

Algorithms in Comparative Genomics

Lecture:

Marília D. V. Braga
Thursdays, 10:15-11:45

Tutorial:

Leonard Bohnenkämper
Fridays, 8:30-10:00

Topics of today:

1. Recall concepts from lecture 01
2. Formal definitions:
 - ▶ genome, family, gene, chromosome
 - ▶ adjacency, telomere
 - ▶ breakpoint model
3. Single-cut-or-join model
4. Computational problems: distance and double distance

Types of genomes - concerning gene content

Singular genome:

each family occurs **exactly once**



Duplicated genome:

each family occurs **exactly twice**



Perfectly duplicated or doubled genome:

duplicated and each adjacency or telomere occurs **exactly twice**



Natural genome:

no restriction on the number of occurrences of families



Definitions / notation (family-based setting)

Given a genome \mathbb{G} :

- ▶ Let $\mathcal{F}(\mathbb{G})$ be the set of gene **families** occurring in \mathbb{G} .
- ▶ For each family $f \in \mathcal{F}(\mathbb{G})$, each **occurrence** of f in \mathbb{G} is called a **gene** and is also represented by f .
- ▶ Each **chromosome** of \mathbb{G} is represented by a sequence of its genes, and each gene f has a sign (f for direct orientation or \bar{f} for reverse orientation) to keep track of the relative orientations.
- ▶ Each **linear** chromosome is flanked by square brackets “[” and “]”, while each **circular** chromosome is flanked by parentheses “(” and “)”.
- ▶ Denote by $\mathcal{G}(\mathbb{G})$ the (multi)set of genes of \mathbb{G} (ignoring orientations).
- ▶ Denote by $\kappa(\mathbb{G})$ the number of linear chromosomes in \mathbb{G} .

Ex1: $\mathbb{G} = [\bar{2} \bar{1} 4 1] (1 3 \bar{2} \bar{5}) \Rightarrow \mathcal{F}(\mathbb{G}) = \{1, 2, 3, 4, 5\}$, $\mathcal{G}(\mathbb{G}) = \{1, 1, 1, 2, 2, 3, 4, 5\}$ and $\kappa(\mathbb{G}) = 1$

Ex2: $\mathbb{S} = [\bar{2} 4 1] [3 \bar{6} \bar{5}] \Rightarrow \mathcal{F}(\mathbb{S}) = \mathcal{G}(\mathbb{S}) = \{1, 2, 3, 4, 5, 6\}$ and $\kappa(\mathbb{S}) = 2$

Obs: If \mathbb{S} is singular, then $\mathcal{F}(\mathbb{S}) = \mathcal{G}(\mathbb{S})$

Ex3: $\mathbb{D} = (\bar{2} 3 1 3 \bar{1} 2) \Rightarrow \mathcal{F}(\mathbb{D}) = \{1, 2, 3\}$, $\mathcal{G}(\mathbb{D}) = \{1, 1, 2, 2, 3, 3\}$ and $\kappa(\mathbb{D}) = 0$

Obs: If \mathbb{D} is duplicated, then $\mathcal{G}(\mathbb{D}) = \mathcal{F}(\mathbb{D}) \cup \mathcal{F}(\mathbb{D})$

Definitions / notation (family-based setting)



Representing \mathbb{G} with sets of adjacencies and telomeres:

- ▶ Each gene (occurrence of a family f) in \mathbb{G} has two extremities: **head** f^h and **tail** f^t
- ▶ If a family f occurs multiple times, we assign **subscripts** to the respective gene extremities to keep track of the correct pairs.
- ▶ Two gene extremities that are next to each other in a chromosome of \mathbb{G} form an **adjacency** of \mathbb{G} .
- ▶ A gene extremity that is at the end of a linear chromosome of \mathbb{G} is called a **telomere** of \mathbb{G} .
- ▶ $\alpha(\mathbb{G})$ is the set of adjacencies in genome \mathbb{G}
- ▶ $\gamma(\mathbb{G})$ is the set of telomeres in genome \mathbb{G}

