

Перестройки

Алгоритмы в биоинформатике

Антон Елисеев
eliseevantoncoon@gmail.com

Что было на прошлой лекции

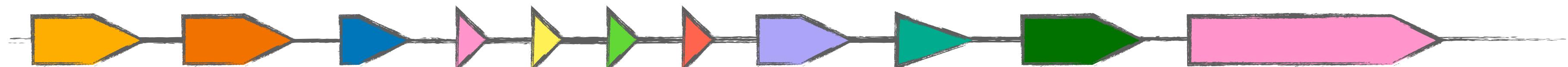
- Эффективный алгоритм поиска ближайшего генома из множества к заданному
- Прямое и обратное преобразование Барроуза-Уилера
- Поиск подстроки в строке при помощи индекса Барроуза-Уилера
- Выравнивание при помощи BWT

Что будет в этой лекции

- Перестановки внутри X хромосомы человека и мыши
- Перестановки в опухолях
- Хотспоты и Random Breakage Models
- Перестройки и reversal distance
- Breakpoint Graphs
- Что такое synteny block

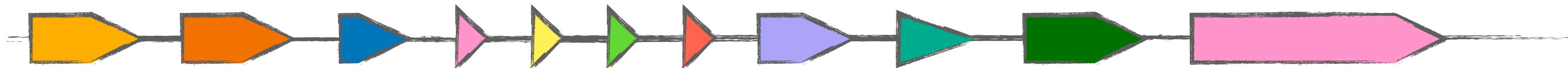
Человек и мышь

Человек (X хромосома 156,040,895 бп)

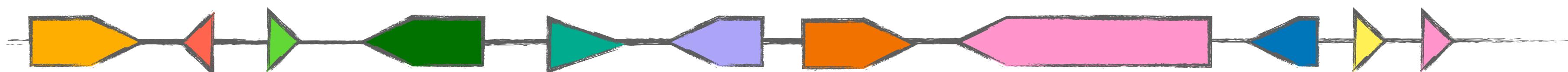


Человек и мышь

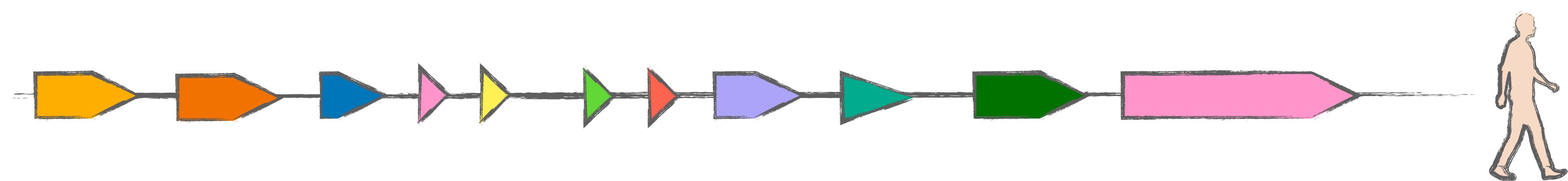
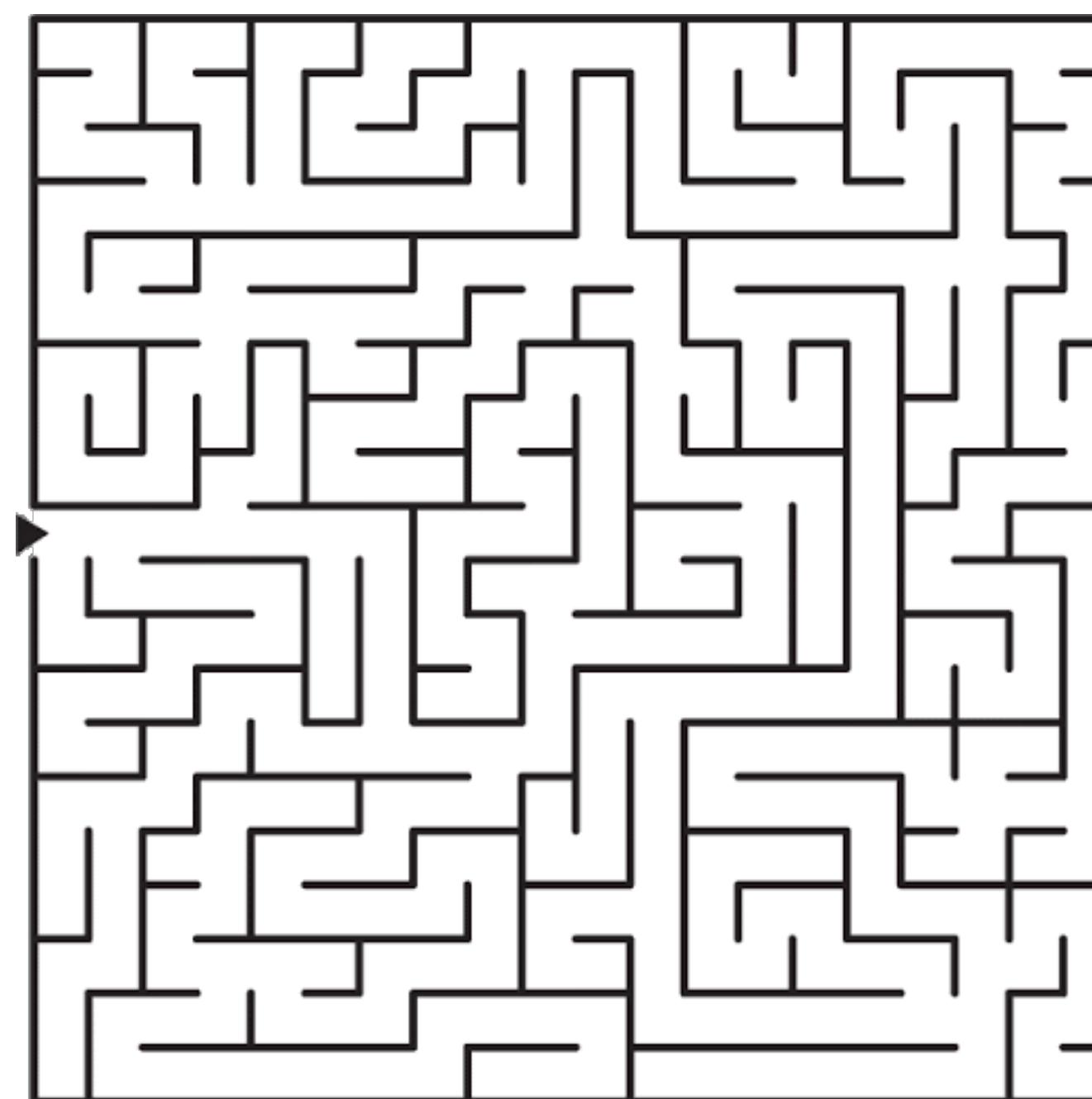
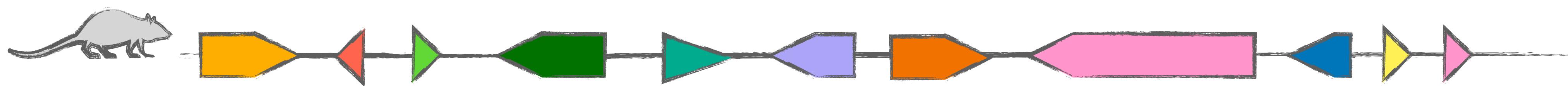
Человек (X хромосома 156,040,895 бп)



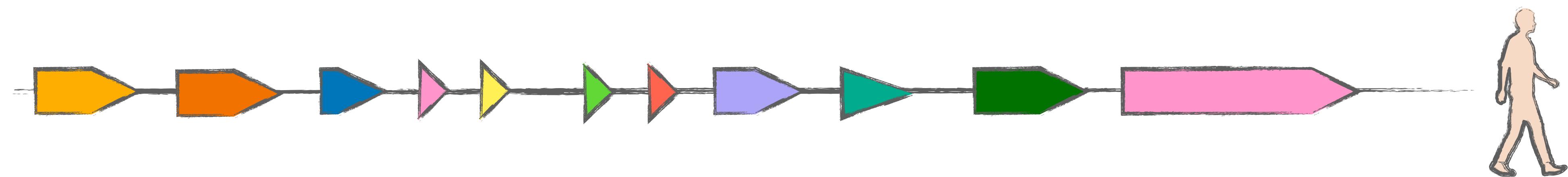
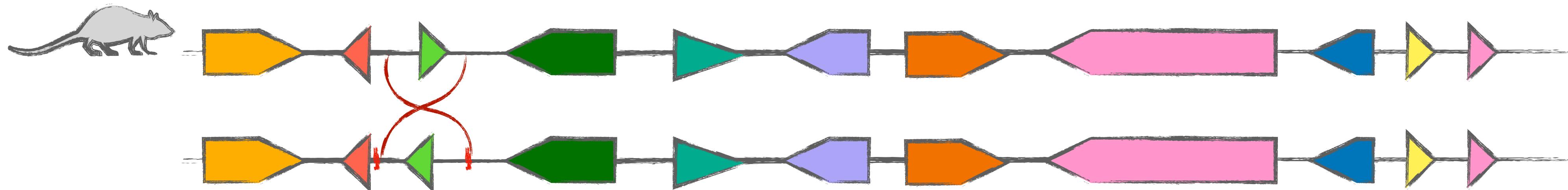
Мышь (X хромосома)



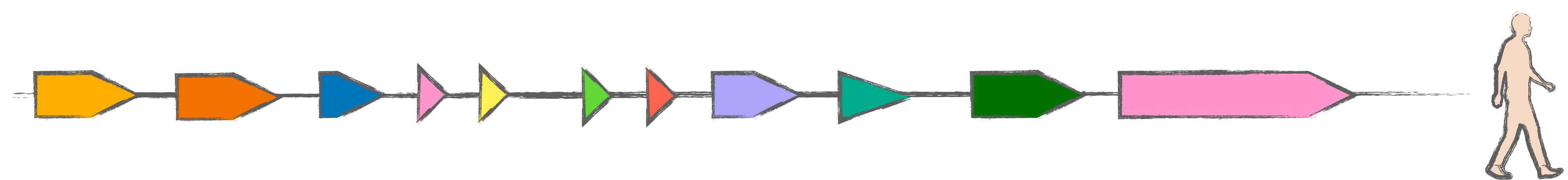
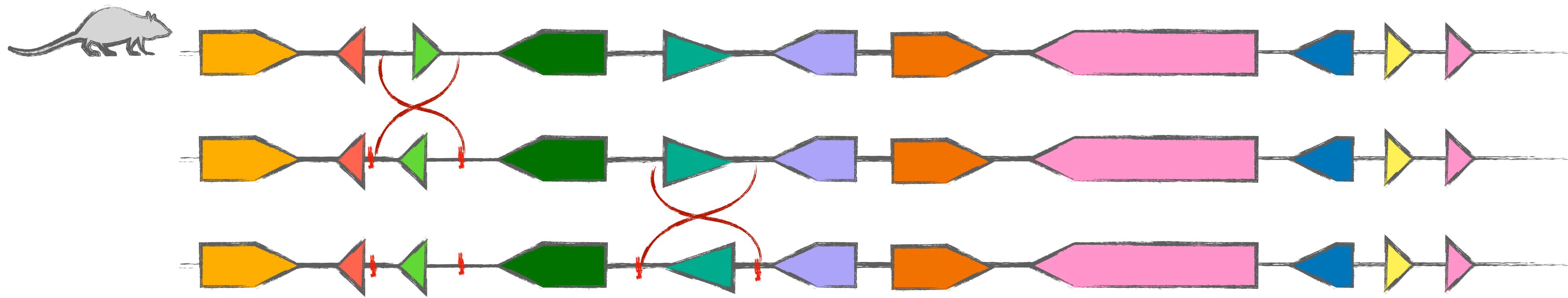
Мышь -> Человек



