BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 2 2013, pages 166–174
doi:10.1093/bioinformatics/bts657

Phylogenetics

Advance Access publication November 18, 2012

Stability analysis of phylogenetic trees

Saad I. Sheikh[1,*], Tamer Kahveci[1], Sanjay Ranka[1] and J. Gordon Burleigh[2]

[1]Department of Computer and Information Science and Engineering, University of Florida, FL, 32611 USA and [2]Department of Biology, University of Florida, FL, 32611 USA

Associate Editor: David Posada

ABSTRACT

Motivation: Phylogenetics, or reconstructing the evolutionary relationships of organisms, is critical for understanding evolution. A large number of heuristic algorithms for phylogenetics have been developed, some of which enable estimates of trees with tens of thousands of taxa. Such trees may not be robust, as small changes in the input data can cause major differences in the optimal topology. Tools that can assess the quality and stability of phylogenetic tree estimates and identify the most reliable parts of the tree are needed.

Results: We define measures that assess the stability of trees, subtrees and individual taxa with respect to changes in the input sequences. Our measures consider changes at the finest granularity in the input data (i.e. individual nucleotides). We demonstrate the effectiveness of our measures on large published datasets. Our measures are computationally feasible for phylogenetic datasets consisting of tens of thousands of taxa.

Availability: This software is available at http://bioinformatics.cise.ufl.edu/phylostab

Contact: sheikh@cise.ufl.edu

Received on May 17, 2012; revised on November 1, 2012; accepted on November 5, 2012

1 INTRODUCTION

Phylogenetics, or reconstructing the evolutionary history of organisms, is important for many areas of biological research. Rapid increases in the amount of available sequence data and computational advances have enabled construction of phylogenetic trees with tens of thousands of species (e.g. Goloboff *et al.*, 2009; Price *et al.*, 2010; Smith *et al.*, 2011). However, it often is difficult to assess the quality and reliability of large phylogenetic trees. An ideal method for assessing tree quality should be computationally tractable for extremely large datasets, quantify the confidence in relationships within the tree and compare the quality of the best tree to alternate trees. In this article, we describe a novel method for computing the stability of phylogenetic trees. We say that a tree is stable if its topology is largely robust to small changes in the input data used to infer the tree. Conversely, unstable trees may be subject to major structural changes with only a few alterations in the input data.

Often the quality of trees is determined using nodal support measures, such as non-parametric bootstrapping (Felsenstein, 1985), jackknifing (Farris *et al.*, 1996), relative support (Goloboff and Farris, 2001) or Bayesian posterior probabilities (Huelsenbeck *et al.*, 2000). These methods assess a large number of trees that represent a range of possible solutions. Perhaps, the most commonly used method is the non-parametric bootstrapping (Felsenstein, 1985). This method builds a collection of trees from pseudoreplicate datasets obtained by sampling with replacement from the original dataset. Because the phylogenetic analysis has to be repeated for each bootstrap replicate, it can be computationally expensive for large datasets; bootstrap analysis of tens of thousands of sequences can take years of processing time (Liu *et al.*, 2009). Also, the bootstrap methods consider only a few of the possible alterations in the input data. The Bremer support or decay index (Bremer, 1988, 1994) measures, for all groups in the tree, the minimum number of additional parsimony steps needed to find a tree without that group. This is similar to a stability measure. However, it does not examine how similar the alternate suboptimal topologies are to the original tree. Several recent methods (Aberer and Stamatakis, 2011; Aberer *et al.*, 2011; Pattengale *et al.*, 2011) seek to identify 'rogue taxa' (Sanderson and Shaffer, 2002), or taxa that cannot be placed reliably anywhere in the tree.

The issue of stability of phylogenetic trees has been proposed in the past (Donoghue and Ackerly, 1996; Giribet, 2003). In other areas of research, 'what-if' analyses have been discussed under the name of 'post-optimality analysis' or 'sensitivity analysis' (see Greenberg, 1997; Sotskov *et al.*, 1995). In this article, we describe a framework of phylogenetic stability measures based on sensitivity and post-optimality analysis. Our methods measure stability of trees with respect to the input data, allowing the data from each taxon to be modified independently. Our framework provides a measure of the minimum change in the input sequences that are required to alter significantly the topology of a given phylogenetic tree. Thus, they assess if and how much a particular tree relies on the accuracy of a small number of character states. The more changes in input data required to alter a tree significantly, the more stable the tree is. Further, nodal measures developed within this framework can be computed quickly. We demonstrate that our stability measures provide a new computationally feasible perspective on the quality of large phylogenetic trees. The main contributions of this work are as follows:

(1) We present an original framework for assessing the stability of phylogenetic trees and define several measures of stability. These measures capture information about stability of trees relative to changes in the input data.

(2) We apply our stability measures to published datasets. In these examples, we identify unstable subtrees and discuss why they are unstable.

*To whom correspondence should be addressed.

Published by Oxford University Press 2012.

Background and preliminary information is presented in Section 2. We then formally define the notions of stability and discuss their relevance in Section 3. In Section 4, we discuss how data obtained from stability measures can be used to analyse subtrees. The computational challenges for computing stability and our solutions to them are described in Section 5. We present experimental results in Section 6 and finally conclude in Section 7.

2 BACKGROUND

In this section, we present the basic definitions necessary to understand this study. We first describe the notation for sequences and trees (Section 2.1) followed by commonly used optimality criteria for trees (Section 2.2). Finally we present the framework for our stability measures (Section 2.3).

2.1 Sequences and trees

The goal of phylogenetics is to infer evolutionary relationships for a set of taxa. Each taxon is typically represented using a sequence, and the phylogenetic tree describes the evolutionary relationship between different sequences. Let $S = \{s_1, s_2, \ldots, s_n\}$ be the set of sequences on which the phylogenetic tree, denoted by **T**, is built. Each leaf node of **T** corresponds to a sequence in S. For example, S shown in Figure 1 can be used to construct a parsimony-based tree $\mathbf{T}_a$ shown in Figure 2. To reconstruct a phylogenetic tree, the input sequences are aligned with each other in a multiple sequence alignment such that each column in the alignment represents a homologous character. In this article, we assume that a multiple sequence alignment and a phylogenetic tree corresponding to this alignment are available.

