Pattern Identification in Biogeography*

Ganeshkumar Ganapathy[1], Barbara Goodson[2], Robert Jansen[2], Vijaya
Ramachandran[1], and Tandy Warnow[1]

[1] Department of Computer Sciences, The University of Texas at Austin, TX 78712;
{gsgk,vlr,tandy}@cs.utexas.edu
[2] Section of Integrative Biology, School of Biological Sciences, The University of Texas at
Austin, TX 78712;
{bgoodson,jansen}@mail.utexas.edu

Abstract. We develop and study two distance metrics for area cladograms (leaf-labeled trees where many leaves can share the same label): the *edge contract-and-refine* metric and the *MAAC* distance metric. We demonstrate that in contrast to phylogenies, the contract-and-refine distance between two area cladograms is not identical to the character encoding distance, and the latter is not a metric. We present a polynomial time algorithm to compute the MAAC distance, based on a polynomial-time algorithm for computing the largest common pruned subtree of two area cladograms. We also describe a linear time algorithm to decide if two area cladograms are identical.

1 Introduction

Biogeography is the study of the spatial and temporal distributions of organisms ([BL98, CKP03]). Biogeographers seek not only to understand ecological processes that influence the distribution of living organism over short periods of time (e.g., climatic stability, effect of area) but also to uncover events occurring in the distant past (e.g., continental drift, glaciation, evolution) which have resulted in the geographic distribution observed today.

Biogeography and Phylogeny. One of the ways of understanding the geographic distribution of species is by studying the *evolutionary history of the species* (see [CLW95, EO05, Jac04b] for instances of this approach). The evolutionary relationships are typically represented as branching tree structures called *phylogenetic trees*, or simply phylogenies. The branching structure of the phylogeny of a set of taxa can be used to differentiate between competing hypotheses concerning the observed geographic distribution of the set of taxa. Moreover, *a consistent pattern* observed in the phylogenies of species from different genera in the same geographic area will imply a stronger evidence for the particular hypotheses suggested by the pattern. As an example of this approach, consider a group of islands, each containing multiple ecological zones (for

example, each island can contain coastal and mountain ecological zones). Suppose our goal is to understand the observed geographic distribution of species on the islands. One hypothesis about the distribution could be that species dispersed from each ecological zone in each island to similar zones in other islands and then differentiated. This process is called *inter-island colonization*. Another hypothesis could be that dispersal *between* islands happened first followed by dispersal to the different ecological zones and differentiation into many species. This process is called *adaptive radiation* (see [JEOH00] for a discussion). The crucial idea is that we might be able to infer which of the above two hypotheses is responsible for the observed distribution: inter-island colonization is suggested by taxa on different islands but the same ecological zone forming a monophyletic group (rooted subtree), and adaptive radiation is suggested if species on the same island in different ecological zones form a monophyletic group (that is, form a rooted subtree in the phylogeny).
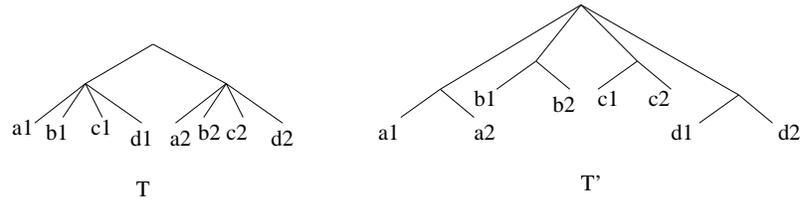

Fig. 1. Two hypothetical phylogenies on eight taxa on four islands (a, b, c, d) with two ecological zones each (1 and 2). T suggests dispersal, and T' suggests adaptive radiation.

Area Cladograms. Before looking for common patterns in the phylogenies of different sets of species in the same geographic area, the phylogeny for each set of species is converted to an *area cladogram*. Area cladograms are rooted or unrooted trees (as are phylogenies) whose leaves are labeled with *geographic areas* instead of taxa (see [Ros78, NP81]). To obtain the area cladogram for a set of species local to a set of areas, we start with the phylogeny for the set of species and, for each leaf, replace the taxon label with the label of the area in which the taxon is found. This process is illustrated in Figure 2. More formally, we define:

Definition 1. *Area Cladogram*
An area cladogram *is an unrooted or rooted leaf-labeled tree T. The leaves are labeled with areas, and many leaves may share the same label.*

In general, it might happen that a single taxon resides in more than one area (such taxa called *widespread taxa*), and this would result in area cladograms with multiply-labeled leaves. We will develop our metrics and algorithms for area cladograms as in Definition 1, but we will show how to apply our results to more general cladograms where leaves can have multiple labels.

