Consistency of Sequence-based Gene Clusters

Roland Wittler[1,2] and Jens Stoye[1]

[1]Department of Mathematics, Simon Fraser University, Burnaby (BC), Canada
[2]Technische Fakultät, Universität Bielefeld, Germany
roland@cebitec.uni-bielefeld.de, stoye@techfak.uni-bielefeld.de

Abstract. In comparative genomics, various combinatorial models can be used to specify gene clusters — groups of genes that are co-located in a set of genomes. Several approaches have been proposed to reconstruct putative ancestral gene clusters based on the gene order of contemporary species. One prevalent and natural reconstruction criterion is *consistency*: For a set of reconstructed gene clusters, there should exist a gene order that comprises all given clusters.
In this paper, we discuss the consistency problem for different gene cluster models on sequences with restricted gene multiplicities. Our results range from linear-time algorithms for the simple model of *adjacencies* to NP completeness for more complex models like *common intervals*.

1 Introduction

The exploration of the ancestral history of different species can give valuable information about their evolution. In whole-genome comparison, one commonly considers the order of the genes or other markers within the genome to study changes and similarities in the structure of different genomes.

Genes belonging to the same gene family are represented by the same identifier. To simplify matters, the term 'gene' will be used to refer to the corresponding gene family identifier. One simple way to model genomes is to use permutations. However, this approach includes the assumption that every gene occurs exactly once in each considered genome. To allow for duplications and deletions, a relaxation to sequences of genes is necessary. A convenient way to account for the orientation of a gene within the genome is to use signed permutations or signed sequences, respectively.

Evolutionary processes can rearrange a gene order. The gene composition of some regions, however, is preserved and can be found in several related genomes. These segments, denoted as *gene clusters*, often contain functionally or evolutionarily associated genes [14, 16]. Whenever the genomes of several species comprise the same gene cluster, it was presumably inherited from a common ancestor. Recent studies [1, 2, 7, 19] build on this idea to reconstruct ancient gene clusters and to infer ancient gene orders. More precisely, the internal nodes of a given phylogenetic tree are labeled with sets of gene clusters, based on the gene orders of contemporary species at the leaves of the tree. Beside the pure identification of gene clusters, such reconstructed scenarios for the origin of the clusters and the development of the gene order can give valuable information about underlying evolutionary processes, the ancestral history of the species, and functional and evolutionary relations of genes.

Proposed reconstruction approaches differ in the underlying models for gene order and gene clusters, and in the applied methodology. However, a general aim is to ensure *consistency*: For a set of putative ancient gene clusters, there should exist at least one gene order that comprises all given clusters. Otherwise, the reconstruction result would be inconsistent with respect to the genome model.

The goal of reconstructing consistent labelings was first introduced by Bergeron *et al*. [2] who presented an algorithm that reconstructs sets of framed common intervals on permutations. Adam *et al*. [1] applied the parsimony principle as an objective function to reconstruct common intervals on permutations. A heuristic is used to reach consistency. Recently, Chauve and Tannier [7] proposed a methodology to reconstruct the gene order of the amniote genome, based on consistent labelings of common intervals and adjacencies. In our previous work [19], we introduced an algorithmical framework that is not restricted to a specific model but instead follows an oracle-based approach to compute most parsimonious consistent labelings for various models.

All of the above methods have been successfully applied to real data and proven to yield reasonable and valuable results. They all rely on permutation-based models, which enable efficient algorithms and data structures. In particular, the verification of consistency can be solved in polynomial time and space using data structures like PQ-trees or PC-trees [10]. Some reconstruction approaches could be easily adapted to the model of *sequences without duplications* which allows genes to be missing in some genomes but still requires each gene to occur at most once in each genome.

In this paper, we discuss consistency for *sequence-based* gene cluster models. Particularly, we consider the simple model of *adjacencies*, the classical model of *common intervals* [20], and two variants of the latter. For each of these models we address the problem: Given a set of gene clusters and a maximum copy number for each gene, decide whether there exists a valid gene order that contains all the clusters. Our results range from algorithms that verify consistency for adjacencies in linear time to the confirmation of NP completeness for the more complex models.

The paper has been organized in the following way. First, we formally introduce the *Consistency Problem* in Section 2. Then, in Section 3, we give an efficient solution for the gene cluster model of both signed and unsigned adjacencies. In Section 4, we present NP completeness results for the model of common intervals and its variants, before we finish with some discussions and conclusions in Section 5. The technical details of the NP completeness proofs can be found in [21, Appendix A].

