On Weighting Schemes for Gene Order Analysis

Matthias Bernt, Nicolas Wieseke, and Martin Middendorf

Parallel Computing and Complex Systems Group, Institute of Computer Science
University Leipzig, Germany
{bernt,wieseke,middendorf}@informatik.uni-leipzig.de

Abstract

Gene order analysis aims at extracting phylogenetic information from the comparison of the order and orientation of the genes on the genomes of different species. This can be achieved by computing parsimonious rearrangement scenarios, i.e. to determine a sequence of rearrangements events that transforms one given gene order into another such that the sum of weights of the included rearrangement events is minimal. In this sequence only certain types of rearrangements, given by the rearrangement model, are admissible and weights are assigned with respect to the rearrangement type. The choice of a suitable rearrangement model and corresponding weights for the included rearrangement types is important for the meaningful reconstruction. So far the analysis of weighting schemes for gene order analysis has not been considered sufficiently. In this paper weighting schemes for gene order analysis are considered for two rearrangement models: 1) inversions, transpositions, and inverse transpositions; 2) inversions, block interchanges, and inverse transpositions. For both rearrangement models we determined properties of the weighting functions that exclude certain types of rearrangements from parsimonious rearrangement scenarios.

1998 ACM Subject Classification F.2.2 Nonnumerical Algorithms and Problems, J.3 Life and Medical Sciences

Keywords and phrases Gene order analysis, maximum parsimony, weighting

Digital Object Identifier 10.4230/OASIcs.GCB.2013.14

1 Introduction

The order of the genes on the chromosomes has changed during evolution by different types of rearrangement operations. For unichromosomal genomes, like most bacterial and mitochondrial genomes, inversions, transpositions, inverse transpositions, and tandem duplication random loss operations modified the order and/or orientation of the genes. In addition deletions, duplications, and horizontal transfer changed the gene content. Multichromosomal genomes, e.g. nuclear genomes, have been subject to additional interchromosomal rearrangement operations (e.g. fission, fusion, translocation, and chromosome duplication).

Gene order data has become an important source of phylogenetic information over the last two decades [14, 22]. The phylogenetic information contained in gene orders can be extracted with methods based on the maximum maximum parsimony principle, i.e. an explanation for given gene order data is sought that uses a minimal number of rearrangement operations (but see also [1]). For a pair of gene orders such an explanation is given by a shortest sequence of rearrangements that transforms one of the given gene orders into the other. If more than two gene orders are given a phylogenetic tree with the given gene orders at the leaves and a minimum number of rearrangements along the edges of the tree, such that a gene order at the root is transformed into the leaf gene orders, serves as an explanation of the data.

Algorithms for pairwise gene order analysis have been studied extensively for separate rearrangement operations. Efficient algorithms for the case that only inversions are considered

German Conference on Bioinformatics 2013 (GCB'13).
Editors: T. Beißbarth, M. Kollmar, A. Leha, B. Morgenstern, A.-K. Schultz, S. Waack, E. Wingender; pp. 14–23
OpenAccess Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

are known [15]. In contrast the problem is NP-hard if only transpositions are allowed [9]. But, in order to get reliable reconstructions all rearrangement operations that played a role during evolution need to be considered. Ideally, different weights should be used for the different types of rearrangements reflecting their importance in gene order evolution.

There are only a few algorithms for weighted gene order analysis considering more than one type of rearrangement operation. `DERANGE` [8, 21] is one of the first algorithms for gene order analysis. It is a 3-approximation algorithm that considers inversions, transpositions, and inverse transpositions. The algorithm explores possible rearrangement scenarios greedily by minimizing the number of breakpoints. Both, the possibility to weight the different types of rearrangements and to weight the operations by the number of affected genes, have been implemented. It was shown empirically that the reconstructions obtained with `DERANGE` are strongly influenced by the choice of the weights [8], see also [12]. Several improved approximation algorithms for this set of rearrangement operations have been introduced [4, 13, 16, 17]. `DERANGE` and the other algorithms assume the same weight for the two types of transpositions. The algorithm from [4] allows a weight of 1 for inversions and a weight in $\in [1:2]$ for transpositions. Other algorithms allow only fixed weights (1:2 [13] or 1:1 [16, 17]). With `CREx` an efficient heuristic for unweighted rearrangement analysis is available that incorporates tandem duplication random loss as a fourth type of rearrangement operation [6]. Also for multiple genome rearrangement analysis a heuristic, that is based on [4], is available that incorporates weights [3].