It should be noted that several methods have been proposed for obtaining area cladograms from phylogenetic trees (see [NP81, Pag88, Bro81, Pag94]). The methods "resolve" the issues of *widespread taxa* (single leaf being labeled by many areas), *redundant taxa* (many leaves being labeled by the same area), and *missing areas* to obtain a

resolved area cladogram where the mapping between leaves and areas is one-one. Unresolved area cladograms are sometimes called *taxon area cladograms* in the literature.
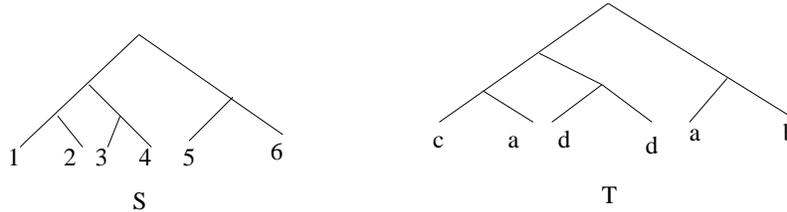

Fig. 2. A phylogeny S and its associated area cladogram T, assuming taxon 1 appears in area c; 2 appears in area a; 3 appears in area d; 4 appears in area d; 5 appears in area a; and 6 appears in area b.

Much of the prior work on area cladograms has focussed on suitable transformation that will result in resolved area cladograms, for which algorithms and metrics for phylogenetic trees apply.

In this paper, we address the problem of directly comparing two area cladograms. We develop distance metrics between area cladograms, and describe algorithms for computing a largest common pruned subtree of two area cladograms and for deciding if two given area cladograms are identical.

Prior Work. Inferring biogeographical history with species and areas is just one instance of the problem of inferring histories of two associated entities: the associated entities may be hosts and parasites, or genes and organisms [Pag94, PC98] (areas are analogous to hosts and organisms, and taxa in biogeography are analogous to parasites and genes). Hence, comparing area cladograms has a long history and a wide variety of applications (see [Jac04a, Jac04b, CLW95, GvVB02, Pag88] for example). Earlier work on comparing area cladograms has included pruning the cladograms until the two cladograms agree on the remaining leaves (see [Ros78, Pag88]), and using similarity metrics such as the *bipartition* metric (also called the *component* metric or the *character encoding* metric in the literature) and the *triplets* metric (see [Pag88]) between area cladograms (the triplets metric only applies when the area cladograms are rooted.)

All such methods apply only to resolved area cladograms. The methods of resolution differ in their interpretation of widespread taxa, redundant taxa and missing areas, and have been called *assumptions* 0, 1 *and* 2 in the literature (see [Pag88, vVZK99]). We will take a different approach to comparing area cladograms: we will compare them *without first resolving them so that the mapping between the leaves and labels is one-one*. This avoids the contentious issues ([Pag90]) surrounding the process of resolution.

Our Contributions. Our contributions are two-fold: we develop both metrics and algorithms for comparing area cladograms. More specifically,

– We show that the equivalence between the edge contract-and-refine metric ("RF-distance") and the bipartition metric ("character-encoding" metric) that holds for

phylogenies *does not hold* for area cladograms. More specifically, we show that the bipartition metric, when extended to area cladograms, is not a metric. For the edge contract-and-refine edit distance between two area cladograms we present a simple, but worst-case exponential-time algorithm. This edit distance can compare only area cladograms that are on the same number of leaves, and when each area labels the same number of leaves in both area cladograms (Section 3).

– We define another metric, the *MAAC* distance metric, for comparing two *rooted* area cladograms, which is based on the size of the largest common pruned sub-tree between the two area cladograms. The MAAC distance metric can compare two arbitrary trees that are not necessarily on the same number of leaves, which is particularly useful when comparing area cladograms (Section 3).

– We present a polynomial time algorithm for computing the MAAC distance be-tween two rooted area cladograms. This algorithm is based on an algorithm we present for computing the largest common pruned subtree, the *maximum agree-ment area cladogram (MAAC)*, of two area cladograms. We also describe a faster, linear-time algorithm to decide if two area cladograms are identical (Section 4).

2   Phylogenies: Distance Metrics and Agreement Subsets

Character Encoding of Phylogenies. Tests for equality between phylogenies are based on the notion of the *character encoding* of phylogenies. Another notion crucial to the study of phylogenies is that of a *bipartition*: removing an edge e from a leaf-labeled tree T induces a bipartition π_e on its set of leaves.

