Topics of today:

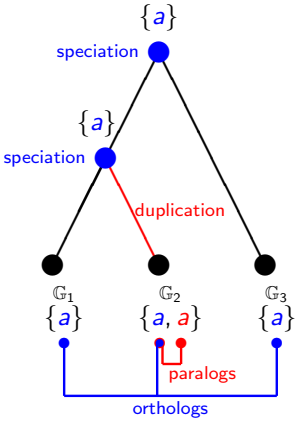
1. Inferring gene families via family-free DCJ-indel distance

2. Quiz-review

Introduction

Gene families, orthologs and paralogs

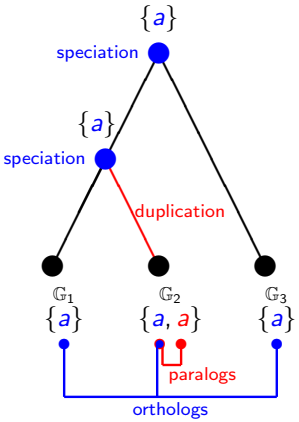
gene family: set of genes sharing a common ancestor



[Fitch, 1970]

Gene families, orthologs and paralogs

gene family: set of genes sharing a common ancestor



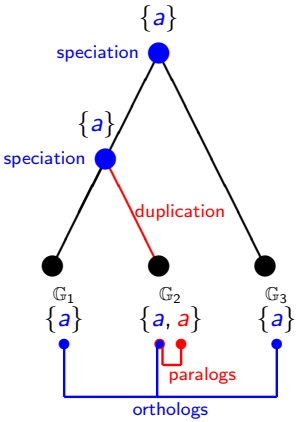
[Fitch, 1970]

# of occurrences

family	$\mathbb{G}_1$	$\mathbb{G}_2$	$\mathbb{G}_3$	
a	1	2	1	ambiguous

Gene families, orthologs and paralogs

gene family: set of genes sharing a common ancestor

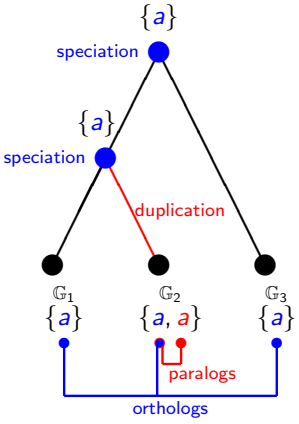


[Fitch, 1970]

	# of occurrences			
family	$\mathbb{G}_1$	$\mathbb{G}_2$	$\mathbb{G}_3$	
a	1	2	1	ambiguous
b	3	0	1	ambiguous
c	0	1	1	resolved
d	1	1	1	resolved

Gene families, orthologs and paralogs

gene family: set of genes sharing a common ancestor



[Fitch, 1970]

	# of occurrences			
family	$\mathbb{G}_1$	$\mathbb{G}_2$	$\mathbb{G}_3$	
a	1	2	1	ambiguous
b	3	0	1	ambiguous
c	0	1	1	resolved
d	1	1	1	resolved

Identifying the orthologs of two genomes within an ambiguous family is often not so easy

↓

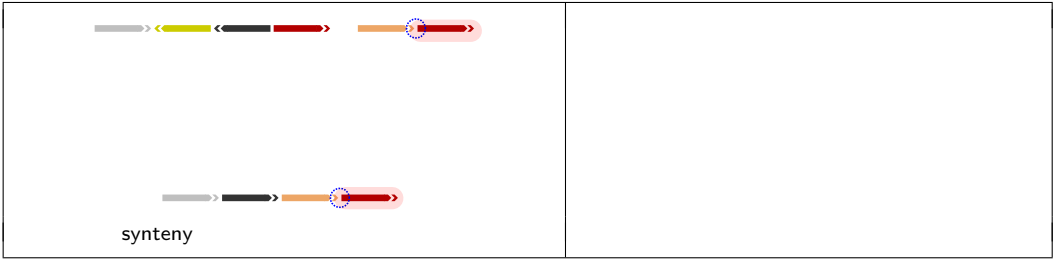
the **genome organizations** might provide some hints ▶

Identifying orthologs of two genomes

within an ambiguous family by comparing **genome organizations**

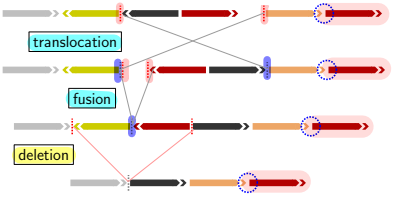
Identifying orthologs of two genomes

within an ambiguous family by comparing **genome organizations** $\left\{ \text{synteny} - \text{e. g. } \boxed{\textsc{Poff}} \right.$ [Lechner *et al.*, 2014]



Identifying orthologs of two genomes

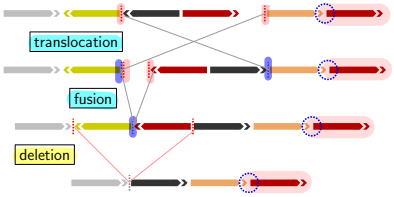
within an ambiguous family by comparing **genome organizations** $\begin{cases} \text{synteny} - \text{e. g. } \boxed{\textsc{Poff}} \text{ [Lechner } \textit{et al.}\text{, 2014]} \\ \textbf{rearrangement distance} \end{cases}$



synteny / **rearrangement distance**

Identifying orthologs of two genomes

within an ambiguous family by comparing **genome organizations** $\begin{cases} \text{synteny – e. g. } \boxed{\textsc{Poff}} \text{ [Lechner } \textit{et al.}\text{, 2014]} \\ \textbf{rearrangement distance} \end{cases}$

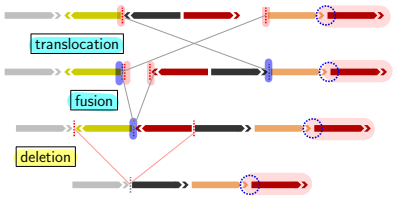


synteny / **rearrangement distance**

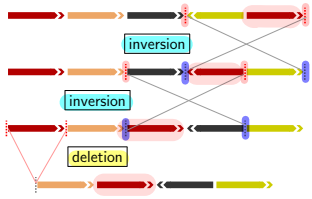
~~synteny~~

Identifying orthologs of two genomes

within an ambiguous family by comparing **genome organizations** $\begin{cases} \text{synteny – e. g. } \boxed{\textsc{Poff}} \text{ [Lechner } \textit{et al.}\text{, 2014]} \\ \textbf{rearrangement distance} \end{cases}$



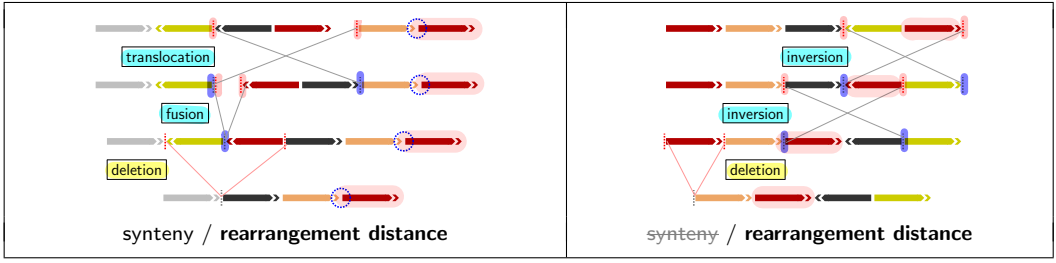
synteny / **rearrangement distance**



~~synteny~~ / **rearrangement distance**

Identifying orthologs of two genomes

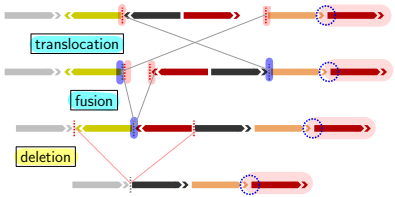
within an ambiguous family by comparing **genome organizations** $\begin{cases} \text{synteny – e. g. } \boxed{\textsc{Poff}} \text{ [Lechner } et\ al.\text{, 2014]} \\ \textbf{rearrangement distance} \end{cases}$



