Algorithms in Comparative Genomics

Lecture notes

Faculty of Technology, Bielefeld University

Winter 2021/22

CHAPTER 1

Basic Definitions

Contents of this chapter: Genome representation, distinguished types of genome (pairs) as well as the general framework of problems that will be examined under the perspective of different models in these lecture notes.

1.1 Representing Genomes

An initial genome representation the reader is likely familiar with are contigs and scaffolds as a result of sequencing projects. These are typically long DNA sequences of which the position in the genome is known. However, genomes can undergo a variety of mutations that are not structural in nature, such as Single Nucleotide Variants (SNV). These mutations act as a local distortion when examining larger, structural changes. Therefore it is advisable to abstract from the concrete DNA sequence and view genomes as comprised of genomic *markers* located on chromosomes. These markers are sometimes also referred to as *genes*, but may also encompass other sequence areas beside biological genes. Markers form the alphabet of the chromosomal sequences, this alphabet being unique to each genome.

Definition 1 *The* set of markers *of a genome* $\mathbb{A}$ *is* $\mathcal{G}(\mathbb{A}) = \{\mathbb{A}[1], ..., \mathbb{A}[n]\}$ *for some* $n \in \mathbb{N}$.

We aim to define a chromosome of genome $\mathbb{A}$ as a sequence of markers. We should first remember though, that biological genes, as well as other possible marker types, are typically located on a DNA double strand. Therefore they can be found either on the forward or backward strand. In order to distinguish these two cases, we introduce the inversion operator signified via a bar over the respective marker. Then $\overline{\mathbb{A}[i]}$ represents a marker on the backward strand, while $\mathbb{A}[i]$ is still used as the representation for a marker on the forward strand. Naturally, applying the operator twice yields a marker on the forward strand again, i.e. $\overline{\overline{\mathbb{A}[i]}} = \mathbb{A}[i]$. We also extend the definition of the operator to sets with $\overline{\{m_1, ..., m_k\}} =$

$\{\overline{m_1}, ..., \overline{m_k}\}$ and sequences with $\overline{m_1...m_k} = \overline{m_k}...\overline{m_1}$. Note that in case of sequences the operator also reverses the order in which characters appear.

We are now able to define chromosomes as strings of oriented markers with the inversion operator indicating orientation. We distinguish linear and circular chromosomes by enclosing them with different brackets.

Definition 2 *A* chromosome *of genome $\mathbb{A}$ is denoted as a string s enclosed by brackets indicating the type of chromosome: $[s]$ if it is linear and as (s) if it is circular where s is an oriented string of markers, $s \in (\mathcal{G}(\mathbb{A}) \cup \overline{\mathcal{G}(\mathbb{A})})^+$, such that no marker appears both in its forward and backward form and no two chromosomes share a marker.*

A visual example on how our abstract notation of chromosomes is derived from a more conventional representation can be found in Figure 1.1 in Subfigures A and B.

As our goal is to compare different genomes to each other, it is necessary to establish some kind of equivalence relation between markers of different genomes or even the same genome. Usually this relation is defined via some measure of sequence similarity between the markers. As the process of creating such an equivalence relation is a complete field of study in itself, we will limit our perspective here and presume that this step has been dealt with. In analogy to biological gene families we assume a *family assignment* f that gives the family $f(m)$ for each marker m. Markers that are assigned the same family are then treated as equivalent. We call a genome $\mathbb{A}$ in combination with a family assignment f an *annotated genome* $\mathbb{A}^f$.