Definition 2. *Character Encoding of a Phylogeny*
The character encoding *of a phylogeny T is the set $C(T) = \{\pi_e : e \in E(T)\}$, which represents the set of bipartitions induced by the edges of T.*

Theorem 1. *Character-Encoding Metric [Bun71]*
Let T and T' be two phylogenies on the same set of taxa. Then $|C(T) \triangle C(T')| = |(C(T) - C(T')) \cup (C(T') - C(T))|$ defines a distance metric.

By Theorem 1, two phylogenies T and T' are isomorphic (with the isomorphism preserving the leaf labels) if and only if $|C(T) \triangle C(T')| = 0$.

A *contraction* operation applied on an edge in a tree collapses that edge and iden-tifies its two end points; a *refinement* operation applied at an unresolved node (i.e., an internal node with degree greater than three) expands that unresolved node into two nodes connected by an edge.

Definition 3. *Robinson-Foulds (RF) Distance*
The Robinson-Foulds distance *between two phylogenies T_1 and T_2 is defined as the number of contractions and refinements necessary to transform T_1 into T_2 (or vice-versa), and is denoted $RF(T_1, T_2)$.*

The RF distance naturally defines a metric since it is an edit distance.

Theorem 2. *[RF81] Let T_1 and T_2 be two phylogenies on the same set of taxa. Then $RF(T_1, T_2) = |C(T_1) \triangle C(T_2)|$.*

Finally, we define the maximum agreement subtree problem for phylogenies. The analogue of this problem for area cladograms is crucial to addressing the problems outlined in Section 1.

Definition 4. *Maximum Agreement Subset (MAST)*
Let $\{T_1, T_2, \ldots, T_k\}$ be a set of phylogenetic trees, on a set L of leaves. A maximum agreement subset (MAST) of trees T_1 through T_k is a set of leaves $L' \subseteq L$ of maximum cardinality such that the restrictions of the trees $T_1, \ldots, T_k$ to the set L' are all isomorphic, with the isomorphism preserving leaf labels.

The maximum agreement subset problem was introduced in [FG85], and has been studied thoroughly since then. The rooted and unrooted versions of MAST are polynomially related since the unrooted MAST problem can be solved by solving a polynomial number of rooted MAST problems. Computing a MAST is NP-hard for three or more trees [AK97]. A $O(n^{2+o(1)})$ time algorithm for the case of two trees on n leaves is given in [FCT94]. For two rooted binary trees, the best known algorithm takes $O(n\log^3 n)$ time ([FCPT95b, FCPT95a]); for two rooted trees which may not be binary, the best known algorithm takes $O(n^{1.5}c^{\sqrt{\log n}})$ time where c is a constant ([FCT94]). For computing a MAST of k rooted trees, an $O(kn^3 + n^d)$ algorithm (with d the maximum degree of a node in any tree) was presented in [FCPT95a].

3   Distance Measures Between Area Cladograms

In this section, we will develop distance metrics for the set of area cladograms. We will first show that the character encoding distance between two different area cladograms can be zero, and hence the character-encoding "distance" is not a metric on area cladograms, and in particular cannot be used as a test of isomorphism. We then propose a metric for comparing area cladograms that is based on computing the size of the largest common pruned subtree of the two area cladograms. We call this the *MAAC* metric, and show how to compute it in Section 4.

While the character-encoding metric for phylogenies does not extend to area cladograms, the contract-and-refine edit distance still defines a metric (because it is an edit distance). We present an algorithm to compute the edge contract-and-refine edit distance between area cladograms. This algorithm is efficient if there are few occurrences of widespread taxa, but it is exponential-time in general. For phylogenies this edit distance which is called the *Robinson-Foulds* distance, can be computed efficiently since it equals the character-encoding distance.

3.1   The Character Encoding Cannot Distinguish Between Area Cladograms

We first define the *extended character encoding* of an area cladogram.

