**Noname manuscript No.**
(will be inserted by the editor)

# Expansion of Gene Clusters and the Shortest Hamiltonian Path Problem

**Sonja J. Prohaska · Christian Höner zu Siderdissen · Peter F. Stadler**

Received: date / Accepted: date

**Abstract** very abstract, this abstract

Insert your abstract here. Include keywords, PACS and mathematical subject classification numbers as needed.

**Keywords** First keyword · Second keyword · More

## 1 Introduction

## 2 Gene Duplications and Genomic Gene Order

We consider a family $X$ of $n = |X|$ paralogous genes whose evolutionary history is given by the tree $T$ (with vertex set $V$, leaf set $X \subset V$, and edge set $E$) and strictly positive branch lengths $\ell : E \to \mathbb{R}^+$. The corresponding

S.J. Prohaska
Computational EvoDevo Group, Department of Computer Science & Interdisciplinary Center for Bioinformatics, Universität Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany.
E-mail: sonja@bioinf.uni-leipzig.de

C. Höner zu Siderdissen
Bioinformatics Group, Department of Computer Science & Interdisciplinary Center for Bioinformatics, Universität Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany; Institute for Theoretical Chemistry, University of Vienna, Währingerstrasse 17, A-1090 Wien, Austria.
E-mail: choener@bioinf.uni-leipzig.de

P.F. Stadler
Bioinformatics Group, Department of Computer Science & Interdisciplinary Center for Bioinformatics, Universität Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany; Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany; RNomics Group, Fraunhofer Institute for Cell Therapy and Immunology, D-04103 Leipzig, Germany; Institute for Theoretical Chemistry, University of Vienna, Währingerstrasse 17, A-1090 Wien, Austria; Santa Fe Insitute, 1399 Hyde Park Rd., Santa Fe NM 87501, USA.
E-mail: studla@bioinf.uni-leipzig.de

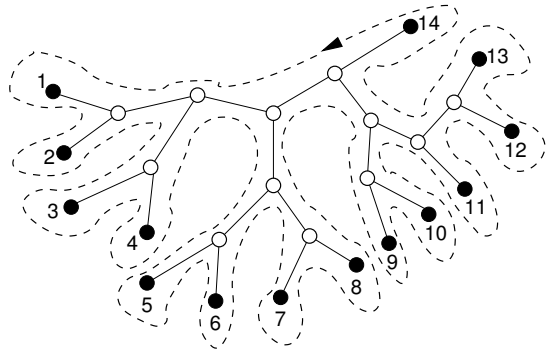


**Fig. 1** Each planar embedding $\breve{T}$ gives rise to a circular ordering of the vertices by following the "outline" around the tree.

genetic distance function $d : X \times X \to \mathbb{R}_0^+$ is given by

$$d_{xy} = \sum_{e \in \wp_{xy}} \ell(e) \tag{1}$$

where $\wp_{xp}$ denotes the unique path connecting $x$ and $y$ in $T$. We write $d_{\max} = \max_{x,y \in X} d_{xy}$ for the maximal distance between two leaves.

Let $\pi : \{1, \dots, n\} \to X$ be a bijection. In other words, $\pi$ defines an ordering of $X$ so that $x \prec y$ iff $\pi^{-1}(x) < \pi^{-1}(y)$. A special ordering $\widehat{\pi}$ is the arrangement of the gene on the genome.

A *circular* (or *cyclic*) *ordering* [17] is a ternary relation $\triangleleft i\, j\, k$ on a set $X$ that satisfied the following five conditions for all $i, j, k \in X$:

(cO1) $\triangleleft i\, j\, k$ implies $i, j, k$ are pairwise distinct. (irreflexive)
(cO2) $\triangleleft i\, j\, k$ implies $\triangleleft k\, i\, j$. (cyclic)
(cO3) $\triangleleft i\, j\, k$ implies $\neg\triangleleft k\, j\, i$ (antisymmetric)
(cO4) $\triangleleft i\, j\, k$ and $\triangleleft i\, k\, l$ implies $\triangleleft i\, j\, l$. (transitive)
(cO5) If $i, j, k$ are pairwise distinct then $\triangleleft i\, j\, k$ or $\triangleleft k\, j\, i$. (total)

A pair of points $(p, q)$ is adjacent in a total circular order on $V$ if there is no $h \in V$ such that $\triangleleft p\, h\, q$. Circular orderings can be linearized by cutting them at any point resulting in a linear order with the cut point as its minimal (or maximal) element [20]. We will write, by abuse of notation, $i \prec j \prec k$ to mean $\triangleleft i\, j\, k$ together with a suitable linearization, i.e., a cut between $k$ and $i$.

It is well known that trees as planar graphs. Let $\breve{T}$ be a fixed planar embedding of $T$. It defines, up to orientation, a unique circular ordering of the leaf set $X$. Any linearization of this circular order defines a linear order which we will refer to as a *$T$-order*, see Fig. 1.

Consider a tree $T = (V, E)$ with leaf-set $X \subset V$ and fix a particular circular order $\pi$ on $X$. Let $E_\pi$ be a set of edges connecting consecutive leaves w.r.t. to $\pi$ and denote by $G_T = (V, E \cup E_\pi)$ the auxiliary graph with the same vertices as $T$ and an edge set extended by $E_\pi$. Thus $G_T$ is a Halin graph [11] whenever $\pi$ is *$T$-order*. A necessary condition for $\pi$ to be a *$T$-order* therefore is that $G_T$ is planar graph.

