Tools for Phylogenetic Postprocessing

by

Nicholas Dylan Pattengale
npcomplete@acm.org

B.S., New Mexico Institute of Mining and Technology, 2001

THESIS

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Master of Science
Computer Science

The University of New Mexico

Albuquerque, New Mexico

May, 2005

npcomplete@acm.org

Dedication

...to that which inspires and fascinates, and to those who leverage inspiration and fascination not only lovingly but toward fostering love itself.

Acknowledgments

Thanks to Bernard M.E. Moret, David A. Bader, Tanya Y. Berger-Wolf (committee).

Also,

- My friends and acquaintances who over the years have intellectually challenged and otherwise entertained me - John Barentine, John Dobson, Steven Goldsmith, Rick Mooney, Cris Moore, David Waggonner, Albert Yu (very incomplete list!).
- My family - Elizabeth, Paul, Kenneth, and Brendan Pattengale, Leona and R.W. Bachmayer, Martha and Paul F. Pattengale, Bingi
- My compass, caretaker y corazon - April Bree Hawkes-Pattengale.

Tools for Phylogenetic Postprocessing

by

Nicholas Dylan Pattengale
npcomplete@acm.org

ABSTRACT OF THESIS

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Master of Science
Computer Science

The University of New Mexico

Albuquerque, New Mexico

May, 2005

Tools for Phylogenetic Postprocessing

by

Nicholas Dylan Pattengale
npcomplete@acm.org

B.S., New Mexico Institute of Mining and Technology, 2001

M.S., Computer Science, University of New Mexico, 2005

Abstract

Phylogenetic reconstruction techniques often produce multiple, competing evolutionary hypotheses. The umbrella term *phylogenetic postprocessing* encompasses methods that attempt to reconcile the ambiguity. Three classes of phylogenetic postprocessing results are presented. (1) A generalized family of metrics on tree space is derived. The metrics can be equipped with sensitivity to edge weights. Two members of the family of metrics are the familiar Robinson-Foulds (RF) metric and the weighted Robinson-Foulds metric. (2) Standard consensus methods are augmented to take edge weight into consideration. A new consensus method based on edge weights is introduced. (3) A sublinear $(1+\epsilon)$ approximation algorithm is derived for computing the RF distance between two trees.

Contents

List of Figures

Chapter 1

Introduction

1.1 Phylogenetics: An Overview

Phylogenetics is the study of evolutionary relationships between organisms. The basic problem statement for computational phylogenetic reconstruction is: *What is the most plausible evolutionary history of a set of taxa*[1] *that are believed to share a common ancestor?*

Technology has enabled extraordinary progress in phylogenetics. One aspect of this is in biology itself. The discovery of DNA and the ability for biologists to sequence DNA has, to say the least, revolutionized the field. Computers have helped enormously as well. Aside from their aiding in sequencing tasks (we could not have assembled the sequence of the human genome without computers), computers facilitate a wide range of phylogenetic problem-solving activities.

The first major contribution of computers is their sheer speed. Realize that given a set of taxa, there are a finite number of ways in which they can be assembled into a

[1] taxa (singular is taxon) is short hand for taxonomical unit, the ambiguous term used to refer to the things that are assumed to be evolutionarily related and are under comparison.

phylogeny. Assuming that the phylogeny takes the form of a tree with n leaves, there are $(2n-5)!! = (2n-5)\cdot(2n-3)\cdot(2n-1)\ldots 5\cdot 3\cdot 1$ possible trees. For small datasets, a computer program can assess every single possible evolutionary hypothesis. This is a very powerful asset.

The second major contribution of computers (or perhaps more appropriately, computer science) is text processing. Since DNA sequences come as a sequence of characters (from the alphabet A,C,T,G), there are many computer science algorithms that can be used as processing tools.

For the sake of discussion assume that $(2n-5)!!$ is a small enough number (true today for $n \leq 15$) such that a computer program can consider every single tree. One of the $(2n-5)!!$ trees must be the real tree. So how to pick the correct one? There needs to be a way of scoring trees so that one can be preferred over another. All such scores, by necessity, are based on mathematical models of evolution or mathematical properties of trees.

Mathematically founded phylogenetic reconstruction methods are typically of three kinds: distance-based methods, likelihood methods, and parsimony methods. The reader is referred to [7, 22] for more detail about each approach. The unfortunate reality addressed in this thesis is that the methods are not guaranteed to produce a single tree. The gene-tree/species-tree dilemma is another reason why more than one tree may arise [30]. The need to reconcile multiple competing hypotheses yields requirements for *phylogenetic postprocessing.*

1.2 Phylogenetic Postprocessing

The basic problem statement is to take multiple trees as input and return "something" that illustrates the [dis]agreement among the input.

Perhaps the simplest of all tree comparisons is a distance function [2, 13, 26, 27]. Distance functions take two trees as input and return a number that indicates dissimilarity between the trees. Distance functions are used extensively in simulation studies to assess the accuracy of reconstruction techniques [34]. Distance functions are also used extensively in clustering techniques [29].

There is another class of postprocessing techniques called consensus methods [9, 12, 25]. Consensus methods take a set of trees defined on the same set of taxa and return a single tree, the so-called consensus tree. There are many consensus methods, each distinguished by the set of properties that it cares about retaining.

The requirement for consensus methods to return a single tree may be overly constraining [4, 14, 24, 28]. Thus there are relaxations of the problem that are allowed to return sets of trees (with hopefully fewer members) [6, 32, 33].

Finally, visualization techniques have been developed [3, 11, 16].

1.3 Road Map

The original inspiration for this work came while I was attending a semester-long seminar in computational molecular biology. After learning about parsimony methods I pondered "nature tends toward optimality, but does not require it." Since parsimony is an optimization problem, why not also consider a set of slightly suboptimal trees? This perspective yielded a back door into the need for phylogenetic postprocessing.