Block interchange is a generalization of the transposition operation. Since pairwise gene order analysis for this operation can be solved efficiently [10] this operation has become an interesting alternative for transpositions. It is an integral part (with inversions and translocations) of the double cut and join frame work [23] that became a highly active research area in the last years. The `SPRING` software [19] and the approach described in [20] allow for reconstructing pairwise rearrangement scenarios based on inversions and block interchanges using a corresponding weight ratio of 1:2. Approximation algorithms for other weighting schemes have been devised in [18]. By allowing for the additional operations tandem duplication and deletion the heuristic presented in [2] can also compute rearrangement scenarios consisting of inversions and block interchanges for gene orders with unequal gene content using a weight of 1 for inversions and 2 for each of the other operations.

All approaches mentioned above assign a (usually two times) larger weight to transpositions or block interchanges than to inversions. This is justified by the larger number of inversions than transpositions that can be observed for several biological data sets. But this is not the case for all data sets, e.g. for metazoan mitogenomes inversions seem to account for only a small proportion of the rearrangements [7].

Besides the possibility to weight rearrangements by the type of the rearrangement also weighting by the length of the affected segments (e.g. [5, 8, 12]), the types of the affected genes, or by other factors that determine the likelihood of a rearrangement (e.g. transcript structures) might be incorporated. Certain constraints as an extreme case can be introduced by allowing for infinite weights which excludes certain types of rearrangements. `CREx` for instance forbids rearrangement that destroy conserved gene clusters [6].

Weighting schemes for genome rearrangement analysis have not been considered in sufficient detail in the literature [14]. Here we analyze weighting schemes for two rearrangement models: (1) Inversions, transpositions, and inverse transpositions (Section 3); (2) Inversions, block interchanges, an inverse transpositions (Section 4). For both cases we derive properties of weighting schemes that exclude one/more of the rearrangement operations from any parsimonious reconstruction of genome rearrangement evolution.

2 Basic Definitions

In the context of this work a gene order is regarded as a *signed permutation* $\pi = (\pi(1), \ldots, \pi(n))$, which is a permutation of the elements $\{1, \ldots, n\}$ where each element has an additional orientation, denoted by a "+" or "−" sign. Each element of a gene order represents one genetic marker, e.g. a gene, and the sign its strandedness. If not stated otherwise, we assume a signed permutation to be directed, i.e. $\pi \neq -\pi = (-\pi(n), \ldots, -\pi(1))$. An *interval* X of a permutation π is a non-empty subset of (unsigned) elements $X \subseteq \{1, \ldots, n\}$ which are consecutive with respect to π, i.e. $\exists i, j \in [1 : n] : X = \{|\pi(k)| : i \leq k \leq j\}$. $\mathfrak{I}(\pi)$ gives the set of all possible intervals of a permutation π. A *rearrangement* ρ is an operation applied to a signed permutation π that changes the position and/or orientation of some of the elements resulting in a new signed permutation denoted as $\pi \circ \rho$. For two rearrangements ρ and ρ', with $\rho \neq \rho'$, and a gene order π it holds that $\pi \circ \rho \neq \pi \circ \rho'$. Let $\mathfrak{R}$ be the set of all $n!2^n$ different rearrangement operations.

Let $w : \mathfrak{R} \to \mathbb{R}_{>0}$ be a weighting function for rearrangement operations. A classification into rearrangement *types* $\mathfrak{T} = \{T_1, \ldots, T_k, T_\epsilon\}$ is a partition of $\mathfrak{R}$ into distinct sets of rearrangements with all $\rho \in T_j$ having the same weight $w(\rho) = w_{T_j}$. The set T_ϵ refers to rearrangements with a weight of $w_{T_\epsilon} = \infty$ and is used for rearrangement operations that are not regarded. The set of valid rearrangement types is denoted as $\mathfrak{T}_{|T_\epsilon} = \mathfrak{T} \setminus T_\epsilon$. A *rearrangement scenario* for a permutation π is a sequence of rearrangements $S = (\rho_1, \ldots, \rho_l)$ such that $\pi \circ \rho_1 \circ \ldots \circ \rho_l = \iota$ and $\forall : i \in [1 : l] : \rho_i \notin T_\epsilon$. The weight of a scenario S is given by $w(S) = \sum_{i=1}^{|S|} w(\rho_i)$. A scenario for π with minimal weight is called *parsimonious*. If not stated otherwise, we consider normalized weights for the admissible rearrangement types, i.e. the weight for one type of rearrangement operation is divided by the sum of the weights of all allowed rearrangement types.