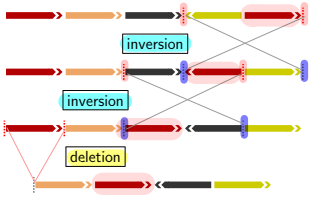
rearrangement distance can identify orthologies that are not "visible" with synteny only [Shi *et al.*, 2010]

Identifying orthologs of two genomes

within an ambiguous family by comparing **genome organizations** $\begin{cases} \text{synteny – e. g. } \boxed{\textsc{Poff}} \text{ [Lechner } et\ al.\text{, 2014]} \\ \textbf{rearrangement distance} \end{cases}$



synteny / **rearrangement distance**



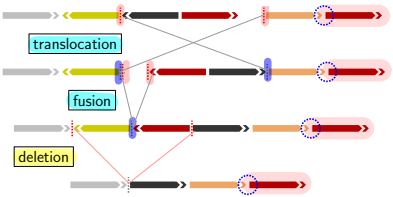
~~synteny~~ / **rearrangement distance**

rearrangement distance can identify orthologies that are not "visible" with synteny only [Shi *et al.*, 2010]

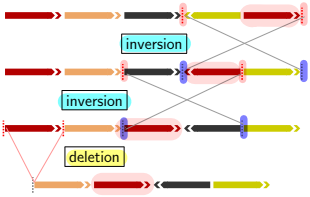
rearrangement types $\begin{cases} \boxed{\text{organizational}} \text{ : change \# of chromosomes, positions and orientations of genes} \end{cases}$

Identifying orthologs of two genomes

within an ambiguous family by comparing **genome organizations**
- synteny – e. g. POFF [Lechner *et al.*, 2014]
- **rearrangement distance**



synteny / **rearrangement distance**



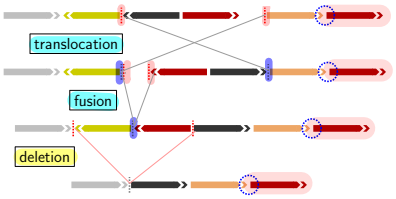
~~synteny~~ / **rearrangement distance**

rearrangement distance can identify orthologies that are not "visible" with synteny only [Shi *et al.*, 2010]

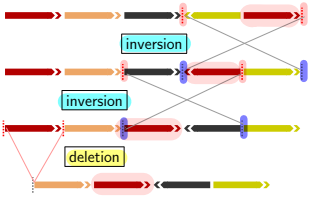
rearrangement types
- organizational : change # of chromosomes, positions and orientations of genes
  - inversions, translocations, fusions, and fissions

Identifying orthologs of two genomes

within an ambiguous family by comparing **genome organizations**
- synteny – e. g. POFF [Lechner *et al.*, 2014]
- **rearrangement distance**



synteny / **rearrangement distance**



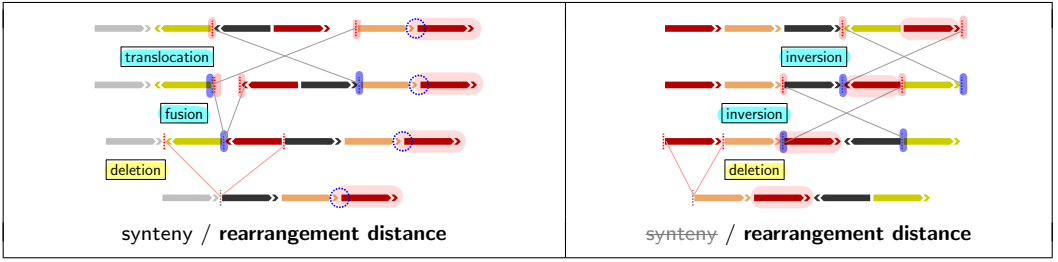
~~synteny~~ / **rearrangement distance**

rearrangement distance can identify orthologies that are not "visible" with synteny only [Shi *et al.*, 2010]

rearrangement types
- organizational : change # of chromosomes, positions and orientations of genes
  - inversions, translocations, fusions, and fissions
  - **double-cut-and-join (DCJ)** operation: two breaks + two repairs [Yancopoulos *et al.*, 2005]

Identifying orthologs of two genomes

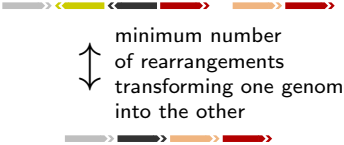
within an ambiguous family by comparing **genome organizations**
- synteny – e. g. POFF [Lechner *et al.*, 2014]
- **rearrangement distance**



rearrangement distance can identify orthologies that are not "visible" with synteny only [Shi *et al.*, 2010]

rearrangement types
- organizational : change # of chromosomes, positions and orientations of genes
  - inversions, translocations, fusions, and fissions
  - **double-cut-and-join (DCJ)** operation: two breaks + two repairs [Yancopoulos *et al.*, 2005]
- content-modifying
  - duplications
  - **insertions and deletions (indels)**

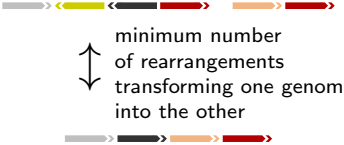
Rearrangement distance



only resolved families: polynomial - e. g.

inversions (+ transl., fus., fis.)	(Hannenhalli and Pevzner, 1995)
DCJ operations	(Bergeron *et al.*, 2006)
DCJ operations and indels	(Braga *et al.*, 2010)

Rearrangement distance

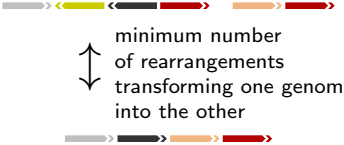


only resolved families: polynomial - e. g.

inversions (+ transl., fus., fis.)	(Hannenhalli and Pevzner, 1995)
DCJ operations	(Bergeron *et al.*, 2006)
DCJ operations and indels	(Braga *et al.*, 2010)

with ambiguous families: NP-hard → find a **matching that minimizes the distance** between the derived genomes **with resolved families**

Rearrangement distance



only resolved families: polynomial - e. g.

- inversions (+ transl., fus., fis.) (Hannenhalli and Pevzner, 1995)
- DCJ operations (Bergeron *et al.*, 2006)
- DCJ operations and indels (Braga *et al.*, 2010)

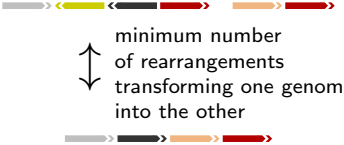
with ambiguous families: NP-hard → find a **matching that minimizes the distance** between the derived genomes **with resolved families**

↘

existing solutions

- MSOAR: heuristic solution (Shi *et al.*, 2010)
  - inversions, translocations, fusions and fissions
  - tandem duplications

Rearrangement distance



only resolved families: polynomial - e. g.

