The SCJ Small Parsimony Problem for Weighted Gene Adjacencies

Nina Luhmann, Manuel Lafond, Annelyse Thévenin, Aïda Ouangraoua, Roland Wittler, and Cedric Chauve

Abstract—Reconstructing ancestral gene orders in a given phylogeny is a classical problem in comparative genomics. Most existing methods compare conserved features in extant genomes in the phylogeny to define potential ancestral gene adjacencies, and either try to reconstruct all ancestral genomes under a global evolutionary parsimony criterion, or, focusing on a single ancestral genome, use a scaffolding approach to select a subset of ancestral gene adjacencies, generally aiming at reducing the fragmentation of the reconstructed ancestral genome. In this paper, we describe an exact algorithm for the Small Parsimony Problem that combines both approaches. We consider that gene adjacencies at internal nodes of the species phylogeny are weighted, and we introduce an objective function defined as a convex combination of these weights and the evolutionary cost under the Single-Cut-or-Join (SCJ) model. The weights of ancestral gene adjacencies can, e.g., be obtained through the recent availability of ancient DNA sequencing data, which provide a direct hint at the genome structure of the considered ancestor, or through probabilistic analysis of gene adjacencies evolution. We show the NP-hardness of our problem variant and propose a Fixed-Parameter Tractable algorithm based on the Sankoff-Rousseau dynamic programming algorithm that also allows to sample co-optimal solutions. We apply our approach to mammalian and bacterial data providing different degrees of complexity. We show that including adjacency weights in the objective has a significant impact in reducing the fragmentation of the reconstructed ancestral gene orders. An implementation is available at http://github.com/nluhmann/PhySca.

Index Terms—Comparative genomics, ancestral reconstruction, parsimony, genome rearrangements

1 Introduction

Reconstructing ancestral gene orders is a long-standing computational biology problem with important applications, as shown in several recent large-scale projects [1], [2], [3]. Informally, the problem can be defined as follows: Given a phylogenetic tree representing the speciation history leading to a set of extant genomes, we want to reconstruct the structure of the ancestral genomes corresponding to the internal nodes of the tree.

Existing ancestral genome reconstruction methods concentrate on two main strategies. *Local* approaches consider the reconstruction of one specific ancestor at a time independently from the other ancestors of the tree. Usually, they do not consider an evolutionary model and proceed in two stages: (1) comparing gene orders of ingroup and outgroup species to define potential ancestral gene adjacencies, and (2) selecting a conflict-free subset of ancestral gene adjacencies-where a conflict is defined as an ancestral gene extremity belonging to more than two potential adjacencies, e.g., due to convergent evolution-, to obtain a set of Contiguous Ancestral Regions (CARs) [4], [5], [6]. The second stage of this approach is often defined as a combinatorial optimization problem aiming to minimize the number of discarded ancestral adjacencies, thus maximizing the number of selected adjacencies [4], [6], [7]. This stage follows principles common in *scaffolding* methods used to obtain gene orders for extant genomes from sequencing data [8], [9]. This approach was recently used to scaffold an ancestral pathogen genome for which ancient DNA (aDNA) sequencing data could be obtained [10]. *Global* approaches on the other hand simultaneously reconstruct ancestral gene orders at all internal nodes of the considered phylogeny, generally based on a parsimony criterion within an evolutionary model. This so called *Small Parsimony Problem* has been studied with several underlying genome rearrangement models, such as the breakpoint distance or the Double-Cut-and-Join (DCJ) distance [11], [12], [13]. While rearrangement scenarios based on complex rearrangement models can give insights into underlying evolutionary mechanisms, from a computational point of view, the Small Parsimony Problem is NP-hard for most rearrangement distances [14]. One

- *N. Luhmann and R. Wittler are with the International Research Training Group "Computational Methods for the Analysis of the Diversity and Dynamics of Genomes" and the Genome Informatics Group, Faculty of Technology and Center for Biotechnology, Bielefeld University, Bielefeld 33615, Germany. E-mail: {nina.luhmann, roland.wittler}@uni-bielefeld.de.*
- *M. Lafond is with the Department of Computer Science and Operational Research, Université de Montréal, Montréal H3C 3J7, Canada. E-mail: lafonman@iro.umontreal.ca.*
- *A. Thévenin is with the Genome Informatics Group, Faculty of Technology and Center for Biotechnology, Bielefeld University, Bielefeld 33615, Germany. E-mail: atheven@cebitec.uni-bielefeld.de.*
- *A. Ouangraoua is with the Department of Computer Science, Université de Sherbrooke, Sherbrooke, QC J1K 2R1, Canada. E-mail: aida.ouangraoua@usherbrooke.ca.*
- *C. Chauve is with the Department of Mathematics, Simon Fraser University, Burnaby, BC V5A 1S6, Canada. E-mail: cedric.chauve@sfu.ca.*

Manuscript received 15 July 2016; accepted 29 Dec. 2016. Date of publication 31 Jan. 2017; date of current version 5 Aug. 2019.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TCBB.2017.2661761

exception is the Single-Cut-or-Join (SCJ) distance, for which linear/circular ancestral gene orders can be found in polynomial time [15], however constraints required to ensure algorithmic tractability yield fragmented ancestral gene orders.

The two approaches outlined above optimize somewhat orthogonal criteria. For example, the underlying goal of the local approach is to maximize the agreement between the resulting ancestral gene order and the set of potential ancestral adjacencies, independently of the other ancestral gene orders. Would it be applied independently to all ancestral nodes, potential ancestral adjacencies exhibiting a mixed profile of presence/absence in the extant genomes might then lead to a set of non-parsimonious ancestral gene orders. The global approach aims only at minimizing the evolutionary cost in the phylogeny and can result in more fragmented ancestral gene orders. Nevertheless, there is little ground to claim that one approach or the other is more accurate or to be preferred, and the work we present is an attempt to reconcile both approaches.

We introduce a variant of the Small Parsimony Problem based on an optimality criterion that accounts for both an evolutionary distance and the difference between the initial set of potential ancestral adjacencies and the final consistent subset of adjacencies conserved at each ancestral node. More precisely we consider that each potential ancestral gene adjacency can be provided with a (prior) non-negative weight at every internal node. The contribution of the discarded adjacencies to the objective function is then the sum of their weights. These adjacency weights can e.g., be obtained as probabilities computed by sampling scenarios for each potential adjacency independently [16] or can be based on ancient DNA sequencing data providing direct prior information assigned to certain ancestral nodes. It follows that the phylogenetic framework we present can then also assist in scaffolding fragmented assemblies of aDNA sequencing data [10], [17].

We prove NP-hardness of the problem variant we introduce and describe an exact exponential time algorithm for reconstructing consistent ancestral genomes under this optimality criterion, based on a mixed Dynamic Programming / Integer Linear Programming approach. We show that this Small Parsimony Problem variant is Fixed-Parameter Tractable (FPT), with a parameter linked to the amount of conflict in the data. Moreover, this also allows us to provide an FPT sampling algorithm for co-optimal solutions,-a problem recently addressed in [18] using a MCMC approach. We evaluate our method on a simulated dataset and compare our results to several other methods reconstructing ancestral genomes. Further, we apply our method to two real data sets: mammalian genomes spanning roughly one million years of evolution, and bacterial genomes (pathogen *Yersinia*) spanning 20,000 years of evolution and for which some aDNA sequencing data is available. We show that we can reduce the fragmentation of ancestral gene orders in both datasets by integrating adjacency weights while reconstructing robust ancestral genomes.

