Cladistics (1990)6:363–367

NUCLEIC ACID SEQUENCE PHYLOGENY AND RANDOM OUTGROUPS

Ward C. Wheeler[1]

[1] *Department of Invertebrates, American Museum of Natural History, Central Park West at 79th St., New York, NY 10024-5192, U.S.A.*

Abstract When divergent taxa are used to root networks, it is assumed that the character states in the outgroup have historical similarity to those in the ingroup. Yet, if the data are nucleic acid sequences, the character states shared by a divergent outgroup may be based not on history but on random similarity. A simple procedure is proposed to test this possibility. In the absence of an appropriate outgroup, root position can be estimated with the use of an asymmetrical character transformation matrix. If the matrix is sufficiently biased, it can supply the polarity information usually derived from an outgroup. This outgroup test and rooting procedure are demonstrated with ADH sequences from the genus *Drosophila*.

Introduction

Rooting undirected networks or topologies can be problematic. The construction of a hypothetical ancestor from a group of taxa, and the use of specific outgroup taxa, depend on the outgroup having information about the polarities of the characters that vary within the ingroup. Specifically, the outgroup can determine which character states shared within the ingroup are pleisiomorphic relative to other states. When the most exclusive sister group is used to root a topology, this assumption about the informative value of the outgroup is likely to be robust. But when appeals are made to more distantly-related groups, it becomes questionable. In fact, if the possible character states are constrained as far as the ingroup is concerned, a distant outgroup taxon may be nothing more than a random collection of character states with little or no historical content. Thus, such a configuration may result in a random root for the topology.

A related problem concerns the unavailability of an informative outgroup—cases where sister groups are too distant—or the possibility that there is no outgroup, as with hypotheses that include both bacterial and eukaryotic taxa. In these instances, some kind of polarity statement is needed to root the topologies and to distinguish among hypotheses that differ only in the placement of the root.

Here, I present a method to examine the implications of using a random-sequence outgroup to root a topology derived from nucleic acid sequence data. I also propose a method to root these topologies without an outgroup, using a combinatorial weight matrix instead (Wheeler, 1990).

Properties of a "Random Root"

I am defining a random root as a series of ancestor—or outgroup—based polarity statements that have no historical information. Due to the great amount of evolution from the outgroup/ancestor to the ingroup, the distribution of presumptive primitive states is indistinguishable from that of a random model.

Since there are only four states in any nucleic acid sequence character (excluding gaps), the probability of two bases having the same character state is readily calculable. The probability of random identity is simply the sum of the square of the occurrence of each of the four bases ($\sum P_i^2$). If the overall frequency of each base is the same (25%),

0748-3007/90/040363 + 05$03.00/0

then there is a 25% chance, given a random sequence, that any one position in an ingroup taxon will match that of the outgroup. Among the variable positions in the sequence, the expected number of matches is equal to the probability of a single match multiplied by the number of these positions. This value is binomially distributed, with the individual match probability designated by the parameter P. The expected position and distance of the outgroup to the ingroup should follow a basic binomial model: this is the crux of the test for a random root.[1]

There are two steps in the examination of a putatively random root. The first step is to determine the placement of a random sequence on a network. If the root is truly random, the probability of the addition of the outgroup sequence to any branch within the ingroup network will be proportional to the length of that branch. Since there are multiple branches, this addition behavior will be multinomially distributed, with P_1, P_2, P_3, etc., representing the length of each branch (1, 2, 3, and so on) divided by the overall length of the network. Hence a random outgroup is expected to join the ingroup on the longest branch.

The next step focuses on determining the evolutionary distance from the outgroup to the ingroup. Again if the sequence is truly random, there should be $\sum P_i^2$ similarity (depending on base frequencies) between the random sequence and the others. The frequency of expected matches multiplied by the number of characters yields the expected branch distance from a random sequence to its addition point on the network. If the branch distance to the outgroup is equal to this expectation (within sampling error), the root is effectively random. The distance between outgroup and ingroup should be binomially distributed. If the expectation of random sequence matching is P, then the variance in the branch length is the length of the sequence multiplied by P and $1-P$:

$$\sigma^2_{\text{distance}} = P(1-P)N_{\text{variable positions}}$$

or if you prefer frequencies:

$$\sigma^2_{\text{distance}} = \sqrt{(Pq/N_{\text{variable positions}})}.$$

An Example

The alcohol dehydrogenase coding sequences for six *Drosophila* species (Bodmer and Ashburner, 1984) were analysed using PAUP (Swofford, 1985). The two outgroup taxa, *affinisdisjuncta* and *mulleri*, were used to determine the relationships of the four ingroup taxa, *orena*, *simulans*, *mauritiana*, and *melanogaster*. Fitch optimization was employed to construct the most parsimonious cladogram via the branch-and-bound algorithm (Fig. 1).

