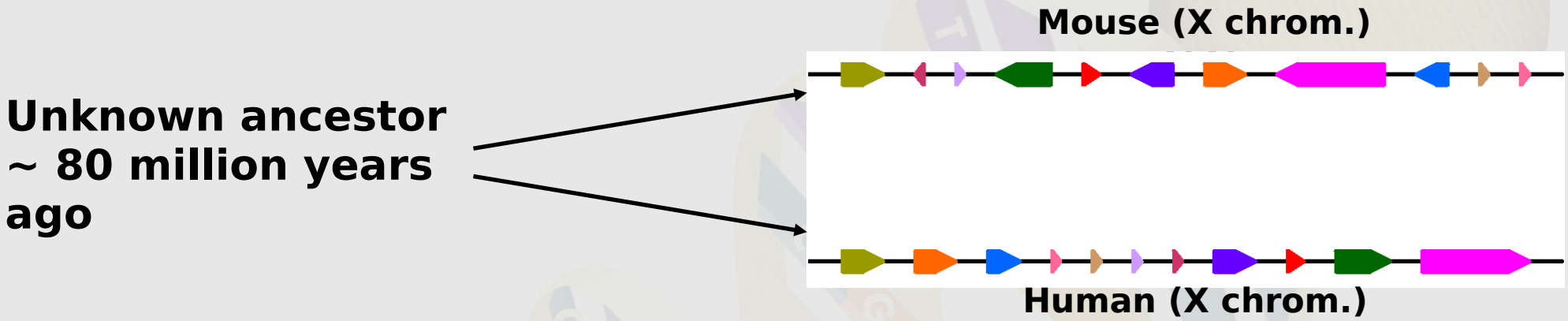
Multi-Break Rearrangements and Chromosomal Evolution

Max Alekseyev

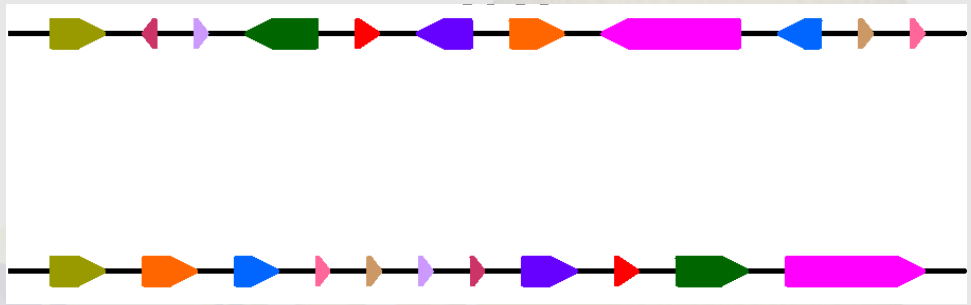
Genome Rearrangements



- What is the architecture of the ancestral genome?
- What is the evolutionary scenario for transforming one genome into the other?

Genome Rearrangements

Mouse X chromosome

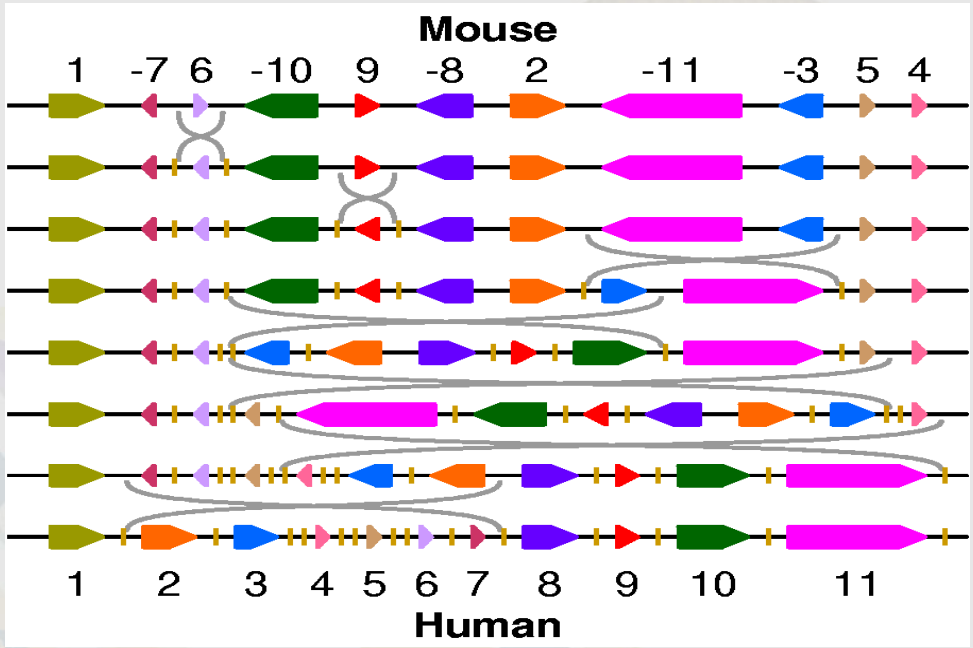


Unknown ancestor
~ 80 M years ago

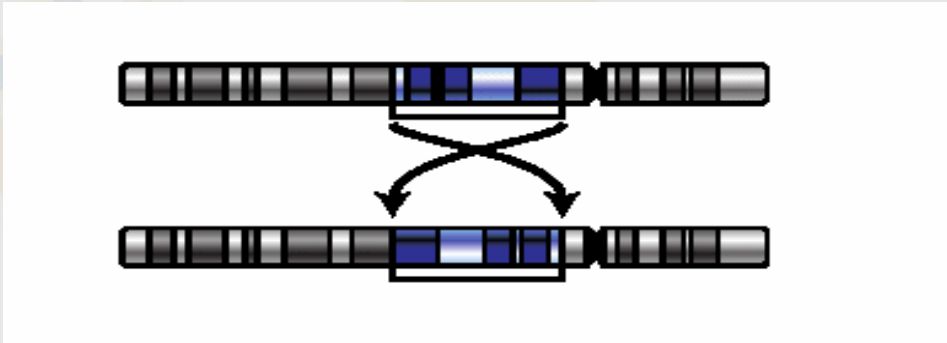
Human X chromosome

Genome Rearrangements: Evolutionary Scenarios

Unknown ancestor
~ 80 M years ago



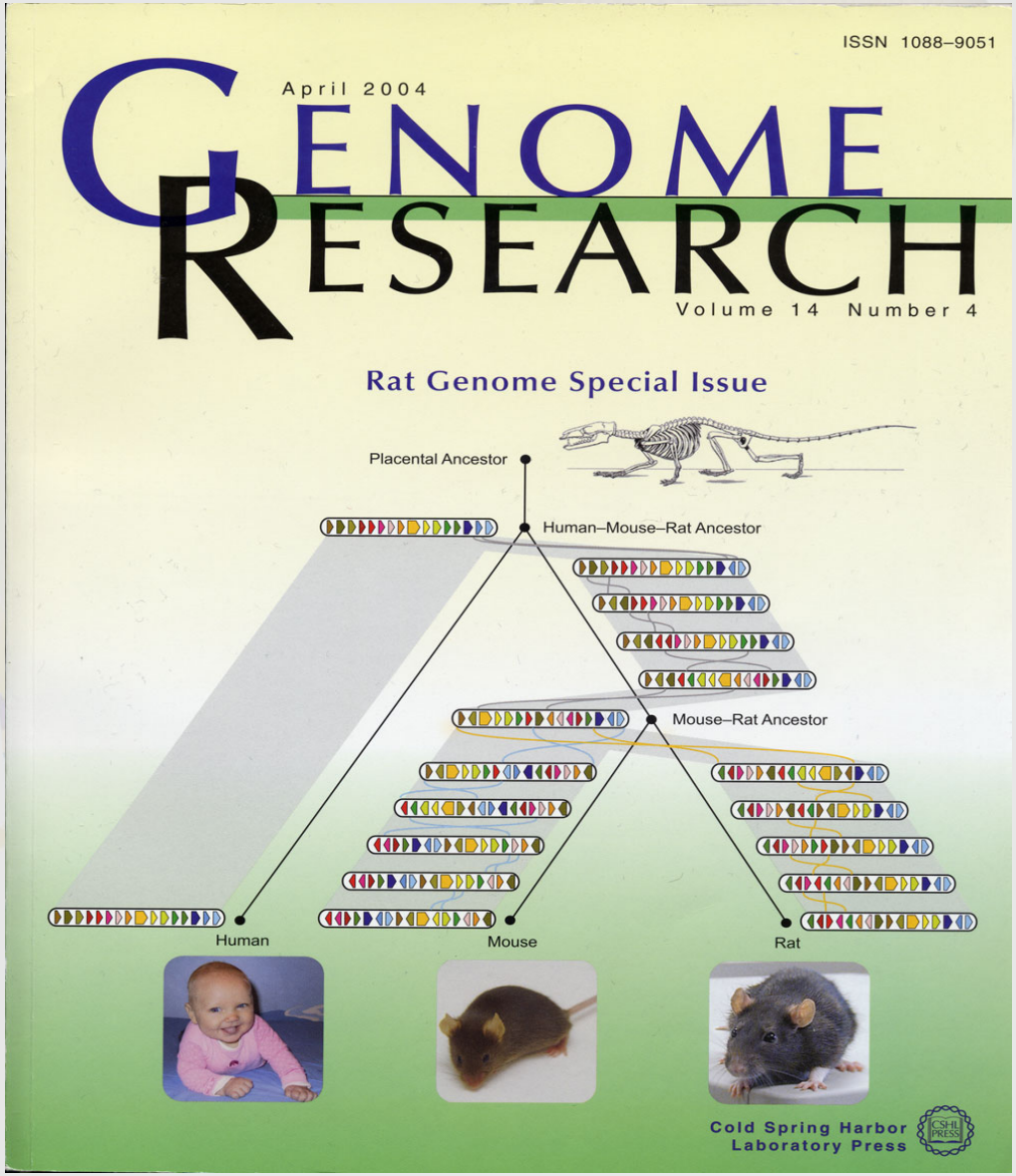
- ✔ What is the evolutionary scenario for transforming one genome into the other?
- ✔ What is the organization of the ancestral genome?
- ✔ Are there any rearrangement hotspots in mammalian genomes?



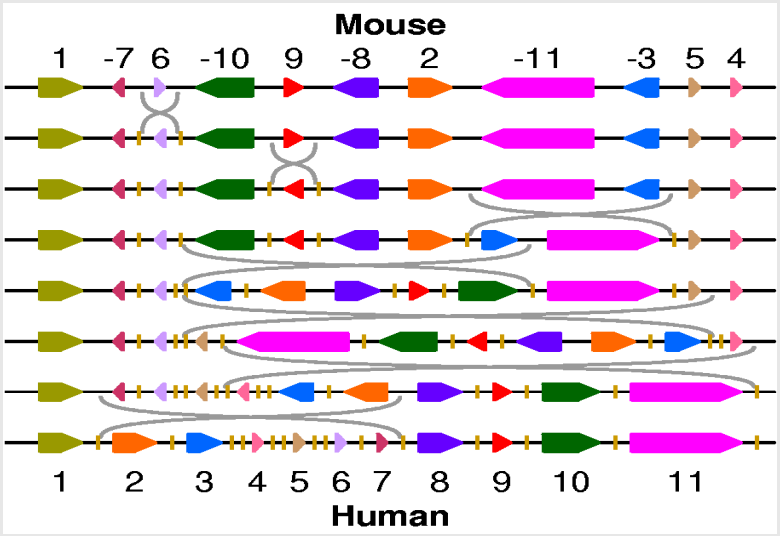
Reversal (inversion) flips a segment of a chromosome

Genome Rearrangements: Ancestral Reconstruction

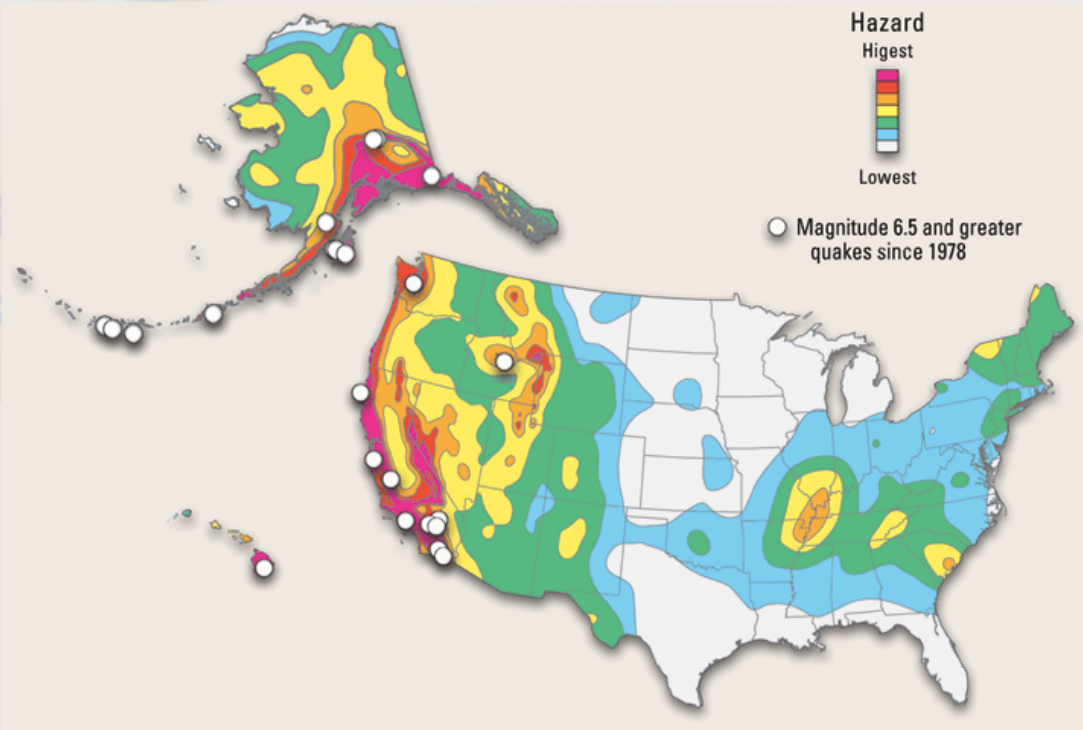
- ✔ What is the evolutionary scenario for transforming one genome into the other?
- ✔ What is the organization of the ancestral genome?
- ✔ Are there any rearrangement hotspots in mammalian genomes?



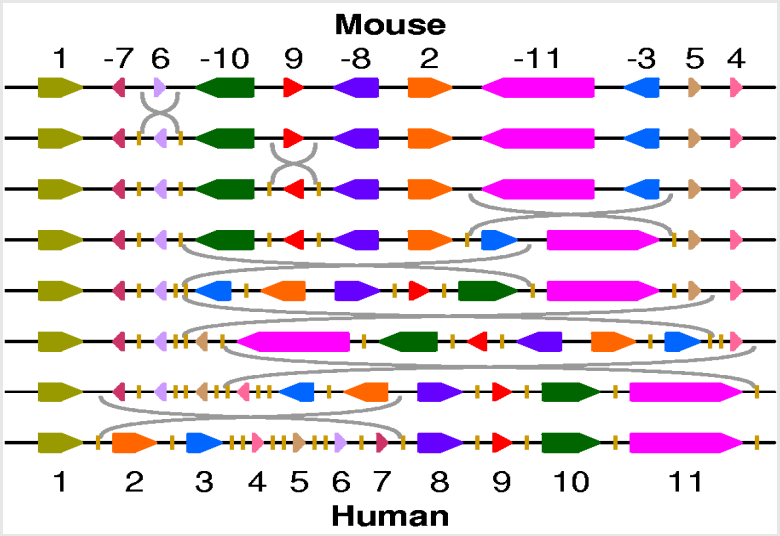
Genome Rearrangements: Evolutionary "Earthquakes"



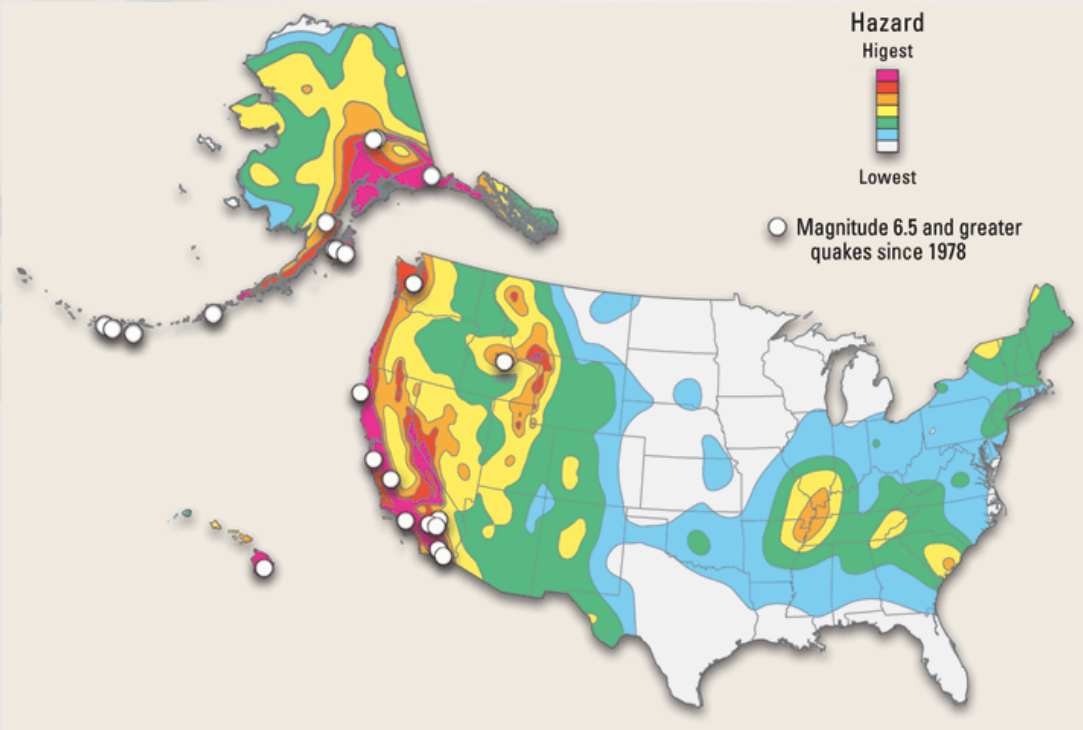
- ✔ What is the evolutionary scenario for transforming one genome into the other?
- ✔ What is the organization of the ancestral genome?
- ✔ Are there any rearrangement hotspots in mammalian genomes? (controversy in 2003-2008)



Genome Rearrangements: Evolutionary "Earthquakes"



- ✔ 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?**

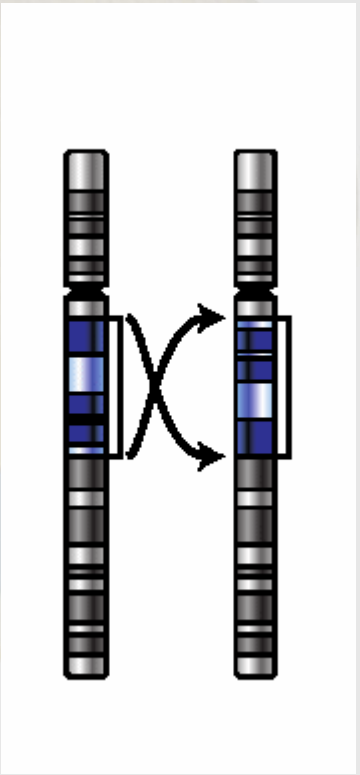


Rearrangement Hotspots in Tumor Genomes

- ✔ Rearrangements may disrupt genes and alter gene regulation.
- ✔ Example: rearrangement in leukemia yields "Philadelphia" chromosome:

- ✔ Thousands of individual rearrangements hotspots known for different tumors.

