# Whole Genome Duplications and Genome Halving Problem

*Max Alekseyev*

2011

# Whole Genome Duplication Hypothesis

- Susumu Ohno, 1970
  - **Whole Genome Duplication Hypothesis:**

    Big leaps in evolution would have been impossible without whole genome duplications.

# Whole Genome Duplication Hypothesis Finally Confirmed After Years of Controversy

- The Whole Genome Duplication hypothesis first met with skepticism and was only recently confirmed.

- Kellis, Birren & Lander, *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**

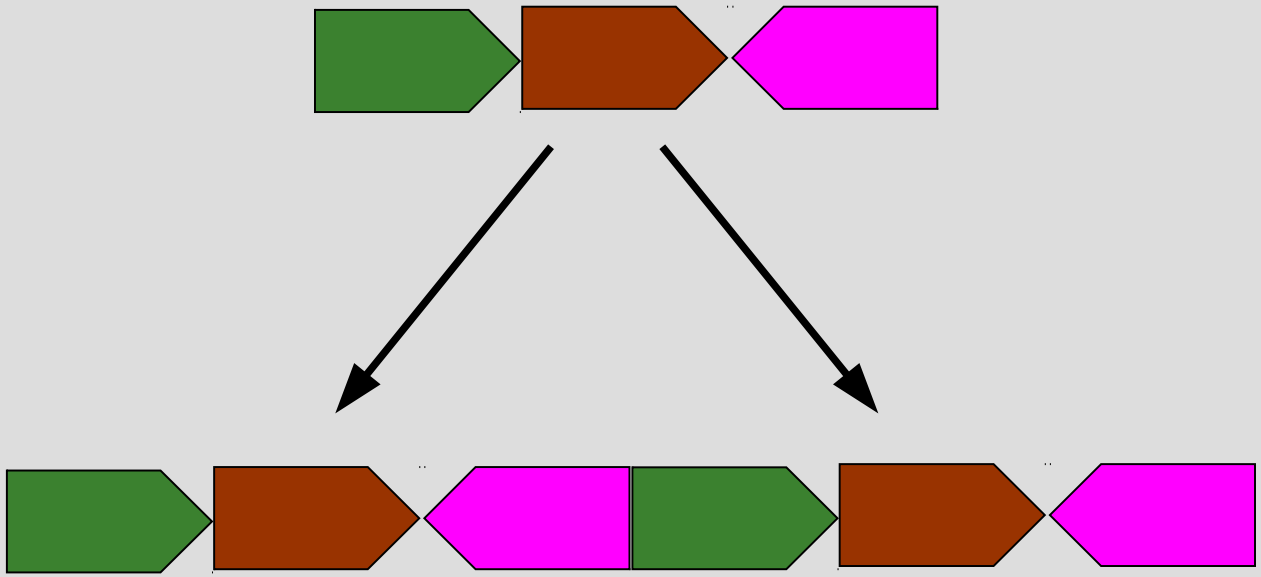
# The Whole Genome Duplication Debate: Algorithmic Challenge

- Kellis, Birren, and Lander, 2004 provided convincing arguments in favor of WGD but did not come up with a reconstruction of pre-duplicated genome and post-WGD evolutionary scenario

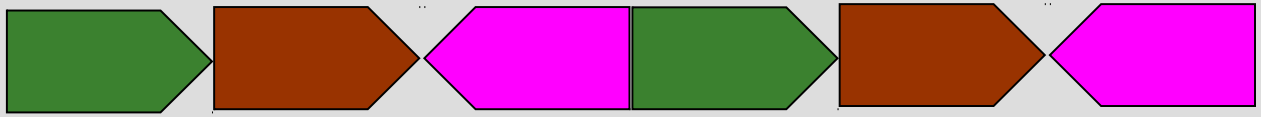
- Recent criticism of WGD is based on the claim that WGD scenario may be less parsimonious than non-WGD scenario

- **To analyze pros and cons of WGD, it would be useful to reconstruct the pre-duplicated genomes**

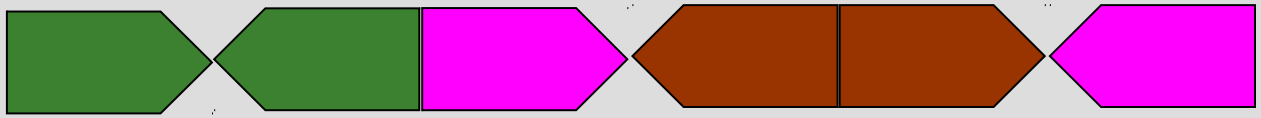
# Whole Genome Duplication



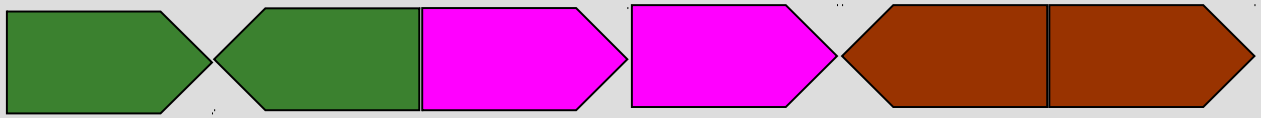
# Genome Rearrangements



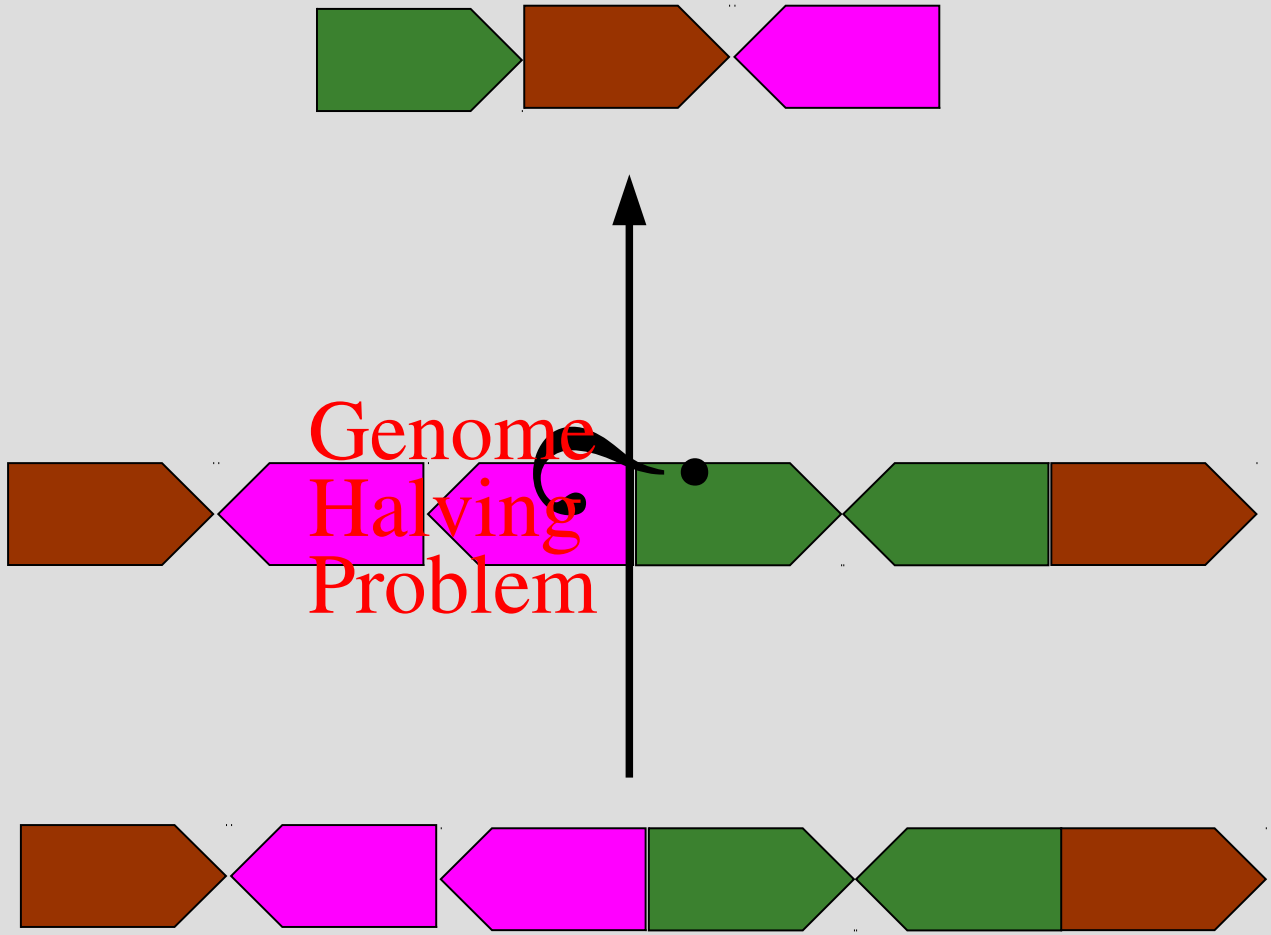
# Genome Rearrangements



# Genome Rearrangements



# Genome Halving Problem

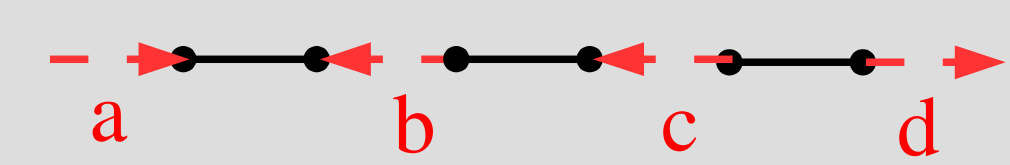


# Questions

- Can we reconstruct the pre-duplicated ancestor?
- What was the sequence of rearrangements after Whole Genome Duplication?

1. **Distance between Unichromosomal Genomes**
2. Distance between Multichomosomal Genomes
3. Breakpoint Graphs for Duplicated Genomes
4. Whole Genome Duplication and Genome Halving Problem
5. Genome Halving Problem for Multichromosomal Genomes
6. A Flaw in El-Mabrouk – Sankoff “Theorem”
7. Classification of Unichromosomal Circular Genomes

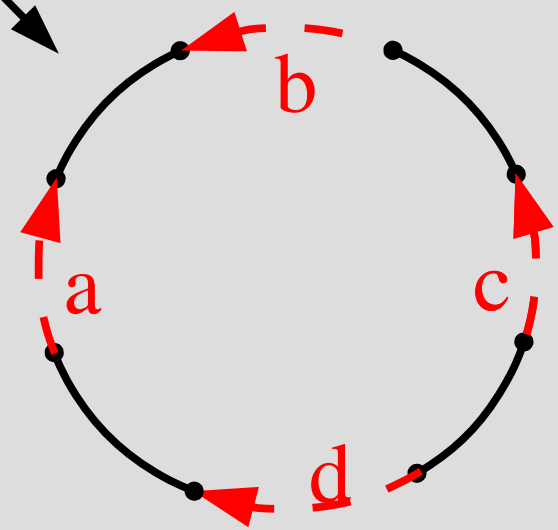
# Two Genome Graph Representations

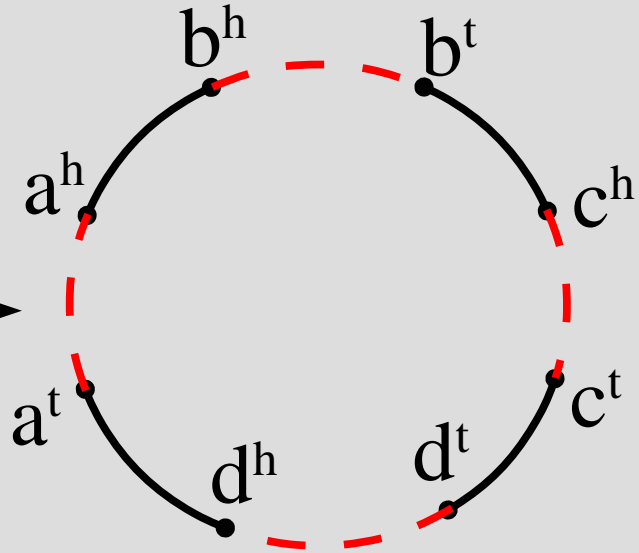



$a^t$ – *"tail"* of a

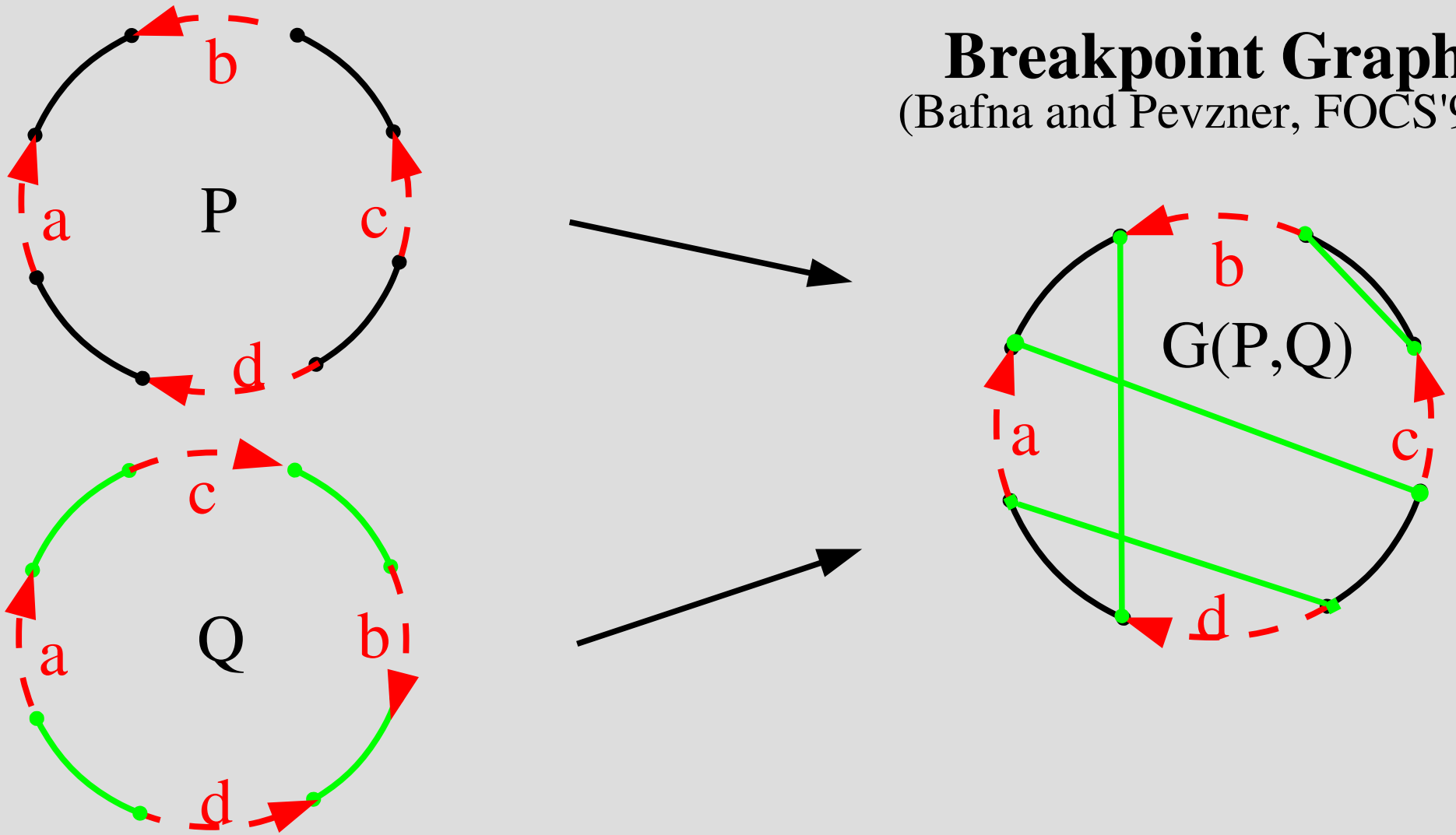
$a^h$ – *"head"* of a

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



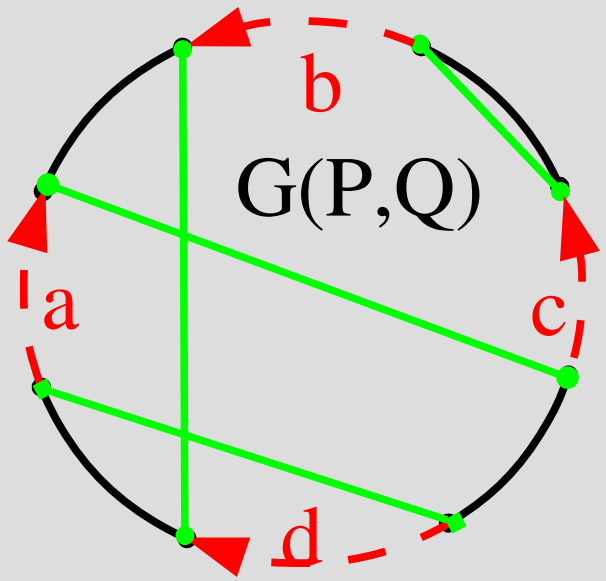



# Breakpoint Graph



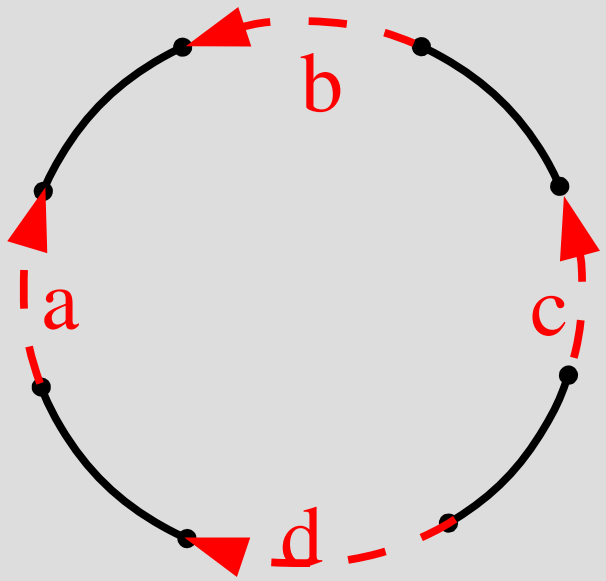
# Breakpoint Graph is Three Matchings

- Breakpoint graph is formed by ***red***, ***black*** and ***green*** matchings.
- Every pair of matchings forms a collection of *alternating* cycles:



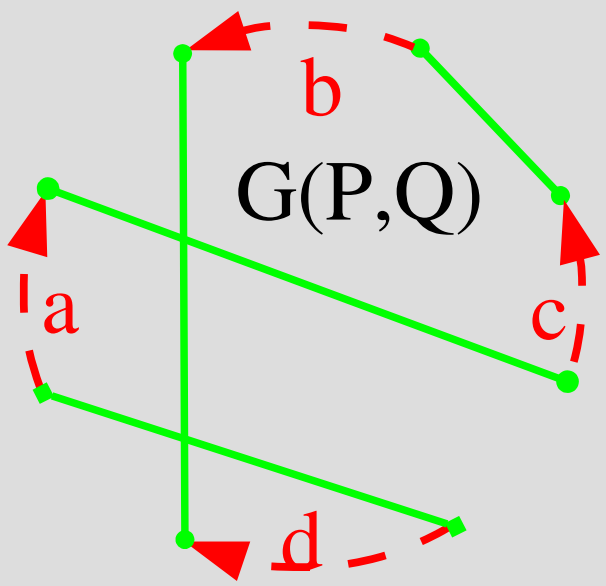
# Black-Red Cycles

- Breakpoint graph is formed by ***red***, ***black*** and ***green*** matchings.
- Every pair of matchings forms a collection of *alternating* cycles:
- ***black-red*** cycles (genome P)