Here we consider the rearrangement operations inversions (I), transpositions (T), inverse transpositions (iT), and block interchanges (BI). In Section 3 the set of valid rearrangement types $\mathfrak{T}_{|T_\epsilon} = \{I, T, iT\}$, with weights w_I, w_T, and w_{iT}, respectively, is considered. In Section 4 block interchange with weight w_{BI} is considered instead of transpositions. Each rearrangement is specified by the intervals it affects. An *inversion* ρ_I on a signed permutation π is defined by the interval $A \in \mathfrak{I}(\pi)$, where in $\pi \circ \rho_I$ the order of the elements from A is reversed and the orientation is switched. A *transposition* ρ_T is defined by two disjoint and consecutive intervals $A, B \in \mathfrak{I}(\pi)$, i.e. $A \cup B \in \mathfrak{I}(\pi), A \cap B = \emptyset$. By a transposition the position of the two intervals is switched in $\pi \circ \rho_T$. Analogous to a transposition an *inverse transposition* ρ_{iT} is also defined by two disjoint and consecutive intervals A and B. It is a combination of transposition and inversion, where after the transposition of A and B an additional inversion of A is performed. A *block interchange* ρ_{BI} is a generalization of the transposition in the way that the two intervals A and B do not have to be consecutive. There might be an interval X in between A and B such that in $\pi \circ \rho_{BI}$ A and B switch their positions with respect to X. In the following we denote the rearrangements $\rho_I, \rho_T, \rho_{iT}$, and ρ_{BI} together with the intervals they affect by $I(A), T(A, B), iT(A, B)$, and $BI(A, B)$, respectively. It holds that $T(A, B) = T(B, A)$ and $BI(A, B) = BI(B, A)$.

3 Inversions, transpositions, and inverse transpositions

First we consider the rearrangement model consisting of the following three types of rearrangements: inversions, transpositions, and inverse transpositions. There are several possibilities to mimic a single rearrangement, i.e. achieve the same effects, by combinations of rearrangements of other types. In the following we derive the different possibilities for

replacing one rearrangement of a certain type by a smallest number of rearrangements of the other type(s).

▶ **Lemma 1.** *For any rearrangement scenario the following replacements are possible within the* $\mathfrak{T}_{|T_\epsilon} = \{I, T, iT\}$ *model:*

1. *A transposition* $T(A, B)$ *can be replaced by each of the following sets of rearrangements*
   - **a.** *three inversions* $I(A)$*,* $I(B)$*, and* $I(A \cup B)$*,*
   - **b.** *one inverse transposition* $iT(A, B)$ *and one inversion* $I(A)$*, or*
   - **c.** *two inverse transpositions when the genome has at least three genes. Then at least one of the following cases holds:*

     Case i. There exists an interval X *such that* $B \cup X$ *is an interval and* $X \cap A = \emptyset = X \cap B$*:* $iT(A, B \cup X)$ *and* $iT(A, X)$*.*

     Case ii. There exists an interval X *such that* $X \cup A$ *is an interval and* $X \cap A = \emptyset = X \cap B$*:* $iT(B, X \cup A)$ *and* $iT(B, X)$

     Case iii. There exists a bipartition of A *into intervals* A_1 *and* A_2 *(i.e.* $|A| \geq 2$*) such that* $A_2 \cup B$ *is an interval:* $iT(B, A_2)$ *and* $iT(B, A_1)$

     Case iv. There is a bipartition of B *into intervals* B_1 *and* B_2 *(i.e.* $|B| \geq 2$*) such that* $B_1 \cup A$ *is an interval:* $iT(A, B_1)$ *and* $iT(A, B_2)$*.*

4. *An inverse transposition* $iT(A, B)$ *can be replaced by each of the following sets of rearrangements:*
   - **a.** *one transposition* $T(A, B)$ *and one inversion* $I(A)$ *or*
   - **b.** *two inversions* $I(A \cup B)$ *and* $I(B)$*.*
3. *An inversion* $I(A)$ *can be replaced by each of the following sets of rearrangements:*
   - **a.** *inverse transposition(s) and one transposition according to the following cases:*

     Case i. When at least one gene is not included in A*, i.e. there exists an interval* X *with* $X \cap A = \emptyset$ *and* $X \cup A$ *is an interval:* $iT(A, X)$ *and* $T(A, X)$*.*

     Case ii. A *includes the whole gene order and can be partitioned into two intervals* A_1 *and* A_2 *(i.e.* $|A| \geq 2$*):* $iT(A_2, A_1)$*,* $iT(A_1, A_2)$*,* $T(A_1, A_2)$*.*
   - **b.** *three inverse transpositions when the gene order has at least three genes. Then at least one of the following cases holds:*

     Case i. There exists a partition of A *into three intervals* A_1*,* A_2*, and* A_3 *(i.e.* $|A| \geq 3$*) such that* $A_1 \cup A_2$ *and* $A_2 \cup A_3$ *are intervals:* $iT(A_1, A_2)$*,* $iT(A_3, A_1)$*, and* $iT(A_2, A_3)$*.*