2.2 Tree measures

Phylogenetic trees are generally inferred based on an optimality criterion. Two commonly used criteria are maximum parsimony and maximum likelihood.

The **parsimony score** is the minimum number of character changes implied by a tree given a multiple sequence alignment. A smaller parsimony score indicates a better tree. Parsimony-based methods count the total number of substitutions in the tree by summing the substitutions between sequences of every pair of adjacent nodes. Sequences for internal nodes may be reconstructed using algorithms such as the Sankoff algorithm (Sankoff, 1975).

2.3 Framework for confidence measures

Our framework for measuring the stability of tree requires computing two types of functions:

(1) A function $\varepsilon()$ that measures the amount of change made to the input sequences

(2) A function $\delta()$ that measures the amount of change in the resulting tree

It is possible to compute either function using a variety of metrics. In the analysis performed in this article, we use the edit distance and the Robinson–Foulds (RF) distance (Robinson and Foulds, 1981), as $\varepsilon()$ and $\delta()$, respectively. An edit operation on a sequence replaces an existing character with a new one, inserts a new character or deletes an existing one. The **edit distance** between two sequences is the number of edit operations needed to convert one of the sequences to the other. We denote the edit distance between two sequences s_a and s_b by $\varepsilon(s_a,s_b)$ or $\varepsilon_{a,b}$. In this article, we use an edit distance based on changing individual sites, which mimics the effects of errors that may emerge from modern sequencing techniques. Note that we could have defined an edit distance based on removing and replicating entire columns in the sequence alignment. This would be a special case of editing each sequence independently.

```
A : AACCCCCTT----
B : AACCC-CTT----
C : AACCT-C-T---G
D : AACCT-C-----G
E : CCTTTT----TTT
F : CCTCTCC-T-CTT
G : ACG----------
```

Fig. 1. DNA Sequences given in the form of multiple sequence alignment

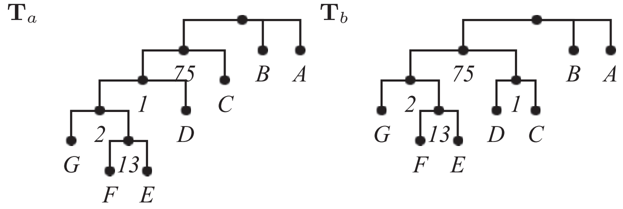


Fig. 2. Trees for sequences in Figure 1 determined by: $\mathbf{T}_a$: Parsimony; $\mathbf{T}_b$: Parsimony (alternate) & maximum likelihood. Bootstrap support values (%) over trees generated by PHYLIP 3.67 (Felsenstein, 2005) are shown

For the tree distance, let **T** be a rooted tree defined on a set of sequences S. Now, for every internal node $\mathfrak{n}$, there exists a set of leaf nodes. These leaf nodes compose a **clade**. Let $\Upsilon_\mathbf{T}$ be the set of all clades in **T**. The **RF** distance between two trees $\mathbf{T}_1$ and $\mathbf{T}_2$ is the number of clades they do not have in common $(\Upsilon_{\mathbf{T}_1} - \Upsilon_{\mathbf{T}_2}) \cup (\Upsilon_{\mathbf{T}_2} - \Upsilon_{\mathbf{T}_1})$. For example, in Figure 2, the **RF** distance between $\mathbf{T}_a$ and $\mathbf{T}_b$ is two. Note that trees do not need to be rooted for the stability measures. However, we assume they are rooted as it enables efficient computation of ancestral sequences.

Again, a large number of measures can be used for either of the functions. For example, we can use a likelihood model to measure the distances between sequences, and a weighted RF for the tree distance. However, to emphasize stability, rather than the specific measures we choose for these functions, we use two commonly used and intuitive measures.

3 STABILITY

We define the stability measures for a tree in Section 3.1. In Section 3.2, we define stability measures that work at a node level. We then analyse these measures and discuss them with the help of an example in Section 3.3.

3.1 Tree measures

We define two measures that capture the stability of the entire dataset. First, we define a measure that captures the minimum change in input (sequences) necessary to cause a major change in the output (tree topology). To define this, a threshold of similarity in tree topology (R) must be defined to identify when the tree topology has changed *significantly*.

Definition 1. *Let* $\boldsymbol{T}_1$ *and* $S_1 = \{s_1, s_2, \ldots, s_n\}$ *be a phylogenetic tree and its set of taxa sequences* (s_i)*, respectively. Let* $S_2 = \{s'_1, s'_2, \ldots, s'_n\}$ *be the set of sequences where* s'_i *for each* $i \in \{1 \ldots n\}$ *is obtained by editing* s_i*. Let* $\mathbb{E}(S_1, S_2)$ *be the total number of edit operations necessary to transform* S_1 *to* S_2*:* $\mathbb{E}(S_1, S_2) = \Sigma_i \varepsilon(s_i, s'_i)$*. Let* $\boldsymbol{T}_2$ *be the optimal phylogenetic tree for* S_2 *under the same evolutionary model as* $\boldsymbol{T}_1$*. Given a tree distance threshold R, we define the* ***edit distance stability*** *(*$\mathcal{EDS}$*) of* $(S_1, \mathbf{T}_1, \mathrm{R})$ *as the minimum value of* $\mathbb{E}(S_1, S_2)$ *necessary such that* $\delta(\mathbf{T}_1, \mathbf{T}_2) \geq \mathrm{R}$.

Formally:

$$\mathcal{EDS}(\mathbf{T}_1, S_1, \mathrm{R}) = \min_{S_2} \{\mathbb{E}(S_1, S_2)\} \quad \text{such that}\, \delta(\mathbf{T}_1, \mathbf{T}_2) \geq \mathrm{R}$$

This measure captures exactly the amount of sequence editing it takes to make the given tree ‘unstable’ with respect to an edit distance threshold. Thus, it provides a guarantee on the amount changes in the data that will not significantly change the tree. For example, the tree $\mathbf{T}_a$ in Figure 2 has an edit distance stability of zero with respect to a tree threshold of one because $\mathbf{T}_b$ is an alternate tree with same parsimony score. However, this is a very low threshold. With a threshold of two, the stability is higher because it requires more editing to make a larger change in the tree. We now define a measure that captures the maximum change in the tree possible if we are allowed to edit the input sequences by a specified amount. To define this, a threshold of editing (E) must be defined to limit the number of changes in the sequences.

