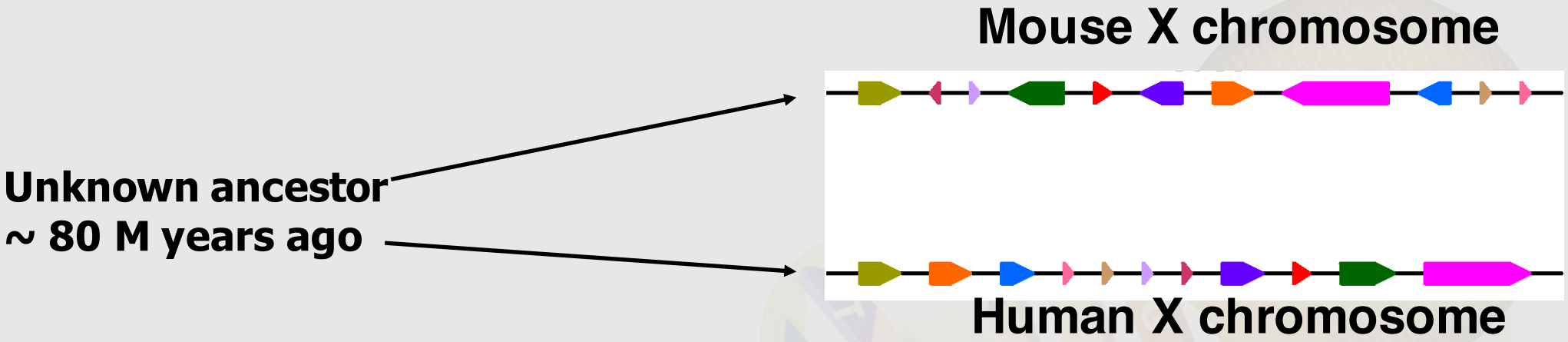
Genome Rearrangements:

from Biological Problem to Combinatorial Algorithms (and back)

Pavel Pevzner

Department of Computer Science, University of California at San Diego

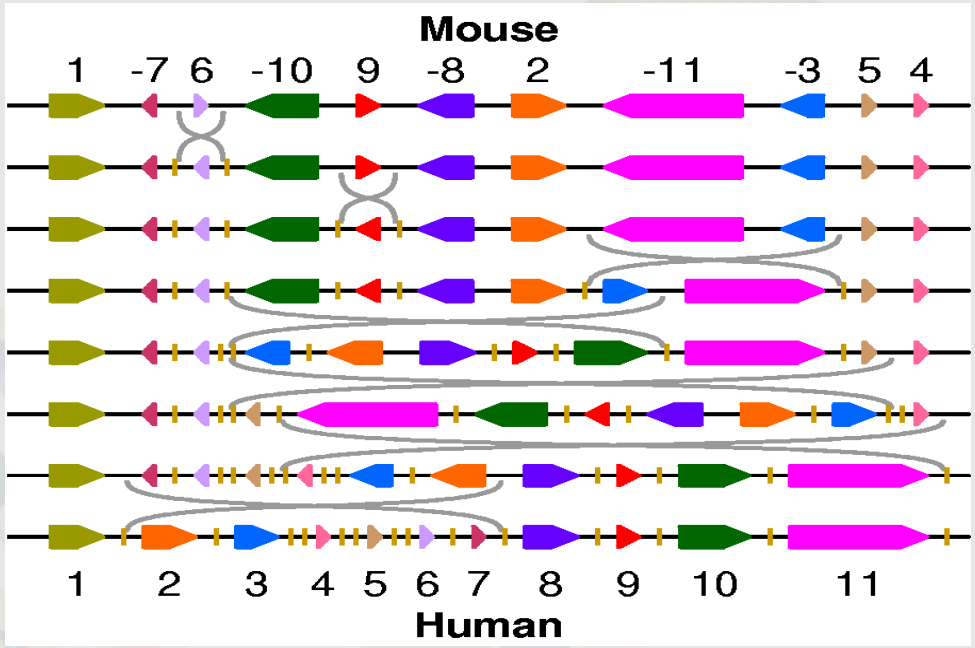
Genome Rearrangements



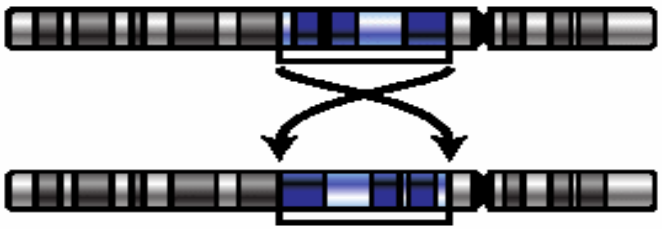
- ✓ What is the evolutionary scenario for transforming one genome into the other?
- ✓ What is the organization of the ancestral genome?
- ✓ Are there any rearrangement hotspots in mammalian genomes?

Genome Rearrangements: Evolutionary Scenarios

Unknown ancestor
~ 80 M years ago



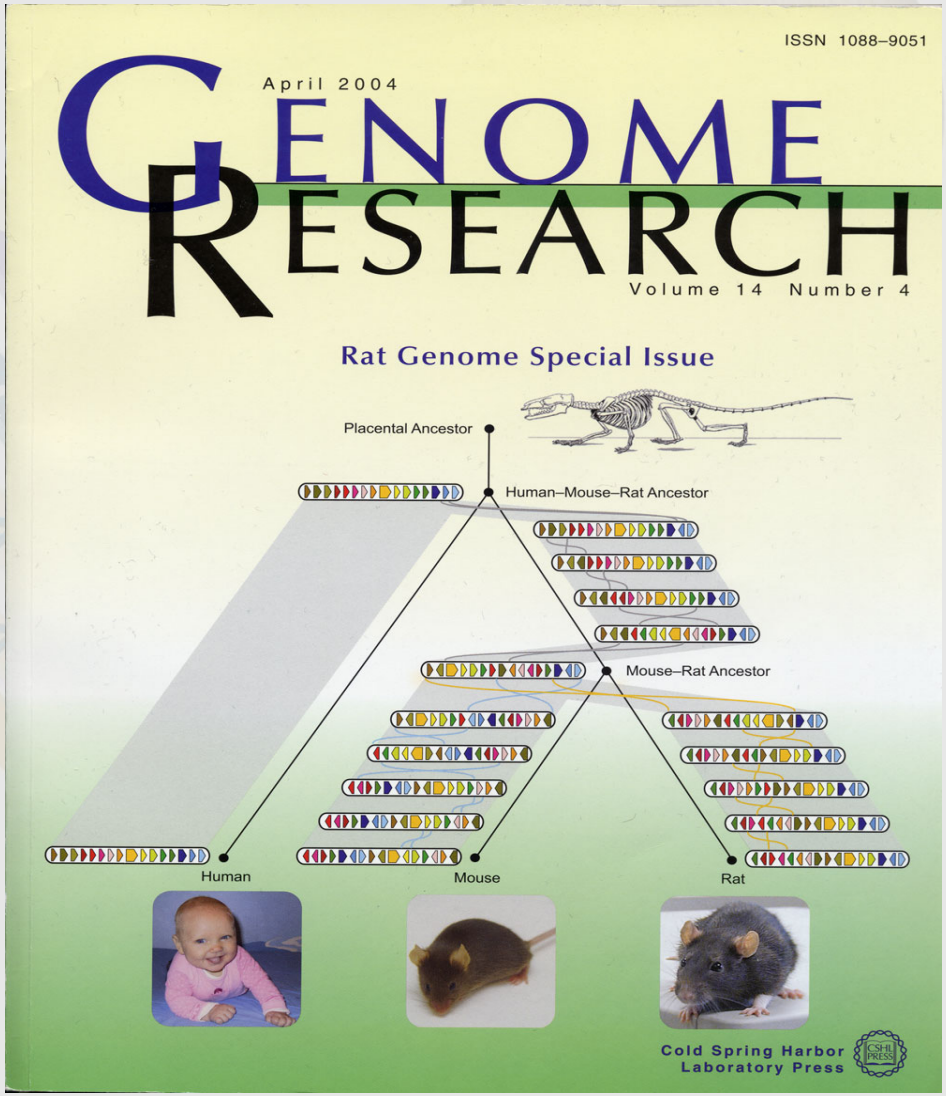
- ✓ What is the evolutionary scenario for transforming one genome into the other?
- ✓ What is the organization of the ancestral genome?
- ✓ Are there any rearrangement hotspots in mammalian genomes?



Reversal flips a segment of a chromosome

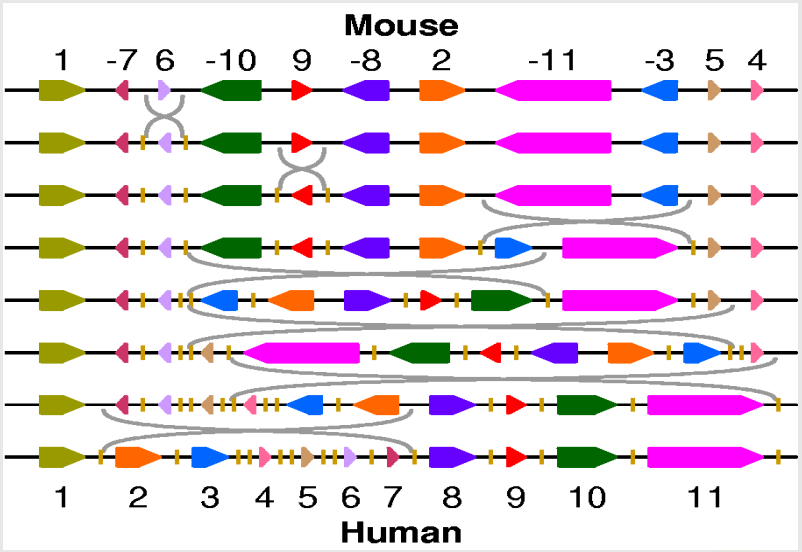
Genome Rearrangements: Ancestral Reconstruction

- ✓ What is the evolutionary scenario for transforming one genome into the other?
- ✓ What is the organization of the ancestral genome?
- ✓ Are there any rearrangement hotspots in mammalian genomes?

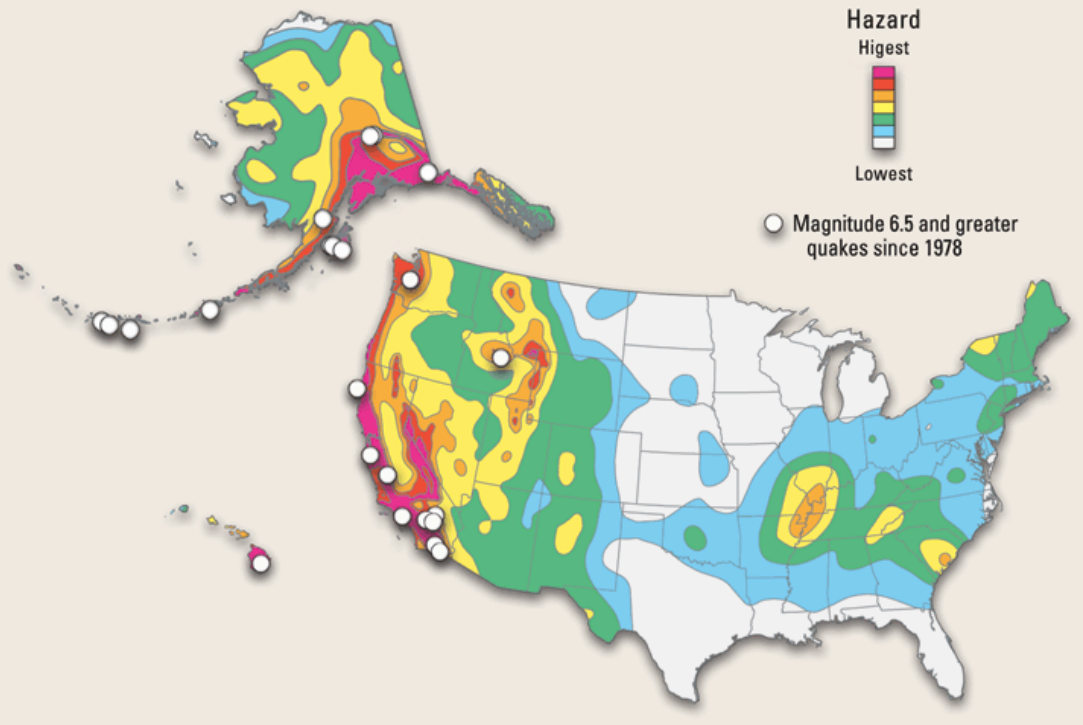


Bourque, Tesler, PP, Genome Res.

Genome Rearrangements: Evolutionary "Earthquakes

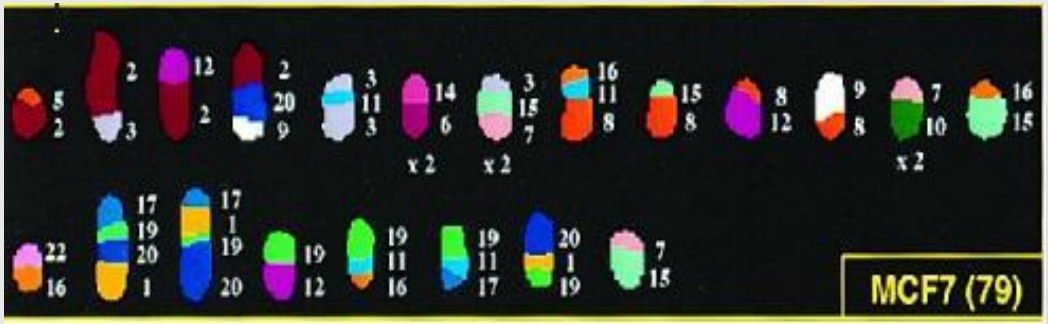


- ✓ What is the evolutionary scenario for transforming one genome into the other?
- ✓ What is the organization of the ancestral genome?
- ✓ Are there any rearrangement hotspots in mammalian genomes?

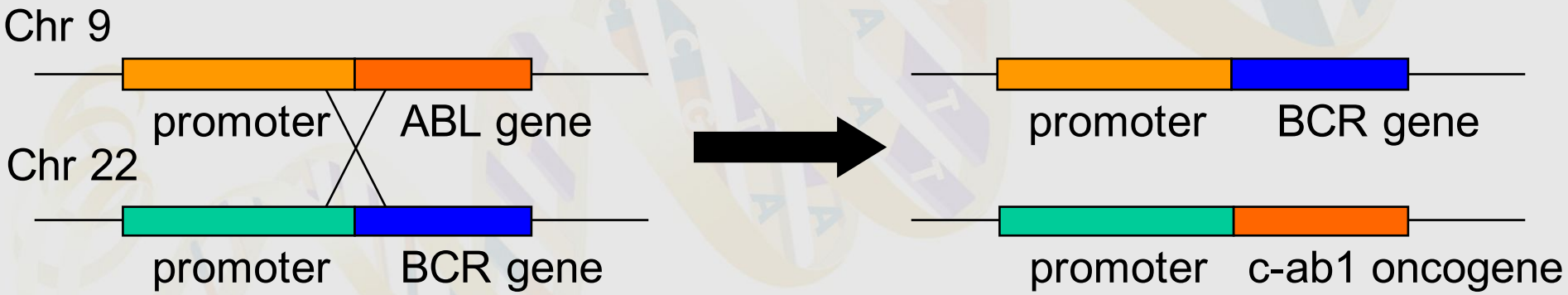


Rearrangement Hotspots in Tumor Genomes

- ✓ Rearrangements may disrupt genes and alter gene regulation.



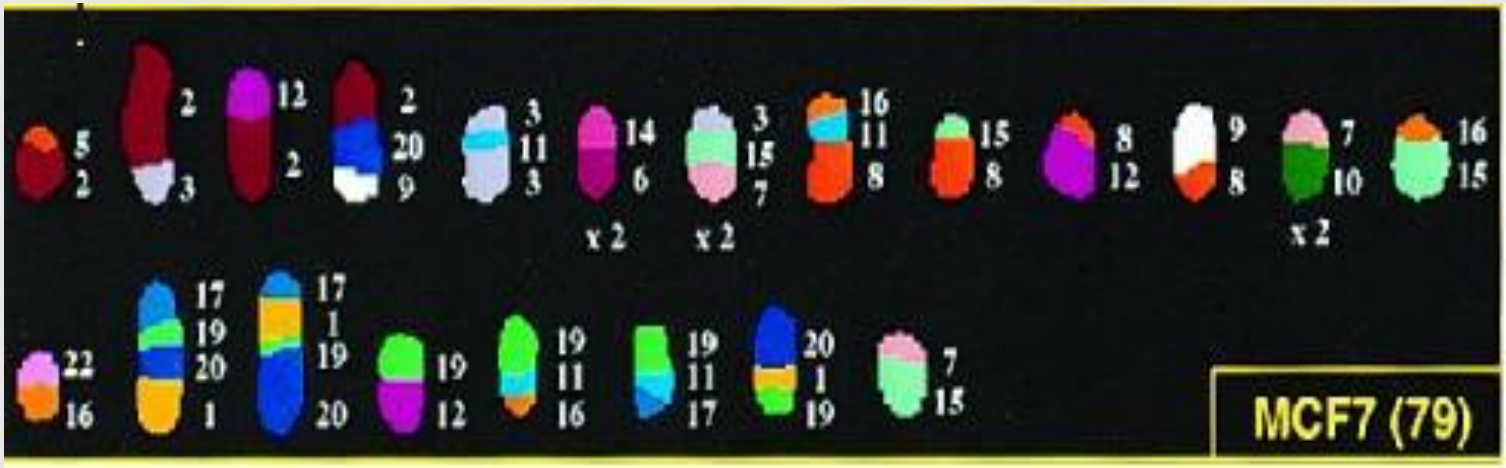
- ✓ Example: rearrangement in leukemia yields "Philadelphia" chromosome:



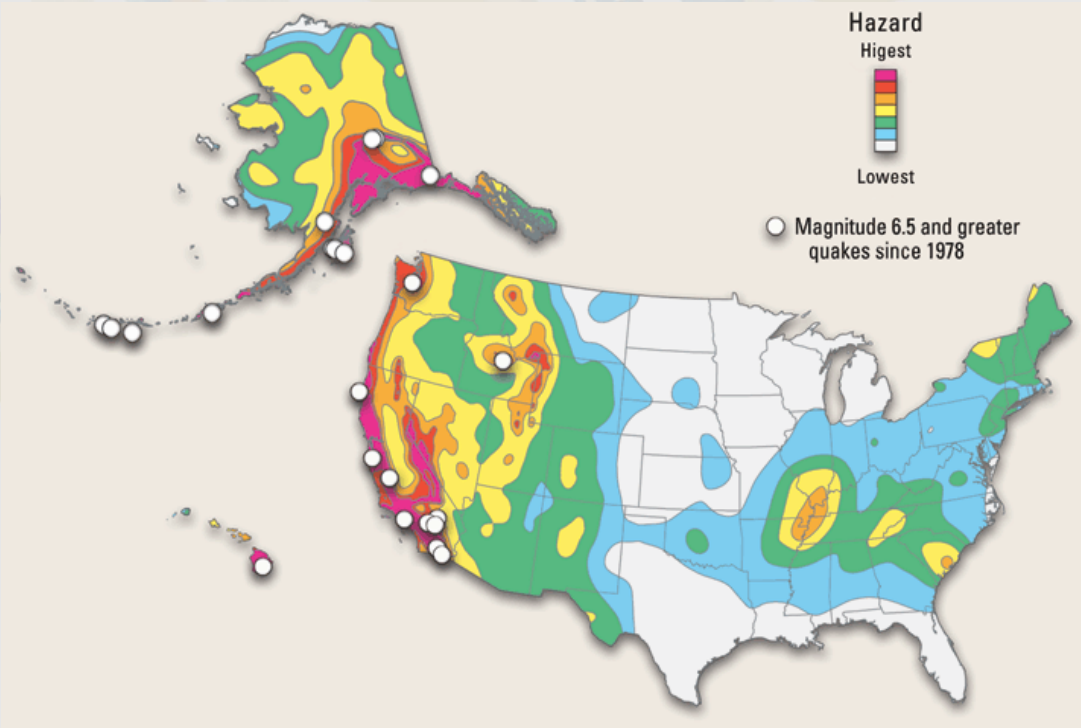
- ✓ Thousands of rearrangements hotspots known for different tumors.

Rearrangement Hotspots in Tumor Genomes

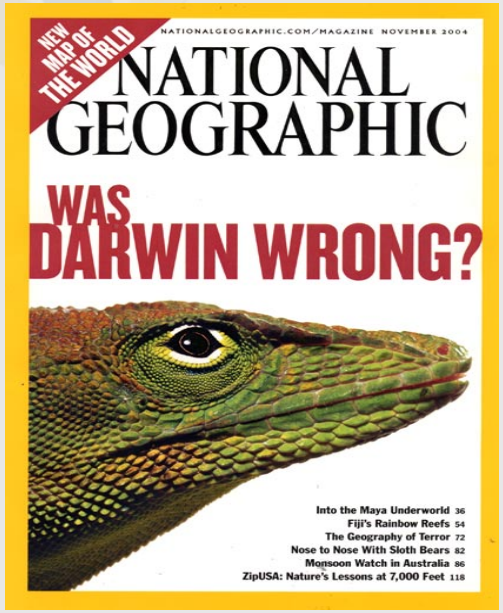
MCF7 breast cancer cell line



- ✓ What is the evolutionary scenario for transforming one genome into the other?
- ✓ What is the organization of the ancestral genome?
- ✓ **Where are the rearrangement hotspots in mammalian genomes?**



Controversy: Evolution vs. Intelligent Design



Three Evolutionary Controversies

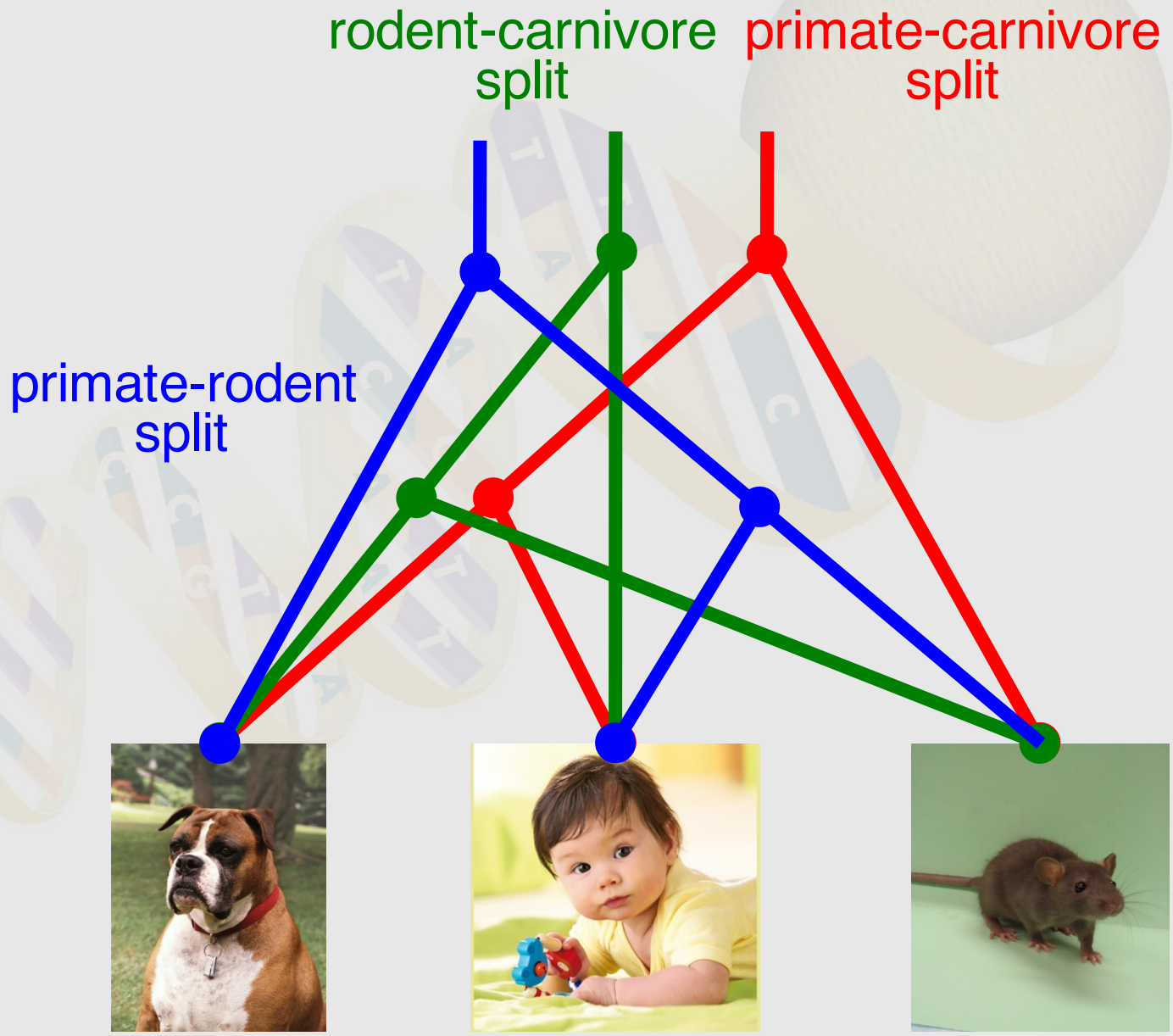
Primate - Rodent - Carnivore Split

Rearrangement Hotspots

Whole Genome Duplications

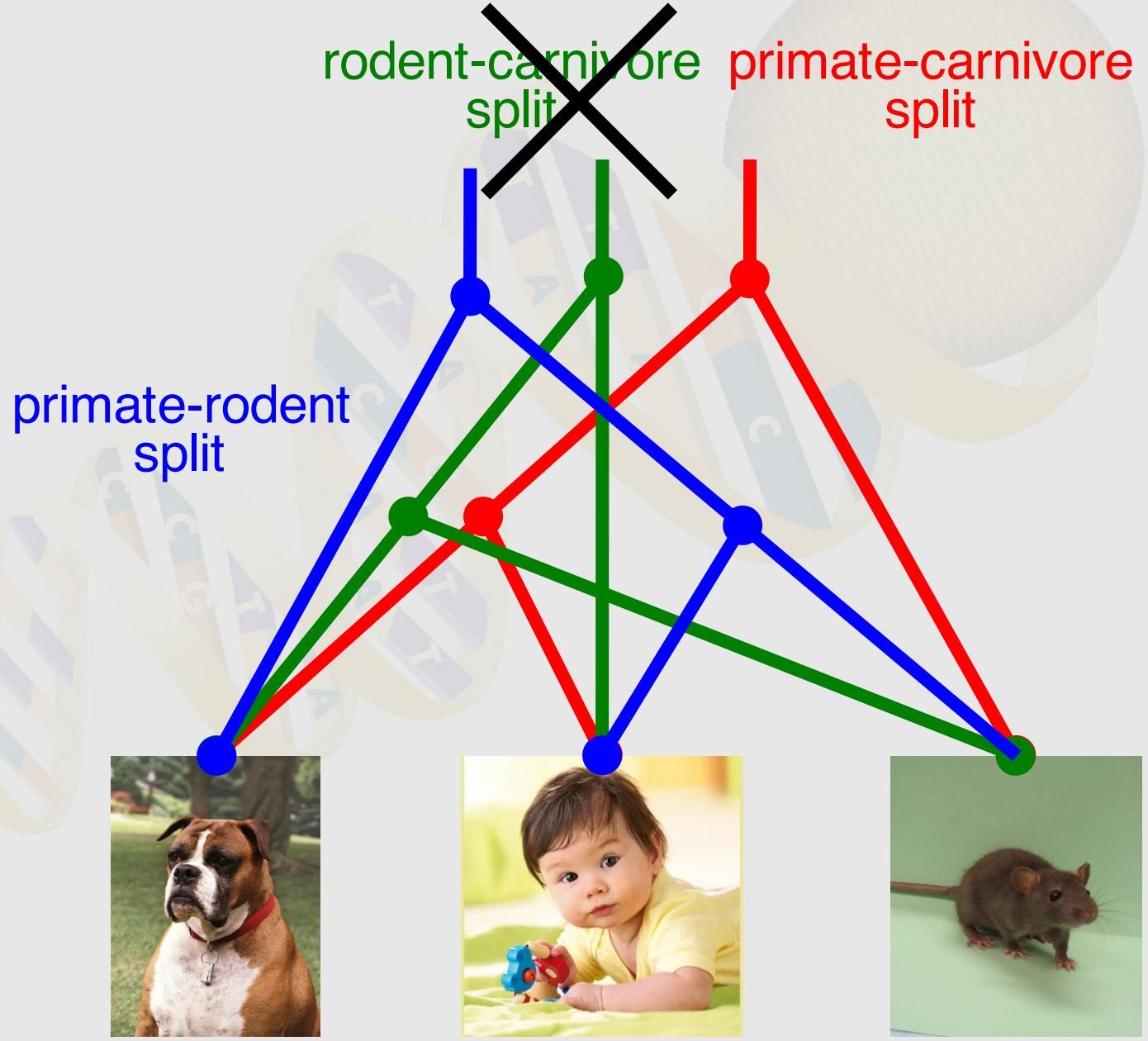
Primate – Rodent – Carnivore Split

Primate - Rodent - Carnivore Split



Primate – Rodent – Carnivore Split (who is "closer" to us: mouse or dog?)

