Biological Journal of the Linnean Society, 2007, **92**, 669–674.

An empirical test of the midpoint rooting method

PABLO N. HESS and CLAUDIA A. DE MORAES RUSSO*

Laboratório de Biodiversidade Molecular, Departamento de Genética, Instituto de Biologia, CCS, Bloco A, Av Pau Brasil 211, Universidade Federal do Rio de Janeiro, Ilha do Fundão, Rio de Janeiro, RJ 21941-S70, Brazil

Received 21 April 2006; accepted for publication 12 February 2007

The outgroup method is widely used to root phylogenetic trees. An accurate root indication, however, strongly depends on the availability of a proper outgroup. An alternate rooting method is the midpoint rooting (MPR). In this case, the root is set at the midpoint between the two most divergent operational taxonomic units. Although the midpoint rooting algorithm has been extensively used, the efficiency of this method in retrieving the correct root remains untested. In the present study, we empirically tested the success rate of the MPR in obtaining the outgroup root for a given phylogenetic tree. This was carried out by eliminating outgroups in 50 selected data sets from 33 papers and rooting the trees with the midpoint method. We were thus able to compare the root position retrieved by each method. Data sets were separated into three categories with different root consistencies: data sets with a single outgroup taxon (54% success rate for MPR), data sets with multiple outgroup taxa that showed inconsistency in root position (82% success rate), and data sets with multiple outgroup taxa in which root position was consistent (94% success rate). Interestingly, the more consistent the outgroup root is, the more successful MPR appears to be. This is a strong indication that the MPR method is valuable, particularly for cases where a proper outgroup is unavailable.

ADDITIONAL KEYWORDS: molecular clock – outgroup rooting – outgroups – phylogenetic trees – systematics – unrooted trees.

INTRODUCTION

Rooting evolutionary trees is usually considered a simple step in phylogenetic construction. Nonetheless, tree building algorithms produce unrooted phylogenetic trees because all the processes leading to the final tree are computed as reversible (Swofford *et al*., 1996; Nei & Kumar, 2000; Sanderson & Shaffer, 2002). Despite its importance, however, rooting is often overlooked in phylogenetic constructions (Swofford *et al*., 1996).

The outgroup method is the most widely used in phylogenetic studies but the correct indication of the root position strongly depends on the availability of a proper outgroup (Hendy & Penny, 1989; Wheeler, 1990; Tarrío, Rodríguez-Trelles & Ayala, 2000). This apparently simple requisite may prove rather limiting when studying viruses (Stavrinides & Guttman, 2004), mostly because of extremely high and diverse evolutionary rates in these organisms. Higher taxonomic groups such as Angiosperms (Qiu *et al*., 2001), birds, and mammals (Holland, Penny & Hendy, 2003) may also be subject to the lack of appropriate extant outgroups.

Additionally, issues such as long-branch attraction (Felsenstein, 1978; Qiu *et al*., 2001; Sanderson & Shaffer, 2002), differences in nucleotide composition between taxa (Tarrío *et al*., 2000), and long-edge attraction (Hendy & Penny, 1989) represent major misleading factors for outgroup rooting.

As previously suggested (Tarrío *et al*., 2000; Sanderson & Shaffer, 2002), the midpoint rooting method (also known as MPR; Farris, 1972) might be useful in these situations because it does not depend on the existence of an outgroup. The MPR method places the root of the tree at the midpoint between the two most divergent operational taxonomic units (OTUs) (Swofford *et al*., 1996; Nei & Kumar, 2000), as measured by the sum of branch lengths between these OTUs. The theoretical basis of MPR relies on the assumption that all OTUs in a given tree should display the same

*Corresponding author. E-mail: claudia@biologia.ufrj.br

average evolutionary rate (Tarrío *et al*., 2000; Huelsenbeck, Bollback & Levine, 2002). Although the midpoint rooting algorithm has been extensively used, the efficiency of this method in retrieving the correct root remains untested.