This paper is an extended version of the work previously presented in [19], particularly including new results on simulated datasets and a hardness proof of the defined problem.

2 Background and Problem Statement

2.1 Genomes and Adjacencies

Genomes consist of chromosomes and plasmids. Each such component can be represented as a linear or circular sequence of oriented markers over a marker alphabet. Markers correspond to homologous sequences between genomes, e.g., genes or synteny blocks. We assume that each marker appears exactly once in each genome, so our model does not consider duplications or deletions. To account for its orientation, each marker x is encoded as a pair of marker extremities (x_h, x_t) or (x_t, x_h).

An *adjacency* is an unordered pair of marker extremities, e.g., $\{x_t, y_h\}$. The order of markers in a genome can be encoded by a set of adjacencies. Two distinct adjacencies are said to be *conflicting* if they share a common marker extremity. If a set of adjacencies contains conflicting adjacencies, it is not *consistent* with a mixed linear/circular genome model. We assume that the set of adjacencies for an extant assembled genome is consistent. The set of adjacencies for one genome naturally defines an *adjacency graph*, where nodes represent marker extremities and edges represent adjacencies. Conflicting adjacencies can be identified as branching nodes in this graph.

2.2 The Small Parsimony Problem and Rearrangement Distances

In a global phylogenetic approach, we are given a phylogenetic tree with extant genomes at its leaves and internal nodes representing ancestral genomes. We denote by $\mathcal{A}$ the set of all adjacencies present in at least one extant genome and assume that every ancestral adjacency belongs to $\mathcal{A}$. Then the goal is to find a labeling of the internal nodes by consistent subsets of $\mathcal{A}$ minimizing a chosen genomic distance over the tree. This is known as the *Parsimonious Labeling Problem*.

Definition 1 (Parsimonious Labeling Problem). *Let $T = (V, E)$ be a tree with each leaf l labeled with a consistent set of adjacencies $\mathcal{A}_l \subseteq \mathcal{A}$, and d a distance between consistent sets of adjacencies. A labeling $\lambda : V \to \mathcal{P}(\mathcal{A})$ with $\lambda(l) = \mathcal{A}_l$ for each leaf is parsimonious for d if none of the internal nodes $v \in V$ contains a conflict and it minimizes the sum $W(\lambda, T)$ of the distances along the branches of T*

$$W(\lambda, T) = \sum_{(u,v)\in E} d\big(\lambda(u), \lambda(v)\big).$$

This problem is NP-hard for most rearrangement distances taken as evolutionary models. The only known exception is the set-theoretic Single-Cut-or-Join distance [15]. It defines a rearrangement distance by two operations: the *cut* and *join* of adjacencies. Given two genomes defined by consistent sets of adjacencies A and B, the SCJ distance between these genomes is

$$d_{SCJ}(A, B) \;=\; |\,A - B\,| + |\,B - A\,|\,.$$

The Small Parsimony Problem under the SCJ model can be solved by computing a parsimonious gain/loss history

for each adjacency separately with the dynamic programming Fitch algorithm [20], [21] in polynomial time. Consistent labelings can be ensured with the additional constraint that in case of ambiguity at the root of the tree, the absence of the adjacency is chosen [15]. As each adjacency is treated independently, this constraint might automatically exclude all adjacencies being part of a conflict to ensure consistency. This results in an unnecessarily sparse reconstruction in terms of reconstructed adjacencies and thus more fragmented genomes higher up in the tree.

2.3 Generalization by Weighting Adjacencies

When considering an internal node v, we define node u as its parent node in T. We assume that a specific adjacency graph is associated to each ancestral node v, whose edges are annotated by a weight $w_{v,a} \in [0,1]$ representing a confidence measure for the presence of adjacency a in species v. Then in a global reconstruction, cutting an adjacency of a higher weight has higher impact in terms of the optimization criterion than cutting an adjacency of lower weight.

Formally, we define two additional variables for each adjacency $a \in \mathcal{A}$ at each internal node $v \in V$: The presence (or absence) of a at node v is represented by $p_{v,a} \in \{0,1\}$, while $c_{v,a} \in \{0,1\}$ indicates a change for the status of an adjacency along an edge (u,v), i.e., $p_{u,a} \neq p_{v,a}$. We consider the problem of optimizing the following objective function, where $\alpha \in [0,1]$ is a convex combination factor.

Definition 2 (Weighted SCJ Labeling Problem). *Let $T = (V,E)$ be a tree with each leaf l labeled with a consistent set of adjacencies $\mathcal{A}_l \subseteq \mathcal{A}$ and each adjacency $a \in \mathcal{A}$ is assigned a given weight $w_{v,a} \in [0,1]$ for each node $v \in V$. A labeling λ of the internal nodes of T with $\lambda(l) = \mathcal{A}_l$ for each leaf is an* optimal weighted SCJ labeling *if none of the internal nodes $v \in V$ contains a conflict and it minimizes the criterion*

$$D(\lambda, T) = \sum_{v,a} \alpha(1 - p_{v,a}) w_{v,a} + (1-\alpha) c_{v,a}.$$

Further, we can state the corresponding co-optimal sampling problem. A sampling method is important to examine different co-optimal rearrangement scenarios that can explain evolution toward the structure of extant genomes.

Definition 3 (Weighted SCJ Sampling Problem). *Given the setting of the Weighted SCJ Labeling Problem, sample uniformly from all labelings λ of the internal nodes of T that are solutions to the Weighted SCJ Labeling Problem.*

2.4 Problem Complexity

Aside of the many heuristics for the Small Parsimony Problem for non-SCJ rearrangement models (see for example [12], [13], [22] for the DCJ distance), there exist a few positive results for the Weighted SCJ Labeling Problem with specific values of α.

If $\alpha = 0$, the objective function corresponds to the Small Parsimony Problem under the SCJ distance and hence a solution can be found in polynomial time [15]. A generalization towards multifurcating, edge-weighted trees including prior information on adjacencies at exactly one internal node of the tree is given in [17]. Recently, Miklós and Smith [18] proposed a Gibbs sampler for sampling optimal labelings under the SCJ model with equal branch lengths. It starts from an optimal labeling obtained as in [15], and then explores the space of co-optimal labelings through repeated constrained parsimonious modifications of a single adjacency evolutionary scenario. This method addresses the issue of the high fragmentation of internal node labelings, but convergence is not proven, and so there is no bound on the computation time.

If $\alpha = 1$, i.e., we do not take evolution in terms of SCJ distance along the branches of the tree into account, we can solve the problem by applying independently a maximum-weight matching algorithm at each internal node [7]. So the extreme cases of the problem are tractable, and while we assume that the general problem is hard, we will now prove it for a small range of α.

Theorem 1. *The Weighted SCJ Labeling Problem is NP-hard for any $\alpha > 33/34$.*

We show the hardness by reduction from the Maximum Intersection Matching Problem, which is defined as follows. Let G_1 and G_2 be two graphs on the same vertex set. Find a perfect matching in G_1 and G_2 such that the number of edges common to both matchings is maximized. We prove NP-hardness of this problem by reduction from 3-Balanced-Max-2-SAT (see appendix for details, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2017.2661761).