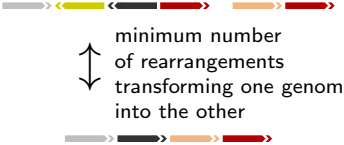
- inversions (+ transl., fus., fis.) (Hannenhalli and Pevzner, 1995)
- DCJ operations (Bergeron *et al.*, 2006)
- DCJ operations and indels (Braga *et al.*, 2010)

with ambiguous families: NP-hard → find a **matching that minimizes the distance** between the derived genomes **with resolved families**

↘ existing solutions

- MSOAR: heuristic solution (Shi *et al.*, 2010)
  - inversions, translocations, fusions and fissions
  - tandem duplications
- BALANCED: **optimal solution** via ILP (Shao *et al.*, 2015)
  - DCJ operations
- DING: **optimal solution** via ILP (Bohnenkämper *et al.*, 2021)
  - DCJ operations
  - indels

Rearrangement distance



all methods below require the **pre-computation of gene families**

only resolved families: polynomial - e. g.

- inversions (+ transl., fus., fis.) (Hannenhalli and Pevzner, 1995)
- DCJ operations (Bergeron *et al.*, 2006)
- DCJ operations and indels (Braga *et al.*, 2010)

with ambiguous families: NP-hard → find a **matching that minimizes the distance** between the derived genomes **with resolved families**

↘ existing solutions

- MSOAR: heuristic solution (Shi *et al.*, 2010)
  - inversions, translocations, fusions and fissions
  - tandem duplications
- BALANCED: **optimal solution** via ILP (Shao *et al.*, 2015)
  - DCJ operations
- DING: **optimal solution** via ILP (Bohnenkämper *et al.*, 2021)
  - DCJ operations
  - indels

Quiz 1

1 Which of the following statements are true?

A An ambiguous family occurs more than once in the same genome.

~~B~~ A resolved family occurs exactly once in each genome.

at most would be true

2 Two occurrences of the same family in the same genome are called...

A orthologs

B paralogs

3 Genes from a resolved family are called...

A orthologs

B paralogs

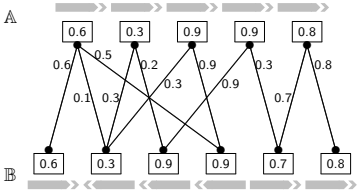
Family-free DCJ-indel distance:

finding orthologs
without pre-computed families

considering simultaneously $\begin{cases} \text{pairwise gene similarities} \\ \text{genome rearrangements} \end{cases}$

Family-free DCJ-indel distance define orthologs of two genomes

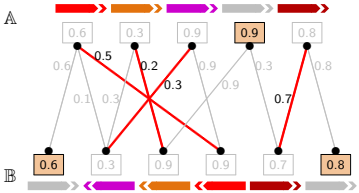
No family assignment, but pairwise normalized similarities



$x = 0.1$

Family-free DCJ-indel distance define orthologs of two genomes

No family assignment, but pairwise normalized similarities



M	$\mathrm{d}^{\mathrm{ID}}_{\mathrm{DCJ}}$	$+$	$\lvert M\rvert$	$-$	$w(M)$	$+$	$w(\widetilde{M})$	$=$	$\mathrm{wd}^{\mathrm{ID}}_{\mathrm{DCJ}}$
■	4	$+$	4	$-$	1.7	$+$	2.3	$=$	8.6

weighted DCJ-indel distance of M

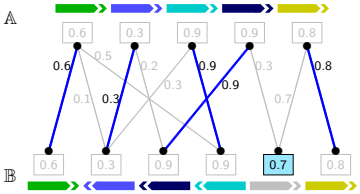
$$\mathrm{wd}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{A}, \mathbb{B}, M) = \mathrm{d}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{A}, \mathbb{B}, M) + |M| - w(M) + w(\widetilde{M})$$

M : matching

complement $= \widetilde{M}$: set of M-unsaturated genes

Family-free DCJ-indel distance define orthologs of two genomes

No family assignment, but pairwise normalized similarities



M	$\mathrm{d}^{\mathrm{ID}}_{\mathrm{DCJ}}$	$+$	$\lvert M\rvert$	$-$	$w(M)$	$+$	$w(\widetilde{M})$	$=$	$\mathrm{wd}^{\mathrm{ID}}_{\mathrm{DCJ}}$
	4	+	4	−	1.7	+	2.3	=	8.6
	3	+	5	−	3.5	+	0.7	=	5.2

weighted DCJ-indel distance of M

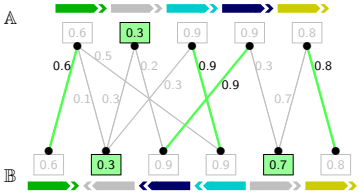
$$\mathrm{wd}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{A}, \mathbb{B}, M) = \mathrm{d}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{A}, \mathbb{B}, M) + |M| - w(M) + w(\widetilde{M})$$

M : matching

complement $= \widetilde{M}$: set of M-unsaturated genes

Family-free DCJ-indel distance define orthologs of two genomes

No family assignment, but pairwise normalized similarities



M	$\mathrm{d}^{\mathrm{ID}}_{\mathrm{DCJ}}$	$+$	$\lvert M\rvert$	$-$	$w(M)$	$+$	$w(\widetilde{M})$	$=$	$\mathrm{wd}^{\mathrm{ID}}_{\mathrm{DCJ}}$
	4	+	4	−	1.7	+	2.3	=	8.6
	3	+	5	−	3.5	+	0.7	=	5.2
	3	+	4	−	3.2	+	1.3	=	5.1

$\subset$

weighted DCJ-indel distance of M

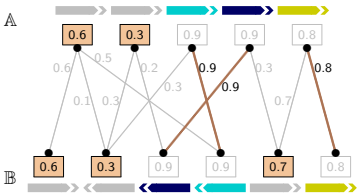
$$\mathrm{wd}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{A}, \mathbb{B}, M) = \mathrm{d}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{A}, \mathbb{B}, M) + |M| - w(M) + w(\widetilde{M})$$

M : matching

complement $= \widetilde{M}$: set of M-unsaturated genes

Family-free DCJ-indel distance define orthologs of two genomes

No family assignment, but pairwise normalized similarities



M	$\mathrm{d}^{\mathrm{ID}}_{\mathrm{DCJ}}$	$+$	$\lvert M\rvert$	$-$	$w(M)$	$+$	$w(\widetilde{M})$	$=$	$\mathrm{wd}^{\mathrm{ID}}_{\mathrm{DCJ}}$
	4	$+$	4	$-$	1.7	$+$	2.3	$=$	8.6
	3	$+$	5	$-$	3.5	$+$	0.7	$=$	5.2
	3	$+$	4	$-$	3.2	$+$	1.3	$=$	5.1
	3	$+$	3	$-$	2.6	$+$	2.5	$=$	5.9

$\subset \quad \subset$

weighted DCJ-indel distance of M

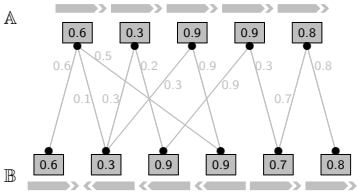
$$\mathrm{wd}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{A}, \mathbb{B}, M) = \mathrm{d}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{A}, \mathbb{B}, M) + |M| - w(M) + w(\widetilde{M})$$

