tests. We hope to address this problem elsewhere.

ACKNOWLEDGMENTS

We thank D. Swofford for open, frank discussions on this topic and for permission to publish results based on PAUP* 4d46 and 4d47. P. Cranston and W. Ballard and his colleagues at the Field Museum provided many helpful comments on an earlier draft of this manuscript.

REFERENCES


Received 15 June 1996; accepted 4 October 1996

Associate Editor: David Cannatella


Statistical Issues Arising in the Analysis of DNA–DNA Hybridization Data

RUDY GUERRA1 AND TERENCE P. SPEED2

1Department of Statistical Science, Southern Methodist University, Dallas, Texas 75275, USA; E-mail: guerra@pearson.stat.smu.edu

2Department of Statistics, University of California, Berkeley, California 94720, USA

DNA–DNA hybridization is a technique that has a long and significant history in the study of molecular systematics. Although many methodological problems have been identified and corrected since its introduction to systematics, few authors have critically examined the statistical issues impacting the design, analysis, and conclusions of DNA–DNA hybridization experiments. In this paper, we used the grebe data set from Ahlquist et al. (1987) to illustrate some of these statistical problems (for a comprehensive overview of the main biological and experimental issues, see the Journal of Molecular Evolution, 1990, 30[3]).

Ahlquist et al. (1987) used DNA–DNA hybridization to measure the divergence between the single-copy nuclear DNA sequences of Aechmophorus occidentalis and A. clarkii, the Western and Clark’s grebes, respectively. Their aims were twofold: (1) to determine the taxonomic status of the Western and Clark’s grebes and (2) to assess the resolving power of the DNA hybridization technique. They concluded that the “DNA hybridization technique is sensitive to differences in sequence complementarity between closely relates species” (Ahlquist et al., 1987:4) and that the mean ΔT50H between the Western and Clark’s grebes is 0.57°C. Bledsoe and Sheldon
(1989:101) discussed the same grebe data in a study of the metric properties of $\Delta T$ values for DNA–DNA hybridization:

When comparisons were made between $A. \text{occidentalis}$ and $A. \text{clarkii}$ (sibling species), 69% of the heterospecific comparisons had a statistically significantly lower mean mode than the homoduplexes. . . . The limit of resolution in the Aechmophorus study is the species level. Below that, most individuals act as a single unit.

In addition, they reported a mean modal difference of 0.73°C (mean of 16 values; Bledsoe and Sheldon, 1989: table 3, column 6) between the two grebes. If the negative values from the same column are excluded, the mean modal difference between the two grebes is 0.85°C.

In this paper, we show that the conclusions of Ahlquist et al. (1987) and Bledsoe and Sheldon (1989) are in fact doubtful because of the experimental design in the grebe study. The genetic distances required to address the aims of these studies cannot be estimated without certain assumptions for which supporting information is unavailable. In addition, the grebe data set is appropriate for showing that estimated distance parameters from DNA–DNA hybridization data should account for known sources of variation and covariation, otherwise the results may be ambiguous.

**THE DATA AND THEIR STRUCTURE**

Sibley and Ahlquist (1981, 1984, 1991) were able to mass-produce DNA hybrids because of an automated thermal elution device, the DNAnalyzer. This particular apparatus is capable of accommodating 25 hybridizations, each in 1 of 25 linearly arranged hydroxyapatite columns along the DNAnalyzer. In most experiments that were run on the DNAnalyzer, replicate hybrids were systematically grouped along the columns. (A detailed account of these methods was provided by Sibley and Ahlquist, 1981.)

Using the notation T/D to indicate that a tracer from species T has been hybridized with a driver from species D, we define three types of hybrids or duplexes: (1) homoduplex $A_i/A_i$, where the tracer is individual $i$ from species A and the driver is also individual $i$ from species A; (2) intraspecific heteroduplex $A_i/A_j$, where the tracer is individual $i$ from species A and the driver is individual $j$ from species A; and (3) interspecific heteroduplex $A_i/B_j$, where the tracer is individual $i$ from species A and the driver is individual $j$ from species B.

