nature ecology & evolution

ARTICLES

PUBLISHED: 19 JUNE 2017 | VOLUME: 1 | ARTICLE NUMBER: 0193

Phylogenetic rooting using minimal ancestor deviation

Fernando Domingues Kümmel Tria[1†], Giddy Landan[1†*] and Tal Dagan[1]

Ancestor–descendent relations play a cardinal role in evolutionary theory. Those relations are determined by rooting phylogenetic trees. Existing rooting methods are hampered by evolutionary rate heterogeneity or the unavailability of auxiliary phylogenetic information. Here we present a rooting approach, the minimal ancestor deviation (MAD) method, which accommodates heterotachy by using all pairwise topological and metric information in unrooted trees. We demonstrate the performance of the method, in comparison to existing rooting methods, by the analysis of phylogenies from eukaryotes and prokaryotes. MAD correctly recovers the known root of eukaryotes and uncovers evidence for the origin of cyanobacteria in the ocean. MAD is more robust and consistent than existing methods, provides measures of the root inference quality and is applicable to any tree with branch lengths.

Phylogenetic trees are used to describe and investigate the evolutionary relations between entities. A phylogenetic tree is an acyclic bifurcating graph, the topology of which is inferred from a comparison of the sampled entities. In the field of molecular evolution, phylogenetic trees are mostly reconstructed from DNA or protein sequences[1]. Other types of data have also been used to reconstruct phylogenetic trees, including phenotypic characteristics of species, biochemical makeup, as well as language vocabularies (for a historical review, see ref. [2]). In most tree reconstruction methods the inferred phylogeny is unrooted, and the ancestral relations between the taxonomic units are not resolved. The determination of ancestor–descendant relations in an unrooted tree is achieved by the inference of a root node, which a priori can be located on any of the branches of the unrooted tree. The root represents the last common ancestor (LCA) from which all operational taxonomic units (OTUs) in the tree descend.

Several root inference methods have been described in the literature, differing in the type of data that can be analysed, the assumptions of the evolutionary dynamics of the data, and their scalability or general applicability. The most commonly used method is the outgroup approach, where OTUs that are assumed to have diverged earlier than the LCA are added to the tree reconstruction procedure[3]. The branch connecting the outgroup to the OTUs of interest—termed ingroup—is assumed to contain the root. Because the ingroup is assumed to be monophyletic in the resulting phylogeny, the choice of an outgroup requires prior knowledge about the phylogenetic relations between the outgroup and the ingroup. Therefore, a wrong assumption regarding the outgroup phylogeny will inevitably lead to an erroneous rooted topology. Another approach, midpoint rooting, assumes a constant evolutionary rate (that is, clock-like evolution) along all lineages, an assumption that in its strongest form, ultrametricity, equates branch lengths with absolute time[4]. In midpoint rooting the path length between all OTU pairs is calculated by the summation of the lengths of the intervening branches, and the root is placed at the middle of the longest path. Midpoint rooting is expected to fail when the requirement for clock-like evolution is violated. Both outgroup and midpoint rooting can be applied independently of the tree reconstruction algorithm or the underlying type of data, with very little computational overhead. For molecular sequences and other character state data, two additional rooting methods include the root position as part of the probabilistic evolutionary models used to infer the tree topology, but at the cost of substantial increase in complexity. In the relaxed clock model approach, the evolutionary rate is allowed to vary among lineages, and the root position is optimized to produce an approximately equal time span between the LCA and all descendants[5]. In the non-reversible model approach the transition probabilities are asymmetric and require a specification of the ancestor–descendant relation for each branch[6]. Again, the root position is optimized to maximize the likelihood of the data. Presently, both probabilistic approaches entail a substantially larger computational cost relative to the inference of unrooted trees by similar probabilistic methods. Given the fundamental role of ancestor–descendant relations in evolutionary theory, the absence of generally applicable and robust rooting methods is notable. This is in stark contrast to the wide range of methods available for the reconstruction of phylogenetic tree topologies.

Here we introduce a new rooting method—the minimal ancestor deviation (MAD) method. MAD rooting operates on unrooted trees of contemporaneous OTUs, with branch lengths as produced by any tree reconstruction algorithm, based on any type of data, and is scalable for large datasets. No outgroup or other prior phylogenetic knowledge is required. While grounded in clock-like reasoning, it quantifies departures from clock-likeness rather than assuming it, making it robust to variation in evolutionary rates among lineages. We assessed the performance of MAD rooting in three biological datasets, one including species from the eukaryotic domain and two prokaryotic datasets of species from the cyanobacteria and proteobacteria phyla. We demonstrate that in the investigated cases, MAD root inference is superior to those of the outgroup, midpoint and the relaxed molecular clock rooting methods.

Results

Algorithm. The MAD method operates on binary unrooted trees and assumes that branch lengths are additive and that OTUs are contemporaneous. MAD estimates the root position by considering all branches as possible root positions, and evaluating the resulting ancestral relationships between nodes.

