Tree Rooting with Outgroups When They Differ in Their Nucleotide Composition from the Ingroup: The Drosophila saltans and willistoni Groups, a Case Study

Rosa Tarrío,* Francisco Rodríguez-Trelles,*†,1 and Francisco J. Ayala*

*Department of Ecology and Evolutionary Biology, University of California, Irvine, California 92697-2525; and †Instituto de Investigaciones Agrobiológicas de Galicia (CSIC), Avenida de Vigo s/n, Apartado 122, 15780 Santiago de Compostela, Spain

Received August 26, 1999; revised April 4, 2000

Rooting is frequently the most precarious step in any phylogenetic analysis. Outgroups can become useless for rooting if they are too distantly related to the ingroup. Specifically, little attention has been paid to scenarios where outgroups have evolved different nucleotide frequencies from the ingroup. We investigate one empirical example that arose seeking to determine the phylogenetic relationship between the saltans and the willistoni groups of Drosophila (subgenus Sophophora). We have analyzed 2085 coding nucleotides from the xanthine dehydrogenase (Xdh) gene in 14 species, 6 from the saltans group and 8 from the willistoni group. We adopt a two-step strategy: (1) we investigate the phylogeny without outgroups, rooting the network by the midpoint method; (2) we reinvestigate the rooting of this phylogeny using predefined outgroups in both a parsimony- and a model-based maximum-likelihood framework. A satisfactory description of the substitution process along the Xdh region calls for six substitution types and substitution rate variation among codon positions. When the ingroup sequences are considered alone, the phylogeny obtained using this description corroborates the known relationships derived from anatomical criteria. Inclusion of the outgroups makes the root unstable, apparently because of differences between ingroups and outgroups in the substitution processes; these differences are better accounted for by a simplified model of evolution than by more complex, realistic descriptions of the substitution process.

Key Words: outgroup and midpoint rooting; Drosophila saltans and willistoni groups; Xdh; nucleotide composition; GC content; maximum-likelihood, among-site rate variation.

Given a set of aligned sequences from four or more taxa, most tree-building methods yield unrooted trees. Unrooted trees, or networks, inform us about the relationships among the taxa, but if we are to distinguish plesiomorphic from synapomorphic changes and decide, accordingly, the path of evolution, the network must be rooted. Usually the only feasible way of determining the polarity of nucleotide character evolution is using one or more outgroups (i.e., taxa that are assumed to lie cladistically outside of a putative monophyletic group), such that a character state shared within the ingroup is assumed to be ancestral if is also present in the outgroup. Still, outgroup rooting is frequently recognized as the most precarious step in phylogenetic reconstruction (Wheeler, 1990; Smith, 1994; Swofford et al., 1996).

The topic of rooting using outgroups has been addressed in a series of papers (Lundberg, 1972; Watrous and Wheeler, 1981; Maddison et al., 1984; Wheeler, 1990; Smith, 1994) focused on situations where the outgroup(s) show similar nucleotide frequencies to the ingroup. In such a case, the most likely source of topological bias at the time of rooting is long-branch attraction (Felsenstein, 1978). This phenomenon arises when the outgroups are distantly related (either because of large divergence time, or increased rates of evolution) to the ingroup, so that homoplastic changes at rapidly evolving sites can lead to artifactual rooting (termed “random rooting” by Wheeler, 1990) along longer ingroup branches (Felsenstein, 1978; Hendy and Penny, 1989; Wheeler, 1990; Maddison et al., 1992). Long-branch-attraction biases are confronted by employing tree-building methods that account for variation in the substitution rate among sites present in the sequences. In this way, the phylogenetic value of each site is given a weight which is inversely related to the rate of evolution of the site, so that rare events are emphasized relative to fast evolving, putatively homoplastic positions. Eventually, the ingroup taxa are analyzed separately, and the outgroup(s) is connected to the resulting network secondarily (Lundberg, 1972; Nixon and Carpenter, 1993). However, no attention has been paid to scenarios...
where the outgroups have diverged in their nucleotide composition from the ingroup. Yet, there is growing evidence that base composition differences among taxa are quite common in nature (e.g., Sueoka, 1962; Bernardi et al., 1988; Rodríguez-Trelles et al., 1999a). It can therefore be expected that, at least for some species groups, the only suitable outgroups for phylogenetic reconstruction are taxa with very different nucleotide frequencies to the ingroup. Here we investigate one example that arose while trying to determine the phylogenetic relationships between the *saltans* and *willistoni* species groups of *Drosophila. The *saltans-willistoni* lineage has evolved highly distinctive nucleotide frequencies (Rodríguez-Trelles et al., 1999a, 2000a) to the extent that the only potential rooting outgroups are compositionally very different.

