Molecular Biology and Evolution

The Prevalence of Multifurcations in Tree-space and Their Implications for Tree-search

Simon Whelan*[,1] and Daniel Money[1]

[1]Computational and Evolutionary Biology, Faculty of Life Sciences, University of Manchester, Manchester, United Kingdom

***Corresponding author:** E-mail: simon.whelan@manchester.ac.uk.

Associate editor: Alexei Drummond

Abstract

Phylogenetic tree-search is a major aspect of many evolutionary studies. Several tree rearrangement algorithms are available for tree-search, but it is hard to draw general conclusions about their relative performance because many effects are data set specific and can be highly dependent on individual implementations (e.g., RAxML or phyml). Using only the structure of the rearrangements proposed by the Nearest Neighbor Interchange (NNI) algorithm, we show tree-search can prematurely terminate if it encounters multifurcating trees. We validate the relevance of this result by demonstrating that in real data the majority of possible bifurcating trees potentially encountered during tree-search are actually multifurcations, which suggests NNI would be expected to perform poorly. We also show that the star-decomposition algorithm is a special case of two other popular tree-search algorithms, subtree pruning and regrafting (SPR) and tree bisection and reconnection (TBR), which means that these two algorithms can efficiently escape when they encounter multifurcations. We caution against the use of the NNI algorithm and for most applications we recommend the use of more robust tree-search algorithms, such as SPR and TBR.

Key words: phylogenetic tree-search, multifurcating trees, algorithms, maximum likelihood.

Phylogenetic trees are critical to many evolutionary studies and maximum likelihood (ML) has proved a popular and effective method of inference (e.g., Felsenstein 2003; Delsuc et al. 2005). The importance of choosing a realistic substitution model has regularly been demonstrated and inadequate models have been shown to result in long-branch attraction artifacts (e.g., Lartillot et al. 2007). Less attention has been focused on the relative quality and efficiency of the rearrangement algorithms used to find the best tree in tree-space (the set of all possible trees). The computational heuristics used in tree-search can only ever provide a "best guess" at the globally optimal ML tree because the only way to reliably identify it is to examine all tree space, which is impractical even for moderate numbers of taxa (e.g., Whelan 2007). Assessing the performance of tree-search algorithms is difficult because it is confounded with their implementation and depends on their relative ability to move through tree space, which is data specific. In this study, we demonstrate that the structure of the popular Nearest Neighbor Interchange (NNI) algorithm means that it may get stuck when it encounters multifurcating trees. We examine a range of real data and show the majority of tree-space tends to consist of multifurcating trees, suggesting that NNI is expected to perform poorly. Note, we do not investigate the impact of our finding on Bayesian inference, where multifurcations are known to affect posterior probabilities (Lewis et al. 2005; Yang 2007).

To explain the problem with NNI, we must first summarize how tree-search algorithms work and how trees are internally represented within programs. Tree-search under ML usually proceeds using the following hill-climbing strategy:

i. Initialize algorithm with a "focus tree" and calculate its likelihood.
ii. Use an algorithm (e.g., NNI) to find the neighbors of the focus tree. (Each neighbor is one possible step on the hillside.)
iii. Calculate the ML score of neighbors.
iv. If the best-scoring neighbor has a higher score than the focus tree: make it the focus tree and go to ii. (An uphill step.) Otherwise: end. (We are at the top of the hill.)

The algorithm in (ii) defines the number of steps between trees, and therefore the topography of tree space and the efficacy of the algorithm. Ideally, algorithms should guarantee that tree space consists of a single hill and all steps upward lead to the summit (the globally optimal ML tree). Real methods tend to produce many hills and it is impossible to identify whether the current hill is higher than other hills without examining all tree space. Most tree-search programs use one or more of a family of tree rearrangement algorithms consisting of NNI, subtree pruning and regrafting (SPR), and tree bisection and reconnection (TBR), each more general than the previous (see Felsenstein 2003; Whelan 2008).

