

The background features a stylized DNA double helix in shades of yellow and blue. A circular chromosome is depicted on the right side, and a linear chromosome is shown in the center, both rendered in a similar color palette. The overall aesthetic is clean and scientific.

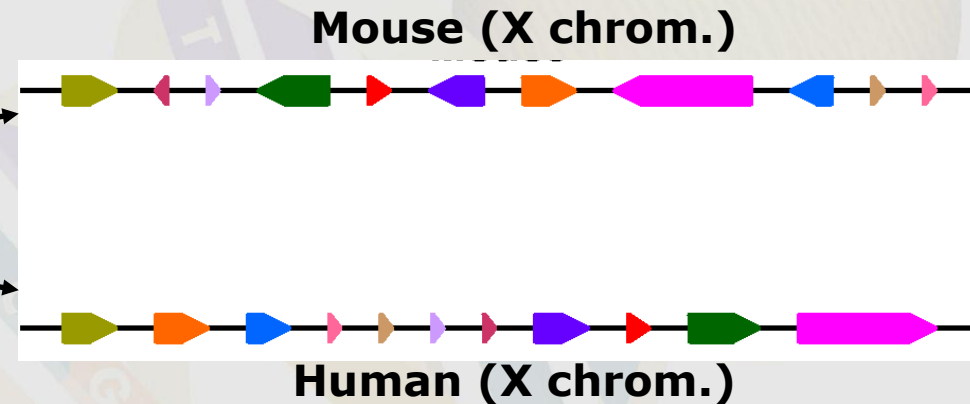
Multi-Break Rearrangements: from Circular to Linear Genomes

Max Alekseyev

University of California, San Diego
2007

Genome Rearrangements

**Unknown ancestor
~ 80 million years
ago**



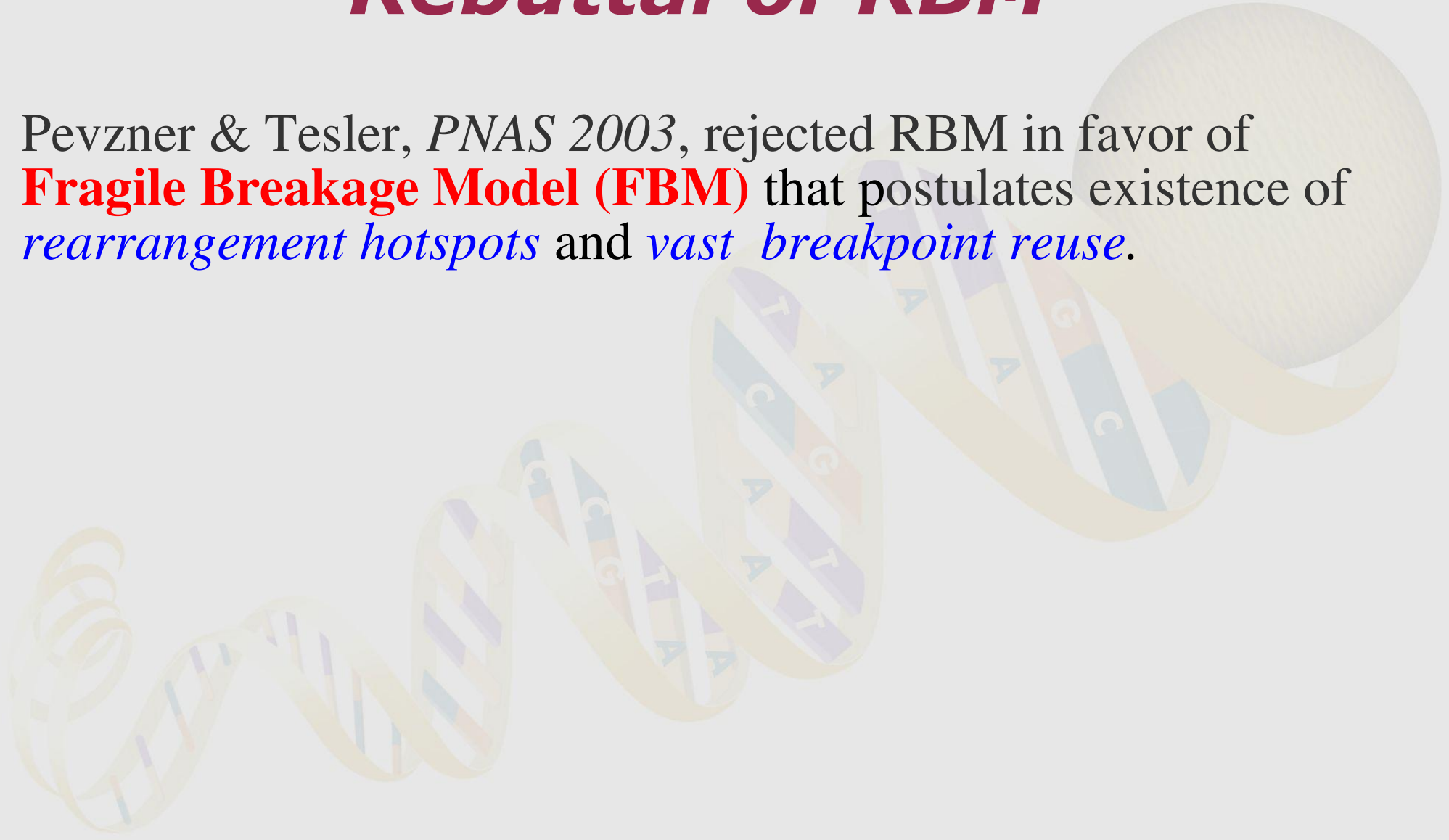
- ✓ What is the evolutionary scenario for transforming one genome into the other?
- ✓ Are there any rearrangement hotspots in mammalian genomes?

Random Breakage Model (RBM)

- ✓ **Random Breakage Model** (*Ohno, 1970*): Genomic architectures are shaped by rearrangements that occur randomly.
- ✓ *Nadeau & Taylor, 1984* gave first convincing arguments in favor of the RBM
- ✓ The random breakage hypothesis was embraced by biologists and has become *de facto* theory of chromosome evolution.
- ✓ RBM implies that there is no rearrangement hotspots

Rebuttal of RBM

- ✓ Pevzner & Tesler, *PNAS* 2003, rejected RBM in favor of **Fragile Breakage Model (FBM)** that postulates existence of *rearrangement hotspots* and *vast breakpoint reuse*.



Rebuttal of the Rebuttal of RBM

- ✓ Pevzner & Tesler, *PNAS* 2003, rejected RBM in favor of **Fragile Breakage Model (FBM)** that postulates existence of *rearrangement hotspots* and *vast breakpoint reuse*.
- ✓ Sankoff & Trinh, 2004, presented arguments against *Fragile Breakage Model*: “... 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.*”

Rebuttal of the Rebuttal of the Rebuttal of RBM

- ✓ Pevzner & Tesler, *PNAS* 2003, rejected RBM in favor of **Fragile Breakage Model (FBM)** that postulates existence of *rearrangement hotspots* and *vast breakpoint reuse*.
- ✓ Sankoff & Trinh, 2004, presented arguments against *Fragile Breakage Model*: “... 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.*”
- ✓ Peng et al., 2006, found an error in Sankoff-Trinh arguments: “*If Sankoff & Trinh fixed their ST-Synteny algorithm, they would confirm rather than reject Pevzner-Tesler's Fragile Breakage Model.*”

Random Breakage Model: Controversy Continues...

The rebuttal of the RBM led to a split among researchers:

- ✓ *Kikuta et al., 2007: “... the Nadeau and Taylor hypothesis is not possible for the explanation of synteny in general.”*
- ✓ *Ma et al., 2006: “Simulations ... suggest that this frequency of breakpoint reuse is approximately what one would expect if breakage was equally likely for every genomic position ...”*

Recent studies supporting 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*

Random Breakage Model and Complex Rearrangements

- ✓ Sankoff, 2006, claims that some complex rearrangement operations (like transpositions) may cause an appearance of fragile regions in the Pevzner-Tesler analysis.
- ✓ The rebuttal of RBM does not apply when there is a significant presence of complex rearrangements (in fact, that was acknowledged in Pevzner & Tesler, *PNAS* 2003).

