Algorithms for Genome Rearrangement by Double Cut and Join

Jens Stoye

Bielefeld University, Germany

Outline

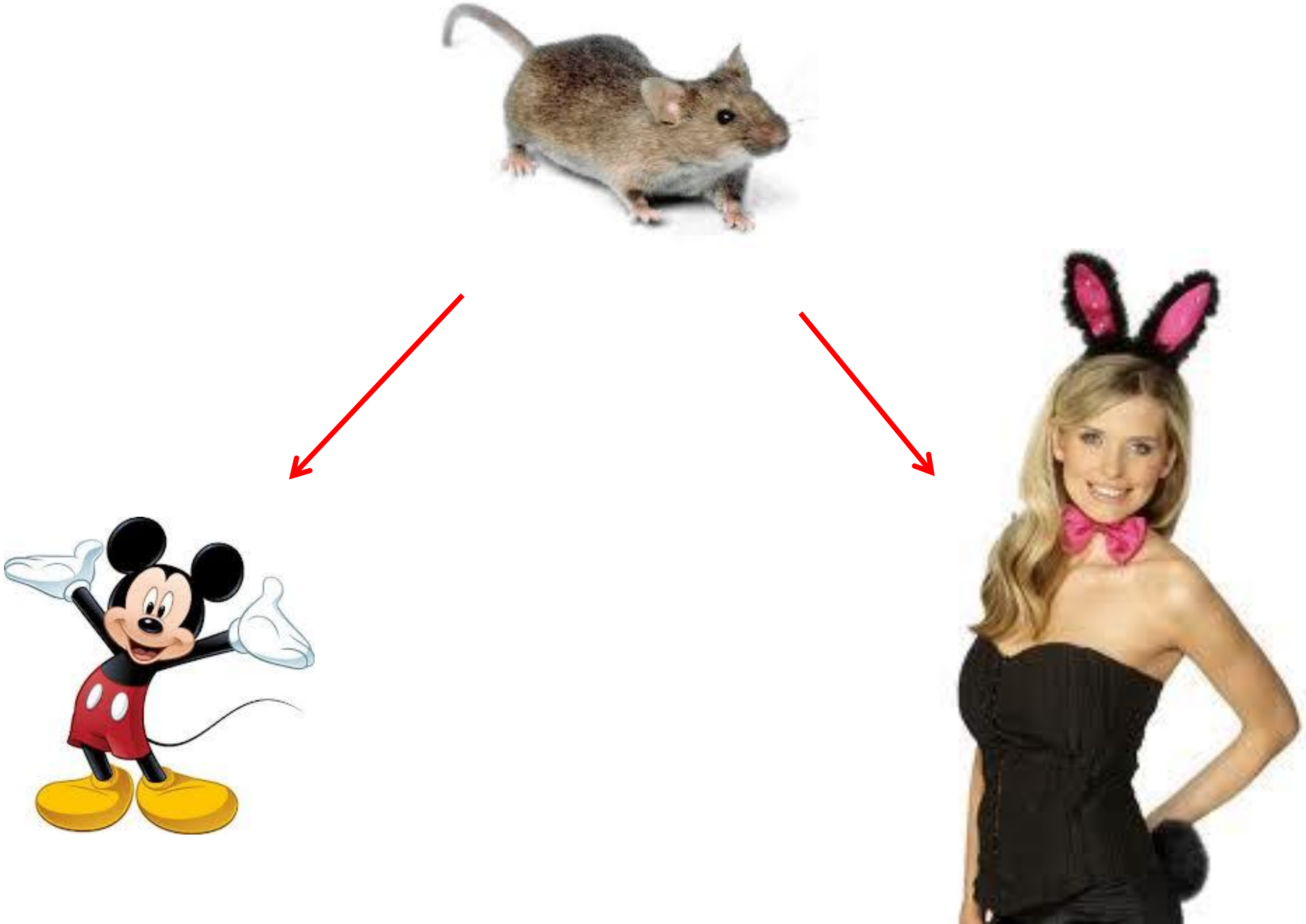
1. Genome evolution

2. Double Cut and Join (DCJ)
3. DCJ distance and sorting
4. Relation to other models
5. Insertions, deletions, substitutions
6. On the weight of indels

7. Summary and Conclusion

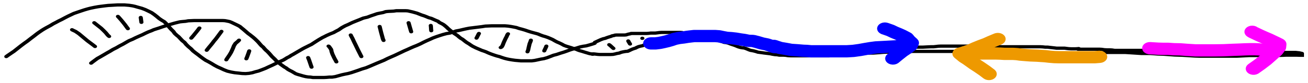
1. Genome evolution

Species change over time.



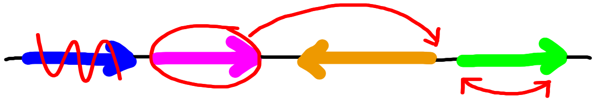
1. Genome evolution

At the molecular level:



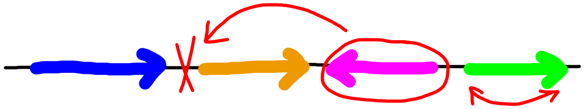
Local vs. global modifications:

- point mutations (sequence analysis)
- large-scale operations (comparative genomics)

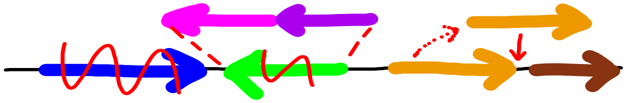


Organizational vs. content-modifying operations:

- rearrangement



- insertion, deletion, substitution, duplication



Motivation

Evolution at the whole genome level:

- Basic understanding of molecular processes at genomic scale
- Evolutionary distances, phylogenetic trees (phylogenomics)
- Ancestral genome reconstruction

- Insights into gene function
- Regulation of genes (e.g. operons in prokaryotic genomes)

- Comparative genome assembly and annotation

- Structural variations, cancer development
- Pathogen evolution, outbreak prediction, vaccination strategies

What happens in detail?

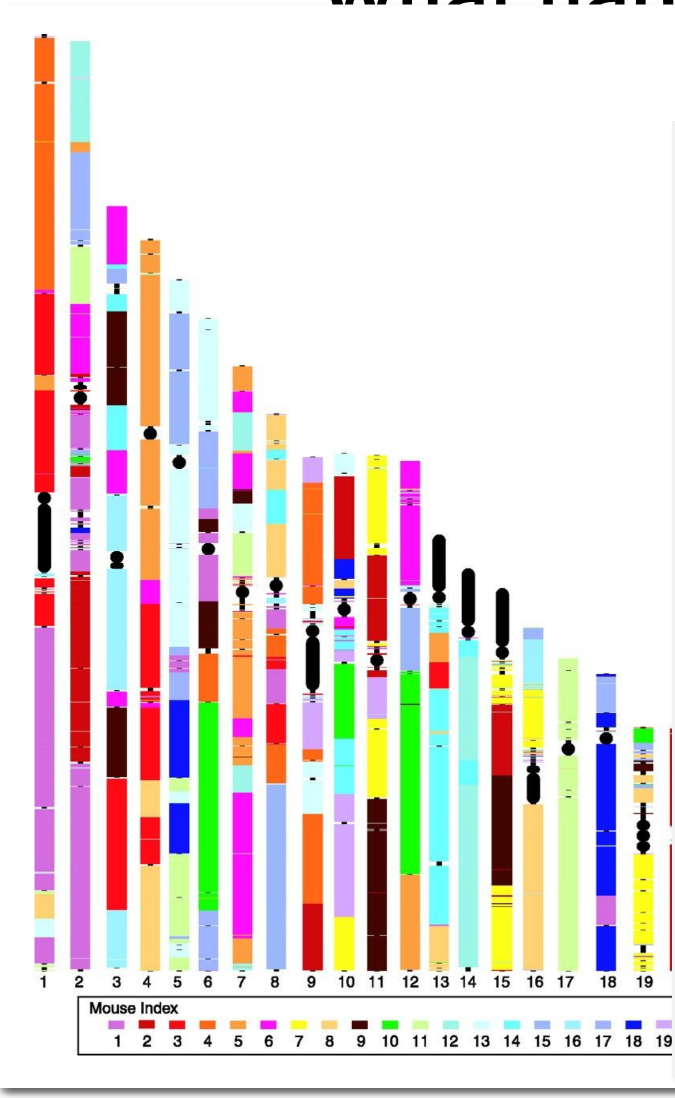
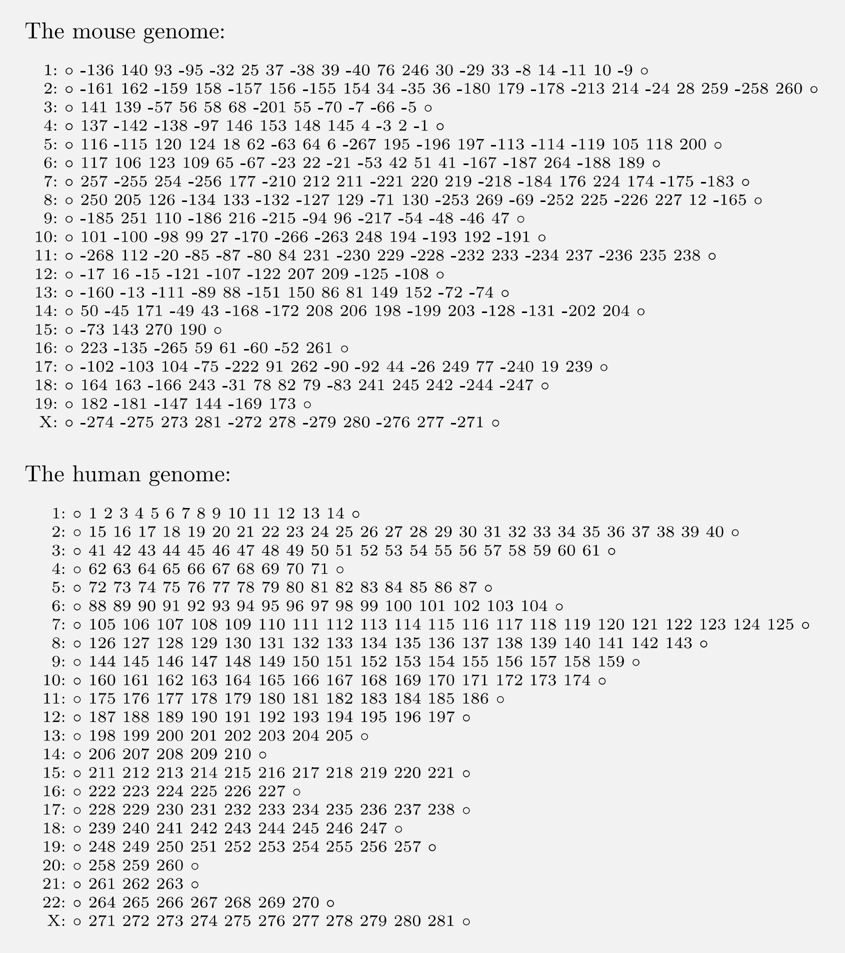