Definition 3 *The* set of families *in an annotated genome $\mathbb{A}^f$ is*

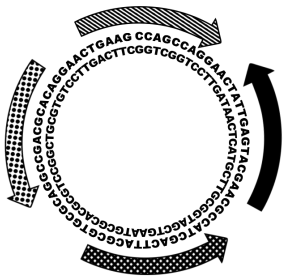
$$\mathcal{F}(\mathbb{A}^f) := \{f(m) : m \in \mathcal{G}(\mathbb{A})\}.$$

We will use this definition throughout these lecture notes, except for the parts in Chapter **??** in which we examine the *family free* approach which integrates the step of establishing equivalence between markers into the formulation of the *genomic distance problem.*

For human readability an oriented marker m or $\overline{m}$ is typically instead identified via its family, e.g. as $f(m)$ or $\overline{f(m)}$ respectively. You can see how this simplifies notation while also making family relations more obvious in Figure 1.1 in the difference between Subfigures B and C. Much of the comparative genomics literature does not make a difference between the identity of a marker and its family and therefore uses only this family based notation. This is due to the fact that in most of the early works in the field only genomes with one marker per family are considered. As we would like to also explore more modern approaches like family-free models later, we use this distinction between marker and family for our definitions. Therefore there might be some discrepancies between these lecture notes and some of the cited works in this regard.

While a notation based on chromosomes as strings of markers is well suited for human readability, it is typically not used in algorithmic contexts, because it entails some redundancies. For example a chromosome may be displayed in forward $[m_1...m_n]$ as well as in backward $[\overline{m_1...m_n}] = [\overline{m_n}...\overline{m_1}]$ direction while remaining functionally the same. A similar problem is found with circular chromosomes, in which one way to "cut" the circular sequence $(m_1m_2...m_n)$ in order to notate it linearly is equivalent to any other way of doing so, i.e. $(m_2...m_nm_1), (m_3...m_1m_2)$,...,$(m_n...m_{n-2}m_{n-1})$. This results in several representations, the

(A)



(B) $(\,\mathbb{A}[1]\,\overline{\mathbb{A}[2]}\,\overline{\mathbb{A}[3]}\,\overline{\mathbb{A}[4]}\,)$ $[\,\mathbb{A}[5]\,\overline{\mathbb{A}[6]}\,\mathbb{A}[7]\,]$

(C) $(\,1\,\bar{3}\,\bar{4}\,\bar{5}\,)$ $[\,1\,\bar{2}\,5\,]$

(D) $\mathbb{A}[1]^h\mathbb{A}[2]^h$ $\mathbb{A}[2]^t\mathbb{A}[3]^h$ $\mathbb{A}[3]^t\mathbb{A}[4]^h$ $\mathbb{A}[4]^t\mathbb{A}[1]^t$ $\mathbb{A}[5]^t$ $\mathbb{A}[5]^h\mathbb{A}[6]^h$ $\mathbb{A}[6]^t\mathbb{A}[7]^t$ $\mathbb{A}[7]^h$

(E) 1^h3^h 3^t4^h 4^t5^h 5^t1^t 1^t 1^h2^h 2^t5^t 5^h

Figure 1.1: The same genome $\mathbb{A}$ with a linear and a circular chromosome described in five different ways: (A) as DNA-sequences annotated with markers, sequence similarity displayed via shading, (B) as a marker based genome, (C) as an annotated genome with family assignment f (see Table 1.1), (D) as a collection of marker based adjacencies and (E) as a collection of family adjacencies and telomeres.

m	$f(m)$
$\mathbb{A}[1]$	1
$\mathbb{A}[2]$	3
$\mathbb{A}[3]$	4
$\mathbb{A}[4]$	5
$\mathbb{A}[5]$	1
$\mathbb{A}[6]$	2
$\mathbb{A}[7]$	5

Table 1.1: Family assignment f for annotated genome $\mathbb{A}^f$ in Figure 1.1

	cl	$\bar{c}l$	$\bar{c}\bar{l}$	$c\bar{l}$
c_1	$(12\bar{3}), [4\bar{5}6]$	$(3\bar{2}\bar{1}), [4\bar{5}6]$	$(3\bar{2}\bar{1}), [\bar{6}5\bar{4}]$	$(12\bar{3}), [\bar{6}5\bar{4}]$
c_2	$(2\bar{3}1), [4\bar{5}6]$	$(\bar{1}3\bar{2}), [4\bar{5}6]$	$(\bar{1}3\bar{2}), [\bar{6}5\bar{4}]$	$(2\bar{3}1), [\bar{6}5\bar{4}]$
c_3	$(\bar{3}12), [4\bar{5}6]$	$(\bar{2}\bar{1}3), [4\bar{5}6]$	$(\bar{2}\bar{1}3), [\bar{6}5\bar{4}]$	$(\bar{3}12), [\bar{6}5\bar{4}]$