A first attempt (before knowledge of things like consensus methods) was to identify collapseable subtrees. A colleague suggested looking at a paper about similarity matching of XML trees [15]. The tree-matching paper uses techniques from high-dimensional geometry [17, 23] by representing trees with "vector sketches." The

vector sketches are constructed such that standard vector-difference operations yield good approximations of "edit distance."

By representing phylogenetic trees as vectors (Chapter 2), it is possible to compute standard distance functions between trees by performing standard vector-difference operations (Chapter 3).

I then learned about consensus methods and was struck to find that they do not, in general, make use of edge weight. This is strange as the score of a tree (the quantity being optimized in a phylogenetic reconstruction) is typically the sum of the edge weights in the tree. A few ways of incorporating edge weights into consensus methods are presented (Chapter 4).

Finally, I made a connection between phylogenetics and another technique from high-dimensional geometry. A technique based on *randomized embedding* of vectors can be used to compact the vectors while maintaining certain properties [1, 17]. The technique yields an asymptotic speedup on a commonly computed distance function between phylogenetic trees (Chapter 5).

Chapter 2

Representing Phylogenetic Trees as Vectors

2.1 Tree Vector Definitions

The set of all possible unrooted, leaf-labeled trees on n taxa is denoted $\mathcal{T}_n$. As mentioned previously $|\mathcal{T}_n| = (2n-5)!!$.

Notice that removing an edge in a phylogenetic tree splits the set of taxa in two. An edge is uniquely identified by the split that it induces. There are

$$b = \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \binom{n}{i} \approx 2^{n-1}$$

ways to split a set of taxa in two. By assigning indices to splits it is possible to represent trees as vectors. Denote the set of splits induced by the edges in tree T as $\Sigma(T)$. Assign indices to splits by using the function

$$f : \bigcup_{T \in \mathcal{T}_n} \Sigma(T) \to \mathbb{N}$$

defined as

$$f(S) = \sum_{i=1}^{|S|-1} \binom{n}{i} + \sum_{j=|S|}^{1} \binom{S[j]-1}{j} + 1$$

where S is a sorted array of the taxa on the side of the split with smaller cardinality. This function [19] simply assigns splits to unique integers on the interval $[1, b]$.

Definition 2.1.1. *The bit-vector representation of a phylogenetic tree T is $v_T \in \mathbb{R}^b$ where each element of v_T is taken as*

$$v_T[i] = \begin{cases} 1 & \text{if } f^{-1}(i) \in T \\ 0 & \text{otherwise} \end{cases}$$

Lemma 2.1.1. *$\forall T \in \mathcal{T}_n$, v_T as defined in definition 2.1.1 is a unique point in $\mathbb{R}^b$.*

Proof. For every pair of trees T_A and T_B in $\mathcal{T}_n$, in order to have $T_A \neq T_B$ one of the trees must have at least one split that the other does not. Call the obligatory split in which they differ S. For the tree that contains S, the element in its bit-vector representation with index $f(S)$ will equal 1. Element $f(S)$ in the other tree will equal 0. □

Corollary 2.1.2. $\bigcup_{T \in \mathcal{T}_n} v_T \subset \mathbb{R}^b$

Most phylogenetic reconstruction methods assign edge weights to tree edges. Use the following definition if it is desirable to retain edge weights in the vector representation. Assume that edge weights for tree T are defined by $w_T : \Sigma(T) \to \mathbb{N}$.

Definition 2.1.2. *The edge-weight vector representation of a phylogenetic tree T is $v_T \in \mathbb{R}^b$ where each element of v_T is defined as*

$$v_T[i] = \begin{cases} w_T(f^{-1}(i)) & \text{if } f^{-1}(i) \in T \\ 0 & \text{otherwise} \end{cases}$$

Notice that tree vectors are very sparse (i.e. contain many zeros). The number of actual edges in a tree is bounded by $2n - 3$. The number of possible edges, b, grows asymptotically faster than the number of actual edges in any constructible tree. Thus tree vectors grow in sparsity as the number of taxa increases.

The following lemma establishes that the space required for a set of m tree vectors on n taxa, using a compact representation, is not prohibitive.

Lemma 2.1.3. *The space required for $\mathcal{T}$, a set of m tree vectors on n taxa, is $m \times |\bigcup_{T \in \mathcal{T}} \Sigma(T)|$.*

Proof. None of the methods presented will be affected by splits that occur in none of the members of $\mathcal{T}$. Thus they are irrelevant and can be left out of the representation. □

See figure 2.1 for an example tree in its bit-vector and weighted-vector representations.

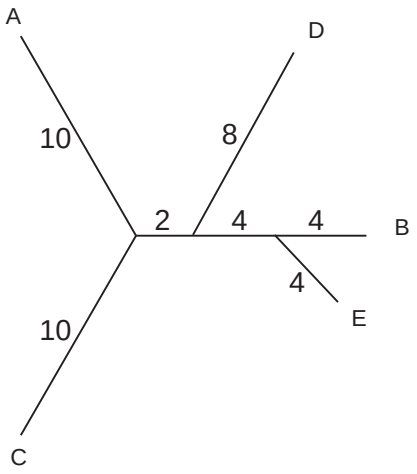
2.2 Summary Vectors

It will prove useful to summarize a set of tree vectors as a single vector containing summary statistics.

The following two definitions hold for both bit-vectors and weighted vectors.