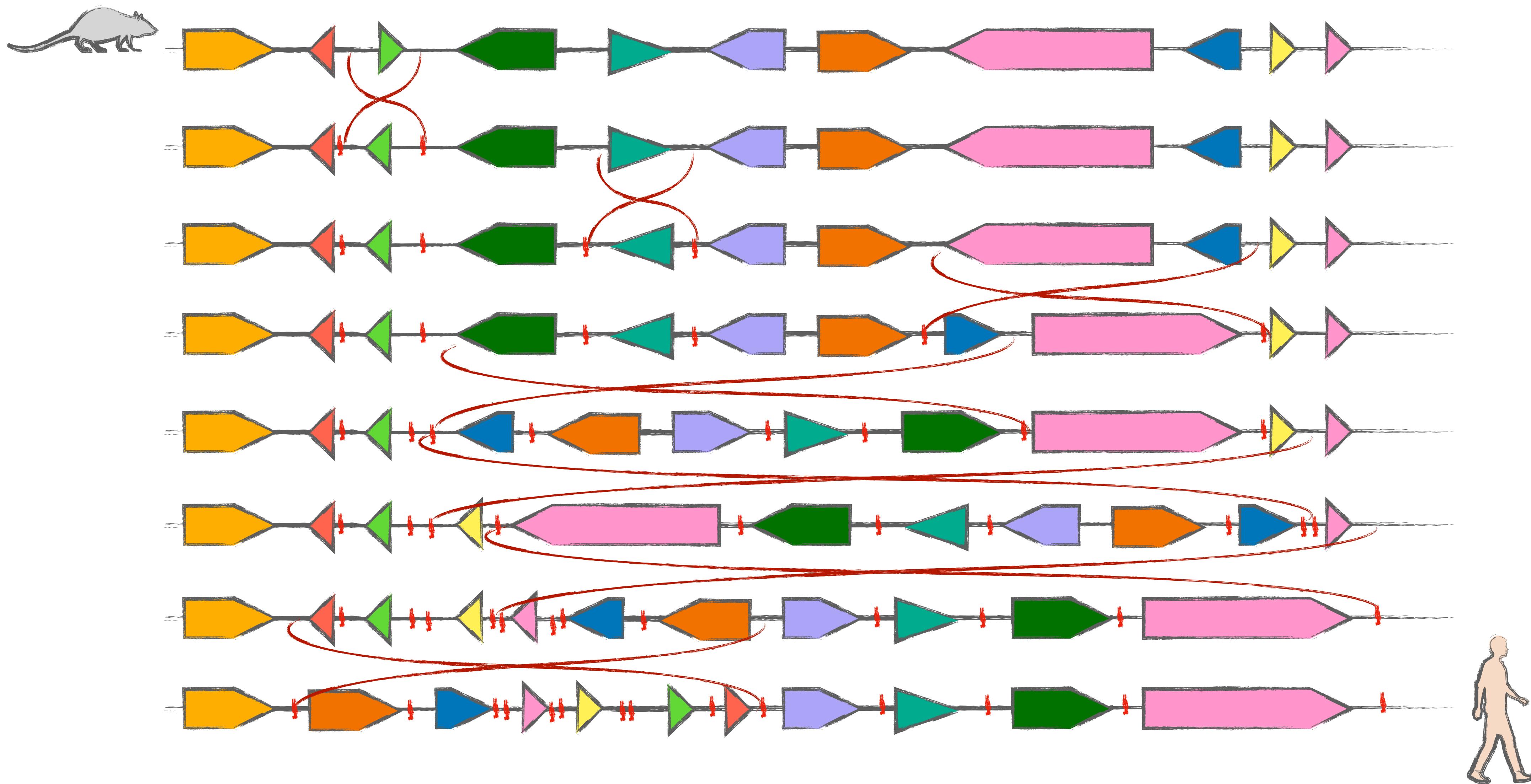
Мышь → Человек



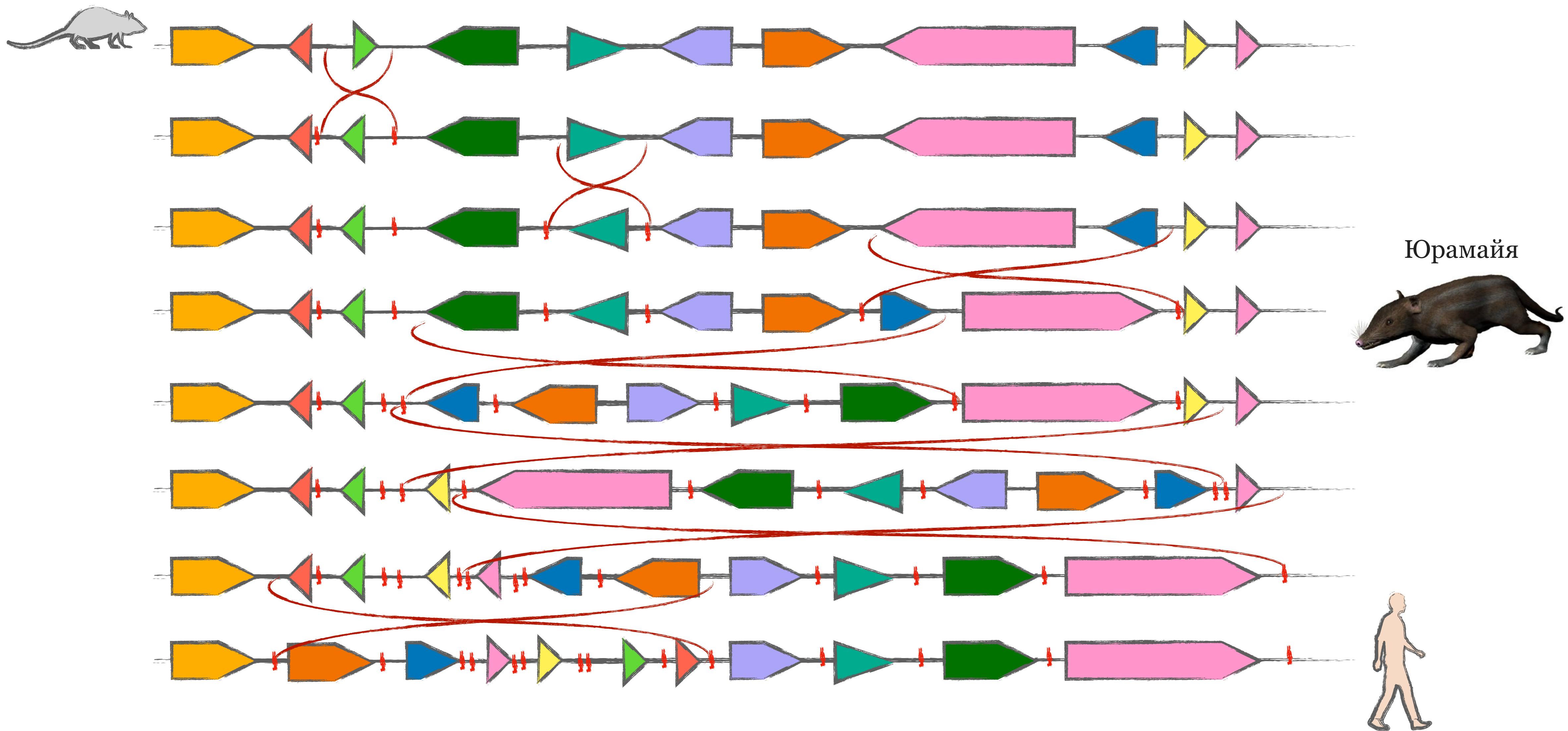
Мышь → Человек



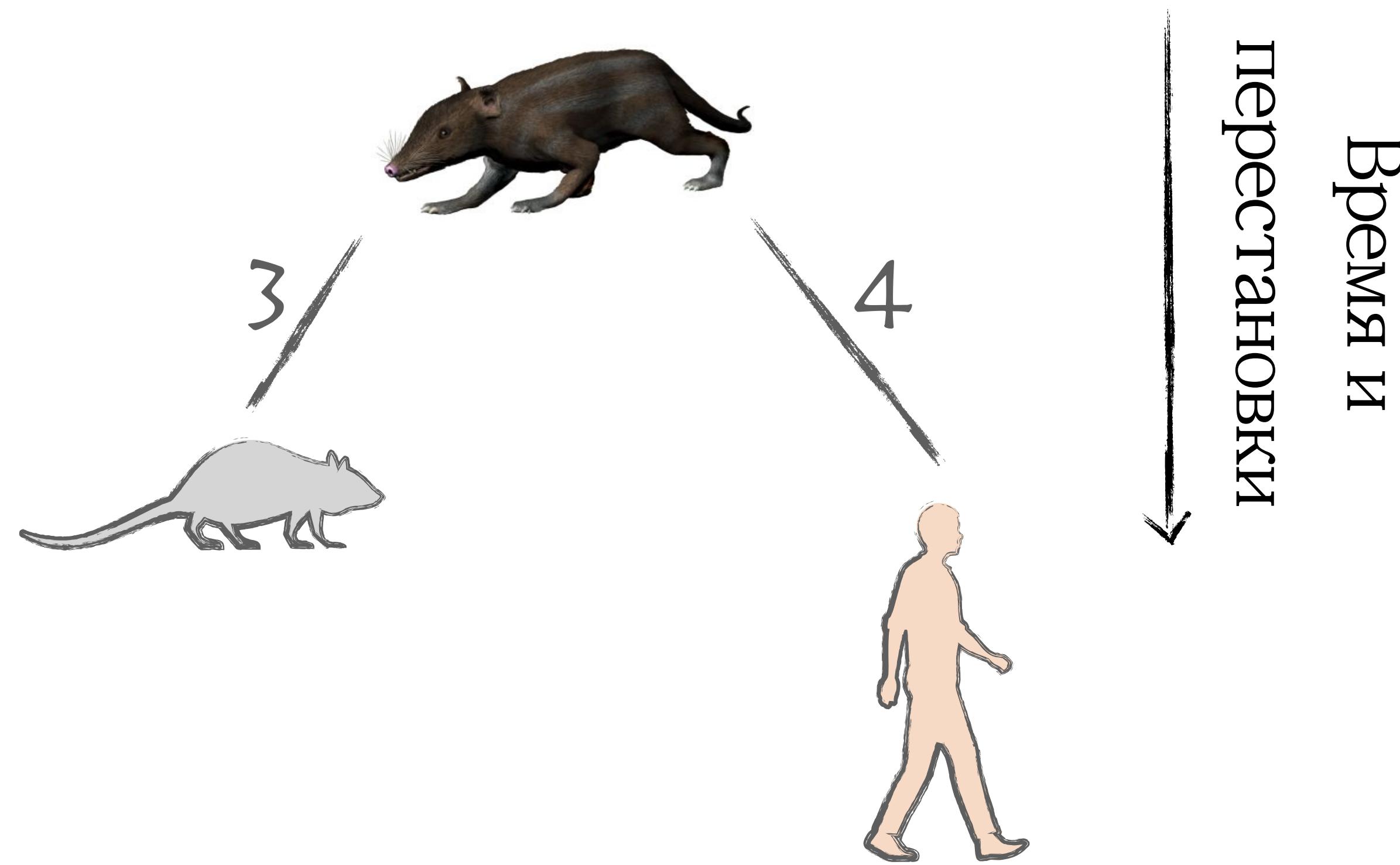
Мышь → Человек



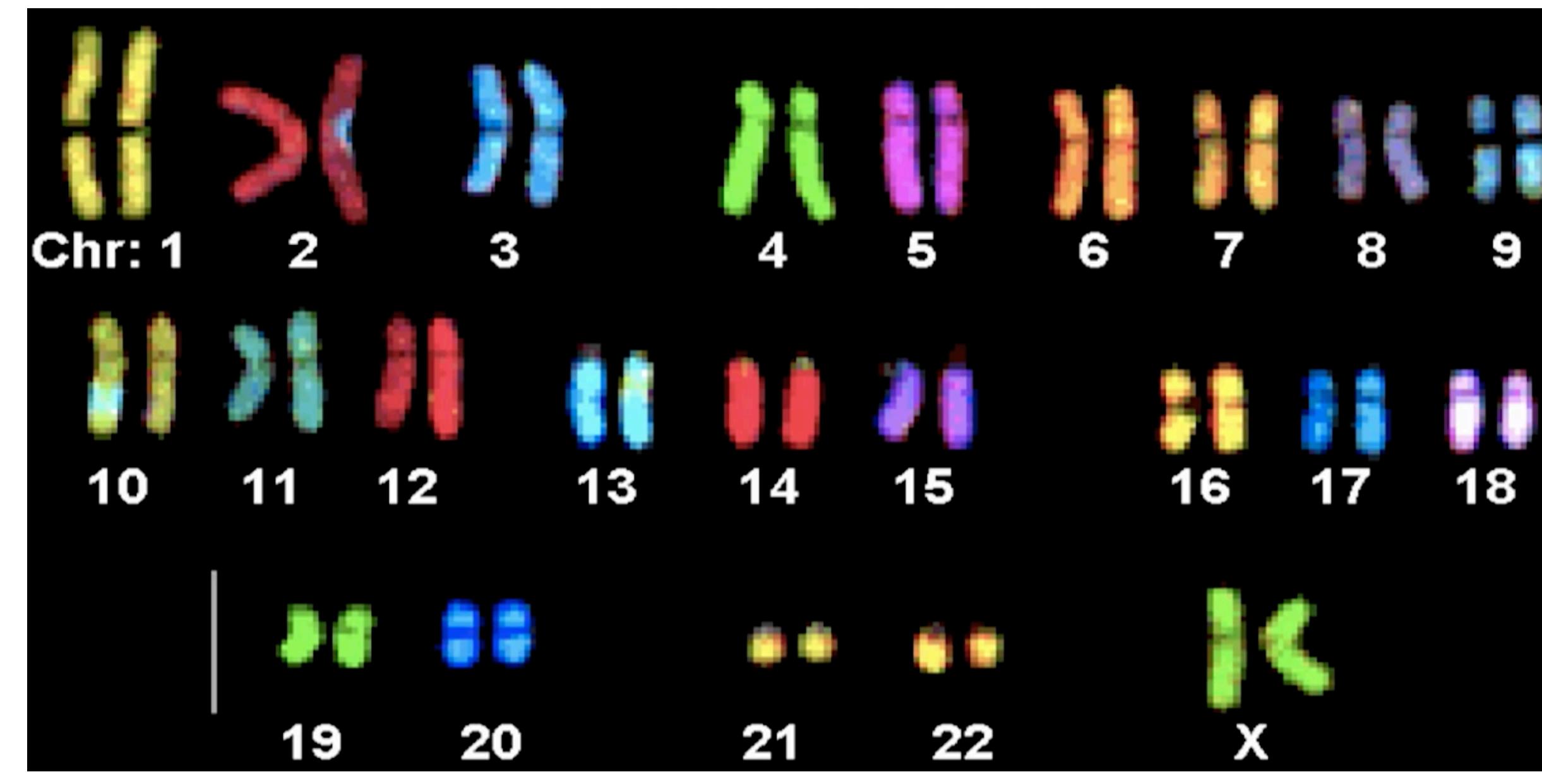
Мышь -> Человек



Мышь → Человек

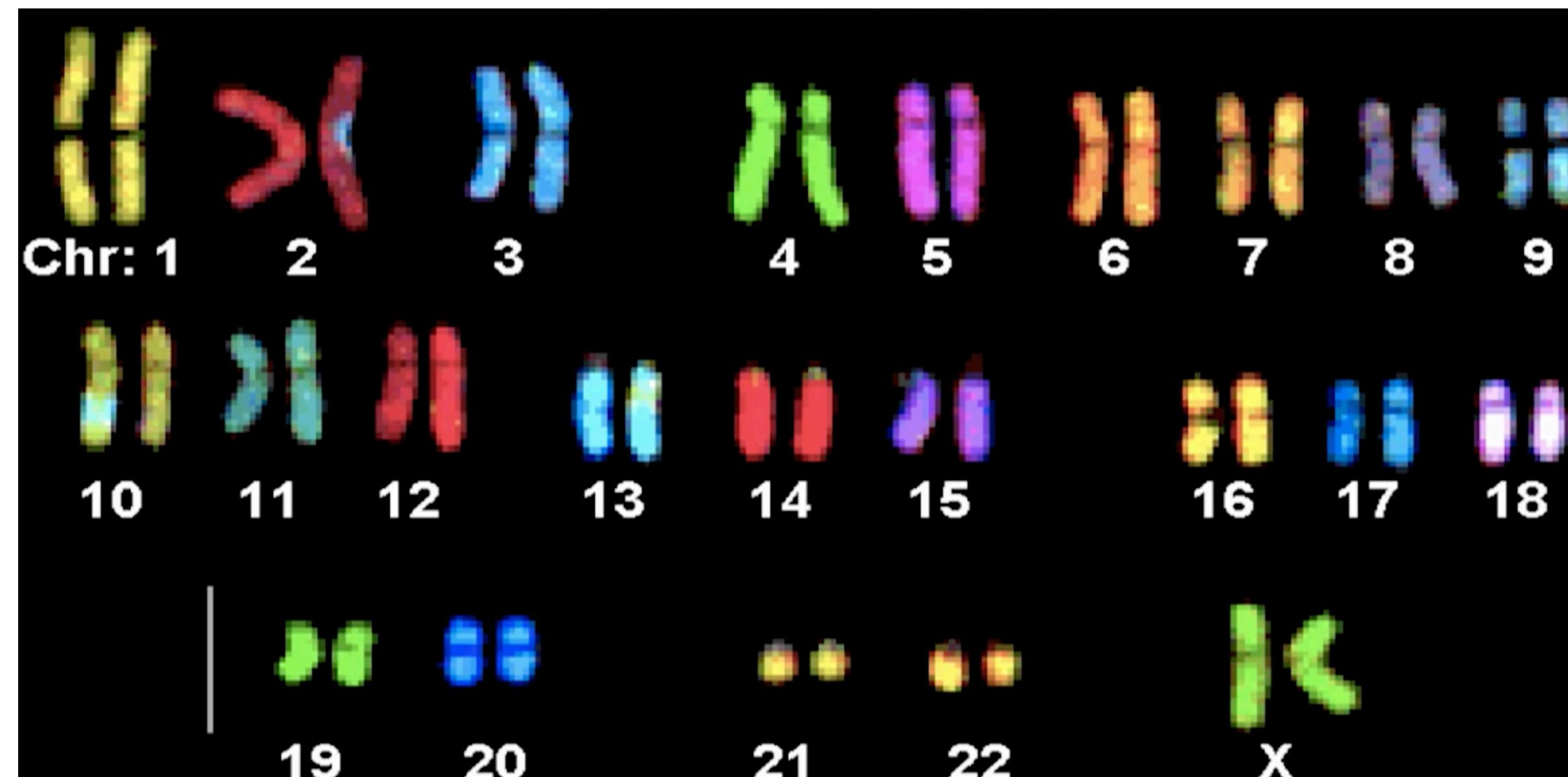


Перестройки в опухолях

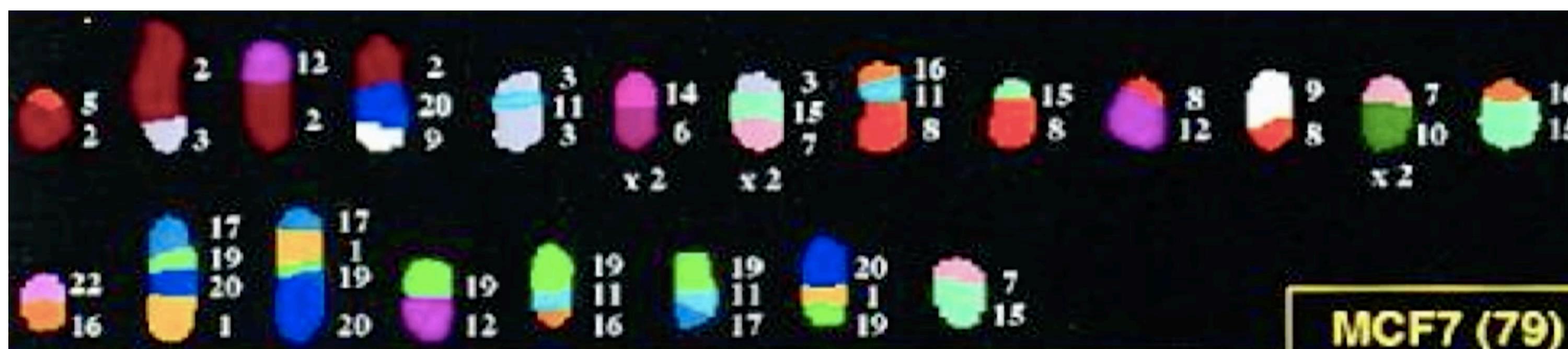


Хромосомы в здоровой клетке

