

## Topics of today:

1. Formalizing the number of occurrences ( $\phi$ ) of families/adjacencies/telomeres
2. Revisiting breakpoint and SCJ double distances
3. SCJ median, halving and guided halving
4. Breakpoint median, halving and guided halving

## Occurrences of families

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

If  $\mathbb{S}$  is singular, then  $\phi(f, \mathbb{S}) = 1$  for each  $f \in \mathcal{F}(\mathbb{S})$ .

If  $\mathbb{D}$  is duplicated, then  $\phi(f, \mathbb{D}) = 2$  for each  $f \in \mathcal{F}(\mathbb{D})$ .

If  $\mathbb{C}_1$  and  $\mathbb{C}_2$  are canonical, then  $\mathcal{F}_* = \mathcal{F}(\mathbb{C}_1) = \mathcal{F}(\mathbb{C}_2)$  and  $\phi(f, \mathbb{C}_1) = \phi(f, \mathbb{C}_2) = 1$  for each  $f \in \mathcal{F}_*$ .

If  $\mathbb{B}_1$  and  $\mathbb{B}_2$  are balanced, then  $\mathcal{F}_* = \mathcal{F}(\mathbb{B}_1) = \mathcal{F}(\mathbb{B}_2)$  and  $\phi(f, \mathbb{B}_1) = \phi(f, \mathbb{B}_2)$  for each  $f \in \mathcal{F}_*$ .

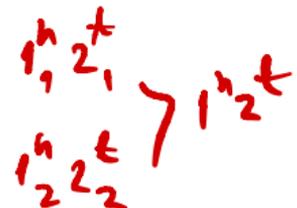
# Occurrences of adjacencies

Given an adjacency  $xy$  and a canonical genome  $\mathbb{C}$ , let  $\phi(xy, \mathbb{C}) = \begin{cases} 1, & xy \in \alpha(\mathbb{C}), \\ 0, & xy \notin \alpha(\mathbb{C}). \end{cases}$

Given an adjacency  $xy$  and a duplicated genome  $\mathbb{D}$ ,

let  $\phi(xy, \mathbb{D})$  be the number of occurrences of adjacencies of type  $x_i y_j$  in  $\alpha(\mathbb{D})$ .

Note that  $\phi(xy, \mathbb{D}) \in \{0, 1, 2\}$ .



Given an adjacency  $xy$  and  $k$  genomes  $G_1, G_2, \dots, G_k$ ,

let  $\phi(xy, G_1, G_2, \dots, G_k) = \phi(xy, G_{1..k}) = \sum_{i=1}^k \phi(xy, G_i)$ .

# Occurrences of telomeres

Given a telomere  $x$  and a canonical genome  $\mathbb{C}$ , let  $\phi(x, \mathbb{C}) = \begin{cases} 1, & x \in \gamma(\mathbb{C}), \\ 0, & x \notin \gamma(\mathbb{C}). \end{cases}$

Given a telomere  $x$  and a duplicated genome  $\mathbb{D}$ ,

let  $\phi(x, \mathbb{D})$  be the number of occurrences of telomeres of type  $x_{[i]}$  in  $\gamma(\mathbb{D})$ .

Note that  $\phi(x, \mathbb{D}) \in \{0, 1, 2\}$ .

Given a telomere  $x$  and  $k$  genomes  $\mathbb{G}_1, \mathbb{G}_2, \dots, \mathbb{G}_k$ ,

let  $\phi(x, \mathbb{G}_1, \mathbb{G}_2, \dots, \mathbb{G}_k) = \phi(x, \mathbb{G}_{1..k}) = \sum_{i=1}^k \phi(x, \mathbb{G}_i)$ .

# Quiz 1

Given genomes  $\mathbb{D} = (1 \underline{2} \underline{3} \underline{4}) [1 \bar{5} \bar{4} \bar{5} \bar{3} \bar{2}]$ ,  $C_1 = [1 \underline{2} \underline{3} \underline{4} \underline{5}]$  and  $C_2 = [\bar{2} \bar{1}] [\bar{4} \bar{3} \bar{5}]$ :

1 Which are the values of  $\phi(3^h 5^t, \mathbb{D})$ ,  $\phi(2^h 3^t, \mathbb{D})$ ,  $\phi(4^h 1^t, \mathbb{D})$ ,  $\phi(1^t, \mathbb{D})$ ?

0  
A 1, 1, 2, 0

B 0, 2, 0, 2

C 0, 2, 1, 1

D 1, 2, 0, 2

2 Which are the values of  $\phi(3^h 5^t, C_1, C_2)$ ,  $\phi(2^h 3^t, C_1, C_2)$ ,  $\phi(1^h 2^t, C_1, C_2)$ ,  $\phi(1^t, C_1, C_2)$ ?