Figure: Eichler & Sankoff 2003



The mouse genome:

```
 1: ○ -136 140 93 -95 -32 25 37 -38 39 -40 76 246 30 -29 33 -8 14 -11 10 -9 ○
 2: ○ -161 162 -159 158 -157 156 -155 154 34 -35 36 -180 179 -178 -213 214 -24 28 259 -258 260 ○
 3: ○ 141 139 -57 56 58 68 -201 55 -70 -7 -66 -5 ○
 4: ○ 137 -142 -138 -97 146 153 148 145 4 -3 2 -1 ○
 5: ○ 116 -115 120 124 18 62 -63 64 6 -267 195 -196 197 -113 -114 -119 105 118 200 ○
 6: ○ 117 106 123 109 65 -67 -23 22 -21 -53 42 51 41 -167 -187 264 -188 189 ○
 7: ○ 257 -255 254 -256 177 -210 212 211 -221 220 219 -218 -184 176 224 174 -175 -183 ○
 8: ○ 250 205 126 -134 133 -132 -127 129 -71 130 -253 269 -69 -252 225 -226 227 12 -165 ○
 9: ○ -185 251 110 -186 216 -215 -94 96 -217 -54 -48 -46 47 ○
10: ○ 101 -100 -98 99 27 -170 -266 -263 248 194 -193 192 -191 ○
11: ○ -268 112 -20 -85 -87 -80 84 231 -230 229 -228 -232 233 -234 237 -236 235 238 ○
12: ○ -17 16 -15 -121 -107 -122 207 209 -125 -108 ○
13: ○ -160 -13 -111 -89 88 -151 150 86 81 149 152 -72 -74 ○
14: ○ 50 -45 171 -49 43 -168 -172 208 206 198 -199 203 -128 -131 -202 204 ○
15: ○ -73 143 270 190 ○
16: ○ 223 -135 -265 59 61 -60 -52 261 ○
17: ○ -102 -103 104 -75 -222 91 262 -90 -92 44 -26 249 77 -240 19 239 ○
18: ○ 164 163 -166 243 -31 78 82 79 -83 241 245 242 -244 -247 ○
19: ○ 182 -181 -147 144 -169 173 ○
 X: ○ -274 -275 273 281 -272 278 -279 280 -276 277 -271 ○
```

The human genome:

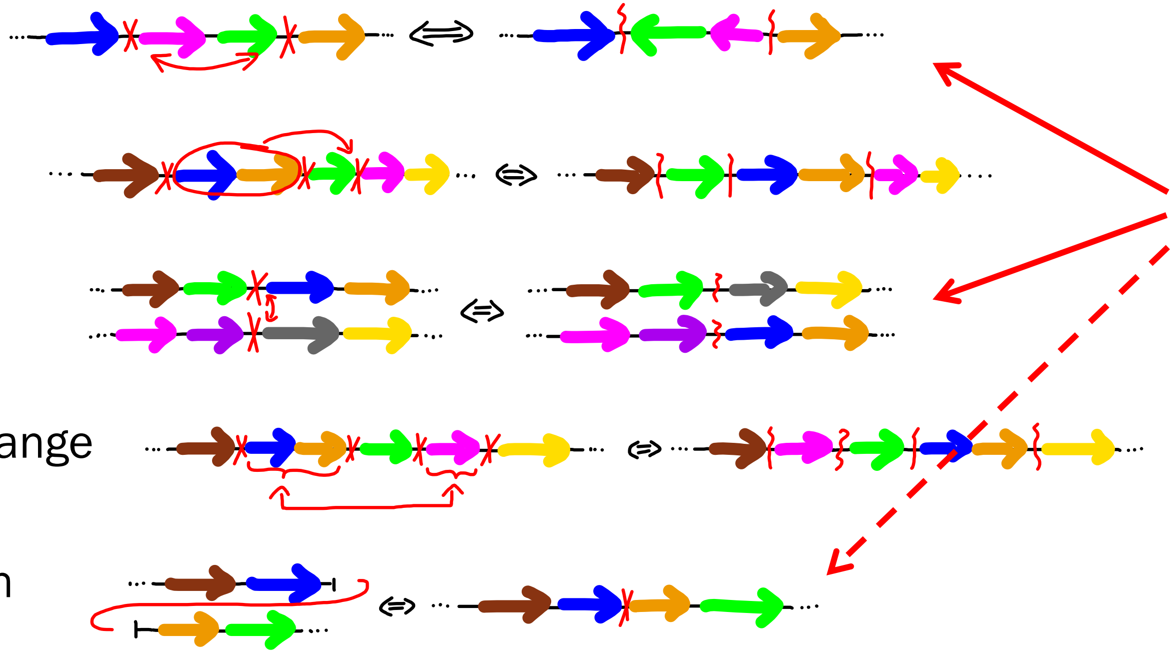
```
 1: ○ 1 2 3 4 5 6 7 8 9 10 11 12 13 14 ○
 2: ○ 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 ○
 3: ○ 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 ○
 4: ○ 62 63 64 65 66 67 68 69 70 71 ○
 5: ○ 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 ○
 6: ○ 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 ○
 7: ○ 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 ○
 8: ○ 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 ○
 9: ○ 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 ○
10: ○ 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 ○
11: ○ 175 176 177 178 179 180 181 182 183 184 185 186 ○
12: ○ 187 188 189 190 191 192 193 194 195 196 197 ○
13: ○ 198 199 200 201 202 203 204 205 ○
14: ○ 206 207 208 209 210 ○
15: ○ 211 212 213 214 215 216 217 218 219 220 221 ○
16: ○ 222 223 224 225 226 227 ○
17: ○ 228 229 230 231 232 233 234 235 236 237 238 ○
18: ○ 239 240 241 242 243 244 245 246 247 ○
19: ○ 248 249 250 251 252 253 254 255 256 257 ○
20: ○ 258 259 260 ○
21: ○ 261 262 263 ○
22: ○ 264 265 266 267 268 269 270 ○
 X: ○ 271 272 273 274 275 276 277 278 279 280 281 ○
```

Data from: Pevzner & Tesler 2003

What happens in detail?

Basic rearrangement operations:

- inversion
- transposition
- translocation
- block interchange
- fusion/fission



Note: 2-cut (double-cut)

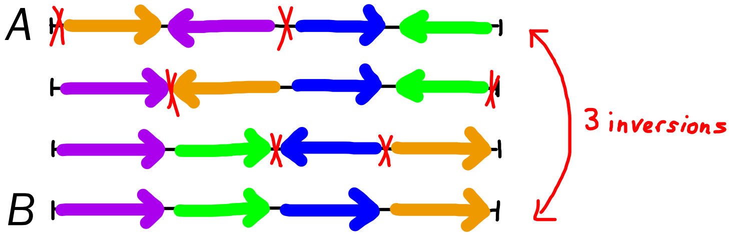
Assumption:

The number of rearrangements needed to transform one genome into another is a measure for the evolutionary distance between two species.

Questions to be asked:

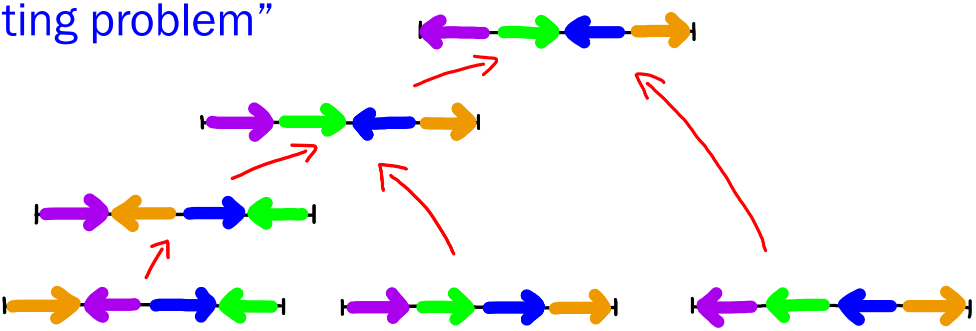
How many rearrangement operations are needed?

- distance *d*(*A*,*B*) ➔ "distance problem"
- diameter problems
- distribution of distances
- halving distance



How much can we reconstruct of the past?