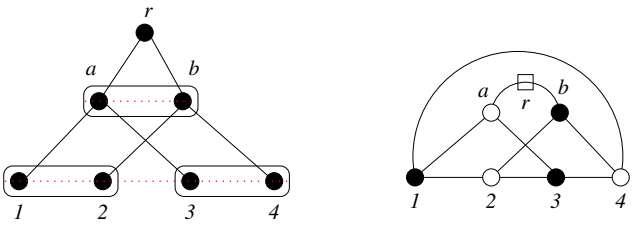


**Fig. 2** Phylogenetic tree arising from a block duplication of two paralogs. The l.h.s. sketches the phylogenetic tree and the genomic ordering of the leaves. The r.h.s. show the corresponding graph $G_T$. After contracting the edge between $a$ and $r$ we are left with a $K_{3,3}$, hence $G_T$ is not planar. Thus the genomic ordering $\widehat{\pi}$ is not a $T$-ordering.

Clearly, the gene family originated exclusively by tandem duplications, then the genomic order $\widehat{\pi}$ is a $T$-order for the gene phylogeny $T$. On the other hand, if a block containing two or more genes is duplicated as a unit, then $\widehat{\pi}$ and the tree are discordant as shown in Fig. 2. Every duplication scenario in which more than single gene duplicated at least once must contain this situation as a subgraph, and thus $K_{3,3}$ as a minor. It follows immediately that $\widehat{\pi}$ is not a $T$-order whenever the evolutionary scenario involves larger block duplications. We remark that gene loss may erase the this signature of block duplications. For instance, the loss of 2 or 3 in Fig. 2 leads back a $T$-order.

## 3 From Trees to Hamiltonian Paths

For an arbitrary order $\pi$ we define the length functional

$$L(\pi) = \sum_{i=2}^{n} d_{\pi(i-1)\pi(i)} \tag{2}$$

$L(\pi)$ can be interpreted as the length of the Hamiltonian path defined by the ordering $\pi$ in the complete graph with vertex set $X$ and edge lengths $d_{xy}$.

**Theorem 1** *Let $d$ be the additive tree metric associated with the tree $T$ and its non-degenerate length function $\ell$. Then $L(\pi)$ is minimal if and only if (ii) $\pi$ is a $T$-order and (ii) $d_{\pi(1)\pi(n)} = d_{\max}$.*

*Proof* We use the abbreviation $\mathcal{L} = \sum_{e\in E} \ell(e)$.

***Claim 1.*** Every order $\pi$ satisfies $L(\pi) \geq 2\mathcal{L} - d_{\max}$.

Denote by $\omega$ the closed walk $\wp_{\pi(1)\pi(2)}\wp_{\pi(2)\pi(3)}\cdots\wp_{\pi(n-1)\pi(n)}\wp_{\pi(n)\pi(1)}$. Its length is $L(\omega) = d_{\pi(n)\pi(1)} + \sum_{i=2}^{n} d_{\pi(i-1)\pi(n)}$. Since $\omega$ connects any two leaves, it contains all edges of $T$. Furthermore, since $T$ contains no cycle, $\omega$ must leave each subtree that it enters along the same edge. Thus $\omega$ covers edge at least twice. Hence $L(\omega) \geq \mathcal{L}$. Since $\omega$ contains exactly one path too many, and the longest possible path had length $d_{\max}$, the claim follows. ◁

***Claim 2.*** If $\pi$ is $T$-order, then $L(\pi) = 2\mathcal{L} - d_{\pi(1)\pi(n)}$.

By construction $\omega$ associated with a $T$-order is the closed walk defined by the "outline" of the tree, *cf.* Fig. 1. Any such walk covers each edge of $T$ exactly twice, once when entering and once when leaving a given subtree. This construction is well known in the literature, see e.g. [18, Thm.5]. The claim follows directly from $L(\pi) = L(\omega) - d_{\pi(1)\pi(n)}$. ◁

Fix an arbitrary leaf 1 as the root of $T$ and a starting and end point of $\omega$ and denote by $n$ the last leaf visited for the first time along $\omega$. Furthermore, for every edge $e$ denote by $T(e)$ the connected component of $T \setminus \{e\}$ that does not contain 0.

***Claim 3.*** If $\omega$ covers every edge of $T$ exactly twice then the leafs contained within every subtree form an interval in $\pi$.

It suffices to note $\omega$ can enter and leaves the subtree $T(e)$ only through $e$. If the edge is covered exactly twice, all leaves of $T(e)$, and only the leaves of $T(e)$ are visited along $\omega$ between the first and the second traversal of $e$. ◁

It follows that, for each edge $e = \{u, v\}$, where $v \in V(T(e))$ and $u \notin V(T(e))$ there is a linear ordering of the children $v_1$, $v_2$, through $v_{d(v)}$ of $v$ so that the subtrees $T(v_1)$, $T(v_2)$, ..., $T(v_{d(v)})$ are traversed by $\omega$ in this order. Consequently, there is a planar layout of $T$ so that the leaves 1 through $n$ are arranged in the order of traversal. In other words, if $\omega$ traverses $T$ so that every edge is covered exactly twice, then $T$ has a planar embedding so that $\omega$ travels along its outline and visits consecutive leaves in the order in which they appear on the outline of the tree.

