Restricted DCJ-indel model revisited

Marília Braga

Inmetro
Brazil

Jens Stoye

Bielefeld University
Germany

BSB 2013

Universität Bielefeld

INMETRO

Overview

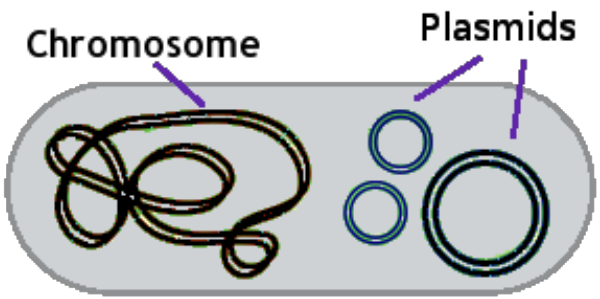
Motivation and Background

Overview

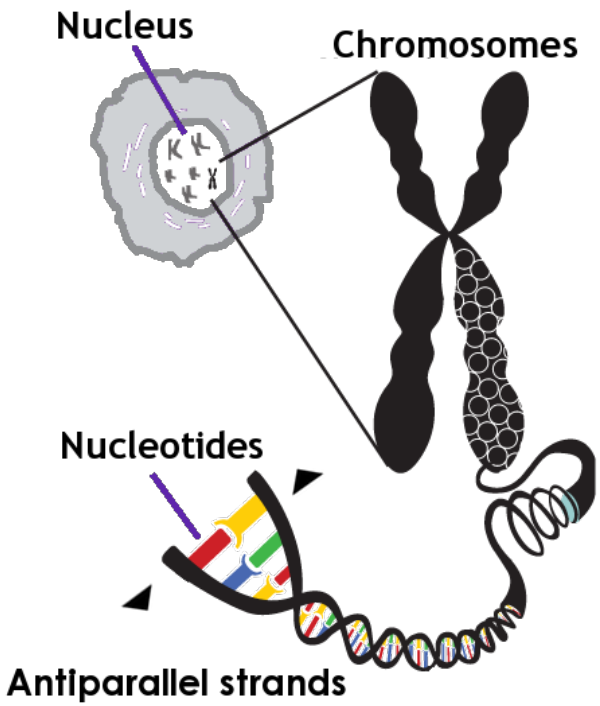
Motivation and Background

Genomes are composed of one or more DNA molecules:

Prokaryotic Cell

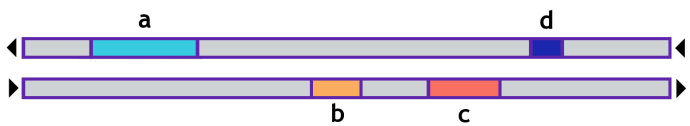


Eukaryotic Cell



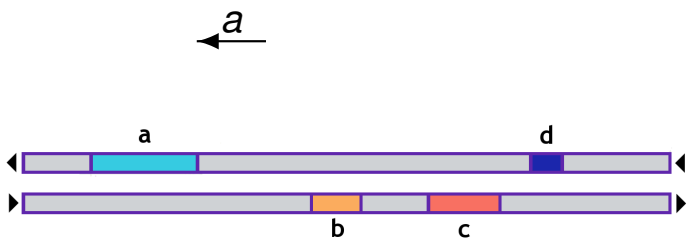
Motivation and Background

Each chromosome contains a set of genes:



Motivation and Background

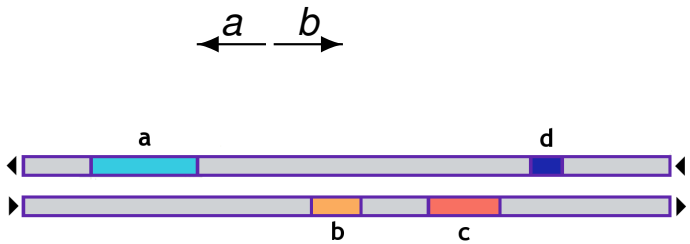
Each chromosome contains a set of genes:



the DNA strand in which each gene lies gives its orientation

Motivation and Background

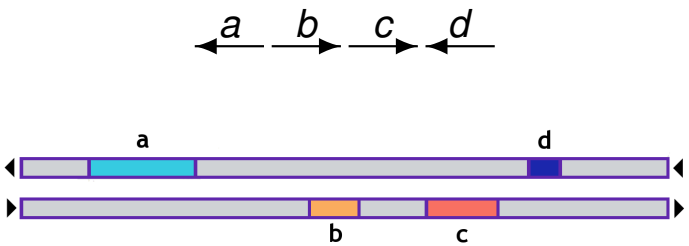
Each chromosome contains a set of genes:



the DNA strand in which each gene lies gives its orientation

Motivation and Background

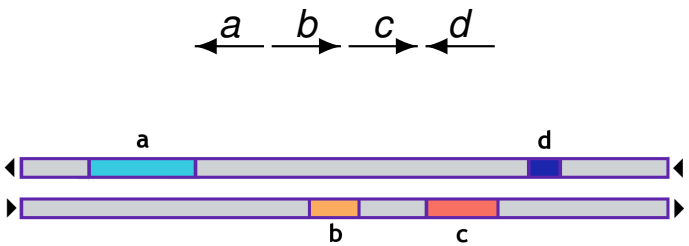
Each chromosome contains a set of genes:



the DNA strand in which each gene lies gives its orientation

Motivation and Background

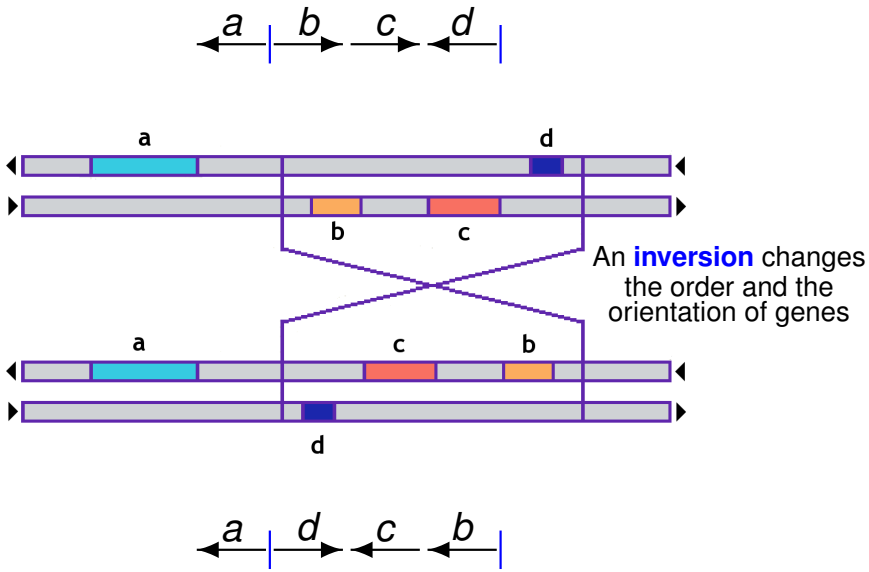
Each chromosome contains a set of genes:



Genomes are subject to large scale mutations...

Motivation and Background

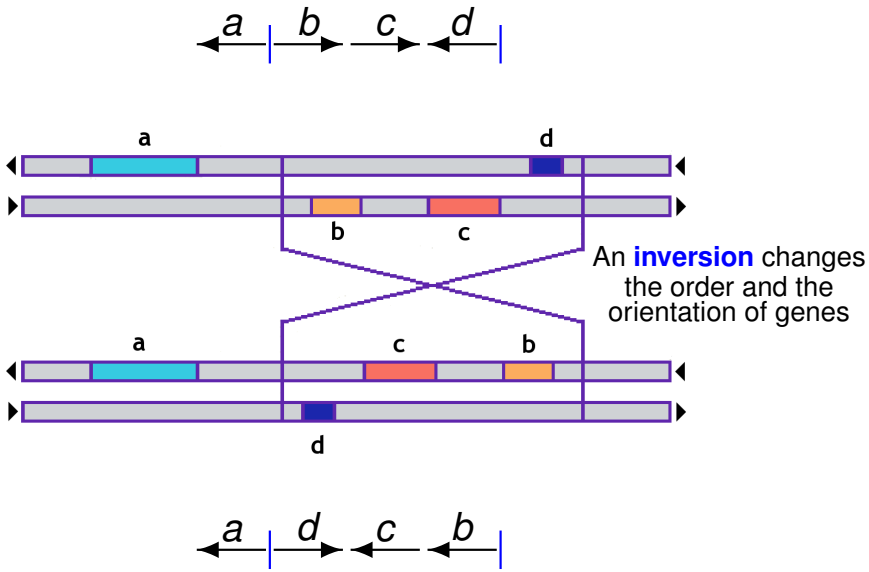
Each chromosome contains a set of genes:



Genomes are subject to large scale mutations...

Motivation and Background

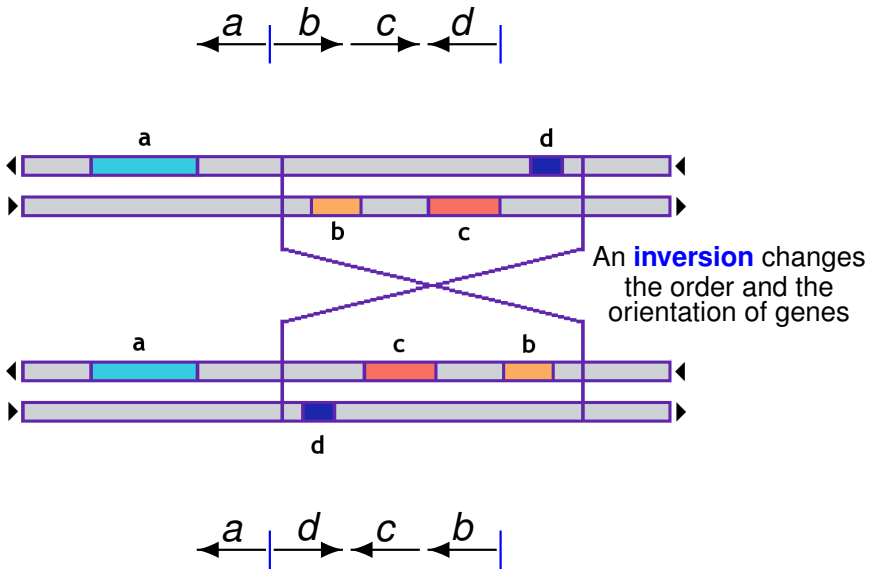
Each chromosome contains a set of genes:



Genomes are subject to large scale mutations...
...inversions, translocations, fusions, fissions...

Motivation and Background

Each chromosome contains a set of genes:



Genomes are subject to large scale mutations...
...inversions, translocations, fusions, fissions...
...insertions, deletions, duplications...

