

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](#)

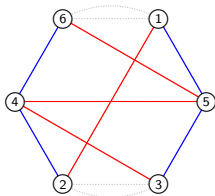
Cropped breakpoint graph of two unsigned canonical chromosomes

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

Unsigned canonical circular chromosomes

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

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

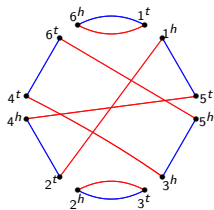


NP-hard problem:

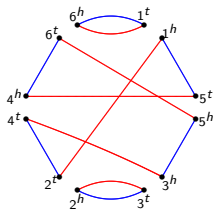
decompose a cropped breakpoint graph into the maximum number of edge-disjoint even cycles alternating colors

⇒ Inversion distance of unsigned chromosomes is NP-hard

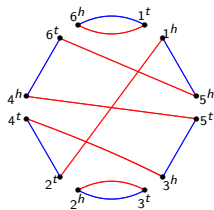
Corresponding breakpoint diagrams of signed canonical chromosomes:



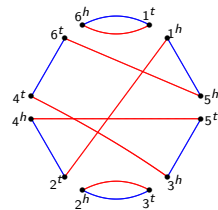
$$\mathbb{A}_1 = (1\ 5\ \bar{3}\ \bar{2}\ \bar{4}\ 6)$$



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



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

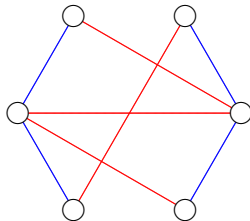


$$\mathbb{A}_4 = (1\ \bar{5}\ \bar{3}\ \bar{2}\ \bar{4}\ 6)$$

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 incident 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

$$d_{\text{DCJ}}(\mathbb{M}, \mathbb{A}) + d_{\text{DCJ}}(\mathbb{M}, \mathbb{B}) + d_{\text{DCJ}}(\mathbb{M}, \mathbb{C})$$

Example:

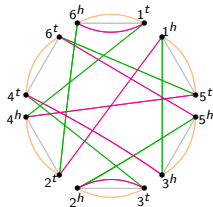
Genomes

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

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

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

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



Median candidate

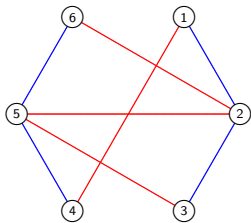
$$\mathbb{M} = \mathbb{A} = (1\ 5\ \bar{3}\ \bar{2}\ \bar{4}\ 6)$$

$$d_{\text{DCJ}}(\mathbb{M}, \mathbb{A}) = 0$$

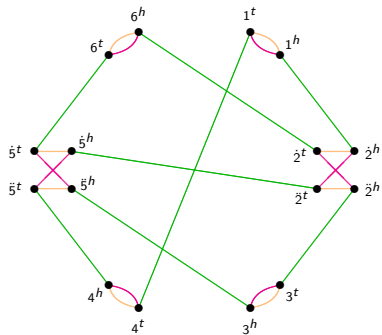
$$d_{\text{DCJ}}(\mathbb{M}, \mathbb{B}) = 6 - 4 = 2$$

$$d_{\text{DCJ}}(\mathbb{M}, \mathbb{C}) = 6 - 2 = 4$$

Reducing BGDEC to the DCJ median of three canonical genomes



⇒



DCJ double distance

DCJ double distance $d_{\text{DCJ}}^2(\mathbb{S}, \mathbb{D})$ of sing-dup-canonical genomes \mathbb{S} and \mathbb{D} :

$$d_{\text{DCJ}}^2(\mathbb{S}, \mathbb{D}) = d_{\text{DCJ}}(2 \cdot \mathbb{S}, \mathbb{D})$$

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

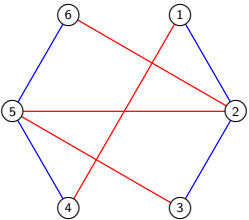
for each family $f \in \mathcal{F}_*$, determine which occurrence of f in $2 \cdot \mathbb{S}$ matches each occurrence of f in \mathbb{D}

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

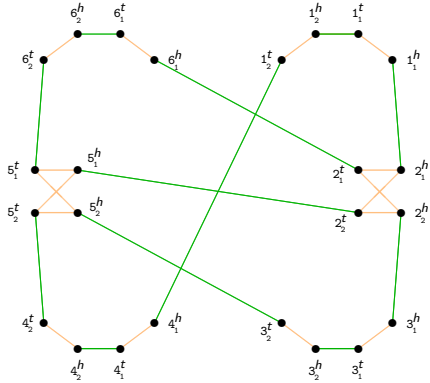
\mathcal{C} : set of all possible pairs of matched canonical genomes obtained from duplicated genomes $2 \cdot \mathbb{S}$ and \mathbb{D}

