Topics of today:

Overview of studied models/problems

NP-hard problems:

1. Decomposing the cropped breakpoint graph of *unsigned* canonical genomes
2. DCJ median problem
3. DCJ double distance
4. DCJ distance of balanced genomes

Overview of models / computational problems - 1995-2020

Edit operations

—— Model ——		Canonical distance	Double distance	Halving	Guided Halving	Median	Balanced distance
Break point	Multi mixed/circular	**P**	**P**	**P**	**P**	**P**	NP?
	Multi linear	**P**	**P**	NP	NP	NP	NP?
	Uni linear/circular	**P**	(open)	(NP)	(NP)	**NP**	NP
SCJ	Multi mixed	**P**	**P**	**P**	**P**	**P**	?
	Multi linear	**P**	**P**	**P**	**P**	**P**	?
	(Multi circular - initial and target)	(**P**)	(**P**)	(**P**)	(**P**)	(**P**)	(?)
	(Uni linear/circular - initial and target)	(**P**)	(open)	(open)	(open)	(open)	(?)
DCJ	Multi mixed/circular	**P**	**NP**	**P**	NP	**NP**	**NP** (ILP)
	Restricted multi linear	**P**	open	open	NP?	NP?	NP?
	Uni linear/circular (**Inversion**)	**P**	open	P	NP?	NP	NP?
	Strict multi linear (**Inv/Trsl/Fus/Fis**)	P	open	open	NP?	NP?	NP?

—— Model ——		Singular genomes	Natural genomes	Family-free genomes
DCJ-indel distance	Multi mixed/circular Restricted multi linear	**P**	**NP** (ILP)	**NP** (ILP)
	Uni linear/circular (**Inversion**)	P	NP?	NP?

previous lectures

this and next lectures

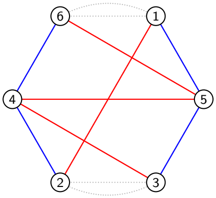
Cropped breakpoint graph of two unsigned canonical chromosomes

Each vertex of a cropped breakpoint graph has degree 0, 2 or 4:

Unsigned canonical circular chromosomes

$\widehat{\mathbb{A}} = (\,1\ 5\ 3\ 2\ 4\ 6\,)$

$\widehat{\mathbb{B}} = (\,1\ 2\ 3\ 4\ 5\ 6\,)$

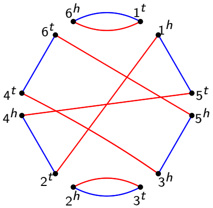


NP-hard problem:

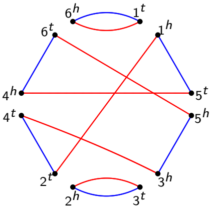
decompose a cropped breakpoint graph into the maximum number of edge-disjoint even cycles alternating colors

⇒ Inversion distance of unsigned chromosomes is NP-hard

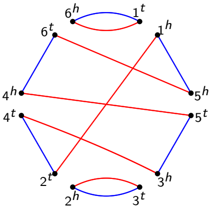
Corresponding breakpoint diagrams of signed canonical chromosomes:



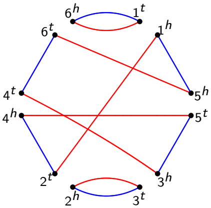
$\mathbb{A}_1 = (\,1\ 5\ \bar{3}\ \bar{2}\ \bar{4}\ 6\,)$



$\mathbb{A}_2 = (\,1\ 5\ \bar{3}\ \bar{2}\ 4\ 6\,)$



$\mathbb{A}_3 = (\,1\ \bar{5}\ \bar{3}\ \bar{2}\ 4\ 6\,)$

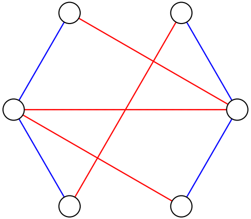


$\mathbb{A}_4 = (\,1\ \bar{5}\ \bar{3}\ 2\ \bar{4}\ 6\,)$

Balanced bicolored graph decomposition ($\textsc{BGDec}$)