Primate - Rodent - Carnivore Split



Primate-Rodent vs. Primate-Carnivore Split

Hutley et al., MBE, **May 07:**

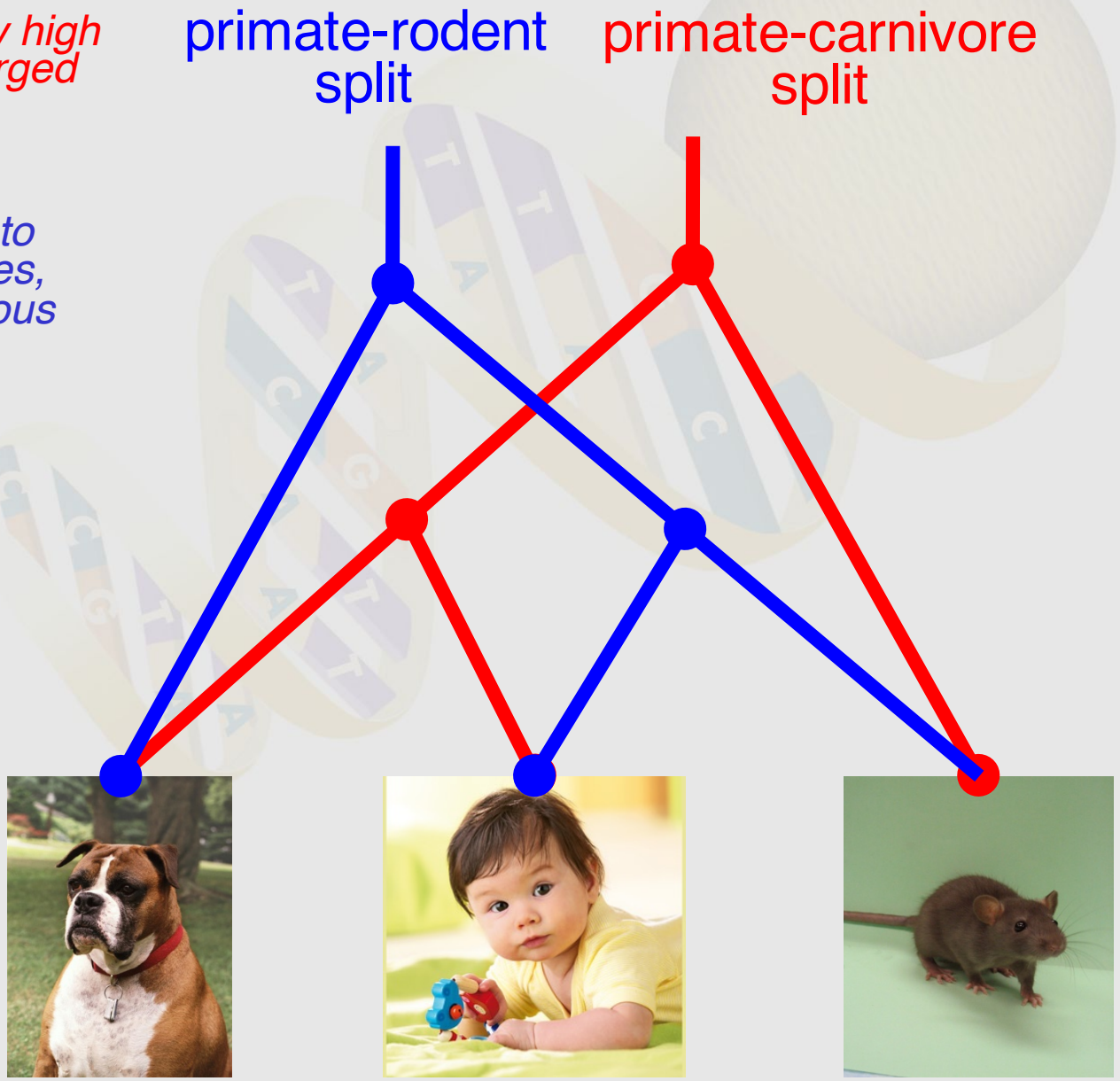
We have demonstrated with very high confidence that the rodents diverged before carnivores and primates

Lunter, PLOS CB, **April 07:**
It appears unjustified to continue to consider the phylogeny of primates, rodents, and canines as contentious

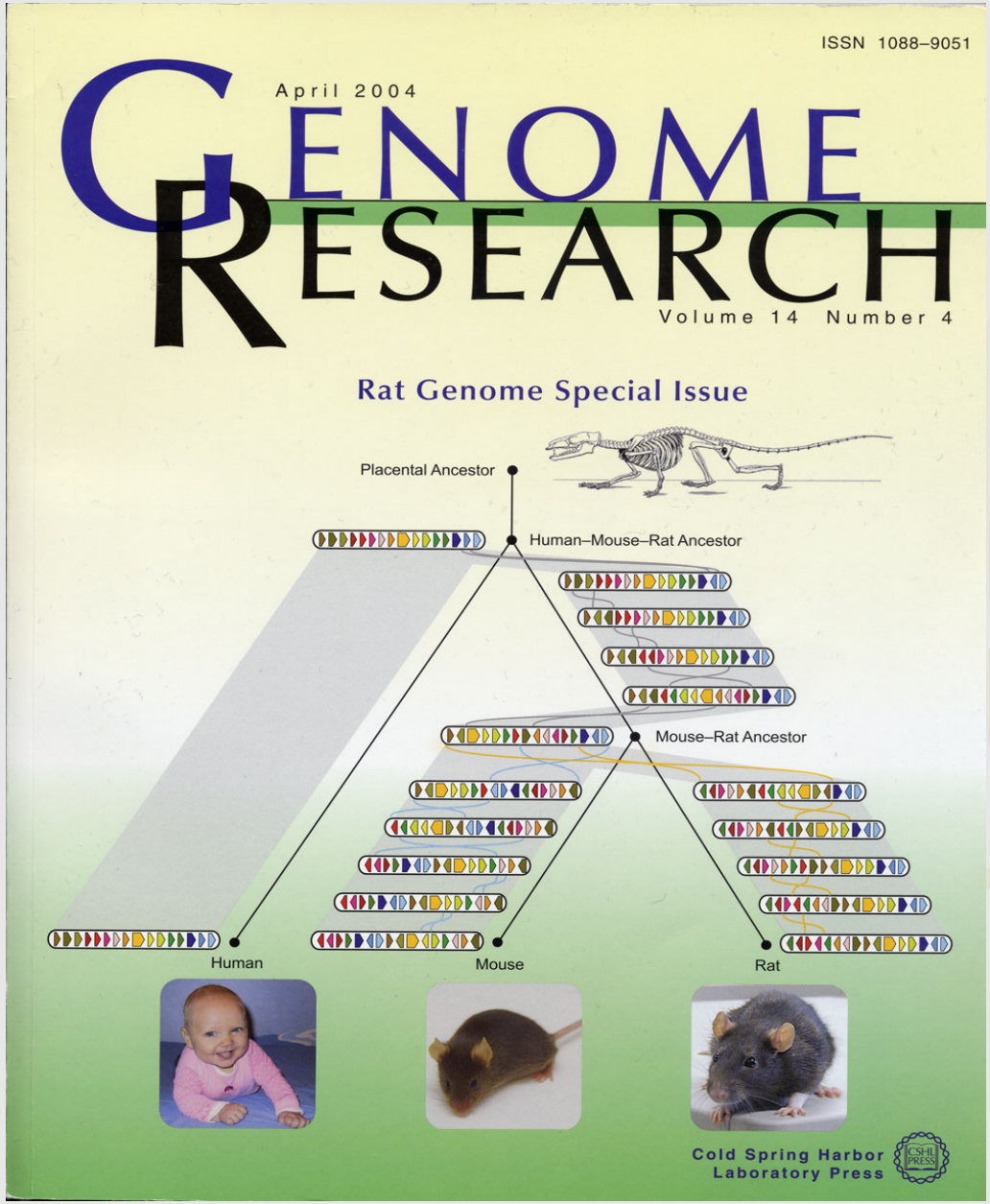
Arnason et al, PNAS 02
Canarozzi, PLOS CB 06

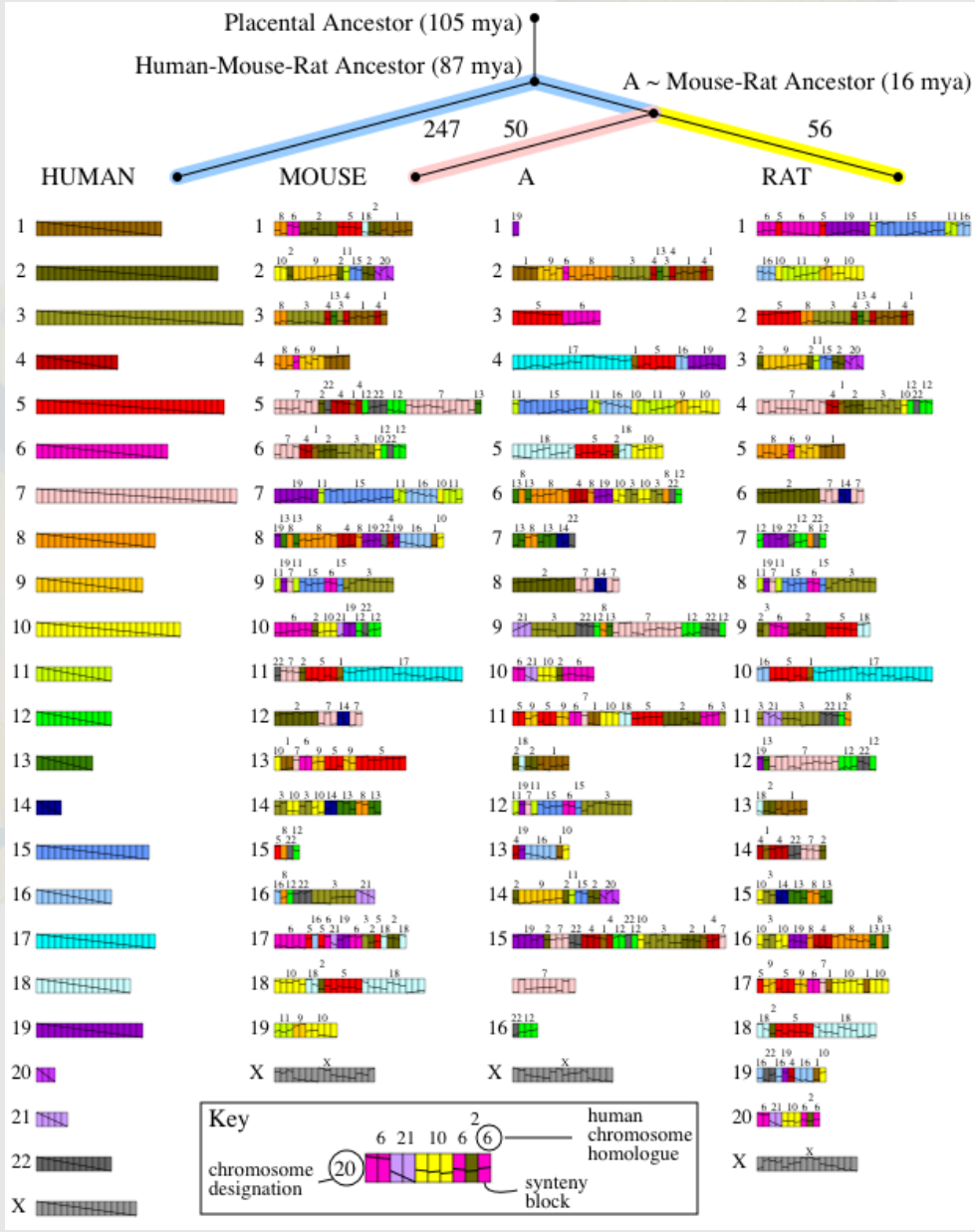
Murphy et al,, Science 01

Kumar&Hedges, Nature, 98

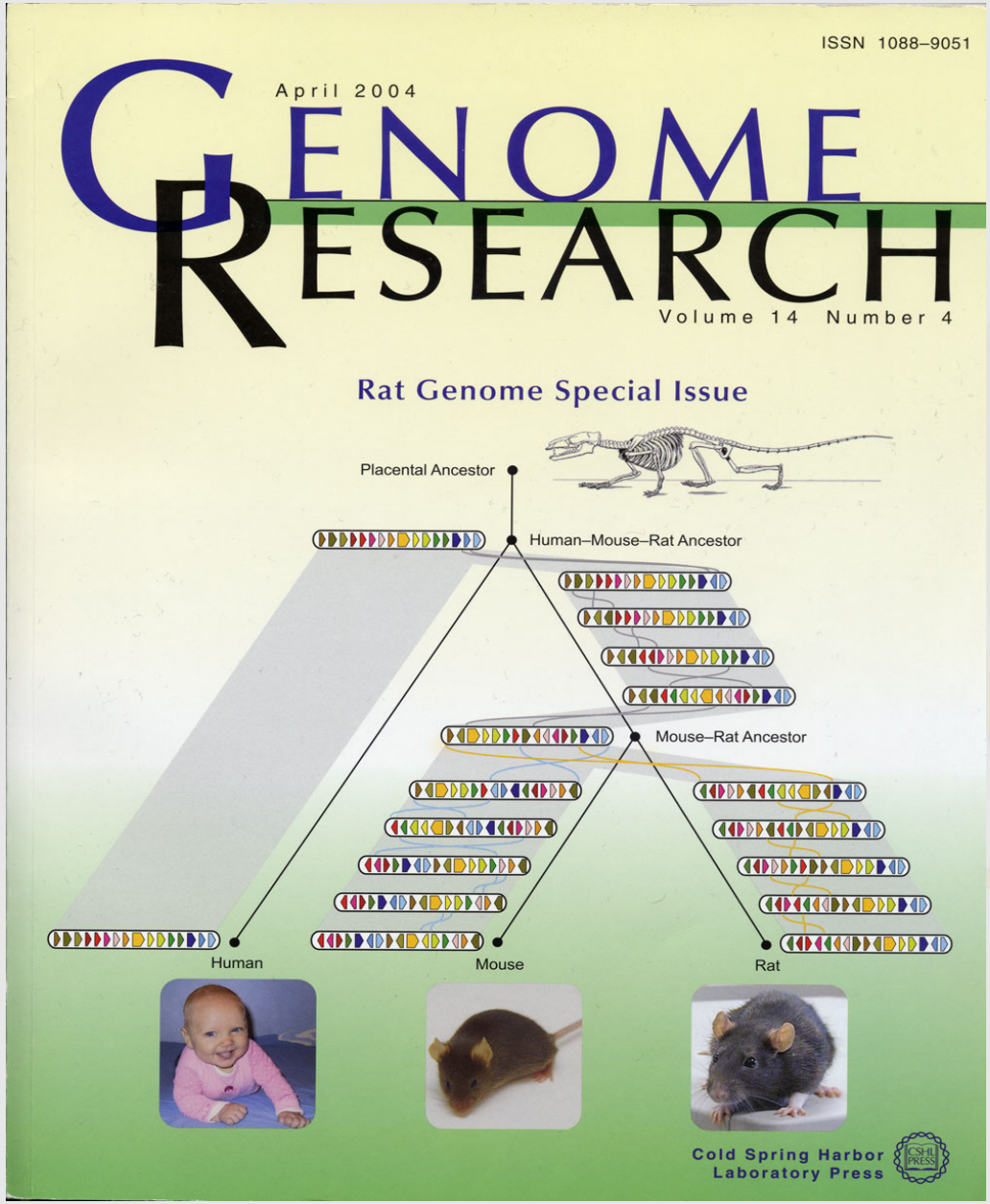


Reconstruction of Ancestral Genomes: Human / Mouse / Rat





Can Rearrangement Analysis Resolve the Primate-Rodent-Carnivore Controversy?



Susumu Ohno: Two Hypothesis

Ohno, 1970, 1973

Rearrangement Hotspots

- ✓ **Random Breakage Hypothesis:** Genomic architectures are shaped by rearrangements that occur randomly (no fragile regions).

Whole Genome Duplications

- ✓ **Whole Genome Duplication (WGD) Hypothesis:** Big leaps in evolution would have been impossible without whole genome duplications.

Random Breakage Model (RBM)

- ✓ The random breakage hypothesis was embraced by biologists and has become *de facto* theory of chromosome evolution.
- ✓ Nadeau & Taylor, *Proc. Nat'l Acad. Sciences 1984*
  - ✓ First convincing arguments in favor of the **Random Breakage Model (RBM)**
  - ✓ RBM implies that there is no rearrangement hotspots
  - ✓ RBM was re-iterated in hundreds of papers

Rearrangement Hotspots

Random Breakage Model (RBM)

Rearrangement Hotspots

- ✓ The random breakage hypothesis was embraced by biologists and has become *de facto* theory of chromosome evolution.
- ✓ **Nadeau & Taylor, *Proc. Nat'l Acad. Sci. 1984***
  - ✓ First convincing arguments in favor of the **Random Breakage Model (RBM)**
  - ✓ RBM implies that there is no rearrangement hotspots
  - ✓ RBM was re-iterated in hundreds of papers
- ✓ **PP & Tesler, *Proc. Nat'l Acad. Sci.* 2003**
  - ✓ Rejected RBM and proposed the **Fragile Breakage Model (FBM)**
  - ✓ Postulated existence of *rearrangement hotspots* and *vast breakpoint re-use*
  - ✓ FBM implies that the human genome is a mosaic of *solid* and *fragile* regions.

Are the Rearrangement Hotspots Real?

- ✓ **The Fragile Breakage Model did not live long:** in 2003 David Sankoff presented convincing arguments against FBM:

"… we have shown that breakpoint re-use of the same magnitude as found in Pevzner and Tesler, 2003 may very well be artifacts in a context where NO re-use actually occurred."

Are the Rearrangement Hotspots Real?

- ✓ **The Fragile Breakage Model did not live long:** in 2003 David Sankoff presented arguments against FBM (*Sankoff & Trinh, J. Comp. Biol, 2005*)

> *"… we have shown that breakpoint re-use of the same magnitude as found in Pevzner and Tesler, 2003 may very well be artifacts in a context where NO re-use actually occurred."*

Before you criticize people, you should walk a mile in their shoes. That way, when you criticize them, you are a mile away. And you have their shoes.

J.K. Lambert

Rebuttal of the Rebuttal of the Rebuttal

- Peng et al., 2006 (*PLOS Computational Biology*) found an error in Sankoff-Trinh rebuttal:

  …If Sankoff & Trinh fixed their ST-Synteny algorithm, they would confirm rather than reject Pevzner-Tesler's Fragile Breakage Model

- Sankoff, 2006 (*PLOS Computational Biology*):

  …Not only did we foist a hastily conceived and incorrectly executed simulation on an overworked RECOMB conference program committee, but worse — *nostra maxima culpa* — we obliged a team of high-powered researchers to clean up after us!

Controversy continues…

- **Random Breakage Model controversy:** While Sankoff acknowledged the flaw in his arguments against RBM, he appears reluctant to acknowledge the rebuttal of RBM, this time arguing that more complex rearrangement events (e.g., transpositions) may create an appearance of breakpoint re-use.

✓ Most recent studies support the Fragile Breakage Model: *van der Wind, 2004, Bailey, 2004, Zhao et al., 2004, Murphy et al., 2005, Hinsch & Hannenhalli, 2006, Ruiz-Herrera et al., 2006, Yue & Haaf, 2006, Mehan et al., 2007, etc*

Kikuta et al., Genome Res. 2007*: "… the Nadeau and Taylor hypothesis is not possible for the explanation of synteny in rat."*

Rebuttal of the Rebuttal of the Rebuttal: Controversy Resolved... Or Was It?

- ✓ Nearly all recent studies support the Fragile Breakage Model: *van der Wind, 2004, Bailey, 2004, Zhao et al., 2004, Murphy et al., 2005, Hinsch & Hannenhalli, 2006, Ruiz-Herrera et al., 2006, Yue & Haaf, 2006, Mehan et al., 2007, etc*

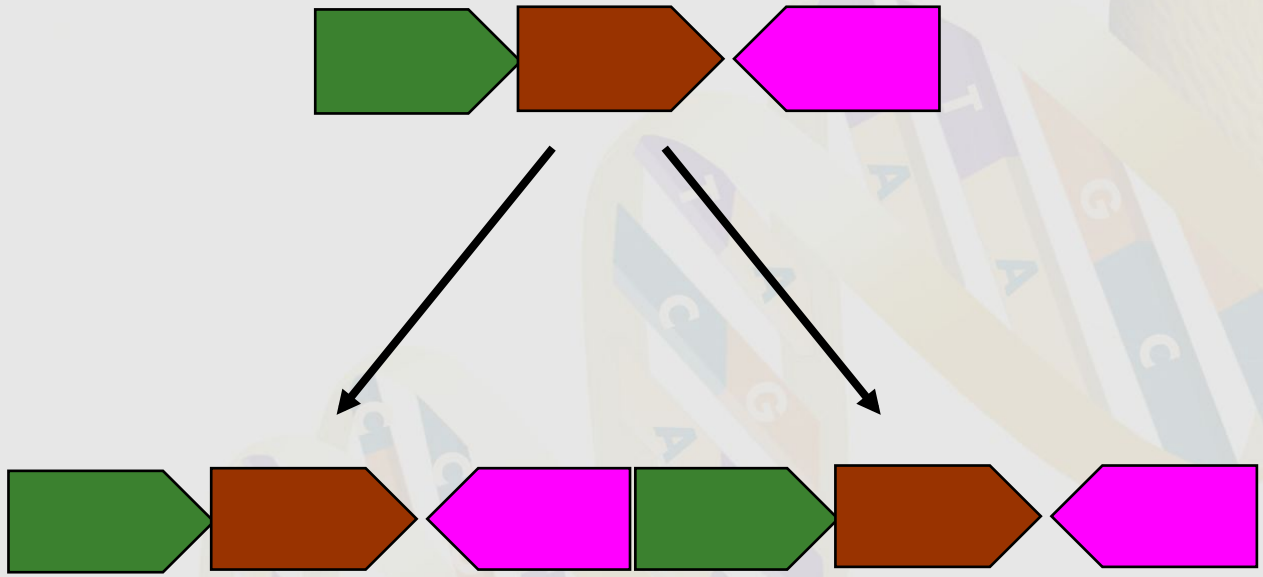
  Kikuta et al., Genome Res. 2007*: "... the Nadeau and Taylor hypothesis is not possible for the explanation of synteny"*

Random vs. Fragile Breakage Debate Continues: Complex Rearrangements

- ✓ *PP & Tesler, PNAS 2003,* argued that every evolutionary scenario for transforming *Mouse* into *Human* genome with **reversals, translocations, fusions, and fissions** must result in a large number of *breakpoint re-uses*, a contradiction to the RBM.

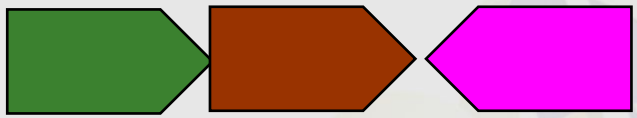
- ✓ *Sankoff, PLoS Comp. Biol. 2006: "We cannot infer whether mutually randomized synteny block orderings derived from two divergent genomes were created ... through processes other than reversals and translocations."*

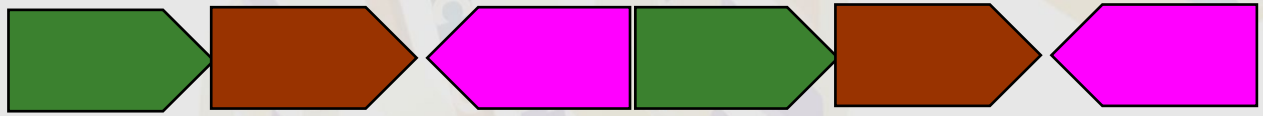
Whole Genome Duplication (WGD)



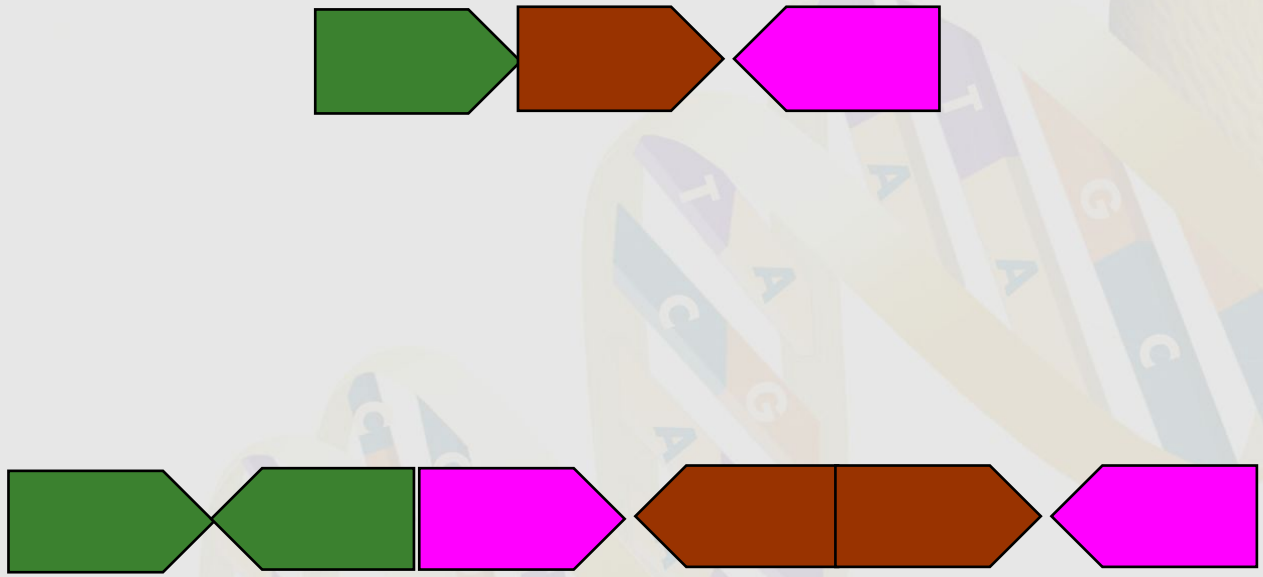
Whole Genome Duplications

Genome Rearrangements After WGD

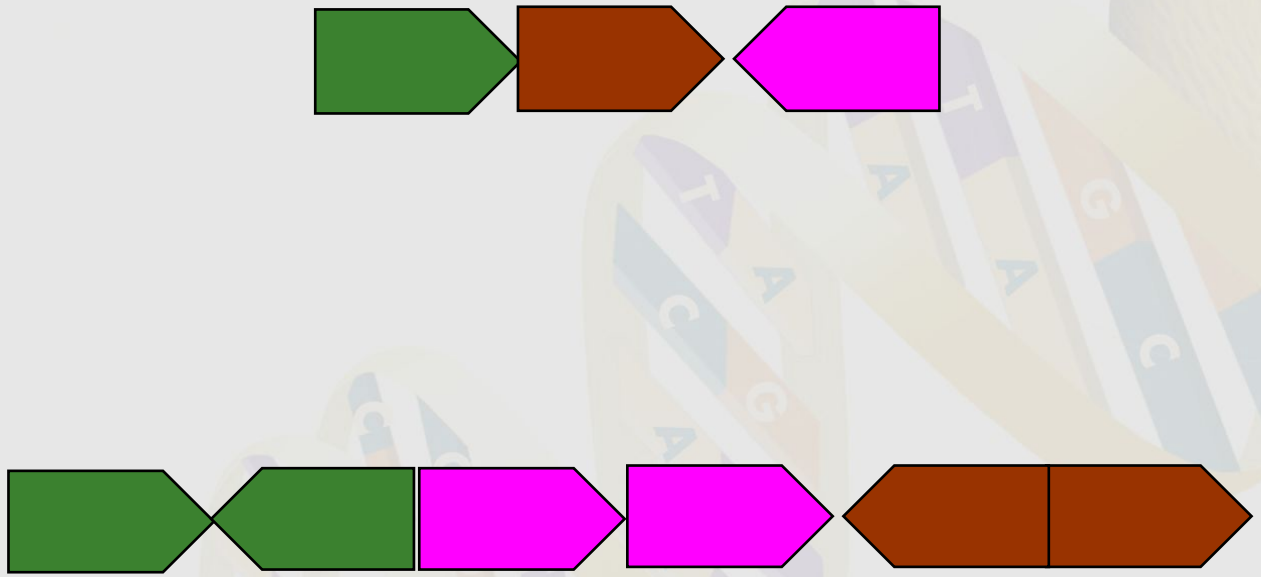




Genome Rearrangements After WGD



Genome Rearrangements After WGD



Whole Genome Duplication Hypothesis Was Confirmed After Years of Controversy

- ✓ **The Whole Genome Duplication** hypothesis first met with skepticism but was finally confirmed by Kellis et al., *Nature 2004: "Our analysis resolves the long-standing controversy on the ancestry of the yeast genome."*

- "There was a whole-genome duplication." **Wolfe, *Nature*, 1997**
- "There was no whole-genome duplication." **Dujon, *FEBS*, 2000**
- "Duplications occurred independently" **Langkjaer, *JMB*, 2000**
- "Continuous duplications" **Dujon, *Yeast* 2003**
- "Multiple duplications" **Friedman, *Gen. Res*, 2003**
- "Spontaneous duplications" **Koszul, *EMBO*, 2004**