Locating the Root of a Compositionally Diverged Ingroup: The Case of the saltans and willistoni Species Groups of *Drosophila*

The saltans and the willistoni groups represent the New World radiation of the Sophophora subgenus (*Throckmorton, 1975; Wotjas et al., 1992; Tatarenkov et al., 1999*). Currently they consist, respectively, of five (*cordata, elliptica, parasaltans, saltans, and sturtevanti*) (Magalhães, 1962), and two (*willistoni and bocainensis*) (see Val et al., 1981) species subgroups (Fig. 1). The two species groups are considered monophyletic because they can unambiguously be distinguished from each other morphologically (*Throckmorton, 1975*), and because of the deletion of an intron of the *Adh* gene specific to the willistoni group (*Anderson et al., 1993; our unpublished results*). Evolutionary relationships have been investigated separately for each group using nucleotide variation (*Gleason et al., 1998; O’Grady et al., 1998; Rodríguez-Trelles et al., 1999b*). None of these studies attempted to combine the two species groups on the same rooted tree. Here we address this issue using already published sequences (*Tarrío et al., 1998; Rodríguez-Trelles et al., 1999b, 2000b*). The data set consists of 2085 aligned coding nucleotides from the xanthine dehydrogenase *(Xdh)* gene in 17 species: 6 saltans and 8 willistoni representatives.
Log-likelihood values are obtained assuming the topology shown in the Fig. 1. 

Different nucleotide frequencies from the

1989). Figure 1 shows that the outgroups exhibit very

10

willistoni species and the outgroups, respectively; 69.7%, for the average GC content across the differences are accounted for by the 3rd (45.5% vs 60.7%;

willistoni, vs 60.7%;

differences between the

1st and 2nd

positions alone contain insufficient phylogenetic information to resolve the short-time-scale relationships involved in the diversification of the salts and willistoni species. We follow a statistical model-fitting approach within the maximum-likelihood (ML) framework of phylogenetic inference (e.g., Ritland and Clegg, 1985; Yang et al., 1995; Cunningham et al., 1998; Rodríguez-Trelles et al., 1999b). As a working topology for model fitting we use a topology that remains stable after applying the computer programs DNAML, DNADIST, and DNAPARS from the PHYLIP package (Felsenstein, 1993), using the default options. This topology coincides with that shown in the Fig. 1; use of alternative topologies is not expected to affect parameter estimates (Yang, 1994). We use three approaches to fit the among-site rate variation into the substitution models: the invariable sites model of Hasegawa et al. (1985) (referred to as INV% model), the discrete gamma distribution method of Yang (1996a) setting eight equal-probability categories of rates (dG models), and the codon-position-specific rate parameter model of Yang (1996b) (C models). In addition to ML, we use distance-based neighbor-joining (NJ), and parsimony criteria for tree making. The estimates of the among-site rate variation and transition bias that we use in distance computation and weighting for parsimony are those obtained by ML in the model-fitting stage, which can be considered the most reliable (Yang, 1996a).