There were 133 informative sites. The most parsimonious cladogram had a length of 184 steps with a consistency index of 92%. Of the length, 80 to 106 steps separated the node joining the two outgroup taxa to the ingroup taxa. This spread was due to multiple, equally parsimonious character state reconstructions on the same cladogram. When the two outgroup taxa were used individually to root the cladogram, only 44 to 55 steps

[1] Only variable positions are included because they are the only nucleotides which are used for polarity assessment; conserved bases have no systematic information. I am not saying that two aligned random sequences would have 25% similarity. In fact, since alignment algorithms augment similarity through the inclusion of gaps, their match frequency would be even greater.

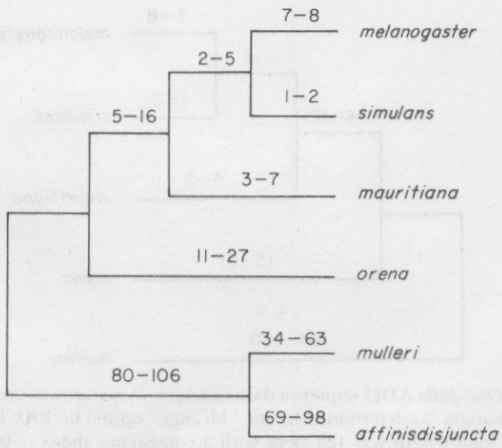

Fig. 1. Cladogram for *Drosophila* ADH sequence data. The number of changes (as determined by the "blrange" option on PAUP) between nodes is shown above each branch. The total length was 282 steps with a consistency index of 0.95 for 236 characters. (When only informative characters were analysed, the length was 184 steps with a consistency index of 0.918 for 133 characters.)

(20–25%) of the total tree length were needed within the ingroup rooted with *affinisdisjuncta* (Fig. 2), while 44 to 50 steps (24–27%) were required when *mulleri* was used (Fig. 3). The fraction of change in the ingroup was not significantly different from either the $29 \pm 7\%$ ($\mu \pm 2\sigma$) expected from the *mulleri* rooting or the $29 \pm 6\%$ expected from *affinisdisjuncta*. Thus the outgroup sequences behave as if they were random. As expected, the outgroup was added to the longest branch in the ingroup—the *orena* branch, along which most ingroup changes occurred.

This is a clear-cut example of an outgroup which is entirely too distant to elucidate ingroup relationships. Any hypothesis of phylogenetic arrangement of the taxa based on these data would be suspect.

The Determination of a Root Without an Outgroup

When situations arise, as with the *Drosophila* example here, in which the possible outgroups are too distant, rooting can only be accomplished through the use of an

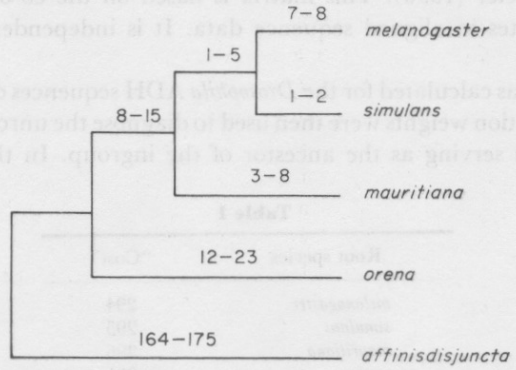

Fig. 2. Cladogram for *Drosophila* ADH sequence data in which *D. affinisdisjuncta* was used to root the four ingroup species. The number of changes between nodes (as determined by the "blrange" option on PAUP) is shown above each branch. The total length was 219 steps with a consistency index of 0.968 for 236 characters.

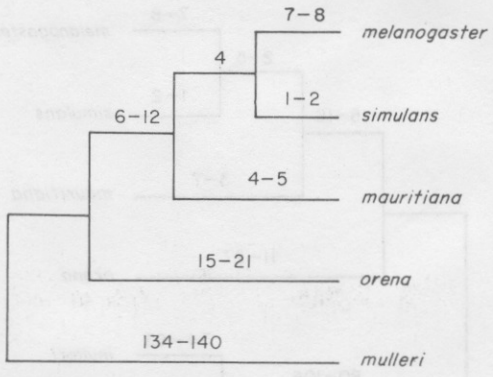

Fig. 3. Cladogram for *Drosophila* ADH sequence data in which *D. mulleri* was used to root the four ingroup species. The number of changes (as determined by the "blrange" option on PAUP) between nodes is shown above each branch. The total length was 184 steps with a consistency index of 0.978 for 236 characters.

