Scaffolding of Ancient Contigs and Ancestral Reconstruction in a Phylogenetic Framework

Nina Luhmann[1], Cedric Chauve[2], Jens Stoye[1], and Roland Wittler[1]

[1] International Research Training Group "Computational Methods for the Analysis of the Diversity and Dynamics of Genomes" and Genome Informatics, Faculty of Technology and Center for Biotechnology, Bielefeld University, Germany
[2] Department of Mathematics, Simon Fraser University, Burnaby (BC), Canada

Abstract. Ancestral genome reconstruction is an important step in analyzing the evolution of genomes. Recent progress in sequencing ancient DNA led to the publication of so-called paleogenomes and allows the integration of this sequencing data in genome evolution analysis. However, the assembly of ancient genomes is fragmented because of DNA degradation over time. Integrated phylogenetic assembly addresses the issue of genome fragmentation in the ancient DNA assembly while improving the reconstruction of all ancient genomes in the phylogeny. The fragmented assembly of the ancient genome can be represented as an assembly graph, indicating contradicting ordering information of contigs.
In this setting, our approach is to compare the ancient data with extant finished genomes. We generalize a reconstruction approach minimizing the Single-Cut-or-Join rearrangement distance towards multifurcating trees and include edge lengths to avoid a sparse reconstruction in practice. When also including the additional conflicting ancient DNA data, we can still ensure consistent reconstructed genomes.

1 Introduction

In comparative genomics, one aim is to analyze the diversity of genomes from present-day species to reconstruct the structure of ancient genomes and shed light on the dynamics of evolutionary processes underlying the development of extant genomes. The speciation history leading to the present-day genomes can be represented as a phylogenetic tree. Genome reconstruction methods aim to infer genomic features, such as gene order, at internal nodes of the tree by comparing conserved features in the extant genomes at its leaves, e. g. under parsimony assumptions. This problem has already been widely studied under different models and distance formulations [1, 2, 4, 6, 9, 12, 13].

Besides the phylogeny and the genome sequences of extant species, a third source of data for reconstruction became available recently. Due to the progress in sequencing technologies, ancient DNA (aDNA) found in conserved remains can be sequenced. One example is the genome of the ancestor of *Yersinia pestis* strains that is understood to be the cause of the Black Death pandemic [3]. However, enviromental conditions influence sources for paleogenomes and result in

degradation and fragmentation of DNA molecules over time, causing sequencing to produce very short reads [5]. This entails the assembly of aDNA to be specifically challenging and leads to a fragmented assembly with many short contigs requiring additional scaffolding. The purpose of the present work is to present a scaffolding method adapted to such datasets, within a phylogenetic framework.

So far, the only existing method specifically targeted at scaffolding aDNA contigs is FPSAC [11]. It follows a local approach concentrating on one internal node representing the ancestor of interest and was able to obtain a single scaffold from a fragmented assembly of the ancient *Yersinia pestis* strain. In this paper, we present a global approach for reconstructing all ancient genomes along a given phylogeny while also scaffolding the aDNA contigs obtained from a preliminary assembly for one internal node of the phylogeny. Contrary to FPSAC, our approach is global and can be described as an extension of the exact small parsimony algorithm minimizing the Single-Cut-or-Join distance described in [6] to the case of multifurcating phylogenetic trees with edge lengths. We show how this allows to handle, still with an exact polynomial time algorithm, constraints from the assembly graph of a sequenced ancestral genome.

2 Background

As a basis of this work, the data representation is described first, before the small parsimony problem under rearrangement distances is introduced.

2.1 Genome Representation

Both extant and ancient genomes are sets of chromosomes, plasmids or contigs. Each such component is represented by a sequence of oriented markers corresponding to homologous sequences, while each marker is contained once. Markers can be defined by alignment of assembled aDNA contigs onto the extant genomes (see [11] for example). To represent the orientation, any marker a has two extremities, a head a_h and a tail a_t. The order of markers in the genome can also be represented by adjacencies, which are unordered pairs of two extremities from neighboring markers, for example $\{a_h, b_t\}$. When one extremity is contained in two different adjacencies, these are said to be *conflicting*. Otherwise the genome can be written as a set of linear or circular sequences of markers and is *consistent*.