Hence there is a $T$-ordering following the outline of $T$ if and only if the corresponding closed walk covers every edge of $T$ exactly twice. By closure of the walk, each edge must be covered an even number of times by $\omega$, so that $\omega$ without the return path from $\pi(n)$ to $\pi(1)$ covers at least one edge thrice, thus $L(\pi) > 2\mathcal{L} - d_{\pi(1)\pi(n)}$ for every $\pi$ that is not a $T$-ordering. □

## 4 Simulating Distance Matrices for Gene Duplications

We show here that genetic distance matrices for models of gene duplications can be simulated directed. This has advantages over the more usual approach of simulating sequence evolution. In particular we can in this manner separate the stochastic noise that may lead to deviations from additive tree metrics.

**Lemma 1** *Let $d : X \times X \to \mathbb{R}$ be an additive tree metric on $X$ and let $\delta_x \geq 0$ for $x \in X$ be arbitrary. Then $d' : X \times X \to \mathbb{R}$ defined as $d'_{xy} = d_{xy} + \delta_x + \delta_y$ for $x \neq y$ is again an additive tree metric.*

*Proof* A metric $d$ is an additive tree metric if and only if every 4-tuple satisfied the "4-point condition" [5,9,10,24], which stipulates that any four leaves can be renamed such that

$$d_{xy} + d_{uv} \leq d_{xu} + d_{yv} = d_{xv} + d_{yu}$$

**Algorithm 1** Simulation of an Additive Tree Metric

**Require:** $n$ {final dimension}
  $V \leftarrow \{1\}$
  **while** $|V| < n$ **do**
    randomly pick $x \in V$, $z \notin V$
    $V \leftarrow V \cup \{z\}$
    $d_{zu} \leftarrow d_{xu}$ for all $u \in V \setminus \{u\}$
    $d_{zx} \leftarrow 0$
    randomly choose $\delta_u \geq$ for all $u \in V$
    **for** $p, q \in V$, $p \neq q$ **do**
      $d_{pq} \leftarrow d_{pq} + \delta_p + \delta_q$
    **end for**
  **end while**

Using the definition of $d'$ immediately yields

$$\begin{aligned} d'_{xy} + d'_{uv} &= d_{xy} + d_{uv} + \delta_x + \delta_y + \delta_u + \delta_v \\ &\leq d_{xu} + d)yv + \delta_x + \delta_y + \delta_u + \delta_v \\ &= d_{xv} + d_{yu} + \delta_x + \delta_y + \delta_u + \delta_v \\ &\leq d'_{xu} + d'_{yv} = d'_{xv} + d'_{yu} \end{aligned}$$

Hence we can propagate time by an increment $\Delta t$ simply by adding setting $\delta_x = r_x \Delta t$ where $r_x$ is the rate of evolution of of taxon $x$. A duplication of gene $x$ can be introduced by simply duplicating the row and column $x$ in the distance matrix $d$, i.e., by setting $d_{zy} = d_{xy}$ for all $y \neq x, z$ and $d_{xz} = 0$. The proceedure is summarized in Algorithm 1.

A rate $r_{x'}$ (and possibly a new rate $r_x$) needs to be chosen. Assuming a constant rate of duplication, we set $\Delta t = 1/n$ und choose one of the leafs at random for duplication. Instead of appending the new leaf $x'$ to the end of the matrix, we insert it explicitly before or after $x$ so that the order $\pi$ of the rows and columns explicitly encodes the genomic order. Duplicating a larger block of rows and columns can immediately be used to simulate the block duplications of any number of adjacent genes.

**Lemma 2** *Every additive tree metric $d'$ can be constructed by Algorithm 1.*

*Proof* If $d'$ is an additive tree metric, then there is a unique additive tree $T$ with edge lengths $\ell : E \to \mathbb{R}_0^+$ representing $d'$. Suppose for the moment that $T$ is binary. Then it has at least one "cherry", i.e., a pair of leaves separated by only a single interior vertex, say $\{p, q\}$. It is easy to check that every cherry in $T$ must satisfy

$$\min_{x,y \in V \setminus \{p,q\}} \left\{ (d'_{px} + d'_{qy}) - (d'_{pq} + d'_{xy}) \right\} > 0 \tag{3}$$

If $\{p, q\}$ is a cherry, then the distances in $T$ from $p$ and $q$ to their last their common ancestor are $\delta_p = (1/2) \min_{u,v \neq p}(d'_{pu} + d'_{pv} - d'_{uv}) \geq 0$ and $\delta_q = (1/2) \min_{u,v \neq q}(d'_{qu} + d'_{qv} - d'_{uv}) \geq 0$, both of which are non-negative as a consequence of the triangle inequality. The reduced distance matrix $d$ on $V \setminus \{q\}$

defined by $d_{xy} = d_{xy}$ for $x, y \notin \{p, q\}$, $d_{xp} = d'_{xp} - \delta_p$ represents the $T$ with the cherry replaced by its last common ancestor, hence it is again an additive distance matrix.

Repeating this construction we arrive a single vertex after $|V| - 1$ steps. Each step identifies a leaf $p$ that is duplicated and the extensions $\delta_p$ and $\delta_q$ of $p$ and its copy $q$. Note that we have set $\delta_x = 0$ for all $x \in V \setminus \{p, q\}$. This reflects that the stepwise elongation of branches of tree modelled in Alg. 1 can subdivided arbitrarily betwenn duplication events that affect a particular branch. Here we simply choose the add the entire length immediately after each duplication event. Thus the construction in this proof backtraces a particular sequence of duplication events in Alg. 1.