- Ancestral genome(s)
- rearrangement scenario(s) ➔ "sorting problem"
- complete phylogenies



Some history (2 genomes)

Inversions (reversals):
Watterson *et al.* 1982; Sankoff 1992; Bafna & Pevzner 1993; Hannenhalli & Pevzner 1995; Kaplan, Shamir & Tarjan 1999; Bader, Moret & Yan 2001; Bergeron 2001; Bergeron, Heber & S 2002; Bergeron, Mixtacki & S 2004

Translocations:
Hannenhalli 1996; Bergeron, Mixtacki & S 2005

Multichromosomal linear ("general HP model"):
Hannenhalli & Pevzner 1995; Tesler 2002; Ozery-Flato & Shamir 2003; Jean & Nikolski 2007; Bergeron, Mixtacki & S 2008; Erdős, Sokoup & S 2011

Double Cut and Join (DCJ):
Yancopoulos, Attie & Friedberg 2005; Bergeron, Mixtacki & S 2006; Kováč, Warren, Braga & S 2011

Other models:
Unsigned inversions: Kececioglu & Sankoff 1993; Christie 1998; Caprara 1999
Transpositions: Meidanis, Walter & Dias, 1997; Elias & Hartman 2006; Bulteau, Fertin, Rusu 2011
Inversions + Transpositions: Walter, Dias & Meidanis 1998; Christie & Irving 2001
Fusion/Fission + Transpositions: Meidanis & Dias 2001
Block interchanges: Christie 1996
Block interchanges + inversions: Mira & Meidanis 2007
Single Cut and Join: Bergeron, Medvedev & S 2010
Single Cut or Join: Feijão & Meidanis 2011

Some history (2 genomes)

All models so far: Strong assumption that all genomes contain exactly the same set of blocks

Inversions + Insertions and Deletions:
El-Mabrouk 2001; Marron, Swenson & Moret 2004

Insertions + Duplications:
Marron, Swenson & Moret 2004

DCJ + Insertions and Deletions:
Yancopoulos & Friedberg 2009; Braga, Willing & S 2010; Braga 2010; Braga, Machado, Ribeiro & S 2011b; Compeau 2012; da Silva, Braga, Machado & Dantas 2012; da Silva, Machado, Dantas & Braga 2012

DCJ + Insertions and Deletions + Duplications:
Yancopoulos & Friedberg 2009

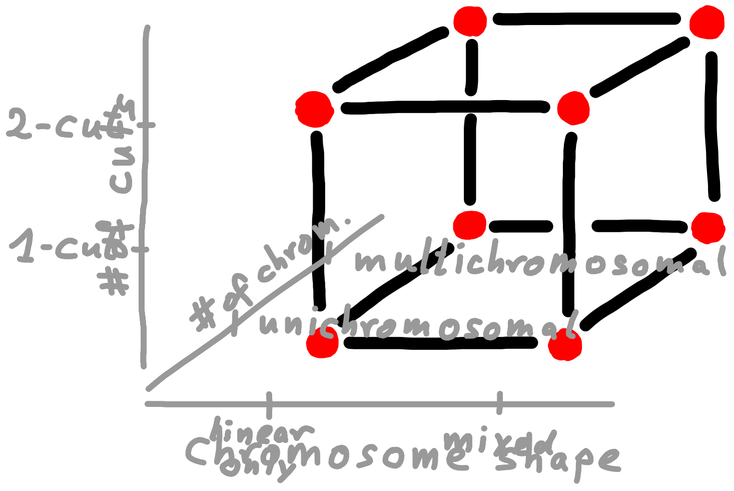
DCJ + Substitutions:
Braga, Machado, Ribeiro & S 2011a

Towards formal modeling

Definitions:
Genome: set of chromosomes
Chromosome: sequence of oriented unique blocks (genes or other markers)

Independent dimensions:

- **Chromosome shapes**
  - linear-only, (circular-only), mixed
- **Number of chromosomes**
  - unichromosomal, multichromosomal
- **Rearrangement operations**
  - single-cut, double-cut, (multi-cut)

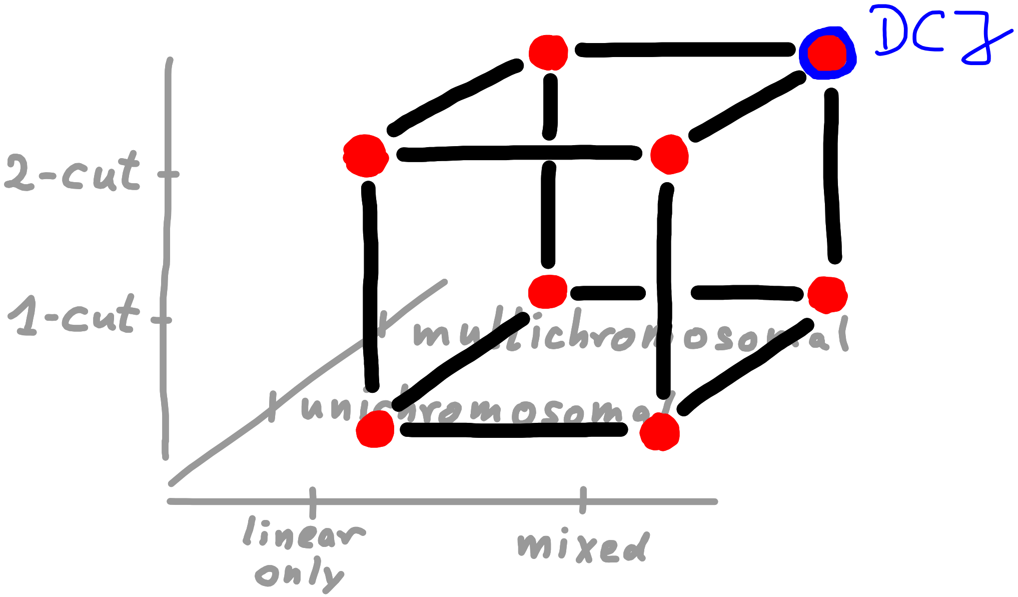


2. Double Cut and Join (DCJ)

(*based on*: Bergeron, Mixtacki & S: *Proc. of WABI* 2006)

The model we will concentrate on:

- mixed linear and circular chromosomes
- multichromosomal genome
- 2-cut operations

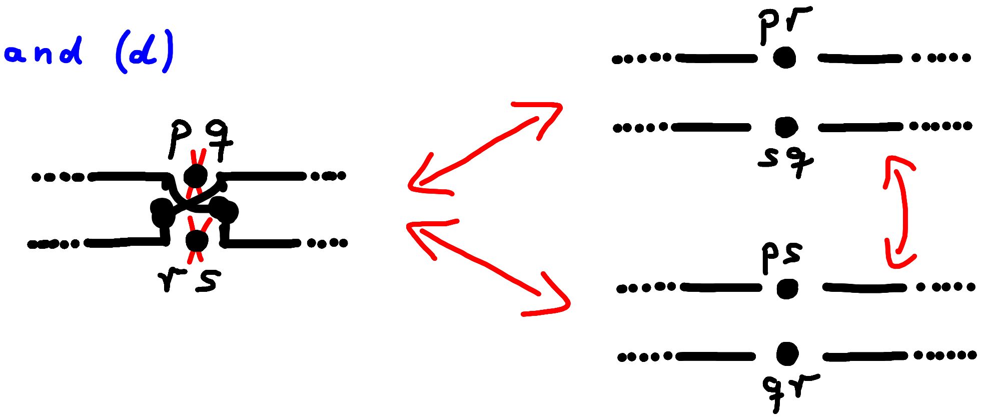


Graphs with vertices of degree one or two

Definition:
The DCJ operation acts on two vertices *u* and *v* of a graph with vertices of degree one or two in one of the following ways:

(a) If both *u* = {*p*,*q*} and *v* = {*r*,*s*} are internal vertices, these are replaced by the two vertices {*p*,*r*} and {*s*,*q*} or by the two vertices {*p*,*s*} and {*q*,*r*}.
(b) If *u* = {*p*,*q*} is internal and *v* = {*r*} is external, these are replaced by {*p*,*r*} and {*q*} or by {*q*,*r*} and {*p*}.
(c) If both *u* = {*q*} and *v* = {*r*} are external, these are replaced by {*q*,*r*}.
(d) A single internal vertex {*q*,*r*} can be replaced by two external vertices {*q*} and {*r*}.

(a) and (d)



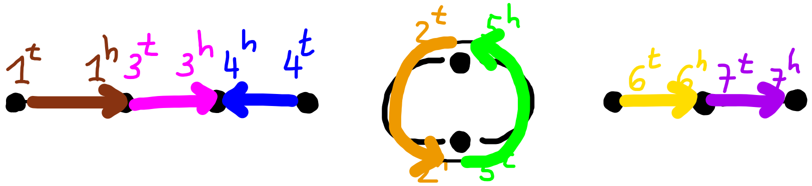
The formal problem

Definitions:

- A block (marker, gene) a is an oriented sequence of DNA that starts with a tail a^t and ends with a head a^h.
- Head and tail are called the extremities of a block.
- An adjacency of two consecutive blocks a and b, depending on their respective orientation, can be of four different types:

$$\{a^h,b^t\}, \{a^h,b^h\}, \{a^t,b^t\}, \{a^t,b^h\}$$

- An extremity that is not adjacent to any other block is called a telomere, represented by a singleton set $\{a^h\}$ or $\{a^t\}$.

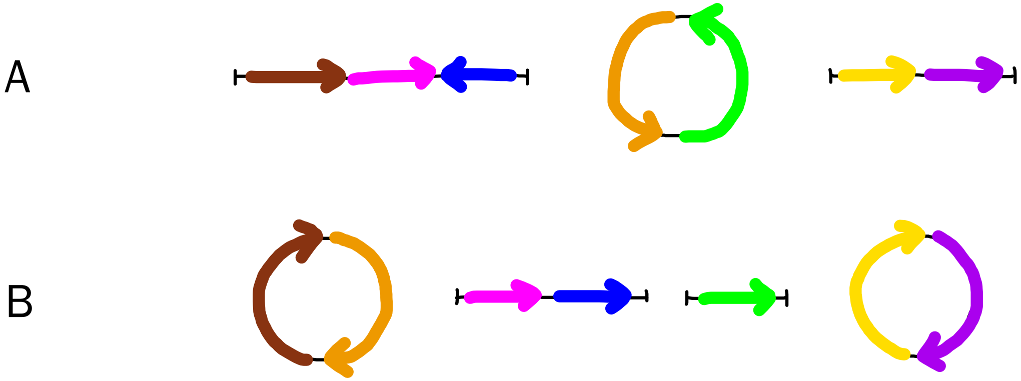


Genome: Set of adjacencies and telomeres such that the tail or head of a block appears in exactly one adjacency or telomere.

$$A = \{ \{1^t\}, \{1^h,3^t\}, \{3^h,4^h\}, \{4^t\}, \{2^h,5^t\}, \{5^h,2^t\}, \{6^t\}, \{6^h,7^t\}, \{7^h\} \}$$

The formal problem

Two genomes:



DCJ Sorting Problem:
Given two genomes *A* and *B* with the same set of blocks, find a shortest sequence of DCJ operations that transforms *A* into *B*. The length of such a sequence is called the DCJ distance between *A* and *B*, denoted by $d^{DCJ}(A,B)$.