Ex1: $\mathbb{G} = [\bar{2} \bar{1} 4 1] \quad [1 3 \bar{2} \bar{5}] \Rightarrow \alpha(\mathbb{G}) = \{2_1^t 1_1^h, 1_1^t 4^t, 4^h 1_2^t, 1_3^h 3^t, 3^h 2_2^h, 2_2^t 5^h, 5^t 1_3^t\}$ and $\gamma(\mathbb{G}) = \{2_1^h, 1_2^h\}$

Obs: $|\mathcal{G}(\mathbb{G})| = |\alpha(\mathbb{G})| + \frac{|\gamma(\mathbb{G})|}{2}$

Ex2: $\mathbb{G} = [\bar{2} 4 1] \quad [3 \bar{6} \bar{5}] \Rightarrow \alpha(\mathbb{G}) = \{2^t 4^t, 4^h 1^t, 3^h 6^h, 6^t 5^h\}$ and $\gamma(\mathbb{G}) = \{2^h, 1^h, 3^t, 5^t\}$

Ex3: $\mathbb{G} = (\bar{2} 3 1 3 \bar{1} 2) \Rightarrow \alpha(\mathbb{G}) = \{2_1^t 3_1^t, 3_1^h 1_1^t, 1_1^h 3_2^t, 3_2^h 1_2^h, 1_2^t 2_2^t, 2_2^h 2_1^h\}$ and $\gamma(\mathbb{G}) = \emptyset$

Obs: If \mathbb{G} is circular, then $|\alpha(\mathbb{G})| = |\mathcal{G}(\mathbb{G})|$ and $\gamma(\mathbb{G}) = \emptyset$

Representing a doubled genome

Given a singular genome \mathbb{S} , let $\mathbb{S} \oplus \mathbb{S}$ or $2 \cdot \mathbb{S}$ be a doubled genome, in which each adjacency/telomere of \mathbb{S} appears twice.

Examples:

$$\mathbb{S}_1 = [\bar{2}13] \text{ and } 2 \cdot \mathbb{S}_1 = [\bar{2}13] [\bar{2}13]$$

LINEAR

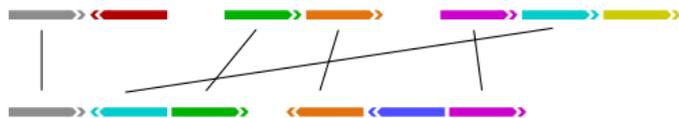
$$\mathbb{S}_2 = (\bar{2}13) \text{ and } 2 \cdot \mathbb{S}_2 = (\bar{2}13) (\bar{2}13) \text{ or } 2 \cdot \mathbb{S}_2 = (\bar{2}13\bar{2}13)$$

CIRCULAR

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

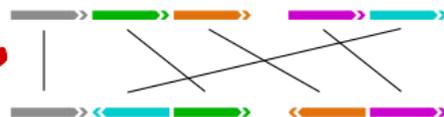


Sing-dup-canonical pair (asymmetric):

one genome is singular and the other is duplicated and the gene families of both genomes are the same



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



Definitions / notation (family-based setting)

Given genomes $\mathbb{G}_1, \mathbb{G}_2, \dots, \mathbb{G}_k$:

- ▶ $\mathcal{F}_* = \mathcal{F}(\mathbb{G}_1) \cap \mathcal{F}(\mathbb{G}_2) \cap \dots \cap \mathcal{F}(\mathbb{G}_k)$ is the set of **common families** (occurring in each \mathbb{G}_i)
- ▶ $\mathcal{G}_* = \mathcal{G}(\mathbb{G}_1) \cap \mathcal{G}(\mathbb{G}_2) \cap \dots \cap \mathcal{G}(\mathbb{G}_k)$ is the (multi)set of **common genes** (genes of each \mathbb{G}_i)
- ▶ $n = |\mathcal{G}_*|$

