Problems with the Cladistic Use of Riboprinting

Mark E. Siddall
Division of Invertebrate Zoology, American Museum of Natural History, Central Park West at 79th Street, New York, New York 10024

Accepted May 20, 2001

A method of character acquisition that is nearly unique to protozoological systematics is riboprinting. Briefly, the method consists of amplifying a portion of the tandemly arranged eukaryotic ribosomal genes with universal primers followed by examination of the banding patterns that result from digestion of the product with an array of restriction enzymes (Van den Bussche, 1991; Clark, 1992, 1993, 1997). The availability of universal primers that allow amplification of homologous sequences from a wide variety of taxa and the fact that the genes occur in multiple copies in the genome are advantageous and contribute to ease of amplification. Moreover, for studies concerned with many taxa, riboprinting may prove to be more economical than direct DNA sequencing and certainly will generate results in less time. However, the use of riboprinting is less straightforward than current applications might suggest and there are difficulties associated with the interpretation of restriction electromorph patterns. For the purpose of species or strain identification, these problems are minimal—either the electromorph patterns are the same or they are not. The extension of their use to phylogenetic applications (e.g., Brown and de Jonckheere, 1994; Clark et al., 1995; Clark, 1997; Pernin and de Jonckheere, 1996; Clark and Diamond, 1997; Xiao and Desser, 2000) does not follow as readily because of how electromorph patterns relate to homology statements.

A character coded into a phylogenetic matrix is a homology statement. Central to the problem of accurate homology statements is the coding of presence or absence of individual bands as though they were separate and independent characters. A variety of pitfalls regarding the use of riboprinting exist and solutions are suggested here.

DIGESTIONS

In their examination of microsporidian relationships, Pomport-Castillion et al. (1997) first amplified a region of the ribosomal repeat that includes SSU rDNA, ITS, and LSU rDNA. Electrophoresis of the uncut amplified products showed inequalities in size across the 12 taxa examined. Specifically, the 4 species of Nosema yielded shorter products than all of the other taxa, and Agamasoma penaei rendered a band of intermediate size. This difference in length is, of course, demonstrative of insertion/deletion events (INDELS) in the history of these taxa. Although this itself is worthy of phylogenetic consideration, it would be unwise to count a single INDEL more than once (Schaal, 1985; Dowling et al., 1996). Figure 1 illustrates this problem for a hypothetical case. Even though each of two enzymes cuts in precisely the same homologous places in 2 taxa they
Problems with the Cladistic Use of Riboprinting

When amplification products to be cut by restriction enzymes are of unequal size due to an INDEL, this single event can eventually be counted more than once. Even though there may be no difference in restriction sites for enzymes (a, b) the presence of one INDEL will cause steps to be contributed for each enzyme used. Additional steps can result from this same INDEL if an enzyme cuts within that region (c).

In Figs. 1a and 1b, two sites for each enzyme are identical for the 2 taxa. However, in both cases, the intervening INDEL (which is a single event) causes a size difference in banding patterns that would, then, be counted independently for each enzyme used. In Fig. 1c, wherein the enzyme finds a canonical site within the INDEL region of the taxon with the longer product, this event also is counted independently. For presence/absence coding of bands, this difference alone would entail an additional three steps using the presence/absence coding strategy. The sum of the number of transformations resulting from presence/absence coding for just these three enzymes is seven steps, and yet only one event has actually happened historically; the single INDEL explains all of the differences between the 2 isolates. Comparison of the banding patterns depicted by Pomport-Castillion et al. (1997) for HincII, AluI, BanI, andMspI digests suggests that this problem may have erroneously contributed extra steps in their analyses.

The simplest solution to the problem of unequal amplification products is to investigate alternative pairs of primers until equal-sized fragments are obtained for all study taxa prior to restriction digestion. An advantage to the use of ribosomal fragments are obtained for all study taxa prior to restriction digestion. An advantage to the use of ribosomal genes is that there are many highly conserved regions that can serve as targets for PCR primers (Hillis and Dixon, 1991). This allows for many possible permutations and many opportunities for obtaining equal-sized products. In those cases in which equal-sized products cannot be obtained, algebraic mapping of homologous fragments (discussed more fully below) may provide a solution; otherwise it would be reasonable to admit that riboprinting cannot be used effectively and that DNA sequencing is the appropriate alternative.