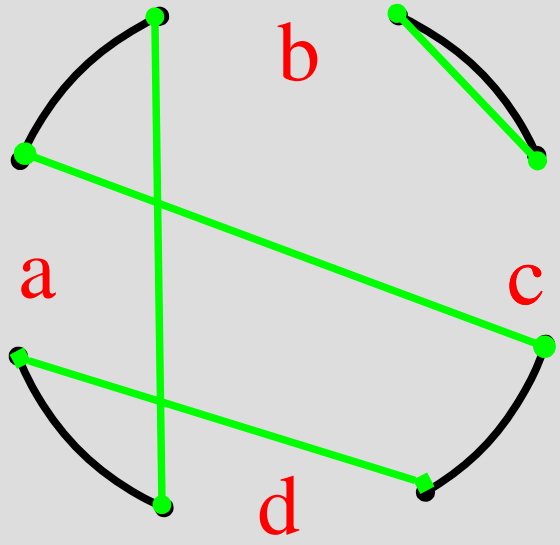
# Green-Red Cycles

- Breakpoint graph is formed by ***red***, ***black*** and ***green*** matchings.
- Every pair of matchings forms a collection of *alternating* cycles:
- ***green-red*** cycles (genome Q)



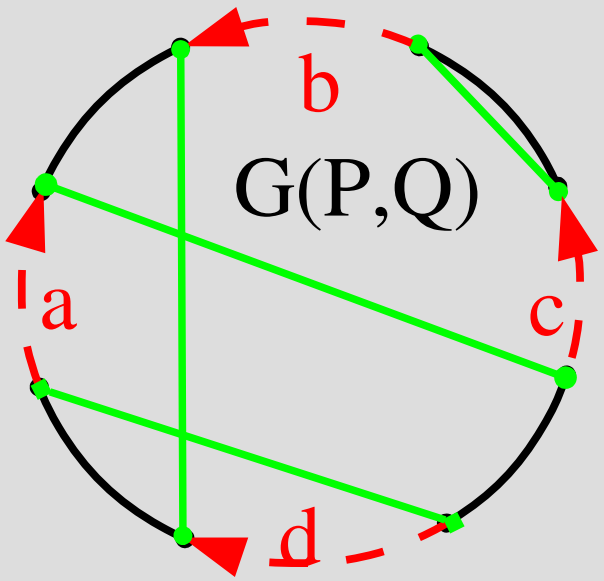
# Black-Green Cycles

- Breakpoint graph is formed by ***red***, ***black*** and ***green*** matchings.
- Every pair of matchings forms a collection of *alternating* cycles:
- ***black-green*** cycles



# Breakpoint Graph

- Breakpoint graph is formed by ***red***, ***black*** and ***green*** matchings.
- Every pair of matchings forms a collection of *alternating* cycles:
- **black-red** cycles (genome P)
- **green-red** cycles (genome Q)
- **black-green** cycles



# Hannenhalli-Pevzner Theorem

- *Reversal Distance* between genomes ***P*** and ***Q***:

$$d(P,Q) = |P| - c(P,Q) + h(P,Q)$$

**|P|=|Q|** = # of blocks (red edges) in P and Q;

***c(P,Q)*** = # of black-green cycles in G(P,Q);

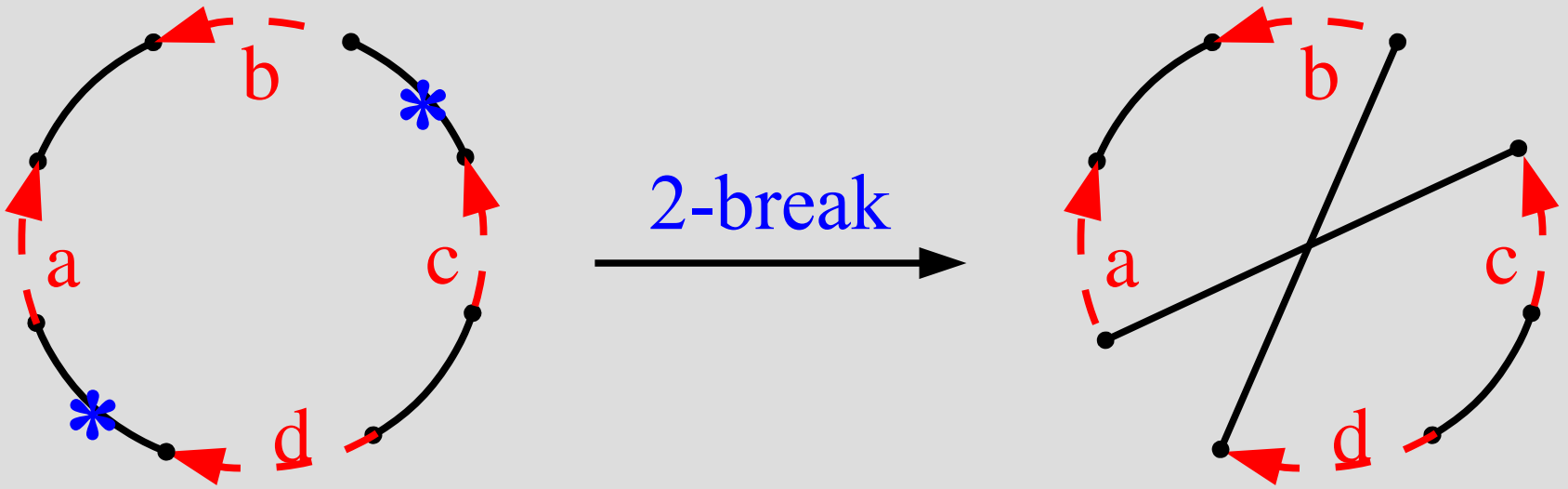
***h(P,Q)*** = combinatorial parameter (usually =0).

1. Distance between Unichromosomal Genomes

**2. Distance between Multichomosomal Genomes**

3. Breakpoint Graphs for Duplicated Genomes
4. Whole Genome Duplication and Genome Halving Problem
5. Genome Halving Problem for Multichromosomal Genomes
6. A Flaw in El-Mabrouk – Sankoff "Theorem"
7. Classification of Unichromosomal Circular Genomes

# 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/fissions represent all different 2-Breaks.

# 2-Break Distance

- The **2-Break distance** $d_2(P,Q)$ between genomes P and Q is the minimum number of 2-Breaks required to transform P into Q.

- The *2-Break Distance* between genomes P and Q:

$$\mathbf{d_2(P,Q) = |P| - c(P,Q)}$$

1. Distance between Unichromosomal Genomes
2. Distance between Multichomosomal Genomes

## 3. Breakpoint Graphs for Duplicated Genomes

4. Whole Genome Duplication and Genome Halving Problem
5. Genome Halving Problem for Multichromosomal Genomes
6. A Flaw in El-Mabrouk – Sankoff "Theorem"
7. Classification of Unichromosomal Circular Genomes

# Distance between duplicated genomes

- **Open Problem 1**: Compute the (reversal or genomic) distance between duplicated genomes.

```
P = +a −a −b +b

Q = +a −b +a +b
```

# Distance between duplicated genomes

- **Open Problem 1**: Compute the (reversal or genomic) distance between duplicated genomes.

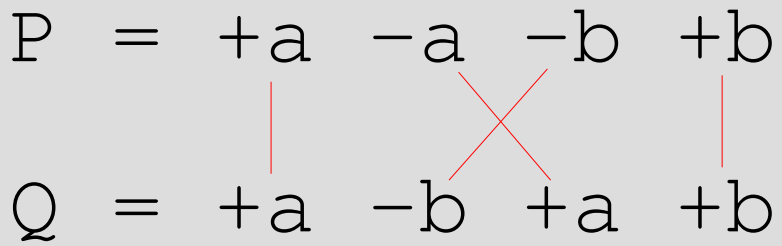
```
P = +a -a -b +b

Q = +a -b +a +b
```

Idea: Find a correspondence between gene copies in P and Q, and apply a known formula for the distance between non-duplicated genomes

# Distance between duplicated genomes

- **Open Problem 1**: Compute the (reversal or genomic) distance between duplicated genomes.




Idea: Find a correspondence between gene copies in P and Q, and apply a known formula for the distance between non-duplicated genomes

# Distance between duplicated genomes

- **Open Problem 1**: Compute the (reversal or genomic) distance between duplicated genomes.

$$P = +a_1\ -a_2\ -b_1\ +b_2$$

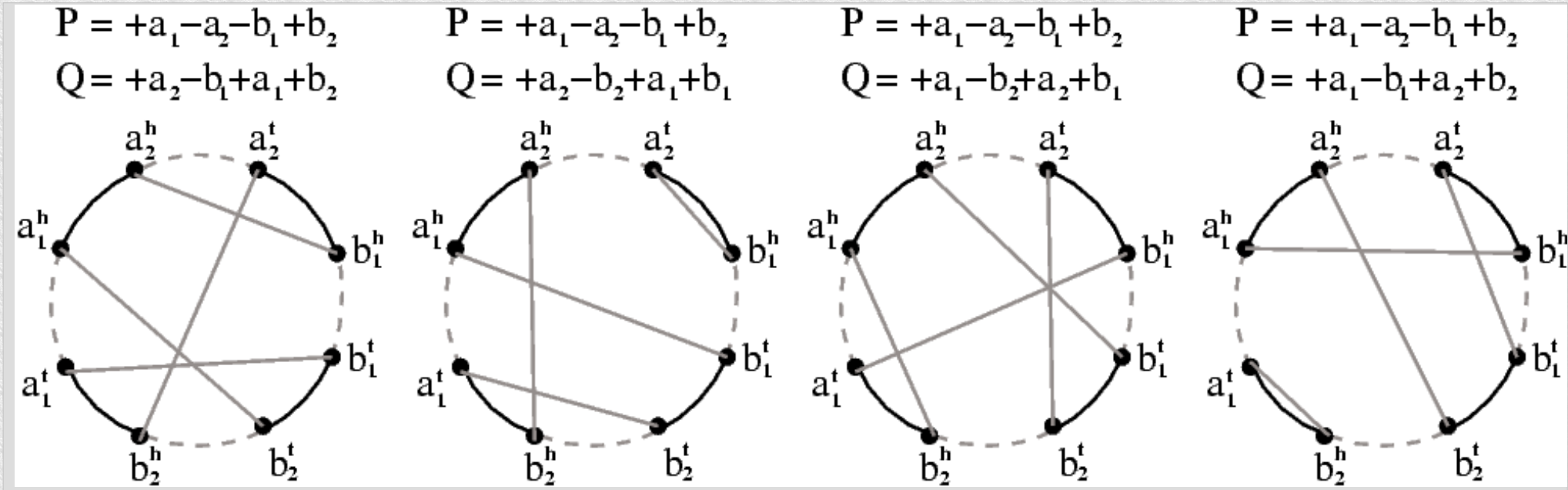
$$Q = +a_1\ -b_1\ +a_2\ +b_2$$

Find the distance between non-duplicated genomes.

# Labellings and Breakpoint Graphs

- Every labelling transforms genomes with duplicated genes into genomes without duplicated genes and enables application of HP algorithm.
- Every labelling corresponds to a breakpoint graph
- Good labellings correspond to breakpoint graphs with large number of cycles.
- Can we construct a labelling corresponding to a large number of cycles?

# Labellings and Breakpoint Graphs




c(G)=1 c(G)=2 c(G)=1 c(G)=2

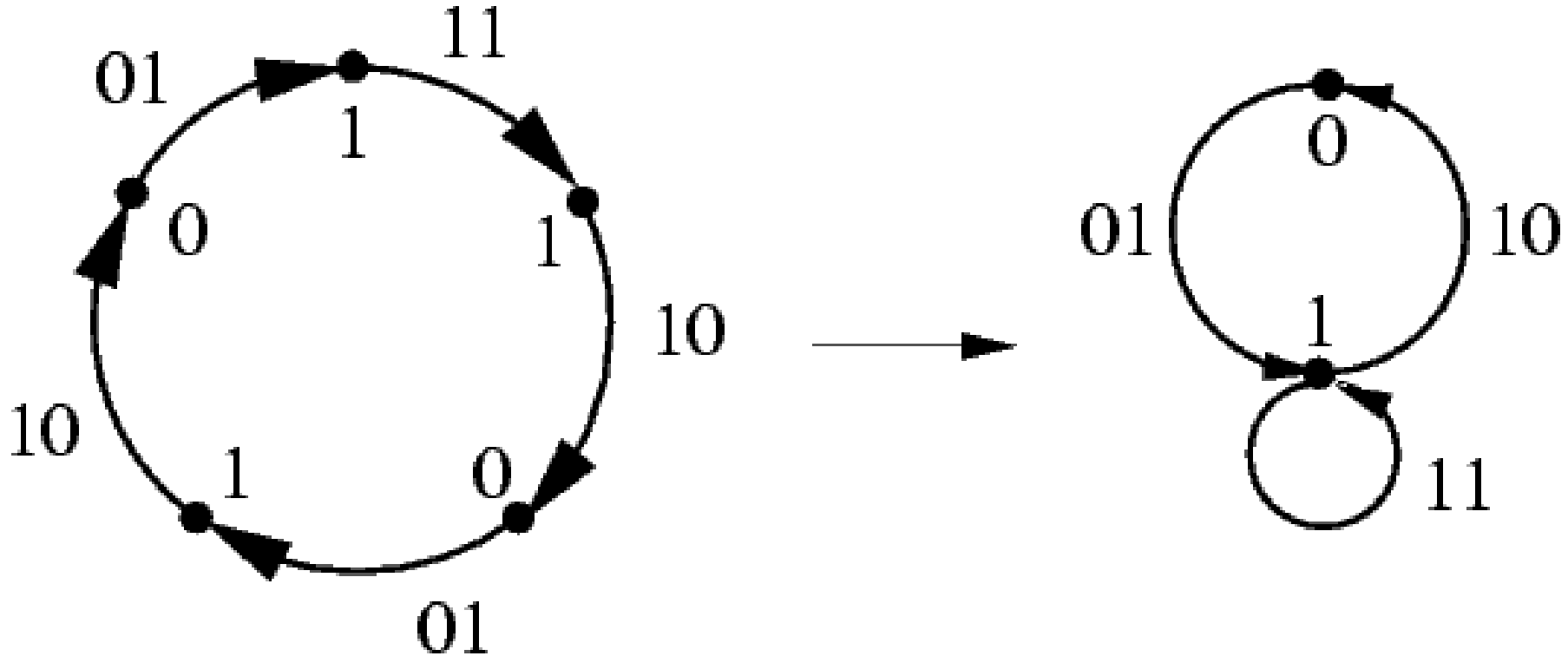
- Three are k! different labellings of k copies of a gene
- One of these labellings is unavoidably an *optimal labelling* corresponding to the optimum rearrangement scenario
- Running time: $(k!)^n$ invocations of HP algorithm for a genome with $n$ genes each present in $k$ copies.

# Challenges

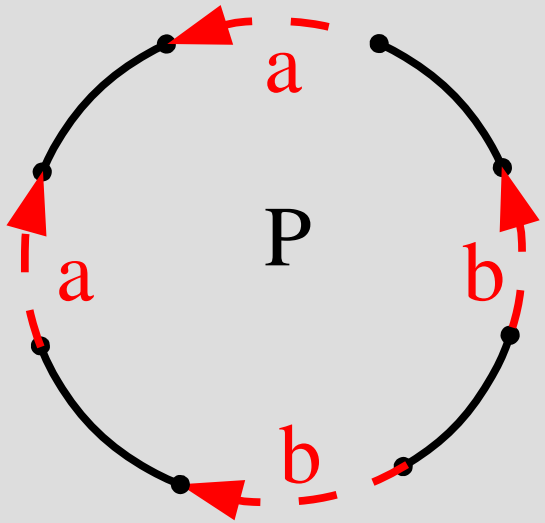
- Breakpoint graphs are not defined for duplicated genomes.
- Can we generalize the notion of breakpoint graph for the case of duplicated genomes?
- Idea: Explore the connection between de Bruijn graphs and breakpoint graphs.

# de Bruijn Graph

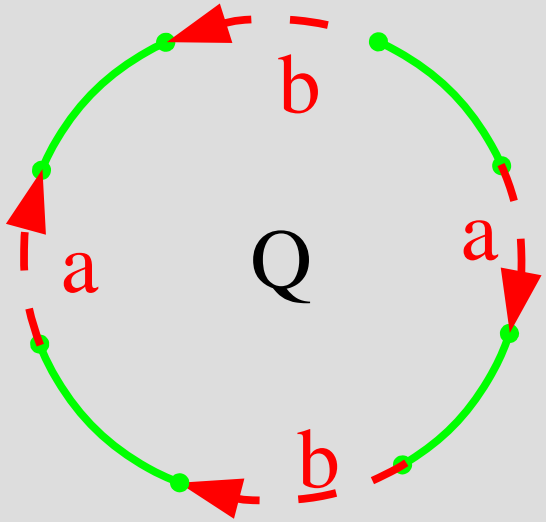
de Bruijn Graph of the circular sequence 10110