Theorem 2. *The Maximum Intersection Matching Problem is NP-complete.*

The relation of the Weighted SCJ Labeling Problem and the Maximum Intersection Matching Problem can be sketched as follows. For a given instance of the Maximum Intersection Matching Problem, G_1 and G_2, we construct a tree that contains the edges of both graphs as potential adjacencies. For $\alpha > 33/34$, an optimal labeling of two internal nodes then corresponds to perfect matchings in G_1 and G_2. Maximizing the number of common edges of the matching then minimizes the SCJ distance between the nodes. A detailed proof is given in the appendix, available online.

3 Methods

In order to find a solution to the Weighted SCJ Labeling Problem, we first show that we can decompose the problem into smaller independent subproblems. Then, for each subproblem containing conflicting adjacencies, we show that, if it contains a moderate level of conflict, it can be solved using the Sankoff-Rousseau algorithm [23] with a complexity parameterized by the size of the subproblem. For a highly conflicting subproblem, we show that it can be solved by an Integer Linear Program (ILP).

3.1 Decomposition into Independent Subproblems

We first introduce a graph that encodes all adjacencies present in at least one internal node of the considered phylogeny (Definition 4 and Supplementary Fig. 7, available online).

As introduced previously, we consider a tree $T = (V, E)$ where each node is augmented with an adjacency graph.

Definition 4 (Global adjacency graph). *The set of vertices V_{AG} of the global adjacency graph AG consists of all marker extremities present in at least one of the adjacency graphs. There is an edge between two vertices $a, b \in V_{AG}$ that are not extremities of a same marker, if there is an internal node in the tree T whose adjacency graph contains the adjacency $\{a, b\}$. The edge is labeled with the list of all internal nodes that contain this adjacency.*

Each connected component C of the global adjacency graph defines a subproblem composed of the species phylogeny, the set of marker extremities equal to the vertex set of C, and the set of adjacencies equal to the edge set of C. According to the following lemma, whose proof is straightforward, it is sufficient to solve each such subproblem independently.

Lemma 1. *The set of all optimal solutions of the Weighted SCJ Labeling Problem is the set-theoretic Cartesian product of the sets of optimal solutions of the instances defined by the connected components of the global adjacency graph.*

To solve the problem defined by a connected component C of the global adjacency graph containing conflicts, we rely on an adaptation of the Sankoff-Rousseau algorithm with exponential time complexity, parameterized by the size and nature of conflicts of C, and thus can solve subproblems with moderate amount of conflict.

3.2 Overview of the Sankoff-Rousseau Algorithm

The Sankoff-Rousseau dynamic programming algorithm [23] solves the general Small Parsimony Problem for discrete characters. Let L be the set of all possible labels of a node in the phylogeny. Then for each node u in the tree, the cost $c(a, u)$ of assigning a label $a \in L$ to this node is defined recursively as follows

$$c(a, u) = \sum_{v \text{ child of } u} \min_{b \in L} \big(c(b, v) + d(a, b)\big),$$

where in our case $d(a, b)$ is defined as in Definition 2. This equation defines a dynamic programming algorithm whose base case is when u is a leaf in which case $c(a, u) = 0$ if $\lambda(u) = a$ and $c(a, u) = \infty$ otherwise. Afterwards, choosing a label with the minimum cost at the root node and backtracking in a top-down traversal of the tree results in a most parsimonious labeling. We refer to [24] for a review on the Sankoff-Rousseau algorithm.

3.3 Application to the Weighted SCJ Labeling Problem

In order to use the Sankoff-Rousseau algorithm to solve the problem defined by a connected component C of the global adjacency graph, we define a label of an internal node of the phylogeny as the assignment of at most one adjacency to each marker extremity. More precisely, let x be a marker extremity in C, v an internal node of T, and $e_1, \ldots, e_{d_x}$ be all edges in the global adjacency graph that are incident to x and whose label contains v (i.e., represent adjacencies in the adjacency graph of node v). We define the set of possible labels of v as $L_{x,v} = \{\emptyset, e_1, \ldots, e_{d_x}\}$. The set of potential labels L_v of node v is then the Cartesian product of the label sets $L_{x,v}$ for all $x \in V(C)$ resulting in a set of discrete labels for v of size $\prod_{x \in V(C)}(1 + d_x)$. Note that not all of these joint labelings are valid as they can assign an adjacency $a = (x, y)$ to x but not to y, or adjacency $a = (x, y)$ to x and $b = (x, z)$ to z thus creating a conflict (see Supplementary Fig. 8, available online for an example).

For an edge (u, v) in the tree, we can then define a cost matrix that is indexed by pairs of labels of L_u and L_v, respectively. The cost is infinite if one of the labels is not valid, and defined by the objective function otherwise. We can then apply the Sankoff-Rousseau approach to find an optimal labeling of all internal nodes of the tree for connected component C.

Note that, if C is a connected component with no conflict, it is composed of two vertices and a single edge, and can be solved in space $O(n)$ and time $O(n)$.

3.4 Complexity Analysis

The time and space complexity of the algorithm is obviously exponential in the size of C. Indeed, the time (resp. space) complexity of the Sankoff-Rousseau algorithm for an instance with a tree having n leaves and r possible labels for each node is $O(nr^2)$ (resp. $O(nr)$) [24]. In our algorithm, assuming n leaves in T (i.e., n extant species), m_C vertices in the global adjacency graph of C and a maximum degree d_C for vertices (marker extremities) in this graph, $(1 + d_C)^{m_C}$ is an upper bound for the size of the label set L_v for a node v. Moreover, computing the distance between two labels of L_v and L_u, where (u, v) is an edge of T, can trivially be done in time and space $O(m_C)$: If both labels are valid, it suffices to check how many common adjacencies are present in both labels, while deciding if a label is not valid can be done by a one-pass examination of the label. Combining this with the Sankoff-Rousseau complexity yields a time complexity in $O(nm_C(1 + d_C)^{2m_C})$ and a space complexity in $O(nm_C(1+ d_C)^{m_C})$.

Given a general instance, i.e., an instance not limited to a single connected component of the global adjacency graph, we can consider each connected component independently (Lemma 1). For a set of N markers and c connected components in the global adjacency graph defining a conflicting instance, we define D as the maximum degree of a vertex and M as the maximum number of vertices in all such components. Then, the complexity analysis above shows that the problem is Fixed-Parameter Tractable.

Theorem 3. *The Weighted SCJ Labeling Problem can be solved in worst-case time $O(nN(1 + D)^{2M})$ and space $O(nN\,(1 + D)^{M})$.*

In practice, the exponential complexity of our algorithm depends on the structure of the conflicting connected components of the global adjacency graph. The dynamic programming algorithm will be effective on instances with either small conflicting connected components or small degrees within such components, and will break down with a single component with a large number of vertices of high degree. For such components, the time complexity is provably high and we propose an ILP to solve them.

