Deformity Index: A Semi-Reference Clade-Based Quality Metric of Phylogenetic Trees

Aritra Mahapatra1 · Jayanta Mukherjee1

Received: 8 September 2020 / Accepted: 20 March 2021 / Published online: 3 April 2021
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

Abstract
Measuring the dissimilarity of a phylogenetic tree with respect to a reference tree or the hypotheses is a fundamental task in the phylogenetic study. A large number of methods have been proposed to compute the distance between the reference tree and the target tree. Due to the presence of unresolved relationships among the species, it is challenging to obtain a precise and an accurate reference tree for a selected dataset. As a result, the existing tree comparison methods may behave unexpectedly in various scenarios. In this paper, we introduce a novel scoring function, called the deformity index, to quantify the dissimilarity of a tree based on the list of clades of a reference tree. The strength of our proposed method is that it depends on the list of clades that can be acquired either from the reference tree or from the hypotheses. We investigate the distributions of different modules of the deformity index and perform different goodness-of-fit tests to understand the cumulative distribution. Then, we examine, in detail, the robustness as well as the scalability of our measure by performing different statistical tests under various models. Finally, we experiment on different biological datasets and show that our proposed scoring function overcomes the limitations of the conventional methods.

Keywords Phylogenetic tree · Semi-reference quality metric · Transfer level · Deformation · Clade deformation · Deformity index

Introduction

Different phylogenetic studies derive various tree topologies considering the Operational Taxonomic Units (OTU). Hence, measuring the numerical dissimilarity of a derived phylogenetic tree with respect to the known hypotheses becomes one of the most fundamental tasks in phylogenetic research (Bogdanowicz and Giaro 2011; Goluch et al. 2020; Sul et al. 2009). Now, different methods have already been proposed to compute the accuracy of the derived phylogenetic trees. The existing methods derive the correctness of a tree by comparing it with a reference tree topology which is an utmost requirement for these methods. Some of the popular methods include Robinson-Fould (RF) distance (Robinson and Foulds 1981), Matching Split (MS) distance (Bogdanowicz and Giaro 2011), Nodal Split (NS) distance (Cardona et al. 2010), Triples (TT) distance (Critchlow et al. 1996), Geodesic Treepath (GTP) distance (Billera et al. 2001; Owen and Provan 2010). However, all these methods have common significant limitations, which are categorized into the following situations:

Case 1 (unresolved relations) A full reference tree for the selected OTUs may not available to us due to the lack of knowledge of the precise relationships among some of them. In such cases, we assume multifurcation among those OTUs whose evolutionary relationships are not clear to us. Existing tree comparison methods are very rigid towards the nature of furcation of the internal nodes in both of the reference tree and the target tree. As a result, the existing methods perform poorly when the reference tree has a multifurcated node whereas the same set of species show bifurcation in the target tree (for details, please refer to Supplementary Sec. 1).

Case 2 (unavailability of an absolute reference tree) Expansion of the phylogenetic study by introducing newly discovered species is necessary for the success of the Tree
of Life Web Project (Goldstein 2010). Moreover, the limited availability of the sequence data is a significant reason to obtain an absolute reference tree for a target OTUs. Nevertheless, the existing tree comparison measures are applied when both of the reference and the target trees have the same set of OTUs. To ensure that, the conventional methods prune the OTUs present in one tree but not in another. In another way, these methods expand the tree(s) by inserting the missing leaf (or leaves). Hence, for this case, all the existing methods may reflect inappropriate results when the target tree has few more species than that of the reference tree (for details, please refer to Supplementary Sec. 1).

Due to these limitations of the existing measures, most of the recent phylogenetic studies discarded the conventional tree comparison methods and measured the accuracy of their derived trees using the following methods:

a. **Manual inspection:** The correctness of the derived tree is checked manually by validating each clade utilizing the existing biological knowledge (Li et al. 2017). However, this technique does not generate any quantitative measure. Apart from that, the manual inspection is very subjective.

b. **Comparison-based method:** In this approach, a phylogenetic tree is formed by applying the widely accepted established methods on the same set of OTUs. This tree is then considered as the reference tree to be compared with the target tree derived from the corresponding dataset (Yin and Yau 2015). However, in this approach, the established methods have to be considered as the gold standard which are not beyond contentions.

c. **Indirect method:** This approach computes different errors, like root mean square error of distance matrices (Xie et al. 2018; Zheng et al. 2019). So this method overlooks the accuracy of the final tree construction steps as well as ignores the correctness of the ultimate derived tree. There are some methods like maximum likelihood, Bayesian estimation, which directly derive trees from the sequences without generating any distance matrices. For these cases, the indirect method is not applicable.

In this regard, we propose a novel measure, called the deformity index, to quantify the dissimilarity of a phylogenetic tree based on the presence of different clades in the reference tree. The computation of the deformity index only depends on the clades acquired from the reference tree. So the precise relationships among the species within a clade do not affect its computation. Hence, the presence of both bifurcation and multifurcation in the target tree does not influence the measure. Deformity index also handles the existence of extra species or OTUs in the target tree appropriately. Additionally, this technique generates a single quantitative measure to represent the correctness of a target tree with respect to the reference tree. This measure is applicable for all types of phylogenetic trees. Hence in this manuscript, we use the term OTU and species interchangeably.

### Methodology

Our proposed measure computes the dissimilarity of a tree with respect to a reference tree by utilizing the clade(s) of the reference tree. Our objective is to compute the degree of deformation for each clade in the target phylogenetic tree with respect to the clades of the reference tree. Next, we define a few terms and notations used to describe the methodology for computing the deformity index (DI). Throughout this paper, we use deformity index and DI interchangeably.

Let us consider a phylogenetic tree $T$ having the set of leaves and the set of internal nodes $V_T$. Thus, the set of all nodes of $T$ is $V = V_T \cup L$. The level of a node $\eta$ is represented as $h(\eta); \forall \eta \in V$. The parent node of $\eta$ is represented as $p(\eta); \forall \eta \in (V - r_T)$ where $r_T$ is the root node of $T$. For a set of leaf nodes $S$ the most recent common ancestor is represented as $M(S)$. The clade at a particular node $\eta$ is represented as the list of all leaves present under $\eta$ and is denoted as $A(\eta), \forall \eta \in V_T$. Please refer to Supplementary Sec. 2 for more details about the definitions and the notations used here.

### Definition 1 Transfer In

The minimum distance required to shift an OTU from its parent node to a selected clade that does not contain the species is called transfer in.

In this study, we consider the unweighted tree; hence the distance between two nodes is considered as the number of edges between them. This movement is applicable only for the leaves of a tree. The transfer in of an OTU is computed when we need to shift the corresponding species to its correct clade from its assigned position. If we require to add an OTU $A$ at node $\eta$, we need to remove $A$ from its parent node, $p(A)$ and insert it in $\eta$. For this case, the transfer in is denoted as $T_{A}(A, \eta)$. Hence,

$$T_{A}(A, \eta) = (h(p(A)) - h(M(L_A))) + (h(\eta) - h(M(L_A)))$$

$$= h(\eta) + h(p(A)) - 2h(M(L_A))$$  \(1\)

where $L_A = \lambda(\eta) \cup A; \forall A \in L$ and $\forall \eta \in V_T$.

A typical example is shown in Fig. 1a. We need to traverse a minimum distance of two edges to shift the species 4 of the tree $T$ from its current position to $\eta$. So, $T_{A}(4, \eta) = 2$. 

© Springer
Definition 2 Transfer Out

The minimum distance required to shift an OTU from its parent node to an ancestor of a selected clade consisting of the OTU is called transfer out.

Like the transfer in, only a leaf node can be shifted from a particular clade to an ancestral clade. The transfer out of an OTU is computed when we need to remove the OTU from a selected clade consisting of this species. For removing an OTU \( B \) from a clade \( \eta \), we need to remove \( B \) from its parent node, \( p(B) \) and insert it to the parent node of \( \eta \), \( p(\eta) \). For this case, the transfer out is denoted as \( TL_{\text{out}}(B, \eta) \). Hence, 
\[
TL_{\text{out}}(B, \eta) = h(\eta) - h(p(\eta)) + 1
\]
where \( \forall B \in L \) and \( \forall \eta \in V_f \).

In the example shown in Fig. 1c, we need to traverse a minimum distance of two edges to move the OTU 8 of the tree \( T \), from its allocated position to the parent of a mentioned node, \( \eta \). So, the transfer out, \( TL_{\text{out}}(8, \eta) = 2 \).

