ELSEVIER

Discrete Applied Mathematics 91 (1999) 215–233

DISCRETE APPLIED MATHEMATICS

Retractions of finite distance functions onto tree metrics

Vincent Moulton[1], Mike Steel*

Biomathematics Research Centre, University of Canterbury, Private Bag 4800, Christchurch, New Zealand

Received 19 June 1997; received in revised form 23 March 1998; accepted 28 July 1998

Abstract

Trees with positively weighted edges induce a natural metric on any subset of vertices, however not every metric is representable in this way. A problem arising in areas of classification, particularly in evolutionary biology, is how to approximate an arbitrary distance function by such a tree metric, and thereby estimate the underlying tree that generated the data. Such transformations, from distances to tree metrics (and thereby to edge-weighted trees) should have some basic properties such as continuity, but this is lacking in several popular methods, for example (as we show) in "neighbor joining." However, a continuous transformation, due to Buneman, frequently leads to uninteresting trees. We show how Buneman's construction can be refined so as to lead to more informative trees without sacrificing continuity, and we provide two simple examples of its use. We also provide a sufficient condition for both the Buneman construction, and its refinement to correctly recover the underlying tree.

AMS classification: 05C05, 92B10

Keywords: Trees; 4-point condition; Retraction; Isolation index

1. Introduction

A distance function d on a finite set S is said to be a *tree metric* if there exists a tree $T = (V,E)$, a map $L : S \to V$, called a *labelling*, and a map $w : E \to \mathbb{R}_{>0}$, called an *edge weighting*, such that for all $x, y \in S$, d_{xy} is the sum of $w(e)$ over all edges e in the unique path in T connecting vertices $L(x)$ and $L(y)$. Thus, a tree metric d is a pseudo-metric (with $d_{xy} = 0$ precisely when $L(x) = L(y)$).

We may assume that the tree T has no vertices in $V - L(S)$ of degree less than or equal to two, since, as is easily seen, any tree metric on S can be realized by such a tree with a suitable edge weighting. We call such a tree T (together with its associated labelling L) an *S-tree*.

* Corresponding author: E-mail: m.steel@math.canterbury.ac.nz.

[1] Thanks to the NZ Lotteries Commission for its support.

0166-218X/99/$ – see front matter
PII: S0166-218X(98)00128-0

S-trees and tree metrics arise in many contexts, particularly in phylogenetic analysis in evolutionary biology (see, for example, [3, 14]).

One classical and simple result is that a tree metric can arise from only one triple (T, L, w) where T is an S-tree, and w is an edge weighting of T [2, 8, 18, 20]. Thus tree metrics are in a natural bijective correspondence with positively edge-weighted S-trees, and, furthermore, there exist fast algorithms for recovering the triple (T, L, w) from d (see, for example, [2, 3, 13]). We refer to T (with its associated labelling L) as the S-tree *associated with* d.

An important problem in applications (such as in biology) is how to take an arbitrary distance function, which is in some sense an estimate of (but not itself) a tree metric, and recover a"nearby" tree metric, and thereby the associated (edge weighted) S-tree. As Buneman [8] pointed out, it is desirable that such a map, from distance functions onto tree metrics, should be *continuous*. That is, a small change in the input distance function should not result in a drastically different edge-weighted tree. This is important for applications where distances are merely estimates obtained from imperfect data, often subject to stochastic effects (in biology, random mutations in DNA sequences). Surprisingly, one of the most popular methods currently in use in phylogenetic analysis – neighbor joining – fails on this count, as we show below in Section 4.2. Some earlier methods which attempt to find a closest tree metric to a given distance function are also discontinuous.

This prompted Buneman [8] to construct a continous map from metrics onto tree metrics, which we recall in Section 4. Buneman (and others subsequently, see [4]) have noticed that such a map applied to real data (particularly when S is large) often leads to highly unresolved "star-like" trees, with few internal edges. Such trees tell a biologist little about the underlying evolutionary relationships. This has led to a preference by practitioners for other (discontinuous) methods as these methods generally construct fully resolved trees, which therefore appear to provide more information about the underlying evolutionary history. Yet, as pointed out in [8], such methods will construct fully resolved trees even if fed completely random data. In this case the evolutionary "information" contained in the tree is completely phantom, and liable to change completely under a small perturbation. Buneman suggests that the non-resolution observed in his tree building method is "the price paid for continuity".

One escape from this dilemma has been to modify Buneman's construction so as to output a graph, rather than necessarily a tree, via the elegant split decomposition theory of Bandelt and Dress [4]. Here we adopt a slightly different approach – by modifying Buneman's construction in an alternative way (see Section 5) we are able to ensure that the output is always a tree, but it will, in general, give a more highly resolved output tree than Buneman's method. This opens up the possibility of constructing still further maps, aimed at extracting as much "tree-like" information from the data as possible, without sacrificing continuity.

Of course any map from distance functions to tree metrics should also have the property that when applied to a distance function d which is already a tree metric it returns d. Two further desirable properties are *homogeneity* and *equivariance* which we

describe below. We call any continuous map which satisfies these last three properties a *good* map.

Before discussing Buneman's good map and its refinement, we first describe, in the following section, two underlying metric structures on the space of tree metrics, and the relationship between them (Theorem 2.1). In Section 3 we define retractions and *good* maps, while in Section 4 we describe Buneman's good map, and other related constructions, along with a proof that the commonly used neighbor-joining method is not a retraction. In Section 5 we describe and establish the claimed properties of our proposed refinement of Buneman's map. Towards the end of this section we provide a "minimal edge" analysis of both the classic and refined Buneman map, which complements a similar result obtained for the neighbor-joining method in [1]. Section 5 concludes by providing two simple examples of the use of the refined Buneman map, including an application to biological data. Section 6 summarizes the paper, and addresses some remaining questions.