     Case ii. There exists a bipartition of A *into two intervals* A_1 *and* A_2 *(i.e.* $|A| \geq 2$*) and there exists an interval* X *with* $A \cap X = \emptyset$ *such that* $A_2 \cup X$ *is an interval:* $iT(A_2 \cup X, A_1)$*,* $iT(X, A_2)$*, and* $iT(A_1, X)$*.*

     Case iii. There exists an interval X *that can be partitioned into intervals* X_1 *and* X_2 *(i.e.* $|X| \geq 2$*) with* $A \cap X = \emptyset$ *such that* AX_1 *is an interval:* $iT(A, X_1)$*,* $iT(X_1, A \cup X_2)$*, and* $iT(X_1, X_2)$*.*

     Case iv. There exist two disjoint intervals X_1 *and* X_2 *with* $X_1 \cap A = \emptyset = X_2 \cap A$ *such that* $X_1 \cup A$ *and* $A \cup X_2$ *are intervals:* $iT(A, X_2)$*,* $iT(X_2, X_1)$*, and* $iT(X_2, X_1 \cup A)$*.*

▶ **Lemma 2.** *Lemma 1 lists all possibilities (with respect to number and type of the rearrangement operations) to replace a single rearrangement of a certain type* $\in \{T, iT, I\}$ *by a smallest number of rearrangements of one or two of the other types of rearrangements* $\in \{T, iT, I\}$*.*

Proof. By definition of the rearrangement operations it is not possible to replace a single operation of one type by any single operation of another type. Observe also that a transposition $T(A, B)$ cannot be replaced by any number of inverse transpositions when the genome has only two genes, i.e. A and B contain only a single gene and there exists no other gene

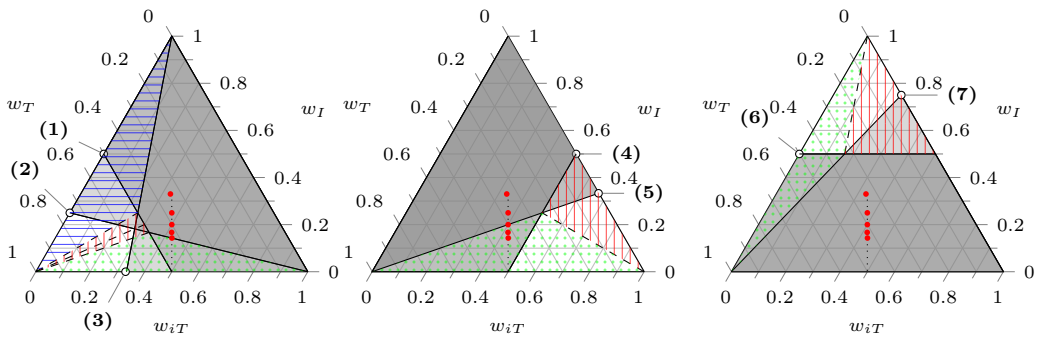

Figure 1 Barycentric plot showing the weighting schemes where transpositions (left), inverse transpositions (middle), and inversions (right) need to be considered; shaded areas indicate for each of the inequalities the valid weighting schemes; darker shading indicates the area where all inequalities hold; the limiting cases, i.e. equality, is annotated by the corresponding equation number given in bold text; dashed lines give the demarcation line between each pair of the alternatives, colored dashed or dotted areas indicate which of the alternatives needs to be considered: blue horizontal lines 2iT, red vertical lines iT+ I, green dots 3I (left) red vertical lines I+T, green dots 2I (middle) red vertical lines T+iT, green dots 3iT (right); the dotted line and red dots indicate the weighting schemes considered in [8]; note that the borders of the plot, i.e. weights of 0, are excluded.

in the genome. Moreover, it is easy to see that a transposition cannot be replaced by two inversions. Hence, the lemma follows with respect to the replacement of transpositions.

Now observe, that any combination of transpositions can neither replace one inversion nor one inverse transposition. This is because both an inversion and an inverse transposition change the sign of at least one gene but a transposition cannot change the sign of a gene. It can also be seen that an inversion cannot be replaced by inverse transpositions when the genome has only one or two genes. It is also not hard to see that an inversion cannot be replaced by two or less inverse transpositions. An inversion of the whole genome cannot be replaced by one inverse transposition plus any number of transpositions. Hence, it follows that for this case at least two inverse transposition plus one transposition are necessary. When the genome has only one gene it is not possible to replace an inversion by any number of inverse transpositions plus transpositions. Thus, the lemma holds also with respect to the replacement of inverse transpositions and inversions. ◀