Motivation and Background

Comparing two distinct genomes

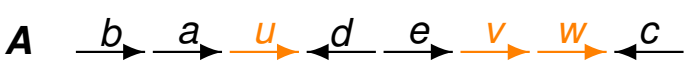
Common genes:

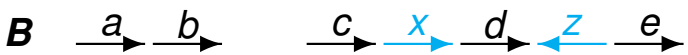
$\mathcal{G} = \{a, b, c, d, e\}$

Unique genes:

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

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



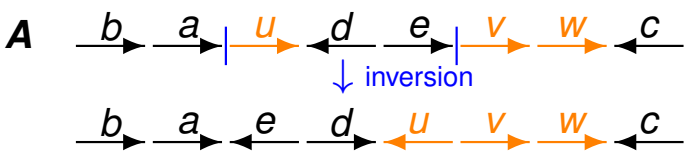


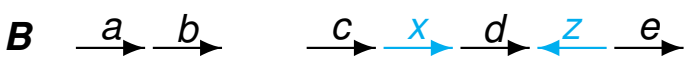
Motivation and Background

Comparing two distinct genomes

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

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





Motivation and Background

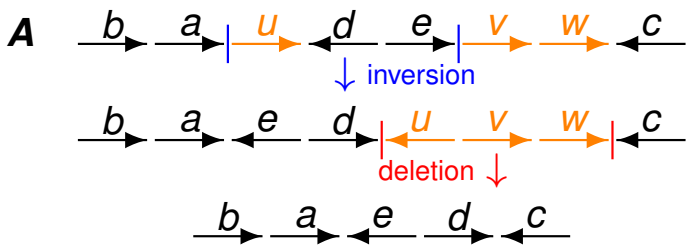
Comparing two distinct genomes

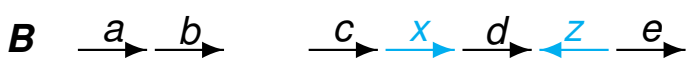
Common genes:

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

Unique genes:

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





Motivation and Background

Comparing two distinct genomes

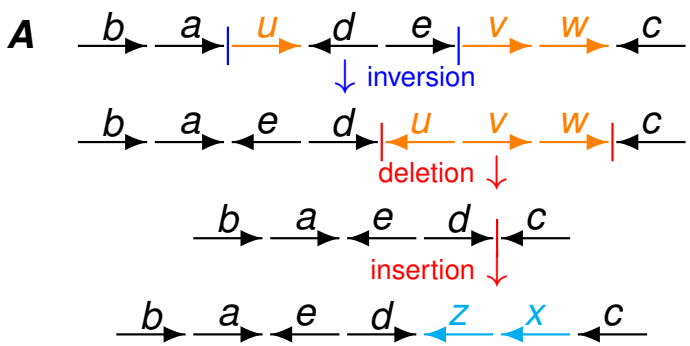
Common genes:

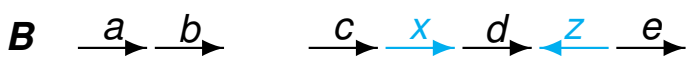
$\mathcal{G} = \{a, b, c, d, e\}$

Unique genes:

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

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





Motivation and Background

Comparing two distinct genomes

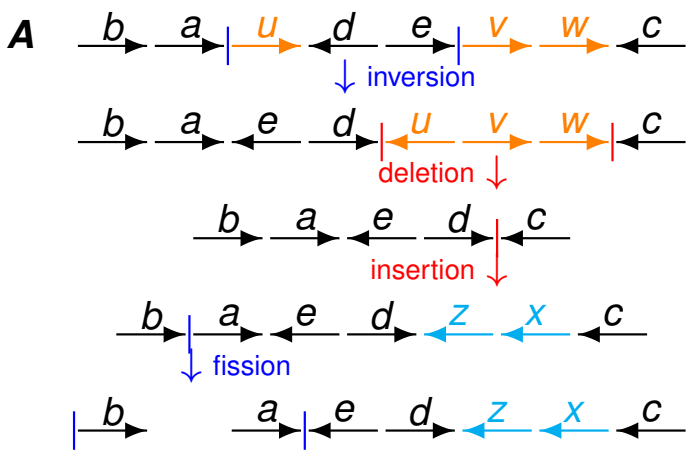
Common genes:

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

Unique genes:

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

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



Motivation and Background

Comparing two distinct genomes

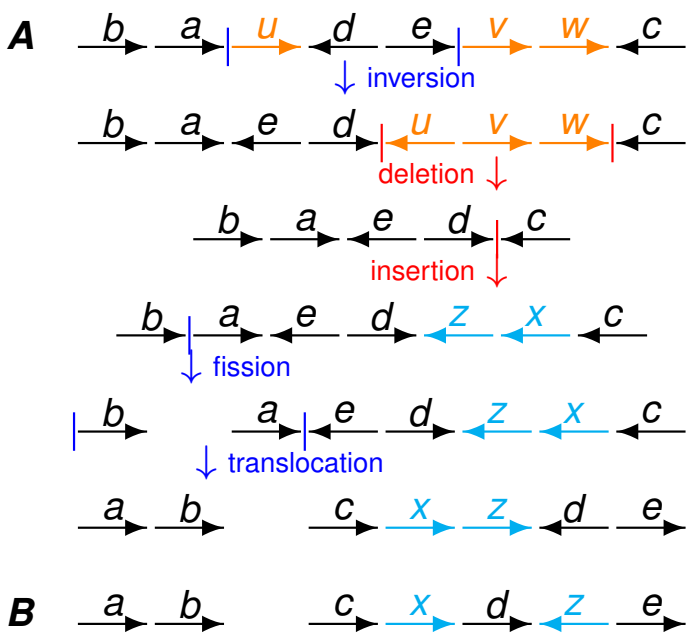
Common genes:

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

Unique genes:

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

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



Motivation and Background

Comparing two distinct genomes

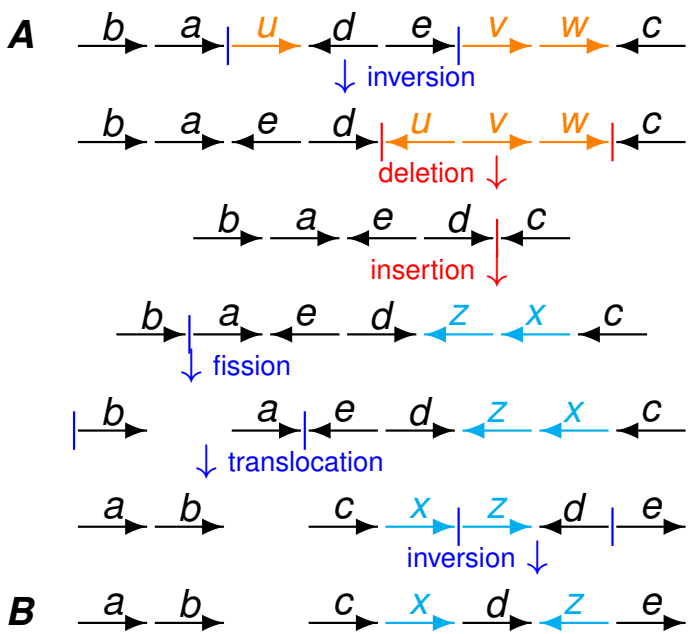
Common genes:

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

Unique genes:

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

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

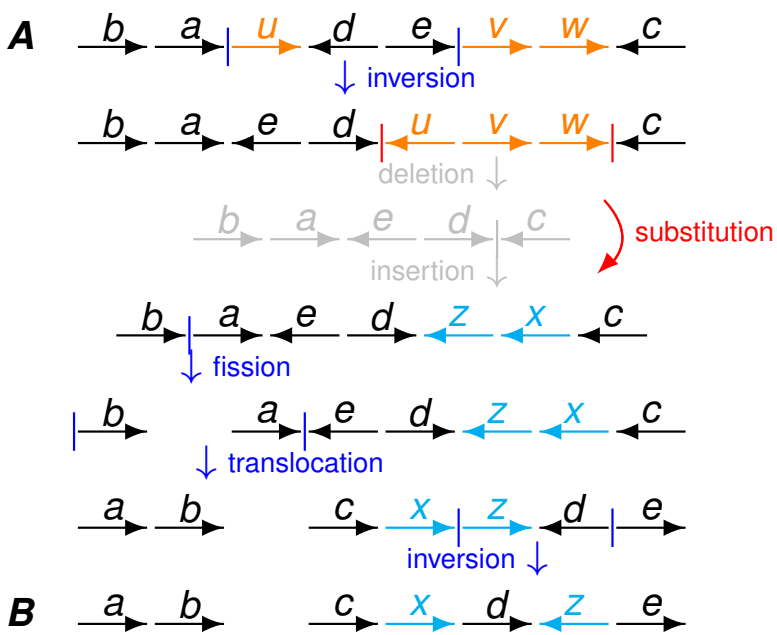


Motivation and Background

Comparing two distinct genomes

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

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

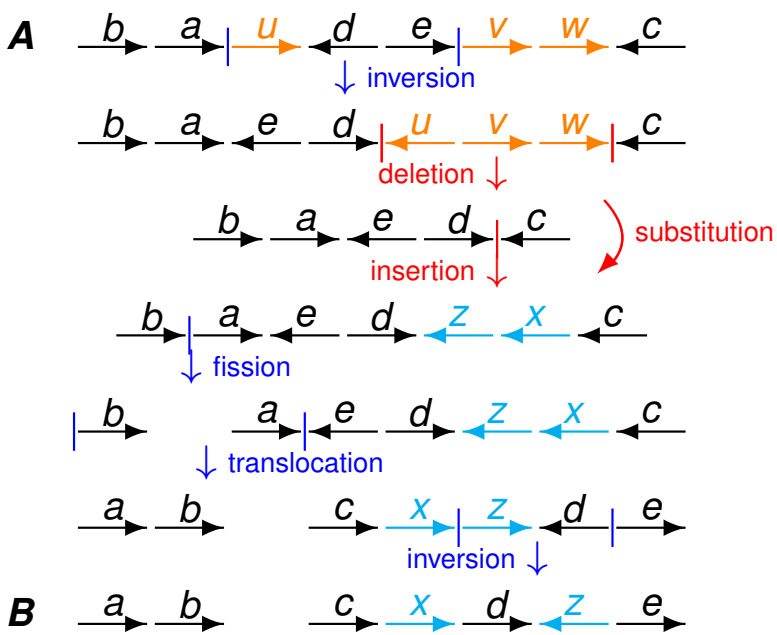


Motivation and Background

Comparing two distinct genomes

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

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



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

Motivation and Background

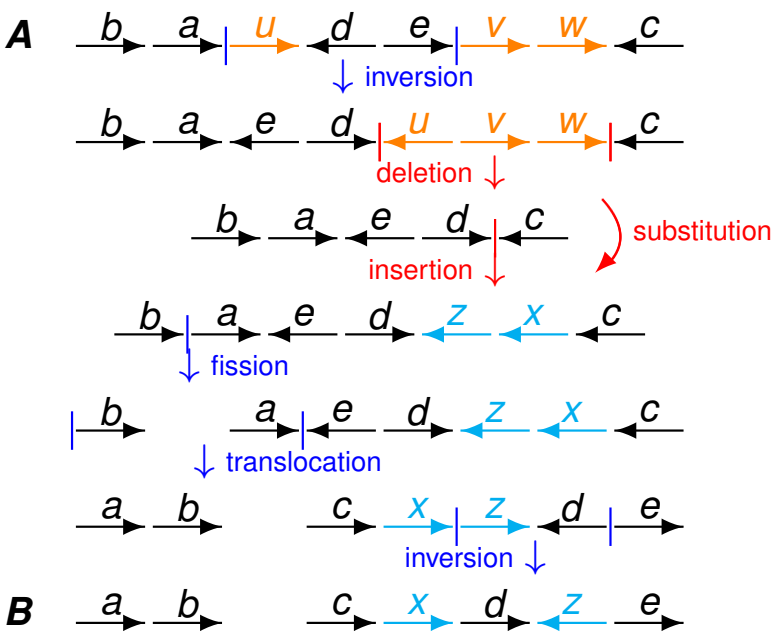
Comparing two distinct genomes

Common genes:

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

Unique genes:

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



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

[Yancopoulos, Attie and Friedberg, 2005]

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

Motivation and Background

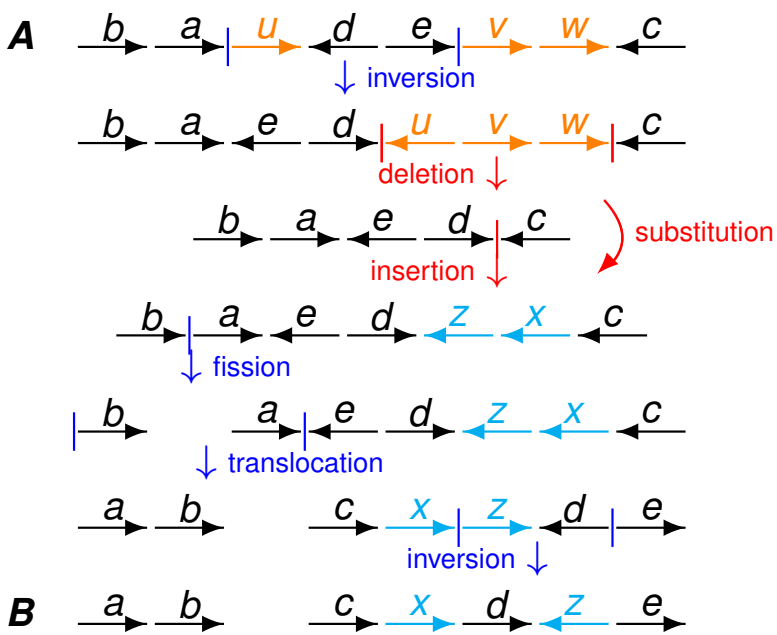
Comparing two distinct genomes

Common genes:

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

Unique genes:

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



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

[Yancopoulos, Attie and Friedberg, 2005]

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

Genomic distance:

minimum number of operations that sort one genome into another

Motivation and Background

Some polynomial models for the genomic distance

Inversion model [Hannenhalli & Pevzner 1995, Meidanis *et al.* 2000]

- unichromosomal genomes with the **same content** and no duplications
- only inversions

Motivation and Background

Some polynomial models for the genomic distance

Inversion model [Hannenhalli & Pevzner 1995, Meidanis *et al.* 2000]

- unichromosomal genomes with the **same content** and no duplications
- only inversions

HP model [Hannenhalli & Pevzner 1995]

- multichromosomal linear genomes with the **same content** and no duplications
- inversions, translocations, fusions and fissions

Motivation and Background

Some polynomial models for the genomic distance

Inversion model [Hannenhalli & Pevzner 1995, Meidanis *et al.* 2000]

- unichromosomal genomes with the **same content** and no duplications
- only inversions

HP model [Hannenhalli & Pevzner 1995]

- multichromosomal linear genomes with the **same content** and no duplications
- inversions, translocations, fusions and fissions

The double cut and join (DCJ) model [Yancopoulos *et al.* 2005, Bergeron *et al.* 2006]

- multichromosomal genomes with the **same content** and no duplications
- inversions, translocations, fusions, fissions, **circular excisions** and **reincorporations**
- **distance** can be easily computed in **linear time**

Motivation and Background

For genomes with unequal contents but no duplications

Inversions with indels [El-Mabrouk 2000, RECOMB-CG 2013]

- unichromosomal, allows **inversions** and **indels** with the **same cost**
- no exact solution yet, but for an important set of instances, the inversion-indel distance can be computed in linear time

Motivation and Background

For genomes with unequal contents but no duplications

Inversions with indels [El-Mabrouk 2000, RECOMB-CG 2013]

- unichromosomal, allows **inversions** and **indels** with the **same cost**
- no exact solution yet, but for an important set of instances, the inversion-indel distance can be computed in linear time

DCJ with indels [Yancopoulos & Friedberg 2008, WABI 2010/2012, Compeau 2012, submit. to LATIN 2014]

- multichromosomal, allows **DCJ operations** and **indels** with **distinct costs**
- **distance** can be computed in **linear time**

Motivation and Background

For genomes with unequal contents but no duplications

Inversions with indels [El-Mabrouk 2000, RECOMB-CG 2013]

- unichromosomal, allows **inversions** and **indels** with the **same cost**
- no exact solution yet, but for an important set of instances, the inversion-indel distance can be computed in linear time

DCJ with indels [Yancopoulos & Friedberg 2008, WABI 2010/2012, Compeau 2012, submit. to LATIN 2014]

- multichromosomal, allows **DCJ operations** and **indels** with **distinct costs**
- **distance** can be computed in **linear time**

DCJ with substitutions [RECOMB-CG 2011, BSB 2012, AMB 2013]

- multichromosomal, allows **DCJ operations** and **substitutions** with **distinct costs**
- more parsimonious than DCJ-indel (substitutions include indels)
- **distance** can be computed in **linear time**

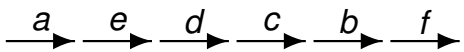
Restricted DCJ model with ER composition

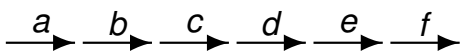
Overview

Restricted DCJ model with ER composition

Sorting linear genomes...

General scenario

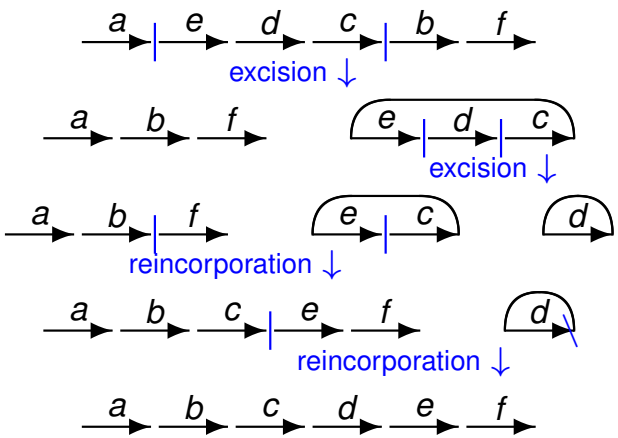




Restricted DCJ model with ER composition

Sorting linear genomes...

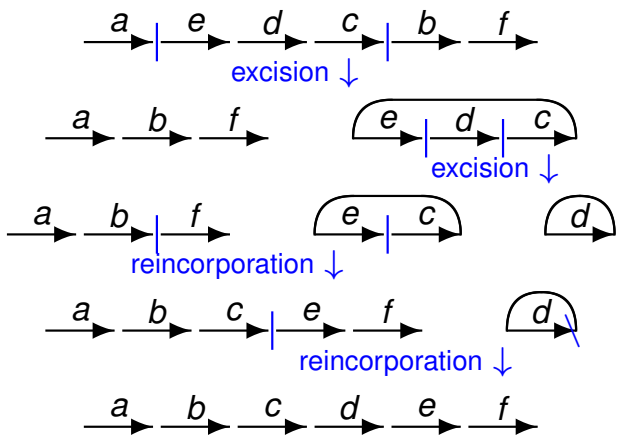
General scenario



Restricted DCJ model with ER composition

Sorting linear genomes...

General scenario

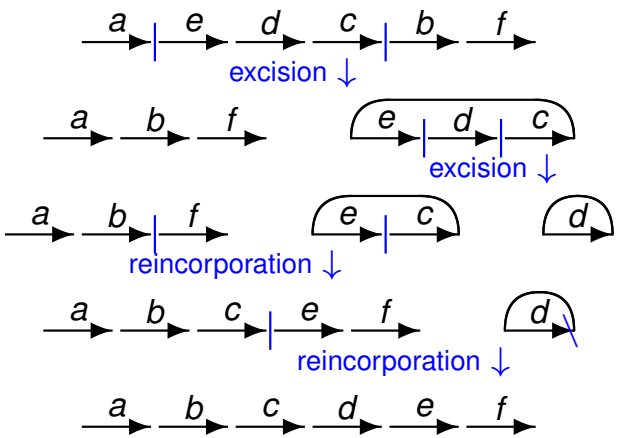


Many circular chromosomes can coexist in the intermediate species.

Restricted DCJ model with ER composition

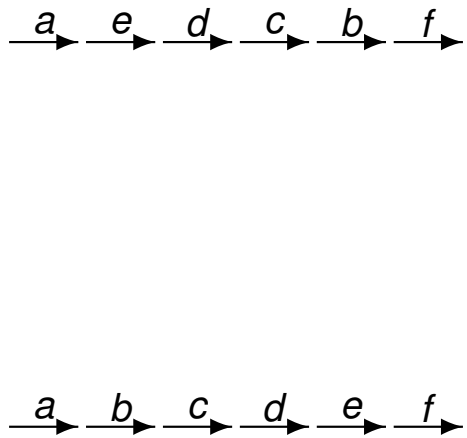
Sorting linear genomes...

General scenario



Many circular chromosomes can coexist in the intermediate species.

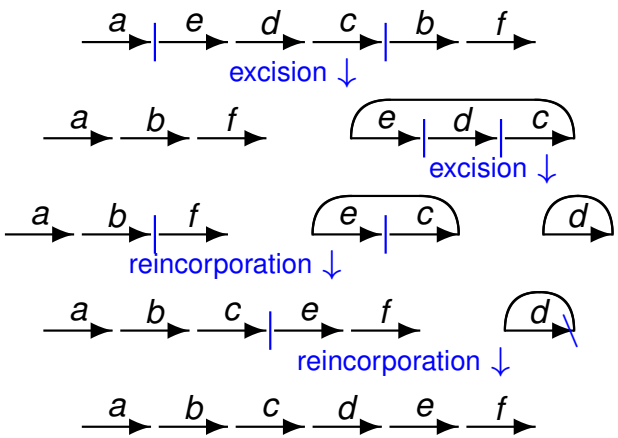
Restricted scenario



Restricted DCJ model with ER composition

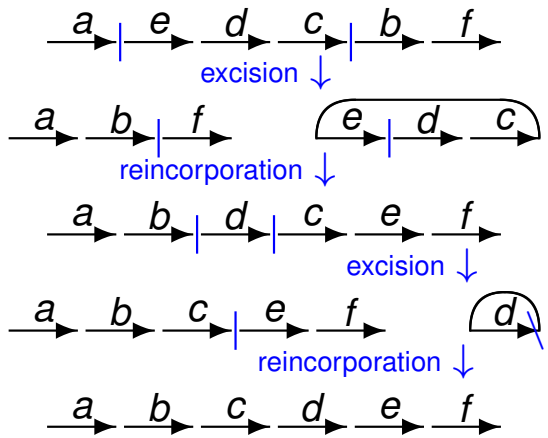
Sorting linear genomes...

General scenario



Many circular chromosomes can coexist in the intermediate species.

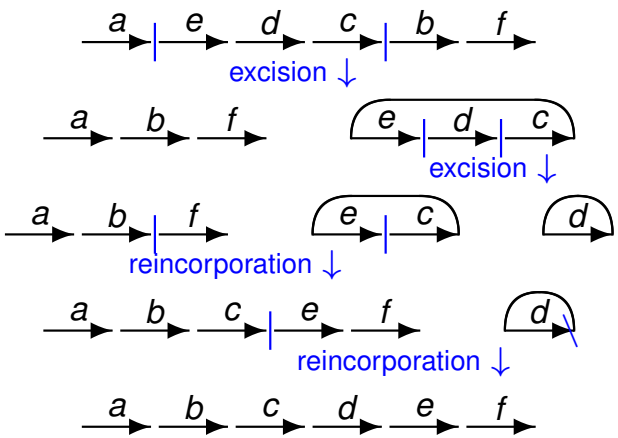
Restricted scenario



Restricted DCJ model with ER composition

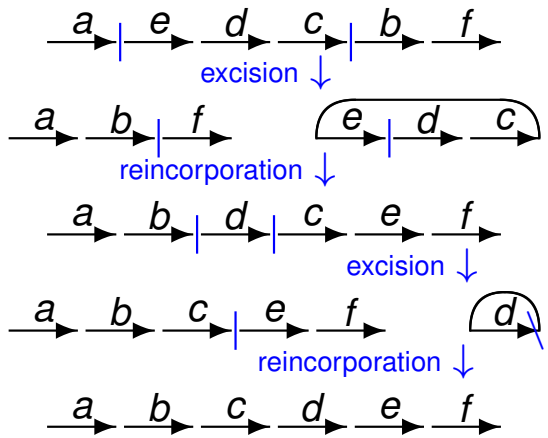
Sorting linear genomes...

General scenario



Many circular chromosomes can coexist in the intermediate species.

Restricted scenario

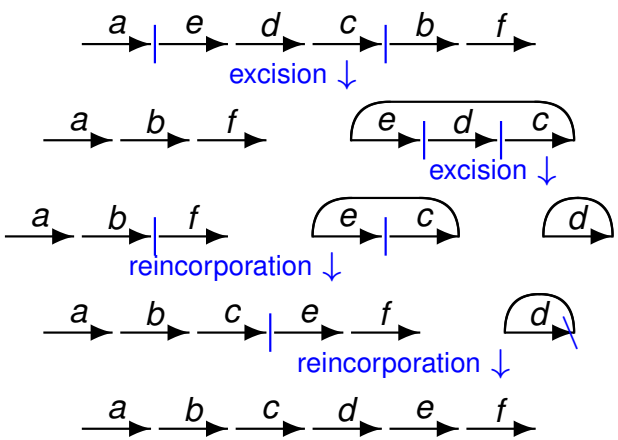


A circular chromosome is immediately reincorporated after its excision **(ER composition)**

Restricted DCJ model with ER composition

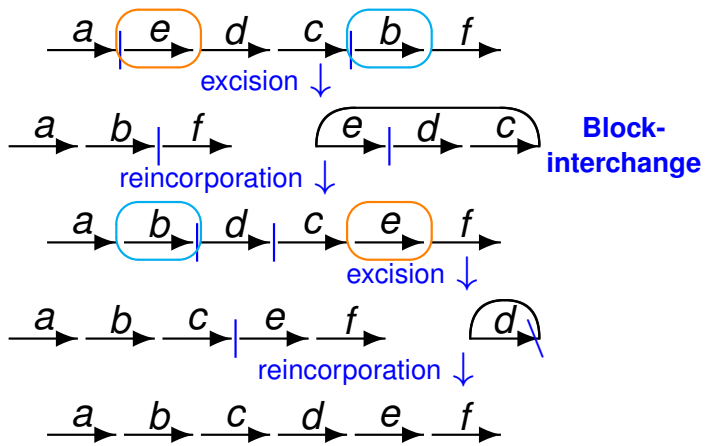
Sorting linear genomes...

General scenario



Many circular chromosomes can coexist in the intermediate species.

Restricted scenario

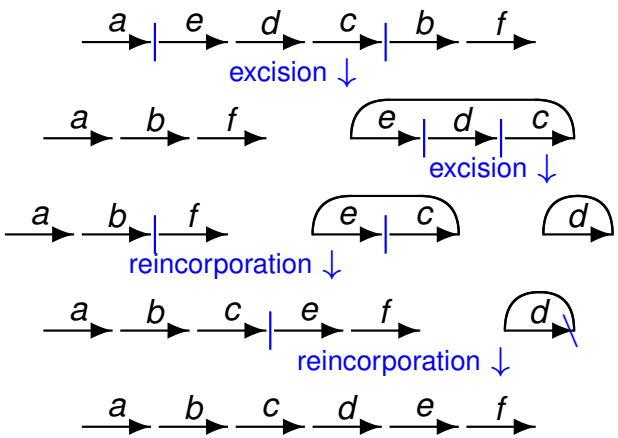


A circular chromosome is immediately reincorporated after its excision **(ER composition)**

Restricted DCJ model with ER composition

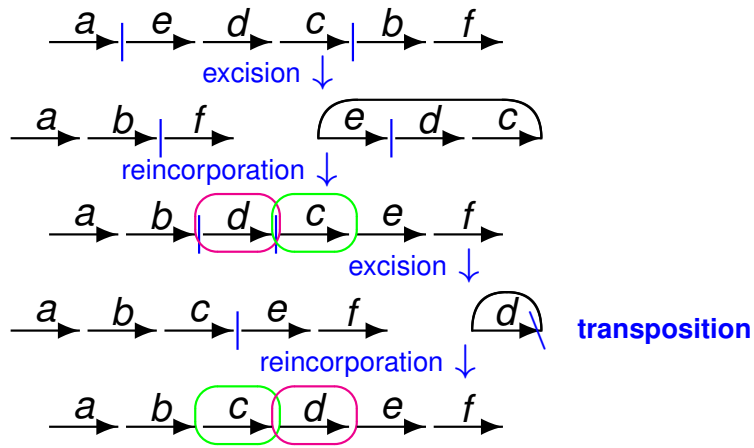
Sorting linear genomes...

General scenario



Many circular chromosomes can coexist in the intermediate species.

Restricted scenario

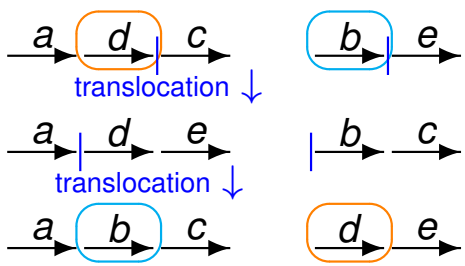


A circular chromosome is immediately reincorporated after its excision **(ER composition)**

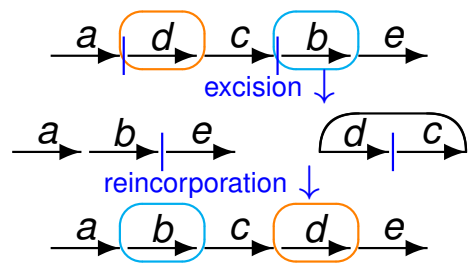
Restricted DCJ model with ER composition

External and Internal Block-Interchange

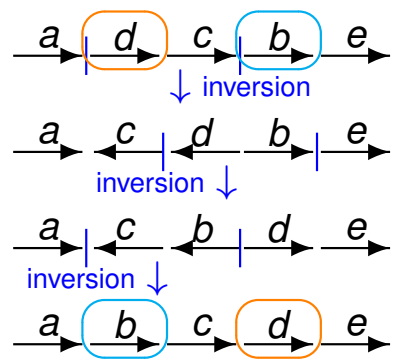
External
(2 translocations)



Internal
(ER composition)



Internal
(3 inversions)



Universität Bielefeld
INMETRO

Restricted DCJ model with ER composition

- inversions, translocations, fusions and fissions:
  mimicked by a single DCJ
- internal transpositions and block-interchanges:
  mimicked by two consecutive DCJs (**ER composition**)
- the distance is the same for both the general and the restricted DCJ models
  [Yancopoulos *et al.* 2005]
- a restricted DCJ scenario can be obtained in $\mathcal{O}(n \log n)$ time
  [Kováč *et al.* 2010]

The restricted DCJ-indel model

Overview

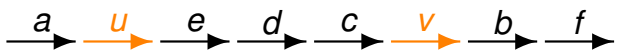
INMETRO Universität Bielefeld

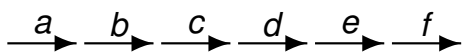
The restricted DCJ-indel model

- each DCJ costs 1
- each indel costs positive w
- inversions, translocations, fusions and fissions:
  mimicked by a single DCJ
- internal transpositions and block-interchanges:
  mimicked by two consecutive DCJs (**ER composition**)