0  
A 0, 1, 1, 2

B 0, 1, 2, 2

C 1, 1, 2, 0

D 1, 2, 0, 2

# Breakpoint model - distance and double distance

Breakpoint distance of canonical genomes  $\mathbb{C}_1$  and  $\mathbb{C}_2$ , with  $n = |\mathcal{G}_\star|$ ,  $a = |\alpha_\star|$  and  $t = |\gamma_\star|$ :

$$d_{\text{BP}}(\mathbb{C}_1, \mathbb{C}_2) = n - a - \frac{t}{2}.$$

Breakpoint double distance of sing-dup-canonical genomes  $\mathbb{S}$  and  $\mathbb{D}$ , with  $\mathcal{G}_\star = \mathcal{G}(\mathbb{S}) \cap \mathcal{G}(\mathbb{D})$  and  $n = |\mathcal{G}_\star|$ :

$$d_{\text{BP}}^2(\mathbb{S}, \mathbb{D}) = d_{\text{BP}}(2 \cdot \mathbb{S}, \mathbb{D}) = n' - a' - \frac{t'}{2} = 2n - a' - \frac{t'}{2},$$

where  $n' = |\mathcal{G}(2 \cdot \mathbb{S}) \cap \mathcal{G}(\mathbb{D})| = 2|\mathcal{G}_\star| = 2n$ ,  $a' = |\alpha(2 \cdot \mathbb{S}) \cap \alpha(\mathbb{D})|$  and  $t' = |\gamma(2 \cdot \mathbb{S}) \cap \gamma(\mathbb{D})|$ .

Since it is possible to find a matching that fulfills each candidate adjacency/telomere between  $2 \cdot \mathbb{S}$  and  $\mathbb{D}$ :

$$a' = \sum_{xy \in \alpha(\mathbb{S})} \phi(xy, \mathbb{D}) \text{ and}$$
$$t' = \sum_{x \in \gamma(\mathbb{S})} \phi(x, \mathbb{D})$$

# SCJ model - distance and double distance

SCJ distance of canonical genomes  $\mathbb{C}_1$  and  $\mathbb{C}_2$ , with  $n = |\mathcal{G}_\star|$  and  $a = |\alpha_\star|$ :

$$d_{\text{SCJ}}(\mathbb{C}_1, \mathbb{C}_2) = 2n - 2a - \kappa(\mathbb{C}_1) - \kappa(\mathbb{C}_2).$$

SCJ double distance of sing-dup-canonical genomes  $\mathbb{S}$  and  $\mathbb{D}$ , with  $\mathcal{G}_\star = \mathcal{G}(\mathbb{S}) \cap \mathcal{G}(\mathbb{D})$  and  $n = |\mathcal{G}_\star|$ :

$$d_{\text{SCJ}}^2(\mathbb{S}, \mathbb{D}) = d_{\text{SCJ}}(2 \cdot \mathbb{S}, \mathbb{D}) = 2n' - 2a' - \kappa(2 \cdot \mathbb{S}) - \kappa(\mathbb{D}) = 4n - 2a' - 2\kappa(\mathbb{S}) - \kappa(\mathbb{D})$$

where  $n' = |\mathcal{G}(2 \cdot \mathbb{S}) \cap \mathcal{G}(\mathbb{D})| = 2|\mathcal{G}_\star| = 2n$  and  $a' = |\alpha(2 \cdot \mathbb{S}) \cap \alpha(\mathbb{D})|$ .

Since it is possible to find a matching that fulfills each candidate adjacency between  $2 \cdot \mathbb{S}$  and  $\mathbb{D}$ :

$$a' = \sum_{xy \in \alpha(\mathbb{S})} \phi(xy, \mathbb{D})$$

# SCJ median of canonical genomes

Given three canonical genomes  $C_1$ ,  $C_2$  and  $C_3$ , find another canonical genome  $M$  that minimizes the sum:

$$s_{SCJ}(M) = d_{SCJ}(M, C_1) + d_{SCJ}(M, C_2) + d_{SCJ}(M, C_3)$$

Recall that:

$$\begin{aligned} d_{SCJ}(M, C_i) &= |\alpha(M) \setminus \alpha(C_i)| + |\alpha(C_i) \setminus \alpha(M)| \\ &= \sum_{xy \in \alpha(M)} (1 - \phi(xy, C_i)) + \sum_{xy \notin \alpha(M)} \phi(xy, C_i) \end{aligned}$$