Unichromosomal Genomes: Reversal Distance



- A *reversal* flips a segment of a chromosome.
- For given genomes P and Q, the number of reversals in a shortest series, transforming one genome into the other, is called the **reversal distance** between P and Q.
- Hannenhalli and Pevzner (*FOCS 1995*) gave a polynomial-time algorithm for computing the reversal distance.

Prefix Reversals

- A *prefix reversal* flips a prefix a permutation.
- Pancake Flipping Problem: sort a given stack (permutation) of pancakes of different sizes with the minimum number of flips of any number of top pancakes.

Discrete Mathematics 27 (1979) 47–57.

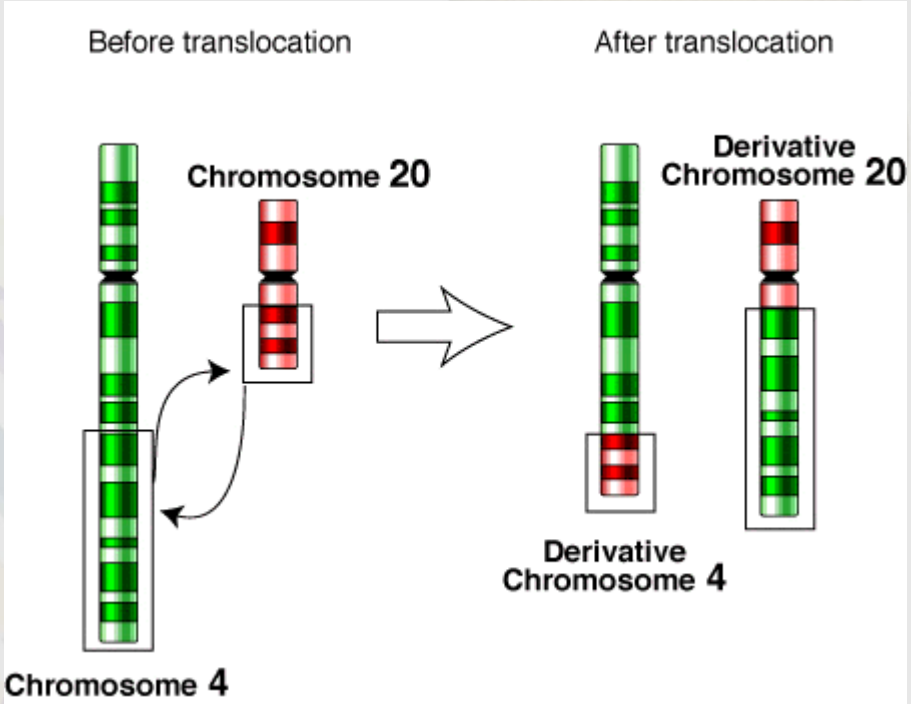
BOUNDS FOR SORTING BY PREFIX REVERSAL

William H. GATES
Microsoft, Albuquerque, New Mexico

Christos H. PAPADIMITRIOU*†
Department of Electrical Engineering, University of California, Berkeley, CA 94720, U.S.A.

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 the other.



- Hannenhalli and Pevzner (STOC 1995) extended their algorithm for computing the reversal distance to computing the genomic distance.

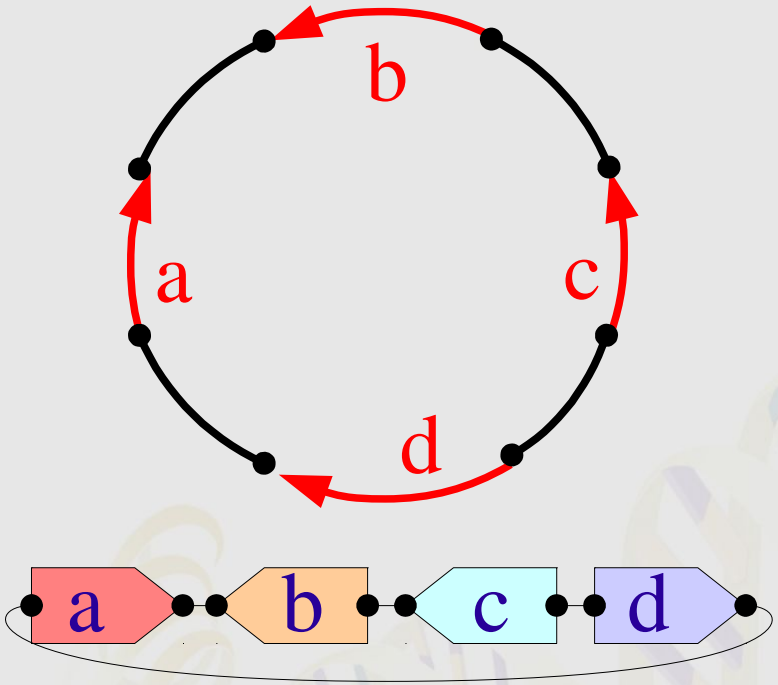
- These algorithms were followed by many improvements: *Kaplan et al. 1999, Bader et al. 2001, Tesler 2002, Ozery-Flato & Shamir 2003, Tannier & Sagot 2004, Bergeron 2001-07, etc.*

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 that makes it difficult to apply in complex setups.

- ✔ To study genome rearrangements in multiple genomes, we use ***2-break*** rearrangements, also known as DCJ (*Yancopoulus et al., Bioinformatics 2005*).

Simplifying HP Theory: Switch from Linear to Circular Chromosomes

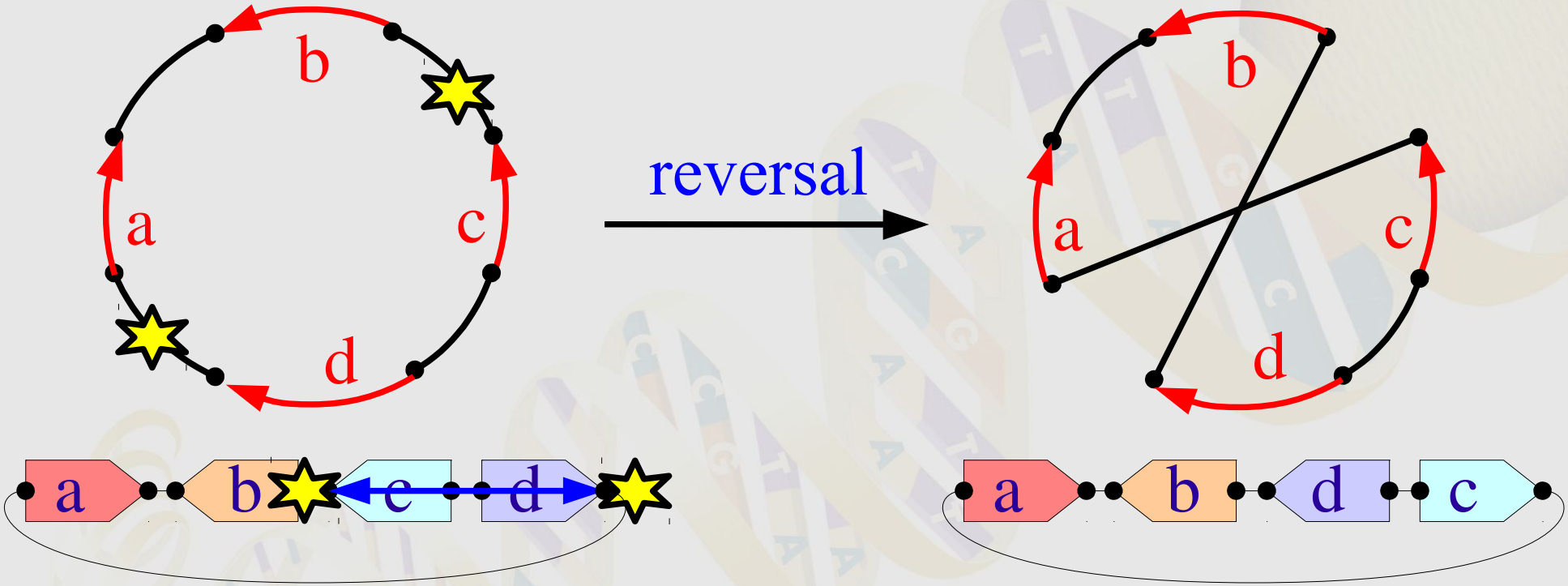


A chromosome can be represented as a *cycle* with *directed red* and ***undirected black*** edges, where:

red edges encode blocks and their directions;

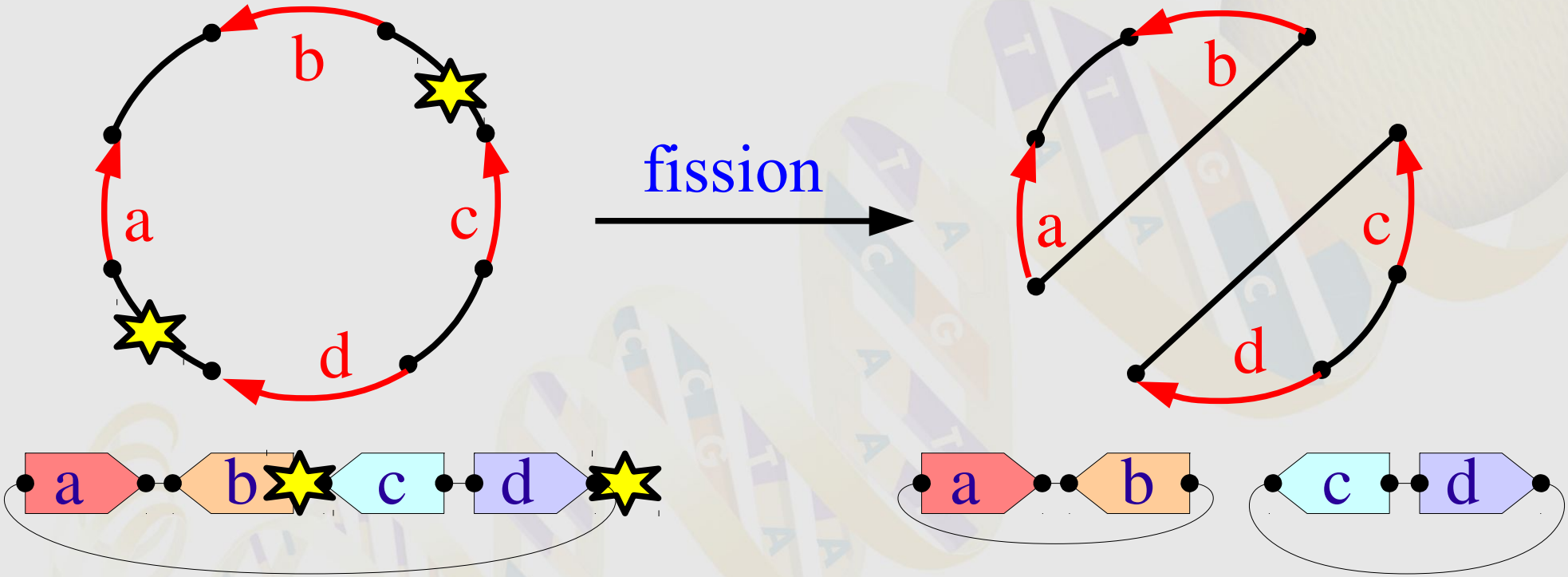
adjacent blocks are connected with black edges.

Reversals on Circular Chromosomes



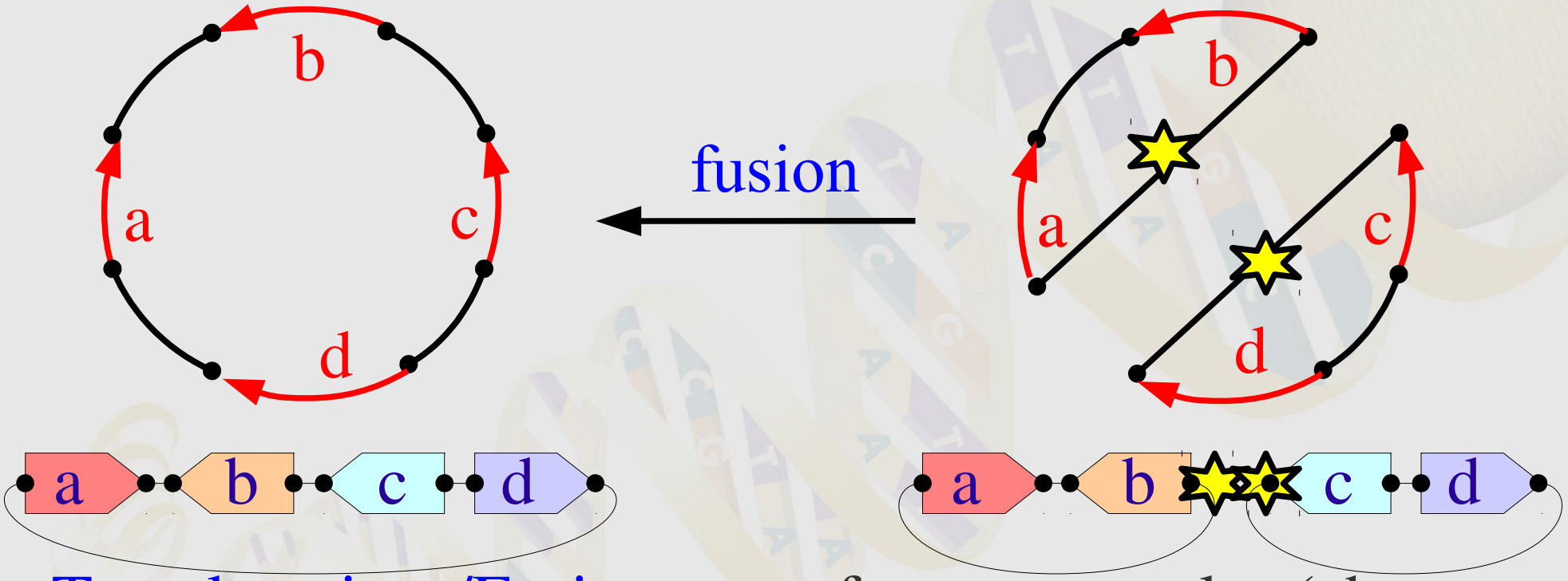
Reversals replace two black edges with two other black edges

Fissions



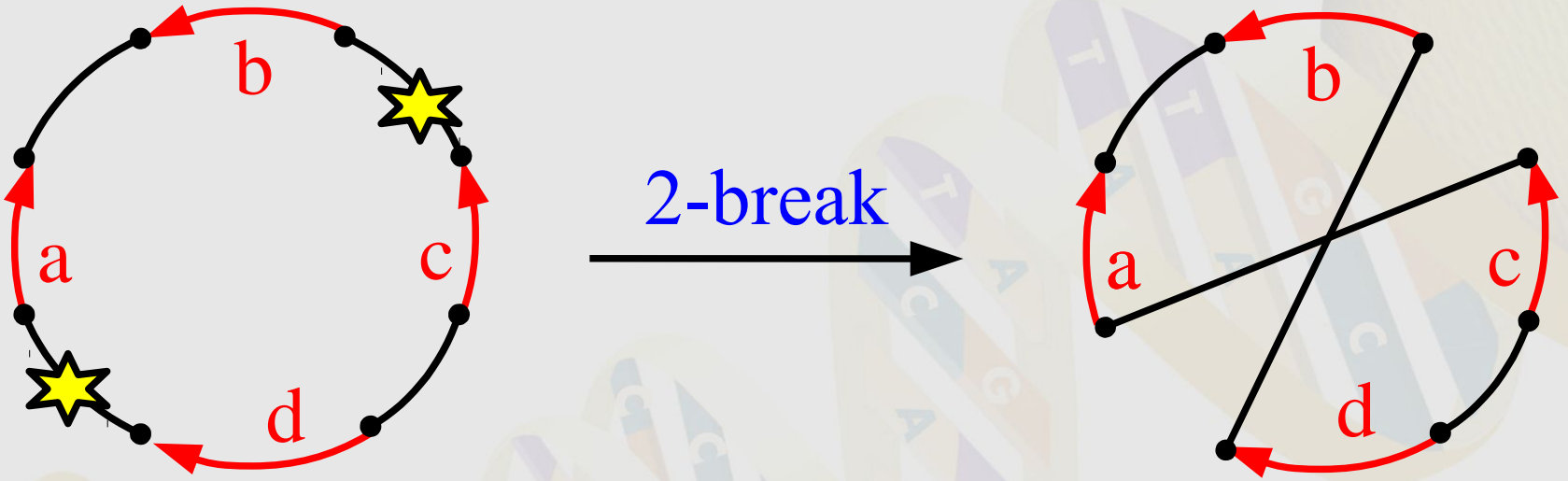
- ✔ Fissions split a single cycle (chromosome) into two.
- ✔ Fissions replace two black edges with two other black edges.

Translocations / Fusions



- ✔ Translocations/Fusions transform two cycles (chromosomes) into a single one.
- ✔ They also replace two black edges with two other black edges.

2-Breaks



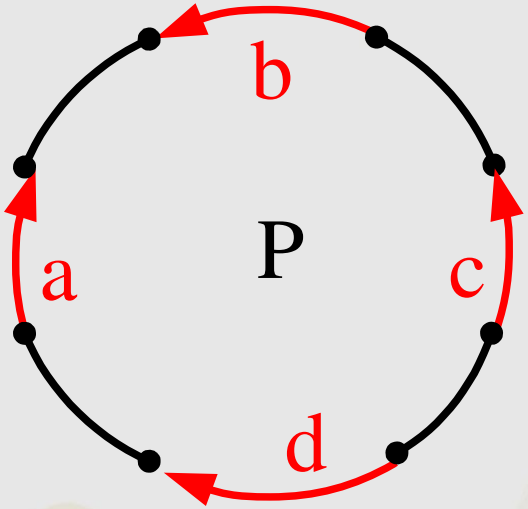
- ✔ 2-Break replaces *any pair* of black edges with another pair forming matching on the same 4 vertices.
- ✔ Reversals, translocations, fusions, and fissions represent all possible types of 2-breaks.

