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

# Expansion of Gene Clusters, Circular Orders, and the Shortest Hamiltonian Path Problem

**Sonja J. Prohaska · Sarah J. Berkemer · Fabian Gärtner · Thomas Gatter · Nancy Retzlaff · The Students of the Graphs and Biological Networks Lab 2017 · Christian Höner zu Siederdissen · Peter F. Stadler**

Received: date / Accepted: date

**Abstract** Clusters of paralogous genes such as the famous HOX cluster of developmental transcription factors tend to evolve by stepwise duplication of its members, often involving unequal crossover. Gene conversion and possibly other mechanisms of concerted evolution further obfuscate the phylogenetic relationships. As a consequence, it is very difficult or even impossible to disentangle the detailed history of gene duplications in gene clusters. In this contribution we show that the expansion of gene clusters by unequal crossover as proposed by Walter Gehring leads to distinctive patterns of genetic distances, namely a subclass of circular split systems. Furthermore, when the gene cluster was left undisturbed by genome rearrangements, the shortest Hamiltonian

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

Sarah J. Berkemer and Nancy Retzlaff
Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany

Fabian Externbrink
Competence Center for Scalable Data Services and Solutions Dresden/Leipzig and Bioinformatics Group, Department of Computer Science, Universität Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany

C. Höner zu Siederdissen and Thomas Gatter
Bioinformatics Group, Department of Computer Science & Interdisciplinary Center for Bioinformatics, Universität Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany;

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

paths w.r.t. genetic distances coincide with the genomic order. This observation can be used to detect ancient genomic rearrangements of gene clusters and to distinguish gene clusters whose evolution was dominated by unequal crossover within genes from those that expanded through other mechanisms.

**Keywords** Evolution of gene clusters · non-homologous recombination · unequal crossover · phylogenetic combinatorics · Kalmanson metrics · Hamiltonian path problems

## 1 Introduction

The genomes of higher eukaryotes typically contain many families of genes with similar DNA sequence. These usually encode similar proteins and share similar function. Their sequence similarity indicates that they have evolved from a single original ancestor by means of multiple rounds of duplication. Such paralogous genes are often, but by no means always, located at the same genomic locus, where they form a gene cluster. In many cases clustered genes are not tied together functionally and the clusters can disintegrate by genome rearrangement without detrimental effects.

However, some gene clusters are evolutionarily old and have retained a very particular organization of their member genes for hundreds of millions of years. Among the best characterized gene clusters are the globin gene clusters, which encode major players in the transport of oxygen within the bloodstream [30] and the homeobox gene clusters, which play a crucial role in the early stages of animal development [18]. In vertebrates, the latter show very low levels of repeats and unrelated open reading frames, and the genes in paralogous clusters share the same order and orientation. Experimental work demonstrated that the consolidated arrangement is crucial and constrained due the necessity of a coordinated regulation orchestrated by enhancer sequences outside the cluster [21,32].

The details of the molecular mechanisms and evolutionary forces that govern the expansion of clusters of paralogous genes are by no means completely understood. Walter J. Gehring, a developmental biologist famous for his studies of the Hox gene cluster in *Drosophila melanogaster* interpreted the fact that the three Hox genes (*abd-B*, *abd-A*, and *Ubx*) appear in a tandem arrangement as evidence for gene duplication by "unequal crossing over". He proposed that the current Hox cluster expanded from two Hox genes by a series of unequal crossing overs between highly similar but mispaired paralogous genes [19]. In this scenario, a new paralog is created as a hybrid of its left and right neighbors as indicated in Fig. 1.

The local gene duplication model constitutes an alternative explanation. Again, unequal crossover is the molecular mechanism resulting in the duplication. However, in this scenario the crossover occurs between genes and thus results in the creation of a faithful copy on the complete gene. Diversification, subfunctionalization, or neofunctionalization then drives the subsequent divergence of the paralogous sequences [38,17].

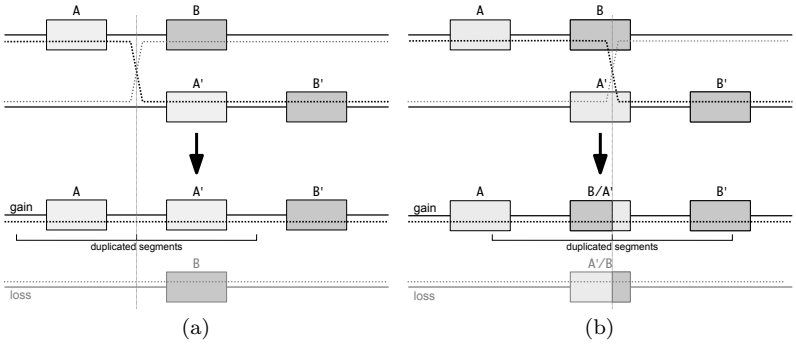


**Fig. 1** Gene cluster expansion by local gene duplication (a) and unequal crossover in Gehring's model (b). During mitosis, when chromatids are paired, unequal crossing leads to a tandem duplication on one chromatid and a deletion on the sister chromatid. The loss of whole genes is considered to be lethal. In Gehring's model the crossover occurs within the gene sequences resulting in hybrid genes. Crossover between intergenic sequences results in duplication of complete genes.

Gehring noted that terminal genes in a Hox cluster are not subject to changes by crossover and that the genes in the middle of the cluster are more similar to the consensus sequence than more distal genes. The paralogs in a cluster most similar to a given gene tend to be its neighbors. A recent analysis of the genetic distances between Hox genes, furthermore, showed that the shortest Hamiltonian path w.r.t. the genetic distance follows the genomic order of the cluster [46]. We ask here if and how these observations can be explained by Gehring's model and the local gene duplication model.

The analysis of the history of a gene family is usually based on the inference of a phylogenetic tree of the paralogous genes in question. However, this is a difficult task and often remains unsuccessful, in particular for the deep branches since several effects conspire to erase the phylogenetic signal. Saturation of the phylogenetic signal limits the power of reconstruction in particular for old events and events separated by relatively short time scales.

Genomic elements that are very similar in sequence and in close proximity, as it is the case in clusters of paralogous genes, are particularly prone to gene conversion and other mechanisms of concerted evolution [7,35]. Last but not least, the very process that introduces additional new members may involve unequal crossover in Gehring's model thus producing a non-tree-like structure of genetic distances to begin with.

The purpose of this contribution is two-fold. First, we investigate the consequences of Gehring's model for gene cluster expansion and show that while the resulting genetic distances are not additive trees, they belong to a special class of circular decomposable metrics. Therefore, they can be represented faithfully by the type of phylogenetic networks produced by the `NeighborNet`[4,5] algorithm. Furthermore, we will see that in the absence of extreme selective pressure they have the Robinson property, so that the Hamiltonian path with

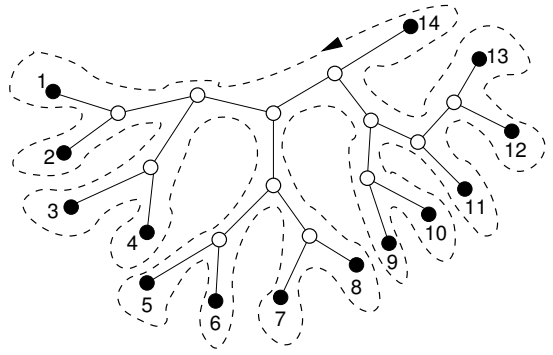


**Fig. 2** Each planar embedding $\breve{T}$ gives rise to a circular ordering of the vertices by following the "outline" around the tree.

the shortest genetic distance between genes is co-linear with the genomic order in the gene cluster. We then use this result to distinguish between gene clusters that likely have evolved under Gehring's model and retained synteny from those that have a different origin or were subjects to a rearrangement of their gene order.

## 2 Trees, Metrics, and Hamiltonian Paths

In this section we introduce the notation and provide some mathematical background information on the connection between tree metrics and Hamiltonian paths. The material presented here is mostly "folklore" and included primarily as an introduction to the more formal development of the following sections. Proofs are included for completeness since we are not aware of any convenient references.