Genomic Microbiology Group, Institute of General Microbiology, Kiel University, Kiel 24118, Germany. †These authors contributed equally to this work. *e-mail: giddy.landan@gmail.com

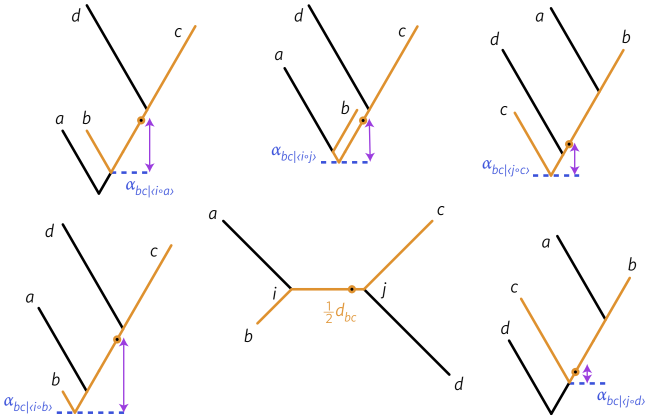
d
c
a
b
$\alpha_{bc|\langle i\circ a\rangle}$
d
a
c
b
$\alpha_{bc|\langle i\circ j\rangle}$
a
d
b
c
$\alpha_{bc|\langle j\circ c\rangle}$
d
c
a
b
$\alpha_{bc|\langle i\circ b\rangle}$
a
c
i
j
b
d
$\frac{1}{2}d_{bc}$
a
b
c
d
$\alpha_{bc|\langle j\circ d\rangle}$

Figure 1 | Schematic illustration of rooting unrooted trees. A four-OTU unrooted tree (bottom centre) and the five rooted trees resulting from placing the root on each of the five branches. Orange lines mark the path between OTUs *b* and *c*, and its midpoint is marked by a dot. A blue dashed line and an *α* mark the ancestor nodes of the OTU pair as induced by the various root positions. Purple arrows mark the deviations between the midpoint and the ancestor nodes.

Before describing the algorithm, let us first define the main features of the problem (Fig. 1). A rooted tree differs from its unrooted version by a single node, the root node, which is the LCA of all the OTUs considered, while internal nodes represent ancestors of partial sets of OTUs. In an n OTUs unrooted tree, one can hypothesize the root node residing in any of the $2n - 3$ branches. Once a branch is selected as containing the root, the ancestral relationships of all nodes in the tree are determined. Note, however, that prior to rooting, ancestral relations are unresolved, and that different root positions can invert the ancestral relations of specific internal nodes.

Under a strict molecular clock assumption (that is, ultrametricity), the midpoint criterion asserts that the middle of the path between any two OTUs should coincide with their LCA. In practice, strict ultrametricity seldom holds, and the midpoint deviates from the actual position of the ancestor node (Fig. 1). The MAD algorithm evaluates the deviations of the midpoint criterion for all possible root positions and all $n(n - 1) / 2$ OTU pairs of the unrooted tree.

Our method estimates the root by: (a) considering each branch separately as a possible root position; (b) deriving the induced ancestor–descendant relationships of all the nodes in the tree; (c) calculating the mean relative deviation from the molecular clock expectation that is associated with the root positioned on the branch. The branch that minimizes the relative deviations is the best candidate to contain the root node.

Let d_{ij} be the distance between nodes i and j. For two OTUs b and c, and an ancestor node α, the distances to the ancestor are $d_{\alpha b}$ and $d_{\alpha c}$, while the midpoint criterion asserts that both should be equal to $\frac{d_{bc}}{2}$. The pairwise relative deviation is then defined as:

$$r_{bc,\alpha} = \left| \frac{2d_{\alpha b}}{d_{bc}} - 1 \right| = \left| \frac{2d_{\alpha c}}{d_{bc}} - 1 \right|$$

(Fig. 1; see the Methods for the complete derivation).

For a putative root in a branch $\langle i \circ j \rangle$ connecting adjacent nodes i and j of the unrooted phylogeny, we define the branch ancestor deviation, $r_{\langle i\circ j\rangle}$, as the root-mean-square of the pairwise relative deviations:

$$r_{\langle i\circ j\rangle} = \left(\overline{r^2_{bc,\alpha}} \right)^{\frac{1}{2}}$$

Branch ancestor deviations take values on the unit interval, with a zero value for exact correspondence of midpoints and ancestors for all OTU pairs, a circumstance attained only by the roots of ultrametric trees.

Branch ancestor deviations quantify the departure from strict clock-like behaviour, reflecting the level of rate heterogeneity among lineages. Wrong positioning of the root will lead to erroneous identification of ancestor nodes, and apparent deviations will tend to be larger. We therefore infer the MAD root as the branch and position that minimizes the ancestor deviation $r_{\langle i\circ j\rangle}$.

We illustrate MAD rooting in Fig. 2a, employing the example of an unrooted tree for 31 eukaryotic species. The minimal ancestor deviation root position is located on the branch separating fungi from metazoa. In this example, existing rooting methods place the inferred root on other branches (Fig. 2b). Moreover, MAD rooting provides explicit values for all branches, thereby describing the full context of the inference. Different definitions of the deviations and averaging strategies give rise to additional MAD variants, described in the Methods.

Performance. We first consider the performance of the proposed MAD method in comparison to other rooting methods in the context of eukaryotic phylogeny. For eukaryotic sequences, we expect uncertainties in root inferences to be mainly owing to methodological or sampling causes rather than biological ones (for example, reticulated evolution). We examined 1,446 trees, which were reconstructed from protein sequences of universal orthologues in 31 opisthokonta species. The root is known to lie between fungi and metazoa[7,8], thereby giving us a clear target for the correct rooted topology. We infer root positions using the MAD method, the traditional midpoint rooting method and the outgroup approach, using ten plant species as the outgroup, all based on maximum-likelihood trees using PhyML[9], while relaxed molecular clock rooting was inferred using MrBayes[10].