2 The Consistency Problem

Assume a set of putative gene clusters, assigned to an ancestral node in a given tree. These ancient clusters in turn imply a set of putative ancient genomes: all those which contain all the given clusters. Depending on the gene cluster model used, this set of genomes can be empty if some of the clusters derived from different contemporary species are in contradiction with others. For example, when we model gene order as permutations, there is no valid gene order comprising the three adjacencies $\{a, b\}$, $\{a, c\}$ and $\{a, d\}$, because, according to the model, gene a can only occur once and thus only be neighbor of two other genes.

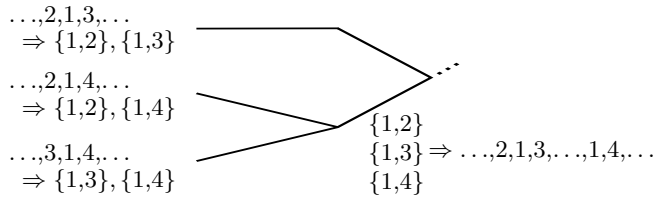

Fig. 1. Example for an artifact that arises when gene clusters are reconstructed on the basis of gene orders, where any gene can occur arbitrarily often. Any most parsimonious labeling would assign all three adjacencies to the lowermost internal node, implying at least two copies of gene 1.

In a more general case, we represent genomes as *sequences* of genes or other genomic markers. In a sequence, any element can occur multiple times or not at all, which in the context of gene order comparison corresponds to paralogous genes and gene deletions, respectively.

If we allow each gene to appear arbitrarily often in any genome, the question for consistency would become redundant: Any set of gene clusters is consistent since there is a valid gene order containing all assigned clusters. For instance, we can simply create a short sequence of genes according to each cluster separately and then concatenate these sequences to an absurd yet valid gene order. Such a construction is possible for any gene cluster model. As a consequence, consistency always holds and does not contribute to a specification of reasonable reconstruction results.

Even if we replace the naive concatenation approach and instead construct preferably compact valid gene orders, we cannot avoid to include some genes multiple times. In some cases, this causes side effects. In the example given in Figure 1, the classical parsimony principle is applied to assign gene clusters to the inner nodes of a given tree, minimizing the number of gains and losses of clusters. Although a gene is contained in all input genomes only once, it is reconstructed to occur multiple times for ancestral nodes. In this simple example, we consider a subsequence of only three genes in each input genome and obtain a segment of five genes for the examined internal node. In general, such artifacts imply unnaturally long genomes for higher levels in the tree.

To preclude this unwanted effect, we refine the concept of consistency. Instead of simply restricting the total length of a genome, we limit the multiplicity of each individual gene.

In the following problem definition, we intentionally refrain from specifying a concrete model of gene clusters and instead use the imprecise notion of a *sequence containing a cluster*. For instance, in the simple model of gene adjacencies, a sequence g *contains* a gene cluster $\{a, b\}$ if and only if the genes a and b occur adjacently in g.

Definition 1 (Consistency Problem). *Let $\mathcal{G}_N := \{1, \ldots, N\}$ be the set of genes and $m : \mathcal{G}_N \to \mathbb{N}$ assign a maximum copy number to each gene. Further, let C be a set of gene clusters. The* consistency problem *is to decide if C is* consistent *with respect to m, i.e. whether there exists a sequence s over $\mathcal{G}_N$ for which the following properties hold:*

(i) s contains each gene g at most $m(g)$ times, and
(ii) s contains all gene clusters $c \in C$.

Whenever we want to consider consistency as a reconstruction criterion, we have to provide a solution for the above problem. As we will see in the following sections, the problem complexity highly depends on the specific cluster definition.

In our framework, we assume that the gene multiplicities are given. Nevertheless, we want to sketch some ways to specify $m(g)$ for the internal nodes in the phylogenetic tree.

For some specific datasets, we can rely on knowledge about the genomic history. For instance, several studies suggest two whole genome duplications in the evolution of the Chordate genome in the teleost fishes lineage [13]. Such information can be used to deduce the ancestral number of genes.

Otherwise, the most accurate but also elaborate approach would be to deploy *gene-tree species-tree reconciliation* [17] to reconstruct the history of the genes in terms of speciation events, gene duplications and gene losses. Less extensive and more suitable for our needs, one could also utilize approaches which do not require any further data or pre-knowledge. Probability-based methods [8] could be applied to effectively and reliably infer ancestral gene multiplicities $m_v(g)$ for all internal nodes v, given the copy number at the leaves. Or, we could apply the concept of parsimony and minimize the amount of copy number differences. A less restrictive solution is to define the multiplicity of a gene g for node u in a bottom-up fashion as the maximum over the multiplicities of its child nodes $v_1, \ldots, v_k$: $m_u(g) := \max_{i=1,\ldots,k} \big(m_{v_i}(g)\big)$.