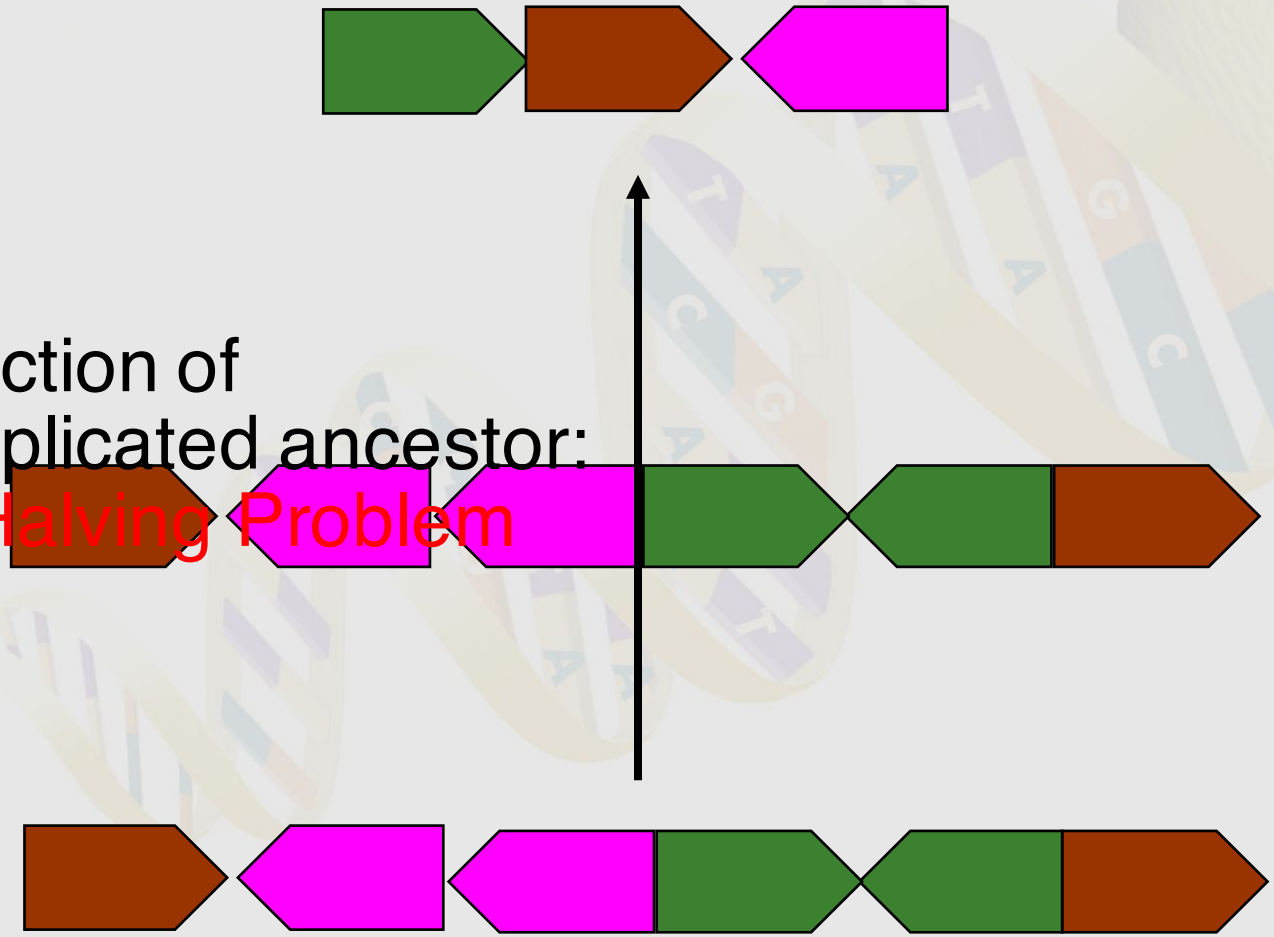
Whole Genome Duplication Hypothesis Was Eventually Confirmed...But Some Scientists Are Not Convinced

- ✓ Kellis, Birren & Lander, *Nature 2004: "Our analysis resolves the long-standing controversy on the ancestry of the yeast genome."*
- ✓ Martin et al., *Biology Direct 2007: "We believe that the proposal of a Whole Genome Duplication in the yeast lineage is unwarranted."*
- ✓ **To address Martin et al., 2007 arguments against WGD, it would be useful to reconstruct the pre-duplicated genomes.**

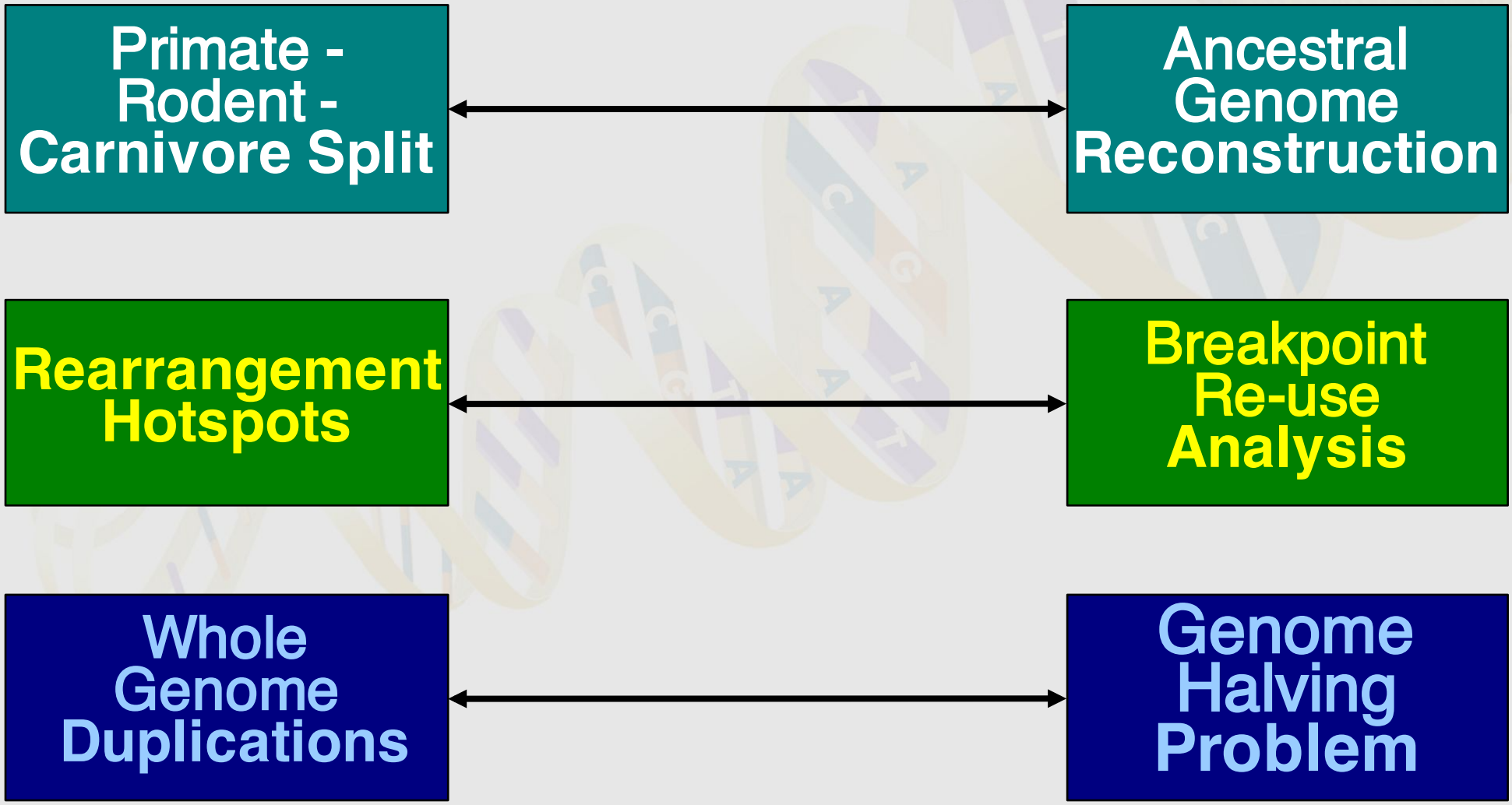
- "There was a whole-genome duplication." **Wolfe, *Nature*, 1997**
- "There was no whole-genome duplication." **Dujon, *FEBS*, 2000**
- "Duplications occurred independently" **Langkjaer, *JMB*, 2000**
- "Continuous duplications" **Dujon, *Yeast* 2003**
- "Multiple duplications" **Friedman, *Gen. Res*, 2003**
- "Spontaneous duplications" **Koszul, *EMBO*, 2004**
- "Our analysis resolved the controversy" **Kellis, *Nature, 2004***
- "WGD in the yeast lineage is unwarranted" **Martin, *Biology Direct, 2007***
- ...

Genome Halving Problem

Reconstruction of
the pre-duplicated ancestor:
Genome Halving Problem



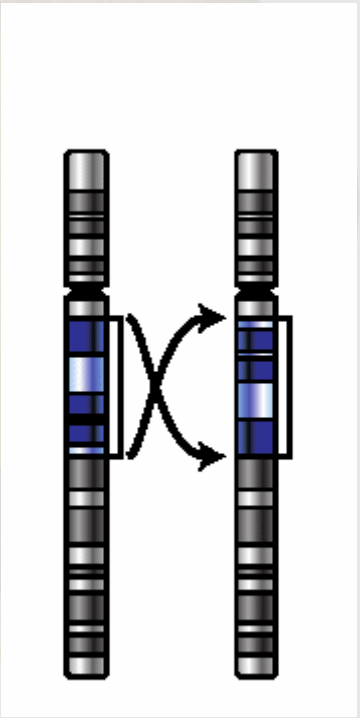
From Biological Controversies to Algorithmic Problems



Algorithmic Background:

Genome Rearrangements and Breakpoint Graphs

Unichromosomal Genomes: Reversal Distance



- **Sorting by reversals:** find the shortest series of reversals transforming one uni-chromosomal genome into another.
- The number of reversals in such a shortest series is the **reversal distance** between genomes.
- Hannenhalli and PP. (*FOCS 1995*) gave a polynomial algorithm for computing the reversal distance.

Step 0:	2	-4	-3	5	-8	-7	-6	1
Step 1:	2	3	4	5	-8	-7	-6	1
Step 2:	-5	-4	-3	-2	-8	-7	-6	1
Step 3:	-5	-4	-3	-2	-1	6	7	8
Step 4:	1	2	3	4	5	6	7	8

Sorting by prefix reversals (pancake flipping problem), Gates and Papadimitriou, Discrete Appl. Math. 1976

Sorting by reversals

Most parsimonious scenarios

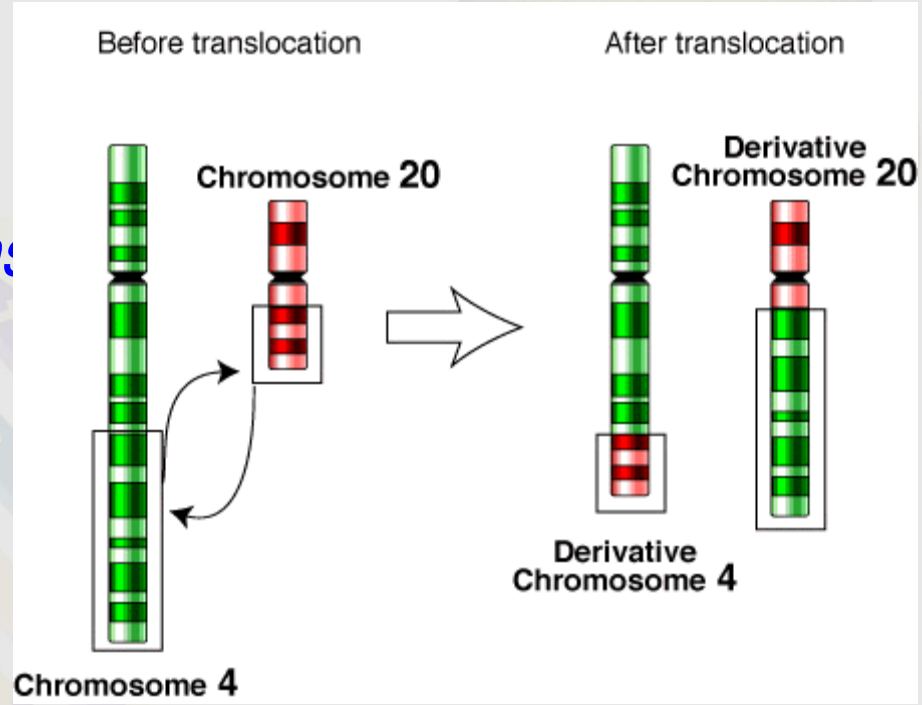
Step 0:	2	-4	-3	5	-8	-7	-6	1
Step 1:	2	3	4	5	-8	-7	-6	1
Step 2:	-5	-4	-3	-2	-8	-7	-6	1
Step 3:	-5	-4	-3	-2	-1	6	7	8
Step 4:	1	2	3	4	5	6	7	8

The *reversal distance* is the minimum number of reversals required to transform one gene order into another.

Here, the distance is 4.

Multichromosomal Genomes: Genomic Distance

- **Genomic Distance** between two genomes is the minimum number of *reversals*, *translocations*, *fusions* and *fissions* required to transform one genome into another.



- Hannenhalli and PP (STOC 1995) gave a polynomial algorithm for computing the genomic distance.

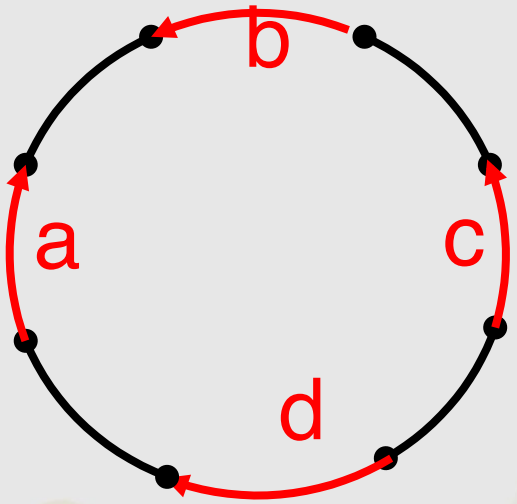
- These algorithms were followed by many improvements: *Kaplan et al. 1999, Bader et al. 2001, Tesler 2002, Ozery-Flato & Shamir 2003, Tannier & Sagot 2004, Bergeron 2001-07, etc.*

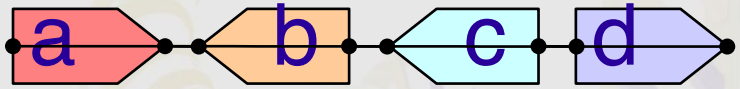
HP Theory Is Rather Complicated: Is There a Simpler Alternative?

- ✓ HP theory is a key tool in most genome rearrangement studies. However, it is rather complicated making it difficult to apply it in complex setups such as the *"RBM vs. FBM"* or *WGD* controversies.

- ✓ To study the outlined evolutionary controversies, we use an alternative (simpler) approach called
  ***k*-break analysis**

Simplifying HP Theory: from Linear to Circular Chromosomes



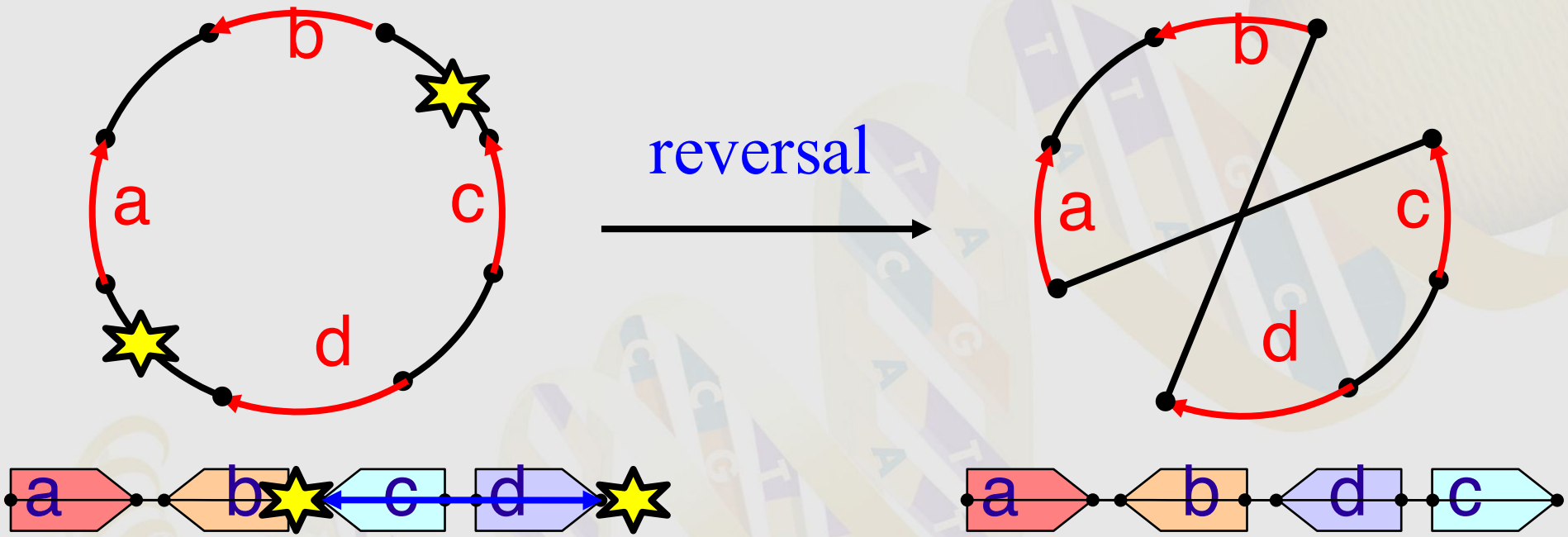


A chromosome is represented as a *cycle* with *directed red* and ***undirected black*** edges:

red edges encode blocks (genes)

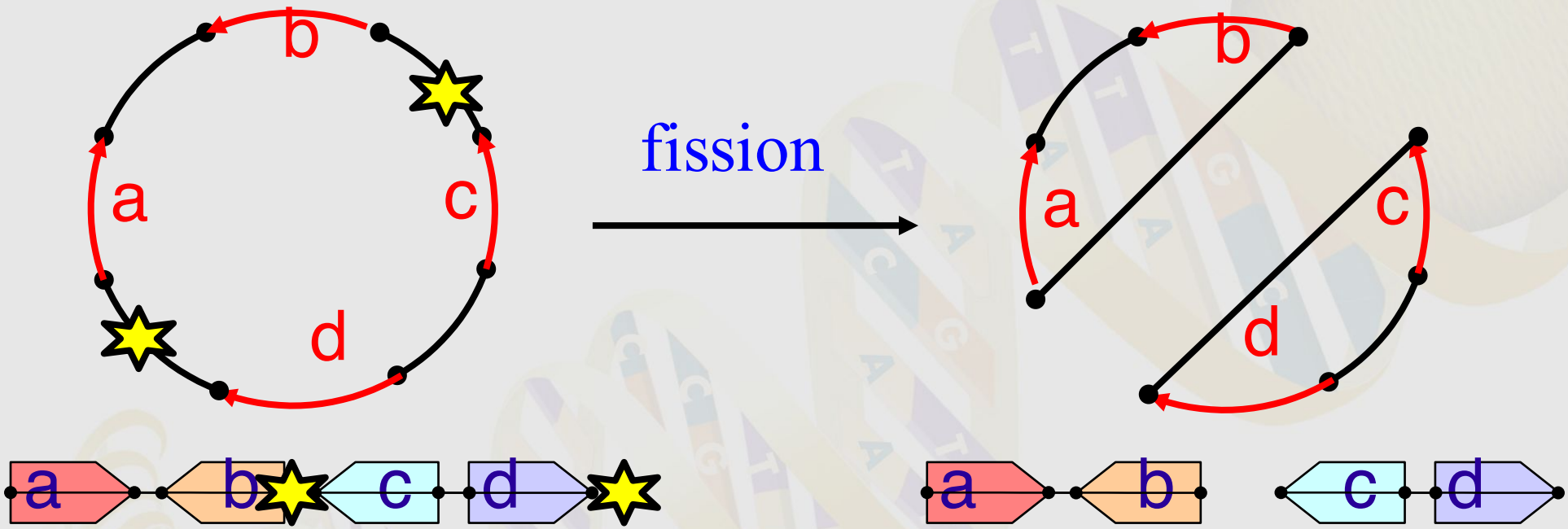
black edges connect adjacent blocks

Reversals on Circular Chromosomes



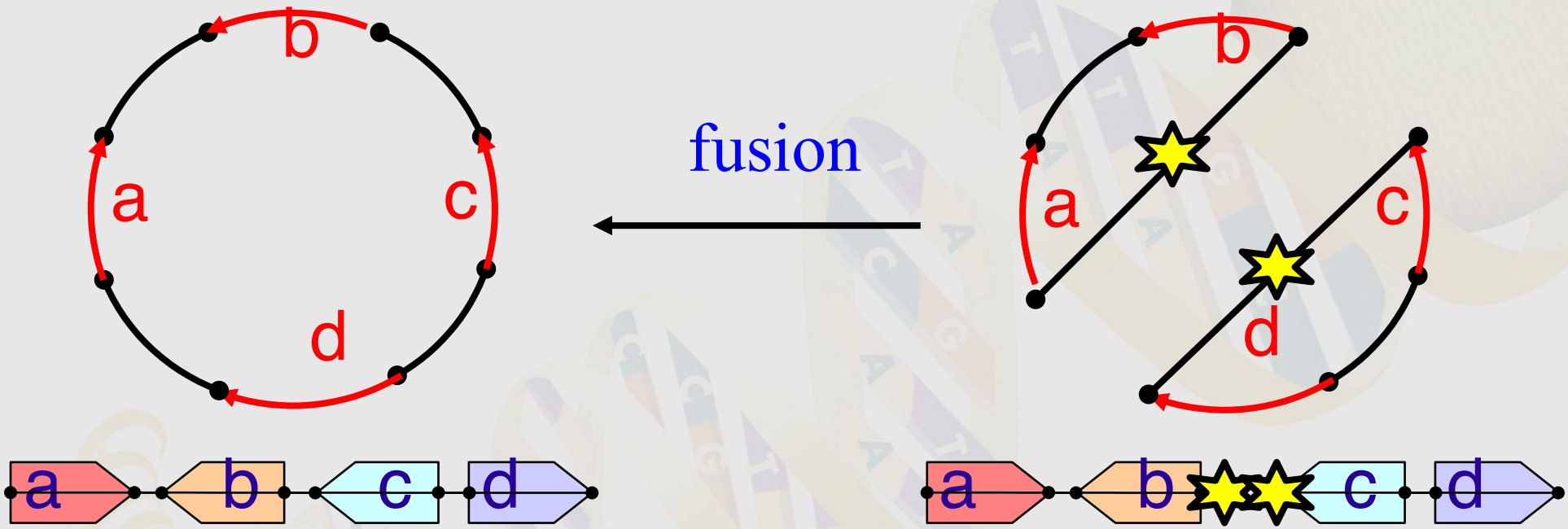
A *reversal* replaces two black edges with two other black edges

Fissions



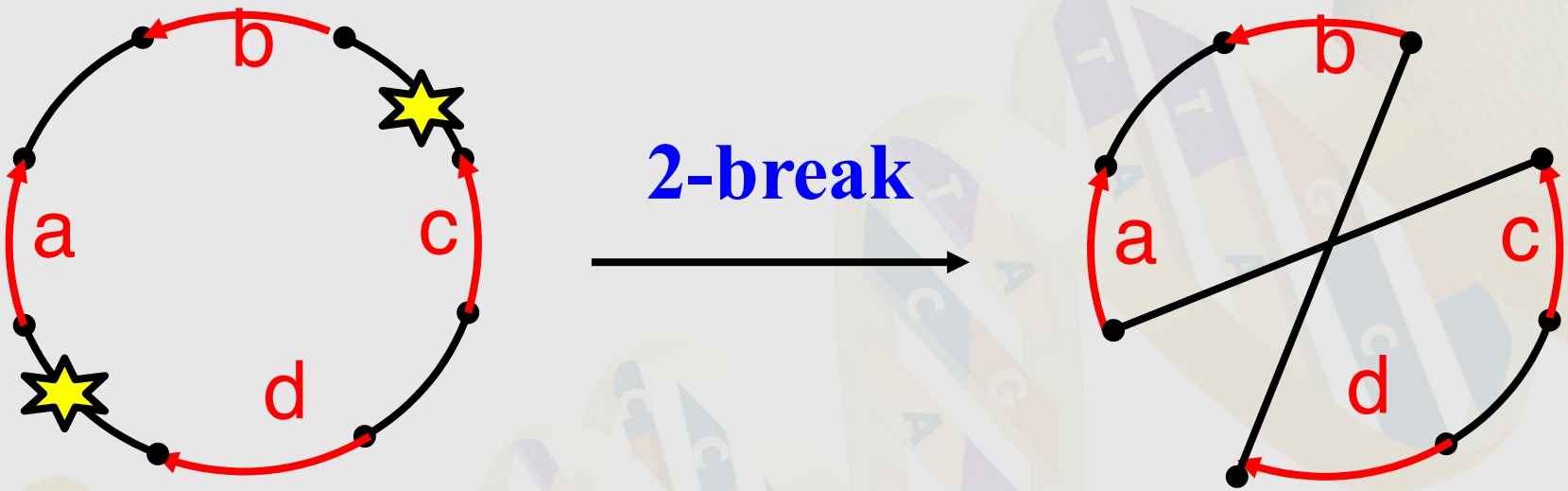
- ✓ A *fission* splits a single cycle (chromosome) into two.
- ✓ A fissions replaces two black edges with two other black edges.

Translocations / Fusions



- ✓ A translocation/fusion merges two cycles (chromosomes) into a single one.
- ✓ They also replace two black edges with two other black edges.

2-Breaks

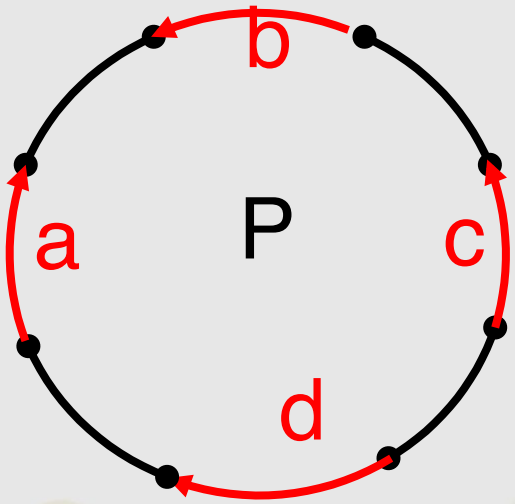


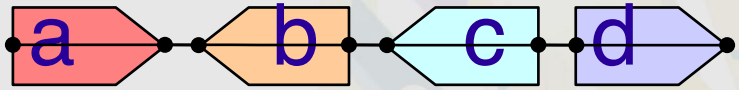
- ✓ A **2-Break** replaces *any* 2 black edges with another 2 black edges forming matching on the same 4 vertices.
- ✓ Reversals, translocations, fusions, and fissions represent all possible types of 2-breaks (introduced as DCJ operations by Yancopoulos et al., 2005)

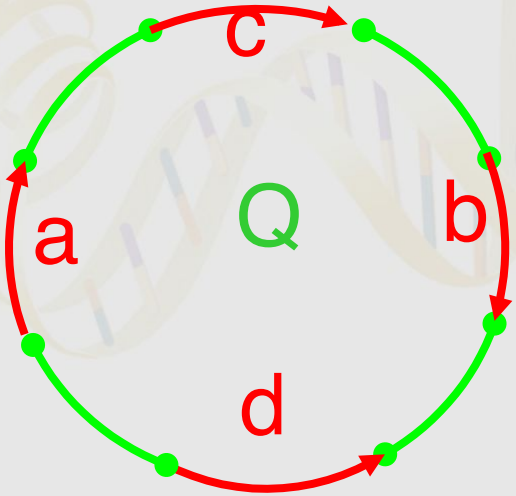
2-Break Distance

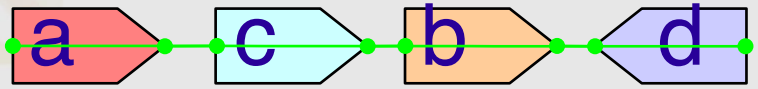
- ✓ The **2-break distance $d_2(P,Q)$** is the minimum number of 2-breaks required to transform ***P*** into ***Q***.
- ✓ In contrast to the genomic distance, the 2-break distance is easy to compute.

