An Overview of Genomic Distances Modeled with Indels

Marília Braga

Inmetro - Brazil

Overview

1. **Motivation**
2. **Relational Diagram:** $R(A, B)$
   - DCJ distance
   - Inversion distance
   - Related graphs
3. **Handling indels: runs and potentials**
4. **Genomic distances modeled with indels**
   - DCJ-indel
   - DCJ-substitution
   - Inversion-indel
5. **Triangular inequality disruption**

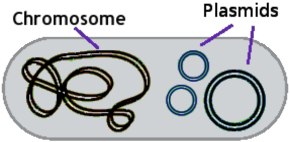
Motivation

Overview

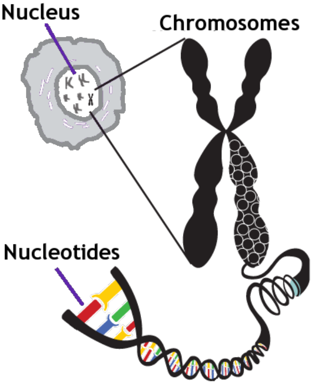
1. Motivation
2. Relational Diagram: $R(A, B)$
   - DCJ distance
   - Inversion distance
   - Related graphs
3. Handling indels: runs and potentials
4. Genomic distances modeled with indels
   - DCJ-indel
   - DCJ-substitution
   - Inversion-indel
5. Triangular inequality disruption

Motivation

Prokaryotic Cell

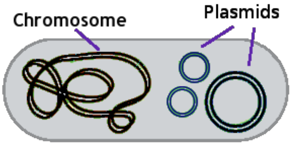


Eukaryotic Cell

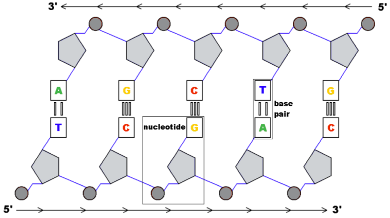


Motivation

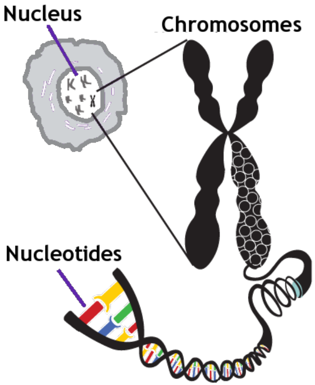
Prokaryotic Cell



DNA antiparallel strands

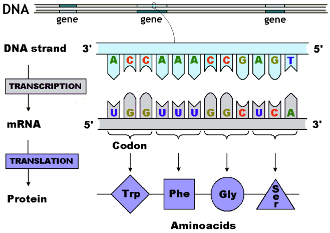


Eukaryotic Cell



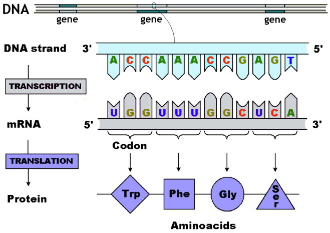
Motivation

Genes are DNA fragments
that code for proteins:

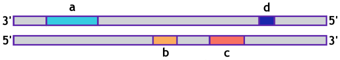


Motivation

Genes are DNA fragments that code for proteins:

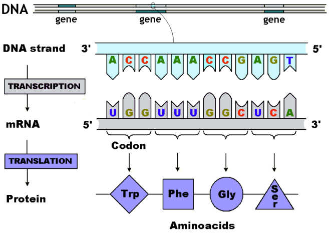


The strand in which each gene lies gives its orientation:

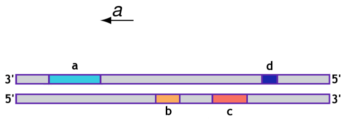


Motivation

Genes are DNA fragments that code for proteins:

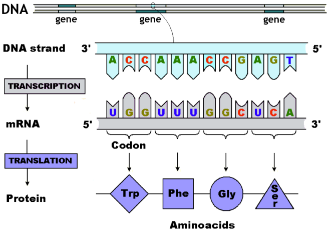


The strand in which each gene lies gives its orientation:

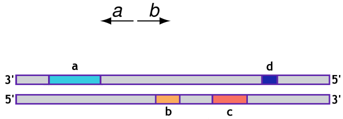


Motivation

Genes are DNA fragments
that code for proteins:

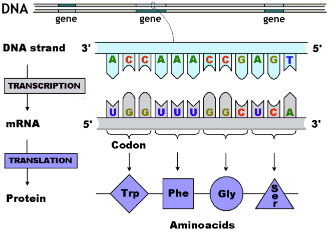


The strand in which each gene
lies gives its orientation:

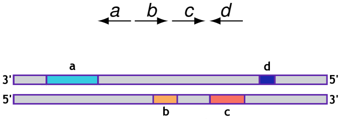


Motivation

Genes are DNA fragments that code for proteins:

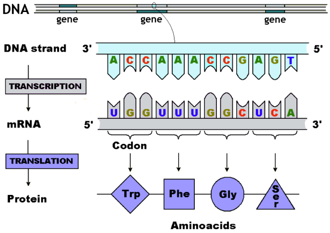


The strand in which each gene lies gives its orientation:

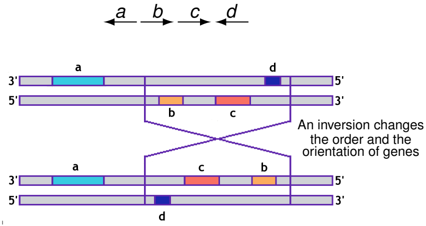


Motivation

Genes are DNA fragments that code for proteins:



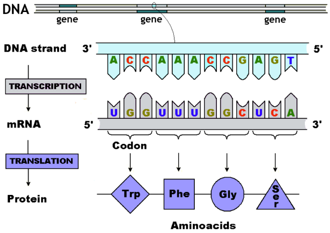
The strand in which each gene lies gives its orientation:



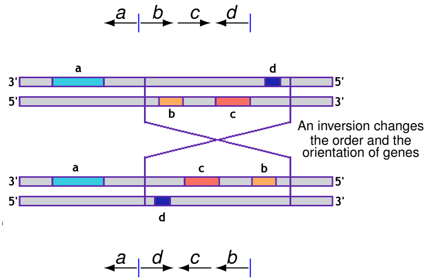
An inversion changes the order and the orientation of genes

Motivation

Genes are DNA fragments that code for proteins:



The strand in which each gene lies gives its orientation:



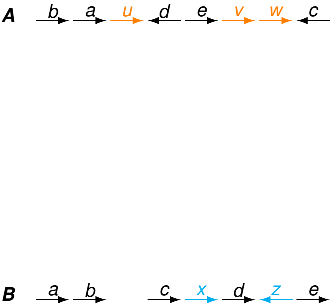
An inversion changes the order and the orientation of genes

Motivation

Comparing genomes with unequal contents

Common genes: $\mathcal{G} = \{a, b, c, d, e\}$

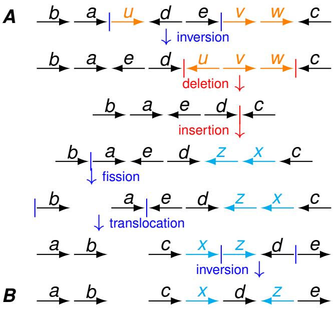
Unique genes: $\mathcal{A} = \{u, v, w\}$ $\mathcal{B} = \{x, z\}$



Comparing genomes with unequal contents

Common genes: $\mathcal{G} = \{a, b, c, d, e\}$

Unique genes: $\mathcal{A} = \{u, v, w\}$ $\mathcal{B} = \{x, z\}$



Motivation

Comparing genomes with unequal contents

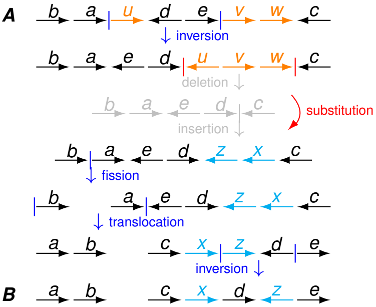
Common genes:

$\mathcal{G} = \{a, b, c, d, e\}$

Unique genes:

$\mathcal{A} = \{u, v, w\}$

$\mathcal{B} = \{x, z\}$



Motivation

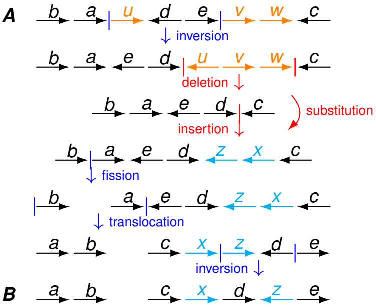
Comparing genomes with unequal contents

Common genes:

$\mathcal{G} = \{a, b, c, d, e\}$

Unique genes:

$\mathcal{A} = \{u, v, w\}$
$\mathcal{B} = \{x, z\}$

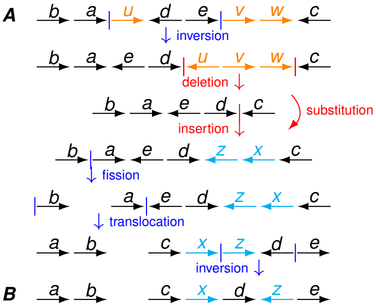


Insertions and *Deletions* - (Indels) or *Substitutions* change the content of the genome

Comparing genomes with unequal contents

Common genes: $\mathcal{G} = \{a, b, c, d, e\}$

Unique genes: $\mathcal{A} = \{u, v, w\}$
$\mathcal{B} = \{x, z\}$



Insertions and *Deletions* - (Indels) or *Substitutions* change the content of the genome

Rearrangements change the organization of the genome and are modeled by the *Double Cut and Join* - (DCJ)

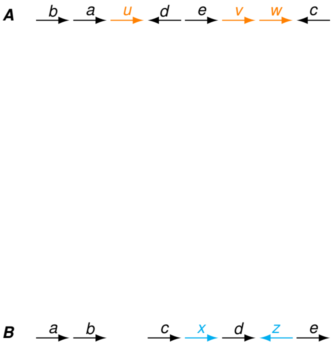
(Yancopoulos, Attie and Friedberg, 2005)

Relational Diagram: $R(A, B)$

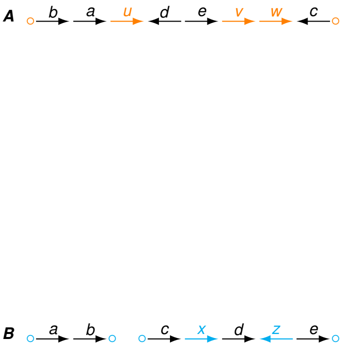
Overview

1. Motivation
2. Relational Diagram: $R(A, B)$
   - DCJ distance
   - Inversion distance
   - Related graphs
3. Handling indels: runs and potentials
4. Genomic distances modeled with indels
   - DCJ-indel
   - DCJ-substitution
   - Inversion-indel
5. Triangular inequality disruption

Relational Diagram: $R(A, B)$

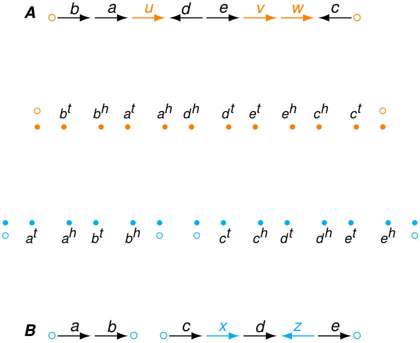


Relational Diagram: $R(A, B)$



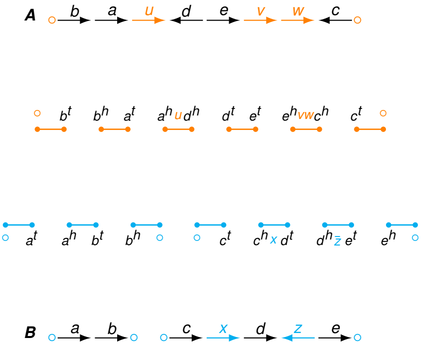
(The symbol ○ represents the telomeres in both genomes.)

Relational Diagram: $R(A, B)$



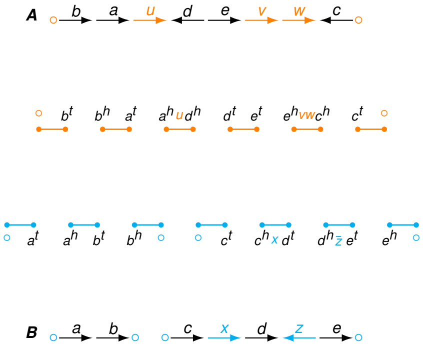
(The symbol ○ represents the telomeres in both genomes.)

Relational Diagram: $R(A, B)$


A ○ b a u d e v w c ○
○ b^t b^h a^t a^h u d^h d^t e^t e^h vw c^h c^t ○
○ a^t a^h b^t b^h ○ ○ c^t c^h x d^t d^h $\bar{z}$ e^t e^h ○
B ○ a b ○ ○ c x d z e ○

(The symbol ○ represents the telomeres in both genomes.)

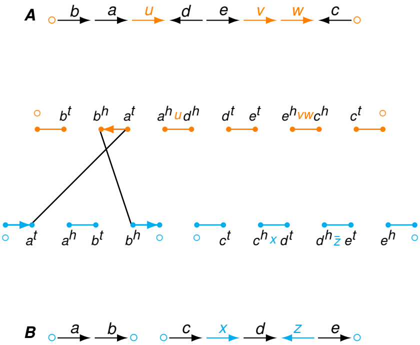
Relational Diagram: $R(A, B)$


A
b a u d e v w c
b^t b^h a^t $a^h u d^h$ d^t e^t $e^h vw c^h$ c^t
a^t a^h b^t b^h c^t $c^h x d^t$ $d^h \bar{z} e^t$ e^h
B
a b c x d z e

Components of $R(A, B)$:

(The symbol ○ represents the telomeres in both genomes.)

Relational Diagram: $R(A, B)$



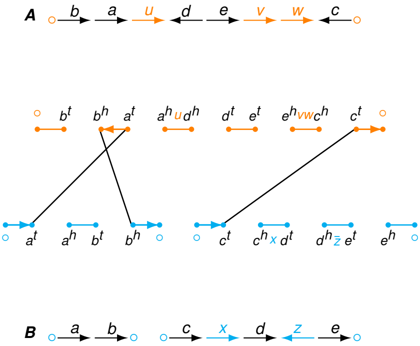
A
b a u d e v w c
○ b^t b^h a^t a^h u d^h d^t e^t e^h vw c^h c^t ○
○ a^t a^h b^t b^h ○ ○ c^t c^h x d^t d^h z e^t e^h ○
B
a b c x d z e

Components of $R(A, B)$:

One clean *BB*-path

(The symbol ○ represents the telomeres in both genomes.)

Relational Diagram: $R(A, B)$



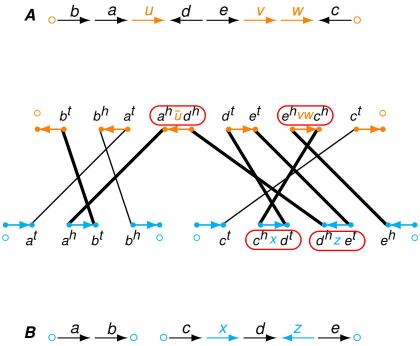
Components of *R*(*A*, *B*):

One clean *BB*-path

One clean *AB*-path

(The symbol ○ represents the telomeres in both genomes.)

Relational Diagram: $R(A, B)$



A
b a u d e v w c
b^t b^h a^t $a^h\bar{u}d^h$ d^t e^t $e^h vw c^h$ c^t
a^t a^h b^t b^h c^t $c^h x d^t$ $d^h z e^t$ e^h
B
a b c x d z e

Components of $R(A, B)$:

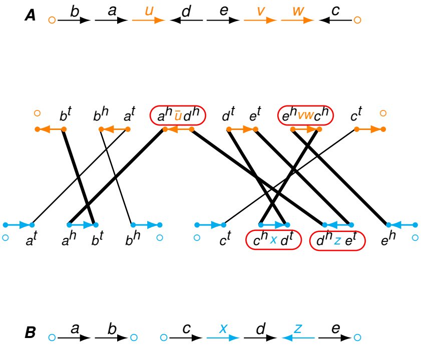
One clean BB-path

One clean AB-path

One AB-path with four labels

(The symbol ○ represents the telomeres in both genomes.)