Definition 2.2.1. *The mean weight vector of a set of tree vectors $\mathcal{V}$ is $v_\mu \in \mathbb{R}^b$, where each element of v_μ is defined as*

$$v_\mu[i] = \frac{\sum_{v_T \in \mathcal{V}} v_T[i]}{|\mathcal{V}|}$$



	A	B	C	D	E	AB	AC	AD	AE	BC	BD	BE	CD	CE	DF
$v_b =$	1	1	1	1	1	0	1	0	0	0	0	1	0	0	0
$v_w =$	10	4	10	8	4	0	2	0	0	0	0	4	0	0	0

Figure 2.1: Example of the bit-vector v_b and weighted vector v_w for the above tree.

Definition 2.2.2. *The edge-weight stability vector of a set of tree vectors $\mathcal{V}$ is $v_{\sigma^2} \in \mathbb{R}^b$, where each element of v_{σ^2} is defined as*

$$v_{\sigma^2}[i] = \frac{\sum_{v_T \in \mathcal{V}} (v_T[i] - v_\mu[i])^2}{|\mathcal{V}|}$$

See figure 2.2 for examples of summary vectors.

	A	B	C	D	E	AB	AC	AD	AE	BC	BD	BE	CD	CE	DF
$v_1 =$	10	4	3	3	3	0	0	0	0	0	6	0	0	7	0
$v_2 =$	10	5	3	4	3	0	0	0	0	0	5	0	0	7	0
$v_3 =$	3	5	3	6	3	0	0	0	0	0	5	0	0	7	0
$v_4 =$	6	5	10	6	3	0	0	0	4	0	5	0	0	0	0
$v_\mu =$	7.25	4.75	4.75	4.75	3	0	0	0	1	0	5.25	0	0	5.25	0
$v_{\sigma^2} =$	8.69	0.19	9.19	1.69	0	0	0	0	3	0	0.19	0	0	9.19	0
$v_{min} =$	3	4	3	3	3	0	0	0	4	0	5	0	0	7	0

Figure 2.2: Summary vectors v_μ, v_{σ^2}, and v_{min} for the set of weighted vectors $v_1..v_4$.

Chapter 3

Standard Tree Comparisons

The usual way of comparing two trees is to count the number of edges in which they differ. This calculation defines the Robinson-Foulds metric [26].

$$d_{RF}(T_A, T_B) = \frac{1}{2}\left(|\Sigma(T_A) - \Sigma(T_B)|\right) + \frac{1}{2}\left(|\Sigma(T_B) - \Sigma(T_A)|\right)$$

where $-$ is set difference, $|.|$ is cardinality, and $+$ is arithmetic.

Figure 3.1 illustrates the RF calculation.

bit-vectors support this operation as

$$d_{RF}(v_A, v_B) = \frac{1}{2}\sum_{i=1}^{b} |v_A[i] - v_B[i]|$$

The weighted RF metric is given as [27]

$$d_{WRF}(T_A, T_B) = \sum_{x \in \Sigma(T_A) \cup \Sigma(T_B)} |w_A(x) - w_B(x)|$$

Weighted vectors support this operation as

$$d_{WRF}(v_A, v_B) = \sum_{i=1}^{b} |v_A[i] - v_B[i]|$$

Notice that we have

$$\sum_{i=1}^{b} |v_A[i] - v_B[i]| = \|v_A - v_B\|_1$$

where $\|.\|_L$ denotes the standard vector L-norm. The L-norm on $\mathbb{R}^b$ is defined as

$$\|v\|_L = \left(\sum_{i=1}^{b} |v[i]|^L \right)^{\frac{1}{L}}$$

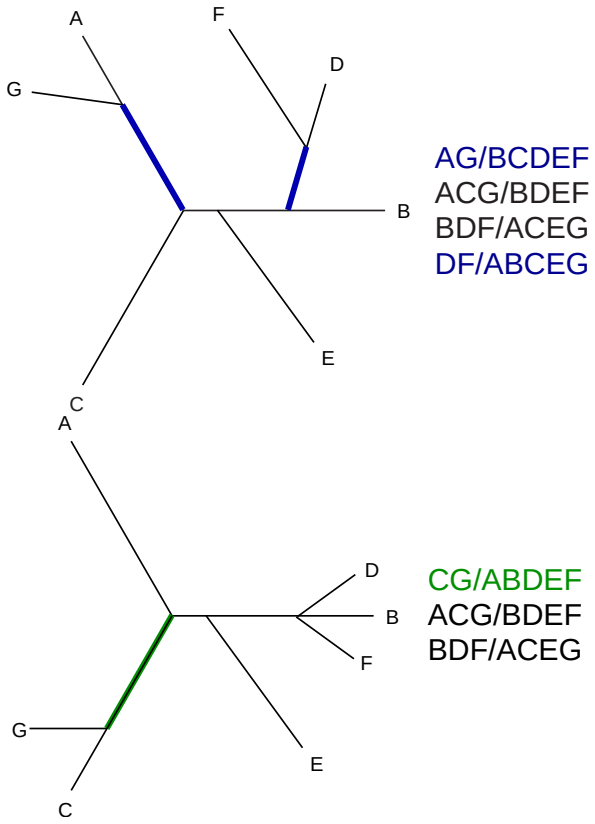
for all $p \geq 1$ and the special case

$$\|v\|_\infty = \max_{i=1}^{b} |v[i]|$$

A distance function is a metric if and only if it satisfies the following three constraints[31]:

- positivity, $d(x, y) \geq 0$ and $d(x, x) = 0$
- symmetry, $d(x, y) = d(y, x)$
- triangle inequality, $d(x, y) + d(y, z) \geq d(x, z)$

It is known that L-norms are metrics in $\mathbb{R}^d$ for all L and all d. By Corollary 2.1.2 all L-norms on tree vectors are metrics in tree space.



$$
\begin{aligned}
d_{RF}(T_A, T_B) &= \frac{1}{2}\left(|\Sigma(T_A) - \Sigma(T_B)|\right) + \frac{1}{2}\left(|\Sigma(T_B) - \Sigma(T_A)|\right) \\
&= \frac{1}{2}(2) + \frac{1}{2}(1) \\
&= \frac{3}{2}
\end{aligned}
$$

Figure 3.1: The RF distance between two trees.