Definition 5. *Let T be an area cladogram. The multi-set $\{\pi_e : e \in E(T)\}$ is called the* extended character encoding *of T, and will be denoted by $C(T)$. Here π_e denotes the bipartition of the multi-set of leaf labels induced by the edge e.*

Contrary to our experience with phylogenetic trees where the mapping between leaves and labels is 1-1, with two area cladograms T_1 and T_2, $C(T_1) = C(T_2)$ does not imply that T_1 and T_2 are isomorphic. We exhibit a pair of such of trees in Figure 3.
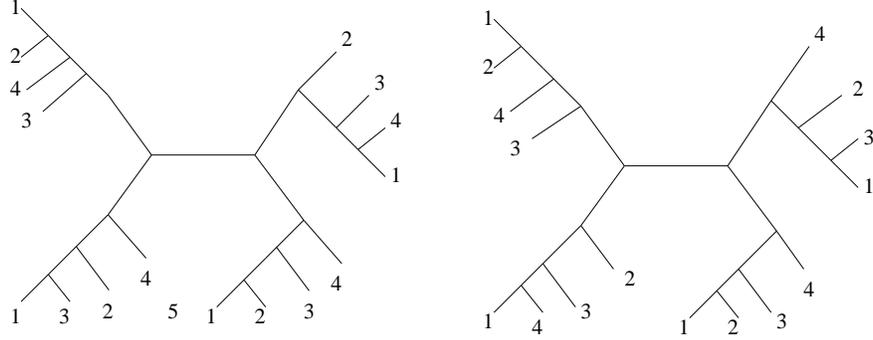
Fig. 3. Two different binary area cladograms that induce the same multi-set of partitions

3.2   The MAAC Distance Metric Between Area Cladograms

In this section we define the problem of computing the largest common pruned subtree of two rooted area cladograms and describe a distance metric based on the size of a largest common pruned subtree. We call a largest common pruned subtree a *Maximum Agreement Area Cladogram (MAAC)* (thus the MAAC is analogous to the maximum agreement subtree of two phylogenies).

Let T be an area cladogram on a set of leaves L. The *restriction* of T to a set of leaves L' is the cladogram obtained by deleting leaves in the set $L - L'$ from T and then suppressing internal nodes of degree two (except the root, if there is one).

We now define *a maximum agreement area cladogram (MAAC)* for a set of rooted area cladograms, and a distance measure between two rooted area cladograms that is based on the size of a MAAC of the two area cladograms.

Definition 6. *Maximum Agreement Area Cladogram (MAAC) and MAAC distance*

Let $\{T_1, T_2, \ldots, T_k\}$ be a set of rooted area cladograms, with L_i the leaf set of tree T_i, for $i = 1, 2, \ldots, k$. Let $\lambda_1 \subseteq L_1$ through $\lambda_k \subseteq L_k$ be sets of leaves of maximum cardinality such that the respective restrictions of the trees $T_1, \ldots, T_k$ to the sets $\lambda_1 \ldots \lambda_k$ are all isomorphic, with the isomorphisms preserving leaf labels. A restriction of any tree T_i to such a subset of leaves λ_i is a maximum agreement area cladogram *(MAAC) for the cladograms T_1 through T_k. The size of the MAAC is defined to be the number of leaves in the maximum agreement area cladogram, and is denoted by $size_{maac}(T_1, T_2, \ldots, T_k)$.*

The MAAC distance *between two trees T_1 and T_2 is $d_M(T_1, T_2) = max(n_1, n_2) - size_{maac}(T, T')$, where n_1 and n_2 are the number of leaves in T_1 and T_2 respectively.*

Note that in the above definition *we do not require that all the given set of trees contain the same number of leaves, or that they be labeled with the same set of areas, or even that they be consistent.* The MAAC distance can be viewed as a generalization of the maximum agreement subtree metric for phylogenies [GKK94], which for two phylogenies on the same set of n labeled leaves was defined as $n - size_{mast}$ where $size_{mast}$ is the size of a maximum agreement subset of the two phylogenies.

Handling Widespread Taxa. For comparing cladograms using maximum agreement area cladograms, leaves labeled by more than one area can be treated thus: each leaf labeled by a group of areas can be split into many separate leaves (all having the same parent), each of which is labeled by a single unique area from the group of areas.
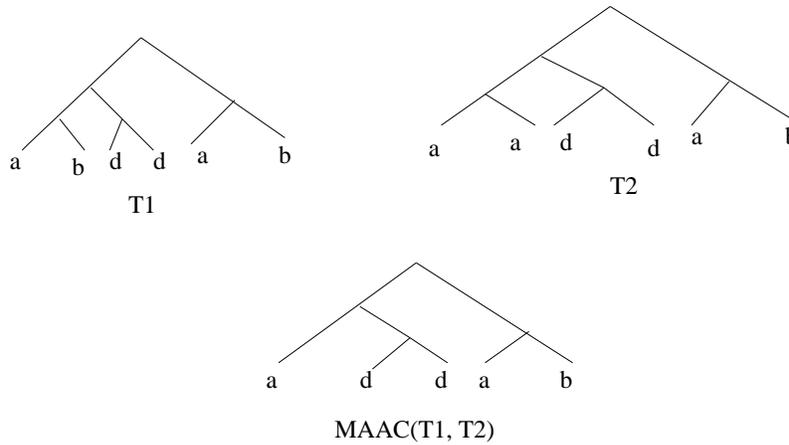