Genomes	Type	
\mathbb{G}_1 and \mathbb{G}_2	singular	$\mathcal{F}_* = \mathcal{G}_*$
\mathbb{G}_1 and \mathbb{G}_2	balanced	$\mathcal{F}_* = \mathcal{F}(\mathbb{G}_1) = \mathcal{F}(\mathbb{G}_2)$ and $\mathcal{G}_* = \mathcal{G}(\mathbb{G}_1) = \mathcal{G}(\mathbb{G}_2)$
\mathbb{G}_1 and \mathbb{G}_2	canonical	$\mathcal{F}_* = \mathcal{F}(\mathbb{G}_1) = \mathcal{F}(\mathbb{G}_2) = \mathcal{G}_* = \mathcal{G}(\mathbb{G}_1) = \mathcal{G}(\mathbb{G}_2)$
\mathbb{S} and \mathbb{D}	sing-dup-canonical	$\mathcal{F}_* = \mathcal{F}(\mathbb{D}) = \mathcal{F}(\mathbb{S}) = \mathcal{G}_* = \mathcal{G}(\mathbb{S})$ and $\mathcal{G}(\mathbb{D}) = \mathcal{G}(\mathbb{S}) \cup \mathcal{G}(\mathbb{S})$

Representing genomes with sets of adjacencies and telomeres:

- ▶ $\alpha_* = \alpha(\mathbb{G}_1) \cap \alpha(\mathbb{G}_2) \cap \dots \cap \alpha(\mathbb{G}_k)$ is the set of **common adjacencies**
- ▶ $\gamma_* = \gamma(\mathbb{G}_1) \cap \gamma(\mathbb{G}_2) \cap \dots \cap \gamma(\mathbb{G}_k)$ is the set of **common telomeres**
- ▶ $a = |\alpha_*|$ and $t = |\gamma_*|$

The sets α_* and γ_* are easy to identify when $\mathbb{G}_1, \mathbb{G}_2, \dots, \mathbb{G}_k$ are canonical, otherwise identifying them implies fixing a matching of the genes of $\mathbb{G}_1, \mathbb{G}_2, \dots, \mathbb{G}_k$

Quiz 1

Given genomes $\mathbb{G}_1 = (1, 4, \bar{6}, 7)$, $[\bar{5}, 2, 3]$, $\mathbb{G}_2 = [12345]$ and $\mathbb{G}_3 = [\bar{2}, \bar{1}]$, $[\bar{4}, \bar{3}, 5, 2]$:

- 1 Which of the following statements are true?
- A Genomes \mathbb{G}_1 and \mathbb{G}_2 are canonical. **F**
- B** Genomes \mathbb{G}_1 and \mathbb{G}_2 are singular. **T**
- C Genomes \mathbb{G}_2 and \mathbb{G}_3 are canonical. **F**
- D** Genomes \mathbb{G}_2 and \mathbb{G}_3 are natural. **T**
- 2 The sets $\alpha(\mathbb{G}_1)$ and $\gamma(\mathbb{G}_1)$ have sizes ...
- A 5 and 4
- B** 6 and 2
- C 6 and 1
- D 7 and 0
- 3 The sets $\mathcal{F}(\mathbb{G}_3)$ and $\mathcal{G}(\mathbb{G}_3)$ have sizes ...
- A** 5 and 6
- B 5 and 5
- C 4 and 5
- D 6 and 6
- 4 The set $\mathcal{F}_* = \mathcal{F}(\mathbb{G}_1) \cap \mathcal{F}(\mathbb{G}_2) \cap \mathcal{F}(\mathbb{G}_3)$ is equal to ...
- A $\{1, 2, 3, 4, 5, 6, 7\}$
- B $\{1, 2, 2, 3, 4, 5\}$
- C** $\{1, 2, 3, 4, 5\}$
- D $\{1, 2, 3, 4, \bar{5}\}$

Breakpoint model - distance and double distance

Breakpoint distance of canonical genomes \mathcal{C}_1 and \mathcal{C}_2 :

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

where $n = |\mathcal{G}_*|$, $a = |\alpha_*|$ and $t = |\gamma_*|$

The distance $d_{BP}(\mathcal{G}_1, \mathcal{G}_2)$ can be easily computed in linear time.