The best description so far of the substitution process along Xdh is attained with the REV + C model (see Table 1), which allows six substitution types (two C ↔ T and A ↔ G transitions, and four C ↔ A, C ↔ G, T ↔ A, and T ↔ G transversions), and substitution rate differences between codon positions. Figure 1 shows the ML tree obtained for the Xdh data set using the REV + C model, with the position of the root (branch labeled 1) inferred by the midpoint rooting method; as noted earlier, this method assumes that the most divergent taxa in the phylogeny evolve at equal rates. To check this premise we tested the somewhat more re-
strictive assumption that the rate of evolution is the same for all taxa (i.e., the molecular clock assumption). Likelihood-ratio tests were performed under two reasonable topologies: one representing saltans and willistoni as monophyletic sister groups, and a similar one except for placing D. subsaltans as an outgroup; in both cases the clock assumption is rejected (\(-2 \log \Lambda = 59.10; 12 \text{ df}, P < 10^{-3}\), and \(-2 \log \Lambda = 67.70; 12 \text{ df}, P < 10^{-6}\), for each topology respectively); however, after removing the relatively slowly evolving D. emarginata, D. sicuinea, and D. capricorni (see Fig. 1), the data set meets the clock assumption (\(-2 \log \Lambda = 9.00; 9 \text{ df}, P = 0.44\), and \(-2 \log \Lambda = 21.90; 9 \text{ df}, P = 0.01\). The ML topology shown in the figure clusters the saltans and willistoni species as separate lineages. If we use the estimates of the rate variation among-sites and transition bias obtained by ML for tree building with distance and maximum parsimony, this result remains unchanged (not shown). So far, analysis of the ingroup alone indicates that saltans and willistoni are monophyletic groups, which agrees with previous studies (Throckmorton, 1975; Anderson et al., 1993); in the following, we will assume this to represent the best corroborated phylogeny.

Table 1 shows the results of the likelihood-ratio tests for the data set combining the saltans-willistoni species and the outgroups. Log-likelihood values are obtained assuming the topology shown in Fig. 1. As it is the case for the ingroup sequences (see Table 1), the best depiction of the substitution process in Xdh is provided by the REV + C model; inclusion of the outgroups has little effect on parameter estimates (results not shown). When inferring the tree, however, the position of the root is unstable, depending on the tree-building criterion. Of the methods assayed, only the NJ, based upon the LogDet + %INV distance correction, places the root along the branch joining the saltans and willistoni groups (branch 1 in Fig. 1); ML assuming the REV + C model and weighted parsimony situate the root in the branch leading to D. subsaltans (branch 2), whereas the distance criterion based on the REV + dG model places the root along the branch leading to D. nebulosa (branch 3).

The reasons why ML, generally more robust to assumption violations than maximum-parsimony and distance methods (see Felsenstein, 1988; Huelsenbeck, 1995; Swofford et al., 1996), fails to find the putatively true location of the root are unclear; more so given the relatively “realistic” model of evolution used to infer the tree (Table 1). Other confounding factors, besides base composition differences among sequences, may be at work. Recent stimulation studies have shown that tree shape, as determined by the length of the branches, can affect the efficiency of ML for recovering the true tree, so that overly simple models often outperform more realistic models of evolution (Yang, 1997b; Bruno and Halpern, 1999). In order to explore this possibility, we repeated the analyses above but ignored the among-site rate variation (i.e., components C, dG, and %NV are dropped from the models; see Table 1). In this case, only unweighted maximum parsimony miss-places the root (it locates it again on the branch leading to D. subsaltans; branch 2 in Fig. 1), whereas the ML, assuming the REV model, and the distance methods (either based upon the LogDet or the REV transformation) root the tree at the junction between the saltans and the willistoni species groups (branch 1).

Phylogenetic Implications

Outgroup sequences can become useless for determining the polarity of ingroup characters if they (i) are distantly related, or (ii) have evolved too fast, so that their sequences have become effectively randomized with respect to the ingroup. Here we provide an empirical example that differences in nucleotide composition between the outgroup(s) and the ingroup can also be a significant source of bias for plesiomorphy assessment. We draw two conclusions: (i) simple models recover the best corroborated topology in most cases, as well as or better than a variety of fairly realistic models of Xdh evolution; and (ii) among the complex models, solely the LogDet distance transformation (either ignoring the among-site rate variation or using the INV% model) succeeds in accurately rooting the tree.