Fig. 4. Two area cladograms T1 and T2, and their MAAC

Due to space constraints, we state the following theorem without proof:

Theorem 3. *The MAAC distance d_M is a metric on the set of all area cladograms.*

Note that twice the MAAC distance between two cladograms is an upper bound on the number of insertions and deletions of leaves necessary to transform one of the cladograms to the other.

In Section 4, we present a polynomial-time algorithm for computing a maximum agreement area cladogram for two area cladograms.

3.3 Contract-and-Refine Distance Metric for Area Cladograms

Though the character-encoding distance fails to extend to area cladograms, the RF distance, being an edit distance, can be extended to unrooted area cladograms to provide a distance metric.

Definition 7. *Robinson-Foulds Distance Between Unrooted Area Cladograms*

The Robinson-Foulds distance *between two unrooted area cladograms T_1 and T_2 is defined to be the number of contractions and refinements necessary to transform T_1 to T_2 (or equivalently, T_2 to T_1).*

Note that if the number of leaves labeled l is different in T_1 and T_2 for some label l, then $RF(T_1, T_2)$ is undefined (i.e., there is no sequence of contractions and refinements that can transform T_1 into T_2). In such cases we define $RF(T_1, T_2)$ to be ∞.

Handling Widespread Taxa. Taxa endemic (resident) to more than one area would result in cladograms with leaves labeled by many areas. Our definition of the Robinson-Foulds distance applies to such cladograms as well: if a leaf is labeled with a set of areas, we can consider that set of areas to be the unique label for that leaf.

As shown in Section 3.1, for area cladograms, the RF distance will not be equal to the extended character-encoding distance. However, we can relate the RF distance between two area cladograms to the RF distance between two associated phylogenies, as we will show. We begin with some definitions.

Definition 8. *Full Differentiation of an Area Cladogram*
Let $T = (t, M)$ be an unrooted area cladogram, where t is an unlabeled tree and M is the mapping assigning labels to the leaves of t. Then, a full differentiation *of T is a leaf-labeled tree $T^* = (t, M^*)$ such that M^* is one-one. In other words, T^* has the same topology as T, but has its leaves labeled uniquely.*

Definition 9. *Consistent Full Differentiations*
Let $T_1 = (t_1, M_1)$ and $T_2 = (t_2, M_2)$ be two unrooted area cladograms with the same set L of leaf labels, and let $T_1^* = (t_1, M_1^*)$ and $T_2^* = (t_2, M_2^*)$ be full differentiations of T_1 and T_2 respectively. T_1^* and T_2^* are consistent full differentiations *if, for each label $l \in L$, the set of labels in assigned to leaves in T_1^* that were labelled l in T_1 is identical to the set of labels assigned to leaves in T_2^* that were labelled l in T_2. Mathematically, this is: $\forall l \in L, \{M_1^*(x) : M_1(x) = l\} = \{M_2^*(x) : M_2(x) = l\}$.*

Due to space constraints, we state the following theorem without proof:

Theorem 4. *Let T_1 and T_2 be two unrooted area cladograms. Then $RF(T_1, T_2)$ is equal to $\max\{RF(T_1^*, T_2^*) : T_1^*$ and T_2^* are mutually consistent full differentiations of T_1 and T_2, respectively$\}$.*

Note that the RF distance between two cladograms T_1 and T_2 is at most the RF distance between *any* consistent full differentiations of T_1 and T_2. Hence this provides a linear-time method for getting an upper bound on the RF distance between two area cladograms T_1 and T_2: we first compute two mutually consistent full differentiations, and then compute their RF distance.