Each vertex of a balanced bicolored graph has degree 0, 2 or 4

The number of red and of blue edges inciding in each vertex is identical



Problem:

Entirely decompose a balanced bicolored graph
into the maximum number of edge-disjoint
alternating even cycles

$\Downarrow$

NP-hard

DCJ median of three canonical genomes

Given three canonical genomes $\mathbb{A}$, $\mathbb{B}$, $\mathbb{C}$, find another canonical genome $\mathbb{M}$ that minimizes the sum

$$\mathsf{d}_{\mathrm{DCJ}}(\mathbb{M}, \mathbb{A}) + \mathsf{d}_{\mathrm{DCJ}}(\mathbb{M}, \mathbb{B}) + \mathsf{d}_{\mathrm{DCJ}}(\mathbb{M}, \mathbb{C})$$

Example:

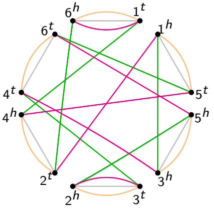
Genomes

$\mathbb{A} = (\,1\ 5\ \bar{3}\ \bar{2}\ \bar{4}\ 6\,)$

$\mathbb{B} = (\,1\ \bar{3}\ 4\,)\ \ (\,2\ \bar{5}\ 6\,)$

$\mathbb{C} = (\,1\ 2\ 3\ 4\ 5\ 6\,)$

Breakpoint graph of $\mathbb{A}$, $\mathbb{B}$ and $\mathbb{C}$



Median candidate

$\mathbb{M} = \mathbb{A} = (\,1\ 5\ \bar{3}\ \bar{2}\ \bar{4}\ 6\,)$

$\mathsf{d}_{\mathrm{DCJ}}(\mathbb{M}, \mathbb{A}) = 0$

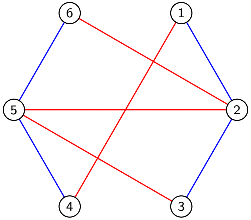
$\mathsf{d}_{\mathrm{DCJ}}(\mathbb{M}, \mathbb{B}) = 6 - \overset{2}{\cancel{4}} = \overset{4}{\cancel{2}}$

$\mathsf{d}_{\mathrm{DCJ}}(\mathbb{M}, \mathbb{C}) = 6 - \underset{4}{\cancel{2}} = \underset{2}{\cancel{4}}$

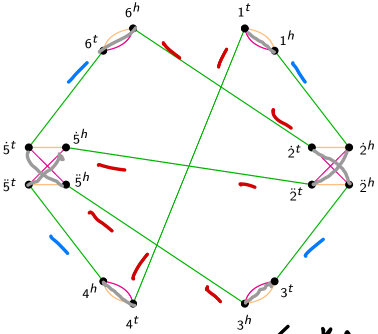
Reducing BGDEC to the DCJ median of three canonical genomes

A, B, C

$n = w_2 + 2w_4$



⇒



⇑

w_2 : # vertices with degree 2

w_4 : # vertices with degree 4

$d(M,A) = \cancel{w_2} + 2w_4 - \cancel{w}_2 - K_A$

$d(M,B) = \cancel{w_2} + 2w_4 - \cancel{w}_2 - K_B$

$K_A + K_B = 3w_4$

$d(M,C) = w_2 + 2w_4 - K_C$

$\sum_{G=\{A,B,C\}} d(M,G) = w_2 + 6w_4 - 3w_4 - K_C = w_2 + 3w_4 - K_C$

DCJ double distance

DCJ double distance $\mathsf{d}^2_{\mathrm{DCJ}}(\mathbb{S},\mathbb{D})$ of sing-dup-canonical genomes $\mathbb{S}$ and $\mathbb{D}$:

$$\mathsf{d}^2_{\mathrm{DCJ}}(\mathbb{S},\mathbb{D}) = \mathsf{d}_{\mathrm{DCJ}}(2{\cdot}\mathbb{S},\mathbb{D})$$