Перестройки в опухолях

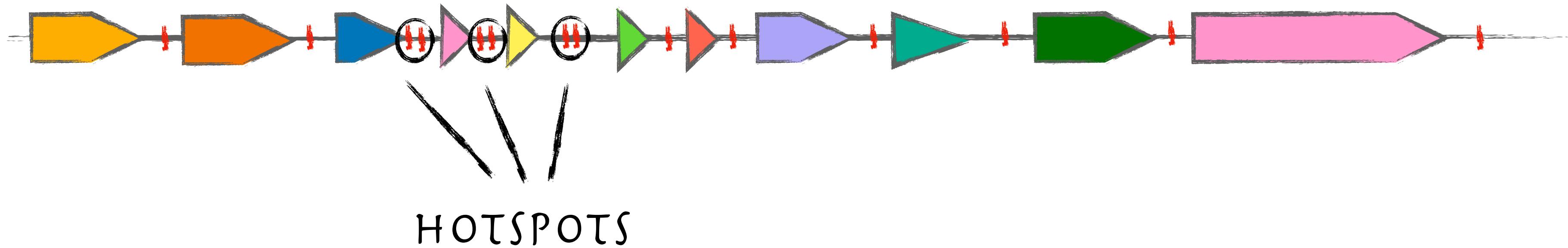


Хромосомы в здоровой клетке



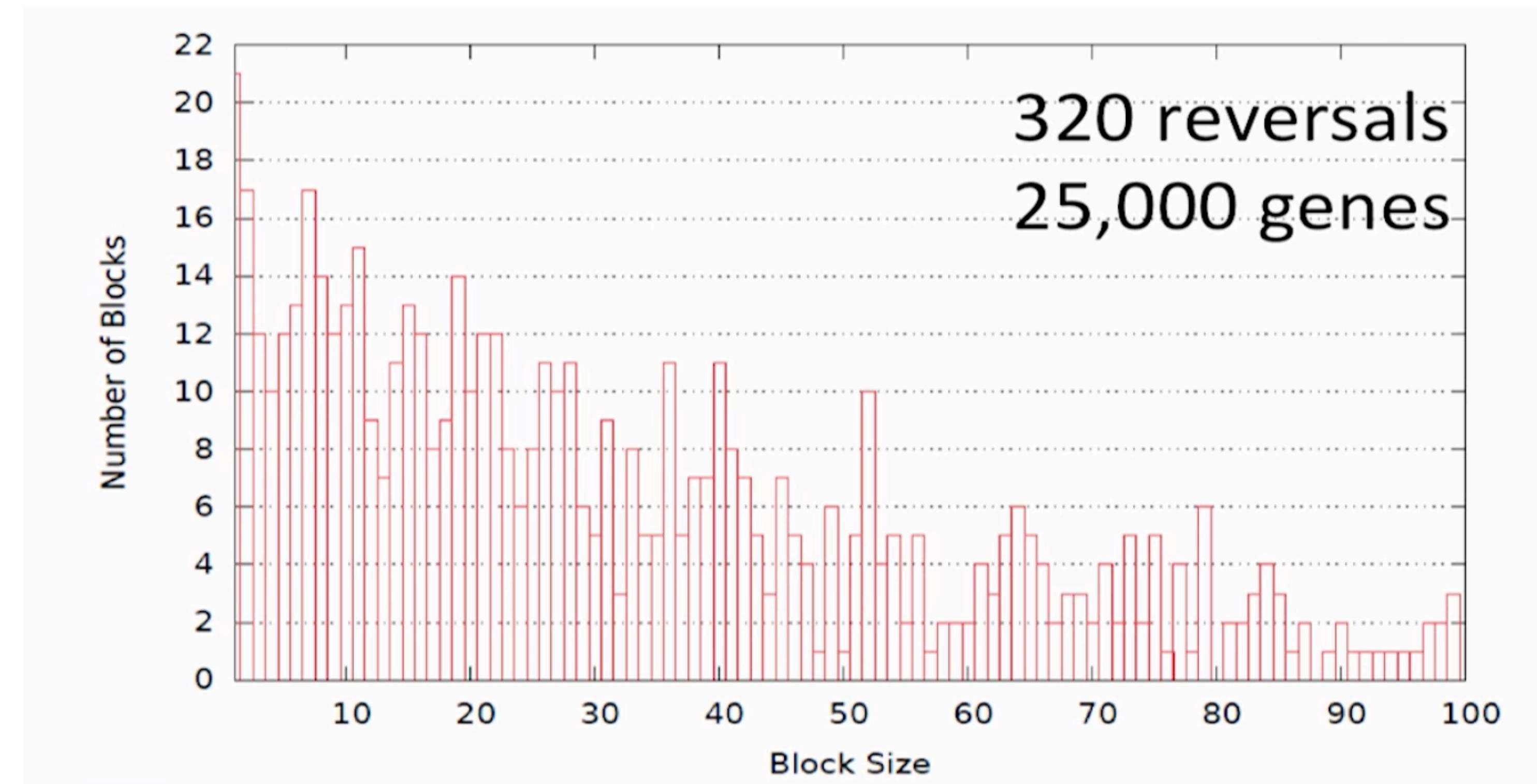
рак молочной железы

Random Breakage Models

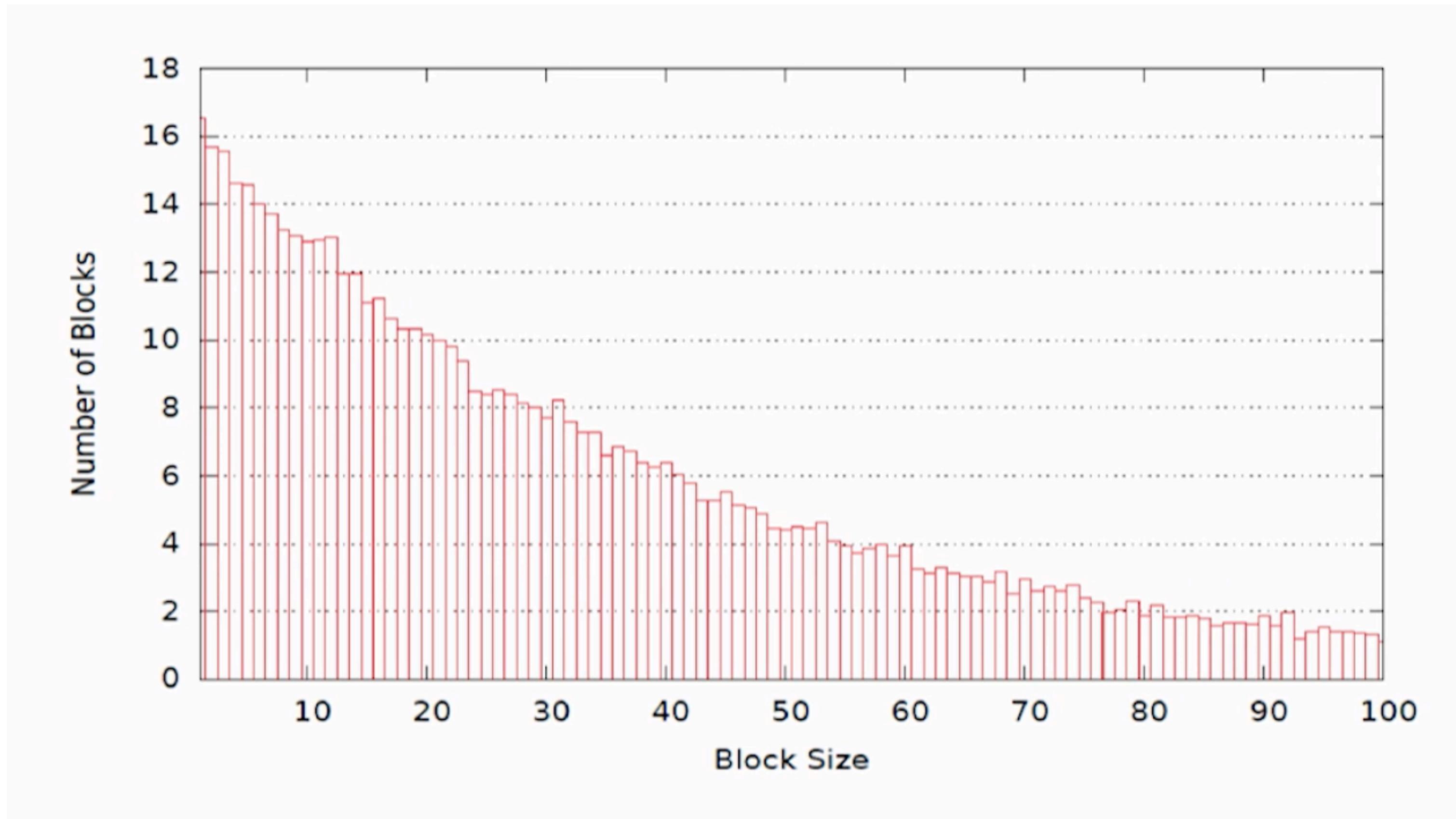


- Сусуму Оно, 1973 – перестановки происходят в случайных местах (нет особых хрупких мест)
- Надеу и Тейлор, 1984 – статистическое подтверждение Random Breakage Models