2. Tree metrics and edge-weighted S-trees

Let $S := \{1,\ldots,n\}$, and define

$$\mathscr{D}(S) := \{d : S \times S \to \mathbb{R}_{\geqslant 0} : d_{xy} = d_{yx},\ d_{xx} = 0 \text{ for all } x, y \in S\}$$

to be the set of *distance functions* on S. Endow $\mathscr{D}(S)$ with the l^p norm, that is, set

$$\|d - d'\|_p = \begin{cases} \left(\sum_{i,j} |d_{ij} - d'_{ij}|^p\right)^{1/p} & p = 1,2,\ldots \\ \max_{i,j} |d_{ij} - d'_{ij}| & p = \infty. \end{cases}$$

Let $\mathscr{T}(S)$ be the subspace of $\mathscr{D}(S)$ consisting of tree metrics, and $\mathscr{S}(S)$ be the set of *splits* of S, that is, bipartitions of S. Note that each edge of an S-tree induces a split of S defined by the two non-empty subsets of S that label the two subtrees of T when e is deleted. We say that this split is a *split of* T and is *associated* to edge e. Notice also that any tree metric $d \in \mathscr{T}(S)$ can be conveniently written in the form

$$d = \sum_{\sigma \in \mathscr{S}(S)} \lambda_\sigma \delta_\sigma, \tag{1}$$

where

$$\lambda_\sigma = \lambda_\sigma(d) := \begin{cases} w(e) & \text{if } \sigma \text{ is associated to } e, \\ 0 & \text{if } \sigma \text{ is not associated to any edge of } T, \end{cases}$$

and where

$$\delta_\sigma(i,j) := \begin{cases} 1 & \text{if } \sigma \text{ separates } i \text{ and } j, \\ 0 & \text{otherwise} \end{cases}$$

($\sigma = \{A,B\}$ *separates* i and j if $i \neq j$, and $|\{i,j\} \cap A| = 1$).

Let $\lambda(d)$ be the vector $[\lambda_\sigma(d)]$ which lies in $\mathbb{R}^{|\mathscr{S}(S)|}$,

$$\mathscr{W}(S) := \{\lambda(d) : d \in \mathscr{T}(S)\},$$

and endow $\mathscr{W}(S) \subseteq \mathbb{R}^{|\mathscr{S}(S)|}$ with the l^p norm. In the next section we define some maps from $\mathscr{S}(S)$ to $\mathbb{R}$, called indices, in order to prove that the map

$$\lambda : \mathscr{T}(S) \to \mathscr{W}(S); d \mapsto \lambda(d)$$

is a homeomorphism.

2.1. Indices

Given $d \in \mathscr{D}(S)$ several useful maps (*indices*) from $\mathscr{S}(S)$ into $\bar{\mathbb{R}}$ can be defined. We review these here, adopting the convenient shorthand xy for d_{xy}. Suppose that $\sigma = \{A, B\}$ is a split of S. Let

$$\mu_\sigma = \mu_\sigma(d) := \frac{1}{2} \min_{a,a' \in A, b,b' \in B} \{\min\{ab + a'b', ab' + a'b\} - (aa' + bb')\},$$

$$\mu_\sigma^+ = \mu_\sigma^+(d) := \max\{0, \mu_\sigma\},$$

$$\alpha_\sigma = \alpha_\sigma(d) := \frac{1}{2} \min_{a,a' \in A, b,b' \in B} \{\max\{ab + a'b', ab' + a'b\} - (aa' + bb')\},$$

$$\alpha_\sigma^+ = \alpha_\sigma^+(d) := \max\{0, \alpha_\sigma\}.$$

The map μ is the *Buneman index* [8], while α^+ is the *isolation index* [4]. Clearly, for any $\sigma \in \mathscr{S}(S)$, we have $\mu_\sigma \leqslant \alpha_\sigma$ and $\mu_\sigma^+ \leqslant \alpha_\sigma^+$. The proof of the following lemma can be found in [4, 8].

Lemma 2.1. *If d is an element of $\mathscr{T}(S)$ with $d = \sum_\sigma \lambda_\sigma \delta_\sigma$, then $\lambda_\sigma = \mu_\sigma^+ = \alpha_\sigma^+$ for all $\sigma \in \mathscr{S}(S)$.*

2.2. $\mathscr{T}(S)$ and $\mathscr{W}(S)$ are homeomorphic

The l^1 norm on the the space $\mathscr{W}(S)$ was proposed in [17] as a natural metric for comparing edge-weighted trees. The following theorem shows that $\mathscr{W}(S)$ and $\mathscr{T}(S)$ are homeomorphic. In particular the question of whether or not a map of $\mathscr{D}(S)$ into $\mathscr{T}(S)$ is good does not depend on whether we view the output as a distance function or as an edge-weighted S-tree.

Theorem 2.1. *For $d, d' \in \mathscr{T}(S)$, we have*

$$\|d - d'\|_\infty \leqslant \|\lambda(d) - \lambda(d')\|_1,$$

$$\|\lambda(d) - \lambda(d')\|_\infty \leqslant 2\|d - d'\|_\infty,$$

and both of these inequalities can be equalities for any S.

Proof. Writing d, d' in the form of Eq. (1) we have

$$\begin{aligned}\|d - d'\|_\infty &= \max_{i,j} |d_{ij} - d'_{ij}| \\ &= \max_{i,j} \Big| \sum_{\{\sigma \in \mathscr{S}(S)\}} (\lambda_\sigma - \lambda'_\sigma)\delta_\sigma(i,j) \Big| \\ &\leqslant \max_{i,j} \sum_{\{\sigma \in \mathscr{S}(S)\}} |\lambda_\sigma - \lambda'_\sigma| \delta_\sigma(i,j) \\ &\leqslant \sum_{\{\sigma \in \mathscr{S}(S)\}} |\lambda_\sigma - \lambda'_\sigma| \max_{i,j} \{\delta_\sigma(i,j)\} \\ &= \|\lambda(d) - \lambda(d')\|_1.\end{aligned}$$

The second inequality in Theorem 2.1 is established in [11]. This completes the proof. □

To see that the inequalities can both be equalities we give the following two examples.

For the first inequality let d be the tree metric induced by the S-tree given by labelling bijectively the degree one vertices of a star tree (a tree having just one vertex of degree larger than 1) by the elements of S, and assigning weight α to each edge. Let d' be defined in the same way, except that we assign one of the edges weight β instead of α. Then we immediately see that

$$\|d - d'\|_\infty = \|\lambda(d) - \lambda(d')\|_1 = |\alpha - \beta|.$$

For the second inequality, take a tree with four leaves, labelled bijectively by S, and with five edges. Let d be the metric on S induced by assigning weight 2 to all five edges; let d' be the metric on S induced by assigning weight 1 to the central edge and 9/4 to the other four edges. Then,

$$\|\lambda(d) - \lambda(d')\|_\infty = 1 = 2\|d - d'\|_\infty.$$

This completes the proof.

