# Topics of today:

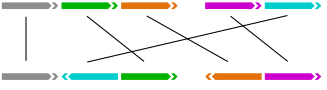
1. Recall concepts from lecture 01
2. Single-cut-or-join model, distance and double-distance
3. Formalizing the number of occurrences ($\phi$) of families/adjacencies/telomeres
4. Other problems: median and halving

# Types of genome pairs/sets

**Pair/set of singular genomes:**
each family occurs **at most once** in each genome

**Pair/set of balanced genomes:**
each family occurs **the same number of times** in each genome

**Pair/set of canonical genomes:**
**singular** and **balanced**



**Singular/duplicated canonical pair:**
one genome is singular, the other is duplicated and the gene families of both genomes are the same

(whole genome duplication)

# Definitions / notation (family-based setting)

Given genomes $\mathbb{G}_1^f, \mathbb{G}_2^f, \ldots, \mathbb{G}_k^f$:

- Set of **common families** (occurring in each $\mathbb{G}_i^f$):

$$\mathcal{F}_\star = \mathcal{F}(\mathbb{G}_1^f) \cap \mathcal{F}(\mathbb{G}_2^f) \cap \ldots \cap \mathcal{F}(\mathbb{G}_k^f)$$

- (Multi)set of annotated **common markers**:

$$\mathcal{G}_\star = \mathcal{G}(\mathbb{G}_1^f) \cap \mathcal{G}(\mathbb{G}_2^f) \cap \ldots \cap \mathcal{G}(\mathbb{G}_k^f)$$

$$|\mathcal{G}_\star| = n$$

| **Type** | | |
|---|---|---|
| singular: | $\mathcal{F}_\star = \mathcal{G}_\star$ | **(a)** |
| balanced: | $\mathcal{F}_\star = \mathcal{F}(\mathbb{G}_1^f) = \mathcal{F}(\mathbb{G}_2^f) = \ldots = \mathcal{F}(\mathbb{G}_k^f)$ **and** $\mathcal{G}_\star = \mathcal{G}(\mathbb{G}_1^f) = \mathcal{G}(\mathbb{G}_2^f) = \ldots = \mathcal{G}(\mathbb{G}_k^f)$ | **(b)** |
| canonical: | both **(a)** and **(b)** | |

# Breakpoint distance

Given genomes $\mathbb{A}^f$ and $\mathbb{B}^f$, let:

- $\Gamma_\star = \Gamma(\mathbb{A}^f) \cap \Gamma(\mathbb{B}^f)$ be the set of **common adjacencies**

  $|\Gamma_\star| = a$

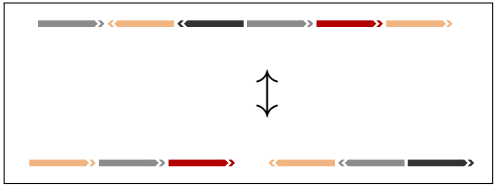
- $\Theta_\star = \Theta(\mathbb{A}^f) \cap \Theta(\mathbb{B}^f)$ be the set of **common telomeres**

  $|\Theta_\star| = t$

The breakpoint distance of **canonical genomes** $\mathbb{A}^f_\triangleright$ and $\mathbb{B}^f_\triangleright$ is defined to be:

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

# Breakpoint distance of balanced genomes

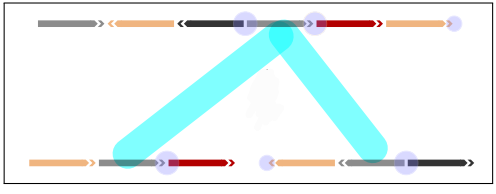


The breakpoint distance of **balanced genomes** $\mathbb{A}^f$ and $\mathbb{B}^f$ is:

$$\mathrm{d}_{\mathrm{BP}}(\mathbb{A}^f, \mathbb{B}^f) = \min_{f_m} \mathrm{d}_{\mathrm{BP}}(\mathbb{A}^{f_m}_{\triangleright}, \mathbb{B}^{f_m}_{\triangleright})$$

where $f_m$ is any function that produces a maximal matching of $f$-families

# Breakpoint distance of balanced genomes



The breakpoint distance of **balanced genomes** $\mathbb{A}^f$ and $\mathbb{B}^f$ is:

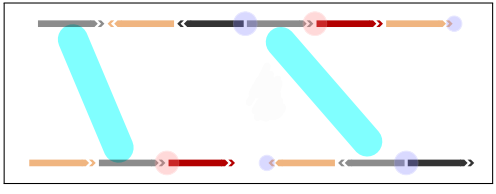
$$\mathrm{d}_{\mathrm{BP}}(\mathbb{A}^f, \mathbb{B}^f) = \min_{f_m} \mathrm{d}_{\mathrm{BP}}(\mathbb{A}^{f_m}_{\triangleright}, \mathbb{B}^{f_m}_{\triangleright})$$

where $f_m$ is any function that produces a maximal matching of $f$-families

Greedy approach: take all common adjacencies/telomeres: $|\mathcal{G}_\star| - a - \frac{t}{2} = 6 - 2 - \frac{1}{2} = 3.5$

↘ may lead to inconsistencies

# Breakpoint distance of balanced genomes



The breakpoint distance of **balanced genomes** $\mathbb{A}^f$ and $\mathbb{B}^f$ is:

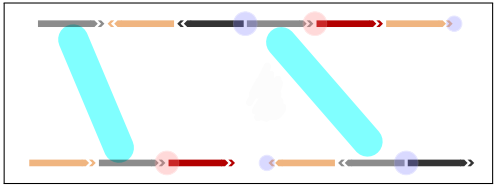
$$\mathrm{d}_{\mathrm{BP}}(\mathbb{A}^f, \mathbb{B}^f) = \min_{f_m} \mathrm{d}_{\mathrm{BP}}(\mathbb{A}^{f_m}_{\triangleright}, \mathbb{B}^{f_m}_{\triangleright})$$

where $f_m$ is any function that produces a maximal matching of $f$-families

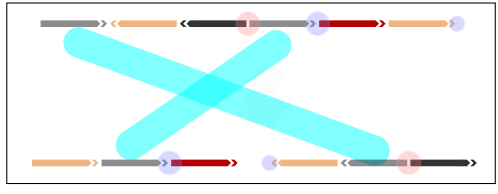
Greedy approach: take all common adjacencies/telomeres: $|\mathcal{G}_\star| - a - \frac{t}{2} = 6 - 2 - \frac{1}{2} = 3.5$