Therefore:

$$\begin{aligned} s_{SCJ}(M) &= \sum_{xy \in \alpha(M)} [ (1 - \phi(xy, C_1)) + (1 - \phi(xy, C_2)) + (1 - \phi(xy, C_3)) ] \\ &\quad + \sum_{xy \notin \alpha(M)} [ \phi(xy, C_1) + \phi(xy, C_2) + \phi(xy, C_3) ] \\ &= \sum_{xy \in \alpha(M)} (3 - \phi(xy, C_{1..3})) + \sum_{xy \notin \alpha(M)} \phi(xy, C_{1..3}) \\ &= \sum_{xy} [ \phi(xy, M) \cdot (3 - \phi(xy, C_{1..3})) + (1 - \phi(xy, M)) \cdot \phi(xy, C_{1..3}) ] \\ &= \sum_{xy} [ 3 \cdot \phi(xy, M) - \phi(xy, M) \cdot \phi(xy, C_{1..3}) + \phi(xy, C_{1..3}) - \phi(xy, M) \cdot \phi(xy, C_{1..3}) ] \\ &= |\alpha(C_1)| + |\alpha(C_2)| + |\alpha(C_3)| + \sum_{xy} [ \phi(xy, M)(3 - 2 \cdot \phi(xy, C_{1..3})) ] \\ &= |\alpha(C_1)| + |\alpha(C_2)| + |\alpha(C_3)| + \sum_{xy \in \alpha(M)} (3 - 2 \cdot \phi(xy, C_{1..3})) \end{aligned}$$

## SCJ median of canonical genomes

$$\begin{aligned} s_{\text{SCJ}}(\mathbb{M}) &= \underbrace{|\alpha(\mathbb{C}_1)|}_{\text{}} + \underbrace{|\alpha(\mathbb{C}_2)|}_{\text{}} + \underbrace{|\alpha(\mathbb{C}_3)|}_{\text{}} + \sum_{xy \in \alpha(\mathbb{M})} (3 - 2 \cdot \phi(xy, \mathbb{C}_{1..3})) \\ &= |\alpha(\mathbb{C}_1)| + |\alpha(\mathbb{C}_2)| + |\alpha(\mathbb{C}_3)| + \omega(\mathbb{M}) \end{aligned}$$

Since  $|\alpha(\mathbb{C}_1)| + |\alpha(\mathbb{C}_2)| + |\alpha(\mathbb{C}_3)|$  is given (does not depend on  $\mathbb{M}$ ), for minimizing  $s_{\text{SCJ}}(\mathbb{M})$  we need to minimize:

$$\omega(\mathbb{M}) = \sum_{xy \in \alpha(\mathbb{M})} \omega(xy) = \sum_{xy \in \alpha(\mathbb{M})} (3 - 2 \cdot \phi(xy, \mathbb{C}_{1..3}))$$

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

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

- ▶ Do not add to  $\mathbb{M}$  any adjacency  $xz$  that have  $\omega(xz) > 0$ :  
this happens when  $\phi(xz, \mathbb{C}_{1..3}) \leq 1$  ( $xz$  occurs in at most one genome among  $\mathbb{C}_1, \mathbb{C}_2$  and  $\mathbb{C}_3$ ).
- ▶ Add to  $\mathbb{M}$  any adjacency  $xy$  that have  $\omega(xy) < 0$ :  
this happens when  $\phi(xy, \mathbb{C}_{1..3}) \geq 2$  ( $xy$  occurs in at least two genomes among  $\mathbb{C}_1, \mathbb{C}_2$  and  $\mathbb{C}_3$ ).
- ▶ 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:

$$\alpha(\mathbb{M}) = \{xy : \phi(xy, \mathbb{C}_{1..3}) \geq 2\}$$

## SCJ median of canonical genomes - intuition

Let  $\mathcal{F}_* = \mathcal{G}_* = \{1, 2, 3, \dots, n\}$

and start with  $\mathbb{M} = [1] [2] \dots [n]$

$\alpha(\mathbb{M}) = \emptyset$  and  $s_{\text{SCJ}}(\mathbb{M}) = |\alpha(\mathbb{C}_1)| + |\alpha(\mathbb{C}_2)| + |\alpha(\mathbb{C}_3)|$