Table 1.2: All 12 ways of notating genome $\{(12\bar{3}), [4\bar{5}6]\}$ by different cuttings of the circular chromosome c_1, c_2, c_3 as well as the four notations via inversions of chromosomes $cl, \bar{c}l, \bar{c}\bar{l}, c\bar{l}$.

number of which is exponential in the number of chromosomes. For a detailed example of this effect, see Table 1.2.

A different, less redundant notation is the *adjacency* notation. In this notation the markers are not viewed as a whole, but instead as their *extremities*, their beginnings and ends. The beginning of a marker m is referred to as the *tail* and notated as m^t, the end of a marker as the *head* with notation m^h. Thus a forward marker m is split into m^t, m^h and a backward marker $\overline{m}$ into m^h, m^t. The structure of the genome is then defined by which of these extremities are in direct vicinity to one another. Trivially, the two extremities of the same marker are always adjacent to one another, so we are primarily interested in whether the extremities of different markers are neighboring. If this is the case, we call the resulting pair of extremities an *adjacency*. As we have seen before, chromosomes can be read in one direction or in the other and therefore this pair is unordered, that is an adjacency $m^x n^y$ stays the same when read backwards as $n^y m^x$. We formally define adjacencies as follows:

Definition 4 *The* set of adjacencies *of a circular chromosome $C = (s)$ is*

$$\Gamma_C = \{m_i^x m_{i+1}^y : m_i^x m_{i+1}^y \textit{ is a substring of } s'\} \cup \{m_1^u m_n^w\}$$

with $s' = m_1^u ... m_n^w = SPLIT(s)$.

The set of adacencies in a linear chromosome $L = [t]$ is

$$\Gamma_L = \{m_i^x m_{i+1}^y : m_i^x m_{i+1}^y \textit{ is a substring of } s'\}$$

with $t' = SPLIT(t)$, where

$$SPLIT(\varepsilon) = \varepsilon,$$

$$SPLIT(a\ s) = a^t\ a^h\ SPLIT(s),$$

$$SPLIT(\overline{a}\ s) = a^h\ a^t\ SPLIT(s)$$

The set of adjacencies of a genome $\mathbb{A}$ is $\Gamma(\mathbb{A}) = \bigcup_{C \text{ chromosome of } \mathbb{A}} \Gamma_C$.

For example the circular chromosome $(\mathbb{B}[1]\,\mathbb{B}[2]\,\overline{\mathbb{B}[3]})$ yields three adjacencies, $\mathbb{B}[1]^h\mathbb{B}[2]^t$, $\mathbb{B}[2]^h\mathbb{B}[3]^h$ and $\mathbb{B}[3]^t\mathbb{B}[1]^t$. The particularly attentive reader might already have noticed that there is a special case, in which extremities of the same marker form an adjacency by the above definition. This occurs if there is a circular chromosome with just one marker. For example the chromosome $(\mathbb{C}[6])$ yields the adjacency $\mathbb{C}[6]^h\mathbb{C}[6]^t$. This is not a bug but a feature of this definition, because in this case the extremities neighbor not just because they are part of the same marker, but because of the structure of the chromosome, which we want to capture using the adjacencies.

Additional to this special case resulting from the markers found on the end of the string representation of a circular chromosome, we find extremities at the ends of linear chromosomes, which do not have any neighboring extremities. We refer to these special extremities as *telomeres*.