↘ may lead to inconsistencies

# Breakpoint distance of balanced genomes



or



The breakpoint distance of **balanced genomes** $\mathbb{A}^f$ and $\mathbb{B}^f$ is:

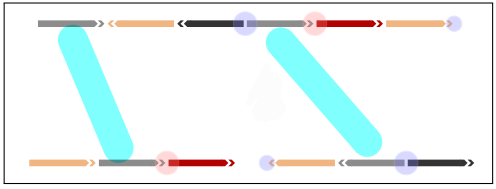
$$d_{BP}(\mathbb{A}^f, \mathbb{B}^f) = \min_{f_m} d_{BP}(\mathbb{A}^{f_m}_{\triangleright}, \mathbb{B}^{f_m}_{\triangleright})$$

where $f_m$ is any function that produces a maximal matching of $f$-families

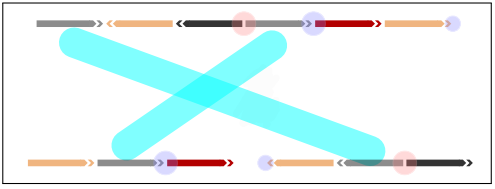
Greedy approach: take all common adjacencies/telomeres: $|\mathcal{G}_\star| - a - \frac{t}{2} = 6 - 2 - \frac{1}{2} = 3.5$

↘ may lead to inconsistencies

# Breakpoint distance of balanced genomes



or



The breakpoint distance of **balanced genomes** $\mathbb{A}^f$ and $\mathbb{B}^f$ is:

$$d_{BP}(\mathbb{A}^f, \mathbb{B}^f) = \min_{f_m} d_{BP}(\mathbb{A}^{f_m}_{\triangleright}, \mathbb{B}^{f_m}_{\triangleright})$$

where $f_m$ is any function that produces a maximal matching of $f$-families

Greedy approach: take all common adjacencies/telomeres: $|\mathcal{G}_\star| - a - \frac{t}{2} = 6 - 2 - \frac{1}{2} = 3.5$

↘ may lead to inconsistencies

Correct distance: $d_{BP}(\mathbb{A}^f, \mathbb{B}^f) = 6 - 1 - \frac{1}{2} = 4.5$

The breakpoint distance of balanced genomes is NP-hard

[Blin, Chauve and Fertin, 2004: The breakpoint distance for signed sequences]

# Breakpoint double distance

For a given singular genome $\mathbb{S}^f_\triangleright$, let $2{\cdot}\mathbb{S}^f_\triangleright$ be the set of doubled genomes derived from $\mathbb{S}^f_\triangleright$.

We define:

- $\mathcal{G}(2{\cdot}\mathbb{S}^f_\triangleright) = \mathcal{G}(\mathbb{S}^f_\triangleright) \oplus \mathcal{G}(\mathbb{S}^f_\triangleright)$ : the multiset of **markers** in any doubled genome from the set $2{\cdot}\mathbb{S}^f_\triangleright$
- $\Gamma(2{\cdot}\mathbb{S}^f_\triangleright) = \Gamma(\mathbb{S}^f_\triangleright) \oplus \Gamma(\mathbb{S}^f_\triangleright)$ : the multiset of **adjacencies** in any doubled genome from the set $2{\cdot}\mathbb{S}^f_\triangleright$
- $\Theta(2{\cdot}\mathbb{S}^f_\triangleright) = \Theta(\mathbb{S}^f_\triangleright) \oplus \Theta(\mathbb{S}^f_\triangleright)$ : the multiset of **telomeres** in any doubled genome from the set $2{\cdot}\mathbb{S}^f_\triangleright$

**Breakpoint double distance**:

$$\mathrm{d}^2_{\mathrm{BP}}(\mathbb{S}^f_\triangleright, \mathbb{D}^f_\diamond) = \min_{\mathbb{P}^f_{\bowtie} \in 2\cdot\mathbb{S}^f_\triangleright} \mathrm{d}_{\mathrm{BP}}(\mathbb{P}^f_{\bowtie}, \mathbb{D}^f_\diamond) \quad \Rightarrow \text{greedy approach is consistent}$$

$$= n' - |\Gamma(\mathbb{P}^f_{\bowtie}) \cap \Gamma(\mathbb{D}^f_\diamond)| - \frac{|\Theta(\mathbb{P}^f_{\bowtie}) \cap \Theta(\mathbb{D}^f_\diamond)|}{2}$$

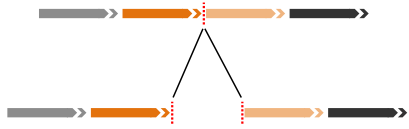
$$= 2n - |\Gamma(2{\cdot}\mathbb{S}^f_\triangleright) \cap \Gamma(\mathbb{D}^f_\diamond)| - \frac{|\Theta(2{\cdot}\mathbb{S}^f_\triangleright) \cap \Theta(\mathbb{D}^f_\diamond)|}{2}$$

$$n = |\mathcal{G}(\mathbb{S}^f_\triangleright)|$$

$$\begin{aligned} n' &= |\mathcal{G}(\mathbb{P}^f_{\bowtie}) \cap \mathcal{G}(2{\cdot}\mathbb{D}^f_\diamond)| \\ &= |\mathcal{G}(2{\cdot}\mathbb{S}^f_\triangleright) \cap \mathcal{G}(2{\cdot}\mathbb{D}^f_\diamond)| \\ &= |\mathcal{G}(2{\cdot}\mathbb{S}^f_\triangleright)| \\ &= 2|\mathcal{G}(\mathbb{S}^f_\triangleright)| \\ &= 2n \end{aligned}$$

# Single-Cut-or-Join (SCJ) model

- A **cut** is an operation that breaks an adjacency of genome $\mathbb{G}$ in two telomeres.
- A **join** is the reverse operation: joins two telomeres of $\mathbb{G}$ into one adjacency.
- Any **single cut** or **single join** is a SCJ operation.