The experimental design of the grebe study, described by Bledsoe and Sheldon (1989), is shown in Table 1. From this design, we can discern the structure of the data. To focus on the main ideas, we have not included the individual identifications in the T/D notation in Table 1. Eight experiments each accommodated up to 25 hybrids in separate tubes. Across experiments certain types of hybrids were placed in the same tube positions. Tubes 1–5 ran homoduplexes. Tubes 6–15 ran intraspecific heteroduplexes, with 6–10 sharing the same driver and 11–15 sharing the same driver but different from that used in tubes 6–10. Tubes 16–25 ran interspecific hetero-
oduplexes, with 16–20 sharing the same driver and 21–25 sharing the same driver but different from that used in tubes 16–20. All heteroduplexes in a given experiment shared the tracer used in tubes 1–5. For any given hybridization, the raw data take the form of radioactive counts. In almost all cases, the counts were used to construct dissociation curves, which in turn were used to estimate genetic distances.

As a result of the experimental design, individual hybrids do not necessarily give independent observations. For example, the 25 hybrids in the first row of Table 1 are in the same experiment, and they all share individual 3 from species \textit{A. occidentalis} (O3) as a tracer. This design gives rise to correlations in the data whenever the experiment or tracer effects are considered as random "error" components that contribute to the observed variation among dissociation curves and subsequently in estimated distances (Felsenstein, 1987a, 1987b).

Well-known distance measures (e.g., \(\Delta T_{\text{mode}}\) and \(\Delta T_{50}\)) were discussed by Bledsoe and Sheldon (1989). In this paper, we make use of \(\Delta T_{\text{mode}}\) and a new measure, SM, proposed by Guerra (1992). Hybridization frequency curves frequently display an exponential decay at the high end of the temperature gradient; SM compares the rates of decay between two curves. As such, it reflects more than a difference of central tendency summaries of curves. Among other things, Guerra (1992) found that SM is well correlated with \(\Delta T_{\text{mode}}\) and is less variable and better distinguishes closely related species than does either \(\Delta T_{\text{mode}}\) or \(\Delta T_{50}\). In the following discussion, we also refer to average homoduplex curves, which means that several homoduplex curves from the same experiment have been averaged to yield a single homoduplex curve, as suggested by Guerra (1992).

\textbf{VARIATION AND COVARIATION}

Individual hybrids do not necessarily give statistically independent observations because of the experimental design. Such nonindependence can lead to optimistic estimates of standard deviations. Consider the 10 observed modes from the interspecific heteroduplexes formed in experiment 393 (Table 1). One expression for the variance of their arithmetic mean is \(\sigma^2/10\), where \(\sigma^2\) is an assumed common error variance. However, suppose that the true model for these data is

\[
\text{observed mode} = \text{true mode} + \text{experiment effect} + \text{residual error},
\]

where the experiment effects have variance \(\sigma^2_E\), the residual error has variance \(\sigma^2_R\), and these two sets of effects are uncorrelated. With this model, \(\sigma^2 = \sigma^2_E + \sigma^2_R\). Now the average of the 10 observed modes can be written

\[
\text{average mode} = \text{true mode} + \text{experiment effect} + \text{average residual error},
\]

and this average has variance \(\sigma^2_E + (\sigma^2_R/10)\), not \((\sigma^2_E + \sigma^2_R)/10\). Thus, it is possible to underestimate the variance of such arithmetic means. Accounting for experiment-to-experiment variation can reduce apparent asymmetry in matrices showing pairwise genetic distances for a given group of species.

\textbf{Experiment Effects}

A basic technique in searching for potential sources of variation is the examination of residuals. Each SM observation involves one of four types of heteroduplex hybrids, O/O, C/C, O/C, or C/O. A residual (observation minus average) is computed for each observation by comparing the observation to the average SM value of the hybrid type to which it belongs. For example, the SM observation in experiment 393, tube position 10 has the average over all eight experiments of the intraspecific heteroduplex O/O observations subtracted from it. Originally, \(8 \times 20 = 160\) heteroduplex hybrids were generated, but because of various experimental problems and obvious gross outliers, 12 were deter-
FIGURE 1. Boxplots of SM residuals grouped according to experiment. The top and bottom of the rectangle are the upper and lower quartiles, respectively, of the residuals. The median of the residuals is indicated by the horizontal line segment within the rectangle. Vertical lines extend from the ends of the box to the tails of the distribution.

mined inadequate for analysis. In this way, we get 148 residuals.