Corollary 2.1. *The map* $\lambda : \mathscr{T}(S) \to \mathscr{W}(S); d \mapsto \lambda(d)$ *is a homeomorphism.*

2.3. δ-hyperbolicity

Given $d \in \mathscr{D}(S)$ and $\delta \geqslant 0$, d is said to be *δ-hyperbolic* if

$$d_{ij} + d_{kl} \leqslant \max\{d_{ik} + d_{jl}, d_{il} + d_{jk}\} + \delta$$

for all $i, j, k, l \in S$. This is a relaxation of the *four-point condition*, in which $\delta = 0$ (for a discussion of this point see [10]). A classical result states that a pseudo-metric d is contained in $\mathscr{T}(S)$ if and only if d is 0-hyperbolic, a result dating back to the

1960s and work by the Russians Zaretsky [20] and Smolensky [18] (see also [2, 8]). More generally, a result originally given in [12], and which is also described in [6], states that if d is δ-hyperbolic, then there exists a $d' \in \mathscr{T}(S)$ with

$$\|d - d'\|_\infty \leqslant (1 + \log_2 n)\delta,$$

where $n = |S|$. Thus, if δ is small, then d is close to a tree metric up to a term that grows slowly in n.

If $d \in \mathscr{D}(S)$ is δ-hyperbolic, then we can relate δ with the Buneman and isolation indices in the following way, which we shall use later.

Lemma 2.2. *If the metric d on S is δ-hyperbolic, and $\sigma \in \mathscr{S}(S)$ then $\alpha_\sigma > \delta/2$ implies that $\mu_\sigma > 0$.*

Proof. For $\sigma = \{A, B\}$, write

$$\mu_\sigma = \tfrac{1}{2}(\min\{P, Q\} - R),$$

where $P = ab' + a'b$, $Q = ab + a'b'$, and $R = aa' + bb'$, for suitably chosen $a, a' \in A$, and $b, b' \in B$. Then

$$\alpha_\sigma \leqslant \tfrac{1}{2}(\max\{P, Q\} - R),$$

and so, if $\alpha_\sigma > \delta/2$, then

$$\max\{P, Q\} - R > v(P, Q, R), \tag{2}$$

where $v(P, Q, R)$ is the difference between the largest and second largest value in the triple P, Q, R. Since $v(P, Q, R) \geqslant 0$, Eq. (2) implies that either P or Q is at least R, and, without loss of generality, we may assume that $P \geqslant R$. But then $Q \geqslant R$ also, for if $P \geqslant R > Q$, then from Eq. (2)

$$P - R > P - R,$$

which is a contradiction.

Thus, we may assume that either $P \geqslant Q \geqslant R$, or $Q \geqslant P \geqslant R$. In the former case Eq. (2) gives the following implications:

$$\begin{aligned} P - R > P - Q &\Rightarrow Q - R > 0 \\ &\Rightarrow \min\{P, Q\} - R > 0 \\ &\Rightarrow \mu_\sigma > 0, \end{aligned}$$

and in the latter case, an analogous argument applies to show that $\mu_\sigma > 0$, thereby completing the proof. □

3. Retractions

A map $\varphi : \mathscr{D}(S) \to \mathscr{D}(S)$ is a *retraction* onto $\mathscr{T}(S)$ if
(i) φ is *continuous*,
(ii) $\varphi(d) \in \mathscr{T}(S)$ for all $d \in \mathscr{D}(S)$, and
(iii) $\varphi(d) = d$ for all $d \in \mathscr{T}(S)$.
Furthermore, if such a retraction φ is *homogeneous*, that is, if

$$\varphi(\lambda d) = \lambda\varphi(d)$$

for all $\lambda > 0$ and $d \in \mathscr{D}(S)$, and if φ is *equivariant*, that is, for all $\tau \in \sum_S$ (the permutation group on S)

$$\varphi(d^\tau) = \varphi(d)^\tau,$$

where

$$(d^\tau)_{ij} = d_{\tau(i)\tau(j)},$$

then we say that φ is *good*. These last two properties are desirable in applications in requiring the method to be independent of the units in which d is measured and the names given to the objects in S, respectively [15, 19].

Regarding homogeneity, we note that in biological applications distance functions are frequently transformed by non-linear (typically logarithmic) functions before being used to reconstruct trees. Clearly such functions are not homogeneous, so the requirement of homogeneity is meant to apply simply to the transformed distances, not to the input distances. Homogeneity is desirable for applications to transformed distances as these distances generally estimate the expected number of mutations that have occurred between pairs of species for sequences that have been undergoing site mutations at some rate over a period of time (see [14]) – homogeneity thus becomes the requirement that the edge weights on the output trees should be proportional to the expected number of mutations on that edge (and so proportional to time, in case the rate is constant).

Define a partial order on the set of retractions as follows. Given two retractions φ_1, φ_2 of $\mathscr{D}(S)$ onto $\mathscr{T}(S)$, and a metric $d \in \mathscr{D}(S)$, let

$$\varphi_i(d) = \sum_{\sigma \in \mathscr{S}(S)} \lambda_\sigma^i(d)\delta_\sigma, \quad i = 1, 2.$$

We say that φ_2 *refines* φ_1, written $\varphi_1 \preceq \varphi_2$, if and only if for all $d \in \mathscr{D}(S)$ we have

$$\lambda_\sigma^1(d) \leqslant \lambda_\sigma^2(d),$$

for all $\sigma \in \mathscr{S}(S)$. As can be easily verified, $\preceq$ is a partial order. Note that if $\varphi_1 \preceq \varphi_2$, and if T_1, T_2 are the S-trees associated with $\varphi_1(d)$, $\varphi_2(d)$, respectively, then T_2 is a *refinement* of T_1, in the sense that T_1 can be obtained from T_2 by collapsing a (possibly empty) subset of edges; also the weights on the edges of T_1 are less than or equal to those on T_2 – thus if we regard T_1, T_2 as edge-weighted trees then T_1 is obtained by shrinking (and sometimes collapsing) certain edges of T_2.

4. Examples of retractions

4.1. The Buneman retraction

Two splits $\sigma = \{A,B\}$, $\sigma' = \{A',B'\}$ in $\mathscr{S}(S)$ are said to be *compatible* if at least one of the intersections $A \cap A'$, $A \cap B'$, $B \cap A'$, $B \cap B'$ is empty. If two splits σ, σ' are not compatible then we say that they are *incompatible*, and denote this by writing $\sigma \perp \sigma'$. Clearly any S-tree gives a set of pairwise compatible splits: just take the set of splits induced by the set of edges of the tree. Moreover in [8] it is shown that a set of pairwise compatible splits gives rise to a unique tree.