A canonical genome $\mathbb{G}^f_\triangleright$ can be represented by its set of adjacencies $\Gamma(\mathbb{G}^f_\triangleright)$

(the set of telomeres $\Theta(\mathbb{G}^f_\triangleright)$ can be derived from $\Gamma(\mathbb{G}^f_\triangleright)$)

Then, SCJ operations can be seen as set operations:

- A cut of an adjacency $xy$: $\Gamma(\mathbb{G}^f_\triangleright) \setminus \{xy\}$.
- A join of an adjacency $xy$: $\Gamma(\mathbb{G}^f_\triangleright) \cup \{xy\}$.

# SCJ distance and sorting of canonical genomes

The SCJ distance $d_{SCJ}(\mathbb{A}^f_\triangleright, \mathbb{B}^f_\triangleright)$ is the minimum number of SCJs that transform $\Gamma(\mathbb{A}^f_\triangleright)$ into $\Gamma(\mathbb{B}^f_\triangleright)$

The only allowed operations are to remove an element from and to include an element in a set

↘ A lower bound is derived from the simple difference between the two given sets:

$$d_{SCJ}(\mathbb{A}^f_\triangleright, \mathbb{B}^f_\triangleright) \geq |\Gamma(\mathbb{A}^f_\triangleright) \setminus \Gamma(\mathbb{B}^f_\triangleright)| + |\Gamma(\mathbb{B}^f_\triangleright) \setminus \Gamma(\mathbb{A}^f_\triangleright)|$$

We can achieve this lower bound by ensuring that all adjacencies that must be included are available (the corresponding involved extremities are "free"):

1. First, remove all elements of $\Gamma(\mathbb{A}^f_\triangleright)$ that are not present in $\Gamma(\mathbb{B}^f_\triangleright)$:
   $$\#\text{ of single cut operations} = |\Gamma(\mathbb{A}^f_\triangleright) \setminus \Gamma(\mathbb{B}^f_\triangleright)|$$

2. Then, include in $\Gamma(\mathbb{A}^f_\triangleright)$ all elements of $\Gamma(\mathbb{B}^f_\triangleright)$ that are not already present in $\Gamma(\mathbb{A}^f_\triangleright)$:
   $$\#\text{ of single join operations} = |\Gamma(\mathbb{B}^f_\triangleright) \setminus \Gamma(\mathbb{A}^f_\triangleright)|$$

**SCJ distance**

$$d_{SCJ}(\mathbb{A}^f_\triangleright, \mathbb{B}^f_\triangleright) = |\Gamma(\mathbb{A}^f_\triangleright) \setminus \Gamma(\mathbb{B}^f_\triangleright)| + |\Gamma(\mathbb{B}^f_\triangleright) \setminus \Gamma(\mathbb{A}^f_\triangleright)|$$

# SCJ sorting of $\mathbb{A}^f_\triangleright$ into $\mathbb{B}^f_\triangleright$

$\Gamma(\mathbb{A}^f_\triangleright) =$ $\{1^h3^h, 3^t2^h, 2^t4^t\}$

$\Gamma(\mathbb{B}^f_\triangleright) =$ $\{1^h2^t, 2^h3^t, 3^h4^t\}$

# SCJ sorting of $\mathbb{A}^f_{\triangleright}$ into $\mathbb{B}^f_{\triangleright}$

$\Gamma(\mathbb{A}^f_{\triangleright}) = \{1^h3^h, 3^t2^h, 2^t4^t\}$

$\Gamma(\mathbb{I}'^f_{\triangleright}) = \Gamma(\mathbb{A}^f_{\triangleright}) \setminus \{1^h3^h\} = \{3^t2^h, 2^t4^t\}$

$\Gamma(\mathbb{I}''^f_{\triangleright}) = \Gamma(\mathbb{I}'^f_{\triangleright}) \setminus \{2^t4^t\} = \{3^t2^h\}$

$\Gamma(\mathbb{I}'''^f_{\triangleright}) = \Gamma(\mathbb{I}''^f_{\triangleright}) \cup \{1^h2^t\} = \{1^h2^t, 2^h3^t\}$

$\Gamma(\mathbb{B}^f_{\triangleright}) = \Gamma(\mathbb{I}'''^f_{\triangleright}) \cup \{3^h4^t\} = \{1^h2^t, 2^h3^t, 3^h4^t\}$

# Alternative formula for the SCJ distance of canonical genomes

$$
\begin{aligned}
d_{SCJ}(\mathbb{A}^f_\triangleright, \mathbb{B}^f_\triangleright) &= |\Gamma(\mathbb{A}^f_\triangleright) \setminus \Gamma(\mathbb{B}^f_\triangleright)| + |\Gamma(\mathbb{B}^f_\triangleright) \setminus \Gamma(\mathbb{A}^f_\triangleright)| \\
&= |\Gamma(\mathbb{A}^f_\triangleright)| - |\Gamma(\mathbb{A}^f_\triangleright) \cap \Gamma(\mathbb{B}^f_\triangleright)| + |\Gamma(\mathbb{B}^f_\triangleright)| - |\Gamma(\mathbb{A}^f_\triangleright) \cap \Gamma(\mathbb{B}^f_\triangleright)| \\
&= |\Gamma(\mathbb{A}^f_\triangleright)| + |\Gamma(\mathbb{B}^f_\triangleright)| - 2|\Gamma(\mathbb{A}^f_\triangleright) \cap \Gamma(\mathbb{B}^f_\triangleright)| \\
&= |\Gamma(\mathbb{A}^f_\triangleright)| + |\Gamma(\mathbb{B}^f_\triangleright)| - 2|\Gamma_\star|
\end{aligned}
$$

Note that: $|\Theta(\mathbb{A}^f_\triangleright)| = 2(n - |\Gamma(\mathbb{A}^f_\triangleright)|) \quad \Rightarrow \quad |\Gamma(\mathbb{A}^f_\triangleright)| = n - \frac{|\Theta(\mathbb{A}^f_\triangleright)|}{2}$, where $n = |\mathcal{G}(\mathbb{A}^f_\triangleright)|$

$$
\begin{aligned}
d_{SCJ}(\mathbb{A}^f_\triangleright, \mathbb{B}^f_\triangleright) &= n - \frac{|\Theta(\mathbb{A}^f_\triangleright)|}{2} + n - \frac{|\Theta(\mathbb{B}^f_\triangleright)|}{2} - 2|\Gamma_\star| \\
&= 2n - 2a - \frac{|\Theta(\mathbb{A}^f_\triangleright)| + |\Theta(\mathbb{B}^f_\triangleright|)}{2} \\
&= 2n - 2a - \kappa(\mathbb{A}) - \kappa(\mathbb{B})
\end{aligned}
$$

where $n = |\mathcal{G}_\star| = |\mathcal{G}(\mathbb{A}^f_\triangleright)| = |\mathcal{G}(\mathbb{B}^f_\triangleright)|$, $a = |\Gamma_\star|$ and
$\kappa(.)$ is the number of linear chromosomes in the respective genome

# Breakpoint distance $\times$ SCJ distance

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

$$\begin{aligned}
\mathrm{d_{SCJ}}(\mathbb{A}^f_{\triangleright}, \mathbb{B}^f_{\triangleright}) &= 2n - 2a - \kappa(\mathbb{A}) - \kappa(\mathbb{B}) \\
&= 2n - 2a - \kappa(\mathbb{A}) - \kappa(\mathbb{B}) - t + t \\
&= 2n - 2a - t - \kappa(\mathbb{A}) - \kappa(\mathbb{B}) + t \\
&= 2(n - a - \frac{t}{2}) - \kappa(\mathbb{A}) - \kappa(\mathbb{B}) + t \\
&= 2\mathrm{d_{BP}}(\mathbb{A}^f_{\triangleright}, \mathbb{B}^f_{\triangleright}) - \kappa(\mathbb{A}) - \kappa(\mathbb{B}) + t
\end{aligned}$$

Note that: $t \leq \kappa(\mathbb{A}) + \kappa(\mathbb{B})$

For circular genomes:

$$\mathrm{d_{SCJ}}(\mathbb{A}^f_{\triangleright}, \mathbb{B}^f_{\triangleright}) = 2\mathrm{d_{BP}}(\mathbb{A}^f_{\triangleright}, \mathbb{B}^f_{\triangleright})$$

In general:

$$\mathrm{d_{BP}}(\mathbb{A}^f_{\triangleright}, \mathbb{B}^f_{\triangleright}) \leq \mathrm{d_{SCJ}}(\mathbb{A}^f_{\triangleright}, \mathbb{B}^f_{\triangleright}) \leq 2\mathrm{d_{BP}}(\mathbb{A}^f_{\triangleright}, \mathbb{B}^f_{\triangleright})$$

# SCJ double distance

The SCJ distance of **balanced genomes** $\mathbb{A}^f$ and $\mathbb{B}^f$ is:

$$\mathrm{d}_{\mathrm{SCJ}}(\mathbb{A}^f, \mathbb{B}^f) = \min_{f_m} \mathrm{d}_{\mathrm{SCJ}}(\mathbb{A}^{f_m}_{\triangleright}, \mathbb{B}^{f_m}_{\triangleright})$$

where $f_m$ is any function that produces a maximal matching of the families defined by $f$

**SCJ double distance**:

$$\mathrm{d}^2_{\mathrm{SCJ}}(\mathbb{S}^f_{\triangleright}, \mathbb{D}^f_{\diamond}) = \min_{\mathbb{P}^f_{\bowtie} \in 2\cdot\mathbb{S}^f_{\triangleright}} \mathrm{d}_{\mathrm{SCJ}}(\mathbb{P}^f_{\bowtie}, \mathbb{D}^f_{\diamond}) = \texttt{greedy approach}$$

$$= |\Gamma(\mathbb{P}^f_{\bowtie}) \setminus \Gamma(\mathbb{D}^f_{\diamond})| + |\Gamma(\mathbb{D}^f_{\diamond}) \setminus \Gamma(\mathbb{P}^f_{\bowtie})|$$

$$= |\Gamma(2\cdot\mathbb{S}^f_{\triangleright}) \setminus \Gamma(\mathbb{D}^f_{\diamond})| + |\Gamma(\mathbb{D}^f_{\diamond}) \setminus \Gamma(2\cdot\mathbb{S}^f_{\triangleright})|$$

Ex: $\mathbb{S} = [\bar{2}\,1\,\bar{3}]$ and $\mathbb{D} = [3\bar{1}\,\bar{2}\,3\,\bar{1}\,2]$

# Quiz 1

Given genomes $\mathbb{G}_1^f = [\bar{2}\,\bar{1}]\ \ [\bar{4}\,\bar{3}\,5]$ , $\mathbb{G}_2^f = [1\,2\,3\,4\,5]$ and $\mathbb{G}_3^f = (1\,2\,3\,4)\ \ [1\,\bar{5}\,\bar{4}\,5\,\bar{3}\,\bar{2}]$ :

1 What is the SCJ distance of $\mathbb{G}_1^f$ and $\mathbb{G}_2^f$?

A 2

B 2.5

C 3

D 4

2 What is the SCJ double distance of $\mathbb{G}_2^f$ and $\mathbb{G}_3^f$?

A 6

B 7

C 7.5

D 8

# Occurrences of families

Given a family $\mathtt{X}$ and a genome $\mathbb{G}^f$, let $\phi(\mathtt{X}, \mathbb{G}^f)$ be the number of occurrences of $\mathtt{X}$ in $\mathcal{G}(\mathbb{G}^f)$.

If genome $\mathbb{S}^f_\triangleright$ is singular, then $\phi(\mathtt{X}, \mathbb{S}^f_\triangleright) = 1$ for each $\mathtt{X} \in \mathcal{F}(\mathbb{S}^f_\triangleright)$.