Definition 2. *Let* $\boldsymbol{T}_1$ *and* $S_1 = \{s_1, s_2, \ldots, s_n\}$ *be a phylogenetic tree and its set of taxa sequences* (s_i)*, respectively. Let* $\mathbb{E}(S_1, S_2)$ *be the total number of edit operations necessary to transform* S_1 *to* S_2*. Let* $\boldsymbol{T}_2$ *be the optimal phylogenetic tree for* S_2 *under the same evolutionary model as* $\boldsymbol{T}_1$*. Given an edit distance threshold E, we define the* ***tree distance stability*** *(*$\mathcal{TDS}$*) of* $(S_1, \mathbf{T}_1, \mathrm{E})$ *as the maximum value of* $\delta(\mathbf{T}_1, \mathbf{T}_2)$ *necessary such that* $\mathbb{E}(S_1, S_2) \leq \mathrm{E}$.

Formally:

$$\mathcal{TDS}(\mathbf{T}_1, S_1, \mathrm{E}) = \max \delta(\mathbf{T}_1, \mathbf{T}_2) \quad \text{such that}\, \mathbb{E}(S_1, S_2) \leq \mathrm{E}$$

This measure captures exactly the maximum change in the tree possible given a threshold to editing. For example, the tree $\mathbf{T}_a$ in Figure 2 has a tree distance stability of one with respect to an edit distance threshold of zero because $\mathbf{T}_b$ is an alternate tree with same parsimony score.

Both of these measures provide explicit guarantees on the stability of the given combination of sequences, model and tree. They also pose challenging optimization problems due to the vast search space of trees. We believe these to be at least as hard as reconstructing a tree.

3.2 Node stability measures

The edit distance and tree distance stability defined in Section 3.1 describe the stability of a given phylogenetic tree from two points of view. However, computing these measures is non-trivial due to the number of combinations of possible edits and resulting trees. Moreover, these definitions fail to capture the overall stability of subtrees and taxa. For example, Definitions 1 and 2 cannot distinguish between $\mathbf{T}_a$ and $\mathbf{T}_b$ in Figure 2. Next, we define two confidence measures that address these problems. These measures focus on modifying individual taxa. As we discuss in Section 4, we also extend the notion to internal nodes by using reconstructed sequences and treating them as taxa. Thus, we use the terms node and taxon interchangeably to define our measures. We compute these measures by comparing (see Fig. 3) each node **x** and make it a sibling of every other node **y** that is not a descendant of **x**. For each pair, we compute: (i) how much editing is required to make **x** the *nearest* node, in terms of edit distance, to **y**; (ii) the distance of the tree from the original tree. In this article, we use the term **move** to refer editing the sequence of the node **x** so that it becomes the node with the sequence closest to that of **y**. Topologically, **x** is moved to a location sister to **y**. This moving is commonly referred to as Subtree Pruning and Regrafting (SPR). Trees obtained through this move are within one SPR distance of the original tree. While editing a sequence to make it closer to another one affects the alignment, we assume that the effect is small. Similarly, we assume moving a node does not affect the entire tree. Section 5.1 discusses these assumptions in detail. To compute ancestral sequences, we assume all trees are rooted.

Next, we define a measure that captures the minimum amount of editing necessary to any *one* sequence in the input sequences to transform the output tree significantly.

Definition 3. *Let* $\boldsymbol{T}_1$ *and* $S_1 = \{s_1, s_2, \ldots, s_{2n-1}\}$ *be a phylogenetic tree and its set of sequences, including the taxa and the inferred ancestral sequences,* (s_i)*, respectively. Let* $\boldsymbol{T}_2$ *be the phylogenetic tree formed by removing* $\boldsymbol{x} = s_i$ *from its original location in* $\mathbf{T}_1$ *and moving it to the sibling position of another node* $\mathbf{y} = s_j \in \mathbf{T}_1$ *by updating* s_i*. Let us denote the updated sequence* s_i *with* s'_i *to make this move.*

We define the ***minimum edit per tree distance*** $(\mathfrak{s})$ *to be the minimum edit distance between* s_i *and* s'_i *such that* $\delta(\mathbf{T}_1, \mathbf{T}_2) \geq \mathrm{R}$.

Formally:

$$\mathfrak{s}(\mathbf{T}_1, S, \mathrm{R}, s_i) = \min_j |\varepsilon(s_i, s'_i)| \quad \text{such that } \delta(\mathbf{T}_1, \mathbf{T}_2) \geq \mathrm{R}$$
$$\text{and } \forall s_k \in S_1\, \varepsilon(s'_i, s_j) \leq \varepsilon(s_j, s_k)$$

This measure (Definition 3) is similar to the original measure in Definition 1. The key difference is that changes in the input are now restricted to one taxon. This provides an explicit guarantee for how different one of the original sequences must be for the tree to change more than a given RF threshold within one *subtree prune and regrafting* (SPR) distance.

We now define a measure that captures the maximum change in the phylogenetic tree possible when editing is allowed on only one taxon.