Instead of performing a separate preprocessing step to fix the thresholds in advance, one could also try to include the gene multiplicity into the overall objective of the reconstruction. However, in general, optimizing for a combination including an original objective, consistency, and the gene copy number would be an intricate task due to the strong interdependencies of the subcriteria.

3 An Efficient Solution for Adjacencies

Probably the simplest formalization of co-localization of genes is the concept of *adjacencies*, i.e. two directly neighboring genes. This elementary pattern of gene order conservation, also known as *gene pairs* or *neighboring genes*, has been widely used in whole genome comparison. Especially in the field of gene order reconstruction, this model is one of the most prevalent concepts [4, 7, 15].

3.1 Unsigned Adjacencies

In the following, we formalize the concept of adjacencies and present a method to efficiently solve the consistency problem for adjacencies on sequences, i.e. to decide if there exists a sequence that contains a set of given adjacencies while considering each gene g at most $m(g)$ times. To model the problem, we use a graph theoretic approach.

Definition 2 (Unsigned Adjacencies on Sequences). *Let $\mathcal{G}_N := \{1, \ldots, N\}$ be a set of genes. An adjacency $\{a, b\}$ of the genes $a, b \in \mathcal{G}_N$ is contained in a sequence s over $\mathcal{G}_N$ if and only if a and b occur adjacently at least once in s.*

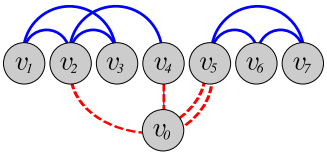

Fig. 2. An example to illustrate the proof of Lemma 1. Consider the set of genes $\mathcal{G}_7$ with the multiplicities $m(g) = 2$ for $g \in \{2, 5\}$ and otherwise $m(g) = 1$, and the set $C = \big\{\{1, 2\}, \{1, 3\}, \{2, 3\}, \{2, 4\}, \{5, 6\}, \{5, 7\}, \{6, 7\}\big\}$ of unsigned adjacencies. The gene order graph $G_7(C)$ is depicted including the extensions described in the proof. The solid edges correspond to the original edges as defined by the given adjacencies, and the dashed lines represent the auxiliary edges. The obtained extended graph contains, for instance, the Eulerian cycle $(v_0, v_2, v_1, v_3, v_2, v_4, v_0, v_5, v_6, v_7, v_5, v_0)$, which corresponds to the valid gene order $(2, 1, 3, 2, 4, 5, 6, 7, 5)$.

Definition 3 (Gene Order Graph). *Let $\mathcal{G}_N = \{1, \ldots, N\}$ be a set of genes and C be a set of pairs $\{a, b\}$ with $a, b \in \mathcal{G}_N$. Then, the* gene order graph of C*, denoted by $G_N(C)$, is the graph with the vertex set $\{v_g \mid g \in \mathcal{G}_N\}$ and the edge set $\big\{\{v_a, v_b\} \mid \{a, b\} \in C\big\}$.*

The gene order graph of a set of adjacencies C can be constructed in $\mathcal{O}(N + |C|)$ time and space. In this process, we keep track of the degree of each node v_g, denoted by $\deg(v_g)$. Then, the following lemma allows us to check for consistency of C in $\mathcal{O}(N)$ steps and thus in a total running time and with a space requirement of $\mathcal{O}(N + |C|)$.

Lemma 1. *Let $\mathcal{G}_N = \{1, \ldots, N\}$ be a set of genes and let $m : \mathcal{G}_N \to \mathbb{N}$ assign a maximum copy number to each gene. Further, let C be a set of pairs $\{a, b\}$ with $a, b \in \mathcal{G}_N$ and $G_N(C) = (V, E)$ be the gene order graph of C. Then, C is consistent with respect to m if and only if the following conditions hold:*

(i) $\deg(v_g) \leq 2m(g)$ *for all vertices $v_g \in V$, and*
(ii) $\sum\limits_{v_g \in c} \big(2m(g) - \deg(v_g)\big) > 0$ *for each connected component c in $G_N(C)$.*

Proof. Assume we have given $\mathcal{G}_N$, m, C and $G_N(C)$ as required by the lemma. We extend the gene order graph $G_N = (V, E)$ to a multigraph $H_N = (V', E')$, where the new vertex set contains one additional node v_0, i.e. $V' = V \cup \{v_0\}$. The multiset of edges E' contains all edges in E with multiplicity one and further auxiliary edges: For each vertex $v_g \neq v_0$ with $deg(v_g) < 2m(g)$ we add the edge $\{v_0, v_g\}$ with multiplicity $2m(g) - deg(v_g)$ to E'.