The four methods recover the fungi–metazoa branch as the most common inferred root position (Fig. 3a and Supplementary Table 1). The MAD method identifies the correct root in 72% of the trees. The midpoint method is less consistent (61%), followed by the outgroup method (57%). The outgroup method could not be applied on 21% of the gene families, either owing to the absence of plant homologues or because of multiple outgroup clusters (Supplementary Table 2). The relaxed molecular clock method identifies the fungi–metazoa branch as the root in 36% of the trees and a neighbouring branch in 34% of the trees. Neighbouring branches are also found as the second most common root position in the other methods, but with much smaller frequencies (Fig. 3a). The eukaryotic dataset serves as a positive control, and it demonstrates that the MAD method is accurate and consistently outperforms the existing rooting methods (see also Supplementary Tables 1, 2 and Supplementary Fig. 1).

Rooting microbial phylogenies is more challenging because of the possibility of reticulated, non tree-like, signals[11]. We consider the case of 130 cyanobacterial species with trees from 172 universal orthologues, using *Gloeobacter violaceus* as an outgroup. *G. violaceus,* a cyanobacterium itself, is assumed to be basal[12] and serves as the traditional outgroup for other cyanobacteria (for example, see ref. [13]). The MAD approach positions the most common root in the branch that separates a *Synechococcus–Prochlorococcus–Cyanobium* (SynProCya) clade from the remaining species, with support from 70% of the trees (Fig. 3b and Supplementary Table 1). The midpoint method detects the same root position with a consistency of 54%. These values are only slightly smaller than those encountered in the eukaryotic dataset, demonstrating the robustness of MAD rooting even for much deeper phylogenetic relations and possible lateral gene transfer. The second most common root position appears in only 9% of the trees, on a neighbouring branch