Breakpoint distance of balanced genomes \mathbb{B}_1 and \mathbb{B}_2 :

$$d_{BP}(\mathbb{B}_1, \mathbb{B}_2) = \min_{(\mathcal{C}_1, \mathcal{C}_2) \in (\mathbb{B}_1, \mathbb{B}_2)} \{d_{BP}(\mathcal{C}_1, \mathcal{C}_2)\}$$

$$\mathbb{B}_1 = [\bar{1} \ 2 \ 3 \ \bar{2}] = \mathcal{C}_1 / \mathcal{C}_1$$
$$\mathbb{B}_2 = [3 \ \bar{1} \ 2 \ \bar{1}] = \mathcal{C}_2 / \mathcal{C}_2$$

Breakpoint double distance of sing-dup-canonical genomes \mathbb{S} and \mathbb{D} :

$$d_{BP}^2(\mathbb{S}, \mathbb{D}) = d_{BI}(2 \cdot \mathbb{S}, \mathbb{D}) = \min_{(\mathcal{C}_1, \mathcal{C}_2) \in (2 \cdot \mathbb{S}, \mathbb{D})} \{d_{BP}(\mathcal{C}_1, \mathcal{C}_2)\}$$

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

The double distance $d_{BP}^2(\mathbb{S}, \mathbb{D})$ can be computed in polynomial time with a greedy approach:

- ▶ There is always a matching of genes that fullfills each candidate common adjacency / telomere between $2 \cdot \mathbb{S}$ and \mathbb{D}

Quiz 2 (from lecture 01)

Given genomes $G_1 = (1,2,3,4) [1\bar{5}\bar{4}5\bar{3}\bar{2}]$, $G_2 = [1,2,3,4,5]$ and $G_3 = [\bar{2},\bar{1}] [\bar{4},\bar{3},5]$
 $[1,2,3,4,5]$

1 Which of the following statements are true?

- A Genome G_1 is linear.
- B Genome G_1 is multichromosomal.
- C Genome G_1 is duplicated.
- D Genome G_1 is doubled.

2 How many families occur in genome G_1 ?

- A 4
- B 5
- C 5,5
- D 6

3 What is the breakpoint distance of G_2 and G_3 ?

- A 1,5
- B 2
- C 2,5
- D 3

4 What is the breakpoint double distance of G_1 and G_2 ?

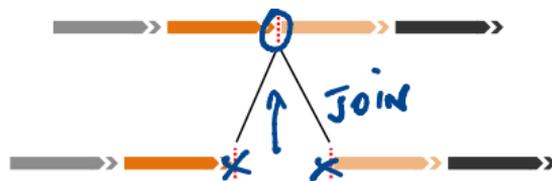
- A 4
- B 4,2
- C 4,5
- D 5

$$d_{BP} = n - a - \frac{k}{2}$$
$$= 5 - 2 - 1 = 2$$

$$d_{BP} = 10 - 5 - \frac{1}{2} = 4,5$$

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: joint two telomeres of \mathbb{G} into one adjacency.
- ▶ Any **single cut** or **single join** is a SCJ.



A genome \mathbb{G} can be represented by its set of adjacencies $\alpha(\mathbb{G})$

(the set of telomeres $\gamma(\mathbb{G})$ can be derived from $\alpha(\mathbb{G})$)

Then, SCJ operations can be seen as set operations:

- ▶ A cut of an adjacency xy : $\alpha(\mathbb{G}) \setminus \{xy\}$.
- ▶ A join of an adjacency xy : $\alpha(\mathbb{G}) \cup \{xy\}$.