Relational Diagram: $R(A, B)$


A
b a u d e v w c
b^t b^h a^t $a^h \bar{u} d^h$ d^t e^t $e^h vw c^h$ c^t
a^t a^h b^t b^h c^t $c^h x d^t$ $d^h z e^t$ e^h
B
a b c x d z e

Components of $R(A, B)$:

One clean BB-path

One clean AB-path

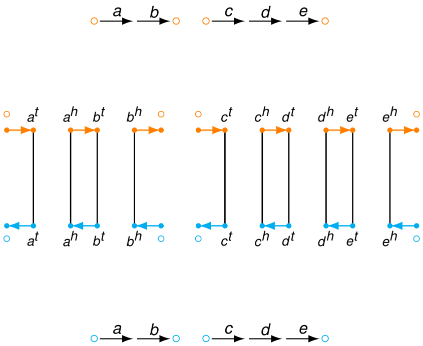
One AB-path with four labels

(collection of paths and cycles;
the number of AB-paths is even)

(The symbol ∘ represents the telomeres in both genomes.)

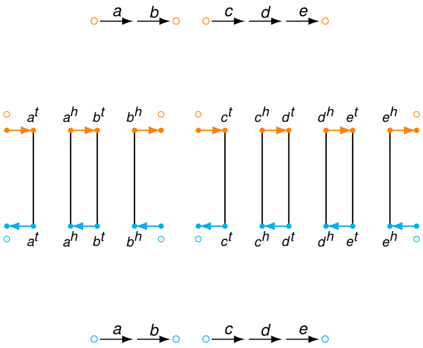
Relational Diagram: $R(A, B)$

For identical (or sorted) genomes...



Relational Diagram: $R(A, B)$

For identical (or sorted) genomes...

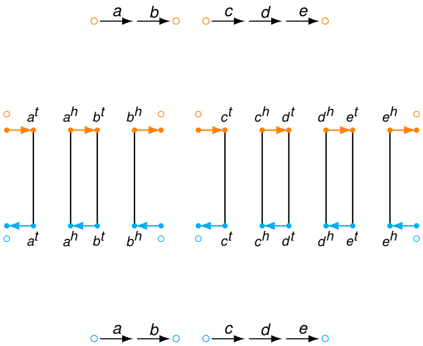


Components of $R(A, B)$:

Only short cycles and *AB*-paths

Relational Diagram: $R(A, B)$

For identical (or sorted) genomes...



Components of *R*(*A*, *B*):

Only short cycles and *AB*-paths

(rearrangements need to increase the number of components)

Relational Diagram: $R(A, B)$

DCJ distance

$\mathcal{G}$: set of common markers of A and B

c: number of cycles in $R(A, B)$

b: number of AB-paths in $R(A, B)$

Types of rearrangements:

rearrangement	effect on $R(A, B)$
optimal (split)	increase c or b
neutral	c and b unchanged
counter-optimal (joint)	decrease c or b

Relational Diagram: $R(A, B)$

DCJ distance

$\mathcal{G}$: set of common markers of A and B

c: number of cycles in $R(A, B)$

b: number of AB-paths in $R(A, B)$

Types of rearrangements:

rearrangement	effect on $R(A, B)$
optimal (split)	increase c or b
neutral	c and b unchanged
counter-optimal (joint)	decrease c or b

Bergeron *et al.* (2006): there is an optimal DCJ at each sorting step.

Relational Diagram: $R(A, B)$

DCJ distance

$\mathcal{G}$: set of common markers of A and B

c: number of cycles in $R(A, B)$

b: number of AB-paths in $R(A, B)$

Types of rearrangements:

rearrangement	effect on $R(A, B)$
optimal (split)	increase c or b
neutral	c and b unchanged
counter-optimal (joint)	decrease c or b

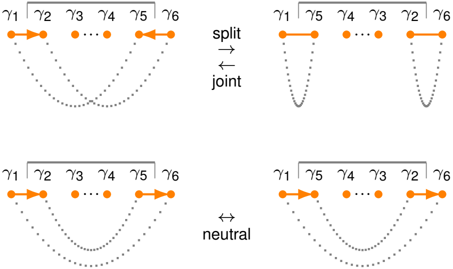
Bergeron *et al.* (2006): there is an optimal DCJ at each sorting step.

DCJ distance of A and B:

$$\boldsymbol{d_{\text{DCJ}}(A, B) = |\mathcal{G}| - (c + \tfrac{b}{2})}$$

Relational Diagram: $R(A, B)$

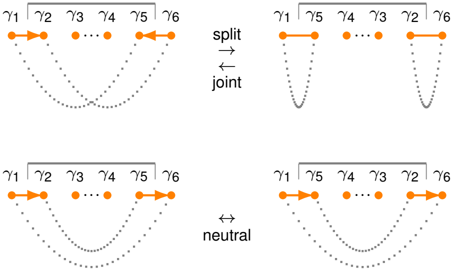
Inversion distance


γ_1 γ_2 γ_3 γ_4 γ_5 γ_6
split
→
←
joint
γ_1 γ_5 γ_4 γ_3 γ_2 γ_6
γ_1 γ_2 γ_3 γ_4 γ_5 γ_6
↔
neutral
γ_1 γ_5 γ_4 γ_3 γ_2 γ_6

An inversion only creates a new cycle if applied to edges of the same component and with opposite orientations.

Relational Diagram: $R(A, B)$

Inversion distance


γ_1 γ_2 γ_3 γ_4 γ_5 γ_6
$\cdots$
split
$\rightarrow$
$\leftarrow$
joint
γ_1 γ_5 γ_4 γ_3 γ_2 γ_6
$\cdots$
γ_1 γ_2 γ_3 γ_4 γ_5 γ_6
$\cdots$
$\leftrightarrow$
neutral
γ_1 γ_5 γ_4 γ_3 γ_2 γ_6
$\cdots$

An inversion only creates a new cycle if applied to edges of the same component and with opposite orientations.

The inversion distance is lower bounded by the DCJ distance:

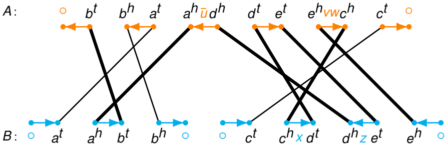
$$\boldsymbol{d_{INV}(A, B) \geq d_{\mathrm{DCJ}}(A, B)}$$

(Hannenhalli and Pevzner (1995): the exact inversion distance can be efficiently computed.)

Relational Diagram: $R(A, B)$

Related graphs

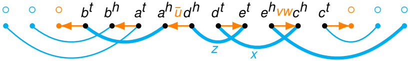
Relational diagram


$A:$
○ b^t b^h a^t a^h $\bar{u}$ d^h d^t e^t e^h v w c^h c^t ○
$B:$
○ a^t a^h b^t b^h ○ ○ c^t c^h x d^t d^h z e^t e^h ○

Relational Diagram: $R(A, B)$

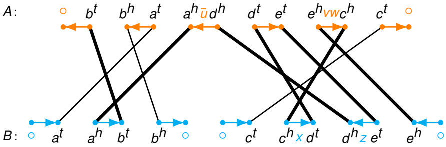
Related graphs

Breakpoint diagram (Bafna and Pevzner, 1993)



b^t b^h a^t a^h $\bar{u}$ d^h d^t e^t e^h vw c^h c^t
z
x

Relational diagram

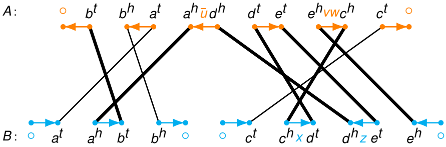


$A:$
b^t b^h a^t a^h $\bar{u}$ d^h d^t e^t e^h vw c^h c^t
$B:$
a^t a^h b^t b^h c^t c^h x d^t d^h z e^t e^h

Relational Diagram: $R(A, B)$

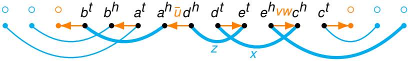
Related graphs

Relational diagram



A:
b^t b^h a^t a^h $\bar{u}$ d^h d^t e^t e^h vw c^h c^t
B:
a^t a^h b^t b^h c^t c^h x d^t d^h z e^t e^h

Breakpoint diagram (Bafna and Pevzner, 1993)