Examination of the residuals reveals that the seven largest are from experiment 674. Moreover, of the 20 residuals from this experiment, 18 are positive. Intuitively, we expect about half the residuals to be positive and half negative. Taking the residuals from experiment 674 to be binomial observations with probability 0.5 of being positive, the chance of observing 18 or more positive residuals out of 20 is about 0.0002. Thus, experiment 674 is contributing some systematic effect, which shows up in both intraspecific and interspecific heteroduplexes. If this were not the case, then it would be difficult to associate the observed effect with the experiment.

We can identify all residuals with their respective experiments and ask, using boxplots (Fig. 1), if other experiments are contributing systematic effects. On average each experiment contributes 18 residuals to its respective boxplot. Figure 1 shows that experiment 553 is also contributing a systematic effect to its hybridizations; all of the residuals are negative. Thus, the effects from experiments 553 and 674 are in opposite directions, with the observations from experiment 553 consistently underestimating the parameters they estimate, whereas those from experiment 674 tend to overestimate. Also, the interquartile ranges of the two distributions, as indicated by the lengths of the boxes, do not overlap. Both of these experiments estimate the A. occidentalis intraspecific heteroduplex distance, denoted $d_\infty$. These two observations suggest that if two separate estimates of $d_\infty$ were obtained, one from experiment 553 and the other from experiment 674, we could expect them to be "significantly" different. In short, Figure 1 provides evidence for the existence of substantial experiment-to-experiment variation in the SM observations. The same conclusion is reached using other distance measures.

A more striking example of experiment-to-experiment variation is illustrated in Figure 2. Each curve in these figures is an average homoduplex profile from one of the eight experiments. For example, the solid curve in Figure 2a is an average of the five homoduplex curves found in experiment 393. Homoduplexes from A. occidentalis (Fig. 2a) and A. clarkii (Fig. 2b) are each provided by four experiments. Figures 2a and 2b show that the A. occidentalis and A. clarkii sets of curves are quite different in character. The A. clarkii curves are more sharply peaked and, with the exception of curve 555, are less variable than the A. occidentalis curves. Homoduplex curves are expected to be very stable, but these data indicate otherwise. Figure 2c illustrates the range of modes associated with all eight curves, about 85-88.5°C. This 3.5°C span is more than four times the mean modal difference between the two varieties of grebes reported by Bledsoe and Sheldon (1989).

Tracer Effects

Another factor that groups of observations share in common is the tracer component in each hybridization. For example, all hybridizations in experiments 663, 664, and 671 share the same tracer individual (C9). Figure 3 shows boxplots of residuals grouped by tracer. The groups that stand
FIGURE 2. Average homoduplex frequency melting profiles. Each curve is an average of five homoduplex curves found in tube positions 1-5 in each experiment. (a) *Aechmophorus occidentalis*. (b) *Aechmophorus clarkii*. (c) *Aechmophorus occidentalis* and *A. clarkii* combined.

out here are those corresponding to tracers O13 and O14. For group O14, however, it is not clear that the observed effects are due to the tracer component because the residuals corresponding to tracer O14 are exactly those that correspond to experiment 553 (Table 1), and we have already noted an apparent systematic effect associated with this experiment. This situation is not unique to this data set. The experimental designs of the Sibley and Ahlquist hominoid studies (1984, 1987) provide many examples of this confounding between tracer and experiment effects. Their experiments typically had a single tracer reference within an experiment, and only rarely was the same tracer used across two or more experiments. In developing statistical models for the distances, accounting for experiment effects usually removes tracer effects, and vice versa. In situations where tracer and experiment are not confounded, both effects may be present.

**Position Effects**

Sarich (pers. comm.) has discussed the possibility of a position effect in the DNAAnalyzer. The residuals discussed above are grouped according to tube position in Figure 4. There is an increasing trend in the boxes corresponding to positions 6-15, which are the intraspecific heteroduplex melts. Focusing on the medians, this conclusion is also apparent in the interspecific heteroduplex positions 18-22. Hence, there is evidence of a systematic effect within the two sets of heteroduplexes. More important is the apparent break between positions 15 and 16, between the intraspecific and interspecific heteroduplexes. The possibility thus arises that any observed differences between intraspecific and interspecific het-
eroduplex distances could be due to a position effect. To examine this possibility, position must be taken into account when estimating distances.