The following lemma (not stated explicitly in [8]), gives the fundamental link between the Buneman index and the notion of compatibility of splits.

Lemma 4.1. *If $\sigma, \sigma' \in \mathscr{S}(S)$ and $\sigma \perp \sigma'$ then*

$$\mu_\sigma + \mu_{\sigma'} \leqslant 0.$$

Proof. The proof of this result is essentially the same as that for Theorem 5.1 which we give later. □

Corollary 4.1 (Buneman [8]). *The set $\{\sigma : \mu_\sigma > 0\}$ is a pairwise compatible collection of splits, and thus gives rise to a unique S-tree.*

The index μ is the basis for the following good map, which is given in [8]. We define the *Buneman retraction* $\varphi_B : \mathscr{D}(S) \to \mathscr{T}(S)$ by setting

$$\begin{aligned}\varphi_B(d) &:= \sum_{\{\sigma : \mu_\sigma > 0\}} \mu_\sigma \delta_\sigma \\ &= \sum_{\sigma \in \mathscr{S}(S)} \mu_\sigma^+ \delta_\sigma.\end{aligned}$$

By the previous corollary and the properties of the Buneman index μ, φ_B is a good map. In addition, from [8], $\varphi_B(d) \leqslant d$, in the sense that

$$\varphi_B(d)_{ij} \leqslant d_{ij} \quad \text{for all } i,j \in S.$$

4.2. Neighbor joining is not a retraction

The neighbor joining method (NJ) is a popular scheme for building up an S-tree (T,L,w) whose induced metric approximates an input distance d. In this section we show that neighbor joining is not continuous, and hence not a retraction.

We first review the NJ method [14, p. 488]. For each $i \in S$, let $r_i = \sum_{k \in S} d_{ik}$, select a pair $\{i,j\}$ to minimize

$$M_{ij} = d_{ij} - \frac{(r_i + r_j)}{(n-2)},$$

where $n = |S|$, and let d' be the distance function defined on

$$S' := (S - \{i,j\}) \cup \{u\}$$

by setting

$$d'_{xy} = d_{xy} \quad \text{if } x, y \neq u,$$

$$d'_{ux} = \tfrac{1}{2}(d_{ix} + d_{jx} - d_{ij}), \ x \neq u,$$

and

$$d'_{uu} = 0.$$

Let (T', L', w') be the edge weighted tree constructed on S' for d'. Then, on S, let T be the tree obtained from T' by making leaves i and j adjacent to leaf u using new edges e_i, e_j and extending the domain of w' to these two new edges by setting

$$w'(e_i) := \tfrac{1}{2}d_{ij} + \frac{(r_i - r_j)}{2(n-2)},$$

and

$$w'(e_j) = d_{ij} - w'(e_i).$$

Consider the weighted graph metric on the set $\{1,\ldots,4\}$ given in Fig. 1. The discontinuity in NJ arises as x tends to 1 from above and below. In the former case we obtain the tree in Fig. 1 with an internal edge weight of $\frac{1}{2}$ (not 0!). In the latter case, as x tends to 1 from below, we obtain a tree with 1 and 3 on the same side of the central edge (since then $M_{13} = M_{24}$ is minimal). Thus the induced tree metrics are different as $x \to 1+$ and $x \to 1-$, and hence we have a discontinuity. Finally, note that when $x = 1$ the S-tree obtained by neighbor joining depends upon the order in which the elements of $\{1,\ldots,4\}$ are chosen.

4.3. *Retractions based on the isolation index*

For applications to data (particularly when n is large) there are typically few (non-trivial) splits with positive Buneman measure, and so φ_B often produces highly unresolved trees. By contrast, the isolation index is typically positive on a much larger set of splits, however these are generally not pairwise compatible and so do not correspond to a tree. Thus the continuous map $\varphi_I : \mathscr{D}(S) \to \mathscr{D}(S)$ defined by setting

$$\varphi_I(d) := \sum_{\sigma} \alpha_\sigma^+ \delta_\sigma,$$

while fixing $\mathscr{T}(S)$, does not map $\mathscr{D}(S)$ into $\mathscr{T}(S)$, but rather into a larger subspace of $\mathscr{D}(S)$ – for details see [4]. In order to obtain a good map using the isolation index, one might instead take some continuous function $f : \mathscr{D}(S) \to \mathbb{R}_{\geqslant 0}$ and set

$$\varphi_f(d) := \sum_{\sigma \in \mathscr{S}(S)} \max\{0, \alpha_\sigma(d) - f(d)\}\delta_\sigma.$$

The proof of the following lemma is straight forward, and is left to the reader.

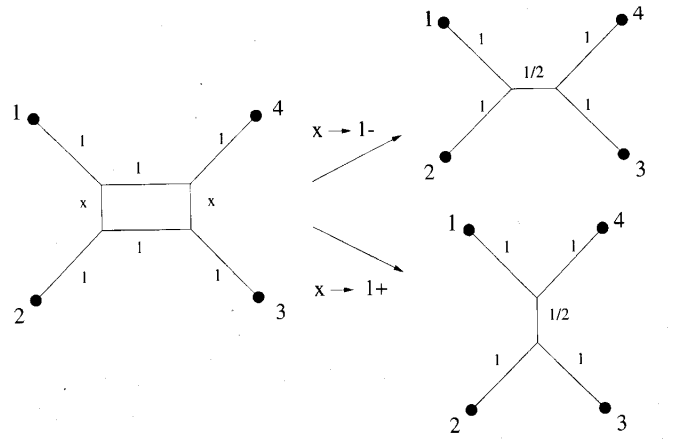

Fig. 1. An example where NJ is not continuous.

Lemma 4.2. *The map φ_f is a good map, provided that*
(i) *f is homogeneous and Σ_S-invariant (i.e. $f(d^\tau) = f(d)$, for all $\tau \in \Sigma_S$),*
(ii) *f is identically zero on $\mathscr{T}(S)$, and*
(iii) *$\{\sigma : \alpha_\sigma(d) > f(d)\}$ is pairwise compatible.*

An example of such a function f is given by $f(d) := \frac{1}{2} hyp(d)$, where $hyp(d)$ is the smallest value of δ such that d is δ-hyperbolic in the sense described in Section 2.3. Then conditions (i) and (ii) of Lemma 4.2 clearly hold, and condition (iii) holds by Lemma 2.2 and Corollary 4.1. However, Lemma 2.2 also shows that choosing $f(d) = \frac{1}{2} hyp(d)$ leads to a no more refined tree than that given by the Buneman retraction.