3.5 An Integer Linear Program

We can formulate the optimization problem as a simple ILP. We consider two variables for any adjacency a and node v, $p_{v,a} \in \{0,1\}$ and $c_{v,a} \in \{0,1\}$, defined as in Section 2.

$$\text{Minimize } \sum_{v,a} \alpha(1-p_{v,a})w_{v,a} + (1-\alpha)c_{v,a}$$

subject to

$$p_{v,a} + p_{u,a} - p_{p,a} \geq 0 \text{ for } (p,u),\ (p,v) \in E(T) \quad (c_1)$$
$$p_{v,a} + p_{u,a} - p_{p,a} \leq 1 \text{ for } (p,u),\ (p,v) \in E(T) \quad (c_2)$$
$$p_{v,a} + p_{u,a} + c_{v,a} \leq 2 \text{ for } (u,v) \in E(T) \quad (c_3)$$
$$p_{v,a} + p_{u,a} - c_{v,a} \geq 0 \text{ for } (u,v) \in E(T) \quad (c_4)$$
$$p_{v,a} - p_{u,a} + c_{v,a} \geq 0 \text{ for } (u,v) \in E(T) \quad (c_5)$$
$$-p_{v,a} + p_{u,a} + c_{v,a} \geq 0 \text{ for } (u,v) \in E(T) \quad (c_6)$$
$$\sum_{a=(x_t,y)} p_{v,a} \leq 1 \text{ and } \sum_{a=(x_h,y)} p_{v,a} \leq 1$$
$$\text{for any marker } x \text{ and node } v. \quad (c_7)$$

The constraints ensure parsimony (c_1 and c_2), consistency of the solution (c_7) and define the correct value for $c_{v,a}$ dependent on the value of p_a along an edge (u,v) (c_{3-6}). This ILP has obviously a size that is polynomial in the size of the problem.

3.6 Sampling Co-Optimal Labelings

The Sankoff-Rousseau DP algorithm can easily be modified to sample uniformly from the space of all optimal solutions to the Weighted SCJ labeling Problem in a forward-backward fashion. The principle is to proceed in two stages: first, for any pair (v,a) we compute the number of optimal solutions under this label for the subtree rooted at v. Then, when computing an optimal solution, if a DP equation has several optimal choices, one is randomly picked according to the distribution of optimal solutions induced by each choice (see Appendix for more details, available in the online supplemental material). This classical dynamic programming approach leads to the following result.

Theorem 4. *The Weighted SCJ Sampling Problem can be solved in worst-case time* $O(nN(1+D)^{2M})$ *and space* $O(nN(1+D)^{M})$.

For subproblems that are too large for being handled by the Sankoff-Rousseau algorithm, the SCJ Small Parsimony Gibbs sampler recently introduced [18] can easily be modified to incorporate prior weights, although there is currently no proven property regarding its convergence.

3.7 Weighting Ancestral Adjacencies

A first approach to assign weights to ancestral adjacencies consist in considering evolutionary scenarios for an adjacency independently of the other adjacencies. An evolutionary scenario for an adjacency is a labeling of the internal nodes of the species phylogeny T by the presence or absence of the adjacency, and the parsimony score of a scenario is the number of gains/losses of the adjacency along the branch of T, i.e., the SCJ score for this single adjacency. For a scenario σ, we denote by $p(\sigma)$ its parsimony score. Its Boltzmann score is then defined as $B(\sigma) = e^{-\frac{p(\sigma)}{kT}}$, where kT is a given constant. If we denote the set of all possible evolutionary scenarios for the adjacency $\{x,y\}$ by $\mathcal{S}(x,y)$, the partition function of the adjacency and its Boltzmann probability are defined as

$$Z(x,y) = \sum_{\sigma \in \mathcal{S}(x,y)} B(\sigma),\ Pr(\sigma) = \frac{B(\sigma)}{Z(x,y)}.$$

The weight of the adjacency at internal node v is then the sum of the Boltzmann probabilities of all scenarios where the adjacency is present at node v. All such quantities can be computed in polynomial time [16].

Parameter kT is useful to skew the Boltzmann probability distribution: If kT tends to zero, parsimonious scenarios are heavily favored and the Boltzmann probability distribution tends to the uniform distribution over optimal scenarios, while when kT tends to ∞, the Boltzmann distribution tends toward the uniform distribution over the whole solution space. In our experiments, we chose a value of $kT = 0.1$ that favors parsimonious scenarios but considers also slightly suboptimal scenarios.

When aDNA sequence data is available for one or several ancestral genomes, markers identified in extant species can be related to assembled contigs of the ancestral genome, as in [10] for example. For an ancestral adjacency in a species for which aDNA reads are available, it is then possible to associate a sequence-based weight to the adjacency-either through gap filling methods (see Section 4, where we use the probabilistic model of GAML [25]), or scaffolding methods such as BESST [26] for example. In comparison to the weighting approach described above, these weights are then not directly based on the underlying phylogeny, but provide an external signal for the confidence of adjacencies at the respective internal node.

4 Results

We evaluated our algorithm on a simulated dataset and compared its sensitivity and precision to several other reconstruction methods. Further, we applied our method to two real datasets: *mammalian* and *Yersinia* genomes. The mammalian dataset was used in the studies [5] and [18]. It contains six mammalian species and two outgroups, spanning over 100 million years of evolution, and five different marker sets of varying resolution (minimal marker length). Our experimental results consider issues related to the complexity of our algorithm, the use of a pure SCJ reconstruction (obtained when the α parameter equals 0) and the relative impact of the value of α on both the total evolutionary cost and the ancestral gene orders fragmentation. Our second dataset contains eleven *Yersinia* genomes, an important human pathogen. This dataset contains contigs from the recently sequenced extinct agent of the Black Death pandemic [27] that occurred roughly 650 years ago. We refer to Supplementary Figs. 9 and 10, available online for the species phylogenies of these two datasets.

4.1 Simulations

We created simulated datasets as described in [28]: with a birth-rate of 0.001 and a death rate of 0, we simulated 20 binary trees with 6 leaves and scaled the branch lengths such that the tree has a diameter $D = 2n$, where n is the

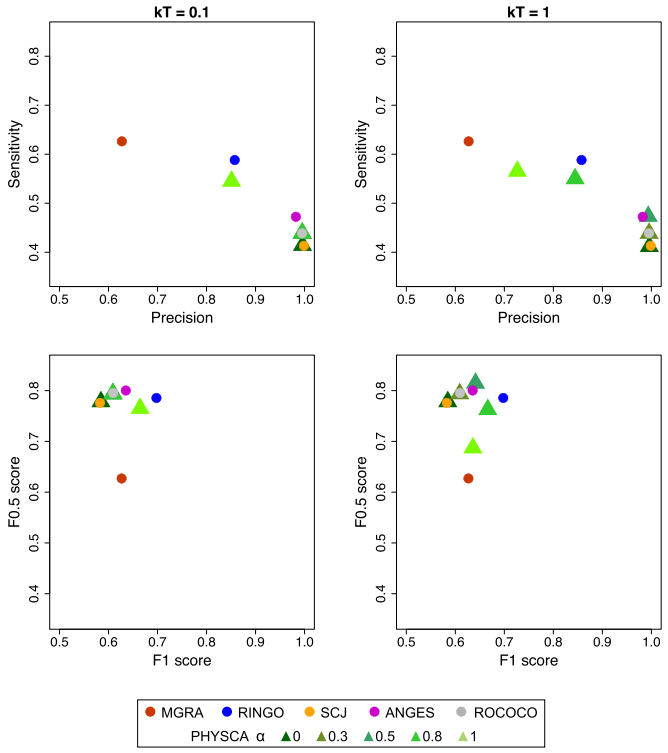

Fig. 1. Average precision and sensitivity (top), and F_1 and $F_{0.5}$ (bottom) of reconstructions on 20 simulated datasets. Adjacency weights have been obtained with parameters $kT = 0.1$ (left) and $kT = 1$ (right).

number of markers in each unichromosomal genome. The root genome with $n = 500$ markers is then evolved along the branches of the tree by applying inversions and translocations with a probability of 0.9 and 0.1 respectively. The number of rearrangements at each branch corresponds to the simulated branch length, the total number of rearrangements ranges from 1,242 to 2,296 in the simulated trees. We compare results of our implementation *PHYSCA* for different values of $\alpha \in \{0, 0.3, 0.5, 0.8, 1\}$ with the tools *RINGO* [28], *MGRA* [29], Fitch-SCJ [30], *ROCOCO* [31], [32] (dense approach for signed adjacencies) and *ANGES* [33] (adjacencies only). We computed adjacency weights as described in Section 3.7 with the software *DeClone* [16] and parameter $kT \in \{0.1, 1\}$.