In the following we will only consider gene orders consisting of at least three elements. Furthermore we exclude the case of the inversion of the complete gene order, i.e. Case 3.b.ii. The seven replacement possibilities that are listed in Lemma 1 imply certain properties that a weighting function for the different types of rearrangement operations has to satisfy in order to make the corresponding rearrangement operation possible for a parsimonious scenario. These properties can be formulated in the form of inequalities between the different weights. A graphical representation of the inequalities and their consequences is given in Fig. 1. The seven inequalities implied by Lemma 1 are:

$$w_T \leq w_I + w_{iT} \quad (1) \qquad w_{iT} \leq w_I + w_T \quad (4) \qquad w_I \leq w_T + w_{iT} \quad (6)$$

$$w_T \leq 3w_I \quad (2) \qquad w_{iT} \leq 2w_I \quad (5)$$

$$w_T \leq 2w_{iT} \quad (3) \qquad w_I \leq 3w_{iT} \quad (7)$$

Each of these inequalities decides if a single rearrangement of a certain type or an alternative more complex, i.e. longer, rearrangement scenario of the other types of rearrangements is parsimonious. The respective single rearrangement operation can occur in a parsimonious scenario only if all of the corresponding inequalities are satisfied, i.e. (1) to (3) for transpositions, (4) and (5) for inverse transpositions, and (6) and (7) for inversions (unless other restrictions exclude one of the corresponding replacement scenarios). If one of these inequalities is violated the corresponding alternative is more parsimonious. For example, when inequality (4) is not satisfied an inverse transposition cannot occur in any parsimonious scenario (unless other restrictions exclude an inversion or a transposition).

In case that a rearrangement operation of a certain type cannot occur in a parsimonious scenario not all of the replacements that are listed in Lemma 1 might be possible in a parsimonious scenario. In the following this aspect is discussed in more detail.

There are three alternatives for a transposition: $iT + I$, $3I$, and $2iT$ with associated weights: $w_{iT} + w_I$, $3w_I$, and $2w_{iT}$. Thus, one can decide between the three alternatives by comparing their weights.

- $iT+I$ needs to be considered only if $w_{iT}+w_I \leq 3w_I$ ($\Leftrightarrow w_{iT} \leq 2w_I$) and $w_{iT}+w_I \leq 2w_{iT}$ ($\Leftrightarrow w_I \leq w_{iT}$).
- $3I$ is a feasible alternative only if $3w_I \leq w_{iT} + w_I$ ($\Leftrightarrow w_{iT} \geq 2w_I$) and $3w_I \leq 2w_{iT}$.
- $2iT$ needs to be considered as alternative only if $2w_{iT} \leq w_{iT} + w_I$ ($\Leftrightarrow w_{iT} \leq w_I$) and $2w_{iT} \leq 3w_I$.

This set of inequalities "partitions" the set of all weighting schemes where transpositions are not parsimonious between the three alternatives (in case of equal weights two or three of the alternatives might be possible). The different sets of the partition are shown as differently patterned areas in Fig. 1.

For the weighting schemes where inverse transpositions are not parsimonious there are the two replacements $2I$ and $I + T$. The former is possible only if $2w_I \leq w_I + w_T$ ($\Leftrightarrow w_I \leq w_T$) holds. Similarly the alternatives for the case that an inversion is not parsimonious, i.e. $T+iT$ and $3iT$, can be chosen on the basis of the comparison $w_T + w_{iT} \leq 3w_{iT}$ ($\Leftrightarrow w_T \leq 2w_{iT}$). As presented in Fig. 1 the remaining weighting schemes are partitioned between the alternatives. It can be readily verified that for each alternative scenario the corresponding necessary rearrangement operations are itself not excluded by any of the other inequalities.

Weighting schemes for the rearrangement model with transpositions, inverse transpositions and inversions as rearrangement operations and the implications of the choice of the weights on the reconstructions that can be obtained with a greedy heuristic that was called `DERANGE` have been discussed in [8]. The weights that have been analyzed in greater detail assumed a fixed weight for inversions and equal weights for transpositions and inverse transpositions that are at least as large as the weight for inversions. In particular, the following (unnormalized) weights have been considered: $w_I = 1$ and weights for transpositions and inverse transpositions that are "somewhat less to somewhat more than" 2, or more exactly $w_T = w_{iT} \in [1 : 3]$. In terms of normalized weights the corresponding set of weights corresponds to a (half open) line in the barycentric plot between the center point of the plot at $w_I = w_T = w_{iT} = 1/3$ and the middle point of the bottom line (i.e. $w_I = 0$) but excluding the middle point itself. This line and five selected points (corresponding to unnormalized inversion weight of 1 and (inverse) transposition weights $1, 1.5, 2, 2.5$, and 3) of it are shown in Fig. 1.

