Statistical Methods for Analyzing DNA-DNA Hybridization Data

By

R. Guerra and T.P. Speed

Department of Statistics
University of California
Berkeley, California 94720

TECH REPORT 277

Introduction

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 Grebe and Clark's Grebe, respectively. Their purpose was two-fold i) to determine the taxonomic status of the Western and Clark's Grebes, and ii) 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 related species", and that the mean $\Delta T_{50}H$ between the Western and Clark's Grebes is 0.57°. Bledsoe and Sheldon (1989) have discussed the same grebe data in the context of studying the metric properties of the traditional ΔT values for DNA-DNA hybridization. They concluded that "the resolution in the *Aechmorphorus* study is the species level. Below that, most individuals act as a single unit." In addition, they report a mean modal difference of 0.7° (mean of 16 values in column 6 of their Table 3) between the two grebes. If one excludes the negative ΔT_{mode} from the same column the mean modal difference between the two grebes is 0.85°.

More recently, Sarich (1990) has also discussed the same grebe data. He concludes that the distance between the two grebes is in fact "zero" and that the "non-zero" distance observed by Bledsoe and Sheldon (1989) is in all, probability an artifact due to "a systematic error introduced by the instrumentation or experimental design." In particular, Sarich argues graphically (his Figures 1 and 2), that a "position effect" in the hybridization apparatus is responsible for the artifactual discrepency between the two grebes.

The above three studies, which will be referred to as the AGS (Ahlquist grebe study), the BGS and SGS respectively, raise several important statistical issues. However, before we summarize these points, we wish to state that although the question concerning the taxonomic status of the two grebes is in itself quite interesting, we will not provide any precise guidelines for deciding on the taxonomic placement of these two grebes, based on DNA hybridization. As stated in the AGS, there are other important matters e.g., widespread sympatry and positive assortative mating, which have to be considered in conjunction with DNA hybridization results, especially at this level of genetic similarity, before any conclusions can be reached. Our interest has been to develop a method of analyzing DNA hybridization data which takes into account the various sources of experimental variation and a number of related statistical issues. We now summarize these issues.

Statistical Issues

Firstly there is the issue of variation. As will be explained below, DNA hybridization experiments such as Sibley and Ahlquist (1981, 1984) and Caccone and Powell (1989) generate data which have quite a complex structure in that there are a number of sources of variation and covariation. Most current methods of analysis do not adequately reflect the complex structure of the data. Felsenstein (1987) discusses the primate data of Sibley and Ahlquist (1987) and uses maximum likelihood with a so-called mixed model to take into account ''major sources of correlation between observations''. However, in view of the fact that his analysis did not begin with the raw data, the radioactive counts, his results cannot be viewed as definitive. Identifying and estimating different sources of variation is important since any estimated distances between species should be appropriately weighted to reflect the variation within and between experiments.

Second, standard methods compare melting curves via single parameter summaries of the curves, e.g., BGS and SGS use T_{mode} and AGS uses $T_{50}H$. It has been pointed out by number authors, e.g., Sarich et. al (1989), Bledsoe and Sheldon (1989), Sheldon and Bledsoe (1989), and Schmid and Marks (1990), that most of these summaries are flawed as measures of phylogenetic distance. In addition these descriptors do not make full use of the available data. In an attempt to make more use of the data we have developed a new measure, called the *slope measure* (SM), which attempts to quantify relative shifts between melting profiles at the high end of the temperature scale. It turns out that SM also avoids some of the flaws of existing measures.

Third, we consider what may be broadly called the quality of the data. Consider for example the BGS. Column 6 of their table 3 shows a mean ΔT_{mode} of -0.29 based on 5 observations of heterospecific heteroduplex hybrids. In theory, this dissimilarity measure should be nonnegative. What should one do about data values that are logically inconsistent with prior expectations? We should point out that here we are using the word ''data'' loosely. When referring to ΔT values as data we are in fact speaking of *derived* data. Consequently, negative ΔT values are in themselves not necessarily ''bad'' in some sense. Instead, they are a reflection of other things, experimental error or a wild homoduplex frequency curve. Below we will discuss ways of approaching ''outlying'' data.