Theorem 4 suggests the following trivial (but expensive) algorithm for computing the RF distance between two area cladograms T_1 and T_2: we simply compute the RF distance between all the possible consistent full differentiations of T_1 and T_2 (in $\Theta(n)$ time per pair, see [Day85]) and choose the minimum. Thus, we have the following theorem:

Theorem 5. *Let T_1 and T_2 be two unrooted area cladograms on n leaves on the same set of areas. For each area a_i appearing at the leaves of T_1 and T_2, let n_i be the number of leaves labeled with area a_i. Then, the RF distance between T_1 and T_2 can be calculated in $\Theta(n\Pi_{i=1}^{k}(n_i)!)$ time.*

4 An Algorithm for the Maximum Agreement Area Cladogram Problem

In this section we describe an algorithm for computing maximum agreement area clado-gram (MAAC) of two given rooted area cladograms. The algorithm is based on a dy-namic programming algorithm for the phylogenetic rooted maximum agreement subtree algorithm from [SW93]. We will first present the maximum agreement subtree algo-rithm. We will then observe that the basic recursion underlying the dynamic-programming algorithm will hold for the maximum agreement area cladogram algorithm as well though the mapping between leaves and their labels may not be one-one in area clado-grams.

The MAST Algorithm from [SW93]. We now give a brief summary of the algorithm in [SW93] for computing the MAST of two rooted binary trees. In our description, the expression $MAST(T, T')$ denotes a maximum agreement subset of two given (rooted binary) phylogenies T and T'.

Let T and T' be two given binary phylogenies on n leaves. Let v be a node in T, and denote by T_v the subtree of T rooted at v. Similarly denote by T'_w the subtree of T' rooted at a node w in T'. The dynamic programming algorithm for MAST operates by computing $MAST(T_v, T'_w)$ for all pairs of nodes (v, w) in $V(T) \times V(T')$ "bottom-up". We now show how to reduce computing $MAST(T_v, T'_w)$ to computing a small number of smaller MAST computations $MAST(S, S')$ where S and S' are subtrees of T_v and T'_w respectively, with at least one of them being a proper subtree.

To begin with, the $MAST(T_v, T'_w)$ is easy to compute when either v or w are leaves. So in the following discussion assume neither v nor w is a leaf.

Let L^* be a MAST of T_v and T'_w, and let T^* be the corresponding MAST tree. Then there exist homeomorphisms mapping T^* to a rooted subtree of T_v and to a rooted subtree of T'_w. Let p be the (not necessarily proper) descendant of v such that the root of T^* is mapped to p. Similarly let q be the descendant of w in T' such that that the root of T^* is mapped to w. Then, $MAST(T_v, T'_w)$ is in fact equal to $MAST(T_p, T'_q)$.

The vertex p may be actually v or it might be a vertex below v. Similarly q may be w or some vertex below w. Based on the location of p and q, we have the following cases.

- *Vertex p is a proper descendent of v.* In this case, T_p is a proper subtree of T_v, and $MAST(T_v, T'_w)$ equals $MAST(T_p, T'_w)$.
- *Vertex q is a proper descendent of w.* In this case, $MAST(T_v, T'_w)$ equals $MAST(T_v, T'_q)$.
- *Vertex p equals v and vertex q equals w.*

In the first two cases, we have reduced the computation of $MAST(T_v, T'_w)$ to a MAST computation on a subproblem. In the last case, let v_1 and v_2 be the children of v, and let w_1 and w_2 be the children of w. Let T^*_1 and T^*_2 be the subtrees of the root of the MAST tree T^*. Then, T^*_1 is homeomorphic to a subtree of T_{v_1} (or to a subtree of T_{v_2}; there is no loss of generality in assuming that it is homeomorphic to a subtree of T_{v_1}). Similarly, T^*_2 is homeomorphic to a subtree of T_{v_2}. It cannot be homeomorphic to a subtree of T_{v_1}, since then T^* would be homeomorphic to a subtree of T_{v_1}, contradicting the assumption that there is no proper descendent p of v such that root of T^* is mapped p. Arguing

similarly, we can conclude that T_1^* and T_2^* are homeomorphic to subtrees of T_{w_1}' and T_{w_2}' respectively. Now, since T^* is a MAST tree, we can conclude that T_1^* is a MAST tree of T_{v_1} and T_{w_2}', and that T_2^* is a MAST tree of T_{v_2} and T_{w_2}'. So in this case we have reduced computing $MAST(T_v, T_w')$ to computing $MAST(T_{v_1}, T_{w_1}')$ and $MAST(T_{v_2}, T_{w_2}')$ and then taking their union.

The above discussion suggests a straightforward dynamic programming algorithm which involves computing $O(n^2)$ subproblems each of which can be solved in $O(1)$ time (for binary trees).

The running time of the above algorithm is $O(n^2)$ for trees of bounded degree. For general rooted phylogenetic trees the running time is $O(n^{2.5} \log n)$.

