Interactive Visual Comparison of Multiple Trees

Sebastian Bremm*
Interactive-Graphics Systems
TU Darmstadt, Germany

Tatiana von Landesberger†
Interactive-Graphics Systems
TU Darmstadt, Germany

Martin Heß‡
Interactive-Graphics Systems
TU Darmstadt, Germany

Tobias Schreck§
Data Analysis and Visualization Group
University of Konstanz, Germany

Philipp Weil¶
Computational Biology
TU Darmstadt, Germany

Kay Hamacher‖
Computational Biology
TU Darmstadt, Germany

ABSTRACT

Traditionally, the visual analysis of hierarchies, respectively, trees, is conducted by focusing on one given hierarchy. However, in many research areas multiple, differing hierarchies need to be analyzed simultaneously in a comparative way – in particular to highlight differences between them, which sometimes can be subtle. A prominent example is the analysis of so-called phylogenetic trees in biology, reflecting hierarchical evolutionary relationships among a set of organisms. Typically, the analysis considers multiple phylogenetic trees, either to account for statistical significance or for differences in derivation of such evolutionary hierarchies; for example, based on different input data, such as the 16S ribosomal RNA and protein sequences of highly conserved enzymes. The simultaneous analysis of a collection of such trees leads to more insight into the evolutionary process.

We introduce a novel visual analytics approach for the comparison of multiple hierarchies focusing on both global and local structures. A new tree comparison score has been elaborated for the identification of interesting patterns. We developed a set of linked hierarchy views showing the results of automatic tree comparison on various levels of details. This combined approach offers detailed assessment of local and global tree similarities. The approach was developed in close cooperation with experts from the evolutionary biology domain. We apply it to a phylogenetic data set on bacterial ancestry, demonstrating its application benefit.

Index Terms: I.3.3 [Computer Graphics]: Viewing Algorithms—[H.5.2]: User Interfaces (D.2.2, H.1.2, I.3.6)—Interaction styles (e.g., commands, menus, forms, direct manipulation); I.3.6 [Computer Graphics]: Methodology and Techniques—Graphics data structures and data types

1 INTRODUCTION

The comparison of multiple trees is an important issue in biology and other application areas. Biologists often are concerned with evolutionary relationships between organisms. Such relationships are typically represented and analyzed by so-called *phylogenetic trees*. These trees show similarities in sequence alignments prepared from biological data such as DNA- or protein sequences. The leaves correspond to the respective organisms, while the branches denote the evolutionary ancestry between them.

*e-mail: sebastian.bremm@gris.tu-darmstadt.de
†e-mail: tatiana.von.landesberger@gris.tu-darmstadt.de
‡e-mail: martin.hess@gris.tu-darmstadt.de
§e-mail: tobias.schreck@uni-konstanz.de
¶e-mail: weil@bio.tu-darmstadt.de
‖e-mail: hamacher@bio.tu-darmstadt.de

The derivation of such phylogenetic trees is usually based on differing assumptions on the evolutionary model; thus the derived trees are highly susceptible to parameter choices [14, 23]. Therefore, it is very important to compare sets of trees flexibly. The simultaneous analysis of multiple trees is expected to lead to more insight into the evolutionary processes and/or to compensate for uncertainties in the model parameterizations.

The **goal** of comparing phylogenetic trees is to find similarities and differences between them at the same time. This is not restricted to global similarity evaluation, but more importantly, also encompasses the assessment of *local patterns*. For example, it is important to examine the conservation of subhierarchies across the trees. The typical tasks include identification of globally interesting trees for reference purpose, finding locally dissimilar structures in trees with high global similarity to a reference tree, or the conservation of a selected subhierarchy in other trees (see Section 4 for more information).

In the analysis, we have to distinguish two different levels of complexity a) the number of trees to compare and b) the number of organisms (leafs) in each tree. A typical research project deals with 10 to 50 phylogenetic trees or even more, while the number of organisms in the analysis spans orders of magnitude, from highly specialized questions on some ten, up to thousands of entities obtained from high-throughput analysis protocols. As numerous projects deal with multiple small trees, new approaches for these questions will have immediate benefit for phylogenetic research. Therefore, our contribution concentrates on these use cases.

Biologists currently have no readily **available advanced methods** for the visual comparison of phylogenetic trees. The tools commonly applied by biologists support visualization of single trees as node link diagrams, e.g., using the FigTree [20] software. For multiple trees, a typical approach is a simple visualization of pairwise tree similarities by a heatmap. However, this approach does not satisfy the analytical needs, in particular, it does not provide structural tree comparison and assessment of local patterns. Therefore, biologists would benefit from flexible, scalable visual analytics tools to mitigate the above mentioned problems.

Comparison of trees has also been addressed in the Information Visualization and Visual Analytics area. The available techniques, however do not support the detailed comparison of multiple trees.

In this paper, we present an interactive visual analytics system capable of comparing multiple trees simultaneously. This approach was developed in close cooperation with the co-authors from biology. As we focus on phylogenetic trees, we assume $n \geq 2$ rooted trees with the same leaf elements (organisms). Although the trees presented in application are predominantly binary, our approach is not restricted to them.

Our main **contributions and application benefits** are:

- We present a new visual analytics approach to *compare multiple trees*, both on global and local levels. To support efficient tree comparison, we combine automatic data analysis with in-

teractive visualization. This combination allows for data analysis on several levels of detail. In particular, the results of automatic analysis are used for highlighting interesting patterns in the data and selecting data for detailed inspection.

- We introduce a new *distance measure* to compare rooted trees. Our measure indicates differences in tree structure better than other available measures.

- Our approach has various *application benefits* for comparing sets of phylogenetic trees (see Section 4). For example, it enables biologists to identify evolutionary stable organism relations, such as the invariance of phylogenies for the two bacteria *Clostridium kluyveri* and *Clostridium beyjerinckii*.

The paper is structured as follows: Section 2 presents related work in the area of visual tree comparison. Section 3 lays out details of our approach. Section 4 demonstrates the usefulness of our approach in a real world application. Section 5 discusses the contributions and limitations of our approach as well as the outline of possible future work. Section 6 concludes.

2 Related Work

An overview of the visual analysis of graphs including their structural comparison is provided in a recent state-of-the-art report [37] presenting an exhaustive survey of work on tree visualization and analysis. Next, building on relevant approaches for interactive tree visualization discussed in Section 2.1, we recall relevant work on visual comparison of multiple trees in Section 2.2.

2.1 Tree Visualization

The main approaches for visual tree analysis include node-link diagrams and treemaps. Node-link diagrams are well suited for the visualization of phylogenetic trees [7, 19, 22]. They allow for the representation of weighted edges and offer an very intuitive representation of binary trees as many users are familiar with them. The usage of links between nodes for larger graphs may be space inefficient. Therefore, specialized layout algorithms have been proposed to increase visualization scalability (see [37] for an overview).

Alternative space efficient techniques, such as treemaps [31], use the whole available space. They recursively lay out child nodes within their respective parent nodes. As this technique employs overlapping of the parent nodes, the users may encounter difficulties with the assessment of the tree structure.

