Topics of today:

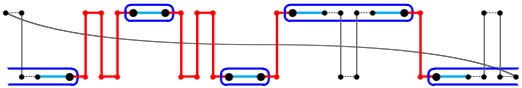
Singular DCJ-indel model: indel-potential of cycles via transitions

Overview of studied problems

Review via exercises/quiz

Indel-potential via transitions

One indel-enclosing cycle:



$\aleph = \Lambda = 4$

indel edges , runs , transitions

$\Lambda(C)$ is the number of **runs** in cycle C

$\aleph(C)$ is the number of **transitions** in cycle C

Λ	$\aleph$	r	λ	
0	0	0	0	cycles
1	0	1	1	cycles and singletons
2	2	1	2	cycles
4	4	1	3	cycles
6	6	1	4	cycles
⋮	⋮	⋮	⋮	⋮

Indel-potential of a component C:

$$\lambda(C) = \begin{cases} 0 & \text{if } \Lambda(C) = 0 \ (C \text{ is indel-free}) \\ 1 & \text{if } \Lambda(C) = 1 \\ \frac{\Lambda(C)}{2} + 1 & \text{if } \Lambda(C) \geq 2 \end{cases}$$

$$\lambda(C) = \frac{\aleph(C)}{2} + r(C)$$

$$r(C) = \begin{cases} 1\,, & \text{component } C \text{ is indel-enclosing} \\ 0\,, & \text{component } C \text{ is indel-free} \end{cases}$$

DCJ-indel distance formula

$$\sum_{\text{each cycle } C} \lambda(C) = \sum_{\text{each cycle } C} \left(\frac{\aleph(C)}{2} + r(C) \right) = e + s + \sum_{\text{each cycle } C} \frac{\aleph(C)}{2}$$

e : number of **indel-enclosing** cycles with length ≥ 2

s : number of (**indel-enclosing**) 0-cycles

DCJ-indel distance formula

$$\sum_{\text{each cycle } C} \lambda(C) = \sum_{\text{each cycle } C} \left(\frac{\aleph(C)}{2} + r(C) \right) = e + s + \sum_{\text{each cycle } C} \frac{\aleph(C)}{2}$$

e : number of **indel-enclosing** cycles with length ≥ 2

s : number of (**indel-enclosing**) 0-cycles

DCJ-indel distance of unbalanced singular **genomes**

$$\mathrm{d(A, B)} = \boxed{p_* + n - c} + \boxed{e + s + \sum \frac{\aleph(C)}{2}}$$

DCJ-indel distance formula

$$\sum_{\text{each cycle } C} \lambda(C) = \sum_{\text{each cycle } C} \left(\frac{\aleph(C)}{2} + r(C) \right) = e + s + \sum_{\text{each cycle } C} \frac{\aleph(C)}{2}$$

e : number of **indel-enclosing** cycles with length ≥ 2

s : number of (**indel-enclosing**) 0-cycles

DCJ-indel distance of unbalanced singular **genomes**

$$\mathrm{d(A,B)} = \boxed{p_* + n - c} + \boxed{e + s + \sum \frac{\aleph(C)}{2}}$$

$$= p_* + n - \tilde{e} + s + \sum \frac{\aleph(C)}{2}$$

c : number of cycles with length ≥ 2

$\tilde{e} = c - e$: number of **indel-free** cycles with length ≥ 2

DCJ-indel distance formula

$$\sum_{\text{each cycle } C} \lambda(C) = \sum_{\text{each cycle } C} \left(\frac{\aleph(C)}{2} + r(C) \right) = e + s + \sum_{\text{each cycle } C} \frac{\aleph(C)}{2}$$

e : number of **indel-enclosing** cycles with length ≥ 2

s : number of (**indel-enclosing**) 0-cycles

DCJ-indel distance of unbalanced singular genomes

$$\begin{aligned} \mathrm{d(A,B)} &= \boxed{p_* + n - c} + \boxed{e + s + \sum \tfrac{\aleph(C)}{2}} \\ &= p_* + n - \tilde{e} + s + \sum \tfrac{\aleph(C)}{2} \end{aligned}$$

c : number of cycles with length ≥ 2

$\tilde{e} = c - e$: number of **indel-free** cycles with length ≥ 2

(everything can be computed in linear time)