Systematic errors are our last point. As demonstrated by the SGS there is evidence suggesting the existence of a systematic position effect in the Yale hybridization apparatus. If in fact the position effect is real, as we believe it is, we then have to consider carefully the consequences of the effect. For this we must include the effect when estimating parameters of interest, and compare in some way the resulting estimates with ''unadjusted'' estimates. Implicit in this procedure is testing for the existence of position effects. This too will be discussed below.

Methods and Data

The idea behind DNA hybridization is fairly simple.

[Insert TS description Re: s230]

The exact steps in a hybridization experiment can be performed in a variety of ways and are summarized in Sheldon and Bledsoe (1989). The raw data for the grebe comparisons were given to us by Vince Sarich in 1989. It came from DNA hybridization experiments carried out by Bledsoe and Sheldon (1989) whose methods seem to have been essentially those of Sibley and Ahlquist (1981). For an understanding of the data, the important point is that the hybrids are fractionated by hydroxylapatite (HAP) chromatography to isolate double-stranded DNA, which is then heated incrementally, and the amount of labelled DNA released at each step is measured by a radioactive counter. Consequently, the raw data takes the form of radioactive counts (see Table 1). In almost all cases, these counts are used to construct frequency or cumulative distribution curves which indicate the amount of disassociation of the hybrids along the temperature gradient (Figures 1 and 2).

Structure of the Data

The structure of the data can be inferred from the experimental design (Table 2). Here we see that individual hybrids are grouped according to experiment and that across experiments different types of hybrids e.g., conspecific heteroduplexes, are grouped according to tube position within the hybridization apparatus. Consequently individual hybrids do not necessarily give statistically independent observations, e.g. all 25 hybrids in the first row of Table 2 have at least two common "components", the fact that they are all in experiment #393, and the fact that all used tracer #3 of A. occidentalis. Most current statistical methods of analysis of hybridization data implicitly assume independence. As examples we cite the use of a t-test on the difference between reciprocal mean delta modes clarkii-occidentalis and occidentalis-clarkii, or the attaching a SE to a mean delta mode. These simple procedures ignore any structure in the data which is implicit in the experimental design. Does this matter? The following example is illuminating. Consider the 10 observed modes from the heterospecific heteroduplexes formed in experiment #393. One expression for their variance of their arithmetic mean is $\sigma^2 / 10$, where σ^2 is the error variance. However, suppose that the true model for these data is

$$\text{observed mode} = \text{true mode} + \text{experiment effect} + \text{error} \qquad (1)$$

where the common experiment effect contributes an amount σ_E^2 to the variance of each observation. Then the variance of their arithmetic mean is

$$\sigma^2/10 + \sigma_E^2. \qquad (2)$$

Comparing the initial expression for the variance with (2), we see that it is possible to grossly underestimate the variance of such arithmetic means. In our view, if units, in this case hybrids, are grouped or associated in various ways, then an analysis of the data should take into this fact account this fact, at least in its initial phases.

Frequency Curves

Our purpose here is to demonstrate, without using any particular distance measures, the existence of ''outlying'' hybrids and/or experiments. Figures 3-6 illustrate the points of interest. Perhaps the most striking is Figure 3. The curves here are average homoduplex profiles from each of the 8 experiments; Figures 3A and 3B show that the occidentalis and clarkii sets of curves are quite different in character. The clarkii curves are more sharply peaked, and with the exception of curve C, are less variable than the occidentalis curves. More important is the fact that all the curves are homoduplex curves. These curves are expected to be fairly stable, Sibley et. al (1990), but the data indicate otherwise. From Figure 3C we see that the range of modes for these curves is about 85°-88.5°.

This 3.5° span is more than four times the mean modal difference between the grebes reported by the BGS. Even when curve C is ignored, the range of ca. 86°-88.5° of the remaining modes is still quite striking. And what about curve C? Are we to assume it is a ''bad'' homoduplex curve? There is certainly reason to suspect such a curve, given our prior expectations, but before ''throwing it out'' it should be compared to the other hybrids in the same experiment. Inconsistencies with our prior expectations does not make the curve bad, and its removal will give a downward bias to our estimates of the variability inherent in the various measures. Its reliability can better be determined by considering it in relation to the other heteroduplex curves found in the same experiment, for any biases in experiment #555 in all likelihood affected every hybrid (in #555) in the same way. With this in mind one can see why simply replacing a bad homoduplex melt with a good one could lead to artifactual biases. A different way around this problem would be to throw out experiment #555, but this would be undesirable since 25 hybrids would then be discarded. We believe that there is useful information in experiment #555, and that its removal would lead to a greatly underestimated error. The point to be made here is that these so-called outlying hybrids and/or experiments are in all likelihood an indication of the (substantial) amount of between and within experiment variability, and that this should be accounted for in the analysis.