If condition (i) of the lemma holds, then all nodes in the obtained extended graph have even degree: All vertices $v_g \neq v_0$ are filled up to a degree of $2m(g)$ and v_0 is incident to $\sum_{v_g \in V} \big(2m(g) - deg(v_g)\big) = \sum_{v_g \in V} 2m(g) - 2|C|$ edges. Further, condition (ii) implies that for each connected component of G_N, in the extended graph, at least one edge connects this subgraph to v_0. Hence, H_N is connected.

Conditions (i) and (ii) imply that H_N is Eulerian. That means, there is a path starting and ending in v_0 which contains all edges, especially the edges of the original gene order graph, exactly once. Since each node $v_g \neq v_0$ has a degree of $2m(g)$, it is traversed

exactly $m(g)$ times. Each such Eulerian path corresponds to a sequence of genes that contains all adjacencies in C and each gene g exactly $m(g)$ times, as exemplified in Figure 2. Thus, C is consistent with respect to m.

On the contrary, if condition (i) is not satisfied, there is at least one gene g that is contained in more given adjacencies than its multiplicity $m(g)$ allows. And, if condition (ii) does not hold for any connected component c, the maximal number of adjacencies of all genes in c is exhausted and the genes cannot be put into a linear order, i.e. a cycle containing v_0, with the remaining genes. In both cases, the existence of a valid gene order is precluded and thus, consistency is disproven. □

3.2 Signed Adjacencies

A slightly more sophisticated variant of the adjacency model is motivated by the observation that the orientation of genes can play a role in co-expression and also in gene order conservation [11].

Definition 4 (Signed Adjacencies on Signed Sequences). *Let $\mathcal{G}_N := \{1, \ldots, N\}$ be a set of genes. A signed adjacency $\{a, b\}$ of the genes $a, b \in \{g, -g \mid g \in \mathcal{G}_N\}$ is contained in a sequence s over $\mathcal{G}_N$ if and only if a is directly followed by $-b$, or b by $-a$ at least once in s.*

Example 1. Consider the model of signed adjacencies for $N = 4$. The signed adjacency $\{2, -3\}$ is contained in both sequences $s_1 = (1, 2, 3, 4)$ and $s_2 = (4, 1, -3, -2)$. No other signed adjacencies of the genes 2 and 3 are contained in any of the two sequences.

We transfer the general idea from the unsigned to the signed case. To this end, we adjust the definition of the gene order graph. Now, each gene g is represented by two nodes in the graph, where each such pair is connected by $m(g)$ edges.

Definition 5 (Signed Gene Order Graph). *Let $\mathcal{G}_N = \{1, \ldots, N\}$ be a set of genes and let $m : \mathcal{G}_N \to \mathbb{N}$ assign a maximum copy number to each gene. Further, let C be a set of pairs $\{a, b\}$ with $a, b \in \{g, -g \mid g \in \mathcal{G}_N\}$. Then, the* signed gene order graph *of C, denoted by $G_N^s(C)$, is the multigraph with the vertex set $\{v_g, v_{-g} \mid g \in \mathcal{G}_N\}$ and the multiset of edges $\big\{\{v_g, v_{-g}\}$ with multiplicity $m(g) \mid g \in \mathcal{G}_N\big\} \cup \big\{\{v_a, v_b\}$ with multiplicity one $\mid \{a, b\} \in C\big\}$.*

Similarly to the unsigned case, we can construct the graph in $\mathcal{O}(N + |C|)$ time.

Lemma 2. *Let $\mathcal{G}_N = \{1, \ldots, N\}$ be a set of genes and let $m : \mathcal{G}_N \to \mathbb{N}$ assign a maximum copy number to each gene. Further, let C be a set of signed adjacencies $\{a, b\}$ with $a, b \in \{g, -g \mid g \in \mathcal{G}_N\}$ and $G_N^s(C) = (V, E)$ be the signed gene order graph of C. Then, C is consistent with respect to m if and only if the following conditions hold:*

(i) $deg(v_g) \leq 2m(|g|)$ *for all vertices $v_g \in V$, and*
(ii) $\sum_{v_g \in c} \big(2m(|g|) - deg(v_g)\big) > 0$ *for each connected component c in $G_N^s(C)$.*

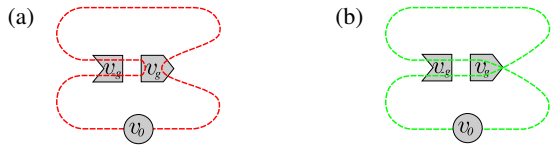

Fig. 3. Illustration of the relation of improper and proper Eulerian cycles in an extended signed gene order graph as described in the proof of Lemma 2.