We have a reference tree \( T_R \) with a list of clades, say \( \Lambda(T_R) \), to compute the DI of a target tree, \( T \). Let us consider a reference clade, \( \Lambda_i \in \Lambda(T_R) \). At any particular node of the target tree, \( \eta \in V_f \), we may need to add a list of OTUs, \( P \), which are present in the reference clade, \( \Lambda_i \), but not in the clade, \( \Lambda(\eta) \), of \( T \). Hence, \( P = \Lambda_i \setminus \Lambda(\eta) \). Also, we may need to remove a list of OTUs, \( Q \), which are present in \( \Lambda(\eta) \) but not in \( \Lambda_i \). So, \( Q = \lambda(\eta) \setminus \Lambda_i \). These operations are necessary to make the clade \( \lambda(\eta) \) of \( T \) concordant to the reference clade, \( \Lambda_i \). For inserting the list of OTUs \( P \), we can compute the transfer in individually for each \( p \in P \). The sum of these individual transfer ins is called total transfer in. Similarly, for deleting the list of OTUs \( Q \), we can compute the transfer out individually for each \( q \in Q \). We call the sum of these transfer outs as the total transfer out. The total transfer in and total transfer out at an internal node \( \eta \) with respect to the reference clade \( \Lambda_i \) are represented as \( TTL_{\text{in}}(\eta, \Lambda_i) \) and \( TTL_{\text{out}}(\eta, \Lambda_i) \), respectively. The sum of \( TTL_{\text{in}}(\eta, \Lambda_i) \) and \( TTL_{\text{out}}(\eta, \Lambda_i) \) is considered as the node deformity at \( \eta \) with respect to \( \Lambda_i \) and is denoted as \( Dn(\eta, \Lambda_i) \). Hence, 
\[
Dn(\eta, \Lambda_i) = \sum_p TL_{\text{in}}(p, \eta) + \sum_q TL_{\text{out}}(q, \eta) ; \forall p \in P \text{ and } \forall q \in Q
\]

Definition 3 Deformation

The deformation at an internal node of the target tree with respect to a reference clade is defined as the normalized node deformity at the same node with respect to the same reference clade.

To compute the deformation at \( \eta \) with respect to \( \Lambda_i \), \( Dn(\eta, \Lambda_i) \) is normalized by the ratio of the size of the reference clade, \( \Lambda_i \), and the size of the target clade, \( \Lambda(\eta) \). Hence,
the deformation at \( \eta \) with respect to \( \lambda_i \) is denoted as \( d(\eta, \lambda_i) \) which is shown in the following expression.

\[
d(\eta, \lambda_i) = \frac{|\Lambda_i|}{|\lambda(\eta)|} D_n(\eta, \lambda_i)
\]

where \( |\Lambda_i| \) and \( |\lambda(\eta)| \) represent the number of members present under the clades \( \lambda_i \) and \( \lambda(\eta) \), respectively.

We can compute the deformation at every internal node of the target tree with respect to the reference clade. Now, our objective is to compute the minimum deformation among all the internal nodes of the target tree, \( T \), with respect to the reference clade, \( \lambda_r \).

\[\text{Definition 4 Clade Deformation}\]

The clade deformation of a target tree is the minimum deformation among all its internal nodes with respect to a reference clade.

Clade deformation with respect to \( \lambda_i \) is denoted as, \( Dc(\lambda_i) \). Hence,

\[
Dc(\lambda_i) = \min \{d(\eta, \lambda_i)\}; \forall \eta \in V_T
\]

For each \( \lambda_i \in \Lambda(T_R) \), we compute the \( Dc(\lambda_i) \). The deformity index (DI) of a tree \( T \) is the average of the clade deformations for all the reference clades. Deformity index of the tree, \( T \), is denoted as \( D(T) \), and is expressed as follows:

\[
D(T) = \frac{1}{|\Lambda(T_R)|} \sum \{Dc(\lambda_i): \forall \lambda_i \in \Lambda(T_R)\}
\]

The minimum possible value of DI is zero which occurs when all the clades of the target tree are consistent with the reference clades. The maximum value is achieved when the tree is a caterpillar tree, and the members of the reference clade are attached at their highest possible level in the caterpillar tree. The formulations, justifications, and detailed explanations on the range of DI are provided in Supplementary Sec. 3. A flow diagram of our proposed methodology for computing the deformity index is shown in Fig. 2.

\[\text{Results and Discussion}\]

In this section, we first compute DI and the other existing measures for a simulated dataset. The performance analysis shows the associativity of DI with the conventional measures. Then, we perform different statistical tests to characterize the distribution of DI. Finally, we compute DI by considering the biological information of clades and illustrate that DI outperforms the traditional techniques of computing the correctness of a tree.