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

Reducing BGDEC to the DCJ double distance



⇒



DCJ distance of balanced genomes

$$\text{Balanced genomes } \mathbb{A} \text{ and } \mathbb{B} \left\{ \begin{array}{l} \mathcal{F}_* = \mathcal{F}(\mathbb{A}) = \mathcal{F}(\mathbb{B}) \\ \mathcal{G}_* = \mathcal{G}(\mathbb{A}) = \mathcal{G}(\mathbb{B}) \\ \text{for each family } f \in \mathcal{F}_*, \Phi(f, \mathbb{A}) = \Phi(f, \mathbb{B}) \end{array} \right.$$

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

for each family $f \in \mathcal{F}_*$, 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}^\dagger and in \mathbb{B}^\dagger

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

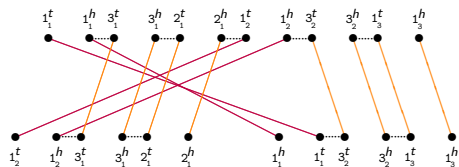
\mathfrak{M} : set of all possible pairs of matched canonical genomes obtained from balanced genomes \mathbb{A} and \mathbb{B}

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

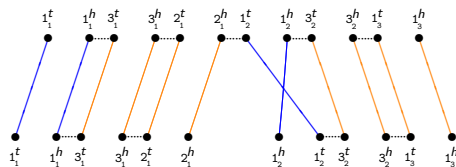
$$d_{\text{DCJ}}(\mathbb{A}, \mathbb{B}) = \min_{(\mathbb{A}^\dagger, \mathbb{B}^\dagger) \in \mathfrak{M}} \{d_{\text{DCJ}}(\mathbb{A}^\dagger, \mathbb{B}^\dagger)\}$$

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

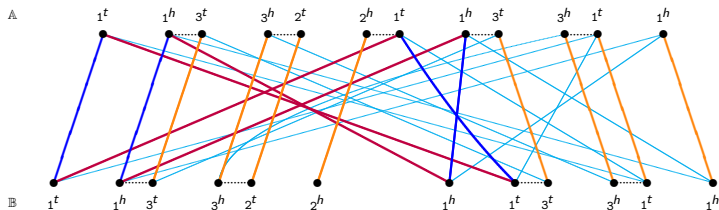
Example: $\mathbb{A} = [132131]$ and $\mathbb{B} = [132] \ [131]$



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



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

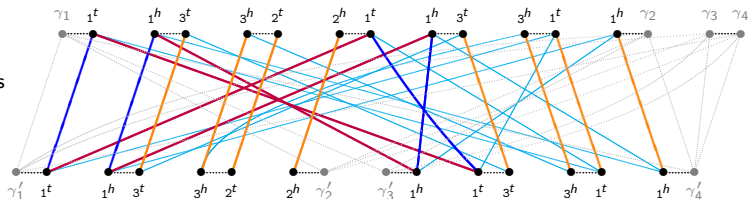


Consistent decompositions of $MRG(\mathbb{A}, \mathbb{B})$

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

Example: $\mathbb{A} = [132131]$ and $\mathbb{B} = [132] \ [131]$, $p_* = \max\{\kappa(\mathbb{A}), \kappa(\mathbb{B})\} = 2$

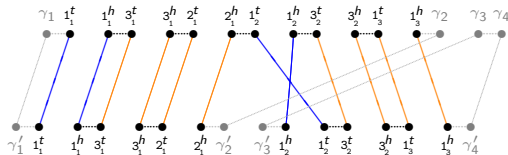
Add $2p_*$ cap extremities to each genome



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

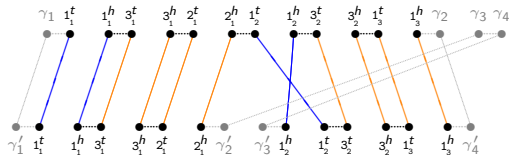
Two distinct cappings of the maximal sibling-set composed of blue + orange edges:

Non-optimal capping



$$d_{DCJ} = n + p_* - c = 6 + 2 - 5 = 3$$

Optimal capping



$$d_{DCJ} = n + p_* - c = 6 + 2 - 6 = 2$$

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

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)