associated with directional selection (see below). Charlesworth et al. (1993) have demonstrated that background selection against deleterious alleles may have a similar effect on linked neutral variation as directional selection and may also explain this correlation (for further references, see Charlesworth [1994] and Hudson [1994]).

Several assumptions have been made in the derivation of formula (1). First, the levels of selective constraints between different gene regions are comparable; that is, \( H_{\text{rej}} \) is approximately equal for all loci included in the analysis. Second, since equation (1) describes an equilibrium hitchhiking scenario, it does not account for individual evolutionary events, such as recent selective sweeps. Third, the effects of other selective forces on DNA polymorphism such as background selection and balancing selection are negligible. The new data set by Kindahl and Aquadro allows us, at least to some extent, to evaluate whether the three criteria for fitting model (2) to data are met. To obtain reliable estimates of \( \alpha v \), ideally one would want to include only those loci in the analysis which meet these criteria. In the absence of divergence measurements, one may use the proportion of coding DNA (of a gene region) as an indicator of selective constraint to satisfy criterion 1. In table 1, the most extreme cases are the \textit{Ubx} region which contains no coding portion, and the \textit{Tl} region which is nearly 100% coding. The \textit{Sod} locus does not seem to meet the second criterion, in that there is evidence for a very recent selective event in this gene region (Hudson et al. 1994). The third criterion addresses forms of natural selection other than directional selection. In order to distinguish directional selection from forces such as balancing selection or background selection, discrimination functions (e.g., the Tajima [1989] statistic) could be used.

Unfortunately, none of the currently available discrimination functions is very powerful. For the data set in table 1, the loci which may not meet criterion 3 seem to be \textit{Hsp26} and \textit{Antp}. Both produced large positive values of the Tajima statistic \( D \) (> 1), whereas loci dominated by hitchhiking associated with directional selection are expected to show negative \( D \) values. But it should be noted that in both cases \( D \) is not statistically different from zero. Similarly, the estimates of \( D \) for the remaining 10 loci (\textit{Lsp}, \textit{Pc}, \textit{Gld}, \textit{tra}, \textit{fz}, \textit{Mle2}, \textit{ry}, \textit{Rh3}, \textit{Est6}, \textit{E(spl)}) are not significantly different from zero; their average value, however, is negative, \(-0.31\). This may indicate that directional selection is operating on at least some of these 10 loci. However, it is very likely that other selective forces such as background selection also work on these genes because the data do not show a statistically significant correlation between Tajima’s \( D \) and recombination rate, as one might expect for the equilibrium hitchhiking model with one selected locus that produced equation (1) (Braverman et al. 1995).

I applied the GM method to model (2) in two ways. First, for all 15 loci of Kindahl and Aquadro’s data set I obtained the following estimates: \( H_{\text{rej}} \) = 0.0079, \( \alpha v \) = 0.128, and the 95% confidence limits for \( \alpha v \) are \( L_1 = 0.048 \) and \( L_2 = 0.209 \). The estimates of \( \alpha v \) and \( L_1 \) and \( L_2 \) are given in units of 1 \text{ACE} (= adjusted coefficient of exchange). To obtain \( \alpha v \) per nucleotide site, we have to rescale the \( X \)-axis as described elsewhere (Wiehe and Stephan 1993). Using E. C. Kindahl and C. F. Aquadro’s (personal communication) scaling factor \( 3.62 \times 10^{-2} \), I found \( \alpha v \) = \( 4.63 \times 10^{-6} \), and for the 95% confidence limits \( L_1 = 1.74 \times 10^{-8} \) and \( L_2 = 7.57 \times 10^{-8} \). This

