

The DCJ-indel model and its potential to improve homology assignment

Marília Braga

Inmetro - Brazil

Overview

1 Motivation

2 DCJ model

- Master graph and its components
- DCJ distance
- Handling indels

3 Using the DCJ model to improve annotation

- (Ongoing work)
- Substitution or missing homology?
- The Rickettsia database
- Resolving duplications

4 Summary

Motivation

Overview

- 1 Motivation**
- 2 DCJ model**
 - Master graph and its components
 - DCJ distance
 - Handling indels
- 3 Using the DCJ model to improve annotation**
 - (Ongoing work)
 - Substitution or missing homology?
 - The Rickettsia database
 - Resolving duplications
- 4 Summary**

Motivation

Comparing genomes

A _____

B _____

Motivation

Comparing genomes

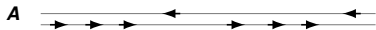
A _____

1. Finding genes

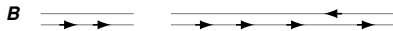
B _____

Motivation

Comparing genomes



1. Finding genes



Motivation

Comparing genomes

A → → → → ← → → → ←

1. Finding genes

B → → → → ← →

Motivation

Comparing genomes

A → → → ← → → → ←

2. Annotation (homology assignment)

B → → → → → ← →

Motivation

Comparing genomes

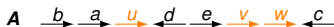
Common genes:

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

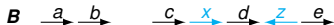
Unique genes:

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

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



2. Annotation (homology assignment)



Motivation

Comparing genomes

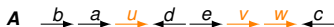
Common genes:

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

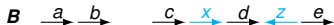
Unique genes:

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

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



3. Computing distance and/or sorting scenario



Motivation

Comparing genomes

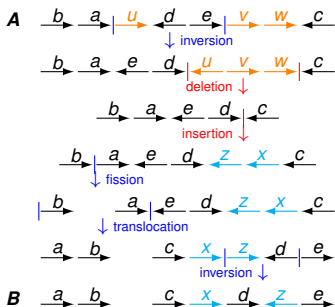
Common genes:

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

Unique genes:

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

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



Motivation

Comparing genomes

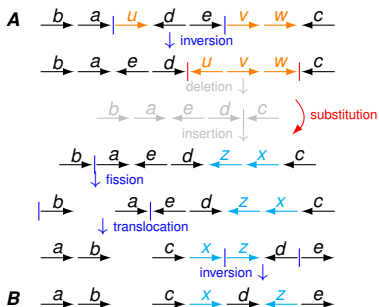
Common genes:

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

Unique genes:

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

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



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

Motivation

Comparing genomes

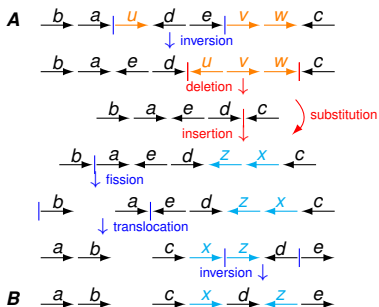
Common genes:

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

Unique genes:

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

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



Insertions and *Deletions* - (Indels)
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)

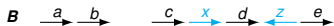
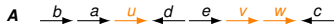
DCJ model

Overview

- 1 Motivation
- 2 **DCJ model**
 - Master graph and its components
 - DCJ distance
 - Handling indels
- 3 **Using the DCJ model to improve annotation**
 - (Ongoing work)
 - Substitution or missing homology?
 - The Rickettsia database
 - Resolving duplications
- 4 **Summary**

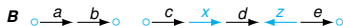
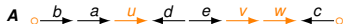
DCJ model

Master graph $R(A, B)$ (no duplicated genes) [Friedberg *et al.*, 2008]



DCJ model

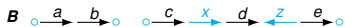
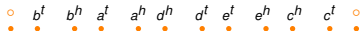
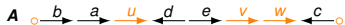
Master graph $R(A, B)$ (no duplicated genes) [Friedberg *et al.*, 2008]



(The symbol \circ represents the telomeres in both genomes.)

DCJ model

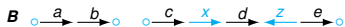
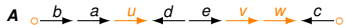
Master graph $R(A, B)$ (no duplicated genes) [Friedberg *et al.*, 2008]



(The symbol \circ represents the telomeres in both genomes.)

DCJ model

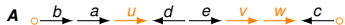
Master graph $R(A, B)$ (no duplicated genes) [Friedberg *et al.*, 2008]



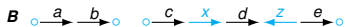
(The symbol \circ represents the telomeres in both genomes.)

DCJ model

Master graph $R(A, B)$ (no duplicated genes) [Friedberg *et al.*, 2008]



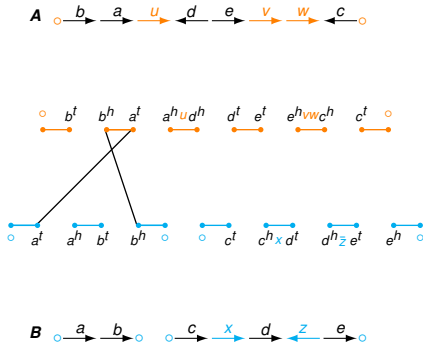
Components of $R(A, B)$:



(The symbol \circ represents the telomeres in both genomes.)

DCJ model

Master graph $R(A, B)$ (no duplicated genes) [Friedberg *et al.*, 2008]



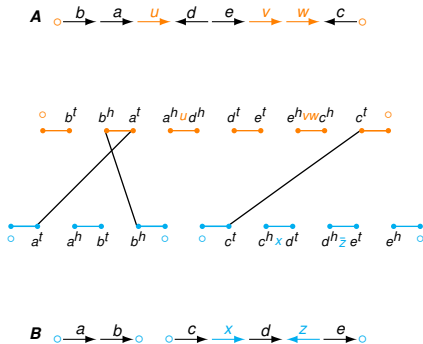
Components of $R(A, B)$:

One clean BB -path

(The symbol \circ represents the telomeres in both genomes.)

DCJ model

Master graph $R(A, B)$ (no duplicated genes) [Friedberg *et al.*, 2008]



Components of $R(A, B)$:

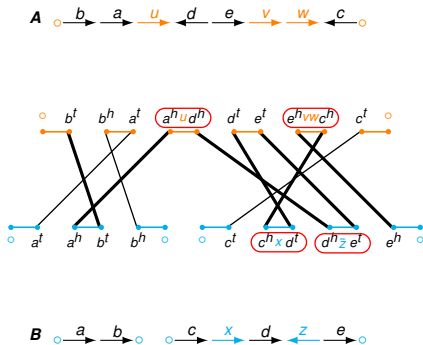
One clean BB -path

One clean AB -path

(The symbol \circ represents the telomeres in both genomes.)

DCJ model

Master graph $R(A, B)$ (no duplicated genes) [Friedberg *et al.*, 2008]



Components of $R(A, B)$:

One clean BB -path

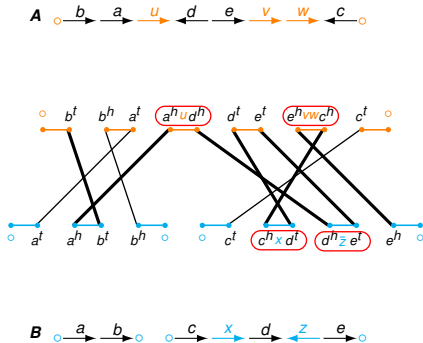
One clean AB -path

One AB -path with four labels

(The symbol \circ represents the telomeres in both genomes.)

DCJ model

Master graph $R(A, B)$ (no duplicated genes) [Friedberg *et al.*, 2008]



Components of $R(A, B)$:

One clean BB -path

One clean AB -path

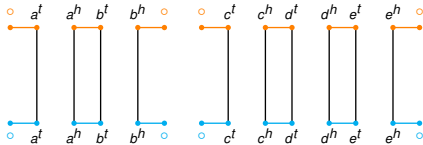
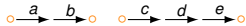
One AB -path with four labels

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

(The symbol \circ represents the telomeres in both genomes.)

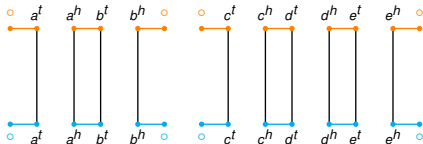
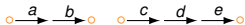
DCJ model

For identical (or sorted) genomes...



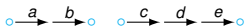
DCJ model

For identical (or sorted) genomes...



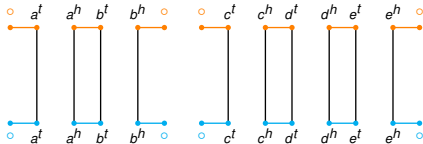
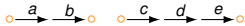
Components of $R(A, B)$:

Only short cycles and short AB -paths



DCJ model

For identical (or sorted) genomes...



Components of $R(A, B)$:

Only short cycles and short AB -paths

(DCJs need to increase the number of components)



DCJ model

DCJ distance

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

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

Types of DCJ operations:

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

DCJ model

DCJ distance

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

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

Types of DCJ operations:

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

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

DCJ model

DCJ distance

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

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

Types of DCJ operations:

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

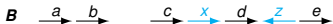
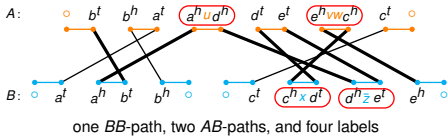
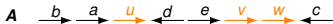
Bergeron *et al.* (2006): there is an optimal DCJ at each sorting step.

$$\text{DCJ distance of } A \text{ and } B: d_{\text{DCJ}}(A, B) = |\mathcal{G}| - (c + \frac{b}{2})$$

(\mathcal{G} : set of common genes of A and B)

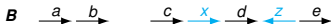
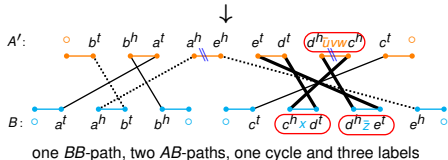
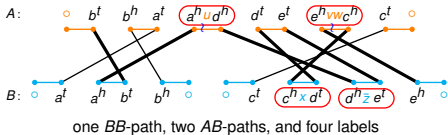
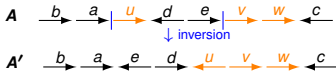
DCJ model

Handling indels - accumulating labels in both genomes:



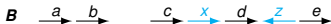
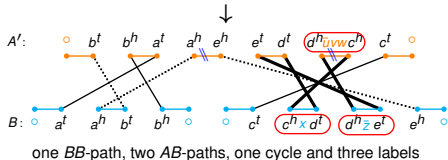
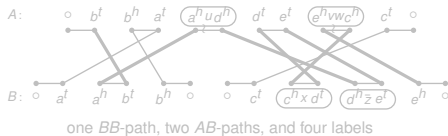
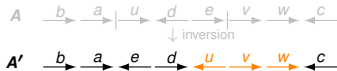
DCJ model

Handling indels - accumulating labels in both genomes:



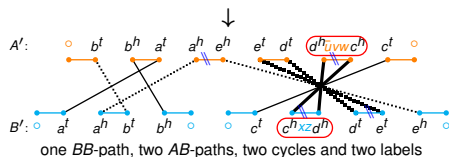
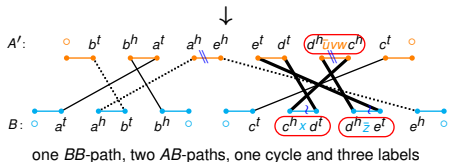
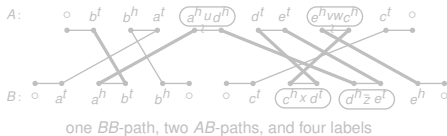
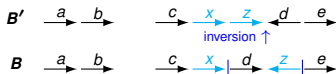
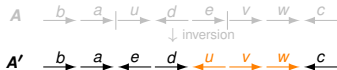
DCJ model

Handling indels - accumulating labels in both genomes:



DCJ model

Handling indels - accumulating labels in both genomes:

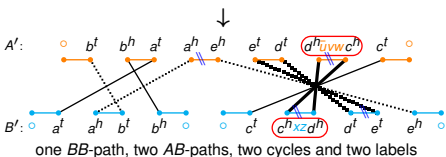
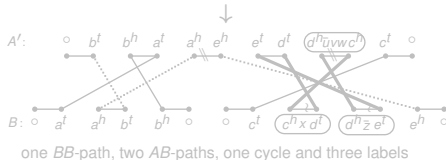
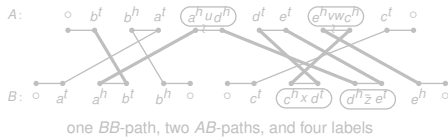
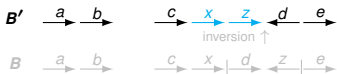


DCJ model

Handling indels - accumulating labels in both genomes:



⋮

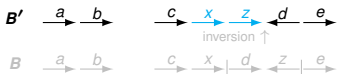


DCJ model

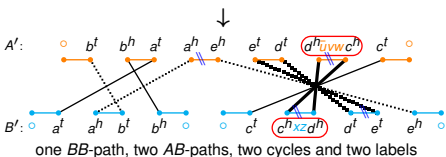
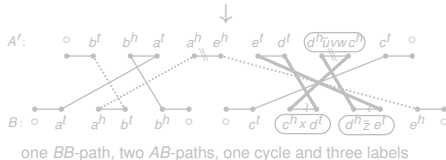
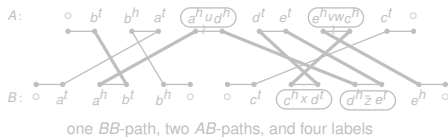
Handling indels - accumulating labels in both genomes:



⋮



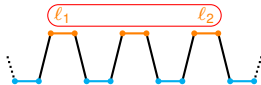
(DCJ operations can increase the number of components and accumulate labels.)



DCJ model

Handling indels - the concept of *run*

Accumulating
labels:

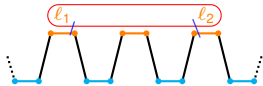


two labels

DCJ model

Handling indels - the concept of *run*

Accumulating
labels:



two labels



one label



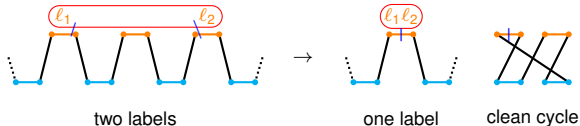
clean cycle

(split DCJ)

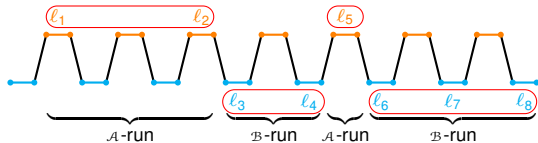
DCJ model

Handling indels - the concept of run

Accumulating labels:



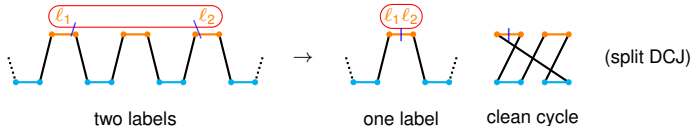
Runs:



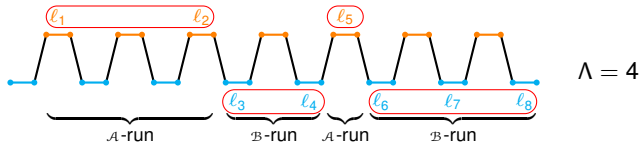
DCJ model

Handling indels - the concept of *run*

Accumulating labels:

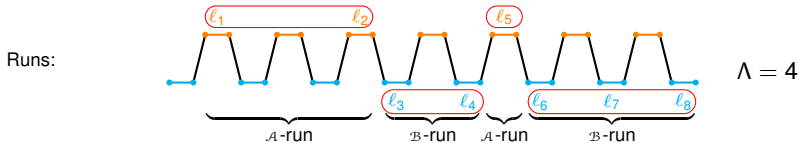


Runs:



DCJ model

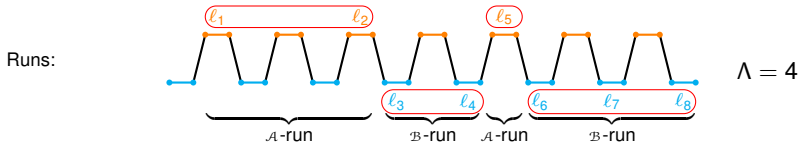
Handling indels - the concept of *run*



Each **run** can be entirely **accumulated** into a single label with split DCJs.

DCJ model

Handling indels - the concept of *run*

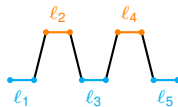


Each **run** can be entirely **accumulated** into a single label with split DCJs.

A split DCJ is always optimal.

DCJ model

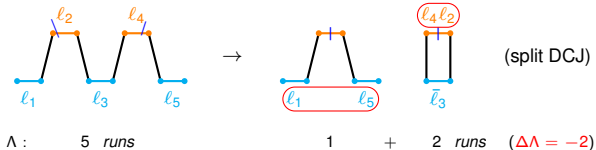
A rearrangement can merge at most two \mathcal{A} -runs and two \mathcal{B} -runs:



Λ : 5 runs

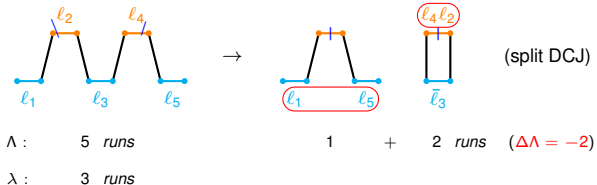
DCJ model

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



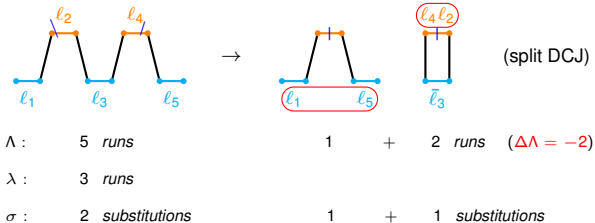
DCJ model

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



DCJ model

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



DCJ model

Handling indels - the concept of *potential*

DCJ model

Handling indels - the concept of *potential*

Indel-potential of a component P [WABI 2010]

Minimum number of **runs** obtained sorting P with **split** DCJs:

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

DCJ model

Handling indels - the concept of *potential*

Indel-potential of a component P [WABI 2010]

Minimum number of **runs** obtained sorting P with **split** DCJs:

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

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

Minimum number of **pairs of runs** obtained sorting P with **split** DCJs:

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

DCJ model

Handling indels - the concept of *potential*

Indel-potential of a component P [WABI 2010]

Minimum number of **runs** obtained sorting P with **split** DCJs:

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

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

Minimum number of **pairs of runs** obtained sorting P with **split** DCJs:

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

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

DCJ model

Distances with indels

DCJ model

Distances with indels

DCJ-indel distance [WABI 2010]

- ▶ An **upper bound** is given by: $d_{DCJ}^{id}(A, B) \leq d_{DCJ}(A, B) + \sum_{P \in R(A, B)} \lambda(P)$
- ▶ The exact distance can be computed in **linear time**.

DCJ model

Distances with indels

DCJ-indel distance [WABI 2010]

- ▶ An **upper bound** is given by: $d_{DCJ}^{id}(A, B) \leq d_{DCJ}(A, B) + \sum_{P \in R(A, B)} \lambda(P)$
- ▶ The exact distance can be computed in **linear time**.

DCJ-substitution distance [RECOMB-CG 2011]

- ▶ An **upper bound** is given by: $d_{DCJ}^{sb}(A, B) \leq d_{DCJ}(A, B) + \sum_{P \in R(A, B)} \sigma(P)$
- ▶ The exact distance can be computed in **linear time**.

Using the DCJ model to improve annotation

Overview

- 1 Motivation
- 2 DCJ model
 - Master graph and its components
 - DCJ distance
 - Handling indels
- 3 **Using the DCJ model to improve annotation**
 - (Ongoing work)
 - Substitution or missing homology?
 - The Rickettsia database
 - Resolving duplications
- 4 Summary

Using the DCJ model to improve annotation

- ▶ The labels in the same component of the master graph seem to be somehow related.
- ▶ This includes, but is not limited to, the case of adjacencies (when the unknown or mis-annotated genes are adjacent to genes of the same family in both genomes).
- ▶ Could this information be used to improve the annotation (missing homology assignment and duplicate disambiguation) of the genomes?

Using the DCJ model to improve annotation

Substitution or homology? *A*-label and *B*-label in **distinct** components

A $\xrightarrow{a} \xrightarrow{d} \xrightarrow{c} \xrightarrow{b} \xrightarrow{x} \xrightarrow{e}$

B $\xrightarrow{a} \xrightarrow{y} \xrightarrow{b} \xrightarrow{c} \xrightarrow{d} \xrightarrow{e}$

Using the DCJ model to improve annotation

Substitution or homology? *A*-label and *B*-label in **distinct** components

A $\xrightarrow{a} \xrightarrow{d} \xrightarrow{c} \xrightarrow{b} \xrightarrow{x} \xrightarrow{e}$

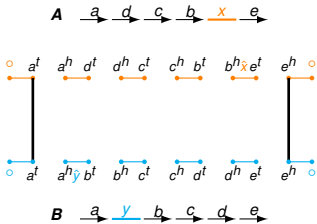
○ $\xrightarrow{a^t}$ $\xrightarrow{a^h d^t}$ $\xrightarrow{d^h c^t}$ $\xrightarrow{c^h b^t}$ $\xrightarrow{b^h \bar{x} e^t}$ $\xrightarrow{e^h}$ ○

○ $\xrightarrow{a^t}$ $\xrightarrow{a^h \bar{y} b^t}$ $\xrightarrow{b^h c^t}$ $\xrightarrow{c^h d^t}$ $\xrightarrow{d^h e^t}$ $\xrightarrow{e^h}$ ○

B $\xrightarrow{a} \xrightarrow{y} \xrightarrow{b} \xrightarrow{c} \xrightarrow{d} \xrightarrow{e}$

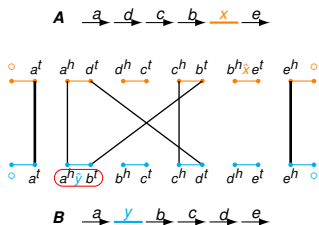
Using the DCJ model to improve annotation

Substitution or homology? *A*-label and *B*-label in **distinct** components



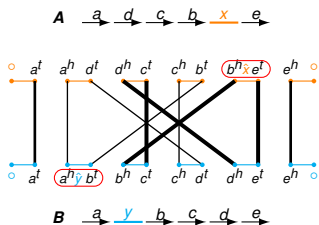
Using the DCJ model to improve annotation

Substitution or homology? *A*-label and *B*-label in **distinct** components



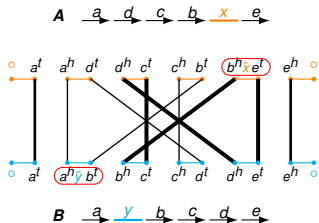
Using the DCJ model to improve annotation