The methods *RINGO* and *MGRA* are global approaches minimizing the DCJ-distance in the tree, while *ANGES* reconstructs specific ancestors locally in the tree and is applied for each node separately. For $\alpha = 0$, our objective is finding a consistent, most parsimonious solution and equals the objectives of Fitch-SCJ and *ROCOCO*, where Fitch-SCJ always finds the most fragmented solution whereas *ROCOCO* and our method aim at reporting least fragmented reconstructions.

We measured sensitivity and precision of the reconstructions based on the comparison of simulated and reconstructed adjacencies by the different methods. A high sensitivity indicates the ability to recover the true marker order of ancestors in the phylogeny, while a high precision denotes few wrongly reconstructed adjacencies. As shown in Fig. 1, our method reaches a high precision of 0.99 for all values of $\alpha \geq 0.5$, while increasing the sensitivity in comparison to the pure Fitch-SCJ solution by reducing the fragmentation of the reconstructed scaffolds. For higher values of α, the influence of the weighting becomes apparent: for $kT = 0.1$, the precision only decreases for $\alpha = 1$, while for $kT = 1$, the precision decreases also for lower values of α, however leading to more complete reconstructions. In comparison, both DCJ-based methods *RINGO* and *MGRA* produce less fragmented solutions by recovering more true adjacencies under the jeopardy of also reconstructing more false adjacencies. The sensitivity and precision of Fitch-SCJ, *ROCOCO* and *ANGES* are comparable to our method for low to medium values of α.

The F_1 score assesses the relation of sensitivity and precision with equal importance. *RINGO* achieves a better F_1 score than all other methods. The $F_{0.5}$ score emphasizes the precision of a method over its sensitivity. With this measure, our method with $kT = 1$ and $\alpha = 0.5$ outperforms the other tools, while *ROCOCO* and *ANGES* also reach similarly good scores.

In general, it can be seen that the equal contribution of global evolution and local adjacency weights in the objective function provides a reliable reconstruction and further a useful tool to explore the solution space under different values of α.

4.2 Mammalian Dataset

We used the markers computed in [5] from whole-genome alignments. The extant species contain a diverse number of chromosomes ranging from 9 chromosomes in *opossum* to 39 chromosomes in *pig*. Unique and universal markers were computed as synteny blocks with different resolution in terms of minumum marker length. Note that all rearrangement breakpoints are therefore located outside of marker coordinates. It results in five different datasets varying from 2,185 markers for a resolution of 100 kb to 629 markers for a resolution of 500 kb.

We considered all adjacencies present in at least one extant genome as potentially ancestral. To weight an adjacency at all internal nodes of the tree, we relied on evolutionary scenarios for each single adjacency, in terms of gain/loss, independently of the other adjacencies (i.e., without considering consistency of ancestral marker orders). We obtain these weights using the software DeClone [16], and we refer to them as *DeClone weights*. We considered two values of the DeClone parameter kT, 0.1 and 1, the former ensuring that only adjacencies appearing in at least one optimal adjacency scenario have a significant DeClone weight, while the latter samples adjacencies outside of optimal scenarios. For the analysis of the ancestral marker orders obtained with our algorithm, we considered the data set at 500 kb resolution and sampled 500 ancestral marker orders for all ancestral species under different values of α.

4.2.1 Complexity

The complexity of our algorithm is dependent on the size of the largest connected component of the global adjacency graph. In order to restrict the complexity, we kept only adjacencies whose weights are above a given threshold x. Fig. 2 shows the expected decrease in computational complexity correlated to threshold x for the five different minimal marker lengths. In most cases, all connected components

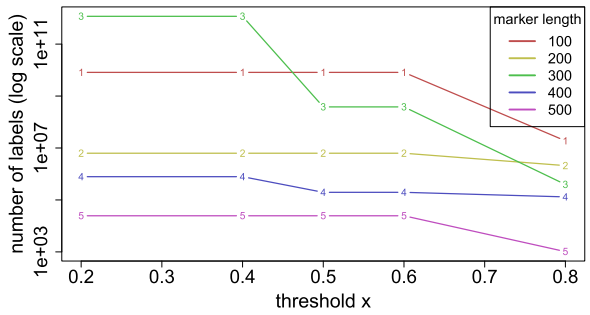

Fig. 2. Number of different labels for the largest connected component in each of the mammalian datasets. This statistic provides an upper bound for the actual complexity of our reconstruction algorithm.

TABLE 1
Frequency of Adjacencies in 500 Samples with $\alpha = 0$ as Percentage of Optimal Labelings They Appear in

Ancestor	Frequency f		
	$f = 100\%$	$100\% > f > 50\%$	$f < 50\%$
Boreoeutheria	94.66	1.07	4.27
Euarchontoglires	95.42	0.88	3.79
Ferungulates	96.53	0.55	2.92
Primates	98.82	0.34	0.84
Rodentia	99.49	0.34	0.17
Theria	97.67	0.89	1.43
root node	92.23	1.23	6.53

are small enough to be handled by our exact algorithm in reasonable time except for very large components in the marker sets with higher resolution under a low threshold x. For the 500 kb dataset with $x = 0.2$ and $kT = 1$, the computation of one solution takes on average 200 s on a 2.6 GHz i5 with 8 GB of RAM. It can be reduced to 30 s when DeClone weights are based on $kT = 0.1$. This illustrates that our algorithm, despite an exponential worst-case time complexity, can process realistic datasets in practice.

4.2.2 Optimal SCJ Labelings

Next, we analyzed the 500 optimal SCJ labelings obtained for $\alpha = 0$, i.e., aiming only at minimizing the SCJ distance, and considered the fragmentation of the ancestral gene orders (number of CARs) and the total evolutionary distance. Note that, unlike the Fitch algorithm used in [15], our algorithm does not favor fragmented assemblies by design but rather considers all optimal labelings. Sampling of co-optimal solutions shows that the pure SCJ criterion leads to some significant variation in terms of number of CARs (Fig. 3). In contrast, Table 1 shows that most observed ancestral adjacencies are present in all sampled scenarios. About 5 percent of adjacencies, mostly located at nodes higher up in the phylogeny, are only present in a fraction of all sampled scenarios, indicating that there is a small number of conflicts between potential adjacencies that can be solved ambiguously at the same parsimony cost.