Random Breakage Model and Complex Rearrangements

- ✓ *Pevzner & Tesler, 2003 argued that every evolutionary scenario for transforming *Mouse* into *Human* genome must have a very large number of *breakpoint re-uses*, a contradiction to RBM.*
- ✓ *Sankoff, 2006: “We cannot infer whether mutually randomized synteny block orderings derived from two divergent genomes were created ... **through processes other than reversals and translocations.**”*

Random Breakage Model Debate: Algorithmic Challenge

- ✓ Algorithmic theory for transpositions remains undeveloped, thus, making analysis of breakpoint reuse difficult.
- ✓ We bypass this challenge by introducing more general multi-break rearrangements and solving the multi-break distance problem.
- ✓ The standard rearrangement operations (reversals, translocations, fusions, and fissions) make *2 breaks* in a genome and glue the resulting pieces in a new order.
- ✓ *k-Break* rearrangement operation makes *k breaks* in a genome and glues the resulting pieces in a new order. Transpositions represent 3-breaks.

Random Breakage Model Debate: Our Contribution

- ✓ **Algorithmic analysis of k-breaks.** The duality theorem for the multi-break distance problem and a linear-time algorithm for the computing multi-break distance.
(**Theoretical Computer Science, 2007, in press**)
- ✓ **Analysis of breakpoint re-use in the case of k-breaks.** For $k=3$ (the most relevant case in evolutionary studies), we showed that even if 3-break rearrangements were frequent, the Pevzner-Tesler argument against RBM still stands.
(**PLoS Computational Biology, 2007, in press**)

Multi-Breaks and Linear Genomes: Algorithmic Challenges

- ✓ The above results address the case of circular genomes (in which every chromosome is circular) but not the (more relevant) case of linear genomes.
- ✓ Multi-break rearrangements in linear genomes are much harder to analyze than in circular genomes.
- ✓ This work extends results from Alekseyev & Pevzner, *Theor. Comput. Sci.* 2007 and *PLoS Comput. Biol.* 2007, to linear genomes.

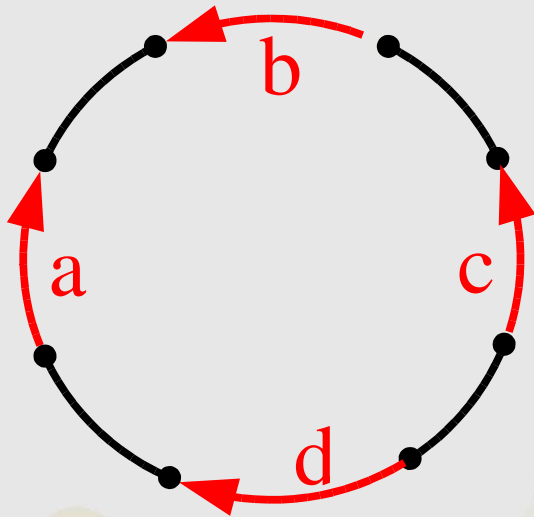


Genome Rearrangements

Genomic Distance Story

- ✓ **Genomic Distance** between two genomes is the minimum number of *reversals*, *translocations*, *fusions*, and *fissions* required to transform one genome into the other.
- ✓ Hannenhalli and Pevzner (FOCS 1995) were first to give a polynomial-time algorithm for computing the reversal distance (between unichromosomal genomes)
- ✓ Hannenhalli and Pevzner (STOC 1995) further extended their algorithm to computing the genomic distance (between multichromosomal genomes)
- ✓ These algorithms were followed by many improvements: *Kaplan et al. 1999, Bader et al. 2001, Tannier & Sagot 2001, Tesler 2002, Ozery-Flato & Shamir 2003, Bergeron 2001-07, etc.*
- ✓ Nearly all rearrangement studies are based on the notion of the **breakpoint graph** introduced by Bafna and Pevzner (FOCS 1994)

Representations of Circular Chromosomes

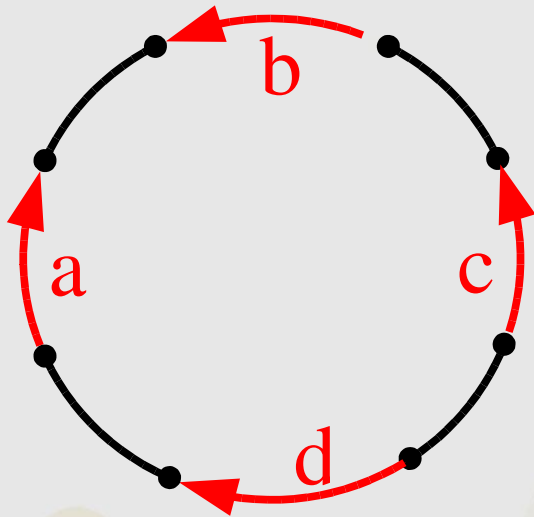


$$P = (+a - b - c + d)$$

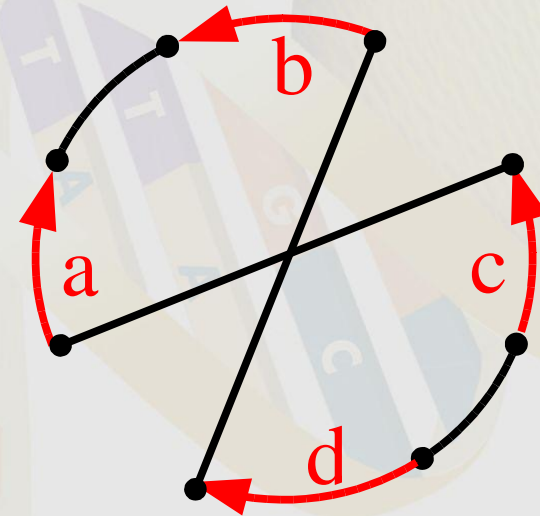
A chromosome can be represented as:

- ✓ a cycle with directed red and undirected black edges, where red edges encode genes and adjacent genes are connected with black edges