For the considered weighting schemes a "phase transition" for the length of the reconstructions, i.e. the number of rearrangements, made with heuristic `DERANGE` was observed at approximately $w_T = w_{iT} = 2$ [8]. The corresponding strong increase of the reconstruction lengths was observed for random data as well as real, i.e. mitochondrial and bacterial, gene

orders. This effect was attributed to the greedy nature of the algorithm that tries to find a move x, with weight w_x removing B_x breakpoints, minimizing $w_x - B_x$ and the observation that B_x is nearly always $1:2$ for inversions vs. (inverse) transpositions (and not the optimal case $2:3$). This leads to the preference of (inverse) transpositions for $w_{iT} = w_T < 2$ and of inversions for $w_T > 2$ inversions. Since more inversions are necessary to remove all breakpoints the rearrangement scenarios are longer in the latter case.

In the light of the analysis presented here we can add a further explanation for the observed "phase transition". Exactly for the unnormalized (inverse) transposition weight of 2 an inverse transposition has equal weight as the alternative consisting of two inversions. For a weight larger than two inverse transpositions cannot be in a parsimonious rearrangement scenario but they must be replaced by two inversions, i.e. twice the number of rearrangements. The other way round inverse transpositions cannot be replaced by this alternative for weights smaller than two. Based on the empirical results Blanchette et al. [8] suggested to that an (inverse) transposition weight of "slightly greater than 2 may be an appropriate value". Our analysis shows that this is not maintainable for any (optimal/suboptimal) solution, since in such a weighting inverse transpositions are excluded as they need to be replaced by the more parsimonious alternative consisting of two inversions. Another "phase transition" should occur for the reconstructions made with `DERANGE` (and must occur for optimal reconstruction) for unweighted (inverse) transposition weights > 3 which makes inversions the only type of rearrangements that can occur in parsimonious rearrangement scenarios.

4 Inversion, inverse transposition, and block interchange

In this section we study the rearrangement model consisting of inversions, inverse transpositions, and block interchanges, i.e. $\mathfrak{T}_{|T_\epsilon} = \{I, BI, iT\}$. It is assumed here that transpositions are a special case of block interchanges. A block interchange $BI(A, B)$ is called *proper* when there exists an interval $X \neq \emptyset$ such that $X \cap A = \emptyset = X \cap B$ and $A \cup X$ and $X \cup B$ form intervals. X is called the *intermediate interval*.

It is clear that for any rearrangement scenario all the replacements that are listed in Lemma 1 also hold for the rearrangement model $\mathfrak{T}_{|T_\epsilon} = \{I, BI, iT\}$ when a transposition $T(A, B)$ is exchanged by a (non-proper) block interchange $BI(A, B)$. In addition the replacements listed in Lemma 3 are relevant for the $\mathfrak{T}_{|T_\epsilon} = \{I, BI, iT\}$ model.

▶ **Lemma 3.** *For any rearrangement scenario the following replacements are possible within the $\mathfrak{T}_{|T_\epsilon} = \{I, BI, iT\}$ model:*

1. *A proper block interchange $BI(A, B)$ with intermediate interval X can be replaced by each of the following sets of rearrangements*
   - **a.** *three inversions: $I(A \cup X)$, $I(A \cup B)$, $I(B \cup X)$,*
   - **b.** *three inverse transpositions: $iT(A \cup X, B)$, $iT(X, A)$, $iT(A, X)$*
   - **c.** *one inverse transposition and two inversions: $iT(A \cup X, B)$, $I(X)$, $I(A)$, or*
   - **d.** *two inverse transpositions and one inversion: $iT(X, B)$, $iT(A, B)$, $I(A \cup X)$.*

▶ **Lemma 4.** *Lemmas 1 and 3 list all possibilities (wrt. number and type of rearrangement operations) to replace a single rearrangement of a certain type $\in \{BI, iT, I\}$ by a smallest number of rearrangements of one or two of the other types of rearrangements $\in \{BI, iT, I\}$.*

Proof. It is clear that for all replacements listed in Lemma 1 the use of a proper block interchange instead of transpositions can not lead to a shorter replacement. Hence, by Lemma 2 it follows that the result holds for all replacements listed in Lemma 1. It remains to consider replacements for a proper block interchange $BI(A, B)$ as considered in Lemma 3.