If genome $\mathbb{D}^f_\diamond$ is duplicated, then $\phi(\mathtt{X}, \mathbb{D}^f_\diamond) = 2$ for each $\mathtt{X} \in \mathcal{F}(\mathbb{D}^f_\diamond)$.

If genomes $\mathbb{S}^f_\triangleright$ and $\mathbb{S'}^f_\triangleright$ are canonical, then

$$\mathcal{F}_\star = \mathcal{F}(\mathbb{S}^f_\triangleright) = \mathcal{F}(\mathbb{S'}^f_\triangleright) \text{ and } \phi(\mathtt{X}, \mathbb{S}^f_\triangleright) = \phi(\mathtt{X}, \mathbb{S'}^f_\triangleright) = 1 \text{ for each } \mathtt{X} \in \mathcal{F}_\star.$$

If genomes $\mathbb{B}^f_1$ and $\mathbb{B}^f_2$ are balanced, then

$$\mathcal{F}_\star = \mathcal{F}(\mathbb{B}^f_1) = \mathcal{F}(\mathbb{B}^f_2) \text{ and } \phi(\mathtt{X}, \mathbb{B}^f_1) = \phi(f, \mathbb{B}^f_2) \text{ for each } \mathtt{X} \in \mathcal{F}_\star.$$

A maximal matching of the genes of two genomes $\mathbb{A}^f_1$ and $\mathbb{A}^f_2$ has size:

$$\sum_{\mathtt{X} \in = \mathcal{F}(\mathbb{A}^f_1) \cup \mathcal{F}(\mathbb{A}^f_2)} \min\{\phi(\mathtt{X}, \mathbb{A}^f_1), \phi(\mathtt{X}, \mathbb{A}^f_2)\}$$

# Occurrences of adjacencies

Given an adjacency $xy$ and a genome $\mathbb{G}^f$, let $\phi(xy, \mathbb{G}^f)$ be the number of occurrences of $xy$ in $\Gamma(\mathbb{G}^f)$.

If genome $\mathbb{S}^f_\triangleright$ is singular, then $\phi(xy, \mathbb{S}^f_\triangleright) = \begin{cases} 1, & xy \in \Gamma(\mathbb{S}^f_\triangleright), \\ 0, & xy \notin \Gamma(\mathbb{S}^f_\triangleright). \end{cases}$

If genome $\mathbb{D}^f_\diamond$ is duplicated, then $\phi(xy, \mathbb{D}) \in \{0, 1, 2\}$.

Given an adjacency $xy$ and a set of $k$ genomes $\mathcal{A}^f = \{\mathbb{A}^f_1, \mathbb{A}^f_2, \ldots, \mathbb{A}^f_k\}$, we define:

$$\phi(xy, \mathcal{A}^f) = \phi(xy, \mathbb{A}^f_{1..k}) = \sum_{i=1}^{k} \phi(xy, \mathbb{A}^f_i)$$

# Occurrences of telomeres

Given a telomere $x$ and a genome $\mathbb{G}^f$, let $\phi(x, \mathbb{G}^f)$ be the number of occurrences of $x$ in $\Theta(\mathbb{G}^f)$.

If genome $\mathbb{S}^f_\triangleright$ is singular, then $\phi(x, \mathbb{S}^f_\triangleright) = \begin{cases} 1, & x \in \Theta(\mathbb{S}^f_\triangleright), \\ 0, & x \notin \Theta(\mathbb{S}^f_\triangleright). \end{cases}$

If genome $\mathbb{D}^f_\diamond$ is duplicated, then $\phi(x, \mathbb{D}) \in \{0, 1, 2\}$.

Given a telomere $x$ and a set of $k$ genomes $\mathcal{A}^f = \{\mathbb{A}^f_1, \mathbb{A}^f_2, \ldots, \mathbb{A}^f_k\}$, we define:

$$\phi(x, \mathcal{A}^f) = \phi(x, \mathbb{A}^f_{1..k}) = \sum_{i=1}^{k} \phi(x, \mathbb{A}^f_i)$$

# Quiz 2

1 Let $\mathbb{D}^f_\diamond = (1\,2\,3\,4)\ \ [1\,\bar{5}\,\bar{4}\,5\,\bar{3}\,\bar{2}]$. Give, respectively, the values of $\phi(3^h5^t, \mathbb{D}^f_\diamond)$, $\phi(2^h3^t, \mathbb{D}^f_\diamond)$, $\phi(4^h1^t, \mathbb{D}^f_\diamond)$, $\phi(1^t, \mathbb{D}^f_\diamond)$:

A 1, 1, 2, 0

B 0, 2, 0, 2

C 0, 2, 1, 1

D 1, 2, 0, 2

2 Let $\mathbb{C}^f_1 = [1\,2\,3\,4\,5]$ and $\mathbb{C}^f_2 = [\bar{2}\,\bar{1}]\ \ [\bar{4}\,\bar{3}\,5]$. Give, respectively, the values of $\phi(3^h5^t, \{\mathbb{C}^f_1, \mathbb{C}^f_2\})$, $\phi(2^h3^t, \{\mathbb{C}^f_1, \mathbb{C}^f_2\})$, $\phi(1^h2^t, \{\mathbb{C}^f_1, \mathbb{C}^f_2\})$, $\phi(1^t, \{\mathbb{C}^f_1, \mathbb{C}^f_2\})$:

A 0, 1, 1, 2

B 0, 1, 2, 2

C 1, 1, 2, 0

D 1, 2, 0, 2

# SCJ model - expressing the double distance via adjacency occurrences

$$
\begin{aligned}
d^2_{\mathrm{SCJ}}(\mathbb{S}^f_{\triangleright}, \mathbb{D}^f_{\diamond}) &= |\Gamma(2{\cdot}\mathbb{S}^f_{\triangleright}) \setminus \Gamma(\mathbb{D}^f_{\diamond})| \quad + \quad |\Gamma(\mathbb{D}^f_{\diamond}) \setminus \Gamma(2{\cdot}\mathbb{S}^f_{\triangleright})| \\
&= \sum_{xy \in \Gamma(\mathbb{H}^f_{\triangleright})} (2 - \phi(xy, \mathbb{D}^f_{\diamond})) \quad + \quad \sum_{xy \notin \Gamma(\mathbb{H}^f_{\triangleright})} \phi(xy, \mathbb{D}^f_{\diamond}) \\
&= \sum_{xy} [\ \phi(xy, \mathbb{H}^f_{\triangleright}) \cdot (2 - \phi(xy, \mathbb{D}^f_{\diamond})) \quad + \quad (1 - \phi(xy, \mathbb{H}^f_{\triangleright})) \cdot \phi(xy, \mathbb{D}^f_{\diamond}) \ ] \\
&= \sum_{xy} [\ 2 \cdot \phi(xy, \mathbb{H}^f_{\triangleright}) - \phi(xy, \mathbb{H}^f_{\triangleright}) \cdot \phi(xy, \mathbb{D}^f_{\diamond}) \quad + \quad \phi(xy, \mathbb{D}^f_{\diamond}) - \phi(xy, \mathbb{H}^f_{\triangleright}) \cdot \phi(xy, \mathbb{D}^f_{\diamond}) \ ] \\
&= |\Gamma(\mathbb{D}^f_{\diamond})| \quad + \quad \sum_{xy} [\ \phi(xy, \mathbb{H}^f_{\diamond})(2 - 2 \cdot \phi(xy, \mathbb{D}^f_{\diamond})) \ ] \\
&= |\Gamma(\mathbb{D}^f_{\diamond})| \quad + \quad \sum_{xy \in \Gamma(\mathbb{H}^f_{\triangleright})} (2 - 2 \cdot \phi(xy, \mathbb{D}^f_{\diamond}))
\end{aligned}
$$

# SCJ halving of a duplicated genome

Given a duplicated genome $\mathbb{D}^f_\diamond$, find a singular genome $\mathbb{H}^f_\triangleright$ that minimizes the SCJ double distance:

$$\mathrm{d}^2_{\mathrm{SCJ}}(\mathbb{H}^f_\triangleright, \mathbb{D}^f_\diamond) = \mathrm{d}_{\mathrm{SCJ}}(2\cdot\mathbb{H}^f_\triangleright, \mathbb{D}^f_\diamond) = |\Gamma(\mathbb{D}^f_\diamond)| + \sum_{xy\in\Gamma(\mathbb{H}^f_\triangleright)} (2 - 2\cdot\phi(xy, \mathbb{D}^f_\diamond))$$

$$= |\Gamma(\mathbb{D}^f_\diamond)| + \omega(\mathbb{H}^f_\triangleright)$$

Since $|\Gamma(\mathbb{D}^f_\diamond)|$ is given (does not depend on $\mathbb{H}^f_\triangleright$), for minimizing $\mathrm{d}^2_{\mathrm{SCJ}}(\mathbb{H}^f_\triangleright, \mathbb{D}^f_\diamond)$ we need to minimize:

$$\omega(\mathbb{H}^f_\triangleright) = \sum_{xy\in\Gamma(\mathbb{H}^f_\triangleright)} \omega(xy) = \sum_{xy\in\Gamma(\mathbb{H}^f_\triangleright)} (2 - 2\cdot\phi(xy, \mathbb{D}^f_\diamond)), \text{ where } \omega(xy) = 2 - 2\cdot\phi(xy, \mathbb{D}^f_\diamond) \in \{-2, 0, +2\}$$

.

For minimizing $\omega(\mathbb{H}^f_\triangleright)$:

- Do not add to $\mathbb{H}^f_\triangleright$ any adjacency $xz$ that have $\omega(xz) > 0$:
  this happens when $\phi(xz, \mathbb{D}^f_\diamond) = 0$ ($xz$ does not occur in $\mathbb{D}^f_\diamond$).
- Add to $\mathbb{H}^f_\triangleright$ any adjacency $xy$ that have $\omega(xy) < 0$:
  this happens when $\phi(xy, \mathbb{D}^f_\diamond) = 2$ ($xy$ occurs twice in $\mathbb{D}^f_\diamond$).
- For $z \neq y$: $\omega(xz) > 0 \Leftrightarrow \omega(xy) < 0$.
- Any adjacency $xy$ with $\omega(xy) = 0$ (occurs once in $\mathbb{D}^f_\diamond$) is optional (can be added to $\mathbb{H}^f_\triangleright$ or not).

Solution with the minimum number of adjacencies: $\Gamma(\mathbb{H}^f_\triangleright) = \{xy : \phi(xy, \mathbb{D}^f_\diamond) = 2\}$

Solution with the maximum number of adjacencies: $\Gamma(\mathbb{H}^f_\triangleright) = \{xy : \phi(xy, \mathbb{D}^f_\diamond) \geq 1\}$

# SCJ median of three canonical genomes

Given three canonical genomes $\mathbb{C}_1^f$, $\mathbb{C}_2^f$ and $\mathbb{C}_3^f$, find another genome $\mathbb{M}_\triangleright^f$ such that:

1. $\mathbb{M}_\triangleright^f$ is canonical with $\mathbb{C}_1^f$, $\mathbb{C}_2^f$ and $\mathbb{C}_3^f$,
2. $\mathbb{M}_\triangleright^f$ minimizes the sum $\mathsf{s}_{\mathrm{SCJ}}(\mathbb{M}_\triangleright^f) = \mathsf{d}_{\mathrm{SCJ}}(\mathbb{M}_\triangleright^f, \mathbb{C}_1^f) + \mathsf{d}_{\mathrm{SCJ}}(\mathbb{M}_\triangleright^f, \mathbb{C}_2^f) + \mathsf{d}_{\mathrm{SCJ}}(\mathbb{M}_\triangleright^f, \mathbb{C}_3^f)$.