# Genomes as alternating cycles

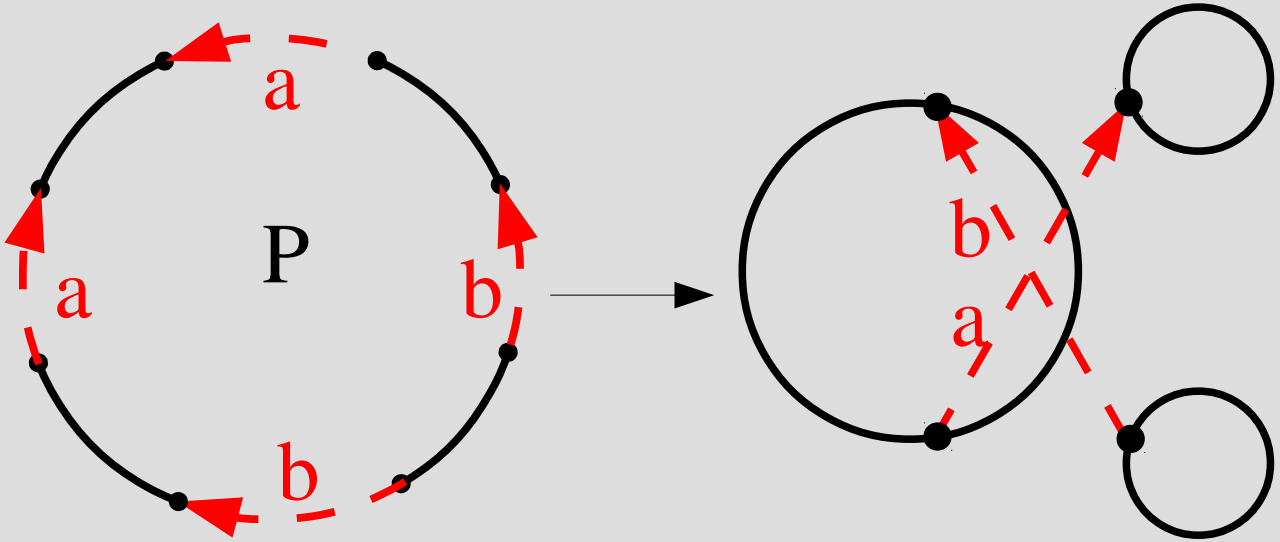



P=+a−a−b+b



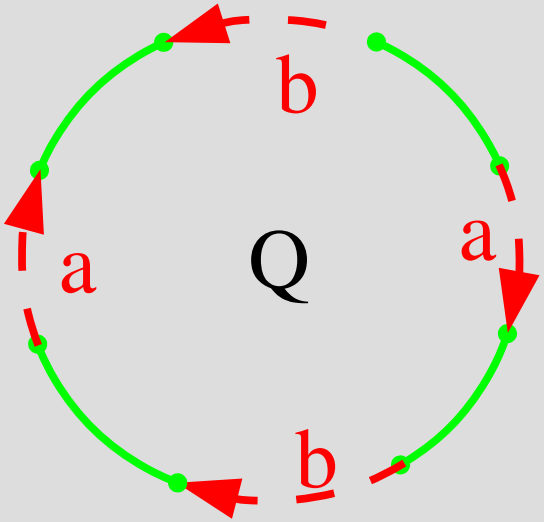

Q=+a−b+a+b

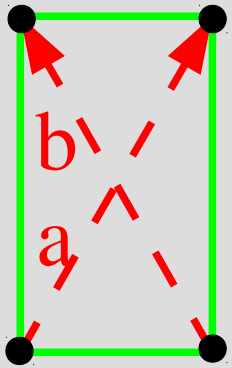
# de Bruijn Graph



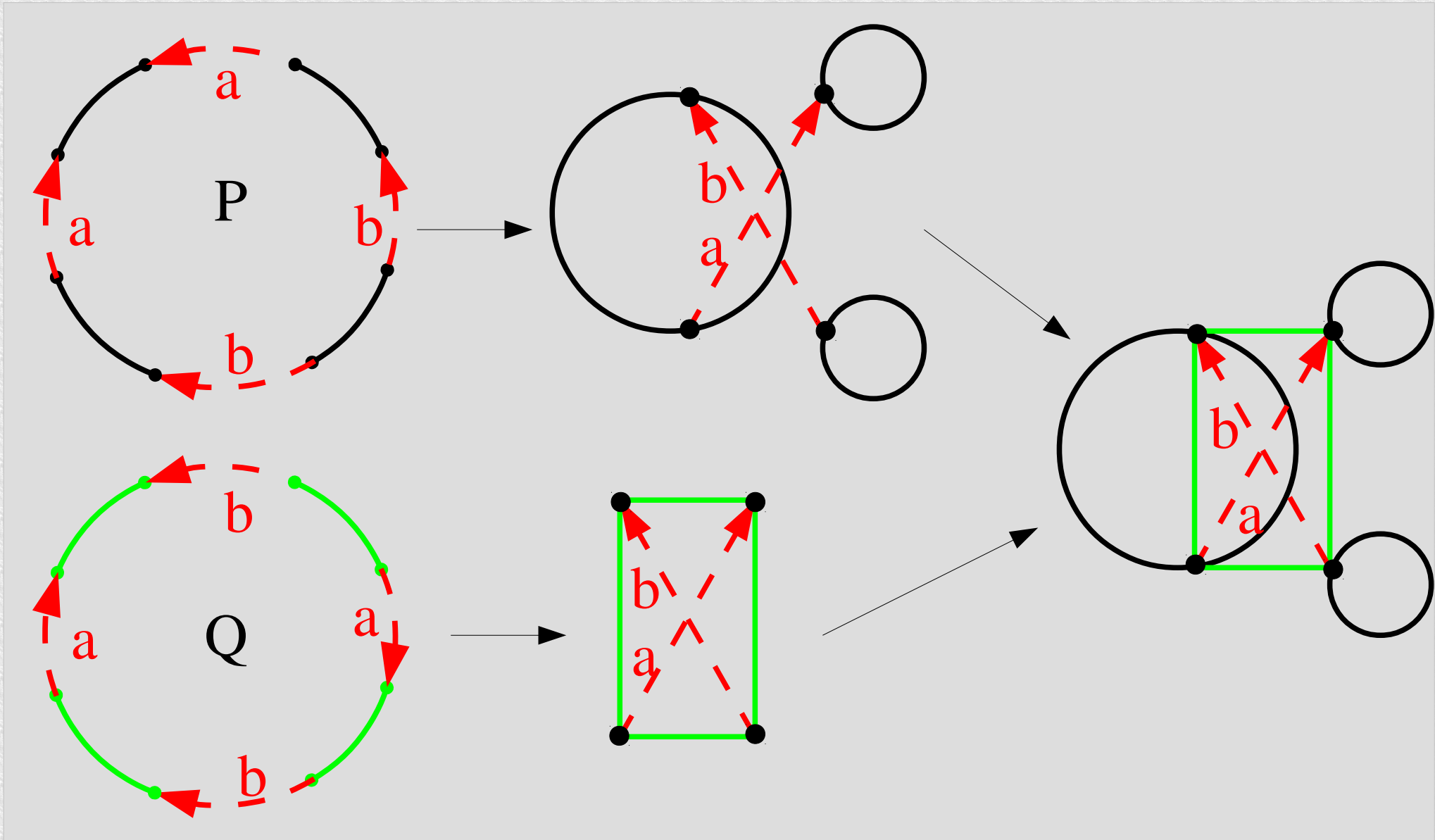




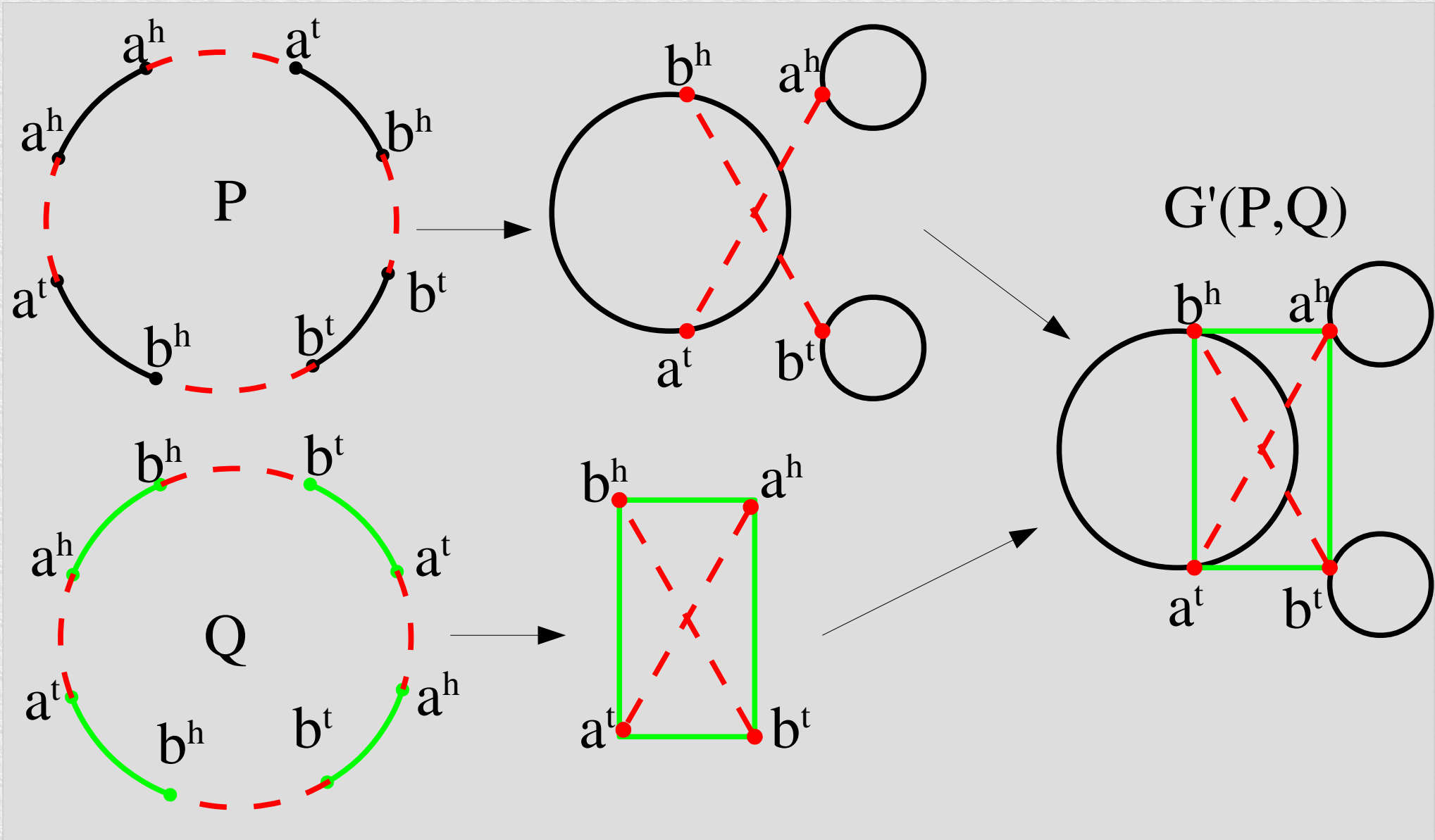


# de Bruijn Graph



# Contracted Breakpoint Graph

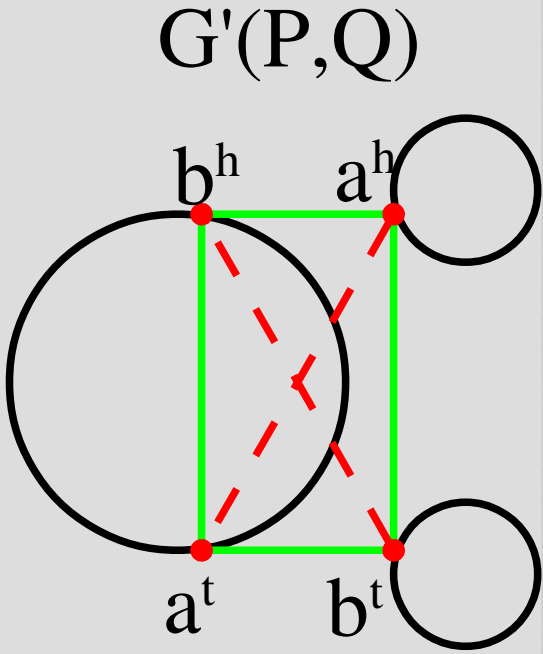


# Contracted Breakpoint Graph

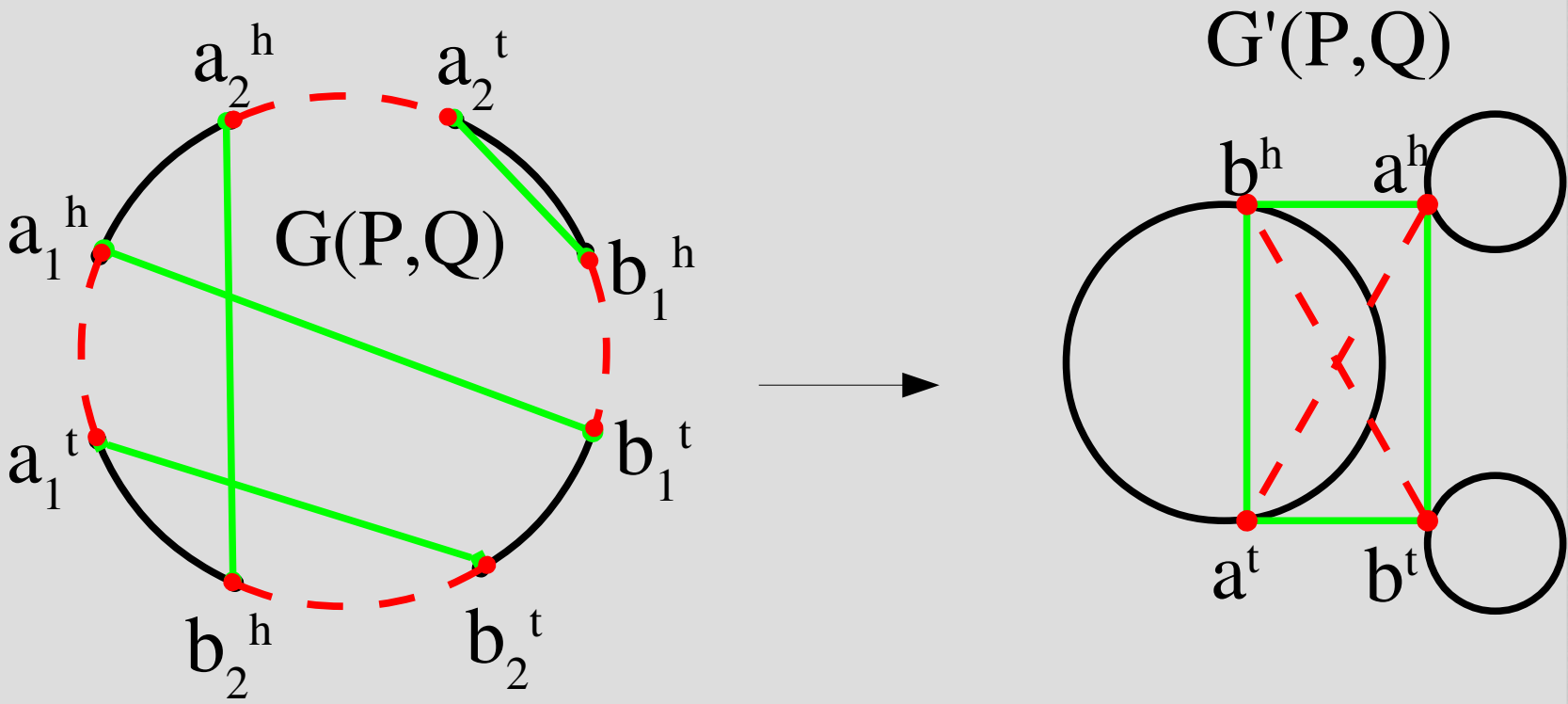
- Edges of 3 colors: ***black***, ***green***, and ***red***.

- Each vertex is incident to **two black** edges, **two green** edges, and a ***pair* of parallel red** edges

- For unichromosomal genomes, the contracted breakpoint graph is connected w.r.t. black and red edges as well as w.r.t. green and red edges

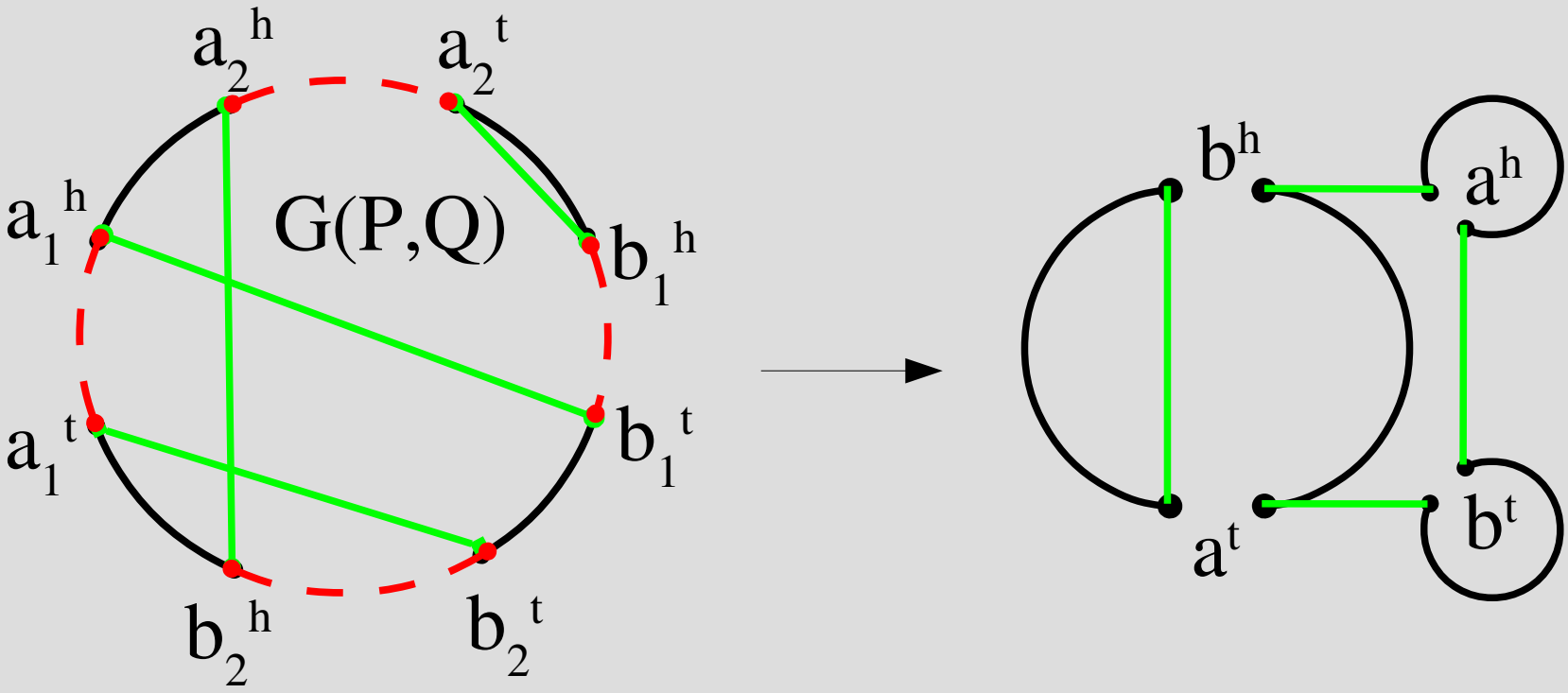


# Contracted Breakpoint Graph



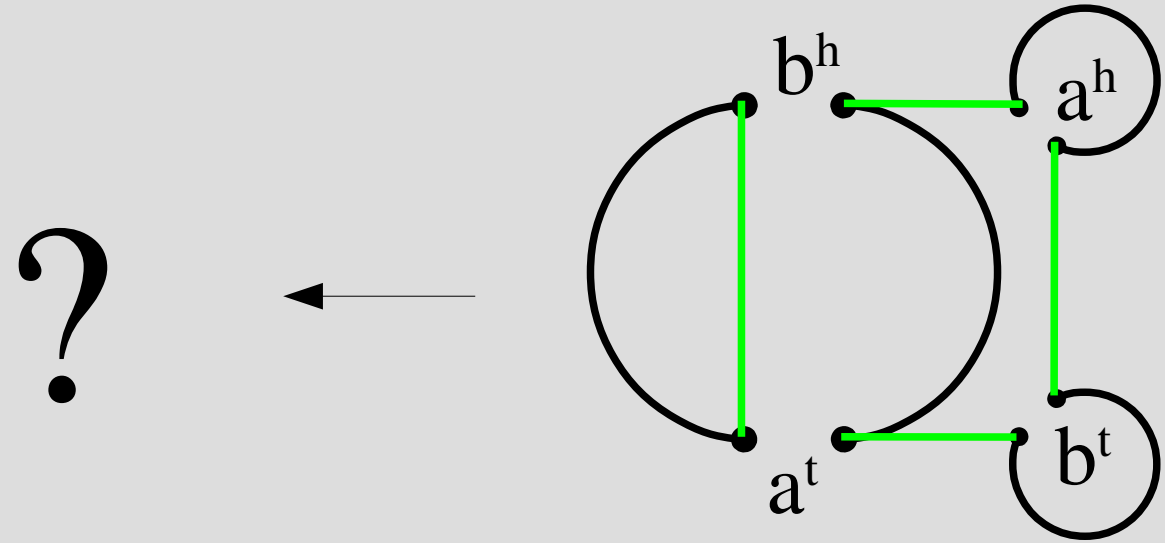
# Induced Black-green Cycle Decomposition

Every breakpoint graph induces a black-green cycle decomposition of the contracted breakpoint graph