Restriction enzymes recognize canonical stretches of DNA (usually palindromic sites) for digestion. However, a variety of conditions can yield bewildering results. For example, in the AluI digestion pattern obtained by Pomport-Castillion et al. (1997) (Fig. 2), for A. panaei (H), the two heaviest fragments are too long to sum to the size of the uncut product. Perhaps A. panaei exhibits heterogeneity in its ribosomal DNA (Bentzen et al., 1988) and there are two mixed products in the restriction digest.

Incomplete digestion will yield more bands than corresponding restriction sites would dictate (Dowling et al., 1996). Unfortunately, so too will overdigestion. Many enzymes exhibit “star formation” or cleavage at

![FIG. 1.](image1)

![FIG. 2.](image2)
noncanonical sites under a variety of conditions. A particularly vexing problem in terms of riboprinting is those enzymes that exhibit star formation in the presence of dimethyl sulfoxide (DMSO). DMSO is used routinely in amplification reactions of rDNAs in order to stabilize the reaction against secondary structure formation (Weiss et al., 1992). The properties of most restriction enzymes are well documented and readily available from suppliers. Those enzymes that are prone to star formation or to variable efficacies are easily avoided. To avoid the problems posed by genomic heterogeneity it is advisable to clone PCR products prior to enzymatic digestion and thus ensure single-copy products (Bentzen et al., 1988). It is paramount that the actual size of each fragment is documented against a standard and that the sum of those fragments approximates the size of the uncut PCR product. If this minimal condition cannot be satisfied, the results from digestion with that enzyme (e.g., the patterns from AluI, AvrII, Ddel, HhaI, and MboII in Pomport-Castillion et al., 1997) should be discarded as spurious.

Various authors (Pomport-Castillion et al., 1997; Xiao and Desser, 2000) indicate that they chose restriction enzymes arbitrarily. Arbitrary choices are problematic because some enzymes have overlapping recognition sites. Thus, a single historical nucleotide change can be counted more than once if it results in two different enzymes cutting at the same place in the sequence (Hugall et al., 1994; Dowling et al., 1996). The simplest form of this redundancy relates to 4-base (tetramer) cutters. For example, MboII, which recognizes ...GATC..., will cut every site that is cut by BgII that recognizes ...AGATCT... However, the form that this redundancy takes cannot always be subtracted a posteriori because of unpredictable overlap. For example, HinII and AccI recognize ...GTYRAC... and ...GTCGAC... respectively; thus a ...GTCGAC... site will be cut by both, whereas a ...GTAAAC... site will be recognized only by HinII. Additional enzymatic redundancies that will have confounded riboprinting studies include NcoI (...CCRG...), MspI (...CCGG...), ApaI (...GGGCC...), HaeIII (...GGCC...), and BanI (...GGYRCC...) with RsaI (...GTAC...). These patterns may be explained by a single restriction site in C, D, and E. However, presence/absence coding methods evaluate this difference as three steps instead of one.

**CODING**

Clark et al. (1995), Clark and Diamond (1997), Pomport-Castillion et al. (1997), and Xiao and Desser (2000) have coded individual electromorphs (fragments) as characters with presence or absence being binary alternative states. This method, though perhaps intuitively easy, repeatedly has been criticized by phylogeneticists (e.g., Mickevich and Mitter, 1981; Adams and Rothman, 1982; Patton and Avise, 1983; Swofford and Olsen, 1990; Murphy, 1993; Dowling et al., 1996; Mishler et al., 1996). Consider Fig. 3 reproducing digestions with ApaI.
Problems with the Cladistic Use of Riboprinting

by Pomport-Castillion et al. (1997). Taxon A has no restriction sites and taxon C has one such site for cleavage. That is, there is a single change. However, coding the presence and absence of bands renders a minimum of three steps (loss of one band and gain of two or loss of two and gain of one, depending on the directionality of change) when logically only one transformation has occurred. This inequality would not be problematic if all events counted three steps. Yet, whereas the difference between no restriction sites and one restriction site is three steps, the difference between no sites and two sites is four steps (not six). In addition, and more difficult still, is that it is possible for two taxa to have the same restriction site, and yet have no electromorphs in common (Fig. 4). It would seem to be erroneous to conclude that two taxa have nothing in common when they do.