Since the LogDet correction is the only one that relaxes the stationarity assumption (i.e., constant nucleotide frequencies) (Lockhart et al., 1994), our second conclusion points to the large base composition differences between ingroup taxa and outgroup taxa as the factor responsible for the inability of the ML method to identify the root. Compositional differences, however, do not appear to be the only determinant, because ML succeeds when a simple model of evolution is assumed. Long-branch attraction (Felsenstein, 1978) can also be at work. This would explain that ML and parsimony locate the root along the branch leading to D. subsaltans (see Fig. 1). This scenario might seem unlikely, inasmuch as our best-fit model accommodates rate differences among sites, so that fast-changing sites, mainly 3rd codon positions, should have little influence on the phylogeny (Yang, 1996a; Cunningham et al., 1998). In our case, however, the amount of among-site rate variation accounted for by the models could be insufficient for phylogenetic methods to overcome long-branch attraction.

This interpretation need not be at odds with the observation that, when the among-site rate variation is ignored, all methods correctly identify the root (except parsimony; see Fig. 1). As we have noted earlier, the Xdh sequences from the saltans group species have a somewhat lower GC content than the willistoni sequences in 3rd codon positions; the latter species are more similar to the outgroups in this respect. Account-
ing for substitution-rate differences among sites would thus reduce the phylogenetic signal contained in the nucleotide composition of 3rd codon positions, which place D. subsaltans within the saltans group. However, the signal would emerge after removal of the among-site rate variation from the model. Long-branch attraction and nucleotide composition differences would thus be exerting antagonistic effects on the phylogeny, their relative importance depending on the parameters included in the model.

Simplified models were already known that can outperform more realistic models when long branches are adjacent in the true phylogeny (Yang, 1997b; Bruno and Halpern, 1999). The present study illustrates that fairly simple models can also perform best when long-branches are not contiguous (the outgroups and D. subsaltans are at opposite ends of an internal branch; see Fig. 1). Our results challenge conclusions from simulation studies about the importance of accommodating among-site rate variation into the substitution models to overcome long-branch attraction problems (Cunningham et al., 1998). Our case may rather add to the list of examples unveiling nonlinearities in the performance of tree-building methods.

Our results corroborate the phylogenetic relationships among the saltans species previously obtained in a ML study combining data from several nuclear (Xdh, Adh, 28SrRNA, and ITS1) and mitochondrial (COI and COII) regions (Rodrı́guez-Trelles et al., 1999b). The study used outgroups from the compositionally highly dissimilar melanogaster and obscura groups. Our inferences, based as they are on the most closely related outgroups of the willistoni group, are expected to be more reliable. This apparent insensitivity of the saltans topology to the base composition of outgroups may happen because of the relatively homogeneous base composition across the species of the group (Fig. 1); accordingly all species would equally be influenced by the composition of the reference taxa. Regarding the willistoni group, the Xdh phylogeny unambiguously places bocainensis as the earliest derived subgroup and shows it to be paraphyletic, with D. nebulosa being more closely related to the willistoni subgroup siblings than to D. sucinae and D. capricorni (Ayala et al., 1975). The relationship among the willistoni siblings are fully consistent with those derived previously from a combined analyses of the per, Adh, and COI loci (Gleason et al., 1998).

ACKNOWLEDGMENTS

We are indebted to Victor DeFilippis for help with the PCR analyses of the Adh intron, and to Rob DeSalle for critical comments. F.R-T. has received support from Consejo Superior de Investigaciones Cientı́ficas (Spain) (Contrato Temporal de Investigación) and Grant XUGA 400001B98 to A. M. Vieitez. Research supported by NIH Grant GM42397 to F.J.A.