Definition 5 *The* set of telomeres *of a linear chromosome* $L = [t]$ *is*

$$\Theta_L = \{m_1^x, m_n^y\},$$

given $m_1^x ... m_n^y = SPLIT(t)$.

The set of telomeres of a genome $\mathbb{A}$ *is* $\Theta(\mathbb{A}) = \bigcup_{\substack{C \text{ linear} \\ \text{chromosome} \\ \text{of } \mathbb{A}}} \Theta_C$

For example the linear chromosome $[\mathbb{D}[4]\,\overline{\mathbb{D}[5]}\,\mathbb{D}[6]]$ yields the telomeres $\mathbb{D}[4]^t$ and $\mathbb{D}[6]^h$. A full example of deriving adjacencies and telomeres of a genome can be found in Figure 1.1 by comparing B and D.

Remark 1 *The definition of a genome by chromosomes as strings of markers and the definition as a set of adjacencies and a set of telomeres are equivalent. This is due to the fact that the different notations can be obtained from each other. The "forward" direction from strings of markers to adjacencies and telomeres is already clear from their definitions (see Def. 4 and 5). For the "backward" direction we note that we can reconstruct each chromosomal sequence by starting at one extremity and following the "breadcrumb trail" of alternatingly chosing the next extremity either from the adjacency or the other extremity from the marker. The type of chromosome is determined by whether the so obtained sequence ends in a telomere or wraps around.*

As mentioned before, one is typically less interested in the markers themselves as in their families. Therefore it is helpful to also define adjacencies and telomeres in terms of family.

Definition 6 *Given a genome* $\mathbb{A}$*, its adjacencies* $\Gamma(\mathbb{A})$ *and a family assignment* f*, the multiset* $\Gamma(\mathbb{A}^f) = \{f(m)^x f(n)^y : m^x n^y \in \Gamma(\mathbb{A})\}$ *is referred to as the* family adjacencies *of* $\mathbb{A}$.

Definition 7 *Given a genome* $\mathbb{A}$*, its telomeres,* $\Theta(\mathbb{A})$ *and a family assignment* f*, the multiset* $\Theta(\mathbb{A}^f) = \{f(m)^x : m^x \in \Theta(\mathbb{A})\}$ *is referred to as the* family telomeres *of* $\mathbb{A}$.

Note that in contrast to the set of families $\mathcal{F}(\mathbb{A}^f)$ these objects here are multisets, that is if an adjacency appears twice in the genome, it is also "counted twice". The simplest example for this behavior is the annotated unichromosomal genome $\{(1\ 1)\}$ with family adjacencies $\{1^h 1^t, 1^h 1^t\}$. We also observe that another annotated genome, namely $\{(1), (1)\}$, has the same family adjacencies. We therefore remark:

Remark 2 *In contrast to marker based adjacencies and telomeres, a genome is not necessarily completely defined by giving its family based adjacencies and telomeres.*

The family adjacencies and telomeres of the example genome we have been following in this section can be found in Figure 1.1.E.

1.2 Types of Genomes and Genome Pairs

Regardless of notation the number of markers per family influences the conception and oftentimes also the complexity of computational problems in comparative genomics. We define it as follows:

Definition 8 *The* number of markers of a family $l \in \mathcal{F}(\mathbb{G}^f)$ *in an annotated genome* $\mathbb{G}^f$ *is*

$$\Phi(l, \mathbb{G}^f) = |\{m \in \mathcal{G}(\mathbb{G}) : \ f(m) = l\}|.$$

We will now distinguish different types of genomes based on number of markers per family. As discussed in Section 1.1, a sizeable portion of the comparative genomics literature considers only genomes with one marker per family. We refer to this simplest type of annotated genome as a *singular genome*.

Definition 9 *An annotated genome* $\mathbb{G}^f$ *is referred to as* singular *if*

$$\forall l \in \mathcal{F}(\mathbb{G}^f) : \ \Phi(l, \mathbb{G}^f) = 1.$$

It may be notated as $\mathbb{G}^f_{\triangleright}$.