2.2 Augmented Phylogenetic Tree

The underlying general data structure for our studies, shown in Figure 1, is a phylogenetic tree $T = (V_T, E_T)$ representing the relations between extant species. Leaves annotated with assembled genome sequences correspond to extant species, internal nodes represent ancestral species. Edges are labeled with lengths describing the evolutionary distances in the tree. Furthermore, we assume that one internal node is augmented with an assembly graph $A = (V_A, E_A)$. We will refer to this augmented node as the *assembly graph node* and to the tree

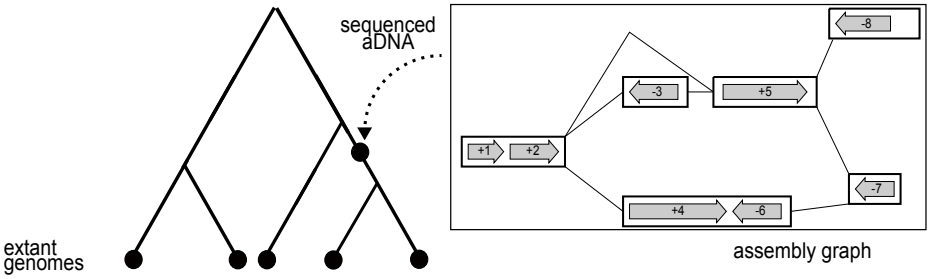

Fig. 1. Phylogenetic tree annotated with extant genomes at its leaves. One internal node is augmented with an assembly graph illustrating the fragmented assembly. It may contain conflicting adjacencies, e. g. $(2_h, 3_h)$ and $(2_h, 4_t)$.

together with the assembly graph as an *augmented phylogenetic tree*. An assembly graph is usually a de Bruijn or string graph connecting contiguous regions in the read data. Paths in the graph are then possible substrings, while branches indicate uncertainty about the exact genome sequence. For our purpose it is important to notice that since branching nodes in the graph connect one extremity with several others, they induce conflicting adjacencies.

2.3 Small Parsimony Problem under Rearrangement Distances

In order to reconstruct ancient genomes, we are starting from consistent genomes at the leaves of the considered phylogeny, represented by sets of adjacencies, and look for an optimal labeling, defined as a labeling minimizing a chosen genomic distance over the tree. This problem is known as the *small parsimony problem*.

Definition 1 (Parsimonious labeling). *Given a tree $T = (V_T, E_T)$ with each leaf l labeled with a state $b_l \in \{0, 1\}$, a labeling $\lambda : V \to \{0, 1\}$ with $\lambda(l) = b_l$ for each leaf l is* parsimonious *if it minimizes the overall distance d in the tree:*

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

In a simple setting, the distance is 0 if the label does not change along an edge and it is 1 otherwise. While for most rearrangement distances, the parsimonious labeling problem is NP-hard, one exception is the Single-Cut-or-Join (SCJ) distance introduced by Feijão and Meidanis [6], a set-theoretic rearrangement distance modeling *cuts* and *joins* of adjacencies.

Definition 2 (SCJ distance [6]). *Given two genomes defined by sets of adjacencies A and B, the SCJ distance between these genomes is*

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

An SCJ minimizing consistent labeling over a given phylogenetic tree can be computed by the Fitch algorithm [7] in polynomial time using a binary encoding, set to 1 if the adjacency is present in the genome and set to 0 otherwise. Reconstructed genomes then contain all adjacencies for which the internal node is labeled 1. Although adjacencies are not independent characters, it has been shown that the reconstruction of every adjacency separately assigns no conflicting adjacencies, provided that labels at all leaves are consistent and 0 is chosen in case of ambiguity at the root. The labeling for the rest of the tree is then unambiguous and provides valid genomes at internal nodes in polynomial time, minimizing the SCJ distance [6].