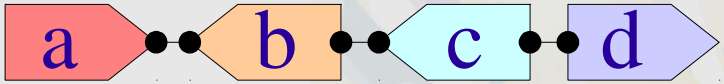
2-Break Distance

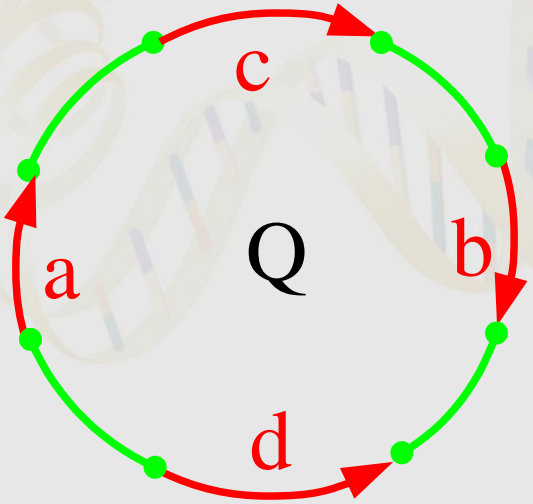
- The **2-Break distance** ***dist(P,Q)*** between genomes ***P*** and ***Q*** is the minimum number of 2-breaks required to transform ***P*** into ***Q***.

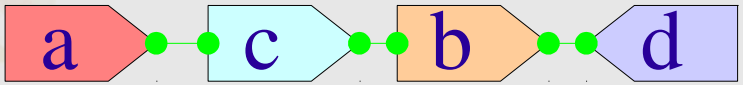
- In contrast to the genomic distance, the 2-break distance is easy to compute.

Two Genomes as Black-Red and Green-Red Cycles

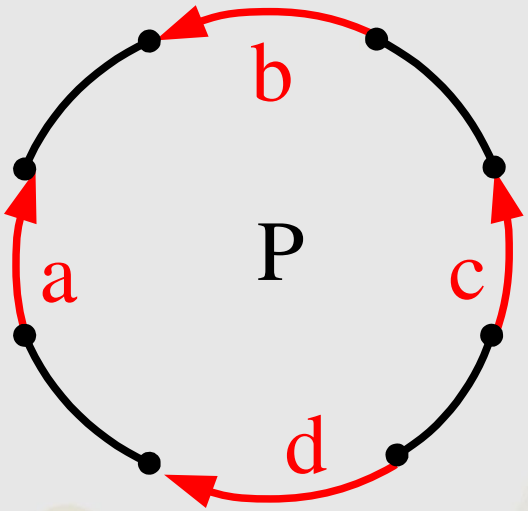


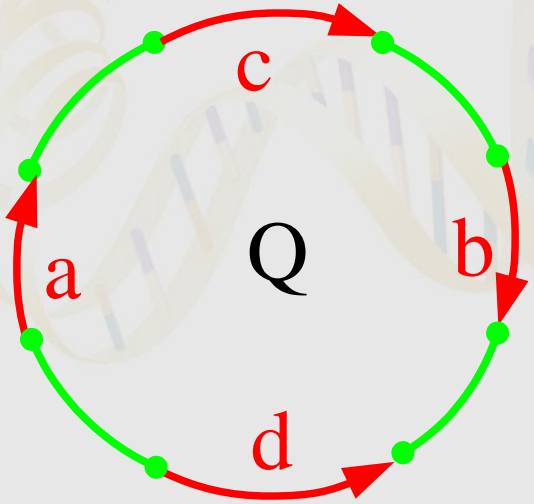


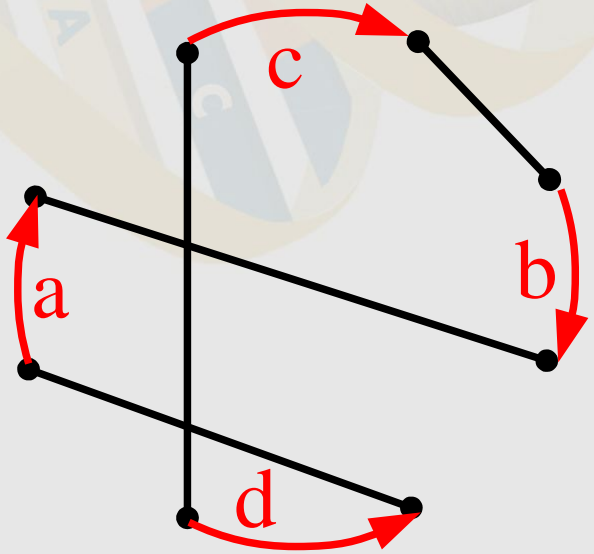




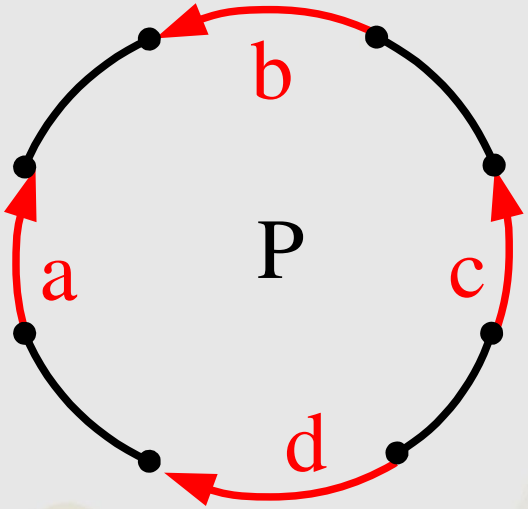
Rearranging P in the Q order

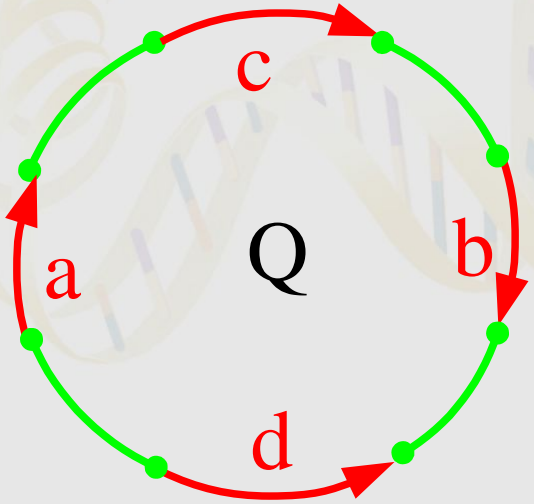




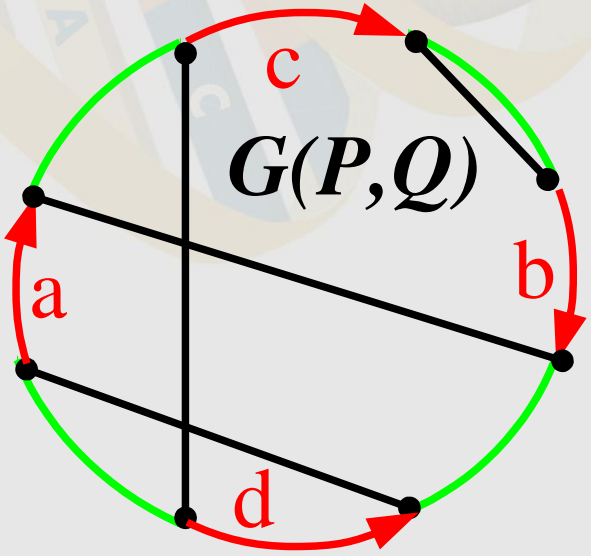


Breakpoint Graph = Superposition of Genome Graphs: Gluing Red Edges with the Same Labels





Breakpoint Graph
(Bafna & Pevzner, FOCS 1994)

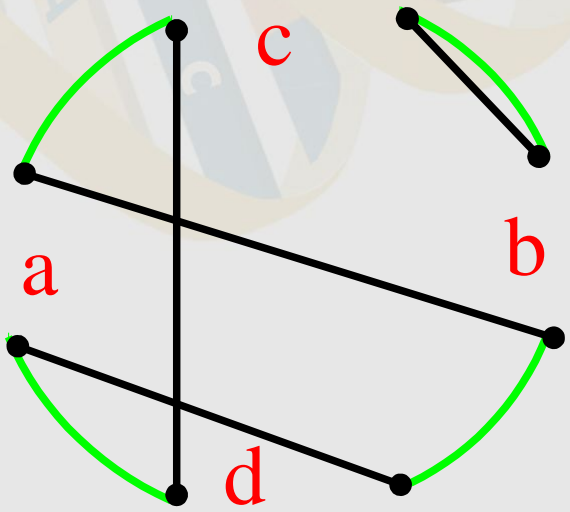


Black-Green Cycles

- Breakpoint graph is formed by ***red***, ***black*** and ***green*** perfect matchings.
- Black and green matchings form a collection of *alternating* ***black-green*** cycles.
- Number of black-green cycles:

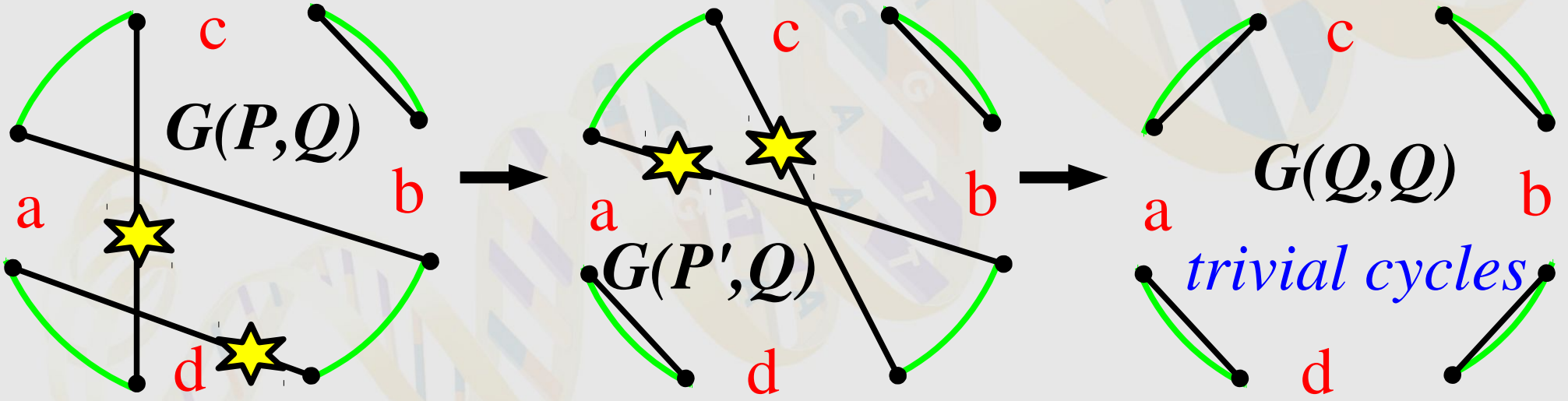
cycle(P,Q)

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



Rearrangements Change Cycles

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



cycles(P,Q) = 2

cycles(P',Q) = 3

cycles(Q,Q) = 4

= blocks(P,Q)

Transforming P into Q 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 $\rightarrow$ ... $\rightarrow$ ***blocks(P,Q)*** cycles

# of black-green cycles increased by
blocks(P,Q) – cycles(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: $\Delta cycles \leq 2\text{-}1 = 1$.

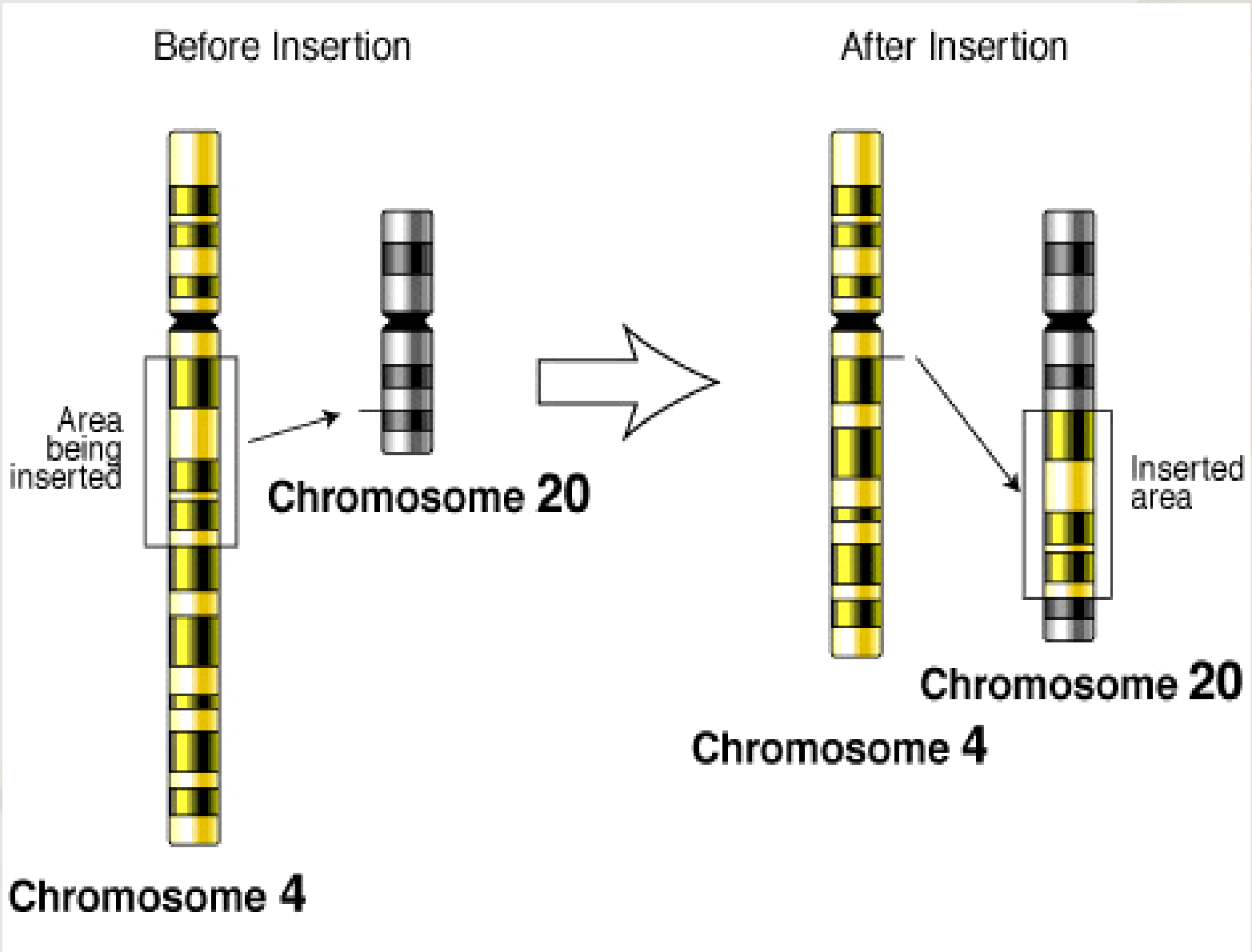
2-Break Distance

- 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(P,Q) - cycles(P,Q)***
- The **2-Break Distance** between genomes *P* and *Q*:

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

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

Complex Rearrangements: Transpositions



Sorting by Transpositions

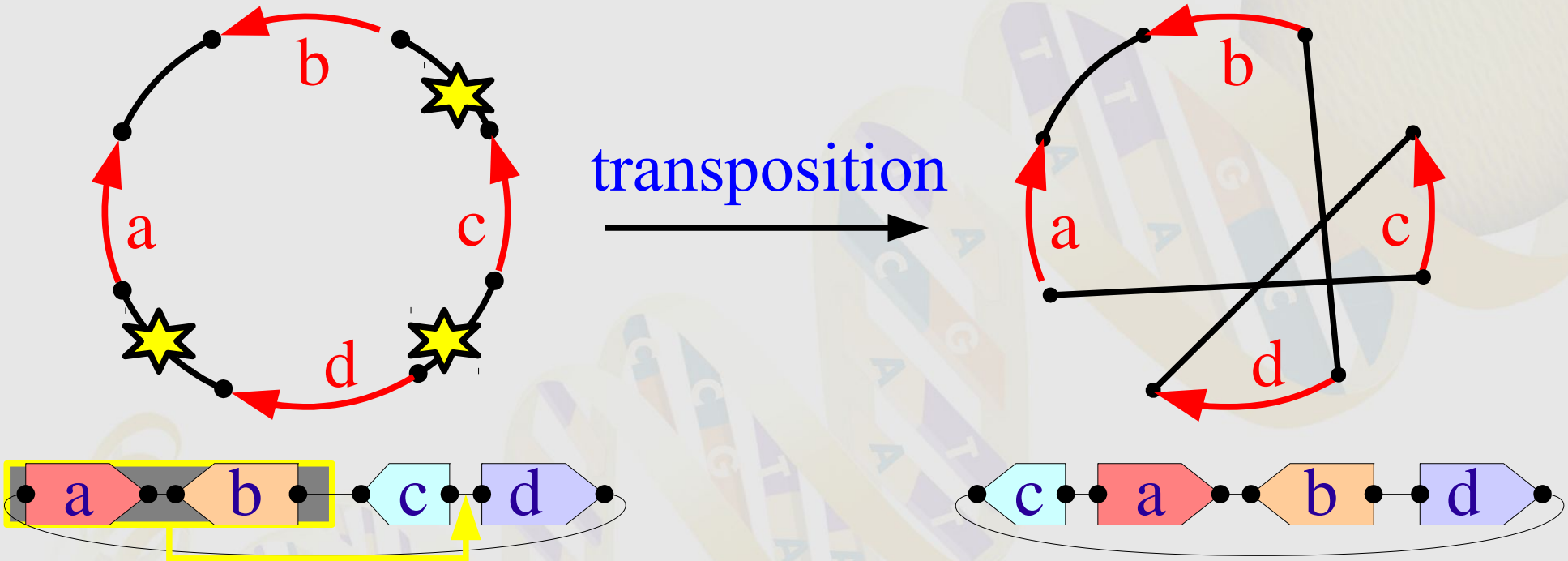
- **Sorting by Transpositions Problem:**
  Given two genomes, find the shortest sequence of transpositions transforming one genome into the other.

- First 1.5-approximation algorithm was given by Bafna and Pevzner (*SODA 1995*).

- Best achievement: 1.375-approximation algorithm due to Elias and Hartman (*WABI 2005*).