We can now understand, why it is not necessary to have separate definitions for marker based and family adjacencies and telomers when considering singular genomes only. Because for each family l holds $\Phi(l, \mathbb{G}^f) = 1$, we know that f is a bijection between the set of markers and the set of families of that genome. Thus an inverse function f^{-1} exists that maps each family back to its marker. Therefore we can also map family adjacencies and telomeres back to marker based adjacencies and telomeres. In conjunction with Remark 1 we conclude the following.

Remark 3 *A singular genome is sufficiently described by giving its family adjacencies and telomeres.*

Note that if f were not bijective this would not be the case as can be seen in Remark 2.

Another type of genome that is also often found in classical comparative genomics literature is the *duplicated genome*. As the name suggests, in a duplicated genome every family occurs exactly twice, thus

Definition 10 *An annotated genome* $\mathbb{G}^f$ *is referred to as* duplicated *if*

$$\forall l \in \mathcal{F}(\mathbb{G}^f) : \ \Phi(l, \mathbb{G}^f) = 2.$$

It may be notated as $\mathbb{G}^f_{\diamond}$.

Note that the word "duplicated" only refers to the occurrences of familes and has nothing to do with the chromosomal structure, for example $\{(13\bar{2}2), [\bar{1}3]\}$ is a duplicated genome although its two chromosomes are not structurally similar. The subclass of duplicated genomes in which the chromosomal structure is also duplicated is called *perfectly duplicated* genomes. In these genomes besides every family occurring exactly twice, every family adjacency and telomere occurs twice.

Definition 11 *A duplicated genome* $\mathbb{G}^f$ *is referred to as* perfectly duplicated *if any family adjacency has multiplicity 2 in the multiset of family adjacencies* $\Gamma(\mathbb{G}^f)$*. It may be notated as* $\mathbb{G}^f_{\bowtie}$*.*

This property causes the genome to be easily divided into two identical sub-genomes and thus a perfectly duplicated genome $\mathbb{A}'^f_{\bowtie}$ can be viewed as the direct result of a whole genome duplication of a singular genome $\mathbb{A}^f_{\triangleright}$. We then refer to $\mathbb{A}'^f_{\bowtie}$ as a *doubled genome* of $\mathbb{A}^f_{\triangleright}$, more precisely:

Definition 12 *A* doubled genome $\mathbb{A}'^f_{\bowtie}$ *of singular genome* $\mathbb{A}^f_{\triangleright}$ *is a perfectly duplicated genome with doubled marker set* $|\mathcal{G}(\mathbb{A}')| = 2|\mathcal{G}(\mathbb{A})|$*, such that* $\Gamma(\mathbb{A}'^f_{\bowtie}) = \Gamma(\mathbb{A}_{\triangleright}{}^f) \oplus \Gamma(\mathbb{A}_{\triangleright}{}^f)$ *and* $\Theta(\mathbb{A}'^f_{\bowtie}) = \Theta(\mathbb{A}_{\triangleright}{}^f) \oplus \Theta(\mathbb{A}_{\triangleright}{}^f)$*, where* $\oplus$ *denotes the sum of two multisets, adding the multiplicities of each element.*

Remark 4 *Notice that there are two ways to double a circular chromosome. For example* $(1\,2\,3)$ *can be doubled as* $(1\,2\,3),(1\,2\,3)$ *or as* $(1\,2\,3\,1\,2\,3)$*. You can check this by comparing the family adjacencies. Therefore, if a genome contains at least one circular chromosome, there are multiple doubled genomes of it.*

To capture this phenomenon, we define a set of these different doublings for later use.