M : matching

complement $= \widetilde{M}$: set of M-unsaturated genes

Family-free DCJ-indel distance define orthologs of two genomes

No family assignment, but pairwise normalized similarities



M	$\mathrm{d}^{\mathrm{ID}}_{\mathrm{DCJ}}$	$+$	$\lvert M\rvert$	$-$	$w(M)$	$+$	$w(\widetilde{M})$	$=$	$\mathrm{wd}^{\mathrm{ID}}_{\mathrm{DCJ}}$
■	4	$+$	4	$-$	1.7	$+$	2.3	$=$	8.6
■	3	$+$	5	$-$	3.5	$+$	0.7	$=$	5.2
■	3	$+$	4	$-$	3.2	$+$	1.3	$=$	5.1
■	3	$+$	3	$-$	2.6	$+$	2.5	$=$	5.9
$\emptyset$	2	$+$	0	$-$	0	$+$	7.7	$=$	9.7

■ $\subset$ ■ $\subset$ ■

weighted DCJ-indel distance of M

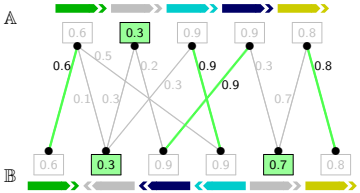
$$\mathrm{wd}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{A}, \mathbb{B}, M) = \mathrm{d}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{A}, \mathbb{B}, M) + |M| - w(M) + w(\widetilde{M})$$

M : matching

complement $= \widetilde{M}$: set of M-unsaturated genes

Family-free DCJ-indel distance define orthologs of two genomes

No family assignment, but pairwise normalized similarities



M	$\mathrm{d}^{\mathrm{ID}}_{\mathrm{DCJ}}$	$+$	$\lvert M\rvert$	$-$	$w(M)$	$+$	$w(\widetilde{M})$	$=$	$\mathrm{wd}^{\mathrm{ID}}_{\mathrm{DCJ}}$
	4	$+$	4	$-$	1.7	$+$	2.3	$=$	8.6
	3	$+$	5	$-$	3.5	$+$	0.7	$=$	5.2
	3	$+$	4	$-$	3.2	$+$	1.3	$=$	5.1
	3	$+$	3	$-$	2.6	$+$	2.5	$=$	5.9
$\emptyset$	2	$+$	0	$-$	0	$+$	7.7	$=$	9.7

$\subset$ $\subset$

weighted DCJ-indel distance of M

$$\mathrm{wd}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{A},\mathbb{B},M) = \mathrm{d}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{A},\mathbb{B},M) + |M| - w(M) + w(\widetilde{M})$$

M : matching

complement $= \widetilde{M}$: set of M-unsaturated genes

family-free DCJ-indel distance

$$\mathrm{ffd}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{A},\mathbb{B}) = \min_{M}\{\mathrm{wd}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{A},\mathbb{B},M)\}$$

NP-hard

optimal solution via ILP with DiffM ▶

DiffM: computing an optimal matching via ILP

Based on the weighted multirelational graph

Objective: Minimize $\sum_{a \in E_\xi} x_a - \sum_{1 \le i \le |V|} \tilde{e}_i + \sum_{k \in K} s_k + \frac{1}{2} \sum_{a \in E} t_a \boxed{- \frac{1}{2} \sum_{a \in E_\xi} w_a x_a + \sum_{c \in E_{\mathrm{ID}}} w_c x_c}$

weights appear only in the objective function

Constraints:

(C.01)	$x_a = 1$	$\forall\, a \in E_\Gamma^A \cup E_\Gamma^B$
(C.02)	$\sum_{\{u,v\} \in E} x_{\{u,v\}} = 2$	$\forall\, u \in V$
(C.03)	$x_a = x_b$	$\forall\, a, b \in E_\xi$ such that a and b are siblings
(C.04)	$\ell_i \le \ell_j + i(1 - x_{\{v_i,v_j\}})$	$\forall\, \{v_i, v_j\} \in E$,
(C.05)	$\ell_i \le i(1 - x_{\{v_i,v_j\}})$	$\forall\, \{v_i, v_j\} \in E_{\mathrm{ID}}^A \cup E_{\mathrm{ID}}^B$
(C.06)	$i \cdot \tilde{e}_i \le \ell_i$	$\forall\, 1 \le i \le \lvert V \rvert$
(C.07)	$r_v \le 1 - x_{\{u,v\}}$	$\forall\, \{u, v\} \in E_{\mathrm{ID}}^A$,
	$r_{v'} \ge x_{\{u',v'\}}$	$\forall\, \{u', v'\} \in E_{\mathrm{ID}}^B$
(C.08)	$t_{\{u,v\}} \ge r_v - r_u - (1 - x_{\{u,v\}})$	$\forall\, \{u, v\} \in E$
(C.09)	$\sum_{a \in E_{\mathrm{ID}}^A,\, a \cap b \neq \varnothing} x_a - t_b \ge 0$	$\forall\, b \in E_\Gamma^A$
(C.10)	$t_a = 0$	$\forall\, a \in E \setminus E_\Gamma^A$
(C.11)	$\sum_{a \in E_{\mathrm{ID}}^k} x_a - \lvert k \rvert + 1 \le s_k$	$\forall k \in K$

Domains:

(D.01)	$x_a \in \{0, 1\}$	$\forall\, a \in E$
(D.02)	$0 \le \ell_i \le i$	$\forall\, 1 \le i \le \lvert V \rvert$
(D.03)	$\tilde{e}_i \in \{0, 1\}$	$\forall\, 1 \le i \le \lvert V \rvert$
(D.04)	$r_v \in \{0, 1\}$	$\forall\, v \in V$
(D.05)	$t_a \in \{0, 1\}$	$\forall\, a \in E$
(D.06)	$s_k \in \{0, 1\}$	$\forall\, k \in K$

(Rubert *et al.*, 2021)

(Bohnenkämper *et al.*, 2021)
(Shao *et al.*, 2015)

Generation of gene families

Generating gene families via genome rearrangements

Inversion of the typical setting $\boxed{\text{gene families} \Rightarrow \text{genome rearrangements}}$

Generating gene families via genome rearrangements

Inversion of the typical setting $\boxed{\text{gene families} \Leftarrow \text{genome rearrangements}}$

Generating gene families via genome rearrangements

Inversion of the typical setting $\boxed{\text{gene families} \Leftarrow \text{genome rearrangements}}$

The optimal matchings given by the ILP $\boxed{\textsc{DiffM}}$ are the basis

for a generator of gene families of a set of genomes

$\Downarrow$

$\boxed{\textsc{DiffMGC}}$

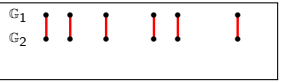
Pipeline of DiffMGC: FFGC + DiffM + integration

n genomes $\longrightarrow$ $\binom{n}{2}$ pairwise comparisons:

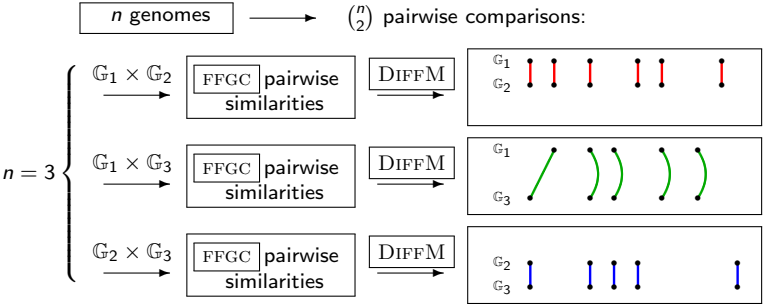
Pipeline of DIFFMGC: FFGC + DIFFM + integration

n genomes $\longrightarrow$ $\binom{n}{2}$ pairwise comparisons:

$n = 3$ $\left\{ \begin{array}{l} \mathbb{G}_1 \times \mathbb{G}_2 \longrightarrow \boxed{\text{FFGC}} \text{ pairwise similarities} \longrightarrow \boxed{\text{DIFFM}} \longrightarrow \end{array} \right.$



Pipeline of DIFFMGC: FFGC + DIFFM + integration

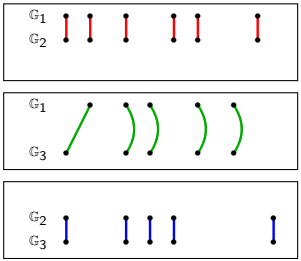


Pipeline of DIFFMGC: FFGC + DIFFM + integration

n genomes → $\binom{n}{2}$ pairwise comparisons:

$n = 3$:

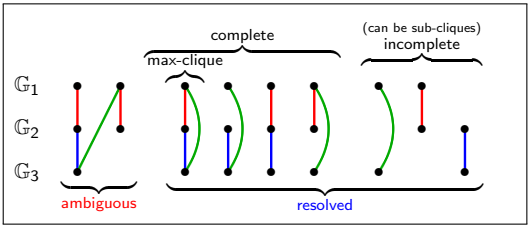
- $\mathbb{G}_1 \times \mathbb{G}_2$ → FFGC pairwise similarities → DIFFM
- $\mathbb{G}_1 \times \mathbb{G}_3$ → FFGC pairwise similarities → DIFFM
- $\mathbb{G}_2 \times \mathbb{G}_3$ → FFGC pairwise similarities → DIFFM



integration of pairwise matchings: DIFFM **families**

⇓

connected components of the 3-partite DIFFM graph



Quiz 2

1 The integration of the pairwise matchings given by DIFFM is....

A a similarity graph

B a multipartite graph

C a multirelational graph

2 Which of the following statements are true?

A Any complete resolved family is a max-clique in the DIFFM graph.

B Any max-clique in the DIFFM graph is a complete resolved family.

C Any clique in the DIFFM graph is a resolved family.

3 A max-clique in the DIFFM graph of n genomes is composed of...

A n^2 vertices and $\frac{n(n-1)}{2}$ edges

B n vertices and $\frac{n(n-1)}{2}$ edges

C n vertices and n^2 edges

n vertices:

each vertex n-1 edges

$\frac{n(n-1)}{2}$

each edge 2 vertices

Experiments

Other genome-scale methods

$\boxed{\text{MSOAR} + \text{MultiMSOAR}}$: $\begin{cases} \text{pairwise heuristic matching via rearrangements (Shi } \textit{et al.}\text{, 2010)} \\ \text{integration of matchings (Shi } \textit{et al.}\text{, 2011)} \end{cases}$

Other genome-scale methods

$\boxed{\text{MSOAR} + \text{MultiMSOAR}}$: $\begin{cases} \cancel{\text{pairwise heuristic matching via rearrangements}} \text{ (Shi } \textit{et al.}\text{, 2010)} \\ \text{integration of matchings (Shi } \textit{et al.}\text{, 2011)} \end{cases}$

Other genome-scale methods

MSOAR + MultiMSOAR: ~~pairwise heuristic matching via rearrangements~~ (Shi *et al.*, 2010)
integration of matchings (Shi *et al.*, 2011)

ProteinOrtho: reciprocal best alignment heuristic (Lechner *et al.*, 2011)

Other genome-scale methods

MSOAR + MultiMSOAR: ~~pairwise heuristic matching via rearrangements~~ (Shi *et al.*, 2010)
integration of matchings (Shi *et al.*, 2011)

ProteinOrtho: reciprocal best alignment heuristic (Lechner *et al.*, 2011)

Poff: extension of ProteinOrtho incorporating synteny (Lechner *et al.*, 2014)

Other genome-scale methods

MSOAR + MultiMSOAR: ~~pairwise heuristic matching via rearrangements~~ (Shi *et al.*, 2010)
integration of matchings (Shi *et al.*, 2011)

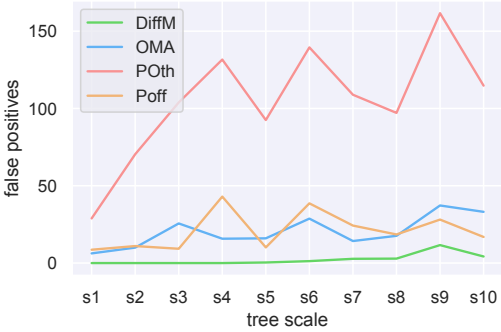
ProteinOrtho: reciprocal best alignment heuristic (Lechner *et al.*, 2011)

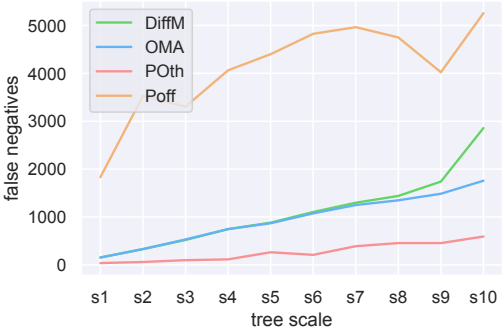
Poff: extension of ProteinOrtho incorporating synteny (Lechner *et al.*, 2014)

Oma: similarities and phylogeny (Dessimoz *et al.*, 2005 & Altenhoff *et al.*, 2019)

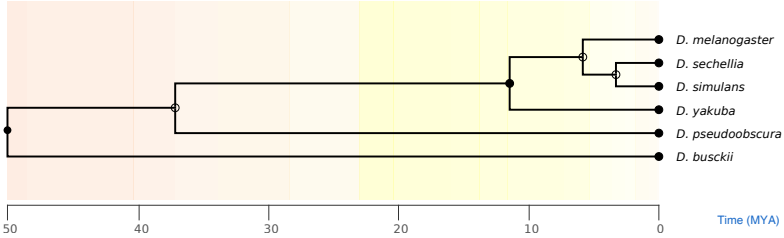
Simulated genomes

- 80 datasets with 10 extant genomes each, with Zombi (Davín *et al.*, 2019)
- High **true positive** rates for all methods (omitted)





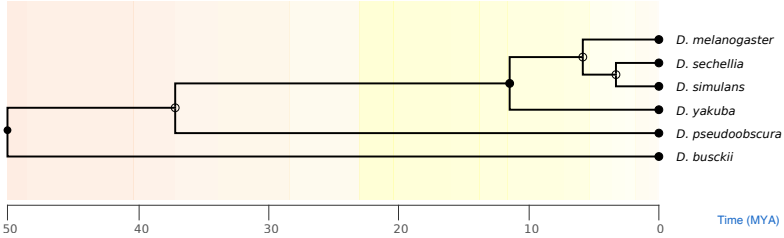
Drosophila genomes



$\sim$ 13,000 genes
5–7 chromosomes

FlyBase: **reference families** for *D. melanogaster*, *D. simulans* and *D. yakuba*