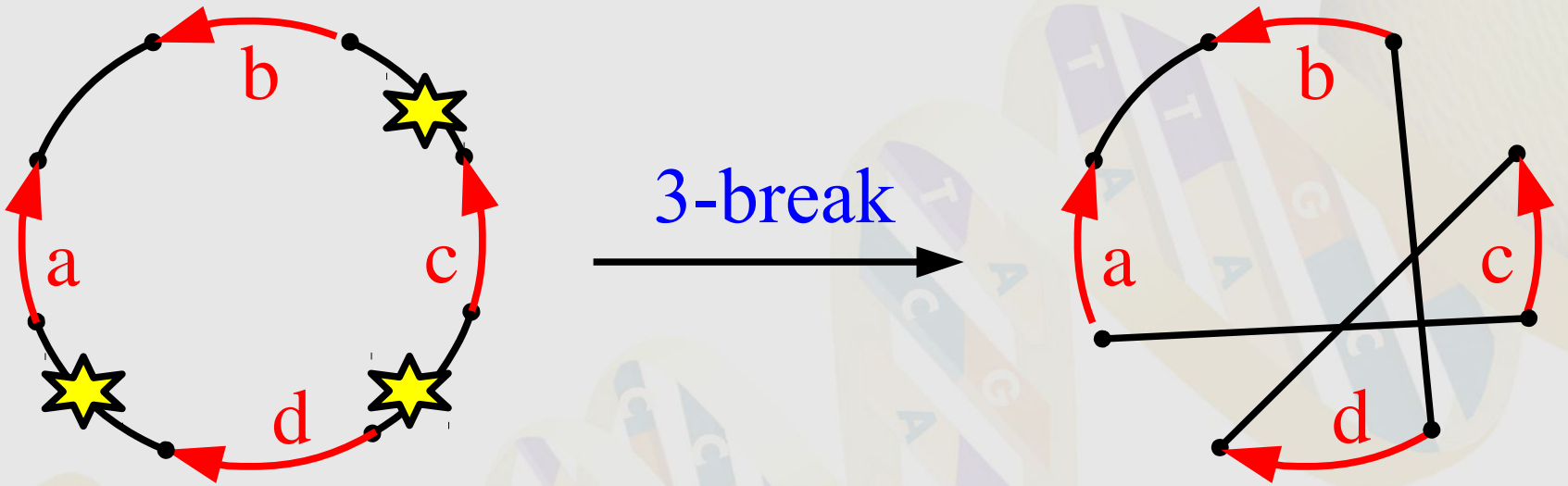
- Proved to be NP-complete by Bulteau, Fertin, and Rusu (*ICALP 2011*).

Transpositions



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

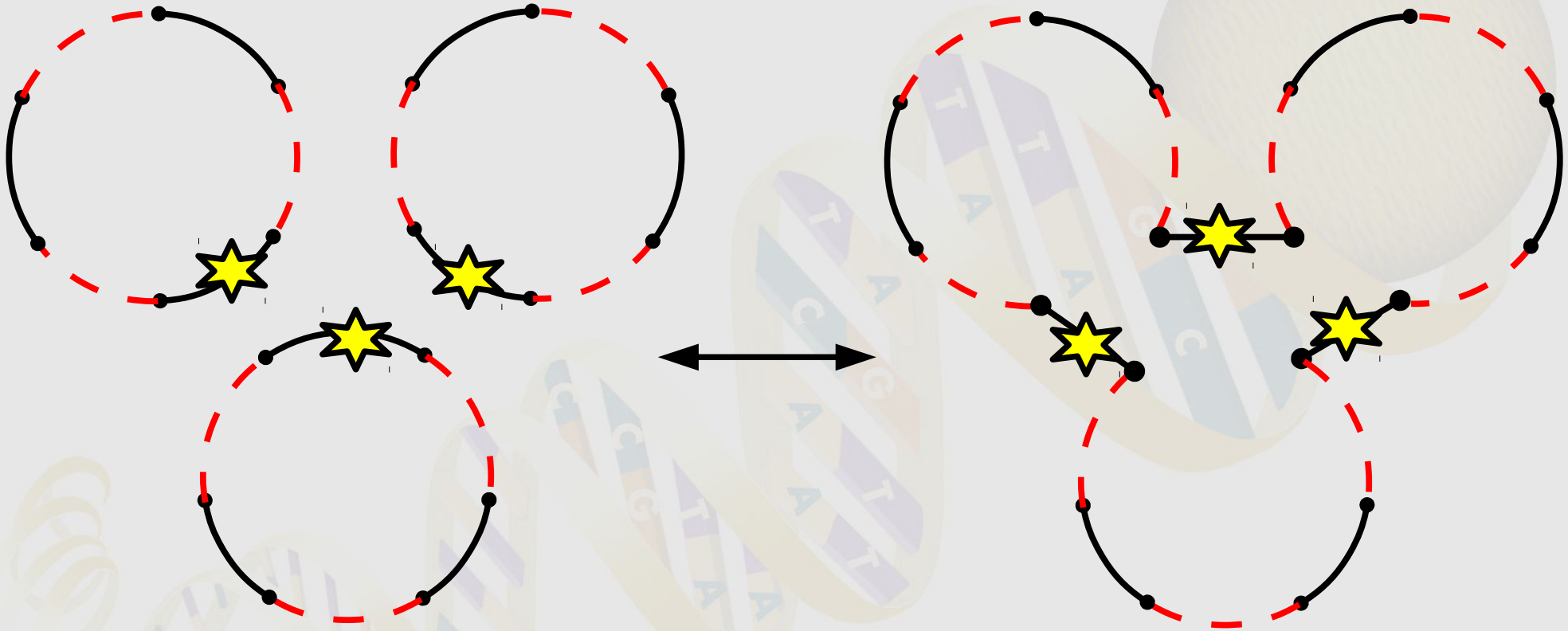
Transpositions Are 3-Breaks



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

- ✔ **Transpositions are 3-Breaks.**

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

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

- 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(P,Q) - cycles(P,Q)***
- The **3-Break Distance** between genomes *P* and *Q*:

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

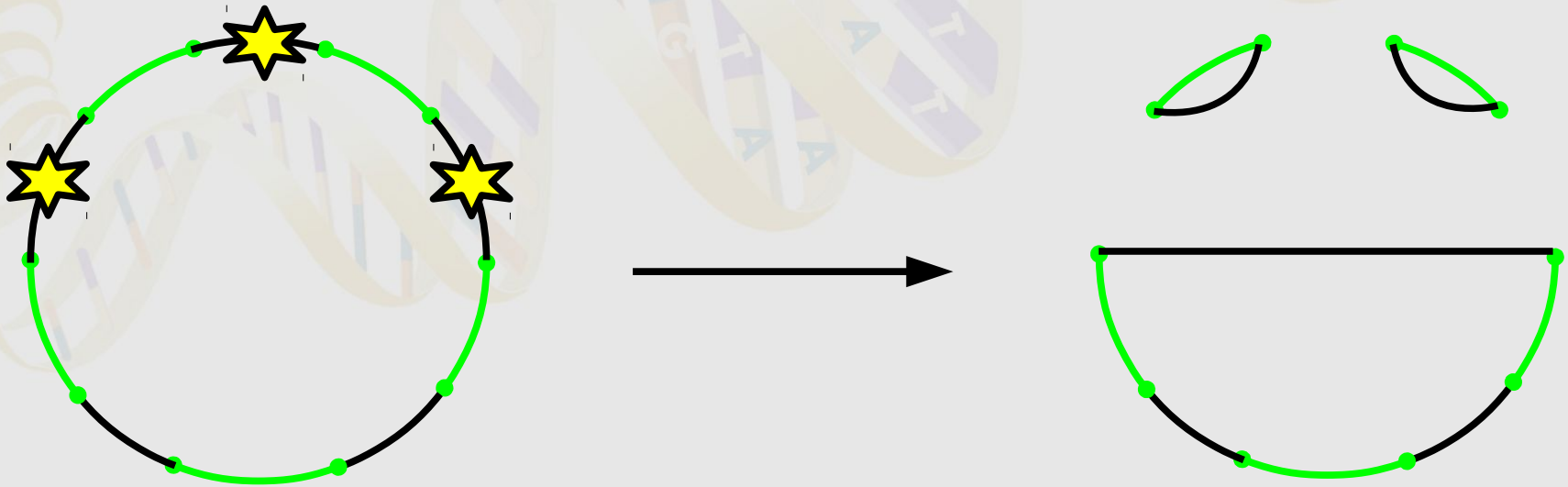
3-Break Distance: OOPS!

- ✔ Any 3-break increases the number of cycles by at most two ($\Delta cycles \leq 2$)
- ✔ **Any non-trivial cycle can be split into three cycles with a 3-break ($\Delta cycles = 2$)**
- ✔ Every sorting by 3-breaks must increase the number of cycles by $blocks(P,Q) - cycles(P,Q)$
- ✔ The **3-Break Distance** between genomes P and Q:

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

3-Break Splitting Odd Cycle

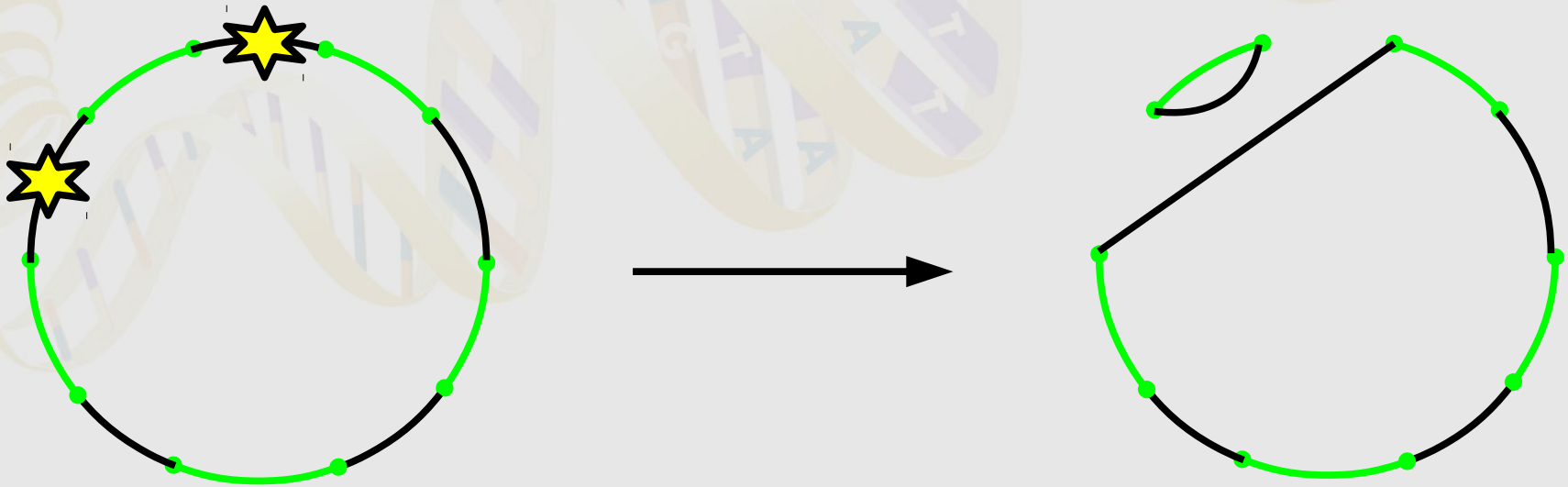
- ✔ Trivial cycles are ***odd*** cycles. 3-break can increase the number of odd cycles by at most 2.
- ✔ A non-trivial odd cycle can be split into three odd cycles with a 3-break ($\Delta cycles^{odd} = 2$)



3-Break Splitting Even Cycle

- Trivial cycles are ***odd*** cycles. 3-break can increase the number of odd cycles by at most 2.

- An even cycle can be split into two odd cycles with a 3-break ($\Delta cycles^{odd} = 2$)



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 ($\Delta cycles^{odd} \leq 2$)
- 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:

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

Multi-Break Rearrangements

- The standard rearrangement operations (*reversals, translocations, fusions, and fissions*) make *2 breakages* in a genome and glue the resulting pieces in a new order.

- *k*-Break rearrangement operation makes *k breakages* 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) may be more common, e.g., complex rearrangements in irradiated human lymphocytes (*Sachs et al., 2004; Levy et al., 2004*).

k-Breaks

- *k*-Break replaces k black edges, forming a matching on $2k$ vertices, with a set of k black edges, forming another matching on the same $2k$ vertices.

- ***k*-Break Distance Problem**: Given two genomes P and Q, find the shortest sequence of k-breaks transforming P into Q.

Number of Cycles

- ✔ Every k-break introduces at most $\boldsymbol{k}$ breakages.
- ✔ A k-break "destroys" at least one cycle
- ✔ A k-break creates at most k new cycles
- ✔ Therefore, k-break can increase the number of cycles by at most k-1, .i.e., $\boldsymbol{\Delta cycles \leq k\text{-}1}$

Computing k-Break Distance

- We need to create $blocks(P,Q) - c^{trivial}(P,Q)$ new trivial cycles

- A k-break can create at most k trivial cycles; *optimal* k-break creates k trivial cycles

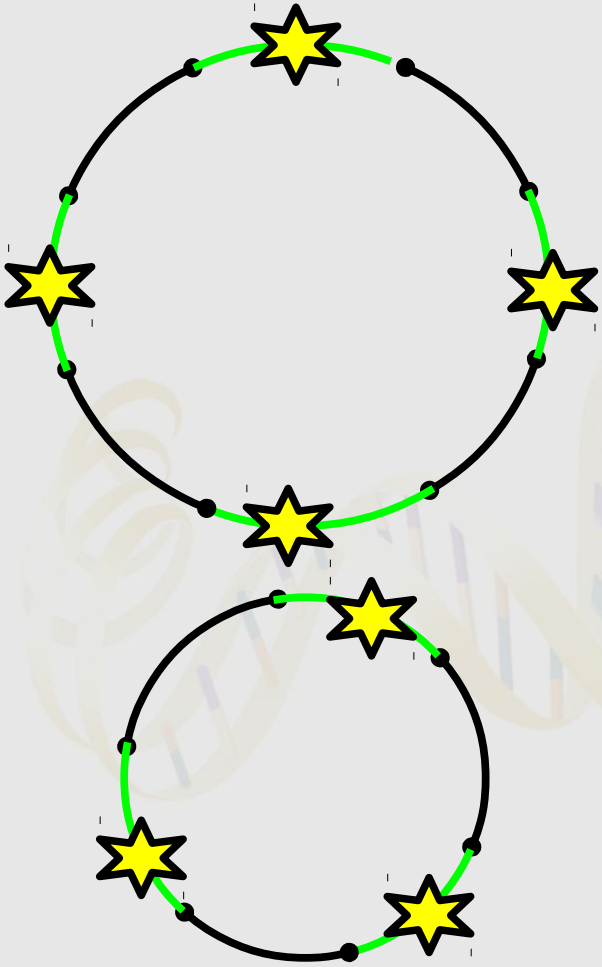
- If there are ≥ k edges in non-trivial cycles, it is always possible to create $k-1$ trivial cycles (*suboptimal* k-break)

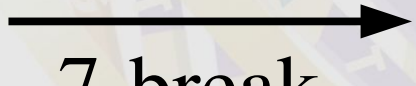
$$ceil((blocks(P,Q) - c^{trivial}(P,Q)) / k) \le d_k(P,Q) \le$$
$$\le ceil((blocks(P,Q) - c^{trivial}(P,Q)) / (k-1))$$

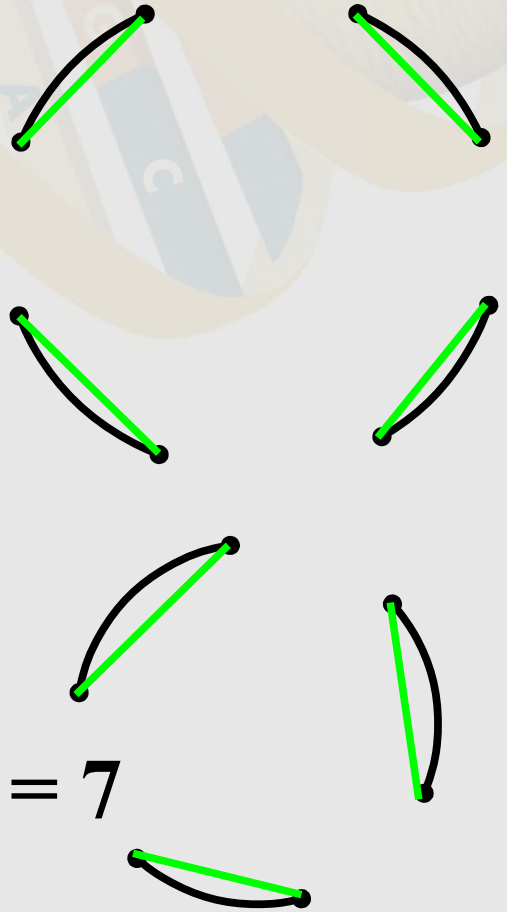
- Goal: maximize the number of optimal k-breaks

Optimal k-Breaks

Lemma. A *k*-break creates *k* trivial cycles if and only if it acts on *all* edges of one or several non-trivial cycles



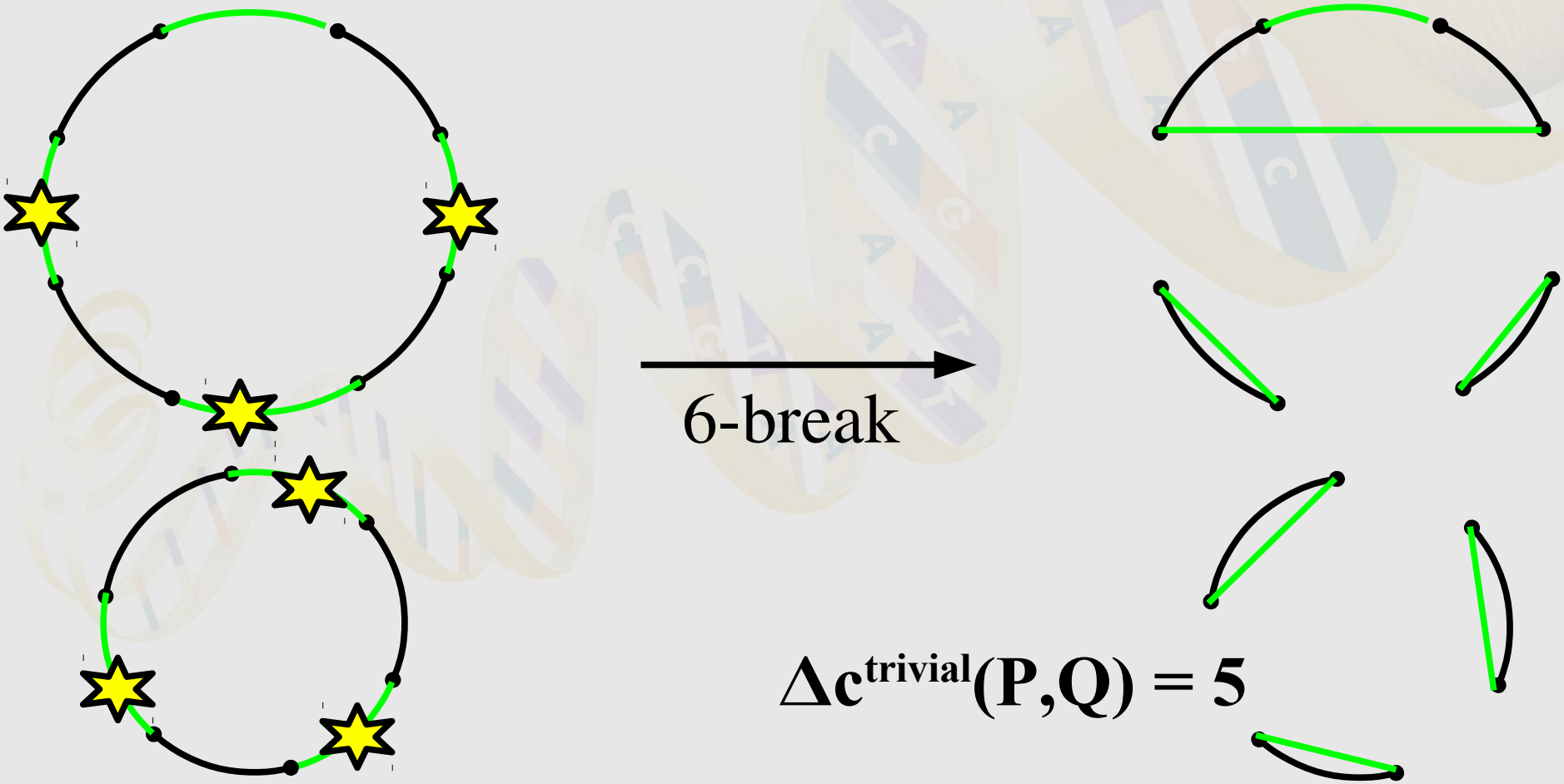




$\Delta c^{trivial}(P,Q) = 7$

Suboptimal k-Breaks

Lemma. A *k*-break creates *k*-1 trivial cycles if and only it acts on a part of edges of one non-trivial cycle and *all* edges of some other non-trivial cycles



Greedy Construction of a Suboptimal k-Break

If there are ≥ k edges in non-trivial cycles, a suboptimal (or even optimal) *k*-break can be constructed in a *greedy* way:

- ✔ Take *all edges* of one cycle, *all edges* of another cycle, and so on while the total number of taken edges is less than *k*.
- ✔ Then take a *part of edges* of yet another cycle, making the total number of taken edges equal *k*.

Breakable Sets

- A set of black-gray cycles is ***breakable*** if the total number of black edges is 1 modulo ***(k-1)***
  - E.g.: for ***k=2***, any set of cycles is breakable
  - E.g.: for ***k=3***, a set of cycles is breakable if and only if it contains odd number of odd cycles

Breaking a Breakable Set into Trivial Cycles

- A breakable set can be broken into trivial cycles using a series of suboptimal k-breaks and at least one optimal k-break

- A breakable set with m black edges can be broken into m trivial cycles with $t=(m-1)/(k-1)$ k-breaks:

  $t-1$ suboptimal k-breaks create $(t-1)\cdot(k-1) = m-k$ trivial cycles leaving k edges in non-trivial cycles that are handled by an optimal k-break

k-Break Distance

- Let $s_k(P,Q)$ be the maximum number of disjoint breakable sets

$$d_k(P,Q) = ceil((blocks(P,Q) - s_k(P,Q)) / (k-1))$$

- E.g.: for $k=2$, every cycle forms a breakable set and thus $s_2(P,Q) = cycles(P,Q)$, implying that
  $d_2(P,Q) = blocks(P,Q) - cycles(P,Q)$
- E.g.: for k=3, every odd cycle forms a breakable set and thus $s_3(P,Q) = cycles^{odd}(P,Q)$, implying that
  $d_3(P,Q) = (blocks(P,Q) - cycles^{odd}(P,Q)) / 2$

More Examples

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.

For ***k=4***, the maximum set of disjoint breakable sets consists of:

- the cycles with 1 modulo 3 black edges each
- the pairs of cycles with 2 modulo 3 black edges each

Computing $s_k(P,Q)$

- Each set of cycles corresponds to a vector $(x_1,x_2,\ldots,x_{k-2})$ where x_j is the number of cycles of length j modulo $(k-1)$.
- A breakable set corresponds to a breakable vector $(x_1,x_2,\ldots,x_{k-2})$ with $1x_1+2x_2+\ldots+(k-2)x_{k-2} \equiv 1 \pmod{k-1}$.
- **Theorem.** Given a vector $c=(c_1,\ldots,c_{k-2})$ where c_j is the number of cycles in $G(P,Q)$ of length j modulo $(k-1)$, find a maximum number of breakable vectors $v^1,v^2,\ldots,v^s$ such that $v^1+v^2+\ldots+v^s \le c$ (component-wise). Then $s = s_k(P,Q)$.

Dynamic Programming

- A breakable vector $(x_1,x_2,\ldots,x_{k-2})$ is ***proper*** if $x_j < k-1$.
  Let V be a set of all proper breakable vectors.
  Clearly, $|V|=(k-1)^{k-3}$.

- Define $s(u) = max\ |\{ v^1, v^2, \ldots, v^t \mid v^1+v^2+\ldots+v^t \le u\}|$.
  Then $s_k(P,Q) = s(c)$.

- $s(c)$ can be computed using dynamic programming:

$$s(u) = max_{v \le u}\ s(u-v) + 1$$

- Size of the dynamic programming table:
  $(c_1+1)\times\ldots\times(c_{k-2}+1) = O((n/k)^{k-2})$
- Overall complexity is $O(n^{k-2})$

Extremal Vectors

- If $\boldsymbol{u} \leq \boldsymbol{v}$ for $\boldsymbol{u}$, $\boldsymbol{v}$ from $\boldsymbol{V}$, then $\boldsymbol{v}$ can be removed from $\boldsymbol{V}$.
- Vector $\boldsymbol{v}$ in $\boldsymbol{V}$ is ***extremal*** if there is no $\boldsymbol{u}$ in $\boldsymbol{V}$ such that $\boldsymbol{u} \leq \boldsymbol{v}$.
- The set of all extremal vectors $\boldsymbol{V'}$ can be computed using *Hilbert bases*.

Extremal Vectors

k	$\lvert V\rvert = (k-1)^{k-3}$	$\lvert H\rvert$	$\lvert V'\rvert$
6	125	27	6
7	1296	39	8
8	16807	83	16
9	262144	117	22
10	4782969	205	37
11	100000000	291	53
12	2357947691	555	92
13	61917364224	634	110
14	1792160394037	1277	201
15	56693912375296	1567	260
16	1946195068359375	2368	376
17	72057594037927936	3315	519
18	2862423051509815793	5740	831
19	121439531096594251776	6228	963
20	5480386857784802185939	11404	1592

Weighted Genomic Distance can hardly impose a bound on the proportion of transpositions

Max Alekseyev

University of South Carolina, Columbia, SC, U.S.A.

2010

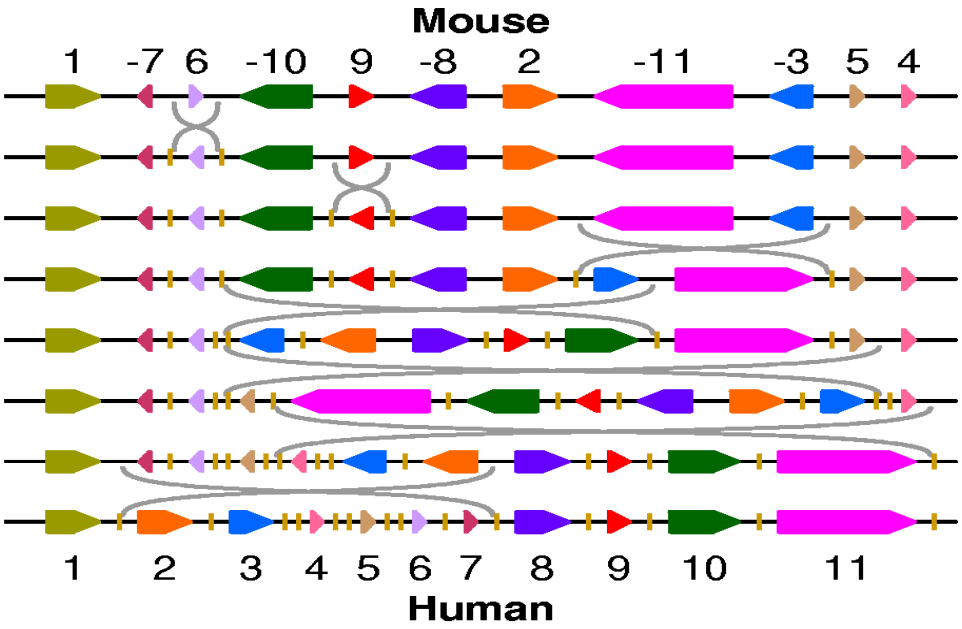
Introduction

- Genome Rearrangements
- Genomic Distance and Breakpoint Graphs
- Transpositions and Weighted Genomic Distance
- Optimal Transformations
- Main Theorem

Genome Rearrangements

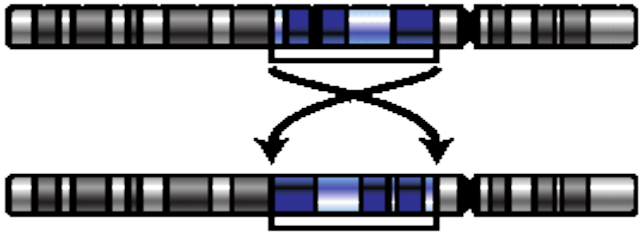
Genome Rearrangements

- ***Genome rearrangements*** are evolutionary events that change genomic architectures.



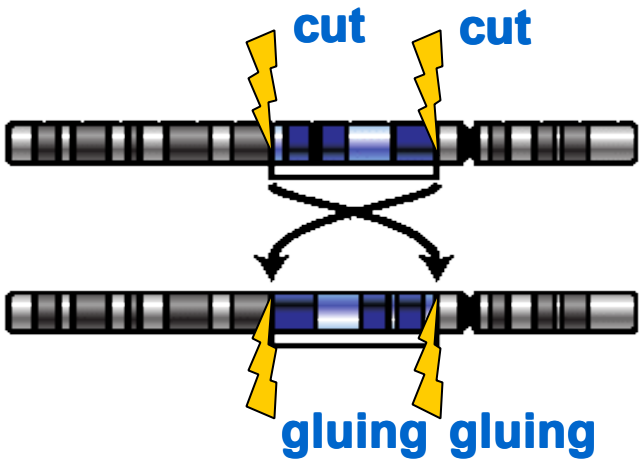
Reversals

- Reversals are frequent genome rearrangements.
- A ***reversal*** flips a segment of a chromosome.



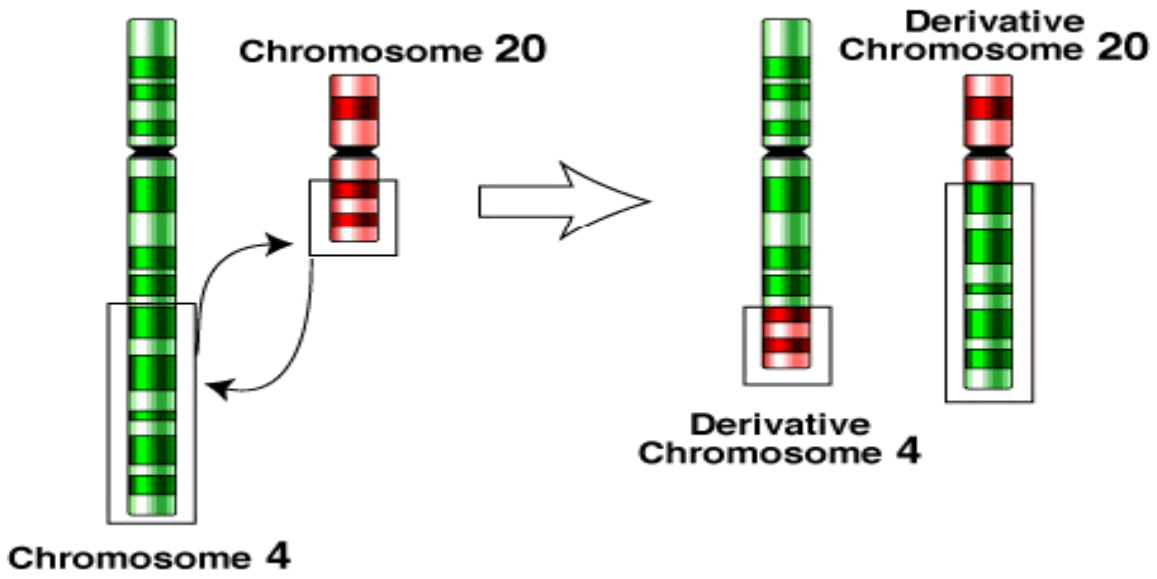
Reversals

- A reversal makes 2 cuts and 2 gluings.



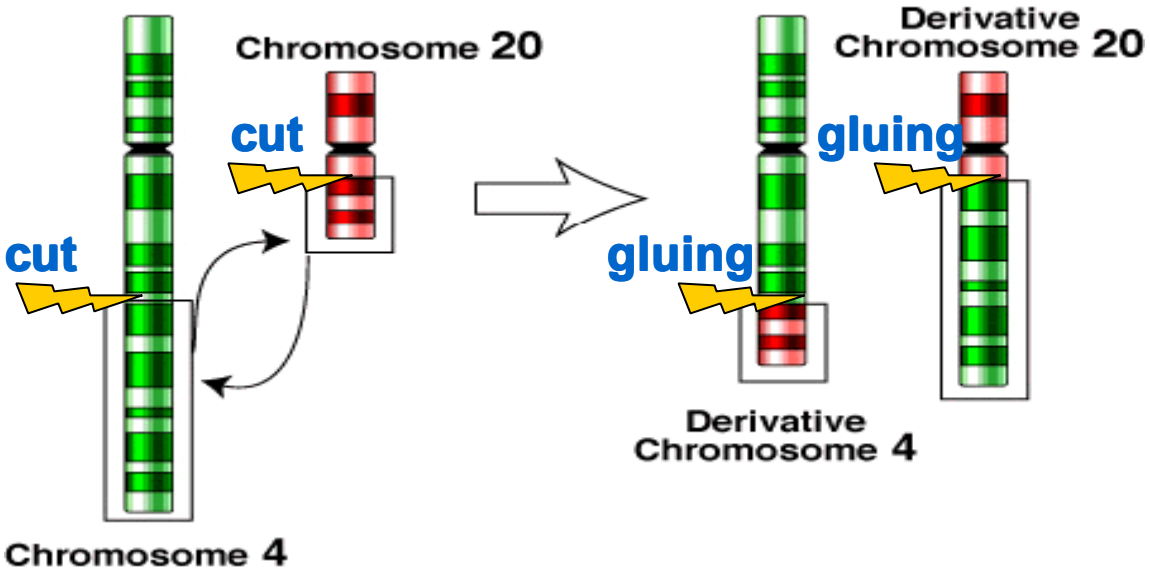
Translocations

- Other common types of rearrangements are translocations, fissions and fusions.
- A ***translocation*** exchanges parts of two chromosomes.



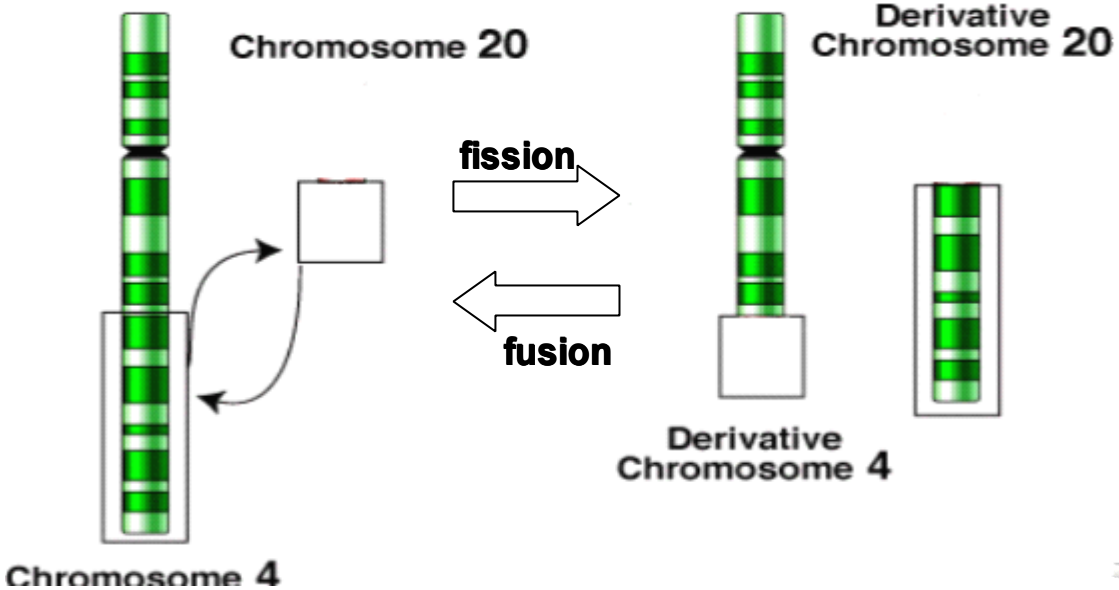
Translocations

- A translocation makes 2 cuts and 2 gluings.



Fissions and Fusions

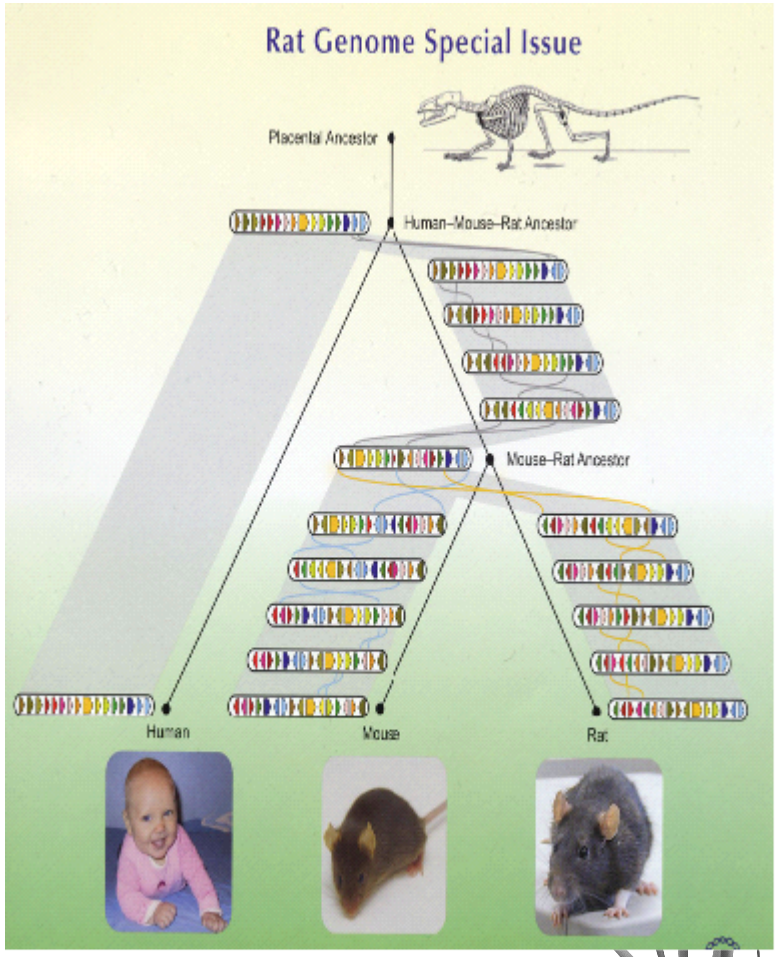
- A ***fission*** splits a chromosome into two.
- A ***fusion*** joins two chromosomes into one.
- Fissions and fusions can be viewed as particular cases of translocations.



Genomic Distance and Breakpoint Graphs

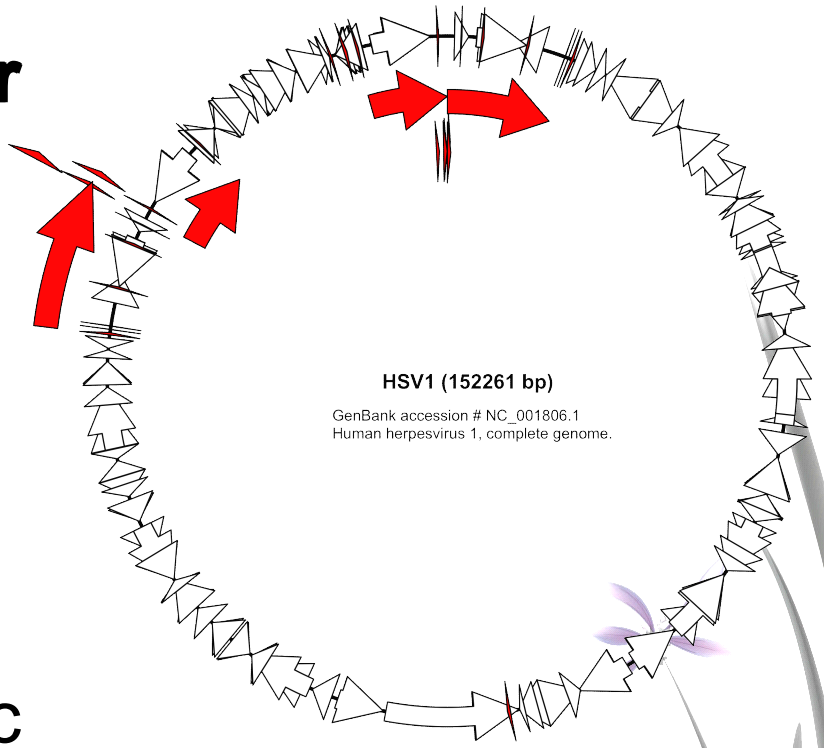
Genomic Distance

- Large-scale rearrangements have dramatic effect on the genomes, but happen rarely.
- The minimal number of rearrangements required to transform one genome into the other is called the ***genomic distance.***
- Genomic distance provides a good measure for evolutionary remoteness of genomes.



Circular Genomes

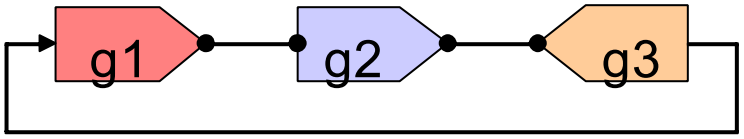
- Rearrangements in **linear genomes** are hard to analyze.
- Rearrangements in ***circular genomes*** are simpler to analyze. They provide reasonable approximation to genomic distance between linear genomes.

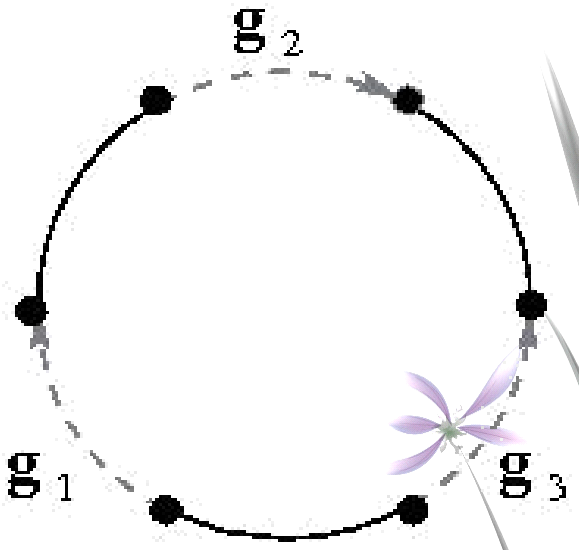


A circular viral genome

Genome Graph

- A chromosome can be represented as a *cycle* with ***directed grey edges*** and ***undirected black edges***, where,
  - ✓ grey edges encode blocks and their directions;
  - ✓ adjacent blocks are connected with black edges.
- Genome is then a collection of cycles representing its chromosomes.

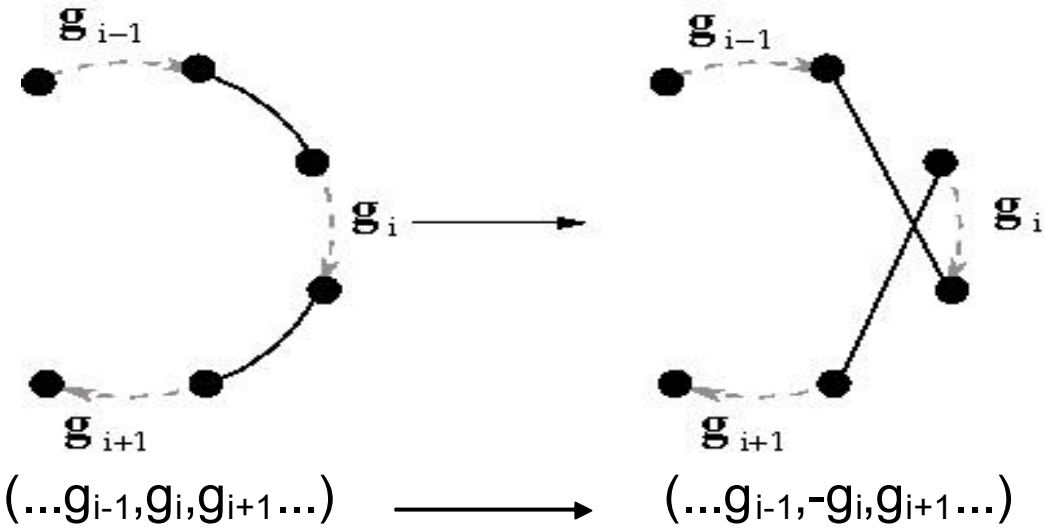




A circular chromosome $(g_1, g_2, -g_3)$

2-break Rearrangement

- ***2-break rearrangement*** is a replacement of 2 black edges in a genome graph with 2 different black edges on the same set of vertices.
- Every reversal, translocation, fission or fusion is a ***2-break***.



A reversal that flips gene g_i

2-break Distance

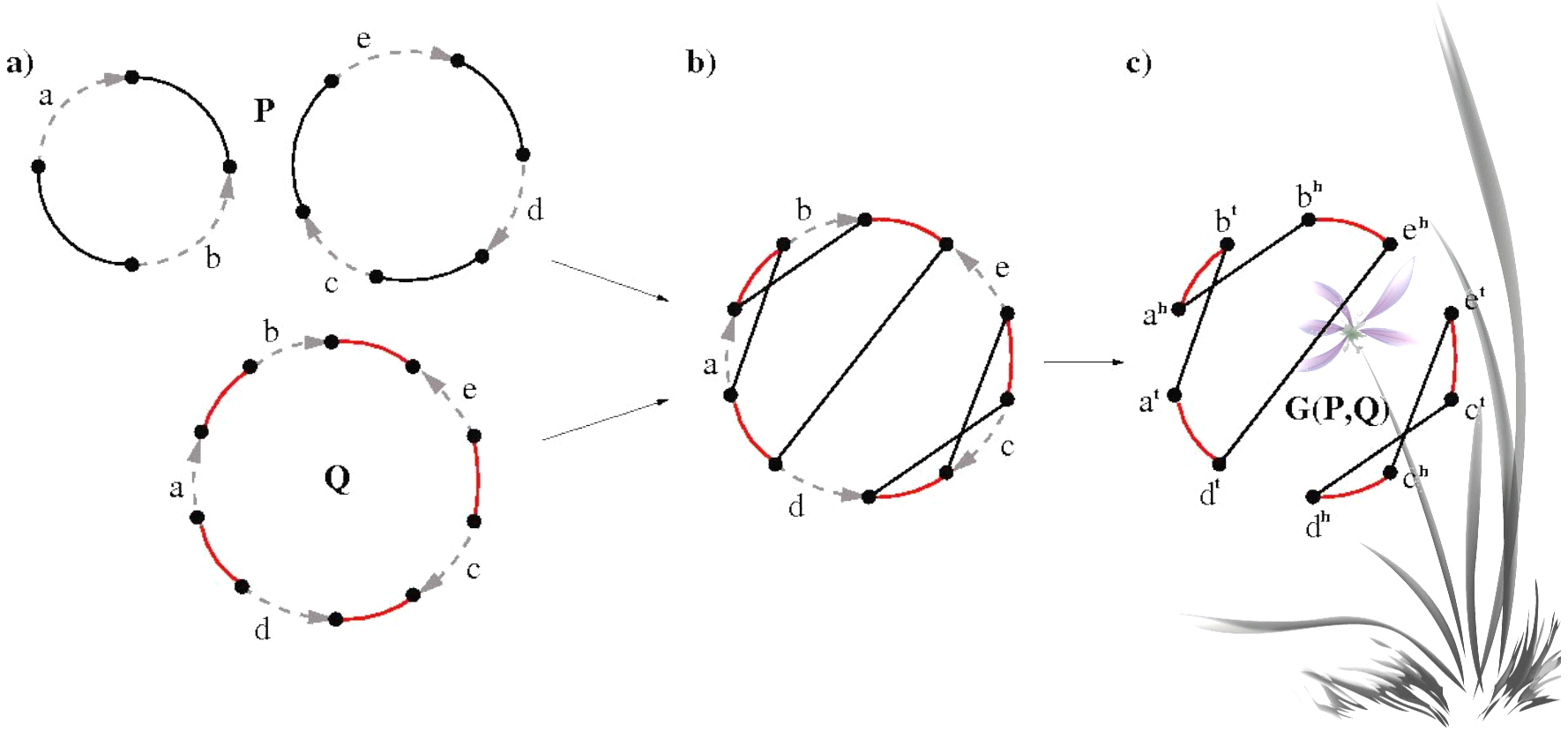
- ***2-break distance*** is the minimal number of 2-breaks required to transform one genome into the other.

- Analyzing distance is a hard problem.

- In contrast to reversal distance, 2-break distance is easier to compute and provides a reasonable approximation to reversal distance.

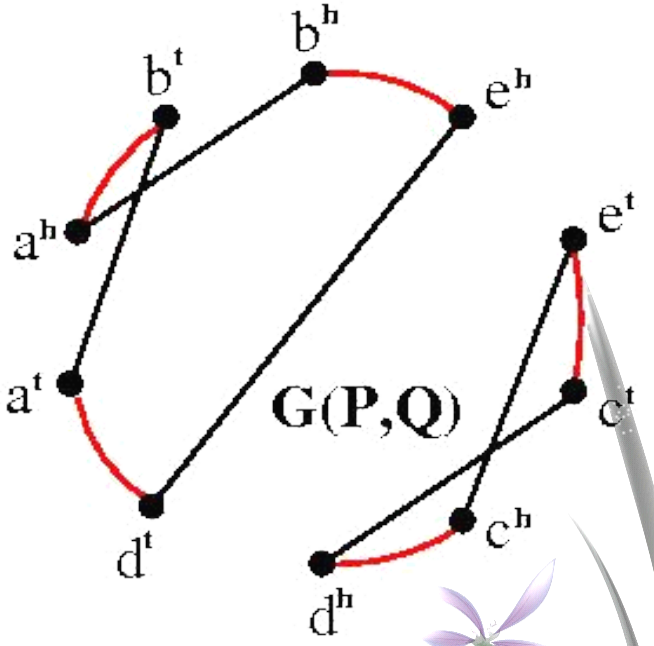
Breakpoint Graphs

Given genomes *P* and *Q*, we define the ***Breakpoint Graph*** *G(P,Q)* as the superposition of the genome graphs of *P* and *Q*.



Breakpoint Graph is a Collection of Cycles

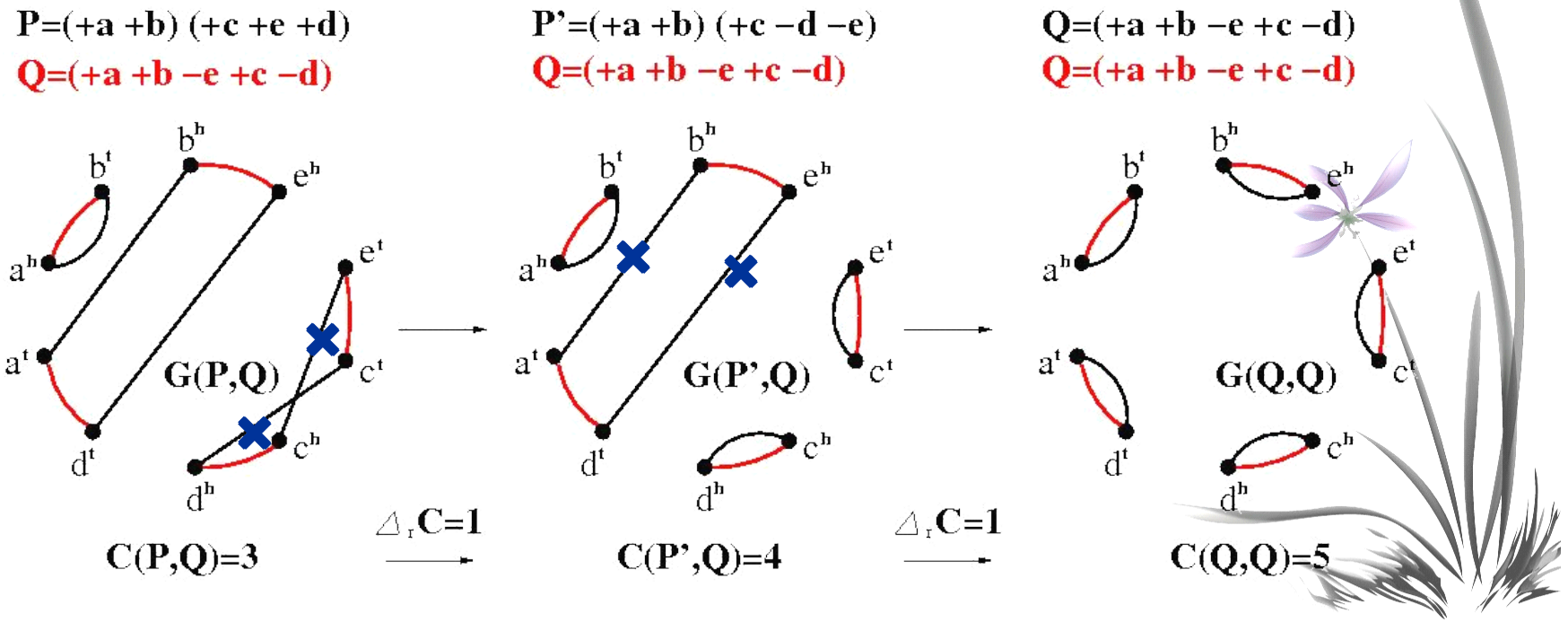
- In the breakpoint graph *G(P,Q)*, the black edges and the red edges form a collection of black-red alternating cycles, where the colors of the edges alternate.



- We denote ***C(P,Q)*** as the number of cycles in the breakpoint graph *G(P,Q)* and refer to the change of *C(P,Q)* caused by any rearrangment *r* as $\Delta_r C$.

Tranformations and Breakpoint Graphs

- Transformation is a sequence of genome rearrangements between two genomes.
- Any transformation of a genome *P* into a genome *Q* corresponds to a transformation of the breakpoint graph *G(P,Q)* into the breakpoint graph *G(Q,Q)* .

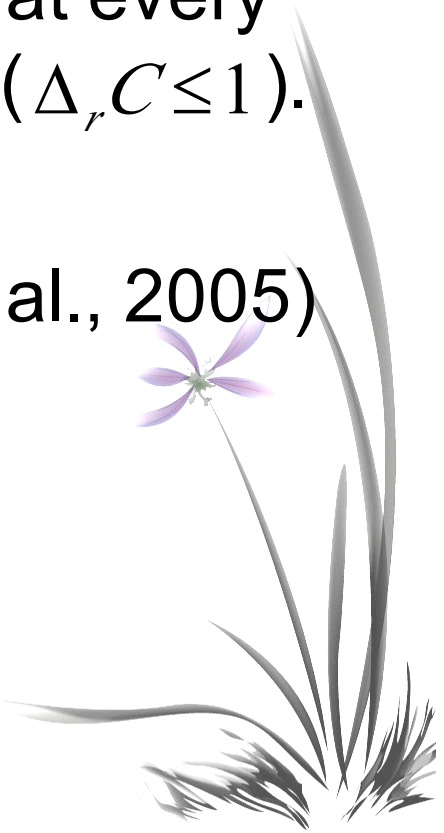


2-break Distance Formula

In any transformation, *C(P,Q)* will be increased to *C(Q,Q)* =|*Q*|, where |*Q*| is the number of genes in genome *Q*. It can be shown that every 2-break increases *C(P,Q)* by at most 1 ($\Delta_r C \leq 1$).