Alternatively, we might let $f(d)$ be the smallest non-negative real number h for which the set

$$\{\sigma \in \mathscr{S}(S) : \alpha_\sigma(d) > h\}$$

is pairwise compatible. Then we have the following:

Proposition 4.1. *The map φ_f is a good map.*

Proof. We check the three conditions in Lemma 4.2. The map f is clearly homogeneous and Σ_S-invariant, so (i) holds. If $d \in \mathscr{T}(S)$, then

$$\{\sigma \in \mathscr{S}(S) : \alpha_\sigma > 0\}$$

is the set of splits of the S-tree that realizes d, and so is pairwise compatible. Thus $f(d) = 0$, and (ii) holds. Condition (iii) holds by the definition of f.

In general φ_f may not necessarily be more refined than, or even comparable with the Buneman tree. Thus, rather than pursuing this approach here, we turn instead to an alternative approach which guarantees a tree at least as refined as the Buneman tree.

5. Refining the Buneman retraction

In this section we define a new index map $\overline{\mu}_\sigma$ which refines the Buneman index, in the sense that $\overline{\mu}_\sigma \geqslant \mu_\sigma$ for all $\sigma \in \mathscr{S}(S)$, with strict inequality holding for certain cases. We assume throughout this section that $n \geqslant 4$.

For a resolved quartet, $q := ab|cd$, of elements $a, b, c, d \in S$, let

$$\beta_q := \tfrac{1}{2}(\min\{ac + bd, ad + bc\} - (ab + cd)).$$

Thus, given a split $\sigma = \{A, B\}$ of S, the Buneman index of σ is given by

$$\mu_\sigma = \min_{a,a' \in A, b,b' \in B} \{\beta_{aa'|bb'}\}.$$

Let Q be the set of quartets $q = aa'|bb'$ consisting of all unordered choices of $a, a' \in A$, and $b, b' \in B$, insisting, furthermore, that if $|A| \geqslant 2$, then $a \neq a'$ and if $|B| \geqslant 2$, then $b \neq b'$. As we shall see, $|Q|$ is greater than or equal to $n - 3$. Now let $q_1, \ldots, q_{|Q|}$ be an ordering of the elements in Q such that $\beta_{q_i} \leqslant \beta_{q_j}$ for all $1 \leqslant i \leqslant j \leqslant |Q|$, and define the *refined Buneman index* by

$$\overline{\mu}_\sigma := \frac{1}{n-3} \sum_{i=1}^{n-3} \beta_{q_i}.$$

Note that, by definition, $\overline{\mu}_\sigma \geqslant \mu_\sigma$ for all $\sigma \in \mathscr{S}(S)$.

5.1. The refined Buneman index gives trees

To show that the refined Buneman index give us trees in a similar way to the Buneman index, we prove the following analogue of Lemma 4.1.

Theorem 5.1. *If $\sigma, \sigma' \in \mathscr{S}(S)$ and $\sigma \perp \sigma'$ then*

$$\overline{\mu}_\sigma + \overline{\mu}_{\sigma'} \leqslant 0.$$

We prove Theorem 5.1 in two steps, the first of which we state as a lemma.

Lemma 5.1. *Suppose that $\sigma = \{A, B\}$, $\sigma' = \{A', B'\}$, and that $\sigma \perp \sigma'$. Let $x := |A \cap A'|$, $y := |A \cap B'|$, $z := |B \cap B'|$, and $w := |A' \cap B|$. Then*

$$xywz \geqslant n - 3.$$

Proof. Clearly,

$$x + y = |A|,$$

and

$$w + z = |B|,$$

and hence $f(x,w) := xwyz = x(|A|-x)w(|B|-w)$. We want to minimize $f(x,w)$ where $0 < x < |A|$, $0 < w < |B|$, $x,w \in \mathbb{N}$, and $|A|, |B| \geqslant 2$ (this last pair of inequalities arise since $\sigma \perp \sigma'$). Using routine calculus, one can see that the minimum value of f under these constraints is equal to $(|A|-1)(|B|-1)$. Furthermore, since $|A|+|B| = n$, and $|A|, |B| \geqslant 2$ we have $(|A|-1)(|B|-1) \geqslant n-3$, which completes the proof. □

Proof of Theorem 5.1. For a split $\tilde{\sigma} = \{C,D\}$, let

$$\mathscr{C}(\tilde{\sigma}) := \{cc'|dd' : c,c' \in C,\ c \neq c',\ \text{and } d,d' \in D,\ d \neq d'\}.$$

Suppose that $\sigma = \{A,B\}$, $\sigma' = \{A',B'\}$ and $\sigma \perp \sigma'$. Consider the quartets $q := xy|wz$, $q' := xw|yz$, such that $x \in A \cap A'$, $y \in A \cap B'$, $w \in B \cap A'$, and $z \in B \cap B'$. Then, by definition,

$$\beta_q \leqslant \tfrac{1}{2}(xw + yz - xy - wz),$$

and,

$$\beta_{q'} \leqslant \tfrac{1}{2}(xy + wz - xw - yz),$$

so that

$$\beta_q + \beta_{q'} \leqslant 0. \tag{3}$$

Note that $q \in \mathscr{C}(\sigma)$ and $q' \in \mathscr{C}(\sigma')$. By Lemma 5.1, there exist at least $n-3$ choices of such q and q', which we denote by $\hat{q}_i, \hat{q}'_i$, $1 \leqslant i \leqslant n-3$. In particular,

$$\overline{\mu}_\sigma + \overline{\mu}_{\sigma'} \leqslant \frac{1}{n-3} \sum_{i=1}^{n-3} (\beta_{\hat{q}_i} + \beta_{\hat{q}'_i}) \leqslant 0$$

by Eq. (3).

Corollary 5.1. *The set* $\{\sigma : \overline{\mu}_\sigma > 0\}$ *is a pairwise compatible collection of splits, and thus gives rise to a unique S-tree.*

5.2. The refined Buneman index gives a good map

In this section we prove that the map $\psi : \mathscr{D}(S) \to \mathscr{D}(S)$, defined by

$$\psi : d \mapsto \sum_{\{\sigma : \overline{\mu}_\sigma > 0\}} \overline{\mu}_\sigma \delta_\sigma,$$

and which we call the *refined Buneman retraction*, is a good map onto the space of tree metrics. First we show that ψ fixes tree metrics. To do this we require the following technical lemma.