Transforming $2{\cdot}\mathbb{S}$ and $\mathbb{D}$ into **matched** canonical genomes $\mathbb{C}_1$ and $\mathbb{C}_2$:

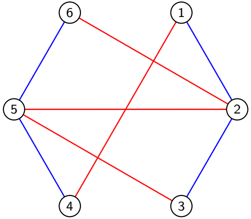
for each family $f \in \mathcal{F}_\star$, determine which occurrence of f in $2{\cdot}\mathbb{S}$ matches each occurrence of f in $\mathbb{D}$

$\Rightarrow$ Matched occurrences receive the same **index** in $\mathbb{C}_1$ and in $\mathbb{C}_2$

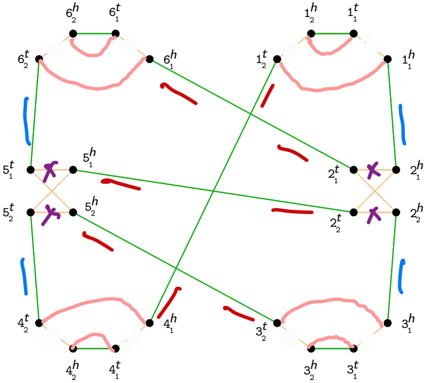
$\mathfrak{C}$: set of all possible pairs of matched canonical genomes obtained from duplicated genomes $2{\cdot}\mathbb{S}$ and $\mathbb{D}$

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

Reducing BGDEC to the DCJ double distance



⇒



GENOMES :

2·S

D

Quiz 1

1 Which of the following statements are true?

~~A~~ The multi mixed/circular DCJ double distance is NP-hard, therefore the multi mixed/circular DCJ halving is also NP-hard.

Halving is P

~~B~~ The multi linear breakpoint double distance is polynomial, therefore the multi linear breakpoint halving is also polynomial.

Multi linear BP halving is NP

[✓ C] The inversion-indel distance can be computed in polynomial time.

2 We prove that DCJ median is NP-hard...

A ... by reducing it to the bicolored graph decomposition.

[✓ B] ... by reducing the bicolored graph decomposition to it.

DCJ distance of balanced genomes

$$\text{Balanced genomes } \mathbb{A} \text{ and } \mathbb{B} \begin{cases} \mathcal{F}_\star = \mathcal{F}(\mathbb{A}) = \mathcal{F}(\mathbb{B}) \\ \mathcal{G}_\star = \mathcal{G}(\mathbb{A}) = \mathcal{G}(\mathbb{B}) \\ \text{for each family } f \in \mathcal{F}_\star,\ \Phi(f, \mathbb{A}) = \Phi(f, \mathbb{B}) \end{cases}$$

Transforming $\mathbb{A}$ and $\mathbb{B}$ into **matched** canonical genomes $\mathbb{A}^\ddagger$ and $\mathbb{B}^\ddagger$:

for each family $f \in \mathcal{F}_\star$, determine which occurrence of f in $\mathbb{A}$ matches each occurrence of f in $\mathbb{B}$

$\Rightarrow$ Matched occurrences receive the same **index** in $\mathbb{A}^\ddagger$ and in $\mathbb{B}^\ddagger$

The number of common genes between any pair of matched genomes $\mathbb{A}^\ddagger$ and $\mathbb{B}^\ddagger$ is $n_* = |\mathcal{G}_\star|$

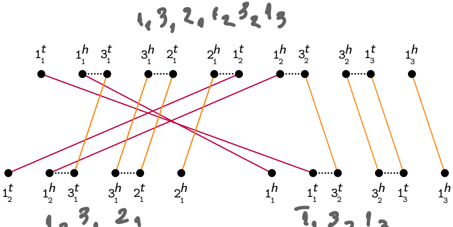
$\mathfrak{M}$: set of all possible pairs of matched canonical genomes obtained from balanced genomes $\mathbb{A}$ and $\mathbb{B}$