Reversals on Circular Chromosomes



reversal

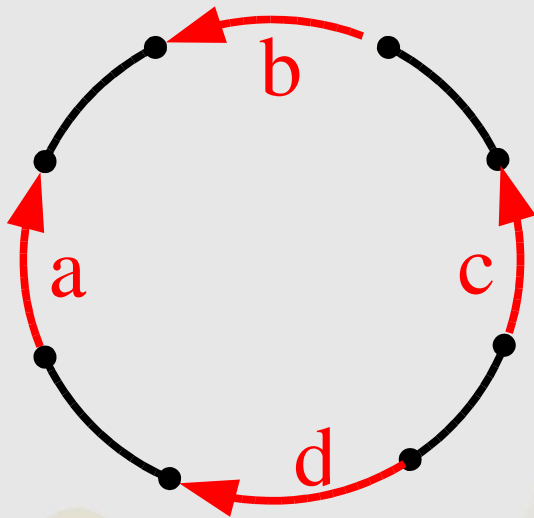


$$P = (+a - b - \underline{c} + d)$$

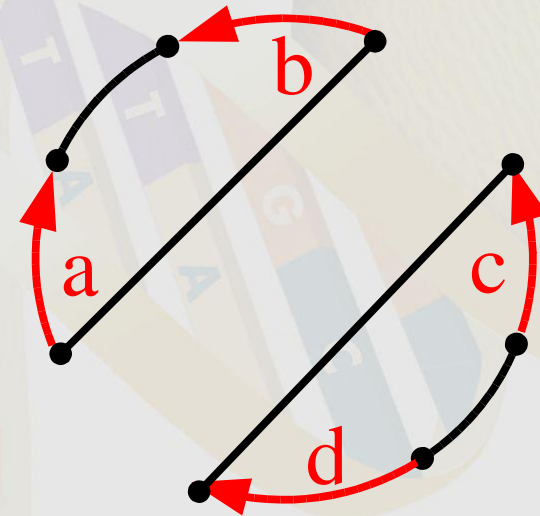
$$Q = (+a - b - d + c)$$

Reversals replace two black edges with two other black edges

Fissions



fission



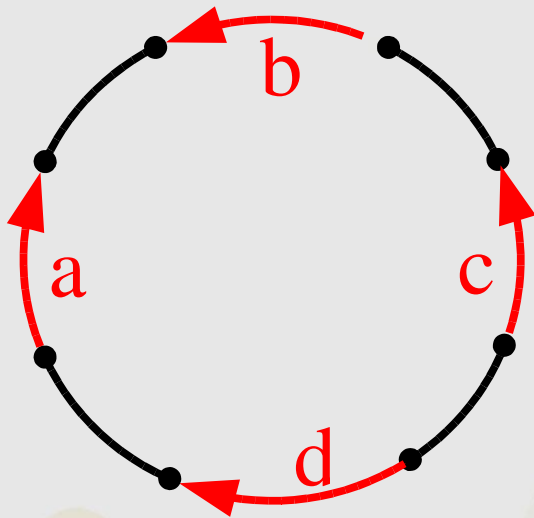
$$P = (+a - b - c + d)$$

$$Q = (+a - b) (-c + d)$$

Fissions split a single cycle (chromosome) into two.

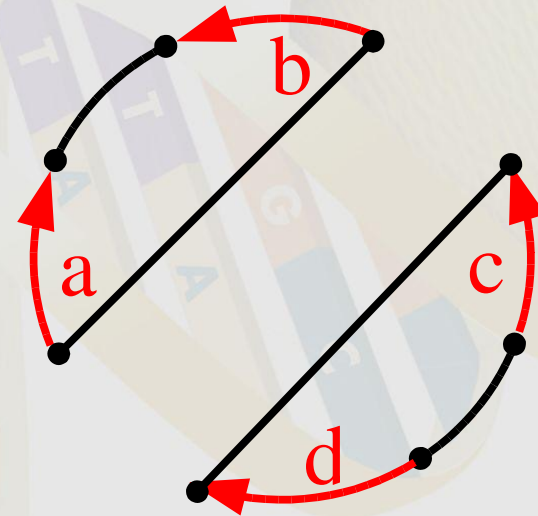
Fissions replace two black edges with two other black edges.

Translocations / Fusions



$$P = (+a - b - c + d)$$

fusion

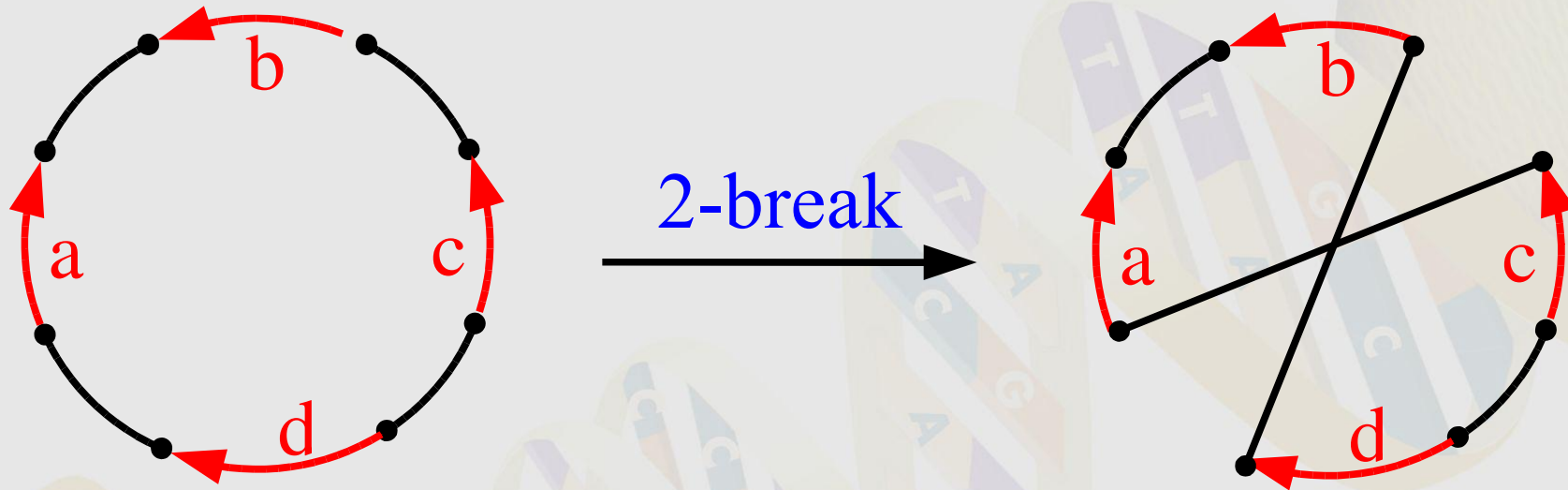


$$Q = (+a - b) (-c + d)$$

Translocations/Fusions transform two cycles (chromosomes) into a single one.

They also replace two black edges with two other black edges.

2-Breaks



$$P = (+a - b - c + d)$$

$$Q = (+a - b - d + c)$$

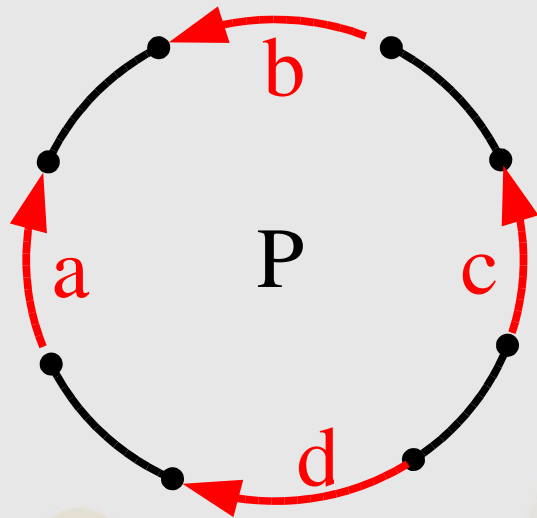
2-Break replaces *any pair* of black edges with another pair forming matching on the same 4 vertices.

Reversals, translocations, fusions, and fissions represent all different types of 2-breaks.

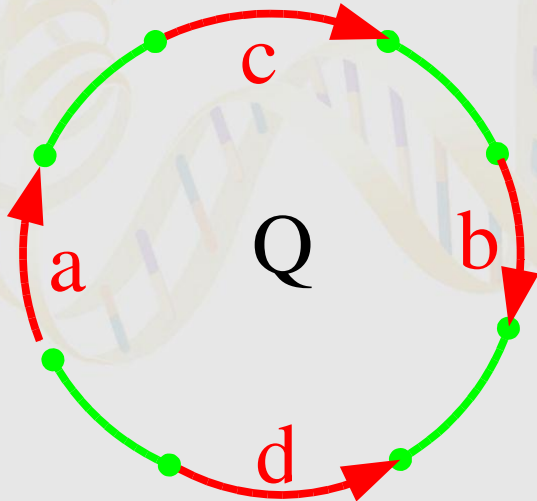
2-Break Distance

- ✓ The **2-Break distance** $d_2(P, Q)$ between circular genomes P and Q is the minimum number of 2-breaks required to transform P into Q .
- ✓ In difference from the genomic distance (between linear genomes), the 2-break distance (between circular genomes) is easy to compute.