Proof. We proceed analogously to the unsigned case described in the proof of Lemma 1: We extend the signed gene order graph $G_N^s = (V, E)$ to a multigraph $H_N^s = (V', E')$, where the new vertex set contains one additional node v_0, i.e. $V' = V \cup \{v_0\}$. The multiset of edges E' contains all edges in E with multiplicity one and further auxiliary edges: For each vertex $v_g \neq v_0$ with $deg(v_g) < 2m(|g|)$ we add the edge $\{v_0, v_g\}$ with multiplicity $2m(|g|) - deg(v_g)$ to E'.

Then, again, the conditions (i) and (ii) of Lemma 2 imply the existence of an Eulerian path in $H_N^s(C)$. But in this case, the correspondence of such a path to a valid gene order is not trivial. When the pair of nodes representing gene g is traversed by a path $(\ldots, v_{-g}, v_g, \ldots)$, this relates to a signed gene order $(\ldots, g, \ldots)$, whereas a path $(\ldots, v_g, v_{-g}, \ldots)$ correlates to a signed gene order $(\ldots, -g, \ldots)$. By definition, $H_N^s(C)$ includes $m(|g|)$ edges $\{v_g, v_{-g}\}$. An Eulerian cycle passes each of these edges, but not necessarily in the above mentioned way. It might also be of the form $(\ldots, v_f, v_{-g}, v_g, v_{-g}, v_h \ldots)$ with $f \neq g \neq h$, which does not represent a signed gene order. In this case, $m(|g|) \geq 2$ and due to the construction of the extended graph, there are $m(|g|)$ edges $\{v_g, v_{-g}\}$ and at least $m(|g|)$ edges $\{v_g, v_h\}$ with $h \neq -g$. Hence, the considered Eulerian cycle has to pass node v_g again in the form $\ldots, v_i, v_g, v_j, \ldots$ with $i \neq -g \neq j$, as shown in Figure 3(a). However, whenever this situation arises, it is always possible to construct an alternative Eulerian cycle $(v_0, \ldots, v_{-g}, v_g, \ldots, v_{-g}, v_g, \ldots, v_0)$, as depicted in Figure 3(b). If this modification is performed for all such improperties, the obtained Eulerian cycle is proper in the sense that it represents a signed gene order $(\ldots, g, \ldots, g, \ldots)$. Thus, conditions (i) and (ii) imply not only the existence of an Eulerian path but also the existence of a valid signed gene order and hence consistency of C with respect to m. The reverse direction of the lemma holds analogously to Lemma 1. □

Based on the definition of a gene order graph, Lemmas 1 and 2 provide algorithms to solve the consistency problem on adjacencies on sequences in time and space linear in the number of genes and in the number of given adjacencies. Both the models and the lemmas can easily be modified to allow one circular gene order or even several circular chromosomes. Only the connectivity requirement has to be relaxed correspondingly.

4 NP Completeness for Common Intervals

To find larger conserved regions, we now address a model for gene clusters that, in contrast to adjacencies, generally span more than two genes: *Common intervals*, segments of the genome containing the same set of genes in an arbitrary order but not interrupted by other genes.

4.1 Basic Common Intervals

In line with other studies, we base our definition on the notion of *character sets*, which enables us to formalize the cluster model in a straightforward way. Since we will utilize this term for models on signed sequences later on, we directly define it for the general, signed case. Although, in our framework, a common interval is defined on a *single* gene order, we stick to the term *common* to not confuse the reader familiar with this gene cluster model by redefining the same concept under a different name.

Definition 6 (Character Set). *Let $s = (a_1, \ldots, a_{|s|})$ be a signed sequence. Then, the* character set *of s, denoted $\mathcal{CS}(s)$, is the set of all elements in s: $\mathcal{CS}(s) := \left\{|a| \mid a \in \{a_1, \ldots, a_{|s|}\}\right\}$.*

Definition 7 (Common Intervals on Sequences). *Let $\mathcal{G}_N := \{1, \ldots, N\}$ be a set of genes. Then, a common interval $c \subseteq \mathcal{G}_N$ with $|c| > 1$ is contained in a sequence s over $\mathcal{G}_N$ if and only if s contains a substring s' such that $\mathcal{CS}(s') = c$.*

Recall that we want to find an answer to the question: Given a set of common intervals C and a multiplicity threshold function m, is there a valid gene order that contains all elements of C and meets the restrictions imposed by m? As we will show now, this problem is NP complete.