Note that:

$$\begin{aligned}\mathsf{d}_{\mathrm{SCJ}}(\mathbb{M}_\triangleright^f, \mathbb{C}_i^f) &= |\Gamma(\mathbb{M}_\triangleright^f) \setminus \Gamma(\mathbb{C}_i^f)| + |\Gamma(\mathbb{C}_i^f) \setminus \Gamma(\mathbb{M}_\triangleright^f)| \\ &= \textstyle\sum_{xy \in \Gamma(\mathbb{M}_\triangleright^f)} (1 - \phi(xy, \mathbb{C}_i^f)) + \sum_{xy \notin \Gamma(\mathbb{M}_\triangleright^f)} \phi(xy, \mathbb{C}_i^f)\end{aligned}$$

Therefore:

$$\begin{aligned}\mathsf{s}_{\mathrm{SCJ}}(\mathbb{M}_\triangleright^f) &= \textstyle\sum_{xy \in \Gamma(\mathbb{M}_\triangleright^f)} [\, 1 - \phi(xy, \mathbb{C}_1^f)) + (1 - \phi(xy, \mathbb{C}_2^f)) + (1 - \phi(xy, \mathbb{C}_3^f)\,] \\ &\quad + \textstyle\sum_{xy \notin \Gamma(\mathbb{M}_\triangleright^f)} [\, \phi(xy, \mathbb{C}_1^f) + \phi(xy, \mathbb{C}_2^f) + \phi(xy, \mathbb{C}_3^f)\,] \\ &= \textstyle\sum_{xy \in \Gamma(\mathbb{M}_\triangleright^f)} (3 - \phi(xy, \mathbb{C}_{1..3}^f)) + \sum_{xy \notin \Gamma(\mathbb{M}_\triangleright^f)} \phi(xy, \mathbb{C}_{1..3}^f) \\ &= \textstyle\sum_{xy} [\, \phi(xy, \mathbb{M}_\triangleright^f) \cdot (3 - \phi(xy, \mathbb{C}_{1..3}^f)) + (1 - \phi(xy, \mathbb{M}_\triangleright^f)) \cdot \phi(xy, \mathbb{C}_{1..3}^f)\,] \\ &= \textstyle\sum_{xy} [\, 3 \cdot \phi(xy, \mathbb{M}_\triangleright^f) - \phi(xy, \mathbb{M}_\triangleright^f) \cdot \phi(xy, \mathbb{C}_{1..3}^f) + \phi(xy, \mathbb{C}_{1..3}^f) - \phi(xy, \mathbb{M}_\triangleright^f) \cdot \phi(xy, \mathbb{C}_{1..3}^f)\,] \\ &= |\Gamma(\mathbb{C}_1^f)| + |\Gamma(\mathbb{C}_2^f)| + |\Gamma(\mathbb{C}_3^f)| + \textstyle\sum_{xy} [\, \phi(xy, \mathbb{M}_\triangleright^f)(3 - 2 \cdot \phi(xy, \mathbb{C}_{1..3}^f))\,] \\ &= |\Gamma(\mathbb{C}_1^f)| + |\Gamma(\mathbb{C}_2^f)| + |\Gamma(\mathbb{C}_3^f)| + \textstyle\sum_{xy \in \Gamma(\mathbb{M}_\triangleright^f)} (3 - 2 \cdot \phi(xy, \mathbb{C}_{1..3}^f))\end{aligned}$$

# SCJ median of three canonical genomes

$$\begin{aligned} s_{\text{SCJ}}(\mathbb{M}^f_{\triangleright}) &= |\Gamma(\mathbb{C}^f_1)| + |\Gamma(\mathbb{C}^f_2)| + |\Gamma(\mathbb{C}^f_3)| \;+\; \textstyle\sum_{xy \in \Gamma(\mathbb{M}^f_{\triangleright})} (3 - 2 \cdot \phi(xy, \mathbb{C}^f_{1..3})) \\ &= |\Gamma(\mathbb{C}^f_1)| + |\Gamma(\mathbb{C}^f_2)| + |\Gamma(\mathbb{C}^f_3)| \;+\; \omega(\mathbb{M}^f_{\triangleright}) \end{aligned}$$

Since $|\Gamma(\mathbb{C}^f_1)| + |\Gamma(\mathbb{C}^f_2)| + |\Gamma(\mathbb{C}^f_3)|$ is given (does not depend on $\mathbb{M}^f_{\triangleright}$), for minimizing $s_{\text{SCJ}}(\mathbb{M}^f_{\triangleright})$ we need to minimize:

$$\omega(\mathbb{M}^f_{\triangleright}) = \sum_{xy \in \Gamma(\mathbb{M}^f_{\triangleright})} \omega(xy) = \sum_{xy \in \Gamma(\mathbb{M}^f_{\triangleright})} (3 - 2 \cdot \phi(xy, \mathbb{C}^f_{1..3}))$$

where $\omega(xy) = 3 - 2 \cdot \phi(xy, \mathbb{C}^f_{1..3}) \in \{-3, -1, +1, +3\}$.

For minimizing $\omega(\mathbb{M}^f_{\triangleright})$:

- Do not add to $\mathbb{M}^f_{\triangleright}$ any adjacency $xz$ that have $\omega(xz) > 0$:
  this happens when $\phi(xz, \mathbb{C}^f_{1..3}) \leq 1$ ($xz$ occurs in at most one genome among $\mathbb{C}^f_1$, $\mathbb{C}^f_2$ and $\mathbb{C}^f_3$).
- Add to $\mathbb{M}^f_{\triangleright}$ any adjacency $xy$ that have $\omega(xy) < 0$:
  this happens when $\phi(xy, \mathbb{C}^f_{1..3}) \geq 2$ ($xy$ occurs in at least two genomes among $\mathbb{C}^f_1$, $\mathbb{C}^f_2$ and $\mathbb{C}^f_3$).
- For $z \neq y$: $\omega(xz) > 0 \Leftrightarrow \omega(xy) < 0$.

