Information Processing Letters 136 (2018) 96–101

Contents lists available at ScienceDirect

Information Processing Letters

www.elsevier.com/locate/ipl

Geometric medians in reconciliation spaces of phylogenetic trees

Katharina T. Huber[a], Vincent Moulton[a,*], Marie-France Sagot[b,c], Blerina Sinaimeri[b,c]

[a] School of Computing Sciences, University of East Anglia, Norwich, NR4 7TJ, UK
[b] Inria Grenoble – Rhône-Alpes; Inovallée 655, avenue de l'Europe, Montbonnot, 38334 Saint Ismier cedex, France
[c] Université de Lyon, F-69000, Lyon; Université Lyon 1; CNRS, UMR5558; 43 Boulevard du 11 Novembre 1918, 69622 Villeurbanne cedex, France

ARTICLE INFO

Article history:
Received 24 May 2017
Received in revised form 28 March 2018
Accepted 2 April 2018
Available online 12 April 2018
Communicated by Kun-Mao Chao

Keywords:
Combinatorial problems
Geometric median
Reconciliation space
Edit-distance
Consensus reconciliation

ABSTRACT

In evolutionary biology, it is common to study how various entities evolve together, for example, how parasites coevolve with their host, or genes with their species. Coevolution is commonly modelled by considering certain maps or *reconciliations* from one evolutionary tree P to another H, all of which induce the same map ϕ between the leaf-sets of P and H (corresponding to present-day associations). Recently, there has been much interest in studying spaces of reconciliations, which arise by defining some metric d on the set $\mathcal{R}(P, H, \phi)$ of all possible reconciliations between P and H.

In this paper, we study the following question: How do we compute a *geometric median* for a given subset Ψ of $\mathcal{R}(P, H, \phi)$ relative to d, i.e. an element $\psi_{med} \in \mathcal{R}(P, H, \phi)$ such that

$$\sum_{\psi' \in \Psi} d(\psi_{med}, \psi') \leq \sum_{\psi' \in \Psi} d(\psi, \psi')$$

holds for all $\psi \in \mathcal{R}(P, H, \phi)$? For a model where so-called host-switches or transfers are not allowed, and for a commonly used metric d called the *edit-distance*, we show that it is possible to compute a geometric median for a set Ψ in $\mathcal{R}(P, H, \phi)$ in polynomial time. We expect that this result could open up new directions for computing a consensus for a set of reconciliations.

1. Introduction

In phylogenetics, the reconciliation problem involves trying to find a map that reconciles one leaf-labelled evolutionary tree with another [4,11]. It has important applications in areas such as ecology and genomics, and arises in various situations. For example, biologists are interested in understanding how parasite and host species [7], genes and species [8], or species and habitats coevolve [12] (in what follows we shall use terminology for host-parasite relationships to keep things concrete).

More formally, a *phylogenetic tree* T is a rooted, binary tree (i.e. every vertex of T that is not the root or a leaf has indegree 1 and outdegree 2), which has root vertex ρ_T (with indegree 0 and outdegree 2). Given a *host-parasite triple* (P, H, ϕ), that is, two phylogenetic trees P and H (the parasite and the host tree, respectively), whose leaf-sets represent present-day species, and a map $\phi : L(P) \rightarrow$

[*] Corresponding author.
E-mail addresses: k.huber@uea.ac.uk (K.T. Huber), v.moulton@uea.ac.uk (V. Moulton), marie-france.sagot@inria.fr (M.-F. Sagot), blerina.sinaimeri@inria.fr (B. Sinaimeri).

https://doi.org/10.1016/j.ipl.2018.04.001
0020-0190/© 2018 Elsevier B.V. All rights reserved.

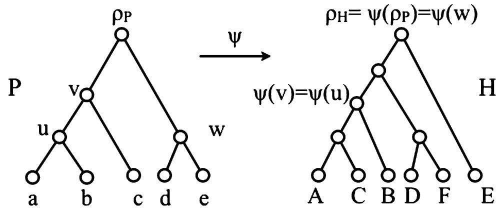

Fig. 1. An example of a reconciliation map. Note that ϕ is given by $\phi(a) = A, \ldots, \phi(e) = E$.

$L(H)$ between their leaf-sets (describing which parasite is currently on which host), a *reconciliation map* is a map $\psi : V(P) \to V(H)$ between their vertex sets which satisfies:

(i) The map ψ restricted to $L(P)$ equals ϕ.
(ii) If v is a vertex in the interior of P, then $\psi(v)$ is either strictly above or equal to $\psi(v')$, for any child v' of v.

We present an example of such a map in Fig. 1. Note that various definitions have been proposed for reconciliation maps (see e.g. [8]). These model evolutionary processes including cospeciation (a host and parasite speciate together), duplication (a parasite speciates on a host), loss (a host speciates but not its parasite) and host-switches (e.g. a parasite switches to another host). In this paper, we are using the definition for a reconciliation map presented in [7,13], with the added assumption that we do not allow host-switches.