The 2-break distance (Yancopoulos et al., 2005) between genomoes *P* and *Q* is

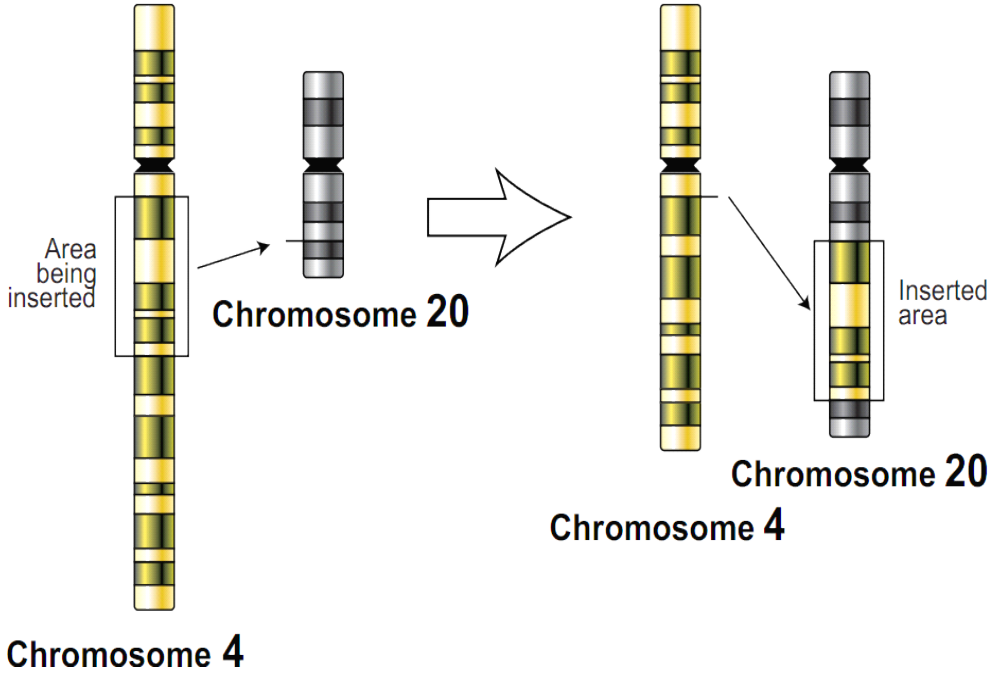
$$d_2(P,Q) = |Q| - C(P,Q)$$



Transpositions and Weighted Genomic Distance

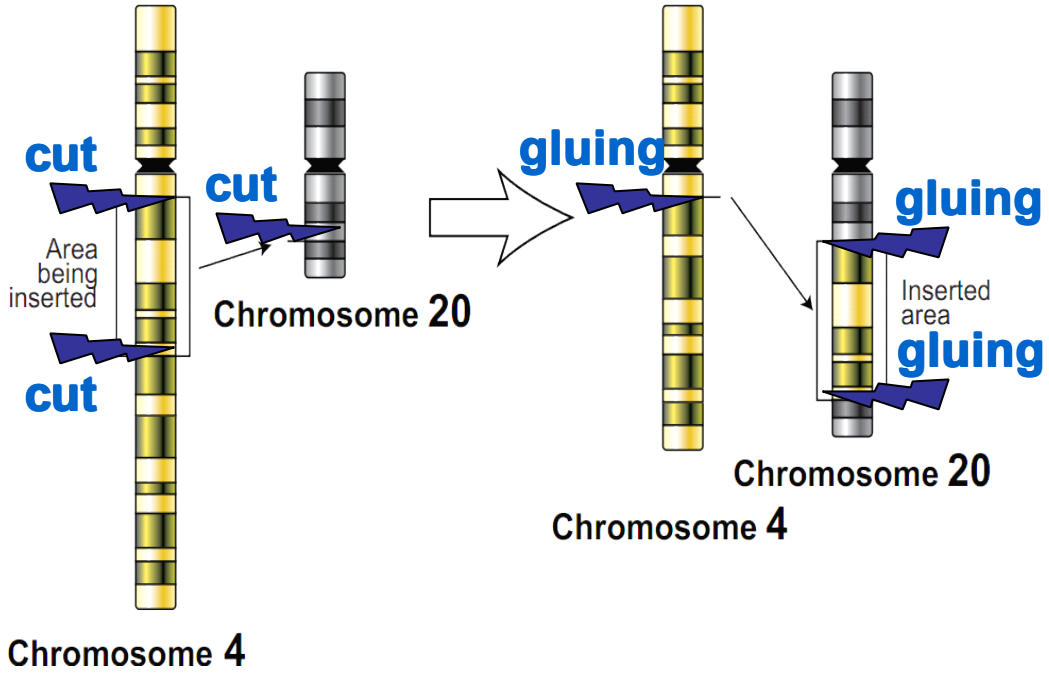
Transpositions

- A ***transposition*** cuts off a segment of a chromosome and inserts it into some other position in the genome.



Transpositions

- A transposition makes 3 cuts and 3 gluings.

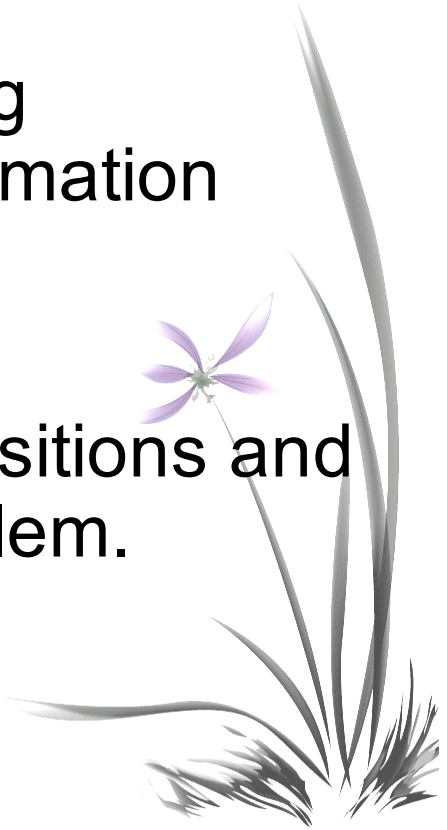


Computing Transposition Distance

- Transpositions are very hard to ananlyze. The complexity of computing transposition ditance remains an open problem.

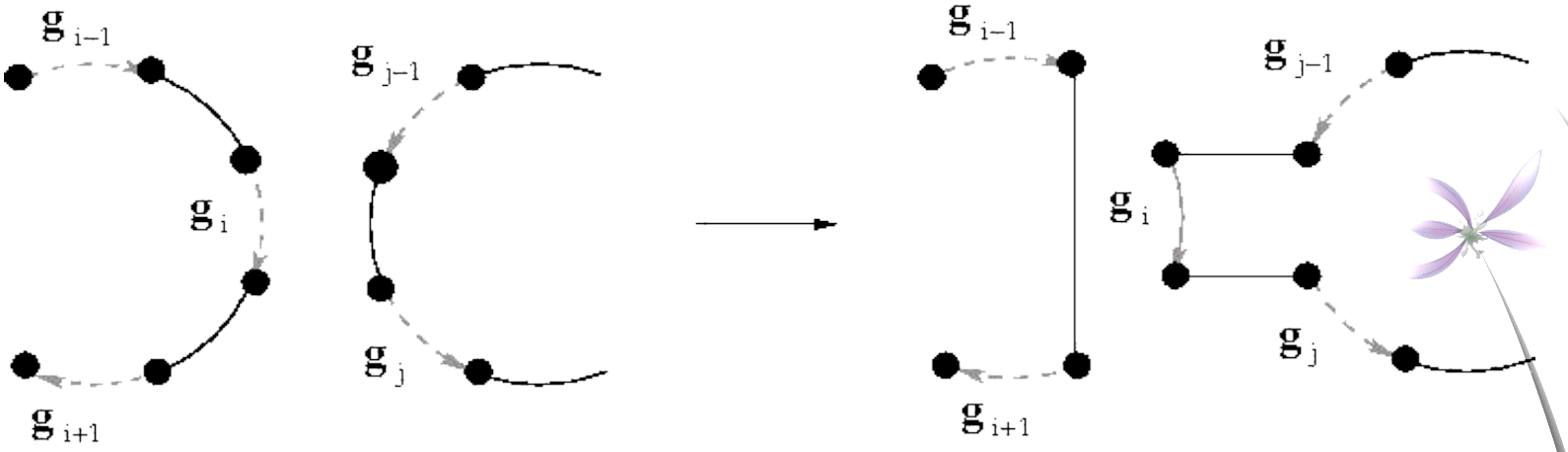
- The best known algorithm for computing transposition distance is 1.375-approximation algorithm by Elias and Hartman, 2006.

- 3-break rearrangements model transpositions and lead to a simpler 3-break distance problem.



3-break Rearrangements

- ***3-break rearrangement*** is a replacement of 3 black edges in a genome graph with 3 different black edges on the same set of vertices.
- 3-breaks model transposition-like rearrangements.

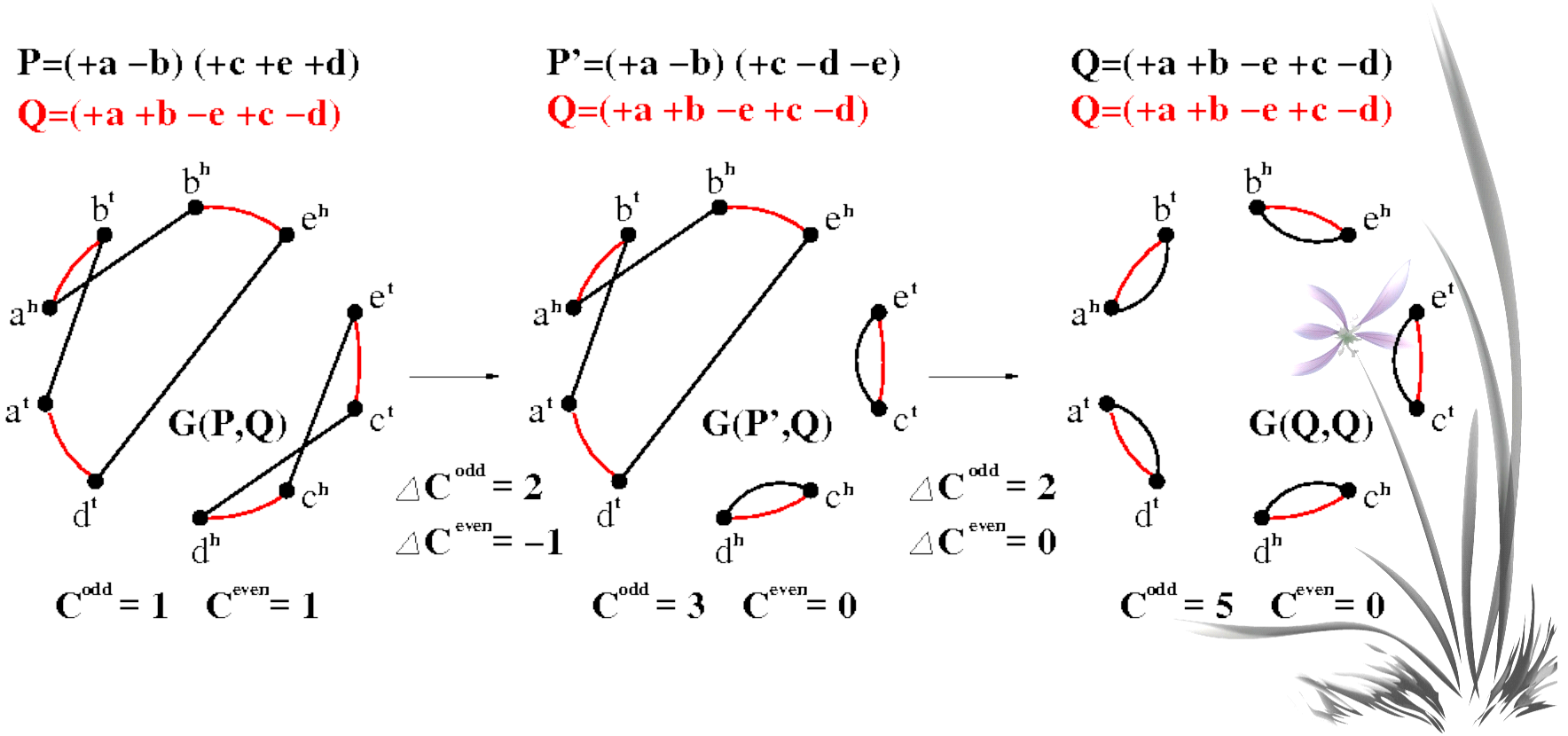


$(\ldots g_{i-1}, g_i, g_{i+1} \ldots)(\ldots g_{j-1}, g_j \ldots) \longrightarrow (\ldots g_{i-1}, g_{i+1} \ldots)(\ldots g_{j-1}, g_i, g_j \ldots)$

Gene g_i is transposed into a different chromosome.

Tranformations with 3-breaks

- We denote the number of ***even cycles*** and ***odd cycles*** in breakpoint graph *G(P,Q)* as $C^{even}(P,Q)$ and $C^{odd}(P,Q)$.
  - ***Even Cycles*** have an even number of black edges.
  - ***Odd Cycles*** have an odd number of black edges.



3-break Distance Formula

- The ***2-break distance*** (Yancopoulos et al., 2005) between genomoes *P* and *Q* is

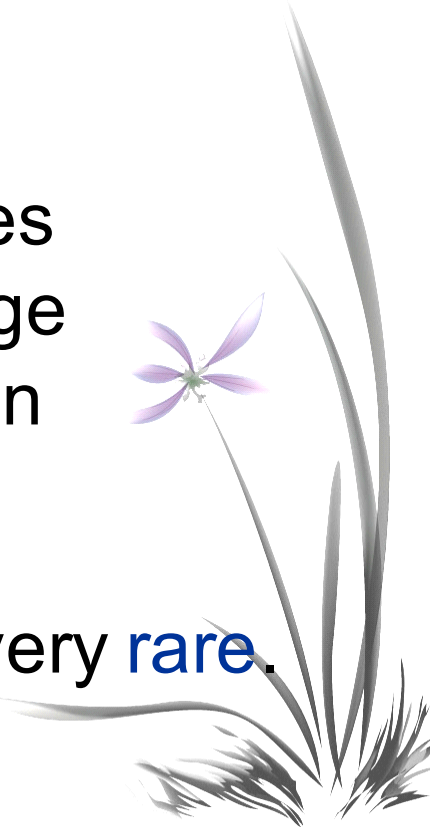
$$d_2(P,Q) = |Q| - C(P,Q)$$

- The ***3-break distance*** (Alekseyev and Pevzner, 2007) between genomoes *P* and *Q* is

$$d_3(P,Q) = \frac{|Q| - C^{odd}(P,Q)}{2}$$

Transpositions are "powerful" but rare

- ***Shortest transformations*** between two genomes have the minimal number of rearrangemments (equal the genomic distance between these genomes).
- When the rearrangement model includes transpositions, they typically have a large proportion in any shortest transformation between two genomes.
- However, in reality, transpositions are very rare.



Weight of a Transformation

- To bound the proportion of transpositions in a transformation, transpositions are assigned a relative weight $\alpha > 1$.

- For a transformation *t*, we define $n_2(t)$ as the number of reversal-like rearrangements (that make 2 cuts) and $n_3(t)$ as the number of transposition-like rearrangements (that make 3 cuts), then the ***weight of the transformation*** *t* is

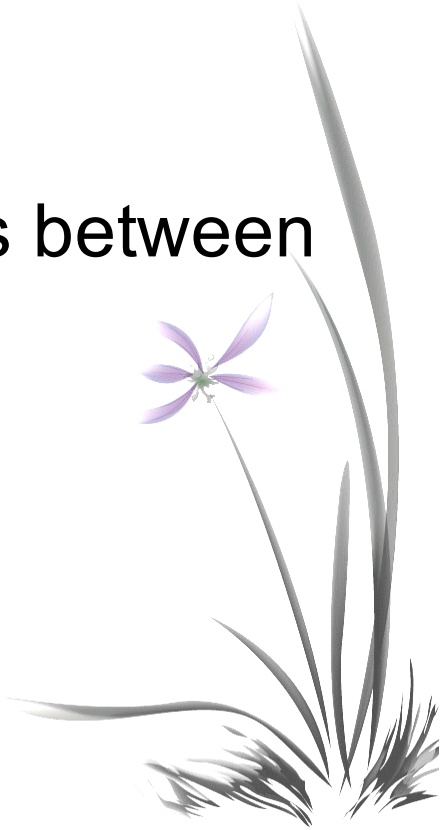
$$W_\alpha(t) = 1 \cdot n_2(t) + \alpha \cdot n_3(t)$$

Weighted Genomic Distance

We define the ***weighted genomic distance*** between two genomes as the minimal possible weight of a transformation between them, that is,

$$\min_{t} W_{\alpha}(t),$$

where t ranges over all transformations between these genomes.

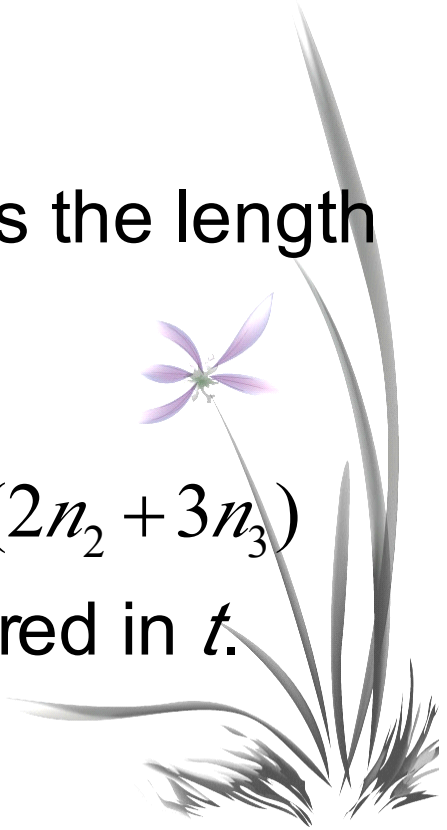


Examples

The weight of a transformation t is

$$W_\alpha(t) = n_2(t) + \alpha \cdot n_3(t)$$

- When $\alpha = 1$, $W_1(t) = n_2(t) + n_3(t)$ represents the length of the transformation t.
- When $\alpha = 3/2$, $W_{3/2}(t) = n_2(t) + 3/2\, n_3(t) = 1/2(2n_2 + 3n_3)$ stands for half of the number of cuts occured in t.



Previous Results

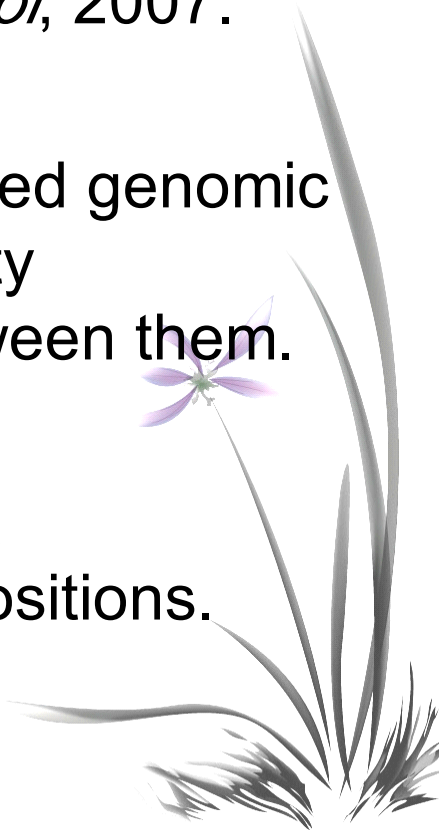
- The complexity of computing weighted genomic distance remains unknown.
- Barder and Ohlebusch, 2007 developed a 1.5-approximation algorithm for computing the weighted genomic distance for $\alpha \in [1,2]$.
- Eriksen, 2001 proposed a $(1+\varepsilon)$ - approximation algorithm for computing the weighted genomic distance for $\alpha = 2, \text{ and any } \varepsilon > 0.$
- Blanchette et al, 1995, observed that if $\alpha \in (1,2),$ then typical transformation still includes large proportion of transpositions.