	A	C	G	T
A	—	1.47	1.05	2.53
C	0.71	—	0.78	0.25
G	0.92	1.39	—	1.90
T	2.21	0.65	1.66	—

Fig. 4. DNA character transformation weight matrix for the *Drosophila* ADH sequences.

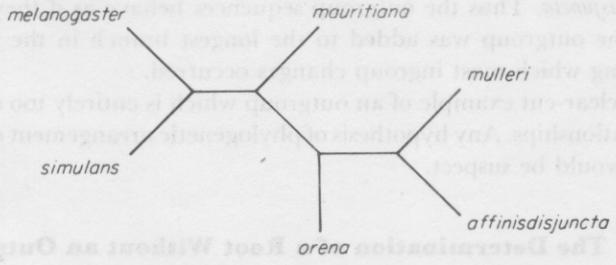

Fig. 5. Unrooted topology for the six *Drosophila* species.

asymmetrical character transformation matrix. The derivation of such a matrix is described by Wheeler (1990). This matrix is based on the co-occurrence of nucleic acid character states in aligned sequence data. It is independent of the cladogram topology.

Such a matrix was calculated for the *Drosophila* ADH sequences described above (Fig. 4). The transformation weights were then used to diagnose the unrooted topology in Fig. 5 with each taxon serving as the ancestor of the ingroup. In this way, all variable

Table 1

Root species	"Cost"
melanogaster	294
simulans	295
mauritiana	296
orena	291
mulleri	289
affinisdisjuncta	273

positions influenced the calculation of cladogram "cost", as described by Wheeler (1990).

The costs of these rootings are shown in Table 1. The selection of *affinisdisjuncta* as the ancestor is clearly the least costly of all the rootings. This difference is "significant" at the 5% level.

Conclusions

Many molecular systematic studies rely on distant outgroups (Miyamoto et al., 1989). In these situations, some knowledge of the polarity information and general quality of the outgroup is necessary to assess hypotheses generated with its sequence. The test presented here offers an avenue to that essential knowledge.

Generally, a random outgroup sequence will join to the longest branch of the ingroup (the largest target); it will be separated from other taxa by a very long branch. When this is the case, the root position is highly unreliable or even meaningless.[2]

However, when an outgroup is either too distant or for some reason unavailable, it may still be possible to gather polarity information. If one can detect asymmetry in the character transformations without topology information, the network can be rooted. The greater the degree of asymmetry, the greater the likelihood of finding a unique, well-supported root.

In the example shown here, the outgroup appeared to be random. However, when a root was determined via a character transformation matrix, the same root position was supported. In fact, this is also the rooting offered by Grimaldi's (1990) work on drosophilid relationships. In other cases, there may not be such agreement. Thus the roots determined by distant outgroups should be suspect.

Acknowledgments

I would like to thank Elise Broach, Kirk Fitzhugh, and Paul Vrana for commenting (extensively) on an earlier version of this manuscript.

REFERENCES

Bodmer, M. and M. Ashburner. 1984. Conservation and change in the DNA sequences coding for alcohol dehydrogenase in sibling species of *Drosophila*. Nature 309: 425–430.

Felsenstein, J. 1981. A likelihood approach to character weighting and what it tells us about parsimony and compatibility. Biol. J. Linn. Soc. 16: 183–196.

Grimaldi, D. A. 1990. A phylogenetic, revised classification of genera in the Drosophilidae (Diptera). Bulletin of the American Museum of Natural History (in press.)

Miyamoto, M. M. and S. M. Boyle. 1989. The potential importance of mitochondrial DNA sequence data to eutherian mammal phylogeny. *In:* B. Fernholm, K. Bremer, and H. Jornvall (eds), The Hierarchy of Life. Elsevier Press, Amsterdam, pp. 437–450.

Swofford, D. L. 1985. PAUP 2.4.1. Program and documentation. Champaign, Illinois.

Wheeler, W. C. 1990. Combinatorial weights in phylogenetic analysis: a statistical parsimony procedure. Cladistics 6: 269–275.

[2] Ingroup taxa may also lack historical similarity to each other and can have analogous, very long branch lengths. Perhaps this procedure would allow for the detection of the branch length problems that Felsenstein (1981) and others attribute to parsimony.

(Received for publication 18 November 1988; accepted 22 May 1990)