Reference

Computing the Rearrangement Distance of Natural Genomes

(Leonard Bohnenkämper, Marília D. V. Braga, Daniel Doerr and Jens Stoye)

JCB, Vol. 28, No. 4 (2021)

Overview of models / computational problems - 1995-2020

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

next lectures

Main references for overview

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)

On the Complexity of Rearrangement Problems under the Breakpoint Distance

(Jakub Kováč)

JCB volume 21, Number 1 (2014)

Double Cut and Join with Insertions and Deletions

(Marília D.V. Braga, Eyla Willing and Jens Stoye)

JCB, Vol. 18, No. 9 (2011)

Quiz 1 - Types of annotated genomes

1 Singular genome

2 Duplicated genome

3 Perfectly duplicated genome

4 Natural genome

5 Canonical genomes

6 Balanced genomes

A Each family occurs the same number of times in each genome

B Each adjacency and each telomere occurs exactly twice

C Each family occurs exactly once in each genome

D Each family occurs exactly twice

E No restriction on the number of occurrences of each family

F Each family occurs exactly once

Quiz 2 - Breakpoint and SCJ distances

1 What are the breakpoint and the SCJ distances between genomes $\begin{cases} \mathbb{A} = \{\ [\bar{1}, \bar{3}, 2]\ \ (4, 5, 6)\ \} \\ \mathbb{B} = \{\ [2, \bar{5}, 3, 1]\ \ (\bar{6}, \bar{4})\ \} \end{cases}$

A 3 and 6

B 2,5 and 5

C 3,5 and 6

D 3,5 and 5

$d_{BP} = n - a - \frac{t}{2} = 6 - 2 - \frac{1}{2} = 3{,}5$

$d_{SCJ} = \#_x = 6$

2 Theoretical bounds for the SCJ distance with respect to the breakpoint distance are:

$\mathbb{A} = [1]\ [2]$ and $\mathbb{B} = [1\ 2]$

A $d_{BP}(\mathbb{A}, \mathbb{B}) < d_{SCJ}(\mathbb{A}, \mathbb{B}) < 2d_{BP}(\mathbb{A}, \mathbb{B})$

B $d_{BP}(\mathbb{A}, \mathbb{B}) \leq d_{SCJ}(\mathbb{A}, \mathbb{B}) < 2d_{BP}(\mathbb{A}, \mathbb{B})$

C $d_{BP}(\mathbb{A}, \mathbb{B}) < d_{SCJ}(\mathbb{A}, \mathbb{B}) \leq 2d_{BP}(\mathbb{A}, \mathbb{B})$

D $d_{BP}(\mathbb{A}, \mathbb{B}) \leq d_{SCJ}(\mathbb{A}, \mathbb{B}) \leq 2d_{BP}(\mathbb{A}, \mathbb{B})$

circular genomes

Quiz 3 - Median of genomes (1)

For each of the following versions of the median problem, say whether its computational complexity is ...

A polynomial B NP-hard C unknown

1. Breakpoint median of general (multi mixed) genomes A
2. Breakpoint multilinear median of multilinear genomes B
3. Breakpoint multicircular median of multicircular genomes A
4. Breakpoint unichromosomal median of unichromosomal genomes B
5. SCJ median of general (multi mixed) genomes A
6. SCJ multilinear median of multilinear genomes A
7. SCJ multicircular median of multicircular genomes A
8. SCJ unichromosomal median of unichromosomal genomes C
9. DCJ median of general (multi mixed) genomes B

Quiz 4 - Median of genomes (2)

