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

—— 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	**P**	**NP** (ILP)	**NP** (ILP)
	Restricted multi linear			
	Uni linear/circular (**Inversion**)	P	NP?	NP?

previous lectures

this and next lectures

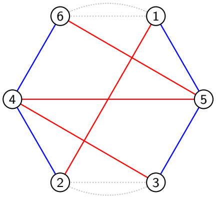
Cropped breakpoint graph of two unsigned canonical chromosomes

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

Unsigned canonical circular chromosomes

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

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

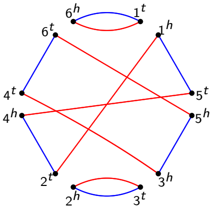


NP-hard problem:

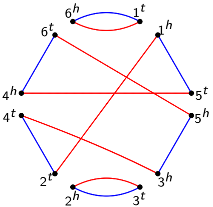
decompose a cropped breakpoint graph into the maximum number of edge-disjoint even cycles alternating colors

⇒ Inversion distance of unsigned chromosomes is NP-hard

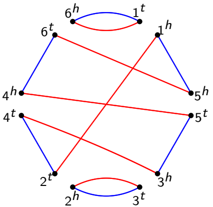
Corresponding breakpoint diagrams of signed canonical chromosomes:



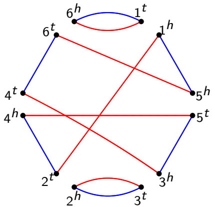
$\mathbb{A}_1 = (\,1\ 5\ \bar{3}\ \bar{2}\ \bar{4}\ 6\,)$



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



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

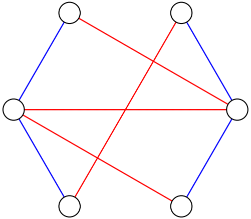


$\mathbb{A}_4 = (\,1\ \bar{5}\ \bar{3}\ \bar{2}\ \bar{4}\ 6\,)$

Balanced bicolored graph decomposition (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

⇓

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:

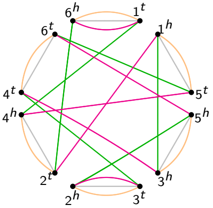
Genomes

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

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

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

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



Median candidate

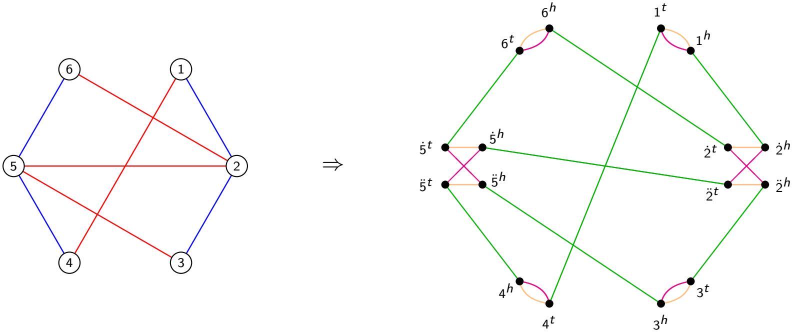
$\mathbb{M} = \mathbb{A} = (\ 1\ 5\ \bar{3}\ \bar{2}\ \bar{4}\ 6\ )$

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

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

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

Reducing BGDec to the DCJ median of three canonical genomes



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 $\mathbb{C}_2$:

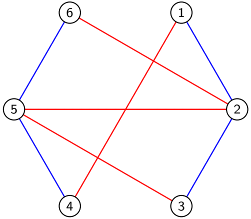
for each family $f \in \mathcal{F}_\star$, determine which occurrence of f in $2\cdot\mathbb{S}$ matches each occurrence of f in $\mathbb{D}$

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

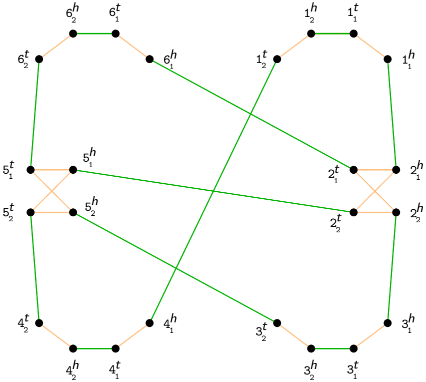
$\mathfrak{C}$: set of all possible pairs of matched canonical genomes obtained from duplicated genomes $2\cdot\mathbb{S}$ and $\mathbb{D}$

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

Reducing BGDEC to the DCJ double distance



⇒



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 $n_* = |\mathcal{G}_\star|$

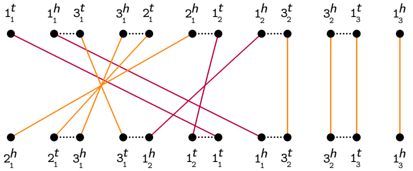
$\mathfrak{M}$: set of all possible pairs of matched canonical genomes obtained from balanced genomes $\mathbb{A}$ and $\mathbb{B}$

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

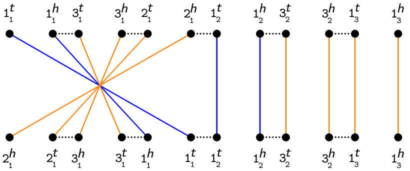
$$\mathrm{d}_{\mathrm{DCJ}}(\mathbb{A}, \mathbb{B}) = \min_{(\mathbb{A}^\ddagger, \mathbb{B}^\ddagger) \in \mathfrak{M}} \{\mathrm{d}_{\mathrm{DCJ}}(\mathbb{A}^\ddagger, \mathbb{B}^\ddagger)\}$$

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

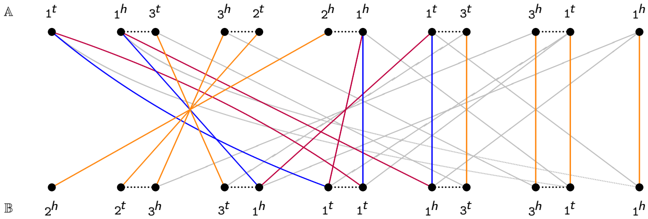
Example: $\mathbb{A} = [1\,3\,2\,1\,3\,1]$ and $\mathbb{B} = [\overline{2}\,\overline{3}\,\overline{1}\,1\,3\,1]$



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



$$\mathrm{d}_{\mathrm{DCJ}} = 6 - 4 - \frac{2}{2} = 2$$



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)