Theorem 1. *The consistency problem for common intervals on sequences is NP complete.*

Proof (Sketch). One can easily formulate an algorithm that verifies a given solution, i.e. a proper gene order, for correctness in polynomial time, which shows that the problem belongs to the complexity class NP.

NP hardness is proven by reduction from the following variant of the Hamiltonian cycle problem: Let $G = (V \cup W, E)$ be a connected, undirected, bipartite graph with $|V| = |W| \geq 3$, $E \subseteq \left\{\{v, w\} \mid v \in V, w \in W\right\}$ and $\deg(u) = |\{e \in E | u \in e\}| \leq 3$ for all $u \in V \cup W$. Decide whether there exists a Hamiltonian cycle in G, i.e. a path in G that starts and ends in the same vertex $v' \in V$ and in-between contains each vertex $v \in V \backslash \{v'\}$ exactly once. This problem is known to be NP complete [12]. By reducing it to the consistency problem for common intervals in polynomial time, we get NP hardness of the latter problem.

For a given graph as stated in the problem definition, we construct an instance of the consistency problem as follows: For each node, depending on its degree, and for each edge, depending on the graph structure, we add certain elements to the set of genes and define certain common intervals. We restrict the multiplicity of the genes such that common intervals have to overlap in a valid gene order in specific substrings. Each of these substrings corresponds to a node or an edge in the graph, and overlapping substrings correspond to paths in the graph. Finally, we show that the substrings can be combined to a complete, valid sequence if and only if there is a Hamiltonian cycle in the graph. Details are given in [21, Appendix A.2]. □

4.2 Variants of Common Intervals

Beside its classical definition, there are different generalizations of common intervals on sequences discussed in the literature, such as r-window clusters and max-gap clusters [9], or approximate gene clusters [6, 18]. Since the consistency problem is NP complete for basic common intervals, any generalization is NP hard as well.

In contrast to generalizations, there are also other cluster models which are restricted variants of common intervals. In the following, we will discuss such models, in particular framed and nested common intervals.

Framed Common Intervals

This gene cluster model, common intervals framed by two genes whose orientations have to be conserved, has first been introduced on permutations as *conserved intervals* [3]. In gene order reconstruction, framed common intervals on permutations have been the first model to formally state the problem of finding putative ancestral sets of gene clusters preserving consistency [2].

Definition 8 (Framed Common Intervals on Signed Sequences). *Let $\mathcal{G}_N := \{1, \ldots, N\}$ be a set of genes. A framed common interval $[a\, I\, b]$ consists of two* extremities *a and b with $|a|, |b| \in \mathcal{G}_N$, and a set of* inner elements *$I \subseteq \mathcal{G}_N$. We say that $[a\, I\, b]$ is contained in a signed sequence s, if and only if in s, a is followed by b or $-b$ by $-a$, and the character set of the substring between the extremities is equal to I.*

According to this definition, a gene can be extremity and inner element, or even left and right extremity at the same time. Apart from that, analogously to basic common intervals, a cluster can occur multiple times in one genome, and one gene can be contained several times in one cluster occurrence, as illustrated by the following example.

Example 2. Consider the model of framed common intervals for $N = 6$ and sequence $s = (5, 4, -2, -1, 2, -3, 6)$. Beside others, the framed common interval $[4\,\{1, 2\}\,{-3}]$, is contained in s as illustrated by the *box diagram* below, where the occurrences of the extremeties and the inner elements are surrounded by rectangles:

$$s = (\ +5, \boxed{+4,} \boxed{-2,\ -1,\ +2,} \boxed{-3,} +6\ ) .$$

The obvious relationship of basic and framed common intervals allows to infer an important correlation of these models with respect to the consistency problem: Any instance of this problem for common intervals can be reduced to an instance for framed common intervals. Based on this, we can deduce the following statement.

Theorem 2. *The consistency problem for framed common intervals on signed sequences is NP complete.*

Proof (Sketch). To show NP hardness, we reduce the consistency problem for common intervals to the consistency problem for framed common intervals in polynomial time. To this end, we introduce two additional genes with a certain multiplicity and, for each

basic common interval, we create a framed common interval including these genes as framing and inner elements. Then, any valid gene order for one problem instance can be transformed into a valid gene order for the other instance by removing occurrences of the additional genes or inserting them, respectively. Details are given in [21, Appendix A.3]. □

Nested Common Intervals

Hoberman and Durand [9] discussed nestedness as a desired property of gene clusters and proposed a first algorithm to identify respective clusters. Recently, *nested common intervals* were formally defined and studied in [5].