Definition 13 *We refer to the* set of doubled genomes of a singular genome $\mathbb{A}^f_{\triangleright}$ *as* $2 \cdot \mathbb{A}^f_{\triangleright}$*.*

As an example, regard the singular genome $\mathbb{A}^f_{\triangleright} = \{(1\,2),(3\,4)\}$. There are four doubled genomes of $\mathbb{A}^f_{\triangleright}$, namely

$$\begin{aligned} 2 \cdot \mathbb{A}^f_{\triangleright} = \{&\{(1\,2),(3\,4),(1\,2),(3\,4)\},\\ &\{(1\,2\,1\,2),(3\,4),(3\,4)\},\\ &\{(1\,2),(1\,2),(3\,4\,3\,4)\},\\ &\{(1\,2\,1\,2),(3\,4\,3\,4)\}\}. \end{aligned}$$

The most general type of genome, in which each family may occur an arbitrary number of times is called *natural genome*. All of the above classes are subclasses of natural genomes. An example of a natural genome that is not part of the other classes is $\{(12\bar{2}\bar{2}),[34]\}$.

As the most fundamental questions in comparative genomics involve exactly two genomes, we will also classify pairs of genomes. An important distinction here is whether two genomes have the same number of markers per family. If this is the case the pair is called *balanced*.

Definition 14 *A pair of annotated genomes* $\mathbb{A}^f, \mathbb{B}^f$ *under the same family assignment* f *is referred to as* balanced *if*

$$\forall l \in \mathcal{F}(\mathbb{A}^f) \cup \mathcal{F}(\mathbb{B}^f) : \Phi(l, \mathbb{A}^f) = \Phi(l, \mathbb{B}^f).$$

When regarding rearrangement scenarios between balanced genomes it is not necessary to involve operations that can substitute, delete or insert markers as they can be transformed into each other by pure rearrangements only. Note that this property is strictly referring to

a pair and meaningless for a single genome. For example $\{(1223)\}, \{(13), (2\bar{2})\}$ is a balanced pair while $\{(1223)\}, \{(123)\}$ is not, although the first genome is the same in each pair.

Another factor, that typically influences problem complexity rather than conception, is whether both genomes are singular. We then refer to the pair as singular.

Definition 15 *A pair of annotated genomes $\mathbb{A}^f, \mathbb{B}^f$ is referred to as* singular *if both $\mathbb{A}^f$ and $\mathbb{B}^f$ are singular genomes.*

Most problems are computationally easier to solve for singular pairs. The most restricted case being a *canonical* pair, which is both singular and balanced.

Definition 16 *A pair of annotated genomes $\mathbb{A}^f, \mathbb{B}^f$ is referred to as* canonical *if it is singular and balanced.*

Again, for the most general class of genomes, we refer to a pair which is not necessarily balanced or singular as a *natural* pair. Many problems in these lecture notes are regarded only for canonical genome pairs as this is usually the first subclass of natural pairs for which an algorithm is developed. In practice oftentimes genomes are heuristically reduced to canonical genomes for example by disregarding markers exclusive to one genome and using only one candidate pair per family.

1.3 Classic Problems in Comparative Genomics

There have been many models in Comparative Genomics, ranging from fundamental comparisons like the Breakpoint model over simple operation-based models like SCJ, DCJ or Inversion to complex models like the DCJ-Indel model, all of which will be examined in these lecture notes. While these models may appear very different from one another, it is essential to remember that they are all attempting to answer the same questions arising from the comparative study of genomes. In this section these questions will be presented in an abstract manner, such that the reader may refer back to it when a certain problem is discussed under a new model later on.