3. DCJ distance and sorting

(*based on*: Bergeron, Mixtacki & S: *Proc. of WABI* 2006; Braga & S: *JCB* 2010)

History of formal studies:

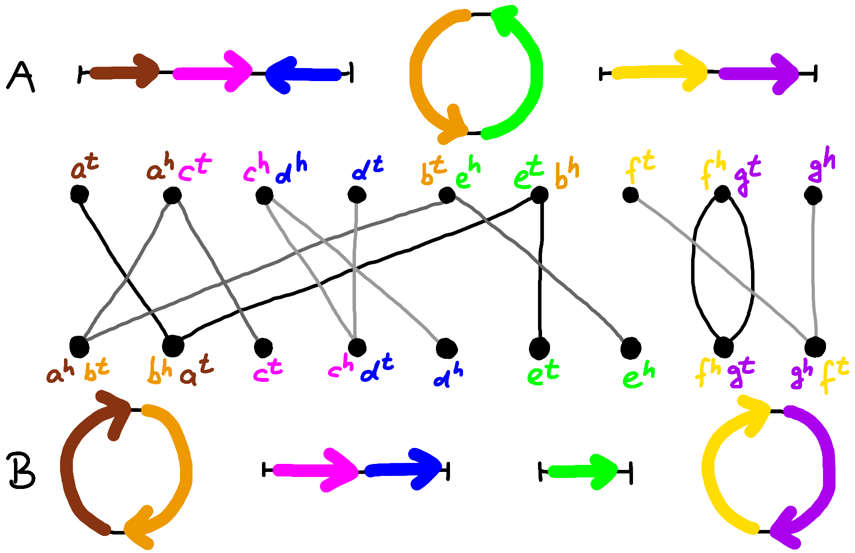
1992 – inversions (INV)
1995 – Hannenhalli-Pevzner (HP) model
1995 – translocations

2005 – DCJ

→ surprisingly simple (in particular compared to the earlier results)

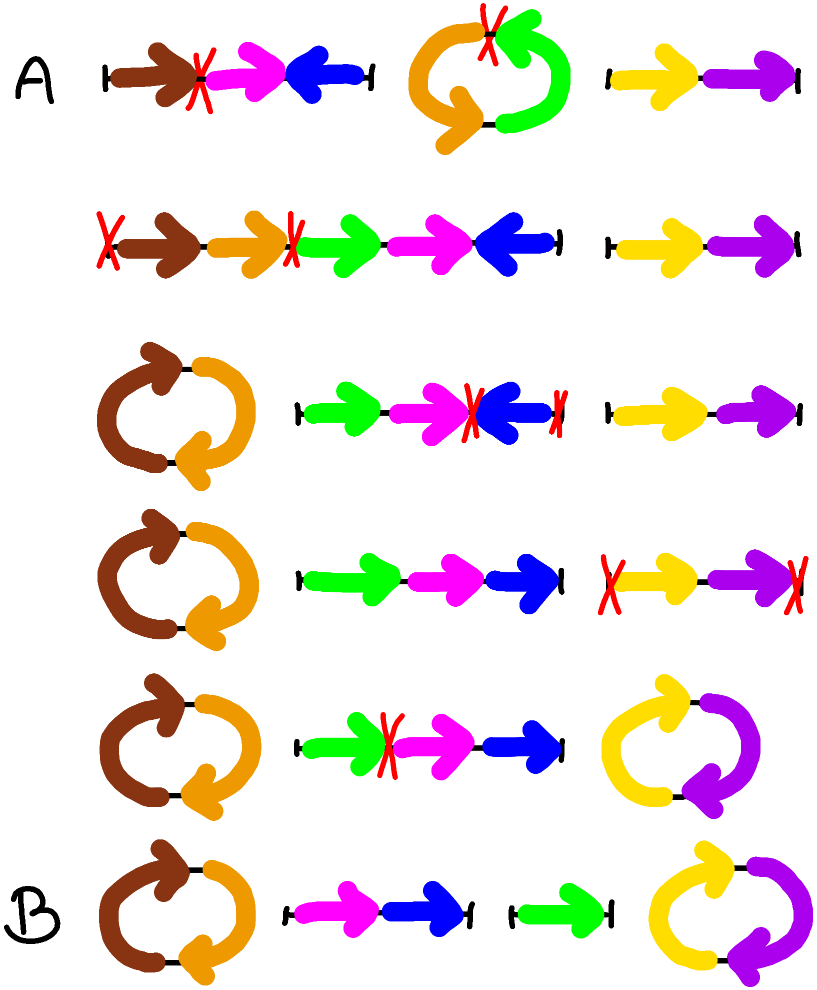
Adjacency graph

Definition:
The adjacency graph *AG*(*A*,*B*) is a graph whose set of vertices are the adjacencies and telomeres of *A* and *B*. For each $u \in A$ and $v \in B$ there are $|u \cap v|$ edges between *u* and *v*.

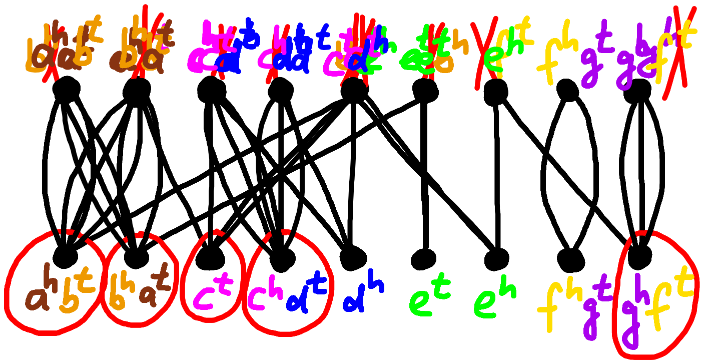


Related to breakpoint graph (Bafna & Pevzner 1993)

Transforming *A* into *B*



Adjacency graph



Algorithm

```
Let $AG(A,B)$ be the adjacency graph of genomes $A$ and $B$

// Generate the adjacencies of $B$ that are not yet present in $A$
for each adjacency $\{p,q\}$ in $B$ do
    let $u$ be the vertex of $A$ that contains $p$
    let $v$ be the vertex of $A$ that contains $q$
    if $u \neq v$ then
        replace vertices $u$ and $v$ in $A$ by $\{p,q\}$ and $(u \setminus \{p\}) \cup (v \setminus \{q\})$
    end if
end for

//Generate the telomeres of $B$ that are not yet present in $A$
for each telomere $\{p\}$ in $B$ do
    let $u$ be the vertex of $A$ that contains $p$
    if $u$ is an adjacency then
        replace vertex $u$ in $A$ by $\{p\}$ and $(u \setminus \{p\})$
    end if
end for
```

Analysis: $O(N)$ time where N = # of blocks

The DCJ distance

Theorem:

Let *A* and *B* be two genomes defined on the same set of *N* blocks, then we have

$$d^{DCJ}(A,B) \;=\; N - (C + I/2)$$

where C = # of cycles and I = # of odd paths in *AG*(*A*,*B*). A sorting sequence can be found in optimal *O*(*N*) time.

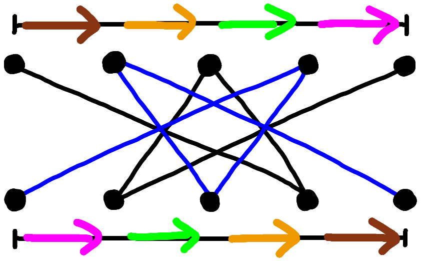
Example (Human-Mouse):

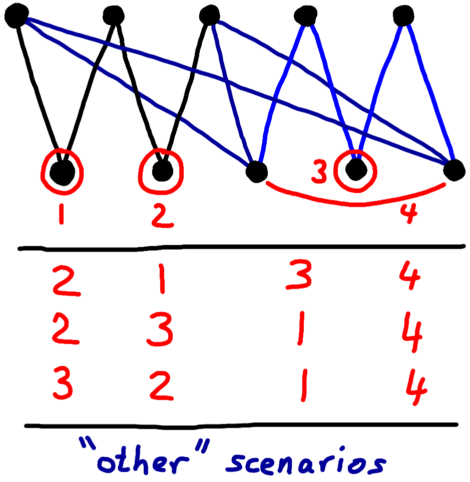
$N = 281$, $C = 27$, $I = 16$ ➔ $d^{DCJ}(\text{Human,Mouse}) = 246$

Note 1: Same as HP distance (no circular chromosomes necessary)
Note 2: Sorting scenarios can be of different types (1-cut vs. 2-cut operations)
Note 3: This can lead to different breakpoint reuse rates $0.89 \leq r \leq 1.51$

The solution space of sorting by DCJ

There are really many rearrangement scenarios for a given pair of genomes:





2	1	3	4
2	3	1	4
3	2	1	4

"other" scenarios

Simplified case (k components with distances $\ell_1,\ldots,\ell_k$):

$$S_{sep} = \frac{(\ell_1+\ell_2+\ldots+\ell_k)!}{\ell_1!\ell_2!\ldots\ell_k!} \times \prod_{i=1}^{k}(\ell_i+1)^{\ell_i-1}$$

General case: more complicated due to recombinations

1 component (distance ℓ)	number of scenarios
1	1
2	3
3	16
4	125
5	1296
6	16807

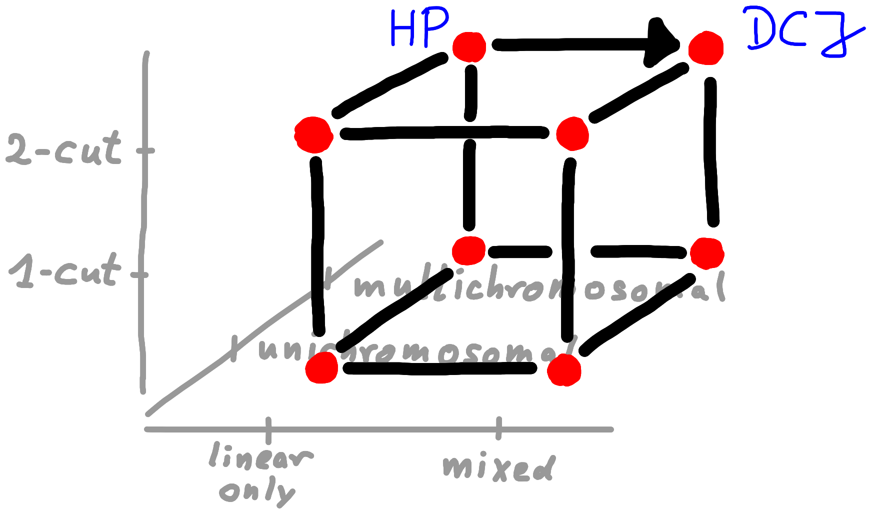
4. Relation to other models

(*based on*: Bergeron, Medvedev & S: *JCB* 2010; Bergeron, Mixtacki & S: *TCS* 2009)