The solution is to code the restriction site, not the restriction fragment. By evaluating the actual sizes of fragments yielded by restriction digestion, it is possible, although difficult, to map the number of transformations that have occurred between any two taxa (Adams and Rothman, 1982; Templeton, 1983; DeBry and Slade, 1985; Avise, 1994; Dowling et al., 1996). In the simplest case, where every taxon renders no more than two fragments, the solution is trivial, but even more complex patterns can be solved algebraically. Consider, for example, the patterns depicted by Pomport-Castillion et al. (1997) for the MboI digestion of Spraguea lophi (A), Glugea stephani (C), Microgemma ovoidea (F), and Nosema costelytrae (I) reproduced here in Fig. 4a. The following homology statements obtain:

\[ A_4 = C_3 = I_4 \]
\[ A_1 = F_1 \]
\[ A_2 = F_2 \]
\[ A_3 = I_3 \text{ and} \]
\[ C_2 = I_2. \]

From these we may infer

\[ \text{if } A_1 = F_1, A_2 = F_2, A_4 = I_4 \text{ and } A_3 = I_3 \]
\[ \text{then } F_3 = A_3 + A_4 = I_3 + I_4 \]

\[ \text{if } A_4 = C_3 \text{ then } C_1 + C_2 = A_1 + A_2 + A_3 \]

\[ \text{if } A_4 = I_4 \text{ and } A_3 = I_3 \]
\[ \text{then } A_1 + A_2 = I_1 + I_2 \text{ and} \]

\[ \text{if } C_2 = I_2 \text{ and } C_3 = I_4 \text{ then } C_1 = I_1 + I_3. \]

As a result, there is only one possible solution to the determination of restriction sites from these algebraic axioms (Fig. 4b). Note that although F and I have no fragments in common, we can infer algebraically that they do share a common restriction site. Coding of presence and absence of restriction fragments would yield the following array of binary characters and states

\[
\begin{align*}
A & \quad 10001011 \\
C & \quad 01010001 \\
F & \quad 10001100 \\
I & \quad 00110011 \\
\end{align*}
\]

and would yield a consistency index of 0.78, whereas coding the restriction sites yields the matrix

\[
\begin{align*}
A & \quad 0111 \\
C & \quad 1001 \\
F & \quad 0110 \\
I & \quad 1011 \\
\end{align*}
\]

and would exhibit no homoplasy whatsoever.

Although the example given proves to be tractable,
it is not likely that this always will be so or necessarily so for all taxa. An “agnostic coding” alternative is to code each enzyme as a character and each unique banding pattern as a separate state. Thus, the agnostic alternative for the four taxa considered would render the unordered matrix

\[
\begin{array}{cc}
A & 0 \\
C & 1 \\
F & 2 \\
I & 4 \\
\end{array}
\]

The problem here, though, is the loss of any relationship among the banding patterns and the information that there is more shared between taxa than the agnostic method allows. There is a generalized algebraic shortcut to determining the maximum possible number of site differences (D) between any two banding patterns i and j

\[
D_{ij} = F_i + F_j - 2(S_{ij}) - 2,
\]

where \(F\) is the total number of restriction fragments for a taxon (i or j) and \(S_{ij}\) is the number of shared fragments. This information can be coded by way of a Sankoff character in PAUP* (Swofford, 1999) or SPA (Goloboff, 1996), for the MboI example, with taxa and states,

\[
\begin{array}{ccc}
A & a \\
C & b \\
F & c \\
I & d \\
\end{array}
\]

and with the symmetrical Sankoff cost matrix,

\[
\begin{array}{cccc}
a & b & c & d \\
a & 0 & 3 & 1 & 2 \\
b & 3 & 0 & 4 & 1 \\
c & 1 & 4 & 0 & 5 \\
d & 2 & 1 & 5 & 0 \\
\end{array}
\]