\[\text{Simulated Dataset}\]

We generate 10,000 trees having the same set of OTUs randomly by considering both the Yule model (Harding 1971) and the uniform model (Semple and Steel 2003). However, the trees estimated from the sequence data do not follow the Yule model and the uniform model in reality. Hence, we also generate 10,000 phylogenetic trees with the same set of OTUs and consider them as the reference tree. For each reference tree, we simulate the DNA sequence of its OTUs by utilizing a tool called INDELible (Fletcher and Yang 2009). Finally, we apply the maximum likelihood method to reconstruct the trees from each set of simulated sequences. In the next sections, we refer to this method as the simulated model. The details of the data simulation are provided in Supplementary Sec. 4. We enumerate the simulations using three different models for different numbers of OTUs (from 10 to 100) separately. Thus, for each set of OTUs, the simulated data contain 10,000 reference and target trees. Now, we compute the RF, MS, NS, TT, GTP, and DI scores of these target trees with respect to the corresponding reference trees. We first use the means of DI of each set of trees to compute the correlation between DI and the traditional methods.

\[\text{Correlation Coefficients}\]

Here, we compute the Pearson correlation coefficients (PCC) (Stigler 1989) between DI and RF, MS, NS, TT, and GTP. The PCC values between DI and the other existing measures, such as RF, MS, NS, TT, and GTP, are very high for all the cases (please refer to Table 1). This phenomenon represents a strong association of DI with the existing measures.

Additionally, our measure deals with the cases where the reference tree and the target tree contain different sets of species and also, the cases where the complete knowledge of the reference tree is not known to us. Again, our proposed method outperforms the existing methods as discussed in Sect. Biological Dataset.

\[\text{Distribution of Deformity Index}\]

As DI is linearly related to clade deformation, the distribution of clade deformation also reflects the distribution of DI. The primary components for computing the clade deformation are \( TTL_{in} \) and \( TTL_{out} \) (please refer to Eq. 5) which depend on the levels of internal vertices and follows the geometric distribution (Steel and McKenzie 2001). Hence, deformation also follows the geometric distribution. Now, to compute clade deformation, the deformation is normalized by the ratio of the size of the reference clade and the number of leaf nodes under a clade of the target tree. The distribution of the number of leaves under a clade follows the beta distribution (Mahmoud and Smythe 1991). Hence,
the distribution of clade deformation is a combination of geometric and beta distribution. However, it is challenging to model it and it remains an open problem at this stage.

**Test for Goodness-of-Fit** The goodness-of-fit test is used to determine whether the observed sample distribution of a given phenomenon is significantly different from the expected probability distribution. Since it is challenging to derive the distribution of DI mathematically, here, we employ the goodness-of-fit test to determine whether this distribution follows a normal distribution or not. We employed two
widely used techniques, *Chi-square* test (Pearson 1992) and Kolmogorov–Smirnov (K–S) test (Marsaglia et al. 2003), to examine whether the distribution of DI follows a normal distribution or not. In our test procedure, we consider the distribution of clade deformation following a normal distribution as the null hypothesis. The Chi-square test accepts the null hypothesis (with a significant level of more than 95%), whereas the K–S test rejects it. We compute the distributions of DI for 10,000 random trees generated under both the Yule model and the uniform model. Now, we perform the goodness-of-fit test on the trees generated from the simulated model as we have described above. We observe that the distribution of DI for the simulated model also follows a normal distribution based on the Chi-square test (with a significant level of more than 95%). However, the K–S test rejects the null hypothesis. Figure 3 represents the histogram of DI for such random trees with 100 leaves. From Fig. 3, we may add that the distribution of clade deformation does not follow a normal distribution but is very close to it.

We compute the means and the standard deviation intervals of DI for 10,000 random trees with different number of leaves generated under both the Yule model and the uniform model (please refer to Fig. 4). It is a fact that trees generated under the uniform model are more distinct than that generated under the Yule model. Means and the standard deviation intervals of DI in the uniform model are larger and grow faster than that of in the Yule model (shown in Fig. 4). Thus, DI shows more distinct values for the uniform model than the Yule model, which satisfies the fact described above. Now, we consider the trees derived from the simulated model and we perform this analysis on them. From Fig. 4, it is observed that the means and the standard deviation intervals of DI for this case grow slower than that of the Yule and the uniform models. These phenomena also claim that the DI is very sensitive to the tree topology.