4.1   The Maximum Agreement Area Cladogram Algorithm.

The difference between the maximum agreement area cladogram and the maximum agreement subset problems is that the former problem takes as input leaf-labeled trees where the mapping between leaves and labels is not one-one. Recall that in the description of the maximum agreement subtree dynamic programming recursion above, p is the *unique* descendant of v such that the homeomorphism mapping T^* to a subtree of T_v maps the root of T^* to p, and q is the *unique* descendant of w such that the homeomorphism mapping T^* to a subtree of T_w maps the root of T^* to q. However, when the map between leaves and labels is not one-one, nodes p and q may not be unique. However, we can remedy this situation by modifying our description thus: in tree T_v, let p be a vertex farthest from v such that the root of T^* is mapped to p, and in T_w', and let q be a vertex farthest from w such that the root of T^* is mapped to q (note that this modification will not affect the actual algorithm at all, only the proof that the algorithm is correct). The rest of dynamic programming recursion uses only the properties of homeomorphisms, and these properties hold true for homeomorphisms between area cladograms as well. Hence, the maximum agreement subtree algorithm from [SW93] works without change as a maximum agreement area cladogram algorithm.

The Running Time of the MAAC Algorithm. The algorithm is same as the maximum agreement subtree algorithm, and hence the running time of the maximum agreement area cladogram algorithm is $O(n^2)$ for trees of bounded degree and $O(n^{2.5} \log n)$ for trees of unbounded degree.

4.2   Testing Isomorphism Between Two Rooted Area Cladograms

The MAAC distance metric between area cladograms gives us a polynomial-time algorithm for testing isomorphism: we apply the maximum agreement area cladogram algorithm from the previous section to compute the MAAC distance between the two area cladograms, and we conclude that the two cladograms are isomorphic if and only if the distance is zero. The algorithm is adapted from the algorithm for testing rooted tree isomorphism from [AHU74].

The input to the algorithm consists of two rooted area cladograms T_1 and T_2 on n leaves (if the number of leaves is different, then clearly they are not isomorphic). We assume that the leaves are labeled with integers from 1 through n, not all distinct. The algorithm is based on assigning to each node u in the tree, an integer, which we call *index*(u). For leaves, the index is just their labels. The algorithm is as follows:

1. Compute the *height*, the maximum distance between the root and a leaf, of the two trees. If the heights are not the same, then the trees are not isomorphic, otherwise, let the height be h.
2. Based on the height, assign level numbers to the nodes of the trees. The level number of a node at a distance of d from the root is set to be $h - d$.
3. For each leaf u at level 0, set *index*$[u]$ to be the leaf-label.
4. Assuming that index has been set for each node at level $i - 1$, calculate the indices at level i thus: for each node v at level i, form a *tuple* (an ordered list) consisting of the indices of its children sorted in ascending order. If v is a leaf, then its tuple consists of just its label. Let L_i be the list of tuples of nodes at level i in T_1. Let L_i' be the corresponding list for T_2. Now lexicographically sort L_i and L_i' to obtain S_i and S_i' respectively.
5. If S_i and S_i' are not identical, then declare T_1 and T_2 to be non-isomorphic and quit. Else, assign *index*$[v]$ for each node v at level i in T_1 thus: *index*$[v]$ is the *rank* of v's tuple in the sorted list S_i. The ranks start from 1, and all identical tuples receive the same rank. Indices for vertices in T_2 are assigned similarly. The level-i indices can now be used to calculate the indices for level $i + 1$.
6. If the roots of T_1 and T_2 are assigned the same index, then the trees are isomorphic, otherwise not.

Proof of Correctness and Running Time: We omit the proof due to space constraints. The running time of the above algorithm for testing isomorphism is $O(n)$, where n is the number of leaves in the input trees (see [AHU74]).

References

[AHU74]   A. V. Aho, J. E. Hopcroft, and J. D. Ullman. *The Design and Analysis of Computer Algorithms*. Addison-Wesley Publishing Company, 1974.

[AK97]   A. Amir and D. Keselman. Maximum Agreement Subtrees in a Set of Evolutionary Trees: Metrics and Efficient Algorithms. *SIAM Journal of Computing*, 26(6):1656–1669, 1997. A preliminary version of this paper appeared in FOCS '94.

[BL98]   J. H. Brown and M. V. Lomolino. *Biogeography*. Sinauer Associates, Sunderland, Massachusetts, second edition, 1998.

[Bro81]   D. R. Brooks. Hennig's Parasitological Method: A Proposed Solution. *Systematic Zoology*, 30:229–249, 1981.