DCJ distance of $\mathbb{A}$ and $\mathbb{B}$:

$$d_{\text{DCJ}}(\mathbb{A}, \mathbb{B}) = \min_{(\mathbb{A}^\ddagger, \mathbb{B}^\ddagger) \in \mathfrak{M}} \{ d_{\text{DCJ}}(\mathbb{A}^\ddagger, \mathbb{B}^\ddagger) \}$$

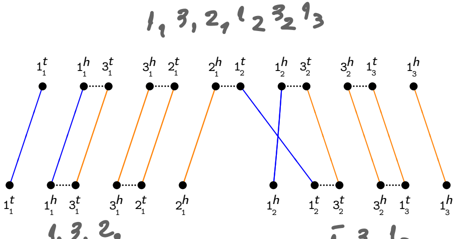
Multi-relational graph $MRG(\mathbb{A}, \mathbb{B})$

Example: $\mathbb{A} = [1\,3\,2\,1\,3\,1]$ and $\mathbb{B} = [1\,3\,2]\ \ [\bar{1}\,3\,1]$


$1_1\,3_1\,2_1\,1_2\,3_2\,1_3$
1_1^t 1_1^h 3_1^t 3_1^h 2_1^t 2_1^h 1_2^t 1_2^h 3_2^t 3_2^h 1_3^t 1_3^h
1_2^t 1_2^h 3_1^t 3_1^h 2_1^t 2_1^h 1_1^h 1_1^t 3_2^t 3_2^h 1_3^t 1_3^h

$1_2\,3_1\,2_1$ $\bar{1}_1\,3_2\,1_3$

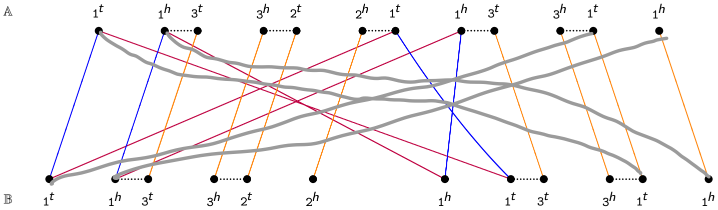
$d_{\text{DCJ}} = 6 - 2 - \frac{2}{2} = 3$


$1_1\,3_1\,2_1\,1_2\,3_2\,1_3$
1_1^t 1_1^h 3_1^t 3_1^h 2_1^t 2_1^h 1_2^t 1_2^h 3_2^t 3_2^h 1_3^t 1_3^h
1_1^t 1_1^h 3_1^t 3_1^h 2_1^t 2_1^h 1_2^h 1_2^t 3_2^t 3_2^h 1_3^t 1_3^h

$1_1\,3_1\,2_1$ $\bar{1}_2\,3_2\,1_3$

$d_{\text{DCJ}} = 6 - 3 - \frac{2}{2} = 2$

MRG :


$\mathbb{A}$
1^t 1^h 3^t 3^h 2^t 2^h 1^t 1^h 3^t 3^h 1^t 1^h
$\mathbb{B}$
1^t 1^h 3^t 3^h 2^t 2^h 1^h 1^t 3^t 3^h 1^t 1^h