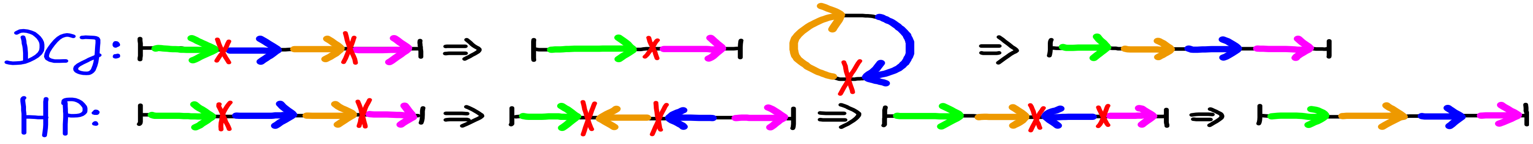
Hannenhalli-Pevzner (HP) model: 2-cut, linear-only, multichromosomal

Observation:
For two linear genomes *A* and *B*, we have that

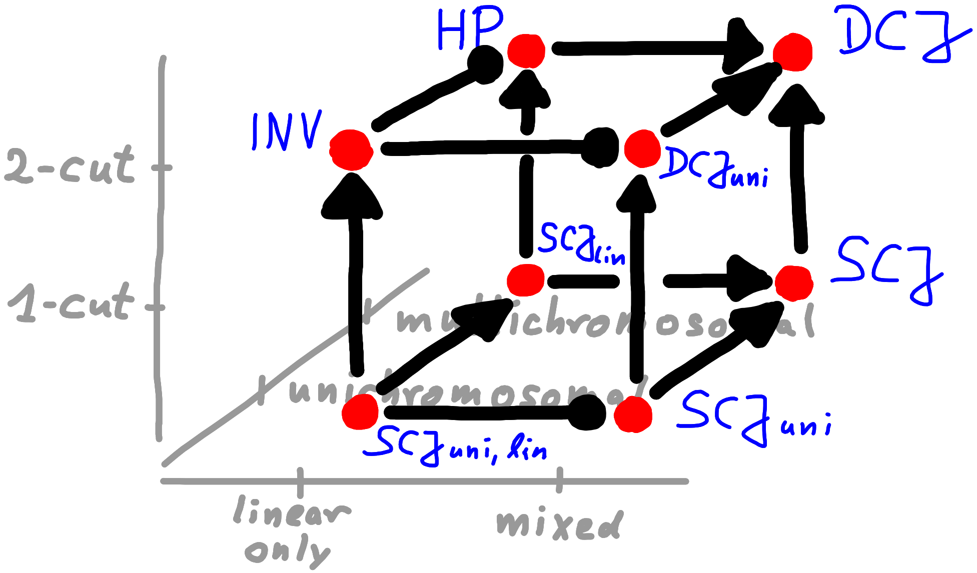
$$d^{DCJ}(A,B) \leq d^{HP}(A,B)$$



In fact, for A = (1,3,2,4) and B = (1,2,3,4) we have $d^{DCJ}(A,B) = 2 < 3 = d^{HP}(A,B)$.

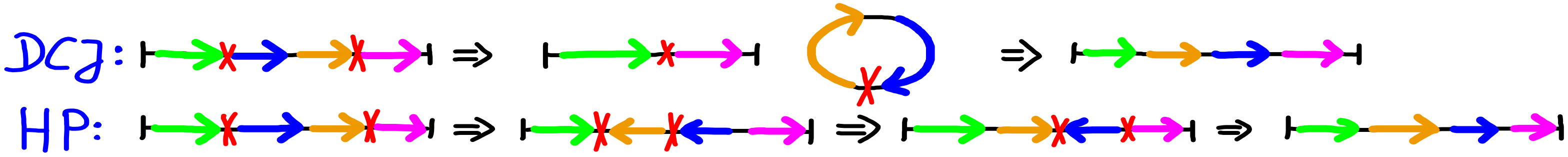


Relationship of distances



Unexpected asymmetry: INV —● HP

General HP distance problem

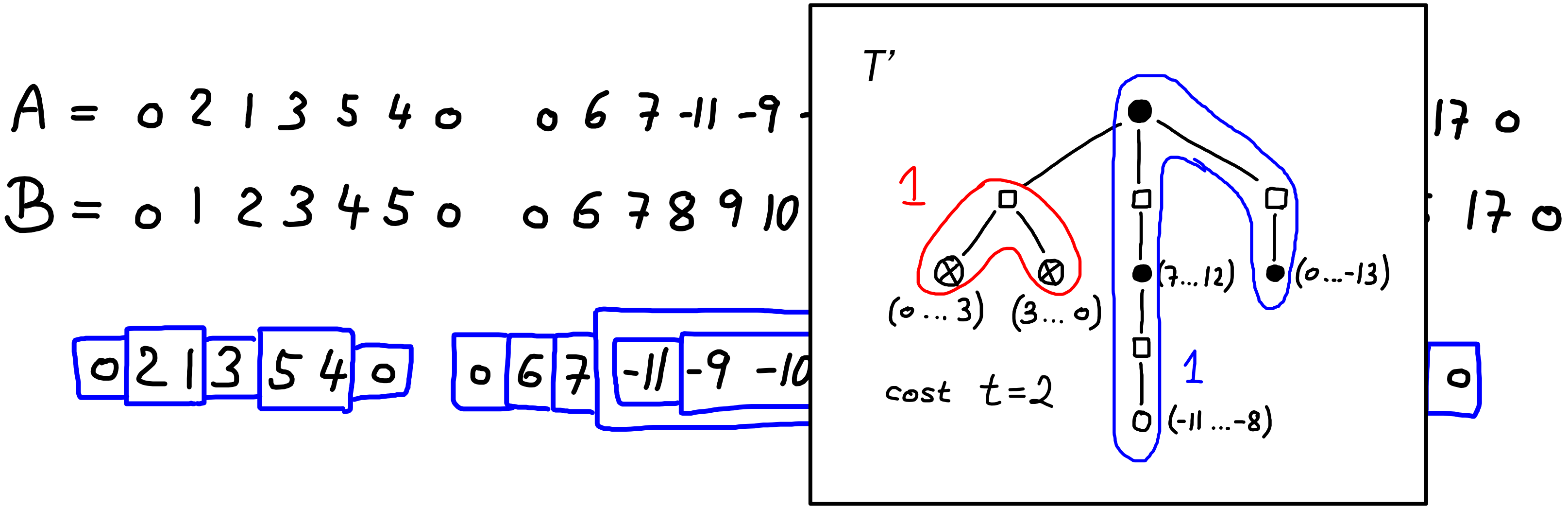


Sometimes HP needs more steps than DCJ: *hurdle*, *fortress*, *knot*, *semi-knot*, *real-knot*, *semi-real-knot*, *weak-fortress-of-real-knots*, etc.

Can we quantify this relative to DCJ?

$$d^{HP}(A,B) \; = \; d^{DCJ}(A,B) + t$$

General HP distance problem



Theorem:

If *t* is the cost of an optimal cover of *T'*, then

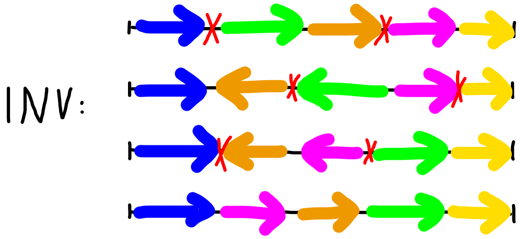
$$d^{HP}(A,B) = d^{DCJ}(A,B) + t$$

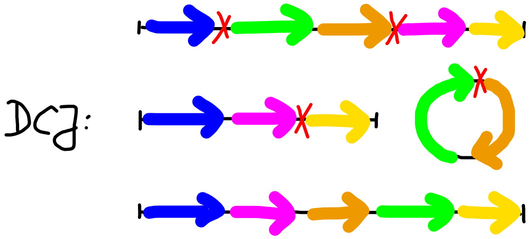
- Closed formula for *t* (Erdős, Soukup & S: *Appl. Math. Lett.* 2011)
- Linear-time algorithm for distance computation (Bergeron, Mixtacki & S: *TCS* 2009)
- Similar result for inversion distance (Bergeron, Mixtacki & S: *Proc. of CPM* 2004)
- Similar result for translocation distance (Bergeron, Mixtacki & S: *JCB* 2006)

Restricted DCJ

(*based on*: Kováč, Warren, Braga & S: *JCB* 2011)

Original motivation for DCJ (Yancopoulos, Attie & Friedberg 2005):
block interchange in 2 steps (instead of 3 as in the INV model)