By eliminating outgroups in data sets that are not problematic in regards to outgroup selection, and rooting the tree with the midpoint method, we were able to compare the root position retrieved by each method. Therefore, in the present study, we empirically tested the success rate of the MPR in obtaining the same root as the outgroup method for a given tree, and verified that it shows a surprisingly high performance.

MATERIAL AND METHODS

To evaluate the success rate of MPR in rooting trees, we selected data sets from the literature. As increases in taxonomic level usually reflect on evolutionary distances between sequences, which in turn lead to progressive violations of the molecular clock assumptions, we restricted our choice to papers focusing on low taxonomic levels of tetrapods (i.e. congeneric species and members of a single species). Furthermore, among those, we also selected papers that analysed mitochondrial genes, to minimize issues caused by paralogy and recombination (Avise *et al*., 1987; Overton & Rhoads, 2004). tRNA-coding segments in the sequences were not used because their sequences were usually incomplete.

Of the 33 papers selected, 13 utilized more than one gene, and thus the total number of individual-gene data sets amounted to 50. The number of OTUs in each data set varied from six to 169 (mean = 25.2, SD = 25.3). As previously explained, all data sets included at least one outgroup, and 28 of them provided more than one (for details, see Supplementary Material).

DNA sequences were retrieved from the NCBI molecular database as indicated by the authors. Protein-coding sequence alignments were performed with ClustalW (Higgins, Thompson & Gibbs, 1994) implementation in DAMBE, version 4.2.13 (Xia & Xie, 2001), based on their respective amino acid products. Noncoding sequences, such as rRNA genes and the mitochondrial D-loop region, were also aligned with the ClustalW implementation present in the DAMBE software. All alignments were performed using default parameters, and they were visually inspected and corrected whenever appropriate.

Phylogenetic and molecular evolutionary analyses were conducted using MEGA, version 2.1 (Kumar *et al*., 2001). The Neighbour-joining method (Saitou & Nei, 1987) was used to reconstruct all phylogenetic trees because of its reliability and computer time limitations for other methods (Kuhner & Felsenstein, 1994; Russo, Takezaki & Nei, 1996; Rosenberg & Kumar, 2001). As expected, intra- and interspecies p-distance measures were small (mean = 0.104, SD = 0.053), a condition that favours the use of the Jukes–Cantor correction (Jukes & Cantor, 1969) due to its smaller variance when compared to more complex evolutionary models (Nei, 1991; Russo, 1997). A bootstrap test (Felsenstein, 1985) with 2000 replicates (Hedges, 1992) was performed on all phylogenetic trees to evaluate statistical branch support (Hillis & Bull, 1993; Sitnikova, Rzhetsky & Nei, 1995).

Thus, we proceeded to the empirical test of the MPR, which required the assignment of an outgroup root. All data sets with a single outgroup (herein termed 'single outgroup data sets'; SO) had their outgroup roots straightforwardly assigned. The other data sets (named 'multiple outgroup data sets'), however, were subject to outgroup root consistency checks (Maddison, Donoghue & Maddison, 1984). Such checks were performed by comparing the root yielded by each of the available outgroups individually. In addition, we also compared these root positions with the one obtained through the simultaneous use of all outgroups. When individual outgroups were inconsistent, but the combination of all outgroups produced a tree in which they were all joined at the same root position, we assigned that position as the outgroup root for MPR comparison purposes. These data sets were named 'multiple outgroup, inconsistently rooted data sets' (MOI). In the two MOI data sets in which the combination of multiple outgroups did not produce a single root position, the final root was based on a majority-rule consensus of individual outgroups.