Given genomes $\begin{cases} \mathbb{A} = [1\ 2\ 3\ 4\ 5] \\ \mathbb{B} = [1\ 4\ 2\ \bar{3}\ 5] \\ \mathbb{C} = [1\ 3\ \bar{4}\ 2\ 5] \end{cases}$

1 Which are breakpoint medians of $\mathbb{A}$, $\mathbb{B}$ and $\mathbb{C}$?

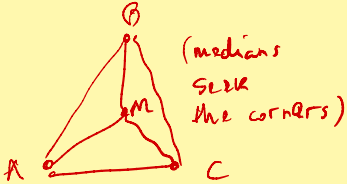
i $[1\ 2]\ [3]\ [\bar{4}\ 5]$

ii $[1\ 2\ 3\ 4\ 5]$

iii $[1\ 4\ 2\ \bar{3}\ 5]$

iv $[1]\ [2]\ [3]\ [4]\ [5]$

v $[1\ 3\ \bar{4}\ 2\ 5]$



2 Which is the SCJ median of $\mathbb{A}$, $\mathbb{B}$ and $\mathbb{C}$?

i $[1\ 2]\ [3]\ [\bar{4}\ 5]$

ii $[1\ 2\ 3\ 4\ 5]$

iii $[1\ 4\ 2\ \bar{3}\ 5]$

iv $[1]\ [2]\ [3]\ [4]\ [5]$

v $[1\ 3\ \bar{4}\ 2\ 5]$

Quiz 5 - Doubling genomes

1 A doubled (or perfectly duplicated) genome has two occurrences of each adjacency and two occurrences of each telomere.

A set $2{\cdot}\mathbb{S}$ of perfectly duplicated genomes (each one with the same set of adjacency and telomeres) can be obtained by doubling a singular genome $\mathbb{S}$.

If $\mathbb{S}$ has ℓ linear and k circular chromosomes, the cardinality of the set set $2{\cdot}\mathbb{S}$ is

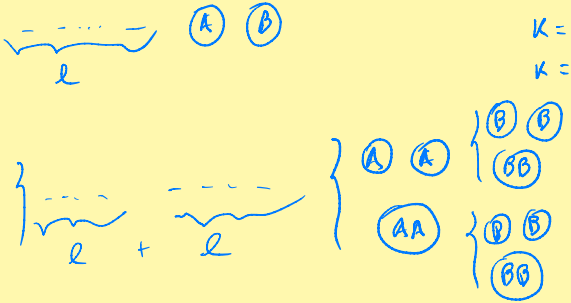
A $2^\ell 2^k$

B $\ell 2^k$

C $(\ell k)^2$

D 2^k

E $k2^\ell$



Quiz 6 - DCJ double distance and halving

1 The computational complexities of computing the DCJ double distance and the DCJ halving are

NP-hard polynomial

A both NP-hard

B both polynomial

C both unknown

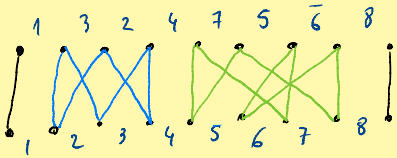
D NP-hard and polynomial

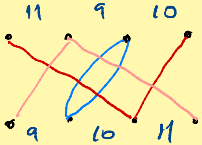
E polynomial and unknown

F polynomial and NP-hard

Quiz 7 - DCJ distance

Given genomes $\begin{cases} \mathbb{A} = [1\ 3\ 2\ 4\ 7\ 5\ \bar{6}\ 8]\ [11\ 9\ 10] \\ \mathbb{B} = [1\ 2\ 3\ 4\ 5\ 6\ 7\ 8]\ [9\ 10\ 11] \end{cases}$





3 2 1 1

1 How many cycles, $\mathbb{AB}$-paths, $\mathbb{AA}$-paths and $\mathbb{BB}$-paths are in the adjacency graph $AG(\mathbb{A}, \mathbb{B})$?