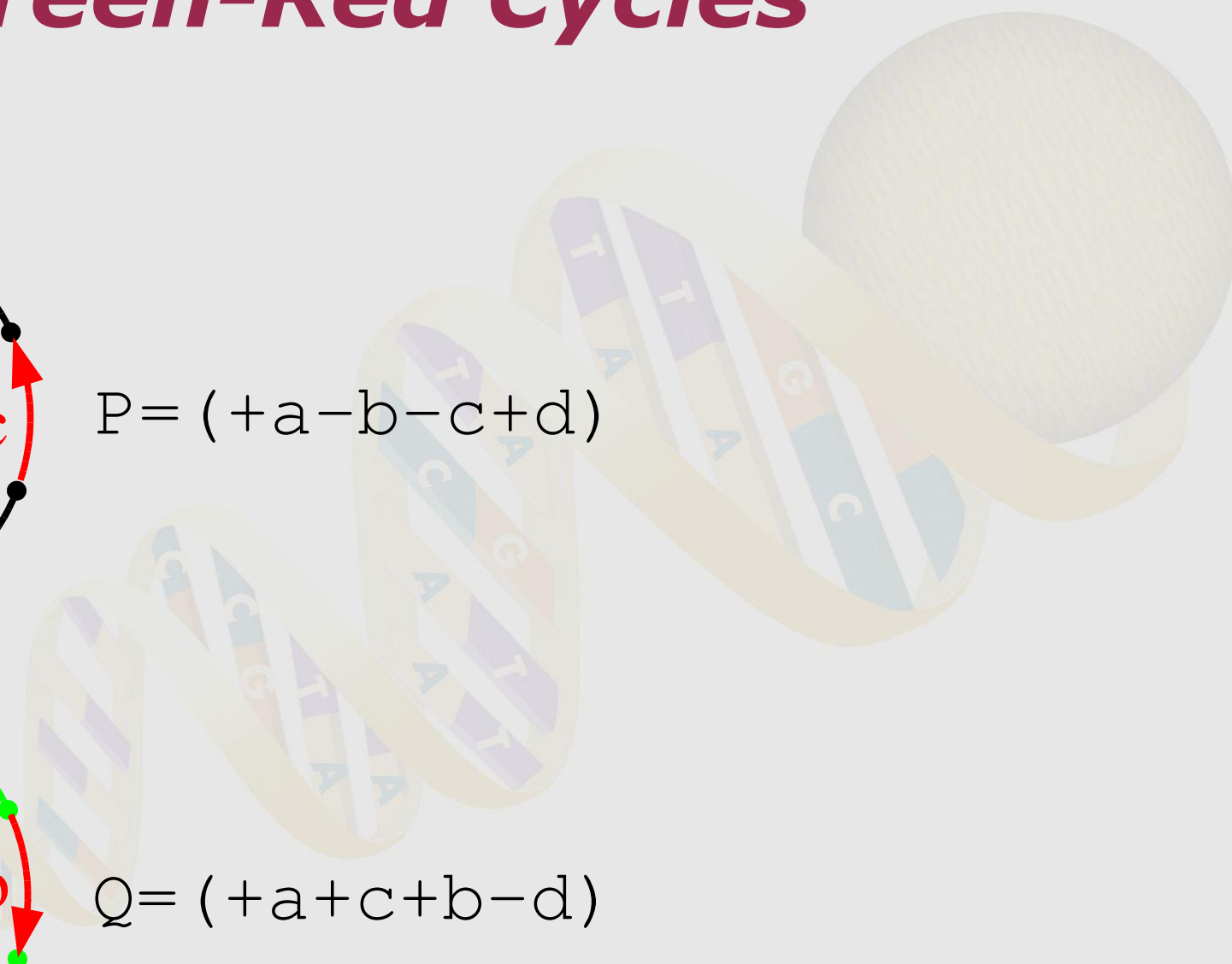
Two Genomes as Black-Red and Green-Red Cycles



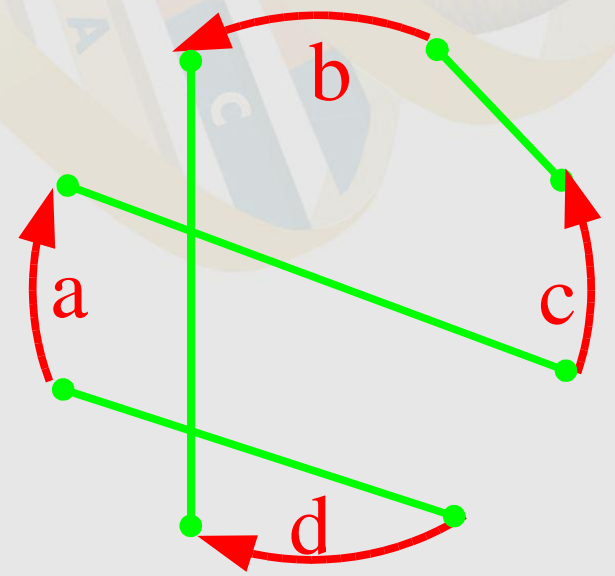
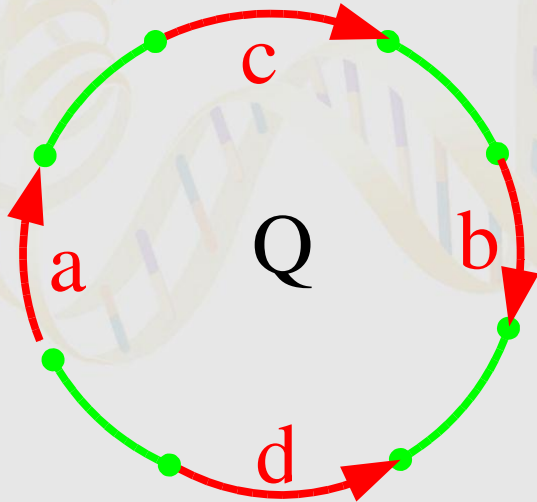
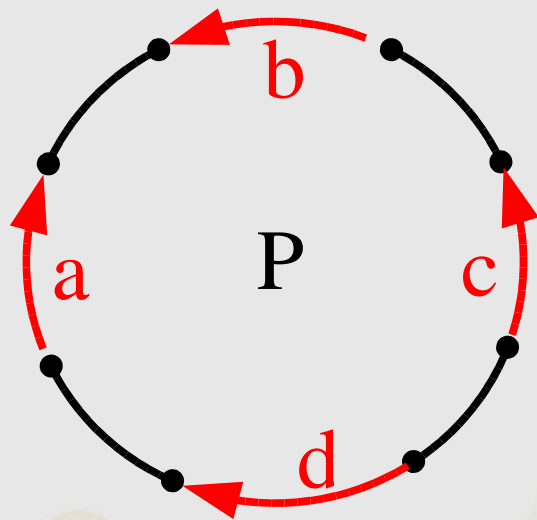
$$P = (+a - b - c + d)$$



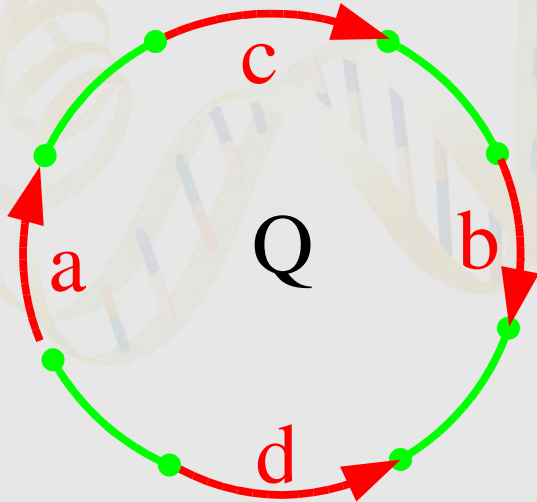
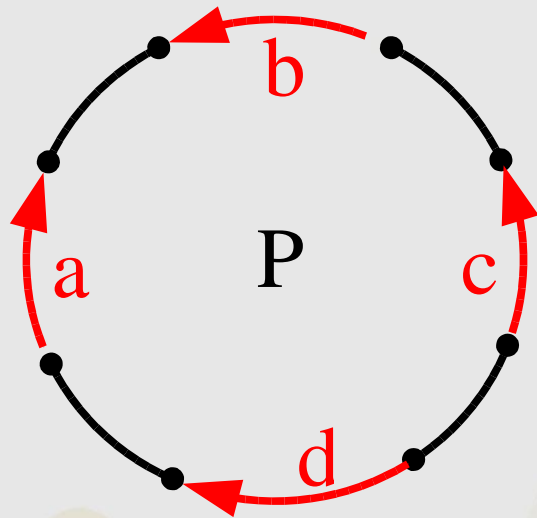
$$Q = (+a + c + b - d)$$