The restricted DCJ-indel model

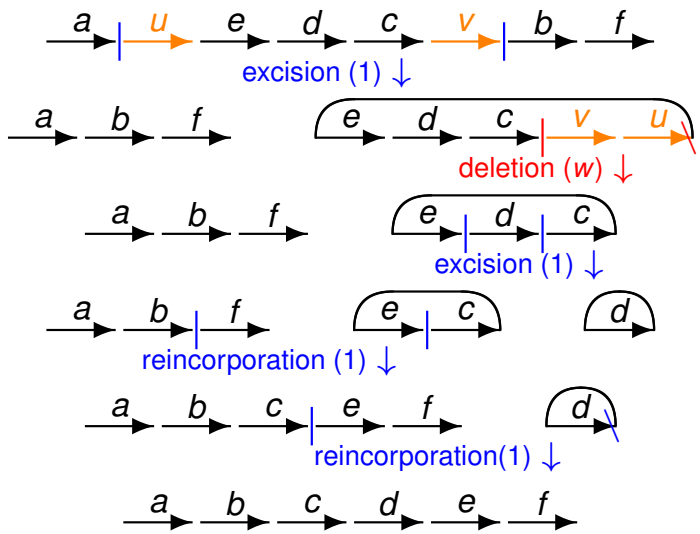
General scenario





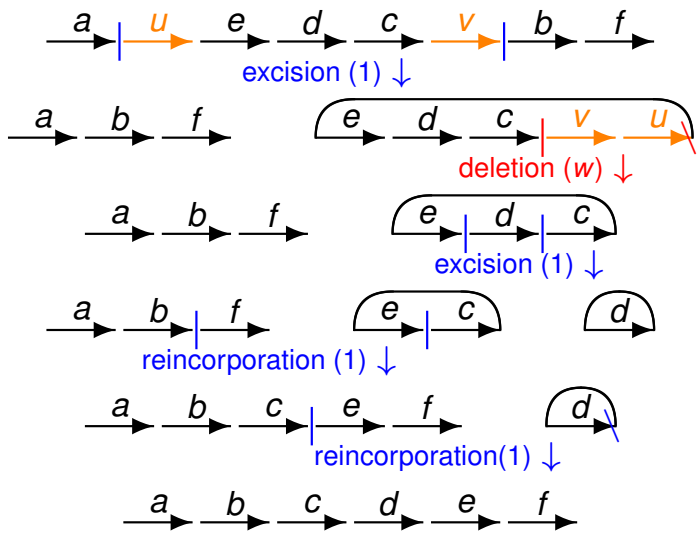
The restricted DCJ-indel model

General scenario



The restricted DCJ-indel model

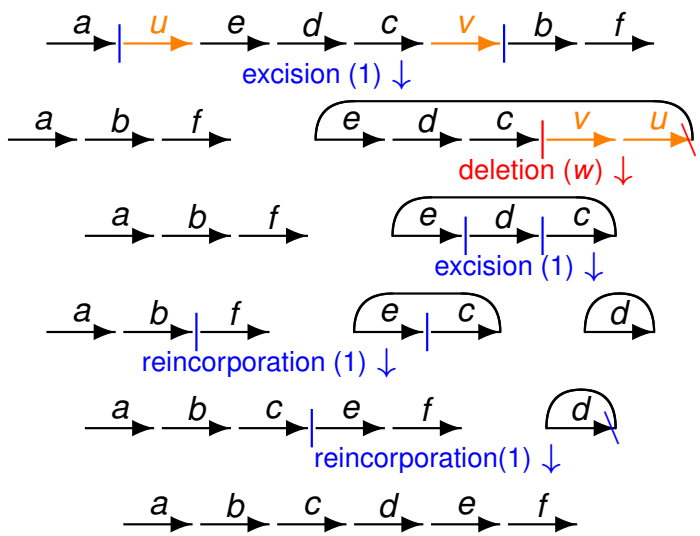
General scenario



Many circular chromosomes can coexist in the intermediate species.

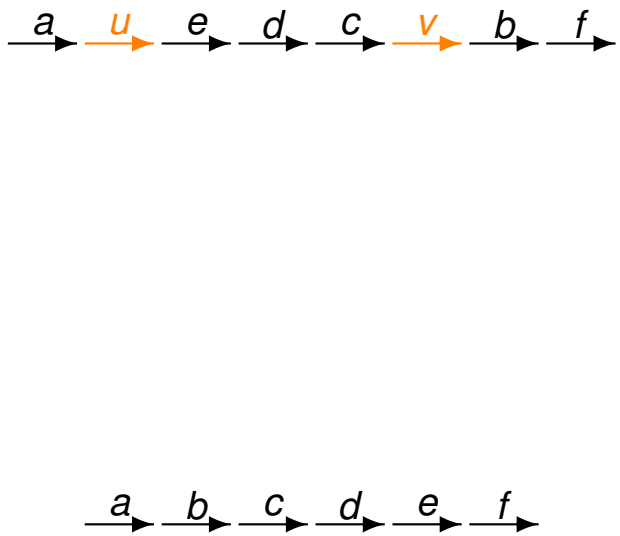
The restricted DCJ-indel model

General scenario



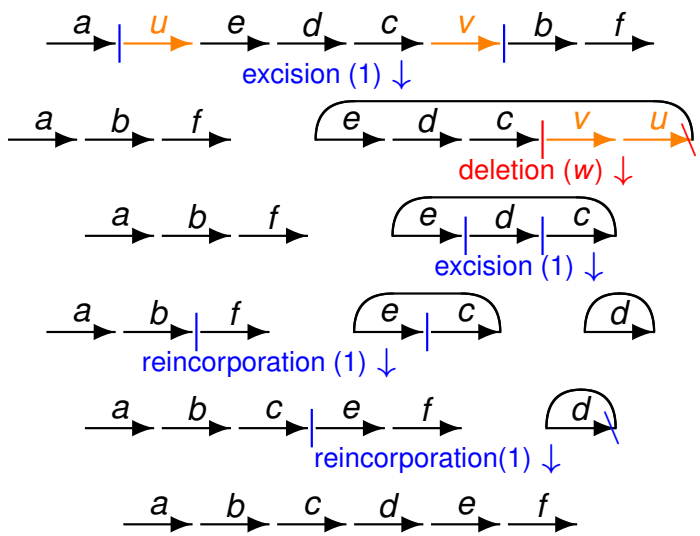
Many circular chromosomes can coexist in the intermediate species.

Restricted scenario

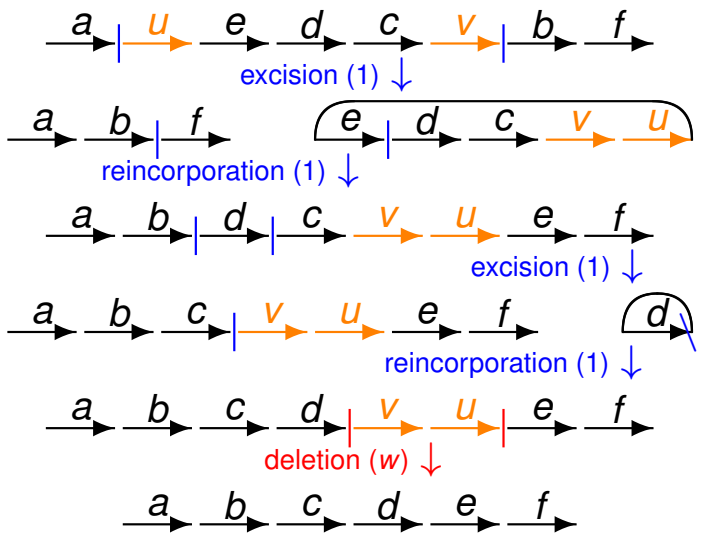


The restricted DCJ-indel model

General scenario



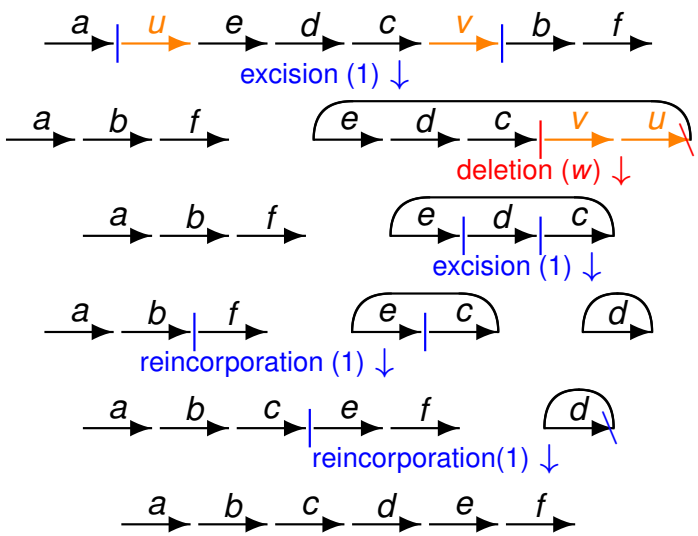
Restricted scenario



Many circular chromosomes can coexist in the intermediate species.

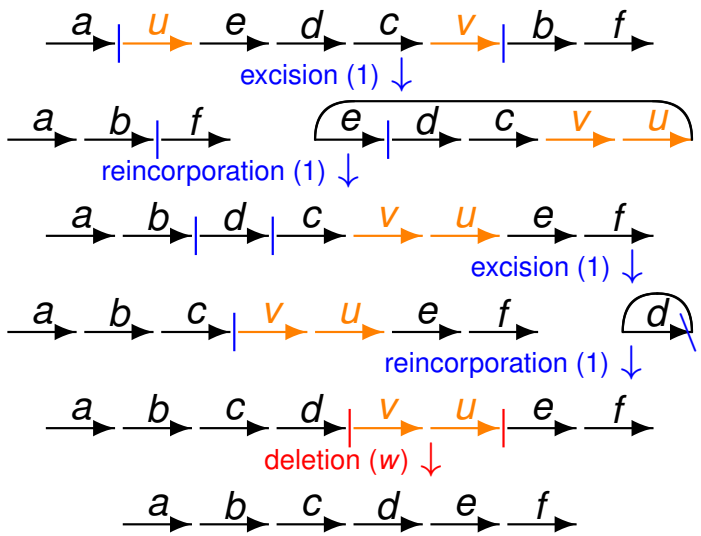
The restricted DCJ-indel model

General scenario



Many circular chromosomes can coexist in the intermediate species.

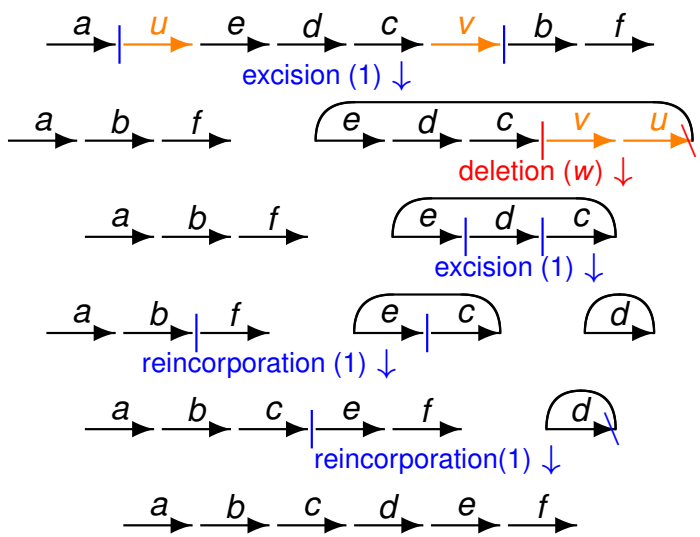
Restricted scenario



A circular chromosome is immediately reincorporated after its excision.
(ER composition)

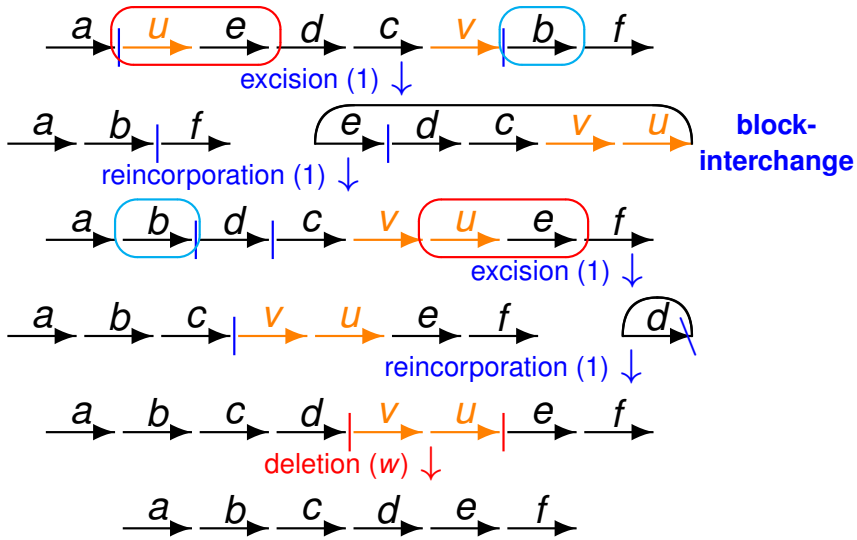
The restricted DCJ-indel model

General scenario



Many circular chromosomes can coexist in the intermediate species.