Finally, the last category of data sets was the 'multiple outgroup, consistently rooted data sets' (MOC), in which all outgroups, either individually or combined, yielded the exact same root position. To test the performance of the MPR based on the outgroup method, one midpoint-rooted tree was constructed for each data set. Naturally, the outgroup was excluded from this analysis. The SYSTAT program, version 11 (available at http://www.systat.com) was used to perform a nonparametric Kruskal–Wallis test to check the homogeneity concerning the numbers of ingroups and outgroups, among and within the three different categories (SO, MOI, and MOC). Additionally, we evaluated the significance of differences in MPR success rates among categories by a chi-square test in SYSTAT, version 11.

RESULTS

In the present study, we assumed that the outgroup method yields the correct root position in every tree. Unfortunately, this assumption may be doubtful in

some cases (Holland *et al*., 2003), yet testing the choice of outgroup by systematists is clearly beyond the scope of our study. Nevertheless, 28 out of 50 analysed data sets (i.e. the MOI and MOC categories) provided multiple outgroups. This allowed us to reduce the potential issue of outgroup root misplacement through ingroup monophyly checks (Maddison *et al*., 1984).

Another issue is how to deal with topological differences caused by the exclusion of the outgroups. This is bound to happen, particularly at this taxonomic level, because the closeness between species produces some short branches with typically low support. We attempted to minimize such problems by analysing root differences in condensed trees. For this, we used 33% and 50% cut-off values for condensing the trees. Cut-off values indicate the minimum support required for a branch to remain uncollapsed. Therefore, the application of a 33% cut-off value to a tree causes every branch with a bootstrap value lower than 33% to be collapsed and become part of a polytomy. Cut-off values higher than those have consistently produced complete polytomies (data not shown).

Outgroup number and root confidence

As previously mentioned, the SO data sets were unsuitable for ingroup monophyly checks, lending this category an uncertain degree of confidence in root placements. Even though the data sets in the MOI category allowed us to test every tree for ingroup monophyly, their inconsistent results also portrayed a doubtful root position. Therefore, we placed an intermediate confidence on the root positions derived by the outgroup method for the data sets in this category. Doubtless confirmation of ingroup monophyly was only possible in the MOC category, which also showed the highest success rate of the MPR amongst all categories.

Generally, data set features such as number of ingroups, number of OTUs, and mean distance between ingroups and outgroups were not significantly distinct among the three categories (SO, MOI, MOC). This result eliminates some potential sources of biases in our analyses. Regarding the 33% cut-off trees, we found significant differences among categories referring to MPR success rates. However, we were unable to establish such a difference between the MOI and MOC categories in the trees condensed at the 50% cut-off limit. Nevertheless, when considered as one category (MOI + MOC), there was a significant difference from the SO category (χ^2: $P = 0.035$ and 0.030, at 33% and 50% cut-off values, respectively).

Condensed trees and root confidence

When trees were condensed at the 33% bootstrap cut-off limit, MPR correctly placed the root in 35 of 50 (70%) data sets. In the SO category, the MPR method retrieved the correct root location in only 54% of such data sets (Table 1) whereas, in the multiple-outgroup data sets (MOI and MOC categories), MPR achieved a much higher (82%) success rate. More specifically, in the 12 MOI data sets, MPR achieved a 67% success rate, whereas the 16 MOC data sets yielded an impressive 94% success rate.

Table 1. Midpoint rooting success rates for the three data set categories

Category	33% cut-off	50% cut-off
SO	54%	64%
MOI	67%	83%*
MOC	94%	94%*

*We failed to assign statistical difference between these values. Only when used in combination (MOI + MOC) did these values show statistical difference from the SO category in the 50% cut-off value.

Percentages indicate cut-off values used for condensing the trees.

MOC, data sets with multiple outgroups available, which showed no consistency issues; MOI, data sets with multiple outgroups available, which showed inconsistencies in rooting the trees; SO, data sets with only one outgroup available.