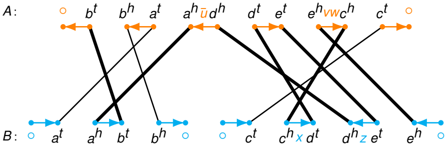
b^t b^h a^t a^h $\bar{u}$ d^h d^t e^t e^h vw c^h c^t
z
x

(Asymmetric, **identifies inversions**)

Relational Diagram: $R(A, B)$

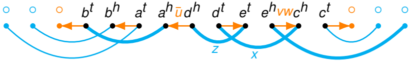
Related graphs

Relational diagram



A:
○ b^t b^h a^t $a^h\bar{u}d^h$ d^t e^t $e^h vw c^h$ c^t ○
B:
○ a^t a^h b^t b^h ○ ○ c^t $c^h x d^t$ $d^h z e^t$ e^h ○

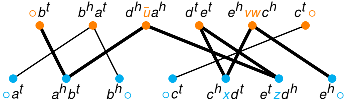
Breakpoint diagram (Bafna and Pevzner, 1993)



○ ○ ○ b^t b^h a^t $a^h\bar{u}d^h$ d^t e^t $e^h vw c^h$ c^t ○ ○ ○
z
x

(Asymmetric, **identifies inversions**)

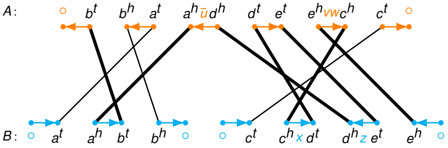
Adjacency graph (Bergeron *et al.*, 2006)



Relational Diagram: $R(A, B)$

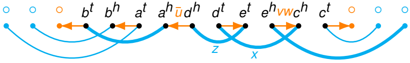
Related graphs

Relational diagram



A:
○ b^t b^h a^t a^h $\bar{u}$ d^h d^t e^t e^h vw c^h c^t ○
B: ○ a^t a^h b^t b^h ○ ○ c^t c^h x d^t d^h z e^t e^h ○

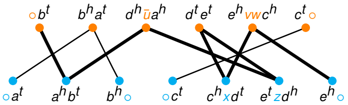
Breakpoint diagram (Bafna and Pevzner, 1993)



○ ○ ○ b^t b^h a^t a^h $\bar{u}$ d^h d^t e^t e^h vw c^h c^t ○ ○ ○
z
x

(Asymmetric, **identifies inversions**)

Adjacency graph (Bergeron *et al.*, 2006)



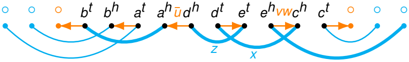
○ b^t b^h a^t d^h $\bar{u}$ a^h d^t e^t e^h vw c^h c^t ○
○ a^t a^h b^t b^h ○ ○ c^t c^h x d^t e^t z d^h e^h ○

(**Symmetric**, does not identify inversions)

Relational Diagram: $R(A, B)$

Related graphs

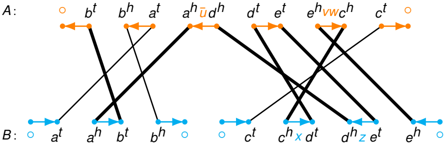
Breakpoint diagram (Bafna and Pevzner, 1993)



b^t b^h a^t a^h $\bar{u}$ d^h d^t e^t e^h vw c^h c^t
z
x

(Asymmetric, **identifies inversions**)

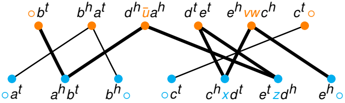
Relational diagram



A:
b^t b^h a^t a^h $\bar{u}$ d^h d^t e^t e^h vw c^h c^t
B:
a^t a^h b^t b^h c^t c^h x d^t d^h z e^t e^h

(**Symmetric**, **identifies inversions**)

Adjacency graph (Bergeron *et al.*, 2006)



b^t $b^h a^t$ $d^h \bar{u} a^h$ $d^t e^t$ $e^h vw c^h$ c^t
a^t $a^h b^t$ b^h c^t $c^h x d^t$ $e^t z d^h$ e^h

(**Symmetric**, does not identify inversions)

The relational diagram has the same components as the breakpoint diagram and the adjacency graph

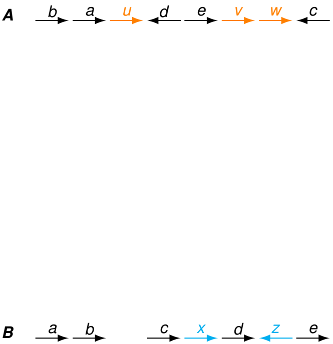
Handling indels: runs and potentials

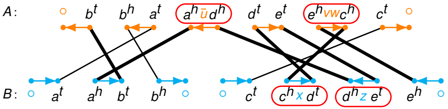
Overview

1. Motivation
2. Relational Diagram: $R(A, B)$
   - DCJ distance
   - Inversion distance
   - Related graphs
3. **Handling indels: runs and potentials**
4. Genomic distances modeled with indels
   - DCJ-indel
   - DCJ-substitution
   - Inversion-indel
5. Triangular inequality disruption

Handling indels: runs and potentials

The symmetry helps to accumulate labels in both genomes:

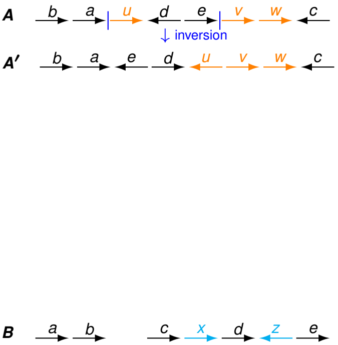


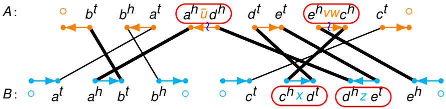


one *BB*-path, two *AB*-paths, and four labels

Handling indels: runs and potentials

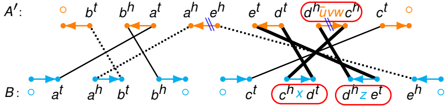
The symmetry helps to accumulate labels in both genomes:




A:
b^t b^h a^t $a^h \bar{u} d^h$ d^t e^t $e^h v w c^h$ c^t
B:
a^t a^h b^t b^h c^t $c^h x d^t$ $d^h z e^t$ e^h

one *BB*-path, two *AB*-paths, and four labels

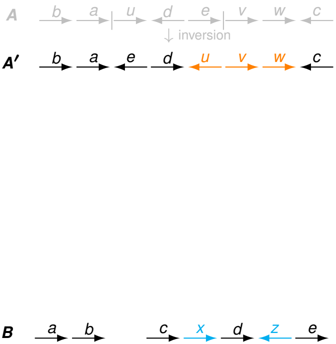
↓

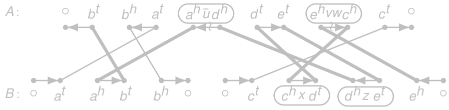

A′:
b^t b^h a^t a^h e^h e^t d^t $d^h \bar{u} v w c^h$ c^t
B:
a^t a^h b^t b^h c^t $c^h x d^t$ $d^h z e^t$ e^h

one *BB*-path, two *AB*-paths, one cycle and three labels

Handling indels: runs and potentials

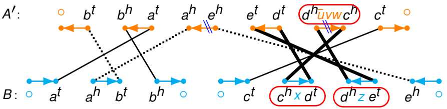
The symmetry helps to accumulate labels in both genomes:




A:
○ b^t b^h a^t $a^h \bar{u} d^h$ d^t e^t $e^h vw c^h$ c^t ○
B:
○ a^t a^h b^t b^h ○ ○ c^t $c^h x d^t$ $d^h z e^t$ e^h ○

one *BB*-path, two *AB*-paths, and four labels

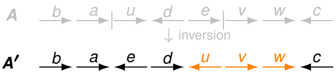
↓

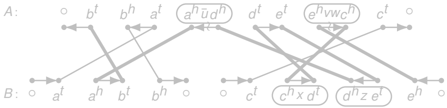

A′:
○ b^t b^h a^t a^h e^h e^t d^t $d^h \bar{u} vw c^h$ c^t ○
B:
○ a^t a^h b^t b^h ○ ○ c^t $c^h x d^t$ $d^h z e^t$ e^h ○

one *BB*-path, two *AB*-paths, one cycle and three labels

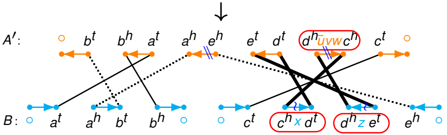
Handling indels: runs and potentials

The symmetry helps to accumulate labels in both genomes:

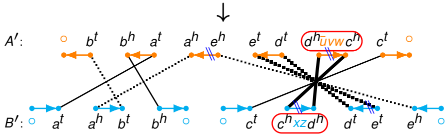



A:
○ b^t b^h a^t $a^h\bar{u}d^h$ d^t e^t e^hvwc^h c^t ○
B:
○ a^t a^h b^t b^h ○ ○ c^t c^hxd^t d^hze^t e^h ○

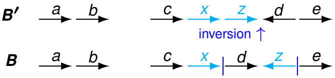
one *BB*-path, two *AB*-paths, and four labels


↓
A′:
○ b^t b^h a^t a^h e^h e^t d^t $d^h\bar{u}vwc^h$ c^t ○
B:
○ a^t a^h b^t b^h ○ ○ c^t c^hxd^t d^hze^t e^h ○

one *BB*-path, two *AB*-paths, one cycle and three labels

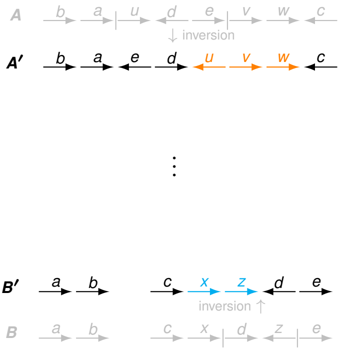

↓
A′:
○ b^t b^h a^t a^h e^h e^t d^t $d^h\bar{u}vwc^h$ c^t ○
B′:
○ a^t a^h b^t b^h ○ ○ c^t c^hxzd^h d^t e^t e^h ○

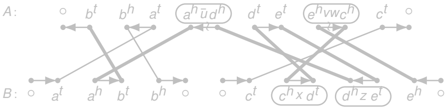
one *BB*-path, two *AB*-paths, two cycles and two labels



Handling indels: runs and potentials

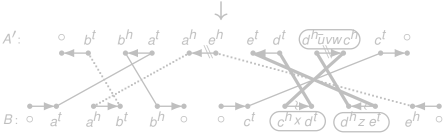
The symmetry helps to accumulate labels in both genomes:





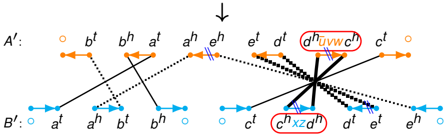
A:
○ b^t b^h a^t $a^h\bar{u}d^h$ d^t e^t e^hvwc^h c^t ○
B:
○ a^t a^h b^t b^h ○ ○ c^t c^hxd^t d^hze^t e^h ○

one *BB*-path, two *AB*-paths, and four labels



↓
A′:
○ b^t b^h a^t a^h e^h e^t d^t $d^h\bar{u}vwc^h$ c^t ○
B:
○ a^t a^h b^t b^h ○ ○ c^t c^hxd^t d^hze^t e^h ○

one *BB*-path, two *AB*-paths, one cycle and three labels

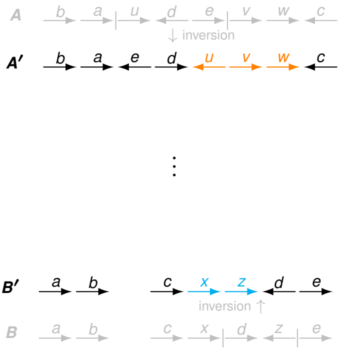


↓
A′:
○ b^t b^h a^t a^h e^h e^t d^t $d^h\bar{u}vwc^h$ c^t ○
B′:
○ a^t a^h b^t b^h ○ ○ c^t c^hxzd^h d^t e^t e^h ○

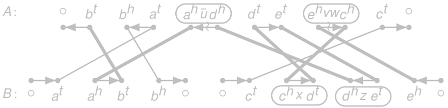
one *BB*-path, two *AB*-paths, two cycles and two labels

Handling indels: runs and potentials

The symmetry helps to accumulate labels in both genomes:

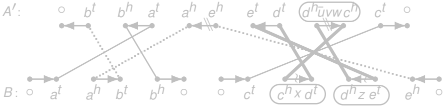


(Rearrangements can increase the number of components and accumulate labels.)


A:
○ b^t b^h a^t $a^h\bar{u}d^h$ d^t e^t e^hvwc^h c^t ○
B:
○ a^t a^h b^t b^h ○ ○ c^t c^hxd^t d^hze^t e^h ○

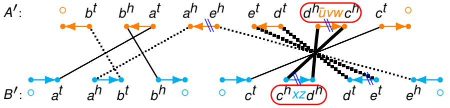
one *BB*-path, two *AB*-paths, and four labels

↓


A′:
○ b^t b^h a^t a^h e^h e^t d^t $d^h\bar{u}vwc^h$ c^t ○
B:
○ a^t a^h b^t b^h ○ ○ c^t c^hxd^t d^hze^t e^h ○

one *BB*-path, two *AB*-paths, one cycle and three labels

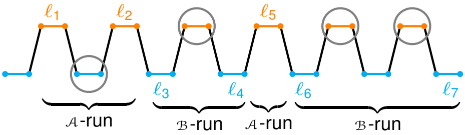
↓


A′:
○ b^t b^h a^t a^h e^h e^t d^t $d^h\bar{u}vwc^h$ c^t ○
B′:
○ a^t a^h b^t b^h ○ ○ c^t c^hxzd^h d^t e^t e^h ○

one *BB*-path, two *AB*-paths, two cycles and two labels

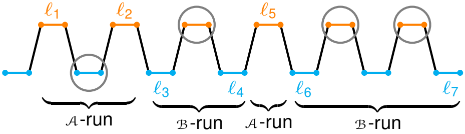
Handling indels: runs and potentials

Runs:



Handling indels: runs and potentials

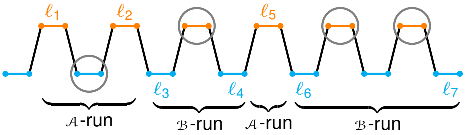
Runs:



$\Lambda = 4$

Handling indels: runs and potentials

Runs:

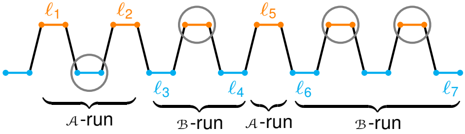


$\Lambda = 4$

(Each **run** can be entirely **accumulated** into a single label with optimal rearrangements.)

Handling indels: runs and potentials

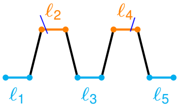
Runs:



$\Lambda = 4$

(Each **run** can be entirely **accumulated** into a single label with optimal rearrangements.)

A rearrangement can merge at most two $\mathcal{A}$-runs and two $\mathcal{B}$-runs:

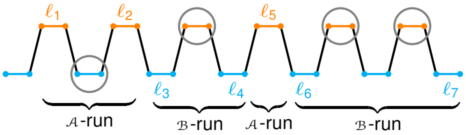


→

Λ : 5 *runs*

Handling indels: runs and potentials

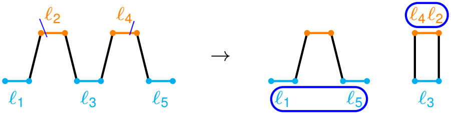
Runs:



$\Lambda = 4$

(Each **run** can be entirely **accumulated** into a single label with optimal rearrangements.)