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

1. If  $xy$  is not present in any genome among  $\{\mathbb{C}_1, \mathbb{C}_2, \mathbb{C}_3\}$ , then  $\Delta s_{\text{SCJ}} = +3$ .
2. If  $xy$  is present in exactly one genome among  $\{\mathbb{C}_1, \mathbb{C}_2, \mathbb{C}_3\}$ , then  $\Delta s_{\text{SCJ}} = +1$ .  
( $\Delta d_{\text{SCJ}}(\mathbb{M}, \mathbb{C}_i) = -1$ , but  $2 \times \Delta d_{\text{SCJ}}(\mathbb{M}, \mathbb{C}_i) = +1$ )
3. If  $xy$  is present in exactly two genomes among  $\{\mathbb{C}_1, \mathbb{C}_2, \mathbb{C}_3\}$ , then  $\Delta s_{\text{SCJ}} = -1$ .  
( $2 \times \Delta d_{\text{SCJ}}(\mathbb{M}, \mathbb{C}_i) = -1$ , but  $\Delta d_{\text{SCJ}}(\mathbb{M}, \mathbb{C}_i) = +1$ )
4. If  $xy$  is present in all three genomes  $\{\mathbb{C}_1, \mathbb{C}_2, \mathbb{C}_3\}$ , then  $\Delta s_{\text{SCJ}} = -3$ .

# SCJ median of $k$ canonical genomes

Given  $k$  canonical genomes  $C_1, C_2, \dots, C_k$ , find another canonical genome  $M$  that minimizes the sum:

$$\begin{aligned} d_{SCJ}(M) &= d_{SCJ}(M, C_1) + d_{SCJ}(M, C_2) + \dots + d_{SCJ}(M, C_k) \\ &= |\alpha(C_1)| + |\alpha(C_2)| + \dots + |\alpha(C_k)| + \omega(M) \end{aligned}$$

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

$$\omega(M) = \sum_{xy \in \alpha(M)} \omega(xy)$$

where  $\omega(xy) = k - 2 \cdot \phi(xy, C_1..k) \in \{-k, -k+2, \dots, +k-2, +k\}$ .

$k=4$   $\{-4, -2, 0, +2, +4\}$   
 $k=5$   $\{-5, -3, -1, +1, +3, +5\}$

For minimizing  $\omega(M)$ :

- ▶ Do not add to  $M$  any adjacency  $xz$  that have  $\omega(xz) > 0$ :  
this happens when  $\phi(xz, C_1..k) < \frac{k}{2}$  ( $xz$  occurs in less than half of the genomes among  $C_1, C_2, \dots, C_k$ ).
- ▶ Add to  $M$  any adjacency  $xy$  that have  $\omega(xy) < 0$ :  
this happens when  $\phi(xy, C_1..k) > \frac{k}{2}$  ( $xy$  occurs in more than half of the genomes among  $C_1, C_2, \dots, C_k$ ).
- ▶ 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:

$$\alpha(M) = \left\{ xy : \phi(xy, C_1..k) > \frac{k}{2} \right\}$$

# SCJ median of $k$ canonical linear genomes

1. Compute the general SCJ median  $\mathbb{M}$  as described above.
2. For each circular chromosome in  $\mathbb{M}$ , remove one adjacency  $xy$  with smallest weight  $\omega(xy)$ .

# SCJ halving of a duplicated genome



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

$$d_{\text{SCJ}}^2(\mathbb{H}, \mathbb{D}) = d_{\text{SCJ}}(2 \cdot \mathbb{H}, \mathbb{D})$$

Therefore:

$$\begin{aligned}
 d_{\text{SCJ}}(2 \cdot \mathbb{H}, \mathbb{D}) &= |\alpha(2 \cdot \mathbb{H}) \setminus \alpha(\mathbb{D})| + |\alpha(\mathbb{D}) \setminus \alpha(2 \cdot \mathbb{H})| \\
 &= \sum_{xy \in \alpha(\mathbb{H})} (2 - \phi(xy, \mathbb{D})) + \sum_{xy \notin \alpha(\mathbb{H})} \phi(xy, \mathbb{D}) + \sum_{xx} \phi(xx, \mathbb{D}) \\
 &= \sum_{xy} [ \phi(xy, \mathbb{H}) \cdot (2 - \phi(xy, \mathbb{D})) + (1 - \phi(xy, \mathbb{H})) \cdot \phi(xy, \mathbb{D}) ] - \sum_{xx} \phi(xx, \mathbb{D}) \\
 &= \sum_{xy} [ 2 \cdot \phi(xy, \mathbb{H}) - \phi(xy, \mathbb{H}) \cdot \phi(xy, \mathbb{D}) + \phi(xy, \mathbb{D}) - \phi(xy, \mathbb{H}) \cdot \phi(xy, \mathbb{D}) ] + \dots \\
 &= |\alpha(\mathbb{D})| + \sum_{xy} [ \phi(xy, \mathbb{H})(2 - 2 \cdot \phi(xy, \mathbb{D})) ] \\
 &= |\alpha(\mathbb{D})| + \sum_{xy \in \alpha(\mathbb{H})} (2 - 2 \cdot \phi(xy, \mathbb{D}))
 \end{aligned}$$