<table>
<thead>
<tr>
<th>Gene Region</th>
<th>ACE</th>
<th>( \pi )</th>
<th>Tajima’s ( D )</th>
<th>Coding Portion (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lsp</td>
<td>0.000</td>
<td>0.0001</td>
<td>(-0.79)</td>
<td>76</td>
</tr>
<tr>
<td>Pc</td>
<td>0.002</td>
<td>0.0010</td>
<td>0.09</td>
<td>34</td>
</tr>
<tr>
<td>Antp</td>
<td>0.004</td>
<td>0.0040</td>
<td>1.32</td>
<td>25-40</td>
</tr>
<tr>
<td>Gld</td>
<td>0.004</td>
<td>0.0022</td>
<td>(-0.14)</td>
<td>75</td>
</tr>
<tr>
<td>Ubx</td>
<td>0.011</td>
<td>0.0084</td>
<td>0.88</td>
<td>0</td>
</tr>
<tr>
<td>tra</td>
<td>0.015</td>
<td>0.0024</td>
<td>(-0.86)</td>
<td>25</td>
</tr>
<tr>
<td>fz</td>
<td>0.016</td>
<td>0.0043</td>
<td>(-0.40)</td>
<td>37</td>
</tr>
<tr>
<td>Mle2</td>
<td>0.020</td>
<td>0.0044</td>
<td>(-0.50)</td>
<td>12</td>
</tr>
<tr>
<td>ry</td>
<td>0.021</td>
<td>0.0048</td>
<td>0.63</td>
<td>74</td>
</tr>
<tr>
<td>Sod</td>
<td>0.026</td>
<td>0.0033</td>
<td>(-1.05)</td>
<td>29</td>
</tr>
<tr>
<td>Tl</td>
<td>0.027</td>
<td>0.0020</td>
<td>(-1.39)</td>
<td>94</td>
</tr>
<tr>
<td>Rh3</td>
<td>0.033</td>
<td>0.0061</td>
<td>(-0.38)</td>
<td>52</td>
</tr>
<tr>
<td>Est6</td>
<td>0.044</td>
<td>0.0070</td>
<td>0.07</td>
<td>80</td>
</tr>
<tr>
<td>E(spl)</td>
<td>0.051</td>
<td>0.0070</td>
<td>(-0.81)</td>
<td>6</td>
</tr>
<tr>
<td>Hsp26</td>
<td>0.062</td>
<td>0.0102</td>
<td>1.20</td>
<td>22</td>
</tr>
</tbody>
</table>

Note.—ACE is the adjusted coefficient of exchange. The value in the last column is the coding portion of a gene region. The data are from E. C. Kindahl and C. F. Aquadro (personal communication).
A genealogical process an outlier. Walthour and Schaeffer (1994) have argued that the differences between the observed levels of nucleotide diversity and the corresponding fitted values. One locus (tra) lies 2.5 standard deviations from the mean of the residuals and may therefore be considered an outlier. Walthour and Schaeffer (1994) have argued a recent selective sweep at or near the locus based on their sequence data. The tra locus may thus not meet criterion 2. I redid the analysis without tra and found the following estimates: \( H_{\text{neu}} = 0.0085, \alpha v = 0.184, L_1 = 0.117, \) and \( L_2 = 0.249. \) The data and equation (1) with these parameters are shown in figure 1 (the X-axis is scaled in units of ACE). On a per-nucleotide-site basis, I found \( \alpha v = 6.66 \times 10^{-8}, \) and for the 95% confidence limits \( L_1 = 4.24 \times 10^{-8} \) and \( L_2 = 9.01 \times 10^{-8}. \) This estimate of \( \alpha v \) is higher, but not significantly different from that for all 15 loci (see above).

If one assumes that most mutations are deleterious and distributed more or less randomly in the genome, then a reduction of neutral variation due to background selection will be occurring throughout the genome and will be stronger in regions of low recombination rates. From this viewpoint, background selection is the null hypothesis rather than neutrality. It is thus important to develop models that incorporate the joint effects of background and directional selection on neutral variation. In large populations, the effect of background selection is to reduce the effective population size, but to otherwise preserve the neutral genealogical process (R. R. Hudson, personal communication). The simplest way to incorporate background selection into model (1) is therefore to replace \( H_{\text{neu}} \) by \( H_{\text{bs}}(\rho), \) where \( H_{\text{bs}}(\rho) \) is a function of the recombination rate \( \rho. \) This model of directional and background selection contains several parameters such as the genomic mutation rate to deleterious alleles, \( U, \) as well as \( \alpha v \) which have to be estimated from the data by curve fitting. However, the estimation of these parameters requires statistical techniques that are beyond the scope of this note because the ensuing regression model is inherently nonlinear. In the present analysis, by neglecting the effects of background selection, we have overestimated the rate of fixation of favorable mutations.

Acknowledgments

I thank C. F. Aquadro, R. R. Hudson, and a reviewer for valuable comments on a previous version of the manuscript. Furthermore, I am grateful to C. F. Aquadro, B. Charlesworth, D. Charlesworth, R. R. Hudson, N. L. Kaplan, E. C. Kindahl, and M. T. Morgan for sharing their results prior to publication. This research was supported by National Science Foundation grant DEB-9407226.

LITERATURE CITED


CHARLES F. AQUADRO, reviewing editor

Received January 30, 1995

Accepted April 6, 1995