In the 50% cut-off trees analysis, the overall (SO + MOI + MOC) number of MPR successes was slightly larger than in the 33% analysis, increasing from 35 (70%) to 39 (78%) out of 50 data sets. Data sets in the SO category were correctly rooted by MPR on 64% of the trees (Table 1). On the other hand, the multiple-outgroup data sets (MOI + MOC) yielded a 89% success rate for MPR. The MOI category alone yielded a 83% success rate, whereas the MOC data sets maintained the 94% rate already achieved through the 33% cut-off condensation.

To ascertain that the collapsing of branches had not artificially increased the success rate of the midpoint method using condensed trees, we also analysed the success rate in noncondensed trees. In this case, the midpoint method successfully retrieved the correct root for 33 out of 50 (66%) data sets. It is interesting to note that, in ten (65%) of the remaining 17 data sets, the midpoint root position was a single node away from that derived by the outgroup method.

To interpret the low MPR success rate (66%) on noncondensed trees, it should not be overlooked that these trees often had their roots placed, by both methods, on branches with very low support values. Consequently, such root positions are highly uncertain themselves. Therefore, the noncondensed trees

ought to remain inconclusive, even though such success rates are higher than the expected by chance (Huelsenbeck, Bollback & Levine, 2002). Condensed trees, on the other hand, allow us to place greater confidence on every branch because poorly supported branches are collapsed. Higher cut-off values are capable of reducing even further the effects of poorly supported branches. By utilizing this approach, we could briefly investigate, in more detail, whether failures in the MPR method were due to phylogenetic reconstruction problems in general.

DISCUSSION

The midpoint method displayed an impressively high success rate, which is especially remarkable in the MOC data sets because these are the situations in which we know the root position with the greater degree of confidence.

Furthermore, on every data set category, MPR offered better results with trees condensed at greater cut-off values. Conversely, for all trees condensed at the same values, the midpoint method achieved greater success in the data sets with higher branch (and thus root) confidence. Again, this is a clear indication that a consistent outgroup root placement also corresponds to an increase in the MPR success rate with the same data. For example, in the SO category, it is possible that MPR retrieved the correct root whereas the single outgroup did not. In the MOC category, the trees were already quite trustworthy at the 33% cut-off threshold, which is demonstrated by the 94% MPR success rate. Results with the higher 50% condensation value corroborate this trend, as the success rate of the midpoint method remained the same.

The MPR success rates in the MOC category are surprisingly high for a rooting method based solely on branch lengths and, hence, highly dependent on the assumption of an untested molecular clock (Holland *et al*., 2003). Nevertheless, in spite of such a high efficiency in placing the root, we would expect that the performance of the midpoint method would be reduced in higher taxa. One would expect this to happen because the main assumption of this method (i.e. homogeneity of substitution rates along the tree, or a clock-like behaviour for the sequences) tends to be progressively violated as biological processes become more distinct between historically distant lineages (Li, 1997).

Interestingly, Huelsenbeck *et al*. (2002) showed that even severe violations of the molecular clock assumptions still allow for a moderate, yet significant, success rate at rooting trees with the direct molecular clock rooting method, which, by definition, strongly depends on clock-like evolution. Therefore, we suggest that the MPR method, which is slightly less dependent on the assumptions of a molecular clock, might as well be successfully applied to higher level phylogenetic reconstructions.

At this point, it is important to mention some issues that may affect outgroup rooting. Long-branch attraction (Felsenstein, 1978; Qiu *et al*., 2001; Sanderson & Shaffer, 2002) is probably the most important and debated source of failure for the outgroup method. In this case, the often long outgroup branch may be attached to other long branches in the tree, thus yielding a wrong root position. Another source of error for the outgroup rooting may be due to differences in nucleotide composition between outgroups and ingroups (Tarrío *et al*., 2000). The difference might confound character polarity, and thus also contribute to outgroups being clustered with OTUs based on sequences compositions rather than on their evolutionary relationships. Finally, long-edge attraction (Hendy & Penny, 1989) may also cause the outgroup to cluster with any external long branch with higher probability than to correctly place the root on one of the short internal branches.