This family of algorithms assumes the focus tree is bifurcating, whereby all branches are fully resolved and internal nodes have three branches leading from them. This assumption is reflected in tree-search programs, which internally use bifurcating trees. A problem arises during hill climbing when the focus tree contains multifurcations (polytomies), where some internal nodes have more than three branches leading from them. Programs represent these multifurcations as bifurcating trees with branches of (approximately) zero length, meaning many equally

http://mbe.oxfordjournals.org/

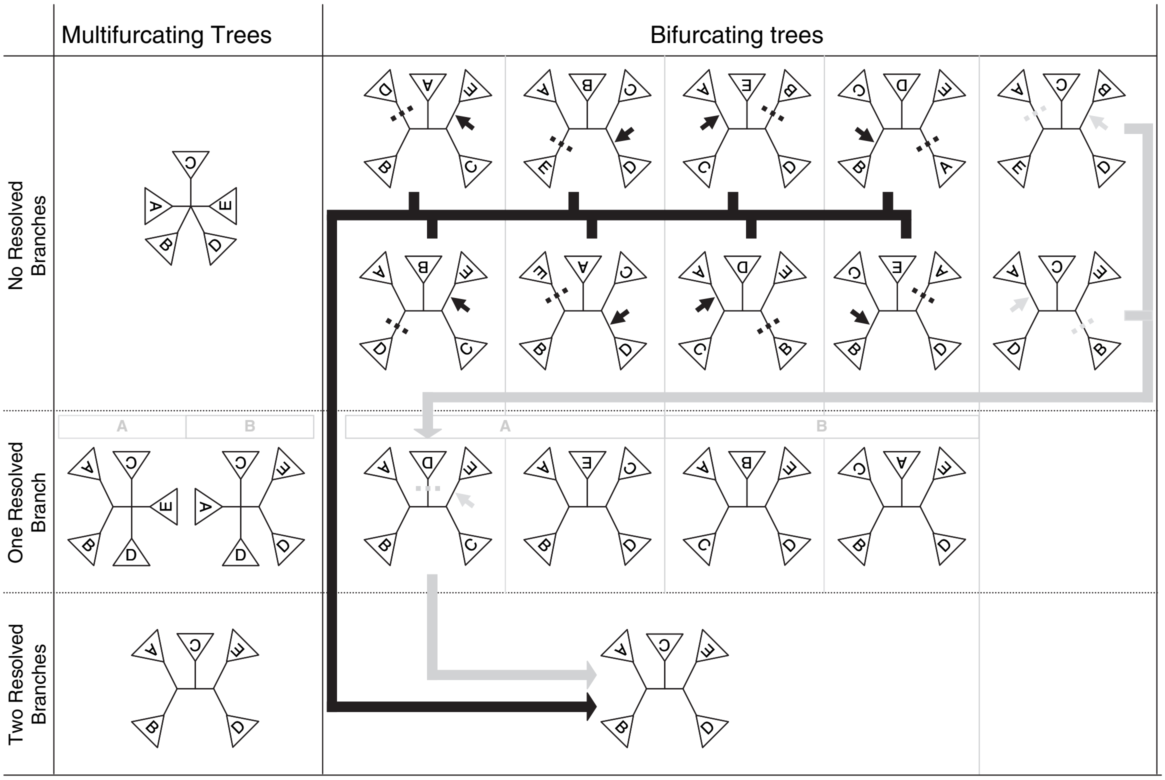

FIG. 1. The effect of a degree 5 multifurcation on NNI and SPR tree search. Bottom row represents the optimal tree relating the five subtrees, with the top and middle rows representing steps toward this tree. The left-hand column shows the steps that star decomposition takes when resolving the multifurcation. The right-hand column and subcolumns describe hill-climbing steps on multifurcating trees. NNI steps move down a row in a subcolumn, with each rearrangement representing the resolution toward a branch in the optimal tree. In the top row, the eight leftmost bifurcating trees can all resolve an internal branch and move down the column. NNI can get stuck at either of the two rightmost trees because they can only move to the four trees level with them in the no resolved branches row, none of which can provide an improvement in likelihood. The arrows show example SPR moves, with all nonresolved bifurcating trees able to reach the optimal tree via intermediates that increases likelihood, either by one step (black arrow) or by two steps (gray arrow; equivalent of star decomposition).

good bifurcating trees can describe a single multifurcating tree. In figure 1, we show the effect a degree 5 multifurcation can have on NNI, where 2/15 bifurcating representations of a multifurcation can get stuck and terminate prematurely during tree-search. This figure is illustrative and assumes only branches present in the resolved tree provide an improvement in likelihood. In real data, suboptimal branches may provide an improvement in likelihood. Supplementary fig. S1 (Supplementary Material online) online shows a clean example of NNI getting stuck during tree-search and the effect of suboptimal branches in multifurcations. Furthermore, we show in supplementary figure S2, Supplementary Material online that for larger multifurcations, the proportion of bifurcating trees that can get stuck increases; for a degree 8 multifurcation, around 53% of the 10,395 bifurcating representations get stuck.