The preceding corresponds reasonably to what the algebraic solution in Fig. 4b suggests but there is still a two-step overestimation of the amount of change between *Mic. ovoida* and *N. costelytrae* (F and I, respectively).

In consideration of the preceding pitfalls that could confound phylogeny reconstruction with respect to riboprinting it should not be surprising that the trees found by Pomport-Castillion et al. (1997) for their two analyses might be subject to revision in light of these alternative methods. Unlike their previous analyses that depicted arbitrarily rooted trees, I have drawn trees as unrooted networks (Figs. 5a and 6a). Use of the agnostic coding method, in which each enzyme is a character and each unique electromorph pattern is a state for that character, provided more resolution for

![FIG. 5. Phylogenetic trees of 12 microsporidians that result from (a) presence/absence coding, (b) agnostic coding, and (c) using the Sankoff character approach.](image-url)
the positions of sites cannot logically be solved, it may be wise to admit to that fact and not include the results of that enzyme, or to try the agnostic and Sankoff approaches.

Ultimately it may be more reasonable to simply sequence the actual nucleotides as opposed to relying on inferential methods. By way of a simple example, Fig. 7a shows the inferred relationships of five species in the phylum Haplosporidia (Haplosporidium nelsoni, Haplosporidium costale, Haplosporidium louisiana, Minchinia teredinis, and Urosporidium crescens) when coding for the presence and absence of restriction sites in their known 18S rDNA sequences using AatII, AflII, BanIII, BclI, BglII, BsiCI, BsmI, BspMI, BssHII, FdiII, NaeI, NarI, NcoI, NheI, PvuI, SmaI, ScalI, BstXI, EagI, HphiI, and KpnI. In this case, because the sequences are known, the results of inferring phylogeny on the basis of restriction sites can be directly compared with the results that would be inferred from the sequenced gene (Fig. 7b). Clearly, these are not in agreement with respect to whether Min. teredinis is closest to H. costale or to H.
nelsoni. Insofar as the restriction sites are just an indirect assessment of sequences, it is disturbing that with only five taxa, and armed with 22 enzymes, the riboprinting method would not properly reflect the nucleotide phylogeny.

Since the riboprinting studies cited here were published, the DNA sequences for most of the organisms concerned have been completed. Comparison of the results is revealing. For example, in their study of Entamoeba species, Clark and Diamond (1997) could not even resolve the relationships of E. chattoni, E. polecki, and E. coli (Fig. 8a). Using actual sequences poses no difficulty in doing so and the results disagree substantially with those from the former method in terms of the relationships of E. hartmanni and E. insolita (Fig. 8b). Similarly, Xiao and Desser’s (2000) results of riboprinting for myxozoans (Fig. 8c) are significantly different (P < 0.0001 using any of the Kishino–Hasegawa, Templeton, or winning sites tests) from those found using the DNA sequence data (Fig. 8d).

Because riboprints are awkward to interpret or code for phylogenetic analysis and because they are obviously prone to yielding spurious results that will only be overturned when someone sequences the rDNA locus, we recommend that they simply be avoided altogether in favor of the now easy and inexpensive methods of DNA sequencing.

ACKNOWLEDGMENTS

I thank Arnold Kluge for comments on an earlier draft and the Cladistics Institute of Harbor Springs Michigan for visiting researcher status during the completion of this work. Data regarding the haplosporidians were obtained with the assistance of a grant from the National Science Foundation (BIO/DEB-9629487) in collaboration with Kim Reece and Eugene Burreson.

REFERENCES


Copyright © 2001 by The Willi Hennig Society
All rights of reproduction in any form reserved


