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?

The mouse genome:

1. -136 140 93 -95 -32 25 37 -38 39 -40 70 246 30 -29 33 -8 14 -11 10 -9 o
3. -141 139 -57 56 58 -96 -201 55 -70 7 6 6 -5 o
4. -137 -142 -138 -97 146 153 148 145 4 3 2 1 o
5. -116 -115 120 124 18 62 -63 64 6 -67 195 -196 197 -113 -114 -119 105 118 200 o
7. -257 -255 254 -256 177 -210 212 211 -221 220 219 -218 -184 176 224 174 -175 -183 o
13. -160 -13 -111 -49 88 -152 110 86 81 149 152 -72 -74 o
15. -73 143 270 190 o
16. -223 -135 -263 59 61 -69 -52 263 o
17. -102 -103 184 -75 -222 91 262 -90 -92 44 -26 249 77 -289 19 239 o
18. 164 163 -166 243 -31 78 82 79 -83 241 245 242 -244 -247 o
19. 182 -181 -147 144 -169 173 o
X: -274 -275 273 281 -272 278 -279 280 -276 277 -271 o

The human genome:

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

Data from: Pevzner & Tesler 2003

Figure: Eichler & Sankoff 2003
What happens in detail?

Basic rearrangement operations:

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

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 $d(A,B) \rightarrow \text{“distance problem”}$
- diameter problems
- distribution of distances
- halving distance

How much can we reconstruct of the past?

- Ancestral genome(s)
- rearrangement scenario(s) $\rightarrow \text{“sorting problem”}$
- complete phylogenies
Some history (2 genomes)

**Inversions (reversals):**

**Translocations:**
Hannenhalli 1996; Bergeron, Mixtacki & S 2005

**Multichromosomal linear ("general HP model"):**

**Double Cut and Join (DCJ):**

**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:

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

DCJ + Substitutions:
Braga, Machado, Ribeiro & S 2011a
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
**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 $\{r\}$.

\[\text{(b) and (d)}\]
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:

A

B

**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)
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 $B$
Algorithm

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

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

// Generate the telomeres of $B$ that are not yet present in $A$
9: for each telomere $\{p\}$ in $B$ do
10: let $u$ be the vertex of $A$ that contains $p$
11: if $u$ is an adjacency then
12: replace vertex $u$ in $A$ by $\{p\}$ and $(u \setminus \{p\})$
13: end if
14: 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):**

$$N = 281, \ C = 27, \ I = 16 \ \rightarrow \ d_{DCJ}(\text{Human,Mouse}) = 246$$

**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:

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

<table>
<thead>
<tr>
<th>1 component (distance $\ell$)</th>
<th>number of scenarios</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>16</td>
</tr>
<tr>
<td>4</td>
<td>125</td>
</tr>
<tr>
<td>5</td>
<td>1296</td>
</tr>
<tr>
<td>6</td>
<td>16807</td>
</tr>
</tbody>
</table>
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: \( \text{INV} \rightarrow \text{HP} \)
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?

\[ d^{HP}(A,B) = d^{DCJ}(A,B) + t \]
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)

\[ d_{\text{rDCJ}}(A,B) = d_{\text{DCJ}}(A,B) \]

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

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

(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:**

**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.
**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) $d^{DCJ-id} (A,B)$ 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 \( B \), \( d^{DCJ-id}(A,B) \) 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:

| Species             | Mbp  | |A| + |B| | \( \Sigma \) | \( \Sigma \) | \( d_{DCJ} \) | \( d^{id}_{DCJ} \) | \( MIN \ DCJs \) (DCJs + indels) | \( MIN \ indels \) (DCJs + indels) |
|---------------------|------|----------------|------------|------------|----------------|----------------|-----------------|----------------|----------------|
| R. felis            | 1.55 | 333            | 241        | 181        | 312           | 493            | 312 + 181       | 406 + 87        |                 |
| R. massiliæ         | 1.36 | 302            | 218        | 172        | 276           | 448            | 276 + 172       | 358 + 90        |                 |
| R. africæ           | 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

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

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

Algorithm:
1. Compute \( d^{DCJ-id}(A,B) \) 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?

A: \[\text{unbalanced diagrams}\]
B: \[\text{unbalanced diagrams}\]
C: \[\text{balanced diagrams}\]

uncorrected distances

A \[\leftrightarrow 3 \leftrightarrow B\]
C \[\leftrightarrow 1 \leftrightarrow C\]

“ghost-DCJ model” (YF 2010)

DCJ-indel model \(d^{DCJ-id}_{1,1}\)
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 $d^{DCJ_{sb}}_{1,k}$ satisfies the triangular inequality if and only if $k \geq 3/4$. 

\[
\begin{align*}
\text{original sequence} & \quad \Rightarrow \quad \text{substituted sequence}
\end{align*}
\]
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
Thank you!