Substitution or homology? *A*-label and *B*-label in **distinct** components



Using the DCJ model to improve annotation

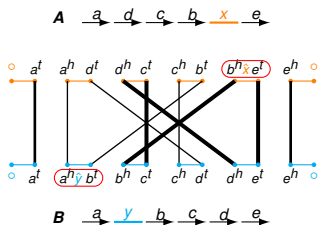
Substitution or homology? *A*-label and *B*-label in **distinct** components



$$\sigma = 1 + 1 \text{ (two substitutions)}$$

Using the DCJ model to improve annotation

Substitution or homology? *A*-label and *B*-label in **distinct** components

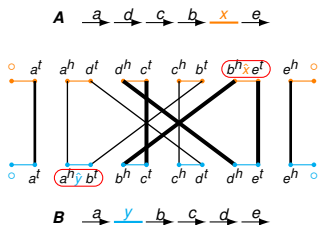


$$\sigma = 1 + 1 \text{ (two substitutions)}$$

$$\text{DCJ distance} = 5 - 2 - 2/2 = 2$$

Using the DCJ model to improve annotation

Substitution or homology? *A*-label and *B*-label in **distinct** components



$$\sigma = 1 + 1 \text{ (two substitutions)}$$

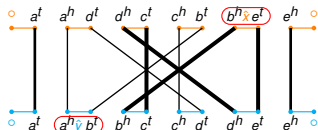
$$\text{DCJ distance} = 5 - 2 - 2/2 = 2$$

$$\text{DCJ-substitution distance} = 4$$

Using the DCJ model to improve annotation

Substitution or homology? *A*-label and *B*-label in **distinct** components

A $\xrightarrow{a} \xrightarrow{d} \xrightarrow{c} \xrightarrow{b} \xrightarrow{x} \xrightarrow{e}$



B $\xrightarrow{a} \xrightarrow{y} \xrightarrow{b} \xrightarrow{c} \xrightarrow{d} \xrightarrow{e}$

A $\xrightarrow{a} \xrightarrow{d} \xrightarrow{c} \xrightarrow{b} \xrightarrow{x} \xrightarrow{e}$

B $\xrightarrow{a} \xrightarrow{x} \xrightarrow{b} \xrightarrow{c} \xrightarrow{d} \xrightarrow{e}$

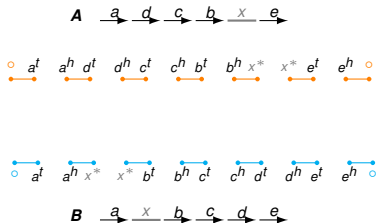
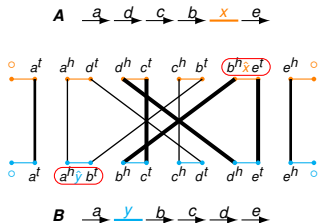
$$\sigma = 1 + 1 \text{ (two substitutions)}$$

$$\text{DCJ distance} = 5 - 2 - 2/2 = 2$$

$$\text{DCJ-substitution distance} = 4$$

Using the DCJ model to improve annotation

Substitution or homology? *A*-label and *B*-label in **distinct** components



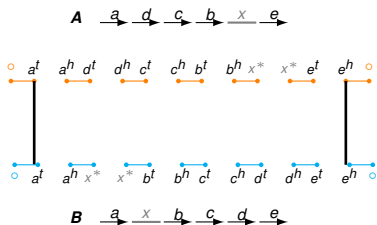
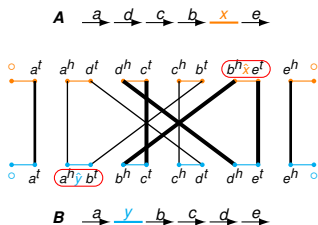
$$\sigma = 1 + 1 \text{ (two substitutions)}$$

$$\text{DCJ distance} = 5 - 2 - 2/2 = 2$$

$$\text{DCJ-substitution distance} = 4$$

Using the DCJ model to improve annotation

Substitution or homology? *A*-label and *B*-label in **distinct** components



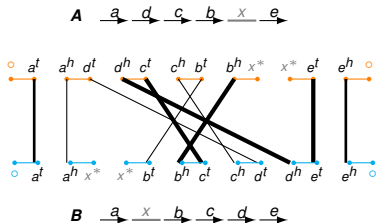
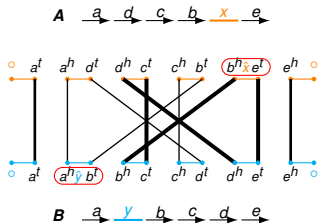
$$\sigma = 1 + 1 \text{ (two substitutions)}$$

$$\text{DCJ distance} = 5 - 2 - 2/2 = 2$$

$$\text{DCJ-substitution distance} = 4$$

Using the DCJ model to improve annotation

Substitution or homology? *A*-label and *B*-label in **distinct** components



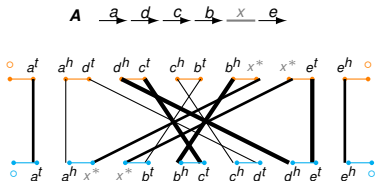
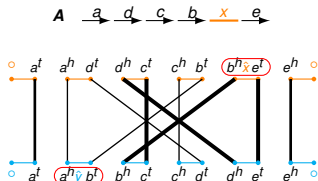
$$\sigma = 1 + 1 \text{ (two substitutions)}$$

$$\text{DCJ distance} = 5 - 2 - 2/2 = 2$$

$$\text{DCJ-substitution distance} = 4$$

Using the DCJ model to improve annotation

Substitution or homology? *A*-label and *B*-label in **distinct** components



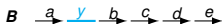
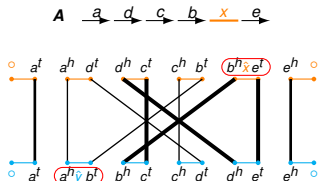
$$\sigma = 1 + 1 \text{ (two substitutions)}$$

$$\text{DCJ distance} = 5 - 2 - 2/2 = 2$$

$$\text{DCJ-substitution distance} = 4$$

Using the DCJ model to improve annotation

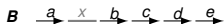
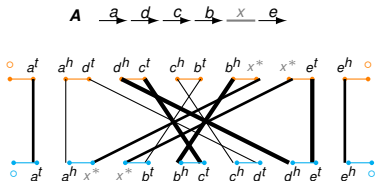
Substitution or homology? *A*-label and *B*-label in **distinct** components



$$\sigma = 1 + 1 \text{ (two substitutions)}$$

$$\text{DCJ distance} = 5 - 2 - 2/2 = 2$$

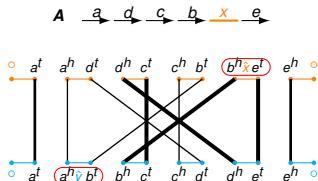
$$\text{DCJ-substitution distance} = 4$$



$$\sigma = 0 \text{ (no substitution)}$$

Using the DCJ model to improve annotation

Substitution or homology? *A*-label and *B*-label in **distinct** components

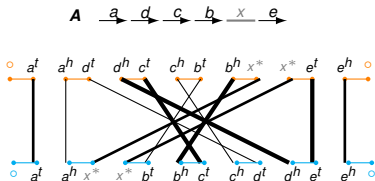


B $\xrightarrow{a} \xrightarrow{y} \xrightarrow{b} \xrightarrow{c} \xrightarrow{d} \xrightarrow{e}$

$$\sigma = 1 + 1 \text{ (two substitutions)}$$

$$\text{DCJ distance} = 5 - 2 - 2/2 = 2$$

$$\text{DCJ-substitution distance} = 4$$



B $\xrightarrow{a} \xrightarrow{x} \xrightarrow{b} \xrightarrow{c} \xrightarrow{d} \xrightarrow{e}$

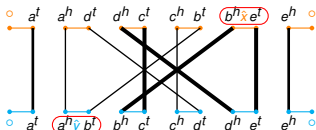
$$\sigma = 0 \text{ (no substitution)}$$

$$\text{DCJ distance} = 6 - 1 - 2/2 = 4$$

Using the DCJ model to improve annotation

Substitution or homology? *A*-label and *B*-label in **distinct** components

A $\xrightarrow{a} \xrightarrow{d} \xrightarrow{c} \xrightarrow{b} \xrightarrow{x} \xrightarrow{e}$



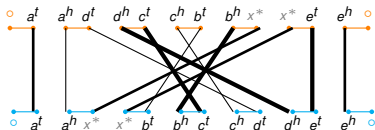
B $\xrightarrow{a} \xrightarrow{y} \xrightarrow{b} \xrightarrow{c} \xrightarrow{d} \xrightarrow{e}$

$$\sigma = 1 + 1 \text{ (two substitutions)}$$

$$\text{DCJ distance} = 5 - 2 - 2/2 = 2$$

$$\text{DCJ-substitution distance} = 4$$

A $\xrightarrow{a} \xrightarrow{d} \xrightarrow{c} \xrightarrow{b} \xrightarrow{x} \xrightarrow{e}$



B $\xrightarrow{a} \xrightarrow{x} \xrightarrow{b} \xrightarrow{c} \xrightarrow{d} \xrightarrow{e}$

$$\sigma = 0 \text{ (no substitution)}$$

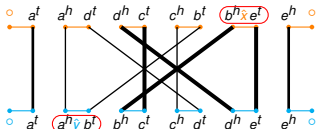
$$\text{DCJ distance} = 6 - 1 - 2/2 = 4$$

$$\text{DCJ-substitution distance} = 4$$

Using the DCJ model to improve annotation

Substitution or homology? *A*-label and *B*-label in **distinct** components

A $\xrightarrow{a} \xrightarrow{d} \xrightarrow{c} \xrightarrow{b} \xrightarrow{x} \xrightarrow{e}$



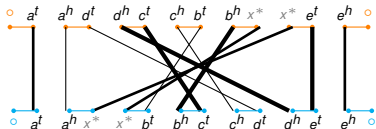
B $\xrightarrow{a} \xrightarrow{y} \xrightarrow{b} \xrightarrow{c} \xrightarrow{d} \xrightarrow{e}$

$$\sigma = 1 + 1 \text{ (two substitutions)}$$

$$\text{DCJ distance} = 5 - 2 - 2/2 = 2$$

$$\text{DCJ-substitution distance} = 4$$

A $\xrightarrow{a} \xrightarrow{d} \xrightarrow{c} \xrightarrow{b} \xrightarrow{x} \xrightarrow{e}$



B $\xrightarrow{a} \xrightarrow{x} \xrightarrow{b} \xrightarrow{c} \xrightarrow{d} \xrightarrow{e}$

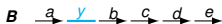
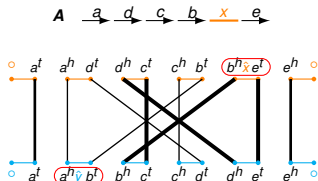
$$\sigma = 0 \text{ (no substitution)}$$

$$\text{DCJ distance} = 6 - 1 - 2/2 = 4$$

$$\text{DCJ-substitution distance} = 4$$

Using the DCJ model to improve annotation

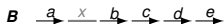
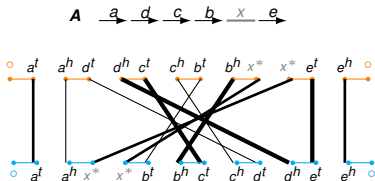
Substitution or homology? *A*-label and *B*-label in **distinct** components



$$\sigma = 1 + 1 \text{ (two substitutions)}$$

$$\text{DCJ distance} = 5 - 2 - 2/2 = 2$$

$$\text{DCJ-substitution distance} = 4$$



$$\sigma = 0 \text{ (no substitution)}$$

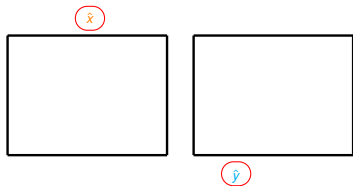
$$\text{DCJ distance} = 6 - 1 - 2/2 = 4$$

$$\text{DCJ-substitution distance} = 4$$

The distance does not decrease if *x* and *y* are homologous, independently of their relative orientations.