Lemma 5.2. *Suppose that* $x_i \in \mathbb{Z}_{\geqslant 0}$, $0 \leqslant i \leqslant r$, *are such that*

$$\sum_{i=0}^{r} x_i = k \geqslant 2.$$

Suppose that $x_i \geqslant 1$, *for* $i > 0$, *and if* $r = 1$ *then* $x_0 \geqslant 1$. *Then*

$$\sum_{i<j;i,j=0,\ldots,r} x_i x_j + \frac{x_0(x_0-1)}{2} \geqslant k-1.$$

Proof. Note that

$$\sum_{i<j;i,j=0,\ldots,r} x_i x_j + \frac{x_0(x_0-1)}{2} = \frac{1}{2}\left\{k^2 - \left(\sum_{i=1}^{r} x_i^2 + x_0\right)\right\}.$$

Hence, it is sufficient to find the maximum value of the function $\sum_{i=1}^{r} x_i^2 + x_0$, subject to the given constraints.

First, we find the maximum value of the function $\sum_{i=1}^{r} x_i^2$, subject to the constraints $x_i \geqslant 1$, $1 \leqslant i \leqslant r$, and $\sum_{i=1}^{r} x_i = k - x_0$. A simple geometric argument shows that this occurs at any of the r vertices of the convex polytope defined by this set of constraints, i.e. at a point $x_j = k - x_0 - (r-1)$ and $x_i = 1$, $i \neq j$. Substituting these values into the intial sum, we see that we need to show that the minimum value of the expression

$$\tfrac{1}{2}\left\{k^2 - \left(((k-x_0)-(r-1))^2 + (r-1) + x_0\right)\right\},$$

subject to the constraint $0 \leqslant x_0 \leqslant k - r$ (where, if $r = 1$, then $x_0 \geqslant 1$), is greater than or equal to $k-1$. A routine check shows that this is in fact true, thus completing the proof of the lemma. □

Theorem 5.2. *If* d *is a tree metric realized by the triple* (T, L, w), *then*

$$\overline{\mu}_\sigma = \mu_\sigma^+ = \begin{cases} 0 & \text{if } \sigma \text{ is not a split of } T, \\ w(e) & \text{if } \sigma \text{ is the split corresponding to edge } e \text{ of } T. \end{cases}$$

Proof. Suppose that $\sigma = \{A, B\}$ corresponds to edge e of T. We divide the argument into two cases: either $\min\{|A|, |B|\} = 1$ or $|A|, |B| \geqslant 2$.

In the first case we may suppose that $A = \{a\}$, which labels a leaf of T which is an endpoint of the edge e. Let e_i, $1 \leqslant i \leqslant k$, denote the edges in T which have a vertex v in common with edge e, let B_i, $1 \leqslant i \leqslant k$, denote the subset of elements $s \in S$ such that the unique path from $L(s)$ to v passes along edge e_i, and let $B_0 := L^{-1}(v)$. Thus, $B = \bigcup_{i=0}^{k} B_i$.

Set $\mathscr{C}(\sigma) := \{aa|bb' : b, b' \in B,\ b \neq b'\}$. Thus, for $q = aa|bb' \in \mathscr{C}(\sigma)$ we have

$$\beta_q = \tfrac{1}{2}(ab + ab' - bb').$$

If $b \in B_i$, $b' \in B_j$ with $i \neq j$, or $b, b' \in B_0$ with $b \neq b'$, then

$$\beta_q = w(e),$$

otherwise $\beta_q > w(e)$. The number of such pairs b, b' satisfying this equation is equal to

$$\sum_{i<j;i,j=0,\ldots,k} |B_i||B_j| + \frac{|B_0|(|B_0|-1)}{2},$$

which, by Lemma 5.2, (subject to the constraint $\sum_{i=0}^{k} |B_i| = |B| = n-1$) is at least $n-2$. Thus, the average of the $n-3$ values of β_q used in the definition of $\overline{\mu}_\sigma$ is equal to $w(e)$.

We now consider the case $|A|, |B| \geqslant 2$. Let v, w denote the endpoints of edge e. Define the sets B_i, $0 \leqslant i \leqslant k$, as in the case where e was a pendant edge. Define the sets A_i, $0 \leqslant i \leqslant l$, in the same way, but this time using vertex w instead of v (so in case w is a leaf, $l = 0$).

Let $q = aa'|bb'$, where $a, a' \in A$ with $a \neq a'$ and $b, b' \in B$ with $b \neq b'$. In the case where

- $a \in A_i$, $a' \in A_j$ with $i \neq j$, or $a, a' \in A_0$ with $a \neq a'$, and
- $b \in B_p$, $b' \in B_q$ with $p \neq q$, or $b, b' \in B_0$ with $b \neq b'$,

we see that

$$\beta_q = w(e),$$

otherwise $\beta_q > w(e)$. Furthermore, the number of such $q \in \mathscr{C}(\sigma)$ is equal to

$$\left(\sum_{i<j;i,j=0,\ldots,l} |A_i||A_j| + \frac{|A_0|(|A_0|-1)}{2}\right)\left(\sum_{p<q;p,q=0,\ldots,k} |B_p||B_q| + \frac{|B_0|(|B_0|-1)}{2}\right)$$

which, by Lemma 5.2, has a minimal value (subject to the constraints $\sum_{i=0}^{k} |A_i| = |A|$, and $\sum_{i=0}^{k} |B_i| = |B|$) of

$$(|A|-1)(|B|-1).$$

This quantity, in turn, is always greater than or equal to $n-3$, and hence the average of the $n-3$ values of β_q used in the defintion of $\overline{\mu}_\sigma$ is again equal to $w(e)$.

It remains to show that if $\sigma = \{A, B\}$ is not a split of T then $\overline{\mu}_\sigma \leqslant 0$. If σ is incompatible with some split $\sigma' = \{A', B'\}$ of T, then by Lemma 5.1 there exist at least $n-3$ quartets $q = ab|a'b'$ for which $a, a' \in A'$, $b, b' \in B'$ and $\{a, b\} \subseteq A$ and $\{a', b'\} \subseteq B$. Hence, it follows that

$$\beta_q \leqslant \tfrac{1}{2}(aa' + bb' - ab - a'b') < 0.$$

If these $n-3$ quartets are labelled $\hat{q}_1, \cdots, \hat{q}_{n-3}$ then

$$\overline{\mu}_\sigma \leqslant \frac{1}{n-3}(\beta_{\hat{q}_1} + \ldots + \beta_{\hat{q}_{n-3}}) < 0,$$

as claimed.