The foregoing discussion of homoduplex curves was based on a visual examination of the frequency plots. In order to get a feel for the data, we suggest looking at all the frequency plots which will enter the analysis. In this way one can get an idea of the quality and reliability of individual hybrids, and of entire experiments made up of many hybridizations.

For example, after looking at the grebe frequency plots we decided that i) experiment #393 appears to be an "ideal" experiment (Figure 4), ii) experiment #663, position #9 required manual intervention, and iii) experiment #555, plot C is Figure 3, could yield unexpected Δ-values, depending on the distance measure used, because of the placement of the homoduplex curve (Figure 6). Given this information one might decide to follow some kind of robust procedure to analyze the data. In short, looking at the data could suggest the direction which the analysis could (should?) take.

Distance Measures

There has been much discussion on the use and interpretation of distance measures for DNA hybridization. Bledsoe and Sheldon (1989) and Sheldon and Bledsoe (1989) are recent and comprehensive studies on dissimilarity measures, and we have used ideas from these papers and also those by Sarich et. al (1989) and Schmid and Marks (1990). [In this article we present and discuss three parallel analyses corresponding to three distance measures.]

Median melting temperature T_m

T_m is the temperature at which 50% of the hybrid strands is dissociated. The calculation of T_m for each hybridization is as follows. Let S be the sum of radioactive counts from 62.5° to 95°. For i = 62.5, 65, . . . , 95 let

$$C_i = \# \text{ counts eluted at temperature } i \tag{3}$$

$$N_i = 100(C_i / S). \tag{4}$$

Using the N_i's construct a cumulative distribution curve of percent dissociation versus temperature. T_m is then the median of this cumulative curve. We used linear interpolation based on the two points immediately above and below the 50% point to estimate T_m. Sample calculations are given by Sheldon and Bledsoe (1989).

Modal melting temperature T_{mode}

T_{mode} is the mode of a normalized frequency distribution based on the counts. Estimating the mode, especially for grouped data as is necessary with hybridization, is not easy. The main difficulty is that the mode may be anywhere in the most frequent group, a 2.5° range. For example, consider Figure 5. The curve with the highest ordinate has what may be called a well-defined mode and common sense tells us to estimate it as 85.5°C. However, the counts corresponding 87.5°C have been eluted throughout a temperature range of 2.5°C (85°-87.5°), and so the mode could be off by as much as 2.5°C. Note however that if the method of estimation is invariant under translations, in the sense that if the estimate based on $X_1, \ldots, X_n$ is m then based on $X_1 + \alpha, \ldots, X_n + \alpha$ it is $m + \alpha$, then this problem disappears.

Another problem is that different estimation procedures for the mode can give quite different estimates. Table 2 of Sheldon and Bledsoe (1989) presents modal estimates based on three different methods: 10th degree polynomials, modified Fermi-Dirac, and 5-point parabolic. Differences exceeding 2°C in modal estimates and 1°C in Δ-mode estimates can be found in the table.

By definition, modes are local properties of curves. Our estimation procedure attempts to reflect this fact: we find the highest point of the frequency curve, (x, y) say, and fit a parabola to (x, y) and its two neighboring points. If the two neighboring points have the same ordinate value, then (x, y) will be the estimate of the mode; otherwise, the estimate will be pulled in the direction of the neighbor with the highest ordinate. This procedure is not only intuitively appealing, it is also objective and simple to program.

The method gives good reproducibility, in that replicates of the same duplex, within an experiment, agree quite well (Table 3). The method also gives realistic estimates. That is, the estimates tend to agree with visual estimates based on frequency curve plots.

We have compared our intraspecific heteroduplex ΔT_{mode} values for the grebe data with those of Bledsoe and Sheldon (1989). In that paper they use the modified Fermi-Dirac curve fitting procedure, presumably because the resulting curve matches the *shape* of the observed frequency curve. As can be seen in Table 4 our results can differ markedly. Sarich (1989) also takes a local approach to estimating the mode. His procedure is to estimate the mode as the intersection of the best fitting straight lines involving the three values to the left of the mode, and the two to the right. For comparison we have applied his method to a number frequency curves and our results tend to agree with his. In particular, his method estimates the ΔT_{mode} (as described in Table 4) for experiment #671, positions 21-25 to be 0.66°C. See Figure 7.