Definition 4. *Let* $\boldsymbol{T}_1$ *and* $S_1 = \{s_1, s_2, \ldots, s_{2n-1}\}$ *be a phylogenetic tree and its set of sequences, including taxa and inferred ancestral sequences,* (s_i)*, respectively. Let* $\boldsymbol{T}_2$ *be the phylogenetic tree*

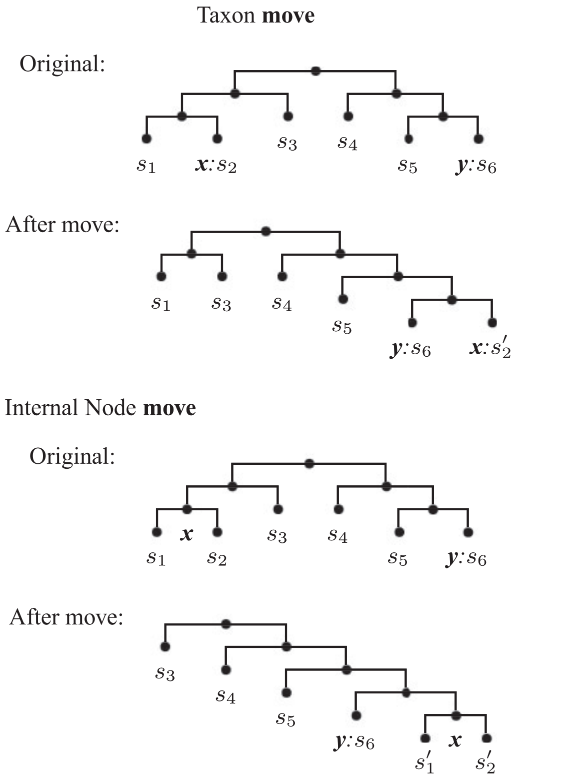


Fig. 3. Comparisons and moves. Every node **x** in a given tree is compared with every node **y** to calculate the cost of editing **x** to make it the *nearest* sequence to **y**; tree (RF) distance between original tree and the new one formed by moving **x** close to **y**: length of the unweighted path between **x** and **y**

formed by removing $\mathbf{x} = s_i$ *from its original location in* $\mathbf{T}_1$ *and moving it to the sibling position close to another node* $y = s_j \in \mathbf{T}_1$ *by updating* s_i*. Let us denote the updated sequence* s_i *with* s_i' *to make this move. We define* ***maximum tree distance per edit*** *as the maximum value of* $\delta(\mathbf{T}_1, \mathbf{T}_2)$ *such* that $\varepsilon(s_i, s_i') \le \mathrm{E}$.

Formally:

$$\tau(\mathbf{T}_1, S, \mathrm{E}, s_i) = \max_j \delta(\mathbf{T}_1, \mathbf{T}_2) \quad \text{such that } \varepsilon(s_i, s_i') \le \mathrm{E}$$
$$\text{and } \forall s_k \in S_1 \quad \varepsilon(s_i', s_j) \le \varepsilon(s_j, s_k)$$

The maximum RF distance per edit distance provides an explicit guarantee on how far a node may move in the input tree within one SPR distance given a tolerance for errors. Thus, this measure highlights the sequences that are more easily moved and therefore need closer inspection. Because the sequence length varies greatly in large-scale analyses, this measure can easily be adjusted to accommodate a percentage of the sequence length rather than the actual sequence.

Definition 5. *Let* ***T****1 and* $S_1 = \{s_1, s_2, \ldots, s_{2n-1}\}$ *be a phylogenetic tree and its set of sequences, including taxa and inferred ancestral sequences, (*s_i*), respectively. Let* **x** *be a node in* ***T****1 such that its sequence, reconstructed or given, is* s_i*. Let* ***T****2 be the phylogenetic tree formed by removing* $\mathbf{x} = s_i$ *from its original location and moving it to the sibling position of another node* $\mathbf{y} = s_j \in \mathbf{T}_1$ *by updating* s_i*. Let us denote the updated sequence* s_i *with* s_i' *to make this move. We say that a node is considered* ***movable*** *if* $\varepsilon(s_i, s_i') \le \mathrm{E}$ *and* $\delta(\mathbf{T}_1, \mathbf{T}_2) \ge \mathrm{R}$.

Definition 5 describes the nodes in the given tree that can be moved by editing sequences. Given thresholds for both RF and edit distance, the number of movable nodes summarizes the stability of the tree. If this number is high, it means that a large number of taxa could potentially be placed in a different part of the tree. In this case, the given tree is not stable.

Next, we examine a small example in detail and discuss how these measures may be used.

3.3 Application and examples

We demonstrate the newly defined measures using the example in Figures 1 and 2. We also compare these measures with the bootstrap support and parsimony scores.

We first consider a set of sequences (Fig. 1) in S_1 for which two equally parsimonious trees may be constructed: $\mathbf{T}_a$ and $\mathbf{T}_b$ (using DNAPARS/PHYLIP 3.6 (Felsenstein, 2005). RAxML (Stamatakis *et al.*, 2008) identifies $\mathbf{T}_b$ as the maximum likelihood tree. Because the two trees have the same parsimony score, the parsimony objective cannot differentiate between these two trees. Also, both trees have the same maximum parsimony and bootstrap scores.

The stability measures described above suggest that $\mathbf{T}_b$ is more stable than $\mathbf{T}_a$. This is because editing up to two positions affects $\mathbf{T}_a$, but it does not $\mathbf{T}_b$. The same threshold of editing allows for more of change in $\mathbf{T}_a$ than $\mathbf{T}_b$. This does not necessarily mean that $\mathbf{T}_b$ is right and $\mathbf{T}_a$ is wrong, but it suggests that $\mathbf{T}_b$ is more robust to potential errors or changes in the data. τ also provides an explicit guarantee that the maximum change that can be caused by editing up to three nucleotides in $\mathbf{T}_b$ is **RF** = 2 within 1 SPR distance. Thus, the stability measures elucidate the parts of the tree that are the least stable (leaves *C* and *D*) and the overall stability of different trees. Importantly, this is done by evaluating the underlying primary data (the sequence alignment).

An important advantage of the measures introduced in Section 3.2 is that they can be used to rank the nodes. We can also create a ranking of nodes at each edit distance using τ, e.g. ranking in a ascending order of instability for E = 6 in $\mathbf{T}_a$: $4 = E, 3 = \{C, G\}, 2 = \{D, B, A\}, 1 = F$. Such a ranking can help prioritize areas in the tree that need further phylogenetic research.

An important question arising at this point is how high a value of ε or δ is too high? To address this question, we propose statistical methods for assessing the stability of clades in the next section. We also discuss this question in Section 7.

4 ANALYSIS OF SUBTREES

Unlike the leaf level taxa, moving an internal node of the tree moves the entire clade under it. In this section, we describe how we analysed subtrees, or internal nodes, to evaluate their stability using sequences of the taxa under them.