Drosophila genomes



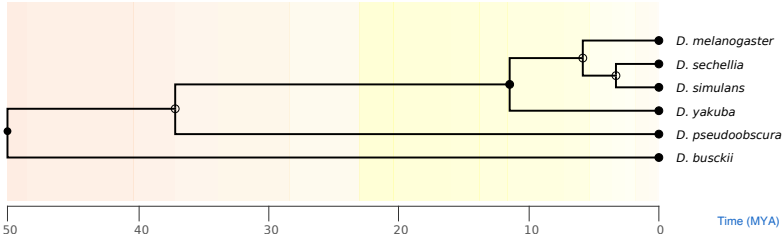
$\sim$ 13,000 genes
5–7 chromosomes

FlyBase: **reference families** for *D. melanogaster*, *D. simulans* and *D. yakuba*

	FlyBase	FlyBase **families inferred by**			
		DiffMGC	Poff	POrtho	Oma
resolved	11659	11515	10487	11396	11383
complete	10809	10713	9745	10594	10594

all methods found at least 90% of all resolved and complete FlyBase families

Drosophila genomes



$\sim$ 13,000 genes
5–7 chromosomes

FlyBase: **reference families** for *D. melanogaster*, *D. simulans* and *D. yakuba*

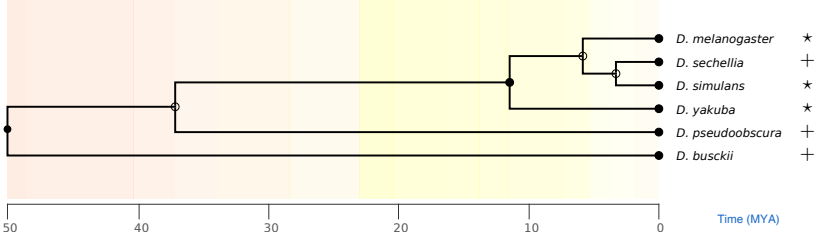
	FlyBase	FlyBase **families inferred by**			
	FlyBase	DiffMGC	Poff	POrtho	Oma
resolved	11659	11515	10487	11396	11383
complete	10809	10713	9745	10594	10594

all methods found at least 90% of all resolved and complete FlyBase families

DiffMGC achieved the highest agreement with FlyBase

99% of DiffMGC complete families are max-cliques in the 3-partite DiffM graph

Results for six *Drosophilas*

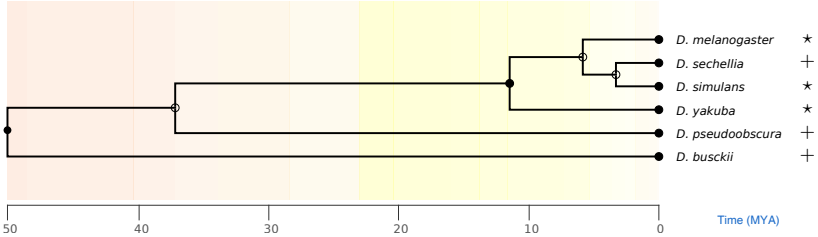


$\sim$ 13,000 genes
5–7 chromosomes

D. sechellia is closely related to *D. melanogaster*, *D. simulans* and *D. yakuba*

D. pseudoobscura and *D. busckii* are phylogenetically more distant

Results for six *Drosophilas*



$\sim$ 13,000 genes
5–7 chromosomes

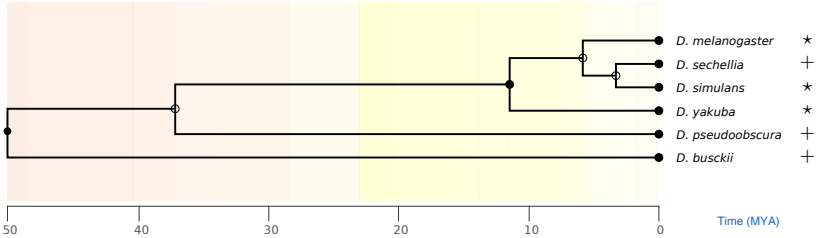
D. sechellia is closely related to *D. melanogaster*, *D. simulans* and *D. yakuba*

D. pseudoobscura and *D. busckii* are phylogenetically more distant

Number of families and of resolved and complete families given by the different methods

	DiffMGC	POff	POrtho	Oma
total	12885	13282	12746	12660
resolved	12549	13050	11844	11848
complete	8010	8894	8429	8387

Results for six *Drosophilas*



$\sim$ 13,000 genes
5–7 chromosomes

D. sechellia is closely related to *D. melanogaster*, *D. simulans* and *D. yakuba*

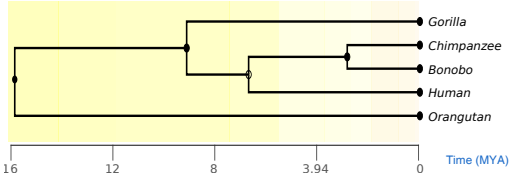
D. pseudoobscura and *D. busckii* are phylogenetically more distant

Number of families and of resolved and complete families given by the different methods

	DiffMGC	POff	POrtho	Oma
total	12885	13282	12746	12660
resolved	12549	13050	11844	11848
complete	8010	8894	8429	8387

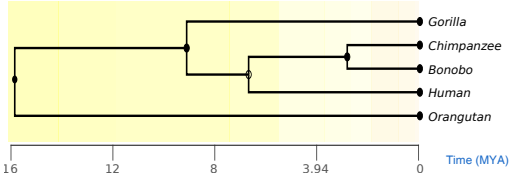
80% of DiffMGC complete families are max-cliques in the 6-partite DiffM graph

X chromosome of five primates



$\sim$ 800 genes

X chromosome of five primates

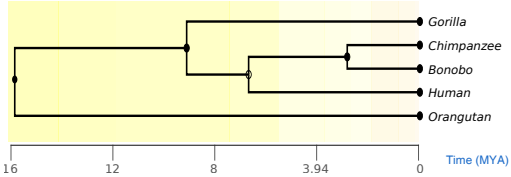


$\sim$ 800 genes

Number of families and of resolved and complete families given by the different methods:

	DiffMGC	POff	POrtho	OMA
total	822	845	782	820
resolved	784	823	715	766
complete	623	609	564	581

X chromosome of five primates



$\sim$ 800 genes

Number of families and of resolved and complete families given by the different methods:

	DiffMGC	POff	POrtho	OMA
total	822	845	782	820
resolved	784	823	715	766
complete	623	609	564	581

98% of DiffMGC complete families are max-cliques in the 5-partite DiffM graph

Running times for six *Drosophilas*

Environment: eight 3GHz cores

ProteinOrtho and Poff → 5 minutes

- pairwise similarities via DIAMOND[1]

[1] (Buchfink *et al.*, 2015)

Oma → 32 hours

- pairwise similarities via Smith-Waterman alignments

DiffMGC → 12 hours

- pairwise similarities via BLAST[2] (9 hours)

[2] (Altschul *et al.*, 1990)

Running times for six *Drosophilas*

Environment: eight 3GHz cores

ProteinOrtho and Poff → 5 minutes

- pairwise similarities via DIAMOND[1]

[1] (Buchfink *et al.*, 2015)

Oma → 32 hours

- pairwise similarities via Smith-Waterman alignments

DiffMGC → 12 hours

- pairwise similarities via BLAST[2] (9 hours)

[2] (Altschul *et al.*, 1990)

- DiffM ILP computations (CPLEX): $\sim$ 3 hours
  - 13 took less than 5 minutes
  - one took 40 minutes
  - one reached the time limit of 2 hours (with a very small opt. gap of 0.25%)