A 2, 2, 1, 0 B 3, 2, 0, 1 C 3, 2, 1, 1 D 3, 1, 1, 2

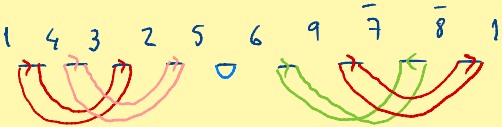
2 What is the DCJ distance $d_{DCJ}(\mathbb{A}, \mathbb{B})$?

A 6 B 8 C 7 D 5

$$d = n - c - \frac{P_{AB}}{2} = 11 - 3 - \frac{2}{2} = 7$$

Quiz 8 - Inversion distance (1)

Given genomes $\begin{cases} \mathbb{A} = (1\ 4\ 3\ 2\ 5\ 6\ 9\ \bar{7}\ \bar{8}) \\ \mathbb{B} = (1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9) \end{cases}$



1 1 3

1 How many trivial cycles, good cycles and bad cycles are in the breakpoint diagram $BD(\mathbb{A}, \mathbb{B})$?

A 1, 1, 3 B 1, 2, 2 C 0, 1, 3 D 2, 2, 1

1 1 1 1

2 How many trivial components, good components, bad components and hurdles are in $BD(\mathbb{A}, \mathbb{B})$?

A 1, 1, 1, 0 B 1, 1, 1, 1 C 0, 1, 2, 1 D 1, 0, 2, 2

3 What is the inversion distance $d_{\text{INV}}(\mathbb{A}, \mathbb{B})$?

A 4 B 6 C 5 D 3

$d = n - c + h + \cancel{f}$

$= 9 - 5 + 1 = 5$

Quiz 9 - Inversion distance (2)

1 The bottleneck of the inversion sorting is

A Finding hurdles

B Identifying a fortress

C Finding safe split inversions

D Identifying super-hurdles

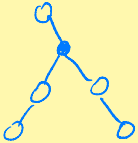
2 A data structure that helps finding safe split inversions is the

A breakpoint diagram

B component tree

C max-flower

D overlap graph



3 All leaves of the component tree are

A good components

B bad components

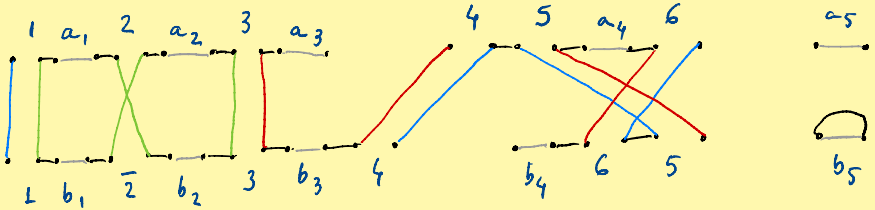
C hurdles

D super-hurdles

leaves in long leaf-branches are super-hurdles

Quiz 10 - DCJ-indel distance (1)

Given genomes $\begin{cases} \mathbb{A} = [1\ a_1\ 2\ a_2\ 3\ a_3]\ \ [4\ 5\ a_4\ 6]\ \ [a_5] \\ \mathbb{B} = [1\ b_1\ \bar{2}\ b_2\ 3\ b_3\ 4]\ \ [b_4\ 6\ 5]\ \ (b_5) \end{cases}$



$$6 - 1 - \frac{2}{2} + \underbrace{3}_{C} + \underbrace{2}_{AA_{AB}} + \underbrace{2}_{BB_{AB}} + 1 + 1 = \underbrace{4}_{DCJ} + 9 = 13$$

1 What is the value of the formula $n - |\mathcal{C}| - \frac{|\mathcal{P}_{\mathbb{AB}}|}{2} + \sum_{C \in RG(\mathbb{A},\mathbb{B})} \lambda(C)$?

A 4 B 9 C 13 D 12

2 Is there any deducting path recombination?

A Yes B No

$$d = 13 - \underbrace{2}_{\delta} = 11$$