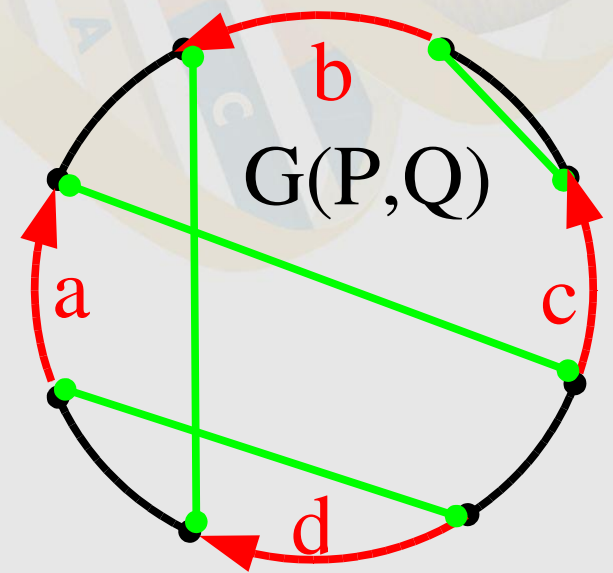
Rearranging Q in the P order



New Interpretation of the Breakpoint Graph: Gluing Red Edges with the Same Labels



Breakpoint Graph

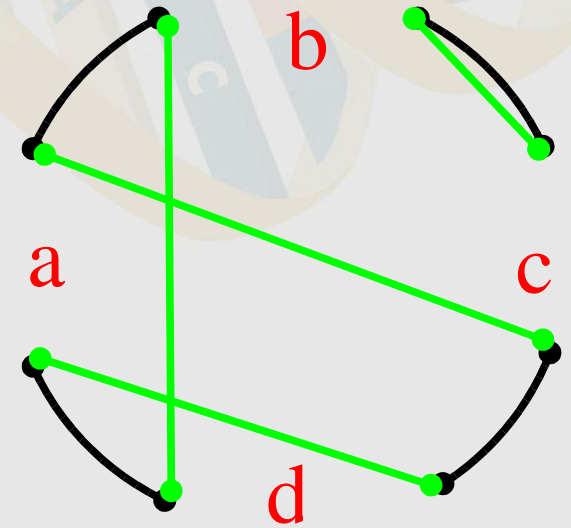


Black-Green Cycles

- ✓ Number of black-green cycles in the breakpoint graph:

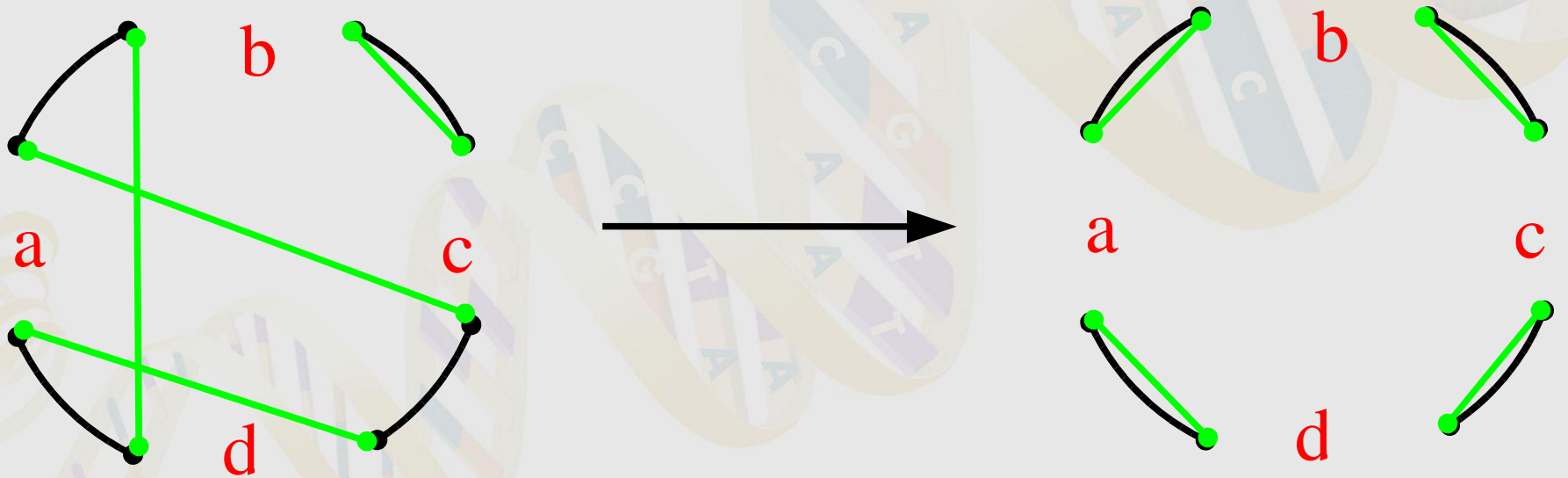
$\text{cycle}(P, Q)$

is the key parameter in computing the 2-break distance.



Rearrangements Change Cycles

Transforming genome Q into genome P by 2-breaks corresponds to transforming black-green cycles in $G(P,Q)$ into *trivial cycles* in $G(P,P)$.



$\text{cycle}(P,Q) = 2$ cycles

$\text{cycle}(P,P) = 4$ *trivial* cycles

Sorting by 2-Breaks

$$Q=Q_0 \xrightarrow{\text{2-breaks}} Q_1 \rightarrow \dots \rightarrow Q_d=P$$

$$G(P,Q) \rightarrow G(P,Q_1) \rightarrow \dots \rightarrow G(P,P)$$

$$\text{cycle}(P,Q) \text{ cycles} \rightarrow \dots \rightarrow |P| \text{ cycles}$$

of black-green cycles increased by $|P| - \text{cycle}(P,Q)$

How much each 2-break can contribute to this increase?

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: $\Delta_{\text{cycle}} \leq 2 - 1 = 1$.

2-Break Distance

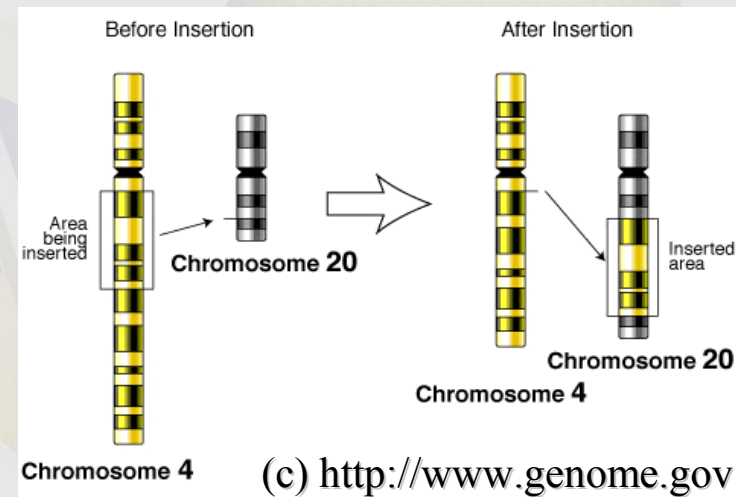
- ✓ Any 2-Break increases the number of cycles by at most one ($\Delta\text{cycle} \leq 1$)
- ✓ Any non-trivial cycle can be split into two cycles with a 2-break ($\Delta\text{cycle} = 1$)
- ✓ Every sorting by 2-break must increase the number of cycles by $|P| - \text{cycle}(P,Q)$
- ✓ The **2-Break Distance** between genomes P and Q:

$$d_2(P,Q) = |P| - \text{cycle}(P,Q)$$

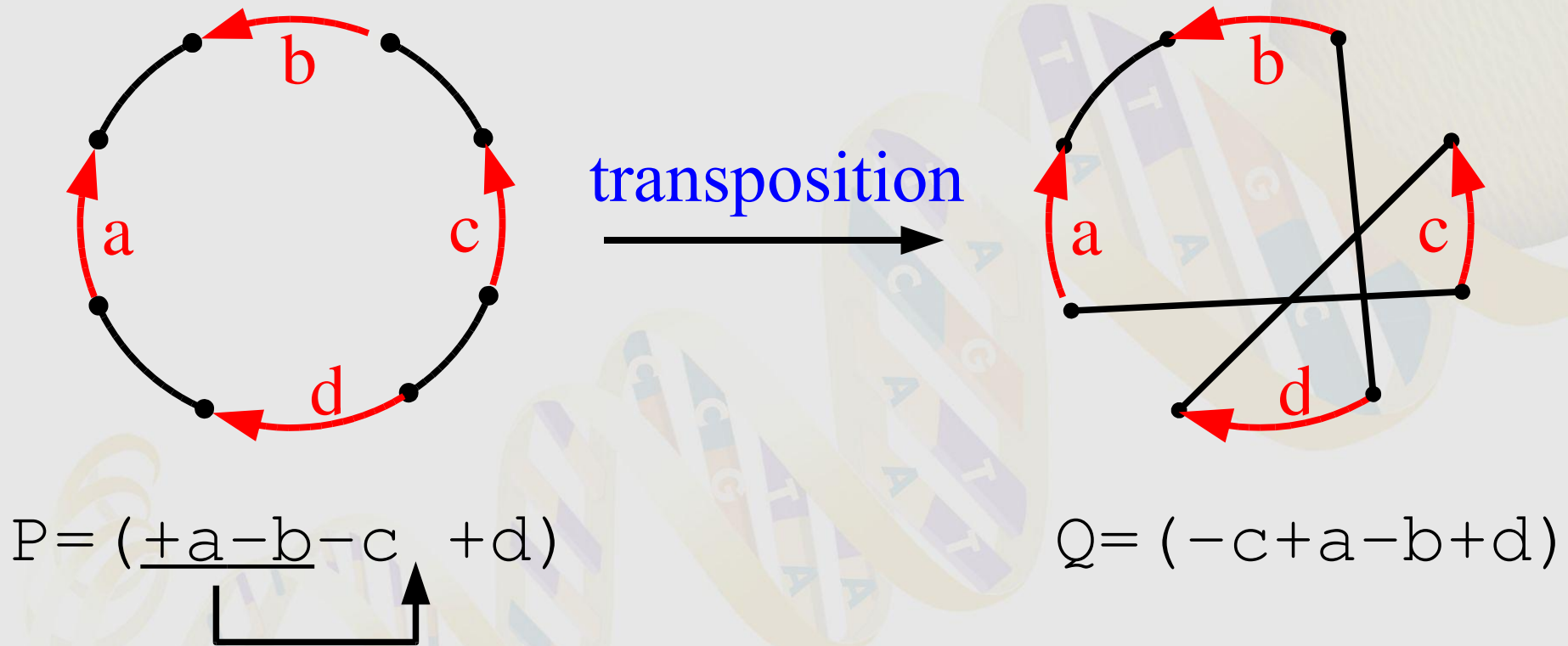
(cp. *Yancopoulos et al., 2005, Bergeron et al., 2006*)

Transpositions

- ✓ **Sorting by Transpositions:**
Given two genomes, find the shortest sequence of transpositions transforming one genome into the other
- ✓ First 1.5-approximation algorithm was given by Bafna and Pevzner (*SODA 1995*)
- ✓ Recent achievement: 1.375-approximation algorithm due to Elias and Hartman (*WABI 2005*)
- ✓ The complexity status remains unknown

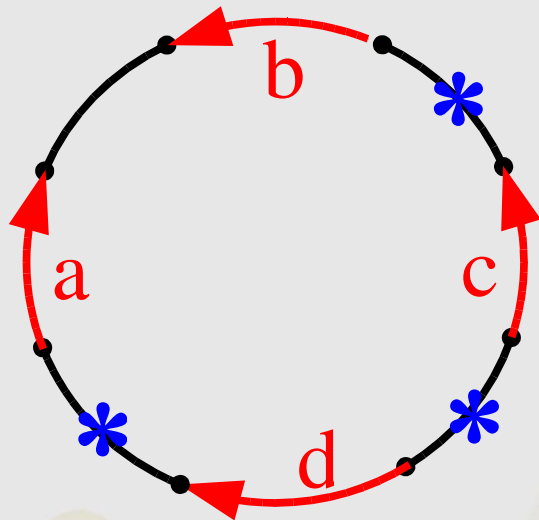


Transpositions

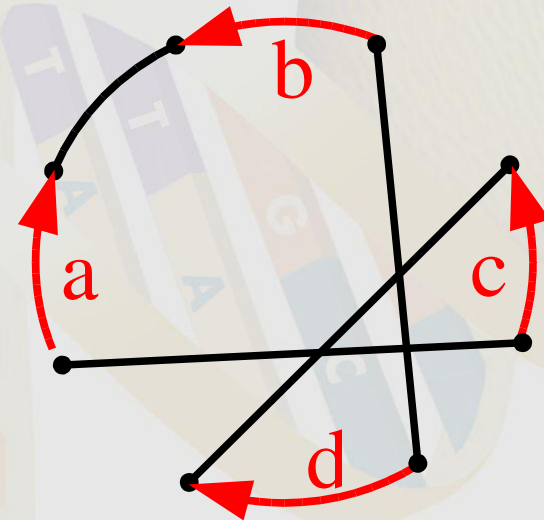


Transpositions cut off a segment of one chromosome and insert it at some position in the same or another chromosome

3-Breaks



3-break



$$P = (+a - b - c + d)$$

$$Q = (-c + a - b + d)$$

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

Transpositions are 3-Breaks.

3-Break Distance

- ✓ The **3-Break Distance** $d_3(P,Q)$ between genomes P and Q is the minimum number of 3-Breaks required to transform P into Q .
- ✓ 3-Break Distance between genomes P and Q (*TCS07*):

$$d_3(P,Q) = (|P| - \text{cycle}^{\text{odd}}(P,Q)) / 2$$

k-Break Distance

- ✓ Exact formulas for $d_k(P, Q)$ becomes complex as k grows, e.g.:

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.

- ✓ We estimate that the formula for the 20-break distance contains over 1,500 terms.
- ✓ We effectively solved the k -break distance problem for an arbitrary k (TCS07).



***From Circular Genomes
to
Linear Genomes***

Circularization of Linear Genomes

- ✓ Graph of a genome P with n linear chromosomes consist of n alternating black-red paths with $2n$ endpoints.
- ✓ If we introduce an arbitrary perfect black matching on these $2n$ endpoints, the resulting graph will represent a circular genome, called circularization or **closure** of P .
- ✓ Rearrangements in a linear genome correspond to 3-breaks in its closure.

Rearrangement Distance between Linear Genomes

- ✓ The k -break distance between closures gives a lower bound for the rearrangement distance between linear genomes P and Q :

$$d_k^{\text{linear}}(P, Q) \geq \max_{P'} \min_{Q'} d_k(P, Q)$$

$$d_k^{\text{linear}}(P, Q) \geq \max_{Q'} \min_{P'} d_k(P, Q)$$

where P' and Q' vary over all possible closures of P and Q respectively.

Rearrangement Distance between Human and Mouse

- ✓ Solving this max-min problem for *Human* and *Mouse* genomes:

$$d_2^{\text{linear}}(\text{H}, \text{M}) \geq 233$$

$$d_3^{\text{linear}}(\text{H}, \text{M}) \geq 137$$

- ✓ The genomic distance between Human and Mouse is $d_2^{\text{linear}}(\text{H}, \text{M}) = 245$ (Pevzner & Tesler, *Genome Res.* 2003).