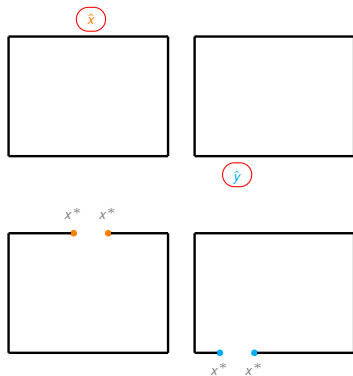
Using the DCJ model to improve annotation

Substitution or homology? *A*-label and *B*-label in **distinct** components



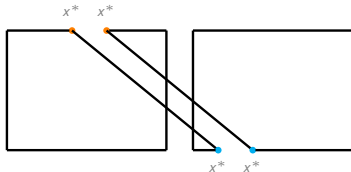
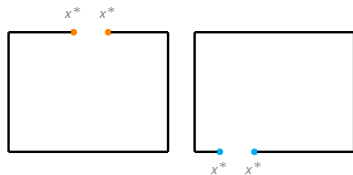
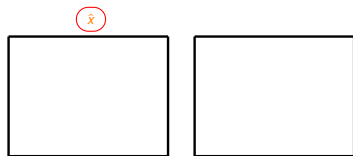
Using the DCJ model to improve annotation

Substitution or homology? *A*-label and *B*-label in **distinct** components



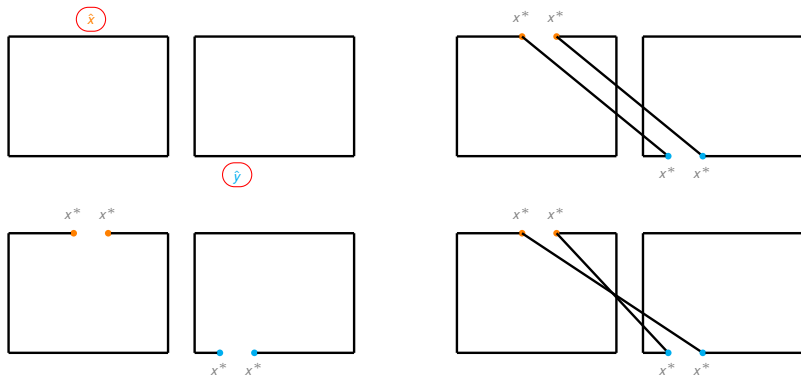
Using the DCJ model to improve annotation

Substitution or homology? *A*-label and *B*-label in **distinct** components



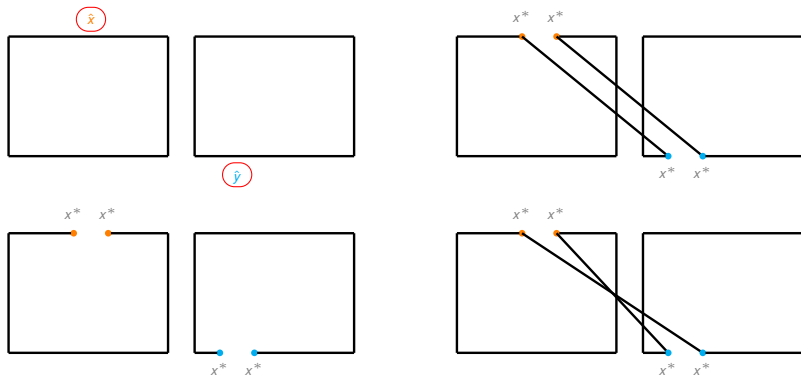
Using the DCJ model to improve annotation

Substitution or homology? *A*-label and *B*-label in **distinct** components



Using the DCJ model to improve annotation

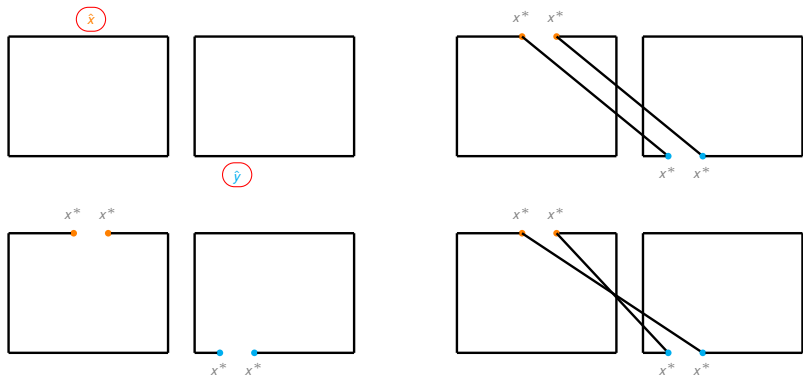
Substitution or homology? *A*-label and *B*-label in **distinct** components



We “remove” two subst., but increase the number of common genes and decrease the number of comp.

Using the DCJ model to improve annotation

Substitution or homology? *A*-label and *B*-label in **distinct** components



We “remove” two subst., but increase the number of common genes and decrease the number of comp.

The distance does not decrease if x and y are homologous, independently of their relative orientations.

Using the DCJ model to improve annotation

Substitution or homology? *A*-label and *B*-label in the **same** component

A $\xrightarrow{a} \xrightarrow{c} \xrightarrow{x} \xrightarrow{b} \xrightarrow{d}$

B $\xrightarrow{a} \xrightarrow{b} \xrightarrow{y} \xrightarrow{c} \xrightarrow{d}$

Using the DCJ model to improve annotation

Substitution or homology? *A*-label and *B*-label in the **same** component

A $\xrightarrow{a} \xrightarrow{c} \xrightarrow{x} \xrightarrow{b} \xrightarrow{d}$

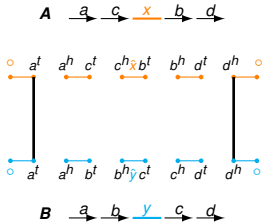
○ $\xrightarrow{a^t}$ $\xrightarrow{a^h} \xrightarrow{c^t}$ $\xrightarrow{c^h} \xrightarrow{x} \xrightarrow{b^t}$ $\xrightarrow{b^h} \xrightarrow{d^t}$ $\xrightarrow{d^h}$ ○

○ $\xrightarrow{a^t}$ $\xrightarrow{a^h} \xrightarrow{b^t}$ $\xrightarrow{b^h} \xrightarrow{y} \xrightarrow{c^t}$ $\xrightarrow{c^h} \xrightarrow{d^t}$ $\xrightarrow{d^h}$ ○

B $\xrightarrow{a} \xrightarrow{b} \xrightarrow{y} \xrightarrow{c} \xrightarrow{d}$

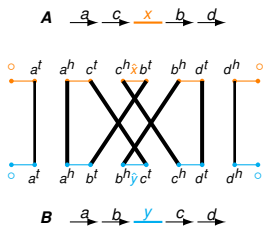
Using the DCJ model to improve annotation

Substitution or homology? *A*-label and *B*-label in the **same** component



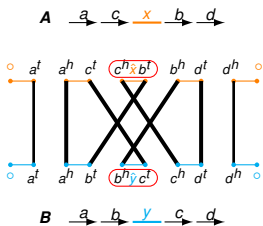
Using the DCJ model to improve annotation

Substitution or homology? *A*-label and *B*-label in the **same** component



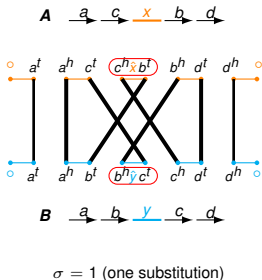
Using the DCJ model to improve annotation

Substitution or homology? *A*-label and *B*-label in the **same** component



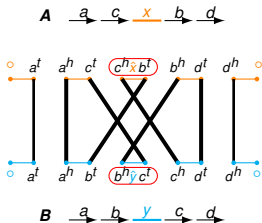
Using the DCJ model to improve annotation

Substitution or homology? *A*-label and *B*-label in the **same** component



Using the DCJ model to improve annotation

Substitution or homology? *A*-label and *B*-label in the **same** component

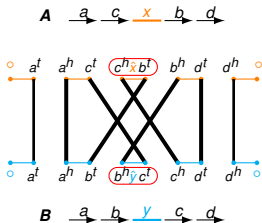


$$\sigma = 1 \text{ (one substitution)}$$

$$\text{DCJ distance} = 4 - 1 - 2/2 = 2$$

Using the DCJ model to improve annotation

Substitution or homology? *A*-label and *B*-label in the **same** component



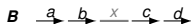
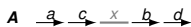
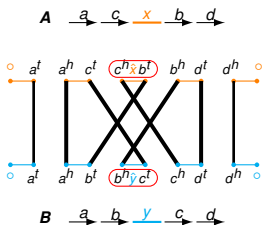
$\sigma = 1$ (one substitution)

DCJ distance = $4 - 1 - 2/2 = 2$

DCJ-substitution distance = 3

Using the DCJ model to improve annotation

Substitution or homology? *A*-label and *B*-label in the **same** component



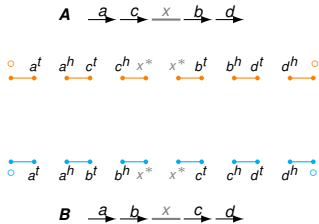
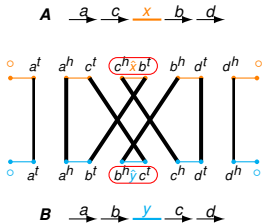
$$\sigma = 1 \text{ (one substitution)}$$

$$\text{DCJ distance} = 4 - 1 - 2/2 = 2$$

$$\text{DCJ-substitution distance} = 3$$

Using the DCJ model to improve annotation

Substitution or homology? *A*-label and *B*-label in the **same** component



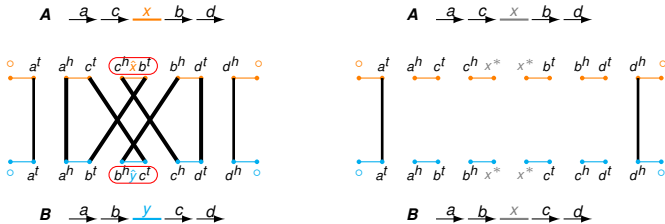
$\sigma = 1$ (one substitution)

DCJ distance = $4 - 1 - 2/2 = 2$

DCJ-substitution distance = 3

Using the DCJ model to improve annotation

Substitution or homology? *A*-label and *B*-label in the **same** component



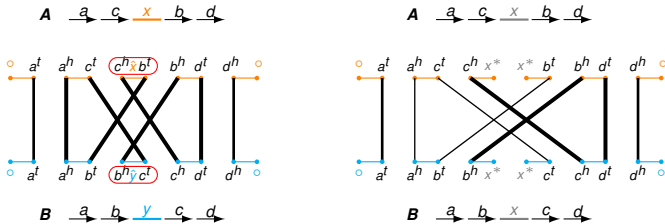
$\sigma = 1$ (one substitution)

DCJ distance = $4 - 1 - 2/2 = 2$

DCJ-substitution distance = 3

Using the DCJ model to improve annotation

Substitution or homology? *A*-label and *B*-label in the **same** component



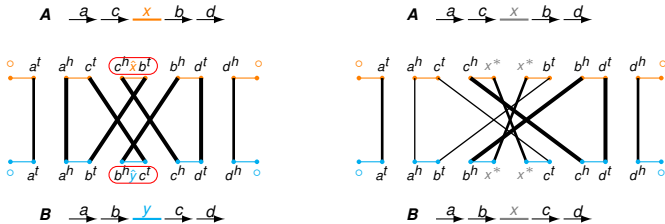
$\sigma = 1$ (one substitution)

DCJ distance = $4 - 1 - 2/2 = 2$

DCJ-substitution distance = 3

Using the DCJ model to improve annotation

Substitution or homology? *A*-label and *B*-label in the **same** component



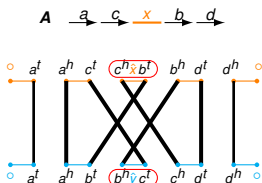
$\sigma = 1$ (one substitution)

DCJ distance = $4 - 1 - 2/2 = 2$

DCJ-substitution distance = 3

Using the DCJ model to improve annotation

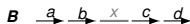
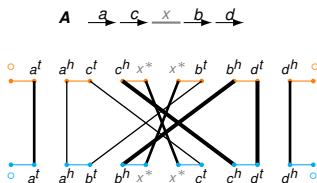
Substitution or homology? *A*-label and *B*-label in the **same** component



$\sigma = 1$ (one substitution)

DCJ distance = $4 - 1 - 2/2 = 2$

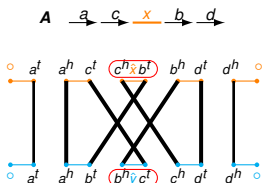
DCJ-substitution distance = 3



$\sigma = 0$ (no substitution)

Using the DCJ model to improve annotation

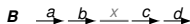
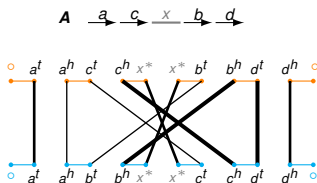
Substitution or homology? *A*-label and *B*-label in the **same** component



$\sigma = 1$ (one substitution)

DCJ distance = $4 - 1 - 2/2 = 2$

DCJ-substitution distance = 3

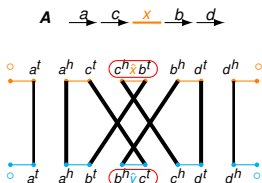


$\sigma = 0$ (no substitution)

DCJ distance = $5 - 1 - 2/2 = 3$

Using the DCJ model to improve annotation

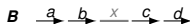
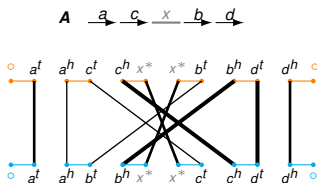
Substitution or homology? *A*-label and *B*-label in the **same** component



$\sigma = 1$ (one substitution)

DCJ distance = $4 - 1 - 2/2 = 2$

DCJ-substitution distance = 3



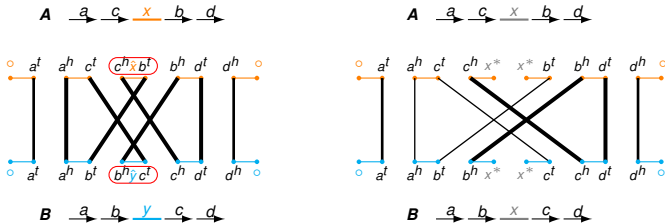
$\sigma = 0$ (no substitution)

DCJ distance = $5 - 1 - 2/2 = 3$

DCJ-substitution distance = 3

Using the DCJ model to improve annotation

Substitution or homology? *A*-label and *B*-label in the **same** component



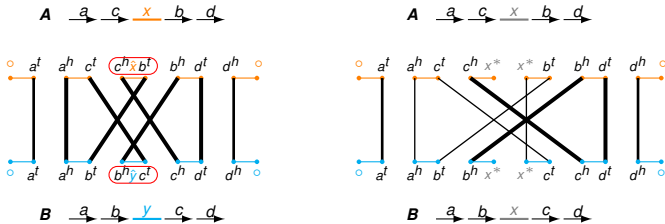
$\sigma = 1$ (one substitution)

DCJ distance = $4 - 1 - 2/2 = 2$

DCJ-substitution distance = 3

Using the DCJ model to improve annotation

Substitution or homology? *A*-label and *B*-label in the **same** component



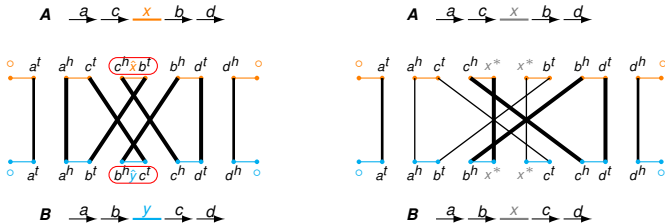
$\sigma = 1$ (one substitution)

DCJ distance = $4 - 1 - 2/2 = 2$

DCJ-substitution distance = 3

Using the DCJ model to improve annotation

Substitution or homology? *A*-label and *B*-label in the **same** component



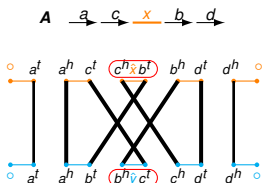
$\sigma = 1$ (one substitution)

DCJ distance = $4 - 1 - 2/2 = 2$

DCJ-substitution distance = 3

Using the DCJ model to improve annotation

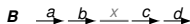
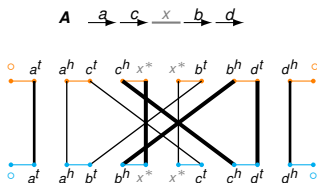
Substitution or homology? *A*-label and *B*-label in the **same** component



$\sigma = 1$ (one substitution)

DCJ distance = $4 - 1 - 2/2 = 2$

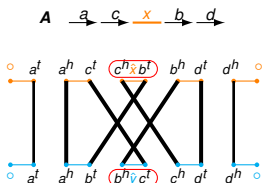
DCJ-substitution distance = 3



$\sigma = 0$ (no substitution)

Using the DCJ model to improve annotation

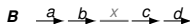
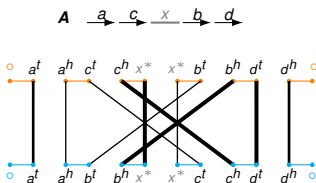
Substitution or homology? *A*-label and *B*-label in the **same** component



$\sigma = 1$ (one substitution)

$$\text{DCJ distance} = 4 - 1 - 2/2 = 2$$

$$\text{DCJ-substitution distance} = 3$$

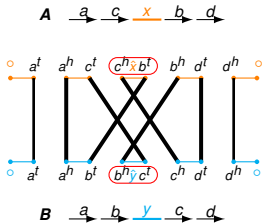


$\sigma = 0$ (no substitution)

$$\text{DCJ distance} = 5 - 2 - 2/2 = 2$$

Using the DCJ model to improve annotation

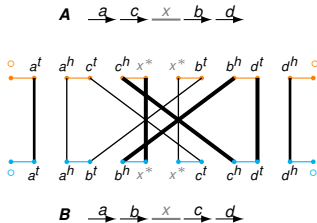
Substitution or homology? *A*-label and *B*-label in the **same** component



$\sigma = 1$ (one substitution)

DCJ distance = $4 - 1 - 2/2 = 2$

DCJ-substitution distance = 3



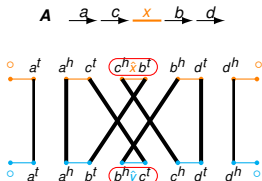
$\sigma = 0$ (no substitution)

DCJ distance = $5 - 2 - 2/2 = 2$

DCJ-substitution distance = 2

Using the DCJ model to improve annotation

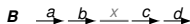
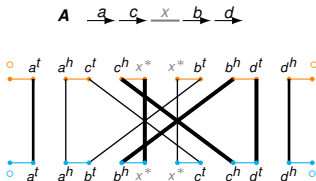
Substitution or homology? *A*-label and *B*-label in the **same** component



$\sigma = 1$ (one substitution)

DCJ distance = $4 - 1 - 2/2 = 2$

DCJ-substitution distance = 3



$\sigma = 0$ (no substitution)

DCJ distance = $5 - 2 - 2/2 = 2$

DCJ-substitution distance = 2

The distance decreases if *x* and *y* are homologous, for one of their two possible relative orientations.

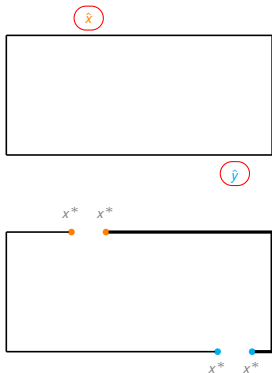
Using the DCJ model to improve annotation

Substitution or homology? *A*-label and *B*-label in the **same** component



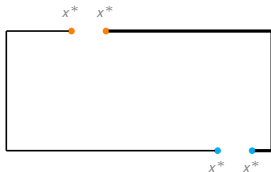
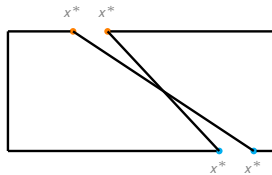
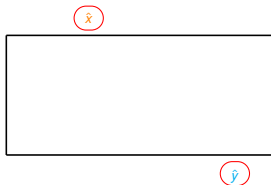
Using the DCJ model to improve annotation

Substitution or homology? *A*-label and *B*-label in the **same** component



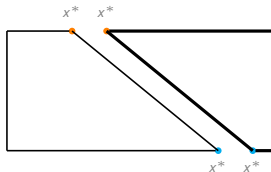
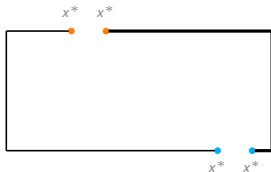
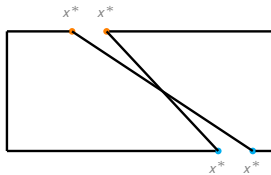
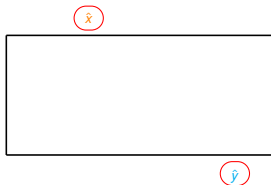
Using the DCJ model to improve annotation

Substitution or homology? *A*-label and *B*-label in the **same** component



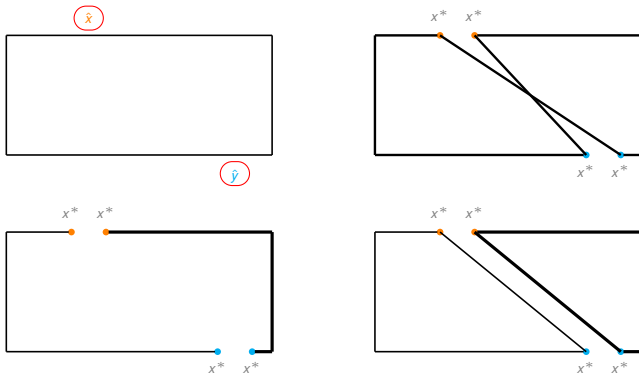
Using the DCJ model to improve annotation

Substitution or homology? *A*-label and *B*-label in the **same** component



Using the DCJ model to improve annotation

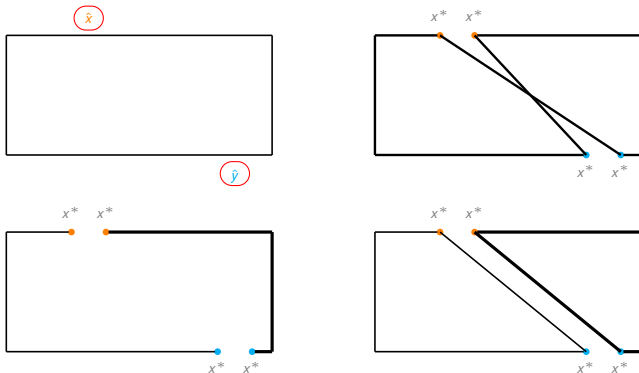
Substitution or homology? *A*-label and *B*-label in the **same** component



We “remove” one subst., increase the number of common genes and may increase the number of comp.

Using the DCJ model to improve annotation

Substitution or homology? *A*-label and *B*-label in the **same** component

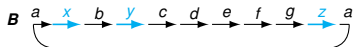
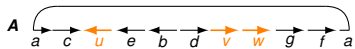


We “remove” one subst., increase the number of common genes and may increase the number of comp.

The distance decreases if x and y are homologous, for one of their two possible relative orientations.