Two Genomes as Black-Red and Green-Red Cycles

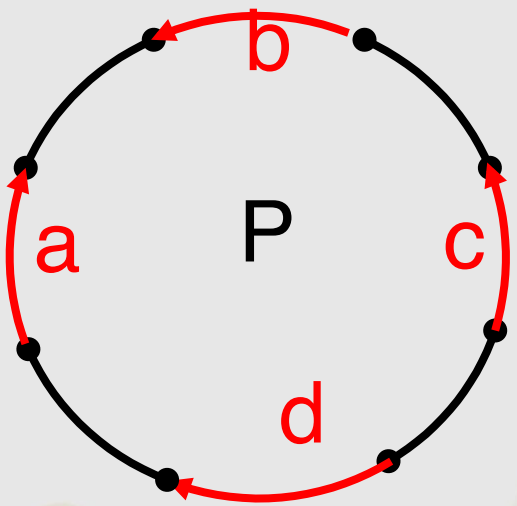


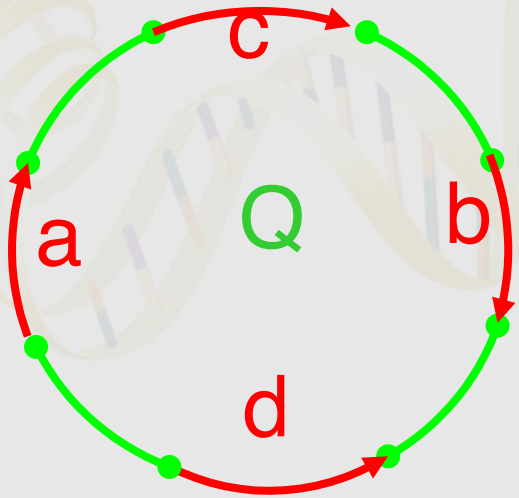


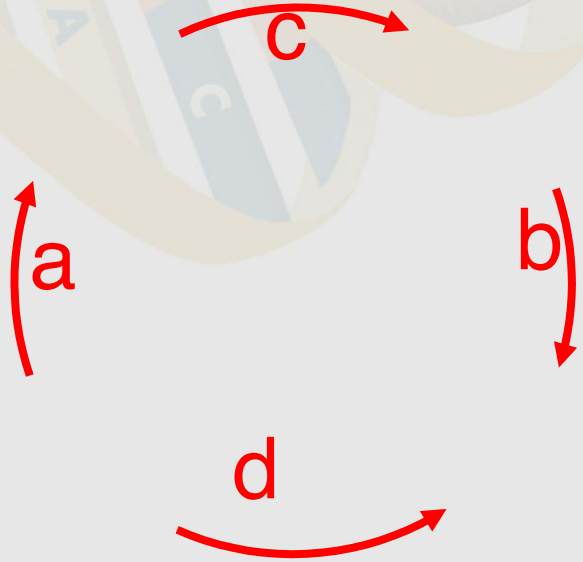




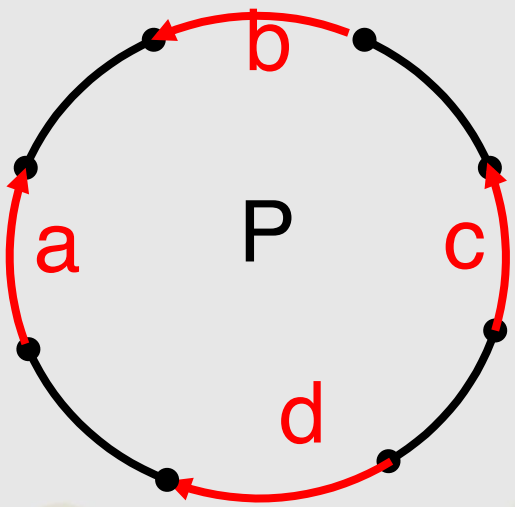
"Q-style" representation of P

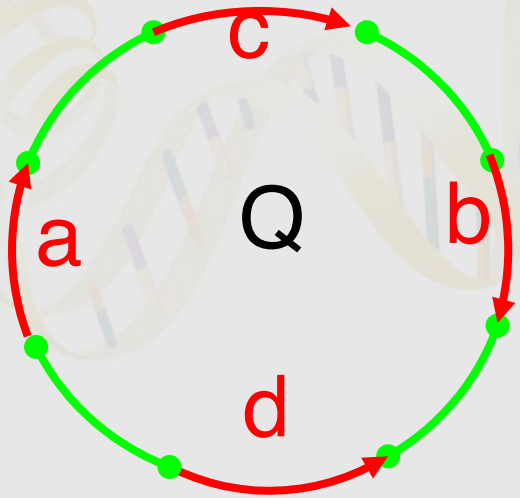


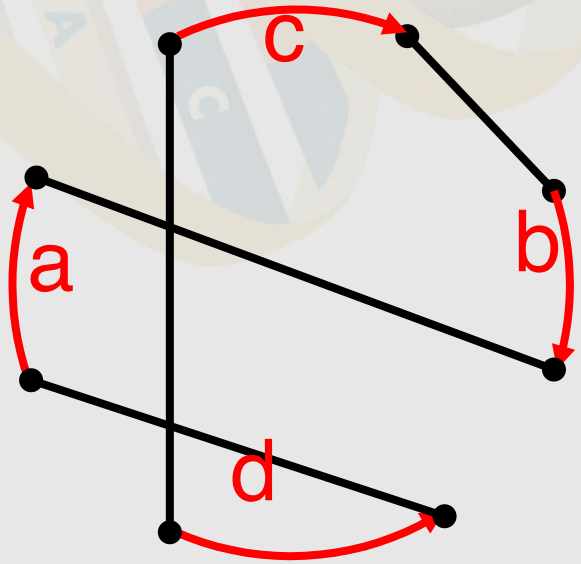




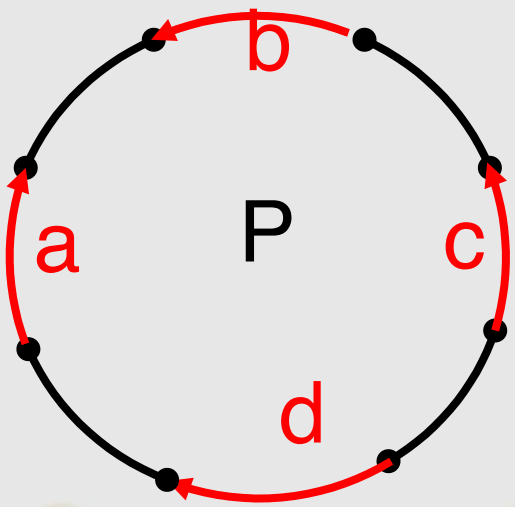
"Q-style" representation of P

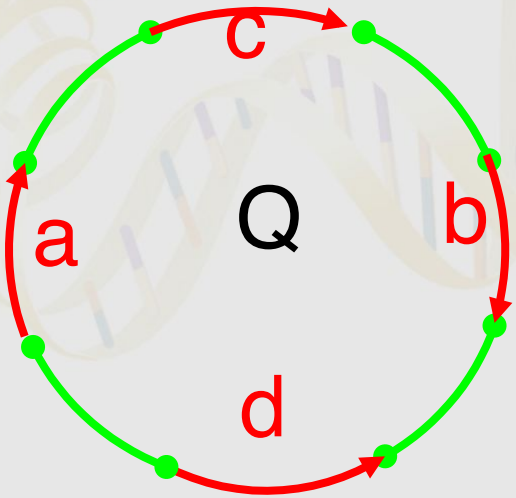






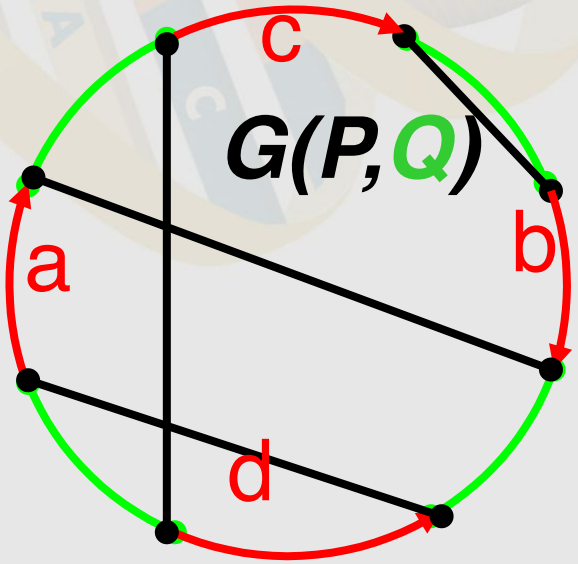
Breakpoint Graph: Superposition of Genome Graphs



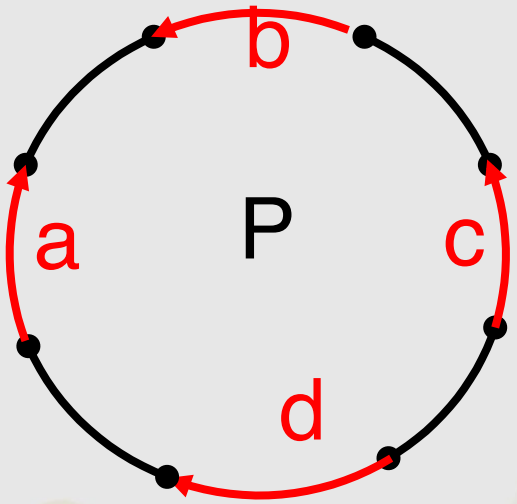


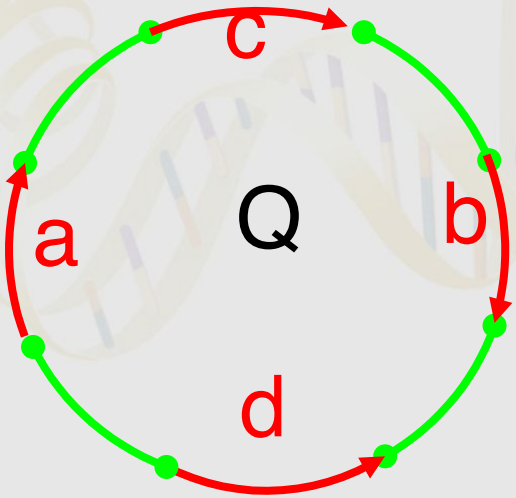
Breakpoint Graph

(Bafna & PP, FOCS 1994)



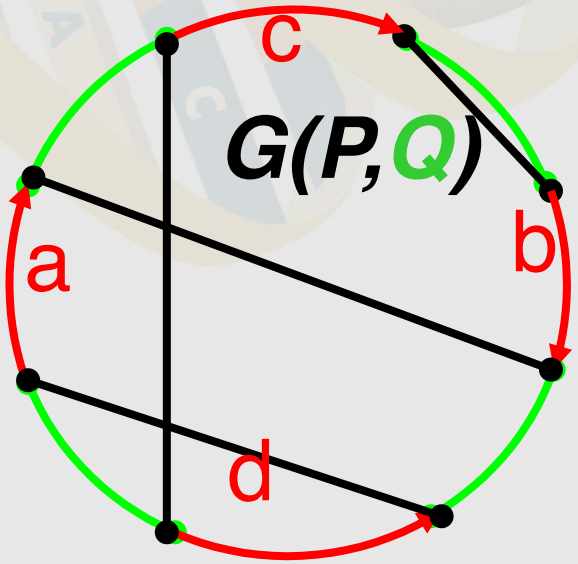
Breakpoint Graph:
GLUING Red Edges with the Same Labels





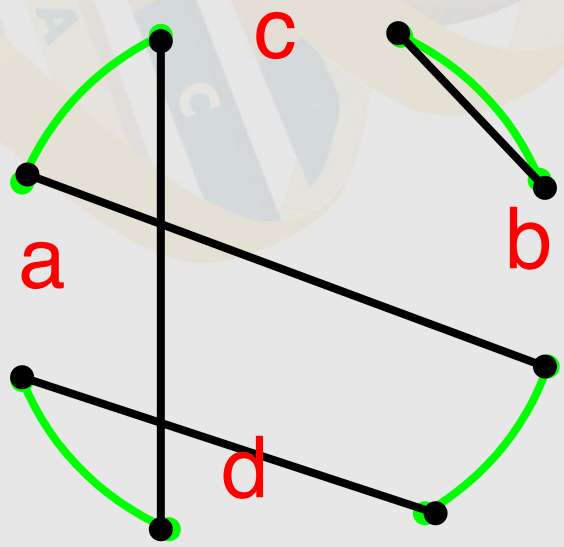
Breakpoint Graph

(Bafna & PP, FOCS 1994)



Black-Green Alternating Cycles

- Black and green edges form *perfect matchings* in the breakpoint graph. Therefore, these edges form a collection of **black-green alternating cycles**

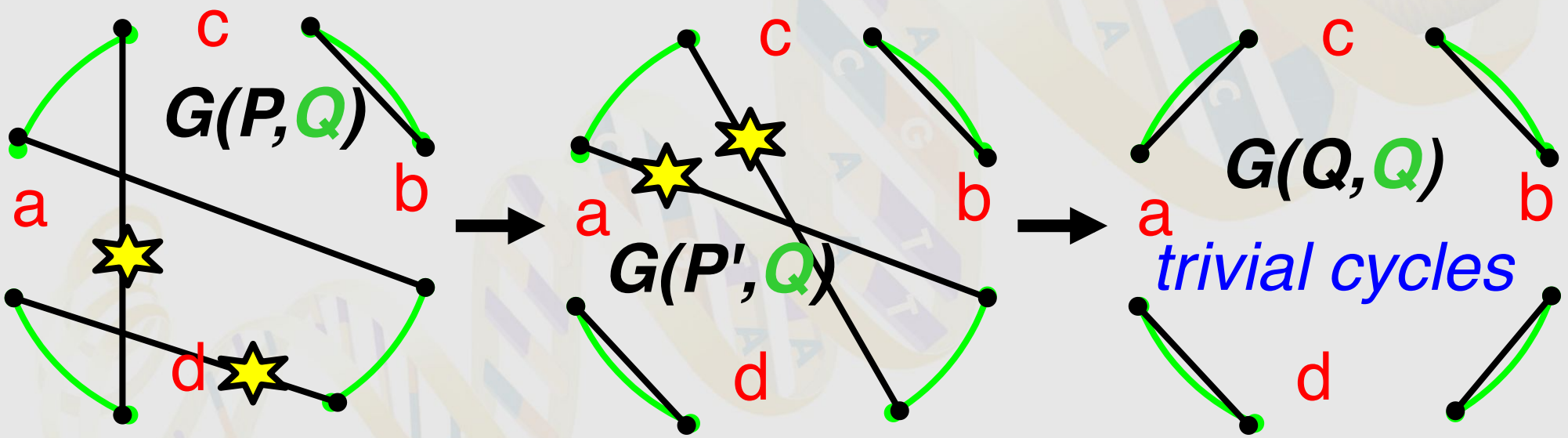


- Let ***cycles(P,Q)*** be the number of black-green cycles.

cycles(P,Q)=2

Rearrangements Change Breakpoint Graphs and cycle(P, Q)

Transforming genome ***P*** into genome ***Q*** by 2-breaks corresponds to transforming the breakpoint graph ***G(P, Q)*** into the *identity breakpoint graph* ***G(Q, Q)***.



cycles(P, Q) = 2 ***cycles(P', Q) = 3*** ***cycles(Q, Q) = 4***

=#blocks

Sorting by 2-Breaks

$$P = P_0 \xrightarrow{\text{2-breaks}} P_1 \rightarrow \ldots \rightarrow P_d = Q$$

$$G(P,Q) \rightarrow G(P_1,Q) \rightarrow \ldots \rightarrow G(Q,Q)$$

cycles(P,Q) cycles → ... → ***#blocks*** cycles

# of black-green cycles increased by
#blocks - cycles(P,Q)

How much each 2-break can contribute to the increase in the number of cycles?

Each 2-Break Increases #Cycles by at Most 1

A 2-break:

- ✓ adds 2 new black edges and thus **creates** at most **2 new** cycles (containing two new black edges)
- ✓ removes 2 black edges and thus **destroys** at least **1 old** cycle (containing two old edges):

  change in the number of cycles: ***Δcycles ≤ 2-1=1.***

2-Break Distance

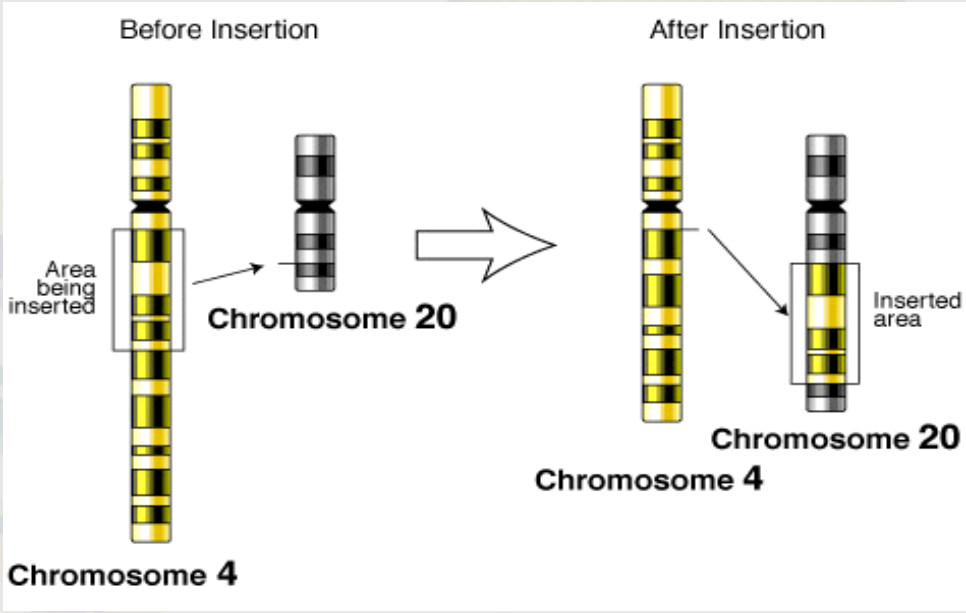
- ✓ Any 2-break increases the number of cycles by at most one (***Δcycles ≤ 1***)
- ✓ Any non-trivial cycle can be split into two cycles with a 2-break (***Δcycles = 1***)
- ✓ Every sorting by 2-breaks must increase the number of cycles by ***#blocks - cycles(P,Q)***
- ✓ The **2-break distance** between genomes ***P*** and ***Q:***

$$d_2(P,Q) = \#blocks - cycles(P,Q)$$

(cp. *Yancopoulos et al., 2005, Bergeron et al., 2006*)

Complex Rearrangements: Transpositions

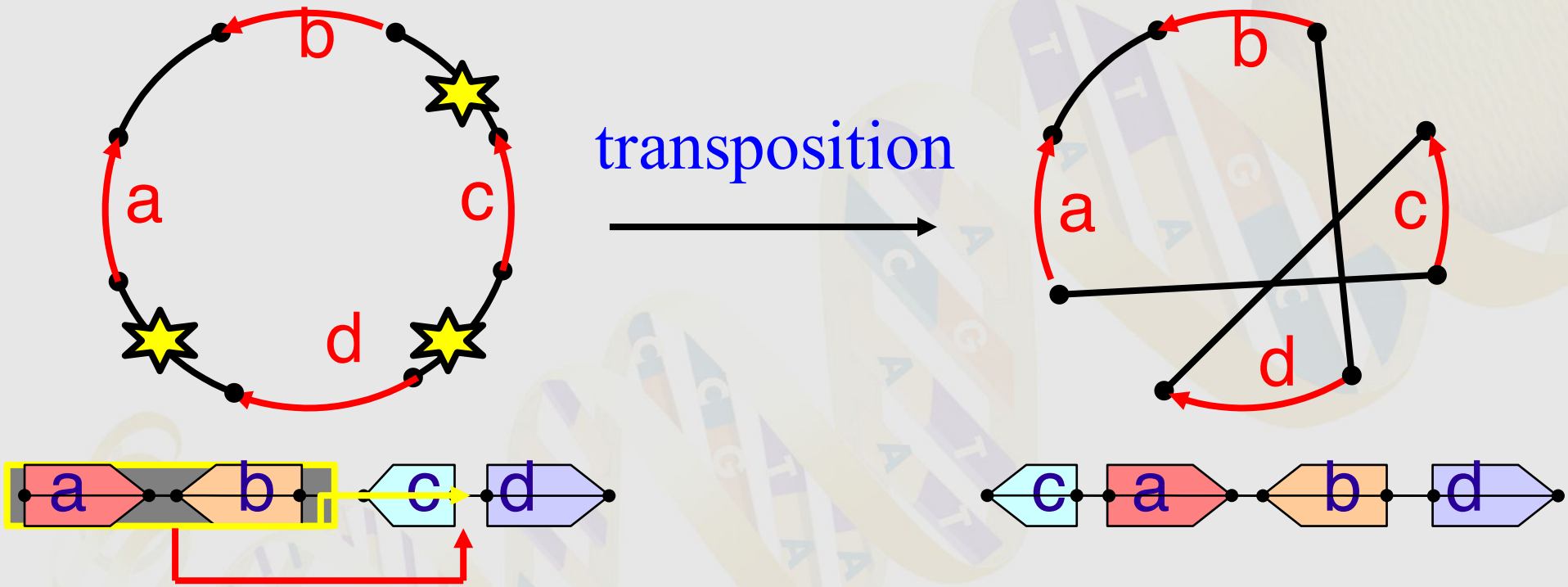
- **Sorting by Transpositions Problem:** find the shortest sequence of transpositions transforming one genome into another.



- First 1.5-approximation algorithm was given by Bafna and P.P. *SODA 1995*. The best known result: 1.375-approximation algorithm of Elias and Hartman, *2005*.

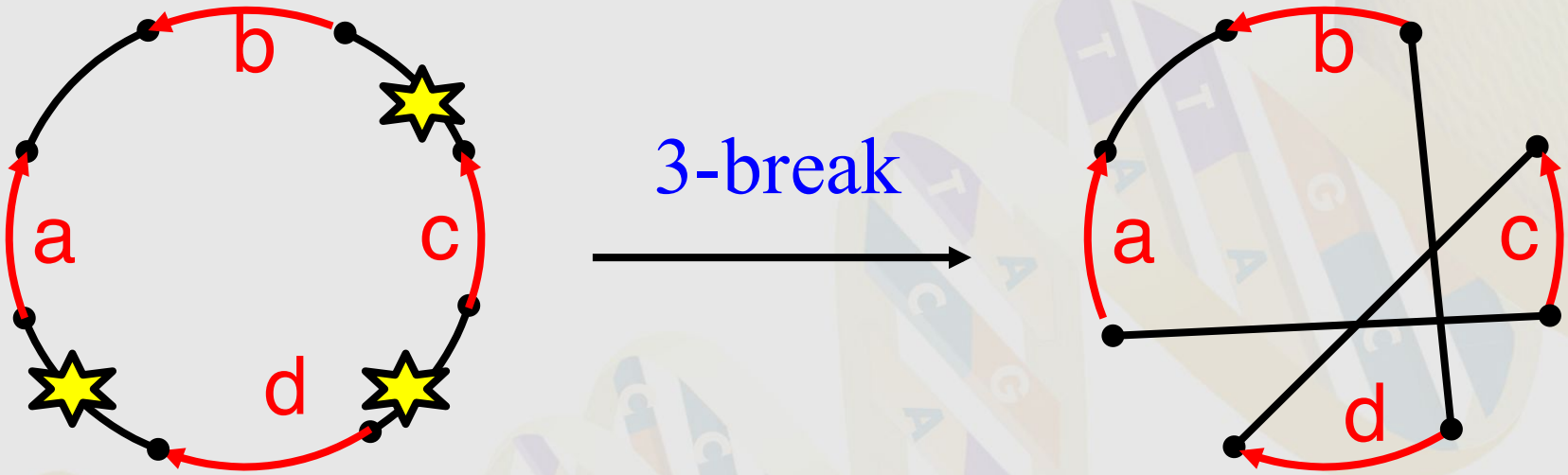
- The complexity status remains unknown.

Transpositions



Transpositions cut off a segment of one chromosome and insert it at some position in the same or another chromosome

Transpositions Are 3-Breaks



- ✓ 3-Break replaces *any triple* of black edges with another triple forming matching on the same 6 vertices.
- ✓ **Transpositions are 3-Breaks.**

Sorting by 3-Breaks

$$P=P_0 \xrightarrow{\text{3-breaks}} P_1 \rightarrow \ldots \rightarrow P_d = Q$$

$$G(P,Q) \rightarrow G(P_1,Q) \rightarrow \ldots \rightarrow G(Q,Q)$$

cycles(P,Q) cycles → ... → ***#blocks*** cycles

# of black-green cycles increased by
#blocks - cycles(P,Q)

How much each 3-break can contribute to this increase?

Each 3-Break Increases #Cycles by at Most 2

A 3-Break:

- ✓ adds 3 new black edges and thus **creates** at most **3 new** cycles (containing three new black edges)
- ✓ removes 3 black edges and thus **destroys** at least **1 old** cycle (containing three old edges):

  change in the number of cycles: ***Δcycle ≤ 3-1=2.***

3-Break Distance

- ✓ Any 3-break increases the number of cycles by at most TWO (***Δcycles ≤2***)
- ✓ Any non-trivial cycle can be split into three cycles with a 3-break (***Δcycles = 2***)
- ✓ Every sorting by 3-breaks must increase the number of cycles by ***#blocks - cycles(P,Q)***
- ✓ The 3-**break distance** between genomes ***P*** and ***Q:***

$$d_3(P,Q) = (\#blocks - cycles(P,Q))/2$$

3-Break Distance

- ✓ Any 3-break increases the number of cycles by at most TWO (***Δcycles ≤2***)
- ✓ **Any non-trivial cycle can be split into three cycles with a 3-break (*Δcycles = 2*) – WRONG STATEMENT**
- ✓ Every sorting by 3-breaks must increase the number of cycles by ***#blocks - cycles(P,Q)***
- ✓ The 3-**break distance** between genomes ***P*** and ***Q:***

$$d_3(P,Q) = (\#blocks - cycles(P,Q))/2$$

3-Break Distance: Focus on Odd Cycles

- ✓ A 3-break can increase the number of *odd* cycles (i.e., cycles with odd number of black edges) by at most 2 (**$\Delta cycles^{odd} \leq 2$**)
- ✓ A non-trivial *odd* cycle can be split into three *odd* cycles with a 3-break (**$\Delta cycles^{odd} = 2$**)
- ✓ An *even* cycle can be split into two *odd* cycles with a 3-break (**$\Delta cycles^{odd} = 2$**)
- ✓ The *3-Break Distance* between genomes ***P*** and ***Q*** is:

$$d_3(P,Q) = (\#blocks - cycles^{odd}(P,Q)) / 2$$

Multi-Break Rearrangements

- ✓ *k*-Break rearrangement operation makes *k breaks* in a genome and glues the resulting pieces in a new order.

- ✓ Rearrangements are rare evolutionary events and biologists believe that *k*-break rearrangements are unlikely for $k>3$ and relatively rare for $k=3$ (at least in the mammalian evolution).

- ✓ Also, in radiation biology, chromosome aberrations for $k>2$ (indicative of chromosome damage rather than evolutionary viable variations) are more common, e.g., complex rearrangements in irradiated human lymphocytes (*Sachs et al., 2004; Levy et al., 2004*).

Polynomial Algorithm for Multi-Break Rearrangements

- ✓ The formulas for ***k***-break distance become complex as ***k*** grows (without an obvious pattern):

Corollary 2. *The 4-break distance between a black matching P and a gray matching Q is*

$$d_4(P, Q) = \left\lceil \frac{|P| - c_1(P, Q) - \lfloor c_2(P, Q)/2 \rfloor}{3} \right\rceil$$

where $c_i(P, Q)$ is the number of black-gray cycles containing i modulo 3 black edges.

Corollary 3. *The 5-break distance between a black matching P and a gray matching Q is*

$$d_5(P, Q) = \left\lceil \frac{|P| - c_1(P, Q) - \min\{c_2(P, Q), c_3(P, Q)\} - \lfloor \max\{0, c_3(P, Q) - c_2(P, Q)\}/3 \rfloor}{4} \right\rceil$$

where $c_i(P, Q)$ is the number of cycles containing i modulo 4 black edges.

The formula for $d_{20}(P,Q)$ contains over 1,500 terms!

Alekseyev & PP (SODA 07, Theoretical Computer Science 08) ***Exact formulas for the k-break distance as well as a linear-time algorithm for computing it.***

Where Do We Go From Here?

Ancestral Genome Reconstruction

Breakpoint Re-use Analysis

Skip

Genome Halving Problem

Breakpoint Re-use Analysis

Algorithmic Problem

Searching for Rearrangement Hotspots (Fragile Regions) in Human Genome

Random vs. Fragile Breakage Debate: Complex Rearrangements

- ✓ *PP & Tesler, PNAS 2003,* argued that every evolutionary scenario for transforming *Mouse* into *Human* genome with reversals and translocations must result in a large number of *breakpoint re-uses*, a contradiction to the RBM.

- ✓ *Sankoff, PLoS CB 2006: "We cannot infer whether mutually randomized synteny block orderings derived from two divergent genomes were created ... through processes other than reversals and translocations."* ***(read: "transpositions" or 3-breaks)***

Random Breakage Theory
Re-Re-Re-Re-Re-Visited

- **Ohno, 1970, Nadeau & Taylor, 1984** introduced RBM
- **Pevzner & Tesler, 2003** (PNAS) argued **against** RBM
- **Sankoff & Trinh, 2004** (RECOMB, JCB) argued **against**
  Pevzner & Tesler, 2003 arguments **against** RBM
- **Peng et al., 2006** (PLOS CB) argued **against**
  Sankoff & Trinh, 2004 arguments **against**
  Pevzner & Tesler, 2003 arguments **against** RBM
- **Sankoff, 2006** (PLOS CB) acknowledged an error in Sankoff&Trinh, 2004
  but came up with a new argument **against**
  Pevzner and Tesler, 2003 arguments **against** RBM
- **Alekseyev and Pevzner, 2007** (PLOS CB) argued **against**
  Sankoff, 2006 new argument **against**
  Pevzner & Tesler, 2003 arguments **against** RBM

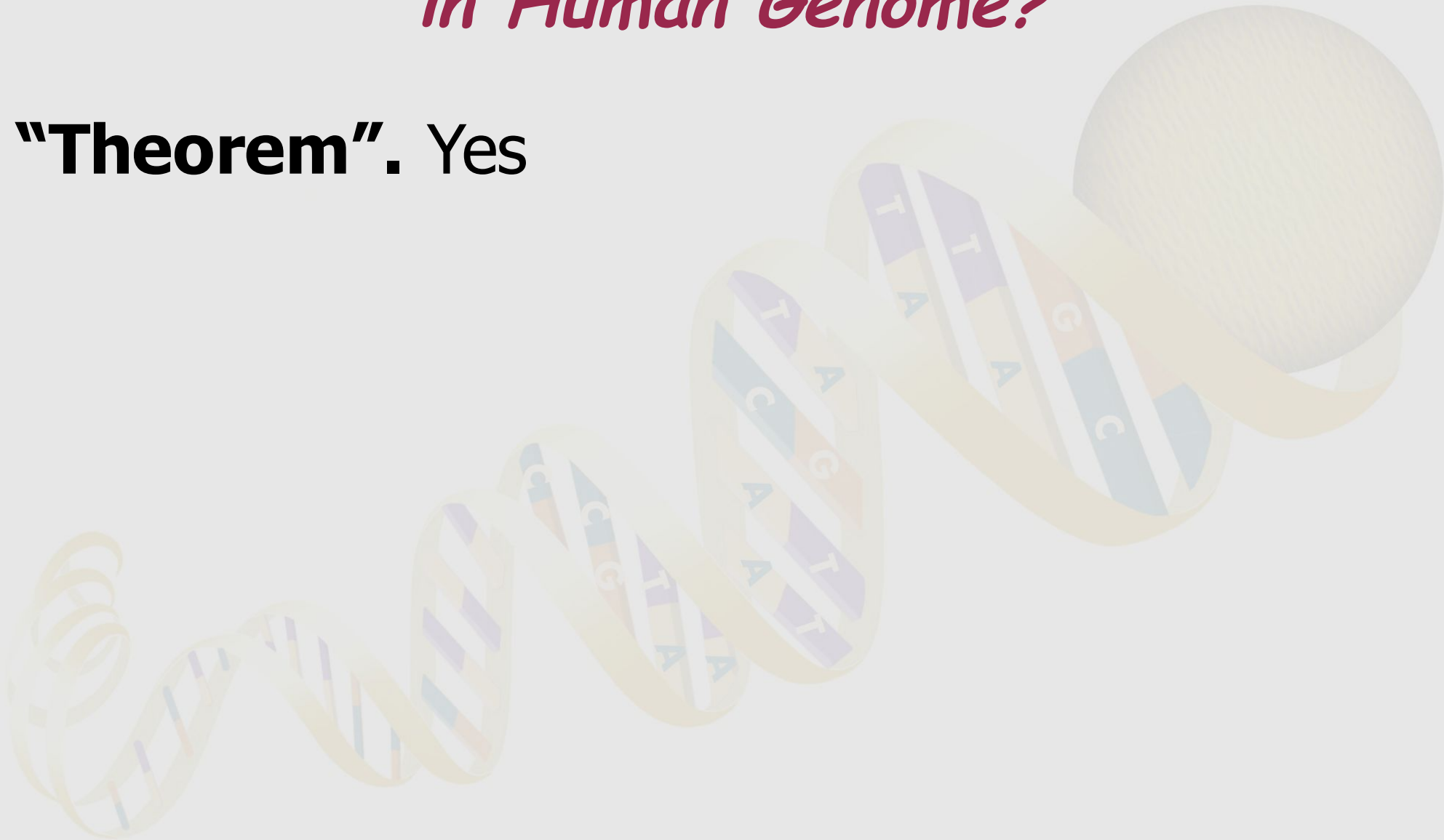
Random Breakage Theory
Re-Re-Re-Re-Re-Visited

- Ohno, 1970, Nadeau & Taylor, 1984 introduced RBM
- Pevzner & Tesler, PNAS 2003 argued **against** RBM
- Sankoff & Trinh, JCB 2004 argued **against**
  Pevzner & Tesler, 2003 arguments **against** RBM
- Peng et al., PLOS CB 2006 argued **against**
  Sankoff & Trinh, 2004 arguments **against**
  Pevzner & Tesler, 2003 arguments **against** RBM
- Sankoff, PLOS CB 2006 acknowledged an error in Sankoff&Trinh, 2004
  but came up with a new argument **against**
  Pevzner and Tesler, 2003 arguments **against** RBM
- Alekseyev and Pevzner, PLOS CB 2007 argued **against**
  Sankoff, 2006 new argument **against**
  Pevzner & Tesler, 2003 arguments **against** RBM

Are There Rearrangement Hotspots in Human Genome?