Since inversions and inverse transpositions change the sign of at least one gene it is clear that $BI(A, B)$ can neither be replaced by a single inversion nor by a single inverse transposition. Assume that it is possible to replace $BI(A, B)$ by two rearrangements from $\{I, iT\}$. Then the sign changes that are made by the first of the two operations has to be reversed by the second operation (and no other sign changes can be made by the second operation). Hence, the interval with the sign changes has to be the same for both operations. Then it is not possible that both operations are inversions (since then one inversion simply reverses the effect of the other inversion). It can also not be the case that one or both rearrangements are inverse transpositions. This is due to the fact that the interval which is inverted is the same for both operations which implies that their effect is equal to the effect of one transposition. ◀

Interestingly block interchanges and transpositions can be replaced with the same number of inversions, i.e. three, but a larger number of rearrangements is necessary if inverse transpositions or mixed rearrangement types are involved. Another difference of block interchanges to transpositions, as discussed in Section 3, is that there are two alternatives consisting of inversions and inverse transpositions.

The replacements given above are captured by a set of inequalities that need to be satisfied if a certain type of rearrangements can be part of a parsimonious rearrangement scenario. Since the replacements for inverse transpositions and inversions are the same as for $\mathfrak{T}_{|T_\epsilon} = \{I, T, iT\}$ also the corresponding inequalities and properties of the weighting schemes are the same when replacing T by BI. Thus, in the following only the case of the block interchange is discussed. Equations (8) to (11) describe the relations of the weights that render block interchanges impossible if one of them is violated. A visual representation is shown in Fig. 2. Note that, a transposition (as a special block interchange) can be replaced by two (instead of three) inverse transpositions and one inversion and inverse transposition (instead of two for one of the rearrangement types). These replacements would induce tighter bounds on the weights for block interchanges. But since these replacements are not possible for proper block interchanges these tighter bounds cannot be applied in general.

$$w_{BI} \leq w_I + 2w_{iT} \quad (8)$$

$$w_{BI} \leq 2w_I + w_{iT} \quad (9)$$

$$w_{BI} \leq 3w_I \quad (10)$$

$$w_{BI} \leq 3w_{iT} \quad (11)$$

For weighting schemes where block interchanges are not possible one or more of the replacement scenarios must be employed. For each of the six pairs of replacements the parsimonious replacement is determined by comparing w_I and w_{iT}. Weighing schemes where these two weights are equal are indicated by a dashed line in Fig. 2. Weighing schemes on this line are a "no man's land" where all four replacements have equal weight. Above this line the replacement by three inverse transpositions and below this line the replacement by three inversions is parsimonious. Interestingly the replacement consisting of inversions and inverse transpositions is not parsimonious in these areas but only on the "no man's land". For $w_I = 0.2, w_{BI} = 0.6, w_{iT} = 0.2$ where all lines in the plot intersect, block interchanges and all of its replacements that are listed in Lemma 3 have equal weights.

5 Conclusion and Discussion

In this paper weighting schemes for two rearrangement models have been analyzed formally: 1) inversions, transpositions, and inverse transpositions; 2) inversions, block interchanges, and inverse transpositions. These rearrangement models are important for the analysis of unichromosomal genomes with equal gene content. For both models inequalities have been

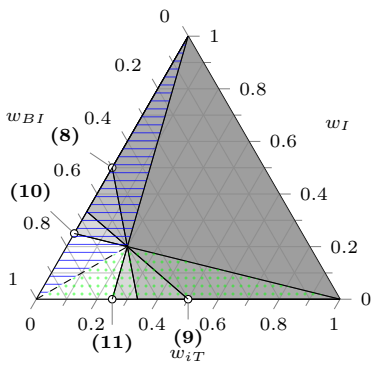

Figure 2 Barycentric plot showing the weighting schemes where block interchanges need to be considered; shaded areas indicate for each of the inequalities the valid weighting schemes; darker shading indicates the area where all inequalities hold; the limiting cases, i.e. equality, is annotated by the corresponding equation number given in bold text; the dashed line gives the demarcation line between the alternative 3I (green dots) and 3iT (blue horizontal lines).

derived that describe weighting schemes for which certain rearrangement types are excluded from parsimonious scenarios. This has been done by analyzing the possibilities to achieve the effects of one rearrangement type by rearrangements of one or more other type(s).

The choice of appropriate weights is an open problem. But, if estimates for the frequency of the different rearrangement operation are available, e.g. from large scale pairwise comparisons [7], it seems to be intuitive to use weights that are inversely (e.g. reciprocal or antiproportional) related to the frequencies. In fact, this is often done to justify chosen weights, e.g. [8]. But then, our results imply hard bounds for the reconstructibility of genome rearrangements by parsimony. If, for instance, inversions are more than three times as frequent as (inverse) transpositions the corresponding reciprocal (unnormalized) weighting scheme ($w_I = 1$ and $w_T, w_{iT} > 3$) forbids an exact reconstruction by parsimony since transpositions and inverse transpositions can not be included in any optimal solution for weighted genome rearrangement problems. These hard bounds might be loosened by using other inverse functional relations of frequency and weight, e.g. by adjusting the factor in case of antiproportionality. Introducing constraints enforcing certain proportions of the frequencies of rearrangement types are another option.