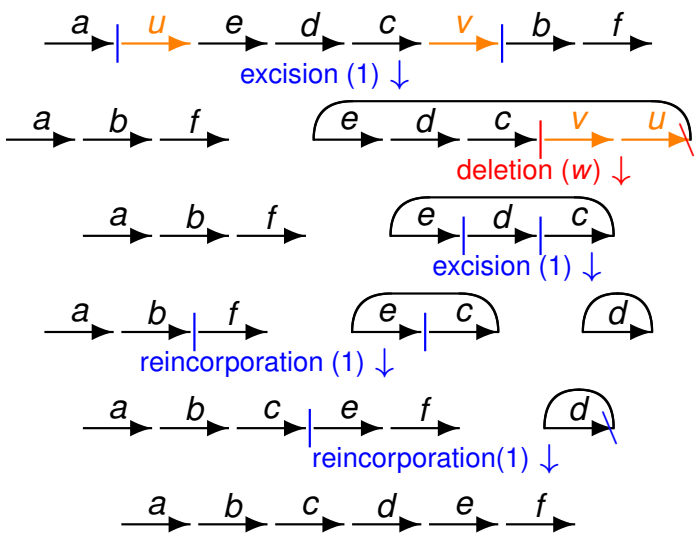
Restricted scenario



A circular chromosome is immediately reincorporated after its excision.
(ER composition)

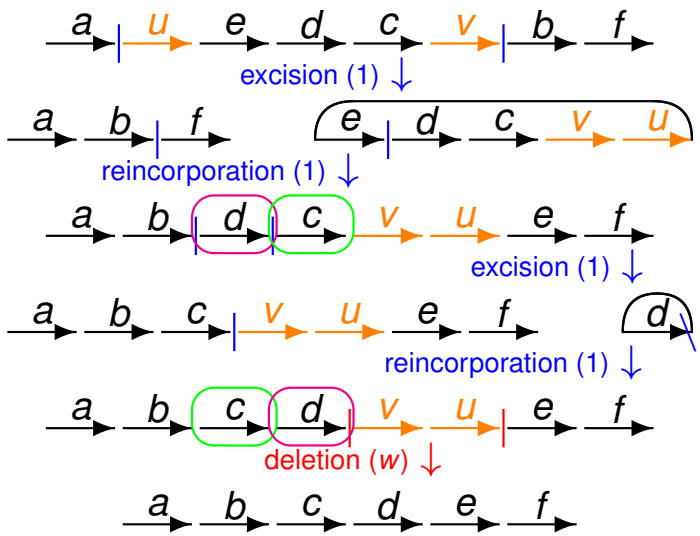
The restricted DCJ-indel model

General scenario



Many circular chromosomes can coexist in the intermediate species.

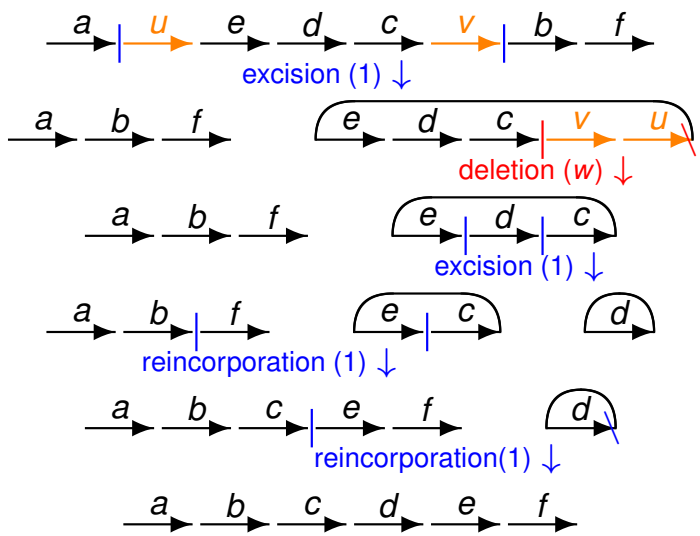
Restricted scenario



A circular chromosome is immediately reincorporated after its excision.
(ER composition)

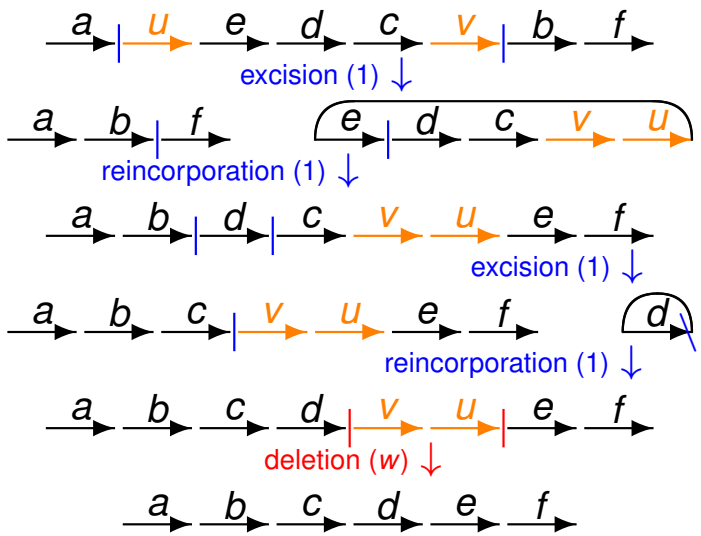
The restricted DCJ-indel model

General scenario



Many circular chromosomes can coexist in the intermediate species.

Restricted scenario



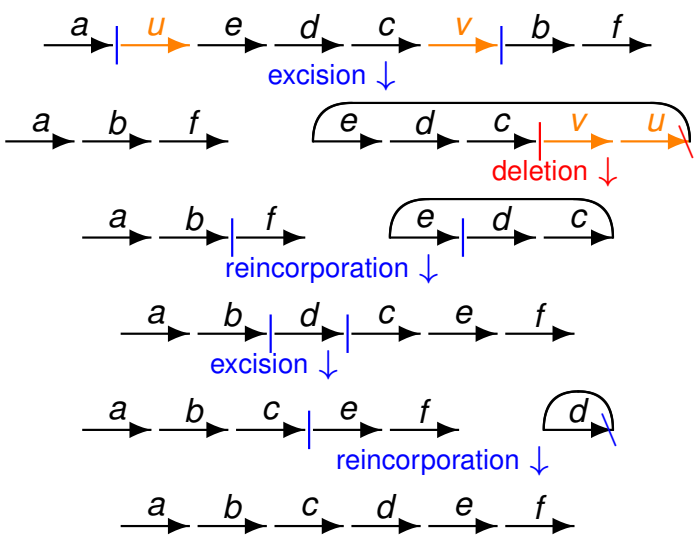
A circular chromosome is immediately reincorporated after its excision.
(ER composition)

Are both the general and the restricted DCJ-indel distances the same?

[RECOMB-CG 2012]

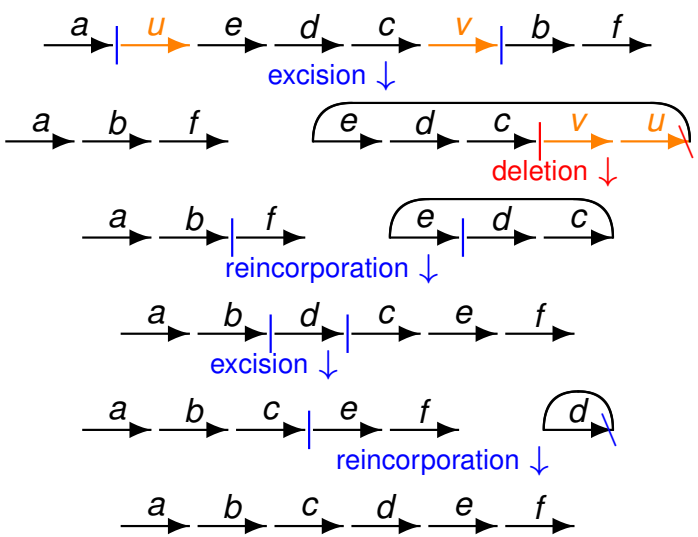
The restricted DCJ-indel model

Moving insertions up and deletions down



The restricted DCJ-indel model

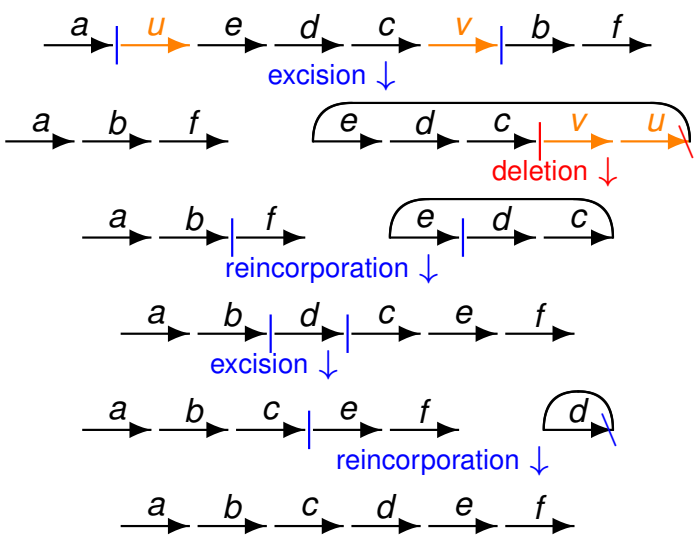
Moving insertions up and deletions down



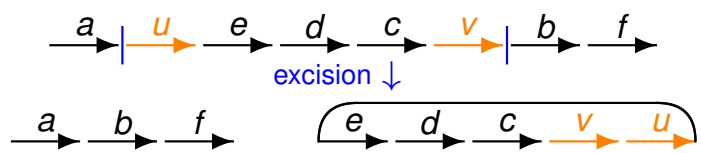
A deletion can always be moved down
[RECOMB-CG 2012]