We analyse internal nodes by looking at the leaf nodes underneath them and counting the number of movable nodes. We use a normal approximation to the binomial distribution to compute a z-score for each subtree in the following fashion.

Let N be the total number of leaf nodes and let K be the number of movable leaf nodes in the tree. The probability of any given leaf node being movable is $\frac{K}{N}$, and the expected number of movable nodes in a subtree of size n is $\frac{Kn}{N}$. The standard deviation, using the approximation, is $\sqrt{n.\frac{Kn}{N}.(1-\frac{Kn}{N})}$. Finally, the z-score for a subtree with k movable nodes, derived using the approximation, is defined as $\frac{k-\frac{Kn}{N}}{\sqrt{n\frac{Kn}{N}(1-\frac{Kn}{N})}}$. Note that the scores of the related internal nodes are correlated as a parent summarizes its children's scores. For instance, the number and/or ratio of movable sequences underneath an internal node indicates the weakness (or strength) of the clade. To use this information further, we sort the internal nodes in a descending order of their z-score. This ranking identifies the least stable subtrees. We also checked the correlation of z-score with the average sequence length, i.e. whether the relatively longer sequences were more stable. Overall, we observed little correlation with the exception of extremely long or short sequences.

We also computed the correlation between z-score and the presence of particular taxonomic families (or any taxonomic group) or genes in the underlying sequence alignment. Let the number of taxa in the tree be N. For each taxonomic family F_i, the number of nodes in the tree classified as that family are M_i. Let the number of nodes classified with family F_i in a given subtree $\mathbf{x}$, with n taxa, be m_i. We compute the probability of observing at least m_i members from family F_i using the hypergeometric distribution as follows:

$$P(F_i, \mathbf{x}) = \Sigma_{m_i \le k \le \min(M_i, n)} \frac{\binom{M_i}{k}\binom{N-M_i}{n-k}}{\binom{N}{n}}$$

We report the negative logarithm of this probability as the enrichment of family in that subtree. Thus, the larger the final number, the more the subtree is **enriched** with that family. For each subtree, we determine the family that maximizes this function. We report this as the **enrichment score** for the subtree.

5 COMPUTATIONAL ASPECTS

In this section, we first discuss basic assumptions made to make the computation possible (Section 5.1). We then describe the computational challenges posed by the internal nodes (Section 5.2). Finally, we discuss the overall complexity of the algorithm used to compute these measures and in Section 5.3.

5.1 Editing sequences and moving taxa

We make two assumptions to make the computation of our measures feasible:

(i) Modifying a small number of characters of a *single* node (which may represent changing one or more sequences in a multiple-sequence alignment, see Section 5.2), in a tree with a large number of nodes and taxa, does not significantly affect the other alignment positions.

(ii) Moving a *single* node (which may represent one or more taxa in a subtree) is unlikely to alter the topology of the rest of the tree significantly.

In the absence of these assumptions, we would have to estimate both sequence alignment and the tree after every small modification of a single sequence. Also, both sequence alignment and phylogeny inference are computationally expensive (NP-hard) problems, and the recomputation is not feasible for the large data sets we consider here.

5.2 Edit distance for clades

To determine the smallest number of edit operations needed to move a given node (leaves or internal) from its position in the tree to the sibling position of another node, we determine a sequence corresponding to each of the internal nodes. In our implementation, we used a parsimony reconstruction algorithm to infer ancestral sequences (Fitch, 1971). However, this can be replaced by any methods to reconstruct ancestral sequences.

It is straightforward to compute the edit distance given a multiple sequence alignment by simply counting the number of positions in alignment where the two sequences have different symbols. However, when considering moving a sequence from one part of the tree to a position sister to a target node (internal or leaf), we must ensure that we move it close to the target sequence while minimizing the editing we have to do. We compute the minimum editing required to move a sequence $\mathbf{x}$ to the sibling position of node $\mathbf{y}$ exactly as follows: For each node of the tree, we pre-compute the edit distance between the sequence of that node and its nodes that are close to it in the tree. We store the smallest edit distance for each node. This number for $\mathbf{y}$ is an upper bound to the distance allowed between $\mathbf{x}$ and $\mathbf{y}$ to move $\mathbf{x}$ next to $\mathbf{y}$. We obtain the minimum editing required for this move by subtracting this value from the original edit distance between the sequences of $\mathbf{x}$ and $\mathbf{y}$. This heuristic can be replaced by finding out exactly how close $\mathbf{x}$ has to be to $\mathbf{y}$ at an additional computational cost. Moreover, under some models of character evolution, two nodes do not need to have relatively similar sequences to be neighbors.

While computing the edit distance is relatively simple for moving a single sequence, it is non-trivial for moving a set of sequences representing a clade. We compute this distance by maintaining at each node a count of each character at each position. When computing the edit distance from a sequence to this clade, we use this information to count the exact number of nucleotides that are different. Thus, we count the exact number of sequences we need to change in leaves, to move an internal node.

5.3 Computing node stability measures and complexity

We compute the node stability measures by traversing the entire tree. We select each node as a putative moveable node $\mathbf{x}$, and compare it with every other node $\mathbf{y}$ in the tree. For every pair of nodes $\mathbf{x}$, $\mathbf{y}$ we compute the edit distance required to make $\mathbf{x}$ the nearest neighbor of $\mathbf{y}$ and the RF distance between the trees before and after this move. We maintain an array each for values of $\mathfrak{s}$ (for $\mathbf{x}$), $\mathfrak{r}$ (for $\mathbf{x}$) for all thresholds from 0 to 1000. After comparing $\mathbf{x}$ to every other node, this information is stored in a file.

For a tree with n nodes, tree height d and alignment length L, the time complexity of the above algorithm is $O[n^2(d+L)]$. This is because we consider $O(n)$ nodes for moving. If we are considering an internal node for a move, we need to compute the frequency of each nucleotide at each position leading to $O(nL)$ time. For each node (leaf or internal), we perform $O(n)$ comparisons to all possible nodes for moving. Each comparison involves comparing sequences in $O(L)$ time and finding out the common ancestor of the two nodes for computing the RF distance in $O(d)$ time. Thus, the time complexity of our algorithm is $O[n^2(d+L)]$. The space complexity is $O(Ln)$ as the memory usage is completely dominated by storing the sequence at every node. This can be reduced by pre-computing all distance comparisons in an offline fashion.