However, this reconstruction is sparse as it finds only the most fragmented under all co-optimal solutions. Some adjacencies will be excluded from the reconstructed genomes, although they could be included without causing conflicts. Furthermore, the Fitch algorithm can only handle binary trees and so excludes phylogenies that are not fully resolved. We will generalize the result of [6] towards multifurcating trees and show how to avoid the sparse approach. Later, in Section 4, we will show how to integrate the constraints of an assembly graph at a single ancestral node of the tree.

3 Edge-weighted Parsimony Problem Minimizing SCJ

Like the Fitch algorithm [7], the Hartigan algorithm [8] consists of a bottom-up and a top-down traversal of the phylogenetic tree. It is a generalization towards multifurcating trees and finds, in contrast to Fitch, *all* optimal parsimonious labelings. However, this more general algorithm induces ambiguity also at internal nodes of the tree. For the small parsimony problem with the SCJ distance, it can easily be shown that choosing 0 whenever possible, also at internal nodes of the tree, results in a consistent labeling, but this could result in an even sparser solution. Conversely, always including an adjacency in case of ambiguity can result in complex conflicts and would therefore require a subsequent conflict resolving step that is mindful of the tree structure. To avoid this, we propose to include edge lengths in the reconstruction and minimize an edge-weighted SCJ distance.

Definition 3 (Edge-weighted SCJ distance labeling problem). *Given a tree $T = (V_T, E_T)$ with each leaf labeled with adjacencies and each edge $e \in E_T$ labeled with an edge length $\ell(e)$, a labeling γ of the internal nodes of T is an* edge-weighted SCJ minimizing consistent labeling *if none of the internal nodes contains a conflict and it minimizes the overall tree distance*

$$D(\gamma, T) = \sum_{(u,v)\in E} \frac{d_{SCJ}(\gamma(u), \gamma(v))}{\ell((u,v))}.$$

3.1 Hartigan Algorithm with Edge Lengths

Consider the reconstruction for one adjacency α in a tree T. A leaf is labeled according to the absence or presence of α in its extant genome. In the bottom-

up phase, every node x with children $C(x)$ is annotated recursively with two candidate sets B_1^α and B_2^α. For both states $b \in \{0,1\}$, consider all children $y \in C(x)$ with $b \in B_1^\alpha(y)$ and let $k_x^\alpha(b) = \sum_{\substack{y \in C(x),\\ b \in B_1(y)}} \frac{1}{\ell((x,y))}$ and $K = max_b\{k_x^\alpha(b)\}$. Then, if e is the edge connecting x to its parent, let

$$B_1^\alpha(x) = \{b \mid k_x^\alpha(b) = K\}, \qquad B_2^\alpha(x) = \{b \mid k_x^\alpha(b) = K - \frac{1}{\ell(e)}\}.$$

In the top-down phase, the root r is assigned with a state $b \in B_1^\alpha(r)$. When processing node x, let b be the state assigned to it in an optimal labeling. Then a child $y \in C(x)$ is labeled as follows:

$$F^\alpha(y) = \begin{cases} \{b\} & \text{if } b \in B_1^\alpha(y) \\ \{b\} \cup B_1^\alpha(y) & \text{if } b \in B_2^\alpha(y) \\ B_1^\alpha(y) & \text{otherwise} \end{cases}$$

We note that the set B_1^α is likely to be of cardinality one and the set B_2^α is likely to be empty in real data sets, where edges are annotated with non-trivial edge lengths such as rational numbers. Therefore the second case rarely occurs and there is often no choice in the other cases. Hence in most real instances, there will be a unique most parsimonious labeling for all adjacencies.

3.2 Reconstructing Consistent Genomes

Following the proof of Lemma 6.1 in [6], we can show that also the edge-weighted Hartigan algorithm assigns consistent genomes. We still assume a sparse variant of the algorithm where the label 0 is chosen during the top-down phase any time there is an ambiguity and call it *sparse edge-weighted Hartigan algorithm.*

Lemma 1. *Given two conflicting adjacencies α and β, for each node x of a tree T labeled according to the sparse edge-weighted Hartigan algorithm, if $B_1^\alpha(x) = \{1\}$, then $B_1^\beta(x) = \{0\}$ if $\nexists$ leaf l with both $B_1^\alpha(l) = \{1\}$ and $B_1^\beta(l) = \{1\}$.*

Proof. The proof is by induction on the height h of a node x in the tree, which is the maximal length from x to any descendant leaf. For $h = 0$, the node is a leaf annotated with a consistent genome, therefore the lemma holds.