Are There Rearrangement Hotspots in Human Genome?

"Theorem". Yes



Are There Rearrangement Hotspots in Human Genome?

Theorem. Yes

Proof:

- The formula for the 3-break distance implies that there were at least $d_3(Human,Mouse) = 139$ rearrangements between human and mouse (including transpositions)

Are There Rearrangement Hotspots in Human Genome?

Theorem. Yes

Proof:

- The formula for 3-break distance implies that there were at least $d_3(Human,Mouse) = 139$ rearrangements between human and mouse (including transpositions)
- Every 3-break creates up to *3* breakpoints

Are There Rearrangement Hotspots in Human Genome?

Theorem. Yes

Proof:

- The formula for 3-break distance implies that there were at least $d_3(Human,Mouse) = 139$ rearrangements between human and mouse (including transpositions)
- Every 3-break creates up to *3* breakpoints
- If there were no breakpoint re-use then after *139* rearrangements we may see 139*3=417 breakpoints

Are There Rearrangement Hotspots in Human Genome?

Theorem. Yes

Proof:

- The formula for 3-break distance implies that there were at least $d_3(Human,Mouse) = 139$ rearrangements between human and mouse (including transpositions)
- Every rearrangement creates up to *3* breakpoints
- If there were no breakpoint re-use then after *139* rearrangements we may see 139*3=417 breakpoints
- But there are only 281 breakpoints between Human and Mouse!

Are There Rearrangement Hotspots in Human Genome?

Theorem. Yes

Proof:

- The formula for 3-break distance implies that there were at least $d_3(Human,Mouse) = 139$ rearrangements between human and mouse (including transpositions)
- Every rearrangement creates up to *3* breakpoints
- If there were no breakpoint re-use then after *139* rearrangements we may see 139*3=417 breakpoints
- But there are only 281 breakpoints between Human and Mouse
- Is 417 larger than 281?

Are There Rearrangement Hotspots in Human Genome?

Theorem. Yes

Proof:

- The formula for 3-break distance implies that there were at least $d_3(Human,Mouse) = 139$ rearrangements between human and mouse (including transpositions)
- Every rearrangement creates up to *3* breakpoints
- If there were no breakpoint re-use then after *139* rearrangements we may see 139*3=417 breakpoints
- But there are only 281 breakpoints between Human and Mouse
- Is 417 larger than 281?
- **Yes, 417 >> 281!**

Are There Rearrangement Hotspots in Human Genome?

Theorem. Yes

Proof:

- The formula for 3-break distance implies that there were at least $d_3(Human,Mouse) = 139$ rearrangements between human and mouse (including transpositions)
- Every rearrangement creates up to *3* breakpoints
- **If there were no breakpoint re-use** then after *139* rearrangements we may see 139*3=417 breakpoints
- But there are only 281 breakpoints between Human and Mouse
- Is 417 larger than 281?
- **Yes, 417 >> 281!**

Transforming Human Genome into Mouse Genome by 3-Breaks

$$Human = P_0 \xrightarrow{\text{3-breaks}} P_1 \rightarrow \ldots \rightarrow P_d = Mouse$$

$$d = d_3(Human, Mouse) = 139$$

✓Our "proof" assumed that each 3-break makes 3 breakages in a genome, so the total number of breakages made in this transformation is ***3 * 139 = 417***.

✓**OOPS!** The transformation may include *2-breaks* (as a particular case of *3-breaks).* If every *3-break* were a *2-break* then the total number of breakages is only ***2*139 = 278 < 281***, in which case there could be no breakpoint re-uses at all.

Minimizing the Number of Breakages

Problem. Given genomes ***P*** and ***Q***, find a series of *k-breaks* transforming ***P*** into ***Q*** and making the *smallest number of breakages.*

Theorem. Any series of *k-breaks* transforming ***P*** into ***Q*** makes at least $d_k(P,Q)+d_2(P,Q)$ breakages

Theorem. There exists a series of $d_3(P,Q)$ *3-breaks* transforming ***P*** into ***Q*** and making $d_3(P,Q)+d_2(P,Q)$ breakages.

$d_2(Human,Mouse) = 246$
$d_3(Human,Mouse) = 139$
minimum number of breakages = 246 + 139 = 385

Are There Rearrangement Hotspots in Human Genome?

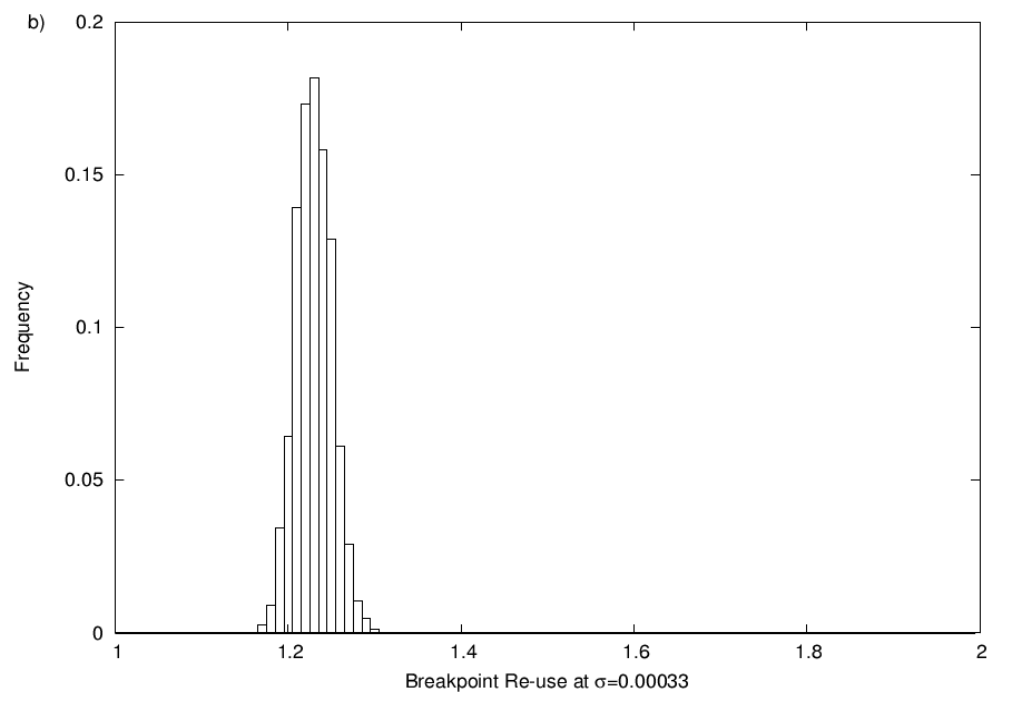
Theorem. Yes

Proof:

- The formula for 3-break distance implies that there were at least $d_3(Human,Mouse) = 139$ rearrangements between human and mouse (including transpositions)
- Every rearrangement creates up to *3* breakpoints
- If there were no breakpoint re-use then after *139* rearrangements we **SHOULD SEE AT LEAST** 385 breakpoints (rather than 417 as before)
- But there are only 281 breakpoints between Human and Mouse
- **Yes, 385 >> 281!**

Breakpoint Re-uses between Human and Mouse Genomes

- ✓ Any transformation of *Mouse* into *Human* genome with *3-breaks* requires at least *385* breakages, while there are *281* breakpoints.
- ✓ So, there are at least *385-281=104* breakpoint re-uses (re-use rate *1.37*) which is significantly higher than statistically expected in the RBM.



- ✓ Mean = *1.23*
- ✓ Standard deviation = *0.02*
- ✓ Maximum breakpoint reuse rate = *1.33* (observed once in *100,000* simulations)

Are there rearrangement hotspots in human genome?

- ✓ We showed that even if *3*-break rearrangements were frequent, the argument against the RBM still stands (**Alekseyev & PP**, ***PLoS Comput. Biol. 2007***)

  We don't believe that 3-breaks are frequent but even if they were, we proved that there exist fragile regions in the human genome. It is time to answer a more difficult question: ***"Where are the rearrangement hotspots in the human genome?"*** (the detailed analysis appeared in **Alekseyev and PP, Genome Biol., Dec.1, 2010**)

Where Do We Go From Here?

Ancestral Genome Reconstruction

Breakpoint Re-use Analysis

Skip

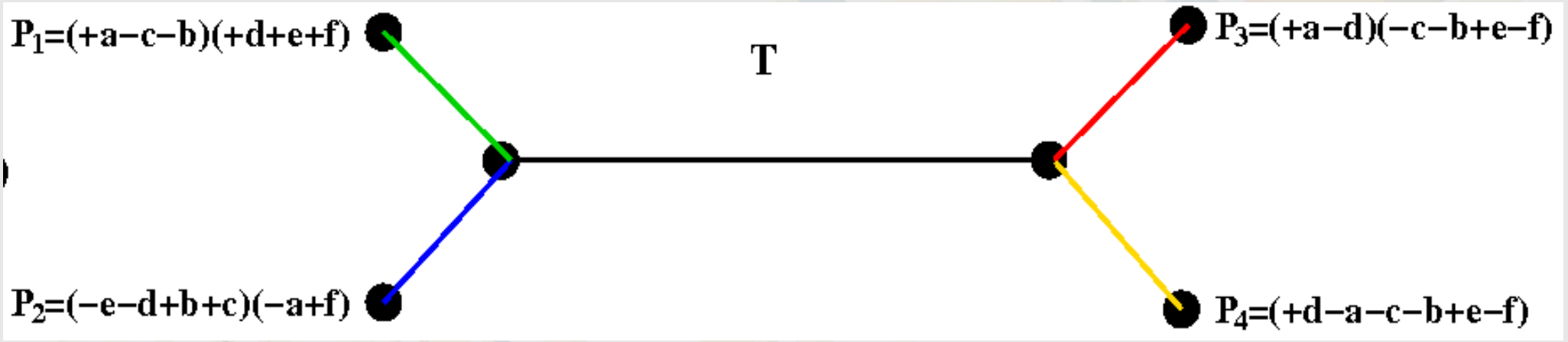
Genome Halving Problem

Algorithmic Problem:

Ancestral Genome Reconstruction and Multiple Breakpoint Graphs

Ancestral Genomes Reconstruction

- ✓ Given a set of genomes, reconstruct genomes of their common ancestors.



Algorithms for Ancestral Genomes Reconstruction

- ✓ **GRAPPA**: *Tang, Moret, Warnow et al . (2001)*
- ✓ **MGR**: *Bourque and PP (Genome Res. 2002)*
- ✓ **InferCARs**: *Ma et al. (Genome Res. 2006)*
- ✓ **EMRAE**: *Zhao and Bourque (Genome Res. 2007)*
- ✓ **MGRA**: *Alekseyev and PP (Genome. Res.2009)*

Ancestral Genomes Reconstruction Problem (with a known tree)

- ✓ **Input:** a set of ***k*** genomes and a phylogenetic tree ***T***
- ✓ **Output:** genomes at the internal nodes of the tree ***T*** that minimize the total sum of the 2-break distances along the branches of ***T***

- ✓ NP-complete in the "simplest" case of ***k=3***.
- ✓ ***What makes it hard?***

Ancestral Genomes Reconstruction Problem (with a known tree)

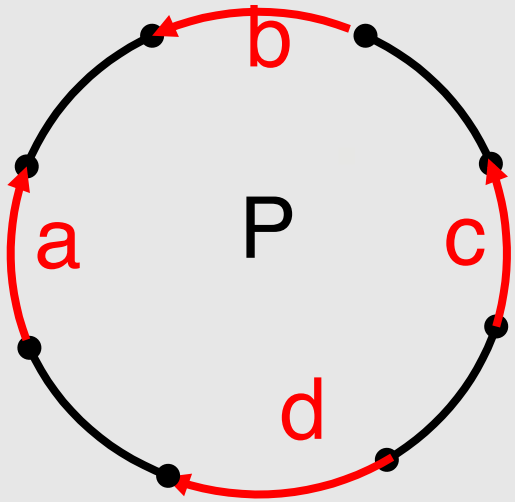
- ✓ **Input:** a set of ***k*** genomes and a phylogenetic tree ***T***
- ✓ **Output:** genomes at the internal nodes of the tree ***T***
- ✓ **Objective**: minimize the total sum of the 2-break distances along the branches of ***T***

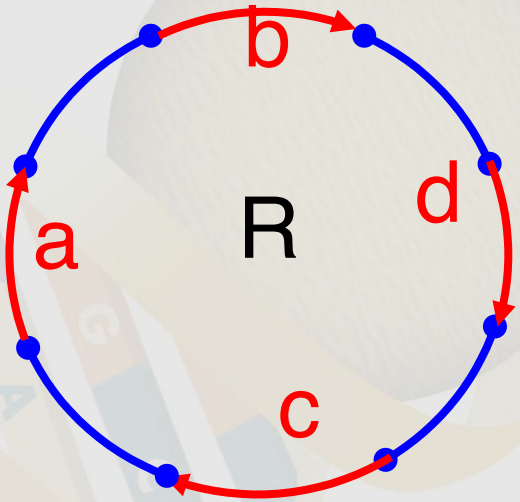
- ✓ NP-complete in the "simplest" case of ***k=3***.
- ✓ ***What makes it hard?*** ***BREAKPOINTS RE-USE***

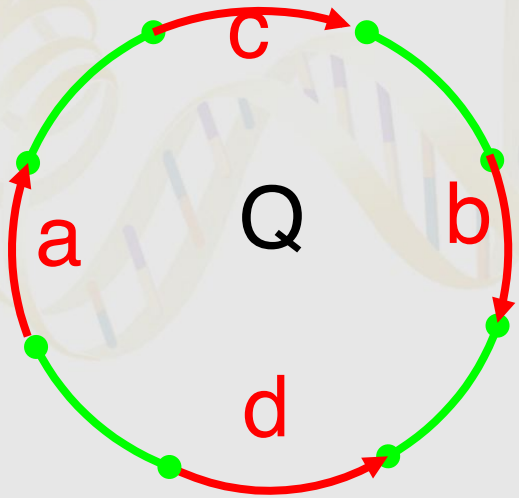
Breakpoints Are "Footprints" of Rearrangements on the "Ground" of Genomes

- ✓ NP-complete in the "simplest" case of ***k=3***.
- ✓ ***What makes it hard?*** ***BREAKPOINTS RE-USES (resulting in messy "footprints")!***
  Ancestral Genome Reconstructions of MANY Genomes (i.e., for **large *k*)** may be easier to solve.

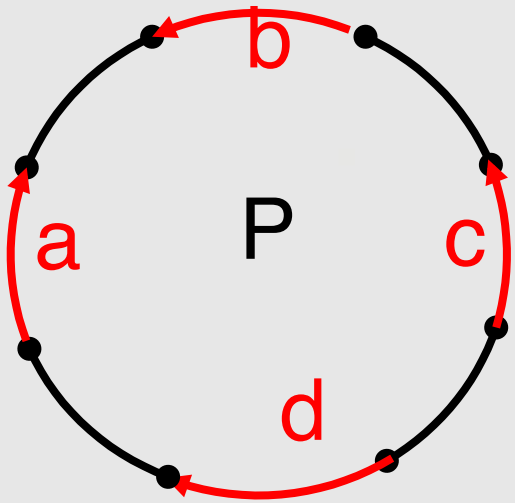
How to Construct the Breakpoint Graph for Multiple Genomes?

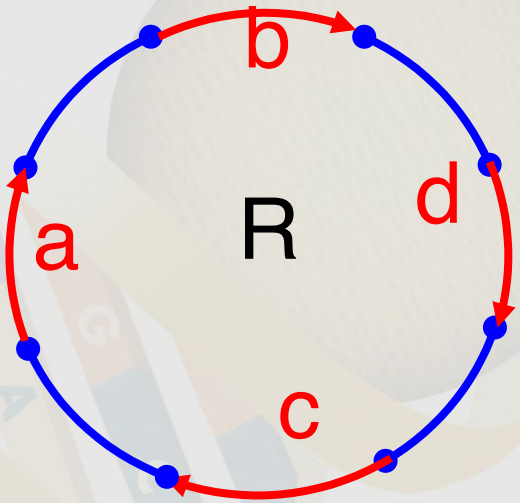


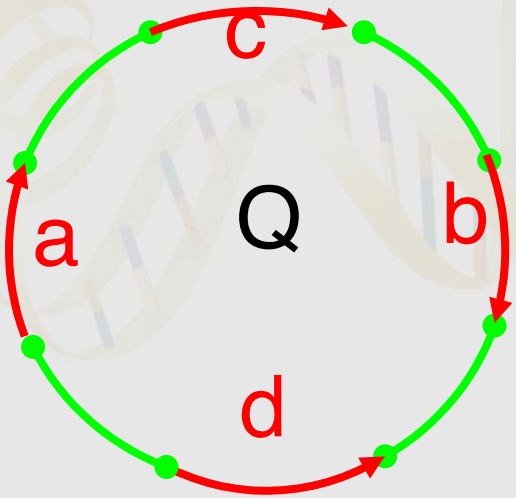


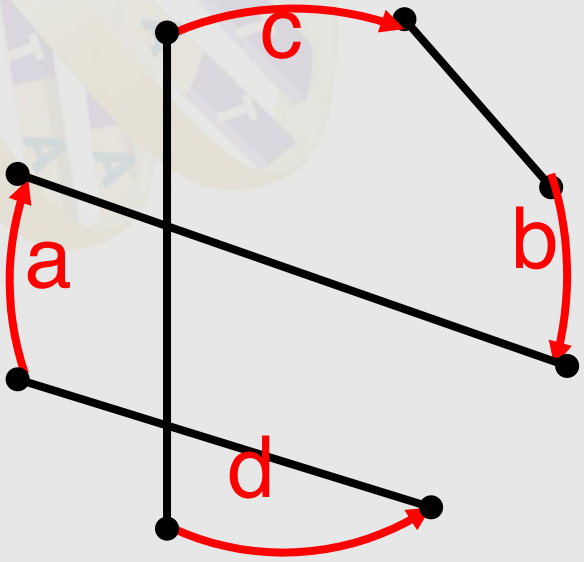


Constructing Multiple Breakpoint Graph: rearranging P in the Q order

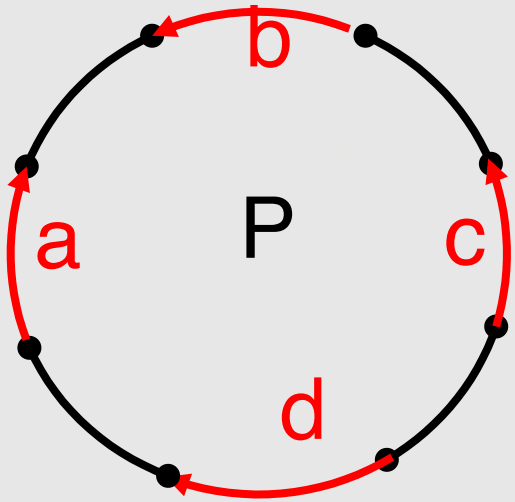


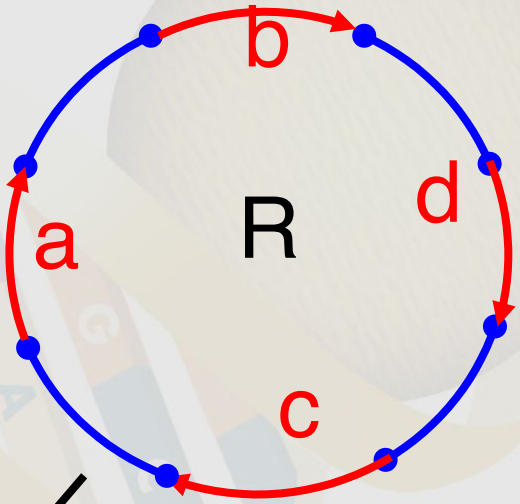


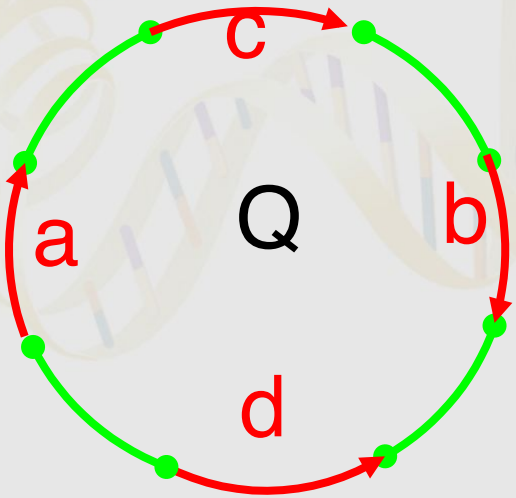


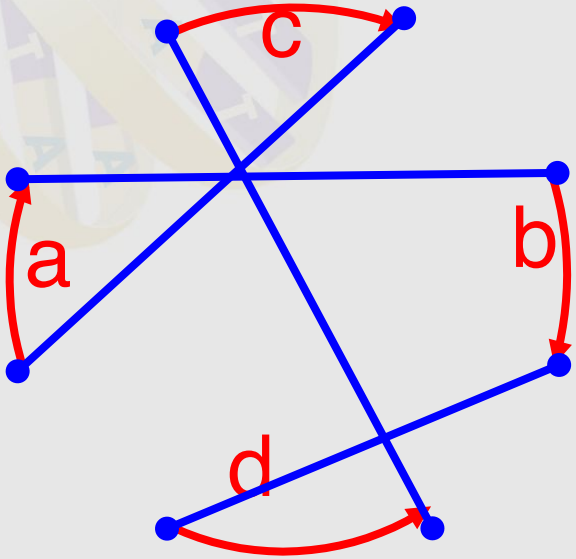


Constructing Multiple Breakpoint Graph: rearranging R in the Q order

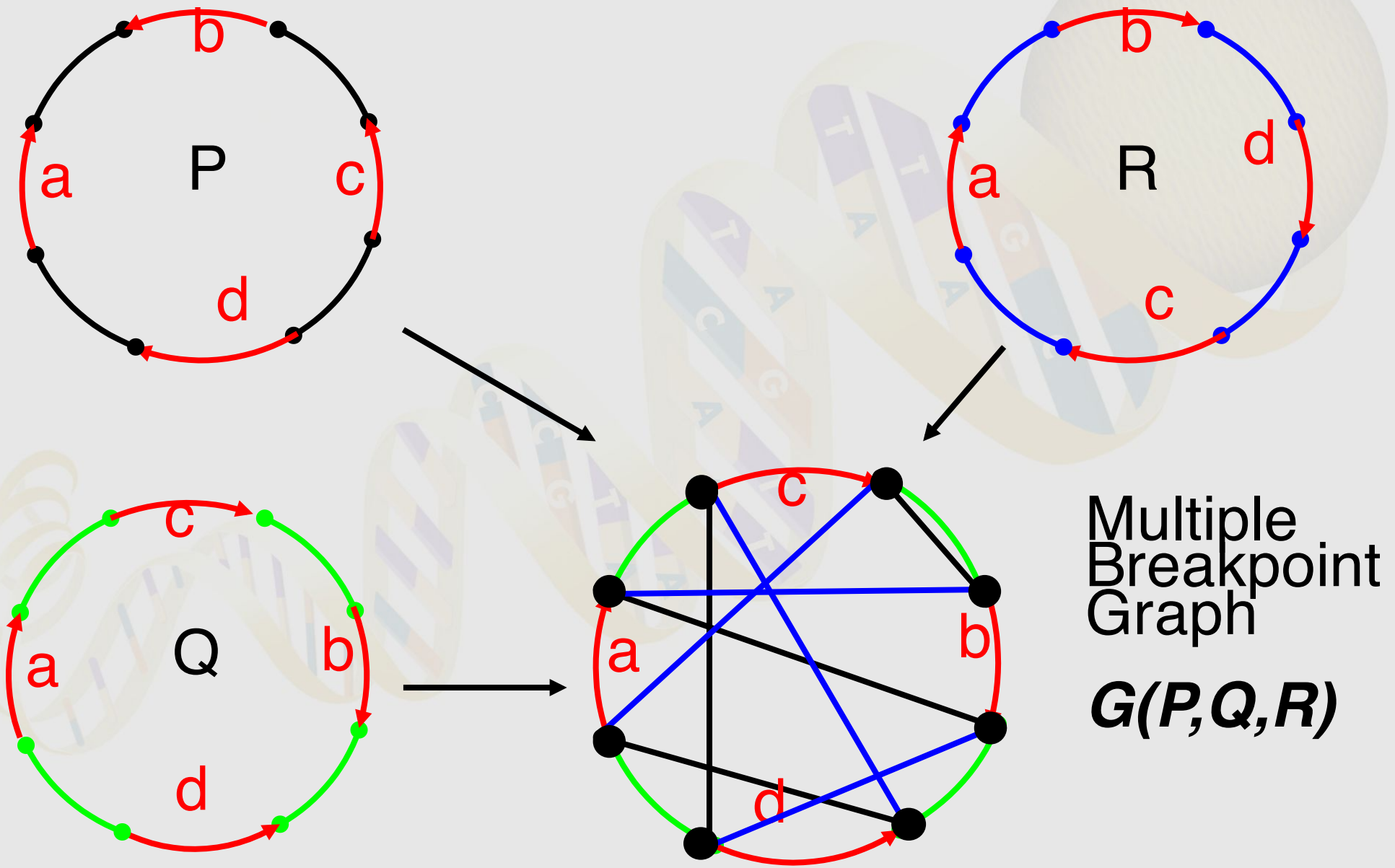








Multiple Breakpoint Graph: Still Gluing Red Edges with the Same Labels



Multiple Breakpoint Graph of 6 Mammalian Genomes

Multiple Breakpoint Graph ***G(M,R,D,Q,H,C)*** of the

M*ouse,*

R*at,*

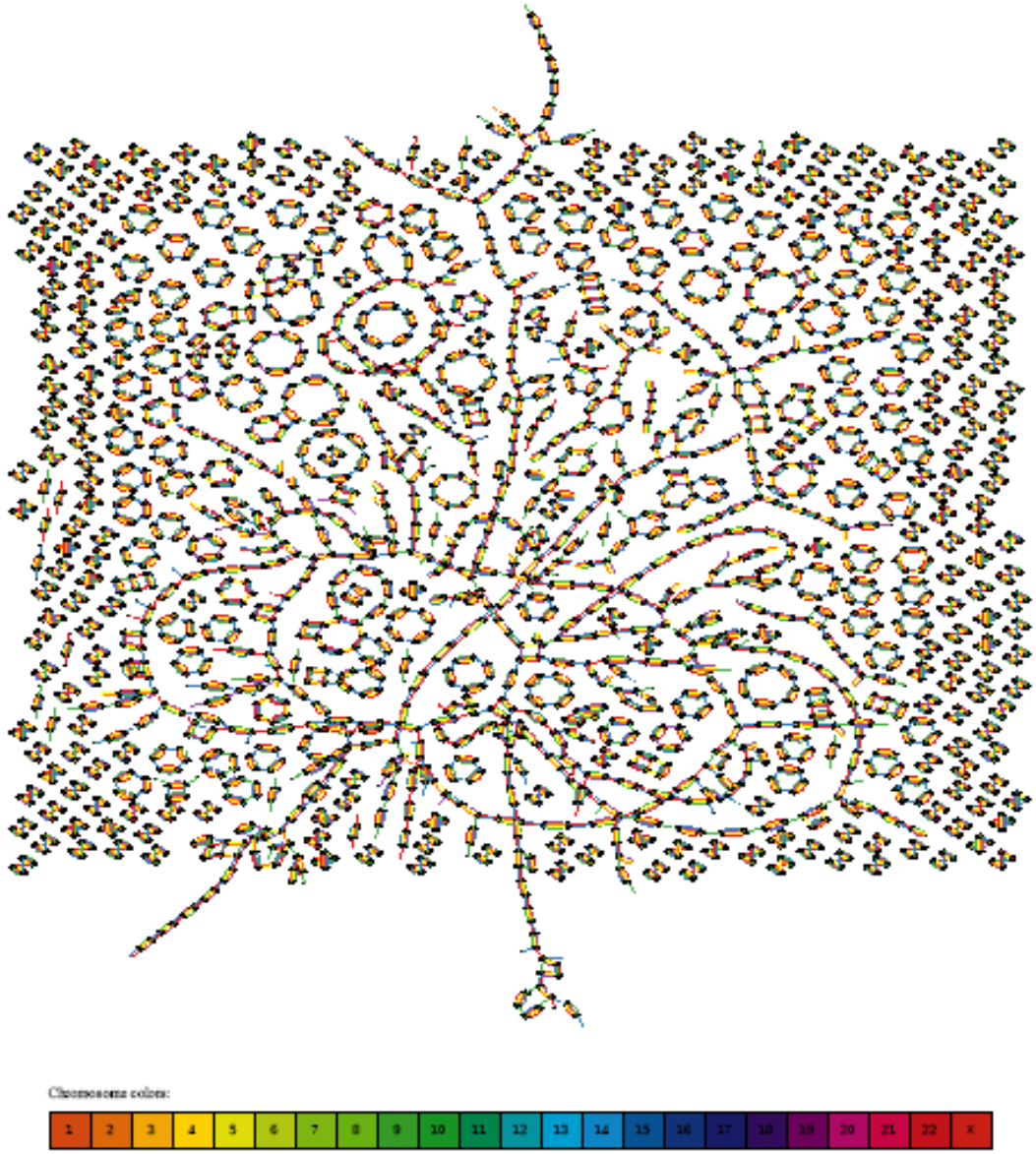
D*og,*

*maca****Q****ue,*

H*uman, and*

C*himpanzee*

genomes.