**Biological Dataset**

There exist various studies for determining the phylogenetic relationships of different sets of OTUs. Hence, there are many observations on the phylogenetic relationships of a selected set of OTUs. In this section, we consider different such hypotheses to analyze the performance of both DI and the conventional measures for each of such hypotheses as well as for a selected reference tree. Here, we consider two datasets of fishes (from order Gadiformes) and mammals. We examine our proposed method on the trees derived by utilizing 14 different methods belonging to the alignment-based method and the alignment-free method. This section

![Fig. 3](image)

**Fig. 3** Histogram of the deformity index for 10,000 random trees with 100 leaves generated under a Yule model, b uniform model, c the trees derived from the simulated sequences by utilizing the maximum likelihood method (Color figure online)
demonstrates the power of DI when the reference tree and the target tree contain different sets of OTUs. Since, GTP (Owen and Provan 2010) executes by considering only the same set of OTUs, we consider RF, MS, NS, and TT methods in this section.

**Gadiformes**

We choose 19 species from eight different taxonomy groups of subfamily from the order Gadiformes and one outgroup species from the order Clupeiformes. The details of the species are given in Supplementary Table 5.1.

To demonstrate the power of DI, we consider a tree as the reference tree (shown in Fig. 5a) and also consider a phylogenetic tree of Gadiformes derived from Euclidean-based dissimilarity measure of k-mers followed by UPGMA method (Lu et al. 2017) (shown in Fig. 5b). Both of the trees have different sets of species. Though these trees have visible differences, all the conventional methods prune the uncommon species from both the trees and provide zero as the comparison score representing that both the trees are equivalent. For this case, DI shows a non-zero value implying that the target tree has some differences with respect to the reference tree. However, a single reference tree may not be sufficient as it may not include complete information about the relationships among the selected species. For these cases, consideration of multiple reference trees can resolve this issue. Nevertheless, the conventional methods consider only a single reference tree while computing the score. However, DI takes account of the knowledge of the relationships among the species from multiple hypotheses and computes a score based on this cumulative knowledge. Here, firstly we demonstrate how to accumulate knowledge of the clades from different hypotheses and use them to compute the DI. Then we present the results showing that DI outperforms the other conventional measures.

There are many conflicts among various hypotheses of the phylogenetic relationships among the species of Gadiformes. Hence, the true reference tree of Gadiformes is still unknown to us. Here we consider the widely accepted hypotheses of the phylogenetic relationships of these families for scoring the derived trees constructed from the different methods. A
brief description of the most accepted hypotheses are given below. Please refer to Supplementary Sec. 5 for the details related to these hypotheses.

HG1 Gadinae subfamily is monophyletic (Roa-Varón and Orti 2009; Shi et al. 2016; Teletchea et al. 2006; von der Heyden and Matthee 2008).

HG2 Subfamily Lotinae is the sister group of Gadinae. So, they form a monophyletic clade (Fahay 1984; Nelson 1984).

HG3 Bregmacerotidae belongs to the sister group of either higher “gadoids” or higher “gadoids” excluding “macruromids” (Nelson 2006; Roa-Varón and Ortí 2009; Shi et al. 2016). Most of the recent studies also accepted the monophyletic relation of Merlucciidae with the Gadinae and Lotinae (Endo 2002; Shi et al. 2016; Teletchea et al. 2006; von der Heyden and Matthee 2008). Hence, Bregmacerotidae and Merlucciidae both form the monophyletic clade with Gadinae and Lotinae.

HG4 Macrourinae is monophyletic (Gaither et al. 2016).

HG5 Macrourinae and Trachyrincinae together form monophyletic clade (Kriwet and Hecht 2008).

HG6 Macrourinae, Macrouroinae, Trachyrincinae, and Bathygadinae together form monophyletic clade (Kriwet and Hecht 2008).

According to these hypotheses, we construct the list of the reference clades. Then we compute different quality metrics such as, DI, RF, MS, NS, and TT of the target trees. We also visually judge the derived trees as “complete”, “partial”, and “marginal”, based on the reference clades.

- “complete” denotes that the tree agrees with the corresponding hypothesis completely.
- “partial” denotes that the tree misses a few relations based on the corresponding hypothesis.
- “marginal” denotes that very few relations of the hypothesis exist in the tree.