Multifurcations are only a problem if they occur regularly in real data. We perform two investigations to ascertain how frequently problematic multifurcations occur during real examples of tree-search. (Full details are given in Supplementary Material online.) First, we take the 106 genes from eight yeast species from Rokas et al (2003), which have sufficiently few taxa to examine tree space exhaustively. For each gene and tree, we compute ML scores under the Jukes and Cantor (JC) and the general time reversible model with Γ-distributed rates across sites $(\mathrm{GTR} + \Gamma)$. To characterize tree space, we calculate: 1) the proportion of multifurcating trees (branch lengths of zero; tolerance of 10^{-3}); 2) the proportion of potentially problematic trees for NNI (trees with two or more adjacent zero length branches); and 3) the proportion of different starting trees where NNI gets stuck at a multifurcation during tree-search. We find all genes have some trees with zero length branches and that 99% (100%) of genes have potentially problematic trees under JC (and $\mathrm{GTR} + \Gamma$), although the total number of trees varies considerably between different genes and models (Supplementary Material online). When performing NNI hill climbing we find that JC rarely gets stuck, occurring in 1/106 genes and for only 0.5% of start trees. In contrast, NNI under $\mathrm{GTR} + \Gamma$ gets stuck in 75/106 genes, with on average 4% of starting trees being affected (range 0–48%).

The yeast data represent what should be an easy tree-search problem, but serve to demonstrate that NNI can function ineffectively even in simple cases. To investigate the prevalence of multifurcations in larger data sets, we take five nucleotide sequence alignments and five amino

Table 1. Numbers of Multifurcations in Real Data.

Data	Taxa	Length	Model	Average Number of Zero Length Branches	Proportion of Problematic Tree
Yeast	8	391–2,995	JC	0.4 (0.0–0.9)	0.03 (0.00–0.17)
			GTR + Γ	2.1 (1.5–2.8)	0.50 (0.26–0.72)
Zeng et al. (2006)	19	451	JC	8.2	1.00
			GTR + Γ	10.5	0.99
Hedges et al. (1990)	27	1,949	JC	19.9	1.00
			GTR + Γ	20.6	1.00
Coleman (2001)	38	799	JC	14.8	1.00
			GTR + Γ	23.9	1.00
Tartar et al. (2002)	41	887	JC	3.52	0.35
			GTR + Γ	17.8	1.00
Hoffman et al. (2007)	43	4,364	JC	2.46	0.21
			GTR + Γ	16.4	1.00
Susko et al. (2003)	13	269	EQU	2.4	0.41
			WAG + F + Γ	5.2	0.95
Ohkuma et al. (2007)	19	331	EQU	2.5	0.41
			WAG + F + Γ	6.4	0.97
Susko et al. (2003)	22	513	EQU	1.66	0.15
			WAG + F + Γ	7.8	0.98
Lartillot and Philippe (2006)	30	719	EQU	0.5	0.03
			WAG + F + Γ	5.8	0.68
Nozaki et al. (2009)	32	291	EQU	6.2	0.83
			WAG + F + Γ	13.6	1.00

acid sequence alignments from TreeBase (Morell 1996), intended to represent the type of data used in phylogenetic studies. For each data set, we randomly sample 100 trees, calculating ML scores under JC and GTR + Γ for nucleotide alignments and equiprobable (EQU) model and Whelan and Goldman (WAG) + F + Γ for amino acid alignments (Felsenstein 2003). Table 1 shows that multifurcating trees occupy the majority of tree space in 19/20 analyses, with the exception being the data from Lartillot and Philippe (2006) where 38% of tree space is multifurcating under EQU. Furthermore, in 14/20 of the analyses, the majority of tree space contained problematic trees, of which we expect a significant portion to affect NNI. We note that the proportion of multifurcating and problematic trees also tends to be higher for the more complex models (GTR + Γ and WAG + F + Γ), suggesting NNI may be less effective when more biologically realistic models are used for tree-search.