6 RESULTS

Implementation details: We have implemented our algorithm using Java. We were able to compute all the taxa stability measures on the datasets described in this article in <72 h. This includes a dataset with >55 000 taxa. Further, this computation can be easily parallelized.

Datasets: In our experiments, we used three datasets that represent the variation found in current large-scale molecular phylogenetic studies.

(1) The Mammals dataset is a phylogeny of 169 species of mammals published in (Meredith *et al.*, 2011). The length of the alignment is 35 603. The sequence length varies from 3893 to 32 201 with an average length of 29 373 and a standard deviation of 3669.

(2) The Saxifragales data is a phylogeny of 950 plant species (D. Soltis *et al.*, submitted for publication). The length of the alignment is 48 465 positions. The sequence length ranges from 227 to 26 863 with an average length of 2511 and a standard deviation of 3298.

(3) The Plants dataset is a phylogeny of 55 473 plant species (Smith *et al.*, 2011). The length of the alignment is 9853 positions. The sequence length ranges from 197 to 9135 with an average length of 1641 and a standard deviation of 1410.

6.1 Mammals dataset

In contrast to the other datasets, the Mammals dataset is highly stable (Fig. 4). There is not a single move (R = 1) possible for an edit distance of up to 530 nucleotides. Even if we place an extremely high limit of E = 1000, the biggest move possible is **RF** = 5. Thus, the stability measures provide an explicit guarantee that there is no move possible for E = 500 and any values of R within 1 SPR distance. This also demonstrates the power of building phylogenies from large densely sampled datasets. We tested our approach on randomly selected subsets of mammals with 80 taxa, and we still found no signs of instability.

6.2 Saxifragales

The Saxifragales dataset contains data from many closely related species. It is very sparse, with nearly 95% missing data in the matrix. In contrast to the mammal data set, it displays much instability (Fig. 5). With an edit distance of 100, it is possible to move nearly two-thirds of the taxa to an RF distance of 3 (Fig. 5a). Much larger changes in the tree topology, however, require a far greater edit distance. Therefore, this dataset demonstrates much local instability, where it is easy to make many small changes, in the midst of some robust phylogenetic structure, where large phylogenetic changes may be difficult.

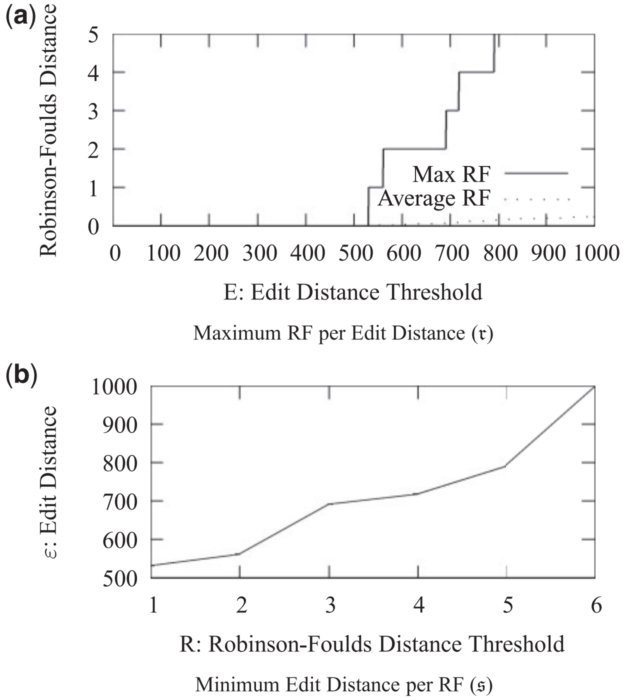
6.3 Plants dataset

We found high levels of instability in the 55 473 species plant tree (Fig. 6). Editing only 50 nucleotides is sufficient to move as many as 10 000 taxa using any of our RF distance thresholds. Figure 6a shows that the number of movable nodes increases sharply for all values of R up to 20. Although we found evidence of instability throughout the tree, for an edit distance threshold of 20, the most unstable taxon was *Ulex europaeus*, which can be moved sister to *Lupinus nootkatensis*, causing a change in the topology with RF distance of 61. This illustrates that very small changes in the underlying sequence data can lead to large changes in the tree topology. Both species, like many of the most unstable species, are part of an extremely large plant family, meaning that it is within a subtree containing many closely related species. Thus, although a RF distance of 61 represents a major change in tree topology, it is still only moving species within the same family. We can account for the distance of evolutionary relationship in our measures of stability by computing a RF distance that is weighted by the branch lengths. This would down-weight topology changes among closely related taxa relative to changes among more distantly related taxa.

As discussed previously in Section 4, we checked if the subtrees with a large ratio of movable taxa are enriched for specific taxonomic families. We found that nodes with some of the highest ranked subtrees were enriched in extremely large families. Out of 20 internal nodes ranked by our algorithm, 11 were enriched in Asteraceae, 4 in Poaceae and 2 in Orchidaceae. These are all among the five largest plant families. We also checked to see if the 11 nodes enriched in Asteraceae were descendants of a deep node enriched in Asteraceae. We found that the least common ancestor to all these nodes is very close to the root, far above the most recent common ancestor of the Asteraceae.

Figure 6b shows that subtrees enriched in families tend to be at either end of the Z-score spectrum, i.e. they are usually very stable or very unstable.