Considering the six hypotheses, we compute DI, RF, MS, NS, and TT for the trees derived from all 14 different methods. The trees are provided in the Supplementary Sec. 6. The visual judgments and the scores for each of the hypotheses are summarized in Supplementary Fig. 7.1. Here, some typical examples of these summarizations are shown in Fig. 6a–d. When visual judgment is “complete”, then the corresponding scores should be zero as the tree completely supports the hypothesis. While for the other cases, the comparison scores should be a positive value. For example, the tree derived by applying Euclidean distance measure on k-mer followed by UPGMA method (please refer to Fig. 5b) has monophyletic clade of the Gadinae subfamily. Hence, it satisfies HG1 “completely”. The DI shows zero score for that case which is represented by the absence of pink bar for HG1 as shown in Fig. 6a. However, all the other methods show the non-zero values because the reference clade contains multifurcation, which is not present in the derived tree (explained in Sect. Introduction). This phenomenon is represented by the presence of other four colored bars for HG1 as shown in Fig. 6a. The same thing happens for HG2 (please refer to HG2 of Fig. 6a). Again, in the same tree (shown in Fig. 5b), the members of subfamily Macrourinae form a paraphyletic clade with the other members of Gadiformes. So it “marginally” satisfies HG4. Hence, according to HG4, a high comparative score is expected. However, only the DI shows a non-zero score which is shown as a pink bar for HG4 as shown in Fig. 6a. All the other scores show zero value because these methods prune all the species which do not belong to the subfamily Macrourinae (explained in Sect. Introduction). This phenomenon is represented by the absence of these four bars in HG4 of Fig. 6a. Apart from that, when the visual judgment is “partial”, then the tree misses few relations based on the corresponding hypothesis. Hence, for a particular target tree, for these cases, the comparison scores should be less than that of when the tree supports the hypothesis “marginally”. Again, for the same tree (shown in Fig. 5b), HG3 is “partially” supported, whereas, both HG4 and HG6 are supported “marginally”. DI score is less for HG3 than that of both HG4 and HG6 (refer to Fig. 6a). Similarly, for all the 14 derived trees, DI shows more logical scores, which are also more associated with the visual inspections than that of the other measures. The detailed scores for each of the cases are provided in Supplementary Table 7.1.

Mammals

We consider 40 mammals from seven orders of class Mammalia. Among the existing hypotheses, we consider some of the widely accepted ones which are listed as follows (details are provided in Supplementary Table 8.1):

HM1 The order Primates forms a monophyletic clade (Hallström et al. 2007; Kriëgs et al. 2006; Li et al. 2017; Murphy et al. 2001a, b; Murphy et al. 2001a, b).

HM2 Human is the sibling of chimpanzee and gorilla is the sister of (human + chimpanzee) (Hallström et al. 2007; Prasad et al. 2008).

HM3 The order Carnivora forms the monophyletic clade (Li et al. 2017; Murphy et al. 2001a, b; Murphy et al. 2001a, b; Springer et al. 2003; Waddell and Shelley 2003).
HM4  The order Rodentia is the sister of Lagomorpha (Murphy et al. 2001a, b; Murphy et al. 2001a, b; Springer et al. 2003; Waddell and Shelley 2003).

HM5  Rodentia and Lagomorpha are sisters of Primates (Murphy et al. 2001a, b; Murphy et al. 2001a, b; Springer et al. 2003).

HM6  The order Perissodactyla is the sister group of order Carnivora (Murphy et al. 2001a, b; Murphy et al. 2001a, b; Springer et al. 2003).

Similar to the Gadiformes, we also consider these reference clades to compute the DI, RF, MS, NS, and TT measures of the trees derived from different methods. The derived trees are provided in Supplementary Sec. 9. We judge these trees visually as "complete", "partial", and "marginal", based on the reference clades, as we have done for the Gadiformes, which are summarized in Supplementary Fig. 10.1. Here, some typical examples are shown in Fig. 6e–h. To show that DI outperforms the other methods for this dataset, we consider the tree derived from the maximum parsimony-based method [as shown in Supplementary Fig. 9.1(d)]. As this tree supports HM3 "completely", hence, a zero comparative score is expected. However, except DI, all the other scoring methods show non-zero positive values. Deformity index (pink) consistently satisfies this property, while the other scoring methods underperform in many cases (Color figure online).
“marginally” and both HM4 and HM6 “partially.” DI shows less scores for both HM4 and HM6 than that of HM2, whereas the other methods show discrepancies in scoring the correctness for these cases. Similarly, based on the other derived trees, we observe that DI outperforms the conventional methods. The detailed scores for all the measures are provided in Supplementary Table 10.1.