The restricted DCJ-indel model

Moving insertions up and deletions down

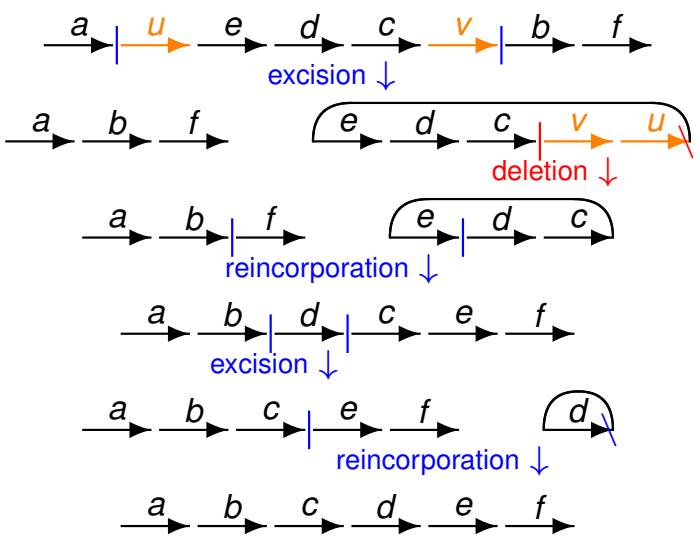


A deletion can always be moved down
[RECOMB-CG 2012]



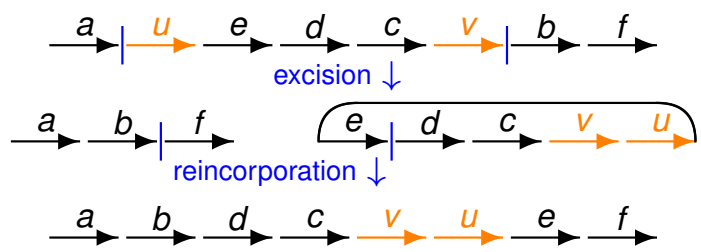
The restricted DCJ-indel model

Moving insertions up and deletions down



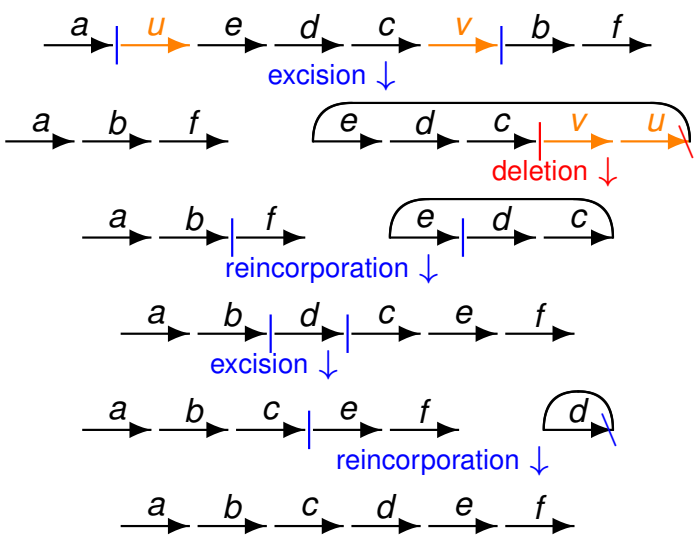
A deletion can always be moved down

[RECOMB-CG 2012]



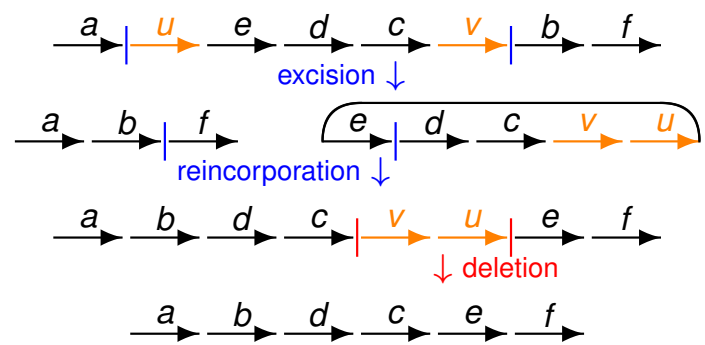
The restricted DCJ-indel model

Moving insertions up and deletions down



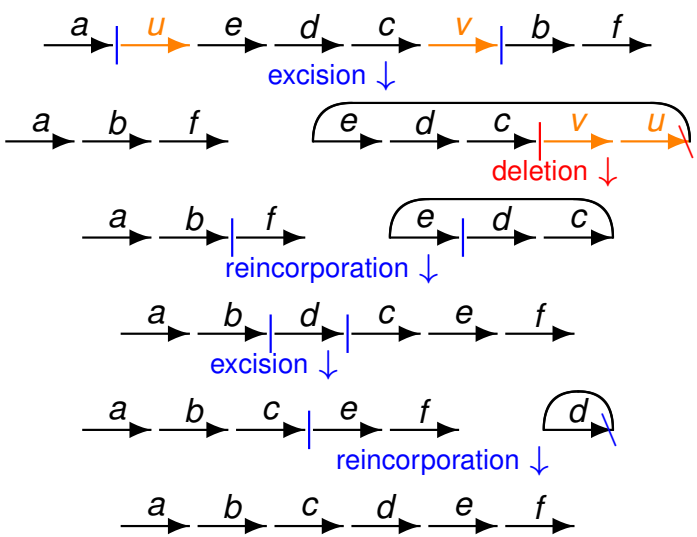
A deletion can always be moved down

[RECOMB-CG 2012]

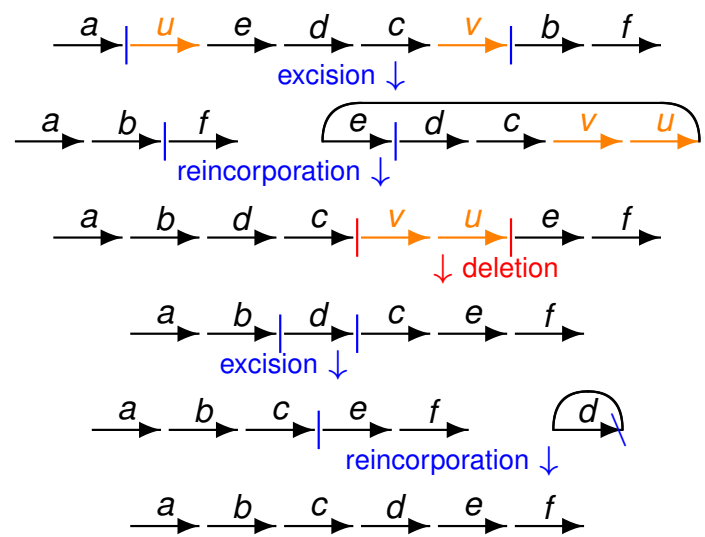


The restricted DCJ-indel model

Moving insertions up and deletions down

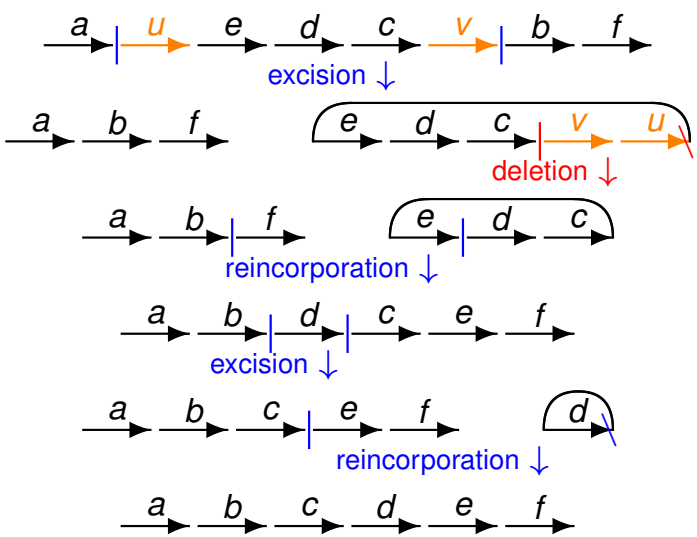


A deletion can always be moved down
[RECOMB-CG 2012]

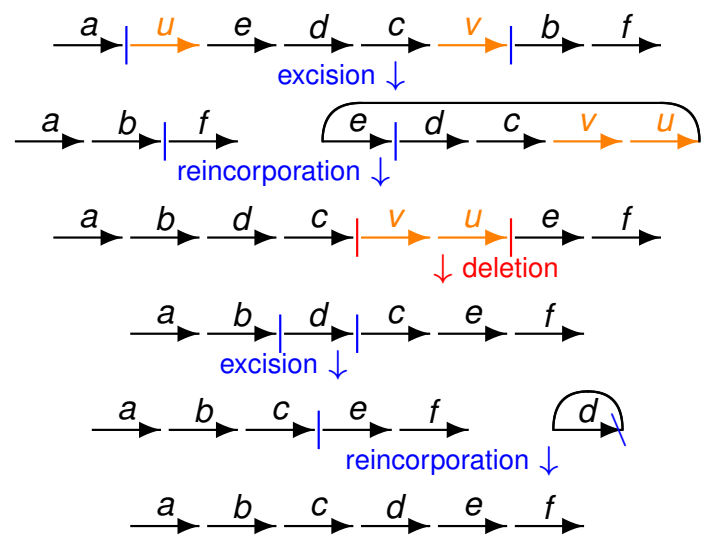


The restricted DCJ-indel model

Moving insertions up and deletions down



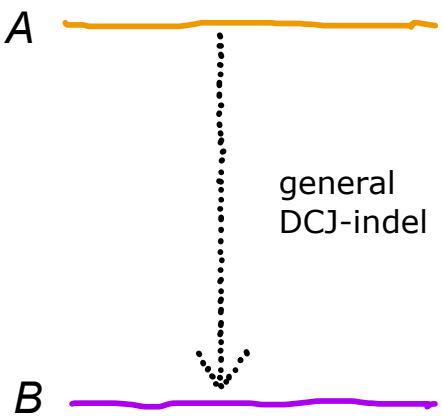
A deletion can always be moved down
[RECOMB-CG 2012]



Analogously, an insertion can always be moved up
[RECOMB-CG 2012]