## SCJ halving of a duplicated genome

$$\begin{aligned}d_{\text{SCJ}}^2(\mathbb{H}, \mathbb{D}) &= |\alpha(\mathbb{D})| + \sum_{xy \in \alpha(\mathbb{H})} (2 - 2 \cdot \phi(xy, \mathbb{D})) \\ &= |\alpha(\mathbb{D})| + \omega(\mathbb{H})\end{aligned}$$

Since  $|\alpha(\mathbb{D})|$  is given (does not depend on  $\mathbb{H}$ ), for minimizing  $d_{\text{SCJ}}^2(\mathbb{H}, \mathbb{D})$  we need to minimize:

$$\omega(\mathbb{H}) = \sum_{xy \in \alpha(\mathbb{H})} \omega(xy) = \sum_{xy \in \alpha(\mathbb{H})} (2 - 2 \cdot \phi(xy, \mathbb{D}))$$

where  $\omega(xy) = 2 - 2 \cdot \phi(xy, \mathbb{D}) \in \{-2, 0, +2\}$ .

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

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

Solution with the minimum number of adjacencies:  $\alpha(\mathbb{H}) = \{xy : \phi(xy, \mathbb{D}) = 2\}$

# SCJ aliquoting of a $k$ -folded genome $\mathbb{K}$ : for each $f \in \mathcal{F}(\mathbb{K})$ , $\phi(f, \mathbb{K}) = k$

Given a  $k$ -folded genome  $\mathbb{K}$ , find a singular genome  $\mathbb{A}$  that minimizes the SCJ  $k$ -folded distance:

$$\underline{d_{\text{SCJ}}^k(\mathbb{A}, \mathbb{K}) = d_{\text{SCJ}}(k \cdot \mathbb{A}, \mathbb{K})}$$

$k \cdot \mathbb{A}$ : each adjacency or  
blomere of  $\mathbb{A}$   
appears  $k$  times

Therefore:

$$\begin{aligned} d_{\text{SCJ}}(k \cdot \mathbb{A}, \mathbb{K}) &= |\alpha(k \cdot \mathbb{A}) \setminus \alpha(\mathbb{K})| + |\alpha(\mathbb{K}) \setminus \alpha(k \cdot \mathbb{A})| \\ &= \sum_{xy \in \alpha(\mathbb{A})} (k - \phi(xy, \mathbb{K})) + \sum_{xy \notin \alpha(\mathbb{A})} \phi(xy, \mathbb{K}) + \sum_{x_2} \phi(x_2, \mathbb{K}) \\ &= \sum_{xy} [ \phi(xy, \mathbb{A}) \cdot (k - \phi(xy, \mathbb{K})) + (1 - \phi(xy, \mathbb{A})) \cdot \phi(xy, \mathbb{K}) ] + \dots \\ &= \sum_{xy} [ k \cdot \phi(xy, \mathbb{A}) - \phi(xy, \mathbb{A}) \cdot \phi(xy, \mathbb{K}) + \phi(xy, \mathbb{K}) - \phi(xy, \mathbb{A}) \cdot \phi(xy, \mathbb{K}) ] + \dots \\ &= |\alpha(\mathbb{K})| + \sum_{xy} [ \phi(xy, \mathbb{A})(k - 2 \cdot \phi(xy, \mathbb{K})) ] \\ &= |\alpha(\mathbb{K})| + \sum_{xy \in \alpha(\mathbb{A})} (k - 2 \cdot \phi(xy, \mathbb{K})) \end{aligned}$$

The solution for the SCJ aliquoting problem of a  $k$ -folded genome is:

$$\alpha(\mathbb{A}) = \left\{ xy : \phi(xy, \mathbb{K}) > \frac{k}{2} \right\}$$

# SCJ guided halving/aliquoting of a $k$ -folded genome

Given a  $k$ -folded genome  $\mathbb{K}$  and a canonical genome  $\mathbb{C}$  find a canonical genome  $\mathbb{A}$  that minimizes the sum:

$$ga_{SCJ}(\mathbb{A}) = d_{SCJ}^k(\mathbb{A}, \mathbb{K}) + d_{SCJ}(\mathbb{A}, \mathbb{C}) = d_{SCJ}(k \cdot \mathbb{A}, \mathbb{K}) + d_{SCJ}(\mathbb{A}, \mathbb{C})$$