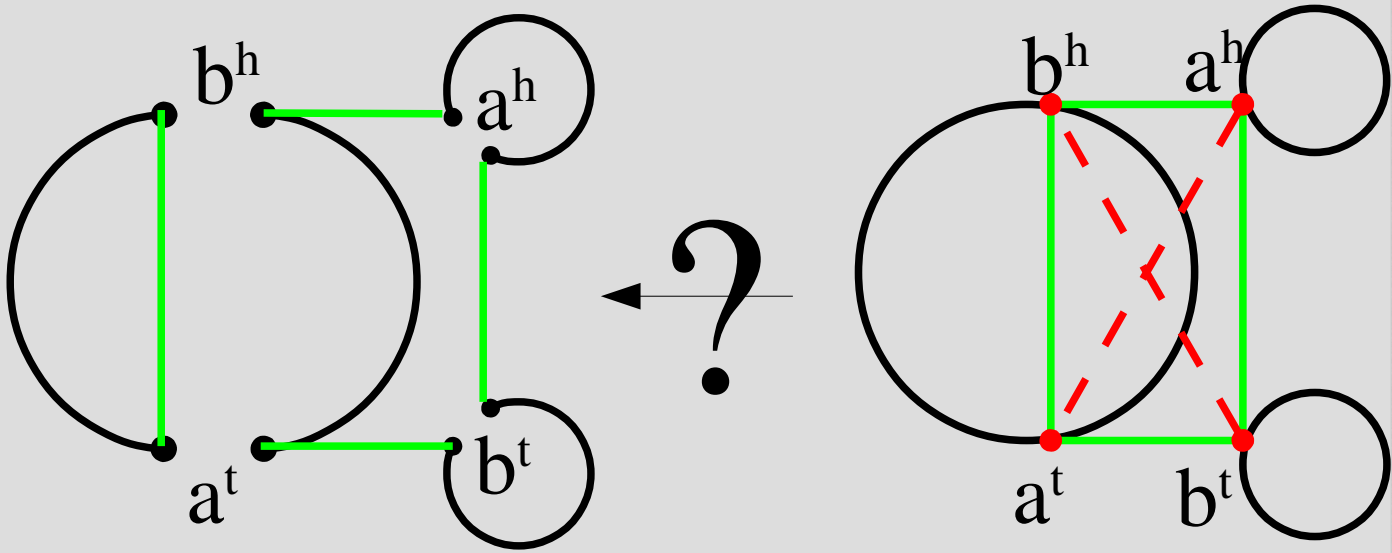
# Labelling Problem

**Open Problem 2**: Find a breakpoint graph G(P,Q) that induces given black-green cycle decomposition of G'(P,Q)

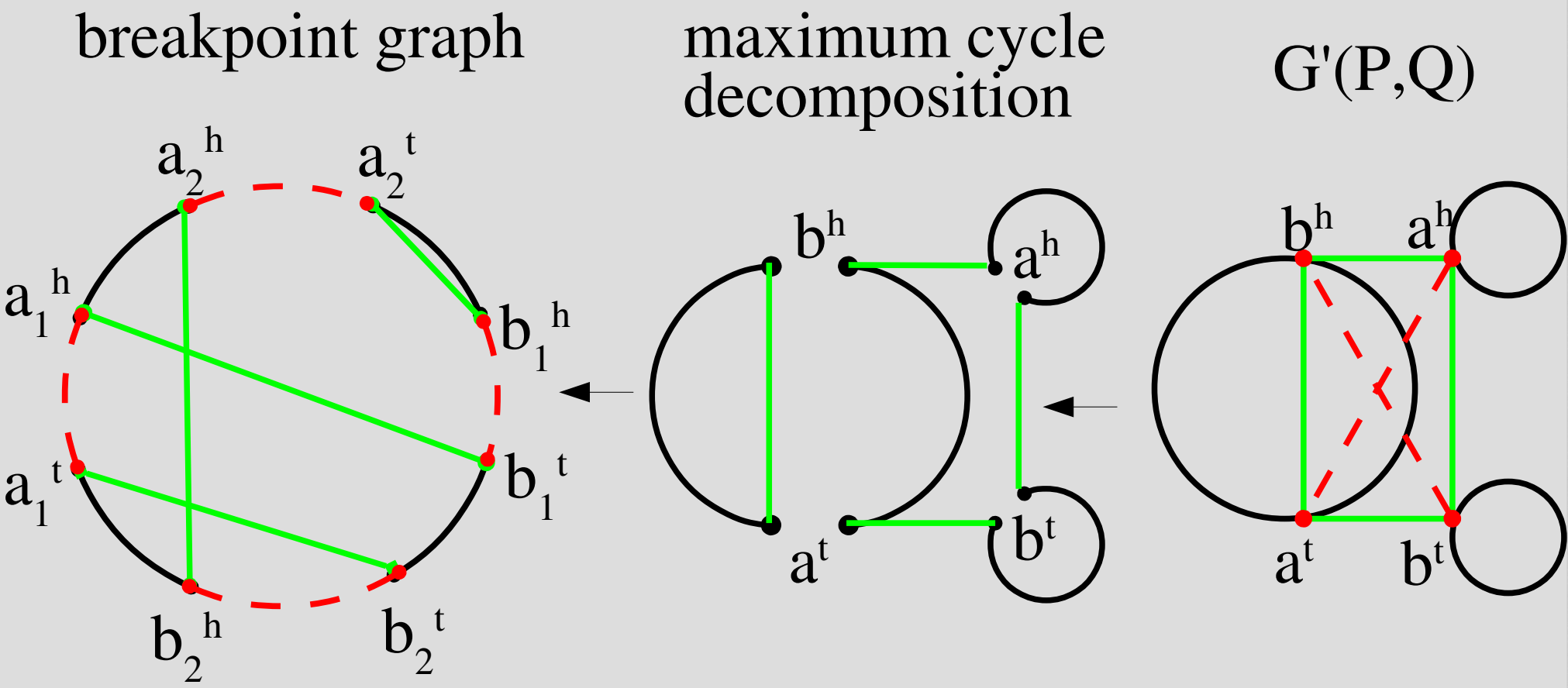


# Maximum Cycle Decomposition Problem

**Open Problem 3**: Given a contracted breakpoint graph, find its maximum black-green cycle decomposition.



# Computing Genomic Distance

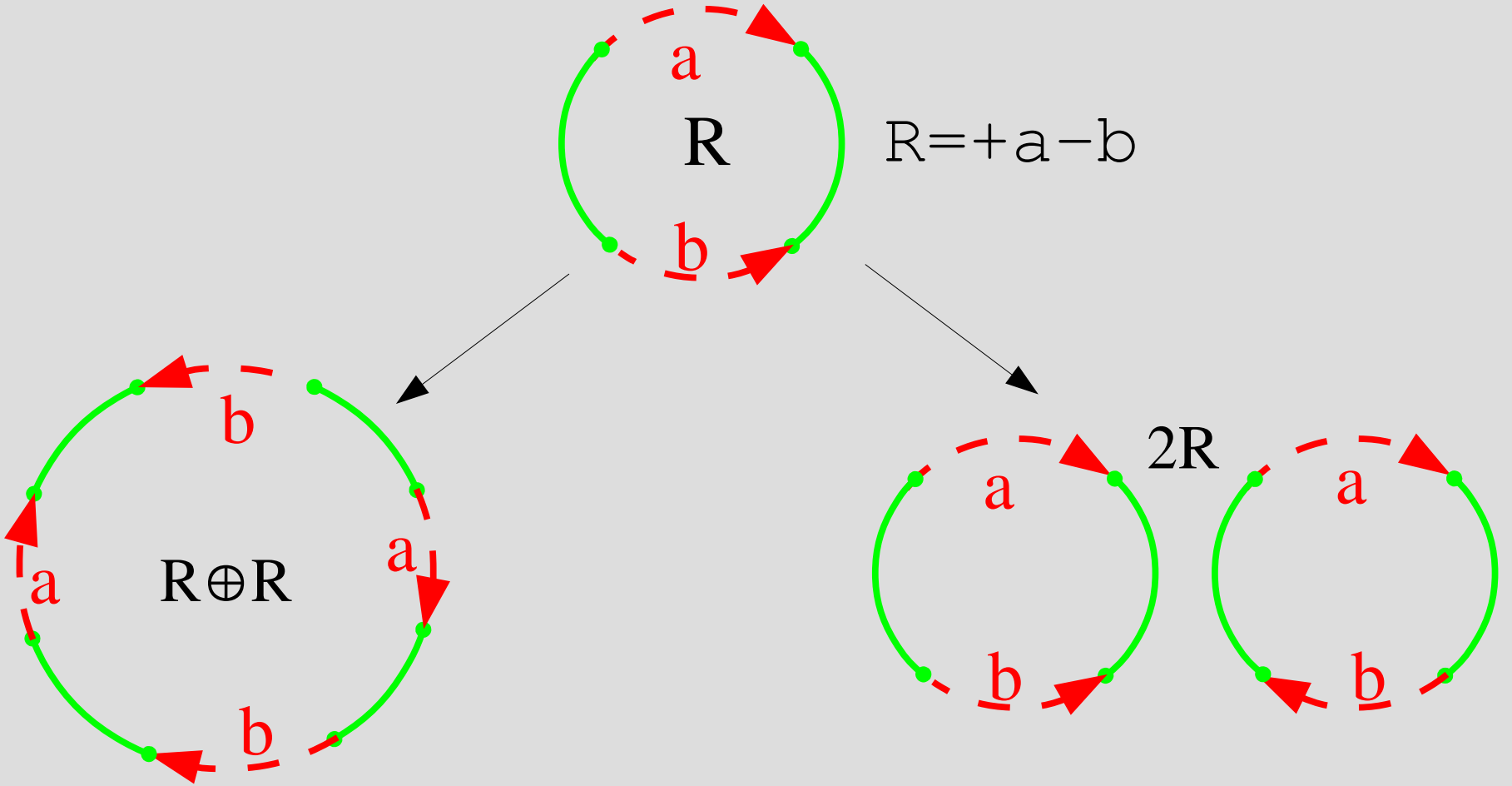


1. Distance between Unichromosomal Genomes
2. Distance between Multichomosomal Genomes
3. Breakpoint Graphs for Duplicated Genomes

**4. Whole Genome Duplication and Genome Halving Problem**

5. Genome Halving Problem for Multichromosomal Genomes
6. A Flaw in El-Mabrouk – Sankoff "Theorem"
7. Classification of Unichromosomal Circular Genomes

# Chromosome Duplication: R+R vs. 2R



# Genome Halving Problem

- WGD results in a perfect duplicated genome Q where each chromosome is of the form R⊕R or 2R.
  A unichromosomal perfect duplicated genome Q=R⊕R.

- Genome Q becomes subject to rearrangements that shuffle genes in Q and result in some rearranged duplicated genome P.

- Problem: reconstruct a perfect duplicated genome Q from a given rearranged duplicated genome *P*.

- Genome Halving Problem: Given a duplicated genome *P*, find a perfect duplicated genome Q minimizing the (reversal or genomic) distance from Q to P.

# Illustration

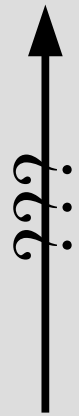
+a −d +e −c +b

+a −d +e −c +b +a −d +e −c +b
+a −d +d −a −b +c −e +e −c +b
+a −d +d +c −e +e −c +b +a +b
+a −d +d −e +e −c −c +b +a +b

+a −d +e −d +e −c −c +b +a +b

# Recover it!

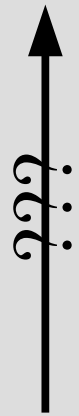
?? ?? ?? ?? ??



+a −d +e −d +e −c −c +b +a +b

# Even worse!

?? ?? ?? ?? ??



+a −d +e −d +e −c −c +b +a +b

# What if...

+a +c +d +e +b

+a +c +d +e +b +a +c +d +e +b

+a −d +e −d +e −c −c +b +a +b

# Too far!

+a +c +d +e +b

+a +c +d +e +b +a +c +d +e +b
+a −e −d −c −a −b −e −d −c +b
+a −e −d −c +c +d +e +b +a +b
+a −e −d −e −d −c +c +b +a +b
+a +d +e +d +e −c +c +b +a +b
+a +d +e +d +e −c −c +b +a +b
+a +d +e −d +e −c −c +b +a +b
+a −d +e −d +e −c −c +b +a +b

# Previous Results

- A series of papers in 1998-2002 by Nadia El-Mabrouk and David Sankoff culminating in solving the Genome Halving Problem in "*The Reconstruction Of Doubled Genome" (SIAM J. of Computing, 2003)*
- The algorithm is rather complicated
- The main "theorem" is incorrect for some genomes

# Previous Results

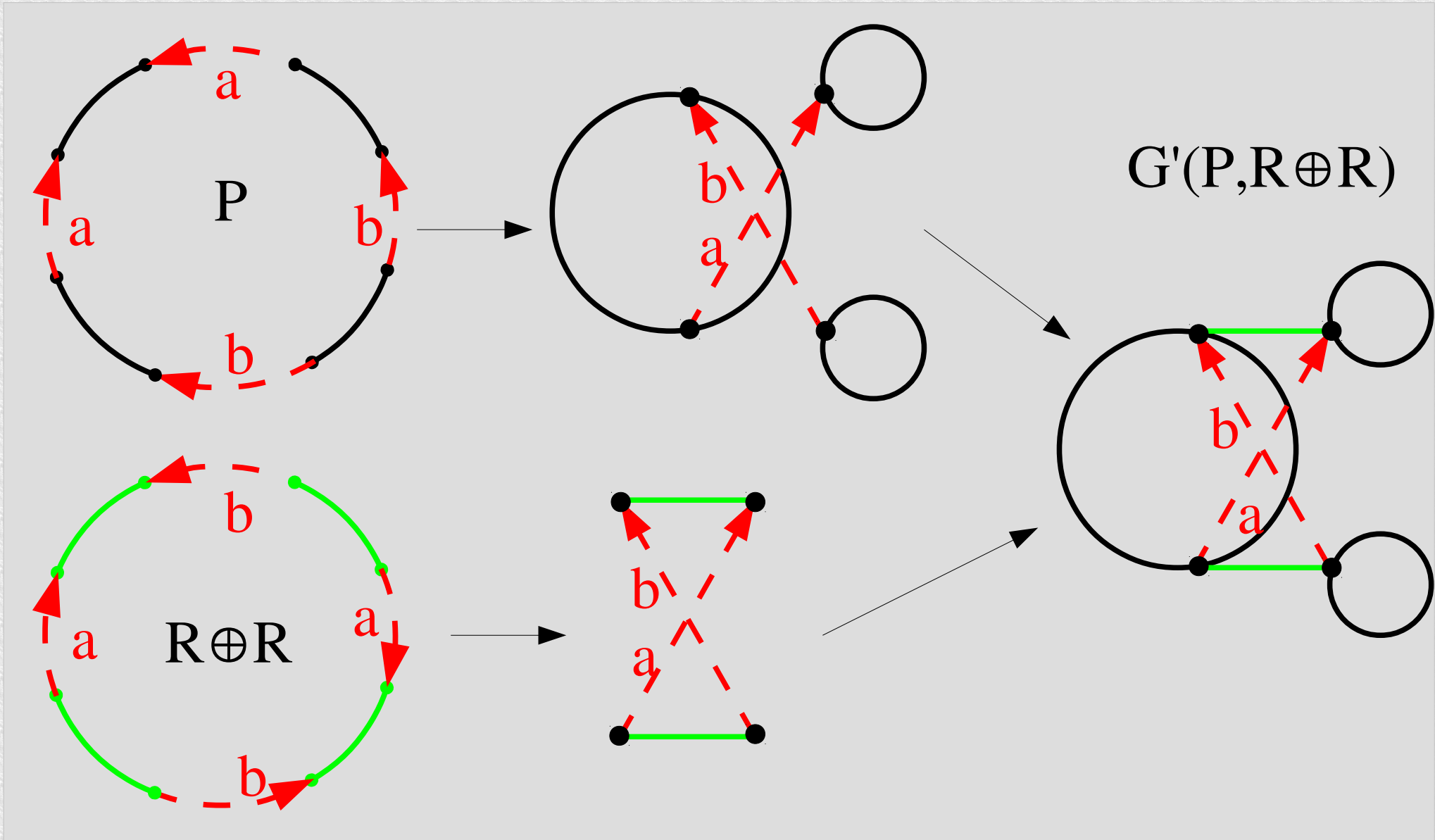
- Recall that $d(P,Q) = |P| - c(P,Q) + h(P,Q)$

- El-Mabrouk and Sankoff showed that for a given genome P, minimizing the reversal distance d(P,Q) over unichromosomal perfect duplicated genomes $Q=R \oplus R$ can be done in a consecutive manner:

  First, find a perfect duplicated genome Q maximizing c(P,Q)

  Second, minimize the term h(P,Q)

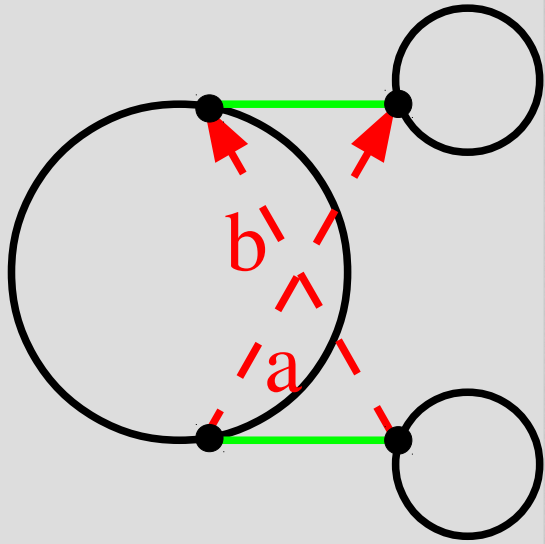
# G'(P,Q) when Q is a Perfect Duplicated Genome



# G'(P,Q) when Q is a Perfect Duplicated Genome

• Green edges form pairs of parallel edges, called **double green** edges

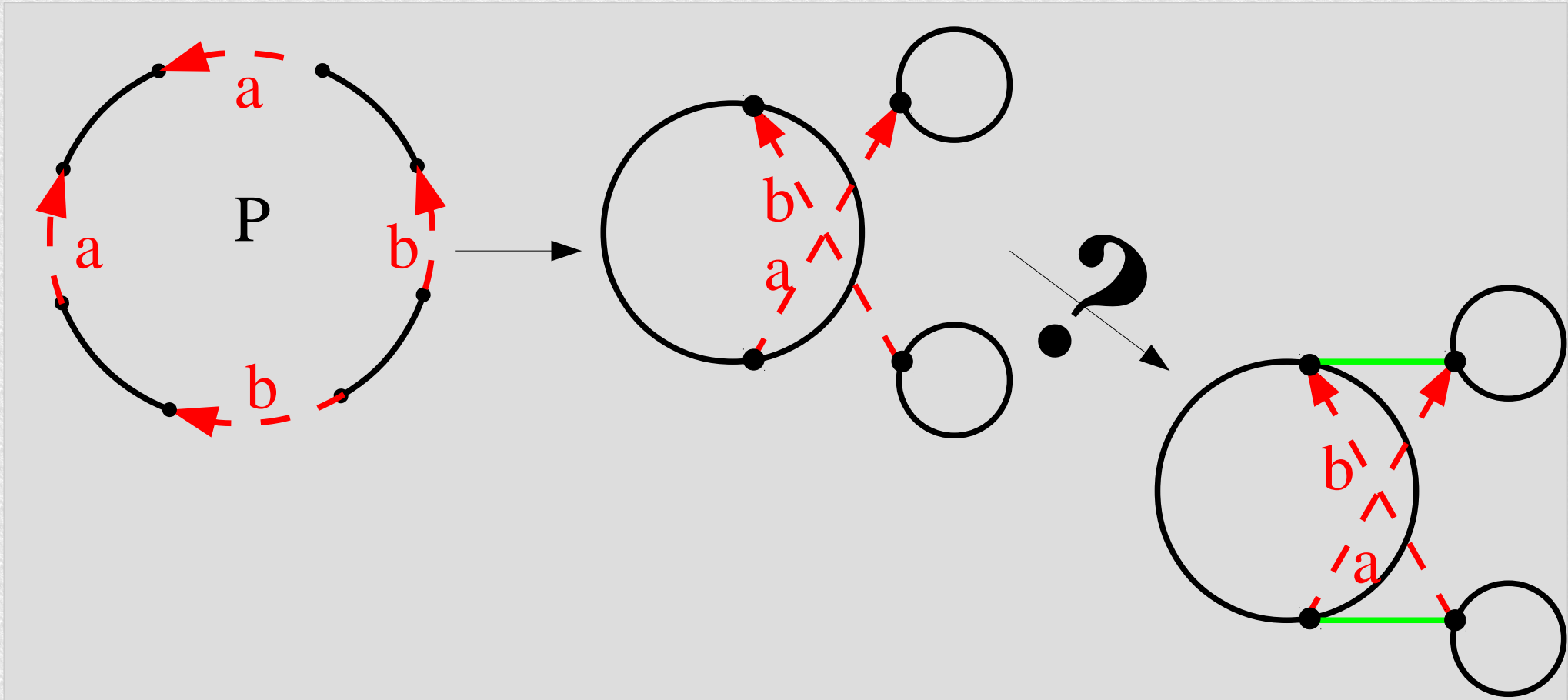
G'(P,R⊕R)