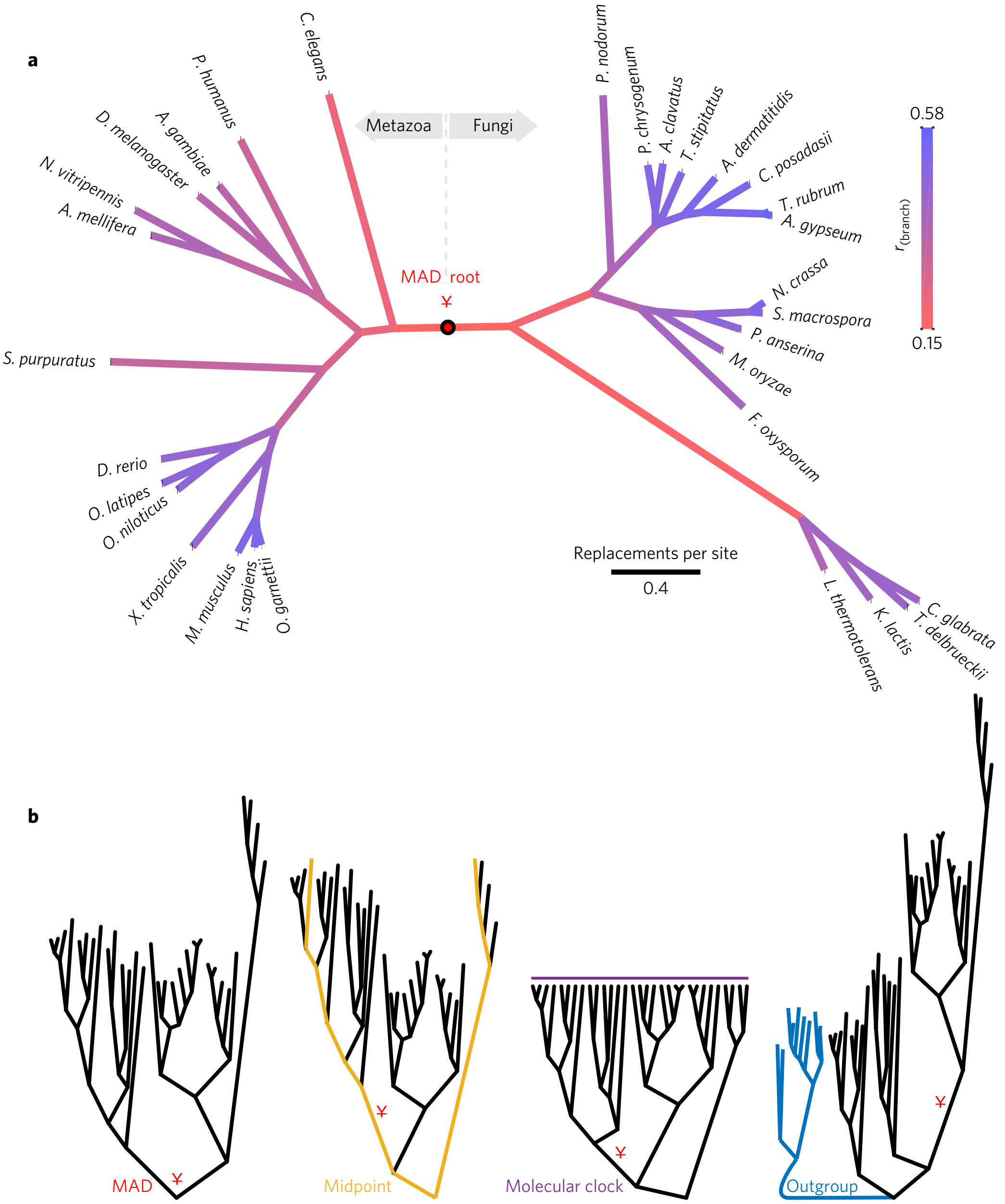

Figure 2 | MAD rooting illustrated with a eukaryotic protein phylogeny. a, An unrooted maximum-likelihood tree of trans-2-enoyl-CoA reductase protein sequences from 14 Metazoa and 17 Fungi species. Full species names are provided in Supplementary Table 3. Branch colours correspond to their ancestor relative deviation $r_{(i\circ j)}$ value. The inferred root position is marked by a black circle and a red symbol (¥). **b**, Rooted phylogenies using four alternative rooting methods, the correct root position is marked by a red symbol (¥). The longest path of the midpoint method is marked in yellow. The molecular clock enforces ultrametricity (purple line). The ten plant outgroup OTUs are marked in blue.

that joins two *Synochococcus elongatus* strains into the SynProCya clade. The Bayesian relaxed clock models support a neighbouring branch that excludes one *Synechococcus* strain from the SynProCya clade in about 15% of the trees and produce unresolved topologies in the root position for 28% of the trees. Using *G. violaceus* as an outgroup produced a unique result by pointing to a branch separating three thermophilic *Synechococcus* strains from the rest of the phylum. This result, which is at odds with all other methods, may well stem from a wrong phylogenetic presumption of *G. violaceus* being an adequate outgroup. Using alternative outgroup species, we find variable support for the two competing root inferences, albeit always with low consistency (Supplementary Tables 1,2).

A more difficult rooting problem is encountered when considering highly diverse phyla. Proteobacteria groups together six taxonomic classes including species with diverse lifestyles and variable trophic strategies. We analysed 130 universal gene families in 72 proteobacteria, using seven Firmicutes species as the outgroup. The MAD method produces the highest consistency, albeit at a support level of 17%, which is much lower than for the previous datasets (Fig. 3c and Supplementary Table 1). The best root position is found on the branch separating epsilonproteobacteria from the remaining classes. The second most frequent branch occurs in 14% of the trees, and the third branch in yet another 8%. All three branches occur next to each other with the second most common branch separating alphaproteobacteria from the other classes, and the third branch joining deltaproteobacteria to the epsilonproteobacteria. These three branches are also the most frequent root branches inferred using the midpoint approach. The relaxed molecular clock approach

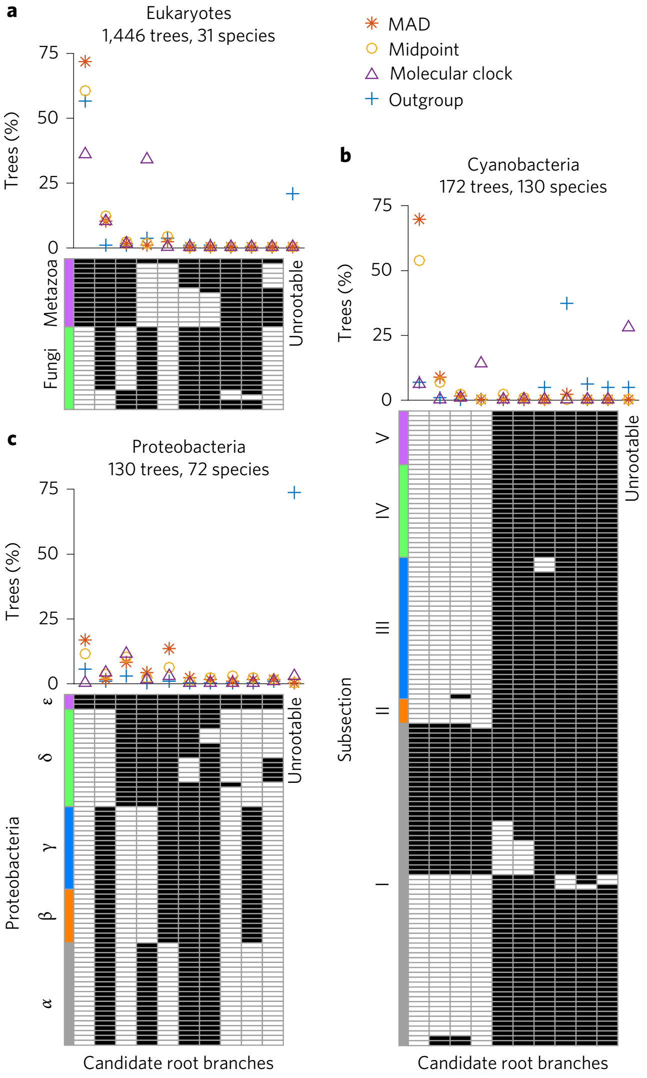

Figure 3 | Root inference by four rooting methods in three datasets. Methods compared are MAD, midpoint, outgroup and molecular clock rooting. **a**–**c**, Rooting of universal protein families are summarized for Eukaryotes (**a**), Cyanobacteria (**b**) and Proteobacteria (**c**). For a complete list of the species that were used, see Supplementary Table 3. Bottom, root branches are reported as OTU splits (black and white checkered columns). The ten most frequently inferred root branches are presented (combined over the four methods). The major taxonomic groups for each dataset are indicated in colour. Top, The percentage of trees with the inferred root positioned in the respective branch for each of the four methods. Rightmost position reports the proportion of unrootable trees (that is, no outgroup orthologues, outgroup OTUs are paraphyletic or unresolved root topology).

most frequently infers only one of these branches as the root, the branch that separates the epsilonproteobacteria and the deltaproteobacteria from the remaining classes. We note that the outgroup approach proved to be inapplicable for this dataset in 74% of the universal gene families.