To overcome space limitations and to support exploration of the tree structure, data analysis, visualization techniques and user interaction is combined. Several tree traversal and expansion techniques [2, 9] can be employed to filter a given tree to the most interesting part. Alternatively, distortion techniques (e.g., fisheye views) allocate more display area to the parts of the tree of more higher interest for the user. They can be based on a degree-of-interest function such as in DOITrees [3, 15] or interactively selected by the user [27, 35]. Multiple coordinated views [5] offer an overview of the main tree structure and a detailed view on the selected parts of the tree. The construction of the overview relies on a score determining the interestingness of the substructure for a more detailed view.

These techniques focus on single trees and therefore form a basis for visual comparison of trees discussed in the following section.

2.2 Visual Tree Comparison

Existing techniques for visual comparison of trees focus on pairwise structural comparison and on comparison of multiple trees. An overview of existing visualization techniques and approaches can be found in [12]. A selection of approaches is laid out in the following section.

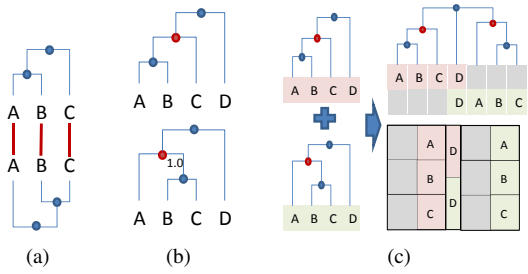


Figure 1: Current pairwise visual tree comparison approaches. a) Approach focusing on leaf node matching [17]. b) Approach focusing on matching most similar structures [27]. The red nodes are the best matches based on a comparison score, yet different subtree structures exist. c) Tree comparison using union trees and contrast treemaps [34].

Pairwise comparison: An overview of general approaches is shown in Fig. 1. Holten et al. [17] proposed an approach matching leaf nodes between two trees. They draw the two trees in opposite parts of the display and link their leaf nodes, while minimizing edge crossing. The crossing of links emphasizes the differences between the tree structures. These visual clues are enhanced by edge bundling. As shown in Figure 1(a), this approach does not reflect all structural differences between trees in some cases. In this case, the leaf nodes are fully aligned, however the tree structure differs. Telea and Auber extended this approach to analyze a sequence of pairwise tree comparisons [33].

Visual tree comparison focusing on the identification of corresponding nodes between the trees was presented by Munzner et al. in the TreeJuxtaposer system [27]. The approach was developed specifically for the analysis of phylogenetic trees. It analyzes and highlights leaf set similarities. This match is performed on demand, when the user clicks on a subtree. In this way, it can be cumbersome to analyze large trees. The matching uses a similarity score based on set overlap of the leaf nodes (see Section 3.1.2). This score regards common groups of nodes without their structural relationship. As seen in Figure 1(b) two nodes (red color) are matched with the highest score, although their subtree structures differ.

Tu and Shen [34] propose a comparison of two trees in a treemap visualization (called "contrast treemap"). It was developed to support a static comparison of dynamic trees in two time points. It unifies the two trees to be compared for structural match. The visualization of the union tree employs specific node coloring and texturing which highlights value differences between leaf nodes. This approach is well suited for comparing value changes, however spotting structural differences above leaf node level is more difficult and is layout dependent. Moreover, the union tree algorithm applied leads to node duplication (creating larger trees), which may complicate visual comparison (see Figure 1(c)).

Comparison of multiple trees: The above mentioned approaches are designed for the comparison of pairs of trees. There are only few techniques dealing with the comparison of multiple graphs (incl. trees). The available techniques for comparing multiple graphs [11, 36] are not specialized on trees and do not allow for explicit visual comparison of tree structures. For trees, the so-called "Trees of Trees" approach constructs a meta-tree by successive union of the underlying trees so that the total distance between tree nodes is minimized [28]. This is a computationally intensive approach. The visualization of the result does not offer direct insights into the inner structural comparisons between the connected trees and does not provide for shared pattern identification. Historically, biologists use simple visualization of pairwise tree simi-

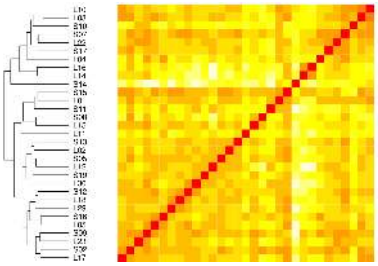


Figure 2: Typical visualization of a multiple tree comparison in biology. Shown are pairwise distances between phylogenetic trees and their hierarchical clustering as computed by the TOPD/FMTS package [30]. Such global analysis neglects any local structure or emergence of patterns in different trees. Note: Sx and Lx entries represent the name convention in microbiology for biomolecular sequences used to derive a particular tree.

larities in a heatmap [29] combined with hierarchic clustering (see Figure 2). It also does not offer structural tree comparison and assessment of local pattern differences.

Hillies et al. presented an approach for the comparison of many trees [16]. Each tree is represented in a scatterplot using Multidimensional scaling to determine its position. For a more detailed comparison, consensus trees are build revealing common substructures. Details about the uncommon structures are not provided.

3 Approach to Visual Comparison of Sets of Trees

Our approach supports the comparison of multiple, rooted trees with identical leaf elements. It is designed to support identification of similarities and differences between these trees. This is not restricted to global similarity evaluation as in Figure 2. Rather, assessment of local patterns is specifically addressed.

Scalability of the analysis with respect to both the number of compared trees and their sizes (measured by the number of leaf nodes) is supported by visual analysis on several levels of detail. In this respect, we combine comparative data visualization and automatic data analysis. The computation of local and global similarities is used for filtering and highlighting interesting data patterns.

Our approach is based on several interlinked views representing multiple levels of detail in the comparison analysis (see Figure 3). An initial overview shows the similarity matrix between all trees (see Section 3.2.1). From there, one reference tree can be selected for a detailed comparison with other trees (see Section 3.2.3 and 3.2.2). All views are supported by integrated calculation of local and global similarity measures (see Section 3.1).

In the following, we describe our approach in more details. We first describe the employed similarity measures and then detail on the interactive visualization.

3.1 Similarity Measures

There are several approaches to calculate the distance between two trees [27, 32, 39]. We develop a novel scoring scheme, specifically suited for phylogenetic analysis. To support different analytical tasks, we also employ two commonly used measures [27, 32]. We first introduce the terminology and then the measures divided into three categories.

3.1.1 Definitions

We are concerned with rooted trees exclusively in this study. A tree T consists of set of undirected edges E that connect pairs of nodes V, formally defined as $T = (V,E); E \subseteq [V^2]$. A tree T is called *rooted*, if one node r is distinguished as a so-called root node: $T = (V,E,r)$. A path in a tree is defined as a unique sequence of connected nodes $p(n_1,n_k) = n_1,n_2,\ldots,n_k$ where $n_i \in V$ and $(n_i,n_{i+1}) \in E$. A weighted tree has edges with associated real numbers as weights $w(e); e \in E$.

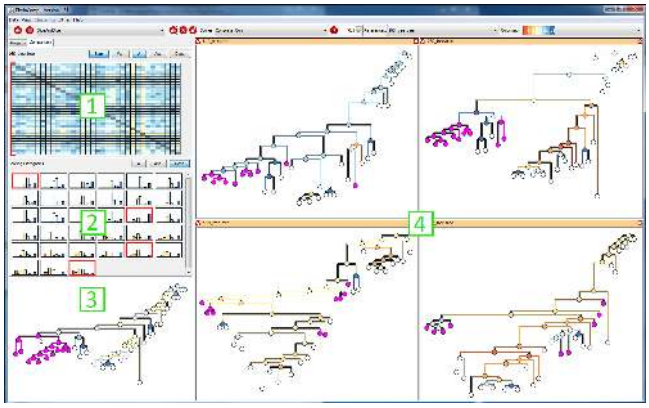


Figure 3: Overview of our approach showing a visualization of multiple levels of detail for tree comparison. 1) Global pairwise tree distance matrix. 2) Score distribution for 1:n tree comparison. 3) Consensus tree for 1:n comparison with average matching scores. 4) Selected trees with scores showing similarity to the reference tree. Selected nodes (pink) are highlighted in all views.