Chapter 4

Consensus Methods

A consensus method is a function $f : \mathcal{T}_n^m \to \mathcal{T}_n$ that summarizes an m-tuple of trees as a single tree. The mean weight vector for a set of bit-vectors is a useful structure from which to compute standard consensus trees. Extending consensus methods to incorporate edge weights is a largely ignored subject. It turns out that v_{σ^2} can be used for this purpose.

4.1 Standard Consensus Methods

Edge frequency is a normalized count of the number of occurrences of an edge in the input set of trees. Four well known consensus methods are based on edge frequency: strict, majority rule, greedy, and asymmetric median. Since v_μ precisely contains edge frequencies, it can be used to calculate standard consensus trees.

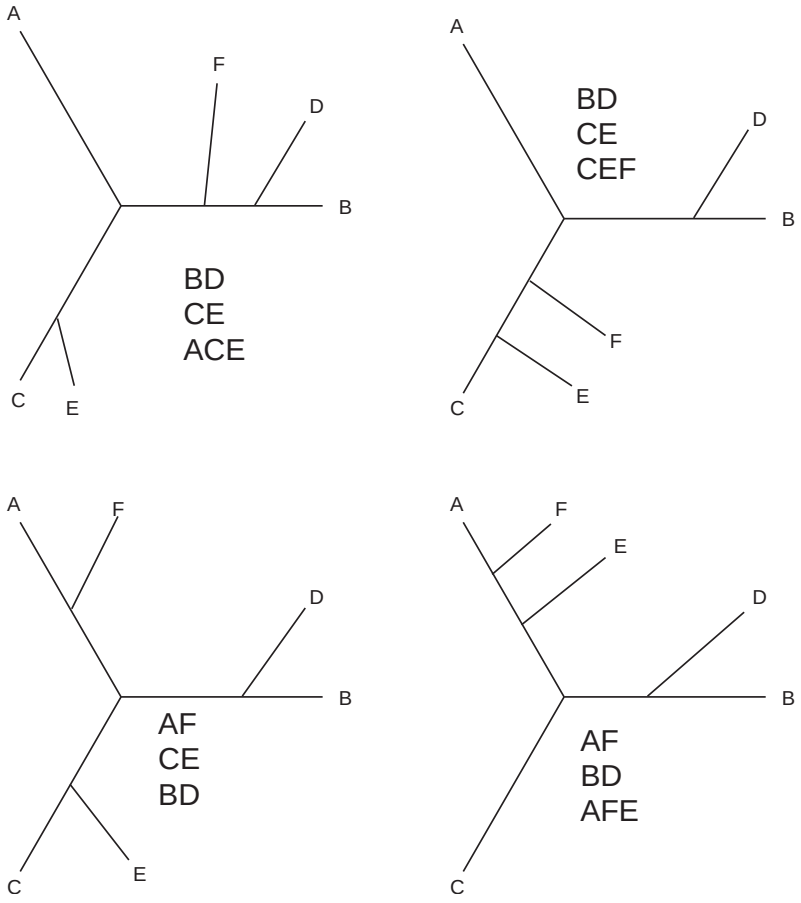
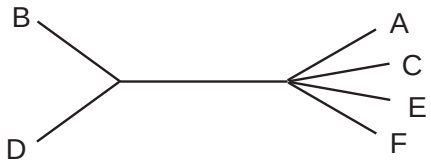


Figure 4.1: Four trees that will be used to illustrate consensus methods.

4.1.1 Strict Consensus Tree

The strict consensus tree is composed of exactly the edges occurring in all input trees. Assume that the input set is composed of bit-vector trees. Edges that occur in all trees will have a value of 1 in the mean weight vector (v_μ). Thus taking the elementwise floor, $\lfloor v_\mu \rfloor$, yields the bit-vector representation of the strict consensus tree.

See figure 4.2 for the strict consensus tree of the trees in figure 4.1.



	AF	BD	CE	ACE	AFE	CEF
$v_\mu =$	$\frac{1}{2}$	1	$\frac{3}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
$\lfloor v_\mu \rfloor =$	0	1	0	0	0	0

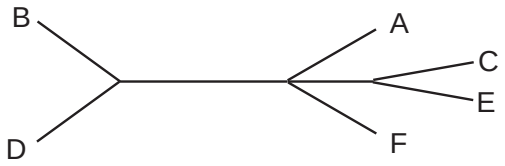
Figure 4.2: The strict consensus tree of the four trees in figure 4.1. The tree is readily computable by taking the floor of v_μ

4.1.2 Majority-Rule Consensus

The majority-rule consensus tree is also readily calculable using the mean weight vector of the input set. The majority-rule tree is composed of all edges that occur in greater than half of the input trees. Thus take each element of v_μ and round up if greater than $\frac{1}{2}$ and otherwise round down.

See figure 4.2 for the majority-rule consensus tree of the trees in figure 4.1.

The two consensus methods presented thus far are actually very similar. The strict consensus method retains edges with support in v_μ equal to 1. The majority-rule consensus method retains edges with support in v_μ greater than $\frac{1}{2}$. By introducing a threshold parameter t to denote which edges to retain from v_μ, strict and majority-rule consensus are the same method with different t values. To minimize information loss while building a consensus tree it is desirable to include as many edges from the input set as possible, which can be accomplished by taking lower and lower values of t. Unfortunately the implications of allowing t to drop below $\frac{1}{2}$ are troublesome.