Why does the MAD approach yield less consistent results for the proteobacteria dataset? One possibility is that this dataset presents an extreme departure from clock-likeness. We evaluate the deviations from clock-likeness of each tree, given the inferred MAD root position, by the coefficient of variation of the distances from the root to each of the OTUs (R_{CCV}, see Methods). The eukaryotic dataset presents the highest level of clock-likeness, but the cyanobacterial dataset—where a consistent root branch is found—presents an even greater departure from clock-likeness than the proteobacteria dataset (Fig. 4a). This shows that the lower consistency is not due to heterotachy alone and that MAD is fairly robust to departures from clock-likeness. The low support observed in proteobacteria is due to three competing branches that together account for 39% of the root inferences. This circumstance is best described as a 'root neighbourhood' rather than a definite root position. To detect competing root positions for a given tree, we define the root ambiguity index, R_{AI}, as the ratio of the MAD value to the second smallest value (see Methods). This ratio will attain the value 1 for ties, that is, two or more root positions with equal deviations, and smaller values in proportion to the relative quality of the best root position. Indeed, comparing the datasets by the distribution of the ambiguity index clearly shows that the eukaryotic dataset is the least ambiguous, whereas most of the trees in the proteobacteria dataset yield very high ambiguity scores (Fig. 4b).

The ambiguity observed can originate from several factors. One source of ambiguity can be due to very close candidate root positions in the tree. This situation would become more acute when the root branch is short and root positions on neighbouring branches can yield comparable ancestor deviation values. Indeed, we find a significant negative correlation between the ambiguity index and the length of the root branch (normalized by tree size, Spearman $\rho = -0.53$; $P = 1.0 \times 10^{-10}$). In other words, short root branches are harder to detect.

Discussion

Our results demonstrate that MAD rooting can outperform previously described rooting methods. Moreover, MAD operates on bifurcating trees with branch lengths, thus it is not dependent on the type of data that underlie the analysis, the tree reconstruction method or the evolutionary models. MAD is also scalable; the running time of MAD is comparable to distance-based tree reconstruction methods. In addition, MAD does not depend on prior phylogenetic knowledge of the outgroup species or on the availability of orthologous sequences of outgroups.

The inferred MAD root for the cyanobacteria phylum implies that the LCA of cyanobacteria was a unicellular organism that lived in a marine environment. This suggests that the basic photosynthesis machinery originated in a marine environment, which contrasts with our earlier conclusions that were based on using *Gloeaobacter* sp. as outgroup[14]. Alternative outgroups reproduce the MAD rooting, albeit with lower support. The cyanobacteria dataset shows that the MAD approach is robust to phylogenetic inference errors and possible lateral gene transfer.

We introduce the concept of 'root neighbourhood' to enable the interpretation of ancestral relations in trees even in the absence of an unambiguous root position. A root neighbourhood can be observed in the proteobacterial dataset, where all highly supported root positions maintain the monophyly of proteobacteria classes. The quantification of ambiguity in root inference is made possible by the evaluation of every branch as a possible root and the comparable magnitude of the ancestor deviation statistic. Thus, the MAD approach supplies a set of statistics that are intrinsically normalized, and are directly comparable between different trees. This enables phylogenomic level application, with implications for the resolution of long standing species–tree conundrums. We note, however, that MAD can infer roots in any type of tree, including trees that differ from the species tree (owing to paralogy or lateral gene transfer, for example).

Midpoint rooting is the ultimate ancestor of the MAD approach. Three elements are new to the MAD formulation. First, the various topological pairings of midpoints to ancestor nodes; second, the exhaustive utilization of metric information from all OTU pairs (instead of just the longest path) and all possible root positions; and finally, heterotachy is embraced and explicitly quantified.

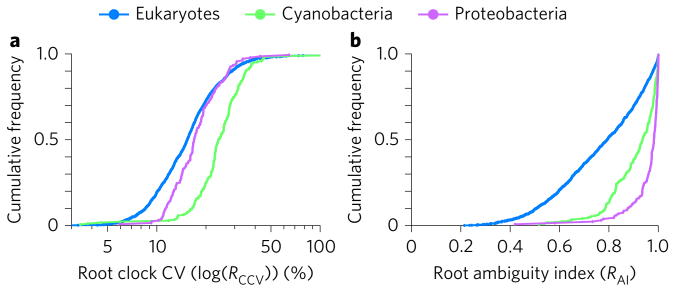

Figure 4 | MAD root clock-likeness and ambiguity statistics in the three datasets. a, Comparison of R_{CCV} distributions, which quantifies the deviation from clock-likeness, or heterotachy, associated with MAD root positions in individual trees. CV, coefficient of variation. **b**, Comparison of the ambiguity index, R_{AI}, distributions for MAD root inferences.