A rearrangement can merge at most two $\mathcal{A}$-runs and two $\mathcal{B}$-runs:



Λ : 5 *runs* 1 $+$ 2 *runs* ($\Delta\Lambda = -2$)

Handling indels: runs and potentials

Potentials:

Handling indels: runs and potentials

Potentials:

Indel-potential of a component P [WABI 2010]

Minimum number of **runs** obtained splitting P with **optimal** rearrangements:

$$\lambda(P) = \left\lceil \frac{\Lambda(P)+1}{2} \right\rceil \qquad (\textit{for } \Lambda(P) \geq 1)$$

Handling indels: runs and potentials

Potentials:

Indel-potential of a component P [WABI 2010]

Minimum number of **runs** obtained splitting P with **optimal** rearrangements:

$$\lambda(P) = \left\lceil \frac{\Lambda(P) + 1}{2} \right\rceil \qquad (\textit{for } \Lambda(P) \geq 1)$$

Substitution-potential of a component P [RECOMB-CG 2011]

Minimum number of **pairs of runs** obtained splitting P with **optimal** rearrangements:

$$\sigma(P) = \left\lceil \frac{\Lambda(P) + 1}{4} \right\rceil \qquad (\textit{for } \Lambda(P) \geq 1)$$

Handling indels: runs and potentials

Potentials:

Indel-potential of a component P [WABI 2010]

Minimum number of **runs** obtained splitting P with **optimal** rearrangements:

$$\lambda(P) = \left\lceil \frac{\Lambda(P)+1}{2} \right\rceil \qquad (\textit{for } \Lambda(P) \geq 1)$$

Substitution-potential of a component P [RECOMB-CG 2011]

Minimum number of **pairs of runs** obtained splitting P with **optimal** rearrangements:

$$\sigma(P) = \left\lceil \frac{\Lambda(P)+1}{4} \right\rceil \qquad (\textit{for } \Lambda(P) \geq 1)$$

$\Lambda(P)$	$\lambda(P)$	$\sigma(P)$
0	0	0
1	1	1
2 3	2 2	1 1
4 5	3 3	2 2
6 7	4 4	2 2
$\vdots$	$\lceil \frac{\Lambda(P)+1}{2} \rceil$	$\lceil \frac{\Lambda(P)+1}{4} \rceil$

Genomic distances modeled with indels

Overview

1. **Motivation**
2. **Relational Diagram:** $R(A, B)$
   - DCJ distance
   - Inversion distance
   - Related graphs
3. **Handling indels: runs and potentials**
4. **Genomic distances modeled with indels**
   - DCJ-indel
   - DCJ-substitution
   - Inversion-indel
5. **Triangular inequality disruption**

Genomic distances modeled with indels

DCJ-indel distance

We can assign distinct costs to DCJ and indel operations, such that the **indel cost** is upper bounded by the **DCJ cost** [WABI 2012]:

DCJ costs **1**

indel costs $\boldsymbol{w} \leq \mathbf{1}$

Genomic distances modeled with indels

DCJ-indel distance

We can assign distinct costs to DCJ and indel operations, such that the **indel cost** is upper bounded by the **DCJ cost** [WABI 2012]:

DCJ costs **1**

indel costs $\boldsymbol{w} \leq \mathbf{1}$

An **upper bound** for the **DCJ-indel distance** is given by:

$$d^{id}_{\mathrm{DCJ}}(A, B) \leq d_{\mathrm{DCJ}}(A, B) \quad + \quad w \sum_{P \in R(A,B)} \lambda(P)$$

Genomic distances modeled with indels

DCJ-indel distance

We can assign distinct costs to DCJ and indel operations, such that the **indel cost** is upper bounded by the **DCJ cost** [WABI 2012]:

DCJ costs **1**

indel costs $\boldsymbol{w} \leq \mathbf{1}$

An **upper bound** for the **DCJ-indel distance** is given by:

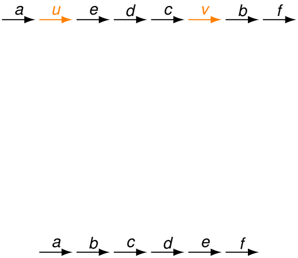
$$d^{id}_{\mathrm{DCJ}}(A, B) \leq d_{\mathrm{DCJ}}(A, B) \quad + \quad w \sum_{P \in R(A,B)} \lambda(P)$$

For any $w \leq 1$, the exact **DCJ-indel distance** can be computed in **linear time**.
[WABI 2010 and 2012]

Genomic distances modeled with indels

DCJ-indel distance

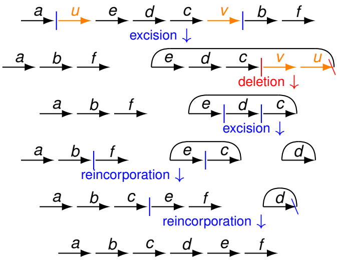
General DCJ-indel model



Genomic distances modeled with indels

DCJ-indel distance

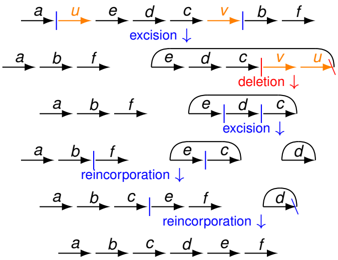
General DCJ-indel model



Genomic distances modeled with indels

DCJ-indel distance

General DCJ-indel model

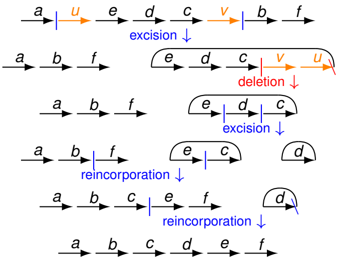


Many circular chromosomes can coexist in the intermediate species.

Genomic distances modeled with indels

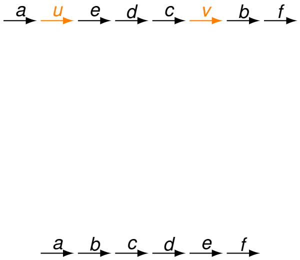
DCJ-indel distance

General DCJ-indel model



Many circular chromosomes can coexist in the intermediate species.

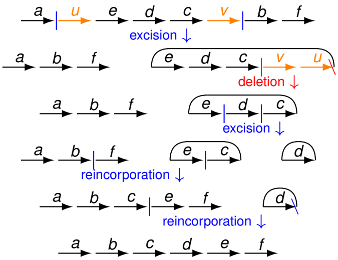
Restricted DCJ-indel model



Genomic distances modeled with indels

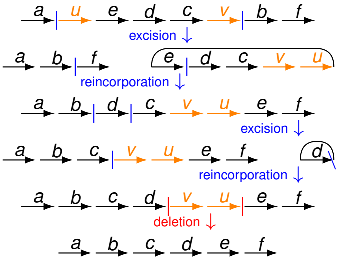
DCJ-indel distance

General DCJ-indel model



Many circular chromosomes can coexist in the intermediate species.

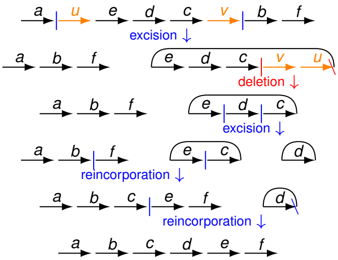
Restricted DCJ-indel model



Genomic distances modeled with indels

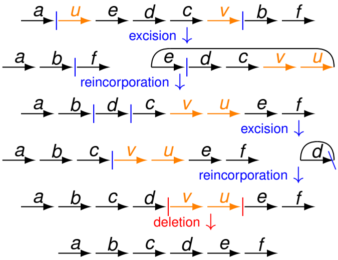
DCJ-indel distance

General DCJ-indel model



Many circular chromosomes can coexist in the intermediate species.

Restricted DCJ-indel model

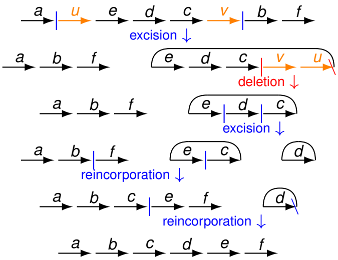


A circular chromosome is immediately reincorporated after its excision.

Genomic distances modeled with indels

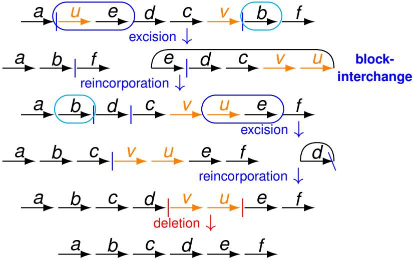
DCJ-indel distance

General DCJ-indel model



Many circular chromosomes can coexist in the intermediate species.

Restricted DCJ-indel model

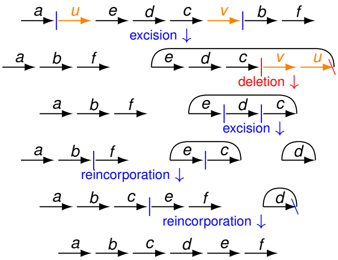


A circular chromosome is immediately reincorporated after its excision.