### 2.1 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 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* [31] is a ternary relation $\triangleleft\, i\, j\, k$ on a set $X$ that satisfied the following five conditions for all $i, j, k \in X$:

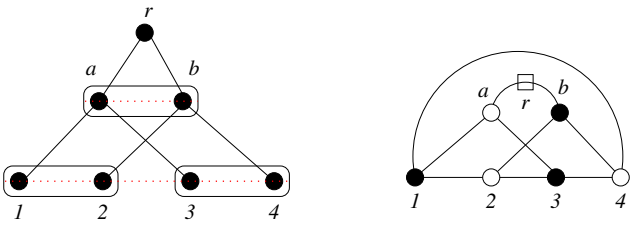


**Fig. 3** Phylogenetic tree arising from a block duplication of two paralogs. The l.h.s. sketches the phylogenetic tree and the genomic ordering of the leaves. The r.h.s. shows the corresponding graph $G_T$. After contracting the edge between $a$ and $r$, we are left with a $K_{3,3}$, hence $G_T$ is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

(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 [37]. 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 are 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. 2.

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 [20] whenever $\pi$ is *$T$-order*. A necessary condition for $\pi$ to be a *$T$-order* therefore is that $G_T$ is a planar graph.

Clearly, if 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. 3. Every duplication scenario in which more than a 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 this signature of block duplications. For instance, the loss of 2 or 3 in Fig. 3 leads back to a $T$-order.

## 2.2 From Trees to Hamiltonian Paths

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

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

$L(\pi)$ 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-degenerative length function $\ell$. Then $L(\pi)$ is minimal if and only if (i) $\pi$ 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 any 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. 2. 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. [33, 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 $e$, $T(e)$ denotes the connected component of $T \setminus \{e\}$ that does not contain 1.

***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 that $\omega$ enters 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 $u \notin V(T(e))$, that is, $T(v) = T(e)$, 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. Now suppose that $\pi$ is not a $T$-ordering. 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 $L(\pi) > 2\mathcal{L} - d_{\pi(1)\pi(n)}$. □

### 2.3 Simulating Distance Matrices for Gene Duplications

We show here that genetic distance matrices for models of gene duplications can be simulated directly. 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 satisfies the "4-point condition" [6,13,15,47], which stipulates that any four leaves can be renamed such that

$$d_{xy} + d_{uv} \leq d_{xu} + d_{yv} = d_{xv} + d_{yu} \tag{3}$$

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 $\delta_x = r_x \Delta t$ where $r_x$ is the rate of evolution of taxon $x$. A duplication of gene $x$ can be introduced by simply duplicating the row and column $x$ in the distance matrix $\mathbf{D}$, i.e., by setting $d_{zy} = d_{xy}$ for all $y \neq x, z$ and $d_{xz} = 0$. The procedure is summarized in Alg. 1.

A rate $r_{x'}$ (and possibly a new rate $r_x$) needs to be chosen. Assuming a constant rate of duplication, we set $\Delta t = 1/n$ and 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 Alg. 1.*

**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**

*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

$$\min_{x,y \in V \setminus \{p,q\}} \left\{ (d'_{px} + d'_{qy}) - (d'_{pq} + d'_{xy}) \right\} > 0 \tag{4}$$

If $\{p, q\}$ is a cherry, then the distances in $T$ from $p$ and $q$ to their last 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 $\mathbf{D}$ on $V \setminus \{q\}$ defined by $d_{xy} = d_{xy}$ for $x, y \notin \{p, q\}$, $d_{xp} = d'_{xp} - \delta_p$ represents $T$ with the cherry replaced by its last common ancestor, hence it is again an additive distance matrix.

Repeating this construction we arrive at 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 the trees' branches modeled in Alg. 1 can be subdivided arbitrarily between duplication events that affect a particular branch. Here we simply choose to 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. □

## 3 Type R Distance Matrices

### 3.1 Construction and Recognition

The model so far corresponds to a mechanism in which unequal crossover occurs only *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

$$\begin{aligned} d_{zu} &= ad_{xu} + (1-a)d_{yu} \\ d_{zx} &= (1-a)d_{xy} \\ d_{zy} &= ad_{xy} \end{aligned} \tag{5}$$

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

$$d'_{pq} = d_{pq} + \delta_p + \delta_q \tag{6}$$

**Definition 1** A distance matrix $\mathbf{D}$ is of *type R* if it is constructed by repeated application of Eqns.(5) and (6)

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 a type R matrix can be recognized efficiently.

We start by observing

$$d'_{xz} + d'_{yz} - d'_{xy} = d_{xz} + d_{yz} - d_{xy} + \delta_x + \delta_z + \delta_y + \delta_z - \delta_x - \delta_y = 2\delta_z \tag{7}$$

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

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

$$\begin{aligned} d'_{uz} - ad'_{ux} - (1-a)d'_{uy} &= \underbrace{d_{uz} - ad_{ux} - (1-a)d_{uy}}_{=0} \\ &\quad + \delta_u + \delta_z - a\delta_u - a\delta_x - \delta_u + a\delta_u - \delta_y + a\delta_y \\ &= \delta_z - a\delta_x - (1-a)\delta_y := f(a) \end{aligned} \tag{8}$$

The key observation is that this expression is independent of $u$. Thus, for $n \geq 5$, there 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,

$$a = \frac{(d'_{uz} + d'_{vy}) - (d'_{vz} + d'_{uy})}{(d'_{ux} + d'_{vy}) - (d'_{vx} + d'_{uy})} \tag{9}$$

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 $n > 5$.

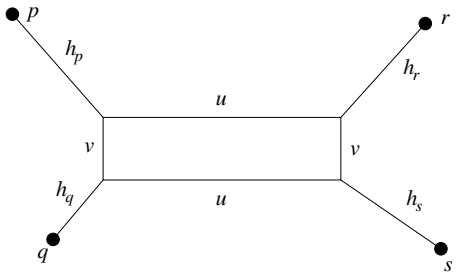


**Fig. 4** Representation of a metric $d$ on 4 points $\{p, q, r, s\}$. Each distance is the sum length of a shortest path in this graph. For instance $d_{pq} = h_p+v+h_q$, $d_{pr} = h_p+u+h_r$, $d_{ps} = h_p+u+v+h_s$.

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

$$\begin{aligned} -a\delta_x - (1-a)\delta_y &= f(a) - \delta_z \\ (1-a)d_{xy} + \delta_x &= d'_{xz} - \delta_z \\ ad_{xy} + \delta_y &= d'_{yz} - \delta_z \end{aligned} \tag{10}$$

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 second and third 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

$$2a(1-a)d_{xy} = (d'_{uz} - ad'_{ux} - (1-a)d'_{uy}) - 2\delta_z + ad'_{xz} + (1-a)d'_{yz} \tag{11}$$

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

$$\begin{aligned} \delta_x &= d'_{xz} - (1-a)d_{xy} - \delta_z \\ \delta_y &= d'_{yz} - ad_{xy} - \delta_z \end{aligned} \tag{12}$$

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 $z$. If $a = 0$ or $a = 1$ in Eqn.(9) then $z$ is a copy of $x$ or $y$, resp. In this case we cannot determine $d_{xy}$ from Eqn.(11) since $2a(1-a) = 0$. By construction, however, we can just remove $z$ from the matrix to obtain the ancestral state.

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 in Gehring's model event.

To see this more clearly, let us first consider the case $n = 4$. It is well known that every metric on four points can be represented as a "box graph" as shown in Fig. 4. The box dimensions can be computed from $2u = (d_{ps} + d_{qr}) - (d_{pq} + d_{rq})$ and $2(u-v) = (d_{rp} + d_{qs}) - (d_{pq} + d_{rs})$. The key ingredients, thus are the three different pairs of distances emphasized by parentheses. For more details see [34]. Now let us start from an arbitrary distance matrix $\mathbf{D}$ on $\{x, y, u\}$ and construct $z$ as a recombinant. In the following, we will use abbreviations for the three pairs of distance sums, thus

$$A = d'_{xz} + d'_{uy} \qquad B = d'_{yz} + d'_{ux} \qquad C = d'_{uz} + d'_{xy}\,. \tag{13}$$