The case of non-binary trees is easily incorporated by observing that it can be represented as binary tree in which an internal branch length of 0 is also allowed.

## 5 Type R Distance Matrices

The model so far corresponds to a mechanism in which non-homologous crossover occurs between the genes of interest. We can, however, also model events in which the genes themselves are recombined. Instead of assuming that $x'$ is a true copy of $x$ we now assume that the newly introduced gene $z$ is a recombinant of two adjacent genes $x$ and $y$. The product is inserted between $x$ and $y$.

Since $z$ is composed of two parts, of relatives sizes $a$ and $(1-a)$, $0 \leq a \leq 1$, that are identical to $x$ and $y$, respectively, we have

$$\begin{aligned} d_{zu} &= a d_{xu} + (1-a) d_{yu} \\ d_{zx} &= (1-a) d_{xy} \\ d_{zy} &= a d_{xy} \end{aligned} \tag{4}$$

After the duplication event, each genes evolves independently with its own rate, so that the genetic distance between $p$ and $q$ again grows by $\delta_p + \delta_q$, i.e.,

$$d'_{pq} = d_{pq} + \delta_p + \delta_q \tag{5}$$

**Definition 1** A distance matrix $d$ is of *type R* if it is constructed by repeated application of eqns.(4) and (5)

Clearly, every additive tree metric is of type R by virtue of setting $a = 0$ (or $a = 1$) in every duplication step. In particular, therefore, for $n = 3$ every distance matrix is of type R. For $n > 3$, however, it is not obvious whether at whether a type R matrix can be recognized efficiently.

We start by observing

$$d'_{xz} + d'_{yz} - d'_{xy} = d_{xz} + d_{yz} - d_{xy} + \delta_x + \delta_z + \delta_y + \delta_z - \delta_x - \delta_y = 2\delta_z \tag{6}$$

since $d_{xz} + d_{yz} = (1-a)d_{xy} + a d_{xy} = d_{xy}$.

For $n \geq 4$, consider the following expression for $u \notin \{x, y, z\}$.

$$\begin{aligned} d'_{uz} - ad'_{ux} - (1-a)d'_{uy} &= \underbrace{d_{uz} - ad_{ux} - (1-a)d_{uy}}_{=0} \\ &\quad + \delta_u + \delta_z - a\delta_u - a\delta_x - \delta_u + a\delta_u - \delta_y + a\delta_y \\ &= \delta_z - a\delta_x - (1-a)\delta_y := f(a) \end{aligned} \tag{7}$$

The key observation is that this expression is independent of $u$. Thus, is $n \geq 5$ there is are distinct leaves $u$, $v$ distinct from $\{x, y, z\}$ so that $d'_{uz} - ad'_{ux} - (1-a)d'_{uy} = f(a) = d'_{vz} - ad'_{vx} - (1-a)d'_{vy}$, which can be rearranged as $d'_{uz} - d'_{uy} - ad'_{ux} + ad'_{uy} = d'_{vz} - d'_{vy} - ad'_{vx} + ad'_{vy}$ and hence, after a short calculation,

$$a = \frac{(d'_{uz} + d'_{vy}) - (d'_{vz} + d'_{uy})}{(d'_{ux} + d'_{vy}) - (d'_{vx} + d'_{uy})} \tag{8}$$

Note that this equation must be satisfied for all $u, v \notin \{x, y, z\}$, hence it restricts the space of type R distance matrices to a submanifold for all $n > 5$.

Once $a$ has been computed, $f(a)$ can also be computed explicitly. Now consider the following system of equations

$$\begin{aligned} -a\delta_x - (1-a)\delta_y &= f(a) - \delta_z \\ (1-a)d_{xy} + \delta_x &= d'_{xz} - \delta_z \\ ad_{xy} + \delta_y &= d'_{yz} - \delta_z \end{aligned} \tag{9}$$

The first line uses the definition of $f(a)$ above, the second and third line are rearrangements of $d'_{xz} = (1-a)d_{xy} + \delta_x + \delta_z$ and $d'_{yz} = ad_{xy} + \delta_y + \delta_z$, resp. Multiplying the 2nd and 3rd line by $a$ and $(1-a)$, resp., and adding up the three equations yields $2a(1-a)d_{xy} = f(a) - 2\delta_z + ad'_{xz} + (1-a)d'_{yz}$. We can now compute $d_{xy}$ from

$$2a(1-a)d_{xy} = (d'_{uz} - ad'_{ux} - (1-a)d'_{uy}) - 2\delta_z + ad'_{xz} + (1-a)d'_{yz} \tag{10}$$

Finally, $\delta_x$ and $\delta_y$ are obtained from

$$\begin{aligned} \delta_x &= d'_{xz} - (1-a)d_{xy} - \delta_z \\ \delta_y &= d'_{yz} - ad_{xy} - \delta_z \end{aligned} \tag{11}$$

In summary, therefore, we can obtain, for $n \geq 5$, complete information on the relative arrangement of the parents $x$ and $y$ and their recombinant offspring $z$.

It remains to determine the values of $\delta_u$ for $u \notin \{x, y, z\}$. This turns out to be not so trivial, since $\delta_u$ is, in contrast to $\delta_x$, $\delta_y$, and $\delta_z$ not uniquely determined by the last unequal crossover event.