**Complexity Analysis**

The deformation is the sum of $TTL_{in}$ and $TTL_{out}$. Its computational complexity depends on the number of species placed under the wrong clade. Considering a reference clade with $c$ leaves and a tree with $n$ leaves, the number of transfer in operations required are $[0, c]$. However, the number of transfer out operations required are $[0, n - c]$. Hence, the total number of $TTL_{in}$ and $TTL_{out}$ is $[0, n]$. Hence, the average time complexity of computing the deformation of a clade is $O(n)$. A tree with $n$ leaves has the maximum of $(n - 1)$ internal nodes (for binary trees). Thus, for each reference clade, the computation of the clade deformation is performed for a maximum of $(n - 1)$ times. Hence, the time complexity of computing the clade deformation of a reference clade is $O(n^2)$. If we have $R$ number of reference clades, then the time complexity for computing the deformity index of the tree is $O(Rn^2)$.

**Conclusion**

In this paper, we propose a novel semi-reference method to measure the quality of a tree using the list of the clades. Deformity index of the tree gives an idea about the correctness of the clades within a tree. As this method only depends on the clades of a reference tree, DI can easily adapt with the present knowledge in biology and provides the quality metric in that context. At the same time, DI can also adapt itself in versatile scenarios where the other conventional tree comparison methods do not provide a meaningful score. Though in this study, we propose the measure for the unweighted tree, this method can also be extended for the weighted tree. We inspect the distributions of different modules of the DI and also perform various statistical tests, such as Chi-square test and K–S test for the goodness-of-fit to understand the distribution of DI. From the statistical tests, we observe that DI is very sensitive to the tree topology. We have given extremal results as well as experimental results for different biological models to characterize the proposed method. Considering two datasets of fishes and mammals, we apply this measure to score the biological trees generated by different state-of-the-art methods. Higher the degree of adherence to the widely accepted hypotheses about various phylogeny, lower is their DI score. Finally, we conclude that deformity index, a flexible, versatile, and scalable tool, outperforms the traditional measures for computing the correctness of a tree and will be useful in the phylogeny research community.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s00239-021-10006-4.

**Acknowledgements** We would like to thank Barnali Das, research scholar, Department of Computer Science and Engineering, Indian Institute of Technology Kharagpur for the valuable comments on this study.

**Funding** This research did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors.

**Code availability** A Python-based tool, named DefIn for computing the Deformity Index is freely available at http://www.facweb.iitkgp.ac.in/~jay/DefIn/HIIARG-DefIn.html. The mirror of this program is also available at https://github.com/ariramhpd/DefIn.git.

**Declarations**

**Conflict of interest** On behalf of all authors, the corresponding author states that there is no conflict of interest.

**Ethical Approval** The ethical approval from the Institute is not applicable for this study.

**Informed Consent** All the data and personal information are collected from a public database. These data have been generated by various researchers since a decade.

**References**

Billera LJ, Holmes SP, Vogtmann K (2001) Geometry of the space of phylogenetic trees. Adv Appl Math 27:733–767

Bogdanowicz D, Giaro K (2011) Matching split distance for unrooted binary phylogenetic trees. IEEE/ACM Trans Comput Biol Bioinf 9:150–160

Cardona G et al (2010) Nodal distances for rooted phylogenetic trees. J Math Biol 61:253–276

Critchlow DE, Pearl DK, Qian C (1996) The triples distance for rooted bifurcating phylogenetic trees. Syst Biol 45:323–334

Endo H (2002) Phylogeny of the order gadiformes (Teleostei, Paracanthopterygii). Mem Grad Sch Fish Sci Hokkaido Univ 49:75–149

Fahay M (1984) Gadiformes: development and relationships. In: Moser HC, Richards WJ, Cohen DM, Fahay MP, Kendall AW, Jr. Richardson SL (eds) Ontogeny and systematics of fishes. American Society of Ichthyologists and Herpetologists, and Allen Press, Lawrence, KS, pp 265–283

Fletcher W, Yang Z (2009) INDELible: a flexible simulator of biological sequence evolution. Mol Biol Evol 26:1879–1888

Gaither MR et al (2016) Depth as a driver of evolution in the deep sea: insights from grenadiers (Gadiformes: Macrouridae) of the genus Coryphaenoides. Mol Phylogenet Evol 104:73–82

Goldstein AM (2010) Exploring phylogeny at the tree of life web project. Evol: Educ Outreach 3:668–674