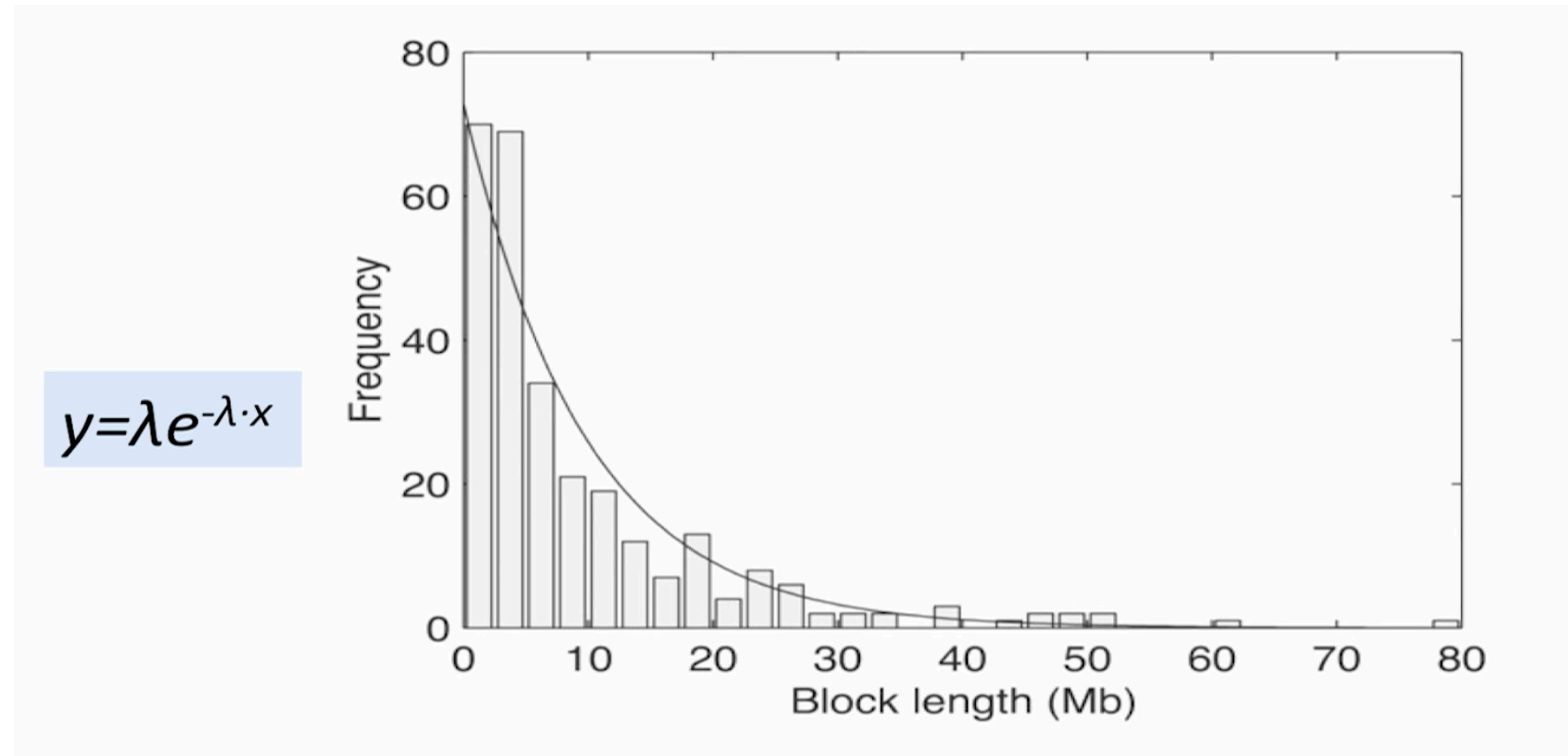
Random Breakage Models



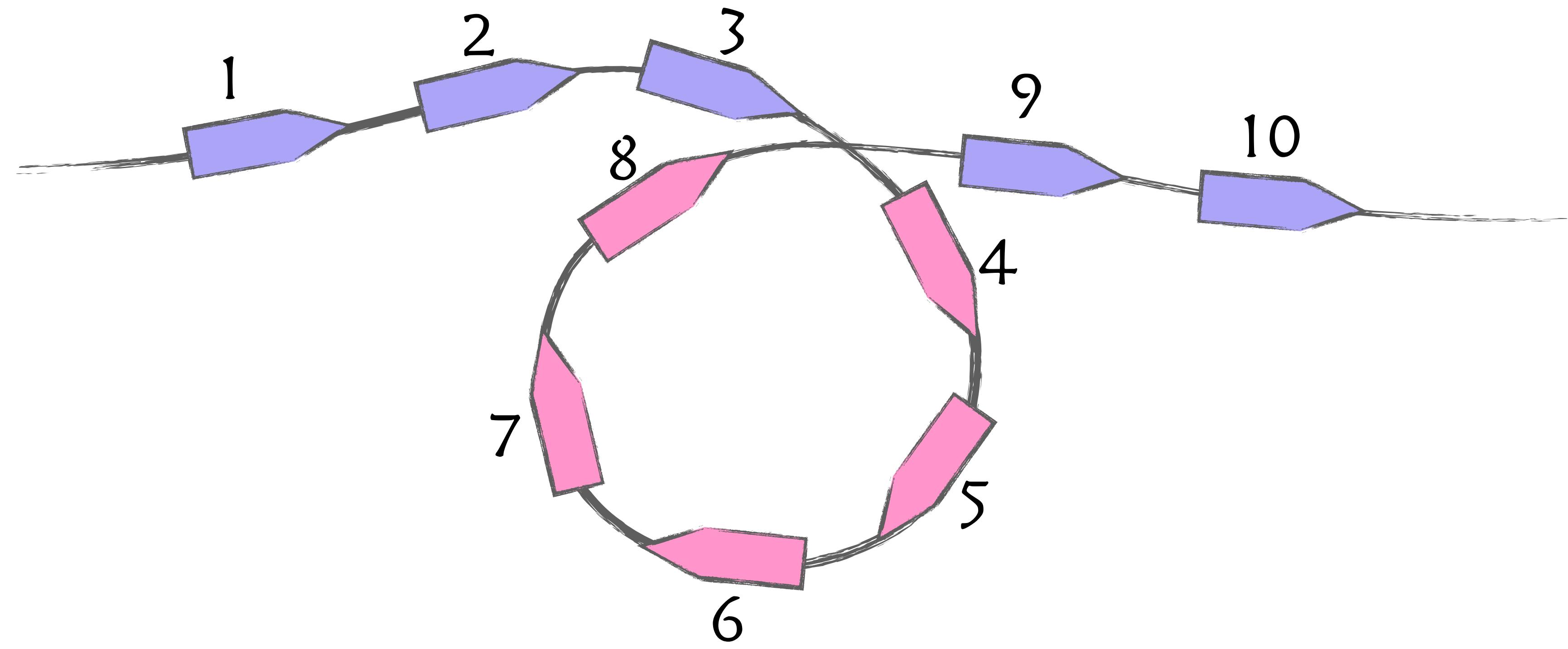
Random Breakage Models



Random Breakage Models

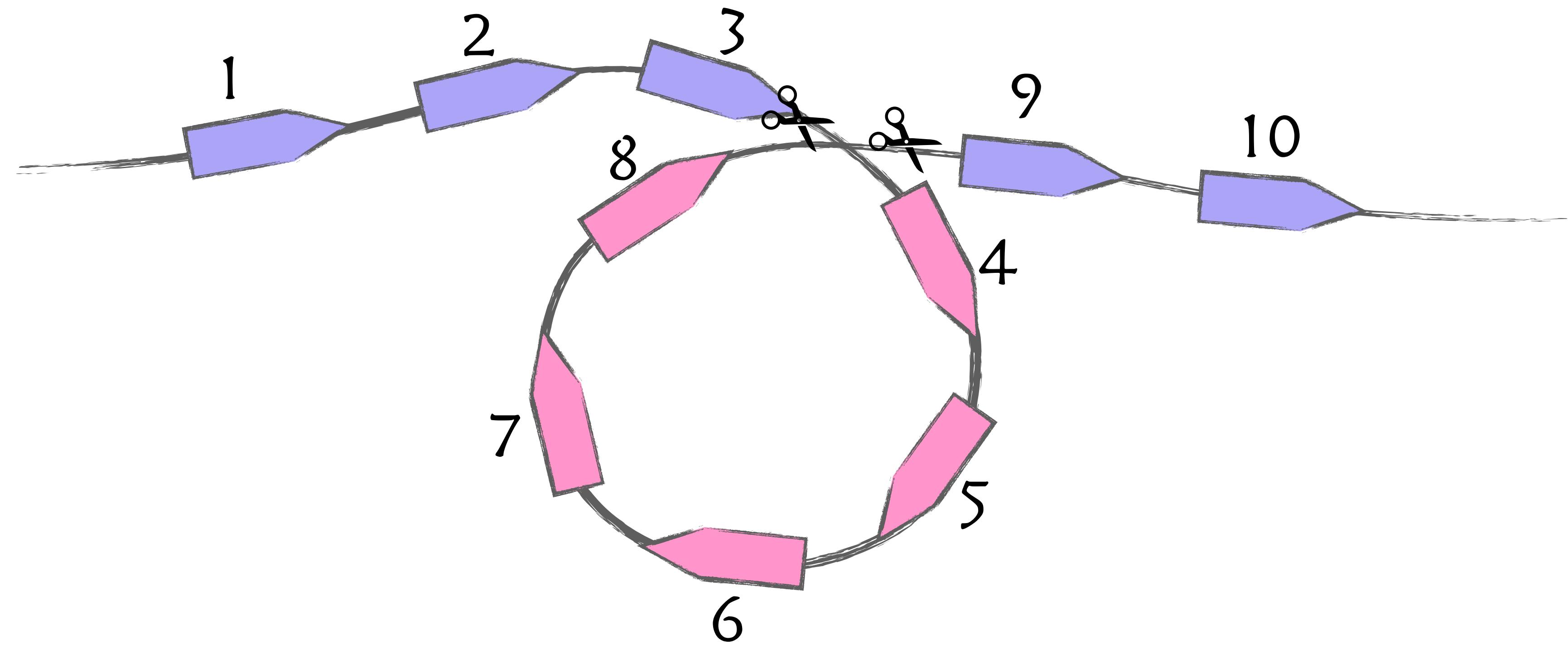


Сортировка разворотами



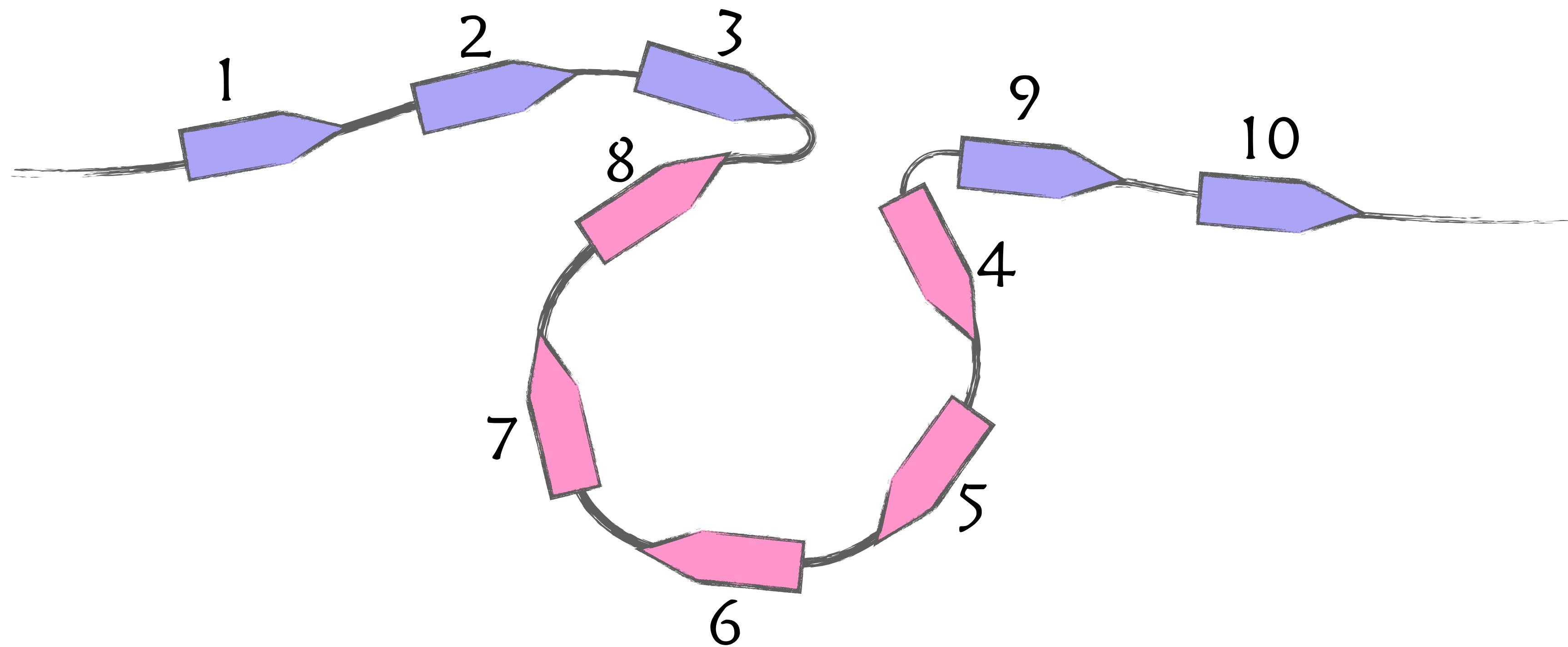
+1 +2 +3 +4 +5 +6 +7 +8 +9 +10

Сортировка разворотами



+1 +2 +3 +4 +5 +6 +7 +8 +9 +10

Сортировка разворотами



+1 +2 +3 -8 -7 -6 -5 -4 +9 +10

2 BREAKPOINTS - 1 REVERSAL

Сортировка разворотами

2 -4 -3 5 -8 -7 -6 1

Сортировка разворотами

2	<u>-4</u>	<u>-3</u>	5	-8	-7	-6	1
2	3	4	5	<u>-8</u>	<u>-7</u>	<u>-6</u>	1
2	3	4	5	6	7	8	<u>1</u>
<u>2</u>	3	4	5	6	7	8	-1
<u>-8</u>	<u>-7</u>	<u>-6</u>	<u>-5</u>	<u>-4</u>	<u>-3</u>	<u>-2</u>	<u>-1</u>
1	2	3	4	5	6	7	8

Сортировка разворотами

2	-4	-3	5	-8	-7	-6	1
<u>2</u>	<u>3</u>	<u>4</u>	<u>5</u>	-8	-7	-6	1
-5	-4	-3	-2	<u>-8</u>	<u>-7</u>	<u>-6</u>	<u>1</u>
<u>-5</u>	<u>-4</u>	<u>-3</u>	<u>-2</u>	<u>-1</u>	6	7	8
1	2	3	4	5	6	7	8