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

**Require:** Distance matrix $\mathbf{D}'$, $n = |V| \geq 4$
  **repeat**
    **for** $(x, y, z) \subseteq V$ **do**
      **for** $\{u, v\} \subseteq V \setminus \{x, y, z\}$ **do**
        compute $a$ using Eqn.(9)
      **end for**
      **if** $a \in [0, 1]$ is the same for all $u, v$ **then**
        **if** $a \neq 0, 1$ **then**
          compute $\delta_z$ using Eqn.(7)
          compute $d_{xy}$ using Eqn.(11)
          compute $\delta_x$, $\delta_y$ using Eqn.(12)
          $\delta_u \leftarrow 0$ for $u \in V \setminus \{x, y, z\}$
          compute $\mathbf{D}$ as $d_{pq} = d'_{pq} - \delta_p - \delta_q$ for all $p, q \in V$
        **end if**
        $\mathbf{D}' \leftarrow \mathbf{D}$ 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**

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

$$\begin{aligned} C - A &= a(d_{xy} + d_{xu} - d_{uy}) \geq 0 \\ C - B &= (1 - a)(d_{xy} + d_{yu} - d_{ux}) \geq 0 \end{aligned} \tag{14}$$

using again the triangle inequality. The terms $C - A$ and $C - B$ correspond to twice the sides of the box in the quadruple graph, shown in Fig 4; 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 $v \neq w \in V$ 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 event. This 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 Alg. 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 duplication event. Note that it suffices to consider $x < y$. If $a = 0$ or $a = 1$, then $z$ was obtained as a faithful copy of $x$ or $y$, resp., and hence it can just be dropped. If a candidate triple $\{x, y, z\}$ is found, the previous distance matrix $\mathbf{D}'$ is computed in quadratic time. Thus Alg. 2 runs in $O(|V|^6)$ time.

For $|V| = 4$ the remaining distance matrix is represented by a unique box as in Fig. 4, which implies a unique circular order of the remaining four nodes, say $u, x, y, z$. The fourth node therefore must be the result of unequal crossover of two nodes that are placed a diagonally opposite corner of the box. Therefore $(u, y : x)$, $(x, z : y)$, $(y, u : z)$ and $(z, x : u)$ are equivalent.

### 3.2 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 $y$.

Alg. 2 identifies triples $(x, y : z)$ so that $z$ was obtained as a recombinant of $x$ and $y$, i.e., that $z$ is located between $x$ and $y$ together with a possible temporal order of these events. It is difficult in general to determine whether a linear order exists that is compatible with an arbitrary collection of betweenness triples: the so-called `Betweenness Sorting Problem` is NP complete [40, 10]. Here, however, we have much more information. We call a type R matrix generic if for every $z$ both parents are uniquely defined. We say that $(u, v : w)$ is a successor of $(x, y : z)$ if $\{u, v\} = \{x, z\}$ or $\{u, v\} = \{y, z\}$. A triple without a successor is a leaf triple.

With a leaf triple $(x, y : z)$ we can associate the path $p_{xy} := x - z - y$. If a triple $(x, y : z)$ has only one successor, say $(x, z : u_1)$, we set $p_{xy} = p_{xz}(z-y)$. If it has two successors, these are of the form $(x, z : u_1)$ and $(z, y : u_2)$, and we set $p_{xy} = p_{xz}p_{zy}$. This is, the paths corresponding to the two "intervals" $x-z$ and $z-y$ are joined at the common vertex $y$. By construction of type R matrices, each triple has at most one predecessor, hence the path $p_{xy}$ is uniquely and completely defined for every triple. A triple $(x, y : z)$ has no predecessor only if $x$ and $y$ are two of the three ancestral nodes. There are at most two such triples by construction of linear type R matrices, which necessarily have one node in common. The paths are joined at this common node. The type R matrix is linear if the final concatenation result is a single path, in which each node appears exactly once. By construction, $z$ is located between $x$ and $y$ for all triples $(x, y : z)$, i.e., the final path encodes the desired linear order of the nodes.

Representing the paths $p_{xy}$ as lists, joining at their end points can be performed in constant time. Any triple $(x, y : z)$ can be a left or right successor to another triple on $(x, y)$, accept a left successor on $(x, z)$, or accept a right successor on $(z, y)$. For each triple, joining to already processed triples and/or generating references for later triples can be achieved in $O(1)$ utilizing these tuples as keys in associative arrays (one per connection type), e.g. using a quadratic array or (sparse) hash-maps. The successor/predecessor relation between the $O(n)$ triples can therefore be established in linear time if the triples that account for duplications are already known. Thus, linearity of a type R matrix can be checked in linear time (see Alg. 3 in the Appendix).

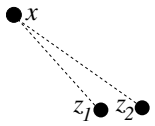


**Fig. 5** Representation of a successor-predecessor tree after two duplications of the same gene $x$: $(x : z_1)$ and $(x : z_2)$. As the time order of duplications to $z_1$ and $z_2$ are unknown, so is their relation in the genome. Both $x - z_1 - z_2$ and $x - z_2 - z_1$ are proper solutions.

This algorithm can also be extended to the non-generic case. Instances with $a = 0$ or $a = 1$ duplications result in $(x : z)$ relations with unknown second flanking gene, which can cause several problems. While the algorithm above can always find one linear configuration, this is no longer unique in the non-generic case. Any pair obtained as "clones" from the same parent have no defined order among themselves, unless a later triple with $0 < a < 1$ can resolve it (see Fig. 5). Hence, the predecessor-successor relationship is no longer binary, but rather any gene might relate to an unlimited number of perfect copies. This requires careful indexing on individual genes, as listing gene tuples would create exponential growth of open references.

Let us now turn to the connection of type R matrices and circular orders.

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

$$\max\{(d_{ij} + d_{kl}), (d_{il} + d_{jk})\} \leq d_{ik} + d_{jl} \tag{15}$$

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

If $(d_{ij})$ satisfies Eqn.(15) then the corresponding TSP is solved by the unit permutations, i.e., $\pi = (1, 2, 3, \ldots, n)$ [23]. Equivalently, if $\lhd$ is a circular ordering of the taxa set $V$ and $\pi$ the permutation of $V$ associated with an arbitrary linearization of $\lhd$, then $(d_{ij})$ is Kalmanson iff

$$\max\{(d_{\pi(i)\pi(j)} + d_{\pi(k)\pi(l)}), (d_{\pi(i)\pi(l)} + d_{\pi(j)\pi(k)})\} \leq d_{\pi(i)\pi(k)} + d_{\pi(j)\pi(l)} \tag{16}$$

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

With each circular ordering $\lhd$ we can associate a set $\mathcal{S}^{\lhd}$ of splits, i.e., non-trivial bipartitions of the set $X$ of taxa. $\{A, X \setminus A\} \in \mathcal{S}^{\lhd}$ 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 $\lhd i\, p\, j$ and $\lhd k\, q\, l$ and (b) $\lhd i\, j\, k$ and $\lhd k\, l\, i$. We write

$$S_{ij} := \{\{\pi(i+1), \pi(i+2), \ldots, \pi(j)\}, \{\pi(j+1), \pi(j+2), \ldots, \pi(i)\}\} \tag{17}$$

with $i, j$ taken $\mod |X|$ for the splits of $\mathcal{S}^{\lhd}$, where $\pi$ is again an arbitrary linearization of $\lhd$. A metric is called *circular decomposable* [2] if there is a circular ordering $\lhd$ (with a corresponding permutation $\pi$), and $\alpha_{ij} \geq 0$, $i \neq j$ so that

$$d_{xy} = \sum_{i<j} \alpha_{ij} \delta_{S_{ij}}(x, y)\,, \tag{18}$$

where the split pseudometric $\delta_{S_{ij}}$ is defined as $\delta_{S_{ij}}(x,y) = 1$ if the split $S_{ij}$ separates $x$ and $y$, and $\delta_{S_{ij}}(x,y) = 0$ otherwise. Such expressions are known as "Crofton formulas" [9]. The *isolation indices* of the splits $S_{ij}$ can be computed as

$$\alpha_{ij} = \alpha(S_{ij}) = \frac{1}{2}\left(d_{\pi(i)\pi(j)} + d_{\pi(i+1)\pi(j+1)} - d_{\pi(i)\pi(j+1)} - d_{\pi(i+1)\pi(j)}\right) \quad (19)$$

It is shown in [11,9] that a metric satisfies the Kalmanson condition if and only if it is circular decomposable. These can represented as so-called split graphs and computed efficiently using the `NeighborNet` algorithm [4,5].

As shown in [26, Thm.37], the solution of the TSP on a *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` [25,4,5]. Note that this is not true for special case of additive tree metrics.