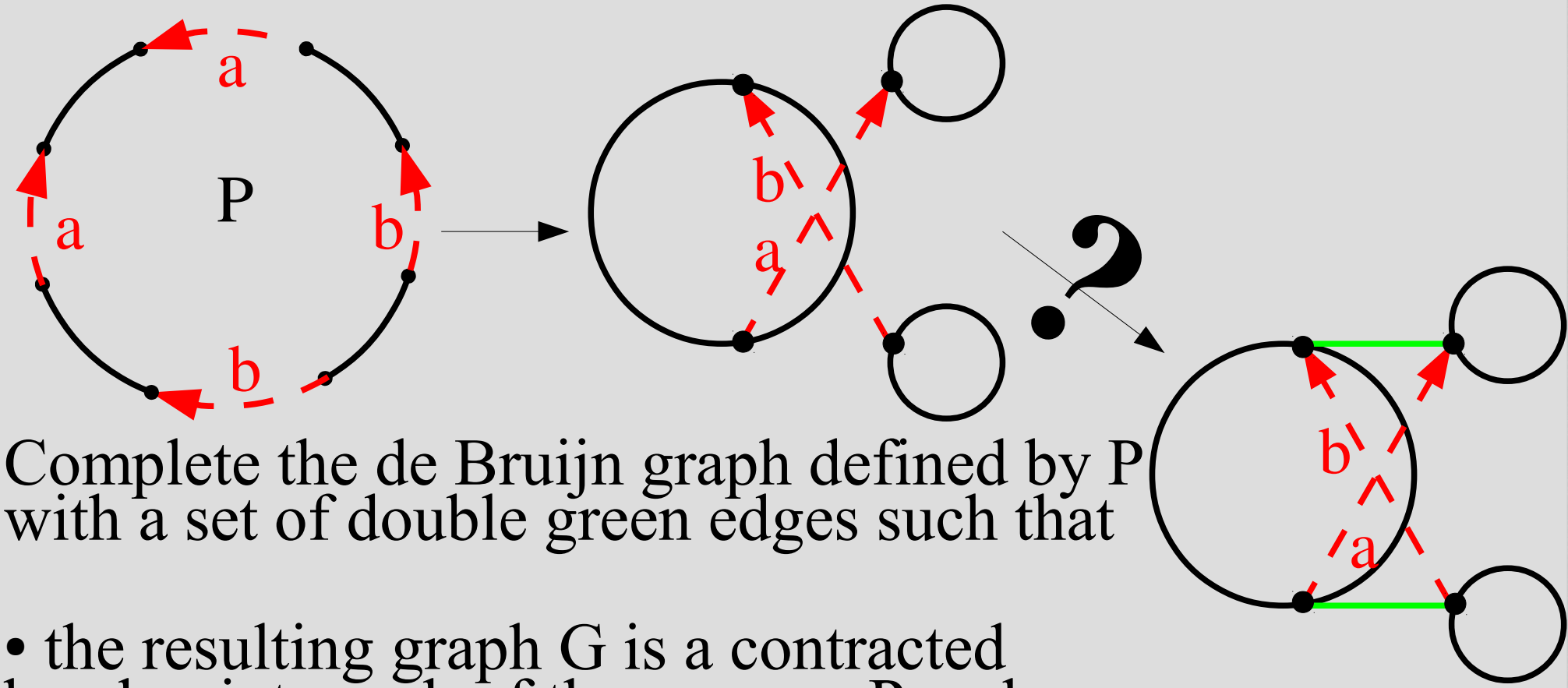

• **black edges** form *black cycles*
• **green edges** form a *matching*
• **red edges** form a *matching*

• For *unichromosomal genomes*, the contracted breakpoint graph has a *single* ***black-red cycle*** and a *single* ***green-red cycle***

# Genome Halving Problem



# Genome Halving Problem




Complete the de Bruijn graph defined by P with a set of double green edges such that

- the resulting graph G is a contracted breakpoint graph of the genome P and some perfect duplicated Q
- $c^{max}(G)$ is maximum

# Solving Maximum Cycle Decomposition Problem

**Theorem**. For a given duplicated genome P and any perfect duplicated genome Q, the number of black-green cycles

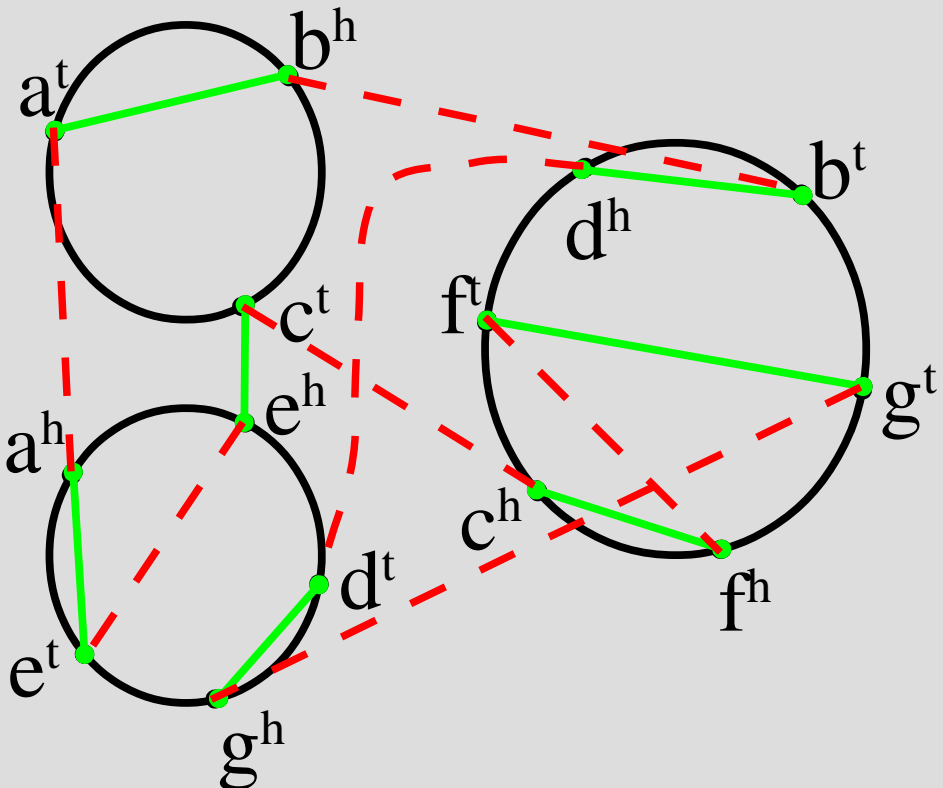
$c(G(P,Q)) \leq |P|/2 + \text{EvenBlackCycles}(P)$

# Solving Maximum Cycle Decomposition Problem

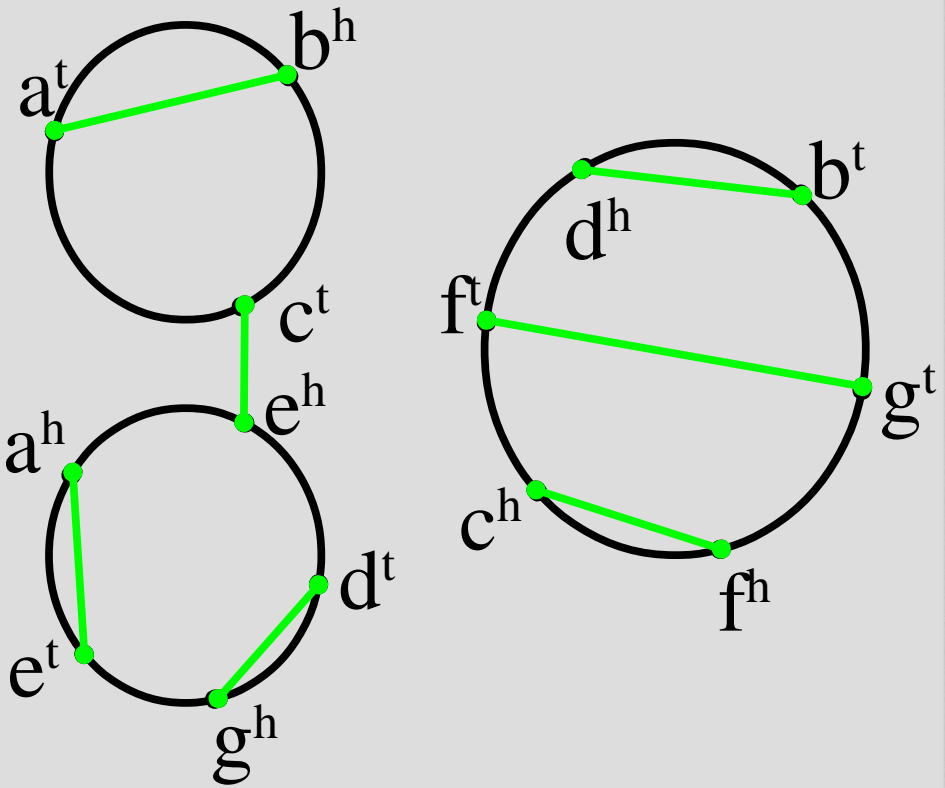
**Theorem**. For non-crossing graphs,

$$c(G) = |P|/2 + \text{EvenBlackCycles}(P)$$

G'(P,R⊕R)




*black-green subgraph*

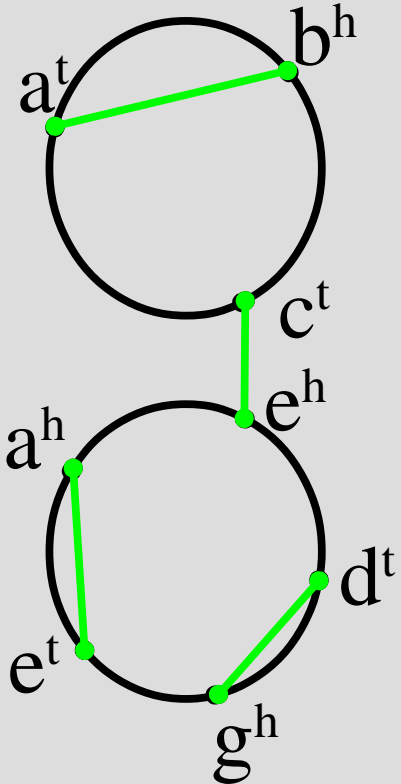


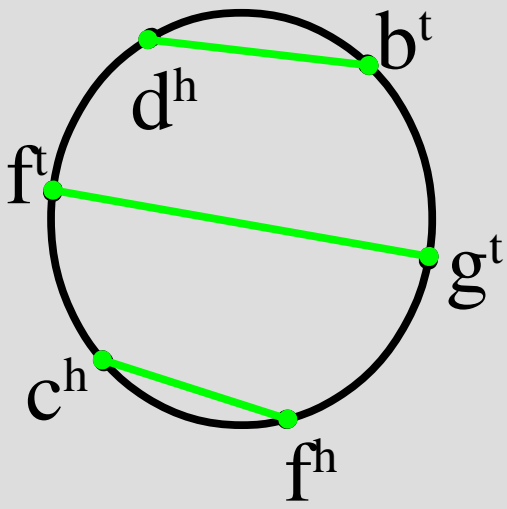
# Solving Maximum Cycle Decomposition Problem

**Theorem**. For non-crossing graphs,

$$c(G) = |P|/2 + \text{EvenBlackCycles}(P)$$

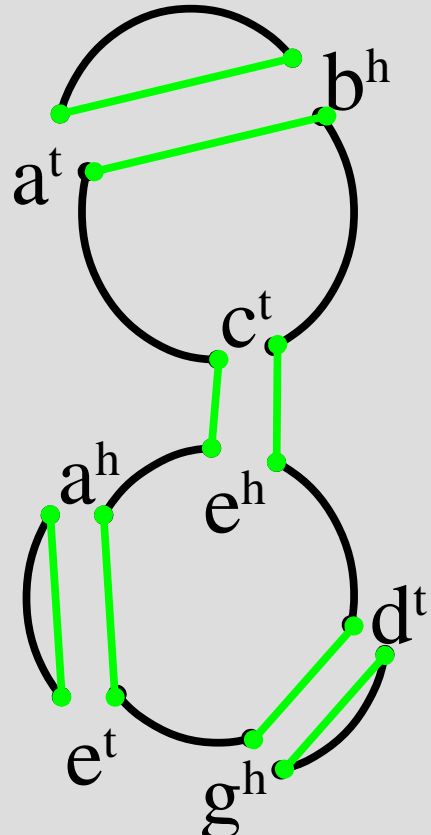
*black-green subgraph*

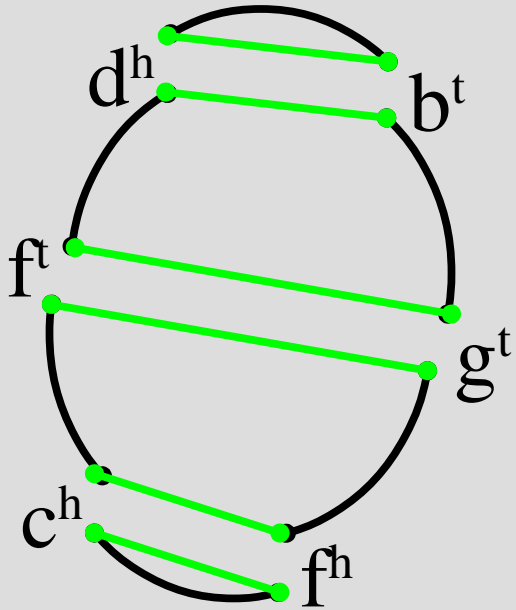



c(G)=8

*cycle decomposition*



# Labelling Problem

**Labelling Problem**. For a given cycle decomposition C of the contracted breakpoint graph G'(P,Q), find a labelling of P and Q that induces C.

- Solution to the Labelling Problem for a maximum cycle decomposition of G'(P,Q), where Q is a perfect duplicated genome, will give a solution to the Genome Halving Problem.

- The Labelling Problem is simpler for multichromosomal genomes than for unichromosomal genomes.

1. Distance between Unichromosomal Genomes
2. Distance between Multichomosomal Genomes
3. Breakpoint Graphs for Duplicated Genomes
4. Whole Genome Duplication and Genome Halving Problem

**5. Genome Halving Problem for Multichromosomal Genomes**

6. A Flaw in El-Mabrouk – Sankoff "Theorem"
7. Classification of Unichromosomal Circular Genomes

# Genome Halving Problem

**Genome Halving Problem (for multichromosomal genomes)**. Given a duplicated genome P, find a perfect duplicated genome Q minimizing the genomic distance $d_2(P,Q)$.

- The genomic distance between *labeled* genomes P and Q is $d_2(P,Q) = |P| - c(G(P,Q))$.

- Minimizing $d_2(P,Q)$ is equivalent to finding a perfect duplicated genome Q and a labelling of genomes P and Q that maximize $c(G(P,Q))$.

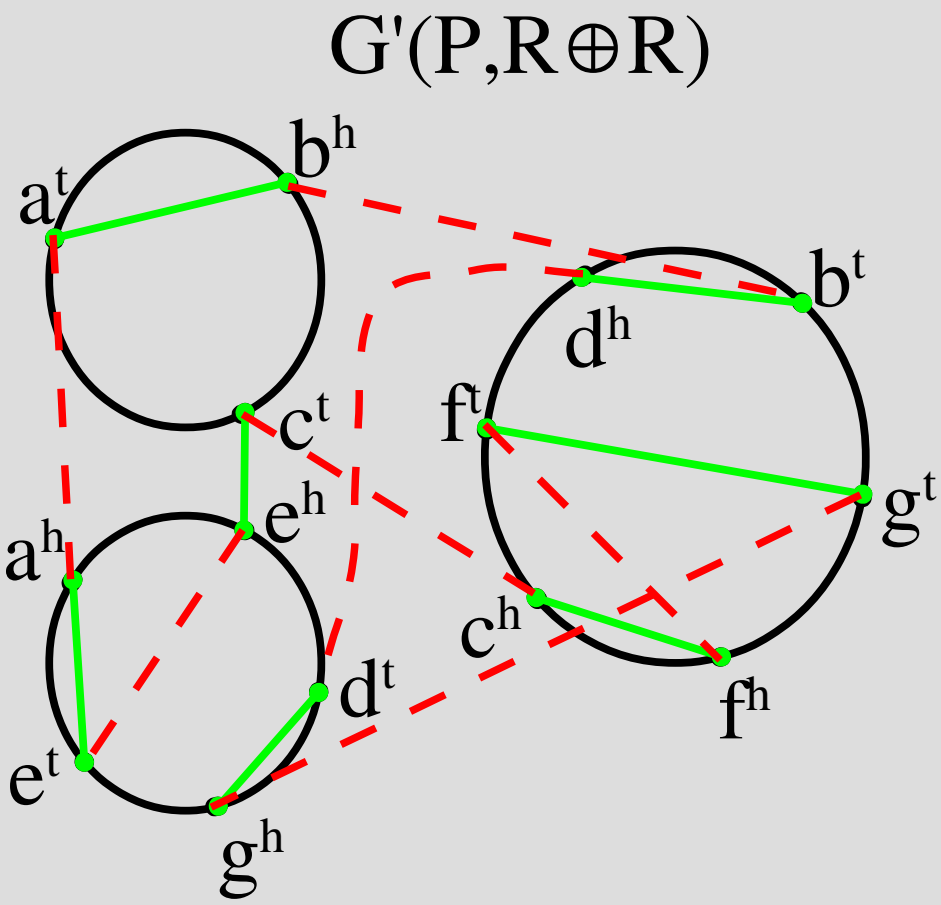
# Genome Halving Problem

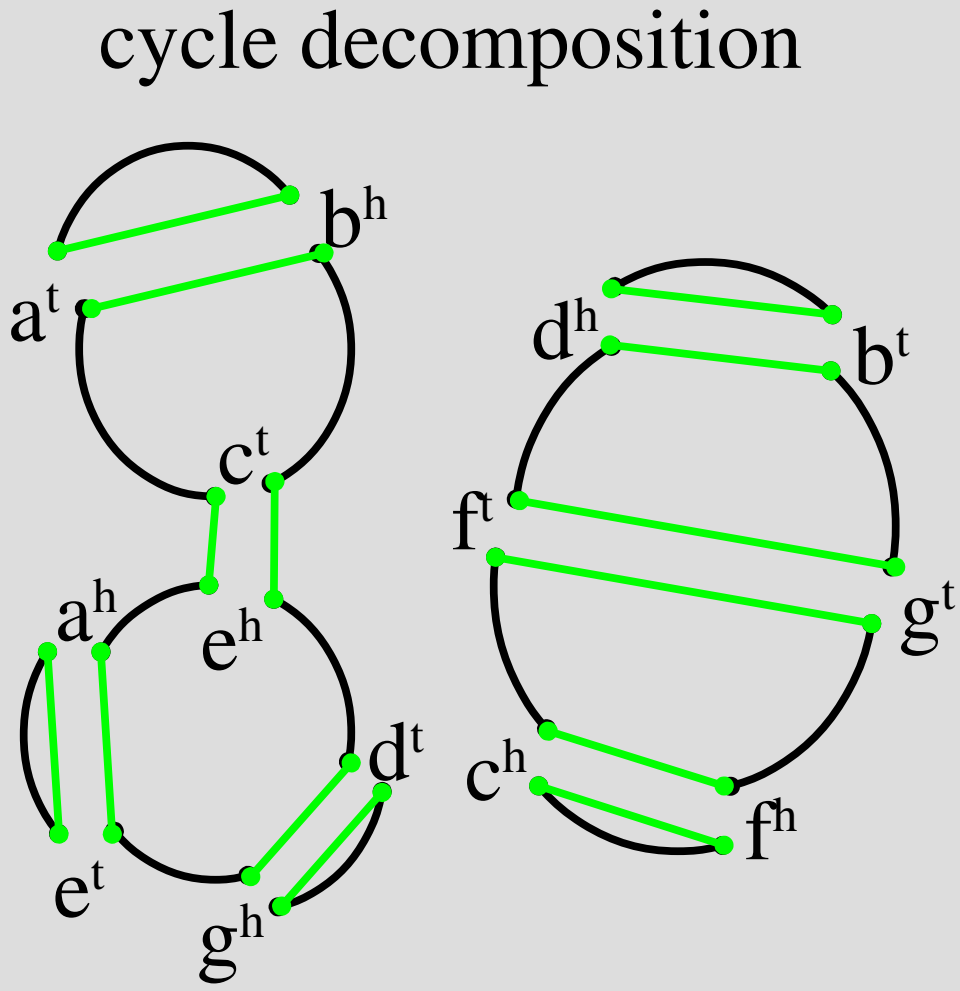
- Minimizing $d_2(P,Q)$ is equivalent to finding a perfect duplicated genome Q and a labelling of genomes P and Q that maximize c(G(P,Q)).

- We know how to solve the Maximum Cycle Decomposition Problem, i.e., to find a perfect duplicated genome Q maximizing $c^{max}(G'(P,Q))$.

- To obtain an optimal labelling of P and Q we need to solve the Labelling Problem.