Будем называть reversal distance – минимальное количество поворотов, необходимых для сортировки.

Сортировка разворотами. Жадно

2	-4	-3	5	-8	-7	-6	1
-1	6	7	8	-5	3	4	-2

Сортировка разворотами. Жадно

2	-4	-3	5	-8	-7	-6	1
-1	6	7	8	-5	3	4	-2
1	6	7	8	-5	3	4	-2

Сортировка разворотами. Жадно

2	-4	-3	5	-8	-7	-6	1
-1	6	7	8	-5	3	4	-2
1	6	7	8	-5	3	4	-2
1	2	-4	-3	5	-8	-7	-6

Сортировка разворотами. Жадно

2	-4	-3	5	-8	-7	-6	1
-1	6	7	8	-5	3	4	-2
1	6	7	8	-5	3	4	-2
1	2	-4	-3	5	-8	-7	-6
1	2	3	4	5	-8	-7	-6
1	2	3	4	5	6	7	8

Даже используя жадный подход мы точно не сделаем более $2n$ операций

Breakpoint Theorem

2 -4 ✓ -3 5 -8 ✓ -7 ✓ -6 1

Breakpoint Theorem

~~x~~ 2 ~~x~~ -4 ✓ -3 ~~x~~ 5 ~~x~~ -8 ✓ -7 ✓ -6 ~~x~~ 1 ~~x~~

Сколько ~~x~~ в отсортированной последовательности?

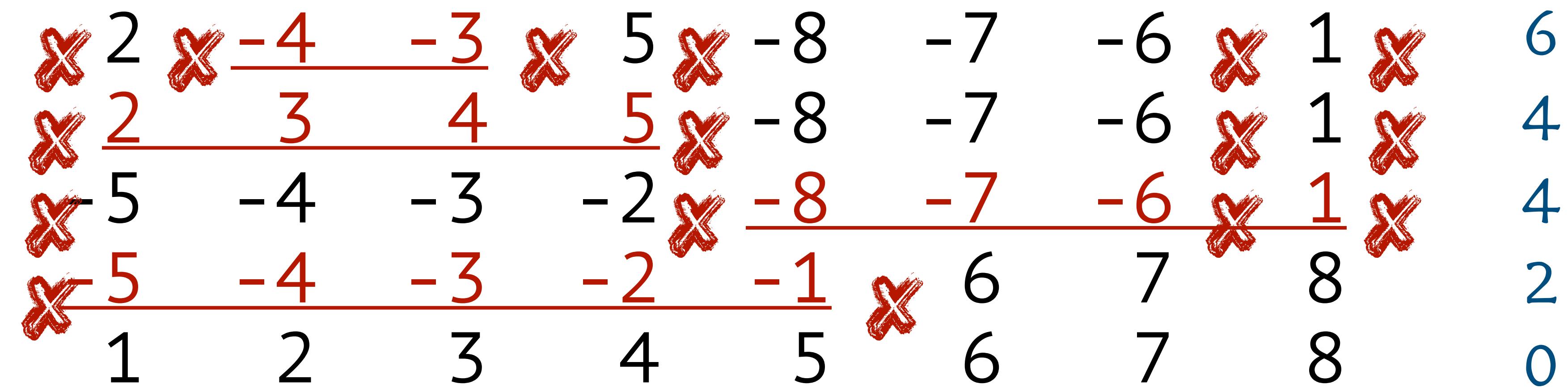
Breakpoint Theorem

~~x~~ 2 ~~x~~ -4 ✓ -3 ~~x~~ 5 ~~x~~ -8 ✓ -7 ✓ -6 ~~x~~ 1 ~~x~~

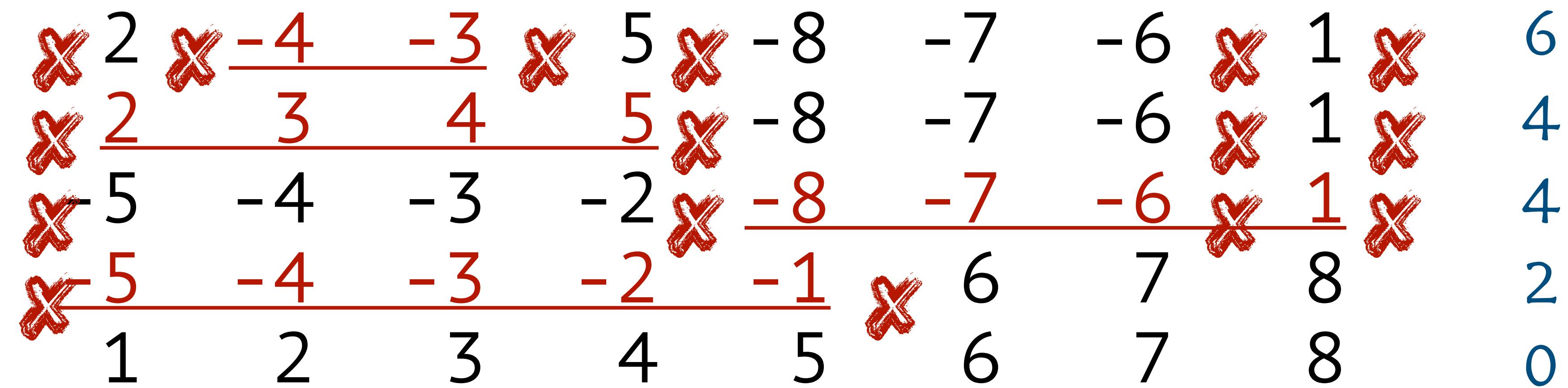
Сколько ~~x~~ в отсортированной последовательности?