Most major animal groups have their internal phylogenetic relationships already stable and trustfully established. Hence, the aforementioned circularity in the outgroup method usually poses no problem for such groups, but it is often a restricting factor for viruses (Stavrinides & Guttman, 2004) because of their high, usually heterogeneous evolutionary rates, and on account of the lack of a priori phylogenetic information on them. Also for some major groups, such as angiosperms (Qiu *et al*., 2001), whose adequate sister-groups are extinct, and, in some situations, even birds and mammals (Holland *et al*., 2003), phylogenies are affected by problems in the application of the outgroup method. In such cases, MPR might become a more valuable method than the outgroup method for retrieving the correct root position in the tree.

When any of these issues are in effect, outgroup rooting usually becomes a less convenient option for rooting (Tarrío *et al*., 2000; Holland *et al*., 2003), and midpoint rooting may be preferred.

Considering the surprisingly high success rates for the midpoint method, we suggest that it should be used as an alternative rooting method, and it could be adopted by default when outgroup rooting is not straightforward and the constructed phylogeny is stable enough.

ACKNOWLEDGEMENTS

We thank Carlos E. G. Schrago and Carolina M. Voloch for valuable suggestions on a former version of the manuscript. This study is part of the MSc thesis of Pablo Nehab-Hess at the Genetics Graduate

Program at the Universidade Federal do Rio de Janeiro. The study was funded by CNPq (Brazilian Science and Technology Ministry) and FAPERJ (Rio de Janeiro Government). P.N.-H. was sponsored by CAPES (Brazilian Education Ministry).

REFERENCES

Avise JC, Arnold J, Ball RM, Bermingham E, Lamb T, Neigel JE, Reeb CA, Saunders NC. 1987. Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. *Annual Review of Ecology and Systematics* **18:** 489–522.

Farris J. 1972. Estimating phylogenetic trees from distance matrices. *American Naturalist* **106:** 645–667.

Felsenstein J. 1978. Cases in which parsimony and compatibility will be positively misleading. *Systematic Zoology* **27:** 401–410.

Felsenstein J. 1985. Confidence-limits on phylogenies – an approach using the bootstrap. *Evolution* **4:** 783–791.

Hedges SB. 1992. The number of replications needed for accurate estimation of the bootstrap-P value in phylogenetic studies. *Molecular Biology and Evolution* **9:** 366–369.

Hendy MD, Penny D. 1989. A framework for the quantitative study of evolutionary trees. *Systematic Zoology* **38:** 297–309.

Higgins D, Thompson J, Gibson T. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22:** 4673–4680.

Hillis DM, Bull JJ. 1993. An empirical-test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology* **42:** 182–192.

Holland BR, Penny D, Hendy MD. 2003. Outgroup misplacement and phylogenetic inaccuracy under a molecular clock – a simulation study. *Systematic Biology* **52:** 229–238.

Huelsenbeck JP, Bollback JP, Levine AM. 2002. Inferring the root of a phylogenetic tree. *Systematic Biology* **51:** 32–43.

Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In: Munro HN, ed. *Mammalian protein metabolism*, Vol. 3. New York, NY: Academic Press, 22–132.

Kuhner MK, Felsenstein J. 1994. Simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular Biology and Evolution* **11:** 459–468.

Kumar S, Tamura K, Jakobsen IB, Nei M. 2001. MEGA2: molecular evolutionary genetics analysis software. *Bioinformatics* **17:** 1244–1245.

Li WH. 1997. *Molecular evolution*. Sunderland, MA: Sinauer Associates.

Maddison WP, Donoghue MJ, Maddison DR. 1984. Outgroup analysis and parsimony. *Systematic Zoology* **33:** 83–103.

Nei M. 1991. Relative efficiencies of different tree-making methods for molecular data. In: Miyamoto MM, Cracraft J, eds. *Phylogenetic analysis of DNA sequences*. New York, NY: Oxford University Press, 90–128.