SCJ Distance and Sorting

Given canonical genomes \mathbb{C}_1 and \mathbb{C}_2 , how many SCJs do we need to transform \mathbb{C}_1 into \mathbb{C}_2 ?

If I have two sets $\alpha(\mathbb{C}_1)$ and $\alpha(\mathbb{C}_2)$ and the only allowed operation is to remove or include elements from the sets, how can I transform $\alpha(\mathbb{C}_1)$ into $\alpha(\mathbb{C}_2)$ with the minimum number of operations?

One way:

- 1. First, remove all elements of $\alpha(\mathbb{C}_1)$ that are not present in $\alpha(\mathbb{C}_2)$.
- 2. Then, include in $\alpha(\mathbb{C}_1)$ all elements of $\alpha(\mathbb{C}_2)$ that are not already in $\alpha(\mathbb{C}_1)$.

In set theory:

- 1. remove $(\alpha(\mathbb{C}_1) \setminus \alpha(\mathbb{C}_2))$ (SCJ: via single cut operations)
- 2. include $(\alpha(\mathbb{C}_2) \setminus \alpha(\mathbb{C}_1))$ (SCJ: via single join operations)

$$d_{\text{SCJ}}(\mathbb{C}_1, \mathbb{C}_2) = |\alpha(\mathbb{C}_1) \setminus \alpha(\mathbb{C}_2)| + |\alpha(\mathbb{C}_2) \setminus \alpha(\mathbb{C}_1)|$$