References

The potential of family-free rearrangements towards gene orthology inference

(Diego P. Rubert, Daniel Doerr and Marília D. V. Braga)

Journal of Bioinformatics and Computational Biology, vol. 19, No. 06, 2140014 (2021)

Review / Quiz

DCJ-indel model - Path recombinations

With respect to the endpoints:

$$\mathbb{A}\text{–}\mathbb{A} \;+\; \mathbb{B}\text{–}\mathbb{B} \begin{cases} \mathbb{A}\text{–}\mathbb{B} \;+\; \mathbb{A}\text{–}\mathbb{B} & \text{(gaining)} \\ \mathbb{A}\text{–}\mathbb{B} \;+\; \mathbb{A}\text{–}\mathbb{B} & \text{(gaining)} \end{cases}$$

$$\mathbb{A}\text{–}\mathbb{B} \;+\; \mathbb{A}\text{–}\mathbb{B} \begin{cases} \mathbb{A}\text{–}\mathbb{A} \;+\; \mathbb{B}\text{–}\mathbb{B} & \text{(losing)} \\ \mathbb{A}\text{–}\mathbb{B} \;+\; \mathbb{A}\text{–}\mathbb{B} & \text{(neutral)} \end{cases}$$

$$\mathbb{A}\text{–}\mathbb{A} \;+\; \mathbb{A}\text{–}\mathbb{A} \begin{cases} \mathbb{A}\text{–}\mathbb{A} \;+\; \mathbb{A}\text{–}\mathbb{A} & \text{(neutral)} \\ \mathbb{A}\text{–}\mathbb{A} \;+\; \mathbb{A}\text{–}\mathbb{A} & \text{(neutral)} \end{cases}$$

With respect to the runs:

$$\mathcal{AB} \;+\; \mathcal{AB} \begin{cases} \mathcal{AA} \;+\; \mathcal{BB} & (\Delta_\lambda = -2) \\ \mathcal{ABBA} \;+\; \varepsilon & (\Delta_\lambda = -2) \end{cases}$$

$$\mathcal{A}(\mathcal{B}) \;+\; \mathcal{A} \begin{cases} \mathcal{AA} \;+\; (\mathcal{B}) & (\Delta_\lambda = -1) \\ \mathcal{AA}(\mathcal{B}) \;+\; \varepsilon & (\Delta_\lambda = -1) \end{cases}$$

$$(\mathcal{A})\mathcal{B} \;+\; \mathcal{B} \begin{cases} (\mathcal{A}) \;+\; \mathcal{BB} & (\Delta_\lambda = -1) \\ (\mathcal{A})\mathcal{BB} \;+\; \varepsilon & (\Delta_\lambda = -1) \end{cases}$$

Deducting path recombinations:

$$\begin{cases} \text{gaining with } \Delta_\lambda = -2 \\ \text{gaining with } \Delta_\lambda = -1 \\ \text{neutral with } \Delta_\lambda = -2 \end{cases}$$

DCJ-indel model - Path recombinations

Putting together (examples):

$$\overset{\mathbb{AB}_{\mathcal{AB}}}{\mathbb{A}\text{–}\mathcal{AB}\text{–}\mathbb{B}} + \overset{\mathbb{AB}_{\mathcal{BA}}}{\mathbb{A}\text{–}\mathcal{BA}\text{–}\mathbb{B}} = \overset{\mathbb{AB}_{\mathcal{A}}}{\mathbb{A}\text{–}\mathcal{ABBA}\text{–}\mathbb{B}} + \overset{\mathbb{AB}_{\varepsilon}}{\mathbb{A}\text{–}\mathbb{B}} \quad \text{(neutral, with } \Delta_\lambda = -2\text{)}$$

$$\overset{\mathbb{AB}_{\mathcal{AB}}}{\mathbb{A}\text{–}\mathcal{ABAB}\text{–}\mathbb{B}} + \overset{\mathbb{AB}_{\mathcal{BA}}}{\mathbb{A}\text{–}\mathcal{BA}\text{–}\mathbb{B}} = \overset{\mathbb{AB}_{\mathcal{A}}}{\mathbb{A}\text{–}\mathcal{ABABBA}\text{–}\mathbb{B}} + \overset{\mathbb{AB}_{\varepsilon}}{\mathbb{A}\text{–}\mathbb{B}} \quad \text{(neutral, with } \Delta_\lambda = -2\text{)}$$

Λ	λ
0	0
1	1
2	2
3	2
4	3
5	3
6	4
7	4

$$\overset{\mathbb{AB}_{\mathcal{AB}}}{\mathbb{A}\text{–}\mathcal{ABAB}\text{–}\mathbb{B}} + \overset{\mathbb{AB}_{\mathcal{BA}}}{\mathbb{A}\text{–}\mathcal{BABA}\text{–}\mathbb{B}} = \overset{\mathbb{AB}_{\mathcal{A}}}{\mathbb{A}\text{–}\mathcal{ABABBABA}\text{–}\mathbb{B}} + \overset{\mathbb{AB}_{\varepsilon}}{\mathbb{A}\text{–}\mathbb{B}} \quad \text{(neutral, with } \Delta_\lambda = -2\text{)}$$

DCJ-indel model - Path recombinations

Putting together (examples):

$$\overset{\mathbb{AB}_{\mathcal{AB}}}{\mathbb{A}\text{–}\mathcal{AB}\text{–}\mathbb{B}} + \overset{\mathbb{AB}_{\mathcal{AB}}}{\mathbb{A}\text{–}\mathcal{AB}\text{–}\mathbb{B}} = \overset{\mathbb{AB}_{\mathcal{AB}}}{\mathbb{A}\text{–}\mathcal{AB}\mathcal{AB}\text{–}\mathbb{B}} + \overset{\mathbb{AB}_{\varepsilon}}{\mathbb{A}\text{–}\mathbb{B}} \quad \text{(neutral, with } \Delta_\lambda = -1\text{)}$$

$$\overset{\mathbb{AB}_{\mathcal{AB}}}{\mathbb{A}\text{–}\mathcal{AB}\text{–}\mathbb{B}} + \overset{\mathbb{AB}_{\mathcal{AB}}}{\mathbb{A}\text{–}\mathcal{AB}\text{–}\mathbb{B}} = \overset{\mathbb{AB}_{\mathcal{AB}}}{\mathbb{A}\text{–}\mathcal{AB}\mathcal{AB}\text{–}\mathbb{B}} + \overset{\mathbb{AB}_{\varepsilon}}{\mathbb{A}\text{–}\mathbb{B}} \quad \text{(neutral, with } \Delta_\lambda = -1\text{)}$$

$$\overset{\mathbb{AB}_{\mathcal{AB}}}{\mathbb{A}\text{–}\mathcal{AB}\text{–}\mathbb{B}} + \overset{\mathbb{AB}_{\mathcal{BA}}}{\mathbb{A}\text{–}\mathcal{BA}\text{–}\mathbb{B}} = \overset{\mathbb{AB}_{\mathcal{A}}}{\mathbb{A}\text{–}\mathcal{ABBA}\text{–}\mathbb{B}} + \overset{\mathbb{AB}_{\varepsilon}}{\mathbb{A}\text{–}\mathbb{B}} \quad \text{(neutral, with } \Delta_\lambda = -2\text{)}$$