# Maximum Cycle Decomposition



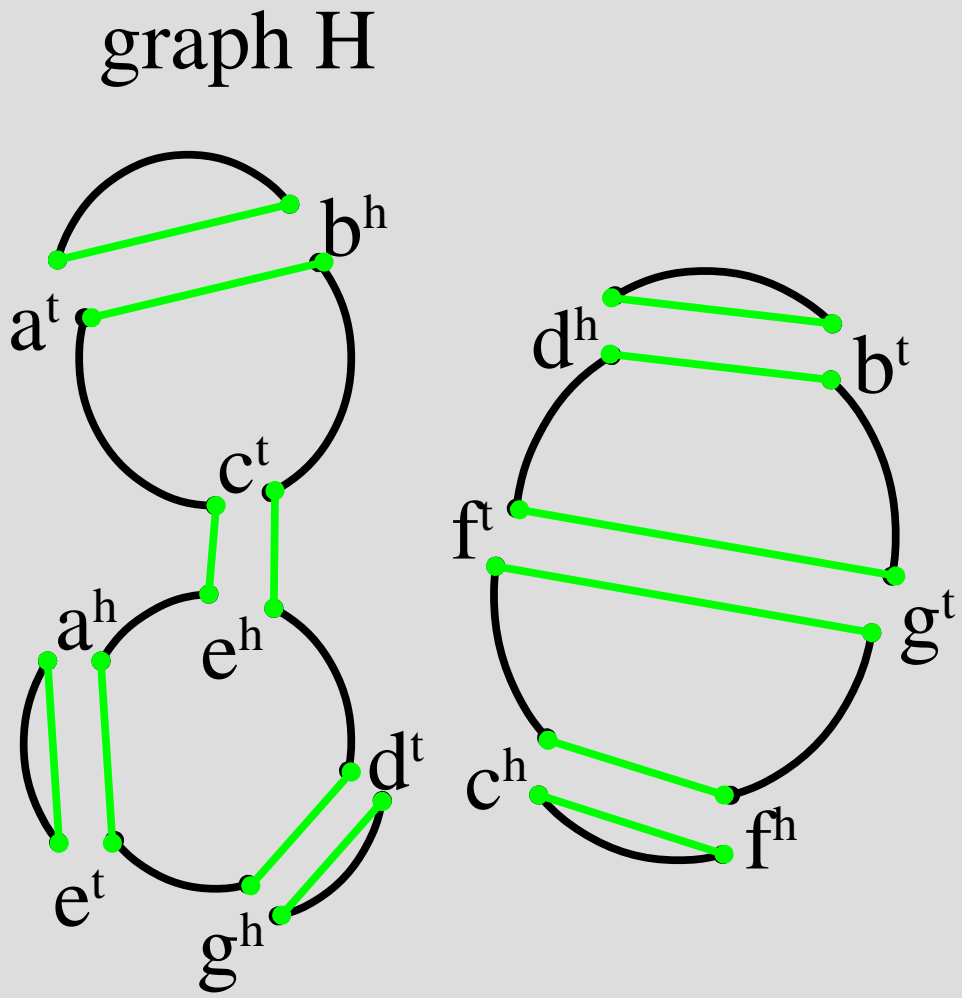




R= (+a +e +c -f +g +d + b)

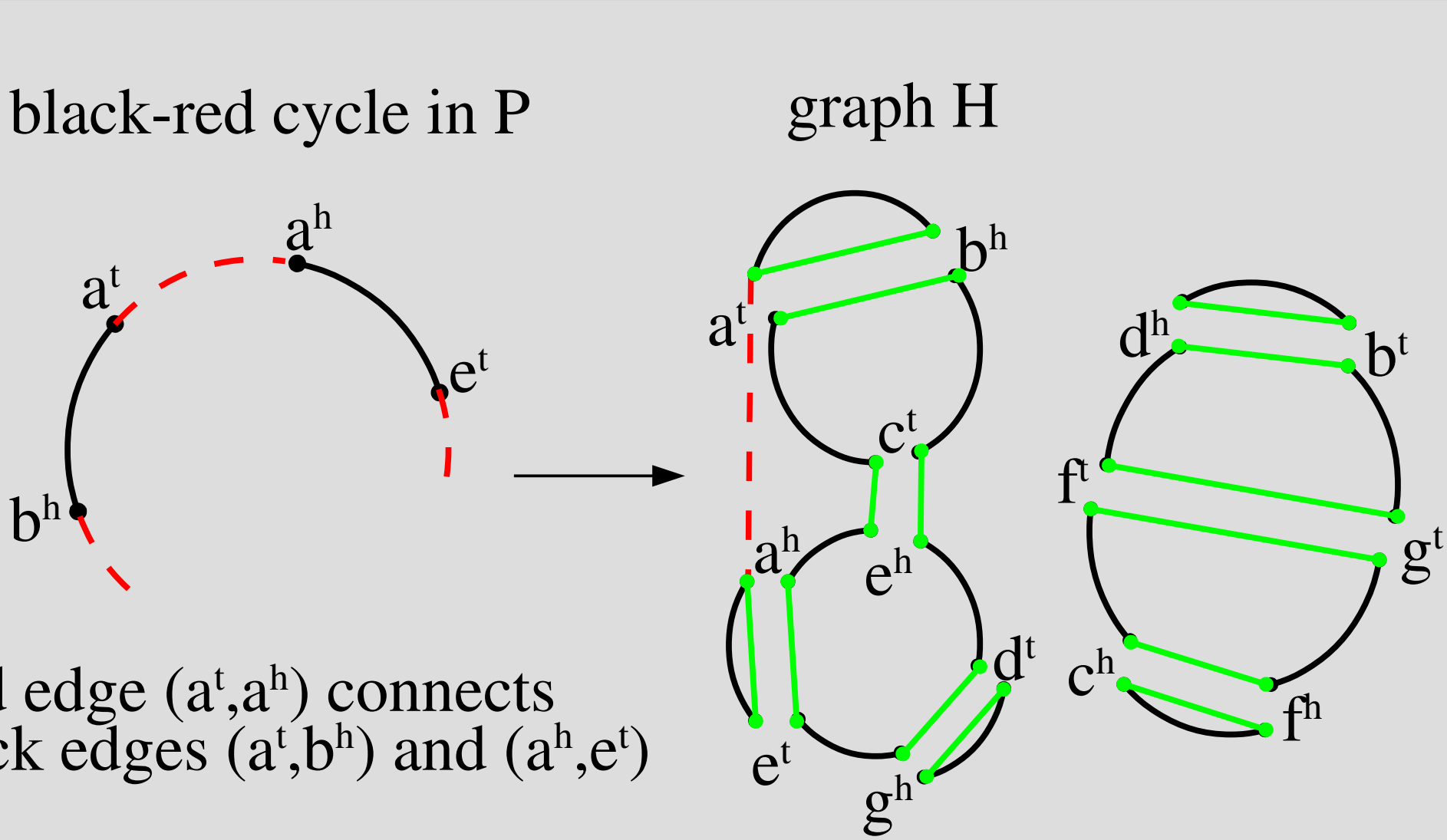
# Labelling Problem

Goal: complete the graph H with a set of red edges such that black-red cycles would represent the genome P.

Then green-red cycles will give the labelling of genome Q.



# Solving Labelling Problem: Genome P Imposes Red Edges in H



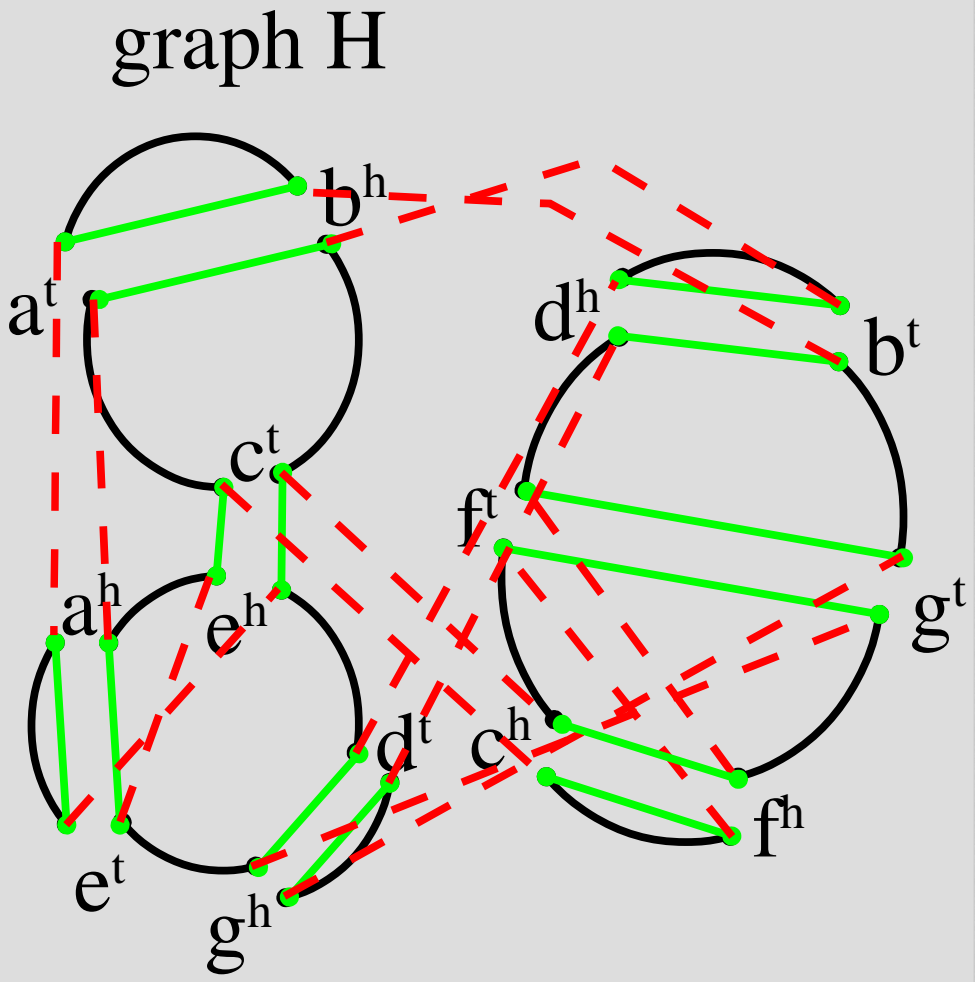

Red edge $(a^t,a^h)$ connects
black edges $(a^t,b^h)$ and $(a^h,e^t)$

# Solving Labelling Problem

Goal achived: the graph H is completed with a set of red edges such that black-red cycles represent the genome P.

green-red cycles define the labelling of genome Q.

**Did we solve the Labelling Problem ?**



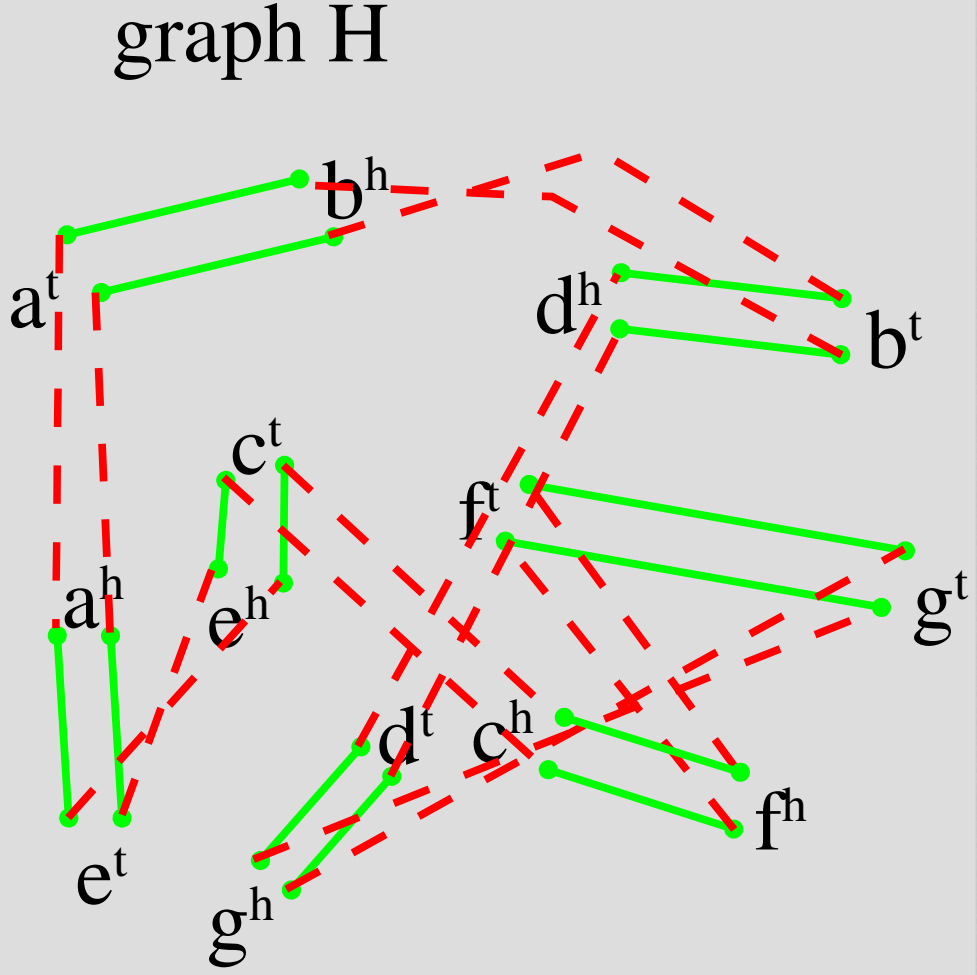
# Solving Labelling Problem

Goal achived: the graph H is completed with a set of red edges such that black-red cycles represent the genome P.

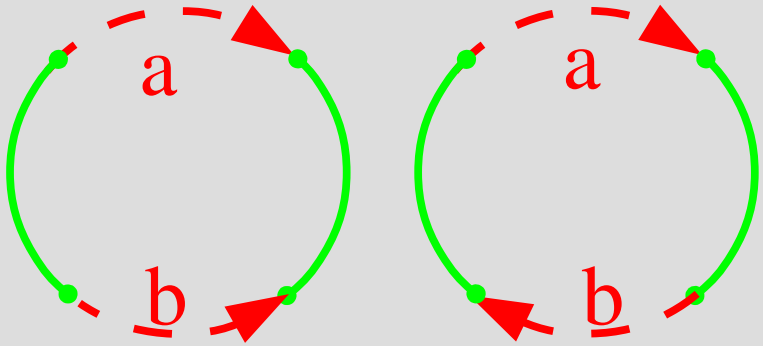
green-red cycles define the labelling of genome Q.

**Did we solve the Labelling Problem ?**

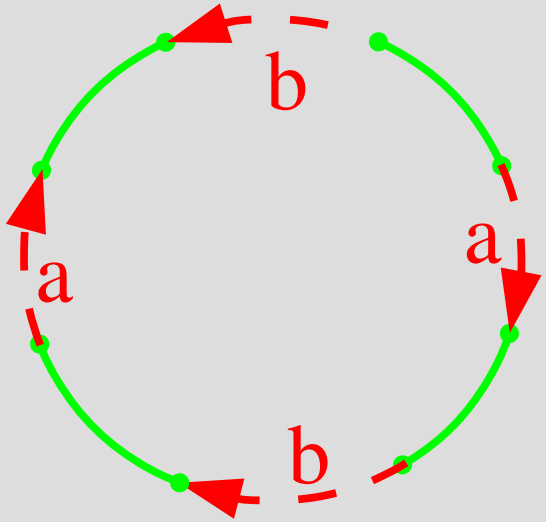
**Not really!** We have got **Q=2R** while we started with G(P,R+R)



# Equivalent Genomes

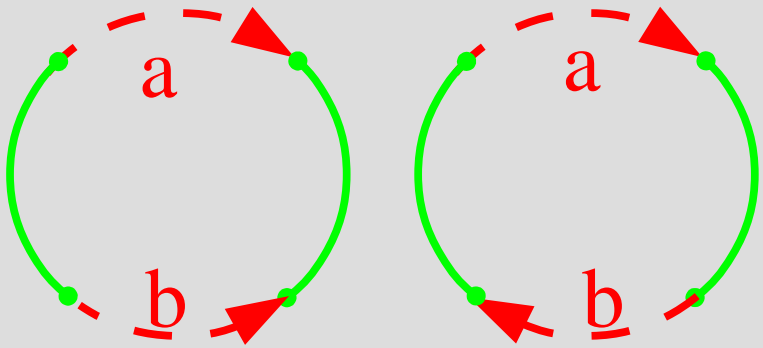



$Q_2 = (+a-b)\ (+a+b)$



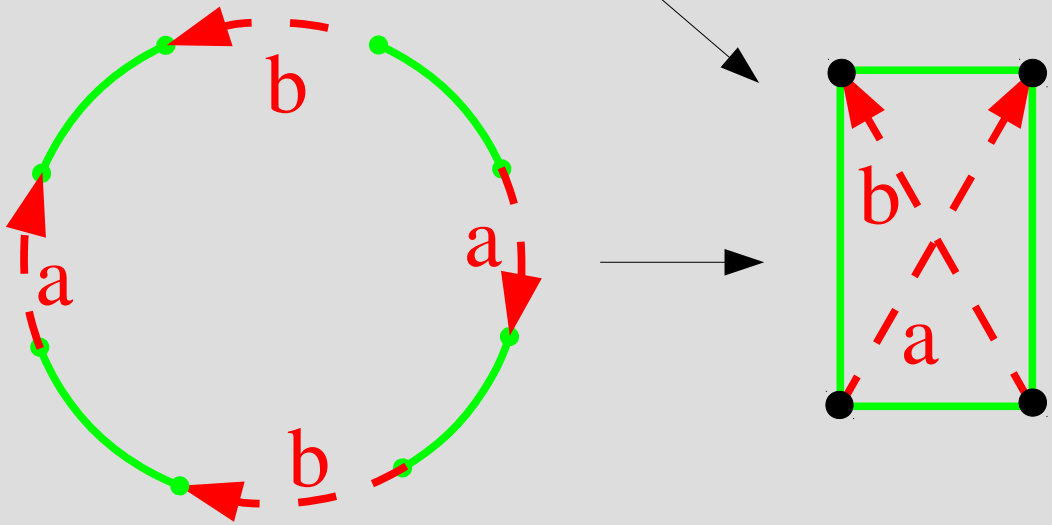

$Q_1 = +a-b+a+b$

# Equivalent Genomes




Two genomes are *equivalent* if their de Bruijn graphs coincide.

If $Q_1 \sim Q_2$, then $G'(P,Q_1)=G'(P,Q_2)$ for any genome P.



# Labelling Problem for Multichromosomal Genomes

**Theorem.** Any black-green cycle decomposition of the contracted breakpoint graph G'(P,Q) is induced by some labeling of P and Q' where genome Q' is equivalent to Q.

1. Distance between Unichromosomal Genomes
2. Distance between Multichomosomal Genomes
3. Breakpoint Graphs for Duplicated Genomes
4. Whole Genome Duplication and Genome Halving Problem
5. Genome Halving Problem for Multichromosomal Genomes

**6. A Flaw in El-Mabrouk – Sankoff "Theorem"**

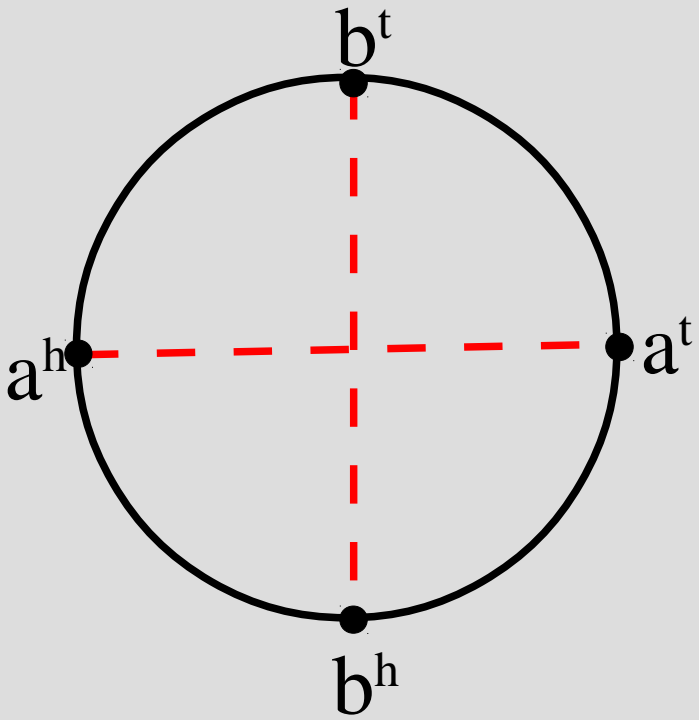
7. Classification of Unichromosomal Circular Genomes

# El-Mabrouk – Sankoff results overview

- They consider multichromosomal and circular unichromosomal genomes.
- Proof of the upper bound: $c(G) \leq U$ (some formula)
- Algorithm (rather complicated) that for a given genome P finds a *multichromosomal* genome Q such that $c(G(P,Q))=U$.
- Unreasonable claim that the same algorithm applies to circular genomes (that is **incorrect!**). There exists such genome P that $c(G(P,R \oplus R))<U$ for any R.

