CHAPTER 2

The Breakpoint and SCJ-Models

Contents of this chapter: The simple models Breakpoint and Single-Cut-or-Join, as well as their application to problems introduced in Section 1.3.

2.1 The Breakpoint Distance

We have seen in Chapter 1 how the order of genes in a genome $\mathbb{A}$ can be expressed by its (family) adjacencies $\Gamma(\mathbb{A}^f)$ and telomeres $\Theta(\mathbb{A}^f)$. Thus it is clear that two annotated genomes $\mathbb{A}^f$ and $\mathbb{B}^f$ are more similar in terms of gene order the more adjacencies and telomeres they have in common. The Breakpoint Distance translates this into a distance measure. We will first examine it for a canonical genome pair.

Definition 18 *The Breakpoint Distance* $\mathrm{d_{BP}}(\mathbb{A}^f_\rhd, \mathbb{B}^f_\rhd)$ *of a canonical genome pair* $\mathbb{A}^f_\rhd, \mathbb{B}^f_\rhd$ *is defined as*

$$\mathrm{d_{BP}}(\mathbb{A}^f_\rhd, \mathbb{B}^f_\rhd) = n - a - \frac{t}{2}$$

with

$n = |\mathcal{G}(\mathbb{A})| = |\mathcal{G}(\mathbb{B})|$ *the number of markers contained in each genome,*

$a = |\Gamma(\mathbb{A}^f) \cap \Gamma(\mathbb{B}^f)|$ *the number of common adjacencies, and*

$t = |\Theta(\mathbb{A}^f) \cap \Theta(\mathbb{B}^f)|$ *the number of common telomeres.*

Note here, that in order for a distance to make sense, it has to fulfill the identity criterion, that is, it has to be zero if the two genomes are identical. In order to check this, we regard the folllowing:

Observation 1 *For any genome* $\mathbb{A}$ *holds*

$$|\mathcal{G}(\mathbb{A})| = |\Gamma(\mathbb{A})| + \frac{|\Theta(\mathbb{A})|}{2}. \tag{2.1}$$

Therefore it is clear that if $\mathbb{A}^f_\triangleright$ and $\mathbb{B}^f_\triangleright$ are identical in terms of gene order, all adjacencies and all telomeres of $\mathbb{A}^f_\triangleright$ have a homologue in $\mathbb{B}^f_\triangleright$, thus

$$a + \frac{t}{2} = |\Gamma(\mathbb{A}^f)| + \frac{|\Theta(\mathbb{A}^f)|}{2} = |\Gamma(\mathbb{A})| + \frac{|\Theta(\mathbb{A})|}{2} = |\mathcal{G}(\mathbb{A})| = n$$

and therefore $\mathrm{d}_{\mathrm{BP}}(\mathbb{A}^f_\triangleright, \mathbb{B}^f_\triangleright) = n - a - \frac{t}{2} = n - n = 0$.

2.1.1 The Breakpoint Double Distance

So far we have only regarded the Breakpoint distance for canonical genomes. When regarding the general formula for the double distance $\mathrm{d}^2(\mathbb{A}^f_\triangleright, \mathbb{B}^f_\diamond) = \min_{\mathbb{A}'^f_\bowtie \in 2\cdot\mathbb{A}^f_\triangleright} \mathrm{d}(\mathbb{A}'^f_\bowtie, \mathbb{B}^f_\diamond)$, we realize that after finding a suitable duplication $\mathbb{A}'^f_\bowtie$ of $\mathbb{A}^f_\triangleright$, it is necessary to calculate the distance between the balanced pair $\mathbb{A}'^f_\bowtie$ and $\mathbb{B}^f_\diamond$. However notice, that so far we have only defined the breakpoint distance for canonical pairs, that is balanced pairs without duplicate markers. There is an important general framework that allows us to reduce a distance problem of non-singular genomes to a distance problem of singular ones, albeit not necessarily in polynomial time. Observe (for example by re-reading Definitions 8 and 9 in Chapter 1) that whether a genome is considered singular or not is mainly dependent on the function f that assigns the families to the markers. The basic idea is now to choose a different function f_m for the assignment such that the genomes are singular. The function f_m is referred to as the *matching*. Of course there are restrictions on which functions are reasonable. Our matching should not group two markers in the same family which according f would not be in the same family and it should actually result in singular genomes. We summarize these minimum requirements in the following definition.

Definition 19 *A function* f_m *is a* ***matching*** *on an annotated genome pair* $\mathbb{A}^f, \mathbb{B}^f$ *if the following conditions apply:*

1. $\forall\,(a,b) \in \mathcal{G}(\mathbb{A}) \times \mathcal{G}(\mathbb{B}) : f_m(a) = f_m(b) \implies f(a) = f(b)$
2. $\forall l \in \mathcal{F}(\mathbb{A}^{f_m}) \cup \mathcal{F}(\mathbb{B}^{f_m}) : \Phi(l, \mathbb{A}^{f_m}) \leq 1, \Phi(l, \mathbb{B}^{f_m}) \leq 1$

Additionally we would like the genome pair's property of being balanced to be preserved under f_m, such that in the end we obtain a canonical pair. For instance, choosing $f_m = \mathrm{id}$, the identity function, correctly creates two singular genomes, but these genomes do not share any families and thus we could not compare them using our previous formula. We therefore require that the maximal number of possible markers should be matched.

Definition 20 *A matching* f_m *on an annotated genome pair* $\mathbb{A}^f, \mathbb{B}^f$ *is referred to as a* ***maximal matching*** *if the following holds:*

$$\begin{aligned}\forall l \in \mathcal{F}(\mathbb{A}^f) \cup \mathcal{F}(\mathbb{B}^f) : |\{g \in \mathcal{G}(\mathbb{A}) : l = f(g), \exists h \in \mathcal{G}(\mathbb{B}) \text{ with } f_m(h) = f_m(g)\}| \\ = |\{g \in \mathcal{G}(\mathbb{B}) : l = f(g), \exists h \in \mathcal{G}(\mathbb{A}) \text{ with } f_m(h) = f_m(g)\}| \\ = \min(\Phi(l, \mathbb{A}^f), \Phi(l, \mathbb{B}^f))\end{aligned}$$

We can now define the breakpoint distance between two balanced genomes as follows

Definition 21 *Given a balanced genome pair $\mathbb{A}^f, \mathbb{B}^f$, the breakpoint distance between them is defined as*

$$\mathrm{d}_{\mathrm{BP}}(\mathbb{A}^f, \mathbb{B}^f) = \min_{f_m \in \mathbb{M}_{max}} \mathrm{d}_{\mathrm{BP}}(\mathbb{A}^{f_m}, \mathbb{B}^{f_m}) \tag{2.2}$$

where $\mathbb{M}_{max}$ is the set of maximal matchings on $\mathbb{A}^f, \mathbb{B}^f$.