Our Contribution

- We first characterize ***optimal transformations*** that at the same time have the shortest length and make the smallest number of cuts in the genomes, first introduced by Alekseyev and Pevzner, *PLoS Comput. Biol*, 2007.

- Then, we prove that for $\alpha \in (1,2]$, the weighted genomic distance between two genomes with necessity corresponds to an optimal tranformation between them.

- In particular, we show that a minimum-weight transformation may entirely consist of transpositions.



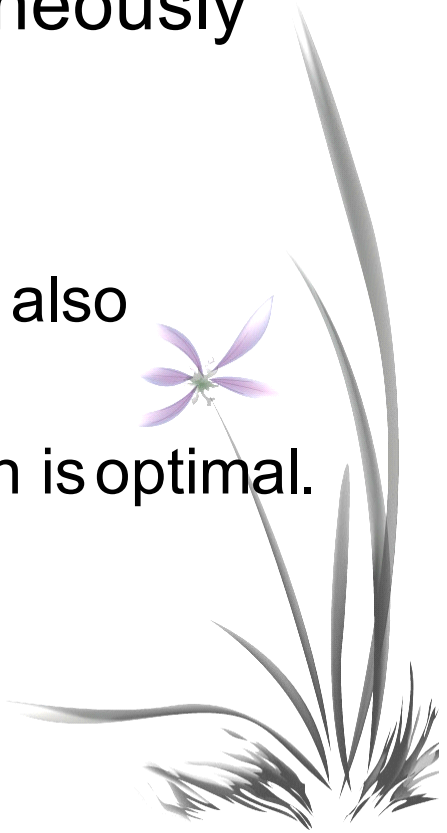
Optimal Transformations

Shortest and Optimal Transformations

- A transformation is ***shortest*** if it minimize $W_1(t)$.

- A transformation is ***optimal*** if it simultaneously minimizes $W_1(t)$ and $W_{3/2}(t)$.

  - ✓ By definition, any optimal transformation is also shortest.
  - ✓ However, not every shortest transformation is optimal.



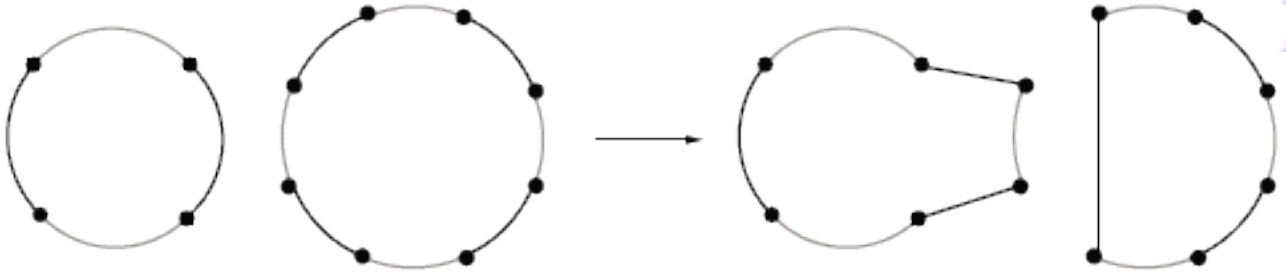
Classification of Shortest Transformations

- Theorem 1.

  A transformation t between two genomes is shortest if and only if for any rearrangement $r \in t$,

  if r is a 2-break, then $\Delta_r C^{odd} = 2$ and $\Delta_r C^{even} = -1$;

  if r is a 3-break, then $\Delta_r C^{odd} = 2$ and $\Delta_r C^{even} = 0$ or $\boxed{\Delta_r C^{even} = -2}$.



A 3-break r with $\Delta_r C^{odd} = 2$ and $\Delta_r C^{even} = -2$

Classification of Optimal Transformations

- Theorem 2.

  A transformation t between two genomes is optimal if and only if for any rearrangement $r \in t$,

  if r is a 2-break, then $\Delta_r C^{odd} = 2$ and $\Delta_r C^{even} = -1$;

  if r is a 3-break, then $\Delta_r C^{odd} = 2$ and $\Delta_r C^{even} = 0$ or ~~$\Delta_r C^{even} = -2$~~.

- Collorary 1.

  An optimal transformation has no rearrangement with $\Delta_r C^{even} = -2$.

Number of 2-breaks and 3-breaks in Optimal Transformations

- Theorem 3.

  A transformation t between genomes P and Q is optimal if and only if, the number of 2-breaks and 3-breaks in t is

$$\begin{cases} n_2(t) = C^{even}(P,Q), \\ n_3(t) = \dfrac{|P| - C^{odd}(P,Q)}{2} - C^{even}(P,Q) \end{cases}$$

Main Theorem

Main Theorem

- Theorem 4.

  For $\alpha \in (1,2]$,

$$\min_t \{W_\alpha(t)\} = W_\alpha(t_0),$$

  where t goes over all transformations between two genomes and t_0 is any optimal transformation between them.

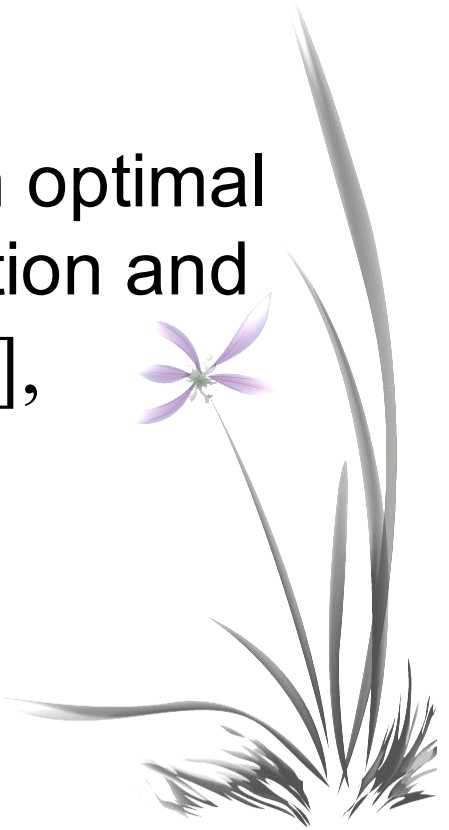
- That is, for any $\alpha \in (1,2]$, the weighted genomic distance between two genomes equals the weight of any optimal transformation between them.

Proof Idea

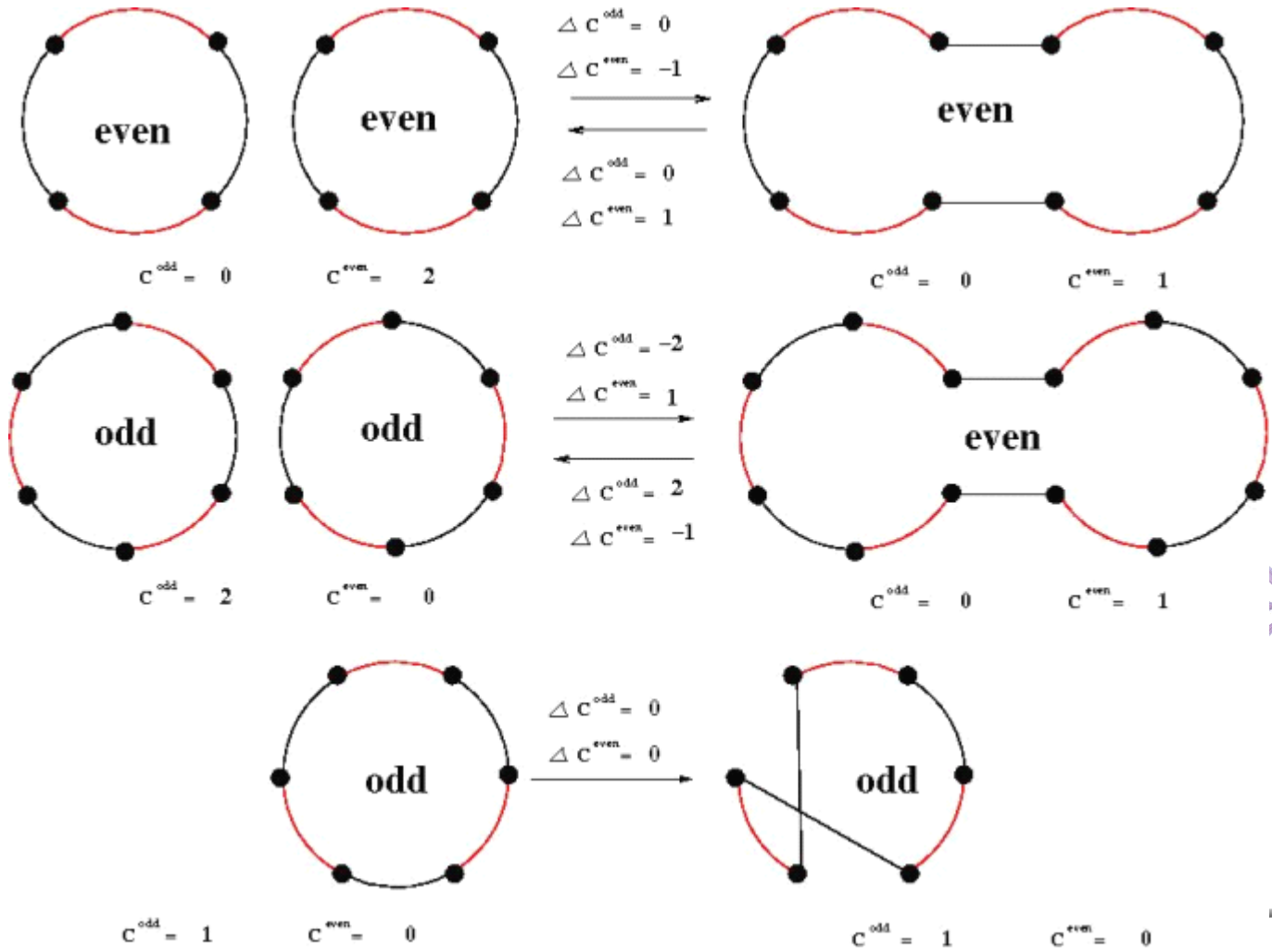
- We classify all possible changes in the numbers of odd and even black-red cycles caused by a rearrangement *r*.

- For an arbitrary transformation *t* and an optimal transformation t_0, we use our classification and Theorem 4 to show that for any $\alpha \in (1,2]$,

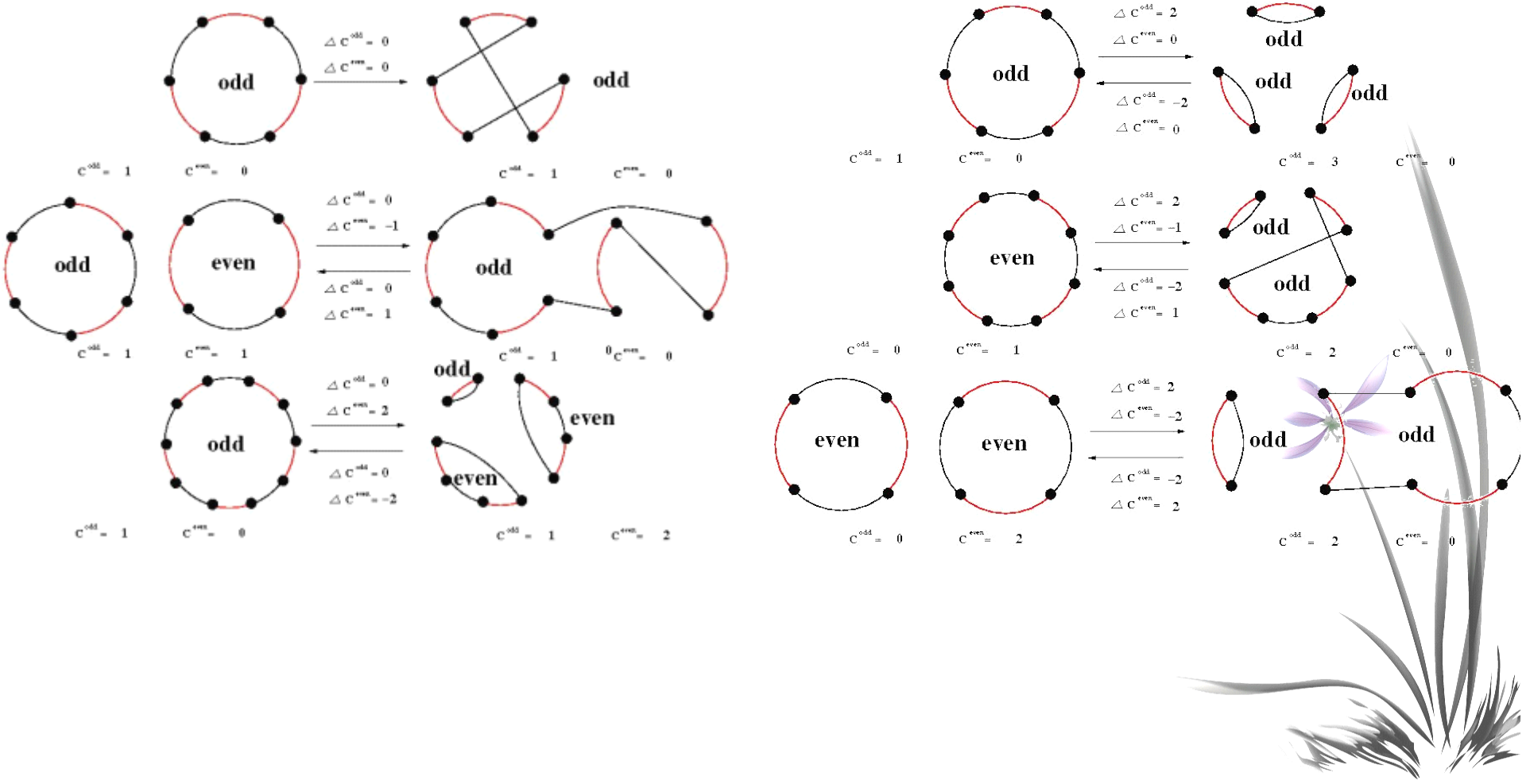
$$W_\alpha(t) \geq W_\alpha(t_0).$$



Classification of Possible $\Delta_r C^{even}$ and $\Delta_r C^{odd}$ for 2-breaks r



Classification of Possible $\Delta_r C^{even}$ and $\Delta_r C^{odd}$ for 3-breaks r



	$n_3(r)=0$					$n_3(r)=1$										
$\Delta_r c^{odd}$	0	0	0	-2	2	0	0	0	0	0	2	2	2	-2	-2	-2
$\Delta_r c^{even}$	0	1	-1	1	-1	0	1	-1	2	-2	0	-1	-2	0	1	2
amount in t	x_1	x_2	x_3	x_4	x_5	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	y_{11}

Given all possible types of rearrangements and their amounts, we have $n_2(t)$ and $n_3(t)$

For the transformation t, we have

$$\begin{cases} n_2(t) = x_1 + x_2 + x_3 + x_4 + x_5, \\ n_3(t) = y_1 + y_2 + y_3 + y_4 + y_5 + y_6 + y_7 + y_8 + y_9 + y_{10} + y_{11}. \end{cases}$$

For the optimal transformation t_0 we derive from Theorem 3 that

$$\begin{cases} n_2(t_0) = -x_2 + x_3 - x_4 + x_5 - y_2 + y_3 - 2y_4 + 2y_5 + y_7 + 2y_8 - y_{10} - 2y_{11}, \\ n_3(t_0) = x_2 - x_3 + y_2 - y_3 + 2y_4 - 2y_5 + y_6 - y_8 - y_9 + y_{11}. \end{cases}$$

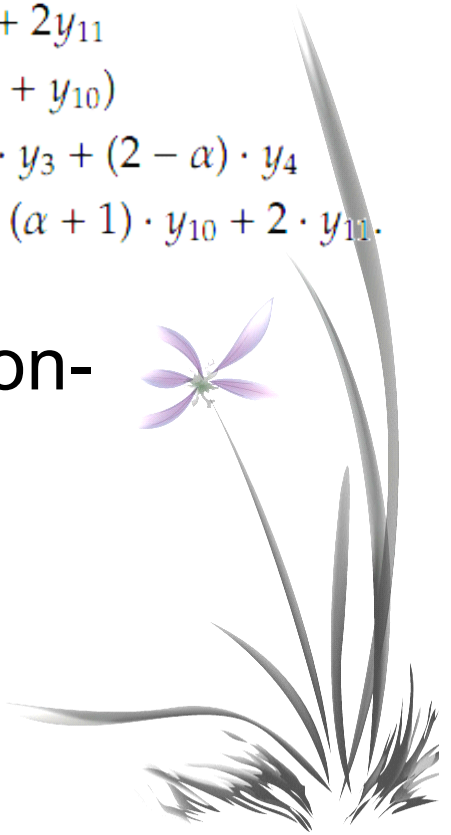
Result

- Now, we can compute $W_\alpha(t) - W_\alpha(t_0)$ as follows:

$$
\begin{aligned}
W_\alpha(t) - W_\alpha(t_0) &= n_2(t) - n_2(t_0) + \alpha \cdot (n_3(t) - n_3(t_0)) \\
&= x_1 + 2x_2 + 2x_4 + y_2 - y_3 + 2y_4 - 2y_5 - y_7 - 2y_8 + y_{10} + 2y_{11} \\
&\quad + \alpha \cdot (-x_2 + x_3 + y_1 + 2y_3 - y_4 + 3y_5 + y_7 + 2y_8 + 2y_9 + y_{10}) \\
&= x_1 + (2-\alpha) \cdot x_2 + \alpha \cdot x_3 + 2x_4 + \alpha \cdot y_1 + y_2 + (2\alpha - 1) \cdot y_3 + (2-\alpha) \cdot y_4 \\
&\quad + (3\alpha - 2) \cdot y_5 + (\alpha - 1) \cdot y_7 + (2\alpha - 2) \cdot y_8 + 2\alpha \cdot y_9 + (\alpha + 1) \cdot y_{10} + 2 \cdot y_{11}.
\end{aligned}
$$

- All coefficients in the last formula are non-negative, implying that $W_\alpha(t) - W_\alpha(t_0) \geq 0$.

- That completes the proof.



Corollary

- For an optimal transformation t between genomes P and Q, we have
  - ✓ (Theorem 3)

$$\begin{cases} n_2(t) = C^{even}(P,Q), \\ n_3(t) = \dfrac{|P| - C^{odd}(P,Q)}{2} - C^{even}(P,Q) \end{cases}$$

  - ✓ (Theorem 4)

    For any $\alpha \in (1,2]$, the weight of t equals the weighted genomic distance between P and Q.

- Corollary 2.

  If $\Delta C^{even}(P,Q) = 0$, then transformation t has $n_2(t) = 0$ and thus consists entirely 3-breaks (transpositions).

Conclusions

- We proved that for $\alpha \in (1,2]$, the minimum-weight transformations include the optimal transformations that may entirely consist of transposition-like rearrangements.

- Thus, the corresponding weighted genomic distance does not actually impose any bound on the proportion of transpositions.