Genomic distances modeled with indels

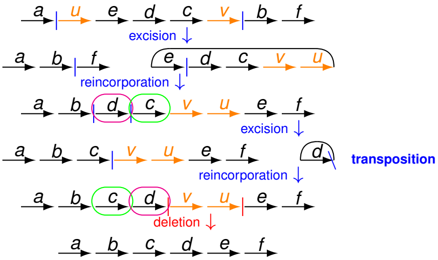
DCJ-indel distance

General DCJ-indel model



Many circular chromosomes can coexist in the intermediate species.

Restricted DCJ-indel model

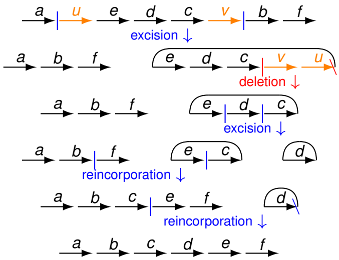


A circular chromosome is immediately reincorporated after its excision.

Genomic distances modeled with indels

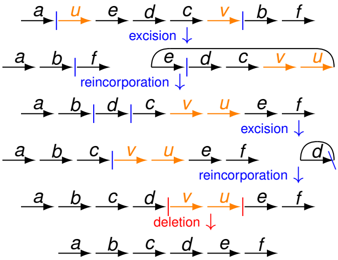
DCJ-indel distance

General DCJ-indel model



Many circular chromosomes can coexist
in the intermediate species.

Restricted DCJ-indel model



A circular chromosome is immediately
reincorporated after its excision.

Both the general and the restricted DCJ-indel distances are the same.
[submitted to BSB 2013]

Genomic distances modeled with indels

DCJ-substitution distance

We can assign distinct costs to DCJ and substitution operations, such that the **substitution cost** is upper bounded by the **DCJ cost** [BSB 2012]:

DCJ costs **1**

substitution costs $\boldsymbol{w} \leq \mathbf{1}$

Genomic distances modeled with indels

DCJ-substitution distance

We can assign distinct costs to DCJ and substitution operations, such that the **substitution cost** is upper bounded by the **DCJ cost** [BSB 2012]:

DCJ costs **1**

substitution costs $\boldsymbol{w} \leq \mathbf{1}$

An **upper bound** for the **DCJ-substitution distance** is given by:

$$\boldsymbol{d^{sb}_{\mathrm{DCJ}}(A, B) \leq d_{\mathrm{DCJ}}(A, B) \quad + \quad w \sum_{P \in R(A,B)} \sigma(P)}$$

Genomic distances modeled with indels

DCJ-substitution distance

We can assign distinct costs to DCJ and substitution operations, such that the **substitution cost** is upper bounded by the **DCJ cost** [BSB 2012]:

DCJ costs **1**

substitution costs $\boldsymbol{w} \leq \mathbf{1}$

An **upper bound** for the **DCJ-substitution distance** is given by:

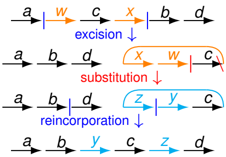
$$d^{sb}_{\mathrm{DCJ}}(A,B) \leq d_{\mathrm{DCJ}}(A,B) \quad + \quad w \sum_{P \in R(A,B)} \sigma(P)$$

For any $w \leq 1$, the exact **DCJ-substitution distance** can be computed in **linear time** [RECOMB-CG 2011 and BSB 2012]

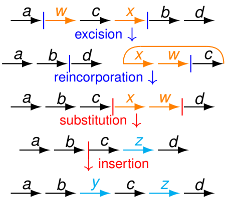
DCJ-substitution distance

The general and the restricted DCJ-substitution distances are not the same:

General DCJ-subtitution model



Restricted DCJ-subtitution model

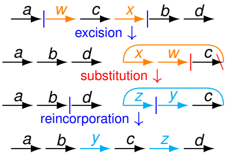


Genomic distances modeled with indels

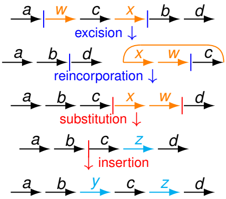
DCJ-substitution distance

The general and the restricted DCJ-substitution distances are not the same:

General DCJ-subtitution model



Restricted DCJ-subtitution model



The restricted version of the DCJ-substitution distance is a complete **open problem**.

Inversion-indel distance

The **same cost** is assigned to **inversions** and **indels**.

Genomic distances modeled with indels

Inversion-indel distance

The **same cost** is assigned to **inversions** and **indels**.

El-Mabrouk, 2001:

- An exact algorithm for the asymmetric case in which only one indel direction is allowed (when we have only insertions or only deletions).
- A heuristic for the symmetric case.

Genomic distances modeled with indels

Inversion-indel distance

The **same cost** is assigned to **inversions** and **indels**.

El-Mabrouk, 2001:

- An exact algorithm for the asymmetric case in which only one indel direction is allowed (when we have only insertions or only deletions).
- A heuristic for the symmetric case.

Our recent results [submitted to RECOMB-CG 2013]:

- With the help of the relational diagram, we developed an exact algorithm for the symmetric case, but only when the genomes can be sorted with split inversions.
- An upper bound for the symmetric case, when the genomes require neutral or joint inversions to be sorted. (An exact algorithm for this case remains an **open problem**.)

Genomic distances modeled with indels

Inversion-indel distance

The **same cost** is assigned to **inversions** and **indels**.

El-Mabrouk, 2001:

- An exact algorithm for the asymmetric case in which only one indel direction is allowed (when we have only insertions or only deletions).
- A heuristic for the symmetric case.

Our recent results [submitted to RECOMB-CG 2013]:

- With the help of the relational diagram, we developed an exact algorithm for the symmetric case, but only when the genomes can be sorted with split inversions.
- An upper bound for the symmetric case, when the genomes require neutral or joint inversions to be sorted. (An exact algorithm for this case remains an **open problem**.)

Extending the model to allow distinct inversion and indel costs has not yet been studied.

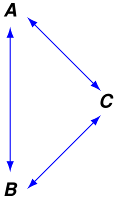
Triangular inequality disruption

Overview

1. **Motivation**
2. **Relational Diagram:** $R(A, B)$
   - DCJ distance
   - Inversion distance
   - Related graphs
3. **Handling indels: runs and potentials**
4. **Genomic distances modeled with indels**
   - DCJ-indel
   - DCJ-substitution
   - Inversion-indel
5. **Triangular inequality disruption**