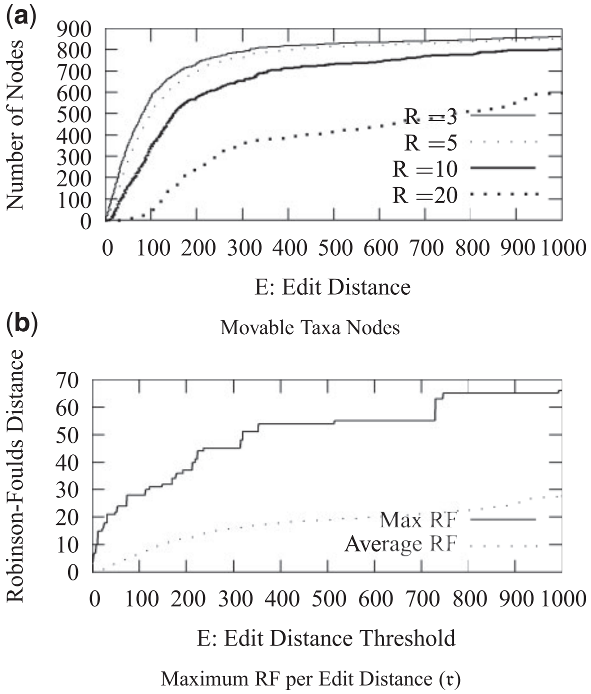
Figure 6c further illustrates the lack of stability of this tree. This figure plots the average and maximum RF distance against E. It also shows that on average, each taxon can be moved to an **RF** distance of 20 with as few as 100 nucleotides changed. With little to no editing, some taxa can be moved to an **RF** distance of 40 or more. Also, Figure 6d plots the minimum edit distance required against R. It shows that it is easy to perform moves with very large **RF** distance with very little editing. The unstable nature of this tree may be due to the presence of many taxa with little nucleotide data and the gappy multiple sequence alignment. We also checked if the reason for unstability is the large number of taxa. We randomly selected 10 subsets each of size 3000, 7000, 12 500 and 25 000. In each of these, we found a very high number of unstable taxa for similarly small thresholds. Thus, simply



Maximum RF per Edit Distance (τ)

Minimum Edit Distance per RF (ς)

Fig. 4. Mammals results. Both (**a**) and (**b**) indicate that it is hard to move nodes. On the average, it is impossible to move a node with up to 1000 edit operations



Movable Taxa Nodes

Maximum RF per Edit Distance (τ)

Fig. 5. Saxifragales results. Both (**a**) and (**b**) indicate a high percentage of locally unstable taxa in the tree, but the dataset is stable for larger RF values

reducing the taxon sampling will not necessarily create a more stable tree.

7 DISCUSSION

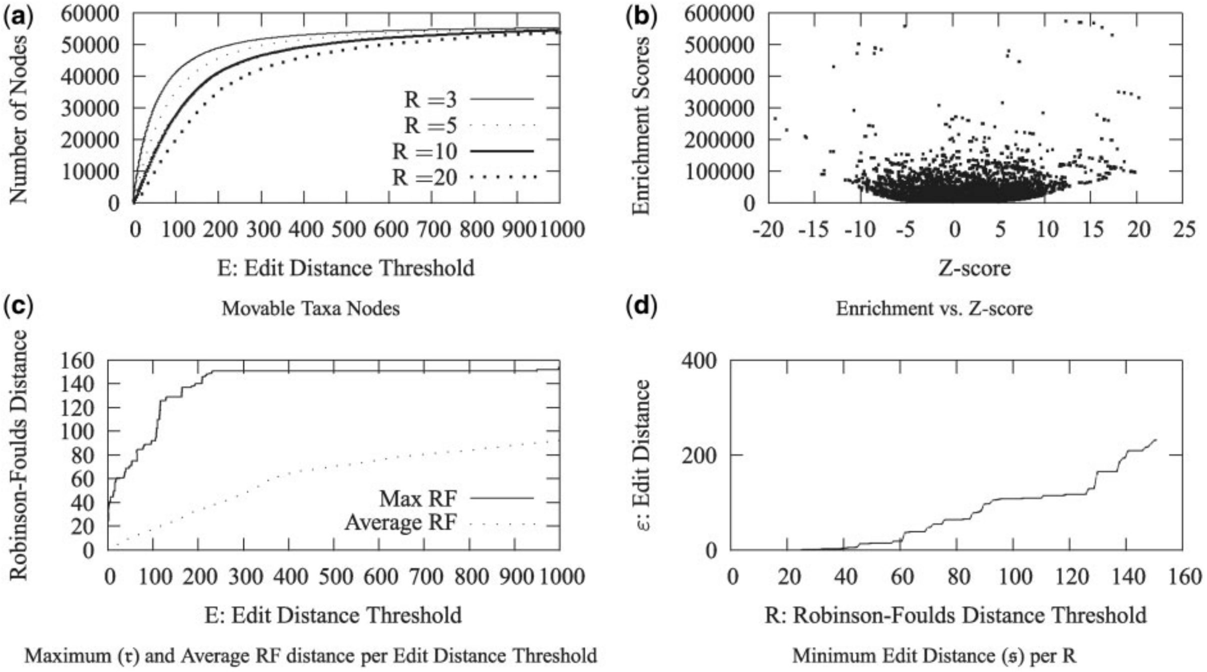
We have presented a novel framework to evaluate the stability of phylogenetic trees. These measures may be used to assess the stability of individual sequences, clades or entire trees. They provide information that may be used to direct further phylogenetic research and also suggest alternate plaubile topologies. Thus, the stability measures presented in this article may be used to improve quality of phylogenetic analyses and the performance of iterative algorithms. Our experiments demonstrate that our stability measures can be easily computed for the largest published empirical datasets.

It is important to note that the success of these measures is dependent on the choice of editing and tree distance measures used. While we have used edit and RF distance measures, any set of measures can be used. We can easily replace the edit distance with a distance estimated using the General Time Reversible model (Tavaré, 1986) or the RF distance with a weighted RF distance. In our current search strategy, SPR or tree bisection and reconnection (TBR) distance measures may not be useful. Criteria for the selection of E and R strongly depends on the choice of distance functions. This is a complex issue, but we briefly suggest how this problem may be approached. A simple approach could be to use percentages of average or current sequence length and tree height for determining E and R respectively. Alternatively, it is possible to rank the nodes in the order of their stability with respect to a E or R. Finally, it is possible to compute the statistical properties of sequence length and move distances to determine outliers with respect to mean and 2–3 standard deviations for range.

Our analyses on published large datasets identified strengths and weaknesses of these phylogenetic hypotheses. Our measures found that the Plants and Saxifragales datasets were highly unstable. Both include many closely related species, and both contain taxa with short sequences and with much missing data. It is therefore not surprising that small changes in the sequences can lead to large tree rearrangements. Our measures also recognized the stability in the Mammals dataset, which has far fewer taxa and much more sequence data.