The optimal SCJ distance in the tree for $\alpha = 0$ is 1,674, while the related DCJ distance in the sampled reconstructions varies between 873 and 904 (Fig. 4). In comparison, we obtained a DCJ distance of 829 with GASTS [22], a small parsimony solver directly aiming at minimizing the DCJ

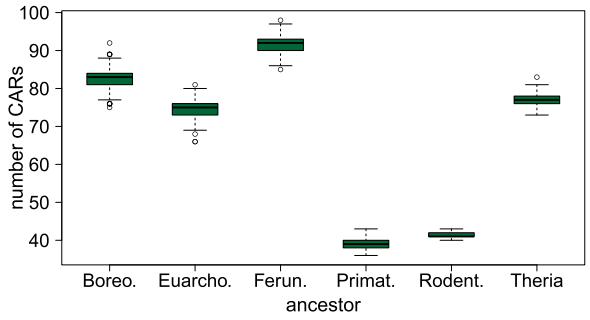

Fig. 3. Number of reconstructed CARs at each internal node in 500 samples for the mammalian dataset with 500 kb resolution, $x = 0.2$ and $\alpha = 0$.

distance. More precisely, over all ancestral nodes, 70 adjacencies found by GASTS do not belong to our predefined set of potential ancestral adjacencies and another 147 appear in the 500 samples with a frequency below 50 percent. This illustrates both a lack of robustness of the pure SCJ optimal labelings, and some significant difference between the SCJ and DCJ distances.

Finally, we compared the Boltzmann probabilities of ancestral adjacencies (DeClone weights) with the frequency observed in the 500 samples. There is a very strong agreement for DeClone weights obtained with $kT = 0.1$ as only 14 ancestral adjacency have a DeClone weight that differs more than 10 percent from the observed frequency in the samples. This shows that, despite the fact that the DeClone approach disregards the notion of conflict, it provides a good approximation of the optimal solutions of the SCJ Small Parsimony Problem.

4.2.3 Ancestral Reconstruction with DeClone Weights and Varying Values of α

For $\alpha > 0$, our method minimizes a combination of the SCJ distance with the DeClone weights of the adjacencies discarded to ensure valid ancestral gene orders. Again, we sampled 500 solutions each for different values of α with the 500 kb data set. We distinguish between DeClone parameter $kT = 0.1$ and $kT = 1$. Figs. 4 and 5 show the respective observed results in terms of evolutionary distance and fragmentation.

For $kT = 0.1$, the optimal SCJ and DCJ distance over the whole tree hardly depends on α. Including the DeClone weights in the objective actually results in the same solution, independent of $\alpha > 0$. In fact, while applying a low weight threshold of $x = 0.2$, the set of potential adjacencies is already consistent at all internal nodes except for a few conflicts at the root that are solved unambiguously for all values of α. This indicates that building DeClone weights on the basis of mostly optimal adjacency scenarios (low kT) results in a weighting scheme that agrees with the evolution along the tree for this dataset. More importantly, Figs. 4 and 5 show that the combination of DeClone weights followed by our algorithm, leads to a robust set of ancestral gene orders.

In comparison, for $kT = 1$, we see an increase in SCJ and DCJ distance for higher α, while the number of CARs at internal nodes decreases, together with a loss of the robustness of the sampled optimal results when α gets close to 1. It

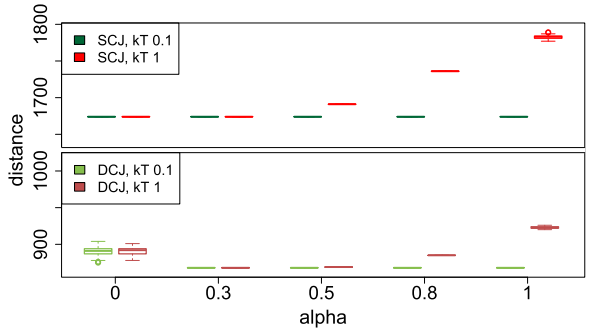

Fig. 4. SCJ distance (upper half) and DCJ (lower half) distance in the whole tree for all samples and selected values of α in the mammalian dataset.

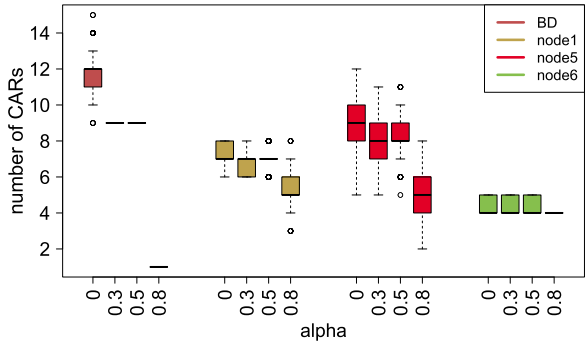

Fig. 6. Reconstructed number of CARs in the *yersinia* dataset with aDNA weights at the BD node and 0 otherwise, for four ancestral nodes.

can be explained by the observation that the weight distribution of ancestral adjacencies obtained with DeClone and $kT = 1$ is more balanced than with $kT = 0.1$ as it considers suboptimal scenarios of adjacencies with a higher probability. It further illustrates that, when the global evolutionary cost of a solution has less weight in the objective function, the algorithm favors the inclusion of an adjacency of moderate weight that joins two CARs while implying a moderate number of evolutionary events (for example an adjacency shared by only a subset of extant genomes). From that point of view, our algorithm-being efficient enough to be run on several values of α-provides a useful tool to evaluate the relation between global evolution and prior confidence for adjacencies whose pattern of presence/absence in extant genomes is mixed.

4.3 *Yersinia pestis* Dataset

We started from fully assembled DNA sequences of seven *Yersinia pestis* and four *Yersinia pseudotuberculosis* genomes. In addition, we included aDNA single-end reads and 2,134 contigs of length > 500 bp assembled from these reads for the Black Death agent, considered as ancestral to several extant strains [27]. We refer to this augmented ancestral node as the *Black Death (BD) node*. The marker sequences for all extant genomes were computed as described in [10], restricting the set of markers to be unique and universal. We obtained a total of 2,207 markers in all extant genomes and 2,232 different extant adjacencies, thus showing a relatively low level of syntenic conflict compared to the number of markers, although it implies a highly dynamic rearrangement history over the short period of evolution [10].

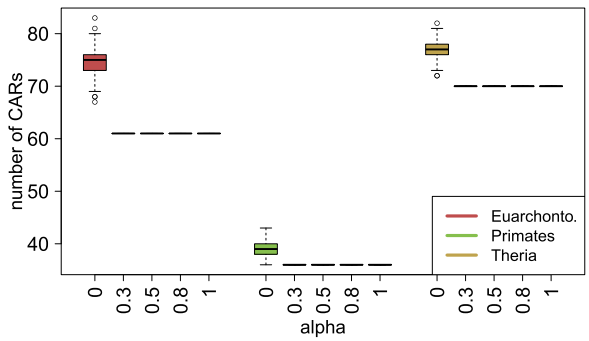

Fig. 5. Number of CARs in the mammalian dataset in all samples at selected internal nodes for different values of α reconstructed with DeClone weights under $kT = 0.1$. While the number of CARs differs in the case of $\alpha = 0$ where the adjacency weights are not considered, the fragmentation stays constant for the other values of α.

