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

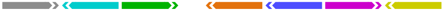
2. Formal definitions:
    - genome, family, gene, chromosome
    - adjacency, telomere
    - breakpoint model

3. Single-cut-or-join model

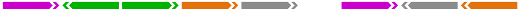
4. Computational problems: distance and double distance

Types of genomes - concerning gene content

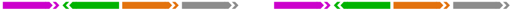
Singular genome:
each family occurs **exactly once**



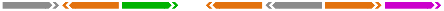
Duplicated genome:
each family occurs **exactly twice**



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: $\mathbb{G} = [\bar{2}\,\bar{1}\,4\,1]\ \ (1\,3\,\bar{2}\,\bar{5}) \Rightarrow \alpha(\mathbb{G}) = \{2^t_1 1^h_1, 1^t_1 4^t, 4^h 1^t_2, 1^h_3 3^t, 3^h 2^h_2, 2^t_2 5^h, 5^t 1^t_3\}$ and $\gamma(\mathbb{G}) = \{2^h_1, 1^h_2\}$

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

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

Ex3: $\mathbb{G} = (\bar{2}\ 3\,1\,3\,\bar{1}\,2) \Rightarrow \alpha(\mathbb{G}) = \{2^t_1 3^t_1, 3^h_1 1^t_1, 1^h_1 3^t_2, 3^h_2 1^h_2, 1^t_2 2^t_2, 2^h_2 2^h_1\}$ 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.

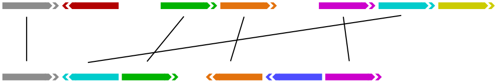
Examples:

$\mathbb{S}_1 = [\bar{2}\,1\,3]$ and $2{\cdot}\mathbb{S}_1 = [\bar{2}\,1\,3]\ \ [\bar{2}\,1\,3]$

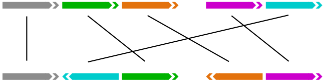
$\mathbb{S}_2 = (\bar{2}\,1\,3)$ and $2{\cdot}\mathbb{S}_2 = (\bar{2}\,1\,3)\ \ (\bar{2}\,1\,3)$ or $2{\cdot}\mathbb{S}_2 = (\bar{2}\,1\,3\,\bar{2}\,1\,3)$

Types of genome pairs/sets

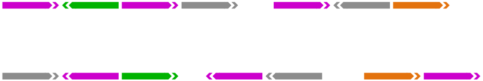
Pair/set of singular genomes:
each family occurs **at most once** in each genome



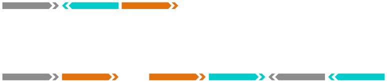
Pair/set of canonical genomes:
singular and **balanced**



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



Definitions / notation (family-based setting)

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

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

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

Representing genomes with sets of adjacencies and telomeres:

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

The sets $\alpha_\star$ and $\gamma_\star$ are easy to identify when $\mathbb{G}_1, \mathbb{G}_2, \ldots, \mathbb{G}_k$ are canonical,
otherwise identifying them implies fixing a matching of the genes of $\mathbb{G}_1, \mathbb{G}_2, \ldots, \mathbb{G}_k$

Breakpoint model - distance and double distance

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

$$\mathrm{d}_{\mathrm{BP}}(\mathbb{C}_1,\mathbb{C}_2) = n - a - \tfrac{t}{2}\,,$$
$$\text{where } n = |\mathcal{G}_\star|,\ a = |\alpha_\star| \text{ and } t = |\gamma_\star|$$

The distance $\mathrm{d}_{\mathrm{BP}}(\mathbb{G}_1,\mathbb{G}_2)$ can be easily computed in linear time.

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

$$\mathrm{d}_{\mathrm{BP}}(\mathbb{B}_1,\mathbb{B}_2) = \min_{(\mathbb{C}_1,\mathbb{C}_2)\in(\mathbb{B}_1,\mathbb{B}_2)} \{\mathrm{d}_{\mathrm{BP}}(\mathbb{C}_1,\mathbb{C}_2)\}$$

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

$$\mathrm{d}^2_{\mathrm{BP}}(\mathbb{S},\mathbb{D}) = \mathrm{d}_{\mathrm{BP}}(2\cdot\mathbb{S},\mathbb{D}) = \min_{(\mathbb{C}_1,\mathbb{C}_2)\in(2\cdot\mathbb{S},\mathbb{D})} \{\mathrm{d}_{\mathrm{BP}}(\mathbb{C}_1,\mathbb{C}_2)\}$$

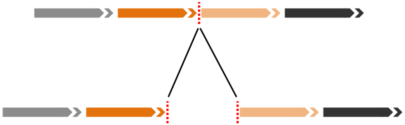
Ex: $\mathbb{S} = [\bar{2}\,1\,\bar{3}]$ and $\mathbb{D} = [3\,\bar{1}\,\bar{2}\,3\,\bar{1}\,2]$

The double distance $\mathrm{d}^2_{\mathrm{BP}}(\mathbb{S},\mathbb{D})$ 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 $\mathbb{D}$

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 $\mathbb{C}_2$, dow many SCJs do we need to tranform $\mathbb{C}_1$ 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 $\alpha(\mathbb{C}_1)$.