ноль!

Breakpoint Theorem



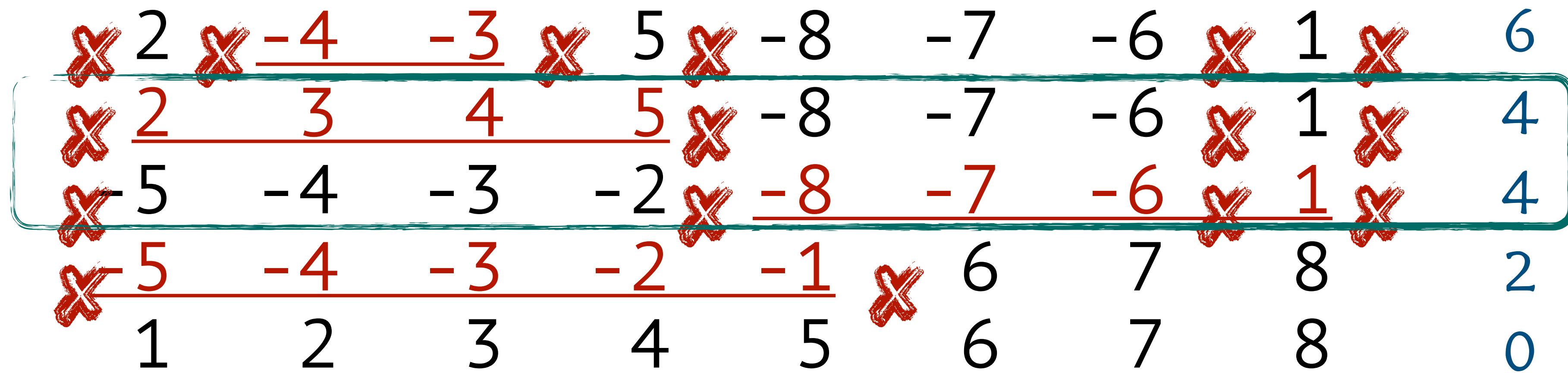
Breakpoint Theorem



Breakpoint Theorem

$$RD(seq) \geq \frac{(\# bp \text{ in } seq)}{2}$$

Breakpoint Theorem



Breakpoint Theorem

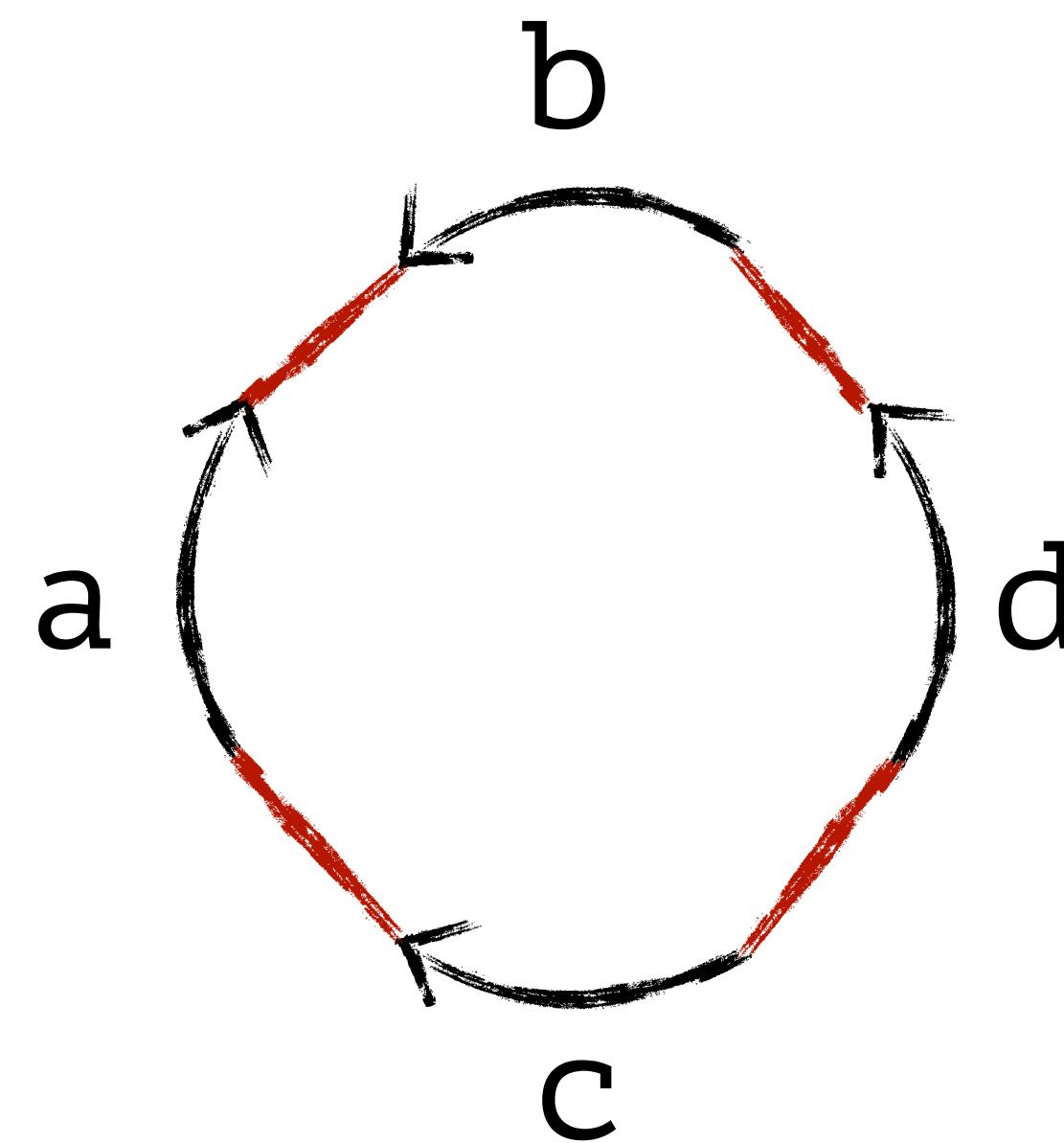
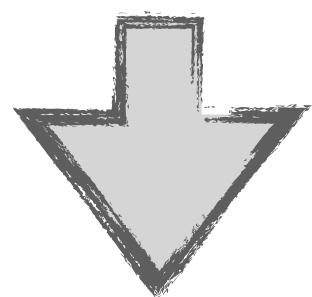
$$RD(seq) \geq \frac{(\# bp \text{ in } seq)}{2}$$