Therefore:

$$\begin{aligned}
 ga_{SCJ}(\mathbb{A}) &= |\alpha(k \cdot \mathbb{A}) \setminus \alpha(\mathbb{K})| + |\alpha(\mathbb{K}) \setminus \alpha(k \cdot \mathbb{A})| + |\alpha(\mathbb{A}) \setminus \alpha(\mathbb{C})| + |\alpha(\mathbb{C}) \setminus \alpha(\mathbb{A})| \\
 &= \sum_{xy \in \alpha(\mathbb{A})} (k - \phi(xy, \mathbb{K})) + \sum_{xy \notin \alpha(\mathbb{A})} \phi(xy, \mathbb{K}) + \sum_{xy \in \alpha(\mathbb{A})} (1 - \phi(xy, \mathbb{C})) + \sum_{xy \notin \alpha(\mathbb{A})} \phi(xy, \mathbb{C}) \\
 &= \sum_{xy \in \alpha(\mathbb{A})} (k + 1 - \phi(xy, \mathbb{K}, \mathbb{C})) + \sum_{xy \notin \alpha(\mathbb{A})} \phi(xy, \mathbb{K}, \mathbb{C}) \\
 &= \sum_{xy} [ \phi(xy, \mathbb{A}) \cdot (k + 1 - \phi(xy, \mathbb{K}, \mathbb{C})) + (1 - \phi(xy, \mathbb{A})) \cdot \phi(xy, \mathbb{K}, \mathbb{C}) ] \\
 &= \sum_{xy} [ (k + 1) \cdot \phi(xy, \mathbb{A}) - \phi(xy, \mathbb{A}) \cdot \phi(xy, \mathbb{K}, \mathbb{C}) + \phi(xy, \mathbb{K}, \mathbb{C}) - \phi(xy, \mathbb{A}) \cdot \phi(xy, \mathbb{K}, \mathbb{C}) ] \\
 &= |\alpha(\mathbb{K})| + |\alpha(\mathbb{C})| + \sum_{xy} [ \phi(xy, \mathbb{A})(k + 1 - 2 \cdot \phi(xy, \mathbb{K}, \mathbb{C})) ] \\
 &= |\alpha(\mathbb{C})| + |\alpha(\mathbb{K})| + \sum_{xy \in \alpha(\mathbb{A})} (k + 1 - 2 \cdot \phi(xy, \mathbb{K}, \mathbb{C}))
 \end{aligned}$$

The solution for the guided SCJ aliquoting problem of a  $k$ -folded genome is:

$$\alpha(\mathbb{A}) = \left\{ xy : \phi(xy, \mathbb{K}, \mathbb{C}) > \frac{k + 1}{2} \right\}$$

## Quiz 2

1 Which of the following statements are true?

- A The SCJ median of four canonical genomes is always unique.
- B The SCJ median of four canonical genomes cannot be unique.
- C The SCJ median of three canonical genomes is always unique.
- D The SCJ linear median of three canonical linear genomes is always unique.
- E The SCJ guided halving problem is equivalent to the SCJ aliquoting problem.
- F The SCJ aliquoting problem can be constrained to linear genomes only.

$$k=4, w(x_4) \in \{-4, -2, 0, 2, 4\}$$

$$w(x_4) \in \{-3, -1, 1, 3\}$$

# Perfect matching and circular canonical genomes

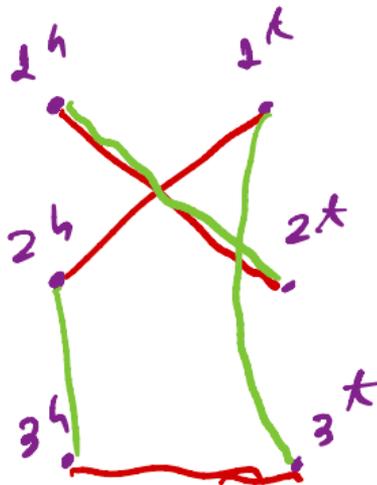
For a given set  $N = \{1, 2, \dots, n\}$ ,

let  $G$  be a complete graph with vertices  $V(G) = \{g^h : g \in N\} \cup \{g^t : g \in N\}$

Ex :  $N = \{1, 2, 3\}$

$(1\ 2)\ (3)$

$(1\ 2\ \bar{3})$



Perfect matching:

set of non-incident edges covering all vertices

A perfect matching  $M$  in  $G$  corresponds to  $|M| = n$  adjacencies and, consequently, defines a circular canonical genome  $\mathbb{C}$ , with  $\mathcal{F}(\mathbb{C}) = \mathcal{G}(\mathbb{C}) = N$  and  $\alpha(\mathbb{C}) = M$ .