Two Genomes:
Two Ways of Sorting by 2-Breaks

Transforming ***P*** into ***Q*** with ***"black" 2-breaks***:

$$P = P_0 \rightarrow P_1 \rightarrow \ldots \rightarrow P_{d-1} \rightarrow P_d = Q$$

$$G(P,Q) \rightarrow G(P_1,Q) \rightarrow \ldots \rightarrow G(P_d,Q) = G(Q,Q)$$

Transforming ***Q*** into ***P*** with ***"green" 2-breaks***:

$$Q = Q_0 \rightarrow Q_1 \rightarrow \ldots \rightarrow Q_d = P$$

$$G(P,Q) \rightarrow G(P,Q_1) \rightarrow \ldots \rightarrow G(P,Q_d) = G(P,P)$$

Let's make a black-green chimeric transformation

Transforming G(P,Q) into (an unknown!) G(X,X) rather than into (a known) G(Q,Q) as before

- ✓ Let ***X*** be *any* genome on a path from ***P*** to ***Q***:

$$P = P_0 \rightarrow P_1 \rightarrow \ldots \rightarrow P_m = X = Q_{m-d} \leftarrow \ldots \leftarrow Q_1 \leftarrow Q_0 = Q$$

- ✓ Sorting by 2-breaks is equivalent to finding a *shortest* transformation of ***G(P,Q)*** into an identity breakpoint graph ***G(X,X)*** of ***a priori unknown*** genome ***X***:

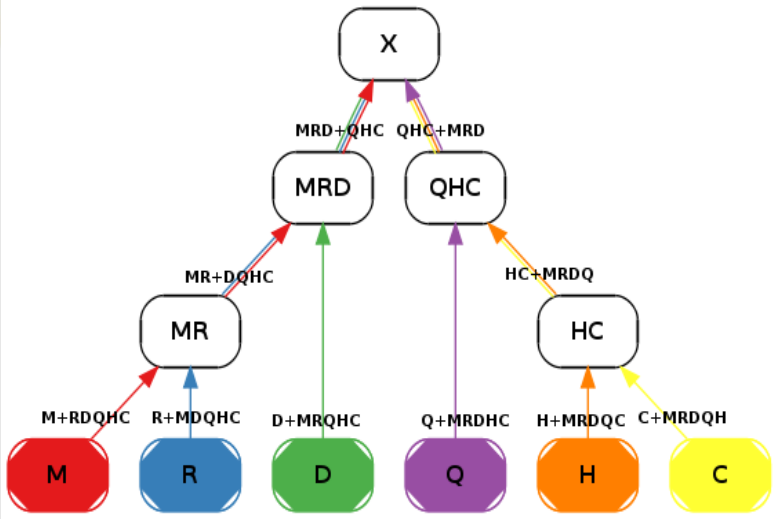
$$G(P_0,Q_0) \rightarrow G(P_1,Q_0) \rightarrow G(P_1,Q_1) \rightarrow G(P_1,Q_2) \rightarrow \ldots \rightarrow G(X,X)$$

- ✓ The **"black"** and **"green"** 2-breaks may arbitrarily alternate.

From 2 Genomes To Multiple Genomes

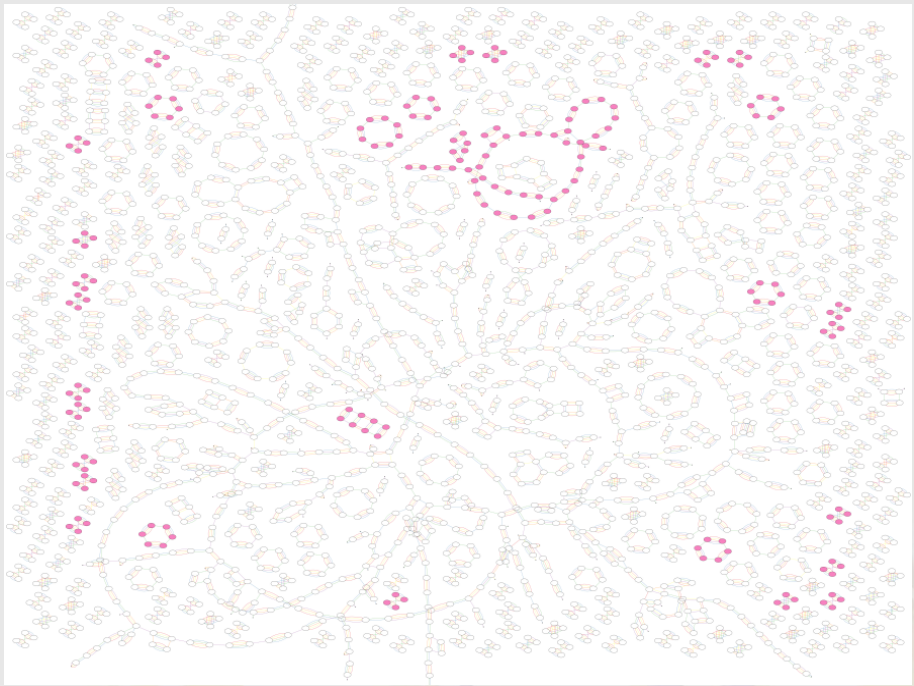
- We find a transformation of the multiple breakpoint graph $G(P_1,P_2,...,P_k)$ into *(a priori unknown!)* identity multiple breakpoint graph $G(X,X,...,X)$:
  $G(P_1,P_2,...,P_k) \rightarrow ... \rightarrow G(X,X,...,X)$

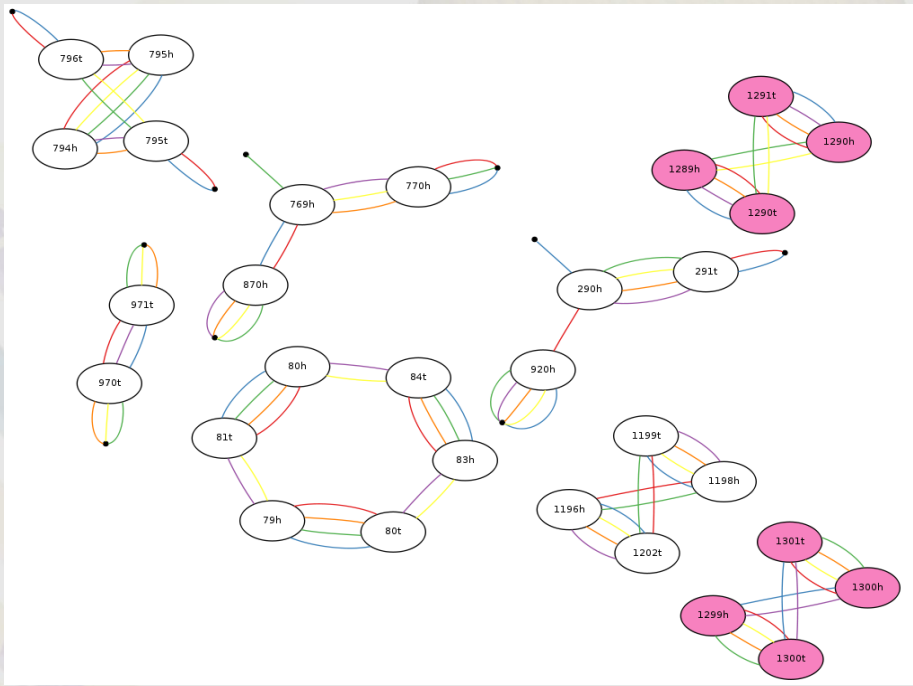
- The evolutionary tree ***T*** defines *groups of genomes* (to which the same 2-breaks may be applied simultaneously).



- For example, the groups ***{M, R}*** (M*ouse and Rat*) and ***{Q,H,C}*** (*macaQue, Human, Chimpanzee*) correspond to branches of ***T*** while the groups ***{M, C}*** and ***{R, D}*** do not.

When All Reliable 2-Breaks Are Identified and "Undone"

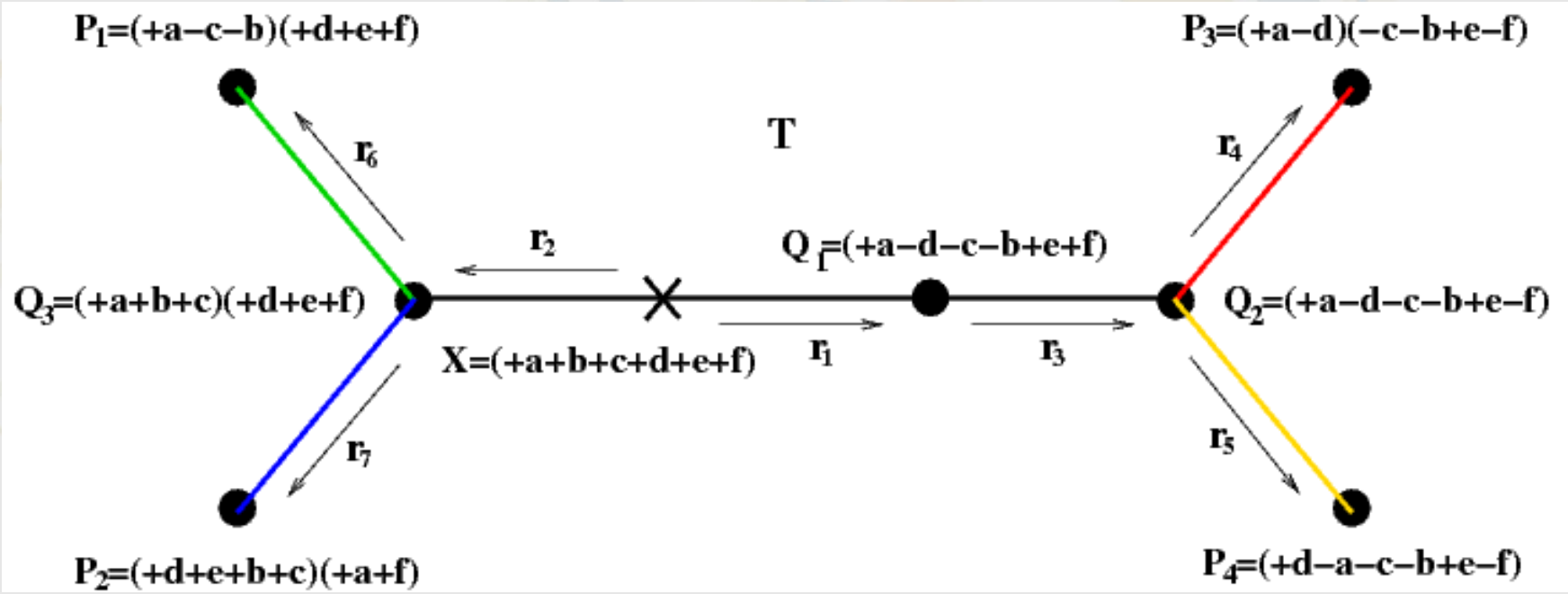




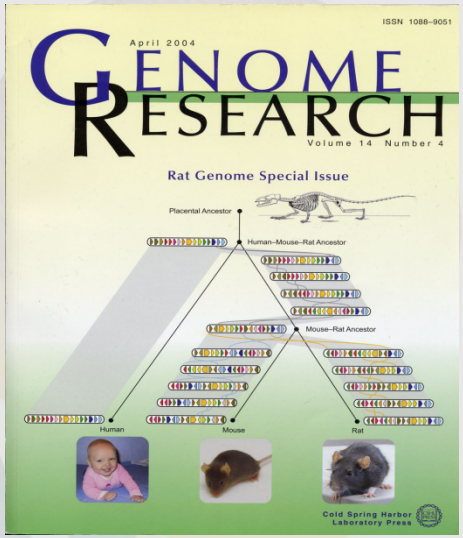
- ✓ The multiple breakpoint graph is reduced dramatically!
- ✓ The remaining (non-trivial) components can be processed manually in the case-by-case fashion.

Reconstruction Of The Ancestral Genomes

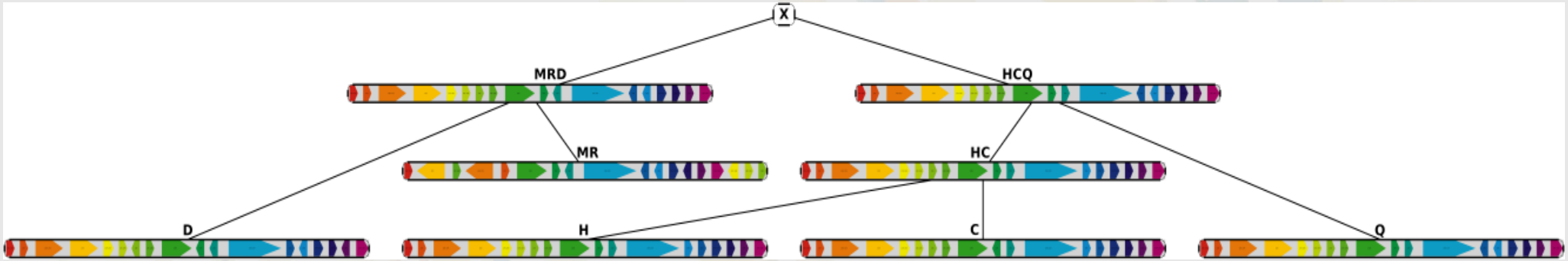
- ✓ The resulting identity breakpoint graph ***G(X,X,...,X)*** defines its underlying genome ***X***.
- ✓ The *reverse transformation* is applied to the genome ***X*** to transform it into each of the original genomes $P_1, P_2, \ldots, P_k$.
- ✓ Since ***X*** is passing through all internal nodes of ***T***, it defines the ancestral genomes at these nodes.



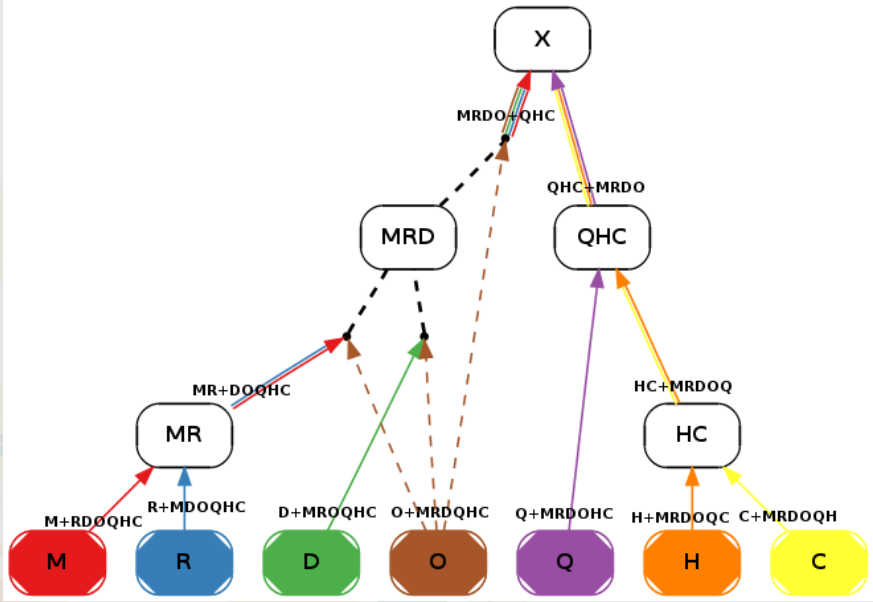
Reconstructed X Chromosomes



- ✓ The *Mouse, Rat, Dog, macaQue, Human, Chimpanzee* genomes and their reconstructed ancestors:



If the Evolutionary Tree Is Unknown

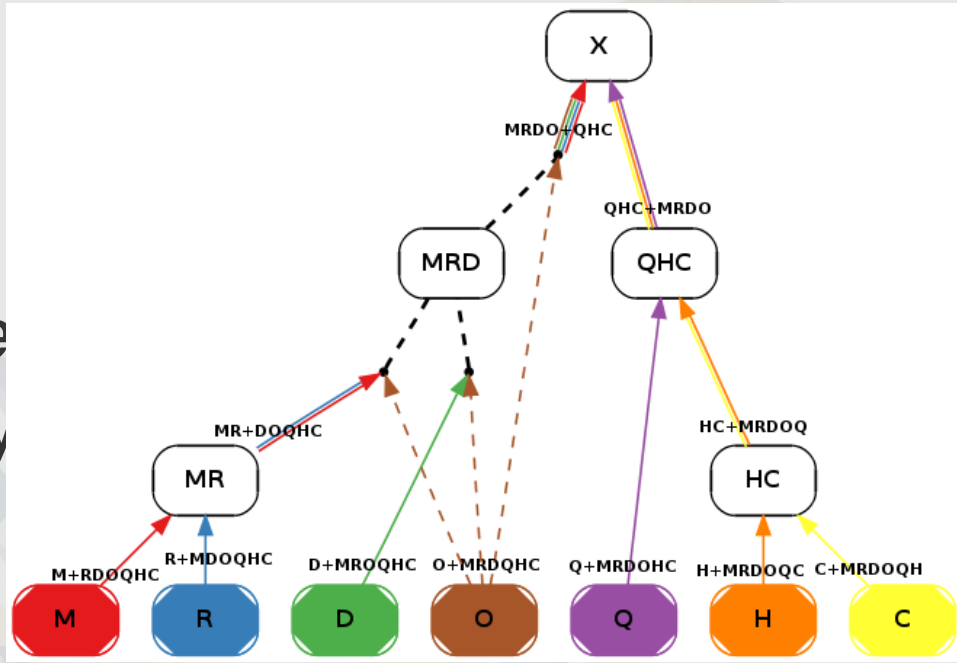


- ✓ For the set of 7 mammalian genomes (*Mouse, Rat, Dog, macaQue, Human, Chimpanzee*, and *Opossum)*, the evolutionary tree ***T*** was subject of enduring debates

- ✓ Depending on the primate – rodent – carnivore split, *three topologies are possible* (only two of them are viable).

If The Evolutionary Tree Is Not Known

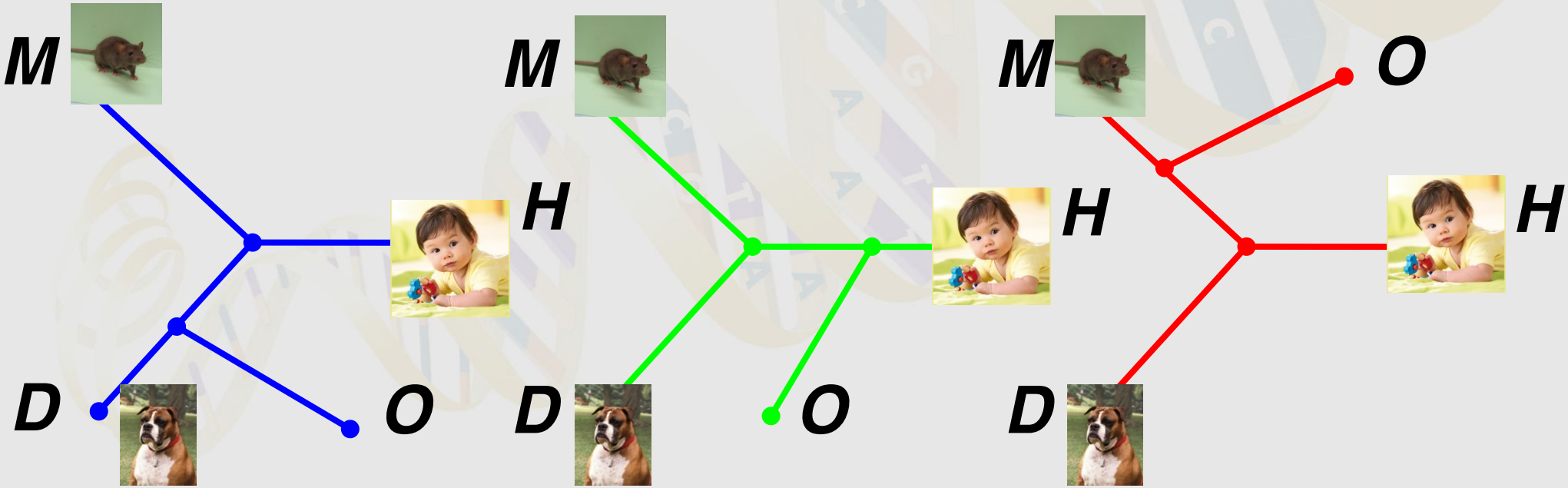
- For the set of 7 mammalian genomes: *Mouse, Rat, Dog, macaQue, Human, Chimpanze* and *Opossum*, the evolutionary tree ***T*** is being debated.



- Depending on the primate – rodent – carnivore split, *three topologies are possible* (only two of them are viable).
- However, these three topologies share many common branches in ***T*** (***confident branches***). We can restrict the transformation only to such branches in order to simplify the breakpoint graph, not breaking an evidence for either of the topologies.

Rearrangement Evidence For The Primate-Carnivore Split

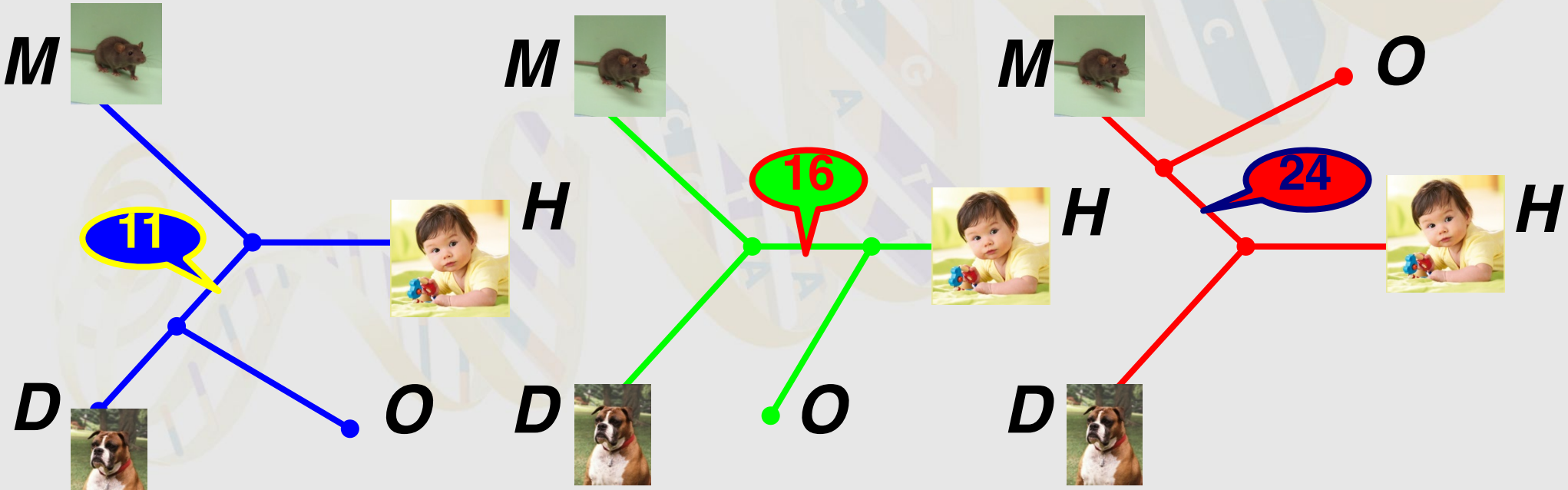
- ✓ Each of these three topologies has an unique branch that supports either blue (primate-rodent), or green (rodent-carnivore), or red (primate-carnivore) hypothesis.



- ✓ Rearrangement analysis supports the primate – carnivore split.

Rearrangement Evidence For The Primate-Carnivore Split

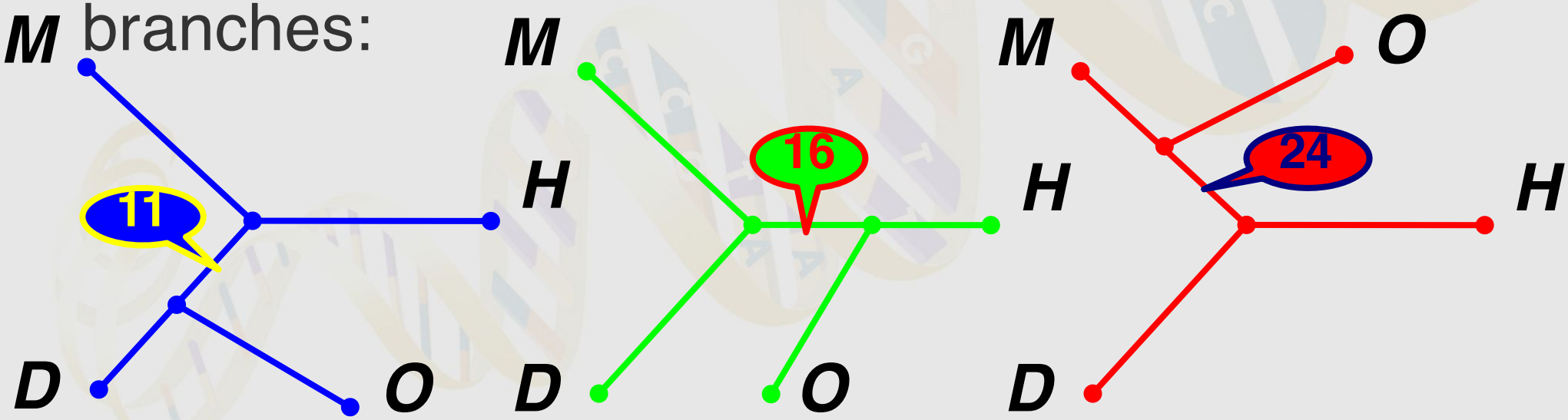
- ✓ Each of these three topologies has an unique branch that supports either blue (primate-rodent), or green (rodent-carnivore), or red (primate-carnivore) hypothesis.



- ✓ Rearrangement analysis supports the primate – carnivore split.

Rearrangement Evidence For The Primate-Carnivore Split

- ✓ Each of these three topologies has an unique branch that supports either blue (primate-rodent), or green (rodent-carnivore), or red (primate-carnivore) hypothesis. We analyze the rearrangements supporting each of these branches:



- ✓ Rearrangement analysis supports the primate – carnivore split.

Where Do We Go From Here?

Ancestral Genome Reconstruction

Breakpoint Re-use Analysis

Skip

Genome Halving Problem

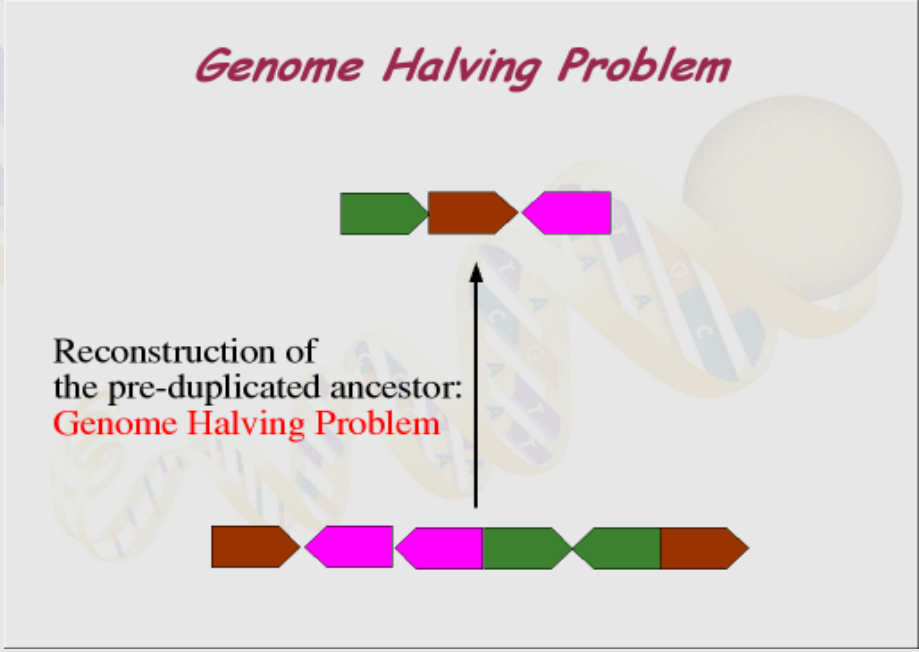
Algorithmic Problem:

Genome Halving Problem

WGD and Genome Halving Problem

- ✓ Whole Genome Duplication (WGD) of a genome ***R*** results in a perfect duplicated genome ***R+R*** where each chromosome is doubled.
- ✓ The genome ***R+R*** is subjected to rearrangements that result in a duplicated genome ***P***.