# Counterexample

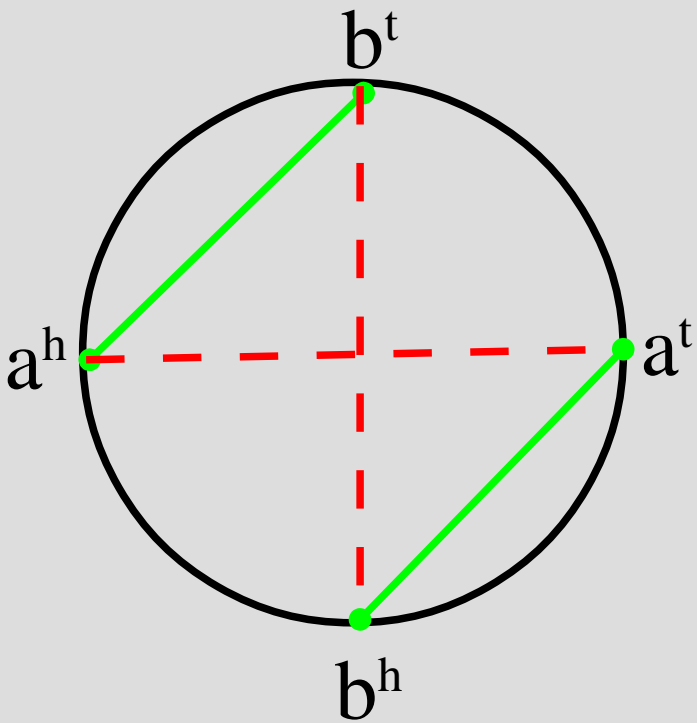
Let's halve genome P = +a +b -a -b

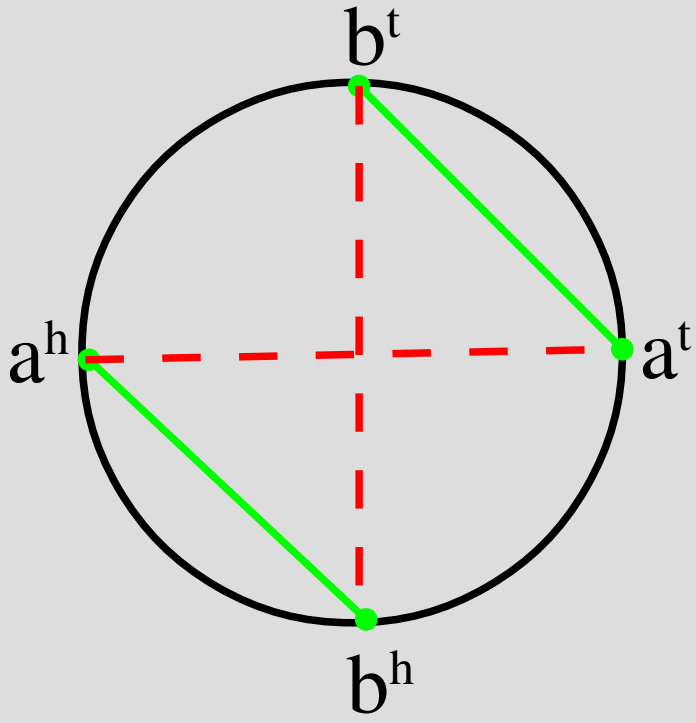



El-Mabrouk – Sankoff
"Theorem" claims existence of a genome R with $c(G(P,R \oplus R))=1$.

# Counterexample

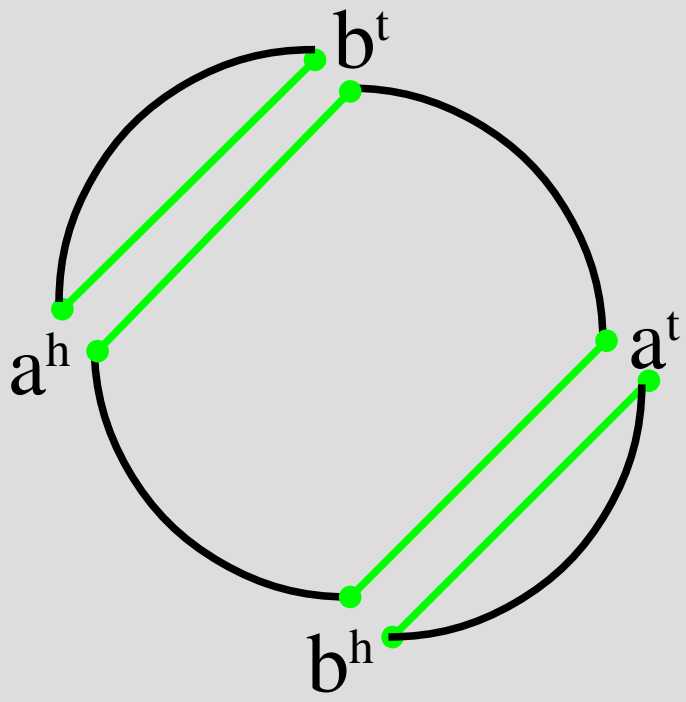
Let's halve genome P = +a +b -a -b
there are two potential contracted breakpoint graphs

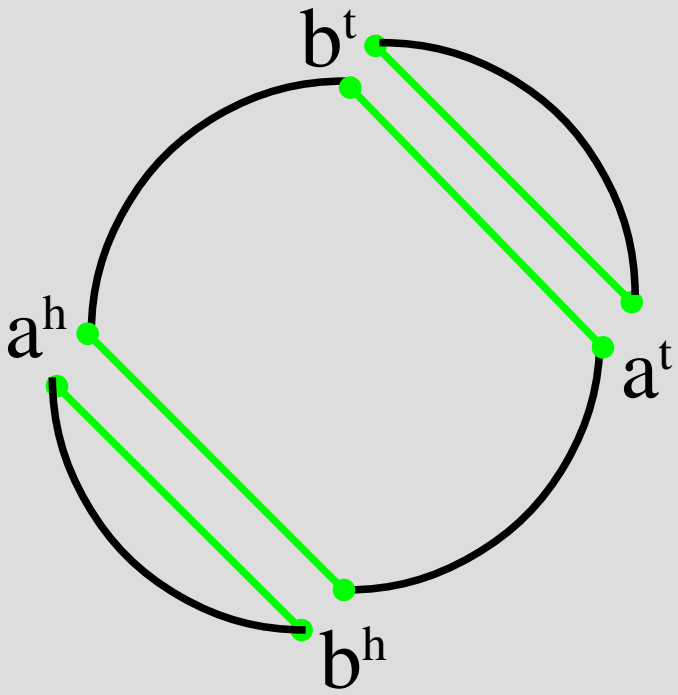


# Counterexample

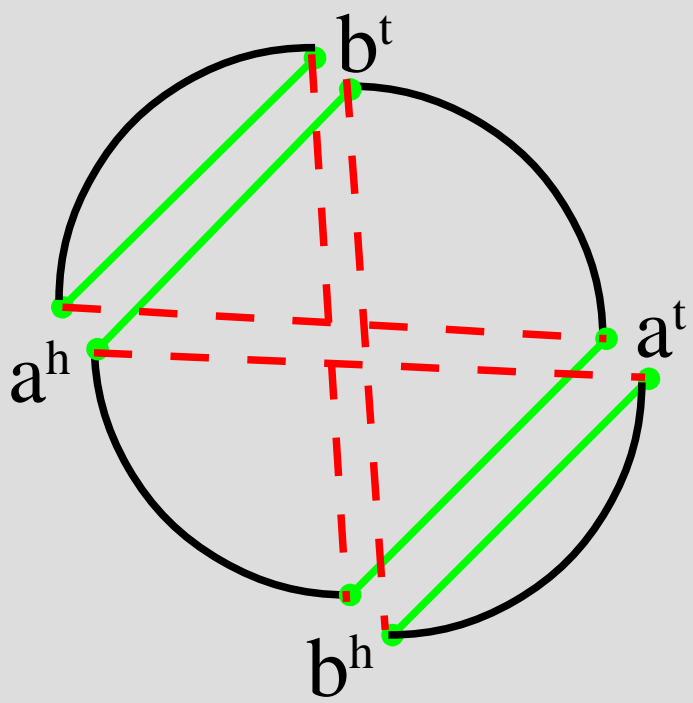
Let's halve genome P = +a +b -a -b
there are two potential contracted breakpoint graphs with maximum black-green cycle decompositions

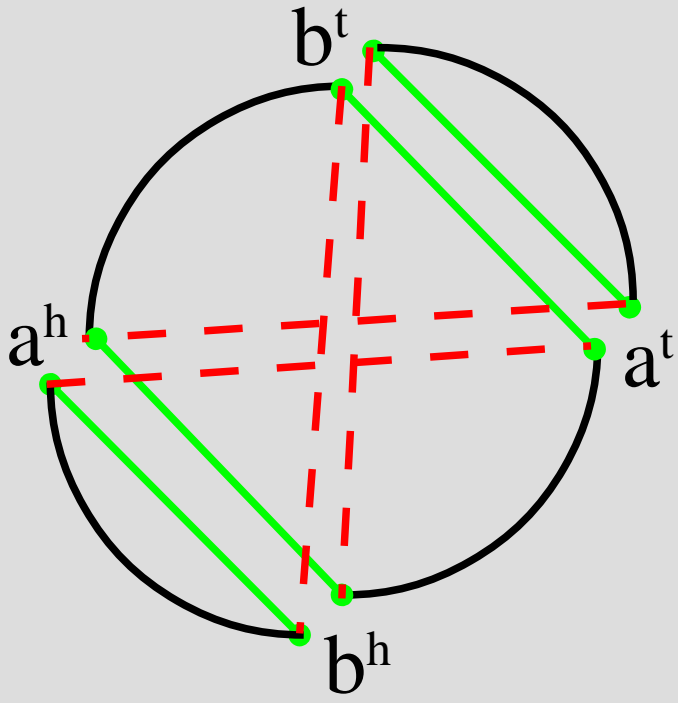


# Counterexample

Let's halve genome P = +a +b -a -b
there are two contracted breakpoint graphs
resulting in the breakpoins graphs

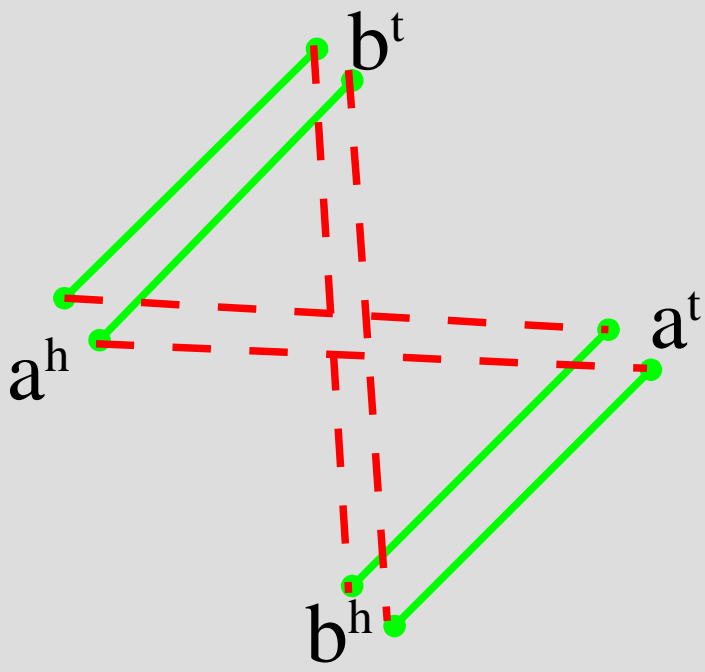


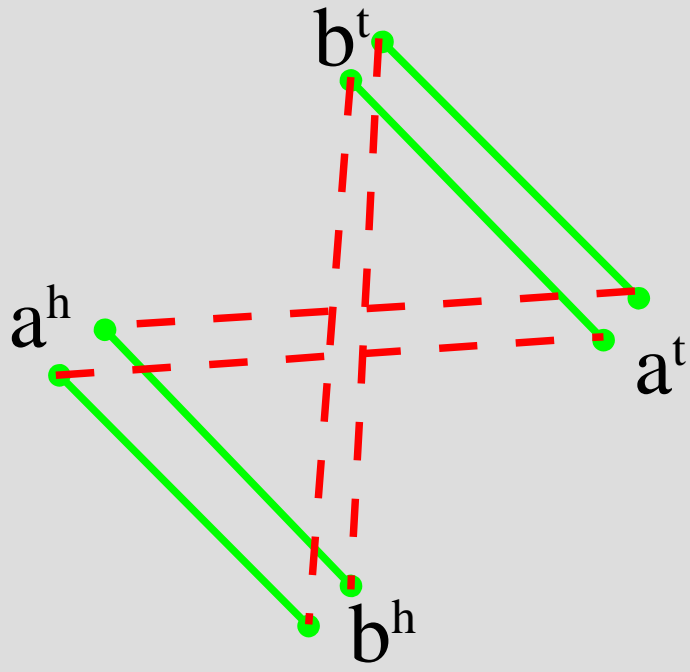
# Counterexample

Let's halve genome P = +a +b -a -b
there are two contracted breakpoint graphs
resulting in two-chromosomal genomes

2R = (+a+b)(+a+b)




2R=(+a-b)(+a-b)



# Labelling Problem for Unichromosomal Genomes

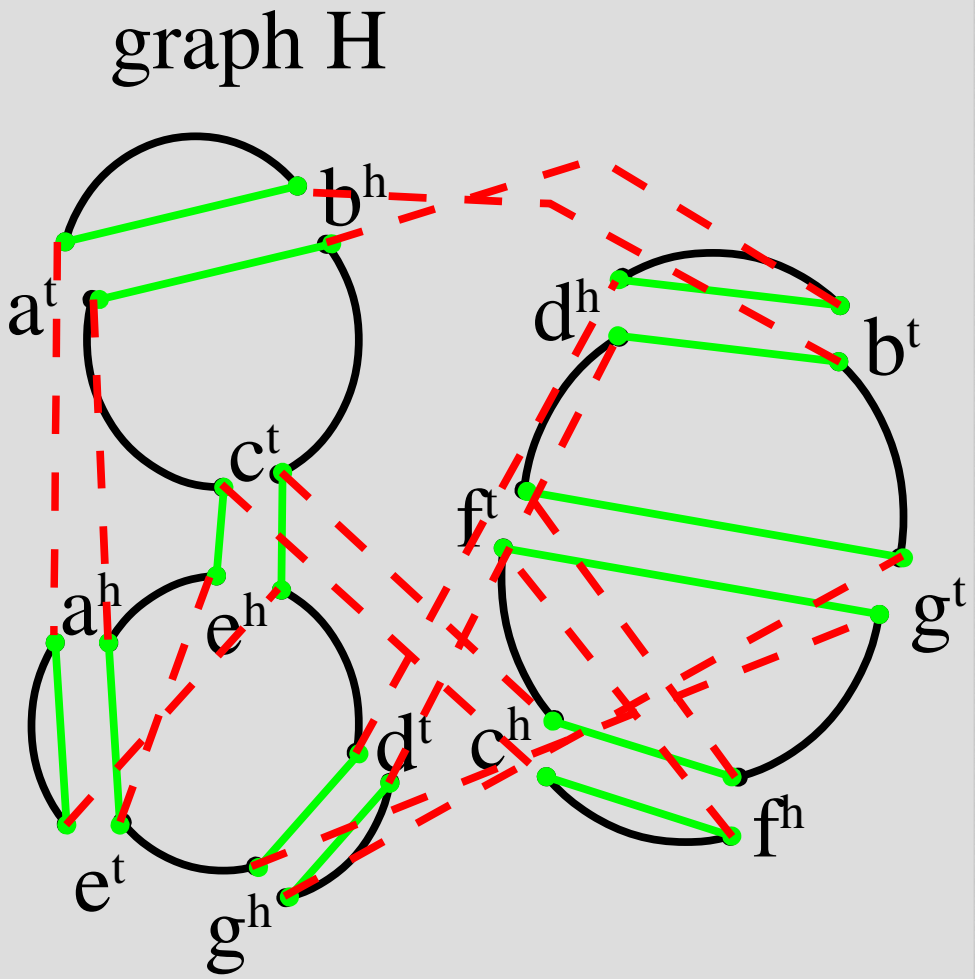
**Theorem (original version).** Any black-green cycle decomposition of the de Bruijn graph G'(P,R⊕R) is induced by some labeling of P and R⊕R.

*It is consistent with El-Mabrouk – Sankoff results but ...*

# Labelling Problem for Unichromosomal Genomes

- there is a single black-red cycle reading the labeled genome P
*(true by contruction)*

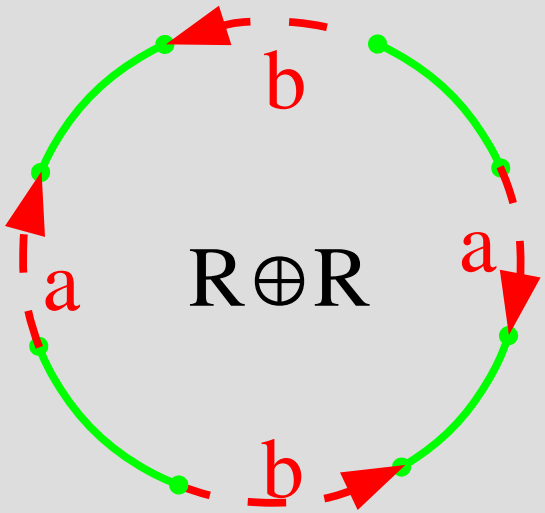
- there is a single green-red cycle reading genome R⊕R
(*may not hold*)



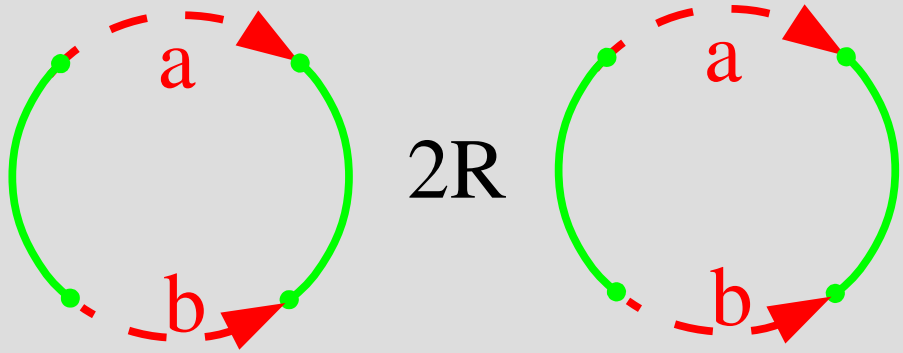
# Genomes R+R and 2R

**Theorem.** Genome R⊕R is equivalent to genome 2R. Moreover, 2R is the only genome equivalent to R⊕R.

$Q_1$=+a−b+a−b




$Q_2$=(+a−b) (+a−b)



# Labelling Problem for Unichromosomal Genomes

**Theorem (original version).** Any black-green cycle decomposition of the de Bruijn graph G'(P,R⊕R) is induced by some labeling of P and R⊕R.

**Theorem (CORRECTED).** Any black-green cycle decomposition of the de Bruijn graph G'(P,R⊕R) is induced by some labeling of P and either R⊕R or 2R.

# R+R vs. 2R

- Which of the following two cases takes place:
  $G=G(P,R \oplus R)$ or $G=G(P,2R)$ ?

- What if $G=G(P,2R)$ ?
  Can we find genome R' such that
  $c(G(P,R' \oplus R')) = c(G(P,2R))$ ?