REFERENCES

Anderson, C. L., Carew, E. A., and Powell, J. R. (1993). Evolution of the Adh locus in the Drosophila willistoni group: The loss of an intron, and shift in codon usage. Mol. Biol. Evol. 10: 605–618.

Ayala, F. J., Tracey, M. L., Hedgecock, D., and Richmond, R. C. (1975). Genetic differentiation during the speciation process in Drosophila. Evolution 28: 576–592.

Bernardi, G. B., Olofsson, B., Filipski, J., et al. (1985). The mosaic genome of warm-blooded vertebrates. Science 228: 953–958.

Bruno, W. J., and Halpern, A. L. (1999). Topological bias and inconsistency of maximum likelihood using wrong models. Mol. Biol. Evol. 16: 564–566.

Cunningham, C. W., Zhu, H., and Hillis, D. M. (1998). Best-fit maximum-likelihood models for phylogenetic inference: Empirical tests with known phylogenies. Evolution 52: 978–987.

Felsenstein, J. (1978). Cases in which parsimony and compatibility will be positively misleading. Syst. Zool. 27: 401–410.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. J. Mol. Evol. 17: 368–376.

Felsenstein, J. (1988). Phylogenies from molecular sequences: inference and reliability. Annu. Rev. Genet. 22: 521–565.

Felsenstein, J. (1993). PHYLIP, phylogenetic inference package an documentation, Version 3.5c. Distributed by the author, Department of Genetics, University of Washington, Seattle.

Gleason, J. M., Griffith, E. C., and Powell, J. R. (1998). A molecular phylogeny of the Drosophila willistoni group: Conflicts between species concepts? Evolution 52: 1093–1103.

Hasegawa, M., Kishino, H., and Yano, T. (1985). Dating the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22: 160–174.

Hendy, M. D., and Penny, D. (1989). A framework for the quantitative studies of evolutionary trees. Syst. Zool. 38: 297–309.

Huelsenbeck, J. P. (1995). The performance of phylogenetic methods in simulation. Syst. Biol. 44: 17–48.

Jukes, T. H., and Cantor, C. R. (1969). Evolution of protein molecules. In "Mammalian protein metabolism" (H. N. Munro Ed.), pp. 21–132, Academic Press, New York.

Lockhart, P. J., Steel, M. A., Hendy, M. D., and Penny, D. (1994). Recovering evolutionary trees under a more realistic model of substitution-rate differences, based as they are on the most closely related species previously obtained in a ML study combining data from several nuclear (Xdh, Adh, 28SrRNA, and ITS1) and mitochondrial (COI and COII) regions (Rodrı́guez-Trelles et al., 1999b). The study used outgroups from the compositionally highly dissimilar melanogaster and obscura groups. Our inferences, based as they are on the most closely related outgroups of the willistoni group, are expected to be more reliable. This apparent insensitivity of the saltans topology to the base composition of outgroups may happen because of the relatively homogeneous base composition across the species of the group (Fig. 1); accordingly all species would equally be influenced by the composition of the reference taxa. Regarding the willistoni group, the Xdh phylogeny unambiguously places bocainensis as the earliest derived subgroup and shows it to be paraphyletic, with D. nebulosa being more closely related to the willistoni subgroup siblings than to D. sucinae and D. capricorni (Ayala et al., 1975). The relationship among the willistoni siblings are fully consistent with those derived previously from a combined analyses of the per, Adh, and COI loci (Gleason et al., 1998).

ACKNOWLEDGMENTS

We are indebted to Victor DeFilippis for help with the PCR analyses of the Adh intron, and to Rob DeSalle for critical comments. F.R-T. has received support from Consejo Superior de Investigaciones Cientı́ficas (Spain) (Contrato Temporal de Investigación) and Grant XUGA 400001B98 to A. M. Vieitez. Research supported by NIH Grant GM42397 to F.J.A.
Rodríguez-Trelles, F., Tarrío, R., and Ayala, F. J. (1999b). Molecular evolution and phylogeny of the Drosophila saltans species group inferred from the Xdh gene. Mol. Phylogenet. Evol. 13: 110–121.