Breakpoint graph

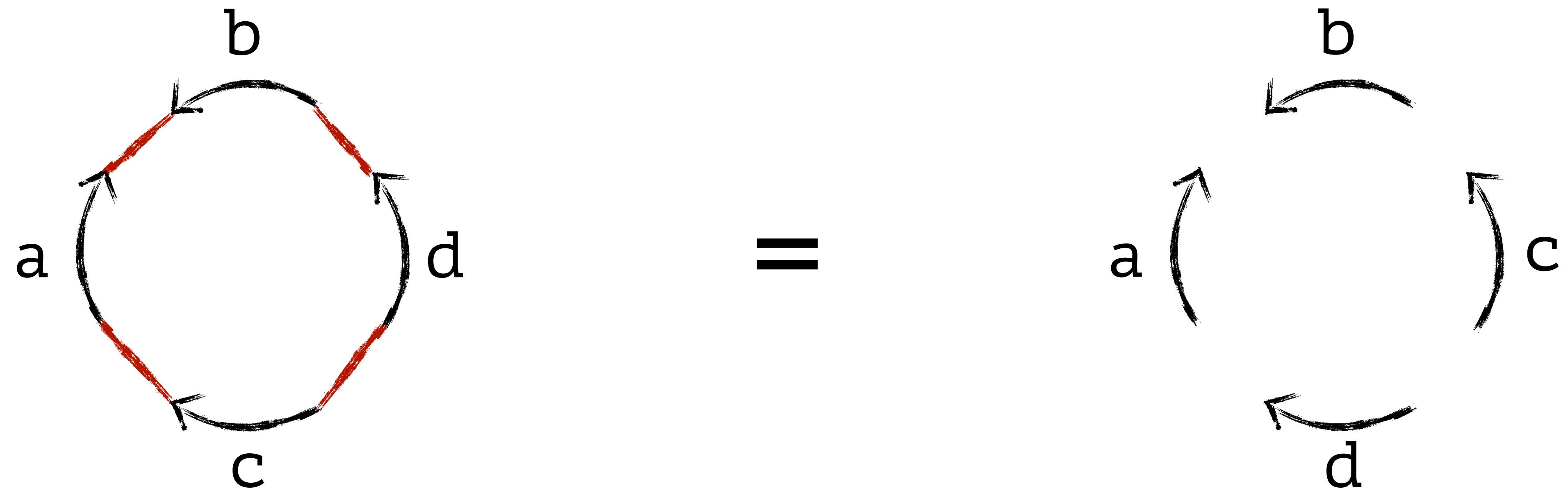
+a -b -d +c

Breakpoint graph

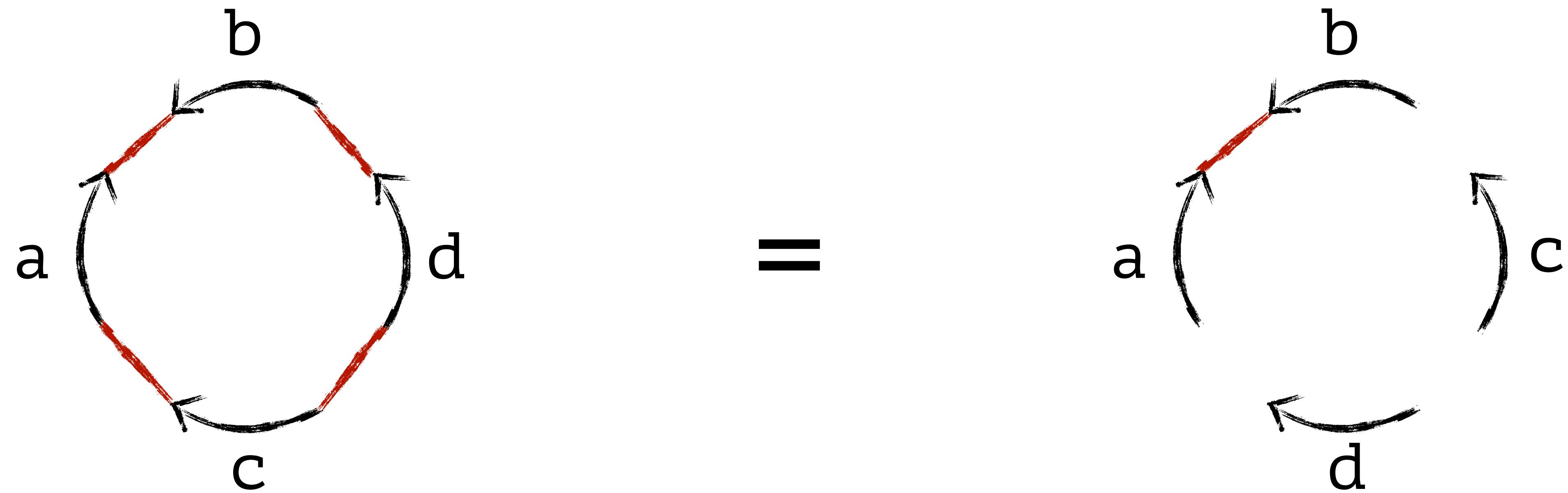
+a -b -d +c



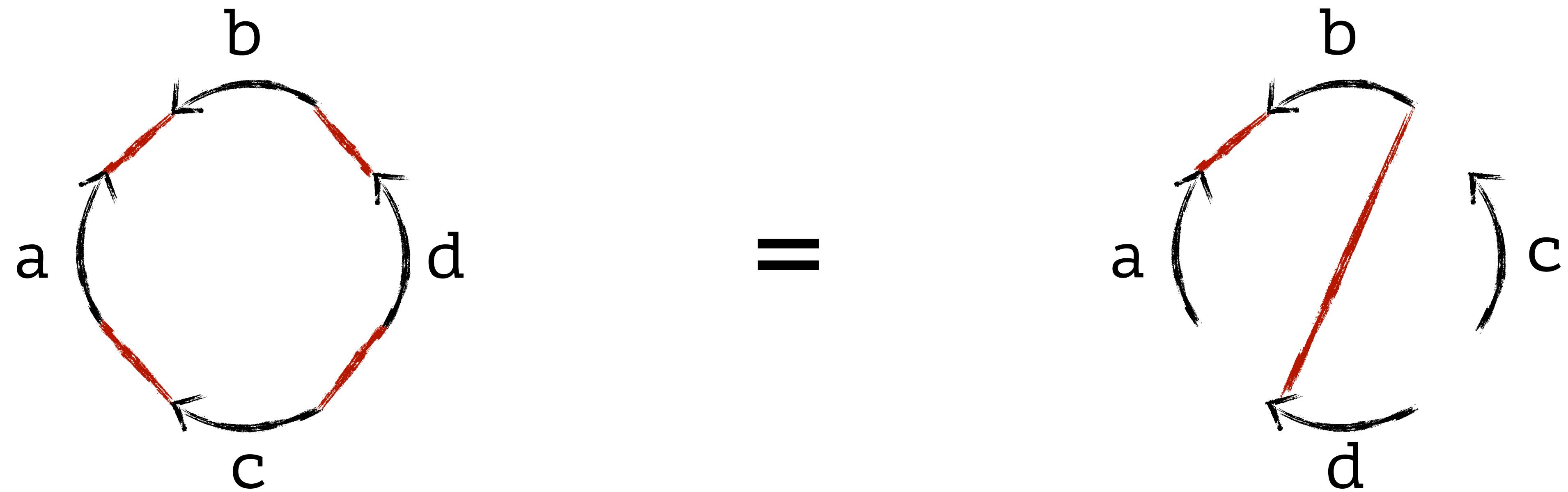
Breakpoint graph



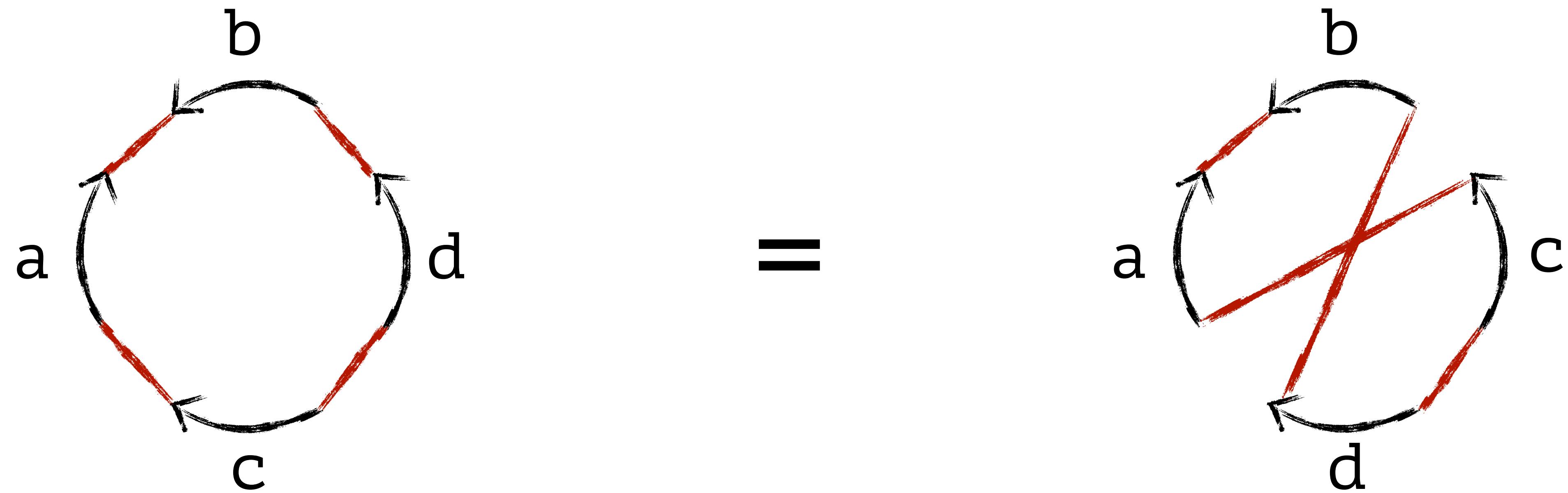
Breakpoint graph



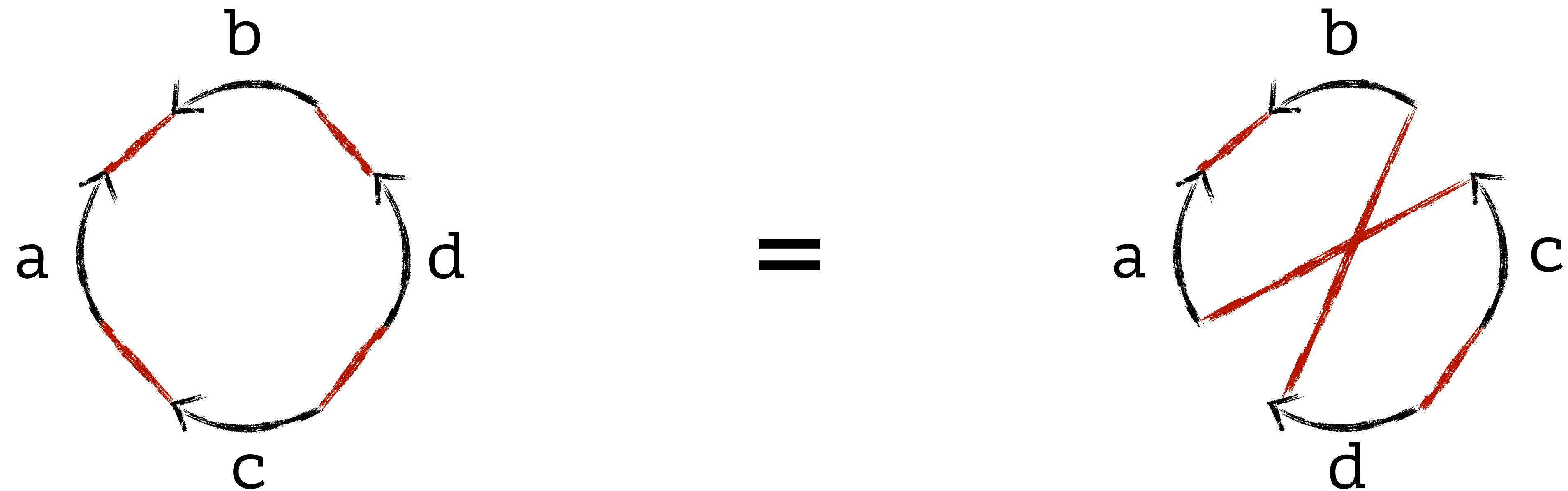
Breakpoint graph



Breakpoint graph

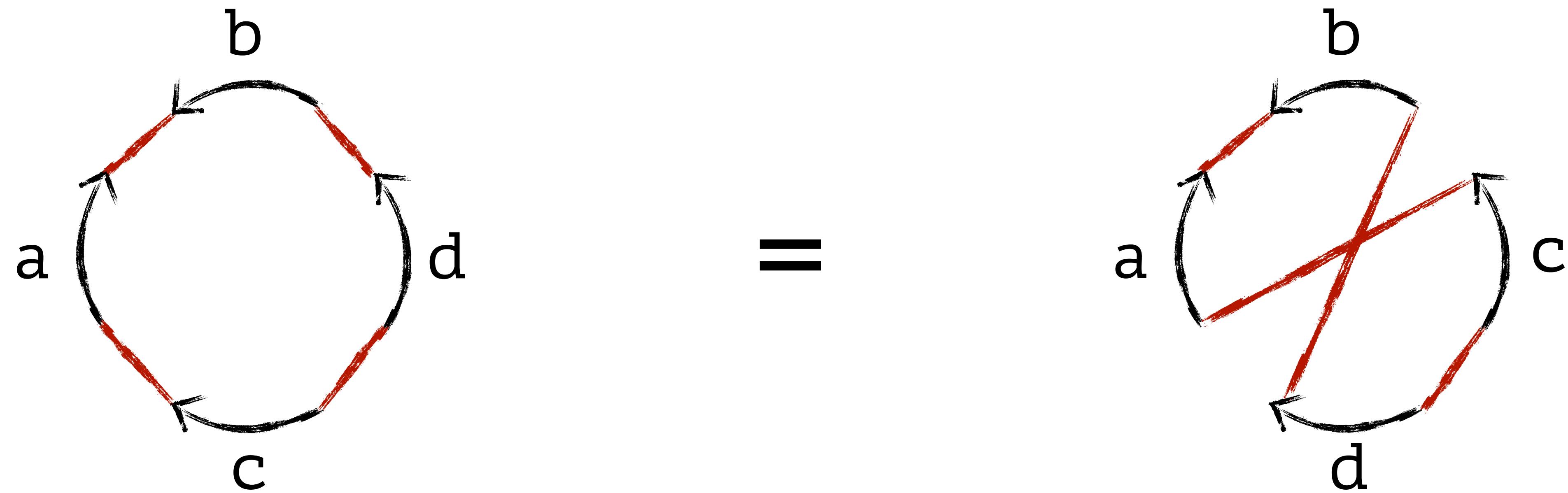


Breakpoint graph



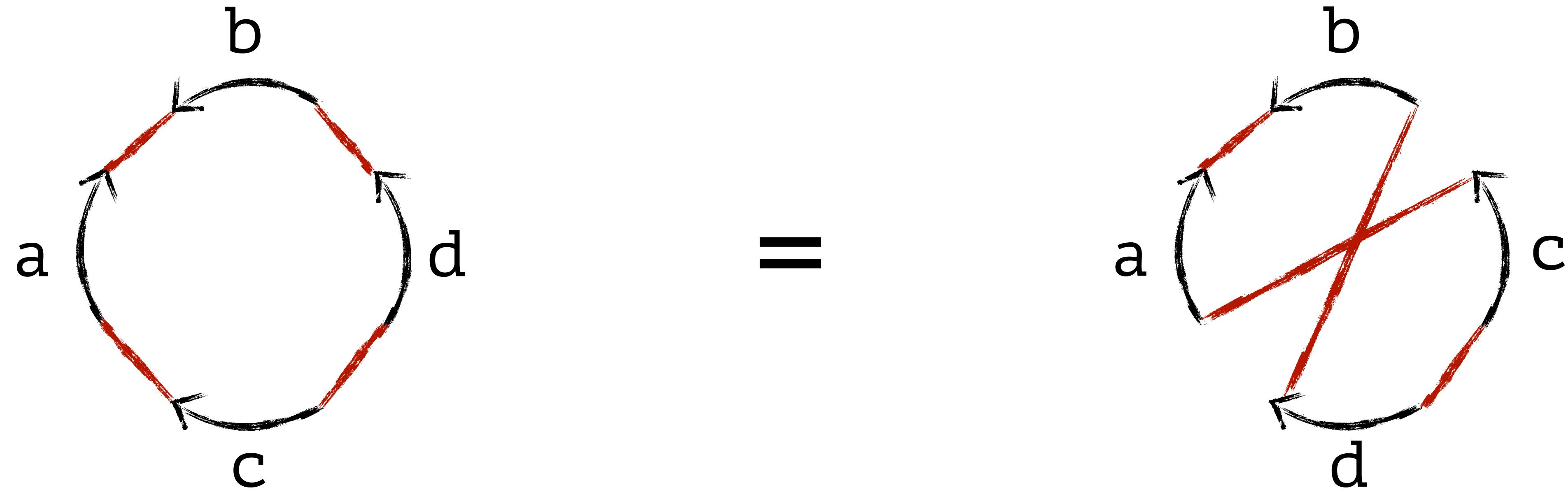
Что такое разворот участка генома в такой интерпретации?