Note that in this paper, we use the term hierarchy synonymously to rooted tree. *Leaf nodes* are nodes, that are connected only to one node. They do not have any child nodes. The set of all leaf nodes of a tree T is denoted as $L(T)$. Non-leaf nodes are referred to as *inner-nodes*. A *binary* rooted tree, is a tree where each node apart from leaves has two child nodes (descendants). A *subtree* $T^n \subset T$ of a tree T is a tree consisting of a node $n \in T$ and all of its descendants in T. The subtree corresponding to the root node is the entire tree.

We denote the *distance* of two trees T_1 and T_2 as $d(T_1,T_2)$ with $0 \leq d(T_1,T_2) \leq 1$, so that small distances (close to 0.0) reflect high similarity of the two trees. The *similarity* is defined as $s(T_1,T_2) = 1.0 - d(T_1,T_2)$. The similarity of two nodes $n_1 \in T_1$ and $n_2 \in T_2$ is defined as the similarity of the subtrees $T_1^{n_1}$ and $T_2^{n_2}$ rooted at nodes n_1 and n_2, respectively. The *score* of the node n is defined as the maximum similarity to all nodes in the compared tree T_2.

The distance of two nodes n_1, n_2 in the same tree is defined as the length of the path connecting them. $d(n_1,n_2) = |p(n_1,n_2)| = |e_i|, e_i \in p(n_1,n_2)$. The weighted distance is defined as: $wd(n_1,n_2) = \sum w(e_i), e_i \in p(n_1,n_2)$.

We define *elements* of a tree as a set of all leaf sets $L(T^n)$ of all subtrees $T^n \subset T$: $Elements(T) = \{\{L(T^n)\}, \forall n \in T\}$.

3.1.2 Leaf-Based Measure

Leaf-based approaches measure the similarity of trees T_1, T_2 based on their contained leaves $L(T_1)$ and $L(T_2)$. We employ a normalized variant of the Robinson-Foulds distance following the strategy of Munzner and Guimbretiere introduced in [27]:
$s(T_1,T_2) = \frac{|L(T_1) \cap L(T_2)|}{|L(T_1) \cup L(T_2)|}$.

As only the leaves are included in the score calculation, the tree structure is ignored. Consequently, two (sub)trees containing the same leafs are classified as similar even if their structure may differ substantially. An example can be seen in Figure 4(a), where all roots have the maximal score. In this case, however, the internal structure of the trees differs significantly. Therefore, in our approach, we calculate the global similarity of two phylogenetic trees as the average score of all subtrees rooted in the inner nodes.

3.1.3 Element-Based Measure

We present a new element-based score, which extends the leaf-based measure, so that it reflects the inner structure of the tree. In contrast to a leaf-based scoring, the inner-nodes of a tree are incorporated in the score.

$s(T_1,T_2) = \frac{|Elements(T_1) \cap Elements(T_2)|}{|Elements(T_1) \cup Elements(T_2)|}$, where

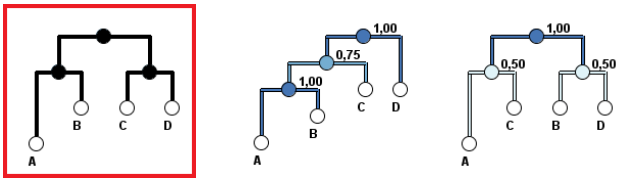
$Elements(T_i) = \{\{L(T^n)\}, \forall n_i \in T_i\}$, for $i \in \{1,2\}$

As an example, the comparison score of the roots of the left (T_l) and center (T_c) tree in Figure 4(b) is shown.

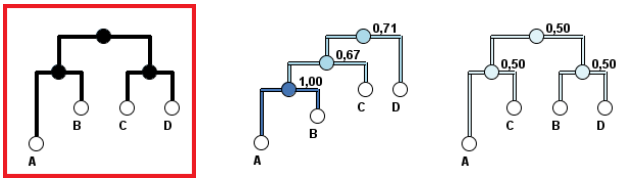
$T_l : \{A,B,C,D,[A,B][C,D]\}$; $T_c : \{A,B,C,D,[A,B][A,B,C]\}$;

$s(T_l,T_c) = \frac{5}{7} \approx 0,71$

This score discriminates structural and node-based differences between trees more profoundly than the leaf-based score (see Figure 8). In many cases, as exemplarily shown in our use case, the score distribution is less skewed and the scores more homogeneously distributed across the value range. This allows for better discrimination of structural differences based on score values. Moreover, as we include also inner-nodes in the calculation, no special score for the root nodes in trees containing the same organisms is needed (see Figures 7 and 4(b)).



(a) Leaf-based score



(b) Element-based score

Figure 4: This figure shows the comparison of two similarity measures. The scores measure similarity of a reference tree (left, with red border) to two other trees (center and right). a) The scores are calculated using the leaf-based measure (Sec. 3.1.2). b) The scores according to element-based measure. The score does not only include the leaves, it also considers the inner-nodes of the subtrees. The color map used is shown in Figure 6.

3.1.4 Edge-Based Measure

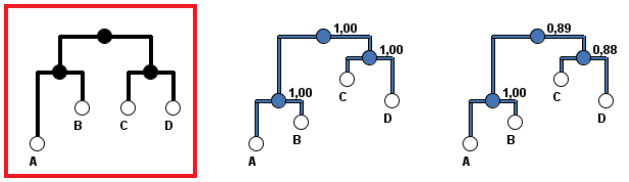
Scores, which rely on nodes only (leaf-based and element-based scores), typically do not take into account the differences in the edge lengths (i.e. weights). However, the edge length encodes important biological information such as evolutionary similarity of species. Therefore, we include a weighted edge-based score, which measures the difference between the sum of path lengths between all pairs of leaves in the compared trees (see Figure 5). This measure is inspired by the approach of Steel & Penny [32], who proposed a metric based on the difference of the number of edges connecting leaves.

$s(T_1,T_2) = 1.0 - \frac{ed(T_1,T_2)}{maxwd(T_1,T_2)}$, where

$ed(T_1,T_2) = \sqrt{\sum (wd(n_i^1,n_j^1) - wd(n_i^2,n_j^2))^2}$, and

$maxwd(T_1,T_2) = \max\{\max\{wd(n_i^1,n_j^1)\}, \max\{wd(n_i^2,n_j^2)\}\}$,

$\forall n_i^1, n_j^1 \in L(T_1), n_i^1 \neq n_j^1$ and $\forall n_i^2, n_j^2 \in L(T_2), n_i^2 \neq n_j^2$



(a) Reference tree (b) Element-based score (c) Edge-based score

Figure 5: A comparison of element-based and edge-based score for measuring similarity between a reference tree (left, red border) and another tree (center, right). a) Element-based scores. Due to equal elements, both trees are regarded as similar. b) An edge-based scoring reveals the differences in edge length.