SCJ sorting of \mathbb{C}_1 into \mathbb{C}_2

$\alpha(\mathbb{C}_1) =$

$\{1^{h3^h}, 3^{t2^h}, 2^{t4^t}\}$

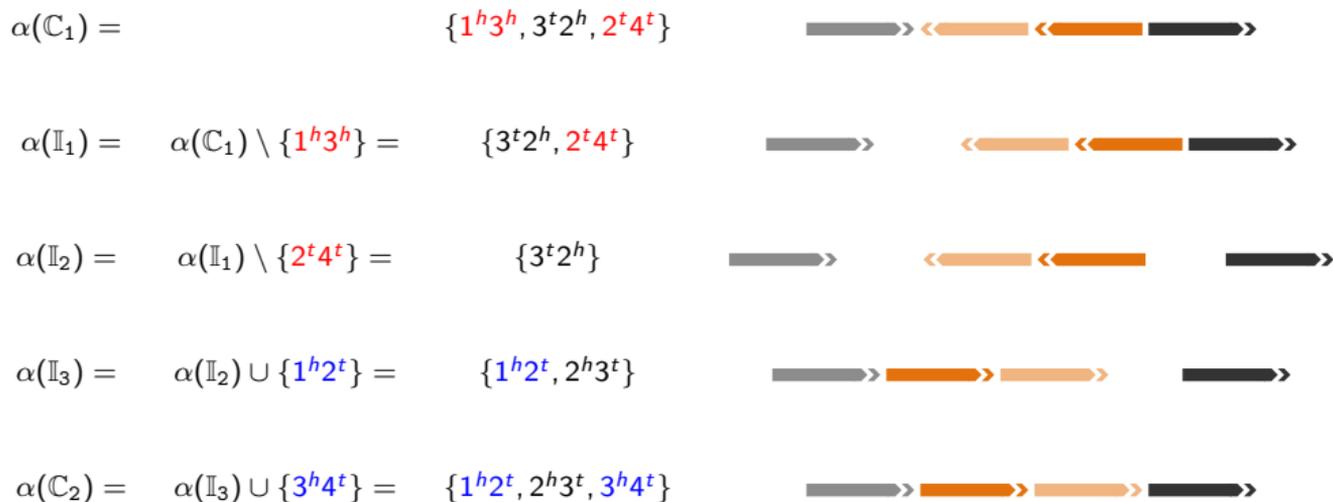


$\alpha(\mathbb{C}_2) =$

$\{1^{h2^t}, 2^{h3^t}, 3^{h4^t}\}$



SCJ sorting of \mathbb{C}_1 into \mathbb{C}_2



SCJ distance of canonical genomes \mathbb{C}_1 and \mathbb{C}_2

$$\begin{aligned}
 d_{\text{SCJ}}(\mathbb{C}_1, \mathbb{C}_2) &= |\alpha(\mathbb{C}_1) \setminus \alpha(\mathbb{C}_2)| + |\alpha(\mathbb{C}_2) \setminus \alpha(\mathbb{C}_1)| \\
 &= |\alpha(\mathbb{C}_1)| - |\alpha(\mathbb{C}_1) \cap \alpha(\mathbb{C}_2)| + |\alpha(\mathbb{C}_2)| - |\alpha(\mathbb{C}_1) \cap \alpha(\mathbb{C}_2)| \\
 &= |\alpha(\mathbb{C}_1)| + |\alpha(\mathbb{C}_2)| - 2|\alpha(\mathbb{C}_1) \cap \alpha(\mathbb{C}_2)|
 \end{aligned}$$

$$= |\alpha(\mathbb{C}_1)| + |\alpha(\mathbb{C}_2)| - 2|\alpha_\star|$$

Since $|\alpha(\mathbb{C}_1)| = n - \frac{|\gamma(\mathbb{C}_1)|}{2}$:

$$\begin{aligned}
 d_{\text{SCJ}}(\mathbb{C}_1, \mathbb{C}_2) &= n - \frac{|\gamma(\mathbb{C}_1)|}{2} + n - \frac{|\gamma(\mathbb{C}_2)|}{2} - 2|\alpha_\star| \\
 &= 2n - 2a - \frac{|\gamma(\mathbb{C}_1)| + |\gamma(\mathbb{C}_2)|}{2} \\
 &= 2n - 2a - \kappa(\mathbb{C}_1) - \kappa(\mathbb{C}_2)
 \end{aligned}$$

where $n = |\mathcal{G}_\star|$ and $a = |\alpha_\star|$

$$|\mathcal{G}(\mathbb{C}_1)| = \frac{|\alpha(\mathbb{C}_1)| + |\gamma(\mathbb{C}_1)|}{2}$$

n

The distance $d_{\text{SCJ}}(\mathbb{C}_1, \mathbb{C}_2)$ can be easily computed in linear time.

Breakpoint distance \times SCJ distance

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

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

$$= 2n - 2a - \kappa(\mathbb{G}_1) - \kappa(\mathbb{G}_2) - t + t$$

$$= 2n - 2a - t - \kappa(\mathbb{G}_1) - \kappa(\mathbb{G}_2) + t$$

$$= 2\left(n - a - \frac{t}{2}\right) - \kappa(\mathbb{G}_1) - \kappa(\mathbb{G}_2) + t$$

$$= 2d_{BP}(\mathbb{G}_1, \mathbb{G}_2) - \kappa(\mathbb{G}_1) - \kappa(\mathbb{G}_2) + t$$

For circular genomes:

$$d_{SCJ}(\mathbb{G}_1, \mathbb{G}_2) = 2d_{BP}(\mathbb{G}_1, \mathbb{G}_2)$$

In general:

$$d_{BP}(\mathbb{G}_1, \mathbb{G}_2) \leq d_{SCJ}(\mathbb{G}_1, \mathbb{G}_2) \leq 2d_{BP}(\mathbb{G}_1, \mathbb{G}_2)$$