# SCJ median of $k$ canonical circular genomes

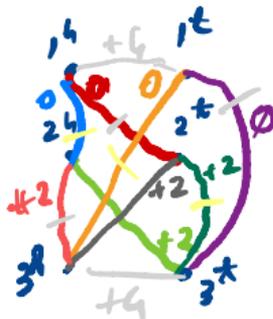
Given  $k$  canonical circular genomes  $C_1, C_2, \dots, C_k$ , find a canonical circular genome  $M$  that minimizes the sum:

$$\begin{aligned} s_{SCJ}(M) &= d_{SCJ}(M, C_1) + d_{SCJ}(M, C_2) + \dots + d_{SCJ}(M, C_k) \\ &= |\alpha(C_1)| + |\alpha(C_2)| + \dots + |\alpha(C_k)| + \omega(M) \\ &= \beta n + \omega(M) \end{aligned}$$

Again, we need to minimize  $\omega(M) = \sum_{xy \in \alpha(M)} \omega(xy)$ , where  $\omega(xy) = k - 2 \cdot \phi(xy, C_{1..k})$ .

1. Build the complete graph  $G$  with vertices  $V(G) = \{g^h : g \in \mathcal{G}_*\} \cup \{g^t : g \in \mathcal{G}_*\}$
2. Assign weights to each edge  $xy$  of  $G$ :  $\omega(xy) = k - 2 \cdot \phi(xy, C_{1..k})$ .

Ex:  $C_1 = (1, 2, 3)$   
 $C_2 = (3, 2, \bar{1})$   
 $C_3 = (3, \bar{1}, \bar{2})$   
 $C_4 = (3, \bar{2}, \bar{1})$



$M_1 = C_3$   
 $M_2 = C_4$

Perfect matching  $M$  in  $G \Leftrightarrow$  Circular genome  $M$  ; with  $\omega(M) = \omega(M)$

Minimum weight matching  $M$  gives a minimum weight circular SCJ median  $M$

$|\alpha(M)| = n$

$(4) (3) (2) (1) (0)$   
 $-4, -2, 0, +2, +4$

# Breakpoint median of canonical circular genomes

Given canonical circular genomes  $C_1$ ,  $C_2$  and  $C_3$ , find a canonical circular genome  $M$  that minimizes the sum:

$$\begin{aligned}
 s_{BP}(M) &= d_{BP}(M, C_1) + d_{BP}(M, C_2) + d_{BP}(M, C_k) \\
 &= n - \sum_{xy \in \alpha(M)} \phi(xy, C_1) + n - \sum_{xy \in \alpha(M)} \phi(xy, C_2) + n - \sum_{xy \in \alpha(M)} \phi(xy, C_3) \\
 &= 3n - \sum_{xy \in \alpha(M)} \phi(xy, C_{1..3}) \\
 &= 3n - \omega'(M)
 \end{aligned}$$

Now we need to maximize  $\omega'(M) = \sum_{xy \in \alpha(M)} \omega'(xy)$ , where  $\omega'(xy) = \phi(xy, C_{1..3})$ .

1. Build the complete graph  $G$  with vertices  $V(G) = \{g^h : g \in \mathcal{G}_*\} \cup \{g^t : g \in \mathcal{G}_*\}$
2. Assign weights to each edge  $xy$  of  $G$ :  $\omega'(xy) = \phi(xy, C_{1..k})$ .

$$\omega(xy) \in \{0, 1, 2, 3\}$$

$$\begin{array}{l}
 C_1 = (1 \ 2 \ 3 \ 4) \\
 C_2 = (4 \ 1) (\bar{3} \ 2) \\
 C_3 = (2 \ 4 \ 1) (3)
 \end{array}
 \begin{array}{cc}
 \cdot & \cdot \\
 \cdot & \cdot \\
 \cdot & \cdot \\
 \cdot & \cdot
 \end{array}$$

Perfect matching  $M$  in  $G \Leftrightarrow$  Circular genome  $M$  ; with  $\omega'(M) = \omega'(M)$

Maximum weight matching  $M$  gives a minimum weight circular breakpoint median  $M$

# Breakpoint median of canonical genomes

Given canonical genomes  $\mathbb{C}_1$ ,  $\mathbb{C}_2$  and  $\mathbb{C}_3$ , find a canonical genome  $\mathbb{M}$  that minimizes the sum:

$$\begin{aligned}
 s_{BP}(\mathbb{M}) &= d_{BP}(\mathbb{M}, \mathbb{C}_1) + d_{BP}(\mathbb{M}, \mathbb{C}_2) + d_{BP}(\mathbb{M}, \mathbb{C}_k) \\
 &= n - \sum_{xy \in \alpha(\mathbb{M})} \phi(xy, \mathbb{C}_1) - \sum_{x \in \gamma(\mathbb{M})} \frac{\phi(x, \mathbb{C}_1)}{2} + n - \sum_{xy \in \alpha(\mathbb{M})} \phi(xy, \mathbb{C}_2) \\
 &\quad - \sum_{x \in \gamma(\mathbb{M})} \frac{\phi(x, \mathbb{C}_2)}{2} + n - \sum_{xy \in \alpha(\mathbb{M})} \phi(xy, \mathbb{C}_3) - \sum_{x \in \gamma(\mathbb{M})} \frac{\phi(x, \mathbb{C}_3)}{2} \\
 &= 3n - \sum_{xy \in \alpha(\mathbb{M})} \phi(xy, \mathbb{C}_{1..3}) - \sum_{x \in \gamma(\mathbb{M})} \frac{\phi(x, \mathbb{C}_{1..3})}{2} \\
 &= 3n - \omega'(\mathbb{M})
 \end{aligned}$$

Now we need to maximize  $\omega'(\mathbb{M}) = \sum_{xy \in \alpha(\mathbb{M})} \omega'(xy) + \sum_{x \in \gamma(\mathbb{M})} \omega'(x)$ ,

where  $\omega'(xy) = \phi(xy, \mathbb{C}_{1..3})$  and  $\omega'(x) = \frac{\phi(x, \mathbb{C}_{1..3})}{2}$ .

1. Build the complete graph  $G$  with vertices  $V(G) = \{g^h : g \in \mathcal{G}_*\} \cup \{g^t : g \in \mathcal{G}_*\}$
2. Assign weights to each edge  $xy$  of  $G$ :  $\omega'(xy) = \phi(xy, \mathbb{C}_{1..k})$ .
3. Build the complete graph  $G_t$  with vertices  $V(G_t) = \{t_{gh} : g \in \mathcal{G}_*\} \cup \{t_{gt} : g \in \mathcal{G}_*\}$
4. Assign weight 0 to each edge of  $G_t$
5. Add one edge connecting each vertex  $x$  in  $G$  to the corresponding vertex  $t_x$  in  $G_t$ , with weight  $\omega'(xt_x) = \frac{\phi(x, \mathbb{C}_{1..k})}{2}$

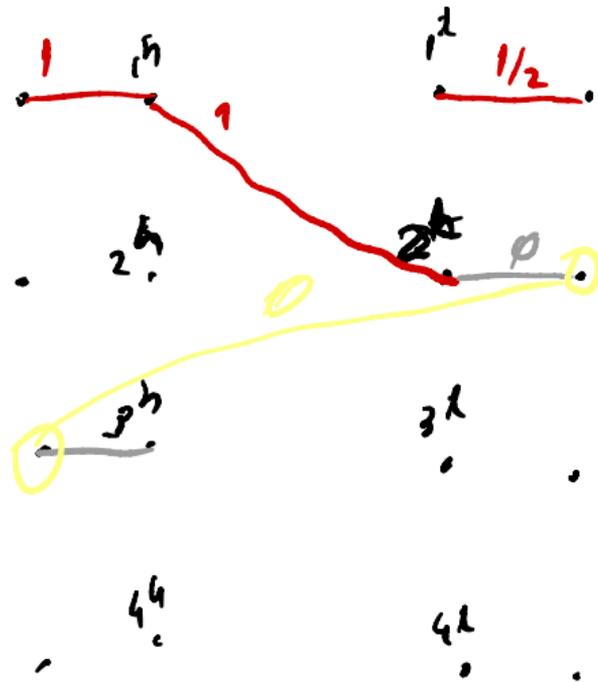
Perfect matching  $M$  in  $G + G_t \Leftrightarrow$  Genome  $\mathbb{M}$  ; with  $\omega'(M) = \omega'(\mathbb{M})$

**Maximum weight matching  $M$  gives a minimum weight breakpoint median  $\mathbb{M}$**

$$C_1 = [1 \cdot 2 \cdot 3 \cdot 4]$$

$$C_2 = [3 \cdot 2 \cdot 1] [4]$$

$$C_3 = [1 \cdot 2 \cdot 4 \cdot 3]$$



Breakpoints } halving  
                  } guided halving

can be computed

in a similar way

# Quiz 3

1 Which of the following statements are true?

- A The SCJ aliquoting problem can be constrained to circular genomes only.
- B The breakpoint median can only be computed for circular genomes.
- C The circular SCJ median is equivalent to the circular breakpoint median of three canonical circular genomes.
- D The breakpoint guided halving is NP-hard.
- E The problem of computing a circular breakpoint halving of a circular duplicated genome is polynomial.

# 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)

On the Complexity of Rearrangement Problems under the Breakpoint Distance

(Jakub Kováč)

JCB volume 21, Number 1 (2014)