	AF	BD	CE	ACE	AFE	CEF
$v_\mu =$	$\frac{1}{2}$	1	$\frac{3}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
$round(v_\mu) =$	0	1	1	0	0	0

Figure 4.3: The majority-rule consensus tree of the four trees in figure 4.1. The tree is readily computable by rounding entries in v_μ up when greater than $\frac{1}{2}$ and down otherwise.

Trouble arises because trees have highly constrained structure. There are edges that simply cannot exist in the same tree. Edges that can simultaneously exist in the same tree are called *compatible*. Equivalently, two splits $A|B$ and $C|D$ are compatible if and only if one of the four set intersections $A \cap C$, $A \cap D$, $B \cap C$, or $B \cap D$ is empty. That is to say, if none of the four intersections are empty, it is impossible to represent both splits in the same tree.

The edges for the strict and majority-rule consensus always form a tree. To see this, consider the following argument: It is known that a set of splits are pairwise compatible if and only if they can be assembled into a unique tree [10]. And,

Lemma 4.1.1. *All edges with support in v_μ greater than $\frac{1}{2}$ are pairwise compatible.*

Proof. This is an application of the pigeonhole principle. If two edges each occur in greater than half of the input trees, then by the pigeonhole principle they appear together in at least one tree. Thus the two edges are compatible. This applies to all pairs of edges where each edge has support in v_μ greater than $\frac{1}{2}$. □

4.1.3 Greedy and Asymmetric Median Consensus

While the strict and majority-rule trees always exist, compatibility can cause problems when the threshold parameter t drops to $\frac{1}{2}$ or lower. Thus decisions need to be made as to which edges (with support $\leq \frac{1}{2}$) should be included in the consensus tree.

Begin by prioritizing inclusion in the consensus tree by frequency in the input set. Then the problem can be phrased in terms of optimization: Maximize the sum of frequencies of edges included in the consensus tree. This is equivalent to choosing the subset of elements of v_μ so that (1) the sum of values from v_μ of the subset is maximized and (2) the edges can be composed into a tree.

The optimal tree is called the *asymmetric median tree* [25]. Unfortunately, optimally solving the problem is $\mathcal{NP}$-hard. Picking edges in a greedy fashion (i.e. pick edges in priority order and keep if compatible) yields the *greedy consensus tree*.

Both methods (asymmetric median and greedy) are based solely on edge frequency. So once again v_μ can clearly be used as the representation from which to calculate these trees.

4.2 Edge Weight Consensus

Edge weights are typically ignored when computing consensus trees. The following section introduces a technique for meaningfully weighting the majority-rule tree. Subsequently, a new weight-sensitive consensus method is proposed that resembles the asymmetric median method.

4.2.1 Weighted Majority-Rule

A median tree for a set of trees $\mathcal{T}$ and distance function $d : \mathcal{T} \times \mathcal{T} \to \mathbb{N}$ is the tree T_M that minimizes

$$\sum_{T \in \mathcal{T}} d(T_M, T)$$

The majority-rule tree is a median tree for the RF distance [5].

Define another summary vector (i.e. in the spirit of v_μ and v_{σ^2}) as follows:

Definition 4.2.1. *The minimum-weight vector of a set of tree vectors $\mathcal{V}$ is $v_{min} \in \mathbb{R}^b$ where each element of v_{min} is defined as*

$$v_{min}[i] = \min_{v_T \in \mathcal{V}} v_T[i]$$

Take the minimum weight vector and set to zero all elements that do not correspond to edges in the majority-rule tree. This yields a weighted majority-rule tree where the weight of an edge is the minimum weight of the edge across the set.

Theorem 4.2.1. *The majority-rule tree with edge weights taken from the minimum weight vector is a median tree for the weighted Robinson-Foulds metric.*

Proof. Consider adding an edge e to a consensus tree T_C such that the weight of e is some value $w_C(e)$. Some trees in $\mathcal{T}$ contain e, call them $\mathcal{T}_{yes} = \{T_i | T_i \in \mathcal{T} \wedge e \in T_i\}$ and some do not, call them $\mathcal{T}_{no} = \{T_i | T_i \in \mathcal{T} \wedge e \notin T_i\}$.

Adding e to T_C causes disagreement between T_C and $\mathcal{T}_{no}$ such that the dissimilarity increases by $\sum_{T_i \in \mathcal{T}_{no}} |w_C(e)|$. On the other hand, adding e to T_C causes agreement between T_C and $\mathcal{T}_{yes}$ such that the dissimilarity decreases by $\sum_{T_i \in \mathcal{T}_{yes}} |w_i(e)| - |w_C(e) - w_i(e)|$. Thus we have

$$\Delta_{add}(e) = \sum_{T_i \in \mathcal{T}_{no}} |w_C(e)| - \sum_{T_i \in \mathcal{T}_{yes}} |w_i(e)| - |w_C(e) - w_i(e)|$$

Removing an edge from T_C yields the symmetric result:

$$\Delta_{remove}(e) = -\Delta_{add}(e)$$

In the case where all weights are equal to 1 (i.e. unweighted trees), this expression collapses to

$$\Delta_{add}(e) = |\mathcal{T}_{no}| - |\mathcal{T}_{yes}|$$

Notice that we just proved majority consensus as a median for the standard RF distance.

So now if we take $w_C(e) = \min_i w_i(e)$, we impose the constraint $\forall i,\ w_i(e) \geq w_C(e)$. Thus we have $|w_C(e) - w_i(e)| = w_i(e) - w_C(e)$ and $\Delta_{add}(e)$ collapses to

$$\Delta_{add}(e) = w_C(e)\,(|\mathcal{T}_{no}| - |\mathcal{T}_{yes}|)$$

□

4.3 Edge-Weight Stability Consensus

Let us add edge-weight considerations into a consensus method like the asymmetric median method. It does not make sense to extend the optimization problem to use the the version of v_μ based on weighted vectors. Such an approach would amount to biasing the prioritization step toward parts of the trees with large amounts of evolution.