As for the mammalian dataset, we considered as potentially ancestral any adjacency that appears in at least one extant genome. However for this dataset, reducing the complexity by applying a weight threshold x was not necessary. For the BD node, adjacency weights can be based on the given aDNA reads for a given potential ancestral adjacency as follows. First, we used FPSAC [10] to compute DNA sequences filling the gaps between any two adjacent marker extremities (obtained by aligning the gap sequences of the corresponding conserved extant adjacencies and reconstructing a consensus ancestral sequence using the Fitch algorithm). Then we computed the weights as a likelihood of this putative gap sequence given the aDNA reads, using the GAML probabilistic model described in [25]. Each adjacency together with its template gap sequence details a proposition for an assembly A as a piece of the real ancestral sequence, and given the aDNA read set R, the model then defines a probability $Pr(R|A) = \prod_{r \in R} Pr(r|A)$ for observing the reads R given that A is the correct assembly. The probability $Pr(r|A)$ can be computed by aligning r to the assembly A while the alignment is evaluated under an appropriate sequencing error model. We refer to [25] for details.

4.3.1 Ancestral Reconstruction with aDNA Weights

Again we sampled 500 solutions for this dataset. We computed the weights at the BD node based on the aDNA data, while adjacencies at all other nodes were given weight 0. Hence we can investigate the influence of including the aDNA sequencing data in the reconstruction while for the rest of the tree, the weights do not impact the objective function. Moreover, this weighting scheme addresses the issue of potential BD adjacencies with a low weight due to the difficulty of sequencing ancient DNA.

As shown in Fig. 6, for selected internal nodes of the phylogeny, the pure SCJ solutions at $\alpha = 0$ result in the highest fragmentation, while the number of CARs decreases as we increase the importance of the adjacency weights in the objective of our method. For the BD node, when including the aDNA weights, the fragmentation is decreasing while the reconstructions for each $\alpha > 0$ are robust. At the other nodes, the applied sequencing weights also reduce the fragmentation except for node6 which is located in the pseudotuberculosis subtree and hence more distant to the BD node. This shows that the aDNA weights not only influence

the reconstructed adjacencies at the BD node, but also other nodes of the tree.

5 Conclusion

Our main contributions are the introduction of the Small Parsimony Problem under the SCJ model with adjacency weights, together with an exact parameterized algorithm for the optimization and sampling versions of the problem. The motivation for this problem is twofold: incorporating sequence signal from aDNA data when it is available, and recent works showing that the reconstruction of ancestral genomes through the independent analysis of adjacencies is an interesting approach [15], [16], [18], [34].

Regarding the latter motivation, we address a general issue of these approaches that either ancestral gene orders are not consistent or are quite fragmented if the methods are constrained to ensure consistency. The main idea we introduce is to take advantage of sampling approaches recently introduced in [16] to weight potential ancestral adjacencies and thus direct, through an appropriate objective function, the reconstruction of ancestral gene orders. Our results on the mammalian dataset suggest that this approach leads to a robust ancestral genome structure. However, we can observe a significant difference with a DCJ-based ancestral reconstruction, a phenomenon that deserves to be explored further. Our algorithm, which is based on the Sankoff-Rousseau algorithm similarly to several recent ancestral reconstruction algorithms [16], [18], [34], is a parameterized algorithm that can handle real instances containing a moderate level of syntenic conflict. Our experimental results on both the mammalian and bacterial datasets suggest that introducing prior weights on adjacencies in the objective function has a significant impact in reducing the fragmentation of ancestral gene orders, even with an objective function with balanced contributions of the SCJ evolution and adjacency weights. For highly conflicting instances, it can be discussed if a reconstruction through small parsimony is the right approach to solve these conflicts or if these should be addressed differently.

Our sampling algorithm improves on the Gibbs sampler introduced in [18] in terms of computational complexity and provides a useful tool to study ancestral genome reconstruction from a Bayesian perspective. Moreover, our algorithm is flexible regarding the potential ancestral gene adjacencies provided as input and could easily be associated with other ideas, such as intermediate genomes for example [28].

There are several research avenues opened by our work. From a theoretical point of view, we know the problem we introduced is tractable for $\alpha = 0$ and $\alpha = 1$, and we show it is hard for $\alpha > 33/34$, but it remains to see whether it is hard otherwise. Further, given that the considered objective is a combination of two objectives to be optimized simultaneously, Pareto optimization is an interesting aspect that should be considered. Our model could also be extended towards other syntenic characters than adjacencies, i.e., groups of more than two markers, following the ancient gene clusters reconstruction approach introduced in [31]. As ancestral gene orders are defined by consistent sets of adjacencies, the principle of our dynamic programming algorithm could be conserved and it would only be a matter of integrating gene clusters into the objective function. From a more applied point of view, one would like to incorporate duplicated and deleted markers into our Small Parsimony Problem. There exist efficient algorithms for the case of a single adjacency [16], [34] that can provide adjacency weights, and natural extensions of the SCJ model to incorporate duplicated genes. However it remains to effectively combine these ideas. Finally, again due to the flexibility and simplicity of the Sankoff-Rousseau dynamic programming algorithm, one could easily extend our method towards the inference of extant adjacencies if some extant genomes are provided in partially assembled form following the general approach described in [35], [36].

This would pave the way towards a fully integrated phylogenetic scaffolding method that combines evolution and sequencing data for selected extant and ancestral genomes.

Acknowledgments

Nina Luhmann and Roland Wittler are funded by the International DFG Research Training Group GRK 1906/1. Cedric Chauve is funded by NSERC grant RGPIN-249834.

References

[1] F. Denoeud, et al., "The coffee genome provides insight into the convergent evolution of caffeine biosynthesis," *Sci.*, vol. 345, 2013, Art. no. 125527.
[2] R. Ming, et al., "The pineapple genome and the evolution of CAM photosynthesis," *Nature Genetics*, vol. 47, pp. 1435–1442, 2015.
[3] D. E. Neafsey, et al., "Highly evolvable malaria vectors: The genome of 16 *Anopheles* mosquitoes," *Sci.*, vol. 347, 2015, Art. no. 1258522.
[4] D. Bertrand, Y. Gagnon, M. Blanchette, and N. El-Mabrouk , "Reconstruction of ancestral genome subject to whole genome duplication, speciation, rearrangement and loss," in *Proc. 10th Int. Workshop Algorithms Bioinf.*, 2010, vol. 6293, pp. 78–89.
[5] 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.*, vol. 4, 2008, Art. no. e1000234.
[6] J. Ma, et al., "Reconstructing contiguous regions of an ancestral genome," *Genome Res.*, vol. 16, pp. 1557–1565, 2006.
[7] J. Mañuch, M. Patterson, R. Wittler, C. Chauve, and E. Tannier, "Linearization of ancestral multichromosomal genomes," *BMC Bioinf.*, vol. 13, no. Suppl 19, 2012, Art. no. S11.
[8] E. Bosi, et al., "MeDuSa: A multi-draft based scaffolder," *Bioinf.*, vol. 31, no. 15, pp. 2443–2451, 2015.
[9] I. Mandric and A. Zelikovsky, "ScaffMatch: Scaffolding algorithm based on maximum weight matching," *Bioinf.*, vol. 31, no. 16, pp. 2632–2638, 2015.
[10] A. Rajaraman, E. Tannier, and C. Chauve, "FPSAC: Fast phylogenetic scaffolding of ancient contigs," *Bioinf.*, vol. 29, pp. 2987–2994, 2013.
[11] M. Alekseyev and P. A. Pevzner, "Breakpoint graphs and ancestral genome reconstructions," *Genome Res.*, vol. 19, pp. 943–957, 2009.
[12] J. Kovác, B. Brejová, and T. Vinar, "A practical algorithm for ancestral rearrangement reconstruction," in *Proc. 11th Int. Workshop Algorithms Bioinf.*, 2011, vol. 6833, pp. 163–174.
[13] C. Zheng and D. Sankoff, "On the pathgroups approach to rapid small phylogeny," *BMC Bioinf.*, vol. 12, no. Suppl 1, 2011, Art. no. S4.
[14] E. Tannier, C. Zheng, and D. Sankoff, "Multichromosomal median and halving problems under different genomic distances," *BMC Bioinf.*, vol. 10, no. 1, 2009, Art. no. 120.
[15] P. Feijão and J. Meidanis, "SCJ: A breakpoint-like distance that simplifies several rearrangement problems," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 8, no. 5, pp. 1318–1329, Sep./Oct. 2011.
[16] C. Chauve, Y. Ponty, and J. Zanetti, "Evolution of genes neighborhood within reconciled phylogenies: An ensemble approach," *BMC Bioinf.*, vol. 16, no. Suppl. 19, 2015, Art. no. S6.