To see this more clearly, let as first consider the case $n = 4$. It is well known that every metric on 4 points can be respresented a "box graph" as shown in Fig. 5. The box dimensions can computed from $2u = (d_{ps} + d_{qr}) - (d_{pq} + d_{rq})$ and $2(u - v) = (d_{rp} + d_{qs}) - (d_{pq} + d_r s)$. The key ingedients, thus are the three different pairs of distances emphasized by parentheses. For more details

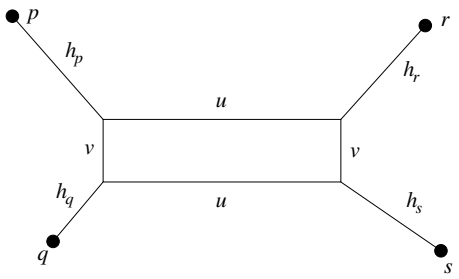


**Fig. 3** Representation of a metric $d$ on 4 points $\{p, q, r, s\}$. Each distance is the sum length of a shortest path in this graph. For instance $d_{pq} = h_p+v+h_q$, $d_{pr} = h_p+u+h_r$, $d_{ps} = h_p+u+v+h_s$.

see [19]. Now let us start from an arbitrary distance matrix $d$ on $\{x, y, u\}$ and construct $z$ as a recombinant. Let us abbreviated the three pairs of distance sums as

$$A = d'_{xz} + d'_{uy} \qquad B = d'_{yz} + d'_{ux} \qquad C = d'_{uz} + d'_{xy}\,. \tag{12}$$

Using the definitions of $d_{xz}$, $d_{yz}$, and $d_{uz}$ we can compute

$$\begin{aligned} C - A &= a(d_{xy} + d_{xu} - d_{uy}) \geq 0 \\ C - B &= (1-a)(d_{xy} + d_{yu} - d_{ux}) \geq 0 \end{aligned} \tag{13}$$

using again the triangle inequality. The terms $C - A$ and $C - B$ correspond to twice the sides of the box in the quadruple graph ; note that they are independent of $\delta_x$, $\delta_y$, $\delta_z$, and $\delta_u$. We obtain a tree whenever the box degenerates to a line, i.e., if $a = 0$ or $a = 1$.

In the general case this becomes $h_u = \delta_u + (1/2)\min_{v,w}(d_{uv} + d_{uw} - d_{vw}) \geq 0$, where the minimum runs over all $v \neq w$ different from 0. It follows that $\delta_u \geq 0$ cannot be determined. Intuitively, this comes from the fact that a contribution $\delta_u + \delta_v$ is added to $d_{uv}$ after every duplication/recombination event. These contribution cannot be divided unambiguously between the individual steps in complete analogy to the situation for additive tree metrics in the previous section.

Hence we can set $\delta_u = 0$ for every $u \notin \{x, y, z\}$ and assume the entire length of $h_u$ stems from previous events. This yields the recursive Algorithm 2 for recognizing type R distance matrices. It requires $O(|V|)$ decomposition steps, each of which needs in the worst case $O(|V|^5)$ computations to identify the triple $\{x, y, z\}$ corresponding to the last recombination event. For this candidate, $\mathbf{D'}$ is computed in quadratic time. Thus Algorithm 2 runs in $O(|V|^6)$ time.

## 6 Linear Type R matrices

**Definition 2** A type-R distance matrix is called *linear* with order $\pi$ if, starting from $V = \{x, y\}$, in each vertex addition step the two parents $x$ and $y$ are adjacent and their offspring $z$ is placed between $x$ and $y$.

**Algorithm 2** Recognition of type R distance matrices

**Require:** Distance matrix $\mathbf{D}'$, $n = |V|$
  **repeat**
    **for** $\{x, y, z\} \subseteq V$ **do**
      **for** $\{u, v\} \subseteq V \setminus \{x, y, z\}$ **do**
        compute $a$ using equ.(8)
      **end for**
      **if** $a \in [0, 1]$ is the same for all $u, v$ **then**
        compute $\delta_z$ using equ.(6)
        compute $d_{xy}$ using equ.(10)
        compute $\delta_x$, $\delta_y$ using equ.(11)
        $\delta_u \leftarrow 0$ for $u \in V \setminus \{x, y, z\}$
        compute $\mathbf{D}$ as $d_{pq} = d'_{pq} - \delta_p - \delta_q$ for all $p, q \in V$
        $\mathbf{D}' \leftarrow \mathbf{D}$ without row and column $z$
        $n \leftarrow n - 1$
      **end if**
    **end for**
    **if** no $\{x, y, z\}$ was found **then**
      return **false**
    **end if**
  **until** $n = 4$
  return **true**

**Definition 3** A distance matrix $\mathbf{D} = (d_{ij})$ satisfies the *Kalmanson condition* if there is a circular order $\triangleleft$ of the points so that the inequality

$$\max\{(d_{ij} + d_{kl}), (d_{il} + d_{jk})\} \leq d_{ik} + d_{jl} \tag{14}$$

for every four points so that $i \prec j \prec k \prec l$.

If $(d_{ij})$ satisfies equ.(14) then the corresponding TSP is solved by the unit permutations, i.e., $\pi = (1, 2, 3, \ldots, n)$ [12]. Equivalently, if $\triangleleft$ is a circular ordering of the taxa set $V$ and $\pi$ the permutation of $V$ associated with an arbitrary linearization of $\triangleleft$, then $(d_{ij})$ is Kalmanson iff

$$\max\{(d_{\pi(i)\pi(j)} + d_{\pi(k)\pi(l)}), (d_{\pi(i)\pi(l)} + d_{\pi(j)\pi(k)})\} \leq d_{\pi(i)\pi(k)} + d_{\pi(j)\pi(l)} \tag{15}$$

for $i < j < k < l$. In this case $L(\pi)$ in equ.(2) is a shortest Hamiltonian cycle for $(d_{ij})$.