In contrast to NNI, the structure of SPR predicts that it should cope well with multifurcating trees. Figure 1 demonstrates that SPR efficiently resolves all degree 5 multifurcations. This occurs because SPR can always resolve a single internal branch by grouping together two leaves (subtrees). The resolved branch is removed from the multifurcation, reducing its degree by one, allowing SPR to iteratively resolve the multifurcation (fig. 1; gray arrows). This move is exactly that of star decomposition, with the upper bound of SPR distances between two trees being equal to the number of steps in the star-decomposition algorithm (Allen and Steel 2001, theorem 2.3). Frequently, SPR will do better than this upper-bound resolving multiple branches simultaneously (fig. 1; black arrows). The presence of this star-decomposition escape step, however, means that SPR will not get stuck at any degree of multifurcation during tree-search providing any individual pairing together of branches can provide an improvement in likelihood. It is not clear whether this requirement is ever violated in real data, but we expect it be much rarer than the problematic multifurcations for NNI. TBR is a generalization of SPR and will therefore also be able to efficiently escape multifurcations.

Our results show that the majority of tree space in typical phylogenetic studies consists of multifurcating trees, that these multifurcations can introduce irresolvable problems for NNI tree-search, and that NNI gets stuck at multifurcations even on simple eight taxa trees. We also show these problems tend to be more prevalent for more realistic substitution models. Based on our findings, we urge careful consideration of the algorithms used to perform tree-search. We caution against using NNI, unless for an individual data set there is evidence that tree space does not contain frequent problematic multifurcations, and recommend the use of algorithms that can successfully escape multifurcations, such as SPR or TBR.

Acknowledgment

D.M. is funded by a Biotechnology and Biological Sciences Research Council Doctoral Training Centre studentship.

References

Allen B, Steel M. 2001. Subtree transfer operations and their induced metrics on evolutionary trees. *Ann Combinatorics*. 5:13.

Coleman AW. 2001. Biogeography and speciation in the Pandorina/Volvulina (Chlorophyta) superclade. *J Phycol*. 37:836–851.

Delsuc F, Brinkmann H, Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet*. 6:361–375.

Felsenstein J. 2003. Inferring phylogenies. Sunderland (MA): Sinauer Associates.

Hedges SB, Moberg KD, Maxson LR. 1990. Tetrapod phylogeny inferred from 18s-ribosomal and 28s-ribosomal RNA sequences and a review of the evidence for amniote relationships. *Mol Biol Evol*. 7:607–633.

Hoffmann K, Discher S, Voigt K. 2007. Revision of the genus Absidia (Mucorales, Zygomycetes) based on physiological, phylogenetic, and morphological characters; thermotolerant Absidia spp. form a coherent group, Mycocladiaceae fam. nov. *Mycol Res*. 111: 1169–1183.

Lartillot N, Brinkmann H, Philippe H. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol Biol*. 7:S4.

Lartillot N, Philippe H. 2006. Computing Bayes factors using thermodynamic integration. *Syst Biol*. 55:195–207.

Lewis PO, Holder MT, Holsinger KE. 2005. Polytomies and Bayesian phylogenetic inference. *Syst Biol*. 54:241–253.

Morell V. 1996. TreeBASE: the roots of phylogeny. *Science* 273:569.

Nozaki H, Maruyama S, Matsuzaki M, Nakada T, Kato S, Misawa K. 2009. Phylogenetic positions of Glaucophyta, green plants (Archaeplastida) and Haptophyta (Chromalveolata) as deduced from slowly evolving nuclear genes. *Mol Phylogenet Evol*. 53: 872–880.

Ohkuma M, Saita K, Inoue T, Kudo T. 2007. Comparison of four protein phylogeny of parabasalian symbionts in termite guts. *Mol Phylogenet Evol*. 42:847–853.

Rokas A, Williams BL, King N, Carroll SB. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* 425:798–804.

Susko E, Field C, Blouin C, Roger AJ. 2003. Estimation of rates-across-sites distributions in phylogenetic substitution models. *Syst Biol*. 52:594–603.

Tartar A, Boucias DG, Adams BJ, Becnel JJ. 2002. Phylogenetic analysis identifies the invertebrate pathogen Helicosporidium sp. as a green alga (Chlorophyta). *Int J Syst Evol Microbiol*. 52:273–279.

Whelan S. 2007. New approaches to phylogenetic tree-search and their application to large numbers of protein alignments. *Syst Biol*. 56:727–740.

Whelan S. 2008. Inferring trees. In: Keith JM, editor. Bioinformatics. Totowa, NJ: Humana Press. p. 287–309.

Yang ZH. 2007. Fair-balance paradox, star-tree paradox, and Bayesian phylogenetics. *Mol Biol Evol*. 24:1639–1655.

Zeng LY, Jacobs MW, Swalla BJ. 2006. Coloniality has evolved once in stolidobranch ascidians. *Integr Comp Biol*. 46: 255–268.