**Theorem 2** *Every linear type R distance matrix satisfies the Kalmanson 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 $z$,

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

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 (omitting the analog of the second and third line above) yield: $\max\{d_{jz} + d_{kl}, d_{jl} + d_{zk}\} \leq a\max\{d_{kl}, d_{jl} + d_{jk}\} + (1-a)\max\{d_{jj'} + d_{kl}, d_{jl} + d_{j'k}\} \leq a(d_{jl} + d_{jk}) + (1-a)(d_{jk} + d_{j'l}) = d_{jk} + d_{zl}$; $\max\{d_{iz} + d_{j'l}, d_{il} + d_{zj'}\} \leq a\max\{d_{ij} + d_{j'l}, d_{il} + d_{jj'} + (1-a)\max\{d_{ij'} + d_{j'l}, d_{il}\} \leq a(d_{ij'} + d_{jl}) + (1-a)(d_{ij'} + d_{j'l}) = d_{ij'} + d_{zl}$; $\max\{d_{ij} + d_{zj'}, d_{ij'} + d_{jz}\} \leq a\max\{d_{ij} + d_{jj'}, d_{ij'}\} + (1-a)\max\{d_{ij}, d_{ij'} + d_{jj'}\} = a(d_{ij} + d_{jj'}) + (1-a)(d_{ij'} + d_{jj'}) = d_{jj'} + d_{iz}$. We conclude that all quadruples involving $z$ satisfy the Kalmanson inequality provided the distances $(d_{ij})$ from a Kalmanson metric 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 that do not involve the new offspring $z$ remain unchanged by the construction principle of type R matrices, we conclude that the distances $(d_{ij})$ on $X \cup \{z\}$ also satisfies the Kalmanson inequalities. □

### 3.3 Robinsonian Distances and Hamiltonian Paths

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 circular decomposable if and only if the Kalmanson conditions also hold for quadruples involving the dummy node, i.e., if and only if

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

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

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

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

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

The so-called serialization problem [44,27] of linearly ordering objects is solved by the order $\pi$ for Robinsonian dissimilarities. This result 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 adjacencies $[\xi(i)\xi(i+1)]$ w.r.t. $\xi$ and the adjacencies $[p, p+1]$ w.r.t. $\iota$ so that $\xi(i) \leq p < p+1 \leq \xi(i+1)]$. 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 adjacency $[\xi'(k)\xi'(k+1)] = [u, v]$ 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 [14,24,43]. So-called quadripolar Robinson dissimilarities that also satisfy the Kalmanson condition are studied in some detail in [12].

**Lemma 4** *Suppose $(d_{ij})$ satisfies Eqn.(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 Eqn.(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 $d_{zj''} = ad_{j'j''} \leq d_{j'j''}$. □

It is important to note that the choice of $\delta_k$ can destroy the inequality: From $\max\{d_{ij}, d_{jk}\} \leq d_{ik}$ we cannot conclude that $\{d_{ij} + \delta_i + \delta_j, d_{ij} + \delta_j + \delta_k\} \leq d_{ik} + \delta_i + \delta_k\}$. 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. The Robinson condition should be satisfied at least in very good approximation if the evolution rates of the offspring are not too different. Gene conversion, which effectively reduces distances, should make it even easier to satisfy Eqn.(21).

## 4 Simulations and Application to Real-Life Data

### 4.1 Inference of Gene Order from Distance Data

The theory outlined above predicts that "well-behaved" gene clusters, i.e., those that (i) evolved by duplication of single genes only and (ii) did not experience rearrangements should be Robinsonian. In other words, the shortest Hamiltonian path w.r.t. the genetic distances between its constituents should be co-linear with the genomic order. It is therefore of interest to study the length distribution of Hamiltonian paths. Associating a pseudo-energy $f(\pi) = \sum_{i=2}^{n} d(\pi_{i-1}, \pi_i)$ with a path/permutation $\pi$ we may construct a probabilistic model where $Prob[\pi] \propto \exp(-\beta f(\pi))$ with an "inverse temperature" parameter $\beta$. In [45,46] we have shown that this model is tractable by a variation of the well-known exponential-time dynamic programming approach to the Travelling Salesman Problem [3]. In brief, the ensemble $(p, A, q)$ of paths starting in $p$, ending in $q$ and running through all elements of $A$ is of the form $(p, A, q) = \bigcup_{u \in A}(p, A \setminus \{u\}, u) \circ (u, q)$. Using a variant of algebraic dynamic programming on sets, this simple decomposition can be used to compute the posterior probabilities of adjacencies in the ensemble of Boltzmann-weighted paths as well as the posterior probabilities of vertices $p$ and $q$ to be endpoints of a Hamiltonian path. Further details on the method can be found in [45,46]. It is implemented in the `Gene Cluster Evolution Determined Order` software package `Gene-CluEDO`. [1]

[1] `http://hackage.haskell.org/package/Gene-CluEDO`

Since the genetic distance matrix is expected to have the Kalmanson properties the `NeighborNet` [4,5] algorithm can be used as an alternative method to infer the expected gene order. The consistency theorem for `NeighborNet` [4,5] in particular guarantees that the correct order will be obtained for ideal input data, i.e., input data that satisfies the Kalmanson condition. In practice, `NeighborNet` has turned out to be rather resilient to noise. Hence, it can be expected to produce good approximations to the gene order also for imperfect, noisy input data. Concurrence of `Gene-CluEDO` and `NeighborNet` can thus be used as support for the correctness of the reconstructed order, see Fig. 6.

### 4.2 Simple Simulation of Gene Cluster Evolution

In order to test whether sequence evolution indeed approximates type R distances we generated artificial amino acid sequence data starting from a random initial sequence of length $N$. For the data reported here we use $N = 1000$ and a uniform distribution of the 21 amino acids (including selenocystein). In each iteration, first a recombinant sequence $z$ is produced from two adjacent parents $x$ and $y$ so that $z$ is placed between $x$ and $y$. To model unequal crossover in Gehring's model we randomly choose a breakpoint position $k$ and produce $z$ as concatenation of $y[1,k]$ and $x[k+1,n]$. In the first step, the initial sequence is simply copied. We also consider the case where the breakpoint is outside the "gene", i.e., instead of producing a recombinant sequence $z$ we use a copy of $x$ or $y$ with probability $\psi$. If $\psi = 1$, we obtain the limit of tree-like evolution.

The second part of each iteration consists of independent mutations applied to all sequences. To this end, we replace with probability $\mu$ the amino acid in each sequence position by a randomly chosen alternative. The per site mutation rate $\mu$ must be chosen large enough to ensure a measurable divergence in each step. On the other hand, the sequence divergence should not saturate after $n$ duplication-mutation steps, i.e., the expected total number of mutations per sites should not substantially exceed 1. Thus $1/N \lesssim \mu \lesssim 1/n$.

Since we do not simulate insertions and deletions, the sequences are already properly aligned. In order to obtain an approximately additive distance matrix from the simulated sequences we use the Jukes-Cantor transformation [22] to account for multiple mutations hitting the same site. Fig. 6 shows data for simulation with only local gene duplications in (a) and with unequal crossover in Gehring's model in each step in (b) to (e).

The gene order in the cluster and the reconstructed order in either the `Gene-CluEDO` or the circular order inferred using `NeighborNet` do not match for tree-like evolution. The reason is that in this case many orders, namely all outlines of any planar embedding of the tree, are equivalently perfect data. The simulated sequence data by construction contain stochastic noise that breaks this symmetry in a random manner. More precisely, distances empirically inferred from sequences will satisfy the equality in equ.(3) only approximately. As a consequence, the tree edge belonging to the split $xy|uv$ will be expanded to narrow box as in Fig. 4. It is completely up the noise, whether the sec-