When $h \geq 1$, we assume that any node with height $g < h$ and therefore all children of x satisfy the lemma. We denote the sum of edge lengths from x to all children $y \in C(x)$ annotated with $B_1^\alpha(y) = \{0\}$ with l_0^α, to children annotated with $B_1^\alpha(y) = \{1\}$ with l_1^α and edge lengths to children annotated with $B_1^\alpha(y) = \{0,1\}$ with l_{01}^α. For the two states 0 and 1, we sum the weights to the appropriate children and have $k_x^\alpha(0) = l_0^\alpha + l_{01}^\alpha$ and $k_x^\alpha(1) = l_1^\alpha + l_{01}^\alpha$. Now we can make some observations about the relation of sets B_1 in the children of x according to the Hartigan algorithm. $B_1^\alpha(x) = \{1\}$ only if $k_x^\alpha(1) > k_x^\alpha(0)$ and thus $l_1^\alpha + l_{01}^\alpha > l_0^\alpha + l_{01}^\alpha \Rightarrow l_1^\alpha > l_0^\alpha$.

Next we consider the second adjacency β that is in conflict with α. As by induction hypothesis any child y satisfies the lemma, at least all children with

$B_1^\alpha(y) = \{1\}$ have to have $B_1^\beta(y) = \{0\}$ and thus $l_0^\beta \geq l_1^\alpha$. On the other hand, only children with $B_1(\alpha, y) = \{0\}$ can have $B_1(\beta, y) = \{1\}$, so $l_1^\beta \leq l_0^\alpha$.

Taking these three observations together, we can derive that $l_0^\beta \geq l_1^\alpha > l_0^\alpha \geq l_1^\beta$, therefore $l_0^\beta + l_{01}^\beta > l_1^\beta + l_{01}^\beta$, which is the same as $k_x^\beta(0) > k_x^\beta(1)$, implying $B_1^\beta(x) = \{0\}$. Therefore when $B_1^\alpha(x) = \{1\}$, we have $B_1^\beta(x) = \{0\}$. □

Lemma 2. *Given two conflicting adjacencies α and β, for each node x of a tree T labeled according to the sparse edge-weighted Hartigan algorithm, if $F^\alpha(x) = \{1\}$, then choosing $F^\beta(x) = \{0\}$ is always possible.*

Proof. As the edge weights are only influencing the bottom-up phase, we do not have to consider them in the top-down phase. Suppose there are internal nodes with value 1 assigned to both α and β. Choose such a node with minimal distance to the root and call it v. We have different possibilities for B_1 and B_2 of v according to α and β, summarized in Table 1. Note that by Lemma 1, $B_1^\alpha(v) = B_1^\beta(v) = \{1\}$ cannot occur. In cases 1–4, choosing 0 for α or β is always

Table 1. Case differentiation for bottom-up sets that fulfill Lemma 1 and could assign label 1 for both adjacencies α and β.

1	2	3	4	5
$B_1^\alpha(v) = \{1\}$	$B_1^\alpha(v) = \{0\}$	$B_1^\alpha(v) = \{0\}$	$B_1^\alpha(v) = \{1,0\}$	$B_1^\alpha(v) = \{1,0\}$
$B_2^\alpha(v) = \{0\} \vee \emptyset$	$B_2^\alpha(v) = \{1\}$	$B_2^\alpha(v) = \{1\}$	$B_2^\alpha(v) = \emptyset$	$B_2^\alpha(v) = \emptyset$
$B_1^\beta(v) = \{0\}$	$B_1^\beta(v) = \{1\}$	$B_1^\beta(v) = \{0\}$	$B_1^\beta(v) = \{0\}$	$B_1^\beta(v) = \{1,0\}$
$B_2^\beta(v) = \{1\}$	$B_1^\beta(v) = \{0\} \vee \emptyset$	$B_1^\beta(v) = \{1\}$	$B_2^\beta(v) = \{1\}$	$B_2^\beta(v) = \emptyset$

possible independent of the parent assignment. In case 5, the parent assignment for both α and β has to be 1 in order to also assign 1 to v. This, however, contradicts the minimality of the depth of v and therefore concludes the proof. □