Definition 9 (Nested Common Intervals on Sequences). *Let $\mathcal{G}_N := \{1, \ldots, N\}$ be a set of genes. The structure of a nested common interval is defined recursively. A nested common interval is either*

(i) an unordered pair of genes $\{a, b\}$ with $a \neq b$, which is contained in a sequence s over $\mathcal{G}_N$ if and only if a and b are adjacent in s, or

(ii) a tupel (c, a) of a nested common interval c and a gene a, which is contained in a sequence s if and only if, in s, a is adjacent to a substring s' of s such that $\mathcal{CS}(c) = \mathcal{CS}(s')$ and c is contained in s',

where the character set of a nested common interval is the set of all contained genes: $\mathcal{CS}(\{a, b\}) := \{a, b\}$ and $\mathcal{CS}((c, a)) := \mathcal{CS}(c) \cup \{a\}$.

Similar to the other cluster models discussed above, any nested common interval may occur multiple times in one genome and one gene may be contained multiple times in the occurrence of a cluster in one genome. Analogously to framed common intervals, one gene may be incorporated in the definition of one cluster several times.

Example 3. Consider the model of nested common intervals for $N = 6$ and sequence $s = (5, 4, 2, 1, 2, 3, 6)$. Then, beside others, the nested common interval $((\{2, 3\}, 1), 4)$ is contained in s as illustrated below, where the occurrences of the subclusters are indicated by lines:

$$(5, \underline{4}, \underline{2, 1}, \underline{2, 3}, 6)\,.$$

Even the strict assumption of nestedness is not strong enough to allow an efficient verification of consistency.

Theorem 3. *The consistency problem for nested common intervals on sequences is NP complete.*

Proof (Sketch). NP hardness is proven similarly to Theorem 1. The common intervals used in that proof can be replaced by certain nested common intervals such that the argumentation holds similarly. Details are given in [21, Appendix A.4]. □

Further Variations and Restrictions

Our NP completeness results also hold for further variations of the above models: defined on circular sequences, restricted in size, basic and nested common intervals containing each gene at most once etc. These results are detailed in [21] and will be discussed elaborately in the full version of this paper.

5 Conclusion

In this paper, we have discussed the consistency problem, i.e. the decision whether there exists a valid gene order comprising a given set of gene clusters. We have discussed this question for different gene cluster models on sequences with restricted gene multiplicities. In summary, we identified a severe border between gene cluster models for which we can verify consistency efficiently and those for which we cannot. The complexity rises drastically from linear for adjacencies to NP hard for more general cluster models, even if they are strongly restricted.

This raises the question for a sequence-based gene cluster model that, on the one hand, allows some degree of flexibility and, on the other hand, offers a polynomial-time algorithm to verify consistency. The integration of such a model into any of the existing reconstruction methods could increase sensitivity. Actually, first results on both simulated and real data indicate that within segments of conserved gene content, the order of the genes is conserved almost exactly [21]. Thus, a model covering only single missing or additional genes, or the reversal of two neighboring genes could already enhance reconstruction results strongly.

Another way to find a practical solution for flexible models is not to be deterred by the NP hardness results. In fact, the reduction procedures on which the proofs are based produce instances of the consistency problem where the multiplicity for some genes grows with the instance size. This suggests to reconsider the problems in the context of *fixed parameter tractability*. However, our first investigations in this direction were not promising. Moreover, the maximum copy number of genes observed in real data can be large in general.

We implemented the gene order graph to model adjacencies on sequences and integrated this gene cluster model into our unified reconstruction framework, presented in [19], available from the web site `bibiserv.techfak.uni-bielefeld.de/rococo/`. An elaborate description of the method and the results can be found in [21]. We refrain from reporting detailed results here because these are concerned more with the reconstruction method than with the general concept of consistency discussed in this paper. Nevertheless, we would like to mention the following overall findings. Simulations showed that estimating the gene multiplicities using the simple maximum approach does not significantly decrease the accuracy of the reconstruction compared to using the "real" simulated copy numbers. Furthermore, we applied our method on genomic data of Corynebacteria using different gene cluster models: Common intervals on permutations and adjacencies on sequences. A comparison of the results revealed a large overlap. Nevertheless, many conserved segments could only be identified by either of the approaches. This highlights the importance of studying gene cluster reconstruction with respect to different, especially flexible models for gene clusters and the relaxed model of sequences for gene order.