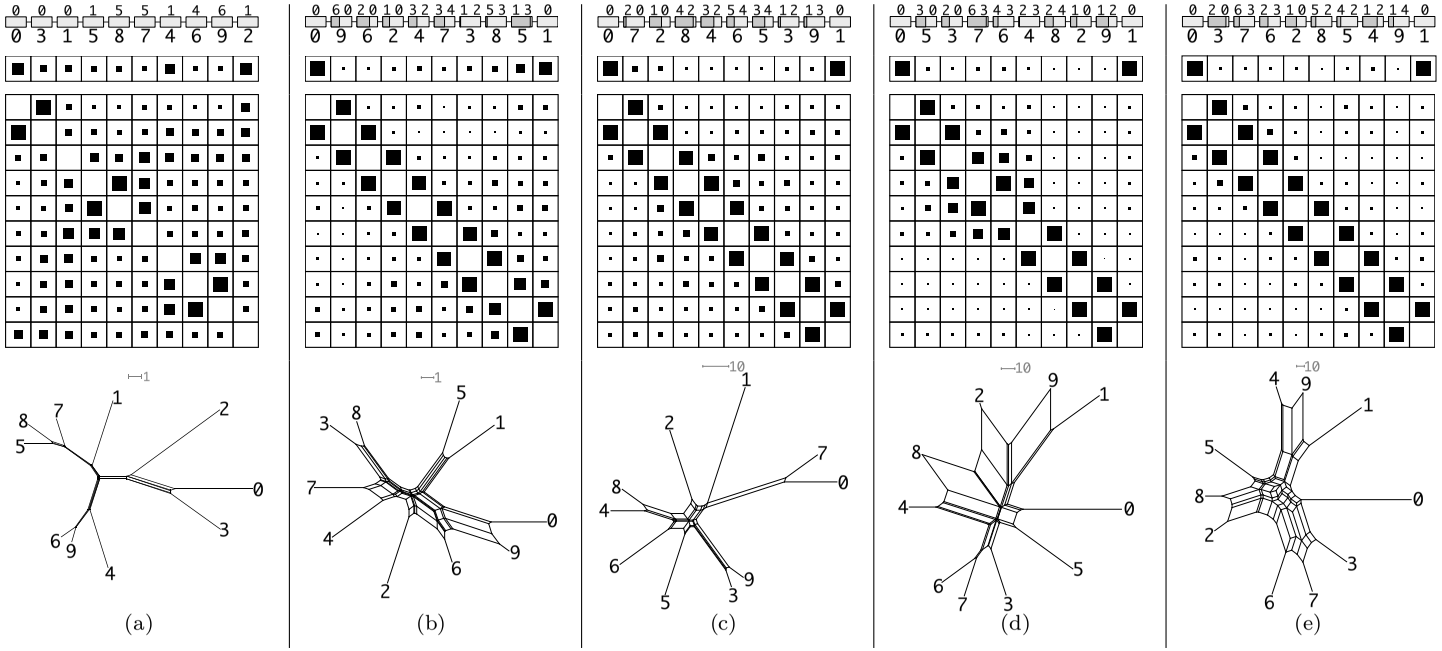


**Fig. 6** Examples of simulated gene clusters (see text for details). The mutation rate $\mu$ in the simulations are: in (a) and (b) 1%, in (c) 5%, in (d) 10%, and in (e) 15%. Only in (a) local gene duplication events are employed as model while in (b)–(e) unequal crossover events as proposed by Gehring's model are used to create the cluster. Each column is a composite of four rows. In the first row the simulated cluster and its genes including their history is shown. Here, in the upper part the ancestral gene/genes are noted while on the bottom the simulated order is depicted. In the case an unequal crossover in Gehring's model, the two parents are the number above the sequence block where the left number contributes the left (dark grey) part of the sequence and the right number the right part (light grey) of the sequence. In the second and third row the results of `Gene-CluEDO` are displayed. They are created with $\beta = 0.01$. The size of the black box in a cell is proportional to the likelihood of this cell. The second row shows the probability that a sequence is on the edge of the cluster. The third row gives the probability that two sequences are adjacent to each other in the cluster. The last row then shows the network that is created with the `NeighborNet` [4,5] algorithm. Note that the `NeighborNet`s may scale differently indicated by a grey scale bar above the net.

ond split is $xu|yv$ or $xv|yu$, an thus, whether the circular order is $x, u, v, y$ or $x, v, u, y$.

In contrast, both `Gene-CluEDO` and circular order reproduce the gene order in the cluster in the vast majority of simulations with unequal crossover in Gehring's model. The choice of the mutation rates $\mu$ makes little difference as long as the genetic distances between the sequences are not saturated.

An exception is Fig. 6(c), where `NeighborNet` "misplaces" sequence 1. A detailed analysis of the data shows that both 3 and 9 are unequal crossover products involving 1, however by chance the breakpoint was located so that only a tiny fraction of 1 was included in 3 and 9. The example thus contains an "almost tree-like" step, which does not retain sufficient ordering information.

## 4.3 Analysis of Gene Clusters

### *4.3.1 Pairwise Distances*

In the following we illustrate the application of the theoretical results to the analysis of several gene clusters. To this end, we retrieved the amino acid sequence data of the annotated proteins from the NCBI data base, constructed and—where necessary—manually curated sequence alignments, and used these to compute the matrices of pairwise genetic distances that are taken as input by both `Gene-CluEDO` and `NeighborNet`.

Multiple sequence alignments were computed with `T-Coffee` [36]. Since highly variable regions in the proteins mostly introduce noise into the alignment and the subsequent reconstruction of the phylogenetic network, we removed highly variable alignment columns using `noisy` [16]. From the processed alignment we then computed the evolutionary distances interpreting gap characters as additional characters. The resulting raw distances are transformed into evolutionary distances using the Jukes-Cantor correction [22]. For the lancelet Hox cluster we obtained an extremely gap-rich alignment. We therefore constructed an alternative alignment using the block-based `dialign` approach [1], which identifies a chain of significant local alignments. We retained only the alignment blocks with a non-zero significance score.

### *4.3.2 Hox gene cluster*

We already showed in previous work [46] that the Hamilton path method implemented in `Gene-CluEDO` can be applied to investigating the ancient evolution of Hox gene clusters. Cephalochordates harbour the largest known single Hox gene clusters, comprising 15 members [41]. The Hox gene clusters are known to have expanded independently in the major deuterostome lineages [42] making them a particularly interesting model system for testing Gehring's model. The results of this analysis are shown in Fig. 7. Over all, the amphioxus cluster behaves as expected. In line with the analysis of Hox clusters from the coelacanth [46], both `Gene-CluEDO` and `NeighborNet` reproduce the genomic

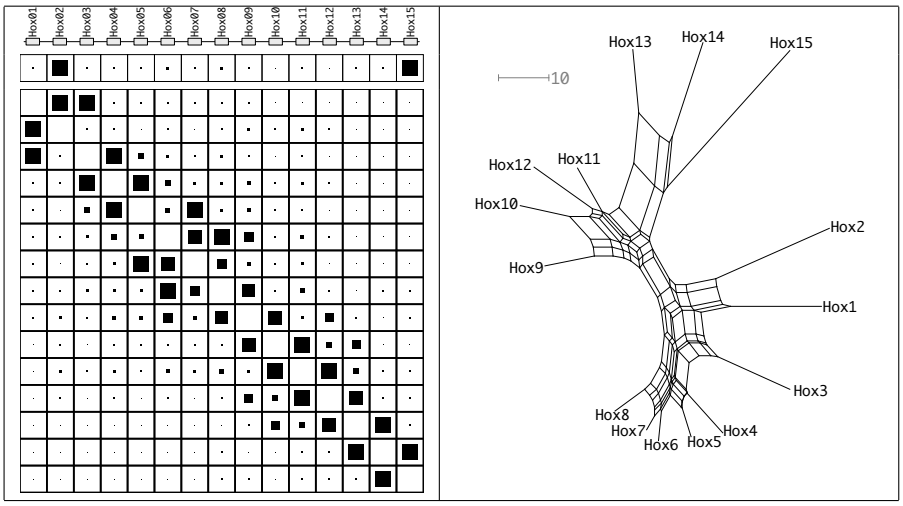


**Fig. 7** The Hox gene cluster of *B. lanceolatum*. How the pairwise distances are created is described in Section 4.3.1. The left site is a composite of three rows. The first row shows the cluster and the order on the genome. In the second and third row the results of `Gene-CluEDO` are displayed. They are created with $\beta = 0.0025$. The size of the black box in a cell coincides with the likelihood of this cell. The second row shows the probability that a sequence is on the edge of the cluster. The third row gives the probability that two sequences are adjacent to each other in the cluster. The right site then shows the network that is created with the `NeighborNet` [4,5] algorithm. The network scale is indicated by a grey scale bar.

arrangement. There are a few notable deviations, however: Both methods report a reversed ordering of HOX1 and HOX2. A `blastp` search, however, confirmed that the sequences of these two genes unambiguously belong to the HOX1 and HOX2 paralog groups that are present in all deuterostomes. We suspect that adaptive evolution of one of these genes may be responsible for the observed discrepancy. `NeighborNet` shows HOX11 and HOX12 in reverse order. However, the splits involved in establishing this ordering have very small weights, suggesting that this reversal is not significant.