Rate heterogeneity among lineages is a real phenomenon stemming from variability in the determinants of evolutionary rates: mutation rates, population dynamics and selective regimes. Thus, it is unrealistic to either assume a molecular clock or to force one by constraining the evolutionary model. The actual levels of heterotachy may appear to be even larger when a wrong position for the root is hypothesized. It is these spurious deviations that are minimized by the MAD method to infer the root position. Withstanding heterotachy is further assisted by the consideration of all OTU pairs and root positions, because lineages with exceptional rates contribute large deviations uniformly to all possible root positions.

In conclusion, MAD holds promise for application also in other fields that rely on evolutionary trees, such as epidemiology and linguistics. MAD rooting provides robust estimates of ancestral relations, the bedrock of evolutionary research.

Methods

Detailed algorithm. In an n OTUs unrooted tree, let d_{ij} be the distance between nodes i and j, calculated as the sum of branch lengths along the path connecting nodes i and j, and thus additive by construction. For simpler exposition we will assume all branches to have a strictly positive length (that is, $d_{ij} > 0\ \forall i \neq j$). For two OTUs b and c, and a putative ancestor node α, the expected distances to the ancestor are $d_{\alpha b}$ and $d_{\alpha c}$, while the midpoint criterion asserts that both should be equal to

$$\frac{d_{bc}}{2}$$

The resulting deviations are

$$\left|d_{\alpha b}-\frac{d_{bc}}{2}\right|=\left|d_{\alpha c}-\frac{d_{bc}}{2}\right|$$

(see Fig. 1). To be able to summarize all OTU pairs on equal footing, we prefer to consider the deviations relative to the pairwise distance d_{bc}, and define the relative deviation as:

$$r_{bc,\alpha}=\left|\frac{2d_{\alpha b}}{d_{bc}}-1\right|=\left|\frac{2d_{\alpha c}}{d_{bc}}-1\right| \qquad (1)$$

which take values on the unit interval, regardless of the magnitude of d_{bc}.

In order to compare ancestor nodes to midpoints for all pairs of OTUs, we first need to identify the LCA of each OTU pair as induced by a candidate root branch. For a branch $\langle i \circ j\rangle$ connecting adjacent nodes i and j, we define the OTU partition $\langle I \circ J\rangle$, as:

$$I=\{\text{terminal node } k: d_{ki}<d_{kj}\}, \quad J=\{\text{terminal node } k: k \notin I\}$$

For any two OTUs lying on the same side of the putative root branch the ancestor is already present as a node in the unrooted tree, and can be identified by:

$$\alpha_{bc|\langle i\circ j\rangle}=k:\{d_{bc}=d_{ib}+d_{ic}-2d_{ik}\} \quad \textit{where} \quad \begin{array}{l} b,c\in I; \\ k\,\textit{a node on the path from } i \textit{ to } b \end{array}$$

and similarly for b, $c \in J$.

For OTU pairs straddling the candidate root branch, $b \in I$, $c \in J$, we first need to introduce a hypothetical ancestor node $o_{\langle i\circ j\rangle}$ with minimal deviations from the midpoints of straddling OTU pairs. Consider all possible positions $o(\rho)$ as parameterized by the relative position ρ, then $d_{io(\rho)} = \rho d_{ij}$ and $d_{jo(\rho)} = (1 - \rho)d_{ij}$, and the sum of squared relative deviations is:

$$r(\rho)=\sum_{b\in I}\sum_{c\in J}\left(\frac{2d_{bo(\rho)}}{d_{bc}}-1\right)^2=\sum_{b\in I}\sum_{c\in J}\left(\frac{2(d_{bi}+\rho d_{ij})}{d_{bc}}-1\right)^2$$

which is minimized by:

$$\rho=\sum_{b\in I}\sum_{c\in J}(d_{bc}-2d_{bi})d_{bc}^{-2} \Big/ \left(2d_{ij}\sum_{b\in I}\sum_{c\in J}d_{bc}^{-2}\right) \qquad (2)$$

Because the minimizing relative position may fall outside the branch, we constrain it to the unit interval:

$$\rho_{\langle i\circ j\rangle}= \min(\max(0,\rho),1)$$

and the position of the node $o_{\langle i\circ j\rangle}$ is given by:

$$d_{io_{\langle i\circ j\rangle}}=\rho_{\langle i\circ j\rangle}d_{ij} \text{ and } d_{jo_{\langle i\circ j\rangle}}=(1-\rho_{\langle i\circ j\rangle})d_{ij}$$

The hypothetical node $o_{\langle i\circ j\rangle}$ serves as the ancestor induced by the branch for all OTU pairs straddling it: $\alpha_{bc|\langle i\circ j\rangle}=o_{\langle i\circ j\rangle}$, $b \in I$, $c \in J$.

For each branch we combine deviations of all OTU pairs into the branch ancestor deviation score, which is defined as the root-mean-square of the relative deviations:

$$r_{\langle i\circ j\rangle}=\left(\overline{r_{bc,\alpha}^2}\right)^{\frac{1}{2}} \qquad (3)$$

where $\alpha=\alpha_{bc|\langle i\circ j\rangle}$ and $b,c\in I\cup J$.

Again, $r_{\langle i\circ j\rangle}$ takes values on the unit interval, with a zero value for exact correspondence of midpoints and ancestors for all OTU pairs, a condition attained only by the root nodes of ultrametric trees.

Next, we compute the ancestor deviation score for all branches. We note that the minimization equation (2), while given as an analytical point solution, can be viewed as a scan of every point in a branch. When applied to all the branches, this amounts to an exhaustive evaluation of all points in the unrooted phylogeny.