Theorem 1. *For a rooted phylogenetic tree T with leaves annotated with consistent genomes containing the same set of markers, the sets $G_v = \{\alpha : F^\alpha(v) = 1\}$ assigned to all internal nodes v with the sparse edge-weighted Hartigan algorithm are consistent genomes and minimize the edge-weighted SCJ distance.*

Proof. According to Theorem 6.3 in [6], including the adjacency α in every node v, where $F^\alpha(v) = 1$, builds genomes that minimize the SCJ distance over T. Lemma 2 shows that also with the sparse edge-weighted Hartigan algorithm no conflicting adjacencies will be assigned to a node v. Therefore the sets G_v minimize the total sum of SCJ cost per edge length. □

4 Integrating aDNA Sequencing Information

The assembly graph based on ancient sequencing reads (cf. Figure 1) defines putative adjacencies between markers on connected contigs. These adjacencies

constrain the reconstruction by providing evidence of the genome structure directly seen at an internal point in the tree. We include these constraints by extending the original tree with an additional leaf attached to the assembly graph node. This leaf will be labeled with the presence or absence of an adjacency in the assembly graph, just like the leaves representing extant genomes. The respective edge length $\ell(e)$ has to be chosen in regard to the other two connected edges a and b of the assembly graph node such that the additional information is relevant at all but not generally dominating. Hence it has to be chosen such that $\frac{1}{\ell(e)}$ is in the interval $[(\frac{1}{\ell(a)} - \frac{1}{\ell(b)}), (\frac{1}{\ell(a)} + \frac{1}{\ell(b)})]$ for $\ell(a) < \ell(b)$, where a smaller edge length gives the assembly graph more importance.

However the set of adjacencies present in the assembly graph is not necessarily consistent and can cause conflicts. Instead of adding a postprocessing step that resolves all the conflicts in the tree after the reconstruction, we propose in Algorithm 1 an approach that integrates the conflicts resolving into the reconstruction process. To resolve conflicts, we rely on the exact polynomial time MAX-ROW-component-mCi1P algorithm described in [10]. It selects a subset of adjacencies that form a set of linear and/or circular chromosomes based on a maximum-weight matching in a graph.

Algorithm 1 Consistent reconstruction integrating aDNA sequencing data

Input: A tree T with edge lengths, extant consistent genomes, aDNA assembly graph
Output: A consistent labeled tree minimizing the edge-weighted SCJ distance.

Attach an additional leaf to the assembly graph node v
Reroot the tree such that v becomes its root
for each adjacency α **do**
  for each internal node x in T **do**
    Compute $B_1^\alpha(x)$ and $B_2^\alpha(x)$ with the sparse edge-weighted Hartigan algorithm
$A = \{\alpha | 1 \in B_1^\alpha(v)\}$
Solve MAX-ROW-component-mCi1P for A
for each adjacency α **do**
  for each internal node x in T **do**
    Compute $F^\alpha(x)$ with the sparse edge-weighted Hartigan algorithm

Theorem 2. *Given an augmented phylogenetic tree, Algorithm 1 computes a consistent labeling integrating the assembly graph information and minimizing the edge-weighted SCJ distance in polynomial time.*

Proof. Acoording to Theorem 1, the sparse edge-weighted Hartigan algorithm assigns consistent, SCJ minimizing genomes when the leaf labels are consistent. Rerooting the tree will not affect the outcome of the reconstruction. In the bottom-up phase, the conflicting leaf will only influence the assignment at the root. All other internal nodes fulfill Lemma 1, as the original leaves are consistently labeled. Therefore they cannot cause a conflicting assignment in the top-down phase when the parent assignment is consistent. As conflicts can thus

only occur at the root node, they have to be resolved with a minimal increase in parsimony costs before propagating the assignment down the tree during the top-down phase. Selecting a maximum cardinality subset of all adjacencies assigned to the root can be done by solving the MAX-ROW-component-mCi1P [10]. With a consistent root labeling, the top-down assignment will be consistent according to Lemma 2.