So instead of prioritizing inclusion by value of v_μ, we use v_{σ^2}. The motivation is to retain portions of input trees that are the same shape and have stable edge weight. Not surprisingly this method is, like the asymmetric median method, computationally hard.

There is a slight twist. Since minimizing variance is desired (whereas in asymmetric median, it is maximizing frequency), the problem is phrased as maximizing the sum of reciprocals of values from v_{σ^2}.

Theorem 4.3.1. *Edge Weight Stability Consensus is $\mathcal{NP}$-hard.*

Proof. The proof is for a decision version of the problem. Given v_{σ^2}, is there a compatible subset of splits where the sum of their reciprocals from v_{σ^2} exceeds some threshold value?

Validating a solution is clearly polynomial time. Verify pairwise compatibility among all selected splits ($O(n^2)$) and sum the v_{σ^2} reciprocals to verify that the sum exceeds the threshold ($O(n)$).

Reduction is from asymmetric median. The instance transformation is trivial. Simply take v_μ from asymmetric median and use as the reciprocal of the v_{σ^2} vector in the transformed instance. It is trivial to see that a solution in the latter constitutes a solution for the former. □

Chapter 5

Sublinear Robinson-Foulds Computation

In this chapter we highlight an advantage of representing trees as vectors. We borrow a technique from geometry to compact vector dimensionality while preserving the Euclidian vector norm of difference vectors. We show that the technique yields a very good approximation of Robinson-Foulds distance while providing an asymptotic speedup over the standard method for computing RF distance.

Consider the following *randomized embedding* of a set of bit-vectors $\mathcal{V} \in \mathbb{R}^b$, $|\mathcal{V}| = m$. The technique is due to Indyk and Motwani [18].

Construct a $b \times \frac{4\ln(m)}{\epsilon^2}$ matrix, f, where each element is a random number taken from the Gaussian distribution with mean 0 and variance 1. Multiply the bit-vector representation of each tree by f. Now the new input set consists of m vectors each in $\frac{4\ln(m)}{\epsilon^2}$ dimensions. First notice that the dimensionality of each tree in the new input set is independent of original tree size and solely dependent on m, the number of trees.

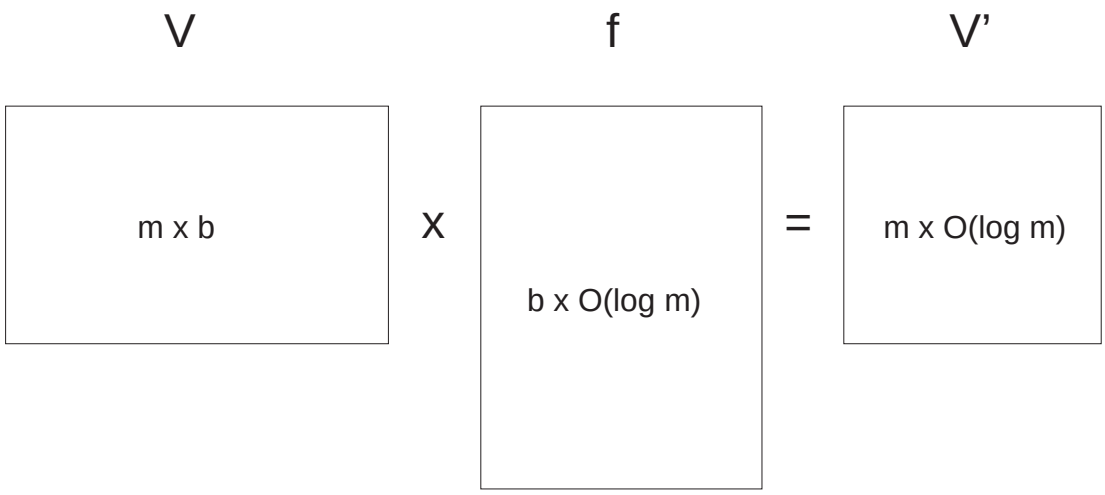


Figure 5.1: A sketch of randomized embedding. Each tree is a row in V. Each row of V' is the embedded representation of the corresponding row vector in V.

The Johnson-Lindenstrauss lemma [17] can be used to prove that the $\|.\|_2$ norm of each embedded point vf, $\forall v \in \mathcal{V}$, obeys

$$(1-\epsilon)\|v\|_2 \leq \|vf\|_2 \leq (1+\epsilon)\|v\|_2$$

with probability $1 - \frac{1}{n^2}$.

See figure 5.1 for a graphical representation of the embedding step.

Recall from chapter 3 that the Robinson-Foulds metric between two bit-vectors is

$$d_{RF}(v_A, v_B) = \frac{1}{2}\|v_A - v_B\|_1$$

Notice the highly constrained structure, as expressed in the following lemma, of a difference vector $v_A - v_B$, where v_A and v_B are bit-vectors.

Lemma 5.0.2. *Given a set of bit-vectors $\mathcal{V}$,*

$$\forall v_A, v_B \in \mathcal{V} \text{ and } 1 \leq i \leq b, (v_A - v_B)[i] = -1 \vee 0 \vee 1$$

Proof. Elements in bit-vectors are 0 or 1. There are only four ways in which these two values can be combined in a subtraction operation. $0 - 0 = 0$, $0 - 1 = -1$, $1 - 0 = 1$, and $1 - 1 = 0$, with only three possible results $-1 \vee 0 \vee 1$. □

There is a simple relationship between the L_1 and L_2 norm of vectors with such structure.