1.3.1 Genomic Distance Problem

The most fundamental question to be asked when given two present day genomes is how much evolutionary time has passed during the development of these two structures from a common ancestor. The *(genomic) distance problem* attempts to answer this question by attributing a number expressing that distance to a given pair of genomes. This number typically represents a number of transformations - like inversions or translocations - that need to be applied to one genome in order to transform it into the other one. This is similar to how the Edit Distance measures how many Indels and Substitutions are needed to transform one sequence to another. Note that this distance always integrates over two paths in the evolutionary tree, one going from the first genome backward in time to the common ancestor and one from the common ancestor forward in time to the second genome. This situation is shown in Figure 1.2. Due to this behaviour on the evolutionary branches, it is

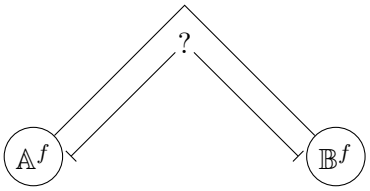


Figure 1.2: The genomic distance problem illustrated on a phylogenetic tree. The annotated genomes $\mathbb{A}^f$ and $\mathbb{B}^f$ are known. The total length of the path in between them ("?") is to be computed.

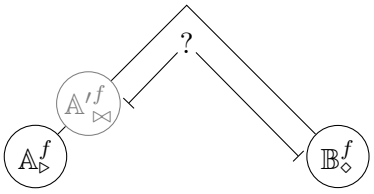


Figure 1.3: The double distance problem illustrated on a phylogenetic tree. $\mathbb{A}^f_{\triangleright}$ and $\mathbb{B}^f_{\diamond}$ are known, $\mathbb{A}'^f_{\bowtie}$ is a doubled genome of $\mathbb{A}^f_{\triangleright}$ that results from a whole genome duplication. The total length of the path between $\mathbb{A}'^f_{\bowtie}$ and $\mathbb{B}^f_{\triangleright}$ is to be computed.

sensible to choose the distance measure to be symmetric. Therefore in all models discussed in these lecture notes, each transformation has an inverse with the same cost.

A variant of the distance problem is the *double distance problem*. Here we are given a singular genome $\mathbb{A}^f_{\triangleright}$ and a duplicated genome $\mathbb{B}^f_{\diamond}$ with the same family sets and are asked to give the evolutionary distance between them. In order to accomplish this, it is necessary to find an optimal doubled genome of $\mathbb{A}^f_{\triangleright}$ and calculate its distance to $\mathbb{B}^f_{\diamond}$. More formally this is written as follows:

Definition 17 *Given a distance measure* d *the* double distance d^2 *between a singular genome* $\mathbb{A}^f_{\triangleright}$ *and a duplicated genome* $\mathbb{B}^f_{\diamond}$ *is defined as*

$$\mathrm{d}^2(\mathbb{A}^f_{\triangleright}, \mathbb{B}^f_{\diamond}) = \min_{\mathbb{A}'^f_{\bowtie} \in 2\cdot\mathbb{A}^f_{\triangleright}} \mathrm{d}(\mathbb{A}'^f_{\bowtie}, \mathbb{B}^f_{\diamond}).$$

The situation in a phylogenetic tree is shown in Figure 1.3. Notice that this approach only makes sense if $\mathbb{A}^f_{\triangleright}$ is very close to or even is an ancestor of $\mathbb{B}^f_{\diamond}$ because while whole genome duplications exist, the reverse, a hypothetical whole genome halving, is extremely unlikely, as it would require a duplicated genome to be sorted to a perfectly duplicated one by pure chance.

1.3.2 Genome Halving Problem

Similar to the genomic distance problem is the *genome halving problem* though it treats only one genome. When observing a duplicated genome $\mathbb{B}^f_{\diamond}$ the obvious hypothesis is that there must have been a whole genome duplication somewhere in its lineage. Therefore it is natural to ask how much time has passed since this duplication and what the singular ancestor might have looked like. This situation is shown in Figure 1.4. In the formal definition one also

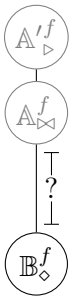


Figure 1.4: The genome halving problem illustrated on a phylogenetic tree. $\mathbb{B}^f_\diamond$ is a known genome, $\mathbb{A}^f_\bowtie$ and with it $\mathbb{A}'^f_\triangleright$ are to be computed, as well as the minimum distance between $\mathbb{B}^f_\diamond$ and $\mathbb{A}^f_\bowtie$.

requires the ancestor to be one of the possible ancestors that minimizes the distance to the present day genome.