We conclude, therefore, that the evolution of the HOX gene cluster most likely followed Gehring's model. Another aspect supporting this conclusion is the placement of splits in the network created by `NeighborNet`. The genes are placed in a nearly perfect circle around the center of the network. Comparing its topology to the topologies of the clusters created by simulating Gehring's model, we can see high similarity in the network structures (see Fig. 6).

### *4.3.3 PSG gene cluster*

The pregnancy-specific glycoproteins (PSG) play an important role in the immune system during pregnancy [8]. They form a well-defined subfamily of the Carcinoembryonales Antigen gene family, which in turn belongs to the immunoglobulin gene superfamily. The PSG family forms a cluster that has independently expanded in some mammalian classes, most prominently in ro-

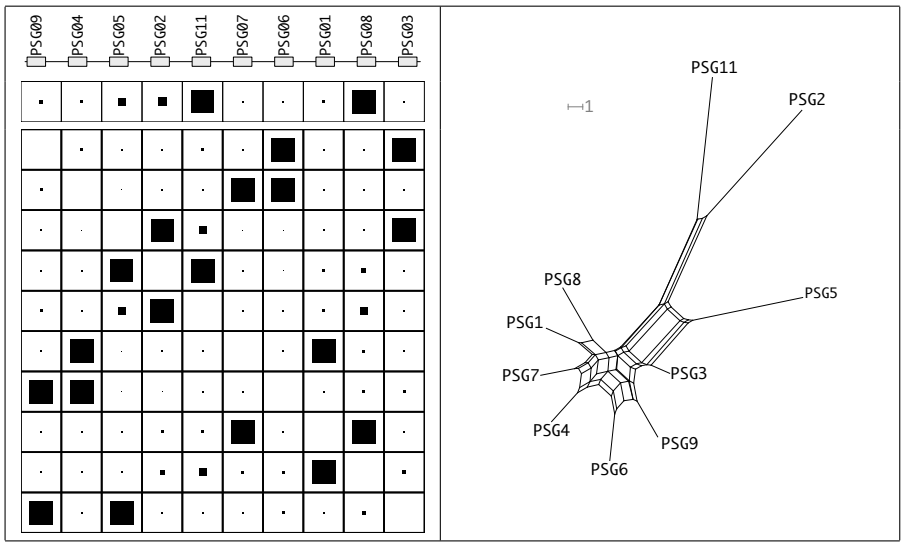


**Fig. 8** The PSG gene cluster of *Homo sapiens*. For additional legends see Fig. 7.

dents and primates. Here we analyzed the *human* PSG gene cluster, which contains ten PSG genes. Five CEACAM pseudogenes are interspersed in the cluster. The results of this analysis are shown in Fig. 8.

The data shows two remarkable properties. Consistent with evolutionarily recent duplications the PSG genes are very similar to each other. The second remarkable property is that the orders inferred with `Gene-CluEDO` and `NeighborNet` do not fit to the real genomic order. In fact only three (`Gene-CluEDO`) or four (`NeighborNet`) genes appear in the order of their genomic positions. The data are not consistent with the prediction from Gehring's model.

Two aspects provide possible explanations. Mouldi *et al.* [48] proposed that the PSG gene cluster in primates evolved under purifying selection for gene conversion. Chang *et al.* [8] proposed that a high number of unequal crossover events had occured in primate evolution. A very large number of duplicates, however, may reduce the selection pressure on single gene copies such that gene loss is no longer lethal. This may lead to missing genes and to large differences in evolution rates of individual copies. The latter may account for a violation of the Robinson property, and thus deviations between the observed genomic gene order and the order inferred by `Gene-CluEDO` from the genetic distances. An observation that supports these explanations is that PSG11 and PSG2 stand out of the other genes as relative diverse (see `NeighborNet` plot). Possibly genes that could close this gap were lost due to unequal crossover.

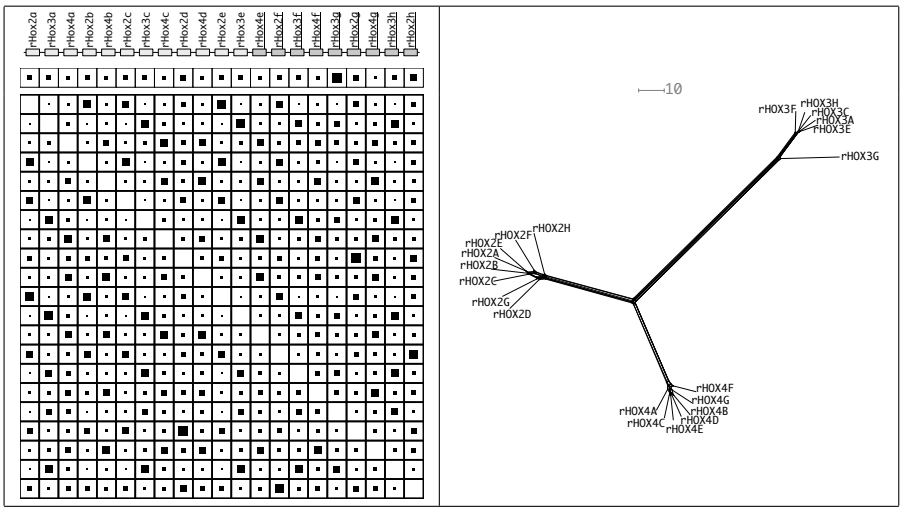


**Fig. 9** The $\alpha$-Rhox gene cluster of *Mus musculus*. Genes oriented in the opposite reading direction are indicated by darker boxes and underlined gene names. For additional legends see Fig. 7.

### *4.3.4 $\alpha$-Rhox gene cluster*

The Rhox genes [29] are expressed during both embryogenesis and in adult reproductive tissues. In the mouse they are located in a single cluster on the X chromosome comprising 33 genes in three subclusters ($\alpha$,$\beta$ and, $\gamma$). The Rhox cluster is notable for its unusually rapid evolution. Here we included 23 well annotated genes of the $\alpha$-Rhox cluster, after removing the pseudogene rHox3d, the highly diverged rHox1 sequence, as well as rHox3b, for which no translation is reported in the NCBI.

Fig. 9 shows that the data set is divided into three groups. All rHox2 genes are in one group (left), all rHox3 genes form the second group (bottom) and all rHox4 genes build the third group (top right). These groups are clearly separated from each other. The $\alpha$-Rhox gene cluster clearly has not evolved conforming to Gehring's model. As described e.g. in [28], the basic unit of tandem duplications is a block comprising an rHox2, rHox3, and rHox4 gene. Subsequent gene losses further restructured the cluster. In addition the cluster was subject to an inversion. Our analysis does not contradict this scenario.

### *4.3.5 ADH gene cluster*

The alcohol dehydrogenases (ADH) family exists in a wide range of taxa, from bacteria to plants and humans [39]. Their main function in animals is to break down alcoholes that are otherwise toxic. Most members of this gene family appear in a well-studied gene cluster. The *Human* ADH gene cluster comprises seven genes, one each belonging to classes 2-5 as well as three paralogous of class 1 ADHs. Here, we find three elements in the cluster, which also cluster

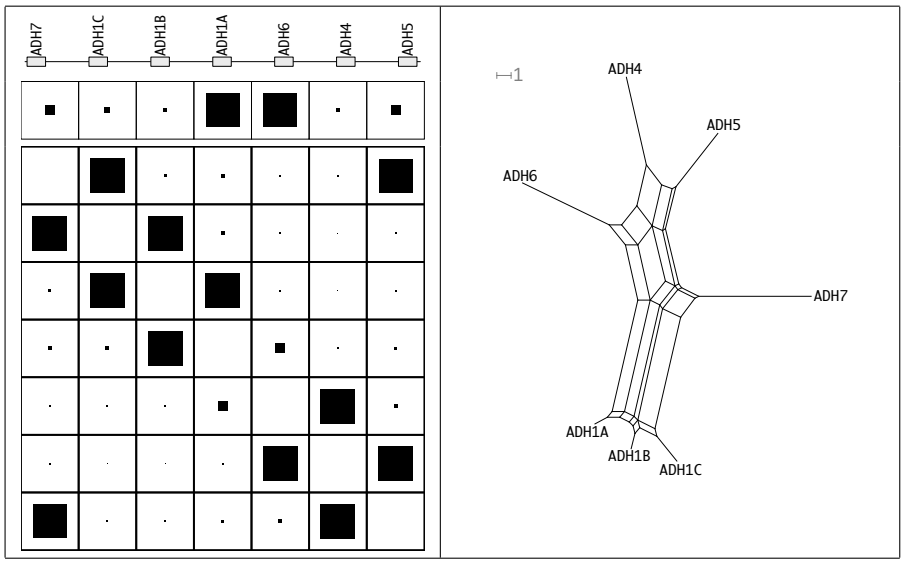


**Fig. 10** The ADH gene cluster of *Homo sapiens*. For additional legends see Fig. 7.

together regarding the results of `Gene-CluEDO` and `NeighborNet`, shown in Fig. 10.