Goluch T, Bogdanowicz D, Giaro K (2020) Visual TreeCmp: comprehensive comparison of phylogenetic trees on the web. Methods Ecol Evol 11:494–499
Hallström BM et al (2007) Phylogenomic data analyses provide evidence that Xenarthra and Afrotheria are sister groups. Mol Biol Evol 24:2059–2068
Harding E (1971) The probabilities of rooted tree-shapes generated by random bifurcation. Adv Appl Probab 3:44–77
Kriwet J, Hecht T (2008) A review of early gadiform evolution and diversification: first record of a rattail fish skull (Gadiformes, Macrouridae) from the Eocene of Antarctica, with otoliths preserved in situ. Naturwissenschaften 95:899–907
Li Y et al (2017) A novel fast vector method for genetic sequence comparison. Sci Rep 7:1–11
Lu YY et al (2017) CAFE: a C celerated A lignment-FrEe sequence analysis. Nucleic Acids Res 45:W554–W559
Mahmoud HM, Smythe RT (1991) On the distribution of leaves in rooted subtrees of recursive trees. Ann Appl Probab. https://doi.org/10.1214/aap/1177005874
Marsaglia G, Tsang WW, Wang J (2003) Evaluating Kolmogorov's distribution. J Stat Softw 8:1–4
Murphy WJ et al (2001a) Molecular phylogenetics and the origins of placental mammals. Nature 409:614–618
Murphy WJ et al (2001b) Resolution of the early placental mammal radiation using Bayesian phylogenetics. Science 294:2348–2351
Nelson JS (1984) Fishes of the world. Wiley, Hoboken
Nelson JS (2006) Fishes of the world. Wiley, Hoboken
Owen M, Provan JS (2010) A fast algorithm for computing geodesic distances in tree space. IEEE/ACM Trans Comput Biol Bioinf 8:2–13
Pearson K (1992) On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. In: Kotz S, Johnson NL (eds) Breakthroughs in statistics: methodology and distribution. Springer, New York, pp 11–28
Prasad AB et al (2008) Confirming the phylogeny of mammals by use of large comparative sequence data sets. Mol Biol Evol 25:1795–1808
Roá-Varón A, Ortí G (2009) Phylogenetic relationships among families of Gadiformes (Teleostei, Paracanthopterygii) based on nuclear and mitochondrial data. Mol Phylogenet Evol 52:688–704
Robinson DF, Foulds LR (1981) Comparison of phylogenetic trees. Math Biosci 53:131–147
Semple C, Steel M (2003) Phylogenetics. Oxford University Press, Oxford
Shi X et al (2016) Characterization of the complete mitochondrial genome sequence of the globose head whiptail Cetorurus globocephs (Gadiformes: Macrouridae) and its phylogenetic analysis. PLoS ONE 11:e0153666
Springer MS et al (2003) Placental mammal diversification and the Cretaceous-Tertiary boundary. Proc Natl Acad Sci 100:1056–1061
Steel M, McKenzie A (2001) Properties of phylogenetic trees generated by Yule-type speciation models. Math Biosci 170:91–112
Stigler SM (1989) Francis Galton's account of the invention of correlation. Stat Sci. https://doi.org/10.1214/ss/1177012580
Sul S-J, Matthews S, Williams TL (2009) Using tree diversity to compare phylogenetic heuristics. BMC Bioinf. https://doi.org/10.1186/1471-2105-10-S4-S3
Teletchea F, Laudet V, Hänni C (2006) Phylogeny of the Gadidae (sensu Svetovidov, 1948) based on their morphology and two mitochondrial genes. Mol Phylogenet Evol 38:189–199
von der Heyden S, Matthee CA (2008) Towards resolving familial relationships within the Gadiformes, and the resurrection of the Lyconidae. Mol Phylogenet Evol 48:764–769
Waddell PJ, Shelley S (2003) Evaluating placental inter-ordinal phylogenies with novel sequences including RAG1, γ-fibrinogen, ND6, and mt-tRNA, plus MCMC-driven nucleotide, amino acid, and codon models. Mol Phylogenet Evol 28:197–224
Xie G-S et al (2018) Graphical representation and similarity analysis of DNA sequences based on trigonometric functions. Acta Biotheor 66:113–133
Yin C, Yao SS-T (2015) An improved model for whole genome phylogenetic analysis by Fourier transform. J Theor Biol 382:99–110
Zheng W et al (2019) SENSE: Siamese neural network for sequence embedding and alignment-free comparison. Bioinformatics 35:1820–1828