In this way, differences between trees containing the same elements in the same structure, but with other edge weight (i.e., length), can be revealed (see Figure 5). One disadvantage of the score is its normalization by the longest weighted path in both trees. As discussed in [32], this leads in large trees often to similarity scores close to 1.0 (see Figure 7(b)). Moreover, this score does not allow for determination of best matching nodes between two trees as only whole trees can be compared. Therefore, the score of a subtree is calculated according to to the corresponding path lengths in the whole compared tree.

3.1.5 Summary

The scores described above capture different tree properties. The proposed element-based score reflects the tree structure and provides a good score distribution. Leaf-based score is useful only in particular cases when global grouping of the leaves is of interest, solely. The edge-based measure captures a signatures of the general structure, and can accommodate edge weights. Nevertheless, its normalization and thereby the scalability to larger trees is an open problem.

When using efficient implementation, in optimal case, all scores exhibit the same computational complexity of $O(|V|^2)$ [25, 27, 32] which is optimal for a pairwise comparison of all subtrees [24].

3.2 Interactive Visualization

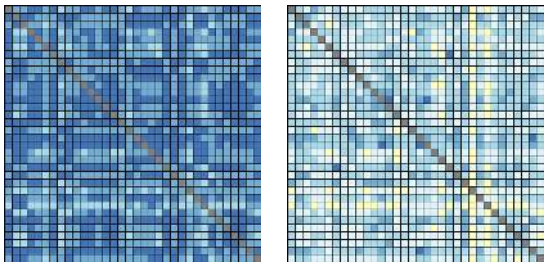
To ensure scalability, several visualizations representing different levels of abstraction are integrated in the phylogenetic tree comparison system (see Figure 3). It thereby enables the user to compare whole trees, subtrees or individual nodes. The visualizations are linked for interactive highlighting of interesting tree parts in all views. The used calculation methods and visual attributes (e.g., colormaps) are consistent in all visual representations and can be chosen interactively. The tree visualization technique was chosen so that it best fits the demands of the biologic use case. As we focused on phylogenetic trees, we employ the commonly used node-link technique reflecting the edge weights [7].

In the following, we present the data views in detail. We use the color map shown in Figure 6, where red represents low and blue high similarity. After testing various color maps from www.colorbrewer.org, we decided to use this color map as it well highlights high and low scores and offers a good score discrimination.

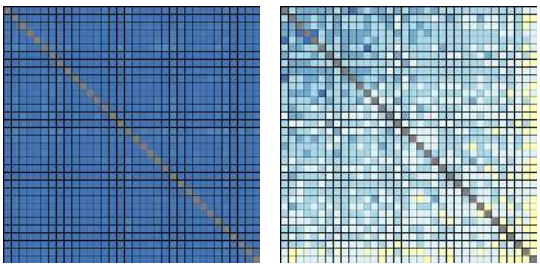
Figure 6: Color map used in the paper. Colors based on www.ColorBrewer.org, by Cynthia A. Brewer, Penn State.

3.2.1 Comparison Overview

An overview of pairwise similarities of all trees in the test data set is presented by the similarity matrix. We included this representation as it is an established approach in biology, presenting a familiar and easy to grasp overview (see Figure 2). Every row/column represents one tree and the cells encode the global similarity evaluation of tree pairs. This view allows for an overview of global similarities among trees and thereby, offers the possibility to select a reference tree for more detailed 1:*n* tree comparison in other views.



(a) Leave-based score (b) Element-based score



(c) Edge-based score (d) Element-based score, sorted

Figure 7: The similarity matrix color codes the pairwise global similarity of the compared trees based on a selected score. a) Leaf-based score b) Edge-based score c) Element-based score d) Element-based score with descending matrix sorting.

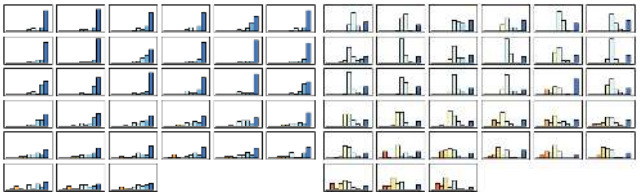
Additionally, the matrix can be sorted according to the sum of the scores (Figure 7). In the sorted map, the most (dis)similar trees to other trees stand out on the sides of the matrix (e.g., two right columns represent most dissimilar trees to all others). To improve the matrix visualization, we also included the variance of the scores to indicate possible local differences. The variance matrix is shown alternatively to the score matrix.

3.2.2 Score Distribution View

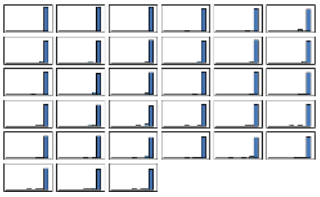
The overview matrix offers only one global score per tree comparison. Therefore, we included a more detailed histogram view on the score distribution in the tree comparison. The histograms offer a compact overview of the score distribution of all nodes in each tree when compared with the selected reference tree (see Figure 8). These views allow the detection of distinct score distribution patterns such as predominantly high scores, bi-modal score distributions (many low and many high scores), or trees with high variance of scores.

The histograms serve also for comparison of various score measures for the analyzed data set (see Section 3.1). Figure 8 shows a comparison of the score distributions for the three used scores. The element-based score shows better discriminative power for tree comparison as the other two measures. For example, the score distribution of the leaf-based measure is strongly skewed towards high scores. In contrast, the new element-based measure has more widely distributed scores.

The histogram view is also used for selecting a subset of trees for a detailed pairwise comparison with the reference tree (see Section 3.2.4).



(a) Leaf-based score (b) Element-based score



(c) Edge-based score

Figure 8: This figure shows the histograms of the scores of the elements of all trees compared with one reference tree, in descending order. The scores are calculated using (a) the leaf-based, (b) the element-based, and (c) and the edge-based score.

3.2.3 Consensus Tree View

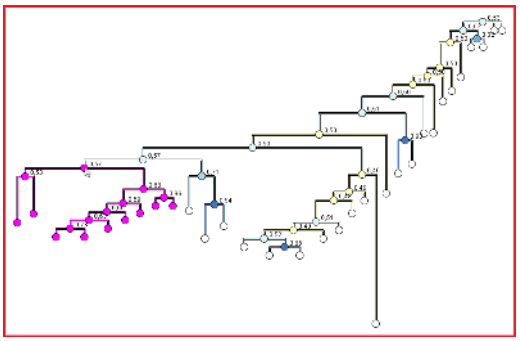
The *consensus tree* provides a compact form of a 1:*n* comparison between one reference tree and all other trees (see Figure 9a). This view shows the conservation of the nodes of the reference tree throughout the data set. Each score is the average of the scores comparing a reference tree node against its best matching unit in all other trees. High scores mean good match across the data set. In this way, the biologist can immediately see, which sub-hierarchies are conserved (Figure 9a).

To gain more details on the composition of the calculated score of a selected node, its distribution in the data set needs to be analyzed. Therefore, we included an interactive functionality, which allows the user to highlight their values in the data set on demand. In particular, the background of all histograms of the compared trees are color coded according to their similarity to a selected node in the reference tree (see Figure 9b)). This is a powerful tool to quickly detect trees with either similar or dissimilar sub-hierarchies.