Rodriguez-Trelles, F., Tarrío, R., and Ayala, F. J. (2000a). Fluctuating mutation bias and the evolution of base composition in Drosophila. J. Mol. Evol. 50: 1–10.

Rodriguez-Trelles, F., Tarrío, R., and Ayala, F. J. (2000b). Disparate evolution of paralogous introns in the Xdh gene of Drosophila. J. Mol. Evol. 50: 123–130.

Rzhetsky, A., and Nei, M. (1995). Tests of the applicability of several substitution models for DNA sequence data. Mol. Biol. Evol. 12: 131–151.

Smith, A. B. (1994). Rooting molecular trees: problems and strategies. Biol. J. Linn. Soc. 51: 279–292.

Strimmer, K., and von Haeseler, A. (1996). Quartet puzzling: A quartet maximum-likelihood method for reconstructing tree topologies. Mol. Biol. Evol. 13: 964–969.

Sueoka, N. (1962). On the genetic basis of variation and heterogeneity of DNA base composition. Proc. Natl. Acad. Sci. USA 48: 582–592.

Swofford, D. L. (1999). PAUP*: Phylogenetic Analysis Using Parsimony (and Other Methods), Version 4.0b2, in progress. Sinauer, Sunderland, MA.

Swofford, D. L., Olsen, G. J., Waddell, P. J., and Hillis, D. M. (1996). Phylogenetic inference. In “Molecular Systematics” (D. M. Hillis, C. Moritz, and B. K. Mable, Eds.), pp. 407–514, Sinauer, Sunderland, MA.

Tamura, K., and Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. 10: 512–526.

Tarrío, R., Rodríguez-Trelles, F., and Ayala, F. J. (1998). New Drosophila introns originate by duplication. Proc. Natl. Acad. Sci. USA 95: 1658–1662.

Tatarinov, A., Kvatowitzki, J., Skarecky, D., Barrio, E., and Ayala, F. J. (1999). On the evolution of Dopa decarboxylase (Ddc) and Drosophila systematics. J. Mol. Evol. 48: 445–462.

Throckmorton, L. H. (1975). The phylogeny, ecology and geography of Drosophila. In “Handbook of Genetics” (R. C. King, Ed.), Vol. 3, pp. 421–436, Plenum, New York.

Val, F. C., Vilela, C. R., and Marques, M. D. (1981). Drosophilidae of the Neotropical region. In “The Genetics and Biology of Drosophilia,” (M. Ashburner, H. L. Carson, and J. N. Thompson, Jr., Eds.), Vol. 3a, pp. 123–168, Academic Press, NY.

Watrous, L. E., and Wheeler, Q. D. (1981). The out-group comparison method of character analysis. Syst. Zool. 30: 1–11.

Wheeler, W. C. (1990). Nucleic acid sequence phylogeny and random outgroups. Cladistics 6: 363–367.

Wojtas, K. M., von Kalm, L., Weaver, J. R., and Sullivan, D. T. (1992). The evolution of duplicate glyceraldehyde-3-phosphate dehydrogenase genes in Drosophila. Genetics 132: 789–797.

Yang, Z. (1994). Estimating the pattern of nucleotide substitution. J. Mol. Evol. 39: 105–111.

Yang, Z. (1996a). The among-site rate variation and its impact on phylogenetic analyses. TREE 11: 367–372.

Yang, Z. (1996b). Maximum likelihood models for combined analyses of multiple data. J. Mol. Evol. 42: 587–596.

Yang, Z. (1997b). How often do wrong models produce better phylogenies? Mol. Biol. Evol. 14: 105–108.

Yang, Z., Launder, I. J., and Lin, H. (1995). Molecular evolution of the Hepatitis B virus genome. J. Mol. Evol. 41: 587–596.