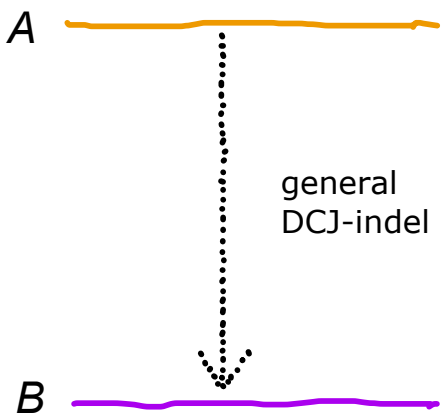
The restricted DCJ-indel model

An algorithm to find a restricted DCJ-indel sorting scenario



The restricted DCJ-indel model

An algorithm to find a restricted DCJ-indel sorting scenario



n_1 DCJs

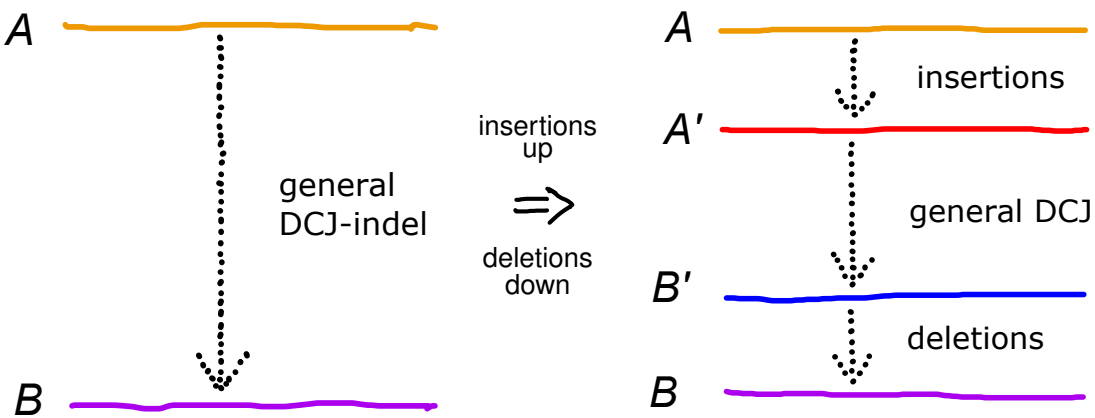
n_2 indels

$d = n_1 + w\,n_2$

$\mathcal{O}(n)$ time

The restricted DCJ-indel model

An algorithm to find a restricted DCJ-indel sorting scenario



n_1 DCJs

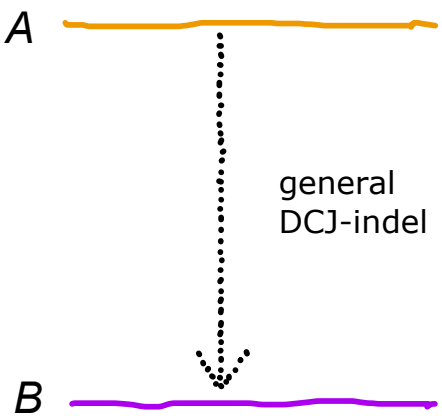
n_2 indels

$d = n_1 + w\,n_2$

$\mathcal{O}(n)$ time

The restricted DCJ-indel model

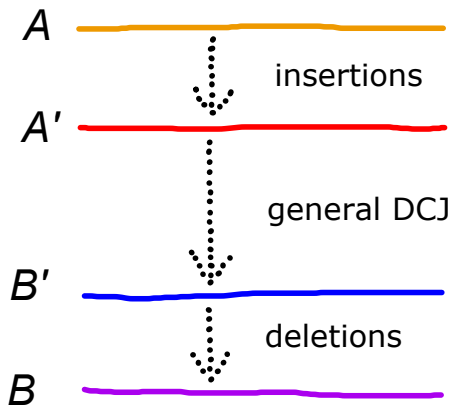
An algorithm to find a restricted DCJ-indel sorting scenario



n_1 DCJs
n_2 indels
$d = n_1 + w\,n_2$

$\mathcal{O}(n)$ time

insertions
up
⇒
deletions
down

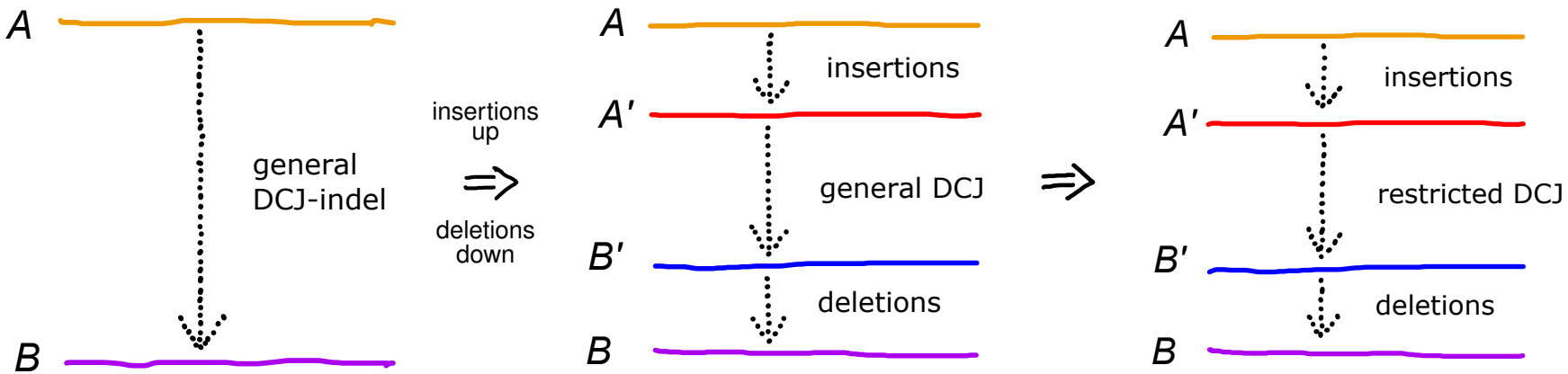


n_1 DCJs
n_2 indels
$d' = d = n_1 + w\,n_2$

$\mathcal{O}(n)$ time

The restricted DCJ-indel model

An algorithm to find a restricted DCJ-indel sorting scenario



n_1 DCJs
n_2 indels
$d = n_1 + w\,n_2$

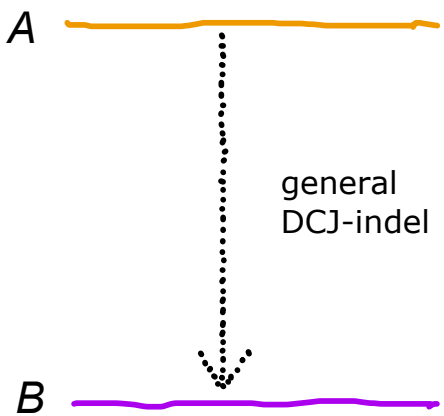
$\mathcal{O}(n)$ time

n_1 DCJs
n_2 indels
$d' = d = n_1 + w\,n_2$

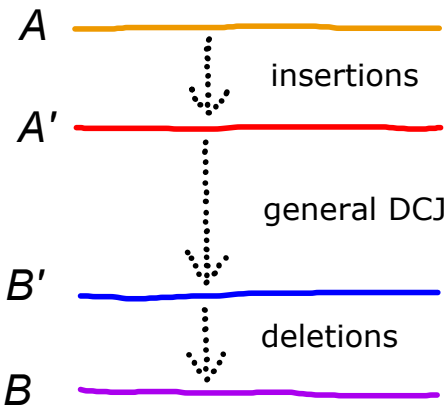
$\mathcal{O}(n)$ time

The restricted DCJ-indel model

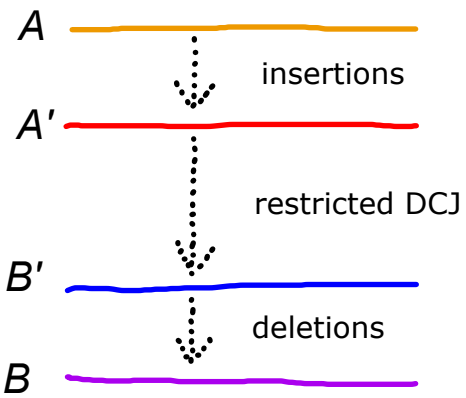
An algorithm to find a restricted DCJ-indel sorting scenario



insertions up ⇒ deletions down



⇒



n_1 DCJs
n_2 indels
$d = n_1 + w\,n_2$

$\mathcal{O}(n)$ time

n_1 DCJs
n_2 indels
$d' = d = n_1 + w\,n_2$

$\mathcal{O}(n)$ time

n_1 DCJs
n_2 indels
$d'' = d' = d = n_1 + w\,n_2$

$\mathcal{O}(n \log n)$ time

Conclusions and Future Work

Overview

Conclusions and Future Work

Summary of results

- The distance is the same for both the general and the restricted DCJ-indel models...

Conclusions and Future Work

Summary of results

- The distance is the same for both the general and the restricted DCJ-indel models...

  ... therefore the distance can be computed in $\mathcal{O}(n)$ time.

Conclusions and Future Work

Summary of results

- The distance is the same for both the general and the restricted DCJ-indel models...

  ... therefore the distance can be computed in $\mathcal{O}(n)$ time.

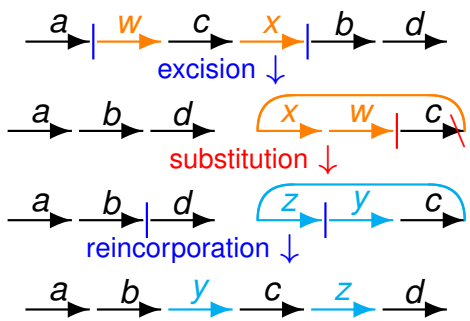
- A restricted scenario can be obtained in $\mathcal{O}(n \log n)$ time.

Conclusions and Future Work

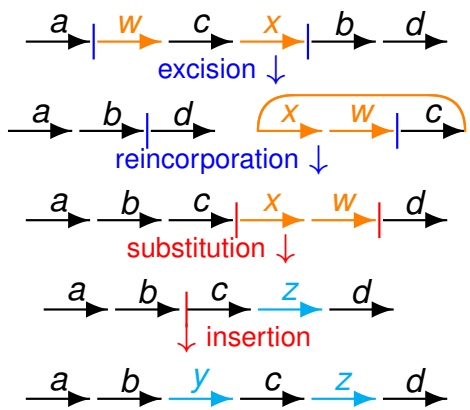
Future work

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

General scenario



Restricted scenario

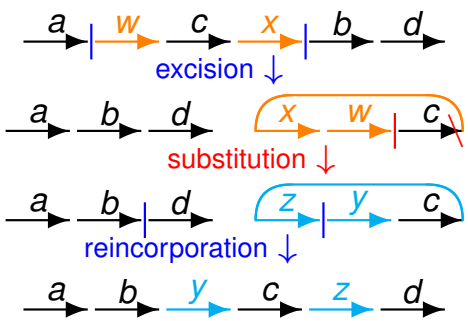


Conclusions and Future Work

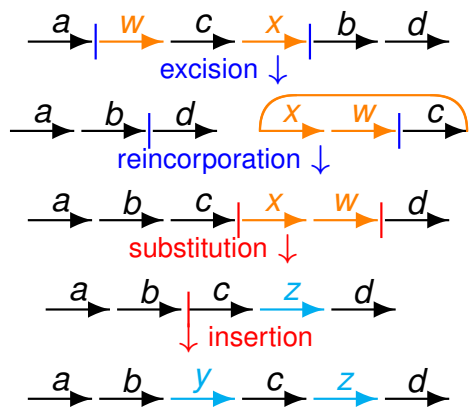
Future work

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

General scenario



Restricted scenario



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

Acknowledgements

This research was supported by:

- Inmetro
- CNPq (grant PROMETRO 563087/2010-2)
- Bielefeld University (Germany)

Acknowledgements

This research was supported by:

- Inmetro
- CNPq (grant PROMETRO 563087/2010-2)
- Bielefeld University (Germany)

Thank you for your attention!