There is no adjacency $xy$ with $\omega(xy) = 0$. Therefore, the SCJ median problem has a unique solution:

$$\Gamma(\mathbb{M}^f_{\triangleright}) = \{xy : \phi(xy, \mathbb{C}^f_{1..3}) \geq 2\}$$

# SCJ median of three canonical genomes - intuition

Let $\mathcal{F}_\star = \mathcal{G}_\star = \{1, 2, 3, ..., n\}$
and start with $\mathbb{M}^f_\triangleright = [1]\ [2]\ \ldots\ [n]$

$\Gamma(\mathbb{M}^f_\triangleright) = \emptyset$ and $\mathsf{s}_{\mathrm{SCJ}}(\mathbb{M}^f_\triangleright) = |\Gamma(\mathbb{C}^f_1)| + |\Gamma(\mathbb{C}^f_2)| + |\Gamma(\mathbb{C}^f_3)|$

Effect of adding an adjacency $xy$ to $\mathbb{M}^f_\triangleright$:

1. If $xy$ is not present in any genome among $\{\mathbb{C}^f_1, \mathbb{C}^f_2, \mathbb{C}^f_3\}$, then $\Delta\mathsf{s}_{\mathrm{SCJ}} = +3$.
2. If $xy$ is present in exactly one genome among $\{\mathbb{C}^f_1, \mathbb{C}^f_2, \mathbb{C}^f_3\}$, then $\Delta\mathsf{s}_{\mathrm{SCJ}} = +1$.
   $(\Delta\mathsf{d}_{\mathrm{SCJ}}(\mathbb{M}^f_\triangleright, \mathbb{C}^f_i) = -1$, but $2 \times \Delta\mathsf{d}_{\mathrm{SCJ}}(\mathbb{M}^f_\triangleright, \mathbb{C}^f_i) = +1)$
3. If $xy$ is present in exactly two genomes among $\{\mathbb{C}^f_1, \mathbb{C}^f_2, \mathbb{C}^f_3\}$, then $\Delta\mathsf{s}_{\mathrm{SCJ}} = -1$.
   $(2 \times \Delta\mathsf{d}_{\mathrm{SCJ}}(\mathbb{M}^f_\triangleright, \mathbb{C}^f_i) = -1$, but $\Delta\mathsf{d}_{\mathrm{SCJ}}(\mathbb{M}^f_\triangleright, \mathbb{C}^f_i) = +1)$
4. If $xy$ is present in all three genomes $\{\mathbb{C}^f_1, \mathbb{C}^f_2, \mathbb{C}^f_3\}$, then $\Delta\mathsf{s}_{\mathrm{SCJ}} = -3$.

# SCJ median of $k$ canonical genomes

Given $k$ canonical genomes $\mathbb{C}_1^f$, $\mathbb{C}_2^f$, $\ldots \mathbb{C}_k^f$, find another canonical genome $\mathbb{M}_\triangleright^f$ that minimizes the sum:

$$\begin{aligned} s_{SCJ}(\mathbb{M}_\triangleright^f) &= d_{SCJ}(\mathbb{M}_\triangleright^f, \mathbb{C}_1^f) + d_{SCJ}(\mathbb{M}_\triangleright^f, \mathbb{C}_2^f) + \ldots + d_{SCJ}(\mathbb{M}_\triangleright^f, \mathbb{C}_k^f) \\ &= |\Gamma(\mathbb{C}_1^f)| + |\Gamma(\mathbb{C}_2^f)| + \ldots + |\Gamma(\mathbb{C}_k^f)| \quad + \quad \omega(\mathbb{M}_\triangleright^f) \end{aligned}$$

Analogously to the median of three genomes, we need to minimize:

$$\omega(\mathbb{M}_\triangleright^f) = \sum_{xy \in \Gamma(\mathbb{M}_\triangleright^f)} \omega(xy)$$

where $\omega(xy) = k - 2 \cdot \phi(xy, \mathbb{C}_{1..k}^f) \in \{-k, -k+2, \ldots, +k-2, +k\}$.

For minimizing $\omega(\mathbb{M}_\triangleright^f)$:

- Do not add to $\mathbb{M}_\triangleright^f$ any adjacency $xz$ that have $\omega(xz) > 0$:
  this happens when $\phi(xz, \mathbb{C}_{1..k}^f) < \frac{k}{2}$ ($xz$ occurs in less than half of the genomes among $\mathbb{C}_1^f$, $\mathbb{C}_2^f$, $\ldots$, $\mathbb{C}_k^f$).
- Add to $\mathbb{M}_\triangleright^f$ any adjacency $xy$ that have $\omega(xy) < 0$:
  this happens when $\phi(xy, \mathbb{C}_{1..3}^f) > \frac{k}{2}$ ($xy$ occurs in more than half of the genomes among $\mathbb{C}_1^f$, $\mathbb{C}_2^f$, $\ldots$, $\mathbb{C}_k^f$).
- For $z \neq y$: $\omega(xz) > 0 \Leftrightarrow \omega(xy) < 0$.
- Any adjacency $xy$ with $\omega(xy) = 0$ is optional (can be added to the median or not). If there is no such an adjacency (e.g., if $k$ is odd), the SCJ median problem has a unique solution.

In general, the following set of adjacencies define a SCJ median of $k$ genomes:

$$\Gamma(\mathbb{M}_\triangleright^f) = \left\{ xy : \phi(xy, \mathbb{C}_{1..k}^f) > \frac{k}{2} \right\}$$

# SCJ linear median of $k$ canonical linear genomes

1. Compute the general SCJ median $\mathbb{M}^f_{\triangleright}$ as described above.

2. For each circular chromosome in $\mathbb{M}^f_{\triangleright}$, remove one adjacency $xy$ with smallest weight $\omega(xy)$.

# Quiz 3

1 Which of the following statements are true?

A The SCJ halving is always satisfied by a unique singular genome.

B The SCJ halving cannot be satisfied by a unique singular genome.

C The SCJ median of four canonical genomes is always unique.

D The SCJ median of four canonical genomes cannot be unique.

E The SCJ median of three canonical genomes is always unique.

F The SCJ linear median of three canonical linear genomes is always unique.

# References

Multichromosomal median and halving problems under different genomic distances

(Eric Tannier, Chunfang Zheng and David Sankoff)

BMC Bioinformatics volume 10, Article number: 120 (2009)

SCJ: A Breakpoint-Like Distance that Simplifies Several Rearrangement Problems

(Pedro Feijão and João Meidanis)

TCBB volume 8 Number: 5 (2011)