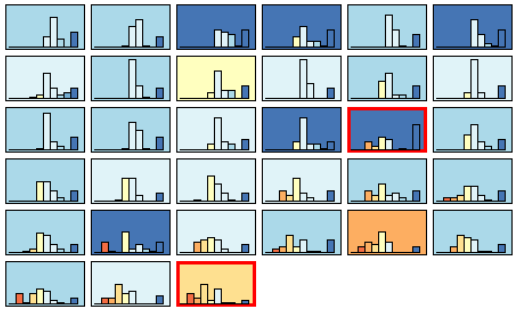
3.2.4 Tree Comparison View

The tree comparison view is the most detailed view in our framework. It contains the comparison tree scored against the reference tree shown in the consensus tree. We extended the well known pairwise comparison view with interactive functions for a better analysis of the data (see Figure 3 (3) and (4)). The used scoring scheme can be chosen from the proposed set of measures (see Section 3.1).

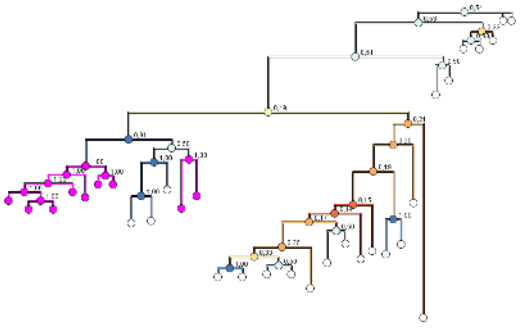
The linked views of tree comparisons allow for highlighting of selected structures and their best matches in all visible trees. This feature has been proven useful for comparing trees [27]. In our analysis setting (comparison of phylogenetic trees), it reveals the particular distribution of the organisms. Thereby, it provides additional information on phylogenetic tree similarity (see Figure 9). The figure shows a reference tree with a highlighted subtree rooted in a user-selected node (pink color). The best matching nodes of the subtree are highlighted in the two compared trees. As expected, the highlighted elements are much more distributed in the dissimilar tree and compact in the similar tree.



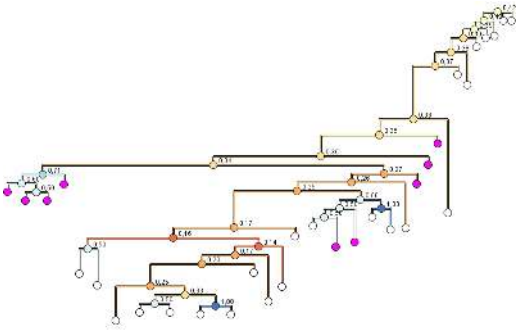
(a) Reference tree



(b) Colored histograms



(c) Compared similar tree

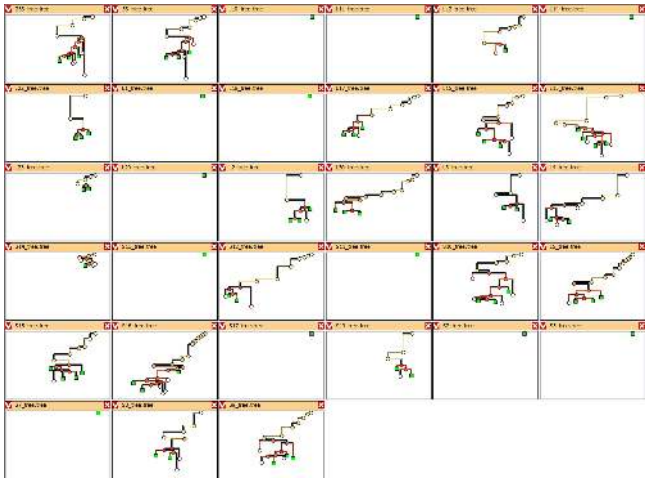


(d) Compared dissimilar tree

Figure 9: Tree comparison view with highlighting. a) The reference tree. One of its subtrees is highlighted by the user (in pink). b) Histogram backgrounds are colored according to selected node in reference tree. c) The compared overall most similar tree. d) The compared most dissimilar tree. c) & d) The leaves of the subtree highlighted in the reference tree are highlighted in the compared trees. Highlighted elements are more diversified in the dissimilar tree and close in the similar tree.

To reduce the complexity of the visualization and improve scalability, the tree representation can be simplified [5]. In our case, those subtrees are collapsed that have element scores below a user-defined threshold (see Figure 10, green rectangles). By this, structures similar to the reference tree are hidden and dissimilarities are pointed out. In analogy, all subtrees above a certain threshold can be collapsed. This allows for focused analysis of similarities among trees. Additionally, branches that do not lead to collapsed nodes can be hidden. The threshold slider allows for interactive setting of the threshold and exploration of the effects of threshold change on the tree collapsing.

(a) Original tree view



(b) Collapsed tree view

Figure 10: Tree comparison view with collapsing of similar elements. a) The original view with all nodes visible. b) The collapsed view, where all elements below a user-defined threshold are collapsed for focused view on tree dissimilarities and better scalability of the view. The green rectangles represent the collapsed nodes.

4 Application to Ribosomal Phylogenies

The application section provides an insight into the usage of the proposed approach for visual comparison of multiple phylogenetic trees in current biologic research activities. The use case has been provided by co-authors from biology.