Lemma 5.0.3. *For an arbitrary vector $v \in \mathbb{R}^b$ where every element is chosen from the set $\{-k, 0, k\}$, the following equality holds:*

$$\|v\|_1 = \frac{(\|v\|_2)^2}{k}$$

Proof. Assume that v has c entries of value $\pm k$ (so there are $b - c$ entries of value 0).

$$\begin{aligned} \|v\|_2 &= \sqrt{\sum_{i=1}^{b} (|v_i|)^2} \\ &= \sqrt{ck^2} \\ &= \sqrt{c}k \end{aligned} \tag{5.1}$$

and

$$\begin{aligned} \|v\|_1 &= \sum_{i=1}^{b} |v_i| \\ &= ck \\ &= \sqrt{c}(\sqrt{c}k) \\ &= \sqrt{c}\|v\|_2 \end{aligned} \tag{5.2}$$

Solving (5.1) for $\sqrt{c}$ yields

$$\sqrt{c} = \frac{\|v\|_2}{k}$$

Finally, plugging into (5.2) yields

$$\|v\|_1 = \frac{(\|v\|_2)^2}{k}$$

□

Thus to calculate the L_1 norm of an arbitrary difference vector between two bit-vector trees it is sufficient to calculate the L_2 norm because we have

$$\|.\|_1 = (\|.\|_2)^2$$

So combining the Johnson-Lindenstrauss lemma, and lemmata 5.0.2 and 5.0.3 yields

Theorem 5.0.4. *A $(1+\epsilon)$ approximation of the Robinson-Foulds distance between any two phylogenetic trees in a set $\mathcal{T}$ (where $|\mathcal{T}| = m$) can be computed in $O(log(m))$, i.e. in time independent of the size of each tree.*

Since the embedding step costs $O(nm\log(m))$, this technique does not yield an asymptotic speedup for an isolated RF distance calculation between two trees (which costs $O(n)$). A more complicated computation such as the $\binom{m}{2}$ pairwise RF distances between all m points traditionally costs $O(m^2 n)$. Embedding the points and subsequently calculating all pairwise distances in the embedded representation costs $O(nm\log(m)) + O(m^2\log(m))$. The first term represents the embedding step and the second represents computing all pairwise distances on the embedded representation.

Chapter 6

Future Work

This has been a fruitful introduction to the field of phylogenetic postprocessing. The next step should be experimentation, although there is always more theory to do as well.

6.1 Vector Norms as Metrics in Tree Space

It is clearly useful to recognize the RF distances as standard L_1 norms. The benefits of having the other norms is unclear as of yet. For example, the higher-order norms for the edge-weight vectors are biased by edges with large edge weights.

6.2 Consensus Methods

There is much theoretical work to be done in order to properly incorporate edge weights into consensus methods. It would also be interesting to pursue edge weightings in supertree methods [8, 20].

Experimentation needs to be performed on the edge-weight stability consensus method. It is suspected that the edge-weight stability consensus method will be useful as a fine discriminator for input sets with high agreement on topology.

6.3 Sublinear Robinson-Foulds

This technique can be readily incorporated into clustering-based methods. Further work would include preserving distances other than RF through dimensionality-reducing vector embeddings (an approach in this spirit can be found in [2]).

An experimental study needs to performed so as to assess the actual performance of the embedding. To this end, [1] shows that the embedding technique works equally well when the random matrix used for embedding has elements taken from a much simpler distribution than the standard normal distribution. This simpler distribution is as follows

$$\begin{aligned} p(X = -1) &= \frac{1}{6} \\ p(X = 0) &= \frac{2}{3} \\ p(X = 1) &= \frac{1}{6} \end{aligned}$$

Multiplying by a matrix with elements taken from this simpler distribution is clearly cheaper in implementation terms. First, it involves no floating point arithmetic. Second, since the expected number of zeros is very high, most of the integer multiplications can be avoided.

6.4 Other Considerations

Another result from the literature on geometric embedding may be relevant to the compact representation of phylogenetic trees. The shortest-path metric is the shortest path between two vertices in a graph. It can be proved that, if tree T has l leaves it is possible to embed the shortest-path metric with no distortion as the L_∞ norm of a vector space in $O(\log l)$ dimensions [23].

Compared with the two phylogenetic representations (parenthized NEWICK trees or distance matrices) the embedded tree offers a middle ground in the tradeoff between space and time. A distance matrix offers constant time lookup of the shortest path metric but requires $\Theta(l^2)$ space. A linear representation, while costing $\Theta(l)$ space, requires traversal to find shortest paths. The embedding offers $\Theta(l \log l)$ space and it costs $\Theta(\log l)$ time to calculate path length.

It may be fruitful to use the embedded representation to do probabilistic structure matching between trees. An approach to this effect is presented in [21] for the domain of graph matching. To sketch the approach: graphs are embedded in a manner similar to that described for trees. This gives rise to a set of points in $\mathbb{R}^d$ per graph. Sets of points are then projected into the same basis. Finally, distances are calculated between weighted sets of points. The technique has yielded success in the practice of many-to-many matchings of object silhouettes.

References

[1] Dimitris Achlioptas. Database-friendly random projections: Johnson lindenstrauss with binary coins. *J. Comput. Syst. Sci.*, 66(4):671–687, 2003.

[2] Benjamin L. Allen and Mike Steel. Subtree transfer operations and their induced metrics on evolutionary trees. *Annals of Combinatorics*, 5(1):1–15, 2001.

[3] Nina Amenta and Jeff Klingner. Case study: Visualizing sets of evolutionary trees. In *INFOVIS '02: Proceedings of the IEEE Symposium on Information Visualization (InfoVis'02)*, page 71. IEEE Computer Society, 2002.

[4] K.J. Arrow. *Social Choice and Individual Values*. John Wiley & Sons, New York, 1951.

[5] J. Barthelemy and F. McMorris. The median procedure for n-trees. *Journal of Classification*, 3:329–334, 1986.