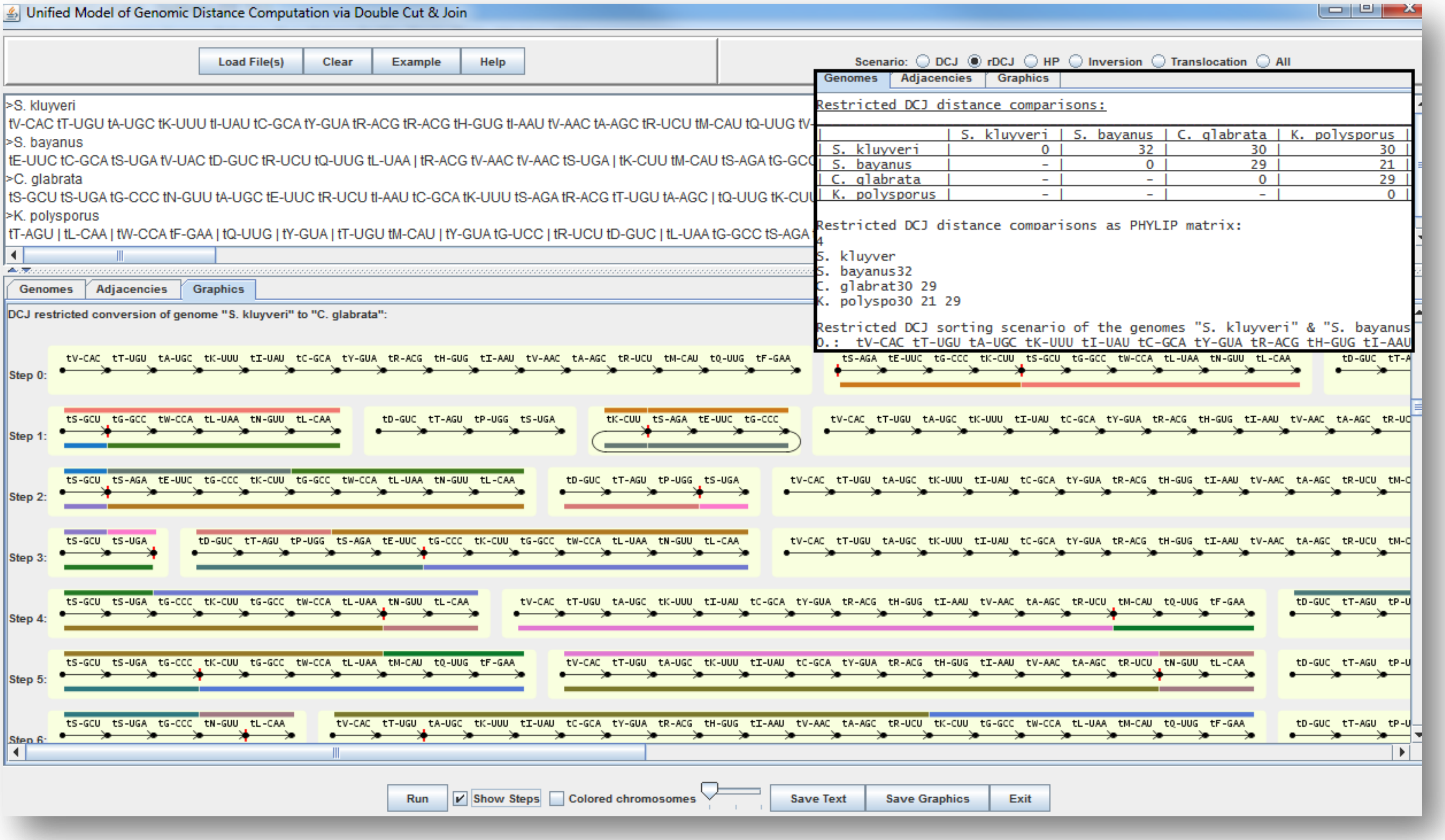




Observation:
We need never more than 1 circular chromosome at a time, $d^{rDCJ}(A,B) = d^{DCJ}(A,B)$.

Algorithmic results: Distance calculation in $O(N)$ time
Sorting in $O(N \log N)$ time [lower bound?]

Software: UNIMoG


Unified Model of Genomic Distance Computation via Double Cut & Join
Load File(s)
Clear
Example
Help
Scenario: ○ DCJ ● rDCJ ○ HP ○ Inversion ○ Translocation ○ All
Genomes
Adjacencies
Graphics

```
>S. kluyveri
tV-CAC tT-UGU tA-UGC tK-UUU tI-UAU tC-GCA tY-GUA tR-ACG tR-ACG tH-GUG tI-AAU tV-AAC tA-AGC tR-UCU tM-CAU tQ-UUG tV-
>S. bayanus
tE-UUC tC-GCA tS-UGA tV-UAC tD-GUC tR-UCU tQ-UUG tL-UAA | tR-ACG tV-AAC tV-AAC tS-UGA | tK-CUU tM-CAU tS-AGA tG-GCC
>C. glabrata
tS-GCU tS-UGA tG-CCC tN-GUU tA-UGC tE-UUC tR-UCU tI-AAU tC-GCA tK-UUU tS-AGA tR-ACG tT-UGU tA-AGC | tQ-UUG tK-CUU
>K. polysporus
tT-AGU | tL-CAA | tW-CCA tF-GAA | tQ-UUG | tY-GUA | tT-UGU tM-CAU | tY-GUA tG-UCC | tR-UCU tD-GUC | tL-UAA tG-GCC tS-AGA
```

```
Restricted DCJ distance comparisons:

|             | S. kluyveri | S. bayanus | C. glabrata | K. polysporus |
| S. kluyveri |           0 |         32 |          30 |            30 |
| S. bayanus  |           - |          0 |          29 |            21 |
| C. glabrata |           - |          - |           0 |            29 |
| K. polysporus |         - |          - |           - |             0 |

Restricted DCJ distance comparisons as PHYLIP matrix:
4
S. kluyver
S. bayanus32
C. glabrat30 29
K. polyspo30 21 29

Restricted DCJ sorting scenario of the genomes "S. kluyveri" & "S. bayanus
0.:  tV-CAC tT-UGU tA-UGC tK-UUU tI-UAU tC-GCA tY-GUA tR-ACG tH-GUG tI-AAU
```

DCJ restricted conversion of genome "S. kluyveri" to "C. glabrata":

Step 0:
tV-CAC tT-UGU tA-UGC tK-UUU tI-UAU tC-GCA tY-GUA tR-ACG tH-GUG tI-AAU tV-AAC tA-AGC tR-UCU tM-CAU tQ-UUG tF-GAA
tS-AGA tE-UUC tG-CCC tK-CUU tS-GCU tG-GCC tW-CCA tL-UAA tN-GUU tL-CAA
tD-GUC tT-A
Step 1:
tS-GCU tG-GCC tW-CCA tL-UAA tN-GUU tL-CAA
tD-GUC tT-AGU tP-UGG tS-UGA
tK-CUU tS-AGA tE-UUC tG-CCC
tV-CAC tT-UGU tA-UGC tK-UUU tI-UAU tC-GCA tY-GUA tR-ACG tH-GUG tI-AAU tV-AAC tA-AGC tR-UC
Step 2:
tS-GCU tS-AGA tE-UUC tG-CCC tK-CUU tG-GCC tW-CCA tL-UAA tN-GUU tL-CAA
tD-GUC tT-AGU tP-UGG tS-UGA
tV-CAC tT-UGU tA-UGC tK-UUU tI-UAU tC-GCA tY-GUA tR-ACG tH-GUG tI-AAU tV-AAC tA-AGC tR-UCU tM-C
Step 3:
tS-GCU tS-UGA
tD-GUC tT-AGU tP-UGG tS-AGA tE-UUC tG-CCC tK-CUU tG-GCC tW-CCA tL-UAA tN-GUU tL-CAA
tV-CAC tT-UGU tA-UGC tK-UUU tI-UAU tC-GCA tY-GUA tR-ACG tH-GUG tI-AAU tV-AAC tA-AGC tR-UCU tM-C
Step 4:
tS-GCU tS-UGA tG-CCC tK-CUU tG-GCC tW-CCA tL-UAA tN-GUU tL-CAA
tV-CAC tT-UGU tA-UGC tK-UUU tI-UAU tC-GCA tY-GUA tR-ACG tH-GUG tI-AAU tV-AAC tA-AGC tR-UCU tM-CAU tQ-UUG tF-GAA
tD-GUC tT-AGU tP-U
Step 5:
tS-GCU tS-UGA tG-CCC tK-CUU tG-GCC tW-CCA tL-UAA tM-CAU tQ-UUG tF-GAA
tV-CAC tT-UGU tA-UGC tK-UUU tI-UAU tC-GCA tY-GUA tR-ACG tH-GUG tI-AAU tV-AAC tA-AGC tR-UCU tN-GUU tL-CAA
tD-GUC tT-AGU tP-U
Step 6:
tS-GCU tS-UGA tG-CCC tN-GUU tL-CAA
tV-CAC tT-UGU tA-UGC tK-UUU tI-UAU tC-GCA tY-GUA tR-ACG tH-GUG tI-AAU tV-AAC tA-AGC tR-UCU tK-CUU tG-GCC tW-CCA tL-UAA tM-CAU tQ-UUG tF-GAA
tD-GUC tT-AGU tP-U
Run
☑ Show Steps
☐ Colored chromosomes
Save Text
Save Graphics
Exit

(Hilker *et al.*: *Bioinformatics* 2012; http://bibiserv.techfak.uni-bielefeld.de/dcj)

Further applications of the DCJ model

Estimating the true evolutionary distance:
Lin & Moret 2008

Perfect rearrangement:
Bérard, Chateau, Chauve, Paul, Tannier 2008

Genome halving:
Warren & Sankoff 2008; Mixtacki 2008; Thomas, Ouangraoua & Varré 2012

DCJ Median:
Xu & Sankoff 2008; Lenne *et al.* 2008; Zhang, Arndt & Tang 2009; Xu 2009; Aganezov & Alekseyev 2012

Multiple genome rearrangement:
Adam & Sankoff 2008; Kováč, Brejová & Vinař 2011

5. Insertions, deletions, substitutions

(*based on*: Braga, Willing & S, *JCB* 2011)

So far: Only organizational operations

Now: Mixture of organizational and content-modifying operations

History:
Inversions + indels: El-Mabrouk 2001; Marron, Swenson & Moret 2004

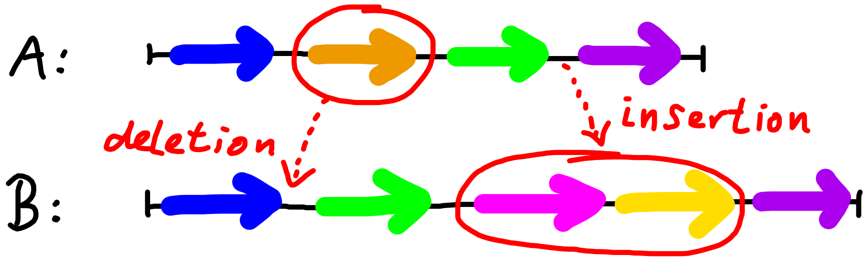
Here:
DCJ + indels: Yancopoulos & Friedberg 2008; Braga, Willing & S 2010; Braga 2010; Braga, Machado, Ribeiro & S 2011b; Da Silva, Braga, Machado & Dantas 2012

Again, the results in the DCJ model are much simpler than in INV or HP.
But we also run into modeling questions, as we will see later.