Acknowledgments

The authors wish to thank Cedric Chauve for discussions on an earlier version of this manuscript. The work of Roland Wittler was supported by the DFG Graduiertenkolleg Bioinformatik (GK 635).

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. Comp. Biol.*, 14(4):436–445, 2007.
2. A. Bergeron, M. Blanchette, A. Chateau, and C. Chauve. Reconstructing ancestral gene order using conserved intervals. In *Proc. of WABI 2004*, 3240 of *LNBI*, pages 14–25, 2004.
3. A. Bergeron and J. Stoye. On the similarity of sets of permutations and its applications to genome comparison. *J. Comp. Biol.*, 13(7):1340–1354, 2006.
4. A. Bhutkar, W. M. Gelbart, and T. F. Smith. Inferring genome-scale rearrangement phylogeny and ancestral gene order: A drosophila case study. *Genome Biol.*, 8(11):R236, 2007.
5. G. Blin and J. Stoye. Finding nested common intervals efficiently. *J. Comp. Biol.*, to appear 2010. (A preliminary version appeared in *Proc. of RECOMB-CG 2009*, 5817 of *LNBI*, pages 59–69, 2009.)
6. S. Böcker, K. Jahn, J. Mixtacki, and J. Stoye. Computation of median gene clusters. *J. Comp. Biol.*, 16(8):1085–1099, 2009.
7. C. Chauve and E. Tannier. A methodological framework for the reconstruction of contiguous regions of ancestral genomes and its application to mammalian genomes. *PLoS Comput. Biol.*, 4(11):e1000234, 2008.
8. M. Csűrös and I Miklós. Mathematical framework for phylogenetic birth-and-death models. *Arxiv preprint arXiv:0902.0970*, 2009.
9. R. Hoberman and D. Durand. § The incompatible desiderata of gene cluster properties. In *Proc. of RECOMB-CG 2005*, 3678 of *LNBI*, pages 73–87, 2005.
10. W.-L. Hsu and R. McConnell. PQ-trees, PC-trees, and planar graphs. In D. P. Mehta and S. Sahni, editors, *Handbook of Data Structures and Applications*. 2004.
11. M. A. Huynen, B. Snel, and P. Bork. Inversions and the dynamics of eukaryotic gene order. *Trends Genet.*, 17(6):304–306, 2001.
12. A. Itai, C. H. Papadimitriou, and J. L. Szwarcfiter. Hamilton paths in grid graphs. *SIAM J. Computing*, 11(4):676–686, 1982.
13. O. Jaillon, J.-M. Aury, F. Brunet, J.-L. Petit, N. Stange-Thomann *et al.* Genome duplication in the teleost fish tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature*, 431(7011):946–957, 2004.
14. J. G. Lawrence and J. R. Roth. Selfish operons: Horizontal transfer may drive the evolution of gene clusters. *Genetics*, 143(4):1843–1860, 1996.
15. J. Ma, L. Zhang, B. B. Suh, B. J. Raney, R. C. Burhans, J. W. Kent, M. Blanchette, D. Haussler, and W. Z. Miller. Reconstructing contiguous regions of an ancestral genome. *Genome Res.*, 16(12):1557–1565, 2006.
16. R. Overbeek, M. Fonstein, M. D'Souza, G. D. Pusch, and N. Maltsev. The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. USA*, 96(6):2896–2901, 1999.
17. R. D. M. Page and M. A. Charleston. From gene to organismal phylogeny: Reconciled trees and the gene tree/species tree problem. *Mol. Phyl. and Evol.*, 7(2):231 – 240, 1997.
18. S. Rahmann and G. W. Klau. Integer linear programs for discovering approximate gene clusters. In *Proc. of WABI 2006*, 4175 of *LNBI*, pages 298–306, 2006.
19. J. Stoye and R. Wittler. A unified approach for reconstructing ancient gene clusters. *IEEE/ACM Trans. Comput. Biol. Bioinf.*, 6(3):387–400, 2009.
20. T. Uno and M. Yagiura. Fast algorithms to enumerate all common intervals of two permutations. *Algorithmica*, 26(2):290–309, 2000.
21. R. Wittler. Phylogeny-based Analysis of Gene Clusters. *Ph.D. Thesis*, Faculty of Technology, Bielefeld University, 2010, Online available from `http://bieson.ub.uni-bielefeld.de/volltexte/2010/1627/`.