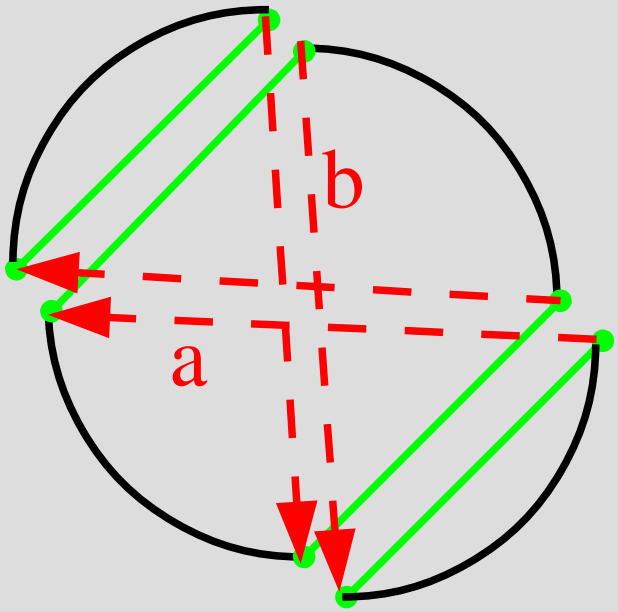
1. Distance between Unichromosomal Genomes
2. Distance between Multichomosomal Genomes
3. Breakpoint Graphs for Duplicated Genomes
4. Whole Genome Duplication and Genome Halving Problem
5. Genome Halving Problem for Multichromosomal Genomes
6. A Flaw in El-Mabrouk – Sankoff "Theorem"

**7. Classification of Unichromosomal Circular Genomes**

# G(P,2R) vs. G(P,R+R)

G(P,2R) contains 2 green-obvere cycles

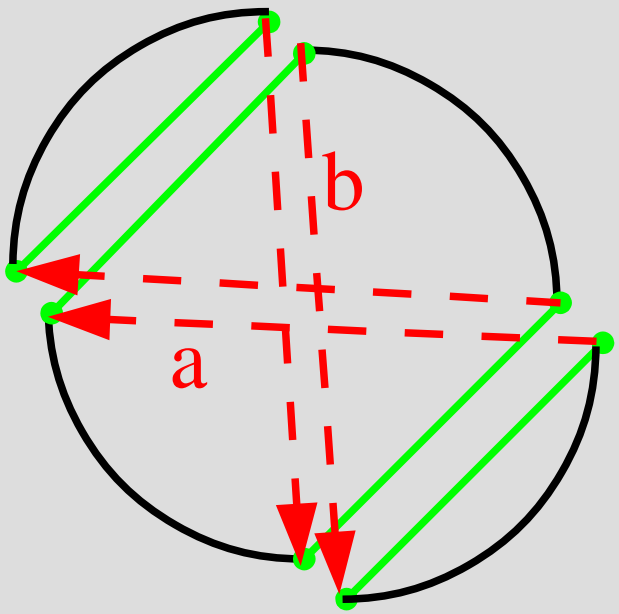
G(P,2R)



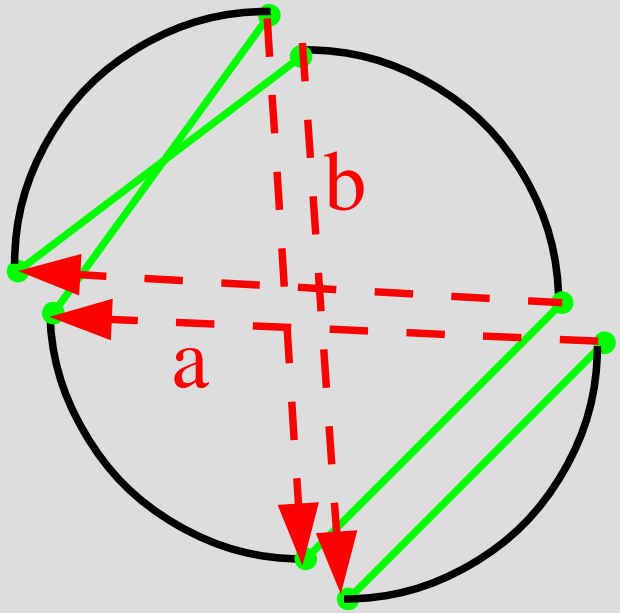
# G(P,2R) vs. G(P,R+R)

G(P,2R) contains two green-red cycles
*criss-cross a pair of counterpart green edges*

G(P,2R)




G(P,R⊕R)



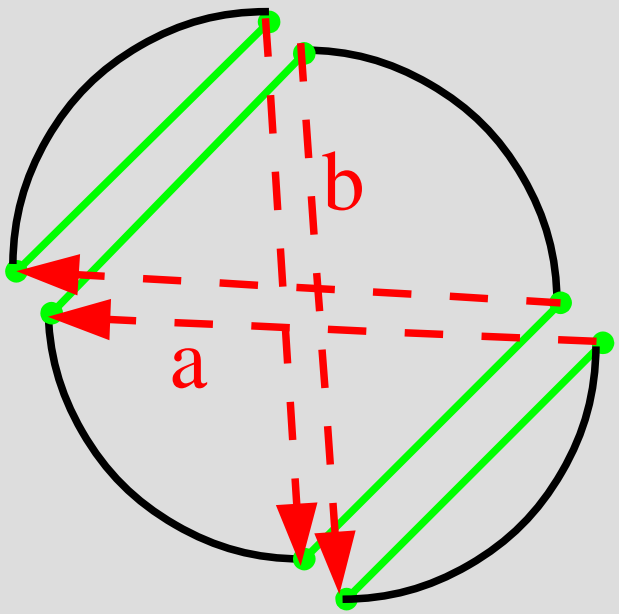
# G(P,2R) vs. G(P,R+R)

G(P,2R) contains two green-red cycles

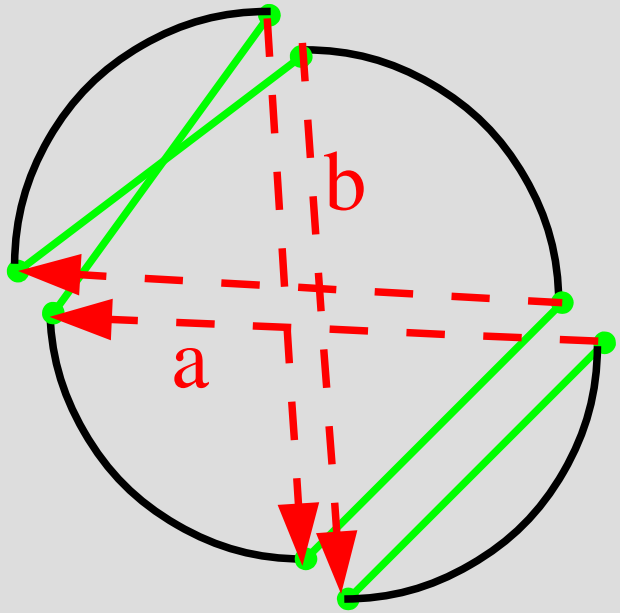
*criss-cross a pair of counterpart green edges*

G(P,R⊕R) contains a single green-red cycle

G(P,2R)




G(P,R⊕R)

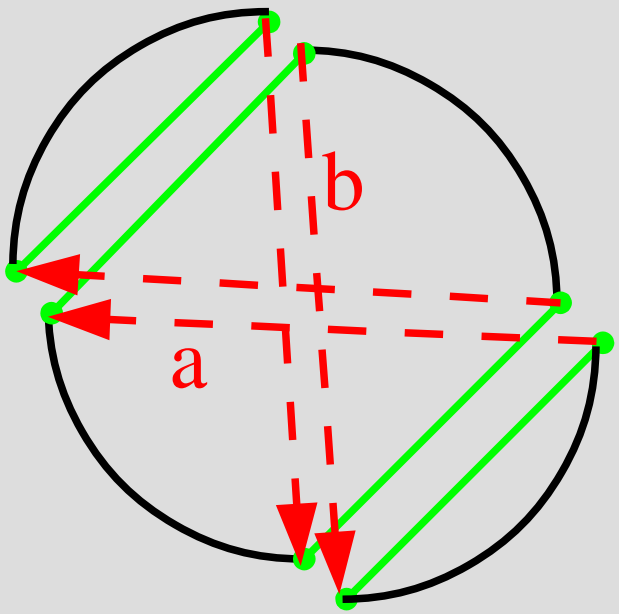


# G(P,2R) vs. G(P,R+R)

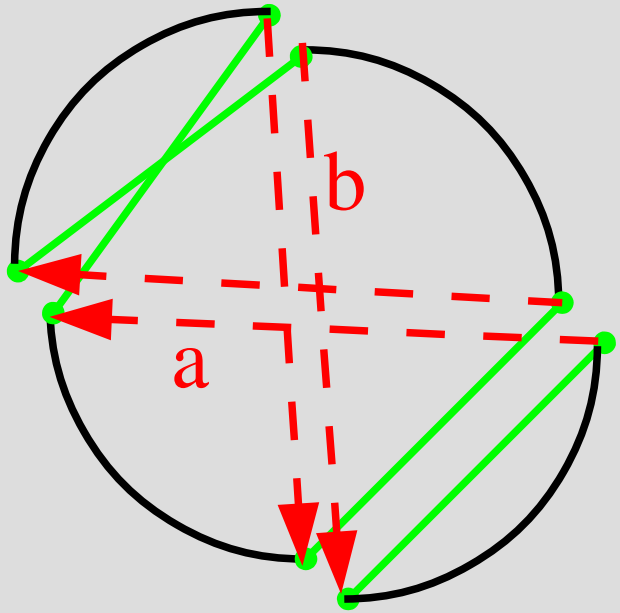
# of black-green cycles:

$$| c(G(P,2R)) - c(G(P,R \oplus R)) | \leq 1.$$

G(P,2R)




G(P,R⊕R)



# Non-Singular Genomes

- We call genome P *singular* if its de Bruijn graph contains only black cycles of *even* size

- **Theorem.** For a non-singular genome P, there exist genomes R' and R'' such that

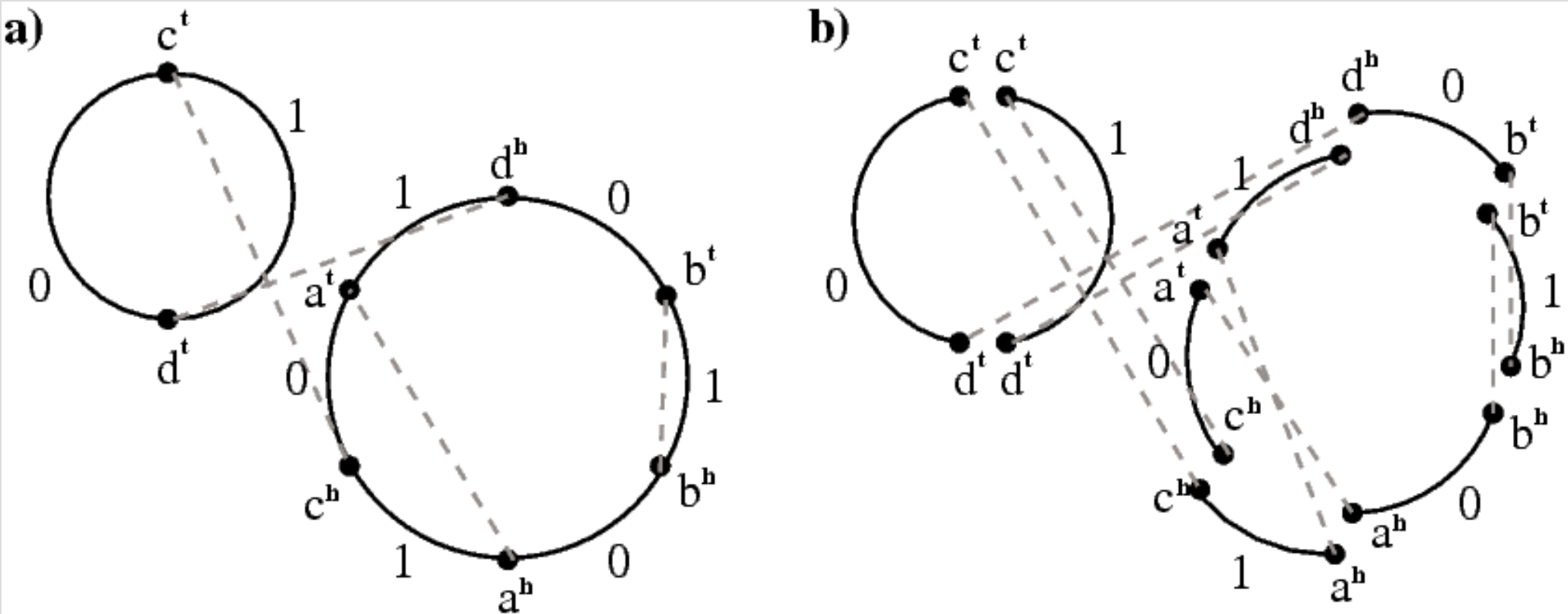
  i) $c(P,R' \oplus R') = c(P,2R'') = |P|/2 + EvenBlackCycles(P)$

  ii) $R' \oplus R'$ and $2R''$ can be transformed into each other by *criss-crossing* a pair of counterpart green edges

# Singular Genomes

Label black cycles of de Bruijn P' graph with {0,1} such that every two adjacent edges have different labels. The labelling is inherited by black-red cycle P.

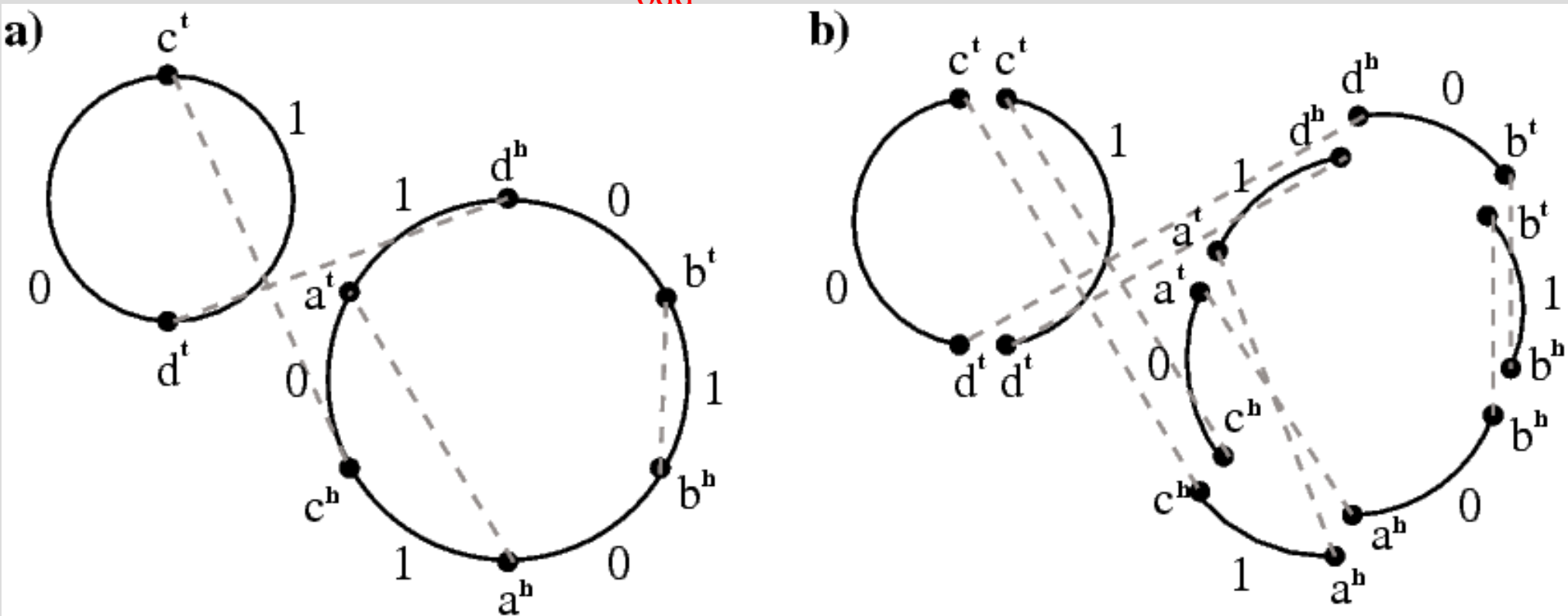
$$P = +a-b-b-d+c-a-d+c$$



# Singular Genomes

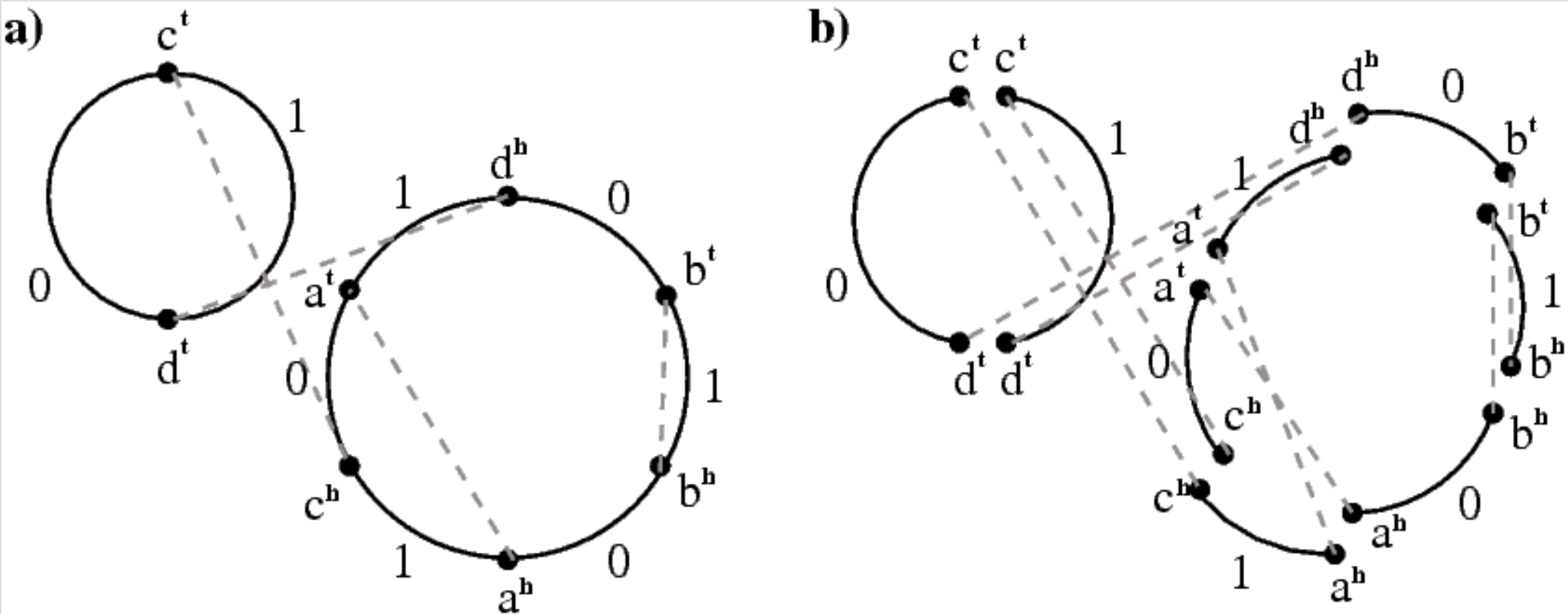
A red edge in P is called even if it connects black edges with the same label, and odd otherwise. The number of such edges $m_{even}$ and $m_{odd}$ are even.

Define parity(P) = $(m_{odd}/2) \bmod 2$.



# Singular Genomes

**Theorem.** For a singular genome P and perfect duplicated genome Q,
if c(G(P,Q)) = |P|/2 + EvenBlackCycles(P) then
Q=R⊕R **iff** parity(P)=1 and Q=2R **iff** parity(P)=0.



# Genome Halving Algorithm

- Construct de Bruijn graph P' of the given genome P
- Complete P' with a set of double green edges such that the resulting graph G' is non-crossing and green-red connected
- Find a maximum black-green cycle decomposition of G' and labeling of the genomes P and Q ($Q=R \oplus R$ or $Q=2R$) inducing this cycle decomposition
- If $Q=R \oplus R$ output R
- If $Q=2R$ and P is non-singular, find R' ...
- If $Q=2R$ and P is singular, then parity(P)=0. Find R' ...

# more Open Problems

**Open Problem 4**. For a genome P, find all pre-duplicated genomes R such that d(P,R⊕R) is minimal.

**Open Problem 5**. Given a genome P having each gene in *even* number of copies, recover the ancestral pre-duplicated genome R minimizing the reversal distance from R⊕R to P (i.e. allow the ancestral pre-duplicated genome to contain duplicated genes).

# That's it for today!