In case σ is compatible with all of the splits of T (but is not one of them), let T^* be the tree obtained from T by adding a new edge e to induce split σ. Let v, w be

the vertices contained in edge e and define the sets A_i, $0 \leqslant i \leqslant l$, and B_i, $0 \leqslant i \leqslant k$, as before. Then any quartet $q = aa'|bb'$, where $a \in A_i$, $a' \in A_j$ with $i \neq j$, or $a, a' \in A_0$ with $a \neq a'$, and $b \in B_p$, $b' \in B_q$, with $p \neq q$, or $b, b' \in B_0$ with $b \neq b'$, gives

$$\beta_q = \tfrac{1}{2}\{\min\{ab + a'b', ab' + a'b\} - (aa' + bb')\} = 0.$$

As before, the number of such q is at least $n-3$ (by Lemma 5.2) and we again have $\overline{\mu}_\sigma \leqslant 0$.

Theorem 5.3. *The map*

$$\psi : d \mapsto \sum_{\{\sigma : \overline{\mu}_\sigma > 0\}} \overline{\mu}_\sigma \delta_\sigma,$$

is a good map, and $\varphi_B \preceq \psi$.

Proof. That ψ is continuous follows from the fact that $\overline{\mu}_\sigma$ is continuous for each $\sigma \in \mathscr{S}(S)$, which can be easily verified using a standard continuity argument. By Theorem 5.1, $\psi(d)$ is contained in $\mathscr{T}(S)$ for all $d \in \mathscr{D}(S)$, and by Theorem 5.2 $\psi(d)$ equals d for all $d \in \mathscr{T}(S)$. The homogeneity and equivariance of ψ follow by the same arguments that apply to the Buneman retraction. Furthermore, $\varphi_B \preceq \psi$ since $\mu_\sigma \leqslant \overline{\mu}_\sigma$ for all $\sigma \in \mathscr{S}(S)$.

5.3. *Identifying S-trees using the Buneman retraction and its refinement*

We show how the Buneman retraction (or its refinement) essentially identifies the underlying S-tree of a tree metric d' when applied to a distance function d that is close enough to d'. Precisely how close is "close enough" depends on the minimal edge weight of the tree defined by d'. The following theorem complements a similar result obtained by Atteson [1] for the neighbor-joining method (for which the proof is much more involved), and who pioneered this type of minimal edge analysis.

Theorem 5.4. *Let* $\varphi = \varphi_B$ *or* ψ *(the Buneman retraction or its refinement). Suppose that* $d' \in \mathscr{T}(S)$ *has associated S-tree* T *and edge weighting* w. *Let*

$$x := \min\{w(e) : e \in T\},$$

and suppose that for $d \in \mathscr{D}(S)$ *one has*

$$\|d - d'\|_\infty < x/2.$$

Then the S-tree, t, *associated to* $\varphi(d)$ *refines* T *and the weight of any edge in* t *that does not correspond to an edge of* T *is less than* x. *In particular, if* T *is a fully resolved tree (i.e. every vertex has degree* 1 *or* 3*), then* $t = T$.

Proof. Suppose $\varphi = \varphi_B$ and let σ be a split of T corresponding to edge e. Then $\mu_\sigma(d') = w(e) \geqslant x$. Now

$$|\mu_\sigma(d) - \mu_\sigma(d')| \leqslant 2\delta, \tag{4}$$

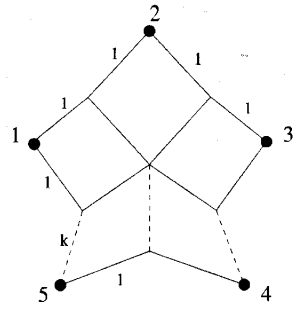

Fig. 2. All edges are weighted 1 except dotted edges, which are weighted k.

where $\delta = ||d - d'||_\infty$, as in the proof of Theorem 2.1. Hence, since $\delta < x/2$,

$$\begin{aligned}\mu_\sigma(d) &\geqslant \mu_\sigma(d') - 2\delta \\ &> x - x = 0,\end{aligned}$$

and so σ is a split of t. Thus t refines T, and in particular, if T is fully resolved then $t = T$.

If T is not fully resolved and σ is a split of t but not T, then by (4)

$$\begin{aligned}\mu_\sigma(d) &\leqslant \mu_\sigma(d') + 2\delta \\ &< 0 + x,\end{aligned}$$

and we deduce that the edge e of t corresponding to σ has weight less than x.

The proof for $\varphi = \psi$ is exactly the same, except that the justification of the analogue of (4) is slightly more involved.

5.4. The refined Buneman retraction is a strict refinement of the Buneman retraction

In this section we give two examples to illustrate that, in certain cases, the refined Buneman retraction gives us a tree which strictly refines the tree given by the Buneman retraction, i.e. $\varphi_B \prec \psi$.

5.4.1. Example 1

For our first example, consider the metric d_k on the set $\{1,\ldots,5\}$ given by the edge-weighted graph metric in Fig. 2, where all edges are weighted length one, except those which are dotted, which all receive edge weight k, for some $k \geqslant 0$.

The Buneman tree for d_k depends upon the value of k. For the case $0 \leqslant k \leqslant 2$ the Buneman tree is simply a vertex. If $k \geqslant 2$, then the Buneman tree consists of one edge of length $k-2$, with its endpoints labelled by $\{1,2,3\}$ and $\{4,5\}$. Thus, in either case, the Buneman tree is highly unresolved (in the sense of [3]).

However, in contrast to this, the refined Buneman tree (i.e. that given by using the refined Buneman index), the topology of which also depends upon k, and which is

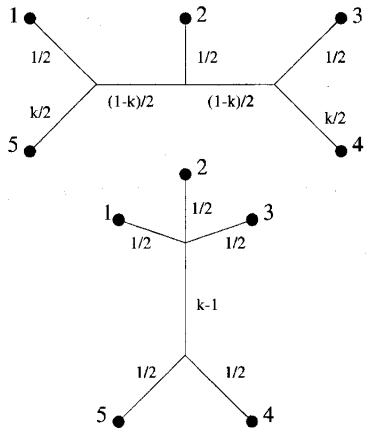

Fig. 3. The refined Buneman tree: the top tree is for the case $0 \leqslant k \leqslant 1$ and the bottom for the case $1 \leqslant k$.

shown in shown in Fig. 3, is fully resolved for $k > 0$, $k \neq 1$. Note that in the case where $k = 1$ we get, as expected, a star tree.