- ✓ **Genome Halving Problem:** Given a duplicated genome ***P***, find a perfect duplicated genome ***R+R*** minimizing the rearrangement distance between ***R+R*** and ***P***.



Genome Halving Problem: Previous Results

- ✓ Algorithm for reconstructing a pre-duplicated genome was proposed in a series of papers by El-Mabrouk and Sankoff (culminating in *SIAM J Comp., 2004*).
- ✓ The proof of the Genome Halving Theorem is rather technical (spans over 30 pages).

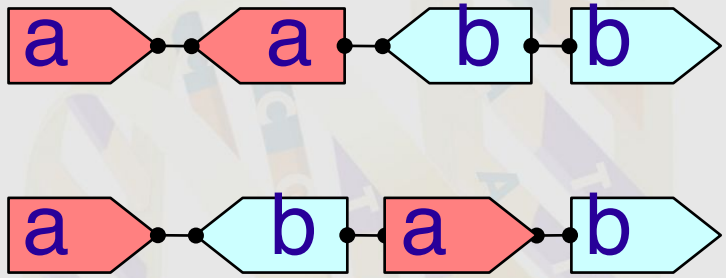
Genome Halving Theorem

- ✓ Found an error in the original Genome Halving Theorem for unichromosomal genomes and proved a theorem that adequately deals with all genomes.
  (Alekseyev & PP, SIAM J. Comp. 2007)
- ✓ Suggested a new (short) proof of the Genome Halving Theorem - our proof is 5 pages long.
  (**Alekseyev & PP, *IEEE Bioinformatics 2007***)
- ✓ Proved the Genome Halving Theorem for the harder case of transposition-like operations (**Alekseyev & PP, *SODA 2007***)

- ✓ Introduced the notion of the ***contracted breakpoint graph*** that makes difficult problems (like the Genome Halving Problem) more transparent.

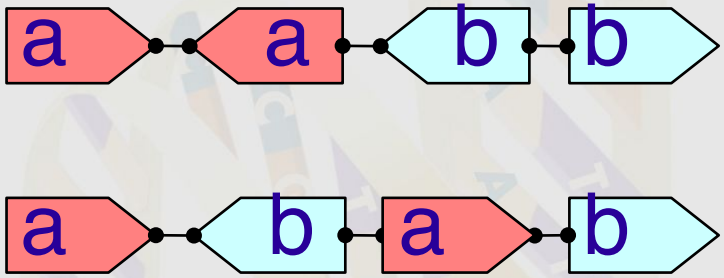
2-Break Distance Between Duplicated Genomes

2-Break Distance between Duplicated Genomes:



2-Break Distance Between Duplicated Genomes

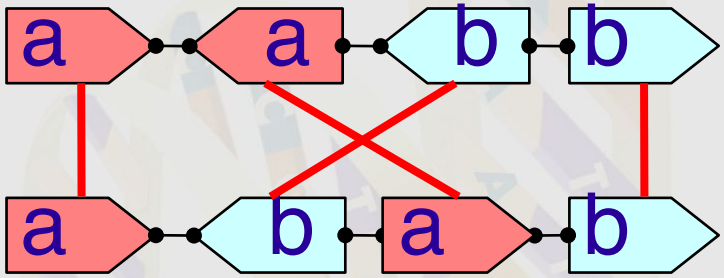
2-Break Distance between Duplicated Genomes:.



Difficulty: The notion of the breakpoint graph is not defined for duplicated genomes (the first ***a*** in ***P*** can correspond to either the first or the second ***a*** in ***Q***).

2-Break Distance Between Duplicated Genomes

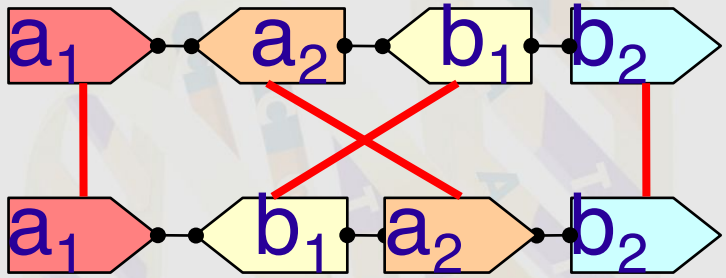
2-Break Distance between Duplicated Genomes:



Idea: Establish a correspondence between the genes in ***P*** and ***Q*** and *"re-label"* the corresponding blocks (e.g., as a_1, a_2, b_1, b_2)

2-Break Distance Between Duplicated Genomes

2-Break Distance between Duplicated Genomes

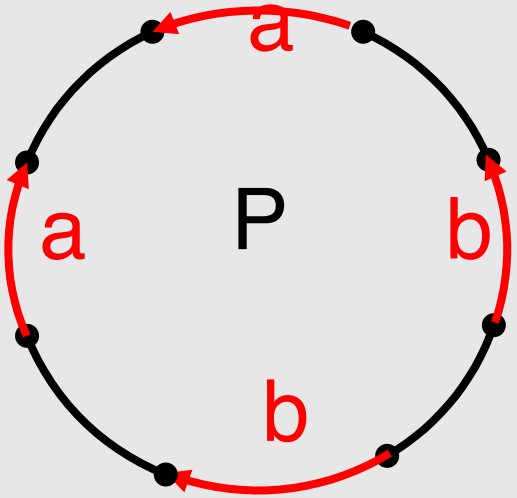


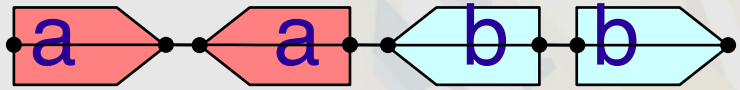
Treat the labeled genomes as non-duplicated, construct the breakpoint graph and compute the 2-break distance between them.

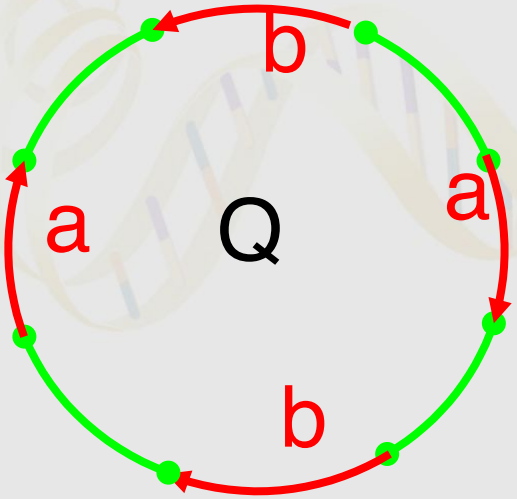
Different Labelings Result in Different Breakpoint Graphs

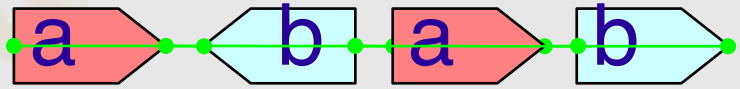
- ✓ There are 2 different labelings of the copies of each block.
- ✓ One of these labelings (corresponding to a breakpoint graph with maximum number of cycles) is an *optimal labeling*.
- ✓ Trying all possible labelings takes exponential time: $2^{\#blocks}$ invocations of the 2-break distance algorithm.
- ✓ **Labeling Problem.** Construct an optimal labeling that results in the maximum number of cycles in the breakpoint graph (hence, the smallest 2-break distance).

Constructing Breakpoint Graph of Duplicated Genomes

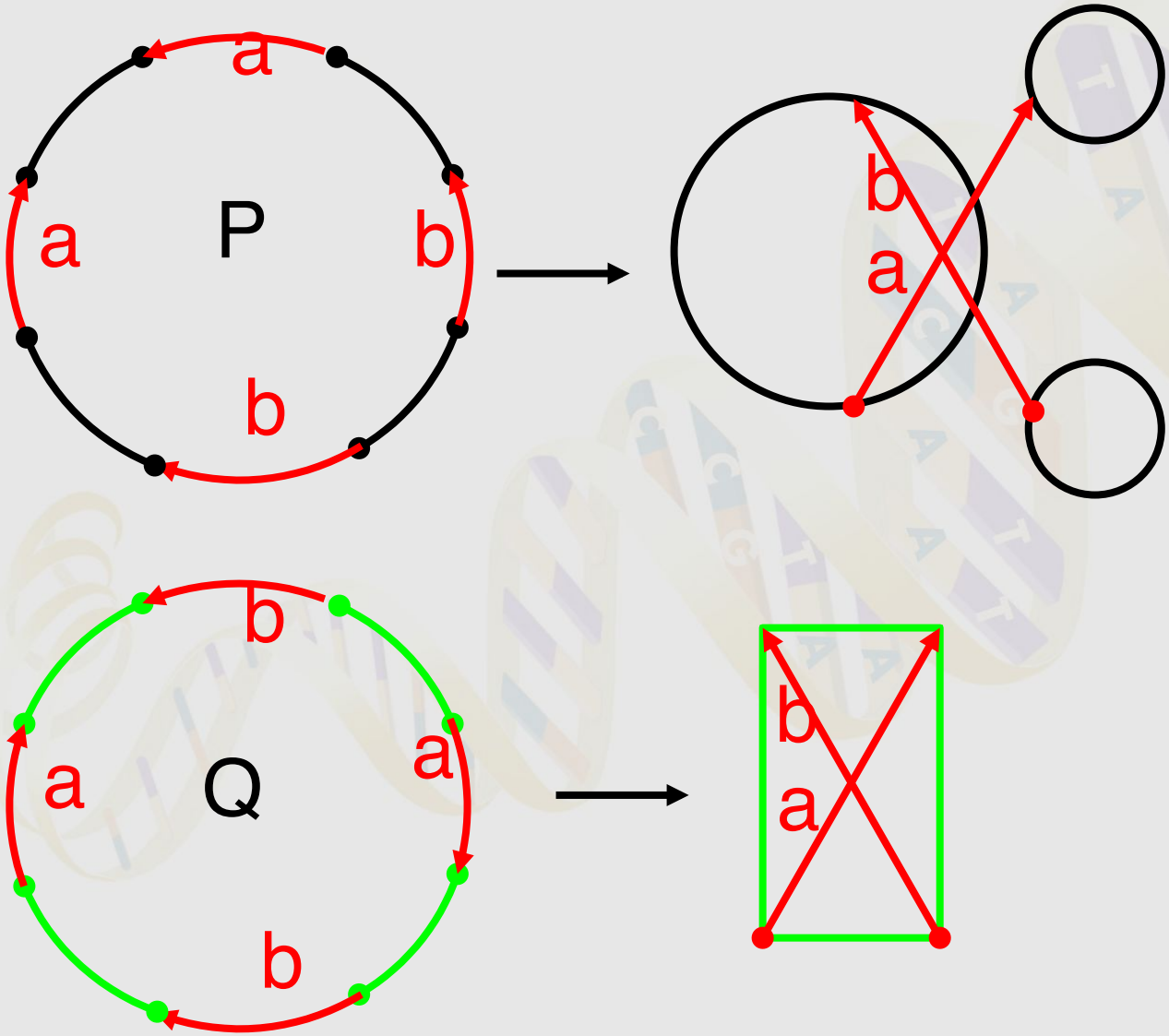




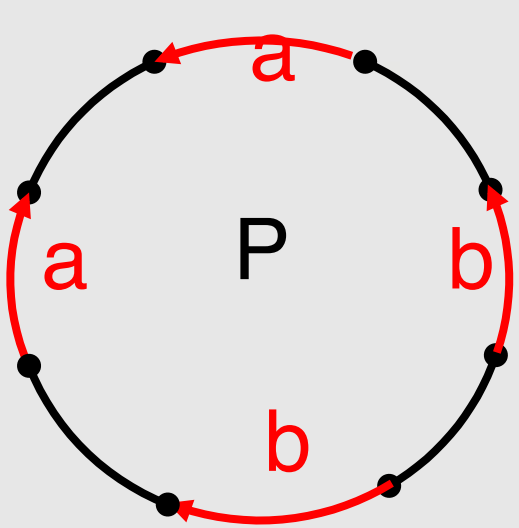


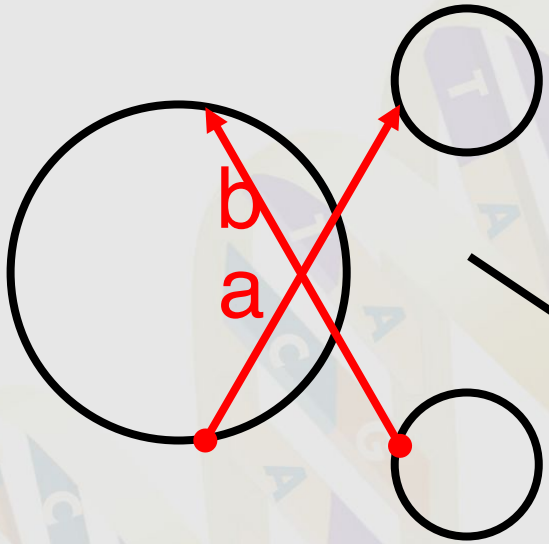


Contracted Genome Graphs: GLUING Red Edges

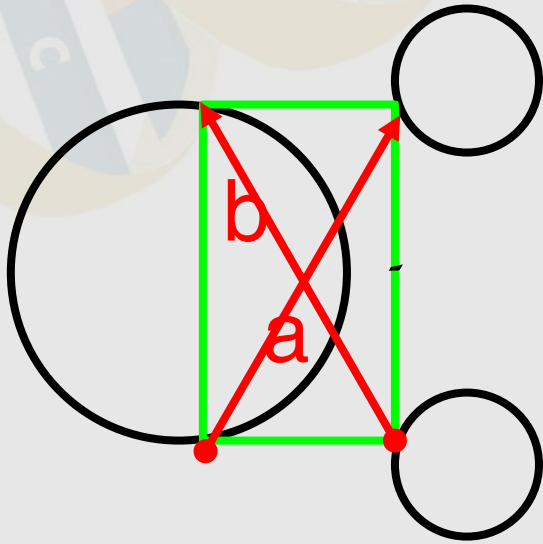


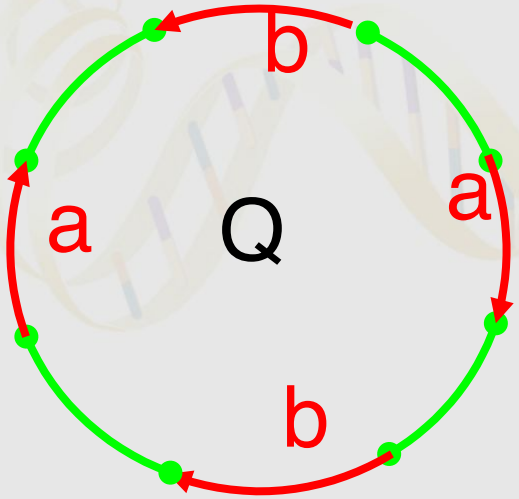
Contracted Breakpoint Graph: Gluing Red Edges Yet Again

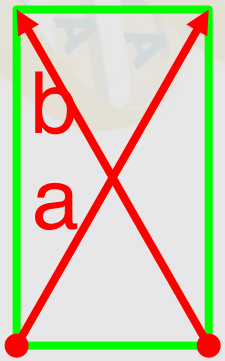




Contracted Breakpoint Graph
G'(P,Q)



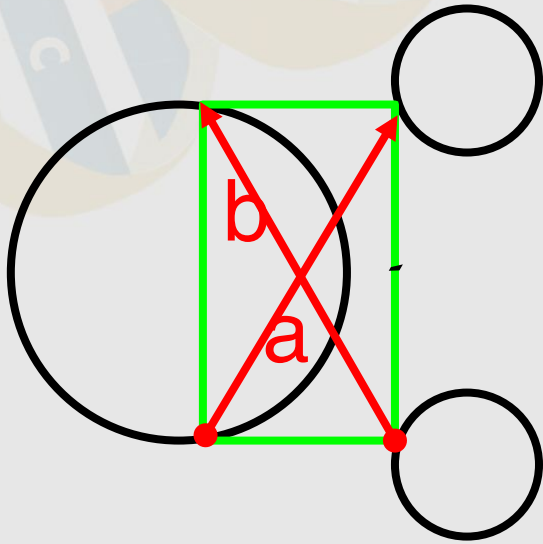




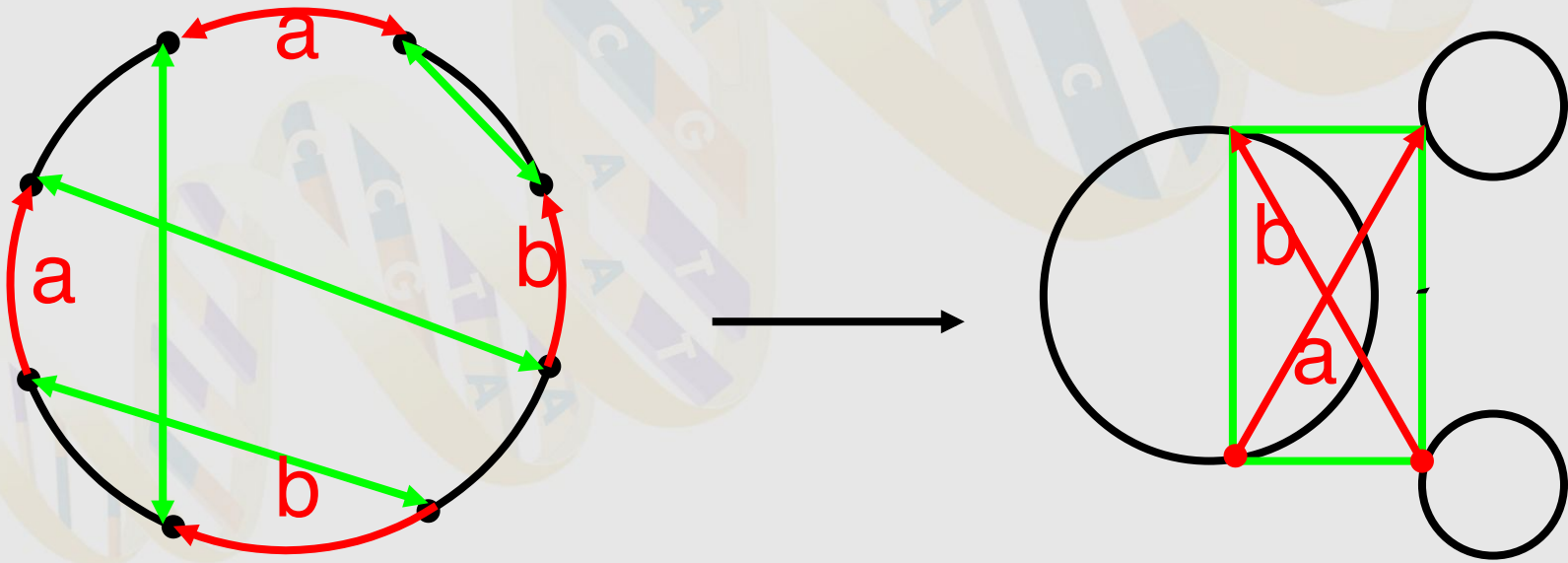
Contracted Breakpoint Graph of Duplicated Genomes

- ✓ Edges of 3 colors: ***black***, ***green,*** and ***red***.

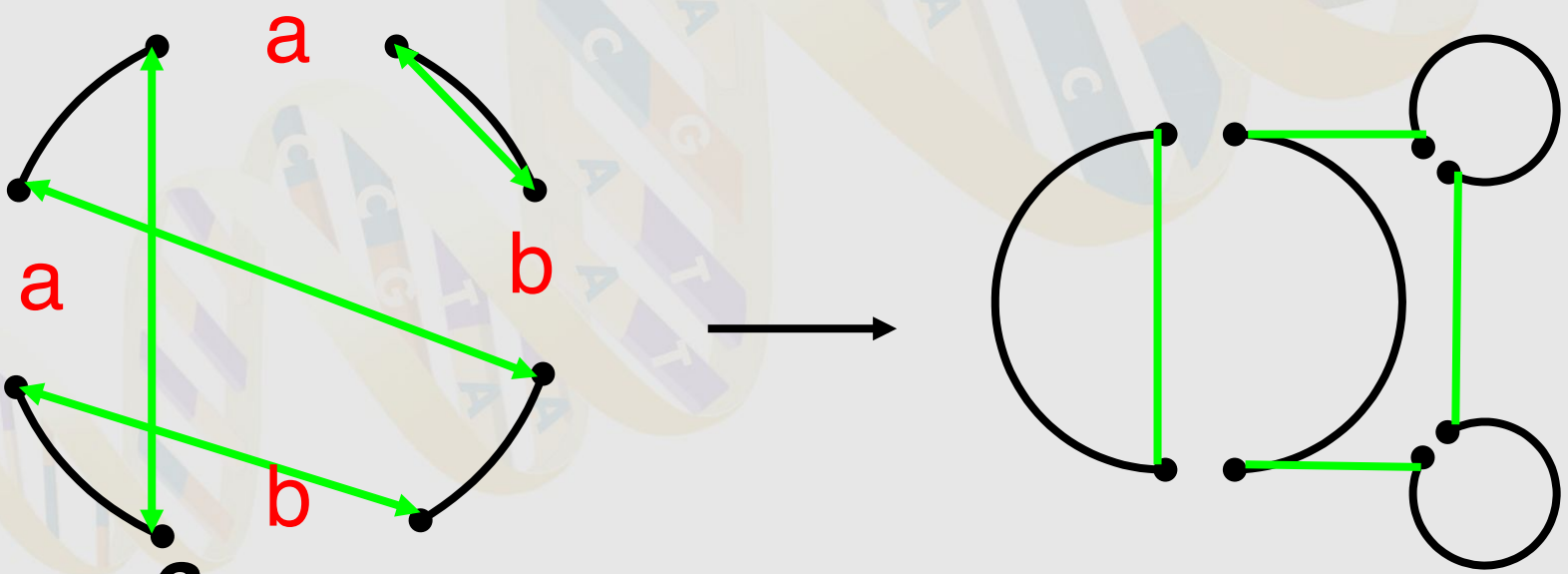
- ✓ *Red edges* form a *matching.*
- ✓ ***Black edges*** form *black cycles.*
- ✓ *Green edges* form *green cycles.*



What is the Relationship Between the Breakpoint Graph and the Contracted Breakpoint Graph?

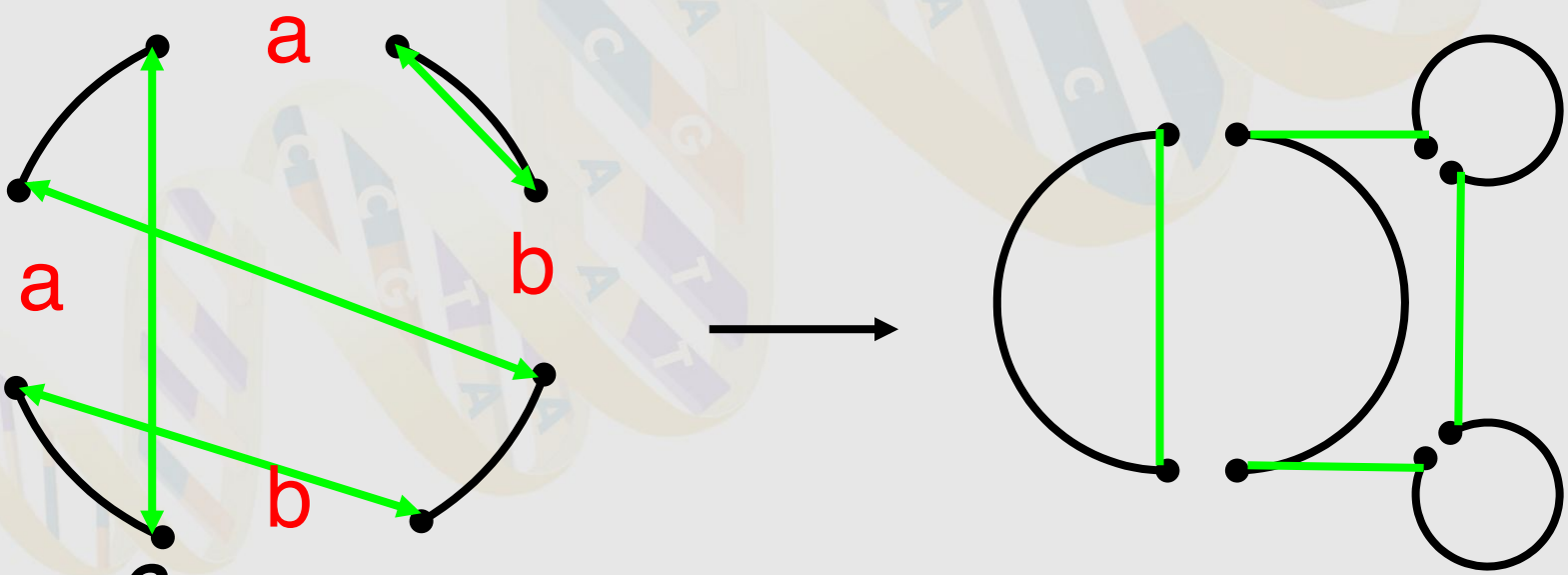


Every Breakpoint Graph Induces a Cycle Decomposition of the Contracted Breakpoint Graph



Why do we care?

Every Breakpoint Graph Induces Cycle Decomposition of the Contracted Breakpoint Graph

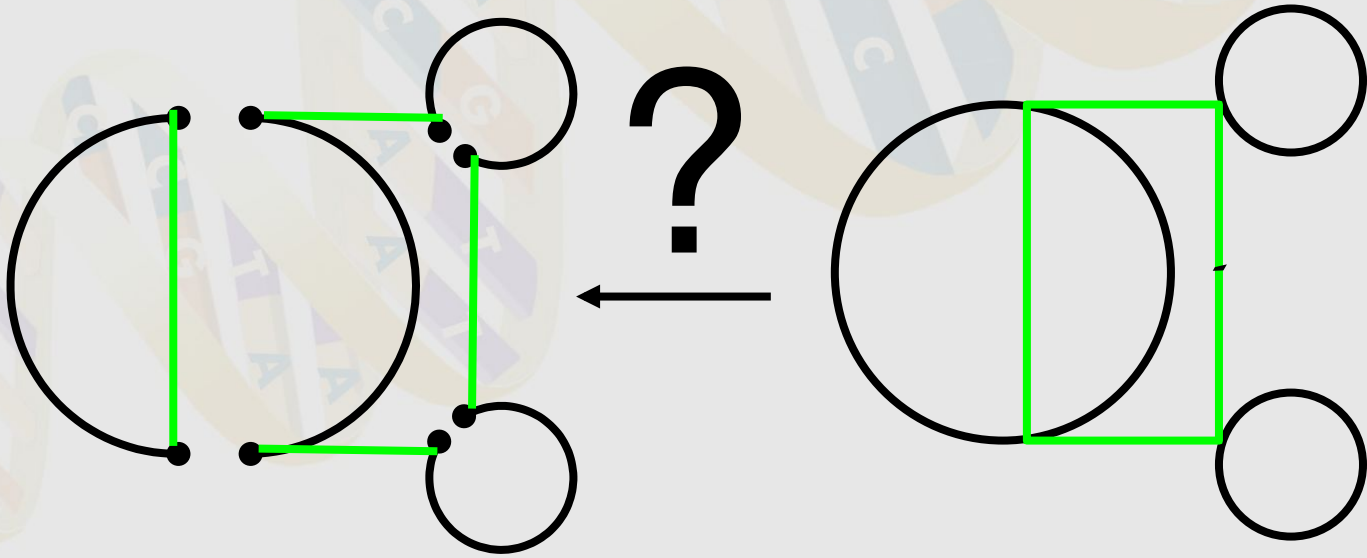


Why do we care?

Because optimal labeling corresponds to a maximum cycle decomposition.

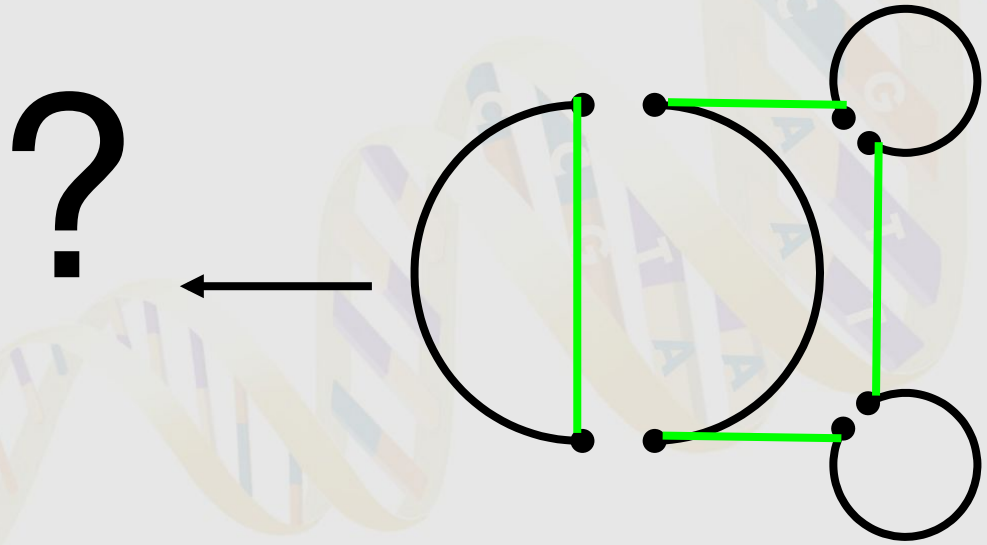
Maximum Cycle Decomposition Problem

Open Problem 1: Given a contracted breakpoint graph, find its maximum black-green cycle decomposition.