It is important to note that the stability measures discussed here complement, rather than replace, existing measures of phylogenetic support. As a sanity check, we performed a limited comparison to a method that identifies rogue taxa running RogueNaRoK (Aberer *et al.*, 2011). While there was some overlap between the unstable taxa and the rogue taxa, the rankings were not obviously correlated. We do not pursue a detailed comparison with the measures of rogue taxa identification as their objectives are different from stability.

The intent of the stability measures is not to invalidate a tree or a method, but to provide explicit guarantees on situations where a *specific* tree on a *give* dataset will *not* be unstable. The placement of any subtrees and taxa identified as unstable should be interpreted with caution. The ranked list of unstable nodes should help prioritize areas of the tree that need further study or data.



Movable Taxa Nodes

Enrichment vs. Z-score

Maximum (τ) and Average RF distance per Edit Distance Threshold

Minimum Edit Distance (ε) per R

Fig. 6. Stability measures results on plants dataset. (**a**) shows that the number of movable taxa increases rapidly; (**b**) enriched nodes are typically highly stable or highly unstable, lower z-score indicates higher instability. (**c**) shows that on the average each node can move with little to no editing; (**d**) shows while short moves are very easy, larger moves in general require a lot of editing, suggesting an overall stable, but locally unstable structure

7.1 Future work

Our framework allows for definition of stability measures beyond those presented in this article. Especially, other forms of editing such as removal or addition of taxa may be informative for stability. In the future, we will examine the stability measures to guide the construction and iterative improvement of trees.

A preliminary analysis of the tree distance measures presented in Section 3.1 suggests computational intractability. In the future, we will explore the computational complexity and approximations to these measures. An important assumption we made is that the global sequence alignment is stable enough such that changing one sequence to move it closer to a different sequence is not going to affect everyone else in the alignment. Similarly, we also assumed that the remaining tree is stable. Both of these assumptions were made to avoid intractability. In the future, we will try to assess the impact this can have on our results and when is it safe to make this assumption. We will also incorporate online methods to update the alignment and the tree.

ACKNOWLEDGEMENT

We would like to thank the following people for their comments and discussion through our research: Tanya Berger-Wolf, Fahad Saeed.

Funding: This work was supported by the NSF grant: CCF-1048217.

Conflict of Interest: none declared.

REFERENCES

Aberer,A.J. and Stamatakis,A. (2011) A simple and accurate method for rogue taxon identification. In *Proceedings of IEEE BIBM 2011*.

Aberer,A.J. *et al.* (2011) RogueNaRok: an efficient and exact algorithm for rogue taxon identification. *Technical Report Exelixis-RRDR-2011-10*. Heidelberg Institute for Theoretical Studies.

Bremer,K. (1988) The limit of amino acid sequence data in angiosperm phylogenetic reconstruction. *Evolution*, **42**, 795–803.

Bremer,K. (1994) Branch support and tree stability. *Cladistics*, **10**, 295–304.

Donoghue,M.J. and Ackerly,D.D. (1996) Phylogenetic uncertainties and sensitivity analyses in comparative biology. *Philos. Trans. Biol. Sci.*, **351**, 1241–1249.

Farris,J.S. *et al.* (1996) Parsimony jackknifing outperforms neighbor-joining. *Cladistics*, **12**, 99–124.

Felsenstein,J. (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, **39**, 783–791.

Felsenstein,J. (2005) Phylip (phylogeny inference package) version 3.6. Distributed by the author. In Department of Genetics, University of Washington, Seattle.

Fitch,W.M. (1971) Toward defining the course of evolution: Minimum change for a specific tree topology. *Syst. Zool.*, **20**, 406–416.

Giribet,G. (2003) Stability in phylogenetic formulations and its relationship to nodal support. *Syst. Biol.*, **52**, 554–564.

Goloboff,P. *et al.* (2009) Phylogenetic analysis of 73 060 taxa corroborates major eukaryotic groups. *Cladistics*, **25**, 211–230.

Goloboff,P.A. and Farris,J.S. (2001) Methods for quick consensus estimation. *Cladistics*, **17**, S26–S34.

Greenberg,H.J. (1997) An annotated bibliography for post-solution analysis in mixed integer programming and combinatorial optimization. *Technical Report*. University of Colorado at Denver, Denver, CO, USA.

Huelsenbeck,J. *et al.* (2000) Accomodating phylogenetic uncertainty in evolutionary studies. *Science*, **288**, 2349–2350.

Liu,K. *et al.* (2009) Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. *Science*, **324**, 1561–1564.

Meredith,R.W. *et al.* (2011) Impacts of the cretaceous terrestrial revolution and kpg extinction on mammal diversification. *Science*, **334**, 521–524.

Pattengale,N.D. *et al.* (2011) Uncovering hidden phylogenetic consensus in large data sets. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **8**, 902–11.

Price,M. *et al.* (2010) Fasttree 2: approximately maximum-likelihood trees for large alignments. *PLoS ONE*, **5**, e9490.

Robinson,D. and Foulds,L. (1981) Comparison of phylogenetic trees. *Math. Biosci.*, **53**, 131–147.

Sanderson,M. and Shaffer,H. (2002) Troubleshooting molecular phylogenetic analyses. *Ann. Rev. Ecol. Syst.*, **33**, 49–72.

Sankoff,D. (1975) Minimal mutation trees of sequences. *SIAM J. Appl. Math.*, **28**, 35–42.

Smith,S.A. *et al.* (2011) Understanding angiosperm diversification using small and large phylogenetic trees. *Am. J. Bot.*, **98**, 404–414.

Sotskov,Y. *et al.* (1995) Some concepts of stability analysis in combinatorial optimization. *Discrete Appl. Math.*, **58**, 169–190.

Stamatakis,A. *et al.* (2008) A rapid bootstrap algorithm for the raxml web-servers. *Syst. Biol.*, **75**, 758–771.

Tavaré,S. (1986) Some probabilistic and statistical problems in the analysis of DNA sequences. *Am. Math. Soc.*, **17**, 57–86.