Problem 1 (Genome halving) *Given a duplicated genome* $\mathbb{B}^f_\diamond$ *and a distance* d*, find a perfectly duplicated ancestor genome* $\mathbb{A}^f_\bowtie$ *with* $\mathcal{F}(\mathbb{A}^f) = \mathcal{F}(\mathbb{B}^f)$*, such that* $\mathrm{d}(\mathbb{A}^f_\bowtie, \mathbb{B}^f_\diamond)$ *is minimal.*

1.3.3 Median Problem

While knowing the distance between genomes or even ancestors for two genomes can be very helpful in building a phylogeny, knowing a possible ancestor and corresponding distances for a group of genomes may yield even more insight. This question is encompassed in the *median problem*. The task here is to find an ancestral annotated genome $\mathbb{M}^f$, such that the combined distance to all present day genomes in the group is minimized. This is illustrated in Figure 1.5. In formal terms the problem can be formulated like this:

Problem 2 (Genomic Median) *Given a collection of present day annotated genomes*

$$\mathcal{A} = \{\mathbb{A}^f_1, ..., \mathbb{A}^f_k\}$$

and a distance d*, find an ancestral genome* $\mathbb{M}^f$ *that minimizes the total distance*

$$\mathrm{s}(\mathbb{M}^f, \mathcal{A}) = \sum_{i=1}^{k} \mathrm{d}(\mathbb{M}^f, \mathbb{A}^f_i).$$

1.3.4 Small Parsimony Problem

While a median can be helpful knowledge when determining possible ancestors, it is oftentimes not realistic as a common ancestor for a set of genomes. This is true especially if the set is large and diverse and there is no weight on how much each individual genome should be considered for median computation. If the phylogeny of the genomes in question is known, it is often possible to reconstruct the ancestral genomes and corresponding distances between the nodes in the tree. This is referred to as the *small parsimony problem*. The situation is illustrated in Figure 1.6. More formally, we write the problem as follows:

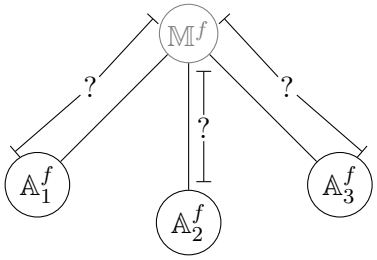


Figure 1.5: The median problem illustrated on a phylogenetic tree as a median of three. The median $\mathbb{M}^f$ is to be computed as well as the (total) length of paths to the known annotated genomes $\mathbb{A}_1^f, \mathbb{A}_2^f$ and $\mathbb{A}_3^f$.

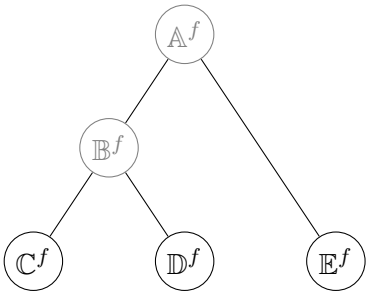


Figure 1.6: The small parsimony problem illustrated for three genomes. The phylogeny as well as the leaf genomes are known. The two ancestral genomes $\mathbb{A}^f$ and $\mathbb{B}^f$ are to be reconstructed as well as the calculation of (total) edge weights in the tree.

Problem 3 (Small Parsimony Problem) *Consider a phylogenetic tree with edges E and vertices $V = L \cup I$ where L are its leaves and I are its inner vertices. Given a distance* d *and annotated genomes $(\mathbb{G}_l^f)_{l \in L}$ associated with each leaf $l \in L$, find ancestor genomes $(\mathbb{G}_v^f)_{v \in I}$ for each inner vertex $v \in I$, such that $\sum_{(u,v) \in E} \mathrm{d}(\mathbb{G}_u^f, \mathbb{G}_v^f)$ is minimized.*