Nei M, Kumar S. 2000. *Molecular evolution and phylogenetics*. New York, NY: Oxford University Press.

Overton LC, Rhoads DD. 2004. Molecular phylogenetic relationships based on mitochondrial and nuclear gene sequences for the Todies (Todus, Todidae) of the Caribbean. *Molecular Phylogenetics and Evolution* **32:** 524–538.

Qiu YL, Lee J, Whitlock BA, Bernasconi-Quadroni F, Dombrovska O. 2001. Was the ANITA rooting of the angiosperm phylogeny affected by long-branch attraction? Amborella, Nymphaeales, Illiciales, Trimeniaceae, and Austrobaileya. *Molecular Biology and Evolution* **18:** 1745–1753.

Rosenberg MS, Kumar S. 2001. Traditional phylogenetic reconstruction methods reconstruct shallow and deep evolutionary relationships equally well. *Molecular Biology and Evolution* **18:** 1823–1827.

Russo CAM. 1997. Efficiencies of different statistical tests in supporting a known vertebrate phylogeny. *Molecular Biology and Evolution* **14:** 1078–1080.

Russo CAM, Takezaki N, Nei M. 1996. Efficiencies of different genes and different tree-building methods in recovering a known vertebrate phylogeny. *Molecular Biology and Evolution* **13:** 525–536.

Saitou N, Nei M. 1987. The Neighbor-Joining method – a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* **4:** 406–425.

Sanderson MJ, Shaffer HB. 2002. Troubleshooting molecular phylogenetic analyses. *Annual Review of Ecology and Systematics* **33:** 49–72.

Sitnikova T, Rzhetsky A, Nei M. 1995. Interior-branch and bootstrap tests of phylogenetic trees. *Molecular Biology and Evolution* **12:** 319–333.

Stavrinides J, Guttman DS. 2004. Mosaic evolution of the severe acute respiratory syndrome coronavirus. *Journal of Virology* **78:** 76–82.

Swofford DL, Olsen GJ, Waddell PJ, Hillis DM. 1996. Phylogenetic inference. In: Hillis DM, Moritz C, Mable BK, eds. *Molecular systematics*. Sunderland, MA: Sinauer Associates, 407–514.

Tarrío R, Rodríguez-Trelles F, Ayala FJ. 2000. Tree rooting with outgroups when they differ in their nucleotide composition from the ingroup: the *Drosophila saltans* and *willistoni* groups, a case study. *Molecular Phylogenetics and Evolution* **16:** 344–349.

Wheeler WC. 1990. Nucleic acid sequence phylogeny and random outgroups. *Cladistics* **6:** 363–367.

Xia X, Xie Z. 2001. DAMBE: data analysis in molecular biology and evolution. *Journal of Heredity* **92:** 371–373.

SUPPLEMENTARY MATERIAL

The following material is available for this article online:

Table S1. General information (# of ingroups, outgroups, OTUs, Mean p-distance between ingroups and between in- and outgroups, and the category – see text) and the performance of the MPR method for each data set.

Figure S1. Trees are shown as rooted by the midpoint method, with trees collapsed at a 33% bootstrap cut-off. Bootstrap support values (only those greater than 50%) are shown above each branch. Grey circles indicate outgroup positions. A circle with a U indicates the single outgroup location, and thus is only present in single-outgroup data sets. In multiple-outgroup data sets, circles with numbers represent the insertion points of individual outgroups, whereas circles with a T indicate the root position as inferred by all combined outgroups when they agree. In cases where multiple combined outgroups disagree on root placement, the circled T is invalidated (as indicated with an X on the circle), and the arrows show where individual outgroups point when combined.

Appendix S1. References used in Supplementary Material.

This material is available as part of the online article from:
http://www.blackwell-synergy.com/doi/abs/10.1111/j.1095-8312.2007.00864.x
(This link will take you to the article abstract).

Please note: Blackwell Publishing are not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.