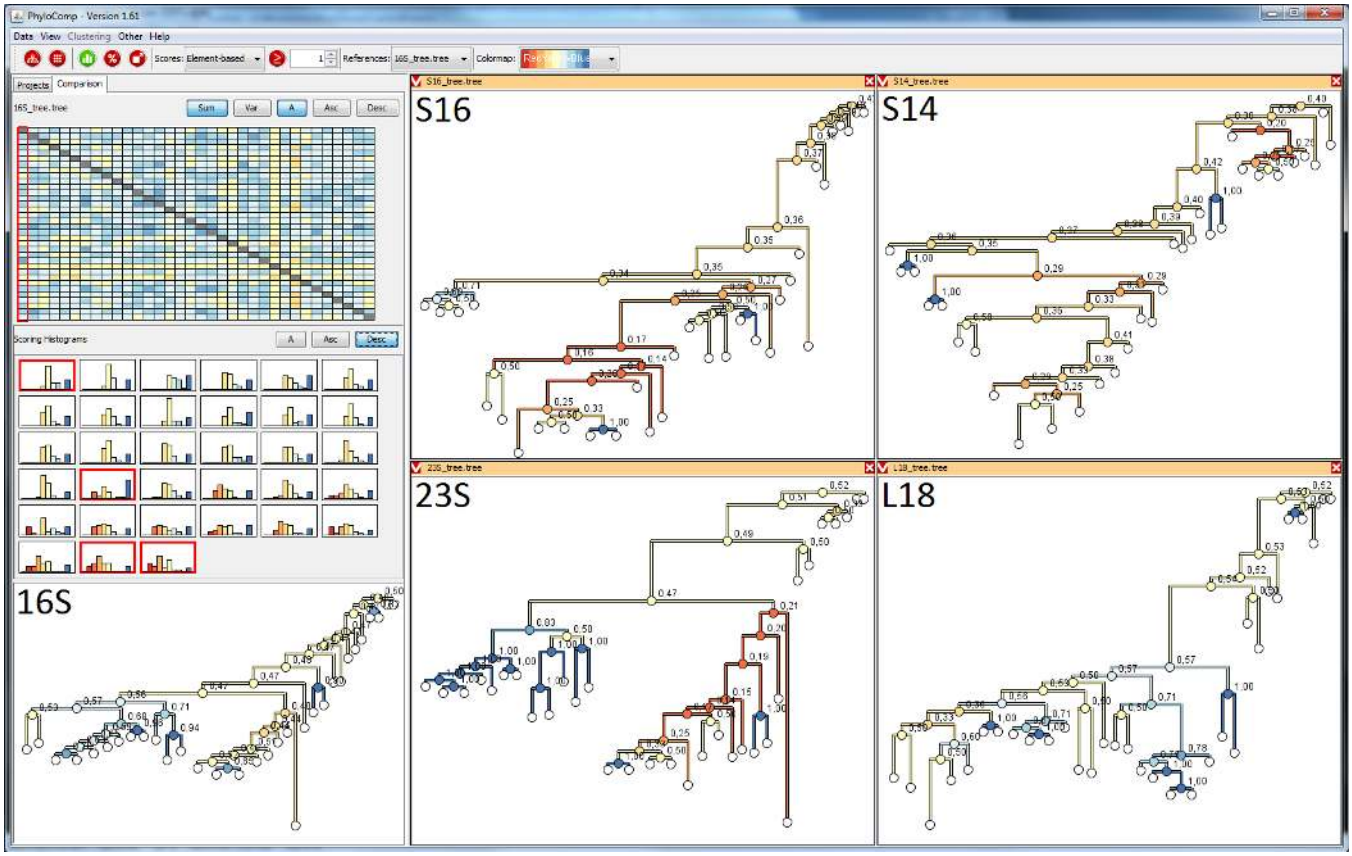


Figure 11: User interface in the application using 34 phylogenetic trees with 32 organisms each. The $16S$ tree was chosen as a reference. Four trees ($L18$, $S14$, $S16$ and $23S$) were chosen for detailed comparison. The selected trees are highlighted in histogram view with red border.

Biologists focus on the analysis of evolutionary relationships among a selection of bacteria. The relationship between the selected bacteria is commonly determined by comparison of the 16*S* rRNA sequences. This comparison leads to the phylogeny approved by the biologic community. Nevertheless such sequence comparison can also be conducted using other sources such as conserved proteins, acknowledged to aid in finding the correct phylogeny. The ribosome is a molecular structure ubiquitous distributed in all living organism. To this respect, our co-workers are interested in the analysis of the effect of the data source on the resulting phylogenies. Thereby, they solve the following **tasks**:

T1 Identification of *globally interesting* trees for reference purposes,

and determination of various *patterns in subhierarchies* of the trees:

T2 The conservation of every subhierachy of the reference tree in all other trees.

T3 Distribution of a selected subhierarchy throughout the dataset.

T4 Trees with high global similarity but low local similarity and vice versa.

T5 All (dis)similar structures in all compared trees with respect to the reference.

The comparison of phylogenetic trees is a starting point for further biological analysis. As this analysis is very time and resource consuming, the biologists need to concentrate on the most promising starting points. The presented tree comparison tool aids biologists to track down these highly promising pointers even in large data sets and to form working hypotheses for research.

4.1 Phylogeny Calculation

The relationship of two organisms is calculated on a common data source. Choosing an appropriate data source for phylogenetic tree calculation is however still an area of ongoing research. In our use case, we have used sequences encoding different parts of the ribosome namely rRNA and proteins.

The input data consists of bacterial genomic data downloaded from GenBank data base (GbDB). The biologists extracted the desired sequences encoding ribosomal parts out of the whole genomes using BioPython [6] and Hidden-Markov-Models (HMM) in the PFAM data base [10]. To calculate the phylogenetic trees, they used the following web based tools of phylogeny.fr [7]. The alignments were performed using the MUSCLE [8] algorithm. After the so-called "curation" of the alignments with GBLOCKS [4] phylogenetic trees have been computed with PhyML [1, 13].

This resulted in 34 phylogenetic trees with 32 organisms each. The trees are named according to the ribosomal parts encoded by source sequence (e.g, 16*S*, *L*4, ...). The employed naming convention is: letters before numbers for proteins and letters after numbers for rRNA. To confirm the suitability of the data sources used, the biologists compared the phylogeny of our calculated 16*S* data to a larger data set already validated by the biologic community [38]. Our subset is in very good agreement with the one acknowledged by the experts. Thus the 16*S* phylogenetic tree can be used for the comparative study described in this section.

4.2 Results from Visual Analysis Process

The presented analysis process started with the selection of a reference tree based on the global similarity of the trees **(T1)**. The comparison of the implemented scores in the overview matrix showed that the element-based score exhibited the best discrimination properties (see Fig. 7 and 8). Therefore, it was used in the following analysis. Potentially interesting reference trees are usually characterized by high or low scores compared to all other trees. These trees were identified in the sorted matrix (Fig. 7 (d)). Following this track, one would choose the *S*16 tree, because it showed up as the most distant tree. However, in this study, we concentrated on tree 16*S* as it refers to a biologically-validated phylogeny [38]. Looking at the global scores in the matrix column corresponding to this reference tree (Fig. 7 (b)), the biologist found out that the derived phylogenetic tree strongly depends on the underlying data set.

The sorted histograms below the matrix allowed the analysts to identify interesting patterns in the score distribution for the elements of each compared tree **(T4)**. Initially, their focus was on the most dissimilar trees. They identified the *S*14 and *S*16 ribosomal proteins as two of the most deviating trees to the reference tree (Figure 12 (d)). Moreover, these trees are most dissimilar to all other trees as well (the two right columns in Figure 7 (d)). This was an interesting finding, because the *S*14 protein is involved in the assembly of the ribosome suggesting a high conserved function and thus a highly conserved sequence [18]. Additionally, the *S*14 tree exhibited a bi-modal score distribution **(T3)**. The distribution was characterized by a large amount of low- and high-scored elements (Figure 12 (d)). It therefore indicated similar and different structures to the reference tree inside the same tree.

The consensus tree offers the possibility to analyze the stability of subtrees across the dataset **(T2)**. Examination of the consensus tree identified two interesting, highly conserved clusters. Such a finding raised attention, because it contradicts the initial statement that "phylogeny computations strongly depend on the underlying data set". The closer analysis of the distribution of the two selected clusters **(T3)** in the trees revealed outliers with a low conservation. Besides the already conspicuous *S*14 and *S*16, the *L*18 tree attracted attention (Figure 12 (a) & (b)). Despite its high global similarity to the reference, it had a very low conservation for both selected substructures (Figure 12).

The identification of extrema throughout the whole data set **(T5)** can only be seen when all trees are simultaneously displayed. For a large number of trees, this led to a cluttered view (see Figure 10 (a)). To reduce the complexity of the visualization, the biologist hid all substructures above a similarity of 0.25. This clearly reveals the most dissimilar structures compared to the reference tree (see Figure 10 (b)). Their closer analysis showed that *Xanthomonas campestris*, a plant pathogen, was present in the opened leaves of most trees. This indicated that the relative position of *Xanthomonas campestris* differed significantly from the approved 16*S* reference tree throughout the dataset.