In general, several algorithms have been developed to compute optimal and suboptimal reconciliations for a pair of trees relative to some predefined cost-function (cf. e.g. [8,9]). When host-switches are not allowed (as in this paper), collections of suboptimal reconciliations can contain thousands of elements [9], and for more complex models (e.g. where host-switches are permitted), this can be the case even for collections of optimal reconciliations [7]. It is thus quite natural to consider properties of the set of all possible reconciliations endowed with some metric which also permits their comparison. These so-called *reconciliation spaces* are of growing importance in the literature [1,3,9,10,14] and permit quantitative analysis of the behaviour of reconciliation maps.

In this paper, we are interested in the problem of computing geometric medians in reconciliation spaces. In general, for Y a finite set endowed with a metric D, and $Y' \subseteq Y$, an element $y^* \in Y$ is a *geometric median* for Y' in Y if

$$\sum_{y' \in Y'} D(y^*, y') = \min\{\sum_{y' \in Y'} D(y, y') : y \in Y\}.$$

Such elements are useful as they can act as an element which summarizes or forms a consensus for the set Y'. Within computational biology, geometric medians (and the closely related concept of *centroids*) have been used in phylogenetics to form a consensus tree for a set of phylogenetic trees [2], and in RNA secondary structure prediction to derive a consensus structure for a set of suboptimal RNA structures [6]. We therefore expect that being able to compute geometric medians in reconciliation spaces should be a useful addition to the theory of reconciliations (e.g. for computing a consensus of a collection of reconciliations).

We now summarize the contents of the rest of the paper. In the next section we present some preliminary definitions and results. This includes the definition of the edit-distance, a metric on the set $\mathcal{R}(P, H, \phi)$ of all reconciliation maps for a host-parasite triple (P, H, ϕ). Variants of this distance have been previously used to quantitatively analyse collections of reconciliations (cf. e.g. [9]). In Section 3, we present some basic observations concerning medians, which we then use in Section 4 to define the concept of a *median reconciliation* for a subset Ψ of $\mathcal{R}(P, H, \phi)$ (Theorem 2). In Section 5, we then show that a median reconciliation is in fact a geometric median for Ψ in $\mathcal{R}(P, H, \phi)$ relative to the edit-distance (Theorem 4). We also explain how to compute a geometric median in polynomial time, even though it should be noted that $\mathcal{R}(P, H, \phi)$ can be exponential in size (see e.g. [7, p.2]). We conclude in Section 6, with a brief discussion of some potential future directions.

2. Preliminaries

For a phylogenetic tree T, denote the set of interior vertices of T by $V^o(T) = V(T) - L(T)$, and the root by ρ_T. If $v \in V^o(T)$, we let $Ch(v)$ denote the set of children of v, and if $v \in V(T) - \{\rho_T\}$, we let $par(v)$ denote the parent of v in T.

We denote by $\succeq_T$ the partial order of $V(T)$ given by T. In case the context is clear, we just use $\succeq$. Also, we say for vertices $x, y \in V(T)$ with $x \succeq y$ that y is *below* x and that x is *above* y. Furthermore, we say that y is *strictly below* x if y is below x and $x \neq y$ and that x is *strictly above* y if x is above y and $x \neq y$. In that case, we also put $x \succ y$. If L is a subset of $L(T)$ of size at least two, we let $lca_T(L) = lca(L)$ denote the *least common ancestor* of the set L, that is, the lowest vertex in T which is above every element of L (with respect to the ordering $\succeq_T$). If $|L| = 1$, then we set $lca_T(L) = x$ where x is the unique element in L.

Now, let (P, H, ϕ) be a host-parasite triple. For $v \in V(P)$, we let

$$m(v) = lca_H(\{\phi(x) : x \in L(P) \text{ and } v \succeq_P x\}).$$

We also let $A(v)$ be the subset of $V(H)$ given by

$$A(v) = \{u \in V(H) : \rho_H \succeq u \succeq m(v)\}.$$

We now make some observations (cf. also [9]) – we prove only (R2) as the rest are straight-forward to check:
(R0) If $v \in V^o(P)$ and $v' \in Ch(v)$, then $m(v) \succeq m(v')$ and $A(v) \subseteq A(v')$.
(R1) If $\psi \in \mathcal{R}(P, H, \phi)$, $x \in L(P)$, $v \in V(P)$ and $v \succeq x$, then $\psi(v) \succeq \psi(x) = \phi(x)$.
(R2) If $\psi \in \mathcal{R}(P, H, \phi)$, then for all $v \in V(P)$ we have $\psi(v) \in A(v)$.

Proof. If $v \in L(P)$ then the statement clearly holds. Suppose now there exist some $v \in V^o(P)$, but $\psi(v) \notin A(v)$. Since $m(v) \in A(v)$, it suffices to consider the following two cases: