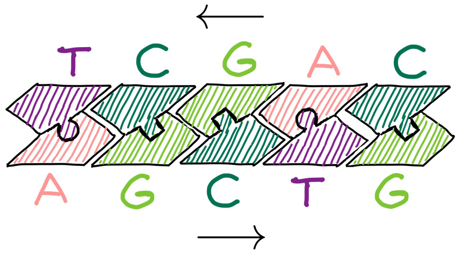
CHAPTER 2

Basic Definitions

A DNA molecule is a chain of anti-parallel complementary base pairs (bp), in which a base of type “**A**” is always paired with a base of type “**T**” and a base of type “**C**” is always paired with a base of type “**G**” (Figure 2.1).



Two complementary anti-parallel strands, linear or circular

Reverse complement:

AGCTG ↔ CAGCT

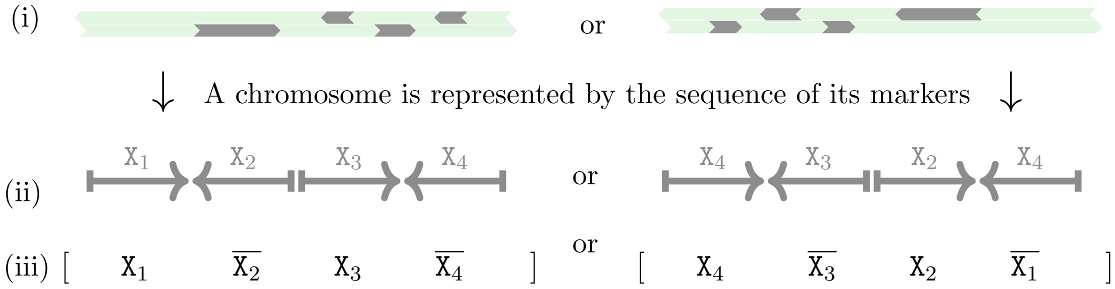
Figure 2.1: A DNA molecule is a chain of oriented *base pairs* (bp), which can potentially be broken at any position (between two consecutive base pairs)

Each DNA molecule that we consider is a *chromosome*, and a *genome* is a collection of chromosomes. But in our studies of large-scale genome rearrangements a high-level view of chromosomes and genomes is adopted.

2.1 Representing Markers, Chromosomes, Genomes

In this high-level view, in each chromosome only particular fragments are considered. Each of these fragments lies on one of the two complementary anti-parallel DNA strands of a chromosome and is called a *marker*. Usually, a marker corresponds to a *gene*, which is a DNA fragment coding for a protein. In our model, chromosomes can only be broken and subsequently repaired between markers or at the chromosome ends, that is, a marker can never be split into pieces. Therefore, the length of a marker is not particularly relevant here.

Each marker can be referred to by a unique identifier. Since a marker is oriented, we need to distinguish its two possible representations, for example by representing it with an arrow labeled with its unique identifier. Another way is to adopt the following textual notation: a marker $\mathtt{X}$ is represented by the symbol $\mathtt{X}$ itself, if it is read in direct orientation, or by the symbol $\overline{\mathtt{X}}$ (the *reverse complement* of $\mathtt{X}$), if it is read in reverse orientation. Yet another way of representing the orientation of a marker $\mathtt{X}$ is by distinguishing its two *extremities*: *head*, denoted by $\mathtt{X}^{\mathtt{h}}$, and *tail*, denoted by $\mathtt{X}^{\mathtt{t}}$. Let the set of markers in a genome $\mathbb{A}$ be denoted by $\mathcal{M}(\mathbb{A})$. Similarly, the set of extremities of markers in $\mathcal{M}(\mathbb{A})$ is $\mathsf{ext}(\mathbb{A}) = \bigcup\limits_{\mathtt{X}_i \in \mathcal{M}(\mathbb{A})} \{\mathtt{X}_i^{\mathtt{h}}, \mathtt{X}_i^{\mathtt{t}}\}$.


(i)
or
(ii)
$\mathtt{X}_1$
$\mathtt{X}_2$
$\mathtt{X}_3$
$\mathtt{X}_4$
or
$\mathtt{X}_4$
$\mathtt{X}_3$
$\mathtt{X}_2$
$\mathtt{X}_1$
(iii) [$\mathtt{X}_1$ $\overline{\mathtt{X}_2}$ $\mathtt{X}_3$ $\overline{\mathtt{X}_4}$]
or
[$\mathtt{X}_4$ $\overline{\mathtt{X}_3}$ $\mathtt{X}_2$ $\overline{\mathtt{X}_1}$]

↓ A chromosome is represented by the sequence of its markers ↓

Figure 2.2: (i) Linear chromosome as a sequence of markers represented with (ii) labeled arrows or (iii) oriented symbols.

Each chromosome is then represented by a sequence of oriented *markers*. In our notation, all markers of a chromosome K are concatenated in a string s that is flanked by square brackets if K is linear, or by parentheses if K is circular. Let $\overline{s}$ be the *reverse complement* of s, that is the string obtained by reverting the order and the orientation of the markers in s. Note that K can be equally represented by s or by $\overline{s}$ and, if K is circular, by any circular rotation of each of these two sequences. As an example, let K be a linear chromosome whose sequence of markers can be equally represented by $[\mathtt{X}_1\,\overline{\mathtt{X}_2}\,\mathtt{X}_3\,\overline{\mathtt{X}_4}]$ or by $[\mathtt{X}_4\,\overline{\mathtt{X}_3}\,\mathtt{X}_2\,\overline{\mathtt{X}_1}]$, as illustrated in Figure 2.2.

The neighborhood between extremities of two consecutive markers in a chromosome is called an *adjacency*. An adjacency is represented by the unordered pair of adjacent extremities. For example, an adjacency between $\mathtt{X}_1^{\mathtt{h}}$ and $\mathtt{X}_2^{\mathtt{h}}$ can be represented as $\{\mathtt{X}_1^{\mathtt{h}}, \mathtt{X}_2^{\mathtt{h}}\}$, or simply abbreviated as $\mathtt{X}_1^{\mathtt{h}}\mathtt{X}_2^{\mathtt{h}}$ ($\equiv \mathtt{X}_2^{\mathtt{h}}\mathtt{X}_1^{\mathtt{h}}$). Therefore, $\mathtt{X}_1^{\mathtt{h}}\mathtt{X}_2^{\mathtt{h}}$, $\mathtt{X}_2^{\mathtt{t}}\mathtt{X}_3^{\mathtt{t}}$, $\mathtt{X}_3^{\mathtt{h}}\mathtt{X}_4^{\mathtt{h}}$ are the adjacencies of chromosome K above. A linear chromosome such as K also has two *telomeres*, which are the marker extremities at the chromosome ends, that do not form adjacencies with other markers. In our example the telomeres are $\mathtt{X}_1^{\mathtt{t}}$ and $\mathtt{X}_4^{\mathtt{t}}$.

A genome $\mathbb{A}$ is then defined by its set $\mathfrak{C}(\mathbb{A})$ of chromosomes. We denote respectively by $\kappa(\mathbb{A})$ and $\theta(\mathbb{A})$ the numbers of linear and of circular chromosomes in genome $\mathbb{A}$. A genome $\mathbb{A}$ is said to be *unichromosomal* when it consists of a single chromosome or *multichromosomal* otherwise. Moreover, $\mathbb{A}$ is said to be *circular* when all its chromosomes are circular, *linear* when all its chromosomes are linear, or *mixed* otherwise. We also adopt the shorter notations *unilinear* for unichromosomal linear, *unicircular* for unichromosomal circular, *multilinear* for multichromosomal linear and *multicircular* for multichromosomal circular. Denote, respectively, by $\mathsf{adj}(\mathbb{A})$ and $\mathsf{tel}(\mathbb{A})$ the sets of adjacencies and telomeres in the chromosomes of genome $\mathbb{A}$.

For example, let a multilinear genome $\mathbb{A}$ be composed of chromosomes $[\mathtt{X}_1\,\overline{\mathtt{X}_2}\,\mathtt{X}_3\,\overline{\mathtt{X}_4}]$ and $[\mathtt{X}_5\,\overline{\mathtt{X}_6}\,\overline{\mathtt{X}_7}]$. A possible representation of $\mathbb{A}$ is then $\mathfrak{C}(\mathbb{A}) = \{[\mathtt{X}_1\,\overline{\mathtt{X}_2}\,\mathtt{X}_3\,\overline{\mathtt{X}_4}], [\mathtt{X}_5\,\overline{\mathtt{X}_6}\,\overline{\mathtt{X}_7}]\}$. Note that $\mathcal{M}(\mathbb{A}) = \{\mathtt{X}_1, \mathtt{X}_2, \mathtt{X}_3, \mathtt{X}_4, \mathtt{X}_5, \mathtt{X}_6, \mathtt{X}_7\}$ and $\mathsf{ext}(\mathbb{A}) = \{\mathtt{X}_1^{\mathtt{h}}, \mathtt{X}_1^{\mathtt{t}}, \mathtt{X}_2^{\mathtt{h}}, \mathtt{X}_2^{\mathtt{t}}, \ldots, \mathtt{X}_7^{\mathtt{h}}, \mathtt{X}_7^{\mathtt{t}}\}$. Furthermore, $\mathsf{adj}(\mathbb{A}) = \{\mathtt{X}_1^{\mathtt{h}}\mathtt{X}_2^{\mathtt{h}}, \mathtt{X}_2^{\mathtt{t}}\mathtt{X}_3^{\mathtt{t}}, \mathtt{X}_3^{\mathtt{h}}\mathtt{X}_4^{\mathtt{h}}, \mathtt{X}_5^{\mathtt{h}}\mathtt{X}_6^{\mathtt{h}}, \mathtt{X}_6^{\mathtt{t}}\mathtt{X}_7^{\mathtt{h}}\}$ and $\mathsf{tel}(\mathbb{A}) = \{\mathtt{X}_1^{\mathtt{t}}, \mathtt{X}_4^{\mathtt{t}}, \mathtt{X}_5^{\mathtt{t}}, \mathtt{X}_7^{\mathtt{t}}\}$.

2.2 Genome rearrangements or mutations

Recall that, in our model, chromosomes cannot be broken within a marker. This means that they can only be broken at adjacencies (between two marker extremities) or at telomeres (next to a marker extremity that is at the end of a linear chromosome). In other words, a *cut* performed on a chromosome K of a genome $\mathbb{A}$ separates two adjacent markers of K, or "opens" one of its telomeres, if K is linear. Each adjacency or telomere of a genome is therefore called a *cutpoint*. Let the set of cutpoints of genome $\mathbb{A}$ be denoted by $\psi(\mathbb{A})$: $\psi(\mathbb{A}) = \mathsf{adj}(\mathbb{A}) \cup \mathsf{tel}(\mathbb{A})$, with cardinality $|\psi(\mathbb{A})| = |\mathsf{adj}(\mathbb{A})| + |\mathsf{tel}(\mathbb{A})| = |\mathcal{M}(\mathbb{A})| + \kappa(\mathbb{A})$.

By breaking chromosomes within cutpoints and subsequently repairing the corresponding open ends, in our studies a genome can be transformed with two types of rearrangements.

The first type are *structural rearrangements*, which change the order, orientations of genes and numbers of chromosomes. These include, for example, fusions and fissions of chromosomes, translocations, reciprocal translocations and intra-chromosomal inversions. The second type are *content-modifying rearrangements*, which can replace, delete or include segments of genes.

In any case, a rearrangement $\rho = (\textcircled{s}_{\mathrm{a}} \to \textcircled{s}_{\mathrm{b}})$ transforms one *starting* state $\textcircled{s}_{\mathrm{a}}$ into another *resulting* state $\textcircled{s}_{\mathrm{b}}$. In the models considered in this text, any rearrangement is reversible. The reverse of ρ is denoted by $\rho^{-1} = (\textcircled{s}_{\mathrm{b}} \to \textcircled{s}_{\mathrm{a}})$. Note that $(\rho^{-1})^{-1} = \rho$.

In the following we describe how exactly structural and content-modifying rearrangements are modeled.

2.2.1 Structural rearrangements: double-cut-and-join (DCJ) operations

A *double-cut and join* or *DCJ* applied on genome $\mathbb{A}$ is the operation that performs cuts in one or two different cutpoints of $\mathbb{A}$. In the case of a single cut, the corresponding cutpoint must be an adjacency. In the case of two cuts, the corresponding cutpoints can be in distinct chromosomes or in the same chromosome of $\mathbb{A}$. In any case, a DCJ creates two to four open ends, and joins these open ends in a different way [99].

For example, let $\mathfrak{C}(\mathbb{A}) = \{[\mathtt{X}_1\overline{\mathtt{X}_2}\mathtt{X}_3\overline{\mathtt{X}_4}], [\mathtt{X}_5\overline{\mathtt{X}_6}\overline{\mathtt{X}_7}]\}$, and consider a DCJ that cuts between extremities $\mathtt{X}_2^{\mathtt{t}}$ and $\mathtt{X}_3^{\mathtt{t}}$ of its first chromosome and between extremities $\mathtt{X}_6^{\mathtt{t}}$ and $\mathtt{X}_7^{\mathtt{h}}$ of its second chromosome, creating segments $\mathtt{X}_1\overline{\mathtt{X}_2}\bullet$, $\bullet\mathtt{X}_3\overline{\mathtt{X}_4}$, $\mathtt{X}_5\overline{\mathtt{X}_6}\bullet$ and $\bullet\overline{\mathtt{X}_7}$ (where the symbols $\bullet$ represent the open ends). If we join the first with the fourth and the second with the third open end, we get $\mathfrak{C}(\mathbb{A}') = \{[\mathtt{X}_1\overline{\mathtt{X}_2}\overline{\mathtt{X}_7}], [\mathtt{X}_5\overline{\mathtt{X}_6}\mathtt{X}_3\overline{\mathtt{X}_4}]\}$, that is, the described DCJ operation is a reciprocal translocation transforming $\mathbb{A}$ into $\mathbb{A}'$.

Indeed, a DCJ operation can correspond not only to a translocation but to several structural rearrangements, such as an inversion, a fusion, a fission, a chromosome circularization or linearization, a circular excision or integration, as we will describe below.

Let γ_i denote a marker extremity. In general, the cutpoints affected by a DCJ operation compose a (*rearrangement*) *DCJ-state* that can be of three types:

- $\textcircled{r}^4$: involves 4 marker extremities forming two adjacencies, e. g. $\gamma_1\gamma_2$ and $\gamma_3\gamma_4$;
- $\textcircled{r}^3$: involves 3 marker extremities forming one adjacency and one telomere, e. g. $\gamma_1\gamma_2$ and γ_3;
- $\textcircled{r}^{2|1}$: involves 2 marker extremities forming one adjacency, e. g. $\gamma_1\gamma_2$;
- $\textcircled{r}^{2|2}$: involves 2 marker extremities forming two telomeres, e. g. γ_1 and γ_2.

A DCJ operation can only transform a state of one type into another state of the same type, with one exception: if the starting state is of type $\textcircled{r}^{2|1}$, then the resulting state is of type $\textcircled{r}^{2|2}$ and *vice-versa.*

The possible DCJ operations are described in detail as follows:

DCJ operation of type 1: involves states of type $\textcircled{r}^4$, whose two cutpoints are adjacencies. The possible DCJ operations in this class can be represented as a circular chain of three states, so that each state can be transformed into each of the other two via a DCJ operation (see Figure 2.3).

$$\begin{tikzcd}
& \boxed{\textcircled{r}^4_{\mathrm{a}} = \{\gamma_1\gamma_2, \gamma_3\gamma_4\}} \arrow[dl, leftrightarrow] \arrow[dr, leftrightarrow] & \\
\boxed{\textcircled{r}^4_{\mathrm{b}} = \{\gamma_1\gamma_3, \gamma_2\gamma_4\}} \arrow[rr, leftrightarrow] & & \boxed{\textcircled{r}^4_{\mathrm{c}} = \{\gamma_1\gamma_4, \gamma_2\gamma_3\}}
\end{tikzcd}$$

Figure 2.3: A DCJ operation of type 1 involves two adjacencies, with two possibilities of rejoining the open extremities in a different way

Note that any of the operations described above can be represented as $\rho = (\textcircled{r}^4_{\mathrm{x}} \to \textcircled{r}^4_{\mathrm{y}})$, whose reverse is $\rho^-1 = (\textcircled{r}^4_{\mathrm{y}} \to \textcircled{r}^4_{\mathrm{x}})$.

1.1: If in $\textcircled{r}^4_{\mathrm{a}}$, the adjacencies $\gamma_1\gamma_2$ and $\gamma_3\gamma_4$ are in distinct linear chromosomes, then in the other two states ($\textcircled{r}^4_{\mathrm{b}}$ and $\textcircled{r}^4_{\mathrm{c}}$) each adjacency is also in a distinct linear chromosome, therefore all six possible DCJ operations represented in this subcase are reciprocal translocations. This is the situation of the example given in the beginning of this section.

1.2: If in $\textcircled{r}_{\mathrm{a}}^{4}$, the adjacencies $\gamma_1\gamma_2$ and $\gamma_3\gamma_4$ are in distinct chromosomes, at least one of the two being circular, then in the other two states ($\textcircled{r}_{\mathrm{b}}^{4}$ and $\textcircled{r}_{\mathrm{c}}^{4}$) both adjacencies are in the same chromosome. In this subcase, both $\rho_1 = (\textcircled{r}_{\mathrm{a}}^{4} \rightarrow \textcircled{r}_{\mathrm{b}}^{4})$ and $\rho_2 = (\textcircled{r}_{\mathrm{a}}^{4} \rightarrow \textcircled{r}_{\mathrm{c}}^{4})$ are circular integrations, both $\rho_1^{-1} = (\textcircled{r}_{\mathrm{b}}^{4} \rightarrow \textcircled{r}_{\mathrm{a}}^{4})$ and $\rho_2{-}1 = (\textcircled{r}_{\mathrm{c}}^{4} \rightarrow \textcircled{r}_{\mathrm{a}}^{4})$ are circular excisions and both $\rho_3 = (\textcircled{r}_{\mathrm{b}}^{4} \rightarrow \textcircled{r}_{\mathrm{c}}^{4})$ and $\rho_3^{-1} = (\textcircled{r}_{\mathrm{c}}^{4} \rightarrow \textcircled{r}_{\mathrm{b}}^{4})$ are inversions.

DCJ operation of type 2: involves states of type $\textcircled{r}^{3}$, where one cutpoint is an adjacency and the other is a telomere. The possible DCJ operations in this class can be represented as a circular chain of three states, so that each state can be transformed into each of the other two via a DCJ operation(see Figure 2.4).

$$\begin{tikzcd}
& \boxed{\textcircled{r}_{\mathrm{a}}^{3} = \{\gamma_1\gamma_2, \gamma_3\}} \arrow[dl, leftrightarrow] \arrow[dr, leftrightarrow] & \\
\boxed{\textcircled{r}_{\mathrm{b}}^{3} = \{\gamma_1\gamma_3, \gamma_2\}} \arrow[rr, leftrightarrow] & & \boxed{\textcircled{r}_{\mathrm{c}}^{3} = \{\gamma_1, \gamma_2\gamma_3\}}
\end{tikzcd}$$

Figure 2.4: A DCJ operation of type 2 involves two adjacencies and one telomere, with two possibilities of rejoining the open extremities in a different way

Note that any of the operations described above can be represented as $\rho = (\textcircled{r}_{\mathrm{x}}^{3} \rightarrow \textcircled{r}_{\mathrm{y}}^{3})$, whose reverse is ρ is $\rho^{-}1 = (\textcircled{r}_{\mathrm{y}}^{3} \rightarrow \textcircled{r}_{\mathrm{x}}^{3})$.

2.1: If in $\textcircled{r}_{\mathrm{a}}^{3}$, the adjacency $\gamma_1\gamma_2$ and the telomere γ_3 are in distinct linear chromosomes, then in the other two states ($\textcircled{r}_{\mathrm{b}}^{3}$ and $\textcircled{r}_{\mathrm{c}}^{3}$) the adjacency and the telomere are also in distinct linear chromosomes, therefore all six possible DCJ operations represented in this subcase are translocations.

2.2: In $\textcircled{r}_{\mathrm{a}}^{3}$ the extremity γ_3 is a telomere, obviously in a linear chromosome. If the adjacency $\gamma_1\gamma_2$ is in a circular chormosome, then in the other two states ($\textcircled{r}_{\mathrm{b}}^{3}$ and $\textcircled{r}_{\mathrm{c}}^{3}$) the adjacency and the telomere are in the same linear chromosome. In this subcase, both $\rho_1 = (\textcircled{r}_{\mathrm{a}}^{3} \rightarrow \textcircled{r}_{\mathrm{b}}^{3})$ and $\rho_2 = (\textcircled{r}_{\mathrm{a}}^{3} \rightarrow \textcircled{r}_{\mathrm{c}}^{3})$ are circular integrations, both $\rho_1^{-1} = (\textcircled{r}_{\mathrm{b}}^{3} \rightarrow \textcircled{r}_{\mathrm{a}}^{3})$ and $\rho_2^{-1} = (\textcircled{r}_{\mathrm{c}}^{3}) \rightarrow \textcircled{r}_{\mathrm{a}}^{3})$ are circular excisions and both $\rho_3 = (\textcircled{r}_{\mathrm{b}}^{3} \rightarrow \textcircled{r}_{\mathrm{c}}^{3})$ and $\rho_3^{-1} = (\textcircled{r}_{\mathrm{c}}^{3} \rightarrow \textcircled{r}_{\mathrm{b}}^{3})$ are inversions.

DCJ operation of type 3: involves states of types $\textcircled{r}^{2|2}$ and $\textcircled{r}^{2|1}$. Let one telomere be γ_1 and the other be γ_2. Then the possible DCJ operations in this class can be represented as a circular chain of two states, so that each state can be transformed into the other via a DCJ operation (see Figure 2.5).

$$\begin{tikzcd}
\boxed{\textcircled{r}^{2|2} = \{\gamma_1, \gamma_2\}} \arrow[d, leftrightarrow] \\
\boxed{\textcircled{r}^{2|1} = \{\gamma_1\gamma_2\}}
\end{tikzcd}$$

Figure 2.5: A DCJ operation of type 3 involves one adjacency or two telomeres, with a single possibility of rejoining the open extremities in a different way

Note that one of the operations described above can be represented as $\rho = (\textcircled{r}^{2|2} \to \textcircled{r}^{2|1})$ and its reverse is $\rho^-1 = (\textcircled{r}^{2|1} \to \textcircled{r}^{2|2})$.

3.1: If in $\textcircled{r}^{2|2}$, the telomeres γ_1 and γ_2 are in distinct linear chromosomes, then in $\textcircled{r}^{2|1}$ they form an adjacency in a linear chromosome. In this subcase, $\rho = (\textcircled{r}^{2|2} \to \textcircled{r}^{2|1})$ is a linear fusion while $\rho^{-1} = (\textcircled{r}^{2|1} \to \textcircled{r}^{2|2})$ is a linear fission.

3.2: If in $\textcircled{r}^{2|2}$, the telomeres γ_1 and γ_2 are in the same linear chromosome, then in $\textcircled{r}^{2|1}$ they form an adjacency in a circular chromosome. In this subcase, $\rho = (\textcircled{r}^{2|2} \to \textcircled{r}^{2|1})$ is a circularization while $\rho^{-1} = (\textcircled{r}^{2|1} \to \textcircled{r}^{2|2})$ is a linearization.

2.2.2 Content-modifying rearrangements: substitutions and indel operations

The content of a chromosome can be modified with *substitutions* of blocks of contiguous markers. Special cases of substitutions are *insertions* and *deletions* of blocks of contiguous markers, collectively called *indel* operations. As an example, consider the simple deletion of block $\overline{\mathtt{X}_6}\,\mathtt{X}_3$ from linear chromosome $[\mathtt{X}_5\overline{\mathtt{X}_6}\,\mathtt{X}_3\,\overline{\mathtt{X}_4}]$, resulting in the shorter linear chromosome $[\mathtt{X}_5\,\overline{\mathtt{X}_4}]$.

A substitution affects a single chromosome, therefore at most one chromosome can be entirely substituted, deleted or inserted at once. In general, the cutpoints of a substitution compose a (*content*) *substitution-state* that includes a block of contiguous markers w and can be of four types:

- $\textcircled{c}^2$: the block w is flanked by 2 marker extremities, e. g. $\gamma_1 w \gamma_2$;
- $\textcircled{c}^1$: the block w is at the end of a linear chromosome, flanked by a marker extremity at one side, e. g. $\gamma_1 w$;
- $\textcircled{c}^{\mathrm{L}}$: the block w is a whole linear chromosome;
- $\textcircled{c}^{\mathrm{C}}$: the block w is a whole circular chromosome.

A substitution transforms one *starting* state into another *resulting* state and both states must be flanked by exactly the same marker extremities. Furthermore, we do not allow the substitution of a linear by a circular chromosome and vice-versa. Therefore a substitution operation can only transform a state of one type into another state of the same type. The possible substitutions are described in detail as follows, assuming that $w \neq \varepsilon$, while w' can be equal to ε.

Substitution of type 1: affects an inner segment of a chromosome. The possible operations in this class can be represented as a circular chain of two states, so that each state can be transformed into the other via a substitution (see Figure 2.6).

$$\begin{tikzcd} \boxed{\textcircled{c}^2_{\mathrm{a}} = \gamma_1 w \gamma_2} \arrow[r, shift left, harpoon, "\text{del if } w'=\varepsilon"] & \boxed{\textcircled{c}^2_{\mathrm{b}} = \gamma_1 w' \gamma_2} \arrow[l, shift left, harpoon, "\text{ins if } w'=\varepsilon"] \end{tikzcd}$$

Figure 2.6: A substitution of type 1 affects an inner segment of a chromosome.

Substitution of type 2: affects a segment at one end of a linear chromosome. The possible operations in this class can be represented as a circular chain of two states, so that each state can be transformed into the other via a substitution (see Figure 2.7).

$$\begin{tikzcd}
\boxed{\textcircled{c}^{1}_{\mathrm{a}} = \gamma_1 w} \arrow[r, shift left, "\text{del if } w'=\varepsilon"] & \boxed{\textcircled{c}^{1}_{\mathrm{b}} = \gamma_1 w'} \arrow[l, shift left, "\text{ins if } w'=\varepsilon"]
\end{tikzcd}$$

Figure 2.7: A substitution of type 2 affects a segment at one end of a linear chromosome.

Substitution of type 3: affects a whole linear chromosome. The possible operations in this class can be represented as a circular chain of two states, so that each state can be transformed into the other via a substitution (see Figure 2.8).

$$\begin{tikzcd}
\boxed{\textcircled{c}^{\mathrm{L}}_{\mathrm{a}} = [w]} \arrow[r, shift left, "\text{del if } w'=\varepsilon"] & \boxed{\textcircled{c}^{\mathrm{L}}_{\mathrm{b}} = [w']} \arrow[l, shift left, "\text{ins if } w'=\varepsilon"]
\end{tikzcd}$$

Figure 2.8: A substitution of type 3 affects a whole linear chromosome.

Substitution of type 4: affects a whole circular chromosome. The possible operations in this class can be represented as a circular chain of two states, so that each state can be transformed into the other via a substitution (see Figure 2.9).

$$\begin{tikzcd}
\boxed{\textcircled{c}^{\mathrm{C}}_{\mathrm{a}} = (w)} \arrow[r, shift left, "\text{del if } w'=\varepsilon"] & \boxed{\textcircled{c}^{\mathrm{C}}_{\mathrm{b}} = (w')} \arrow[l, shift left, "\text{ins if } w'=\varepsilon"]
\end{tikzcd}$$

Figure 2.9: A substitution of type 4 affects a whole circular chromosome.

Any of the operations described above can be represented as $\rho = (\textcircled{c}^{\tau}_{\mathrm{x}} \to \textcircled{c}^{\tau}_{\mathrm{y}})$, whose reverse is $\rho^{-}1 = (\textcircled{c}^{\tau}_{\mathrm{y}} \to \textcircled{c}^{\tau}_{\mathrm{x}})$, where τ denotes one of the four described types.

In the special cases of indels, we have the particular situations as follows: the inner block must be non-empty in the start state of a deletion and in the final state of an insertion, while it must be empty in the start state of an insertion and in the final state of a deletion. Therefore, if ρ is an insertion, then ρ^{-1} is a deletion and *vice versa*.

In Figure 2.10 we show an example of (i) a sorting scenario and (ii) its reverse; both with two DCJs and one content-modifying operation.

2.3 Family-annotated genomes

The markers of a genome can be grouped into *families*. Each marker must belong to a single family and the markers in the same family are considered to be equivalent.

For example, consider the linear genome $\mathfrak{C}(\mathbb{A}) = \{[\mathtt{X}_1\overline{\mathtt{X}_2}\mathtt{X}_3\overline{\mathtt{X}_4}], [\mathtt{X}_5\overline{\mathtt{X}_6}\overline{\mathtt{X}_7}]\}$, whose first chromosome is represented in Figure 2.11, and suppose that marker $\mathtt{X}_1$ belongs to family $\mathtt{1}$, markers $\mathtt{X}_2$, $\mathtt{X}_4$ and $\mathtt{X}_5$ belong to family $\mathtt{2}$, markers $\mathtt{X}_3$ and $\mathtt{X}_6$ belong to family $\mathtt{3}$ and marker $\mathtt{X}_7$ belongs to family $\mathtt{4}$. The set of families that occur in $\mathbb{A}$, denoted by $\mathcal{F}(\mathbb{A})$ is then $\mathcal{F}(\mathbb{A}) = \{\mathtt{1}, \mathtt{2}, \mathtt{3}, \mathtt{4}\}$.

Let $\eta(\mathtt{X})$ be a function that gives the family-annotation of a marker $\mathtt{X}$. The chromosome representation of $\mathbb{A}$ according to η is then denoted by $\eta(\mathfrak{C}(\mathbb{A}))$, which in our example is $\eta(\mathfrak{C}(\mathbb{A})) = \{[\mathtt{1}\overline{\mathtt{2}}\mathtt{3}\overline{\mathtt{2}}], [\mathtt{2}\overline{\mathtt{3}}\overline{\mathtt{4}}]\}$. We can now define the multiset of family-annotated markers $\mathcal{G}(\mathbb{A}) = \eta(\mathcal{M}(\mathbb{A}))$ and the multiset of family-annotated marker extremities $\mathsf{EXT}(\mathbb{A}) =$

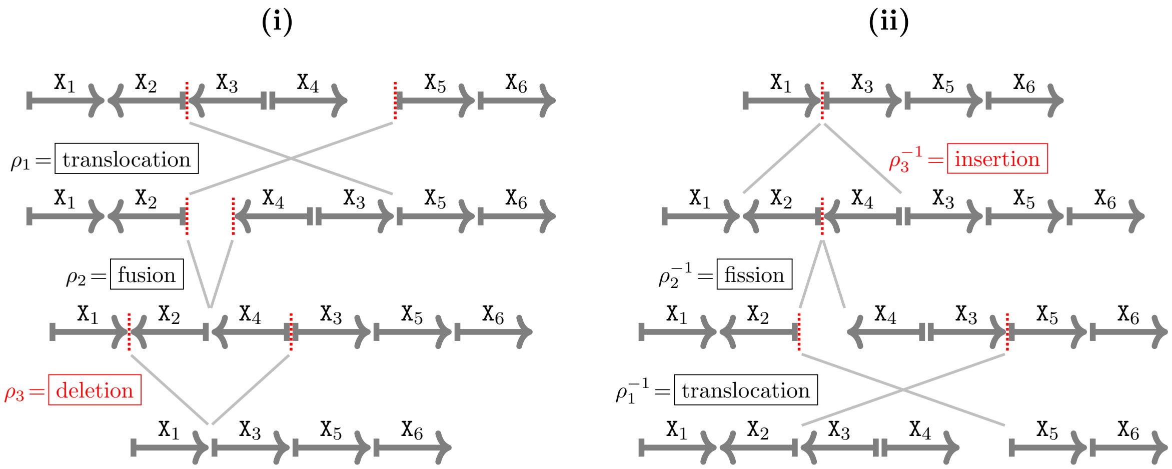
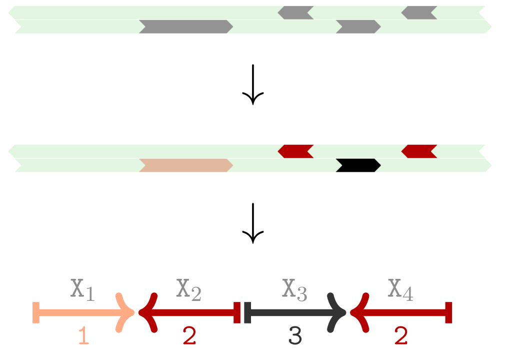


Figure 2.10: (i) A sequence $s = \rho_1\rho_2\rho_3$ whose three subsequent steps are $\rho_1 = (\{\mathtt{X}_2{}^{\mathtt{t}}\mathtt{X}_3{}^{\mathtt{h}}, \mathtt{X}_5{}^{\mathtt{t}}\} \rightarrow \{\mathtt{X}_5{}^{\mathtt{t}}\mathtt{X}_3{}^{\mathtt{h}}, \mathtt{X}_2{}^{\mathtt{t}}\})$, $\rho_2 = (\{\mathtt{X}_2{}^{\mathtt{t}}, \mathtt{X}_4{}^{\mathtt{h}}\} \rightarrow \{\mathtt{X}_2{}^{\mathtt{t}}\mathtt{X}_4{}^{\mathtt{h}}\})$ and $\rho_3 = (\mathtt{X}_1{}^{\mathtt{h}}\overline{\mathtt{X}}_2\overline{\mathtt{X}}_4\mathtt{X}_3{}^{\mathtt{t}} \rightarrow \mathtt{X}_1{}^{\mathtt{h}}\mathtt{X}_3{}^{\mathtt{t}})$. (ii) Reversing s gives $s^{-1} = \rho_3^{-1}\rho_2^{-1}\rho_1^{-1}$.



$$\begin{array}{lccccr} [& \mathtt{X}_1 & \overline{\mathtt{X}_2} & \mathtt{X}_3 & \overline{\mathtt{X}_4} &] \\ [& 1 & \overline{2} & 3 & \overline{2} &] \end{array}$$

Figure 2.11: Family-annotated linear chromosome. The colors indicate to which family each marker belongs.

$\eta(\mathsf{ext}(\mathbb{A}))$. Although $\mathcal{G}(\mathbb{A})$ and $\mathsf{EXT}(\mathbb{A})$ are defined as multisets, each of their elements "remembers" to which element it corresponds in the set of markers or in the set of marker extremities:

$$\begin{array}{rclcccccccl} \mathcal{G}(\mathbb{A}) & = & \{ & 1, & 2, & 3, & 2, & 2, & 3, & 4 & \} \\ \eta\uparrow & & & \updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow & \\ \mathcal{M}(\mathbb{A}) & = & \{ & \mathtt{X}_1, & \mathtt{X}_2, & \mathtt{X}_3, & \mathtt{X}_4, & \mathtt{X}_5, & \mathtt{X}_6, & \mathtt{X}_7, & \} \end{array}$$

$$\begin{array}{rclcccccccl}
\mathsf{EXT}(\mathbb{A}) & = & \{ & 1^{\mathtt{h}},1^{\mathtt{t}}, & 2^{\mathtt{h}},2^{\mathtt{t}}, & 3^{\mathtt{h}},3^{\mathtt{t}}, & 2^{\mathtt{h}},2^{\mathtt{t}}, & 2^{\mathtt{h}},2^{\mathtt{t}}, & 3^{\mathtt{h}},3^{\mathtt{t}}, & 4^{\mathtt{h}},4^{\mathtt{t}} & \} \\
\eta\uparrow & & & \updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow & \\
\mathsf{ext}(\mathbb{A}) & = & \{ & \mathtt{X}_1^{\mathtt{h}},\mathtt{X}_1^{\mathtt{t}}, & \mathtt{X}_2^{\mathtt{h}},\mathtt{X}_2^{\mathtt{t}}, & \mathtt{X}_3^{\mathtt{h}},\mathtt{X}_3^{\mathtt{t}}, & \mathtt{X}_4^{\mathtt{h}},\mathtt{X}_4^{\mathtt{t}}, & \mathtt{X}_5^{\mathtt{h}},\mathtt{X}_5^{\mathtt{t}}, & \mathtt{X}_6^{\mathtt{h}},\mathtt{X}_6^{\mathtt{t}}, & \mathtt{X}_7^{\mathtt{h}},\mathtt{X}_7^{\mathtt{t}}, & \}
\end{array}$$

Indeed, the original unambiguous marker name can be understood as a “hidden” property of an annotated marker, and this hidden property can be acessed at any time. The same unambiguous correspondence exists for the (multi) set of annotated adjacencies $\mathsf{ADJ}(\mathbb{A}) = \eta(\mathsf{adj}(\mathbb{A}))$ and for the (multi) set of annotated telomeres $\mathsf{TEL}(\mathbb{A}) = \eta(\mathsf{tel}(\mathbb{A}))$:

$$\begin{array}{rclcccccl}
\mathsf{ADJ}(\mathbb{A}) & = & \{ & 1^{\mathtt{h}}2^{\mathtt{h}}, & 2^{\mathtt{t}}3^{\mathtt{t}}, & 3^{\mathtt{h}}2^{\mathtt{h}}, & 2^{\mathtt{h}}3^{\mathtt{h}}, & 3^{\mathtt{t}}4^{\mathtt{h}} & \} \\
\eta\uparrow & & & \updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow & \\
\mathsf{adj}(\mathbb{A}) & = & \{ & \mathtt{X}_1^{\mathtt{h}}\mathtt{X}_2^{\mathtt{h}}, & \mathtt{X}_2^{\mathtt{t}}\mathtt{X}_3^{\mathtt{t}}, & \mathtt{X}_3^{\mathtt{h}}\mathtt{X}_4^{\mathtt{h}}, & \mathtt{X}_5^{\mathtt{h}}\mathtt{X}_6^{\mathtt{h}}, & \mathtt{X}_6^{\mathtt{t}}\mathtt{X}_7^{\mathtt{h}}, & \}
\end{array}$$

$$\begin{array}{rclccccl}
\mathsf{TEL}(\mathbb{A}) & = & \{ & 1^{\mathtt{t}}, & 2^{\mathtt{t}}, & 2^{\mathtt{t}} & 4^{\mathtt{t}} & \} \\
\eta\uparrow & & & \updownarrow & \updownarrow & \updownarrow & \updownarrow & \\
\mathsf{tel}(\mathbb{A}) & = & \{ & \mathtt{X}_1^{\mathtt{t}} & \mathtt{X}_4^{\mathtt{t}}, & \mathtt{X}_5^{\mathtt{t}} & \mathtt{X}_7^{\mathtt{t}} & \}
\end{array}$$

For a given family $\mathtt{N} \in \mathcal{F}(\mathbb{A})$, we denote by $\Phi(\mathtt{N},\mathbb{A})$ the number of occurrences of $\mathtt{N}$ in $\mathbb{A}$. In our example above, $\Phi(2,\mathbb{A}) = 3$, $\Phi(3,\mathbb{A}) = 2$, and $\Phi(1,\mathbb{A}) = \Phi(4,\mathbb{A}) = 1$.

Any genome is *natural* and may contain duplicates from the same family (such as in the example given above). However, if a genome $\mathbb{A}$ contains a single marker from each family, that is, if $\Phi(\mathtt{N},\mathbb{A}) = 1$ for each $\mathtt{N} \in \mathcal{F}(\mathbb{A})$, then genome $\mathbb{A}$ is said to be *singular*. In this case, the sets of family-annotated markers $\mathcal{G}(\mathbb{A})$, marker extremities $\mathsf{EXT}(\mathbb{A})$, adjacencies $\mathsf{ADJ}(\mathbb{A})$ and telomeres $\mathsf{TEL}(\mathbb{A})$ are simple sets.

2.4 Types of pairs of family-annotated genomes

Given a pair of genomes $\mathbb{A}$ and $\mathbb{B}$, if the markers of $\mathbb{A}$ and $\mathbb{B}$ are collectively grouped into families, we can then compare the structural organization of $\mathbb{A}$ and $\mathbb{B}$. This is due to the fact that, if a marker $\mathtt{X}$ from genome $\mathbb{A}$ and a marker $\mathtt{Y}$ from genome $\mathbb{B}$ belong to the same family, markers $\mathtt{X}$ and $\mathtt{Y}$ are considered to be equivalent.

Any pair of family-annotated genomes $\mathbb{A}$ and $\mathbb{B}$ is said to be *natural*. If, however, genomes $\mathbb{A}$ and $\mathbb{B}$ share some common property according to their annotated markers, we may assign more specific classifications to the pair:

- If $\mathcal{F}(\mathbb{A}) = \mathcal{F}(\mathbb{B})$ and, for each $\mathtt{N} \in \mathcal{F}(\mathbb{A})$, we have $\Phi(\mathtt{N},\mathbb{A}) = \Phi(\mathtt{N},\mathbb{B})$, then the pair of genomes $\mathbb{A}$ and $\mathbb{B}$ is said to be *balanced*. Otherwise, if there is any $\mathtt{N}' \in \mathcal{F}(\mathbb{A}) \cup \mathcal{F}(\mathbb{B})$, such that $\Phi(\mathtt{N}',\mathbb{A}) \neq \Phi(\mathtt{N}',\mathbb{B})$, the pair is said to be *unbalanced*. Balanced genomes are said to have the *same content*, while unbalanced genomes are said to have *unequal contents*.
- If both genomes $\mathbb{A}$ and $\mathbb{B}$ are singular, then the pair of genomes $\mathbb{A}$ and $\mathbb{B}$ is said to be *singular*. A pair of singular genomes can be unbalanced but is always *unambiguous*, meaning that the correspondence of genes can be established in a unique way.

- A pair of genomes that is singular and balanced is said to be *canonical*. Two genomes in a canonical pair are therefore unambiguous and have the same content.

2.5 Comparing family-annotated genomes: sorting and distance

A sequence s of k operations that can be (structural) DCJs and (content-modifying) substitutions and indels transforming a genome $\mathbb{A}$ into another genome $\mathbb{B}$ is called a *sorting scenario* whose length is k. An example is given in Figure 2.12. The *sorting* problem consists of finding a parsimonious (minimum-length) scenario transforming one genome into the other.

Closely related to the sorting is the *genomic distance* problem, which consists of determining the length of any parsimonious sequence to sort one genome into the other. We denote by $\mathrm{d}_{\mathrm{DCJ}}^{\mathfrak{C}}(\mathbb{A}, \mathbb{B})$ the genomic distance of $\mathbb{A}$ and $\mathbb{B}$, where $\mathfrak{C}$ represents content-modifying operations that either correspond to substitutions (including indels) or are restricted to indels only.

Let ρ be a content-modifying or DCJ operation and let $\rho\mathbb{A}$ be the genome obtained after applying ρ to a genome $\mathbb{A}$. The operation ρ is said to be *optimal* with respect to the target genome $\mathbb{B}$ when $\mathrm{d}_{\mathrm{DCJ}}^{\mathfrak{C}}(\rho\mathbb{A}, \mathbb{B}) = \mathrm{d}_{\mathrm{DCJ}}^{\mathfrak{C}}(\mathbb{A}, \mathbb{B}) - 1$. Similarly, a sequence of k operations transforming $\mathbb{A}$ into $\mathbb{A}'$ is *optimal* with respect to $\mathbb{B}$, if $\mathrm{d}_{\mathrm{DCJ}}^{\mathfrak{C}}(\mathbb{A}', \mathbb{B}) = \mathrm{d}_{\mathrm{DCJ}}^{\mathfrak{C}}(\mathbb{A}, \mathbb{B}) - k$.

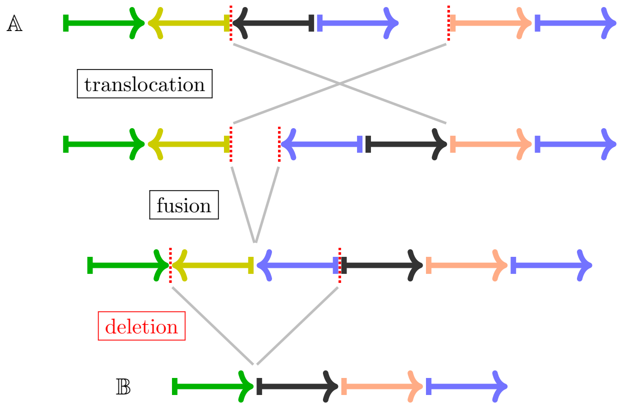


Figure 2.12: A scenario of length 3, composed of two DCJ operations (a translocation and a fusion) and one indel (a deletion), sorting multilinear annotated genome $\mathbb{A}$ into unilinear annotated genome $\mathbb{B}$.

2.5.1 Restriction on substitutions and indels

For equivalent markers `X` and `Y`, our model does not allow the deletion of `X` followed by the insertion of `Y`, nor the insertion of `X` followed by the deletion of `Y`. This restriction prevents the *free lunch* artifact of sorting one genome into the other by simply substituting their contents, or deleting the chromosomes of the first and inserting the chromosomes of the second, ignoring their common parts [100]. It implies that, in the substitution of a block of markers w by another block w', no marker in w can be equivalent to a marker in w'.

2.5.2 Complexity overview

Note that, if annotated genomes $\mathbb{A}$ and $\mathbb{B}$ are an unbalanced pair, they have some content differences. In this case one can only be completely sorted into the other with DCJ and content-modifying operations. Otherwise the genome pair is balanced, so that the genomes have the same content. For balanced genomes no content-modifying operation is allowed and only (structural) DCJ operations can be used for sorting. In this case we can denote the distance by $\mathrm{d}_{\mathrm{DCJ}}(\mathbb{A}, \mathbb{B})$.

As we will see along this text, when no ambiguity is present, that is, in the case of a singular or a canonical pair of genomes, both distance and sorting problems can be solved in linear time [10, 22]. In contrast, when some ambiguity is present, that is, in the case of a balanced or natural pair of genomes containing duplicates, both distance and sorting problems are NP-hard but optimal solutions can be computed via ILP [15,88]. A summary of these results is given in Table 2.1.

Table 2.1: Complexity of the DCJ-indel model for distinct types of family-annotated inputs

	# of occurrences of each family ≤ 1	**a family can occur ≥ 2 times**
balanced	canonical → linear [10]	non-canonical → NP-hard, ILP [88]
unbalanced	singular → linear [22]	natural → NP-hard, ILP [15]

2.5.3 Simplified notation for singular genomes

If genomes $\mathbb{A}$ and $\mathbb{B}$ form a singular pair, their sets of annotated markers $\mathcal{G}(.)$, annotated marker extremities $\mathsf{EXT}(.)$, annotated adjacencies $\mathsf{ADJ}(.)$ and annotated telomeres $\mathsf{TEL}(.)$ are simple sets. In this case, we refer to an annotated marker simply as marker. Furthermore, since there is no ambiguity, we ignore the original unannotated sets and assume a simpler notation in which $\mathbb{A}$ and $\mathbb{B}$ represent $\eta(\mathfrak{C}(\mathbb{A}))$ and $\eta(\mathfrak{C}(\mathbb{B}))$.

CHAPTER 3

DCJ model of canonical genomes

Let $\mathbb{A}$ and $\mathbb{B}$ be two annotated genomes and note that, if $\mathbb{A}$ and $\mathbb{B}$ form a canonical pair, then $\mathcal{F}(\mathbb{A}) = \mathcal{G}(\mathbb{A}) = \mathcal{G}(\mathbb{B}) = \mathcal{F}(\mathbb{B})$. We then denote by n_* the cardinality of all these sets: $n_* = |\mathcal{F}(\mathbb{A})| = |\mathcal{G}(\mathbb{A})| = |\mathcal{G}(\mathbb{B})| = |\mathcal{F}(\mathbb{B})|$. Recall that, in this case, only DCJ operations are used for sorting one genome into the other, and the corresponding DCJ distance is denoted by $\mathrm{d}_{\mathrm{DCJ}}(\mathbb{A}, \mathbb{B})$.

3.1 Relational graph of canonical genomes

Finding sorting DCJ operations and computing the DCJ distance between two canonical genomes $\mathbb{A}$ and $\mathbb{B}$ can be achieved with the help of the *relational graph* of $\mathbb{A}$ and $\mathbb{B}$ [18], denoted by $\mathsf{G}_{\mathrm{R}}(\mathbb{A}, \mathbb{B}) = (V, E)$, whose sets of vertices and edges are defined as follows:

1. The set of vertices is $V = V(\mathbb{A}) \cup V(\mathbb{B})$, where

   $V(\mathbb{A})$ contains a vertex for each extremity of each marker in $\mathcal{M}(\mathbb{A})$ and
   $V(\mathbb{B})$ contains a vertex for each extremity of each marker in $\mathcal{M}(\mathbb{B})$.

   Each vertex v has an identifier corresponding to the unannotated marker extremity it represents, and a label $\eta(v)$, corresponding to the annotated marker extremity it represents. Note that there are $4n_*$ vertices in $\mathsf{G}_{\mathrm{R}}(\mathbb{A}, \mathbb{B})$, $2n_*$ per genome.

2. The set of edges is $E = E_{\mathsf{adj}}(\mathbb{A}) \cup E_{\mathsf{adj}}(\mathbb{B}) \cup E_{\mathsf{ext}}$, where the *adjacency edges* are sets

$$\begin{aligned} E_{\mathsf{adj}}(\mathbb{A}) = \quad & \{uv : u, v \in V(\mathbb{A}) \text{ and } uv \in \mathsf{adj}(\mathbb{A})\} \text{ and} \\ E_{\mathsf{adj}}(\mathbb{B}) = \quad & \{uv : u, v \in V(\mathbb{B}) \text{ and } uv \in \mathsf{adj}(\mathbb{B})\}, \end{aligned}$$

and the set of *extremity edges* (whose cardinality is $2n_*$) is

$$E_{\mathsf{ext}} = \{uv : u \in V(\mathbb{A}) \text{ and } v \in V(\mathbb{B}) \text{ and } \eta(u) = \eta(v)\}.$$

Since any vertex in $\mathsf{G}_{\mathrm{R}}(\mathbb{A}, \mathbb{B})$ has exactly one extremity edge and at most one adjacency edge, its degree is one or two. Therefore, $\mathsf{G}_{\mathrm{R}}(\mathbb{A}, \mathbb{B})$ is a collection of paths and cycles. A vertex (marker extremity) that has no adjacency edge corresponds to a telomere and is therefore also called *telomere*. Each cutpoint of each genome is represented in the graph either as an adjacency edge or as a telomere. Recall that the number of cutpoints in $\mathbb{A}$ (respectively $\mathbb{B}$) is $n_* + \kappa(\mathbb{A})$ (respectively $n_* + \kappa(\mathbb{B})$).

Each connected component of the graph alternates between extremity edges and cutpoints, and we define the *length* of a component Γ to be the number of extremity edges in Γ. An *i-cycle* and an *i-path* denote respectively a cycle and a path of length i. Note that all cycles have even length, while paths start and end with extremity edges and can have even or odd lengths, called *even* and *odd paths* respectively.

A cycle can be simply denoted by $\mathbb{O}$. An odd path has one endpoint in a telomere from $\mathbb{A}$ and the other endpoint in a telomere from $\mathbb{B}$ and is called an $\mathbb{AB}$-path, simply denoted by $\mathbb{AB}$. Even paths have either both endpoints in $\mathbb{A}$, being an $\mathbb{AA}$-path, simply denoted by $\mathbb{AA}$, or both endpoints in $\mathbb{B}$, being a $\mathbb{BB}$-path, simply denoted by $\mathbb{BB}$. Even paths can also be called *unbalanced paths*, while odd paths are also called *balanced paths*. Let the sets of cycles, $\mathbb{AB}$-, $\mathbb{AA}$- and $\mathbb{BB}$-paths be respectively denoted by $\mathcal{C}$, $\mathcal{P}_{\mathbb{AB}}$, $\mathcal{P}_{\mathbb{AA}}$ and $\mathcal{P}_{\mathbb{BB}}$. Now let $\Upsilon(\Gamma)$ give the type of component Γ. For example, if Γ is a cycle, then $\Upsilon(\Gamma) = \mathbb{O}$. We can then explicitly write the above mentioned sets as:

$$\mathcal{C} = \{\Gamma \mid \Gamma \in \mathsf{G}_{\mathrm{R}}(\mathbb{A}, \mathbb{B}) \text{ and } \Upsilon(\Gamma) = \mathbb{O}\}\ ,$$

$$\mathcal{P}_{\mathbb{AB}} = \{\Gamma \mid \Gamma \in \mathsf{G}_{\mathrm{R}}(\mathbb{A}, \mathbb{B}) \text{ and } \Upsilon(\Gamma) = \mathbb{AB}\}\ ,$$
$$\mathcal{P}_{\mathbb{AA}} = \{\Gamma \mid \Gamma \in \mathsf{G}_{\mathrm{R}}(\mathbb{A}, \mathbb{B}) \text{ and } \Upsilon(\Gamma) = \mathbb{AA}\} \text{ and}$$
$$\mathcal{P}_{\mathbb{BB}} = \{\Gamma \mid \Gamma \in \mathsf{G}_{\mathrm{R}}(\mathbb{A}, \mathbb{B}) \text{ and } \Upsilon(\Gamma) = \mathbb{BB}\}.$$

Recall that $\kappa(\mathbb{A})$ and $\kappa(\mathbb{B})$ are the numbers of linear chromosomes in genomes $\mathbb{A}$ and $\mathbb{B}$. The endpoints of paths and chromosomes are the same telomeres, therefore we have $\kappa(\mathbb{A}) + \kappa(\mathbb{B}) = |\mathcal{P}_{\mathbb{AB}}| + |\mathcal{P}_{\mathbb{AA}}| + |\mathcal{P}_{\mathbb{BB}}|$. Furthermore, the numbers of telomeres in each genome are even. Since each $\mathbb{AA}$- or $\mathbb{BB}$-path takes either zero or two telomeres per genome and each $\mathbb{AB}$-path takes one telomere per genome, the number of $\mathbb{AB}$-paths must be even.

Related graphs. As illustrated in Figure 3.1, the relational graph has the same properties of two simpler graphs that were proposed earlier:

1. The first is the so-called *breakpoint graph*, originally proposed in the seminal studies of the *inversion sorting and distance* [7]. It can be derived from the relational graph by contracting each extremity edge e of $\mathsf{G}_{\mathrm{R}}(\mathbb{A}, \mathbb{B}) = (V, E)$ and assigning to the resulting single vertex the common annotation of the vertices that are connected by e. In the breakpoint graph there are only adjacency edges. Furthermore, cycles also have even length, while $\mathbb{AB}$-paths are even and $\mathbb{AA}$- and $\mathbb{BB}$-paths are odd.

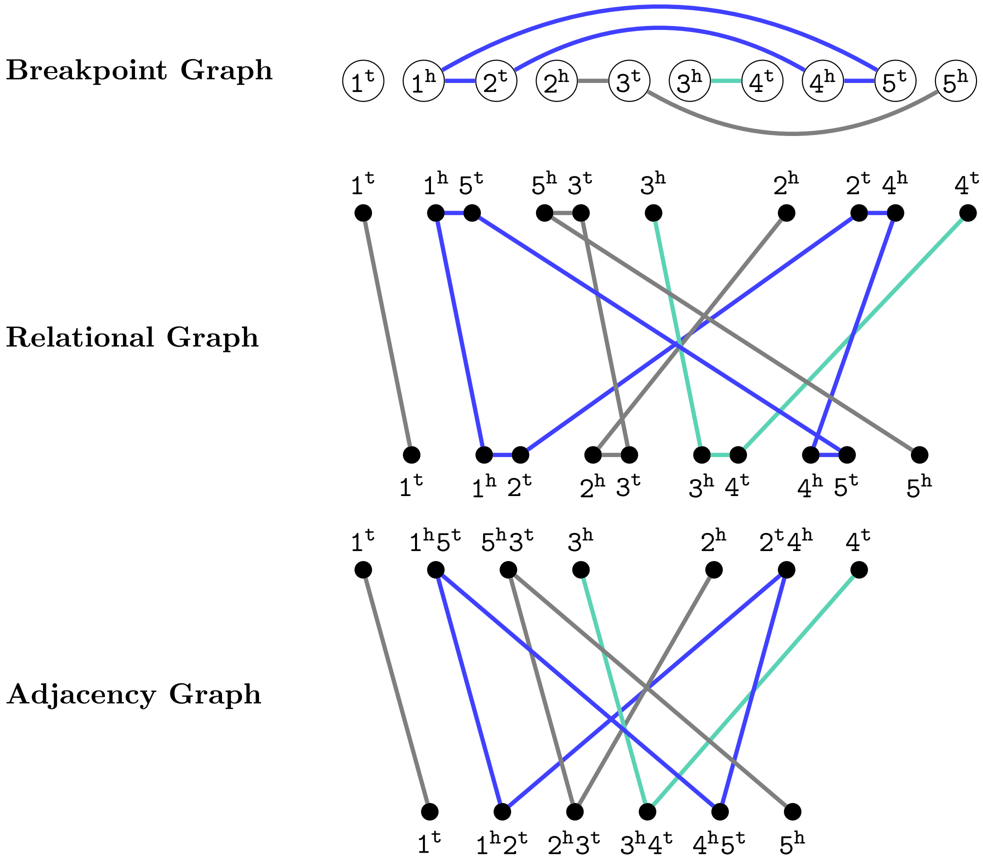

Breakpoint Graph
1t 1h 2t 2h 3t 3h 4t 4h 5t 5h
Relational Graph
1t 1h 5t 5h 3t 3h 2h 2t 4h 4t
1t 1h 2t 2h 3t 3h 4t 4h 5t 5h
Adjacency Graph
1t 1h5t 5h3t 3h 2h 2t4h 4t
1t 1h2t 2h3t 3h4t 4h5t 5h

Figure 3.1: For a canonical pair formed by multilinear genome $\mathbb{A} = \{\ [\texttt{1 5 3}],\ [\bar{\texttt{2}}\ \bar{\texttt{4}}]\ \}$ and unilinear genome $\mathbb{B} = \{\ [\texttt{1 2 3 4 5}]\ \}$, where $n_* = 5$, we represent the relational graph (in the middle) surrounded by the breakpoint graph (top) and by the adjacency graph (bottom). Note that the number of vertices in the breakpoint graph and the numbers of edges in both relational and adjacency graphs are equal to $2n_*$. In all graphs we have a (blue) 4-cycle, a (red) $\mathbb{AA}$-path and two $\mathbb{AB}$-paths.

2. The second is the so-called *adjacency graph*, which is bipartite and was originally proposed in the formalization of the *DCJ sorting and distance* [10]. It can be derived from the relational graph by contracting each adjacency edge a of $\mathsf{G}_{\mathrm{R}}(\mathbb{A}, \mathbb{B}) = (V, E)$, concatenating in the label of the resulting single vertex the annotations of the vertices that are connected by a. In other words, the vertices of the adjacency graph are the adjacencies and telomeres of $\mathbb{A}$ and $\mathbb{B}$ and all edges are extremity edges. Similarly to the relational graph, in the adjacency graph cycles have even length, $\mathbb{AB}$-paths are odd and $\mathbb{AA}$- and $\mathbb{BB}$-paths are even.

Relational graph of sorted and unsorted genomes. The smallest components that can occur in $\mathsf{G}_{\mathrm{R}}(\mathbb{A}, \mathbb{B})$ are 2-cycles and ($\mathbb{AB}$) 1-paths, denoted *short components*. A cycle whose length is greater than 2 or a path whose length is greater than 1 is called a *long component*. When canonical genomes $\mathbb{A}$ and $\mathbb{B}$ are identical (or *sorted*), their relational graph is a collection of short components: identical genomes have the same sets of adjacencies and telomeres, and each common adjacency corresponds to a 2-cycle while each common telomere corresponds to a 1-path in $\mathsf{G}_{\mathrm{R}}(\mathbb{A},\mathbb{B})$. Recall that the length of a component corresponds to its number of extremity edges and that $\mathsf{G}_{\mathrm{R}}(\mathbb{A},\mathbb{B})$ has $2n_*$ extremity edges. Therefore, for sorted genomes we have $2n_* = 2|\mathcal{C}| + |\mathcal{P}_{\mathbb{AB}}|$ and, consequently, $n_* = |\mathcal{C}| + \frac{|\mathcal{P}_{\mathbb{AB}}|}{2}$. Otherwise, when canonical genomes $\mathbb{A}$ and $\mathbb{B}$ are distinct (or *unsorted*), their relational graph contains at least one long component. Therefore, in this case $n_* > |\mathcal{C}| + \frac{|\mathcal{P}_{\mathbb{AB}}|}{2}$. With these observations we can already see that the DCJ operations that transform $\mathbb{A}$ into $\mathbb{B}$ must increase the numbers of cycles and/or of $\mathbb{AB}$-paths in $\mathsf{G}_{\mathrm{R}}(\mathbb{A},\mathbb{B})$. In the following we will present the results from Bergeron *et al.* [10], explaining how this can be achieved.

3.2 Types of DCJ operation with respect to the relational graph

Note that, with respect to its effect on the relational graph, a DCJ ρ cuts one or two components, and rejoins the open ends to transform them into one or two new components. A DCJ operation ρ is said to be *internal* (INT) to a single component Γ, when ρ cuts only at cutpoint(s) that are in Γ. The result of an internal DCJ ρ can be one component (distinct from Γ) or two components. In contrast, A DCJ operation ρ is said to be a *recombination* (REC) when ρ cuts at cutpoints of two distinct components Γ and Γ'. The result of a recombination ρ can be either a single component or two components (distinct from Γ and Γ').

In order to describe all possible DCJ operations, we adopt the following notation to represent the types of components and their lengths:

- $\mathbb{O}^{\langle \imath \rangle}$: even cycle with length $\imath \in \{2, 4, 6, \ldots\}$;
- $\mathbb{AB}^{\langle \imath \rangle}$: balanced path with length $\imath \in \{1, 3, 5, \ldots\}$;
- $\mathbb{AA}^{\langle \imath \rangle}$: unbalanced $\mathbb{AA}$-path with length $\imath \in \{2, 4, \ldots\}$;
- $\mathbb{BB}^{\langle \imath \rangle}$: unbalanced $\mathbb{BB}$-path with length $\imath \in \{2, 4, \ldots\}$;
- $\Gamma^{\langle \imath \rangle}$: component of any type with length $\imath \geq 1$.

The possible types of DCJ operation applied on cutpoints of genome $\mathbb{A}$ are described in Table 3.1 [10]. Note that each DCJ operation must be of one of three types:

Gaining DCJ. Either increases $|\mathcal{C}|$ by one or increases $|\mathcal{P}_{\mathbb{AB}}|$ by two.

Neutral DCJ. Does not change the cardinalities of the sets $\mathcal{C}$ and $\mathcal{P}_{\mathbb{AB}}$.

Losing DCJ. Either decreases $|\mathcal{C}|$ by one or decreases $|\mathcal{P}_{\mathbb{AB}}|$ by two.

Table 3.1: Effects of DCJ operations (applied on cutpoints of $\mathbb{A}$) on the relational graph

	$\boxed{\textbf{component(s)}}$ **[state]**	**DCJ**	$\boxed{\textbf{component(s)}}$ **[state]**		**affects...**
1.	$\boxed{\Gamma^{\langle \imath+\jmath \rangle}}$ $[\textcircled{r}^4 \,/\, \textcircled{r}^3]$	$\xrightleftharpoons[\text{losing (REC)}]{\text{gaining (INT)}}$	$\boxed{\begin{matrix}\mathbb{O}^{\langle \imath \rangle} \\ \Gamma^{\langle \jmath \rangle}\end{matrix}}$ $[\textcircled{r}^4 \,/\, \textcircled{r}^3]$	with $\begin{cases} \imath \in \{2,4,6,\ldots\} \\ \jmath \geq 1 \\ \Upsilon(\Gamma^{\langle \imath+\jmath \rangle}) = \Upsilon(\Gamma^{\langle \jmath \rangle}) \end{cases}$	$\Delta\lvert\mathcal{C}\rvert$ by ±1
2.	$\boxed{\Gamma^{\langle \imath+\jmath \rangle}}$ $[\textcircled{r}^4 \,/\, \textcircled{r}^3]$	$\xrightleftharpoons[\text{neutral (INT)}]{\text{neutral (INT)}}$	$\boxed{\check{\Gamma}^{\langle \imath+\jmath \rangle}}$ $[\textcircled{r}^4 \,/\, \textcircled{r}^3]$	with $\begin{cases} \imath \in \{2,4,6,\ldots\} \\ \jmath \geq 1 \\ \Gamma^{\langle \imath+\jmath \rangle}) \neq \check{\Gamma}^{\langle \imath+\jmath \rangle} \\ \Upsilon(\Gamma^{\langle \imath+\jmath \rangle}) = \Upsilon(\check{\Gamma}^{\langle \imath+\jmath \rangle}) \end{cases}$	
3.	$\boxed{\begin{matrix}\mathbb{AA}^{\langle \imath+\jmath \rangle} \\ \mathbb{BB}^{\langle \imath'+\jmath' \rangle}\end{matrix}}$ $[\textcircled{r}^4 \,/\, \textcircled{r}^3]$	$\xrightleftharpoons[\text{losing (REC)}]{\text{gaining (INT)}}$	$\boxed{\begin{matrix}\mathbb{AB}^{\langle \imath+\imath' \rangle} \\ \mathbb{AB}^{\langle \jmath+\jmath' \rangle}\end{matrix}}$ $[\textcircled{r}^4 \,/\, \textcircled{r}^3]$	with $\begin{cases} \imath \in \{0,2,4,\ldots\} \\ \jmath \in \{2,4,6,\ldots\} \\ \imath',\jmath' \in \{1,3,5,\ldots\} \end{cases}$	$\Delta\lvert\mathcal{P}_{\mathbb{AB}}\rvert$ by ±2
4.	$\boxed{\begin{matrix}\mathbb{AB}^{\langle \imath+\jmath' \rangle} \\ \mathbb{AB}^{\langle \jmath+\imath' \rangle}\end{matrix}}$ $[\textcircled{r}^4 \,/\, \textcircled{r}^3]$	$\xrightleftharpoons[\text{neutral (REC)}]{\text{neutral (REC)}}$	$\boxed{\begin{matrix}\mathbb{AB}^{\langle \imath+\imath' \rangle} \\ \mathbb{AB}^{\langle \jmath+\jmath' \rangle}\end{matrix}}$ $[\textcircled{r}^4 \,/\, \textcircled{r}^3]$	with $\begin{cases} \text{all } \mathbb{AB}\text{-paths distinct} \\ \imath \in \{0,2,4,\ldots\} \\ \jmath \in \{2,4,6,\ldots\} \\ \imath',\jmath' \in \{1,3,5,\ldots\} \end{cases}$	
5.	$\boxed{\begin{matrix}\mathbb{AA}^{\langle \imath+\jmath \rangle} \\ \mathbb{AB}^{\langle \imath'+\jmath' \rangle}\end{matrix}}$ $[\textcircled{r}^4 \,/\, \textcircled{r}^3]$	$\xrightleftharpoons[\text{neutral (REC)}]{\text{neutral (REC)}}$	$\boxed{\begin{matrix}\mathbb{AA}^{\langle \imath+\imath' \rangle} \\ \mathbb{AB}^{\langle \jmath+\jmath' \rangle}\end{matrix}}$ $[\textcircled{r}^4 \,/\, \textcircled{r}^3]$	with $\begin{cases} \text{all paths distinct} \\ \imath,\imath' \in \{2,4,6,\ldots\} \\ \jmath \in \{0,2,4,\ldots\} \\ \jmath' \in \{1,3,5,\ldots\} \end{cases}$	
6.	$\boxed{\begin{matrix}\mathbb{BB}^{\langle \imath+\jmath \rangle} \\ \mathbb{AB}^{\langle \imath'+\jmath' \rangle}\end{matrix}}$ $[\textcircled{r}^4 \,/\, \textcircled{r}^3]$	$\xrightleftharpoons[\text{neutral (REC)}]{\text{neutral (REC)}}$	$\boxed{\begin{matrix}\mathbb{BB}^{\langle \imath+\jmath' \rangle} \\ \mathbb{AB}^{\langle \imath'+\jmath \rangle}\end{matrix}}$ $[\textcircled{r}^4 \,/\, \textcircled{r}^3]$	with $\begin{cases} \text{all paths distinct} \\ \imath,\jmath,\jmath' \in \{1,3,5,\ldots\} \\ \imath' \in \{0,2,4,\ldots\} \end{cases}$	
7.	$\boxed{\begin{matrix}\mathbb{AA}^{\langle \imath+\jmath' \rangle} \\ \mathbb{AA}^{\langle \jmath+\imath' \rangle}\end{matrix}}$ $[\textcircled{r}^4 \,/\, \textcircled{r}^3]$	$\xrightleftharpoons[\text{neutral (REC)}]{\text{neutral (REC)}}$	$\boxed{\begin{matrix}\mathbb{AA}^{\langle \imath+\imath' \rangle} \\ \mathbb{AA}^{\langle \jmath+\jmath' \rangle}\end{matrix}}$ $[\textcircled{r}^4 \,/\, \textcircled{r}^3]$	with $\begin{cases} \text{all } \mathbb{AA}\text{-paths distinct} \\ \imath \in \{0,2,4,\ldots\} \\ \jmath,\imath',\jmath' \in \{2,4,6,\ldots\} \end{cases}$	
8.	$\boxed{\begin{matrix}\mathbb{BB}^{\langle \imath+\jmath' \rangle} \\ \mathbb{BB}^{\langle \jmath+\imath' \rangle}\end{matrix}}$ $[\textcircled{r}^4]$	$\xrightleftharpoons[\text{neutral (REC)}]{\text{neutral (REC)}}$	$\boxed{\begin{matrix}\mathbb{BB}^{\langle \imath+\imath' \rangle} \\ \mathbb{BB}^{\langle \jmath+\jmath' \rangle}\end{matrix}}$ $[\textcircled{r}^4]$	with $\begin{cases} \text{all } \mathbb{BB}\text{-paths distinct} \\ \imath,\jmath,\imath',\jmath' \in \{1,3,5,\ldots\} \end{cases}$	
9.	$\boxed{\mathbb{AA}^{\langle \imath \rangle}}$ $[\textcircled{r}^{2\mid 2}]$	$\xrightleftharpoons[\text{losing (INT)}]{\text{gaining (INT)}}$	$\boxed{\mathbb{O}^{\langle \imath \rangle}}$ $[\textcircled{r}^{2\mid 1}]$	with $\imath \in \{2,4,6,\ldots\}$	$\Delta\lvert\mathcal{C}\rvert$ by ±1
10.	$\boxed{\mathbb{BB}^{\langle \imath+\jmath \rangle}}$ $[\textcircled{r}^{2\mid 1}]$	$\xrightleftharpoons[\text{losing (REC)}]{\text{gaining (REC)}}$	$\boxed{\begin{matrix}\mathbb{AB}^{\langle \imath \rangle} \\ \mathbb{AB}^{\langle \jmath \rangle}\end{matrix}}$ $[\textcircled{r}^{2\mid 2}]$	with $\imath,\jmath \in \{1,3,5,\ldots\}$	$\Delta\lvert\mathcal{P}_{\mathbb{AB}}\rvert$ by ±2
11.	$\boxed{\mathbb{AA}^{\langle \imath+\jmath \rangle}}$ $[\textcircled{r}^{2\mid 1}]$	$\xrightleftharpoons[\text{neutral (REC)}]{\text{neutral (INT)}}$	$\boxed{\begin{matrix}\mathbb{AA}^{\langle \imath \rangle} \\ \mathbb{AA}^{\langle \jmath \rangle}\end{matrix}}$ $[\textcircled{r}^{2\mid 2}]$	with $\imath,\jmath \in \{2,4,6,\ldots\}$	
12.	$\boxed{\mathbb{AB}^{\langle \imath+\jmath \rangle}}$ $[\textcircled{r}^{2\mid 1}]$	$\xrightleftharpoons[\text{neutral (REC)}]{\text{neutral (INT)}}$	$\boxed{\begin{matrix}\mathbb{AA}^{\langle \imath \rangle} \\ \mathbb{AB}^{\langle \jmath \rangle}\end{matrix}}$ $[\textcircled{r}^{2\mid 2}]$	with $\begin{cases} \imath \in \{2,4,6\ldots\} \\ \jmath \in \{1,3,5,\ldots\} \end{cases}$	

Bibliography

[1] Mark D. Adams, Susan E. Celniker, Robert A. Holt, Cheryl A. Evans, Jeannine D. Gocayne, Peter G. Amanatides, Steven E. Scherer, Peter W. Li, Roger A. Hoskins, Richard F. Galle, Reed A. George, Suzanna E. Lewis, Stephen Richards, Michael Ashburner, Scott N. Henderson, Granger G. Sutton, Jennifer R. Wortman, Mark D. Yandell, Qing Zhang, Lin X. Chen, Rhonda C. Brandon, Yu-Hui C. Rogers, Robert G. Blazej, Mark Champe, Barret D. Pfeiffer, Kenneth H. Wan, Clare Doyle, Evan G. Baxter, Gregg Helt, Catherine R. Nelson, George L. Gabor, Miklos, Josep F. Abril, Anna Agbayani, Hui-Jin An, Cynthia Andrews-Pfannkoch, Danita Baldwin, Richard M. Ballew, Anand Basu, James Baxendale, Leyla Bayraktaroglu, Ellen M. Beasley, Karen Y. Beeson, P. V. Benos, Benjamin P. Berman, Deepali Bhandari, Slava Bolshakov, Dana Borkova, Michael R. Botchan, John Bouck, Peter Brokstein, Phillipe Brottier, Kenneth C. Burtis, Dana A. Busam, Heather Butler, Edouard Cadieu, Angela Center, Ishwar Chandra, J. Michael Cherry, Simon Cawley, Carl Dahlke, Lionel B. Davenport, Peter Davies, Beatriz de Pablos, Arthur Delcher, Zuoming Deng, Anne Deslattes Mays, Ian Dew, Suzanne M. Dietz, Kristina Dodson, Lisa E. Doup, Michael Downes, Shannon Dugan-Rocha, Boris C. Dunkov, Patrick Dunn, Kenneth J. Durbin, Carlos C. Evangelista, Concepcion Ferraz, Steven Ferriera, Wolfgang Fleischmann, Carl Fosler, Andrei E. Gabrielian, Neha S. Garg, William M. Gelbart, Ken Glasser, Anna Glodek, Fangcheng Gong, J. Harley Gorrell, Zhiping Gu, Ping Guan, Michael Harris, Nomi L. Harris, Damon Harvey, Thomas J. Heiman, Judith R. Hernandez, Jarrett Houck, Damon Hostin, Kathryn A. Houston, Timothy J. Howland, Ming-Hui Wei, Chinyere Ibegwam, Mena Jalali, Francis Kalush, Gary H. Karpen, Zhaoxi Ke, James A. Kennison, Karen A. Ketchum, Bruce E. Kimmel, Chinnappa D. Kodira, Cheryl Kraft, Saul Kravitz, David Kulp, Zhongwu Lai, Paul Lasko, Yiding Lei, Alexander A. Levitsky, Jiayin Li, Zhenya Li, Yong Liang, Xiaoying Lin, Xiangjun Liu, Bettina Mattei, Tina C. McIntosh, Michael P. McLeod, Duncan McPherson, Gennady Merkulov, Natalia V. Milshina, Clark Mobarry, Joe Morris, Ali Moshrefi, Stephen M. Mount, Mee Moy, Brian Murphy, Lee Murphy, Donna M. Muzny, David L. Nelson, David R. Nelson, Keith A. Nelson, Katherine Nixon, Deborah R. Nusskern, Joanne M. Pacleb, Michael Palazzolo, Gjange S. Pittman, Sue Pan, John Pollard, Vinita Puri, Martin G. Reese, Knut Reinert, Karin Remington, Robert D. C. Saunders, Frederick Scheeler, Hua Shen, Bixiang Christopher Shue, Inga Sidén-Kiamos, Michael Simpson,

Marian P. Skupski, Tom Smith, Eugene Spier, Allan C. Spradling, Mark Stapleton, Renee Strong, Eric Sun, Robert Svirskas, Cyndee Tector, Russell Turner, Eli Venter, Aihui H. Wang, Xin Wang, Zhen-Yuan Wang, David A. Wassarman, George M. Weinstock, Jean Weissenbach, Sherita M. Williams, Trevor Woodage, Kim C. Worley, David Wu, Song Yang, Q. Alison Yao, Jane Ye, Ru-Fang Yeh, Jayshree S. Zaveri, Ming Zhan, Guangren Zhang, Qi Zhao, Liansheng Zheng, Xiangqun H. Zheng, Fei N. Zhong, Wenyan Zhong, Xiaojun Zhou, Shiaoping Zhu, Xiaohong Zhu, Hamilton O. Smith, Richard A. Gibbs, Eugene W. Myers, Gerald M. Rubin, and J. Craig Venter. The genome sequence of *Drosophila melanogaster*. *Science*, 287:2185–2195, 2000.

[2] Max Alekseyev and Pavel A. Pevzner. Colored de Bruijn graphs and the genome halving problem. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 4(1):98–107, 2008.

[3] Adrian M. Altenhoff and Christophe Dessimoz. Phylogenetic and functional assessment of orthologs inference projects and methods. *PLOS Computational Biology*, 5(1):1–11, 01 2009.

[4] Adrian M. Altenhoff, Jeremy Levy, Magdalena Zarowiecki, Bartłomiej Tomiczek, Alex Warwick Vesztrocy, Daniel A. Dalquen, Steven Müller, Maximilian J. Telford, Natasha M. Glover, David Dylus, and Christophe Dessimoz. OMA standalone: orthology inference among public and custom genomes and transcriptomes. *Genome Research*, 29:1152–1163, 2019.

[5] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–410, 1990.

[6] Sébastien Angibaud, Guillaume Fertin, Irena Rusu, Annelyse Thévenin, and Stéphane Vialette. On the approximability of comparing genomes with duplicates. *J Graph Algo App*, 13(1):19–53, 2009.

[7] Vineet Bafna and Pavel A. Pevzner. Genome rearrangements and sorting by reversals. In *Proceedings of FOCS 1993*, pages 148–157, 1993.

[8] Anne Bergeron. A very elementary presentation of the hannenhalli-pevzner theory. In *Proc. of CPM*, volume 2089 of *LNCS*, pages 106–117, 2001.

[9] Anne Bergeron, Steffen Heber, and Jens Stoye. Common intervals and sorting by reversals: a marriage of necessity. *Bioinformatics*, 18(Suppl. 2):S54–G63, 2002.

[10] Anne Bergeron, Julia Mixtacki, and Jens Stoye. A unifying view of genome rearrangements. In *Proc. of WABI*, volume 4175 of *Lecture Notes in Bioinformatics*, pages 163–173, 2006.

[11] Anne Bergeron, Julia Mixtacki, and Jens Stoye. A new linear time algorithm to compute the genomic distance via the double cut and join distance. *Theoretical Computer Science*, 410(51):5300–5316, 2009.

[12] Matthias Bernt, Daniel Merkle, and Martin Middendorf. Genome rearrangement based on reversals that preserve conserved intervals. *IEEE/ACM Trans Comput Biol Bioinform*, 3(3):275–288, 2006.

[13] Priscila Biller, Laurent Guéguen, Carole Knibbe, and Eric Tannier. Breaking Good: Accounting for Fragility of Genomic Regions in Rearrangement Distance Estimation. *Genome Biology and Evolution*, 5(8):1427—-1439, 2016.

[14] Leonard Bohnenkämper. The Floor Is Lava: Halving Natural Genomes with Viaducts, Piers, and Pontoons. *J Comput Biol*, 31(4):294–311, 2024. A preliminary version appeared in the Proc. of Recomb-CG 2023.

[15] Leonard Bohnenkämper, Marília D. V. Braga, Daniel Doerr, and Jens Stoye. Computing the rearrangement distance of natural genomes. *J Comput Biol*, 28(4):410–431, 2021. A preliminary version appeared in Proc. of RECOMB 2020, LNCS 12074, 3–18.

[16] Leonard Bohnenkämper. Recombinations, chains and caps: resolving problems with the DCJ-indel model. *Algorithms Mol Biol*, 19(8), 2024. A preliminary version appeared in the Proc. of WABI 2023.

[17] Jeffrey L. Boore. The duplication/random loss model for gene rearrangement exemplified by mitochondrial genomes of deuterostome animals. In David Sankoff and Joseph H. Nadeau, editors, *Comparative Genomics*, pages 133–148. Springer, 2000.

[18] Marília D. V. Braga. An overview of genomic distances modeled with indels. In *Proc. of CiE*, volume 7921 of *LNCS*, pages 22–31. Springer, 2013.

[19] Marília D. V. Braga, Leonie R. Brockmann, Katharina Klerx, and Jens Stoye. Investigating the complexity of the double distance problems. *Algorithms Mol Biol*, 19(1), 2024. Preliminary versions appeared in the Proc. of WABI 2022 and RECOMB-CG 2023.

[20] Marília D. V. Braga, Cedric Chauve, Daniel Doerr, Katharina Jahn, Jens Stoye, Annelyse Thévenin, and Roland Wittler. The potential of family-free genome comparison. In C. Chauve, N. El-Mabrouk, and E. Tannier, editors, *Models and Algorithms for Genome Evolution*, volume 19 of *Computational Biology Series*, chapter 13, pages 287–307. Springer, Berlin, 2013.

[21] Marília D. V. Braga and Jens Stoye. The solution space of sorting by DCJ. *J Comput Biol*, 17(9):1145–1165, 2010. A preliminary version appeared in the Proc. of RECOMB-CG 2009.

[22] Marília D. V. Braga, Eyla Willing, and Jens Stoye. Double cut and join with insertions and deletions. *J Comput Biol*, 18(9):1167–1184, 2011. A preliminary version appeared in the Proc. of WABI 2010.

[23] Marília D. V. Braga, Daniel Doerr, Diego P. Rubert, and Jens Stoye. Family-free genome comparison. In João Carlos Setubal, Peter F. Stadler, and Jens Stoye, editors, *Comparative Genomics: Methods and Protocols*, volume 2802 of *Methods in Molecular Biology*, pages 57–72. Springer, New York, 2024. Second edition; for the first edition see [42, from 2018].

[24] Marília D. V. Braga, Raphael Machado, Leonardo C. Ribeiro, and Jens Stoye. Genomic distance under gene substitutions. *BMC Bioinform*, 12(Suppl 9):S8, 2011.

[25] Marília D. V. Braga, Raphael Machado, Leonardo C. Ribeiro, and Jens Stoye. On the weight of indels in genomic distances. *BMC Bioinformatics*, 12(Suppl. 9):S13, 2011.

[26] Marília D. V. Braga and Jens Stoye. Sorting linear genomes with rearrangements and indels. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 12(3):500–506, 2015.

[27] David Bryant. The complexity of calculating exemplar distances. In David Sankoff and Joseph H. Nadeau, editors, *Comparative Genomics*, volume 1 of *Computational Biology Series*, pages 207–211. Kluver Academic Publishers, London, 2000.

[28] Benjamin Buchfink, Chao Xie, and Daniel H. Huson. Fast and sensitive protein alignment using DIAMOND. *Nat Methods*, 12:59–60, 2015.

[29] Sèverine Bérard, Annie Chateau, Cedric Chauve, Christophe Paul, and Eric Tannier. Perfect DCJ rearrangement. In *Proceedings of RECOMB-CG*, volume 5267 of *LNBI*, page 158–169, 2008.

[30] Cedric Chauve. Personal communication in Dagstuhl Seminar no. 18451 - Genomics, Pattern Avoidance, and Statistical Mechanics, November 2018.

[31] Xin Chen. On sorting unsigned permutations by double-cut-and-joins. *J Comb Optim*, 25(1):339–351, 2013.

[32] Xin Chen, Jie Zheng, Zheng Fu, Peng Nan, Yang Zhong, S. Lonardi, and Tao Jiang. Assignment of orthologous genes via genome rearrangement. *IEEE/ACM Trans Comput Biol Bioinform*, 2(4):302–315, 2005.

[33] David A. Christie. Sorting permutations by block-interchanges. *Inf Process Lett*, 60(4):165–169, 1996.

[34] M. Chrobak, T. Szymacha, and A. Krawczyk. A data structure useful for finding hamiltonian cycles. *Theor Comput Sci*, 71(3):419–424, 1990.

[35] Andrew G. Clark, Michael B. Eisen, Douglas R. Smith, Casey M. Bergman, Brian Oliver, Therese A. Markow, Thomas C. Kaufman, Manolis Kellis, and Drosophila12GenomesConsortium. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature*, 450:203–218, 2007.

[36] Philip E. C. Compeau. DCJ-indel sorting revisited. *Algorithms Mol Biol*, 8(1), 2013. A preliminary version appeared in the Proc. of WABI 2012.

[37] Poly H. da Silva, Raphael Machado, Simone Dantas, and Marília D. V. Braga. Restricted DCJ-indel model: sorting linear genomes with DCJ and indels. *BMC Bioinformatics*, 13(Suppl. 19):S14.

[38] Poly H. da Silva, Raphael Machado, Simone Dantas, and Marília D. V. Braga. DCJ-indel and DCJ-substitution distances with distinct operation costs. *Algorithms Mol Biol*, 8(21), 2013. A preliminary version appeared in Proc. of WABI 2012.

[39] Poly H. da Silva, Raphael Machado, Simone Dantas, and Marília D.V. Braga. Genomic distance with high indel costs. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 14(3):728–732, 2017.

[40] C. Dessimoz, G. Cannarozzi, M. Gil, D. Margadant, A. C. J. Roth, A. Schneider, and G. H. Gonnet. OMA, a comprehensive, automated project for the identification of orthologs from complete genome data: introduction and first achievements. In *Proc. of RECOMB-CG*, volume 3678 of *Lecture Notes in Bioinformatics*, pages 61–72, 2005.

[41] Daniel Doerr and Cedric Chauve. Small parsimony for natural genomes in the DCJ-indel model. *J Bioinform Comput Biol*, 19(6):2140009, 2021.

[42] Daniel Doerr, Pedro Feijão, and Jens Stoye. Family-free genome comparison. In João C. Setubal, Jens Stoye, and Peter F. Stadler, editors, *Comparative Genomics: Methods and Protocols*, volume 1704 of *Methods in Molecular Biology*, pages 331–342. Springer Nature, New York, 2018. First edition; for the second edition see [23, from 2024].

[43] Daniel Doerr, Annelyse Thévenin, and Jens Stoye. Gene family assignment-free comparative genomics. *BMC Bioinform*, 13(Suppl 19):S3, 2012.

[44] Nadia El-Mabrouk. Sorting signed permutations by reversals and insertions/deletions of contiguous segments. *Journal of Discrete Algorithms*, 1(1):105–122, 2001.

[45] Nadia El-Mabrouk and David Sankoff. The reconstruction of doubled genomes. *SIAM Journal on Computing*, 32(3):754–792, 2003.

[46] Péter L. Erdős, Lajos Soukup, and Jens Stoye. Balanced vertices in trees and a simpler algorithm to compute the genomic distance. *Applied Mathematics Letters*, 24(1):82–86, 2011.

[47] Jianxing Feng and Daming Zhu. Faster algorithms for sorting by transpositions and sorting by block interchanges. *ACM Trans Algorithms*, 3(3), 2007.

[48] Paul Feyerabend. *Against Method.* New Left Books, 1975.

[49] Paul Feyerabend. *Farewell to Reason.* New Left Books, 1987.

[50] Walter M. Fitch. Distinguishing homologous from analogous proteins. *Syst Zool*, 19:99–113, 1970.

[51] Shmuel Friedland. An upper bound for the number of perfect matchings in graphs, 2008.

[52] Marija Gimbutas. *The Goddesses and Gods of Old Europe - Myths and Cult Images 6500 - 3500 BC.* University of California Press, 1982.

[53] David Graeber and David Wengrow. *The Dawn of Everything, a new history of humanity.* Allen Lane Imprint, 2021.

[54] P. Hall. On representatives of subsets. *Journal of the London Mathematical Society*, s1-10(1):26–30, 1935.

[55] Sridhar Hannenhalli and Pavel A. Pevzner. Transforming men into mice (polynomial algorithm for genomic distance problem). In *Proc. of FOCS*, pages 581–592, 1995.

[56] Sridhar Hannenhalli and Pavel A. Pevzner. Transforming cabbage into turnip: polynomial algorithm for sorting signed permutations by reversals. *Journal of the ACM*, 46(1):1–27, 1999. A preliminary version appeared in the Proc. of ACM Symposium on the Theory of Computing 1995.

[57] Haim Kaplan and Elad Verbin. Sorting signed permutations by reversals, revisited. *J Comput Syst Sci*, 70(3):321–341, 2005.

[58] Eugene V. Koonin. Orthologs, paralogs, and evolutionary genomics. *Annual Review of Genetics*, 39(1):309–338, 2005.

[59] Davi Kopenawa and Bruce Albert. *The Falling Sky: Words of a Yanomami Shaman.* Harvard University Press, 2013. Originally published in French in 2010.

[60] Jakub Kováč, Robert Warren, Marília D. V. Braga, and Jens Stoye. Restricted dcj model: Rearrangement problems with chromosome reincorporation. *J. Comput. Biol.*, 18:1231–1241, 2011. A preliminary version appeared in Proc. of RECOMB-CG 2010.

[61] Ailton Krenak. *Ideas to Postpone the End of the World.* House Of Anansi Press Ltd, 2019.

[62] Sudhir Kumar, Glen Stecher, Michael Li, Christina Knyaz, and Koichiro Tamura. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol Biol Evol*, 35(6):1547–1549, 2018.

[63] Sudhir Kumar, Glen Stecher, Michael Suleski, and S. Blair Hedges. Timetree: a resource for timelines, timetrees, and divergence times. *Mol Biol Evol*, 34(7):1812–1819, 2017.

[64] Marcus Lechner, Sven Findeiß, Lydia Steiner, Manja Marz, Peter F. Stadler, and Sonja J. Prohaska. Proteinortho: Detection of (co-)orthologs in large-scale analysis. *BMC Bioinform*, 12(124), 2011.

[65] Marcus Lechner, Maribel Hernandez-Rosales, Daniel Doerr, Nicolas Wieseke, Annelyse Thévenin, Jens Stoye, Roland K. Hartmann, Sonja J. Prohaska, and Peter F. Stadler. Orthology detection combining clustering and synteny for very large datasets. *PLoS One*, 9(8:e105015), 2014.

[66] Fábio V. Martinez, Pedro Feijão, Marília D. V. Braga, and Jens Stoye. On the family-free DCJ distance and similarity. *Algorithms Mol Biol*, 13(10), 2015. A preliminary version appeared in the Proc. of WABI 2014.

[67] Humberto R. Maturana and Francisco J. Varela. *The Tree of Knowledge: The Biological Roots of Human Understanding.* Shambhala Publications, 1998.

[68] Compiled by T. C. McLuhan. *Touch the Earth: a self-portrait of Indian existence.* Abacus, 1989.

[69] J. Meidanis, M. E. M. T. Walter, and Z. Dias. Reversal distance of signed circular chromosomes. Relatório Técnico IC-00-23, Institute of Computing, University of Campinas, Brazil, 2000.

[70] Carolyn Merchant. *The Death of Nature: Women, Ecology and the Scientific Revolution.* New York: Harper and Row, 1980.

[71] Julia Mixtacki. Genome halving under DCJ revisited. In *Proceedings of COCOON 2008*, volume 5092 of *LNCS*, pages 276–286, 2008.

[72] C. Moritz, T. E. Dowling, and W. M. Brown. Evolution of animal mitochondrial dna: relevance for population biology and systematics. In *Annu. Rev. Ecol. Syst*, volume 18, pages 269–292, 1987.

[73] Mihai Nadin. *Disrupt Science: The Future Matters.* Springer, 2023.

[74] Aïda Ouangraoua and Anne Bergeron. Combinatorial structure of genome rearrangements scenarios. *J Comput Biol*, 17(9):1129–1144, 2010. A preliminary version appeared in the Proc. of RECOMB-CG 2009.

[75] Catia Pesquita, Daniel Faria, Hugo Bastos, António EN Ferreira, André O Falcão, and Francisco M Couto. Metrics for GO based protein semantic similarity: a systematic evaluation. *BMC Bioinformatics*, 9(Suppl 5):S4, 2008.

[76] Stephen Richards, Yue Liu, Brian R. Bettencourt, Pavel Hradecky, Stan Letovsky, Rasmus Nielsen, Kevin Thornton, Melissa J. Hubisz, Rui Chen, Richard P. Meisel, Olivier Couronne, Sujun Hua, Mark A. Smith, Peili Zhang, Jing Liu, Harmen J. Bussemaker, Marinus F. van Batenburg, Sally L. Howells, Steven E. Scherer, Erica Sodergren, Beverly B. Matthews, Madeline A. Crosby, Andrew J. Schroeder, Daniel Ortiz-Barrientos, Catharine M. Rives, Michael L. Metzker, Donna M. Muzny, Graham Scott, David Steffen, David A. Wheeler, Kim C. Worley, Paul Havlak, K. James Durbin, Amy Egan, Rachel Gill, Jennifer Hume, Margaret B. Morgan, George Miner, Cerissa Hamilton, Yanmei Huang, Lenée Waldron, Daniel Verduzco, Kerstin P. Clerc-Blankenburg, Inna Dubchak, Mohamed A.F. Noor, Wyatt Anderson, Kevin P. White, Andrew G. Clark, Stephen W. Schaeffer, William Gelbart, George M. Weinstock, and Richard A. Gibbs. Comparative genome sequencing of *Drosophila pseudoobscura*: Chromosomal, gene, and cis-element evolution. *Genome Res*, 15:1–18, 2005.

[77] Robert Rosen. *Life Itself: A Comprehensive Inquiry into the Nature, Origin, and Fabrication of Life.* Columbia University Press,, 1991.

[78] Robert Rosen. *Essays on Life Itself.* Columbia University Press, 2000.

[79] Alexander C. J. Roth, Gaston H. Gonnet, and Christophe Dessimoz. Algorithm of OMA for large-scale orthology inference. *BMC Bioinform*, 9(518), 2008.

[80] Diego P. Rubert and Marília D. V. Braga. Efficient gene orthology inference via large-scale rearrangements. *Algorithms Mol Biol*, 18(14), 2023. A preliminary version appeared in the Proc. of WABI 2022.

[81] Diego P. Rubert, Daniel Doerr, and Marília D. V. Braga. The potential of family-free rearrangements towards gene orthology inference. *J Bioinform Comput Biol*, 19(6):2140014, 2021.

[82] Diego P. Rubert, Fábio V. Martinez, and Marília D. V. Braga. Natural Family-Free Genomic Distance. *Algorithms Mol Biol*, 16(4), 2021. A preliminary version appeared in the Proc. of WABI 2020.

[83] Marie-France Sagot and Eric Tannier. Perfect sorting by reversals. In *Proc. of COCOON*, volume 3595 of *LNCS*, pages 42–51, 2005.

[84] Naruya Saitou and Masatoshi Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, 4(4):406–425, 1987.

[85] David Sankoff. Edit distance for genome comparison based on non-local operations. In *Proc. of CPM*, volume 644 of *Lecture Notes in Computer Science*, pages 121–135, 1992.

[86] David Sankoff. Genome rearrangement with gene families. *Bioinformatics*, 15(11):909–917, 1999.

[87] João C. Setubal and Peter F. Stadler. Gene phylogenies and orthologous groups. In João C. Setubal, Jens Stoye, and Peter F. Stadler, editors, *Comparative Genomics: Methods and Protocols*, pages 1–28. Springer, New York, 2018.

[88] Mingfu Shao, Yu Lin, and Bernard Moret. An exact algorithm to compute the double-cut-and-join distance for genomes with duplicate genes. *J Comput Biol*, 22(5):425–435, 2015.

[89] Guanqun Shi, Meng-Chih Peng, and Tao Jiang. MultiMSOAR 2.0: an accurate tool to identify ortholog groups among multiple genomes. *PLoS One*, 6(6:e20892), 2011.

[90] Guanqun Shi, Liqing Zhang, and Tao Jiang. MSOAR 2.0: Incorporating tandem duplications into ortholog assignment based on genome rearrangement. *BMC Bioinform*, 11(10), 2010.

[91] Alfred H. Sturtevant. A case of rearrangement of genes in drosophila. In *Proc. of Natl Acad Sci USA*, volume 7, pages 235–237, 1921.

[92] Krister M. Swenson, Vaibhav Rajan, Yu Lin, and Bernard M.E. Moret. Sorting signed permutations by inversions in $O(n \log n)$ time. *J Comput Biol*, 17(3):489–501, 2010.

[93] Eric Tannier, Chunfang Zheng, and David Sankoff. Multichromosomal median and halving problems under different genomic distances. *BMC Bioinformatics*, 10:120, 2009.

[94] Tamir Tassa. Finding all maximally-matchable edges in a bipartite graph. *Theoretical Computer Science*, 423:50–58, 2012.

[95] Roman L. Tatusov, Eugene V. Koonin, and David J. Lipman. A genomic perspective on protein families. *Science*, 278(5338):631–637, 1997.

[96] Fredj Tekaia. Inferring orthologs: Open questions and perspectives. *Genomics Insights*, 9:GEI.S37925, 2016.

[97] Eyla Willing, Jens Stoye, and Marília D. V. Braga. Computing the inversion-indel distance. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 18(6):2314–2326, 2021.

[98] Eyla Willing, Simone Zaccaria, Marília D. V. Braga, and Jens Stoye. On the Inversion-Indel Distance. *BMC Bioinformatics*, 14(Suppl. 15):S3, 2013.

[99] Sophia Yancopoulos, Oliver Attie, and Richard Friedberg. Efficient sorting of genomic permutations by translocation, inversion and block interchange. *Bioinformatics*, 21(16):3340–3346, 2005.

[100] Sophia Yancopoulos and Richard Friedberg. DCJ path formulation for genome transformations which include insertions, deletions, and duplications. *J Comput Biol*, 16(10):1311–1338, 2009.

[101] Zhaoming Yin, Jijun Tang, Stephen W. Schaeffer, and David A. Bader. Exemplar or matching: modeling DCJ problems with unequal content genome data. *Journal of Combinatorial Optimization*, 32(4):1165–1181, 2016.

[102] Doron Zeilberger. Yet another proof of cayley's formula for the number of labelled trees: `http://www.math.rutgers.edu/~zeilberg/mamarim/mamarimPDF/cayley.pdf`, included in Appendix B.

[103] Qi Zhou and Doris Bachtrog. Ancestral chromatin configuration constrains chromatin evolution on differentiating sex chromosomes in *Drosophila. PLoS Genet*, 11(6), 2015.

APPENDIX A

Symbols used in the adopted formalism

$\mathbb{A}, \mathbb{B}, \mathbb{C}$	genomes
$\eta(.)$	family-annotation of marker, marker extremity, adjacency, telomere or chromosome
$\mathsf{adj}(\mathbb{A})$	set of unannotated adjacencies in genome $\mathbb{A}$
$\mathsf{ADJ}(\mathbb{A})$	$= \eta(\mathsf{adj}(\mathbb{A}))$: set of annotated adjacencies in genome $\mathbb{A}$
$\mathsf{ext}(\mathbb{A})$	set of unannotated marker extremities in genome $\mathbb{A}$
$\mathsf{EXT}(\mathbb{A})$	$= \eta(\mathsf{ext}(\mathbb{A}))$: set of annotated marker extremities in genome $\mathbb{A}$
$\mathsf{tel}(\mathbb{A})$	set of unannotated telomeres in genome $\mathbb{A}$
$\mathsf{TEL}(\mathbb{A})$	$= \eta(\mathsf{tel}(\mathbb{A}))$: set of annotated telomeres in genome $\mathbb{A}$
$\psi(\mathbb{A})$	set of cutpoints in genome $\mathbb{A}$
$\mathfrak{C}(\mathbb{A})$	set of unannotated chromosomes in genome $\mathbb{A}$
$\mathcal{F}(\mathbb{A})$	set of families occurring in annotated genome $\mathbb{A}$
$\mathcal{G}(\mathbb{A})$	$= \eta(\mathcal{M}(\mathbb{A}))$: set of annotated markers of genome $\mathbb{A}$
$\mathcal{M}(\mathbb{A})$	set of unannotated markers of genome $\mathbb{A}$
$\Phi(\mathtt{F}, \mathbb{A})$	number of occurrences (markers) of family $\mathtt{F}$ in genome $\mathbb{A}$

In comparison of two annotated genomes $\mathbb{A}$ and $\mathbb{B}$:

$\mathcal{F}_*$	set of common families of genomes $\mathbb{A}$ and $\mathbb{B}$
$\mathcal{A}$	set of families exclusive to genome $\mathbb{A}$ (see also table below)
$\mathcal{B}$	set of families exclusive to genome $\mathbb{B}$ (see also table below)
$\mathcal{G}_*$	(multi)set of common annotated markers of genomes $\mathbb{A}$ and $\mathbb{B}$

Graphs and respective derived properties:

G_{R}	relational graph of canonical or singular genomes
G_{MR}	multi-relational graph of balanced or natural genomes
$\mathsf{G}^{\mathsf{w}}_{\mathrm{FFR}}$	family-free multi-relational graph of unannotated genomes
$\mathrm{J}(.)$	set of junctions in the whole relational graph or in one of its connected components
$\mathsf{G}_{\mathcal{F}_*}$	family graph of natural genomes
$\mathsf{G}^{\mathsf{w}}_{\sigma}$	similarity graph
$\mathsf{G}^{\mathsf{w}}_{\mathfrak{C}}$	shared-content graph
$\widehat{\mathsf{G}}^{\mathsf{w}}_{\mathfrak{C}}$	perfect shared-content graph

Γ	component (path or cycle) of a relational graph
$\Gamma^{\langle i \rangle}$	component whose length (given by the number of extremity edges) is $i \geq 0$
$\mathbb{O}$	cycle of a relational graph
$\mathbb{O}^{\langle i \rangle}$	cycle of (even) length $i \geq 0$
$\mathbb{AA}$	path of a relational graph starting and ending in a telomere of genome $\mathbb{A}$
$\mathbb{AA}^{\langle i \rangle}$	$\mathbb{AA}$-path of (even) length $i \geq 0$
$\mathbb{AB}$	path of a relational graph starting in a telomere of $\mathbb{A}$ and ending in a telomere of $\mathbb{B}$
$\mathbb{AB}^{\langle i \rangle}$	$\mathbb{AB}$-path of (odd) length $i \geq 1$
$\mathbb{BB}$	path of a relational graph starting and ending in a telomere of genome $\mathbb{B}$
$\mathbb{BB}^{\langle i \rangle}$	$\mathbb{BB}$-path of (even) length $i \geq 0$

$\Upsilon(\Gamma)$	type of component Γ
$\mathcal{S}$	$= \{\Gamma \mid \Gamma \in \mathsf{G}_{\mathrm{R}}(\mathbb{A}, \mathbb{B}) \text{ and } \Upsilon(\Gamma) = \mathbb{O}^{\langle 0 \rangle}\}$: set of (circular) singletons
$\mathcal{C}$	$= \{\Gamma \mid \Gamma \in \mathsf{G}_{\mathrm{R}}(\mathbb{A}, \mathbb{B}) \text{ and } \Upsilon(\Gamma) = \mathbb{O}^{\langle 2,4,\ldots \rangle}\}$: set of cycles of length at least 2
$\mathcal{P}_{\mathbb{AA}}$	$= \{\Gamma \mid \Gamma \in \mathsf{G}_{\mathrm{R}}(\mathbb{A}, \mathbb{B}) \text{ and } \Upsilon(\Gamma) = \mathbb{AA}\}$: set of $\mathbb{AA}$-paths
$\mathcal{P}_{\mathbb{AB}}$	$= \{\Gamma \mid \Gamma \in \mathsf{G}_{\mathrm{R}}(\mathbb{A}, \mathbb{B}) \text{ and } \Upsilon(\Gamma) = \mathbb{AB}\}$: set of $\mathbb{AB}$-paths
$\mathcal{P}_{\mathbb{BB}}$	$= \{\Gamma \mid \Gamma \in \mathsf{G}_{\mathrm{R}}(\mathbb{A}, \mathbb{B}) \text{ and } \Upsilon(\Gamma) = \mathbb{BB}\}$: set of $\mathbb{BB}$-paths

$\mathcal{A}$	as a path-subscript: odd sequence of runs starting and ending in an $\mathcal{A}$-run
$\mathcal{B}$	as a path-subscript: odd sequence of runs starting and ending in an $\mathcal{B}$-run
$\mathcal{AB}$	path-subscript: even sequence of runs starting in an $\mathcal{A}$-run and ending in a $\mathcal{B}$-run
$\Lambda(\Gamma)$	number of runs in component Γ
$\aleph(\Gamma)$	number of transitions in component Γ
$\lambda(\Gamma)$	indel-potential of component Γ
$\sigma(\Gamma)$	substitution-potential of component Γ

$\mathcal{O}$	ortholog-set
$\widetilde{\mathcal{O}}$	complement of ortholog-set $\mathcal{O}$
$\mathcal{H}$	ortholog-maxset
$\widetilde{\mathcal{H}}$	complement of ortholog-maxset $\mathcal{H}$
$\mathcal{L}$	sibling-set
$\widetilde{\mathcal{L}}$	complement of sibling-set $\mathcal{L}$
$\mathcal{J}$	sibling-maxset
$\widetilde{\mathcal{J}}$	complement of sibling-maxset $\mathcal{J}$
$\mathcal{Q}$	capping-maxset