As the genes are relatively similar to each other, genetic distances are small. The reconstructed cycle order inferred with both `Gene-CluEDO` and `NeighborNet` is the same as the genomic gene order. `Gene-CluEDO` identified ADH1A and ADH6 as the extreme ends in terms of genetic distance. These two genes are located adjacent to each other in the middle of the cluster. This may be an artefact of the small distances, since ADH5 and ADH7, for instance, have more or less the same distance to the split point inferred by `Gene-CluEDO`.

Our analysis thus suggests that the cluster evolved in line with Gehring's model. The order is perfectly reconstructed. It is argued in [39] based on the observation that different exons of the genes resulted in different maximum parsimony trees that the ADH1 genes have not been subject to gene conversion [39]. This observation is also consistent with the assumption of unequal crossover within the gene as mechanism underlying the duplications: in this scenario, duplicate genes are composed of two parts of two distinct genes, with different evolutionary history. Gene duplication following Gehring's model therefore provides an explaining for the differences in exon-specific tree reconstructions as observed vor ADH gene clusters.

## 5 Conclusions

In this contribution we have investigated in some detail a model of gene cluster evolution that goes beyond identical tandem copies. Based on Walter Gehring's

ideas, we saw that unequal crossing over events produce genes that are hybrids of their adjacent genes. The distances between the members of a gene cluster therefore are not expected to be tree-like. Instead they form a distinctive subclass of circular decomposable (Kalmanson) distances, which we have termed here type R. As a consequence, the genomic gene order matches the circular order associated with the Kalmanson-type genetic distance matrix. The `NeighborNet` algorithm [4], a commonly used tool for the inference of phylogenetic networks, readily infers this order. This provides a simple method to check whether a gene cluster evolves according to Gehring's model or not. To better characterize type R distances, we showed that they are recognizable in polynomial time and that the sequence of unequal crossover events can be inferred from a given type R distance matrix.

Additive tree metrics, which arise if the crossover breakpoints are located between genes, are a special case of type R distances. In this case, the circular order is ambiguous since an arbitrary decision can be made at each interior vertex of the phylogenetic tree. More precisely, all planar embeddings of the phylogenetic tree yield a valid circular order.

The genetic distances of gene clusters evolving according to Gehring's model of unequal crossover within genes also satisfy the Robinson condition, at least as long as selective pressures and thus evolutionary rates on paralogous members are not too different. This implies that shortest Hamiltonian paths w.r.t. the genetic distance should be co-linear with the genomic order of genes. Numerical simulations show that this type of co-linearity can be used to distinguish clusters that evolve through unequal crossover within genes from clusters where unequal crossing over occurs (mostly) between genes. The tree-like evolution in the latter case yields equivalent solutions of the shortest Hamiltonian path problem, again corresponding to arbitrary planar embeddings of the tree. Small amounts of noise in the data then typically yield optimal solutions that differ substantially from co-linearity with the genomic arrangement.

We tested these ideas using well-studied gene clusters as examples. The Hox cluster of the lancelet, for instance, essentially follows Gehring's paradigm. This is also true to a certain extent for the ADH gene cluster. Other clusters, such as the cluster of rodent Rhox genes or the PSG immunoglobulines, however, show little or no indication of unequal crossover within genes, and drastic deviations from co-linearity between gene orders inferred from genetic distances and their actual genomic arrangements.

The work presented here focused on the mathematical foundations and the demonstration that genetic distance matrices are informative about the mode of gene cluster evolutions. Several open problems remain, in particular related to practical applications. The recognition algorithm Alg. 2 requires an exact type R structure. Since the conditions for a metric to be type R involves equalities, an empirically determined distance matrix generically will not be type R due to noise. This begs the question how a best-fitting type R matrix can be identified, and how the deviation from a type R matrix should be quantified most appropriately. Together with the approximation of a type

R matrix it would be useful to compute the most likely sequence of unequal crossovers.

In this contribution we have considered only the special case that unequal crossover is restricted to adjacent genes. This assumption does not cover all cases of biological interest, as the case of the Rhox cluster shows: there, the unit of duplication is a sequence of three genes. It will be interesting to see, whether unequal crossover events that lead to the duplication of larger subclusters leads to similar mathematical structures, and whether such events could be inferred from a careful analysis of the genetic distance matrix.

**Acknowledgements** SJP gratefully acknowledges a series of conversations with Walter Gehring that sparked the idea to this project, which unfortunately was realized only after he passed away. The simulations and the analysis of real-life gene clusters reported here were the topic of a bioinformatics computer lab course at U. Leipzig in the winter term 2016/17. The following students contributed their observations: Adarelys Andrades, Yves Annanias, Marius Brunnert, Alexander Engler, Maik Fröbe, Christian Heide, Felix Helfer, Ulrike Klotz, Stefan Krämer, Sebastian Luhnburg, Florian Mäschle, Markus Michaelis, Michael Rode, Jeremias Schebera, Alexander Scholz, Stephan Thönes, Kathleen Wende, Marcel Winter, Jan Witte, Sophie Wolf, Anastasia Wolschewski.
Part of this work was funded by the German Federal Ministry of Education and Research within the project Competence Center for Scalable Data Services and Solutions (ScaDS) Dresden/Leipzig (BMBF 01IS14014B) and the Deutsche Forschungsgemeinschaft (DFG STA 850/19-2 within SPP 1738).

## References

1. Al Ait, L., Yamak, Z., Morgenstern, B.: DIALIGN at GOBICS – multiple sequence alignment using various sources of external information. Nucleic Acids Res. **41**, W3–W7 (2013)
2. Bandelt, H.J., Dress, A.W.M.: A canonical decomposition theory for metrics on a finite set. Adv. math. **92**, 47 (1992)
3. Bellman, R.: Dynamic programming treatment of the travelling salesman problem. J. ACM **9**, 61–63 (1962)
4. Bryant, D., Moulton, V., Spillner, A.: NeighborNet: An agglomerative method for the construction of planar phylogenetic networks. Mol. Biol. Evol. **21**, 255–265 (2004)
5. Bryant, D., Moulton, V., Spillner, A.: Consistency of the NeighborNet algorithm. Alg. Mol. Biol. **2**, 8 (2007)
6. Buneman, P.: A note on the metric property of trees. J. Combin. Theory Ser. B **17**, 48–50 (1974)
7. Carson, A.R., Scherer, S.W.: Identifying concerted evolution and gene conversion in mammalian gene pairs lasting over 100 million years. BMC Evol Biol **9**, 156 (2009)
8. Chang, C.L., Semyonov, J., Cheng, P.J., Huang, S.Y., Park, J.I., Tsai, H.J., Lin, C.Y., Grützner, F., Soong, Y.K., Cai, J.J., et al.: Widespread divergence of the CEACAM/PSG genes in vertebrates and humans suggests sensitivity to selection. PloS one **8**, e61,701 (2013)
9. Chepoi, V., Fichet, B.: A note on circular decomposable metrics. Geometriae Dedicata **69**, 237–240 (1998)
10. Chor, B., Sudan, M.: A geometric approach to betweenness. SIAM J Discr Math **11**, 511–523 (1998)
11. 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)