[6] Tanya Y. Berger-Wolf. Properties of compatibility and consensus sets of phylogenetic trees. *UNM Computer Science Technical Report*, TR-CS-2004-24, 2004.

[7] L. J. Billera, S. P. Holmes, and K. Vogtmann. Geometry of the space of phylogenetic trees (vol 27, pg 733, 2001). *Advances in Applied Mathematics*, 29(1):136 – 136, Jul 2002.

[8] S. Bocker, D. Bryant, A. W. M. Dress, and M. A. Steel. Algorithmic aspects of tree amalgamation. *Journal of Algorithms*, 37(2):522 – 537, Nov 2000.

[9] D. Bryant. A classification of consensus methods for phylogenies. *Lapointe, F.-J., McMorris, F.R., Mirkin, B., Roberts, F.S. (eds) BioConsensus, DIMACS. AMS.*, pages 163 – 184, 2003.

[10] P. Buneman. The recovery of trees from measures of dissimilarity. *In Hodson, Kendall, and Tautu, editors, Mathematics in the Archaeological and Historical Sciences*, 1971.

[11] Savrina F. Carrizo. Phylogenetic trees: an information visualisation perspective. In *CRPIT '29: Proceedings of the second conference on Asia-Pacific bioinformatics*, pages 315–320. Australian Computer Society, Inc., 2004.

[12] G. Cucumel and Lapointe F.J. The average consensus procedure for weighted trees. *Syst. Biol.*, 46:306–312, 1997.

[13] B. DasGupta, X. He, T. Jiang, M. Li, J. Tromp, and L. Zhang. On computing the nearest neighbor interchange distance. *in D. Z. Du, P M. Pardalos and J. Wang (eds.), Proceedings of the DIMACS Workshop on Discrete Problems with Medical Applications, DIMACS Series in Discrete Mathematics and Theoretical Computer Science, American Mathematical Society*, 55:125 – 143, 2000.

[14] William H.E. Day and F.R. McMorris. Axiomatics in group choice and bioconsensus. *Lapointe, F.-J., McMorris, F.R., Mirkin, B., Roberts, F.S. (eds) BioConsensus, DIMACS. AMS.*, pages 3 – 35, 2003.

[15] Minos N. Garofalakis and Amit Kumar. Correlating xml data streams using tree-edit distance embeddings. In *PODS*, pages 143–154. ACM, 2003.

[16] D. H. Huson. Splitstree: analyzing and visualizing evolutionary data. *Bioinformatics*, 14(1):68 – 73, 1998.

[17] P. Indyk. Algorithmic applications of low-distortion geometric embeddings. In *FOCS '01: Proceedings of the 42nd IEEE symposium on Foundations of Computer Science*, page 10. IEEE Computer Society, 2001.

[18] Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. pages 604–613, 1998.

[19] Jaap. Indexing combinations. *From Jaap's Puzzle Page*, http://www.geocities.com/jaapsch/puzzles/compindx.htm.

[20] J. Jansson, J. H. K. Ng, K. Sadakane, and W. K. Sung. Rooted maximum agreement supertrees. *LATIN 2004: Theoretical Informatics*, 2976:499 – 508, 2004.

[21] Yakov Keselman, Ali Shokoufandeh, M. Fatih Demirci, and Sven J. Dickinson. Many-to-many graph matching via metric embedding. In *CVPR (1)*, pages 850–857. IEEE Computer Society, 2003.

[22] J. Kim and T. Warnow. Tutorial on phylogenetic tree estimation. *Intelligent Systems for Molecular Biology, Heidelberg*, 1999.

[23] Nathan Linial, Eran London, and Yuri Rabinovich. The geometry of graphs and some of its algorithmic applications. *Combinatorica*, 15:215–245, 1995.

[24] Fred R. McMorris and Robert C. Powers. Consensus functions on trees that satisfy independence axiom. *Discrete Applied Mathematics*, 47(1):47–55, 1993.

[25] C. Phillips and T.J. Warnow. The asymmetric median tree - a new model for building consensus trees. *Discrete Applied Mathematics*, 71:311–335, 1996.

[26] D. Robinson and L. Foulds. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53:131–147, 1981.

[27] D.F. Robinson and L.R. Foulds. Comparison of weighted labeled trees. *Lecture Notes in Mathematics*, 748:119–126, 1978.

[28] M. Steel, A. W. M. Dress, and S. Bocker. Simple but fundamental limitations on supertree and consensus tree methods. *Systematic Biology*, 49(2):363 – 368, JUN 2000.

[29] Cara Stockham, Li-San Wang, and Tandy Warnow. Statistically based postprocessing of phylogenetic analysis by clustering. In *ISMB*, pages 285–293, 2002.

[30] N. Takahata. Gene genealogy in three related populations: consistency probability between gene and population trees. *Genetics*, 122:957–966, 1989.

[31] Eric W. Weisstein. "metric.". *From MathWorld–A Wolfram Web Resource*, http://mathworld.wolfram.com/Metric.html.

[32] M. Wilkinson. Common cladistic information and its consensus representation - reduced adams and reduced cladistic consensus trees and profiles. *Systematic Biology*, 43(3):343 – 368, SEP 1994.

[33] Mark Wilkinson and Joseph L. Thorley. Reduced consensus. *Lapointe, F.-J., McMorris, F.R., Mirkin, B., Roberts, F.S. (eds) BioConsensus, DIMACS. AMS.*, pages 195 – 203, 2003.

[34] Tiffani L. Williams, Tanya Berger-Wolf, Bernard M.E. Moret, Usman Roshan, and Tandy Warnow. The relationship between maximum parsimony scores and phylogenetic tree topologies. *UNM Computer Science Technical Report*, TR-CS-2004-04, 2004.