Insertion/Deletion

Extended model: Genomes with possibly unequal gene content

Unique blocks: Blocks only occurring in one of the two genomes



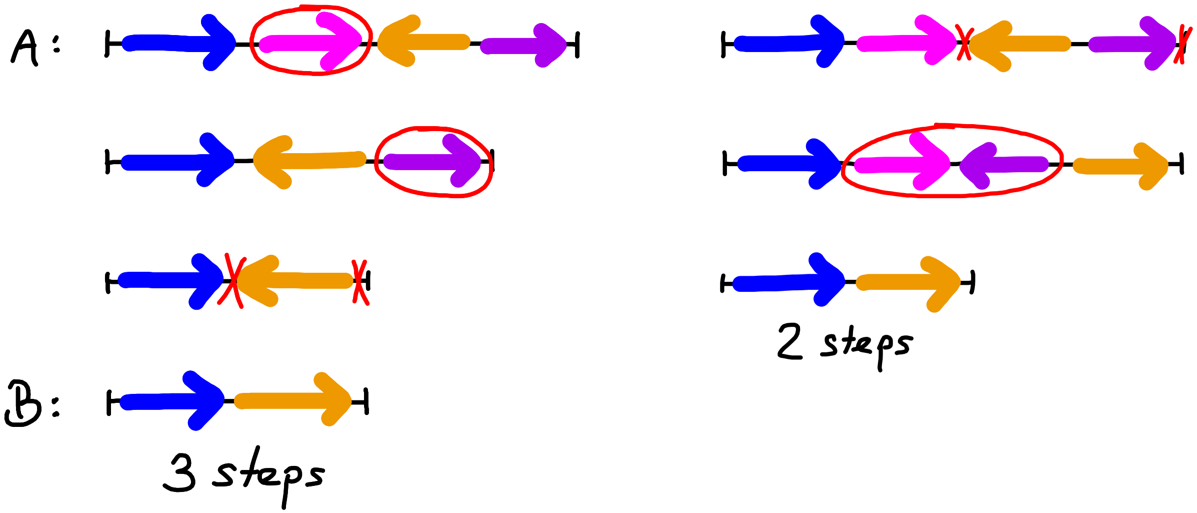
DCJ-indel distance:

Given two genomes *A* and *B*, find the minimum number of steps (DCJ **and** indel operations) $d^{DCJ\text{-}id}(A,B)$ necessary to sort *A* into *B*.

We consider: cost for 1 insertion = cost for 1 deletion = cost for 1 DCJ

The DCJ-indel model

Saving indel operations:



Group unique blocks during sorting ➔ less indel operations

The DCJ-indel model

Result:

$$d^{DCJ-id}(A,B) \;=\; d^{DCJ}(A,B) \;+\; \sum_{C \in AG(A,B)} \lambda(C) \;-\; W$$

Theorem:
Given two genomes *A* and *B*, $d^{DCJ\text{-}id}(A,B)$ and a shortest sorting scenario can be computed in linear time O(|*A*|+|*B*|).

In fact, indels can be traded for DCJ operations, for example:

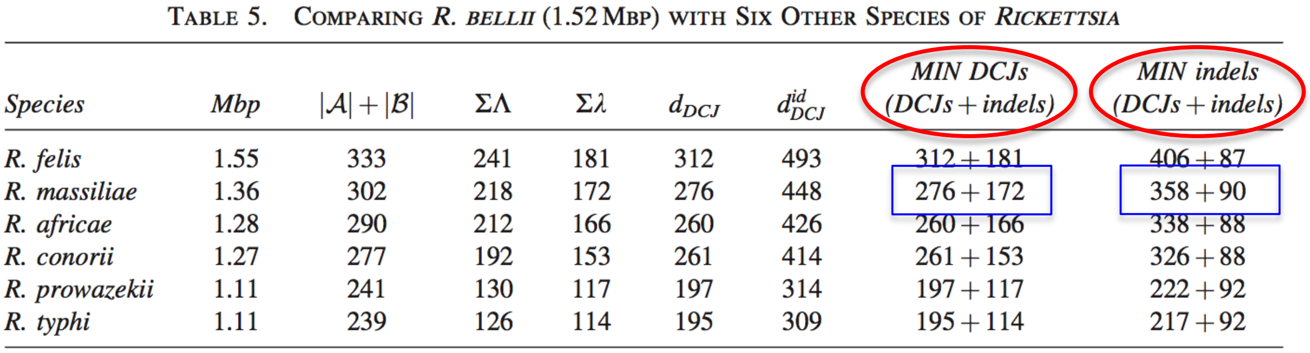
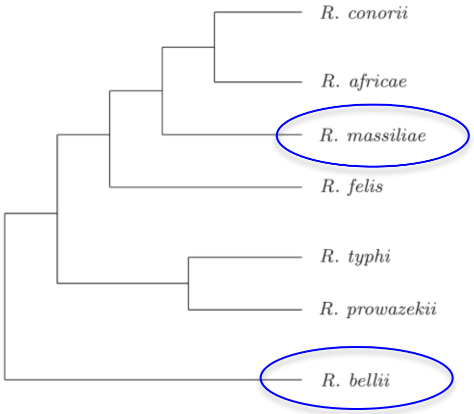


TABLE 5. COMPARING *R. BELLII* (1.52 MBP) WITH SIX OTHER SPECIES OF *RICKETTSIA*

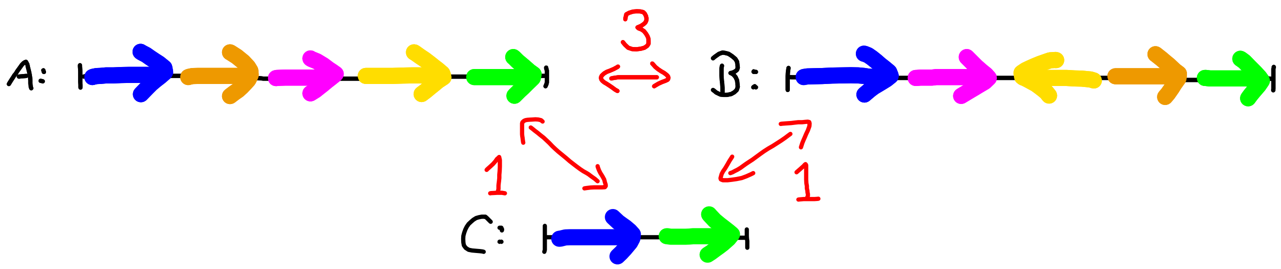
Species	*Mbp*	$\lvert\mathcal{A}\rvert + \lvert\mathcal{B}\rvert$	$\Sigma\Lambda$	$\Sigma\lambda$	d_{DCJ}	d^{id}_{DCJ}	*MIN DCJs (DCJs + indels)*	*MIN indels (DCJs + indels)*
R. felis	1.55	333	241	181	312	493	312 + 181	406 + 87
R. massiliae	1.36	302	218	172	276	448	276 + 172	358 + 90
R. africae	1.28	290	212	166	260	426	260 + 166	338 + 88
R. conorii	1.27	277	192	153	261	414	261 + 153	326 + 88
R. prowazekii	1.11	241	130	117	197	314	197 + 117	222 + 92
R. typhi	1.11	239	126	114	195	309	195 + 114	217 + 92



6. On the weight of indels

(*based on*: Braga, Machado, Ribeiro & S: *BMC Bioinformatics* 2011b)

Observation (Yancopoulos & Friedberg 2008):
When indel operations of multiple blocks are allowed, the triangle inequality may be disrupted.



$$d(A,B) > d(A,C) + d(C,B)$$

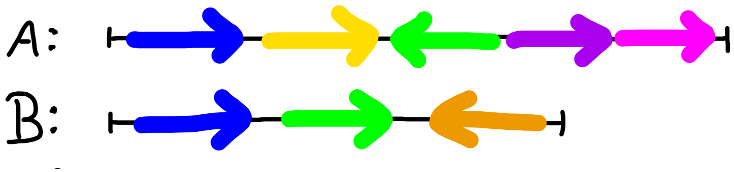
Question: Is there a distance definition that does not disrupt the triangle inequality?

A posteriori correction

Lemma:
Applying an *a posteriori* correction, the triangle inequality holds for the function

$$d_{1,k}^{DCJ\text{-}id}(A,B) = d^{DCJ\text{-}id}(A,B) + k \cdot u(A,B)$$

and for any constant $k \geq 1$, where $u(A,B)$ = # of unique markers in A and B.



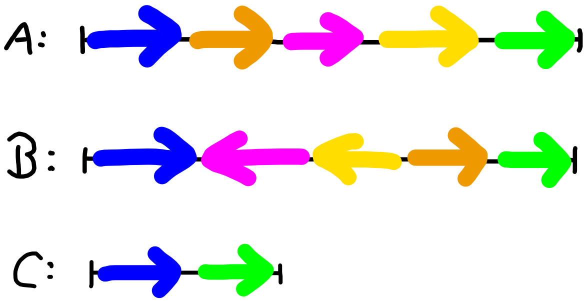
$$\left.\begin{array}{l} d^{DCJ\text{-}id}(A,B) = 3 \\ u(A,B) = 4 \end{array}\right\} \; d_{1,1}^{DCJ\text{-}id}(A,B) = 7$$

Algorithm:

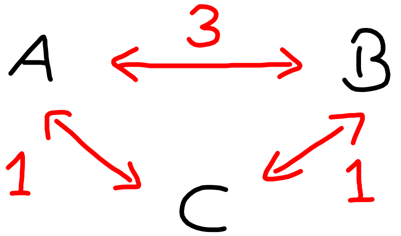
1. Compute $d^{DCJ\text{-}id}(A,B)$ by the standard algorithm
2. Add $k \cdot u(A,B)$ to obtain the corrected metric distance

Question: What is the best choice of k ?

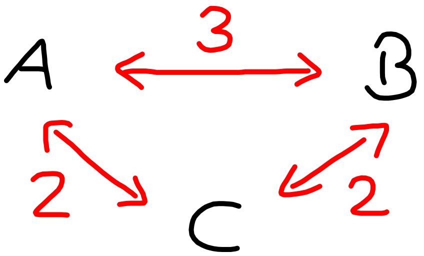
More plausible distances?