Considering rearrangement models for the case of undirected gene orders (i.e. $\pi = -\pi$), distinguishing between transpositions and proper block interchanges, or incorporating multichromosomal rearrangements (e.g. [11, 23]) is future work.

References

1 Z. Adam, M. Turmel, C. Lemieux, and D. Sankoff. Common intervals and symmetric difference in a model-free phylogenomics, with an application to streptophyte evolution. *J Comput Biol*, 14(4):436–445, 2007.

2 M Bader. Sorting by reversals, block interchanges, tandem duplications, and deletions. *BMC Bioinformatics*, 10(Suppl 1):S9, 2009.

3 M. Bader, M. Abouelhoda, and E. Ohlebusch. A fast algorithm for the multiple genome rearrangement problem with weighted reversals and transpositions. *BMC Bioinformatics*, 9:516, 2008.

4 M. Bader and E. Ohlebusch. Sorting by weighted reversals, transpositions, and inverted transpositions. *J Comput Biol*, 14(5):615–636, 2007.

5 M. A. Bender, D. Ge, S. He, H. Hu, R. Y. Pinter, S. Skiena, and F. Swidan. Improved bounds on sorting by length-weighted reversals. *J Comput Syst Sci*, 74(5):744–774, 2008.

6 M. Bernt, D. Merkle, K. Ramsch, G. Fritzsch, M. Perseke, D. Bernhard, M. Schlegel, P. F. Stadler, and M. Middendorf. CREx: inferring genomic rearrangements based on common intervals. *Bioinformatics*, 23(21):2957–2958, 2007.

7 M. Bernt and M. Middendorf. A method for computing an inventory of metazoan mitochondrial gene order rearrangements. *BMC Bioinformatics*, 12(Suppl 9):S6, 2011.

8 M. Blanchette, T. Kunisawa, and D. Sankoff. Parametric genome rearrangement. *Gene*, 172(1):11–17, 1996.

9 L. Bulteau, G. Fertin, and I. Rusu. Sorting by transpositions is difficult. In *ICALP*, number 6755 in LNCS, pages 654–665, 2011.

10 D. A. Christie. Sorting permutations by block-interchanges. *Inform Process Lett*, 60(4):165–169, 1996.

11 Z. Dias and J. Meidanis. Genome rearrangements distance by fusion, fission, and transposition is easy. *SPIRE'2001*, pages 250–253, 2001.

12 N. Eriksen. Combinatorics of genome rearrangements and phylogeny, 2001.

13 N. Eriksen. $(1+\epsilon)$-approximation of sorting by reversals and transpositions. *Theor Comput Sci*, 289(1):517–529, 2002.

14 G. Fertin, A. Labarre, I. Rusu, E. Tannier, and S. Vialette. *Combinatorics of Genome Rearrangements*. MIT Press, 2009.

15 Sridhar Hannenhalli and Pavel A. Pevzner. Transforming cabbage into turnip: polynomial algorithm for sorting signed permutations by reversals. *JACM*, 46(1):1–27, 1999.

16 T. Hartman and R. Sharan. A 1.5-approximation algorithm for sorting by transpositions and transreversals. *J Comput Syst Sci*, 70(3):300–320, 2005.

17 G-H. Lin and G. Xue. Signed genome rearrangement by reversals and transpositions: Models and approximations. In *COCOON*, volume 1627 of *LNCS*, pages 71–80. 1999.

18 Y. Lin, C-Y. Lin, and C. Lin. Sorting by reversals and block-interchanges with various weight assignments. *BMC Bioinformatics*, 10(1):398, 2009.

19 Y-C. Lin, C-L. Lu, Y-C. Liu, and C-Y. Tang. Spring: a tool for the analysis of genome rearrangement using reversals and block-interchanges. *Nucleic Acids Res*, 34(suppl 2):W696–W699, 2006.

20 C. Mira and J. Meidanis. Sorting by block-interchanges and signed reversals. In *ITNG*, pages 670–676, 2007.

21 D. Sankoff. Edit distance for genome comparison based on non-local operations. In *CPM*, volume 644 of *LNCS*, pages 121–135. Springer, 1992.

22 G. A. Watterson, W. J. Ewens, T. E. Hall, and A. Morgan. The chromosome inversion problem. *J Theor Biol*, 99(1):1–7, 1982.

23 S. Yancopoulos, O. Attie, and R. Friedberg. Efficient sorting of genomic permutations by translocation, inversion and block interchange. *Bioinformatics*, 21(16):3340–3346, 2005.