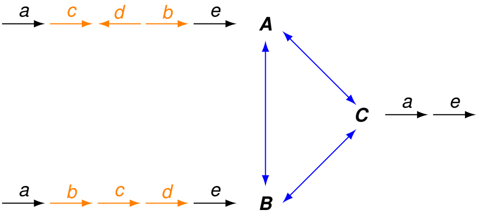
Triangular inequality disruption

Triangular inequality: $d(A,B) \leq d(A,C) + d(B,C)$



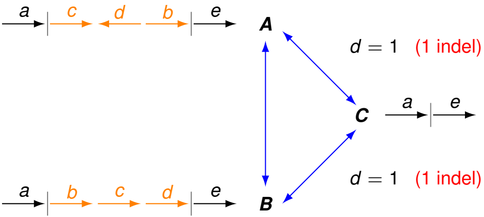
Triangular inequality disruption

Triangular inequality: $d(A, B) \leq d(A, C) + d(B, C)$



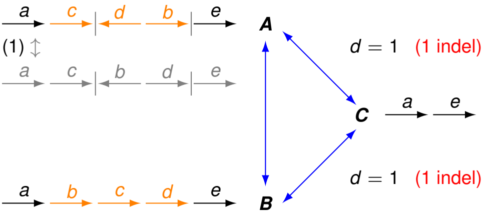
Triangular inequality disruption

Triangular inequality: $d(A, B) \leq d(A, C) + d(B, C)$



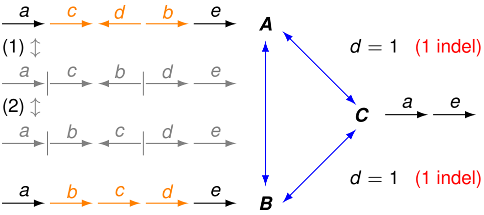
Triangular inequality disruption

Triangular inequality: $d(A,B) \leq d(A,C) + d(B,C)$



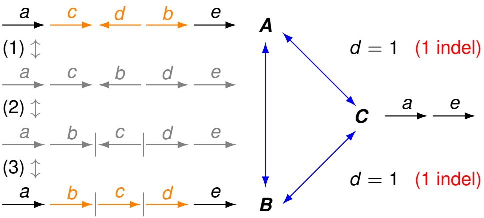
Triangular inequality disruption

Triangular inequality: $d(A, B) \leq d(A, C) + d(B, C)$



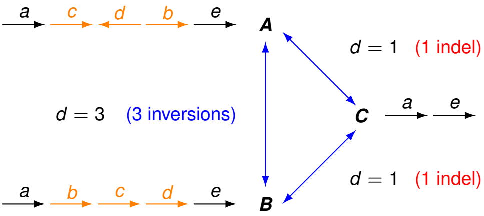
Triangular inequality disruption

Triangular inequality: $d(A,B) \leq d(A,C) + d(B,C)$



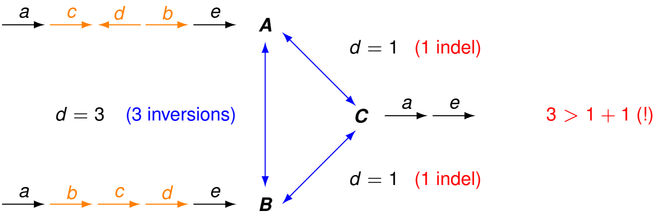
Triangular inequality disruption

Triangular inequality: $d(A, B) \leq d(A, C) + d(B, C)$



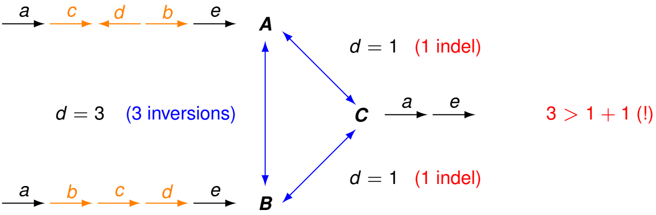
Triangular inequality disruption

Triangular inequality: $d(A, B) \leq d(A, C) + d(B, C)$



Triangular inequality disruption

Triangular inequality: $d(A, B) \leq d(A, C) + d(B, C)$



- **Adjustment:** the inequality holds for $\boldsymbol{m(A, B) = d(A, B) + k \cdot u(A, B)}$, where $u(A, B)$ is the number of unique markers between A and B.

Triangular inequality disruption

Calculating the diameter of the DCJ-indel distance

$|\boldsymbol{P}|$: # of **orange and blue edges** in P

$\lvert P\rvert$	$d_{\mathrm{DCJ}}(P)$	$\max \Lambda(P)$	$\max \lambda(P)$
1	0	1	1
2	0	2	2
3	1	3	2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\lvert P\rvert$	$\left\lfloor \frac{\lvert P\rvert-1}{2} \right\rfloor$	$\lvert P\rvert$	$\left\lceil \frac{\lvert P\rvert+1}{2} \right\rceil$

DCJ costs **1**
indel costs $\boldsymbol{w \leq 1}$

Let genomes A and B be **unichromosomal and linear**. The number of **orange and blue edges** in R(A,B) is $\mathbf{2(|\mathcal{G}| + 1)}$.

1. **The diameter of a component:**

$$\begin{aligned} d_{\mathrm{DCJ}}^{id}(P) &= d_{\mathrm{DCJ}}(P) + w\lambda(P) \\ &\leq \left\lfloor \frac{|P|-1}{2} \right\rfloor + w \left\lceil \frac{|P|+1}{2} \right\rceil \\ &\leq \frac{(w+1)|P|}{2} + \frac{w-1}{2} \\ &\leq \frac{(w+1)|P|}{2} \text{ , since } \frac{w-1}{2} \leq 0 \end{aligned}$$

2. **The diameter of the DCJ-indel distance:**

$$\begin{aligned} d_{\mathrm{DCJ}}^{id}(A,B) &\leq \textstyle\sum_{P \in R(A,B)} d_{\mathrm{DCJ}}^{id}(P) \\ &\leq \textstyle\sum_{P \in R(A,B)} \frac{(w+1)|P|}{2} \\ &= \frac{(w+1)}{2} \textstyle\sum_{P \in R(A,B)} |P| \\ &= \frac{(w+1)}{2} 2(|\mathcal{G}| + 1) \end{aligned}$$

$$\boldsymbol{d_{\mathrm{DCJ}}^{id}(A,B) \leq (w+1)(|\mathcal{G}|+1)}$$

Triangular inequality disruption

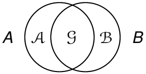
Finding the lower bound of k for the DCJ-indel distance

DCJ costs **1**
indel costs $\boldsymbol{w} \leq \mathbf{1}$

For unichr. linear genomes:

$$d^{id}_{\mathrm{DCJ}}(A, B) \leq (w+1)(|\mathcal{G}|+1)$$

Worst case: C is an empty genome.



$$C = \emptyset$$

$$d^{id}_{\mathrm{DCJ}}(A, C) = d^{id}_{\mathrm{DCJ}}(B, C) = w$$

$$\begin{aligned} m(A, B) &= d^{id}_{\mathrm{DCJ}}(A, B) + k(|\mathcal{A}| + |\mathcal{B}|) \\ m(A, C) &= d^{id}_{\mathrm{DCJ}}(A, C) + k(|\mathcal{A}| + |\mathcal{G}|) \\ m(B, C) &= d^{id}_{\mathrm{DCJ}}(B, C) + k(|\mathcal{B}| + |\mathcal{G}|) \end{aligned}$$

The following inequality has to be satisfied:

$$m(A, C) + m(B, C) \geq m(A, B)$$

$$2w + k(2|\mathcal{G}| + |\mathcal{A}| + |\mathcal{B}|) \geq (w+1)(|\mathcal{G}|+1) + k(|\mathcal{A}| + |\mathcal{B}|)$$

$$2w + k(2|\mathcal{G}|) \geq (w+1)(|\mathcal{G}|+1)$$

$$2w + 2k|\mathcal{G}| \geq w|\mathcal{G}| + w + |\mathcal{G}| + 1$$

$$2k|\mathcal{G}| \geq |\mathcal{G}|(w+1) - w + 1$$

$$k \geq \tfrac{w+1}{2} + \tfrac{1-w}{2|\mathcal{G}|}$$

$$\boldsymbol{k} \geq \boldsymbol{\tfrac{w+1}{2}}$$

Triangular inequality disruption

Summary: the lower bound of *k*

- **Adjustment:** the inequality holds for $\boldsymbol{m(A, B) = d(A, B) + k \cdot u(A, B)}$, where $u(A, B)$ is the number of unique markers between A and B.
- **DCJ** costs **1**
- **indel** costs $\boldsymbol{w \leq 1}$

Distance	*k*	References
DCJ-indel distance	$k \geq \frac{w+1}{2}$	WABI 2010, RECOMB-CG 2011b, WABI 2012
DCJ-substitution distance	$k \geq \frac{w+2}{4}$	RECOMB-CG 2011a and 2011b, to appear in AMB 2013
inversion-indel distance	**open**	

Acknowledgements

Many thanks to:

- Simone Dantas (Universidade Federal Fluminense /Brazil)
- Raphael Machado (Inmetro /Brazil)
- Leonardo Ribeiro (Inmetro /Brazil)
- Poly da Silva (Universidade Federal Fluminense and Inmetro /Brazil)
- Jens Stoye (Universität Bielefeld /Germany)
- Eyla Willing (Universität Bielefeld /Germany)
- Simone Zaccaria (Università di Milano-Bicocca /Italy)

This research was supported by:

- Inmetro /Brazil
- Brazilian research agency CNPq (grant PROMETRO 563087/2010-2)

Acknowledgements

Many thanks to:

- Simone Dantas (Universidade Federal Fluminense /Brazil)
- Raphael Machado (Inmetro /Brazil)
- Leonardo Ribeiro (Inmetro /Brazil)
- Poly da Silva (Universidade Federal Fluminense and Inmetro /Brazil)
- Jens Stoye (Universität Bielefeld /Germany)
- Eyla Willing (Universität Bielefeld /Germany)
- Simone Zaccaria (Università di Milano-Bicocca /Italy)

This research was supported by:

- Inmetro /Brazil
- Brazilian research agency CNPq (grant PROMETRO 563087/2010-2)

Thank you for your attention!