***Complex Rearrangements
and
Breakpoint Re-use***

Transforming Circularized Mouse into Human by 3-Breaks

$$\text{Mouse} = Q_0 \xrightarrow{\text{3-breaks}} Q_1 \rightarrow \dots \rightarrow Q_d = \text{Human}$$

$$d = d_3(\text{Human}, \text{Mouse}) = 139$$

- ✓ Each 3-break makes 3 breaks, hence the total number of breaks in this transformation is $3 * 139 = 417$.
- ✓ 417 is much larger than 281 (the observed number of breakpoints between *Human* and *Mouse* genomes), implying that there is high breakpoint re-use inconsistent with RBM.

Transforming Circularized Mouse into Human by 3-Breaks

$$\text{Mouse}=\text{Q}_0 \xrightarrow{\text{3-breaks}} \text{Q}_1 \rightarrow \dots \rightarrow \text{Q}_d=\text{Human}$$

$$d = d_3(\text{Human}, \text{Mouse}) = 139$$

- ✓ Each 3-break makes 3 breaks, hence the total number of breaks in this transformation is $3*139 = 417$.
- ✓ **OOPS!** Some of these 3-breaks may be actually 2-breaks making only 2 breaks each.
If every 3-break in the series were a 2-break then the total number of breaks is $2*139=278 < 281$, in which case there could be no breakpoint re-uses at all.

Minimum Number of Breaks

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

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

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

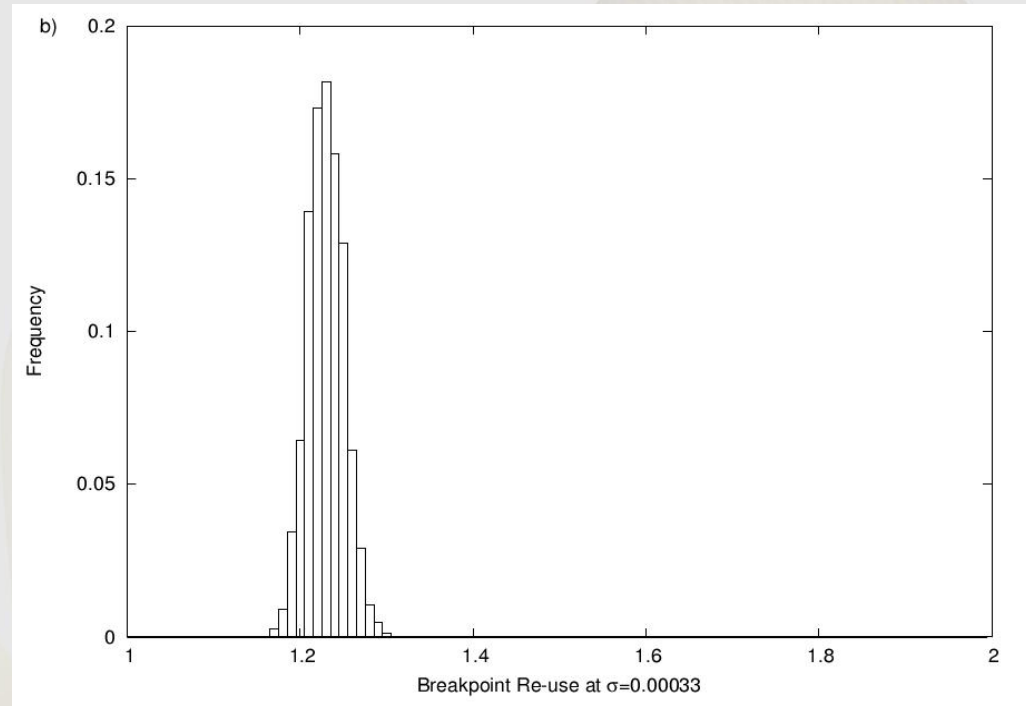
$$d_2(\text{Human}, \text{Mouse}) = 246$$

$$d_3(\text{Human}, \text{Mouse}) = 139$$

$$\text{minimum number of breaks} = 385$$

Breakpoint Re-use between Human and Mouse Genomes

✓ Any transformation of *Mouse* into *Human* genome with 3-breaks requires at least 385 (370 in the linear case) breaks, 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 RBM.

✓ Mean = 1.23

✓ Standard deviation = 0.02

✓ Maximum breakpoint reuse rate = 1.33 (observed once in 100,000 simulations)

Breakpoint Re-use between Linear Genomes

- ✓ The number of breaks in a series of rearrangements transforming a linear genome P into a linear genome Q is bounded as:

$$\#breaks(P,Q) \geq \max_{P'} \min_{Q'} d_2(P,Q) + d_3(P,Q)$$

$$\#breaks(P,Q) \geq \max_{Q'} \min_{P'} d_2(P,Q) + d_3(P,Q)$$

where P' and Q' vary over all possible closures of P and Q respectively.

- ✓ This max-min problem was solved exactly.