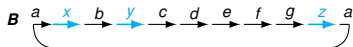
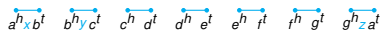
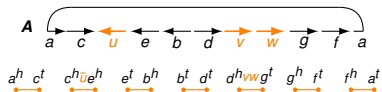
Using the DCJ model to improve annotation

Finding missing homologies: a more complex example



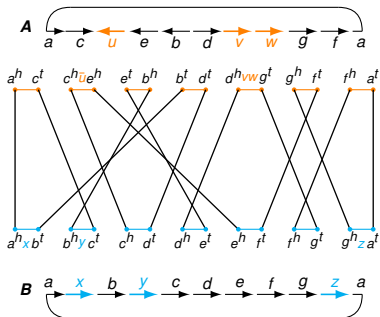
Using the DCJ model to improve annotation

Finding missing homologies: a more complex example



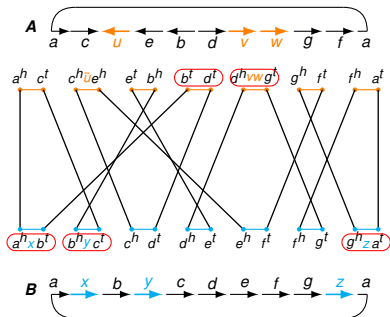
Using the DCJ model to improve annotation

Finding missing homologies: a more complex example



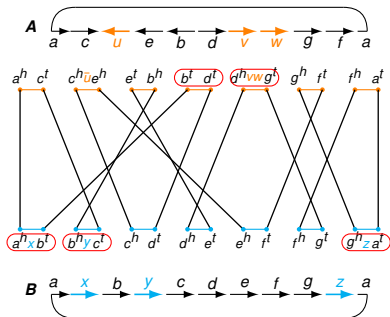
Using the DCJ model to improve annotation

Finding missing homologies: a more complex example



Using the DCJ model to improve annotation

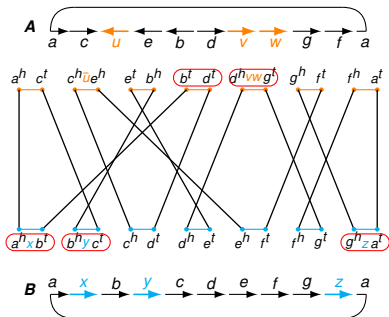
Finding missing homologies: a more complex example



$$\Lambda = 4; \sigma = 2 \text{ (two subst.)}$$

Using the DCJ model to improve annotation

Finding missing homologies: a more complex example

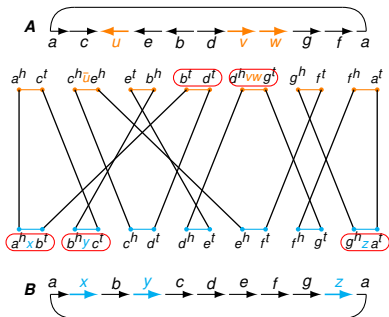


$$\Lambda = 4; \sigma = 2 \text{ (two subst.)}$$

$$\text{DCJ distance} = 7 - 1 = 6$$

Using the DCJ model to improve annotation

Finding missing homologies: a more complex example



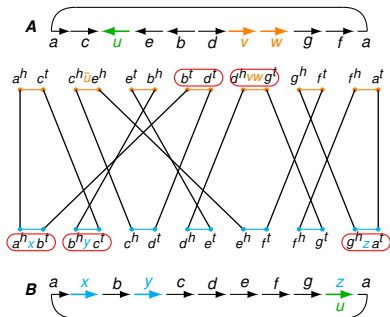
$$\Lambda = 4; \sigma = 2 \text{ (two subst.)}$$

$$\text{DCJ distance} = 7 - 1 = 6$$

$$\text{DCJ-substitution distance} = 8$$

Using the DCJ model to improve annotation

Finding missing homologies: a more complex example



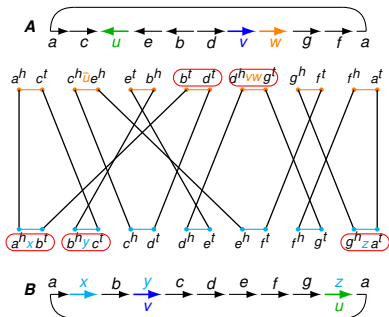
$\Lambda = 4; \sigma = 2$ (two subst.)

DCJ distance = $7 - 1 = 6$

DCJ-substitution distance = 8

Using the DCJ model to improve annotation

Finding missing homologies: a more complex example



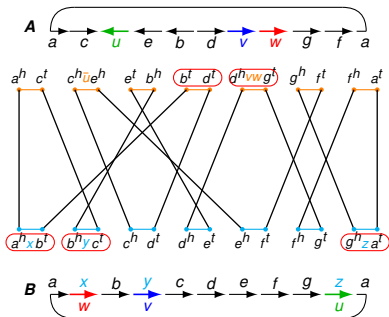
$\Lambda = 4$; $\sigma = 2$ (two subst.)

DCJ distance = $7 - 1 = 6$

DCJ-substitution distance = 8

Using the DCJ model to improve annotation

Finding missing homologies: a more complex example



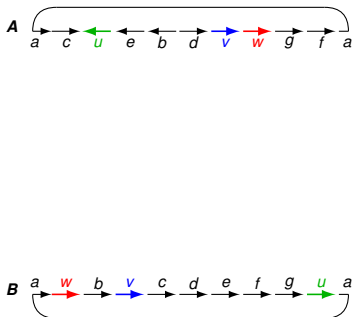
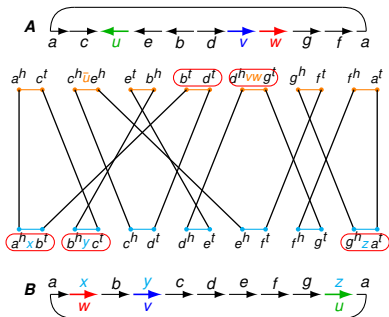
$\Lambda = 4$; $\sigma = 2$ (two subst.)

DCJ distance = $7 - 1 = 6$

DCJ-substitution distance = 8

Using the DCJ model to improve annotation

Finding missing homologies: a more complex example



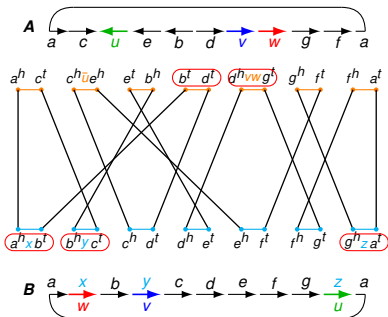
$$\Lambda = 4; \sigma = 2 \text{ (two subst.)}$$

$$\text{DCJ distance} = 7 - 1 = 6$$

$$\text{DCJ-substitution distance} = 8$$

Using the DCJ model to improve annotation

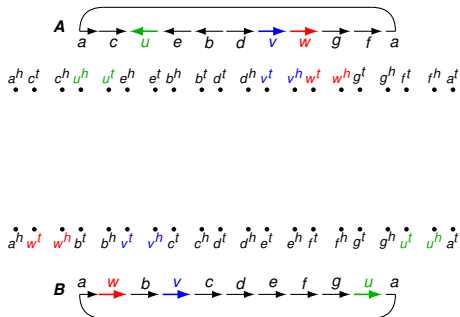
Finding missing homologies: a more complex example



$\Lambda = 4$; $\sigma = 2$ (two subst.)

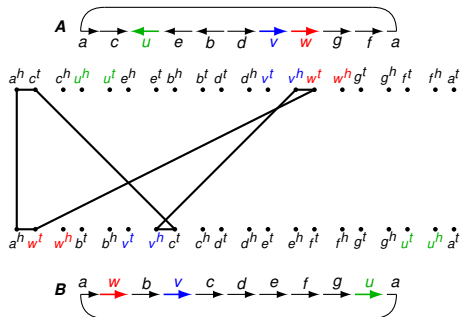
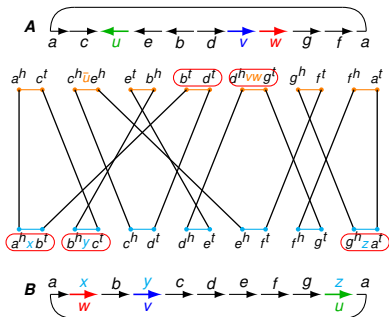
DCJ distance = $7 - 1 = 6$

DCJ-substitution distance = 8



Using the DCJ model to improve annotation

Finding missing homologies: a more complex example



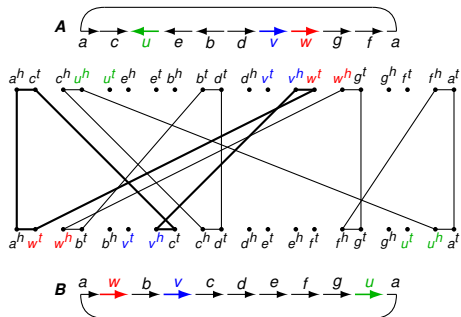
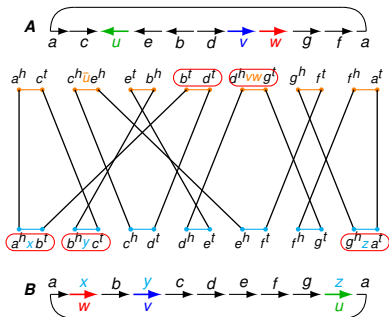
$\Lambda = 4$; $\sigma = 2$ (two subst.)

DCJ distance = $7 - 1 = 6$

DCJ-substitution distance = 8

Using the DCJ model to improve annotation

Finding missing homologies: a more complex example



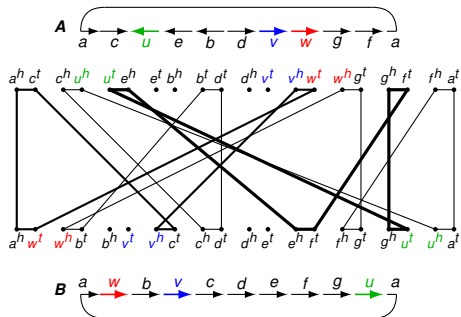
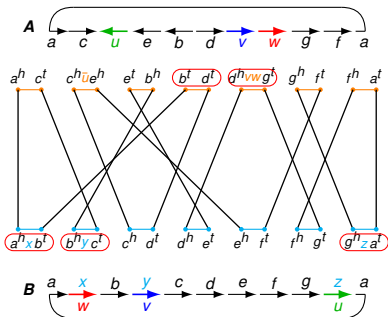
$$\Lambda = 4; \sigma = 2 \text{ (two subst.)}$$

$$\text{DCJ distance} = 7 - 1 = 6$$

$$\text{DCJ-substitution distance} = 8$$

Using the DCJ model to improve annotation

Finding missing homologies: a more complex example



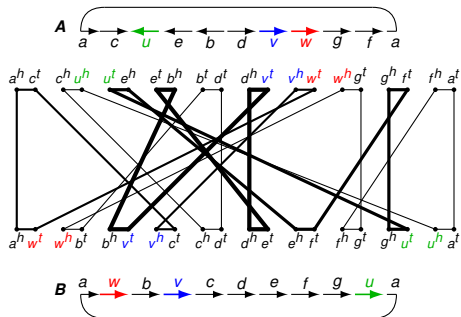
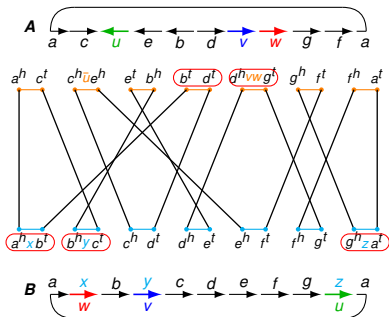
$$\Lambda = 4; \sigma = 2 \text{ (two subst.)}$$

$$\text{DCJ distance} = 7 - 1 = 6$$

$$\text{DCJ-substitution distance} = 8$$

Using the DCJ model to improve annotation

Finding missing homologies: a more complex example



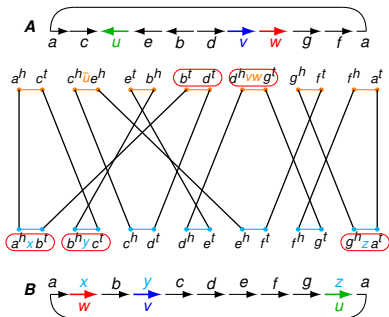
$$\Lambda = 4; \sigma = 2 \text{ (two subst.)}$$

$$\text{DCJ distance} = 7 - 1 = 6$$

$$\text{DCJ-substitution distance} = 8$$

Using the DCJ model to improve annotation

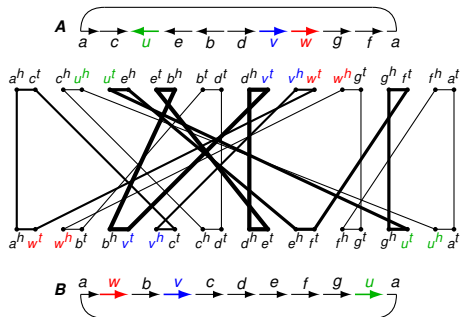
Finding missing homologies: a more complex example



$$\Lambda = 4; \sigma = 2 \text{ (two subst.)}$$

$$\text{DCJ distance} = 7 - 1 = 6$$

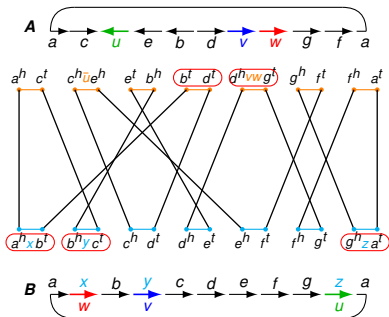
$$\text{DCJ-substitution distance} = 8$$



$$\sigma = 0 \text{ (no substitution)}$$

Using the DCJ model to improve annotation

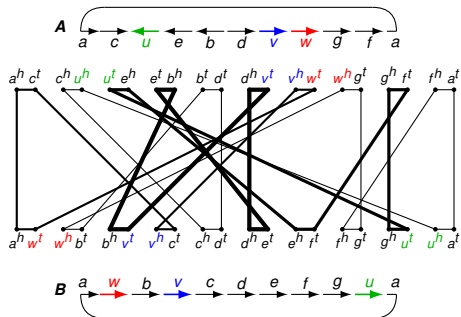
Finding missing homologies: a more complex example