Finally, MAD infers the root of the tree as residing on the branch(es) with the minimal induced ancestor deviation. Let $\{\beta_1 \cdots \beta_{2n-3}\}$ be the set of branches sorted by their ancestor deviation statistic $r_{\langle\beta\rangle}$, then the root branch is β_1 and the inferred root node is:

$$^{\text{MAD}}R=o_{\langle\beta_1\rangle}$$

with a position as defined in equation (2).

Formally, the minimal value can be attained by more than one branch, but in practice ties are very rare (not one tie in the 1,748 trees analysed here). Close competition, however, is common and can be quantified by the root ambiguity index:

$$R_{\text{AI}}=\frac{r_{\langle\beta_1\rangle}}{r_{\langle\beta_2\rangle}}$$

which take the value 1 for ties, and smaller values with increasing separation between the minimal ancestor deviation value to the second smallest value.

Since the MAD method evaluates departures from ultrametricity, it is useful to quantify the clock-likeness of the inferred root position. We define the root clock coefficient of variation (CV) as:

$$R_{\text{CCV}}=\text{CV}\left(d_{o_{\langle\beta_1\rangle}b}\right) \qquad (4)$$

with $b\in\{1\cdots n\}$ OTUs.

Several elements in the preceding formulation can be modified to yield slightly different variants of MAD. We evaluated the following variants and their several combinations:

A is the definition of the pairwise deviation:

A1 is the relative deviation, see equation (1) and equation (2);

A2 is the absolute deviation, not normalized by the pairwise distance d_{bc}, with

$$r_{bc,\alpha}=\left|d_{\alpha b}-\frac{d_{bc}}{2}\right|=\left|d_{\alpha c}-\frac{d_{bc}}{2}\right| \quad \text{and} \quad \rho=\sum_{b\in I}\sum_{c\in J}(d_{bc}-2d_{bi})\,/\,(2d_{ij}\cdot|I|\cdot|J|)$$

replacing equation (1) and equation (2).

B is the averaging of the squared pairwise deviations:

B1 is a simple mean of all $n(n-1)/2$ squared deviations, equation (3);

B2 averaging occurs separately at each ancestor node for all pairs straddling it. The final score is taken as the mean of the $(n-1)$ ancestor values.

Yet other rooting variants within the conceptual framework of MAD are produced by ignoring the magnitude of deviations. For the 'minimal clock coefficient-of-variation' (MCCV) method, hypothetical ancestor nodes $o_{\langle i\circ j\rangle}$ are retained and the resulting variation in clock-likeness, similarly to equation (4), is used as the branch score. Again, the branch minimizing the score is selected as the inferred root branch. For the 'pairwise midpoint rooting' (PMR) variant, we omit even $o_{\langle i\circ j\rangle}$, and enumerate all pairwise paths that traverse a given branch. The branch score is then the percentage of paths with midpoints falling within the branch:

$$D_{io} = \{d_{io|bc} : 0 \le d_{io|bc} \le d_{ij}\}$$

where

$$d_{io|bc} = \frac{d_{bc}}{2} - d_{ib},\ \ b \in I, c \in J$$

and

$$\mathrm{PMR}_{\langle i\circ j\rangle} = \frac{|D_{io}|}{|I|\cdot|J|}$$

In this variant, the branch maximizing the score is the inferred root branch. Essentially, the PMR is the simplest extension of the midpoint rooting method to integrate the information from all pairwise paths.

The performances of the PMR method, the MCCV method, and of the four combinations of variants A and B are reported in Supplementary Table 1.

Data. Universal protein families for the eukaryotic and proteobacteria datasets were extracted from EggNOG version 4.5 (ref. [15]). The cyanobacteria protein families were constructed from completely sequenced genomes available in the RefSeq database[16] (version May 2016), except the Melainabacteria Zag 1 genome, which was downloaded from IMG[17]. Species in the three datasets were selected from the available genomes so that the number of represented taxa were as large as possible and genus-level redundancy was reduced. The datasets are: Eukaryotes (31 opisthokonta with 10 outgroup plant species), Proteobacteria (72 species with 7 outgroup Firmicutes species), and Cyanobateria (130 species with 6 outgroup bacterial species) (See Supplementary Table 3 for the complete list of species). Outgroup species were selected according to the accepted taxonomic knowledge. EggNOG clusters with complete ingroup species-set representation were extracted, resulting in 1,446 eukaryotic protein families and 130 proteobacterial protein families. For the construction of cyanobacteria protein families, at the first stage, all protein sequences annotated in the genomes were blasted all-against-all using stand-alone BLAST[18] version 2.2.26. Protein sequence pairs that were found as reciprocal best BLAST hits[19] with a threshold of $E \le 1\times10^{-5}$ were further compared by global alignment using needle[20]. Sequence pairs that had ≥30% identical amino acids were clustered into protein families using the Markov clustering algorithm (MCL)[21] version 12-135 with the default parameters. Protein families with complete ingroup species-set representation were retained, resulting in 172 cyanobacterial protein families.