There are two points to be made here. The first is that locally based estimates will tend to agree with one another. Secondly, locally based estimates can substantially differ from global ones for the obvious reason that global procedures, such as high order polynomials and the modified Fermi-Dirac, by their nature attempt to capture the *entire* shape of the curve. As a result, global procedures can overfit the data or be highly influenced by quirks in the data. In short, we do not see any reason to use points for removal from the mode in estimating it.

The slope measure SM

This measure essentially compares the rate of decay, at the high temperature and of the melt, between homoduplex and heteroduplex profiles. Let

$$\begin{aligned} g(t) &= \text{homoduplex profile} \\ f(t) &= \text{heteroduplex profile.} \end{aligned} \tag{5}$$

Plots of hybridization frequency curves indicate that the high temperature part of the profiles decays approximately exponentially. Assume that beyond a certain temperature, say t_0, g and f take the form

$$\begin{aligned} g(t) &= e^{-ct} \\ f(t) &= e^{-c't}. \end{aligned} \tag{6}$$

Then under certain constraints we can show that

$$\ln\left[\frac{g(t)}{f(t)}\right] = (c' - c)(t - b) \tag{7}$$

for some b. In other words, the log of the ratio of the two curves is a line with slope $c' - c$. The closer the curves to each other the smaller the slope is; in fact, the slope approaches zero. Our idea is to estimate this slope and use it as a measure of distance between the two curves. The estimate is called the *slope measure* (SM) estimate of the distance between the duplex curves.

We have found that for the grebe data the exponential decay model fits the profiles quite well in the temperature range 87.5°-95°C, so in practice our method is as follows:

Let S = sum of counts from 60°-95°C

$$N_t = \frac{1}{S}(\text{count at temperature } t).$$

Now compute

$$s(t) = \log\left[\frac{N_t \text{ for homoduplex}}{N_t \text{ for heteroduplex}}\right], \qquad t = 87.5\,(2.5)\,95$$

and find the 'best' fit line to the points (t, s(t)) for t = 87.5 (2.5) 95.

The slope of the best fit line is SM for the given comparison.

Figure 8 provides an example of the method. We should point that we are computing the log of the ratio of two *normalized* counts. If for a given set of comparisons, say, within an experiment, there is only one homoduplex, then it suffices to compute the log of the ratio of rare counts. However, if there are several homoduplexes available within the experiment, then either one homoduplex must be selected to be the 'standard' homoduplex, and everything else compared to it, including the other homoduplexes, or an average homoduplex is constructed and all heteroduplexes compared to it. In the second case it is convenient to work with ratios of normalized counts.

So far we have presented what is essentially a mathematical procedure to compare the curves. We now discuss the biological aspects of the method. As has been pointed out by Sheldon and Bledsoe (1989) it is desirable to use characteristics of melting profiles which reflect overall hybrid stability rather than percent mismatches. The reason is that the melting curves reflect the ''dissociation of many duplexes spanning a [wide] range of complementarity (and their dissociation temperatures)''. The mode is such a characteristic. We argue that the SM is also an indicator of duplex stability. Consider two curves, one a homoduplex and the other an interspecific heteroduplex. In theory the heteroduplex curve is expected to be shifted toward the incubation temperature relative to the homoduplex. In turn the mode of the heteroduplex is shifted towards the lower end of the temperature scale. As previously mentioned, frequency plots show that the high temperature side of the profiles can be modelled by exponential functions. To be more precise, for any individual frequency plot, the exponential model describes that part of the curve which lies to the right of the mode. One more note, *all* types of duplex curves are expected to have an ordinate of zero at 95°C — the last point of the temperature gradient. Hence, if the mode of the heteroduplex is m_0 and that of the homoduplex is $m_0 + \Delta$ $(\Delta > 0)$, then the exponential rate of decay for the heteroduplex must be greater than the rate for the homoduplex. It is exactly this comparison which equation (7) subsumes. It is now clear that $c' - c$ increases as Δ increases. In short, the SM 'tracks' Δ and it follows that SM is a measure of duplex stability.