$\Lambda = 4$; $\sigma = 2$ (two subst.)

DCJ distance = $7 - 1 = 6$

DCJ-substitution distance = 8

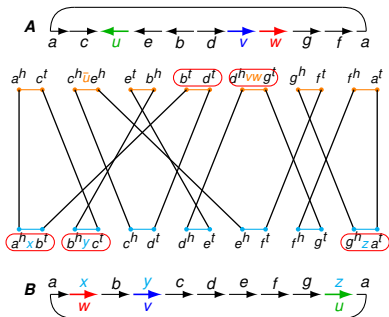


$\sigma = 0$ (no substitution)

DCJ distance = $10 - 4 = 6$

Using the DCJ model to improve annotation

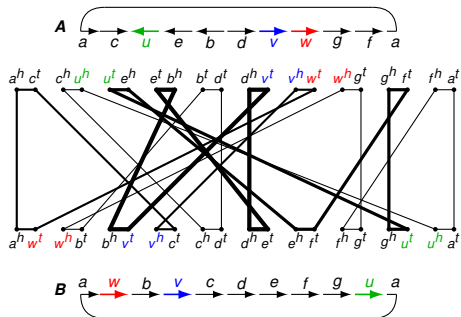
Finding missing homologies: a more complex example



$\Lambda = 4$; $\sigma = 2$ (two subst.)

DCJ distance = $7 - 1 = 6$

DCJ-substitution distance = 8



$\sigma = 0$ (no substitution)

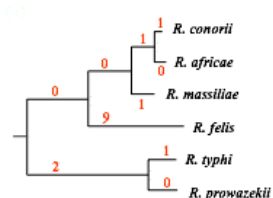
DCJ distance = $10 - 4 = 6$

DCJ-substitution distance = 6

Using the DCJ model to improve annotation

The *Rickettsia* database

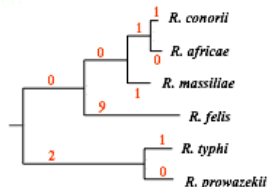
Phylogenetic tree with DCJ
distance in its branches
(Blanc *et al.*, 2007)



Using the DCJ model to improve annotation

The *Rickettsia* database

Phylogenetic tree with DCJ distance in its branches
(Blanc *et al.*, 2007)



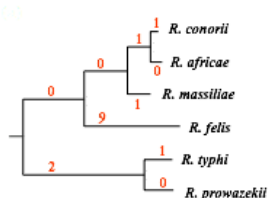
Comparison	D	SC	LC	$\lambda=1$	$\lambda \geq 2$
<i>R.pr.</i> x <i>R.ty.</i>	1	797	1	1	0
<i>R.co.</i> x <i>R.af.</i>	1	874	1	1	0
<i>R.co.</i> x <i>R.ma.</i>	3	867	2	9	0
<i>R.af.</i> x <i>R.ma.</i>	2	868	2	10	0
<i>R.pr.</i> x <i>R.co.</i>	4	789	1	38	1
<i>R.ty.</i> x <i>R.co.</i>	5	787	2	37	1
<i>R.pr.</i> x <i>R.af.</i>	3	788	1	39	1
<i>R.ty.</i> x <i>R.af.</i>	4	786	2	38	1
<i>R.pr.</i> x <i>R.ma.</i>	3	786	3	43	1
<i>R.ty.</i> x <i>R.ma.</i>	4	784	4	42	1
<i>R.pr.</i> x <i>R.fe.</i>	11	777	4	59	2
<i>R.ty.</i> x <i>R.fe.</i>	12	775	5	58	2
<i>R.co.</i> x <i>R.fe.</i>	11	844	3	38	2
<i>R.af.</i> x <i>R.fe.</i>	10	845	3	39	2
<i>R.ma.</i> x <i>R.fe.</i>	10	851	3	37	4

D = DCJ distance; SC = short cycle; LC = long cycle

Using the DCJ model to improve annotation

The *Rickettsia* database

Phylogenetic tree with DCJ distance in its branches
(Blanc *et al.*, 2007)



Comparison	D	SC	LC	$\lambda=1$	$\lambda \geq 2$
<i>R.pr.</i> x <i>R.ty.</i>	1	797	1	1	0
<i>R.co.</i> x <i>R.af.</i>	1	874	1	1	0
<i>R.co.</i> x <i>R.ma.</i>	3	867	2	9	0
<i>R.af.</i> x <i>R.ma.</i>	2	868	2	10	0
<i>R.pr.</i> x <i>R.co.</i>	4	789	1	38	1
<i>R.ty.</i> x <i>R.co.</i>	5	787	2	37	1
<i>R.pr.</i> x <i>R.af.</i>	3	788	1	39	1
<i>R.ty.</i> x <i>R.af.</i>	4	786	2	38	1
<i>R.pr.</i> x <i>R.ma.</i>	3	786	3	43	1
<i>R.ty.</i> x <i>R.ma.</i>	4	784	4	42	1
<i>R.pr.</i> x <i>R.fe.</i>	11	777	4	59	2
<i>R.ty.</i> x <i>R.fe.</i>	12	775	5	58	2
<i>R.co.</i> x <i>R.fe.</i>	11	844	3	38	2
<i>R.af.</i> x <i>R.fe.</i>	10	845	3	39	2
<i>R.ma.</i> x <i>R.fe.</i>	10	851	3	37	4

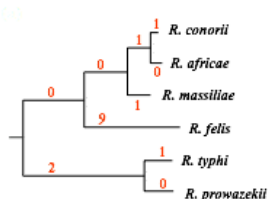
D = DCJ distance; SC = short cycle; LC = long cycle

With a quick look, we could find:

Using the DCJ model to improve annotation

The *Rickettsia* database

Phylogenetic tree with DCJ distance in its branches
(Blanc *et al.*, 2007)



Comparison	D	SC	LC	$\lambda=1$	$\lambda \geq 2$
<i>R.pr.</i> x <i>R.ty.</i>	1	797	1	1	0
<i>R.co.</i> x <i>R.af.</i>	1	874	1	1	0
<i>R.co.</i> x <i>R.ma.</i>	3	867	2	9	0
<i>R.af.</i> x <i>R.ma.</i>	2	868	2	10	0
<i>R.pr.</i> x <i>R.co.</i>	4	789	1	38	1
<i>R.ty.</i> x <i>R.co.</i>	5	787	2	37	1
<i>R.pr.</i> x <i>R.af.</i>	3	788	1	39	1
<i>R.ty.</i> x <i>R.af.</i>	4	786	2	38	1
<i>R.pr.</i> x <i>R.ma.</i>	3	786	3	43	1
<i>R.ty.</i> x <i>R.ma.</i>	4	784	4	42	1
<i>R.pr.</i> x <i>R.fe.</i>	11	777	4	59	2
<i>R.ty.</i> x <i>R.fe.</i>	12	775	5	58	2
<i>R.co.</i> x <i>R.fe.</i>	11	844	3	38	2
<i>R.af.</i> x <i>R.fe.</i>	10	845	3	39	2
<i>R.ma.</i> x <i>R.fe.</i>	10	851	3	37	4

D = DCJ distance; SC = short cycle; LC = long cycle

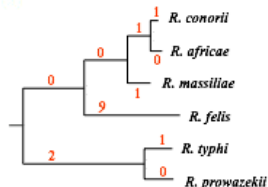
With a quick look, we could find:

- ▶ two pairs of genes that could be homologous between *R. felis* and the three species *R. conorii*, *R. africae* and *R. massiliae*.

Using the DCJ model to improve annotation

The *Rickettsia* database

Phylogenetic tree with DCJ distance in its branches
(Blanc *et al.*, 2007)



Comparison	D	SC	LC	$\lambda=1$	$\lambda \geq 2$
<i>R.pr.</i> x <i>R.ty.</i>	1	797	1	1	0
<i>R.co.</i> x <i>R.af.</i>	1	874	1	1	0
<i>R.co.</i> x <i>R.ma.</i>	3	867	2	9	0
<i>R.af.</i> x <i>R.ma.</i>	2	868	2	10	0
<i>R.pr.</i> x <i>R.co.</i>	4	789	1	38	1
<i>R.ty.</i> x <i>R.co.</i>	5	787	2	37	1
<i>R.pr.</i> x <i>R.af.</i>	3	788	1	39	1
<i>R.ty.</i> x <i>R.af.</i>	4	786	2	38	1
<i>R.pr.</i> x <i>R.ma.</i>	3	786	3	43	1
<i>R.ty.</i> x <i>R.ma.</i>	4	784	4	42	1
<i>R.pr.</i> x <i>R.fe.</i>	11	777	4	59	2
<i>R.ty.</i> x <i>R.fe.</i>	12	775	5	58	2
<i>R.co.</i> x <i>R.fe.</i>	11	844	3	38	2
<i>R.af.</i> x <i>R.fe.</i>	10	845	3	39	2
<i>R.ma.</i> x <i>R.fe.</i>	10	851	3	37	4

D = DCJ distance; SC = short cycle; LC = long cycle

With a quick look, we could find:

- ▶ two pairs of genes that could be homologous between *R. felis* and the three species *R. conorii*, *R. africae* and *R. massiliae*.
- ▶ two pairs of genes that could be homologous between *R. prowazekii* and *R. typhi* and the four species *R. felis*, *R. conorii*, *R. africae* and *R. massiliae*.

Using the DCJ model to improve annotation

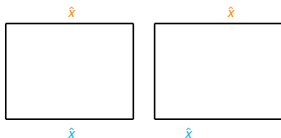
Resolving duplications

- ▶ The master graph is only defined for genomes without duplicated genes.
- ▶ However, duplicates could be represented as labels in the components of the graph.
- ▶ The information of the components could help to disambiguate the duplications.