With each circular ordering $\triangleleft$ we can associate a set $\mathcal{S}^{\triangleleft}$ of splits, i.e., nontrival bipartitions of the set $X$ of taxa. $\{A, X \setminus A\} \in \mathcal{S}^{\triangleleft}$ if and only if (i) $A \neq \emptyset$, (ii) $A \neq X$, (iii) there is $i, j \in A$ and $k, l \in X \setminus A$ so that (a) for all $p \in A$ and $q \in X \setminus A$ holds $\triangleleft\, i\, p\, j$ and $\triangleleft\, k\, q\, l$ and (b) $\triangleleft\, i\, j\, k$ and $\triangleleft\, k\, l\, i$. We write

$$S_{ij} := \{\{\pi(i+1), \pi(i+2), \ldots, \pi(j)\}, \{\pi(j+1), \pi(j+2), \ldots, \pi(i)\}\} \tag{16}$$

with $i, j$ taken $\mod |X|$ for the splits of $\mathcal{S}^{\triangleleft}$, where $\pi$ is again an arbitrary linearization of $\triangleleft$. A metric is called *circular decomposable* [1] if there is a circular ordering $\triangleleft$ (with a corresponding permutation $\pi$), and $\alpha_{ij} \geq 0$, $i \neq j$ so that

$$d_{xy} = \sum_{i<j} \alpha_{ij} \delta_{S_{ij}}(x, y)\,, \tag{17}$$

where the split pseudometric $\delta_{S_{ij}}$ is defined as $\delta_{S_{ij}}(x,y) = 1$ is the split $S_{ij}$ separates $x$ and $y$, and $\delta_{S_{ij}}(x,y) = 0$ otherwise. The *isolation indices* of the splits $S_{ij}$ can be computed as

$$\alpha_{ij} = \frac{1}{2}\left(d_{\pi(i)\pi(j)} + d_{\pi(i+1)\pi(j+1)} - d_{\pi(i)\pi(j+1)} - d_{\pi(i+1)\pi(j)}\right) \tag{18}$$

It is shown in [7,6] that a metric satisfies the Kalmanson condition if and only it is circular decomposable. These can represented as so-called split graphs and computed efficiently using the `NeighborNet` algorithm [3,4].

As shown in [6], circular decomposable metrics satisfy a "Crofton formula" of the form

$$d_{xy} = \sum \{\alpha(S_{ij} | S_{ij} \text{ separates } \pi(x) \text{ and } \pi(y)\} \tag{19}$$

As shown in [15, Thm.37], the solution of the TSP on a generic generic circular decomposable metric is unique. Thus, one can use the TSP solutions of $(d_{xy})$ directly for finding circular orderings to be used in `Neighbornet` [14].

**Theorem 2** *Every linear type R distance matrix satisfied the Kalmansson condition.*

*Proof* We only need to show that the distance matrix on $X \cup \{z\}$ is Kalmanson provided the distance matrix on $X$ is Kalmanson. Suppose $z$ is the recombinant of $j$ and $j'$. In the general case we have $i \prec j \prec z \prec j' \prec k \prec l$, since by circularity of the ordering it does not matter whether we duplicate $i$, $j$, $k$, or $l$. In addition to the general case we have to consider the special cases with $i = j$ and/or $j' = k$. The proof repeatedly makes use of the simple observation that $\max(a+p, b+q) \leq \max(a,b) + \max(p,q)$.

We assume that the Kalmanson inequalities hold for all quadruples in $X$ with an appropriate circular order. For the general case we have, by substituting the definition of the distances involving the recombinant vertex $z$,

$$\begin{aligned}
&\max\{d_{iz} + d_{kl}, d_{il} + d_{zk}\} \\
=&\max\{a(d_{ij} + d_{kl}) + (1-a)(d_{ij'} + d_{kl}), a(d_{il} + d_{jk}) + (1-a)(d_{il} + d_{j'j})\} \\
\leq& a \max\{d_{ij} + d_{kl}, d_{il} + d_{jk}\} + (1-a)\max\{d_{ij'} + d_{kl}, d_{il} + d_{j'k}\} \\
\leq& a(d_{ik} + d_{jl}) + (1-a)(d_{ik} + d_{j'l}) = d_{ik} + a d_{jl} + (1-a) d_{j'l} \\
=& d_{ik} + d_{zl}
\end{aligned}$$