We have identified several sources of variation in the hybridization framework of Sibley and Ahlquist. Felsenstein (1987b) incorporated tracer and experiment effects as sources of variation in estimating distance parameters in phylogenetic trees. We have complemented his discussion by graphically illustrating the effects, identifying the potential confounding between tracer and experiment effects, identifying position effects, and demonstrating via a simple example the consequences of ignoring sources of variation on standard deviations. We propose a variance components model to estimate distance parameters. The model is similar to that suggested by Felsenstein (1987a, 1987b) but with mean structure parameters corresponding to tree branch lengths replaced by untransformed distance measure parameters.

**MODELING**

Although there are other factors that contribute to the observed variation in hybridization data (e.g., average tracer fragment length), experiment, tracer, and position are the ones for which we have information. Accounting for these factors should improve our estimates of genetic distances. A model proposed to estimate SM distance parameters is

$$\text{observed SM} = \text{true SM} + \text{position effect} + \text{experiment effect} + \text{tracer effect} + \text{residual error}. \quad (3)$$

In this model, random effects of the same type have mean zero and common variance: the experiment effects have mean zero and variance $\sigma^2_{\alpha}$, the tracer effects have mean zero and variance $\sigma^2_{\tau}$, the residual errors have mean zero and variance $\sigma^2_{R}$, and these three sets of random effects are assumed uncorrelated. The tube position effects are assumed to be fixed effects or biases, causing the SM observations to be systematically too high or too low from experiment to experiment. The distributions of the experiment, tracer, and residual effects have purposely not been specified. In some situations, the usual Gaussian assumptions might suffice, whereas other situations might require different assumptions. For example, $t$ distri-

---

**FIGURE 4.** Boxplots of SM residuals grouped according to tube position. The top and bottom of the rectangle are the upper and lower quartiles, respectively, of the residuals. The median of the residuals is indicated by the horizontal line segment within the rectangle. Vertical lines extend from the ends of the box to the tails of the distribution. Extreme points are plotted individually (●).
butions might be more appropriate in the presence of outliers, as discussed by Guerra (1992). In the grebe study, experiment and tracer effects are almost completely confounded, and inclusion of both in the model might not be necessary. This might not be the case in other studies.

Felsenstein (1987b) considered a similar model to estimate phylogenetic trees. Specifically, he used observed Δ values to estimate tree branch lengths. We estimate Δ parameters per se, not tree branch lengths. Our resulting estimates can then be entered into a distance matrix that might be used to estimate a phylogenetic tree. The rationale for the change of emphasis is that in the traditional process of going from the raw counts to trees, via distance matrix methods, inadequate attention has been given to the critical step of properly estimating the entries of these matrices. Felsenstein clearly recognized the limitations in standard practice and incorporated important components into his model, but for a purpose different from ours.

A PROBLEM OF CONFOUNDING: EXPERIMENTAL DESIGN IN THE GREBE STUDY

In the grebe study, there are four parameters of interest: the true "distances" for the intraspecific heteroduplex pairs, denoted \( d_{cc} \) for \( c_{clarkii} / c_{clarkii} \) and similarly \( d_{00} \) and \( d_{co} \). Estimates of these parameters and their associated standard deviations should provide all the necessary information to address such issues as resolution and symmetry. Our goal is to obtain estimates that reflect the way the data were obtained. The confounding problem is not a function of choice of distance measure, and therefore the qualitative conclusions hold for any measure; in the following we use SM. To explain the confounding problem, we begin by assuming that all of the 160 SM results (8 experiments × 20 heteroduplex hybrids) are available for analysis. Although this is not quite true, the difference due to about 7% missing data is unlikely to be substantial. Now let \( p_1, \ldots, p_{15}, p_{16}, \ldots, p_{25} \) denote the true position effects in the distances corresponding to tubes 6, ..., 15, 16, ..., 25, respectively.

As indicated in the previous section, we assume an additive model:

\[
\text{observed distance} = d + p + \text{error},
\]

where \( d \) is one of the four true distance parameters, \( p \) is a tube position effect, and the error may well have further structure, as in Equation 3. What we want to address is just what can and cannot be distinguished under the additive model (Eq. 4).

