DOI: 10.1534/genetics.104.027052

Letter to the Editor

Maximum Likelihood *vs.* Minimum Distance: Searching for Hills in the Plain

Aurora García-Dorado[1] and Araceli Gallego

Departamento de Genética, Facultad de Biología, Universidad Complutense de Madrid, 28040, Madrid, Spain

Manuscript received February 2, 2004
Accepted for publication February 13, 2004

COMPARING maximum likelihood (ML; see KEIGHTLEY 1998) and minimum distance (MD; see GARCÍA-DORADO 1997) methods to infer the properties of deleterious spontaneous mutation from simulated mutation accumulation data, GARCÍA-DORADO and GALLEGO (2003) found that ML estimates of the deleterious mutation rate (U) had higher mean square errors (MSE) due to occasional ML outlier estimates. Thus, relative $\mathrm{MSE}^{1/2}$ for ML U estimates was about six times larger than that for MD U estimates obtained under comparable conditions, although the MSEs for the estimates of the average mutational effect were very similar. In his *Letter to the Editor*, KEIGHTLEY (2004, p. 551) claims that this result is due to the fact that MD nonconvergence is "declared if the profile of distance . . . changes nonsignificantly over a range of three times the parameter value" and that the comparison excludes "all MD replicates that fail to converge and any ML replicates for which the ML U estimate exceeds 50" and proposes that this difference in the exclusion criteria allows outliers for ML estimates of U (U_{ML}) but not for MD ones (U_{MD}).

As explained in our original article (GARCÍA-DORADO and GALLEGO 2003, p. 810), the Fortran MD program searches the minimum distance in a grid for (U, P_a, α) specified by the user (where P_a is the probability that the mutational effect is advantageous, and α is the shape parameter of the gamma distribution assumed for the deleterious mutational effect) to obtain a profile of the distance, minimized with respect to α and P_a, against U. For these three parameters, we started with intervals that were iteratively widened when required. In particular, the MD estimate for U (U_{MD}) was initially searched between $U/20$ and $10U$, but when, occasionally, the minimum distance corresponded to U values close to (or at) the edge of this interval, the grid was moved accordingly until a minimum for the distance was graphically identified (corresponding to U_{MD}) or the distance profile became flat.

Because the theoretical distribution used to compute the distance was empirically approximated through the simulation of 10^4 mutation-accumulation line means, distances have a random component. To deal with this component objectively, after a minimum was identified for the distance, we used regression to test the significance of the increase in distance within an interval ($U_{\mathrm{MD}} < U \leq U_{\mathrm{up}}$). This regression interval should be inferred from inspection of a previous profile obtained using a different seed for the generation of random variables. For our particular sets of parameters, it turned out that ($U_{\mathrm{MD}} < U \leq 3U_{\mathrm{MD}}$) was an appropriate interval, as it contains a relevant proportion of the observed increase in distance. This provides a non-ad hoc criterion to determine the interval for each new data set and avoids repeating each profile with a new random seed to test the significance of the minimum. Replicates with nonsignificant slopes were classified as providing no global MD estimate. Keightley notes that 15 of 62 replicates were classified as giving no global estimate using MD, while only 6 of 60 fell into this category using ML, and he suggests that outliers might have been removed from the MD results because of the above significance test.

To assess Keightley's concern, we reexamined those cases that did not give a global MD estimate. Twelve of 15 gave profiles where the distance decreased with increasing U until the profile, represented at a graphical precision high enough as to reveal the distance's random dispersion, became flat (*i.e.*, it could not be distinguished from horizontal at either a statistical or a visual level). On the contrary, the 6 excluded ML profiles did not become flat, but the likelihood monotonically increased with U at least up to $U = 50$. The remaining 3 of 15 MD excluded replicates had a graphical minimum, but it was not significant. However, their exclusion scarcely affected the results as they were not outliers; if we assume that these minima reflect valid estimates for U and include them in our study, the average $\mathrm{MSE}^{1/2}$ for U_{MD} increases by only 5% and remains at 0.28 times the $\mathrm{MSE}^{1/2}$ for U_{ML}.

Thus, ML estimates were searched until the likelihood showed a maximum or while it increased with U, up to

[1] *Corresponding author:* Departamento de Genética, Facultad de Biología, Universidad Complutense de Madrid, 28040, Madrid, Spain. E-mail: augardo@bio.ucm.es

a boundary of $U = 50$. MD estimates were searched until the distance showed a minimum or while it decreased with U, the search being stopped when the distance profile became flat. Setting aside the significance test, the criteria for ML and MD can be considered similar, excluding profiles that either become flat or reach a boundary $U = 50$. Despite this, the relative MSE$^{1/2}$ of U_{ML} was 3.72 times larger than that of U_{MD} for the cases quoted by Keightley. Note also that using a more flexible alternative (CS-MD; see García-Dorado and Gallego 2003), the MSE$^{1/2}$ for U_{ML} is 5.95 times larger than that for U_{MD} and just 8 of 62 replicates gave no MD global estimate.

Superiority for MD can be expected from its robustness properties against departures from the model (Parr and Schucany 1980; Woodward *et al.* 1984). Nevertheless, the smaller mean square errors for MD estimates were unexpected, because they correspond to cases where the model used in the simulation was the same model assumed in our ML and MD analyses. We agree that this is likely to be due to the algorithm used to locate the maximum, and we proposed this explanation in our article (García-Dorado and Gallego 2003, see p. 816, second column, first paragraph). However, it is not due to exploring different parameter ranges or to the occasional rejection of nonsignificant minima. Rather, the key difference was that ML identified U_{ML} outlier estimates in virtually flat profiles (Keightley 1998; García-Dorado and Gallego 2003). Such profiles occur when, for increasing U values, the remaining parameter can be adjusted as to continue accounting for the data with virtually the same likelihood. This means that those data do not contain enough information to estimate both U and the remaining parameter. Our implementation of the MD method has the practical advantage that it discriminates against such data, as it does not detect minima in virtually flat profiles: if there are any underlying outliers, they correspond to distance minima shallow enough as to remain hidden within profile regions that appear flat, due to the random component of distance. Therefore, it seems not worth removing this tiny random component by numerically computing the values of the theoretical probability distribution. This would allow using MD to look for hills in the plain, but it would be unlikely to improve the estimates from data. As we suggested previously (García-Dorado and Gallego 2003, see p. 817, last paragraph), when using ML "some criterion should be developed to indicate whether the ML surface is too flat, in which case the parameter estimates should be ignored." This criterion might be based, for example, on the wideness of the estimate's support limits.

We are grateful to C. López-Fanjul and to S. Otto for helpful comments.

LITERATURE CITED

García-Dorado, A., 1997 The rate and effects distribution of viability mutation in Drosophila: minimum distance estimation. Evolution **51:** 1130–1139.

García-Dorado, A, and A. Gallego, 2003 Comparing analysis methods for mutation-accumulation data: a simulation study. Genetics **164:** 807–819.

Keightley, P. D., 1998 Inference of genome-wide mutation rates and distributions of mutation effects for fitness traits: a simulation study. Genetics **150:** 1283–1293.

Keightley, P. D., 2004 Comparing analysis methods for mutation-accumulation data. Genetics **167:** 551–553.

Parr, W. C., and W. R. Schucany, 1980 Minimum distance and robust estimation. J. Am. Stat. Assoc. **75:** 616–624.

Woodward, W. A., W. C. Parr, W. R. Schucany and H. Lindsley, 1984 A comparison of minimum distance and maximum likelihood estimation of a mixture proportion. J. Am. Stat. Assoc. **79:** 590–598.

Communicating editor: S. P. Otto