[Bun71]   P. Buneman. The Recovery of Trees from Measures of Dissimilarity. *Mathematics in the Archaelogical and Historical Sciences*, pages 387–395, 1971.

[CKP03]   J. V. Crisci, L. Katinas, and P. Posadas. *Historical Biogeography: An Introduction*. Harvard University Press, Cambridge, Massachusetts, 2003.

[CLW95]   M. Crisp, H. P. Linder, and P. Weston. Cladistic Biogeography of Plants in Australia and New Guinea: Congruent Pattern Reveals Two Endemic Tropical Tracts. *Systematic Biology*, 44(4):457–473, 1995.

[Day85]   W.H.E. Day. Optimal Algorithms for Comparing Trees with Labeled Leaves. *Journal of Classification*, 2:7–28, 1985.

[EO05]   B. C. Emerson and P. Oromi. Diversification of the Forest Beetle Genus Tarphius on the Canary Islands, and the Evolutionary Origins of Island Endemics. *Evolution*, 59(3):586–598, 2005.

[FCPT95a]   M. Farach-Colton, T.M. Przytycka, and M. Thorup. On the Agreement of Many Trees. *Information Processing Letters*, 55:297–301, 1995.

[FCPT95b]   M. Farach-Colton, T.M. Przytycka, and M. Thorup. The Maximum Agreement Subtree Problem for Binary Trees. 1995. Manuscript.

[FCT94]   M. Farach-Colton and M. Thorup. Sparse Dynamic Programming for Evolutionary Tree Comparison. In *Proc. of the 35th Annual Symp. on the Foundations of Computer Science*, pages 770–779, 1994.

[FG85]   C.R. Finden and A.D. Gordon. Obtaining Common Pruned Trees. *Journal of Classification*, 2:255–276, 1985.

[GKK94]   W. D. Goddard, E. Kubicka, and G. Kubicki. The Agreement Metric for Labeled Binary Trees. *Mathematical Biosciences*, 123:215–226, 1994.

[GvVB02]   M. D. Green, M.G.P. van Veller, and D. R. Brooks. Assessing Modes of Speciation: Range Asymmetry and Biogeographical Congruence. *Cladistics*, 18(1):112–124, 2002.

[Jac04a]   A. P. Jackson. Cophylogeny of the Ficus Microcosm. *Biological Review*, 79(4):751–768, 2004.

[Jac04b]   A. P. Jackson. Phylogeny and Biogeography of the Malagasy and Australasian Rainbowfishes (Teleostei : Melanotaenioidei): Gondwanan Vicariance and Evolution in Freshwater. *Molecular Phylogenetics and Evolution*, 33(3):719–734, 2004.

[JEOH00]   C. Juan, B.C. Emerson, P. Oromi, and G. M. Hewitt. Colonization and Diversification: Towards a Phylogeographic Synthesis for the Canary Islands. *Trends in Ecology and Evolution*, 15(3):104–109, 2000.

[NP81]   G. Nelson and N. Platnick. *Systematics and Biogeograpy*. Columbia University Press, New York, 1981.

[Pag88]   R. D. M. Page. Quantitative Cladistic Biogeograpy: Constructing and Comparing Area Cladograms. *Systematic Zoology*, 37:254–270, 1988.

[Pag90]   R. D. M. Page. Temporal Congruence and Cladistic Analysis of Biogeography and Cospeciation. *Systematic Zoology*, 39:205–226, 1990.

[Pag94]   R.D.M. Page. Maps Between Trees and Cladistic Analysis of Historical Associations Among Genes. *Systematic Biology*, 43(1):58–77, 1994.

[PC98]   R.D.M. Page and M. Charleston. Trees Within Trees: Phylogenies and Historical Associations. *Trends in Ecology and Evolution*, 13(9):356–359, 1998.

[RF81]   D.F. Robinson and L.R. Foulds. Comparison of Phylogenetic Trees. *Mathematical Biosciences*, 53:131–147, 1981.

[Ros78]   D. E. Rosen. Vicariant Patterns and Historical Explanation in Biogeograpy. *Systematic Zoology*, 27:159–188, 1978.

[SW93]   M. Steel and T. Warnow. Kaikoura Tree Theorems: Computing the Maximum Agreement Subtree. *Information Processing Letters*, 48:77–82, 1993.

[vVZK99]   M.G.P. van Veller, M. Zandee, and D. J. Kornet. Two Requirements for Obtaining Common Patterns Under Different Assumptions in Vicariance Biogeography. *Cladistics*, 15:393–405, 1999.