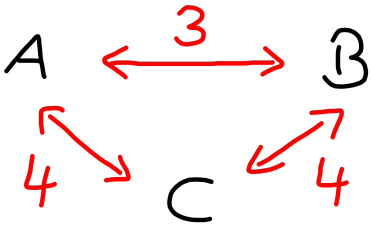
uncorrected distances



“ghost-DCJ model” (YF 2010)



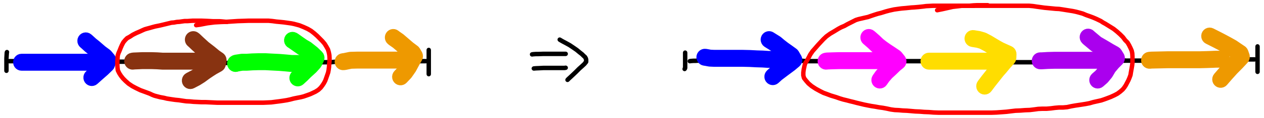
DCJ-indel model $d_{1,1}^{DCJ\text{-}id}$



DCJ with substitutions

(*based on*: Braga, Machado, Ribeiro & S: *BMC Bioinformatics* 2011a)

Consider the simultaneous substitution of $m \geq 0$ markers by $n \geq 0$ markers.



- subsumes the DCJ-indel model
- distances become slightly smaller

Lemma:

The corrected DCJ-substitution distance $d_{1,k}^{DCJ\text{-}sb}$ satisfies the triangular inequality if and only if $k \geq 3/4$.

7. Summary and Conclusion

- Genome evolution, rearrangement
- DCJ, distance and sorting, restricted DCJ
- Relation to HP, INV, translocation models
- DCJ + indels, DCJ + substitutions, indel/substitution weights

- Power of DCJ: simple + tractable, generalizable
- More advanced questions can be asked
- (not talked about median, but there is a lot)

- More formal/algorithmic than biological results → typical for the field
- Analysis is still very manual, e.g. no software where I can upload a few genomes …
- But the field is changing, more and more biological studies are upcoming

Acknowledgments

Anne Bergeron

Marília D. V. Braga

Paul Medvedev

Julia Mixtacki

Eyla Willing

RECOMB Comparative Genomics 2012
October 17-19 — Niterói, Brazil

Thank you!

References

AP 2007 – Alekseyev, Pevzner: Colored de Bruijn graphs and the genome halving problem. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 4 (1): 98-107, 2007.

AS 2008 – Adam, Sankoff: The ABCs of MGR with DCJ. *Evol. Bioinf. Online* 4: 69-74, 2008.

BCCPT 2008 – Bérard, Chateau, Chauve, Paul, Tannier: Perfect DCJ rearrangement. *Proc. of RECOMB-CG 2008*, 158-169, 2008.

BFR 20xx – Bulteau, Fertin, Rusu: Sorting by transpositions is difficult. *SIAM J. Discrete Math.*, to appear.

BHS 2002 – Bergeron, Heber, S: Common intervals and sorting by reversals: a marriage of necessity. *Bioinformatics* 18 (Suppl. 2): S54-S63, 2002.

BMRS 2011a – Braga, Machado, Ribeiro, S: Genomic distance under gene substitutions. *BMC Bioinformatics* 12 (Suppl. 9): S8, 2011.

BMRS 2011b – Braga, Machado, Ribeiro, S: On the weight of indels in genomic distances. *BMC Bioinformatics* 12 (Suppl. 9): S13, 2011.

BMS 2004 – Bergeron, Mixtacki, S: Reversal distance without hurdles and fortresses. *Proc. of CPM 2004*, 388-399, 2004.

BMS 2005 – Bergeron, Mixtacki, S: On sorting by translocations. *Proc. of RECOMB 2005*, 615-629, 2005.

BMS 2006 – Bergeron, Mixtacki, S: A unifying view of genome rearrangements. *Proc. of WABI 2006*, 163-173, 2006.

BMS 2008 – Bergeron, Mixtacki, S: HP distance via Double Cut and Join distance. *Proc. of CPM 2008*, 56-68, 2008.

BMS 2010 – Bergeron, Medvedev, S: Rearrangement models and single-cut operations. *J. Comput. Biol.* 17 (9): 1213-1225, 2010. (Extended version of MS 2009)

BMY 2001 – Bader, Moret. Yan: A linear-time algorithm for computing inversion distance between signed permutations with an experimental study. *J. Comput. Biol.* 8 (5): 483-491, 2001.

BP 1993 – Bafna, Pevzner: Genome Rearrangements and Sorting by Reversals. *Proc of FOCS 1993*, 148-157, 1993. (Preliminary version of BP1996)

BP 1996 – Bafna, Pevzner: Genome Rearrangements and Sorting by Reversals. *SIAM J. Computing* 25 (1), 272-289, 1996. (Extended version of BP 1993)

BP 1998 – Bafna, Pevzner: Sorting by transpositions. *SIAM J. Discrete Math.* 11 (2): 224-240, 1998.

BS 2009 – Braga, S: Counting All DCJ Sorting Scenarios. *Proc. of RECOMB-CG* 2009, 36-47, 2009. (Preliminary version of BS 2010)

BS 2010 – Braga, S: The Solution Space of Sorting by DCJ. J. Comp. Biol. 17 (9): 1145-1165, 2010. (Extended version of BS 2009)

BWS 2010 – Braga, Willing, S: Genomic distance with DCJ and indels. *Proc. of WABI 2010*, 90-101, 2010.

C 1996 – Christie: Sorting permutations by block-interchanges. *Inf. Process. Lett.* 60 (4): 165-169, 1996.

C 1998 – Christie: A 3/2 approximation algorithm for sorting by reversals. *Proc. of SODA 1998*, 244-252, 1998.

E-M 2001 – El-Mabrouk: Sorting signed permutations by reversals and insertions/deletions of contiguous segments. *J Discrete Alg.* 1, 105–122, 2001.

E-M 2002 – El-Mabrouk: Reconstructing an ancestral genome using minimum segments duplications and reversals. *J. Comput. Syst. Sci.* 65 (3): 442-464, 2002.

E-MNS – El-Mabrouk, Nadeau, Sankoff: Genome Halving. Proc. of CPM 1998, 235-250, 1998.

E-MS 2003 – El-Mabrouk, Sankoff: The reconstruction of doubled genomes. *SIAM J. Comput.* 32 (3): 754-792, 2003.

ES 2003 – Eichler, Sankoff: Structural dynamics of eukaryotic chromosome evolution. *Science* 301: 793-797, 2003.

ESS 2011 – Erdős, Soukup, S: Balanced vertices in trees and a simpler algorithm to compute the genomic distance. *Appl. Math. Lett.* 24 (1): 82-86, 2011.

H 1996 – Hannenhalli: Polynomial-time algorithm for computing translocation distance between genomes. Discrete Appl. Math. 71 (1): 137-151, 1996.

HP 1995 – Hannenhalli, Pevzner: Transforming men into mice (polynomial algorithm for genomic distance problem). Proc. of FOCS 1995, 581-592, 1995.

JN 2007 – Jean, Nikolski: Genome rearrangements: a correct algorithm for optimal capping. Inform. Process. Lett. 104 (1): 14-20, 2007.

KS 1994 – Kececioglu, Sankoff: Efficient bounds for oriented chromosome inversion distance. Proc. Of CPM 1994, 307-325, 1994.

KST 1999 – Kaplan, Shamir, Tarjan: A faster and simpler algorithm for sorting signed permutations by reversals. *SIAM J. Comput.* 29 (3): 880-892, 1999.

LM 2008 – Lin, Moret: Estimating true evolutionary distances under the DCJ model. *Bioinformatics* 24 (13): i114-i122, 2008.

M 2008 – Mixtacki: Genome Halving under DCJ revisited. *Proc. of COCOON 2008*, 276-286, 2008.

MD 2002 – Meidanis, Dias. Genome rearrangements distance by fusion, fission, and transposition is easy. *Proc. of SPIRE'2001*, 2001.

MS 2009 – Medvedev, S: Rearrangement models and single-cut operations. *Proc. of RECOMB-CG 2009*, 84-97, 2009. (Preliminary version of BMS 2010)

MSM 2004 – Marron, Swenson, Moret: Genomic distances under deletions and insertions. *Theor. Comput. Sci.* 325, 347–360, 2004.

MWD 1997 – Meidanis , Walter, Dias: Transposition distance between a permutation and its reverse. *Proc. of WSP 1997*, 70-79, 1997.

MWD 2002 – Walter, Dias, Meidanis: A lower bound on the reversal and transposition diameter. *J. Comp. Biol.* 9 (5) 743–745, 2002. (Extended version of WDM 1998)

O-FS 2003 – Two notes on genome rearrangement. *J. Bioinf. Comput. Biol.* 1 (1): 71-94, 2003.

PH 1988 – Palmer, Herbon: Plant mitochondrial DNA evolved rapidly in structure, but slowly in sequence. *J. Mol. Evol.* 28 (1-2): 87-97, 1988.

PT 2003 – Pevzner, Tesler: Genome rearrangements in mammalian evolution: lessons from human and mouse genomes. *Genome Res.* 13: 37-45, 2003.

S 1992 – Sankoff: Edit distance for genome comparison based on non-local operations. *Proc. of CPM 1992*, 121-135, 1992.

T 2002 – Tesler: Efficient algorithms for multichromosomal genome rearrangements. *J. Comput. Syst. Sci.* 65 (3): 587-609, 2002.

WDM 1998 – Walter, Dias, Meidanis: Reversal and transposition distance of linear chromosomes. *Proc. of SPIRE 1998*, 96-102, 1998. (Preliminary version of WDM 2002)

WEHM 1982 – Watterson, Ewens, Hall, Morgan: The chromosome inversion problem. *J. Theor. Biol.* 99 (1): 1-7, 1982.

WS 2009 – Waren, Sankoff: Genome aliquoting with double cut and join. *BMC Bioinformatics* 10 (Suppl. 1): S2, 2009.

YAF 2005 – Yancopoulos, Attie, Friedberg: Efficient sorting of genomic permutations by translocation, inversion and block interchange. *Bioinformatics* 21 (16): 3340-3346, 2005.

YF 2008 – Yancopoulos, Friedberg: Sorting genomes with insertions, deletions and duplications by DCJ. *Proc. of RECOMB-CG 2008*, 170-183, 2008. (Preliminary version of YF 2010)

YF 2009 – Yancopoulos, Friedberg: DCJ path formulation for genome transformations which include insertions, deletions, and duplications. *J. Comput. Biol.* 16 (10): 1311-1338, 2009. (Extended version of YF 2008)

ZAT 2009 – Zhang, Arndt, Tang: An exact solver for the DCJ median problem. *Proc. of PSB 2009*, 138-149, 2009.