Labeling Problem

Open Problem 2: Find a breakpoint graph that induces a given cycle decomposition of the contracted breakpoint graph.

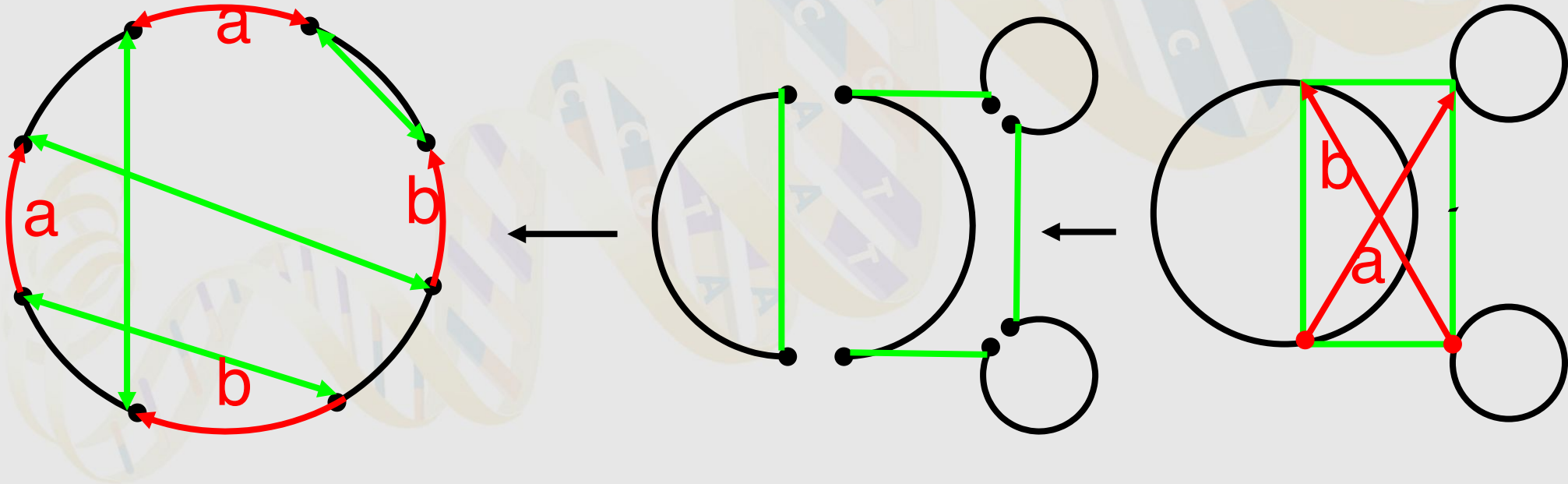


Computing 2-Break Distance Between Duplicated Genomes: Two Problems

breakpoint graph inducing max cycle decomposition

maximum cycle decomposition

contracted breakpoint graph

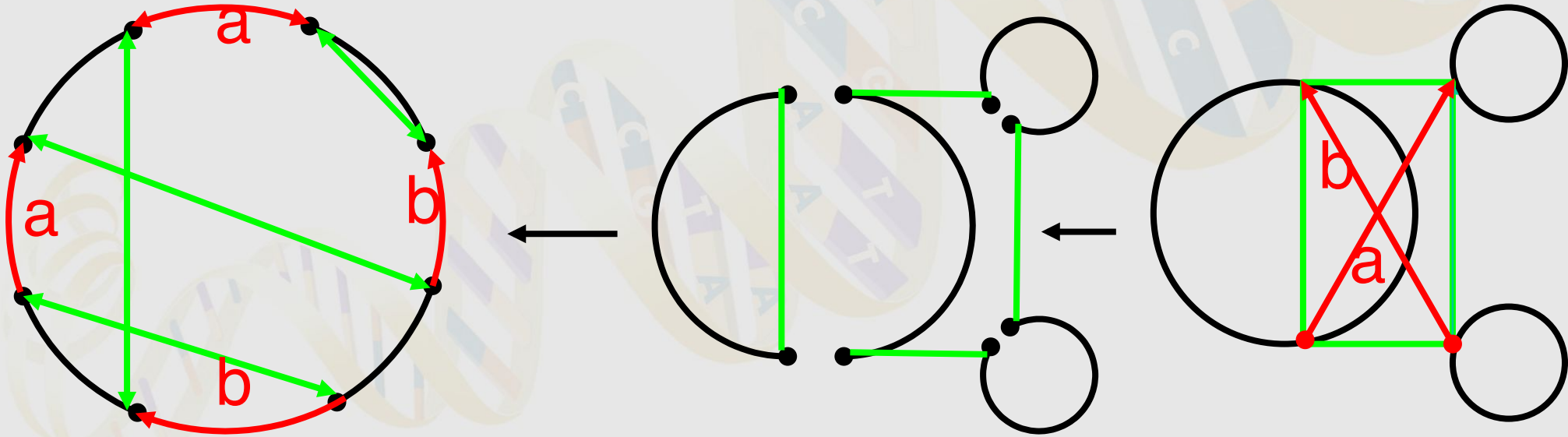


Computing 2-Break Distance Between Duplicated Genomes: Two HARD Problems

breakpoint graph inducing max cycle decomposition

maximum cycle decomposition

contracted breakpoint graph



These problems are difficult and we do not know how to solve them.

However, we solved them in the case of the Genome Halving Problem.

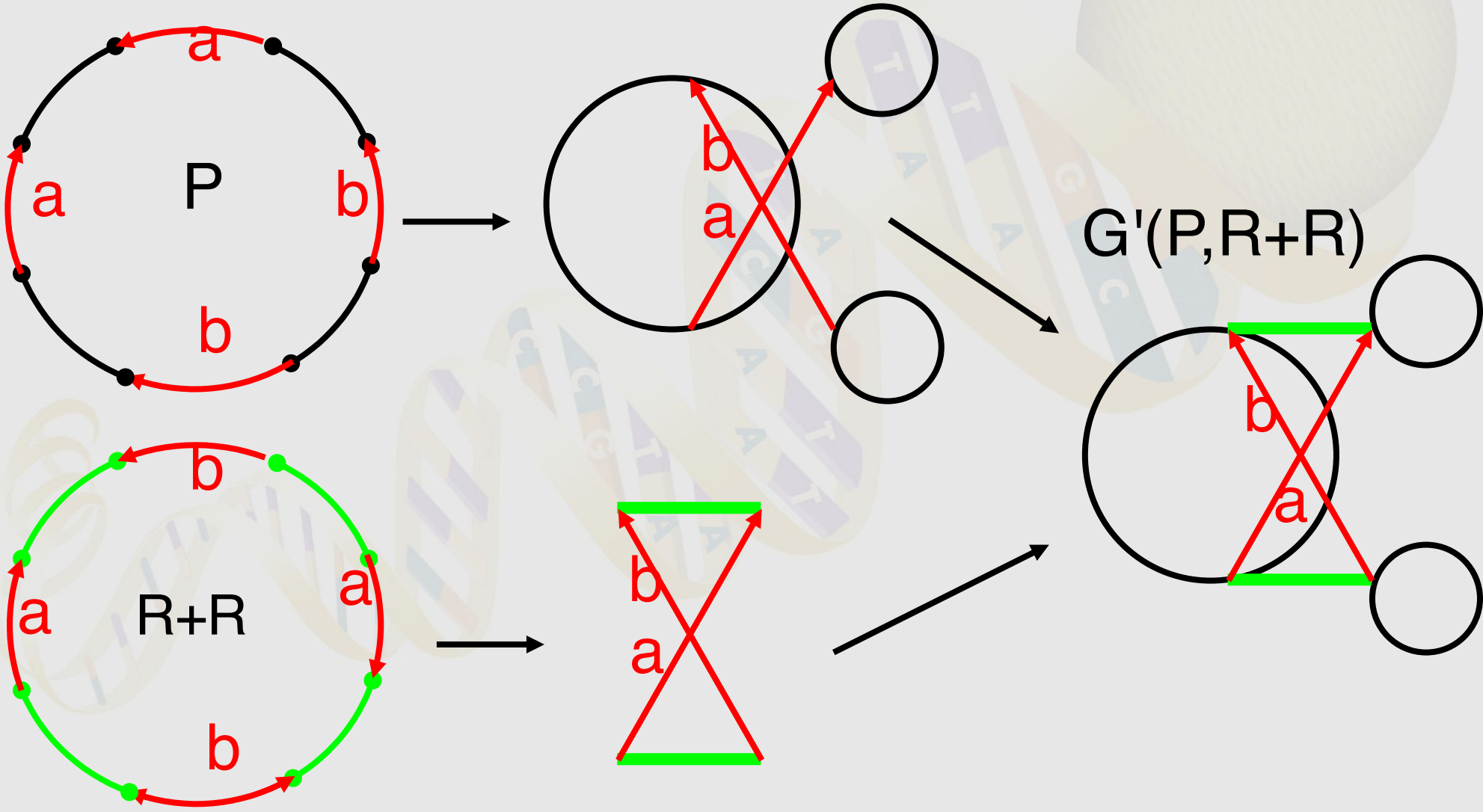
2-Break Genome Halving Problem

2-Break Genome Halving Problem: Given a duplicated genome ***P***, find a perfect duplicated genome ***R+R*** minimizing the 2-break distance:

$$d_2(P,R+R) = \#blocks - cycles(P,R+R)$$

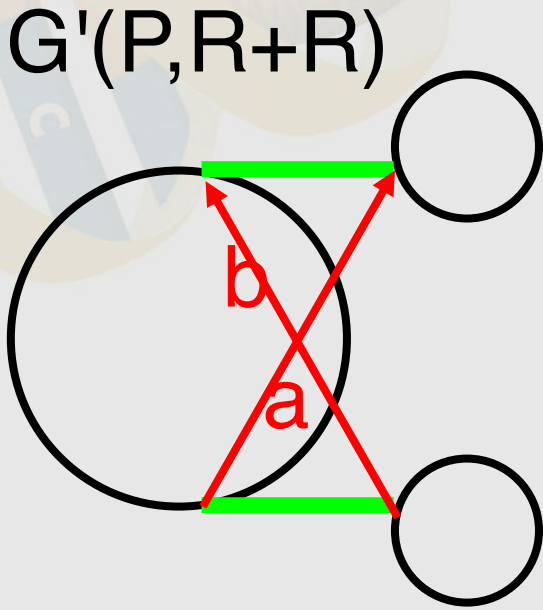
Minimizing $d_2(P,R+R)$ is equivalent to finding a perfect duplicated genome ***R+R*** and a labeling of ***P*** and ***R+R*** that maximizes ***cycles(P,R+R)***.

Contracted Breakpoint Graph of a Perfect Duplicated Genome

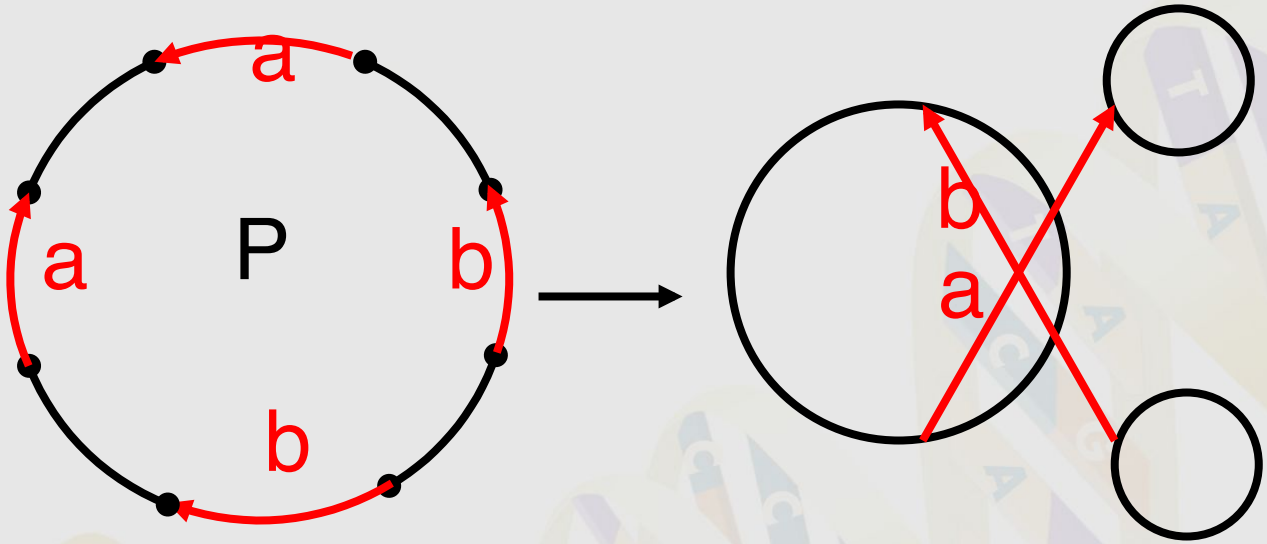


Contracted Breakpoint Graph of a Perfect Duplicated Genome has a Special Structure

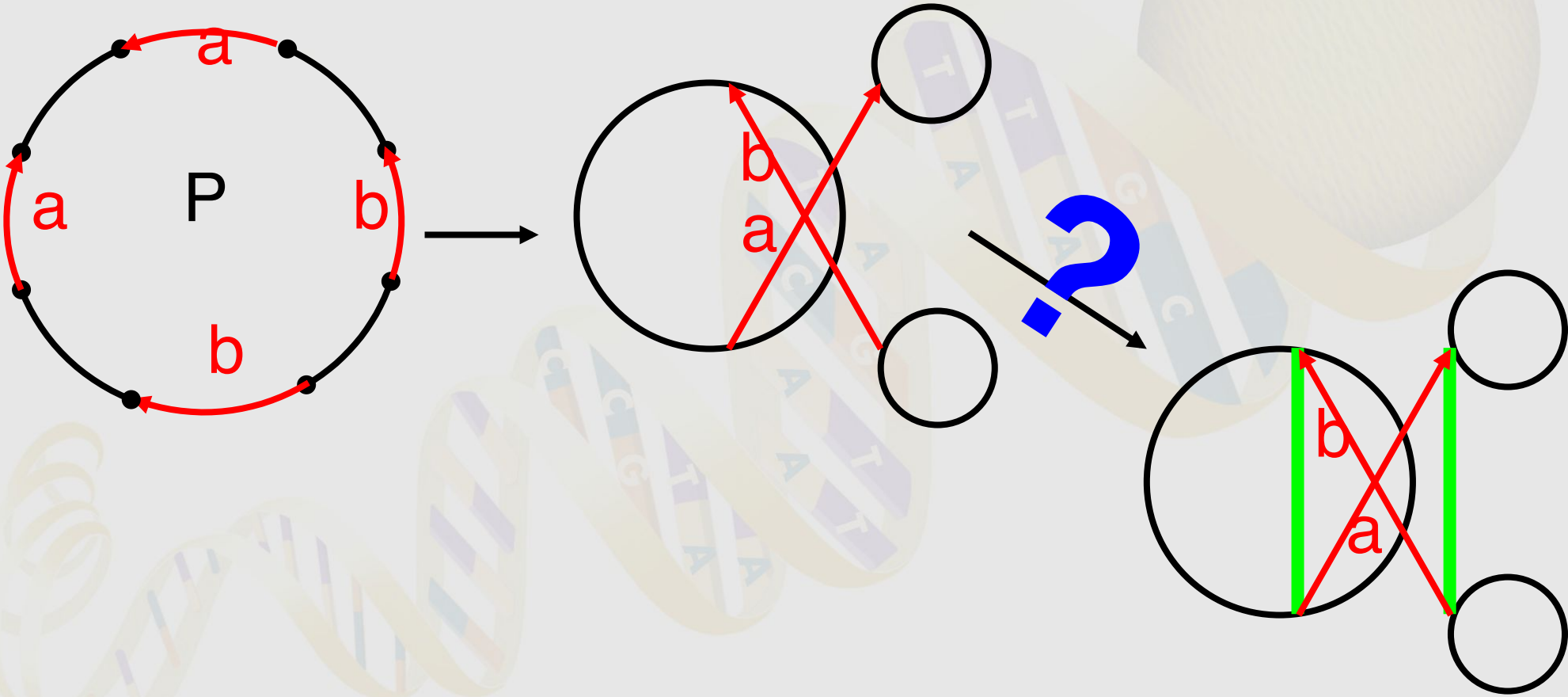
- ✓ *Red edges* form a *matching.*
- ✓ ***Black edges*** form *black cycles.*
- ✓ ***Green edges*** form cycles in general case but now (for R+R) these cycles are formed by parallel (double) green edges forming a matching



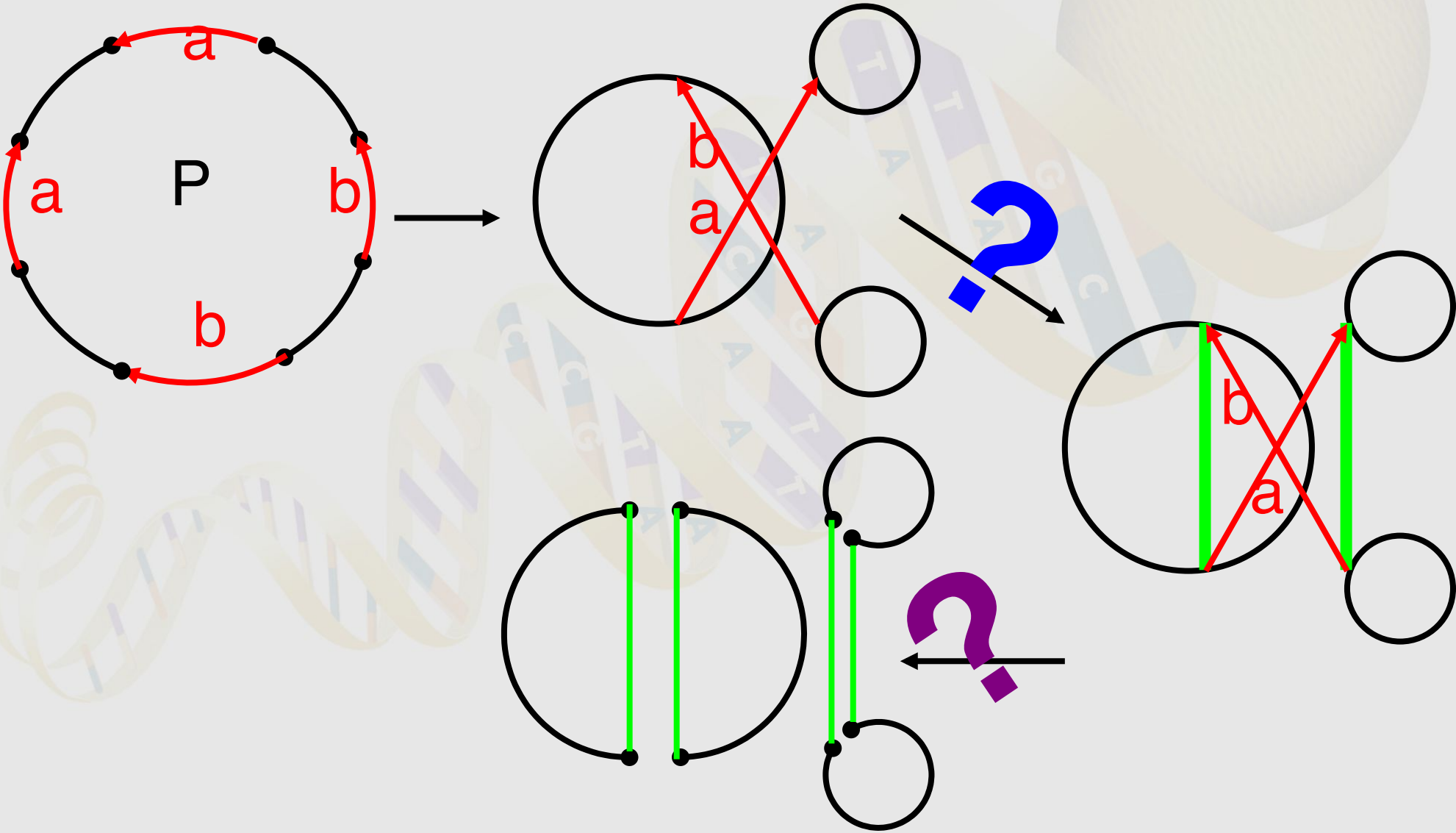
Finding the "Best" Contracted Breakpoint Graph for a Given Genome



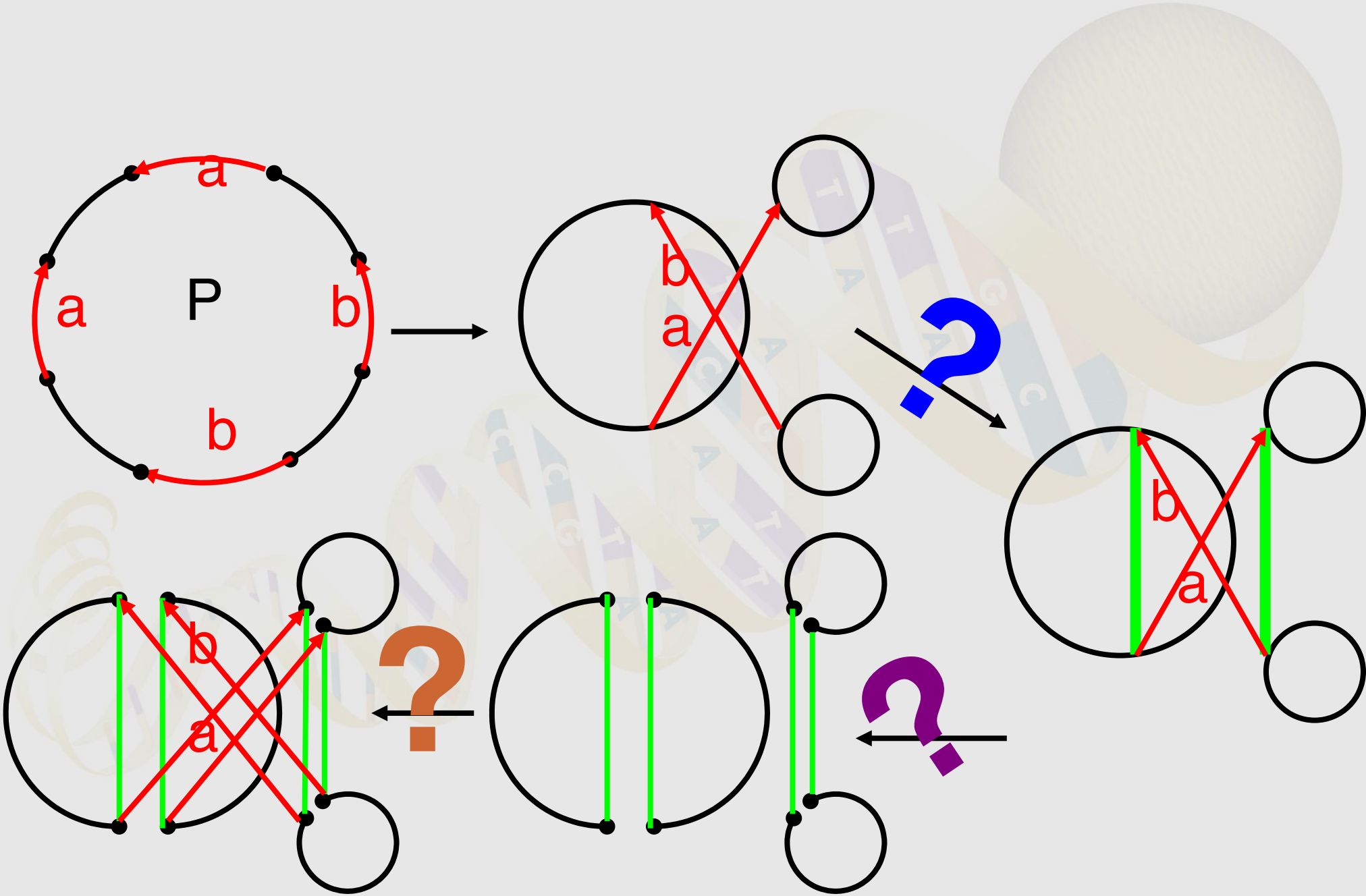
"Blue" Problem: Optimal Contracted Breakpoint Graph



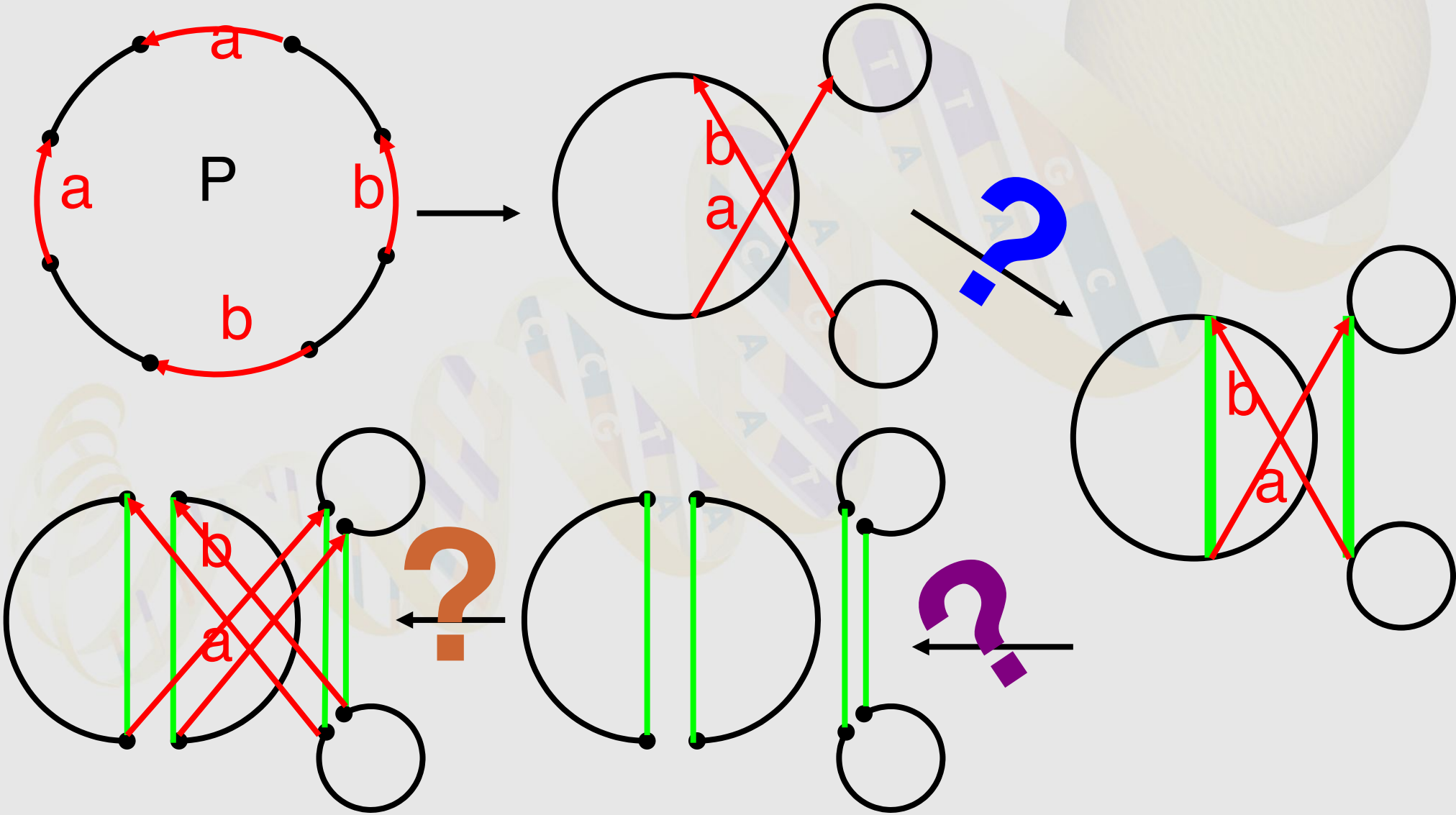
"Magenta" Problem: Maximum Cycle Decomposition



"Orange" Problem: Labeling



Solutions to ?, ?, and ? are too complex to be presented in this talk (SIAM J. Comp.2007)



Genome Halving Problem: What is Left Behind

(1) 2-Break Genome Halving Problem for Multichromosomal Genomes (*partially covered in this talk*)

(2) 2-Break Genome Halving Problem for Unichromosomal Genomes (more difficult)

(3) A Flaw in the El-Mabrouk–Sankoff Theorem for Unichromosomal Genomes

(4) Classification of Unichromosomal Genome (fixing the flaw in the El-Mabrouk–Sankoff Theorem)

(5) 3-Break Genome Halving Problem

Acknowledgments

Max Alekseyev

Acknowledgments

Max Alekseyev

David Sankoff

"If you are not criticized, you may not be doing much."

Donald Rumsfeld

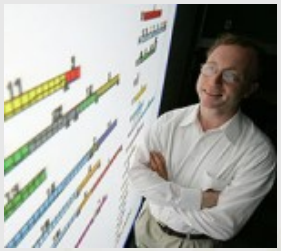
Acknowledgments

Max Alekseyev

David Sankoff

"If you are not criticized, you may not be doing much."

Donald Rumsfeld



Glenn Tesler, Math. Dept., UCSD

Qian Peng