Consider experiments 555–671, those corresponding to \( d_{cc} \) and \( d_{co} \). Suppose that all tubes in all four experiments ran exactly the same tracer/driver pair. This assumption corresponds to no genetic difference between the two grebes. Then an estimate of

\[
(p_{16} + \ldots + p_{25}) - (p_6 + \ldots + p_{15})
\]

would be obtained by adding all 40 values from tubes 16–25, subtracting from this the sum of all 40 values from tubes 6–15, and dividing the result by 4.

Conversely, suppose that Expression 5 is equal to zero, i.e., the average position effect from tubes 16–25 is equal to that from tubes 6–15, and that the tracer/driver allocations were just as in the design given in Table 1. Then an estimate of

\[
d_{co} - d_{cc}
\]

would be obtained by adding all 40 values from tubes 16–25, subtracting from this the sum of all 40 values from tubes 6–15, and dividing the result by 40. Therefore we see that the data expression

\[
(\text{sum of values in tubes 16–25})

- (\text{sum of values in tubes 6–15})
\]

can be used in the estimation of Expression 5 when we assume Expression 6 is negligible and in the estimation of Expression 6 when we assume Expression 5 is negligible. So how are we to interpret the magnitude of Expression 7? The point is that the value of Expression 7 can be big because Expression 5 is big and Expression 6 small, or vice versa or anything in between; similar reasoning applies when Expression 7 is small. Consequently, when
we cannot assume either Expression 5 or Expression 6 to be negligible, it is impossible without prior information to distinguish their individual contributions to Expression 7. In statistical terminology, Expression 5 and Expression 6 are parameter combinations that are completely aliased. This situation is very unsatisfactory because Expression 6 is just what the experiments were meant to measure.

The aliasing problem remains when we consider the entire set of eight experiments. The total confounding of the linear combination of position effects (Exp. 5) and

\[(d_{oc} - d_{cc}) + (d_{oc} - d_{oo})\]  

(8)

does not allow for the estimation of either of the two parts in Expression 8.

The explanation of why the data cannot tell us whether Expression 7 represents primarily the difference between species or the difference between position effects, or some unknown combination of both, will now be illustrated numerically. For the full set of eight experiments, we discuss two extreme cases: attributing the magnitude of Expression 7 to Expression 5, and attributing the magnitude of Expression 7 to Expression 8. The results have taken into account experiment and tracer random effects following the proposed model. A statistically robust analysis to accommodate outliers is not needed at this point because the confounding problem does not disappear with robust methodology. Also, the data used are the observed 148 SM values.

The estimation procedure employs generalized least squares for the estimation of fixed effects (d and p) and restricted maximum likelihood (REML) for variance components associated with experiment, tracer, and residual random effects. REML is preferred over maximum likelihood for estimating variance component parameters because it yields unbiased estimates in general linear models, including the one with which we work. As a simple example, consider the sample variance of a data set \(x_1, \ldots, x_n\) computed as \(\Sigma(x_i - \bar{x})^2 / d\), where \(\bar{x}\) is the sample mean and \(d\) is either \(n\) (maximum likelihood) or \(n - 1\) (REML), both under the same Gaussian distributional assumption. The REML estimate is the well-known unbiased estimate of \(\sigma^2\), the variance of the underlying population (for standard references to REML, see Patterson and Thompson, 1971; Harville, 1977). All parameter estimates and standard deviations were obtained from Program REML (Robinson, 1987).

First, suppose that the magnitude of Expression 7 is due to species differences as given by the linear combination in Expression 8. In this case, the estimated distances (in SM units \(\times 10^3\)) are those found in Table 2 (column 3). We also have \(d_{oc} - d_{00} = 29 \pm 10\) and \(d_{oc} - d_{cc} = 51 \pm 10\), suggesting that there is a significant difference between the two varieties of grebes. The standard deviations are almost twice as large as those based on the usual arithmetic mean analysis because we have accounted for experiment and tracer as random effects.

Now suppose that the magnitude of Expression 7 is due to position effects as given by Expression 5. In this case the estimated distances (in SM units \(\times 10^3\)) are given in Table 2 (column 4). In addition, we have \(d_{oc} - d_{00} = -9 \pm 23\) and \(d_{oc} - d_{cc} = 12 \pm 23\), which suggests no difference between the two varieties of grebes, contradicting the previous results.