We will now discover a way to simultaneously find the best duplication $\mathbb{A}'^f_{\bowtie} \in 2 \cdot \mathbb{A}^f_{\triangleright}$ and the optimal maximal matching f_m on $\mathbb{A}'^f_{\bowtie}$ and $\mathbb{B}^f_{\diamond}$. In order accomplish this and similar goals in comparative genomics, it is oftentimes helpful to investigate lower bounds on the distance, or in more casual terms to "find out, what's the best we can hope for". Following from the definition of a doubled genome, we know that each $\mathbb{A}'^f_{\bowtie} \in 2 \cdot \mathbb{A}^f_{\triangleright}$ has the same family adjacencies and telomeres, namely

$$\Gamma(\mathbb{A}'_{\bowtie}{}^f) = \Gamma(\mathbb{A}_{\triangleright}{}^f) \oplus \Gamma(\mathbb{A}_{\triangleright}{}^f)$$

and

$$\Theta(\mathbb{A}'_{\bowtie}{}^f) = \Theta(\mathbb{A}_{\triangleright}{}^f) \oplus \Theta(\mathbb{A}_{\triangleright}{}^f).$$

In the best case, all of the family adjacencies and telomeres that are shared with $\Gamma(\mathbb{B}_{\diamond}{}^f)$ and $\Theta(\mathbb{B}_{\diamond}{}^f)$ under f are preserved under f_m. This gives us the lower bound

$$\begin{aligned}\mathrm{d}^2_{\mathrm{BP}}(\mathbb{A}^f_{\triangleright}, \mathbb{B}^f_{\diamond}) &= \min_{\substack{\mathbb{A}'^f_{\bowtie} \in 2 \cdot A \\ f_m \in \mathbb{M}_{max}}} \mathrm{d}_{\mathrm{BP}}(\mathbb{A}'^{f_m}, \mathbb{B}^{f_m}) && (2.3) \\ &= \min_{\substack{\mathbb{A}'^f_{\bowtie} \in 2 \cdot A \\ f_m \in \mathbb{M}_{max}}} |\mathcal{G}(\mathbb{B})| - |\Gamma(\mathbb{A}'^{f_m}) \cap \Gamma(\mathbb{B}^{f_m})| - \frac{|\Theta(\mathbb{A}'^{f_m}) \cap \Theta(\mathbb{B}^{f_m})|}{2} \\ &\geq \min_{\mathbb{A}'^f_{\bowtie} \in 2 \cdot A} |\mathcal{G}(\mathbb{B})| - |\Gamma(\mathbb{A}'^f) \cap \Gamma(\mathbb{B}^f)| - \frac{|\Theta(\mathbb{A}'^f) \cap \Theta(\mathbb{B}^f)|}{2} \\ &= |\mathcal{G}(\mathbb{B})| - |(\Gamma(\mathbb{A}^f) \oplus \Gamma(\mathbb{A}^f)) \cap \Gamma(\mathbb{B}^f)| - \frac{|(\Theta(\mathbb{A}^f) \oplus \Theta(\mathbb{A}^f)) \cap \Theta(\mathbb{B}^f)|}{2} && (2.4)\end{aligned}$$

If we can now find a matching and a doubled genome, such that this lower bound is reached, we know this must be one of the optimal choices for f_m and $\mathbb{A}'^f_{\bowtie}$. Conveniently, there is a general strategy that allows us to construct the family adjacencies of $\mathbb{A}'$ under f_m in a way, such that whenever possible, two adjacencies are homologous under f_m.

Let $a, b \in \mathcal{G}(\mathbb{B}), a \neq b$ be the two markers of a family l in $\mathbb{B}^f_{\diamond}$ ($f(a) = f(b) = l$). Then let f_m arbitrarily assign different families to these markers by adding a subscript:

- $f_m(a) = f(a)_1$

- $f_m(b) = f(b)_2$

Now we regard the family adjacencies and telomeres of $\mathbb{B}^{f_m}$ and construct $\mathbb{A}'^{f_m}$, such that a potential common adjacency is always realized as such. For each family adjacency $m^x n^y$ in $\mathbb{A}^f$ proceed as follows.

(i) If $m_1^x n_1^y \in \Gamma(\mathbb{B}^{f_m})$ or $m_2^x n_2^y \in \Gamma(\mathbb{B}^{f_m})$, choose $m_1^x n_1^y$ and $m_2^x n_2^y$ to be in $\Gamma(\mathbb{A}'^{f_m})$.

(ii) If $m_1^x n_2^y \in \Gamma(\mathbb{B}^{f_m})$ or $m_2^x n_1^y \in \Gamma(\mathbb{B}^{f_m})$, choose $m_1^x n_2^y$ and $m_2^x n_1^y$ to be in $\Gamma(\mathbb{A}'^{f_m})$.

(iii) Otherwise label the adjacency arbitrarily.

Let us regard an example for this procedure: $\mathbb{A}_\triangleright^f = \{(12), (34)\}$ and $\mathbb{B}_\diamond^f = \{(343124), (1\overline{2})\}$ is the pair for which we want to find a matching. The family adjacencies are $\mathcal{F}(\mathbb{A}^f) = \{1^h 2^t, 2^h 1^t,\ 3^h 4^t, 4^h 3^t\}$ and $\mathcal{F}(\mathbb{B}^f) = \{3^h 4^t, 4^h 3^t, 3^h 1^t, 1^h 2^t, 2^h 4^t, 4^h 3^t, 1^h 2^h, 2^t 1^t\}$. Then

$$|(\mathcal{F}(\mathbb{A}^f) \oplus \mathcal{F}(\mathbb{A}^f)) \cap \mathcal{F}(\mathbb{B}^f)| = |\{1^h 2^t, 3^h 4^t, 4^h 3^t, 4^h 3^t\}| = 4$$

meaning the best distance we can hope for under f_m is $n - a - \frac{t}{2} = 8 - 4 - 0 = 4$.

We proceed by arbitrarily assigning matchings to markers of $\mathbb{B}_\diamond^f$, that is

$$\mathbb{B}^{f_m} = \{(3_1 4_1 3_2 1_1 2_1 4_2), (1_2 \overline{2_2})\}$$

The family adjacencies under f_m now read

$$\Gamma(\mathbb{B}^{f_m}) = \{3_1^h 4_1^t, 4_1^h 3_2^t, 3_1^h 1_1^t, 1_1^h 2_1^t, 2_1^h 4_2^t, 4_2^h 3_1^t, 1_2^h 2_2^h, 2_2^t 1_2^t\}.$$

We now distinguish Cases i-iii for each of the possible adjacencies for the doubled genome:

Adjacency	Case	Resulting doubling
$1^h 2^t$	i	$1_1^h 2_1^t, 1_2^h 2_2^t$
$2^h 1^t$	iii	arbitrary
$3^h 4^t$	i	$3_1^h 4_1^t, 3_2^h 4_2^t$
$4^h 3^t$	ii	$4_1^h 3_2^t, 4_2^h 3_1^t$

For adjacency $2^h 1^t$ we will arbitrarily choose to include $2_1^h 1_1^t$ and $2_2^h 1_2^t$ although we could have also included $2_1^h 1_2^t$ and $2_1^h 1_2^t$. Reconstructing yields the genome $\{(1_1\, 2_1), (1_2\, 2_2), (3_1\, 4_1\, 3_2\, 4_2)\}$ under f_m which is $\{(1\, 2), (1\, 2), (3\, 4\, 3\, 4)\}$ - a perfectly duplicated genome - under f. We can also check that all common adjacencies are still matched under f_m. What happens, if we choose $2_1^h 1_2^t$ and $2_1^h 1_2^t$ instead as a matching is left as an exercise to the reader.

It is easily checked, that by this procedure all common adjacencies $m^x n^y$ under f always remain common adjacencies under f_m. Note that common telomeres are always preserved, due to $\mathbb{A}'^{f_m}$ having both versions t_1, t_2 of each telomere t in $\mathbb{A}^f$. Because $\mathbb{A}'^{f_m}$ is singular, we have fully described it by only giving its family adjacencies and telomeres (see Remark 3). It is also clear (see Definition 12) that $\mathbb{A}'^f$ is a doubled genome of $\mathbb{A}^f$ under f.

As we have now shown that the lower bound detailed in Equation 2.4 can always be attained, we can conclude

Theorem 1 *The Breakpoint double distance between a singular genome $\mathbb{A}^f_\triangleright$ and a duplicated genome $\mathbb{B}^f_\diamond$ with the same families ($\mathcal{F}(\mathbb{A}^f) = \mathcal{F}(\mathbb{B}^f)$) is*

$$\mathrm{d}^2_{\mathrm{BP}}(\mathbb{A}^f_\triangleright, \mathbb{B}^f_\diamond) = 2n - a - \frac{t}{2} \tag{2.5}$$

with

$$\begin{aligned} 2n &= 2|\mathcal{G}(\mathbb{A})| = |\mathcal{G}(\mathbb{B})| \\ a &= |(\Gamma(\mathbb{A}^f) \oplus \Gamma(\mathbb{A}^f)) \cap \Gamma(\mathbb{B}^f)| \\ t &= |(\Theta(\mathbb{A}^f) \oplus \Theta(\mathbb{A}^f)) \cap \Theta(\mathbb{B}^f)|. \end{aligned}$$

2.2 The SCJ Distance

We have so far only seen a purely quantitative distance measure that does not involve operations in any way. The SCJ model will be the first that involves a concrete rearrangement operation. We will also see its similarity to the Breakpoint distance.

2.2.1 The SCJ Operations

SCJ stands for *single cut or join*, meaning the allowed operations are either to cut between two markers, creating two new telomeres or to join two telomeres, thereby creating a new adjacency. It is clear that by using these operations one can sort one genome into the other, given that the two form a balanced pair.

As an example, take the genome $\{[1\ 2\ 3], (4\ 5)\}$ and the genome $\{[3\ 1\ 2\ \bar{5}\ \bar{4}]\}$. Denoting joins by $*$ and cuts by $|$, one sorting sequence is the following:

$[*1\ 2\ 3*], (4\ 5)$

$\rightarrow$ $(1\ 2\ 3), (4\ 5|)$

$\rightarrow$ $(1\ 2\,|\,3), [4\ 5]$

$\rightarrow$ $[3\ 1\ 2*], [4\ 5*]$

$\rightarrow$ $[3\ 1\ 2\ \bar{5}\ \bar{4}]$

Formally we will define the SCJ operation via adjacencies and telomeres.

Definition 22 *Given a genome $\mathbb{G}$ an SCJ operation transforms it into a genome $\mathbb{G}'$ by either*

(i) cutting an adjacency $m^x n^y \in \Gamma(\mathbb{G})$:

- $\Gamma(\mathbb{G}') = \Gamma(\mathbb{G}) \setminus \{m^x n^y\}$,
- $\Theta(\mathbb{G}') = \Theta(\mathbb{G}) \cup \{m^x, n^y\}$

or

(ii) joining two telomeres $m^x, n^y \in \Theta(\mathbb{G})$:

- $\Theta(\mathbb{G}') = \Theta(\mathbb{G}) \setminus \{m^x, n^y\}$,
- $\Gamma(\mathbb{G}') = \Gamma(\mathbb{G}) \cup \{m^x n^y\}$

Based on this definition, we can also regard the example from earlier in adjacency-notation:

Adjacencies	Telomeres
$\{1^h2^t, 2^h3^t, 4^h5^t, 5^h4^t\}$	$\{1^t*, 3^h*\}$
$\{1^h2^t, 2^h3^t, 4^h5^t, 5^h\|4^t, 1^t3^h\}$	$\{\}$
$\{1^h2^t, 2^h\|3^t, 4^h5^t, 1^t3^h\}$	$\{5^h, 4^t\}$
$\{1^h2^t, 4^h5^t, 1^t3^h\}$	$\{5^h*, 4^t, 2^h*, 3^t\}$
$\{1^h2^t, 4^h5^t, 1^t3^h, 2^h5^h\}$	$\{4^t, 3^t\}$

Note that because the genomes in question are singular, we can notate family adjacencies and telomeres for easier readability, although the definition specifies marker adjacencies and telomeres.

2.2.2 The SCJ Distance

With the possible operations defined, the question is now, how many operations are necessary to sort one genome into the other, that is, we want to know the transformation distance between two genomes under the SCJ model. Perhaps surprisingly, the derivation of the genomic distance oftentimes follows the same abstract pattern, even for more complicated distances. This general pattern is described in Algorithm 2.1 for future reference.

Algorithm 2.1 A cooking recipe for determining genomic rearrangement distances

1. Find a suitable data structure that represents the genome pair and shows (somewhat) consistent behavior when an operation of the model is applied.
2. Find a quantity q in the data structure, that the operations of the model can change by at most 1.

   $\implies |\Delta q| \leq 1$ with $\Delta q = q - q'$, where q is the quantity before and q' is the quantity after the operation was applied.
3. Identify the state $q^\star$ of this quantity that occurs if and only if the sorting is complete.

   $\implies d \geq |q - q^\star|$
4. Find a way of sorting that decreases $|q - q^\star|$ by 1 in every step, thus reaching the lower bound.

   $\implies d = |q - q^\star|$

Of course, when investigating an unfamiliar model, the process is rarely this streamline. For example, it is oftentimes not possible to find an algorithm that is capable of reducing $|q - q^\star|$ by 1 in every step. Sometimes there is no distinct quantity $q^\star$ in the sorted case. Therefore

one would typically make several passes through this procedure, each time modifying an aspect of it, be it the data structure or the quantity q.

However, in the case of the here examined SCJ distance, the data structures and quantities needed are relatively straightforward, such that the reader might even be capable to perform some steps on their own. Therefore, they are invited to pause reading the following section after each step and with the help of Algorithm 2.1 try and solve the next step themselves.

We want to solve the distance problem for a canonical pair of genomes $\mathbb{A}^f_\triangleright, \mathbb{B}^f_\triangleright$. Therefore we regard the sequence of intermediate genomes $\mathbb{A}^f_\triangleright = \mathbb{A}^f_0, \mathbb{A}^f{}_2, ..., \mathbb{A}^f_{n-1}, \mathbb{A}^f_n = \mathbb{B}^f_\triangleright$ with $\mathbb{A}^f_{i+1}$ being the result of an SCJ operation being applied to $\mathbb{A}^f_i$ for all $0 \leq i < n$. Notice that because there is no SCJ operation to create or remove markers, every intermediate genome $\mathbb{A}^f_i$ is also singular and forms a canonical pair with every other genome $\mathbb{A}^f_j$ of the sequence. For Step 1 of Algorithm 2.1 our currently known data structures, that is family adjacencies and telomeres, suffice.

For Step 2 it is instructive to experiment with some examples or at least re-read the example from earlier this chapter. The first idea one might come up with is that the number of adjacencies in the genome to be sorted always changes by 1 when sorting. Thus one could derive the quantity $q_i = |\Gamma(\mathbb{A}_i)|$. However, while there is a defined state $q^\star$ with $q^\star = |\Gamma(\mathbb{B})|$ that q reaches once $\mathbb{A}^f_\triangleright$ is sorted into $\mathbb{B}^f_\triangleright$, this number of adjacencies and telomeres could also be attained earlier. In fact, if $\mathbb{A}^f_\triangleright$ and $\mathbb{B}^f_\triangleright$ had the same number of linear chromosomes, it would even be reached in the beginning of the sorting. Nonetheless, family adjacencies are a good metric as their number can only be changed by at most one by every operation. We regard the number of adjacencies that $\mathbb{A}^f_\triangleright$ and $\mathbb{B}^f_\triangleright$ do **not** have in common, that is the quantity $q_i = |\Gamma(\mathbb{A}^f_i) \setminus \Gamma(\mathbb{B}^f)| + |\Gamma(\mathbb{B}^f) \setminus \Gamma(\mathbb{A}^f_i)|$. Both summands can only be changed by 1 as each operation can at best create or remove one adjacency. Similarly, only one term can change at once as there is never more than one adjacency affected by any operation in the SCJ model.

Step 3 is easy to solve: When we are done sorting, all adjacencies are shared between the genomes, thus $q^\star = 0$. We still need to determine, if this is the only case, in which q_i can be 0. We presume a genome $\mathbb{A}^f_i$ with the same markers as $\mathbb{A}^f_\triangleright$ that is potentially different from $\mathbb{B}^f_\triangleright$, but fulfills that $q_i = |\Gamma(\mathbb{A}^f_i) \setminus \Gamma(\mathbb{B}^f)| + |\Gamma(\mathbb{B}^f) \setminus \Gamma(\mathbb{A}^f_i)| = 0$. We therefore know that $\Gamma(\mathbb{A}^f_i) = \Gamma(\mathbb{B}^f)$. Because we can derive the family telomeres of $\mathbb{A}^f_i$ by just finding the extremities that are not yet present in family adjacencies, we know that also $\Theta(\mathbb{A}^f_i) = \Theta(\mathbb{B}^f)$. As both genomes $\mathbb{A}^f_i$ and $\mathbb{B}^f_\triangleright$ are singular they are fully described just by giving their family adjacencies and telomeres (see Remark 3). Therefore $\mathbb{A}^f_i = \mathbb{B}^f_\triangleright$ must hold.

Step 4 is to find a way to sort $\mathbb{A}^f_\triangleright$ into $\mathbb{B}^f_\triangleright$, such that

$$|q_i - q^\star| = |q_i - 0| = q_i = |\Gamma(\mathbb{A}^f_i) \setminus \Gamma(\mathbb{B}^f)| + |\Gamma(\mathbb{B}^f) \setminus \Gamma(\mathbb{A}^f_i)|$$

is reduced by one in each step. The simplest way to do so, is to first remove all adjacencies exclusive to $\mathbb{A}^f_\triangleright$ by cutting each adjacency belonging to $|\Gamma(\mathbb{A}^f) \setminus \Gamma(\mathbb{B}^f)|$ and then creating each adjacency in $|\Gamma(\mathbb{B}^f) \setminus \Gamma(\mathbb{A}^f)|$ by joining telomeres. Thus we have shown that

Theorem 2 *The SCJ distance between a canonical genome pair $\mathbb{A}^f_\triangleright, \mathbb{B}^f_\triangleright$ is*

$$\mathrm{d}_{\mathrm{SCJ}}(\mathbb{A}^f_\triangleright, \mathbb{B}^f_\triangleright) = |\Gamma(\mathbb{A}^f) \setminus \Gamma(\mathbb{B}^f)| + |\Gamma(\mathbb{B}^f) \setminus \Gamma(\mathbb{A}^f)|.$$

2.2.3 Relationship to the Breakpoint distance

One might already suspect that the SCJ and Breakpoint distances are somewhat similar to each other as the respective distance formulas mainly depend on shared or non-shared adjacencies. To explore this, we first apply some transformations from standard set theory to the SCJ distance formula.

$$\begin{aligned}
\mathrm{d}_{\mathrm{SCJ}}(\mathbb{A}^f_{\triangleright}, \mathbb{B}^f_{\triangleright}) &= |\Gamma(\mathbb{A}^f) \setminus \Gamma(\mathbb{B}^f)| + |\Gamma(\mathbb{B}^f) \setminus \Gamma(\mathbb{A}^f)| \\
&= |\Gamma(\mathbb{A}^f)| - |\Gamma(\mathbb{A}^f) \cap \Gamma(\mathbb{B}^f)| + |\Gamma(\mathbb{B}^f)| - |\Gamma(\mathbb{B}^f) \cap \Gamma(\mathbb{A}^f)| \\
&= |\Gamma(\mathbb{A}^f)| + |\Gamma(\mathbb{B}^f)| - 2 \cdot |\Gamma(\mathbb{A}^f) \cap \Gamma(\mathbb{B}^f)|
\end{aligned}$$

Here we already notice that the formula looks fairly similar to that of the breakpoint distance as it already has some positive term followed by a negative term entailing the number of common adjacencies, although the latter is weighted by two instead of one. The first term still looks quite different to the equivalent in the Breakpoint distance as it depends on the number of adjacencies instead of the number of genes. In order to obtain a term with the number of genes, we use the identity $|\mathcal{G}(\mathbb{A})| = |\Gamma(\mathbb{A})| + \frac{|\Theta(\mathbb{A})|}{2}$ (see Observation 1) which in case of singular genomes also holds for family adjacencies. Solved for the number of family adjacencies the equation reads

$$|\Gamma(\mathbb{A}^f)| = |\mathcal{G}(\mathbb{A})| - \frac{|\Theta(\mathbb{A}^f)|}{2},$$

which we substitute into the distance formula, yielding

$$\begin{aligned}
&\Gamma(\mathbb{A}^f)| + |\Gamma(\mathbb{B}^f)| - 2 \cdot |\Gamma(\mathbb{A}^f) \cap \Gamma(\mathbb{B}^f)| \\
&= |\mathcal{G}(\mathbb{A})| + |\mathcal{G}(\mathbb{B})| - 2 \cdot |\Gamma(\mathbb{A}^f) \cap \Gamma(\mathbb{B}^f)| - \frac{|\Theta(\mathbb{A}^f)|}{2} - \frac{|\Theta(\mathbb{B}^f)|}{2}.
\end{aligned}$$

If we substitute in $n = \mathcal{G}(\mathbb{A}) = \mathcal{G}(\mathbb{B})$, $a = |\Gamma(\mathbb{A}^f) \cap \Gamma(\mathbb{B}^f)|$ and $t = |\Theta(\mathbb{A}^f) \cap \Theta(\mathbb{B}^f)|$ from the definition of the breakpoint distance (see Def. 18), we obtain

$$\begin{aligned}
&2 \cdot n - 2 \cdot a - \frac{|\Theta(\mathbb{A}^f)|}{2} - \frac{|\Theta(\mathbb{B}^f)|}{2} \\
&\quad = 2 \cdot n - 2 \cdot a - \frac{|\Theta(\mathbb{A}^f)|}{2} - \frac{|\Theta(\mathbb{B}^f)|}{2} - t + t \\
&\quad = 2 \cdot (n - a - \frac{t}{2}) - \frac{|\Theta(\mathbb{A}^f)|}{2} - \frac{|\Theta(\mathbb{B}^f)|}{2} + t \\
&\quad = 2 \cdot \mathrm{d}_{\mathrm{BP}}(\mathbb{A}^f_{\triangleright}, \mathbb{B}^f_{\triangleright}) - \frac{|\Theta(\mathbb{A}^f)|}{2} - \frac{|\Theta(\mathbb{B}^f)|}{2} + t.
\end{aligned}$$

If we assume that the number of linear chromosomes and therewith telomeres is low relative to the number of markers overall, an assumption typically satisfied by real genomes, the SCJ distance is roughly double the Breakpoint distance. More generally holds

$$\mathrm{d}_{\mathrm{BP}}(\mathbb{A}^f_{\triangleright}, \mathbb{B}^f_{\triangleright}) \leq \mathrm{d}_{\mathrm{SCJ}}(\mathbb{A}^f_{\triangleright}, \mathbb{B}^f_{\triangleright}) \leq 2 \cdot \mathrm{d}_{\mathrm{BP}}(\mathbb{A}^f_{\triangleright}, \mathbb{B}^f_{\triangleright}) \tag{2.6}$$

though we will not prove it here. Using this close relationship, we will be able to solve problems in one model using the knowledge we obtained from studying the other.

2.3 SCJ Double Distance and Genome Halving

As a first result, we realize that in Section 2.1.1 when we gave the algorithm for finding a matching as well as the doubling for the Breakpoint double distance calculation, we never needed to account for telomeres as the maximal number of them would be automatically matched. Therefore, we can apply the same technique to finding a matching under the SCJ distance and conclude the following:

Theorem 3 *The SCJ double distance between a singular genome $\mathbb{A}^f_\triangleright$ and a duplicated genome $\mathbb{B}^f_\diamond$ with the same families ($\mathcal{F}(\mathbb{A}^f) = \mathcal{F}(\mathbb{B}^f)$) is*

$$\mathrm{d}^2_{\mathrm{BP}}(\mathbb{A}^f_\triangleright, \mathbb{B}^f_\diamond) = |(\Gamma(\mathbb{A}^f) \oplus \Gamma(\mathbb{A}^f)) \setminus \Gamma(\mathbb{B}^f)| + |\Gamma(\mathbb{B}^f) \setminus (\Gamma(\mathbb{A}^f) \oplus \Gamma(\mathbb{A}^f))|. \tag{2.7}$$

We will now regard the Genome Halving Problem we know from Section 1.3.2 under the SCJ model. Conceptually, the Genome Halving Problem is not that different from the double distance, only that now the singular ancestor $\mathbb{A}^f_\triangleright$ is unknown and part of the optimization. As we have already seen that an optimal doubling and matching can be easily found, once the family adjacencies are set, we can formulate the problem as finding a singular genome that minimizes the double distance

$$\min_{\mathbb{A}^f_\triangleright} \mathrm{d}^2_{\mathrm{SCJ}}(\mathbb{A}^f_\triangleright, \mathbb{B}^f_\diamond) = \min_{\mathbb{A}^f_\triangleright} |(\mathcal{F}(\mathbb{A}^f_\triangleright) \oplus \mathcal{F}(\mathbb{A}^f_\triangleright)) \setminus \mathcal{F}(\mathbb{B}^f_\diamond)| + |\mathcal{F}(\mathbb{B}^f_\diamond) \setminus (\mathcal{F}(\mathbb{A}^f_\triangleright) \oplus \mathcal{F}(\mathbb{A}^f_\triangleright))|.$$

Conceptually, we can think of creating $\mathbb{A}^f_\triangleright$ by starting with a genome without any adjacencies, meaning each gene is insular on its own linear chromosome and then joining the telomeres to create the adjacencies that form the optimal ancestor. In order to find beneficial adjacencies, we have to examine the double distance with respect to individual adjacencies. We therefore regard the number of times $\phi(m^x n^y, \mathbb{G}^f)$ an adjacency $m^x n^y$ appears in the family adjacencies of a genome $\mathbb{G}^f$. Expressed in this notation, our double distance formula reads

$$\min_{\mathbb{A}^f_\triangleright} \mathrm{d}^2_{\mathrm{SCJ}}(\mathbb{A}^f_\triangleright, \mathbb{B}^f_\diamond) = \min_{\mathbb{A}^f_\triangleright} \sum_{\substack{m^x n^y \\ \in \Gamma(\mathbb{A}^f)}} (2 - \phi(m^x n^y, \mathbb{B}^f)) + \sum_{\substack{m^x n^y \\ \notin \Gamma(\mathbb{A}^f)}} \phi(m^x n^y, \mathbb{B}^f).$$

In order to unify the two sums, we can apply a more generally useful trick. Note that since $\mathbb{A}^f_\triangleright$ is a singular genome, $\phi(m^x n^y, \mathbb{A}^f)$ is a binary variable with values $\{0, 1\}$, therefore we can use it similar to a boolean. Multiplying with $\phi(m^x n^y, \mathbb{A}^f)$ filters out all terms of a sum, for which $m^x n^y$ is not an adjacency of $\mathbb{A}^f$. Similarly, $(1 - \phi(m^x n^y, \mathbb{A}^f))$ acts just like a negated boolean and multiplying with it eliminates those terms for which $m^x n^y$ is an adjacency of $\mathbb{A}^f$. We obtain

$$\begin{aligned}
\mathrm{d}^2_{\mathrm{SCJ}}(\mathbb{A}^f_\triangleright, \mathbb{B}^f_\diamond) &= \sum_{m^x n^y} \phi(m^x n^y, \mathbb{A}^f) \cdot (2 - \phi(m^x n^y, \mathbb{B}^f)) \\
&\quad + \sum_{m^x n^y} (1 - \phi(m^x n^y, \mathbb{A}^f)) \cdot \phi(m^x n^y, \mathbb{B}^f) \\
&= \sum_{m^x n^y} (\phi(m^x n^y, \mathbb{B}^f) + \phi(m^x n^y, \mathbb{A}^f) \cdot (2 - \phi(m^x n^y, \mathbb{B}^f))).
\end{aligned}$$

It is clear that the sum $\sum_{m^x n^y} \phi(m^x n^y, \mathbb{B}^f)$ simply counts the number of adjacencies in $\mathbb{B}^f$. Applying this fact and resetting the sum to only contain adjacencies of $\mathbb{A}^f$ we get

$$\mathrm{d}^2_{\mathrm{SCJ}}(\mathbb{A}^f_\triangleright, \mathbb{B}^f_\diamond) = |\Gamma(\mathbb{B})| + \sum_{\substack{m^x n^y \\ \in \mathbb{A}^f}} (2 - \phi(m^x n^y, \mathbb{B}^f)).$$

Examining this formula more closely, we find that the first term $|\Gamma(\mathbb{B})|$ does not depend on the ancestor we want to optimize and therefore does not influence this optimization. We therefore need not regard it any further. The summation over the adjacencies however portrays an interesting behavior. Each adjacency $m^x n^y$ independently contributes a value $s(m^x n^y) = (2 - \phi(m^x n^y, \mathbb{B}^f))$ to the sum. We can view this value as a kind of score expressing how bad it would be if we included the adjacency in $\mathbb{A}^f_\triangleright$. We see that adjacencies that do not occur in $\mathbb{B}^f_\diamond$ at all have score 2, adjacencies that occur once have score 0 and adjacencies that occur twice have score -2. An easy way of minimizing the sum is thus to only include adjacencies into $\mathbb{A}^f_\triangleright$ if they occur twice in $\mathbb{B}^f_\diamond$, thus only getting negative contributions to the sum. However, we need to make sure that these adjacencies actually form a singular genome, that is there cannot be any *conflict* between family adjacencies. A conflict occurs when two different family adjacencies share the same extremity, i.e. $m^x n^y$ and $n^y l^z$ with $m^x \neq l^z$. In this case the set of adjacencies does not yield a defined singular genome. We can easily see though that if we had a conflict, say $m^x n^y$ and $n^y l^z$ would be both included in $\mathbb{A}^f_\triangleright$, then $\mathbb{B}^f_\diamond$ would need to have these conflicting adjacencies twice, making it impossible for $\mathbb{B}^f_\diamond$ to be a duplicated genome - a contradiction! Therefore there cannot be any conflict if we choose only adjacencies occurring twice in $\mathbb{B}^f_\diamond$.

As an example, let us regard the genome $\mathbb{B}^f_\diamond = \{(1\,2\,3), (3\,4\,\bar{2}\,\bar{1}), [4]\}$. We list the number of occurrences of each family adjacency as well as its score:

$m^x n^y$	$\phi(m^x n^y, \mathbb{B}^f_\diamond)$	$s(m^x n^y)$
$1^h 2^t$	2	-2
$2^h 3^h$	1	0
$3^t 1^t$	2	-2
$3^h 4^t$	1	0
$4^h 2^h$	1	0

Thus we choose to include $1^h 2^t$ and $3^t 1^t$, obtaining $\mathbb{A}^f_\triangleright = \{[\bar{3}\,1\,2], [4]\}$ as the singular ancestor. Note that we could also include some of the adjacencies with score 0 without changing the distance, however they are not guaranteed to be free of conflicts. For an example, try including both $2^h 3^h$ and $3^h 4^t$ in $\mathbb{A}^f_\triangleright$. You will find that this is impossible if $\mathbb{A}^f_\triangleright$ is supposed to be singular. To find the correct perfectly duplicated ancestor, we would technically need to solve a double distance problem. In this case, this is easy as there is only one way to duplicate $\mathbb{A}^f_\triangleright$ and we obtain $\mathbb{A}'^f_\bowtie = \{[\bar{3}\,1\,2], [4], [\bar{3}\,1\,2], [4]\}$ as the perfectly duplicated ancestor.

2.4 SCJ and Breakpoint Median

In this section we will examine the Median problem under the SCJ and Breakpoint distance. Because distances for non-singular or even natural genomes are a relatively recent devel-

opment in the field, higher-level problems like medians or parsimonies have been mainly studied for collections of genomes, in which the genomes form canonical pairs. Therefore, we regard the median problem only for collections of canonical genomes.

2.4.1 SCJ Median for canonical genomes

We want to find the Median $\mathbb{M}^f$ for a collection of genomes $\mathcal{A} = \{\mathbb{A}_1^f, ..., \mathbb{A}_k^f\}$, in which each two genomes $\mathbb{A}_i^f, \mathbb{A}_j^f$ form a canonical pair. We know that the median must minimize the total distance

$$\mathrm{s}_{\mathrm{SCJ}}(\mathbb{M}^f, \mathcal{A}) = \sum_{\mathbb{A}_i^f \in \mathcal{A}} \mathrm{d}_{\mathrm{SCJ}}(\mathbb{M}^f, \mathbb{A}_i^f).$$

If we substitute in the distance formula for the SCJ model, we obtain

$$\begin{aligned}\mathrm{s}_{\mathrm{SCJ}}(\mathbb{M}^f, \mathcal{A}) &= \sum_{\mathbb{A}_i^f \in \mathcal{A}} (|\Gamma(\mathbb{M}^f) \setminus \Gamma(\mathbb{A}_i^f)| + |\Gamma(\mathbb{A}_i^f) \setminus \Gamma(\mathbb{M}^f)|) \\ &= \sum_{\mathbb{A}_i^f \in \mathcal{A}} |\Gamma(\mathbb{M}^f) \setminus \Gamma(\mathbb{A}_i^f)| + \sum_{\mathbb{A}_i^f \in \mathcal{A}} |\Gamma(\mathbb{A}_i^f) \setminus \Gamma(\mathbb{M}^f)|.\end{aligned}$$

Again, in this case we can think of creating a median by starting with a genome without adjacencies and finding the best set of adjacencies to form a median. Expressed in the notation from earlier, our total distance formula reads

$$\mathrm{s}_{\mathrm{SCJ}}(\mathbb{M}^f, \mathcal{A}) = \sum_{\mathbb{A}_i^f \in \mathcal{A}} \sum_{\substack{m^x n^y \\ \in \Gamma(\mathbb{M}^f)}} (1 - \phi(m^x n^y, \mathbb{A}_i^f)) + \sum_{\mathbb{A}_i^f \in \mathcal{A}} \sum_{\substack{m^x n^y \\ \notin \Gamma(\mathbb{M}^f)}} \phi(m^x n^y, \mathbb{A}_i^f).$$

We can further simplify the formula by extending the definition of ϕ to also work on sets of genomes, that is for a collection of genomes $\mathcal{S}$ we have $\phi(m^x n^y, \mathcal{S}) = \sum_{\mathbb{G}^f \in \mathcal{S}} \phi(m^x n^y, \mathbb{G}^f)$. Using this we obtain

$$\begin{aligned}\mathrm{s}_{\mathrm{SCJ}}(\mathbb{M}^f, \mathcal{A}) &= \sum_{\substack{m^x n^y \\ \in \Gamma(\mathbb{M}^f)}} \sum_{\mathbb{A}_i^f \in \mathcal{A}} (1 - \phi(m^x n^y, \mathbb{A}_i^f)) + \sum_{\substack{m^x n^y \\ \notin \Gamma(\mathbb{M}^f)}} \sum_{\mathbb{A}_i^f \in \mathcal{A}} \phi(m^x n^y, \mathbb{A}_i^f) \\ &= \sum_{\substack{m^x n^y \\ \in \Gamma(\mathbb{M}^f)}} (k - \phi(m^x n^y, \mathcal{A})) + \sum_{\substack{m^x n^y \\ \notin \Gamma(\mathbb{M}^f)}} \phi(m^x n^y, \mathcal{A}).\end{aligned}$$

Again because $\mathbb{M}^f$ is a singular genome we can apply the trick from earlier to unify the sums.

$$\begin{aligned}\mathrm{s}_{\mathrm{SCJ}}(\mathbb{M}^f, \mathcal{A}) &= \sum_{m^x n^y} \phi(m^x n^y, \mathbb{M}^f) \cdot (k - \phi(m^x n^y, \mathcal{A})) + \sum_{m^x n^y} (1 - \phi(m^x n^y, \mathbb{M}^f)) \cdot \phi(m^x n^y, \mathcal{A}) \\ &= \sum_{m^x n^y} (k \cdot \phi(m^x n^y, \mathbb{M}^f) - \phi(m^x n^y, \mathbb{M}^f)\phi(m^x n^y, \mathcal{A}) \\ &+ \phi(m^x n^y, \mathcal{A}) - \phi(m^x n^y, \mathbb{M}^f)\phi(m^x n^y, \mathcal{A})) \\ &= \sum_{\mathbb{A}_i^f \in \mathcal{A}} \Gamma(\mathbb{A}_i^f) + \sum_{m^x n^y} (k \cdot \phi(m^x n^y, \mathbb{M}^f) - 2 \cdot \phi(m^x n^y, \mathcal{A}))\end{aligned}$$

Again, we see a kind of score $s(m^x n^y) = k - 2 \cdot \phi(m^x n^y, \mathcal{A})$ for how bad it would be to include adjacency $m^x n^y$ into the median $\Gamma(\mathbb{M}^f)$. We observe that including an adjacency in our median that occurs in half or fewer of the genomes of $\mathcal{A}$ will yield a positive contribution, while including an adjacency that occurs in more than half of the genomes of $\mathcal{A}$ will yield a negative one. As we want to minimize the total distance, we can again follow a greedy approach and choose to include an adjacency if it has a negative contribution, that is if it occurs in more than half of the genomes in $\mathcal{A}$. This approach, too, cannot lead to conflicting adjacencies. Proving this fact is an exercise left to the reader that will be part of the tutorials.

To demonstrate the method, we regard a small example for a median of four. Let the collection of genomes be $\mathcal{A} = \{\mathbb{A}_1^f, \mathbb{A}_2^f, \mathbb{A}_3^f, \mathbb{A}_4^f\}$ with $\mathbb{A}_1^f = \{[1\,2\,3]\}$, $\mathbb{A}_2^f = \{(2\,3), [1]\}$, $\mathbb{A}_3^f = \{[3\,1\,2]\}$ and $\mathbb{A}_4^f = \{(1\,2\,3)\}$. We observe the following counts for the adjacencies:

$m^x n^y$	$\phi(m^x n^y, \mathcal{A})$	$s(m^x n^y)$
$1^h 2^t$	3	-2
$2^h 3^t$	3	-2
$3^h 1^t$	2	0
$3^h 2^t$	1	2

We therefore know to include $1^h 2^t$ and $2^h 3^t$ in our median, yielding $\mathbb{M}^f = \{[1\,2\,3]\}$. Notice that because the contribution of $3^h 1^t$ is 0, we could again include this adjacency without increasing the total distance and we would have an alternative median $\mathbb{M}'^f = \{(1\,2\,3)\}$. Such adjacencies that are neutral in terms of contribution to the total distance can only occur when we calculate the median for an even number of genomes. For odd numbers of genomes, all adjacencies have either positive or negative score. Therefore, for odd numbers of genomes the SCJ median is unique.

SCJ linear median of canonical genomes

As we have seen, the structure of the median we might obtain by this procedure is not necessarily guaranteed. That is, we might for example obtain a circular median for a set of linear genomes. Therefore, one sometimes wishes to restrict, which types of genomes qualify as a valid median.

Definition 23 *The linear genomic median of a set of linear genomes* $\mathcal{A} = \{\mathbb{A}_0^f, ..., \mathbb{A}_k^f\}$ *on a distance measure* d *is a genome* $\mathbb{M}_l^f$ *consisting only of linear chromosomes that minimizes*

$$\mathrm{s}(\mathbb{M}_l^f, \mathcal{A}) = \sum_{\mathbb{A}_i^f \in \mathcal{A}} \mathrm{d}(\mathbb{M}_l^f, \mathbb{A}_i^f).$$

Fortunately, there is an easy strategy that allows us to derive a linear median from a non-linear one in the SCJ model. As each adjacency $m^x n^y$ has a score $s(m^x n^y)$, we can simply remove the highest scoring adjacency from each circular chromosome in the median. This way, each of the circular chromosomes will be linearized and we obtain a linear genome. To see that this genome - let us refer to it as $\mathbb{G}^f$ - is in fact a linear median, we consider the general median $\mathbb{M}_g^f$ we used to construct $\mathbb{G}^f$ and a true linear median $\mathbb{M}_l^f$ of $\mathcal{A}$. Without loss of generality, we can assume that $\Gamma(\mathbb{M}_l^f) \subseteq \Gamma(\mathbb{M}_g^f)$ as we can simply remove adjacencies with

score 0 from $\Gamma(\mathbb{M}_l^f)$ without changing the total distance. We then know that because $\mathbb{M}_l^f$ minimizes the total distance for linear genomes, all adjacencies of linear chromosomes and all but one adjacency of each circular chromosome must be shared with $\mathbb{M}_g^f$. Otherwise one could include the additional adjacencies from $\mathbb{M}_g^f$ and obtain a linear genome that further decreases the total distance. We know that an adjacency missing in $\mathbb{M}_l^f$ that was part of a circular chromosome in $\mathbb{M}_g^f$ must have the highest score of all adjacencies of that chromosome. Otherwise one could include that adjacency in $\mathbb{M}_l^f$, remove a higher scoring adjacency from the chromosome and obtain another linear genome with a shorter total distance. Therefore the total distances of $\mathbb{M}_l^f$ and $\mathbb{M}_g^f$ must be the same.

2.4.2 SCJ and Breakpoint Median for canonical circular genomes

As we have seen the linear SCJ median for canonical genomes, it makes sense to also investigate its circular counterpart. In fact, one could argue that a circular median is even more worthy of consideration for the SCJ model as medians tend to fragment into many linear chromosomes under this model if the genomes in question have only few common adjacencies. We define the circular median as follows.

Definition 24 *The circular genomic median of a set of circular genomes* $\mathcal{A} = \{\mathbb{A}_0^f, ..., \mathbb{A}_k^f\}$ *on a distance measure* d *is a genome* $\mathbb{M}_c^f$ *consisting only of circular genomes that minimizes*

$$\mathrm{s}(\mathbb{M}_c^f, \mathcal{A}) = \sum_{\mathbb{A}_i^f \in \mathcal{A}} \mathrm{d}(\mathbb{M}_c^f, \mathbb{A}_i^f).$$

In order to find a median, we can use a structure that is more generally useful to observe multiple possible genomes in one structure. We define a complete graph, in which every extremity is a vertex, the edges represent all possible adjacencies and the edge weights are the scores of the respective adjacencies, that is $V = \{g^h : g \in \mathcal{F}(\mathbb{M}_c^f)\} \cup \{g^t : g \in \mathcal{F}(\mathbb{M}_c^f)\}$ and $w((m^x, n^y)) = s(m^x n^y)$. Notice that every perfect matching in this graph defines a circular genome. The reverse is also true: Every singular circular genome with this set of families defines a matching in the graph. Therefore, a perfect matching with minimal weight will be a circular median. There exist algorithms to accomplish this in polynomial time, though we will not discuss them here. In most small examples, one can easily find such a matching manually.

Let us regard such a small example for the three genomes

$$\mathcal{A} = \{\{(1\,2\,3)\}, \{(1\,2), (3)\}, \{(1), (2\,3)\}\}.$$

We show the corresponding graph in Figure 2.1. We can see that the maximal matching has at least weight -1 and we find a matching meeting this lower bound in Figure 2.2. The resulting circular median is $\mathbb{M}_c^f = \{(1\,2\,3)\}$. Note that the general median $\mathbb{M}^f = \{[1\,2\,3]\}$ has a better score, but is not a perfect matching in the graph.

As we have seen, the Breakpoint distance is very similar to the SCJ distance, especially for circular genomes. We can therefore also use this graph structure to find a circular Breakpoint median. First, we deduce the weight for each adjacency.

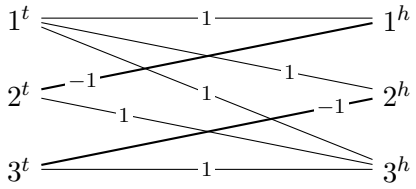

Figure 2.1: Circular median graph for the genomes $\{(1\,2\,3)\}$, $\{(1\,2),(3)\}$ and $\{(1),(2\,3)\}$. Edges with minimum weight (here -1) are highlighted. Edges with the maximum weight (here 3) are not displayed, but they exist in the definition of the graph and connect each pair of vertices that is not yet connected by an edge of lower weight.

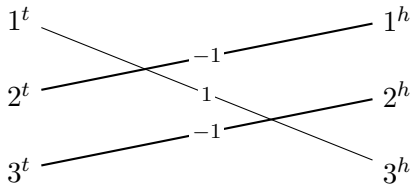

Figure 2.2: Minimum weight maximal matching on the circular median graph of $\{(1\,2\,3)\}$, $\{(1\,2),(3)\}$ and $\{(1),(2\,3)\}$.

$$\begin{aligned}
\mathrm{s}_{\mathrm{BP}}(\mathbb{M}^f, \mathcal{A}) &= \sum_{\mathbb{A}_i^f \in \mathcal{A}} \mathrm{d}_{\mathrm{BP}}(\mathbb{M}^f, \mathbb{A}_i^f) \\
&= \sum_{\mathbb{A}_i^f \in \mathcal{A}} (n - \sum_{\substack{m^x n^y \\ \in \Gamma(\mathbb{M}^f)}} \phi(m^x n^y, \mathbb{A}_i^f)) \\
&= k \cdot n - \sum_{\substack{m^x n^y \\ \in \Gamma(\mathbb{M}^f)}} \phi(m^x n^y, \mathcal{A}))
\end{aligned}$$

We see, that we should choose weight $w((m^x, n^y)) = -\phi(m^x n^y, \mathcal{A})$ to compute a Breakpoint median. If one were to apply this scoring scheme and calculate the circular Breakpoint median for the above example, one would find the same circular median. Executing this and answering the question why the same median arises is an exercise left to the reader.

2.4.3 Breakpoint Median for canonical genomes

We would like to use this new graph structure to calculate a general Breakpoint median. Notice that the reason the graph only yields circular genomes is because we considered only perfect matchings, whereas the matchings that do not necessarily cover every vertex are a description of all possible singular genomes with this set of families, not just circular ones. We might therefore be tempted to simply search for any matching with minimum weight. However if we extend the definition of ϕ to count the amount of occurrences $\phi(m^x, \mathcal{S})$ of telomere m^x in genome set $\mathcal{S}$, we can regard the total distance of the Breakpoint median for the general case, that is

$$\begin{aligned}
\mathrm{s}_{\mathrm{BP}}(\mathbb{M}^f, \mathcal{A}) &= \sum_{\mathbb{A}_i^f \in \mathcal{A}} \mathrm{d}_{\mathrm{BP}}(\mathbb{M}^f, \mathbb{A}_i^f) \\
&= \sum_{\mathbb{A}_i^f \in \mathcal{A}} (n - \sum_{\substack{m^x n^y \\ \in \Gamma(\mathbb{M}^f)}} \phi(m^x n^y, \mathbb{A}_i^f) - \sum_{\substack{m^x \\ \in \Theta(\mathbb{M}^f)}} \frac{\phi(m^x, \mathbb{A}_i^f)}{2}) \\
&= k \cdot n - \sum_{\substack{m^x \\ \in \Gamma(\mathbb{M}^f)}} \phi(m^x n^y, \mathcal{A})) - \frac{1}{2} \cdot \sum_{\substack{m^x \\ \in \Theta(\mathbb{M}^f)}} \phi(m^x, \mathcal{A})),
\end{aligned}$$

and we see that we need to also give a weight to the telomeres. We therefore add an additional vertex $\varnothing_{m^x}$ for extremity m^x that is connected to it via an edge. The edge being chosen in a matching then signifies that the extremity is a telomere. We therefore score this edge with the telomere score $s(m^x) = -\frac{1}{2}\phi(m^x, \mathcal{A})$. In order to still use the perfect matching framework, we also introduce edges between these additional vertices with weight 0. In more formal terms we define a weighted graph (V, E, w) with the following:

- $V = V_e \cup V_\varnothing$ with
  - $V_e = \{g^h : g \in \mathcal{F}(\mathbb{M}^f)\} \cup \{g^t : g \in \mathcal{F}(\mathbb{M}^f)\}$,
  - $V_\varnothing = \{\varnothing_{m^x} : m^x \in V_e\}$
- $E = \{(m^x, n^y) : m^x, n^y \in V_e, m^x \neq n^y\} \cup \{(m^x, \varnothing_{m^x}) : m^x \in V_e\}$
  $\cup\{(\varnothing_k, \varnothing_l) : \varnothing_k, \varnothing_l \in V_\varnothing, k \neq l\}$ with
  - $w(m^x, n^y) = -\phi(m^x n^y, \mathcal{A})$,
  - $w(m^x, \varnothing_{m^x}) = -\frac{1}{2}\phi(m^x, \mathcal{A})$ and
  - $w(\varnothing_k, \varnothing_l) = 0$.

A perfect matching on this graph then defines a general Breakpoint median for the set of genomes $\mathcal{A}$.

As an example regard the following three genomes

$$\mathcal{A} = \{\{[1\,2\,3]\}, \{(1\,2\,3)\}, \{[1], [2\,3]\}\}.$$

You can see the graph for these genomes in Figure 2.3. One maximal matching can be found in Figure 2.4. You can see that only with the inclusion of the telomere vertices, we can accurately score the median. It is also clear that there are now several ways to match up the remaining telomere vertices $\varnothing_{1^h}$, $\varnothing_{2^h}$, $\varnothing_{2^t}$ and $\varnothing_{3^t}$ (all with score 0) that do not change the genome we chose as median.

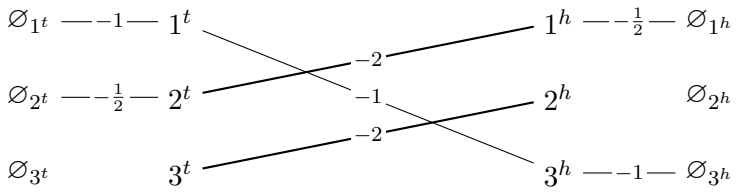

Figure 2.3: Median graph for the genomes $\{[1\,2\,3]\}$, $\{(1\,2\,3)\}$ and $\{[1],[2\,3]\}$. Edges with minimum weight (here -2) are highlighted. Edges with the maximum weight (here 0) are not displayed.

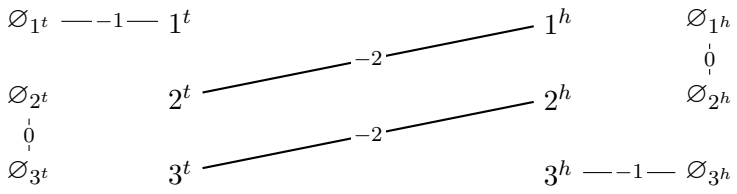

Figure 2.4: Minimum weight maximal matching on the median graph for the genomes $\{[1\,2\,3]\}$, $\{(1\,2\,3)\}$ and $\{[1],[2\,3]\}$.