[17] N. Luhmann, C. Chauve, J. Stoye, and R. Wittler, "Scaffolding of ancient contigs and ancestral reconstruction in a phylogenetic framework," in *Proc. 9th Brazilian Symp. Bioinf. Comput. Biol.*, 2014, vol. 8826, pp. 135–143.
[18] I. Miklós and H. Smith, "Sampling and counting genome rearrangement scenarios," *BMC Bioinf.*, vol. 16, no. Suppl 14, 2015, Art. no. S6.
[19] N. Luhmann, A. Thévenin, A. Ouangraoua, R. Wittler, and C. Chauve, "The SCJ small parsimony problem for weighted gene adjacencies," in *Proc. Int. Symp. Bioinf. Res. Appl.*, 2016, pp. 200–210.
[20] W. Fitch, "Toward defining the course of evolution: Minimum change for a specific tree topology," *Syst. Biol.*, vol. 20, pp. 406–416, 1971.
[21] J. A. Hartigan, "Minimum mutation fits to a given tree," *Biometrics*, vol. 29, no. 1, pp. 53–65, 1973.
[22] A. Xu and B. Moret, "GASTS: Parsimony scoring under rearrangements," in *Proc. 11th Int. Workshop Algorithms Bioinf.*, 2011, vol. 6833, pp. 351–363.
[23] D. Sankoff and P. Rousseau, "Locating the vertices of a steiner tree in an arbitrary metric space," *Math. Program.*, vol. 9, pp. 240–246, 1975.
[24] M. Csűrös, "How to infer ancestral genome features by parsimony: Dynamic programming over an evolutionary tree," in *Proc. Models Algorithms Genome Evol.*, 2013, pp. 29–45.
[25] V. Boža, B. Brejová, and T. Vinař, "GAML: Genome assembly by maximum likelihood," *Algorithms Mol. Biol.*, vol. 10, 2015, Art. no. 18.
[26] K. Sahlin, F. Vezzi, B. Nystedt, J. Lundeberg, and L. Arvestad, "BESST-efficient scaffolding of large fragmented assemblies," *BMC Bioinf.*, vol. 15, 2014, Art. no. 281.
[27] K. I. Bos, et al., "A draft genome of yersinia pestis from victims of the black death," *Nature*, vol. 478, pp. 506–510, 2011.
[28] P. Feijão and E. Araujo, "Fast ancestral gene order reconstruction of genomes with unequal gene content," *BMC Bioinf.*, vol. 17, no. 14, 2016, Art. no. 187.
[29] P. Avdeyev, S. Jiang, S. Aganezov, F. Hu, and M. A. Alekseyev, "Reconstruction of ancestral genomes in presence of gene gain and loss," *J. Comput. Biol.*, vol. 23, no. 3, pp. 150–164, 2016.
[30] P. Biller, P. Feijão, and J. Meidanis, "Rearrangement-based phylogeny using the single-cut-or-join operation," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 10, no. 1, pp. 122–134, Jan./Feb. 2013.
[31] J. Stoye and R. Wittler, "A unified approach for reconstructing ancient gene clusters," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 6, no. 3, pp. 387–400, Jul.-Sep. 2009.
[32] R. Wittler, Phylogeny-based analysis of gene clusters. PhD dissertation, Faculty Technol., Bielefeld Univ., Bielefeld, Germany, 2010.
[33] B. R. Jones, A. Rajaraman, E. Tannier, and C. Chauve, "ANGES: Reconstructing ancestral genomes maps," *Bioinf.*, vol. 28, no. 18, pp. 2388–2390, 2012.
[34] S. Bérard, C. Gallien, B. Boussau, G. J. Szöllősi, V. Daubin, and E. Tannier, "Evolution of gene neighborhoods within reconciled phylogenies," *Bioinf.*, vol. 28, pp. 382–388, 2012.
[35] S. Aganezov Jr , N. Sitdykova, and M. Alekseyev, "Scaffold assembly based on genome rearrangement analysis," *Comput. Biol. Chemistry*, vol. 57, pp. 46–53, 2015.
[36] Y. Anselmetti, V. Berry, C. Chauve, A. Chateau, E. Tannier, and S. Bérard, "Ancestral gene synteny reconstruction improves extant species scaffolding," *BMC Genomics*, vol. 16, no. Suppl 10, 2015, Art. no. S11.
[37] P. Berman and M. Karpinski, "On some tighter inapproximability results," in *Proc. Int. Colloq. Automata Languages Program.*, 1999, pp. 200–209.

Nina Luhmann studied computer science in the natural sciences at Bielefeld University, Germany. She has been working toward the PhD degree in the Genome Informatics Group at Bielefeld University and defended her PhD thesis in December 2016. She is working on integrated assembly of paleogenomes using both comparative ancestral reconstruction and ancient DNA data.

Manuel Lafond received the PhD degree from the University of Montreal, Canada, in 2016. He is now a NSERC postdoctoral fellow in the Department of Mathematics and Statistics, University of Ottawa, Canada, where he is developing new models and methods to distinguish orthologous genes from paralogous genes. His research interests mainly reside in the intersection of computational biology, algorithms, and graph theory.

Annelyse Thévenin received the PhD degree in the field of comparative genomics from the University of Paris-South, in 2009. She was a postdoctoral fellow in the group of Prof. Ron Shamir in the School of Computer Science, Tel Aviv University, Israel, working on linear and spatial cancer genome organizations and in the group of Prof. Jens Stoye in the Faculty of Technology, Bielefeld University, Germany, working on comparative genomics with gene family assignment-free. She has been a school teacher in France since last year.

Aïda Ouangraoua received the master's and PhD degrees in computer science from the Université de Bordeaux, France, in 2004 and 2007, respectively. She did her postdoctoral training at Simon Fraser University and the Université du Québec à Montréal, Canada. She was a researcher with INRIA Lille, France, from 2009 to 2014. She is currently a professor in the Department of Computer Science of Université de Sherbrooke, Québec, Canada. Her research interests include algorithms and computational molecular biology.

Roland Wittler studied computer science in the natural sciences and received the PhD degree in 2010, on a topic related to phylogeny-based gene-order analysis, from Bielefeld University. After a postdoctoral position at Simon Fraser University, Burnaby, BC, Canada, he became staff researcher with the Genome Informatics Group and scientific coordinator of a bioinformatics graduate program with Bielefeld University, Germany.

Cedric Chauve is a professor in mathematics at Simon Fraser University. His research interests include algorithms design and analysis, comparative genomics, genome rearrangements, and ancestral genomes reconstruction.