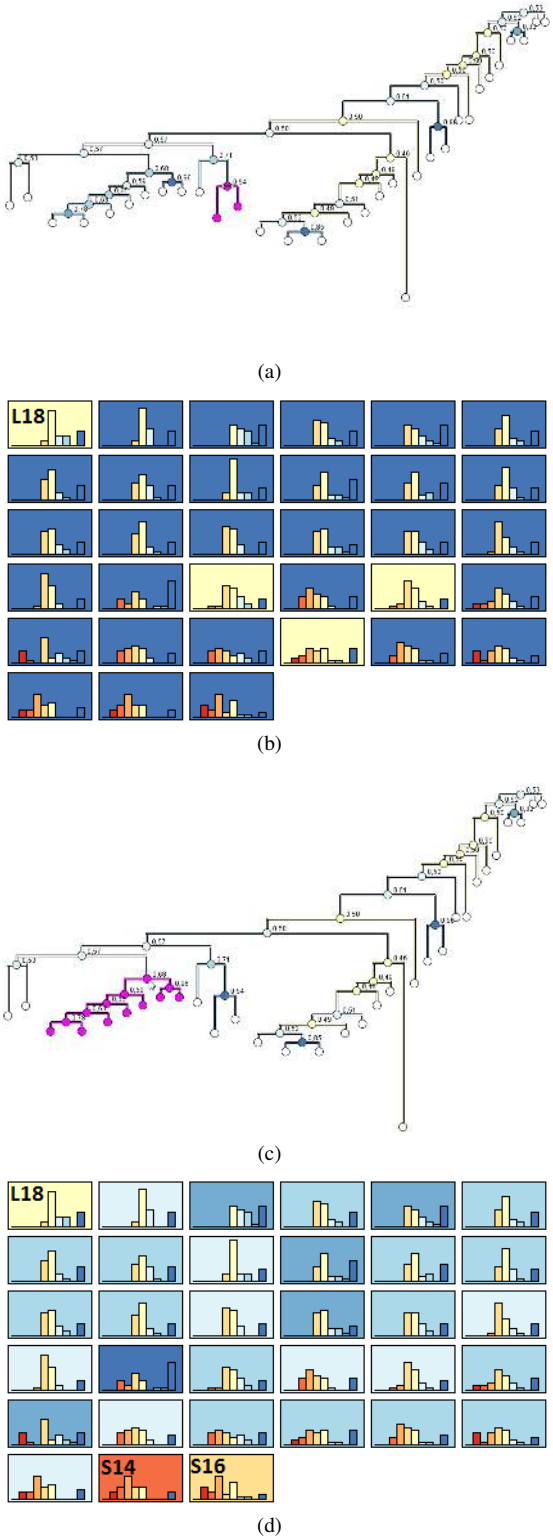
Our study led biologists to take a closer look on function and structure of the ribosomal proteins *S*14, *S*16 and *L*18 for the identified species. They now especially concentrate on the question, why the phylogenies resulting from those protein sequences exhibited such organism clusters. This research could lead to previously unknown connections between the analyzed species.

5 Discussion

The presented approach allows for comparison of multiple trees on several levels of detail. This is enabled via a combination of algorithmic analysis based on a new tree similarity score and interactive visualization employing linked views. As shown, this work is very useful for analysis of multiple trees in biologic research, it however has some limitations and potential for extensions.

In our work, we focused on rooted trees with the identical leaf nodes. This was inspired by the underlying biologic application. For comparison of rooted trees with varying sets of leaf nodes, the employed similarity scores need to be extended. Note that although the use case considered binary trees, the proposed similarity score and visualization are applicable also to general trees.

In order to address the scalability issues, we developed an approach analyzing trees on multiple levels. The scalability with regard to the number of trees is reflected in the selection of the refer-



(a) (b) (c) (d)

Figure 12: 16*S* consensus tree and score histograms with background coloring according to the selected elements. (a) & (b) The selection of a highly conserved pair of organisms. The globally similar *L*18 tree does not contain the marked substructure as one of few. (c) & (d) The selection of a subtree of seven organisms and their scores across the data set.

ence tree and selection of the compared trees in the histogram view. In this way, the users can focus on details of only a small set of important trees to be compared. This approach, however, relies on pairwise global tree comparison, which has quadratic complexity both with regard to the number of trees as well as the number of leaf nodes (i.e., tree size). The selection of the reference tree plays an important role in the analysis. It is supported by a similarity matrix. Given the time constraints, it is assumed that only a small set of reference trees is selected by the user for the analysis. Another limitation regards the number and the size of trees visualized for detailed comparison given the available screen size. As shown, this constraint can be diminished by similarity-based tree simplification (see collapsing of nodes in Section 3.2.4). It thereby allows for simultaneous comparison of dozens of selected trees with hundreds of leafs. Our approach however needs to be extended to accommodate for large trees with thousands of nodes.

6 Conclusion and Future Work

We presented a new approach for visual comparison of multiple trees. It supports evaluating both global and local patterns among trees. We combined algorithmic calculation with interactive data visualization. The computation of (sub)tree similarity is based on a new distance measure, which takes into account tree structure and enhances discrimination in a large set of trees. The interactive visualization encompasses multiple views on similarity providing several abstraction levels – global tree, subtrees, and individual nodes.

We applied our approach to biological data on evolutionary relationship of bacteria. The new method is useful for various analytical tasks such as identification of evolutionary stable organism relations. As it allows to compare patterns in multiple trees simultaneously, it triggers new hypotheses about evolutionary ancestry between organisms, which could not be found using only pairwise comparison. The presented approach can be used utilized not only for comparing phylogenetic trees but also for other problems such as evaluating results of hierarchic clustering with differing parameter settings.

Our work can be extended in various ways. In our use case, the tree roots were determined by the phylogeny algorithm. According to the biologists, it would also be interesting to be able assessing the effect of the root position on tree similarity. It will allow them to analyze new phylogenetic hypotheses. Therefore, we would like to extend our approach to the comparison of various tree roots. Moreover, we would like to adapt our approach to $n-$ary trees, as well as to accommodate for trees with varying sets of leaf nodes. This would allow us to cover a wider range of use cases. We would also like to test our system with data sets of various sizes in order to assess its limitations with respect to the number of trees and tree sizes.

Acknowledgements

We are grateful to the Fonds der chemischen Industrie for their financial support. This work was partly supported by the German Research Foundation (DFG) through the Graduate School 1657 [26] and the Strategic Research Initiative on Scalable Visual Analytics (SPP 1335) [21].

References

[1] M. Anisimova and O. Gascuel. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Systematic Biology*, 55(4):539–552, Aug. 2006. PMID: 16785212.

[2] C. Appert and J.-D. Fekete. Naviguer dans des grands arbres avec controltree. In *Proceedings of International Conference of the Association Francophone d'Interaction Homme-Machine*, pages 139–142, 2007.

[3] S. K. Card and D. Nation. Degree-of-interest trees: a component of an attention-reactive user interface. In *AVI '02: Proceedings of the*

Working Conference on Advanced Visual Interfaces, pages 231–245, New York, NY, USA, 2002. ACM.

[4] J. Castresana. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution*, 17(4):540–552, Apr. 2000. PMID: 10742046.

[5] J. Chen, A. MacEachren, and D. Peuquet. Constructing Overview+ Detail Dendrogram-Matrix Views. *Visualization and Computer Graphics, IEEE Transactions on*, 15(6):889–896, 2009.

[6] P. J. A. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, and M. J. L. de Hoon. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics (Oxford, England)*, 25(11):1422–1423, June 2009. PMID: 19304878.

[7] A. Dereeper, V. Guignon, G. Blanc, S. Audic, S. Buffet, F. Chevenet, J. Dufayard, S. Guindon, V. Lefort, M. Lescot, J. Claverie, and O. Gascuel. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Research*, 36(Web Server):W465–W469, 2008.