Thus, the aliasing does not allow us to decide on the basis of these data whether or not the two grebe varieties are separated by a genetic distance significantly different from zero.

**Estimated Position Effects**

Consider the position effect estimates based on the two extreme assumptions.

<table>
<thead>
<tr>
<th>Tracer</th>
<th>Driver</th>
<th>Species</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>O</td>
<td>O</td>
<td>22</td>
<td>41</td>
</tr>
<tr>
<td>O</td>
<td>C</td>
<td>51</td>
<td>32</td>
</tr>
<tr>
<td>C</td>
<td>O</td>
<td>61</td>
<td>41</td>
</tr>
<tr>
<td>C</td>
<td>C</td>
<td>10</td>
<td>29</td>
</tr>
</tbody>
</table>
Figure 5a shows a plot of these estimates assuming that the average position effect from tubes 6–15 is equal to that from tubes 16–25. The peculiar feature here is the break between positions 15 and 16. If the given assumption is correct, i.e., if Expression 5 is equal to zero, then an explanation for this break is needed. It seems unlikely but not impossible that some physical phenomenon such as an abrupt change at the observed position, e.g., a sudden temperature change in the bath of the apparatus, is responsible for the break. A more likely explanation would be a gradual change of temperature along the apparatus. In this case, there might be an increasing or decreasing trend of position effects when plotted against tube order in the apparatus. This trend is exactly what is observed under the assumption that there are no species differences (Fig. 5b). It seems more likely that this assumption approximates the truth in the grebe situation, although we cannot be sure.

We began this section by wanting to estimate distance parameters according to how the data were generated, taking account of known sources of variation and co-variance. In particular, we were interested in estimating the differences $d_\infty - d_o$ and $d_\infty - d_c$. We demonstrated that these differences are confounded with position effects and therefore are not estimable. With some a priori information on position effects, it might be possible to provide estimates for the differences. However, such information for the apparatus used during the grebe study (the DNAnalyzer) appears to be unavailable. Thus, the results of Ahlquist et al. (1987) and Bledsoe and Sheldon (1989) on the resolving power of DNA–DNA hybridization are questionable.

**Guidelines for Experimental Design**

First, there should be appropriate levels of replication of all relevant comparisons so that adequate estimates of the errors of estimation can be obtained. Second, the location of particular tracer/driver combinations must be assigned randomly along any apparatus such as the DNAnalyzer. If different portions of experiments are to be conducted over periods of several hours or across different days, then further randomization must be adopted. The details of such arrangements are necessarily connected to the manner in which any given experiment can be divided into subexperiments consisting of 25 hybrids, and it may be appropriate to employ some form of restricted randomization.

When running homoduplexes such as $A_i/A_i$ in experiments such as those described above, as well as intraspecific heteroduplexes $A_i/A_j$ and interspecific heteroduplexes $A_i/B_j$, it is desirable to have equal numbers of dual duplexes, e.g., of $A_j/A_j$, $A_j/A_i$, and $B_j/A_i$, respectively, to achieve the highest degree of balance in the design. This design will facilitate the statistical analysis and improve the efficiency of estimation, provided the replications are chosen appropriately. Unfortunately, the question of how many of each kind of duplex there should be is dif-
ficult. The answer will depend on each specific context and on the constraints on time and resources, the likely magnitude of the effects being estimated, and the variability being experienced.

CONCLUSIONS

The hybridization literature has traditionally been concerned with the construction of phylogenetic trees. In going from the raw counts and melting curves to the trees, insufficient attention has been given to how any particular Δ parameter is actually estimated. This practice is unsatisfactory. The confidence placed in the trees is ultimately a function of the quality of the Δ estimates found in the distance matrices, and these quantities should be properly estimated. This last statement is, of course, meaningless if what one wishes to estimate is not estimable, as with dco in the grebe investigation.

The issues presented in this paper are not limited to the grebe data. More generally, the points we have raised apply to any DNA–DNA hybridization data generated via a system with known major sources of variation and covariation or without the random assignment of particular hybrids.

ACKNOWLEDGMENTS

We thank Charles Sibley and Vincent Sarich for providing us with data and discussing DNA–DNA hybridization.

REFERENCES


Received 19 January 1995; accepted 21 July 1996

Associate Editor: Charles Marshall