In the fourth line we use that the Kalmanson inequality holds for $i \prec j \prec k \prec l$ and $i \prec j' \prec k \prec l$ by assumption, the last line used the definition of $d_{zl}$. Analogous computations for the three special cases (ommitting the analog of

the 2nd and 3rd line above yield

$$\begin{aligned}
&\max\{d_{jz}+d_{kl}, d_{jl}+d_{zk}\}\\
\le& a\max\{d_{kl}, d_{jl}+d_{jk}\} + (1-a)\max\{d_{jj'}+d_{kl}, d_{jl}+d_{j'k}\\
\le& a(d_{jl}+d_{jk}) + (1-a)(d_{jk}+d_{j'l} = d_{jk}+d_{zl}\\
&\max\{d_{iz}+d_{j'l}, d_{il}+d_{zj'}\}\\
\le& a\max\{d_{ij}+d_{j'l}, d_{il}+d_{jj'}\} + (1-a)\max\{d_{ij'}+d_{j'l}, d_{il}\}\\
\le& a(d_{ij'}+d_{jl}) + (1-a)(d_{ij'}+d_{j'l}) = d_{ij'}+d_{zl}\\
&\max\{d_{ij}+d_{zj'}, d_{ij'}+d_{jz}\}\\
\le& a\max\{d_{ij}+d_{jj'}, d_{ij'}\} + (1-a)\max\{d_{ij}, d_{ij'}+d_{jj'}\}\\
=& a(d_{ij}+d_{jj'}) + (1-a)(d_{ij'}+d_{jj'}) = d_{jj'}+d_{iz}
\end{aligned}$$

We conclude that all quadruples involving $z$ satisfy the Kalmanson inequality provided $(d_{ij})$ is Kalmanson matrix on $V$: we have used the Kalmanson conditions for $i \prec j \prec k \prec l$ as well as the triangle inequality in our proof. As the distances not involving the new offspring $z$ remain unchanged by the construction principle of type R matrices, we conclude that $(d_{ij})$ on $X \cup \{z\}$ satisfies the Kalmanson inequalities.

The basic idea of converting a TSP into a shortest Hamiltonian path problem is folklore. One simply adds a dummy node 0 between 1 and $n$ with $d_{0\pi(i)} = c$ large enough. Then a shortest Hamiltonian path will use 0 as an endpoint to avoid using $2c$ in the solution. The resulting expanded distance matrix $(d_{ij})$ on $V \cup \{0\}$ is cirular decomposable if, and only if, the Kalmanson conditions also hold for quadruples involving the dummy node, i.e., if and only if

$$\max\{d_{0i}+d_{jk}, d_{0k}+d_{ij}\} \le d_{0j}+d_{ik}\} \tag{20}$$

holds for all $0 \prec i \prec j \prec k$. Since $d_{0i} = c$ this simplified to the condition

$$\max\{d_{ij}, d_{jk}\} \le d_{ik} \qquad \text{for all } i < j < k\,. \tag{21}$$

A dissimilarity $d$ is called Robinsonian if there is a permutation $\pi$ so that

$$\max\{d_{\pi(i)\pi(j)}, d_{\pi(j)\pi(k)}\} \le d_{\pi(i)\pi(k)} \qquad \text{for all } i < j < k. \tag{22}$$

The so-called serialization problem [23,16] of linearly ordering objects is solved by the order $\pi$ for Robinsonian dissimilarities. This results appears to be folklore, we have not found a simple direct proof.

**Lemma 3** *If $d$ is Robinsonian, then $\pi$ is a shortest Hamiltonian path.*

*Proof* W.l.o.g. we assume $\pi = \iota = (1, 2, \ldots, n)$. Consider an arbitrary permutation $\xi$. Then there is a bijection $\varphi$ between the adjacenies $[\xi(i)\xi(i+1)]$ w.r.t. to $xi$ and the adjacencies $[p, p+1]$ w.r.t. $\iota$ so that $\xi(i) \le p < p+1 \le \xi(i+1)]$. To see this we argue by induction. For $n = 2$ the statement is trivial. In general $\xi$ is either (1) the extension of a permutation $\xi'$ on $\{1, 2, \ldots, n-1\}$ by one of the adjacencies $[1, n]$ or $[n-1, n]$, or (2) $\xi$ is obtained by inserting $n$ into

the adjaceny $[\xi'(k)\xi'(k+1)] = [u, v]$ with $u = \min(\xi'(k), \xi'(k+1))$ and $v = \max(\xi'(k), \xi'(k+1))$ In case (1) $\varphi$ is the extensions of $\varphi'$ by $[1, n] \mapsto [n-1, n]$ or $[n-1, n] \mapsto [n-1, n]$. In case (2) we obtain $\varphi$ from $\varphi'$ by replacing $[u, v] \mapsto [p, p+1]$ with $[u, n] \mapsto [p, p+1]$ and adding $[v, n] \mapsto [n-1, n]$. The Robinson condition (21) implies $d_{\xi(i),\xi(i+1)} \geq d_{p,p+1}$ for $\varphi([\xi(i)\xi(i+1)]) = [p, p+1]$ and hence $L(\xi) \geq L(\iota)$, i.e., $\iota$ is a shortest Hamiltonian path.

The Robinson property also plays an important role in cluster analysis, where it characterizes certain generalizations of hierarchies [21,13,22]. So-called quadripolar Robinson dissimilarities, that also satisfy the Kalmansson condition are studied in some detail in [8].