[8] R. C. Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792 –1797, Mar. 2004.

[9] N. Elmqvist and J.-D. Fekete. Hierarchical aggregation for information visualization: Overview, techniques, and design guidelines. *IEEE Transactions on Visualization and Computer Graphics*, 99, 2009.

[10] R. D. Finn, J. Tate, J. Mistry, P. C. Coggill, S. J. Sammut, H. Hotz, G. Ceric, K. Forslund, S. R. Eddy, E. L. L. Sonnhammer, and A. Bateman. The pfam protein families database. *Nucleic Acids Research*, 36(Database):D281–D288, 2007.

[11] M. Freire, C. Plaisant, B. Shneiderman, and J. Golbeck. ManyNets: an interface for multiple network analysis and visualization. In *Proceedings of international conference on Human factors in computing systems*, pages 213–222, New York, NY, USA, 2010. ACM.

[12] M. Graham and J. Kennedy. A survey of multiple tree visualisation. *Information Visualization*, 9(4):235–252, 2009.

[13] S. Guindon and O. Gascuel. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, 52(5):696–704, Oct. 2003. PMID: 14530136.

[14] K. Hamacher. Protein domain phylogenies - information theory and evolutionary dynamics. In A. Fred, J. Filipe, and H. Gamboa, editors, *Bioinformatics*, pages 114–122, 2010.

[15] J. Heer and S. K. Card. Doitrees revisited: scalable, space-constrained visualization of hierarchical data. In *AVI '04: Proceedings of the working conference on Advanced visual interfaces*, pages 421–424, New York, NY, USA, 2004. ACM.

[16] D. Hillis, T. Heath, and K. John. Analysis and visualization of tree space. *Systematic Biology*, 54(3):471, 2005.

[17] D. Holten and J. Van Wijk. Visual comparison of hierarchically organized data. *Computer Graphics Forum*, 27(3):759–766, 2008.

[18] N. Hulo, A. Bairoch, V. Bulliard, L. Cerutti, B. A. Cuche, E. de Castro, C. Lachaize, P. S. Langendijk-Genevaux, and C. J. A. Sigrist. The 20 years of PROSITE. *Nucleic Acids Research*, 36(Database issue):D245–D249, Jan. 2008. PMID: 18003654 PMCID: 2238851.

[19] D. Huson, D. Richter, C. Rausch, T. Dezulian, M. Franz, and R. Rupp. Dendroscope: An interactive viewer for large phylogenetic trees. *Bmc Bioinformatics*, 8(1):460, 2007.

[20] Institute of Evolutionary Biology, University of Edinburgh. Figtree. `http://tree.bio.ed.ac.uk/software/figtree/`.

[21] D. Keim, T. Ertl, H. Ritter, G. Weikum, and S. Wrobel. Strategic Research Initiative 1335: Scalable Visual Analytics - Interactive Visual Analysis Systems of Complex Information Spaces. `http://www.visualanalytics.de/`.

[22] I. Letunic and P. Bork. Interactive tree of life (itol): an online tool for phylogenetic tree display and annotation. *Bioinformatics*, 23(1):127, 2006.

[23] M. Li, J. Badger, C. Xin, S. Kwong, and P. e. Kearney. An information based sequence distance and its application to whole mitochondrial genome phylogeny. *Bioinformatics*, 17:149–154, 2001.

[24] Y. Lin and T. Hsu. Efficient algorithms for descendent subtrees comparison of phylogenetic trees with applications to co-evolutionary classifications in bacterial genome. *Algorithms and Computation*, pages 339–351, 2003.

[25] Y.-L. Lin and T.-S. Hsu. Efficient algorithms for descendent subtrees comparison of phylogenetic trees with applications to co-evolutionary classifications in bacterial genome. In T. Ibaraki, N. Katoh, and H. Ono, editors, *Algorithms and Computation*, volume 2906 of *Lecture Notes in Computer Science*, pages 339–351. Springer Berlin / Heidelberg, 2003.

[26] M. Löbrich and G. Thiel. DFG Graduiertenkolleg 1657 "Molecular and cellular reactions on ionizing radiation" . `http://www.grk1657.de/`.

[27] T. Munzner, F. Guimbretière, S. Tasiran, L. Zhang, and Y. Zhou. Treejuxtaposer: scalable tree comparison using focus+context with guaranteed visibility. *ACM Trans. Graph.*, 22:453–462, July 2003.

[28] T. Nye. Trees of trees: an approach to comparing multiple alternative phylogenies. *Systematic biology*, 57(5):785, 2008.

[29] S. Pape, F. Hoffgaard, and K. Hamacher. Distance-dependent classification of amino acids by information theory. *Proteins: Structure, Function, and Bioinformatics*, 78(10):2322–2328, 2010.

[30] P. Puigbo, S. Garcia-Valle, and J. O. McInerney. TOPD/FMTS: a new software to compare phylogenetic trees. *Bioinformatics*, 23(12):1556 –1558, June 2007.

[31] B. Shneiderman. Tree visualization with tree-maps: 2-d space-filling approach. *ACM Transactions on graphics (TOG)*, 11(1):92–99, 1992.

[32] M. Steel and D. Penny. Distributions of tree comparison metrics–some new results. *Systematic Biology*, 42(2):126–141, 1993.

[33] A. Telea and D. Auber. Code flows: Visualizing structural evolution of source code. In *Computer Graphics Forum*, volume 27, pages 831–838. Wiley Online Library, 2008.

[34] Y. Tu and H. Shen. Visualizing changes of hierarchical data using treemaps. *IEEE Transactions on Visualization and Computer Graphics*, pages 1286–1293, 2007.

[35] Y. Tu and H.-W. Shen. Balloon focus: a seamless multi-focus+context method for treemaps. *Visualization and Computer Graphics, IEEE Transactions on*, 14(6):1157–1164, Nov.-Dec. 2008.

[36] T. von Landesberger, M. Görner, and T. Schreck. Visual analysis of graphs with multiple connected components. In *Proceedings of IEEE Symposium on Visual Analytics Science and Technology*, pages 155–162, 2009.

[37] T. von Landesberger, A. Kuijper, T. Schreck, J. Kohlhammer, J. van Wijk, J.-D. Fekete, and D. Fellner. Visual analysis of large graphs: State-of-the-art and future research challenges. *Computer Graphics Forum*, pages no–no, 2011.

[38] D. Wu, P. Hugenholtz, K. Mavromatis, R. Pukall, E. Dalin, N. N. Ivanova, V. Kunin, L. Goodwin, M. Wu, B. J. Tindall, S. D. Hooper, A. Pati, A. Lykidis, S. Spring, I. J. Anderson, P. Dhaeseleer, A. Zemla, M. Singer, A. Lapidus, M. Nolan, A. Copeland, C. Han, F. Chen, J.-F. Cheng, S. Lucas, C. Kerfeld, E. Lang, S. Gronow, P. Chain, D. Bruce, E. M. Rubin, N. C. Kyrpides, H.-P. Klenk, and J. A. Eisen. A phylogeny-driven genomic encyclopaedia of bacteria and archaea. *Nature*, 462:1056 – 1060, 2009.

[39] Y. Zhong, C. Meacham, and S. Pramanik. A general method for tree-comparison based on subtree similarity and its use in a taxonomic database. *Biosystems*, 42(1):1–8, 1997.