The traversal of the tree with n leaves and a adjacencies takes $O(an)$ time. The MAX-ROW-component-mCi1P can be solved in $O(a^{3/2})$ [10]. Therefore the overall running time is in $O(an + a^{3/2})$. □

5 Conclusion

We have described a generalization of the exact algorithm solving the small parsimony problem under the SCJ rearrangement distance. Computing the labeling of internal nodes with the Hartigan algorithm enables handling multifurcating trees. Including edge lengths still ensures the reconstruction of valid genomes and is also expected to provide a unique optimal solution under non-trivial edge lengths in practice. Building upon this result, we presented an integrated phylogenetic assembly approach. It includes aDNA sequencing information in the reconstruction of other ancient genomes in the phylogeny and also scaffolds the fragmented assembly while minimizing the SCJ distance.

Among the questions our work raises, it would be interesting to study model variants that allow to integrate copy numbers or unequal marker content. Another question of interest would be to design efficient heuristics or parameterized algorithms to augment an initial parsimonious consistent labeling with extra adjacencies that preserve both parsimony and consistency.

Acknowledgements. NL and RW are funded by the International DFG Research Training Group GRK 1906/1.

References

[1] Bergeron, A., Blanchette, M., Chateau, A., Chauve, C.: Reconstructing ancestral gene orders using conserved intervals. In: Proc. of WABI 2004, LNBI, vol. 3240, pp. 14–25. Springer (2004)

[2] Bertrand, D., Gagnon, Y., Blanchette, M., El-Mabrouk, N.: Reconstruction of ancestral genome subject to whole genome duplication, speciation, rearrangement and loss. In: Proc. of WABI 2012, LNBI, vol. 6293, pp. 78–89. Springer (2010)

[3] Bos, K.I., Schuenemann, V.J., Golding, G.B., et al.: A draft genome of yersinia pestis from victims of the black death. Nature 478(7370), 506–510 (2011)

[4] Chauve, C., Tannier, E.: A methodological framework for the reconstruction of contiguous regions of ancestral genomes and its application to mammalian genomes. PLoS Computational Biology 4(11), e1000234 (2008)

[5] Drancourt, M., Raoult, D.: Palaemicrobiology: current issues and perspectives. Nature Rev Microbiol 3, 23–35 (2005)

[6] Feijão, P., Meidanis, J.: SCJ: a breakpoint-like distance that simplifies several rearrangement problems. IEEE/ACM Trans. Comput. Biol. and Bioinf. 8(5), 1318–1329 (2011)
[7] Fitch, W.M.: Toward defining the course of evolution: minimum change for a specific tree topology. Systematic Biology 20(4), 406–416 (1971)
[8] Hartigan, J.A.: Minimum mutation fits to a given tree. Biometrics pp. 53–65 (1973)
[9] Ma, J., Zhang, L., Suh, B.B., et al.: Reconstructing contiguous regions of an ancestral genome. Genome Res. 16(12), 1557–1565 (2006)
[10] Maňuch, J., Patterson, M., Wittler, R., Chauve, C., Tannier, E.: Linearization of ancestral multichromosomal genomes. BMC Bioinformatics 13(19), S11 (2012)
[11] Rajaraman, A., Tannier, E., Chauve, C.: FPSAC: fast phylogenetic scaffolding of ancient contigs. Bioinformatics 29(23), 2987–2994 (2013)
[12] Stoye, J., Wittler, R.: A unified approach for reconstructing ancient gene clusters. IEEE/ACM Trans. Comput. Biol. Bioinf. 6(3), 387–400 (2009)
[13] Zheng, C., Sankoff, D.: On the pathgroups approach to rapid small phylogeny. BMC Bioinformatics 12(S-1), S4 (2011)