SCJ - double distance

SCJ distance of balanced genomes \mathbb{B}_1 and \mathbb{B}_2 :

$$d_{\text{SCJ}}(\mathbb{B}_1, \mathbb{B}_2) = \min_{(C_1, C_2) \in (\mathbb{B}_1, \mathbb{B}_2)} \{d_{\text{SCJ}}(C_1, C_2)\}$$

marking

SCJ double distance of sing-dup-canonical genomes \mathbb{S} and \mathbb{D} :

$$d_{\text{SCJ}}^2(\mathbb{S}, \mathbb{D}) = d_{\text{SCJ}}(2 \cdot \mathbb{S}, \mathbb{D}) = \min_{(C_1, C_2) \in (2 \cdot \mathbb{S}, \mathbb{D})} \{d_{\text{SCJ}}(C_1, C_2)\}$$

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

[Handwritten: $\mathbb{S} = [2 \ 1 \ 3]$ and $\mathbb{D} = [3 \ 1 \ 2 \ 3 \ 1 \ 2]$ with red 'x' marks on the second and fourth elements of \mathbb{D}]

The double distance $d_{\text{SCJ}}^2(\mathbb{S}, \mathbb{D})$ can be computed in polynomial time with a greedy approach:

- ▶ There is always a matching of genes that fullfills each candidate common adjacency between $2 \cdot \mathbb{S}$ and \mathbb{D}



$$d_{\text{SCJ}}^2(\mathbb{S}, \mathbb{D}) = 12 - \frac{2n - 2a - k(\mathbb{S}) - k(\mathbb{D})}{6} - 2 - 1$$

$$= 3$$

(Handwritten calculation with red annotations: 12 - 6 - 2 - 1 = 3)

Quiz 3

Given genomes $\mathbb{G}_1 = (1234) [1\bar{5}\bar{4}5\bar{3}\bar{2}]$, $\mathbb{G}_2 = [12345]$ and $\mathbb{G}_3 = [\bar{2}\bar{1}] [\bar{4}\bar{3}5]$:

1 What is the SCJ distance of \mathbb{G}_2 and \mathbb{G}_3 ?

A 2

B 2,5

C 3

D 4

2 What is the SCJ double distance of \mathbb{G}_1 and \mathbb{G}_2 ?

A 6

B 7

C 7,5

D 8

Other computational problems

Halving

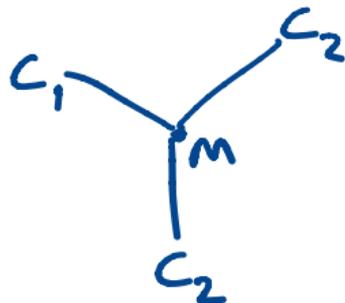
Given a duplicated genome D ,
find a singular genome S that minimizes
the double distance:

$$\underline{d^2(S, D) = d(2 \cdot S, D)}$$

Median

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

$$d(M, C_1) + d(M, C_2) + d(M, C_3)$$



SCJ Median

$$s(M) = d(C_1, M) + d(C_2, M) + d(C_3, M)$$

$$\text{If } xy \in M \text{ and } \begin{cases} xy \notin C_1, C_2, C_3 : \Delta s(M) = +3 \\ xy \in C_1, \notin C_2, C_3 : \Delta s(M) = +1 \\ xy \in C_1, C_2 \notin C_3 : \Delta s(M) = -1 \\ xy \in C_1, C_2, C_3 : \Delta s(M) = -3 \end{cases}$$

Adjacencies with $\Delta s < 0$ are good

SCT Median

$M = \{xy : xy \text{ is present in at least two of the input genomes}\}$

$$C_1 = [1 \cdot 2 \cdot 3 \cdot 4 \cdot 5]$$

$$C_2 = [5 \cdot 1 \cdot 2 \cdot 3 \cdot 4]$$

$$C_3 = [4 \cdot 5 \cdot 1 \cdot 2 \cdot 3]$$

$$M = (1 \cdot 2 \cdot 3 \cdot 4 \cdot 5)$$

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)