Breakpoint Re-use between Human and Mouse

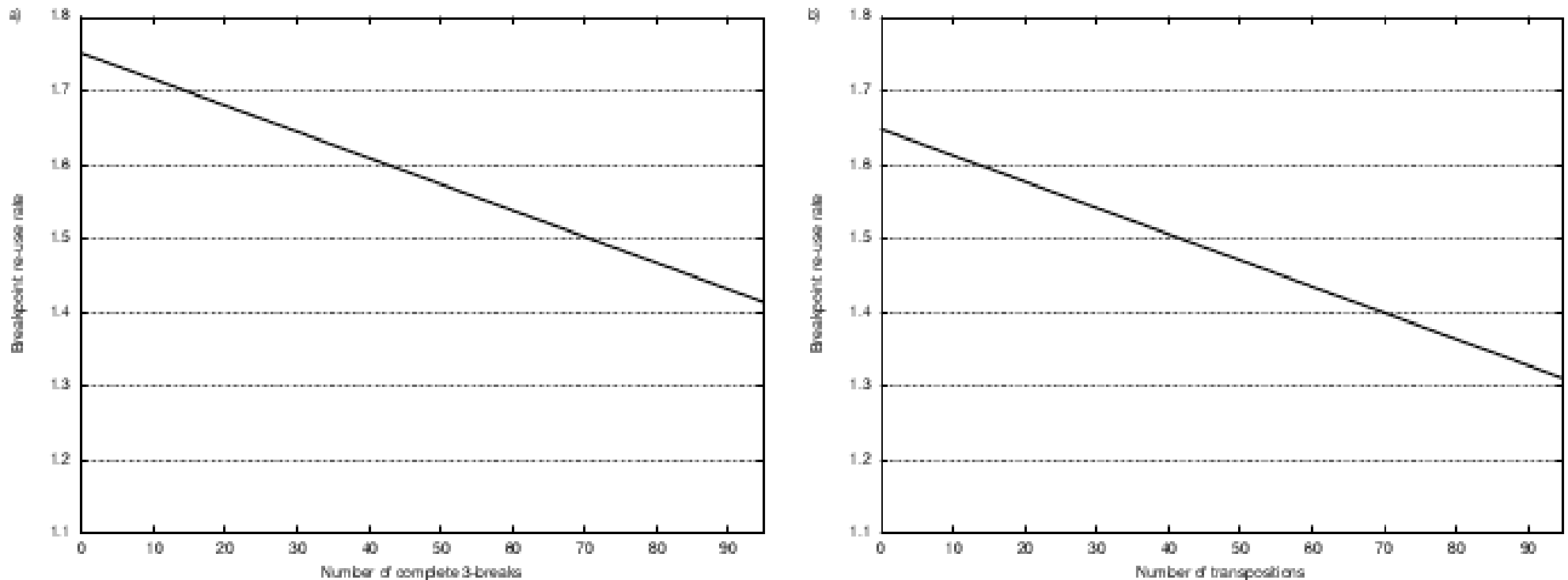
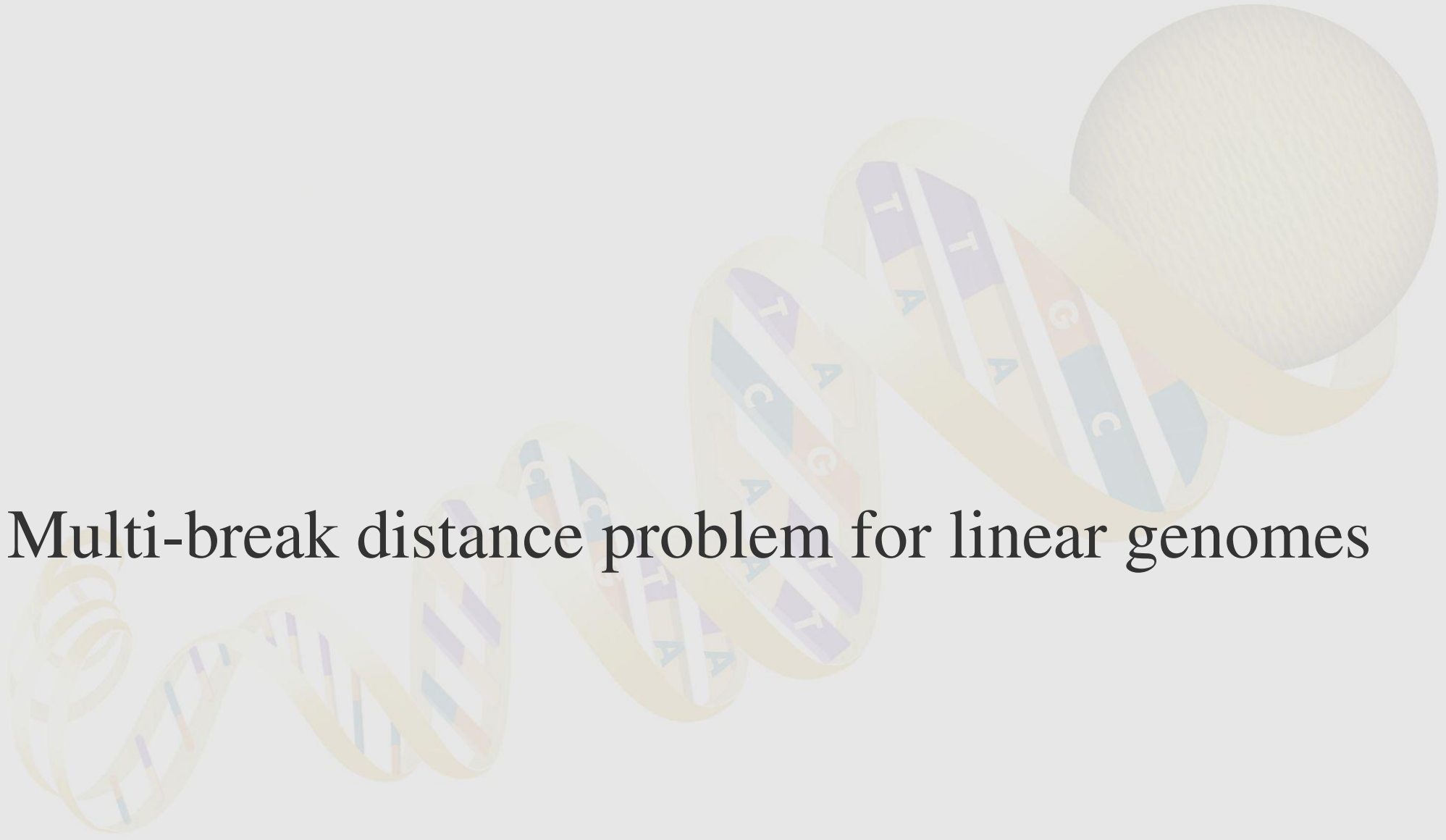


Fig. 3: The lower bound on the breakpoint re-use rate for human and mouse genomes based on 281 synteny blocks from [40]. The lower bound is represented as a function of a) the number of complete 3-breaks in a series of 3-breaks between the circularized human and mouse genomes. (Reproduced from [2]). b) the number of transpositions in a series of rearrangements between the linear human and mouse genomes.

Open Problem

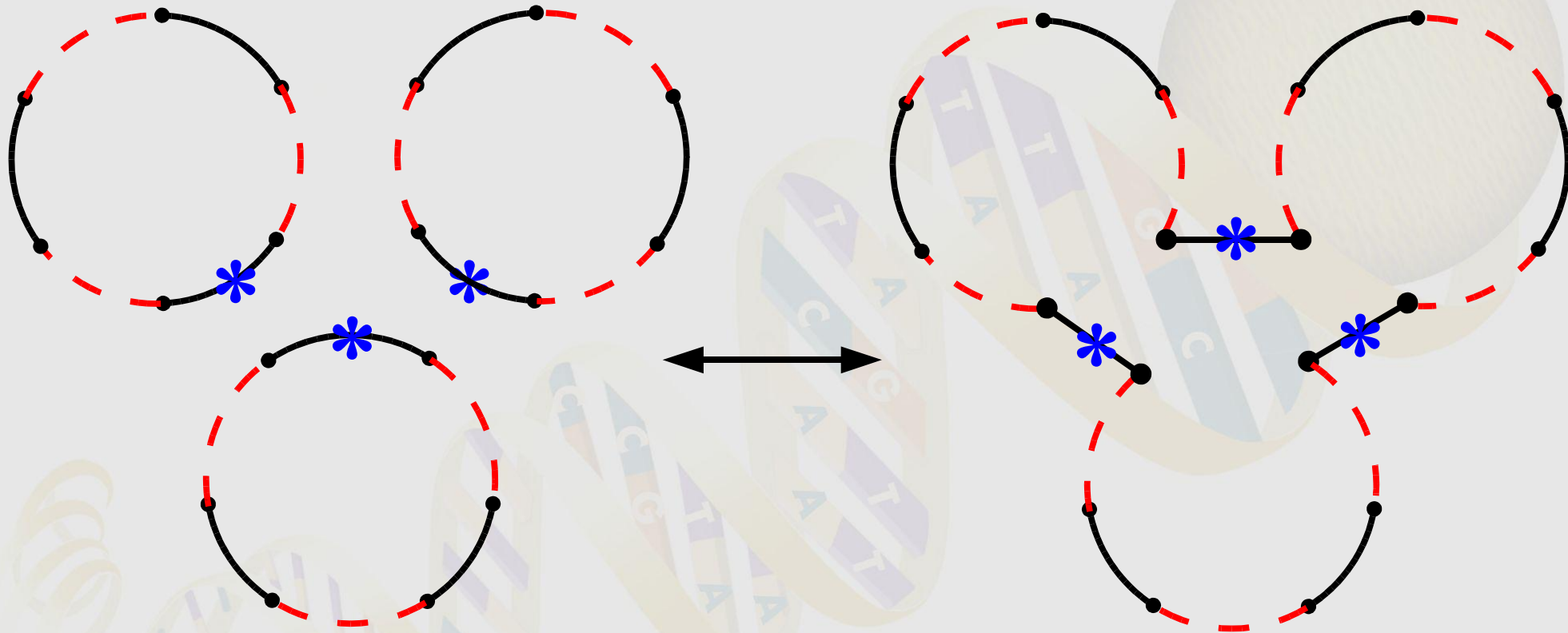
- ✓ Multi-break distance problem for linear genomes



A 3D illustration of a DNA double helix structure. The two strands are represented by yellow ribbons that spiral around each other. Between the strands are horizontal bars representing nitrogenous bases, colored in purple, blue, orange, and light blue. Some of these bars have white letters 'A', 'T', 'C', and 'G' on them. At the right end of the helix, there is a large, textured, light-brown sphere. The entire scene is set against a light gray background.

Thank You!

3-Breaks include 3-Way Fusions and Fissions



3-Breaks can merge three chromosomes into a single one as well as split a single chromosome into three ones.

3-Break Distance: Focus on Odd Cycles

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

$$d_3(P, Q) = (|P| - \text{cycle}^{\text{odd}}(P, Q)) / 2$$