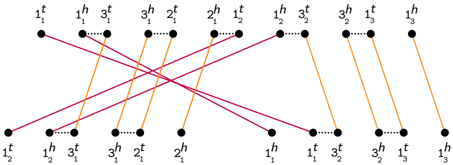
$\phi(1) = 3$

$\phi(2) = 1$

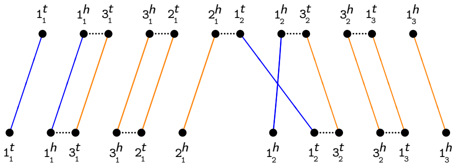
$\phi(3) = 2$

Multi-relational graph $MRG(\mathbb{A}, \mathbb{B})$

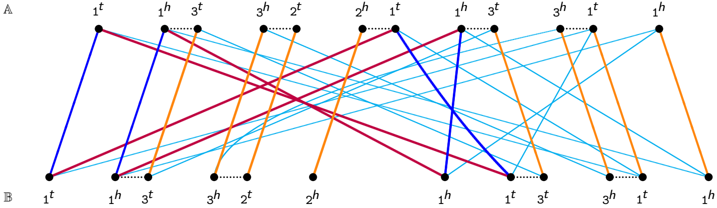
Example: $\mathbb{A} = [1\,3\,2\,1\,3\,1]$ and $\mathbb{B} = [1\,3\,2]\ \ [\bar{1}\,3\,1]$


1_1^t 1_1^h 3_1^t 3_1^h 2_1^t 2_1^h 1_2^t 1_2^h 3_2^t 3_2^h 1_3^t 1_3^h
1_2^t 1_2^h 3_1^t 3_1^h 2_1^t 2_1^h 1_1^h 1_1^t 3_2^t 3_2^h 1_3^t 1_3^h

$\mathrm{d}_{\mathrm{DCJ}} = 6 - 2 - \frac{2}{2} = 3$


1_1^t 1_1^h 3_1^t 3_1^h 2_1^t 2_1^h 1_2^t 1_2^h 3_2^t 3_2^h 1_3^t 1_3^h
1_1^t 1_1^h 3_1^t 3_1^h 2_1^t 2_1^h 1_2^h 1_2^t 3_2^t 3_2^h 1_3^t 1_3^h

$\mathrm{d}_{\mathrm{DCJ}} = 6 - 3 - \frac{2}{2} = 2$


$\mathbb{A}$
1^t 1^h 3^t 3^h 2^t 2^h 1^t 1^h 3^t 3^h 1^t 1^h
$\mathbb{B}$
1^t 1^h 3^t 3^h 2^t 2^h 1^h 1^t 3^t 3^h 1^t 1^h

Consistent decompositions of $MRG(\mathbb{A}, \mathbb{B})$

S: Sibling set: matching of extremity edges

$D[S]$: decomposition induced by a $\underline{\underline{\text{maximal}}}$ sibling set S

+ all adjacency edges

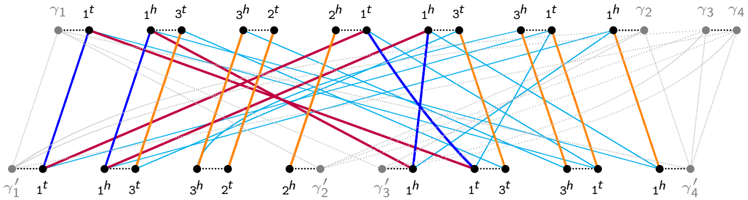
$$D[S] : C_D + P_D \quad \Rightarrow \quad d_{DCJ}(D[S]) : n - C_D - \frac{P_D}{2}$$

$$d_{DCJ}(\mathbb{A}, \mathbb{B}) = \min_{S \in S_{MAX}} \left\{ d_{DCJ}(D[S]) \right\}$$

Capped multi-relational graph $CMRG(\mathbb{A}, \mathbb{B})$

Example: $\mathbb{A} = [1\,3\,2\,1\,3\,1]$ and $\mathbb{B} = [1\,3\,2]\ \ [\bar{1}\,3\,1]$, $p_* = \max\{\kappa(\mathbb{A}), \kappa(\mathbb{B})\} = 2$

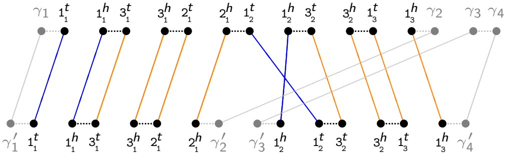
Add $2p_*$ cap extremities to each genome