Finally, note that the tree obtained from d_k by using the retraction defined by the good map φ_f in Theorem 4.1 is the same as the Buneman tree, except that the edge appears for $k \geqslant 1$, and has length equal to $k - 1$. Thus, for $1 < k < 2$ the tree $\varphi_f(d_k)$ refines the Buneman tree. Also, the *splitstree graph* [5, 8] of the edge weighted graph in Fig. 2, given by considering all those splits $\sigma \in \mathscr{S}$ with isolation index $\alpha_\sigma > 0$, is in fact the graph itself.

5.4.2. Example 2

As in [16] we analyzed the 16S rRNA sequences of seven chloroplasts and the cyanobacterium *Anacystis nidulans* by using as our metric d the logdet values which correct for multiple site substitutions under a non-stationary Markov model (for futher details see [16]). The Buneman tree, and refined Buneman tree for d were then constructed using the computer package *SplitsTree 2.2* [5]. The Buneman tree is highly unresolved, with a vertex of degree six, and two internal edges. By contrast, the refined Buneman tree has just one vertex of degree four, and all other non-leaf vertices have degree 3, indeed it is essentially the tree reported in [16] with just one edge collapsed.

6. Conclusion

We have shown that our extension of Buneman's construction is valid, and leads to a map which is more refined, at least on certain inputs. It would be interesting to see if there are other such refinements. It would also be useful to find ways of scaling $\psi(d)$ so that it matches d more closely (in Example 5.4.2, for instance, the refined Buneman tree distances underestimate d). One possibility would be to let M be the maximum

(or average) value that d takes, M' be the maximum (or average) distance in the tree $\psi(d)$, and then to multiply each value of $\psi(d)$ by M/M'. In ([8]) it is shown that the Buneman tree distances are always less than or equal to the input distances, and it is an interesting question whether this theorem also holds for the modified Buneman map. Finally, we remark that it is very useful for applications in cases where n is large to be able to compute the refined Buneman tree by an algorithm whose running time grows polynomially with n. Fortunately such an algorithm has recently been described, and we refer the interested reader to [7].

Acknowledgements

This work was supported in part by a New Zealand Marsden Fund research grant (UOC-516). The authors would like to thank David Bryant and Chris Tuffley for some helpful comments, and Daniel Huson for programming. SplitsTree 2.2 is available via ftp at ftp.uni-bielefeld.de/pub/math/splits/splitstree2. We also thank the two anonymous referees for numerous helpful comments.

References

[1] K. Atteson, The performance of neighbor-joining algorithms of phylogeny reconstruction, Proc.3rd Ann. Int. Comput. Combinatorics Conf. 1997, pp. 101–110.
[2] H.-J. Bandelt, Recognition of tree metrics, SIAM J. Discrete Math. 3 (1990) 1–6.
[3] H.-J. Bandelt, A. Dress, Reconstructing the shape of a tree from observed dissimilarity data, Adv. Appl. Math. 7 (1986) 309-343.
[4] H.-J. Bandelt, A. Dress, A canonical decomposition theory for metrics on a finite set, Adv. Math. 92 (1992) 47-105.
[5] H.-J. Bandelt, A. Dress, A new and useful approach to phylogenetic analysis of distance data, Mol. Phyl. Evol. 1 (1992) 242-252.
[6] B. Bowditch, Notes on Gromov's hyperbolicity criterion for path-metric spaces, in: E. Ghys, A. Haefliger, A. Verjovsky (Eds.), Group Theory from a Geometrical Viewpoint, World Scientific Publishing Co. Pte. Ltd., Singapore, 1991, pp. 64–167.
[7] D. Bryant, and V. Moulton, A polynomial time algorithm for constructing the refined Buneman tree, Appl. Math. Lett. (1998), in press.
[8] P. Buneman, The recovery of trees from measures of dissimilarity, in: F. Hodson, D. Kendall, P. Tautu (Eds.), Mathematics in the Archaeological and Historical Sciences, Edinburgh University Press, Edinburgh, 1971, pp. 387–395.
[9] A. Dress, D. Huson, V. Moulton, Analyzing and visualizing sequence and distance data using SplitsTree, Discrete Appl. Math. 71 (1996) 95–110.
[10] A. Dress, V. Moulton, W. Terhalle, T-Theory – an overview, Europ. J. Combin. 17 (1996) 161–175.
[11] M. Farach, S. Kannan, Efficient algorithms for inverting evolution, Proc. 1996 ACM Symp. on the Foundations of Computer Science.
[12] M. Gromov, Hyperbolic Groups, in: S. Gerstin (Ed.), Essays in Group Theory, MSRI Publ. vol. 8, Springer, Berlin, 1987, pp. 75–263.
[13] D. Gusfield, Efficient algorithms for inferring evolutionary trees, Networks 21 (1991) 19–28.
[14] D. Hillis, C. Moritz, K. Barbara, Molecular Systematics, 2nd. ed., Sinauer Associates Inc., 1996.
[15] N. Jardine, R. Sibson, The construction of hierarchic and non-hierarchic classifications, Comput. J. 11 (1968) 177-184.
[16] P.J. Lockhart, M.A. Steel, M.D. Hendy, D. Penny, Recovering evolutionary trees under a more realistic model of sequence evolution, Mol. Biol. Evol. 11 (1994) 605-612.

[17] D. Robinson, L. Foulds, Comparison of weighted labelled trees, in: Lecture Notes in Mathematics, vol. 748, Springer, Berlin, 1979, pp. 119–129.
[18] Y. A. Smolensky, A method for linear recording of graphs, USSR Comput. Math. Phys. 2 (1969) 396–397.
[19] K. Wolf, P. Degens, On properties of additive tree algorithms, Conceptual and Numerical Analysis of Data, Proc. 13th Conf. of the Gesellschaft für Klasification, Springer, Berlin, 1989.
[20] K. A. Zaretsky, Reconstruction of a tree from the distances between its pendant vertices, Uspekhi Math. Nauk, Russian Mathematical Surveys, 20 (1965) 90–92 (in Russian).