Using the DCJ model to improve annotation

Resolving duplications - pairs from the same or from distinct components

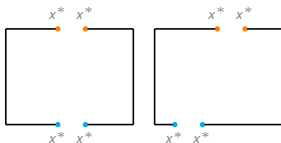
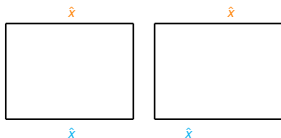
Two cycles:



Using the DCJ model to improve annotation

Resolving duplications - pairs from the same or from distinct components

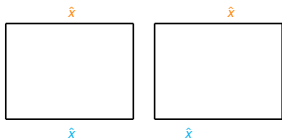
Two cycles:



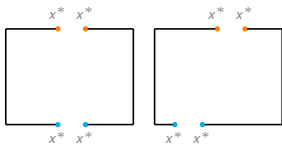
Using the DCJ model to improve annotation

Resolving duplications - pairs from the same or from distinct components

Two cycles:



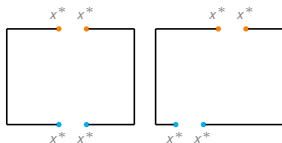
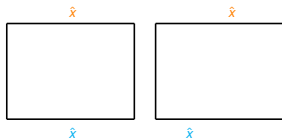
Pairs from distinct cycles



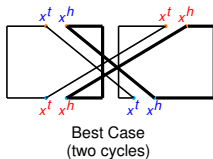
Using the DCJ model to improve annotation

Resolving duplications - pairs from the same or from distinct components

Two cycles:



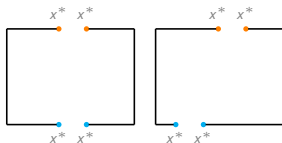
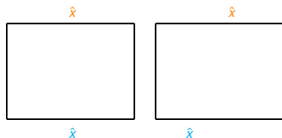
Pairs from distinct cycles



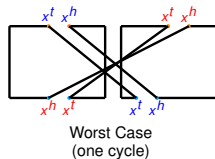
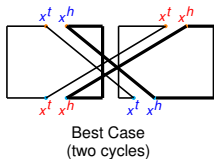
Using the DCJ model to improve annotation

Resolving duplications - pairs from the same or from distinct components

Two cycles:



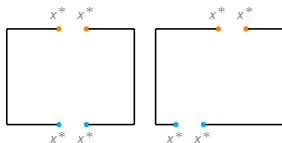
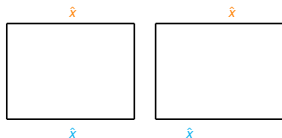
Pairs from distinct cycles



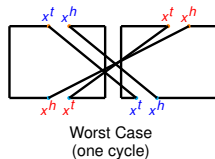
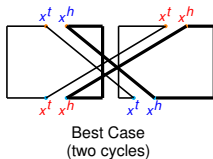
Using the DCJ model to improve annotation

Resolving duplications - pairs from the same or from distinct components

Two cycles:



Pairs from distinct cycles

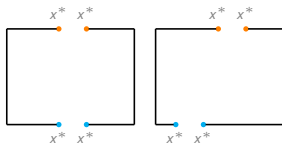
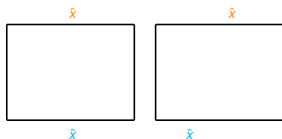


Pairs from the same cycle

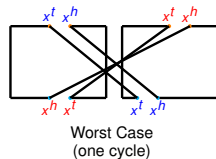
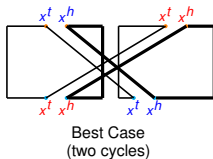
Using the DCJ model to improve annotation

Resolving duplications - pairs from the same or from distinct components

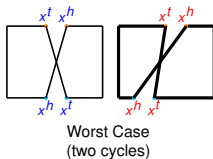
Two cycles:



Pairs from distinct cycles



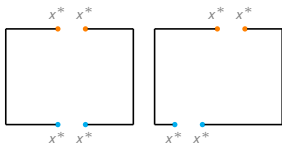
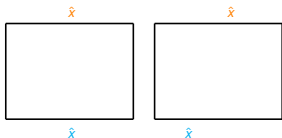
Pairs from the same cycle



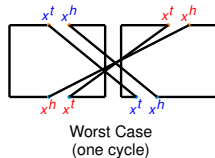
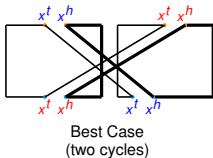
Using the DCJ model to improve annotation

Resolving duplications - pairs from the same or from distinct components

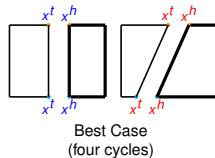
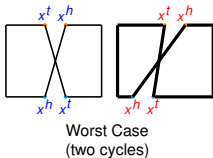
Two cycles:



Pairs from distinct cycles



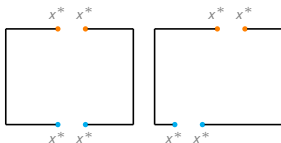
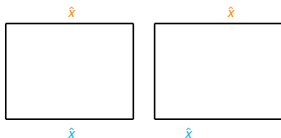
Pairs from the same cycle



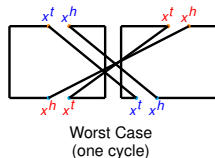
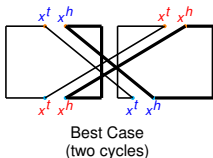
Using the DCJ model to improve annotation

Resolving duplications - pairs from the same or from distinct components

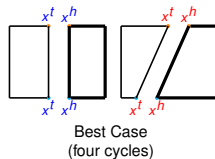
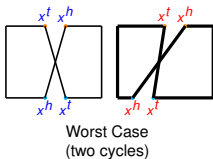
Two cycles:



Pairs from distinct cycles



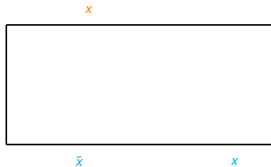
Pairs from the same cycle



Assigning pairs in the same cycle is better or at least as good as assigning pairs in distinct cycles.

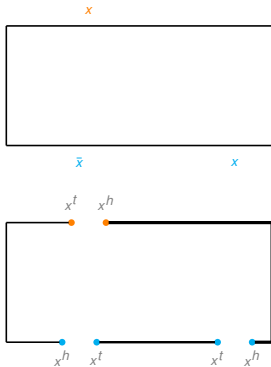
Using the DCJ model to improve annotation

Resolving duplications - more labels in the same component



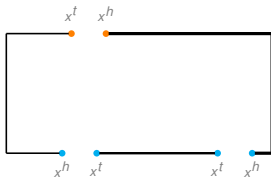
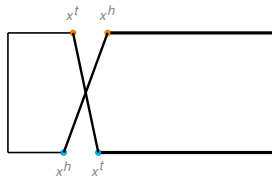
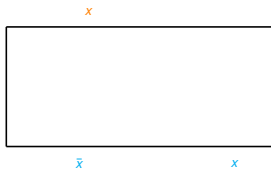
Using the DCJ model to improve annotation

Resolving duplications - more labels in the same component



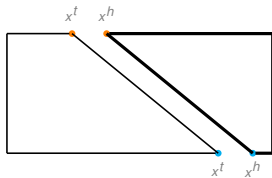
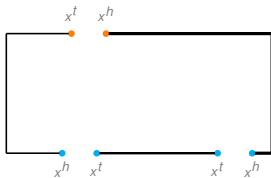
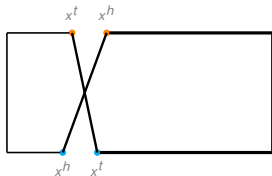
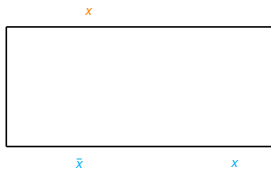
Using the DCJ model to improve annotation

Resolving duplications - more labels in the same component



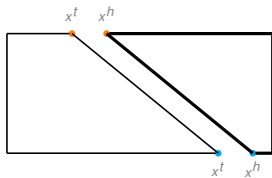
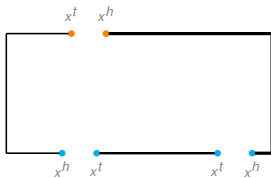
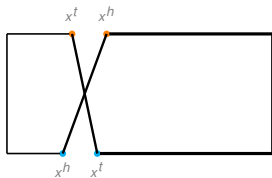
Using the DCJ model to improve annotation

Resolving duplications - more labels in the same component



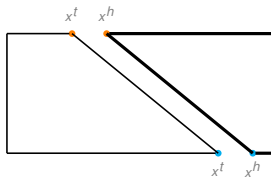
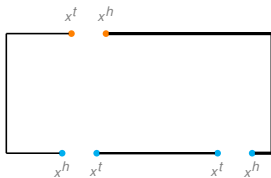
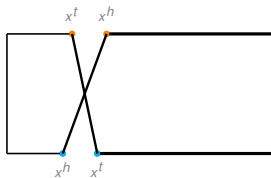
Using the DCJ model to improve annotation

Resolving duplications - more labels in the same component



Using the DCJ model to improve annotation

Resolving duplications - more labels in the same component



It may be possible to find the optimal pair(s).

Summary

Overview

- 1 Motivation
- 2 DCJ model
 - Master graph and its components
 - DCJ distance
 - Handling indels
- 3 Using the DCJ model to improve annotation
 - (Ongoing work)
 - Substitution or missing homology?
 - The Rickettsia database
 - Resolving duplications
- 4 Summary

Summary

- ▶ In genome rearrangements, the analysis usually has three main steps:
 1. Find genes in the given genomes
 2. Annotate genes
 3. Compute distance according to some rearrangement model

Summary

- ▶ In genome rearrangements, the analysis usually has three main steps:
 1. Find genes in the given genomes
 2. Annotate genes
 3. Compute distance according to some rearrangement model
- ▶ In the development of approaches to solve step (3), it is often assumed that steps (1) and (2) are given.

Summary

- ▶ In genome rearrangements, the analysis usually has three main steps:
 1. Find genes in the given genomes
 2. Annotate genes
 3. Compute distance according to some rearrangement model
- ▶ In the development of approaches to solve step (3), it is often assumed that steps (1) and (2) are given.
- ▶ Here we have shown that the graph structure used in step (3) for the DCJ model, that actually requires some annotation of the genomes, can be used to improve the annotation itself.

Summary

- ▶ In genome rearrangements, the analysis usually has three main steps:
 1. Find genes in the given genomes
 2. Annotate genes
 3. Compute distance according to some rearrangement model
- ▶ In the development of approaches to solve step (3), it is often assumed that steps (1) and (2) are given.
- ▶ Here we have shown that the graph structure used in step (3) for the DCJ model, that actually requires some annotation of the genomes, can be used to improve the annotation itself.
- ▶ However, finding candidates for homology in a component of the graph can be difficult, if the component is long and with many labels.

Summary

- ▶ In genome rearrangements, the analysis usually has three main steps:
 1. Find genes in the given genomes
 2. Annotate genes
 3. Compute distance according to some rearrangement model
- ▶ In the development of approaches to solve step (3), it is often assumed that steps (1) and (2) are given.
- ▶ Here we have shown that the graph structure used in step (3) for the DCJ model, that actually requires some annotation of the genomes, can be used to improve the annotation itself.
- ▶ However, finding candidates for homology in a component of the graph can be difficult, if the component is long and with many labels.
- ▶ Fortunately, for some datasets (in particular closely related genomes such as *Rickettsia*), the components are usually short and have few labels.

Summary

- ▶ In genome rearrangements, the analysis usually has three main steps:
 1. Find genes in the given genomes
 2. Annotate genes
 3. Compute distance according to some rearrangement model
- ▶ In the development of approaches to solve step (3), it is often assumed that steps (1) and (2) are given.
- ▶ Here we have shown that the graph structure used in step (3) for the DCJ model, that actually requires some annotation of the genomes, can be used to improve the annotation itself.
- ▶ However, finding candidates for homology in a component of the graph can be difficult, if the component is long and with many labels.
- ▶ Fortunately, for some datasets (in particular closely related genomes such as *Rickettsia*), the components are usually short and have few labels.
- ▶ There is a potential in the use of this graph to disambiguate duplicate genes.

Acknowledgements

This research is supported by the Brazilian research agency CNPq
(grant PROMETRO 563087/2010-2)

Acknowledgements

This research is supported by the Brazilian research agency CNPq
(grant PROMETRO 563087/2010-2)

Thank you for your attention!