Because we are only interested in universal families of orthologues in this study, we sorted out the paralogues from the protein families as previously described in ref. [22]. Of the universal protein families, 1,339 eukaryotic, 85 proteobacterial and 64 cyanobacterial families contained paralogous sequences, and were condensed as follows. Sequences of the protein families were aligned using MAFFT version 7.027b[23] with the L-INS-i alignment strategy, and the percentage of identical amino acids between all sequence pairs was calculated. Next we clustered the sequences by amino acid identity using the single-linkage algorithm, and the largest cluster with at most a single sequence for each species was selected as a seed. Species not represented in the seed cluster were included by the addition of the sequence with the maximal median identity to the seed cluster.

Protein sequences of the resulting universal protein families were aligned using MAFFT version 7.027b with the L-INS-i alignment strategy. Phylogenetic trees were reconstructed using PhyML version 20120412 (ref. [9]) with the following parameters: -b -4 -v e -m LG -c 4 -s SPR. MAD rooting and midpoint rooting were performed using in-house MATLAB scripts. Molecular clock roots were inferred from phylogenies reconstructed with MrBayes version 3.2.3 (ref. 10) with the following parameters: lset rates = invgamma ngammacat = 4; prset aamodelpr = fixed(wag) brlenspr = clock:uniform clockvarpr = igr; sumt contype = allcompat. Outgroup rooting was inferred from PhyML trees reconstructed from independent MAFFT alignments that include the outgroup sequences.

Code availability. Implementations of MAD in R and in MATLAB used in this study are available through our institutional website at: https://www.mikrobio.uni-kiel.de/de/ag-dagan/ressourcen.

Data availability. The datasets used in this study are available through our institutional website at: https://www.mikrobio.uni-kiel.de/de/ag-dagan/ressourcen.

Received 26 January 2017; accepted 16 May 2017; published 19 June 2017

References

1. Fitch, W. M. & Margoliash, E. Construction of phylogenetic trees. *Science* **155,** 279–284 (1967).
2. Ragan, M. A. Trees and networks before and after Darwin. *Biol. Direct* **4,** 43 (2009).
3. Kluge, A. G. & Farris, J. S. Quantitative phyletics and the evolution of anurans. *Syst. Biol.* **18,** 1–32 (1969).
4. Farris, J. S. Estimating phylogenetic trees from distance matrices. *Am. Nat.* **106,** 645–668 (1972).
5. Lepage, T., Bryant, D., Philippe, H. & Lartillot, N. A general comparison of relaxed molecular clock models. *Mol. Biol. Evol.* **24,** 2669–2680 (2007).
6. Williams, T. A. *et al.* New substitution models for rooting phylogenetic trees. *Phil. Trans. R. Soc. B* **370,** 20140336 (2015).
7. Stechmann, A. & Cavalier-Smith, T. Rooting the eukaryote tree by using a derived gene fusion. *Science* **297,** 89–91 (2002).
8. Katz, L. A., Grant, J. R., Parfrey, L. W. & Burleigh, J. G. Turning the crown upside down: gene tree parsimony roots the eukaryotic tree of life. *Syst. Biol.* **61,** 653–660 (2012).
9. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59,** 307–321 (2010).
10. Ronquist, F. *et al.* MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **61,** 539–542 (2012).
11. Bapteste, E. *et al.* Prokaryotic evolution and the tree of life are two different things. *Biol. Direct* **4,** 34 (2009).
12. Turner, S., Pryer, K. M., Miao, V. P. W. & Palmer, J. D. Investigating deep phylogenetic relationships among cyanobacteria and plastids by small subunit rRNA sequence analysis. *J. Eukaryot. Microbiol.* **46,** 327–338 (1999).
13. Shih, P. M. *et al.* Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. *Proc. Natl Acad. Sci. USA* **110,** 1053–1058 (2013).
14. Dagan, T. *et al.* Genomes of stigonematalean cyanobacteria (subsection V) and the evolution of oxygenic photosynthesis from prokaryotes to plastids. *Genome Biol. Evol.* **5,** 31–44 (2013).
15. Huerta-Cepas, J. *et al.* eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.* **44,** D286–D293 (2016).
16. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44,** D733–D745 (2016).
17. Markowitz, V. M. *et al.* IMG 4 version of the integrated microbial genomes comparative analysis system. *Nucleic Acids Res.* **42,** D560–D567 (2014).
18. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).
19. Tatusov, R. L., Koonin, E. V. & Lipman, D. J. A genomic perspective on protein families. *Science* **278,** 631–637 (1997).
20. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European molecular biology open software suite. *Trends Genet.* **16,** 276–277 (2000).
21. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30,** 1575–1584 (2002).
22. Thiergart, T., Landan, G. & Martin, W. F. Concatenated alignments and the case of the disappearing tree. *BMC Evol. Biol.* **14,** 266 (2014).
23. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30,** 772–780 (2013).

Acknowledgements

We thank D. Bryant, A. Kupczok and M. Wilkinson for critical comments on the manuscript. We acknowledge support from the European Research Council (grant no. 281375) and CAPES (Coordination for the Improvement of Higher Education Personnel–Brazil).

Author contributions

T.D., G.L. and F.D.K.T. conceived the study. F.D.K.T. and G.L. developed and implemented the method. F.D.K.T. performed all analyses. T.D., G.L. and F.D.K.T. wrote the manuscript.

Additional information

Supplementary information is available for this paper.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to G.L.

How to cite this article: Tria, F. D. K., Landan, G. & Dagan, T. Phylogenetic rooting using minimal ancestor deviation. *Nat. Ecol. Evol.* **1,** 0193 (2017).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Competing interests

The authors declare no competing financial interests.