γ_1 1^t 1^h 3^t 3^h 2^t 2^h 1^t 1^h 3^t 3^h 1^t 1^h γ_2 γ_3 γ_4
γ'_1 1^t 1^h 3^t 3^h 2^t 2^h γ'_2 γ'_3 1^h 1^t 3^t 3^h 1^t 1^h γ'_4

$CMRG(\mathbb{A}, \mathbb{B})$ includes all possible cappings of each maximal sibling-set

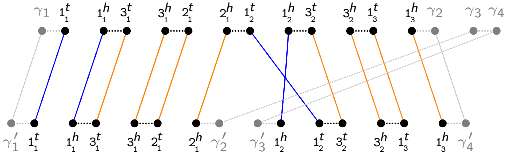
Two distinct cappings of the maximal sibling-set composed of blue + orange edges:

Non-optimal capping


γ_1 1^t_1 1^h_1 3^t_1 3^h_1 2^t_1 2^h_1 1^t_2 1^h_2 3^t_2 3^h_2 1^t_3 1^h_3 γ_2 γ_3 γ_4
γ'_1 1^t_1 1^h_1 3^t_1 3^h_1 2^t_1 2^h_1 γ'_2 γ'_3 1^h_2 1^t_2 3^t_2 3^h_2 1^t_3 1^h_3 γ'_4

$\mathrm{d}_{\mathrm{DCJ}} = n + p_* - c = 6 + 2 - 5 = 3$

Optimal capping


γ_1 1^t_1 1^h_1 3^t_1 3^h_1 2^t_1 2^h_1 1^t_2 1^h_2 3^t_2 3^h_2 1^t_3 1^h_3 γ_2 γ_3 γ_4
γ'_1 1^t_1 1^h_1 3^t_1 3^h_1 2^t_1 2^h_1 γ'_2 γ'_3 1^h_2 1^t_2 3^t_2 3^h_2 1^t_3 1^h_3 γ'_4

$\mathrm{d}_{\mathrm{DCJ}} = n + p_* - c = 6 + 2 - 6 = 2$

Consistent decompositions of $CMRG(\mathbb{A}, \mathbb{B})$

E_0 = set of cap extremity edges

P : capping set : matchings of edges in E_0

$D[S, P]$: decomposition induced by a maximal sibling set and a maximal capping set

+ all adjacency edges

$$d_{DCJ}(A, B) = \min_{\substack{S \in S_{MAX} \\ P \in P_{MAX}}} \{ d_{DCJ}(D[S, P]) \}$$

Quiz 2

1 Which of the following statements are true?

~~A~~ The multi-relational graph is a collection of paths and cycles.

[✓ B] A consistent decomposition of the multi-relational graph is a collection of paths and cycles.

[✓ C] There is a bijection between consistent decompositions of $MRG(\mathbb{A}, \mathbb{B})$ and pairs of matched canonical genomes.

2 Given that $\Phi(f, \mathbb{A}, \mathbb{B})$ is the number of occurrences of family f in $\mathbb{A}$ and in $\mathbb{B}$, the number of pairs of matched canonical genomes derived from balanced genomes $\mathbb{A}$ and $\mathbb{B}$ is...

[✓ A] $\prod_{f \in \mathcal{F}_\star} \Phi(f, \mathbb{A}, \mathbb{B})!$

B $2 \sum_{f \in \mathcal{F}_\star} \Phi(f, \mathbb{A}, \mathbb{B})!$

3 The number of distinct caping sets is

A $2p_*$

[✓ B] $(2p_*)!$

C $(2p_*)^2$

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)

An Exact Algorithm to Compute the Double-Cut- and-Join Distance for Genomes with Duplicate Genes

(Mingfu Shao, Yu Lin, and Bernard M. E. Moret)

JCB, vol. 22, no. 5, pp 425–435 (2015)