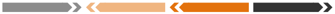
In set theory:

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)

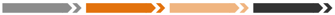
$$\mathsf{d}_{\mathrm{SCJ}}(\mathbb{C}_1, \mathbb{C}_2) \quad = \quad |\alpha(\mathbb{C}_1) \setminus \alpha(\mathbb{C}_2)| \quad + \quad |\alpha(\mathbb{C}_2) \setminus \alpha(\mathbb{C}_1)|$$

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

$\alpha(\mathbb{C}_1) =$ $\{1^h3^h, 3^t2^h, 2^t4^t\}$

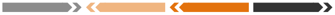


$\alpha(\mathbb{C}_2) =$ $\{1^h2^t, 2^h3^t, 3^h4^t\}$

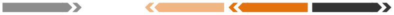


SCJ sorting of $\mathbb{C}_1$ into $\mathbb{C}_2$

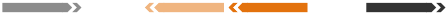
$\alpha(\mathbb{C}_1) = \{1^h3^h, 3^t2^h, 2^t4^t\}$



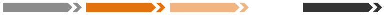
$\alpha(\mathbb{I}_1) = \alpha(\mathbb{C}_1) \setminus \{1^h3^h\} = \{3^t2^h, 2^t4^t\}$



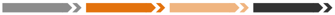
$\alpha(\mathbb{I}_2) = \alpha(\mathbb{I}_1) \setminus \{2^t4^t\} = \{3^t2^h\}$



$\alpha(\mathbb{I}_3) = \alpha(\mathbb{I}_2) \cup \{1^h2^t\} = \{1^h2^t, 2^h3^t\}$



$\alpha(\mathbb{C}_2) = \alpha(\mathbb{I}_3) \cup \{3^h4^t\} = \{1^h2^t, 2^h3^t, 3^h4^t\}$



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

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

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

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

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

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

Breakpoint distance × SCJ distance

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

$$\begin{aligned}
\mathrm{d}_{\mathrm{SCJ}}(\mathbb{G}_1, \mathbb{G}_2) &= 2n - 2a - \kappa(\mathbb{G}_1) - \kappa(\mathbb{G}_2) \\
&= 2n - 2a - \kappa(\mathbb{G}_1) - \kappa(\mathbb{G}_2) - t + t \\
&= 2n - 2a - t - \kappa(\mathbb{G}_1) - \kappa(\mathbb{G}_2) + t \\
&= 2(n - a - \frac{t}{2}) - \kappa(\mathbb{G}_1) - \kappa(\mathbb{G}_2) + t \\
&= 2\mathrm{d}_{\mathrm{BP}}(\mathbb{G}_1, \mathbb{G}_2) - \kappa(\mathbb{G}_1) - \kappa(\mathbb{G}_2) + t
\end{aligned}$$

For circular genomes:

$$\mathrm{d}_{\mathrm{SCJ}}(\mathbb{G}_1, \mathbb{G}_2) = 2\mathrm{d}_{\mathrm{BP}}(\mathbb{G}_1, \mathbb{G}_2)$$

In general:

$$\mathrm{d}_{\mathrm{BP}}(\mathbb{G}_1, \mathbb{G}_2) \leq \mathrm{d}_{\mathrm{SCJ}}(\mathbb{G}_1, \mathbb{G}_2) \leq 2\mathrm{d}_{\mathrm{BP}}(\mathbb{G}_1, \mathbb{G}_2)$$

SCJ - double distance

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

$$\mathsf{d}_{\text{SCJ}}(\mathbb{B}_1, \mathbb{B}_2) = \min_{(\mathbb{C}_1,\mathbb{C}_2)\in(\mathbb{B}_1,\mathbb{B}_2)} \{\mathsf{d}_{\text{SCJ}}(\mathbb{C}_1, \mathbb{C}_2)\}$$

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

$$\mathsf{d}^2_{\text{SCJ}}(\mathbb{S}, \mathbb{D}) = \mathsf{d}_{\text{SCJ}}(2\cdot\mathbb{S}, \mathbb{D}) = \min_{(\mathbb{C}_1,\mathbb{C}_2)\in(2\cdot\mathbb{S},\mathbb{D})} \{\mathsf{d}_{\text{SCJ}}(\mathbb{C}_1, \mathbb{C}_2)\}$$

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

The double distance $\mathsf{d}^2_{\text{SCJ}}(\mathbb{S}, \mathbb{D})$ 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 $\mathbb{D}$

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)