12. 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)
13. Cunningham, P.: Free trees and bidirectional trees as representations of psychological distance. J. Math. Psych. **17**, 165–188 (1978)
14. Diday, E.: Orders and overlapping clusters in pyramids. In: J. De Leeuw, W.J. Heiser, J.J. Meulman, F. Critchley (eds.) Multidimensional Data Analysis, pp. 201–234. DSWO Press, Leiden, NL (1986)
15. Dobson, A.J.: Unrooted trees for numerical taxonomy. J. Appl. Probab. **11**, 32–42 (1974)
16. Dress, A.W., Flamm, C., Fritzsch, G., Grünewald, S., Kruspe, M., Prohaska, S.J., Stadler, P.F.: Noisy: identification of problematic columns in multiple sequence alignments. Alg. Mol. Biol. **3**, 7 (2008)
17. Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.l., Postlethwait, J.: Preservation of duplicate genes by complementary, degenerative mutations. Genetics **151**, 1531–1545 (1999)
18. Garcia-Fernàndez, J.: The genesis and evolution of homeobox gene clusters. Nature Rev Genet **6**, 881–892 (2005). DOI 10.1038/nrg1723
19. Gehring, W.J.: Master Controle Genes in Development and Evolution: The Homeobox Story. Yale University Press, New Haven and London (1998)
20. Halin, R.: Studies on minimally $n$-connected graphs. In: Combinatorial Mathematics and its Applications, pp. 129–136. Academic Press, London, UK (1971)
21. Hardison, R., Slightom, J.L., Gumucio, D.L., Goodman, M., Stojanovic, N., Miller, W.: Locus control regions of mammalian $\beta$-globin gene clusters: Combining phylogenetic analyses and experimental results to gain functional insights. Gene **205**, 73–94 (1997)
22. Jukes, T.H., Cantor, C.R.: Evolution of protein molecules. In: H.N. Munro (ed.) Mammalian Protein Metabolism, pp. 21–132. Academic Press, New York (1969)
23. Kalmanson, K.: Edgeconvex circuits and the traveling salesman problem. Canadian J. Math. **27**, 1000–1010 (1975)
24. Kleinman, A., Harel, M., Pachter, L.: Affine and projective tree metric theorems. Ann. Comb. **17**, 205–228 (2013)
25. Korostensky, C., Gonnet, G.: Using traveling salesman problem algorithms for evolutionary tree construction. Bioinformatics **16**, 619–627 (2000)
26. Levy, D., Pachter, L.: The neighbor-net algorithm. Adv. Appl. Math. **47**, 240–258 (2011)
27. Liiv, I.: Seriation and matrix reordering methods: An historical overview. Statistical Analysis & Data Mining **3**, 70–91 (2010)
28. MacLean, J.A., Lorenzetti, D., Hu, Z., Salerno, W.J., Miller, J., Wilkinson, M.F.: Rhox homeobox gene cluster: recent duplication of three family members. Genesis **44**, 122–129 (2006)
29. MacLean 2nd, J.A., Wilkinson, M.F.: The Rhox genes. Reproduction **140**, 195–213 (2010)
30. Maniatis, T., Fritsch, E.F., Lauer, J., Lawn, R.M.: The molecular genetics of human hemoglobins. Annual Rev. Genetics **14**, 145–178 (1980)
31. Meggido, N.: Partial and complete cyclic orders. Bull. Am. Math. Soc. **82**, 274–276 (1976)
32. Montavon, T., Duboule, D.: Chromatin organization and global regulation of Hox gene clusters. Phil. Trans. R. Soc. B **368**(1620), 20120,367 (2013)
33. 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)
34. Nieselt-Struwe, K.: Graphs in sequence spaces: a review of statistical geometry. Biophys. Chem. **66**, 111–131 (1997)
35. Noonan, J.P., Grimwood, J., Schmutz, J., Dickson, M., Myers, R.M.: Gene conversion and the evolution of protocadherin gene cluster diversity. Genome Res **14**, 354–366 (2004)
36. Notredame, C., Higgins, D.G., Heringa, J.: T-Coffee: A novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. **302**, 205–217 (2000)
37. Novák, V.: Cuts in cyclically ordered sets. Czech. Math. J. **34**, 322–333 (1984)
38. Ohno, S.: Evolution by Gene Duplication. Springer-Verlag, Berlin, Heidelberg (1970)

39. Oota, H., Dunn, C.W., Speed, W.C., Pakstis, A.J., Palmatier, M.A., Kidd, J.R., Kidd, K.K.: Conservative evolution in duplicated genes of the primate class I ADH cluster. Gene **392**(1), 64–76 (2007)
40. Opatrny, J.: Total ordering problem. SIAM J Computing **8**, 111–114 (1979)
41. Pascual-Anaya, J., Adachi, N., Álvarez, S., Kuratani, S., DAniello, S., Garcia-Fernàndez, J.: Broken colinearity of the amphioxus Hox cluster. EvoDevo **3**, 28 (2012)
42. Pascual-Anaya, J., DAniello, S., Kuratani, S., Garcia-Fernàndez, J.: Evolution of *Hox* gene clusters in deuterostomes. BMC Developmental Biology **13**, 26 (2013)
43. Préa, P., Fortin, D.: An optimal algorithm to recognize Robinsonian dissimilarities. J. Classification **31**, 1–35 (2014)
44. Robinson, W.S.: A method for chronologically ordering archaeological deposits. Amer. Antiquity **16**, 293–301 (1951)
45. Höner zu Siederdissen, C., Prohaska, S.J., Stadler, P.F.: Dynamic programming for set data types. In: S. Campos (ed.) Advances in Bioinformatics and Computational Biology: BSB 2014, *Lect. Notes Comp. Sci.*, vol. 8826, pp. 57–64 (2014)
46. Höner zu Siederdissen, C., Prohaska, S.J., Stadler, P.F.: Algebraic dynamic programming over general data structures. BMC Bioinformatics **16**, 19:S2 (2015)
47. Simões-Pereira, J.M.S.: A note on the tree realizability of a distance matrix. J. Combin. Theory **6**, 303–310 (1969)
48. Zid, M., Drouin, G.: Gene conversions are under purifying selection in the carcinoembryonic antigen immunoglobulin gene families of primates. Genomics **102**(4), 301–309 (2013)

## Predecessor Relation of Crossover Events

Alg. 3 utilizes the associative properties of gene identifiers that allow constant time mapping between pairs of genes and recombinant triples. If triples are derived from a linear type R matrix, they form a natural binary tree that is to be established by the algorithm, where each triple $(x, y : z)$ can be a left or right successor to another triple on $(x, y)$, accept a left successor on $(x, z)$, or accept a right successor on $(z, y)$.

Each triple is added in turn, checking for connections to already added triples using associative arrays (map) for each connection type. If a connected triple was already added, an open entry is found in the corresponding map, else a new entry will be added to the according inverse map. For instance, an added left successor needs to look at an open predecessor. If a single tree was created, the linear order of genes can be found by traversing the tree. If no linear order exists, multiple trees will be created, as necessary connectors are either never added to an open map or have been removed since entries in open maps are only used for a single connection.

**Algorithm 3** Establishes the successor/predecessor relation of triples

**Require:** set $T = \{t_1, .., t_n\}$ of triples in the form $t_i = (x_i, y_i : z_i)$
  Initialize Map open_predecessor
  Initialize Map open_left_successor
  Initialize Map open_right_successor
  **for** $t_i = (x_i, y_i : z_i) \in T$ **do**
    **if** $x_i z_i$ in open_predecessor **then**
      $t_i$.left_child $\leftarrow$ open_predecessor[$x_i z_i$]
      remove open_predecessor[$x_i z_i$]
    **else**
      open_left_successor[$x_i z_i$] $\leftarrow t_i$
    **end if**
    **if** $z_i y_i$ in open_predecessor **then**
      $t_i$.right_child $\leftarrow$ open_predecessor[$z_i y_i$]
      remove open_predecessor[$z_i y_i$]
    **else**
      open_right_successor[$z_i y_i$] $\leftarrow t_i$
    **end if**
    **if** $x_i y_i$ in open_left_successor **then**
      open_left_successor[$x_i y_i$].left_child $\leftarrow t_i$
      remove open_left_successor[$x_i y_i$]
    **else if** $x_i y_i$ in open_right_successor **then**
      open_right_successor[$x_i y_i$].right_child $\leftarrow t_i$
      remove open_right_successor[$x_i y_i$]
    **else**
      open_predecessor[$x_i z_i$] $\leftarrow t_i$
    **end if**
  **end for**
  traverse tree for order