$$\overset{\mathbb{AB}_{\mathcal{AB}}}{\mathbb{A}\text{–}\mathcal{AB}\text{–}\mathbb{B}} + \overset{\mathbb{AB}_{\mathcal{BA}}}{\mathbb{A}\text{–}\mathcal{BA}\text{–}\mathbb{B}} = \overset{\mathbb{AB}_{\mathcal{A}}}{\mathbb{A}\text{–}\mathcal{AA}\text{–}\mathbb{B}} + \overset{\mathbb{AB}_{\mathcal{B}}}{\mathbb{A}\text{–}\mathcal{BB}\text{–}\mathbb{B}} \quad \text{(neutral, with } \Delta_\lambda = -2\text{)}$$

$$\overset{\mathbb{AB}_{\mathcal{AB}}}{\mathbb{A}\text{–}\mathcal{AB}\text{–}\mathbb{B}} + \overset{\mathbb{AB}_{\mathcal{BA}}}{\mathbb{A}\text{–}\mathcal{BA}\text{–}\mathbb{B}} = \overset{\mathbb{AB}_{\mathcal{B}}}{\mathbb{A}\text{–}\mathcal{BAAB}\text{–}\mathbb{B}} + \overset{\mathbb{AB}_{\varepsilon}}{\mathbb{A}\text{–}\mathbb{B}} \quad \text{(neutral, with } \Delta_\lambda = -2\text{)}$$

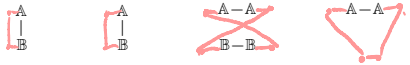
$$\overset{\mathbb{AB}_{\mathcal{AB}}}{\mathbb{A}\text{–}\mathcal{AB}\text{–}\mathbb{B}} + \overset{\mathbb{AB}_{\mathcal{BA}}}{\mathbb{A}\text{–}\mathcal{BA}\text{–}\mathbb{B}} = \overset{\mathbb{AB}_{\mathcal{B}}}{\mathbb{A}\text{–}\mathcal{BB}\text{–}\mathbb{B}} + \overset{\mathbb{AB}_{\mathcal{A}}}{\mathbb{A}\text{–}\mathcal{AA}\text{–}\mathbb{B}} \quad \text{(neutral, with } \Delta_\lambda = -2\text{)}$$

Λ	λ
0	0
1	1
2	2
3	2
4	3
5	3
6	4
7	4

DCJ and DCJ-indel models - Capping

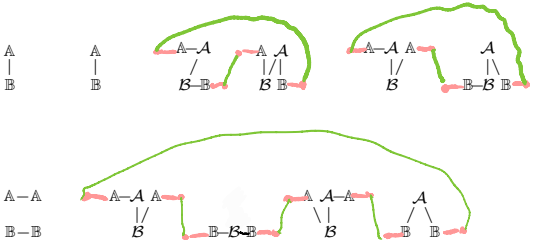
Add caps to close all paths of the graph into cycles, preserving the distance

Canonical capping (no indel edges): maximizes the number of cycles



paths	linking cycle	Δn	Δc	$\Delta(2\mathbb{AB})$	Δ_{DCJ}
$\mathbb{AB}$	$(\mathbb{AB})$	$+0.5$	$+1$	-0.5	0
$\mathbb{AA}+\mathbb{BB}$	$(\mathbb{AA},\mathbb{BB})$	$+1$	$+1$	0	0
$\mathbb{AA}$	$(\mathbb{AA},\Gamma_{\mathbb{B}})$	$+1$	$+1$	0	0
$\mathbb{BB}$	$(\mathbb{BB},\Gamma_{\mathbb{A}})$	$+1$	$+1$	0	0

Singular capping (with indel edges): optimizes the number of cycles and of runs at the same time



paths	linking cycle	Δn	Δc	$\Delta(2\mathbb{AB})$	$\Delta\lambda$	$\Delta_{\mathrm{DCJ}}^{\lambda}$
$\mathbb{AA}_{\mathcal{AB}}+\mathbb{BB}_{\mathcal{AB}}$	$(\mathbb{AA}_{\mathcal{AB}},\mathbb{BB}_{\mathcal{BA}})$	$+1$	$+1$	0	-2	-2
$2\times\mathbb{AA}_{\mathcal{AB}}+\mathbb{BB}_{\mathcal{A}}+\mathbb{BB}_{\mathcal{B}}$	$(\mathbb{AA}_{\mathcal{AB}},\mathbb{BB}_{\mathcal{B}},\mathbb{AA}_{\mathcal{BA}},\mathbb{BB}_{\mathcal{A}})$	$+2$	$+1$	0	-4	-3
$\mathbb{AB}_{\mathcal{AB}}+\mathbb{AB}_{\mathcal{BA}}$	$(\mathbb{AB}_{\mathcal{AB}},\mathbb{AB}_{\mathcal{BA}})$	$+1$	$+1$	-1	-2	-1
$\mathbb{AB}$	$(\mathbb{AB})$	$+0.5$	$+1$	-0.5	0	0
$\mathbb{AA}+\mathbb{BB}$	$(\mathbb{AA},\mathbb{BB})$	$+1$	$+1$	0	0	0

DCJ-indel model 1

1 The indel-potential is defined as...

A the number of runs in a component

B the smallest number of runs that can be obtained after sorting with internal gaining DCJs

C the number of indel-edges in a component

2 The indel-potential of a component depends on..

A its number of runs

B its number of indel-edges

C its length

3 The number of runs in a cycle can be...

A 0,1,2,4,6,8,...

B any non-negative integer

C any positive integer

4 The number of runs in a path can be...

A 0,1,2,4,6,8,...

B any non-negative integer

C any positive integer

DCJ-indel model 2

runs Λ	indel-pot. λ
0	0
1	1
2	2
3	2
4	3
5	3

1 A recombination can reduce the overall number of runs by at most...

A 1 B 2 C 3

2 A recombination can reduce the overall overall indel-potential by at most...

A 1 B 2 C 3

3 A recombination involving a cycle is...

A gaining

B neutral

C losing

$\Delta_{DCJ} = +2$

$\Delta^{ID} \geq -2$

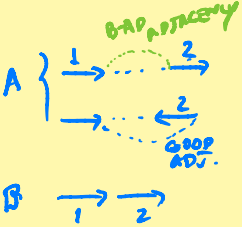
4 A recombination involving a cycle can be...

A deducting $\Delta^{ID}_{DCJ} < 0$

B part of an optimal sorting sequence

C none of those

Review-Quiz 1: Inversion model



1 In the DCJ model any target adjacency can be reconstructed with an optimal sorting step, but the same is not true for the inversion model because...

- A a target adjacency can be bad
- B a target adjacency can be already present in the genome
- C reconstructing a target adjacency can be unsafe

2 A cycle is bad when...

- A it cannot be sorted by inversions
- B it interleaves another bad cycle
- C it contains only bad target adjacencies

3 Which data structure helps finding safe inversions?

- A relational diagram
- B overlap graph
- C component tree

4 A bad component can be fixed...

- A with a neutral inversion
- B with a split inversion
- C with a safe inversion

Inversion model 2

1 Each leaf of the component tree represents...

A a bad component

B a hurdle

C a fortress

2 The cost of covering a component tree can be expressed in terms of...

A the number of bad nodes

B the length of the longest traversal of the tree

C the number of leaves

cost = { # leaves
# leaves + 1 (fortress)

3 Fixing a super hurdle with a neutral inversion

A is a good strategy

B creates a new hurdle

C destroys a good component

4 Merging two good (or trivial) components..

A can merge bad components into a good one

B creates a new bad component

C is never recommended