**Lemma 4** *Suppose $(d_{ij})$ satisfies equ.(21) on $V$. Then the distance matrix on $V \cup \{z\}$ obtained by inserting the recombinant node $z$ between adjacent parents $j'$ and $j''$ also satisfies the Robinson condition equ.(21).*

*Proof* Suppose $j = z$ is the new node derived from parents $j' \prec z \prec j''$. Then for $i < j'$ and $k > j''$ we have $d_{iz} = ad_{ij'} + (1-a)d_{ij''}$ and $d_{zk} = ad_{kj'} + (1-a)d_{j''k}$. Thus $\max\{d_{iz}, d_{z,j}\} \leq a \max\{d_{ij'} + d_{j'k}\} + (1-a)(d_{ij''} + d_{j''k}) \leq d_{ik}$. The special case $i = j'$, $k > j''$ yields: $d_{j'z} = (1-a)d_{j'j''}$ and thus $\max\{(1-a)d_{j'j''}, ad_{j'k} + (1-a)d_{j'',k}\} \leq ad_{j'k} + (1-a)\max\{d_{j'j''}, d_{j'',k}\} \leq ad_{j'k} + (1-a)d_{j'k} = d_{j'k}$. An analogous computation works for $i < j'$ and $j'' = k$. Finally, for $i = j'$ and $k = j''$ we have, by construction $d_{j'z} = (1-a)d_{j'j''} \leq d_{j'j''}$ and $d_{zj''} = ad_{j'j''} \leq d_{j'j''}$.

However, the choice of the $\delta_k$ can destroy the inequality: From $\max\{d_{ij}, d_{jk}\} \leq d_{ik}$ we cannot conclude that $\{d_{ij} + \delta_i + \delta_j, d_{ij} + \delta_j + \delta_k\} \leq d_{ik} + \delta_i + \delta_k\}$. However, we can assume that the condition still works in very good approximation if the evolution rates of the offsprings are approximately the same. Hence very uneven evolution rates or a mechanism that makes the "middle" genes in a gene cluster evolve much faster can destroy the betweenness conditions. On the other hand, gene conversion makes it easier to satisfy equ.(21).

Dynamic programming for TSP [2].

## 7 Conclusions

**Acknowledgements**

## References

1. Bandelt, H.J., Dress, A.W.M.: A canonical decomposition theory for metrics on a finite set. Adv. math. **92**, 47 (1992)
2. Bellman, R.: Dynamic programming treatment of the travelling salesman problem. J. ACM **9**, 61–63 (1962)
3. Bryant, D., Moulton, V., Spillner, A.: `NeighborNet`: An agglomerative method for the construction of planar phylogenetic networks. Mol. Biol. Evol. **21**, 255–265 (2004)

4. Bryant, D., Moulton, V., Spillner, A.: Consistency of the neighbornet algorithm for constructing phylogenetic networks. Alg. Mol. Biol. **2** (2007)
5. Buneman, P.: A note on the metric property of trees. J. Combin. Theory Ser. B **17**, 48–50 (1974)
6. Chepoi, V., Fichet, B.: A note on circular decomposable metrics. Geometriae Dedicata **69**, 237–240 (1998)
7. Christopher, G., Farach, M., Trick, M.: The structure of circular decomposable metrics. In: J. Diaz, M. Serna (eds.) Algorithms ESA'96, Lect. Notes Comp. Sci., pp. 406–418. Springer, New York (1996)
8. Critchley, F.: On quadripolar Robinson dissimilarity matrices. In: E. Diday, Y. Lechevallier, M. Schader, P. Bertrand, B. Burtschy (eds.) New Approaches in Classification and Data Analysis, pp. 93–101. Springer, Heidelberg (1994)
9. Cunningham, P.: Free trees and bidirectional trees as representations of psychological distance. J. Math. Psych. **17**, 165–188 (1978)
10. Dobson, A.J.: Unrooted trees for numerical taxonomy. J. Appl. Probab. **11**, 32–42 (1974)
11. Halin, R.: Studies on minimally $n$-connected graphs. In: Combinatorial Mathematics and its Applications, pp. 129–136. Academic Press, London, UK (1971)
12. Kalmanson, K.: Edgeconvex circuits and the traveling salesman problem. Canadian J. Math. **27**, 1000–1010 (1975)
13. Kleinman, A., Harel, M., Pachter, L.: Affine and projective tree metric theorems. Ann. Comb. **17**, 205–228 (2013)
14. Korostensky, C., Gonnet, G.: Using traveling salesman problem algorithms for evolutionary tree construction. Bioinformatics **16**, 619–627 (2000)
15. Levy, D., Pachter, L.: The neighbor-net algorithm. Adv. Appl. Math. **47**, 240–258 (2011)
16. Liiv, I.: Seriation and matrix reordering methods: An historical overview. Statistical Analysis & Data Mining **3**, 70–91 (2010)
17. Meggido, N.: Partial and complete cyclic orders. Bull. Am. Math. Soc. **82**, 274–276 (1976)
18. Moret, B.M.E., Tang, J., Wang, L.S., Warnow, T.: Steps toward accurate reconstructions of phylogenies from gene-order data. J. Comp. Syst. Sci. **65**, 508–525 (2002)
19. Nieselt-Struwe, K.: Graphs in sequence spaces: a review of statistical geometry. Biophysical Chemistry **66**, 111–131 (1997)
20. Novák, V.: Cuts in cyclically ordered sets. Czech. Math. J. **34**, 322–333 (1984)
21. Orders, overlapping clusters in pyramids: Diday, e. In: J. De Leeuw, W.J. Heiser, J.J. Meulman, F. Critchley (eds.) Multidimensional Data Analysis, pp. 201–234. DSWO Press, Leiden, NL (1986)
22. Préa, P., Fortin, D.: An optimal algorithm to recognize Robinsonian dissimilarities. J. Classification **31**, 1–35 (2014)
23. Robinson, W.S.: A method for chronologically ordering archaeological deposits. Amer. Antiquity **16**, 293301 (1951)
24. Simões-Pereira, J.M.S.: A note on the tree realizability of a distance matrix. J. Combin. Theory **6**, 303–310 (1969)