Breakpoint graph



Что такое разворот участка генома в такой интерпретации?
Это удаление 2 красных ребер и добавление 2 новых!

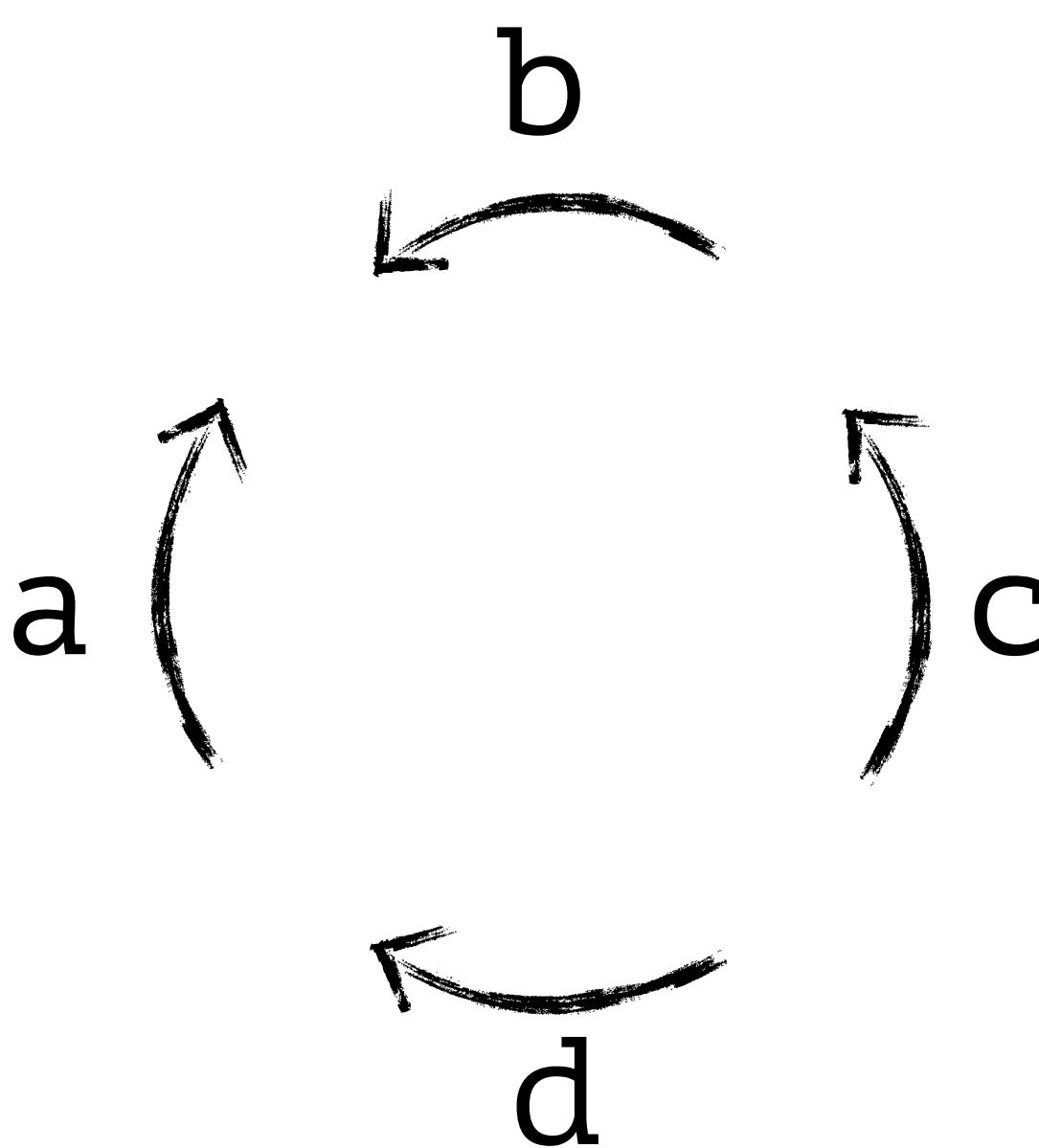
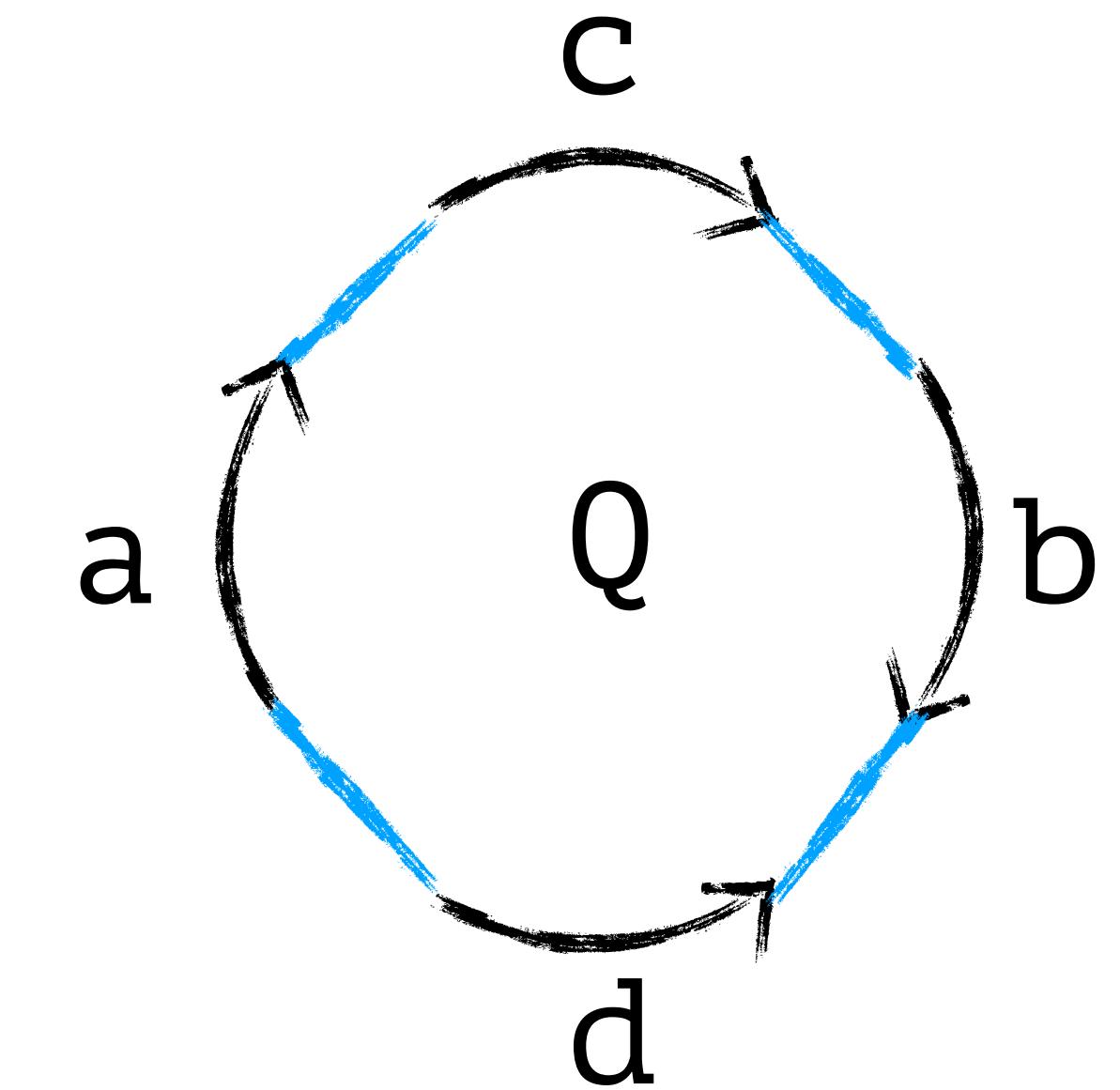
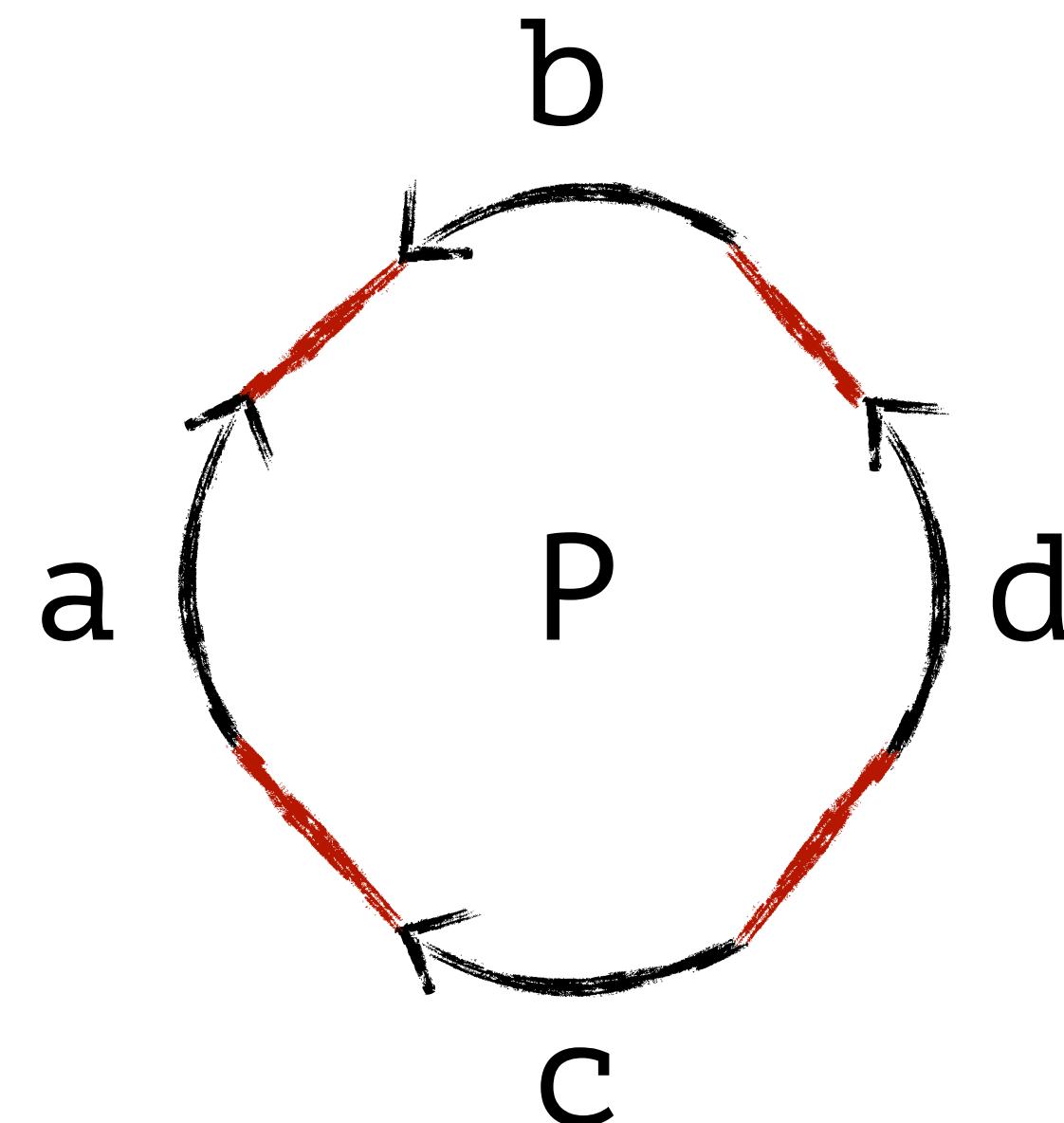
Breakpoint graph



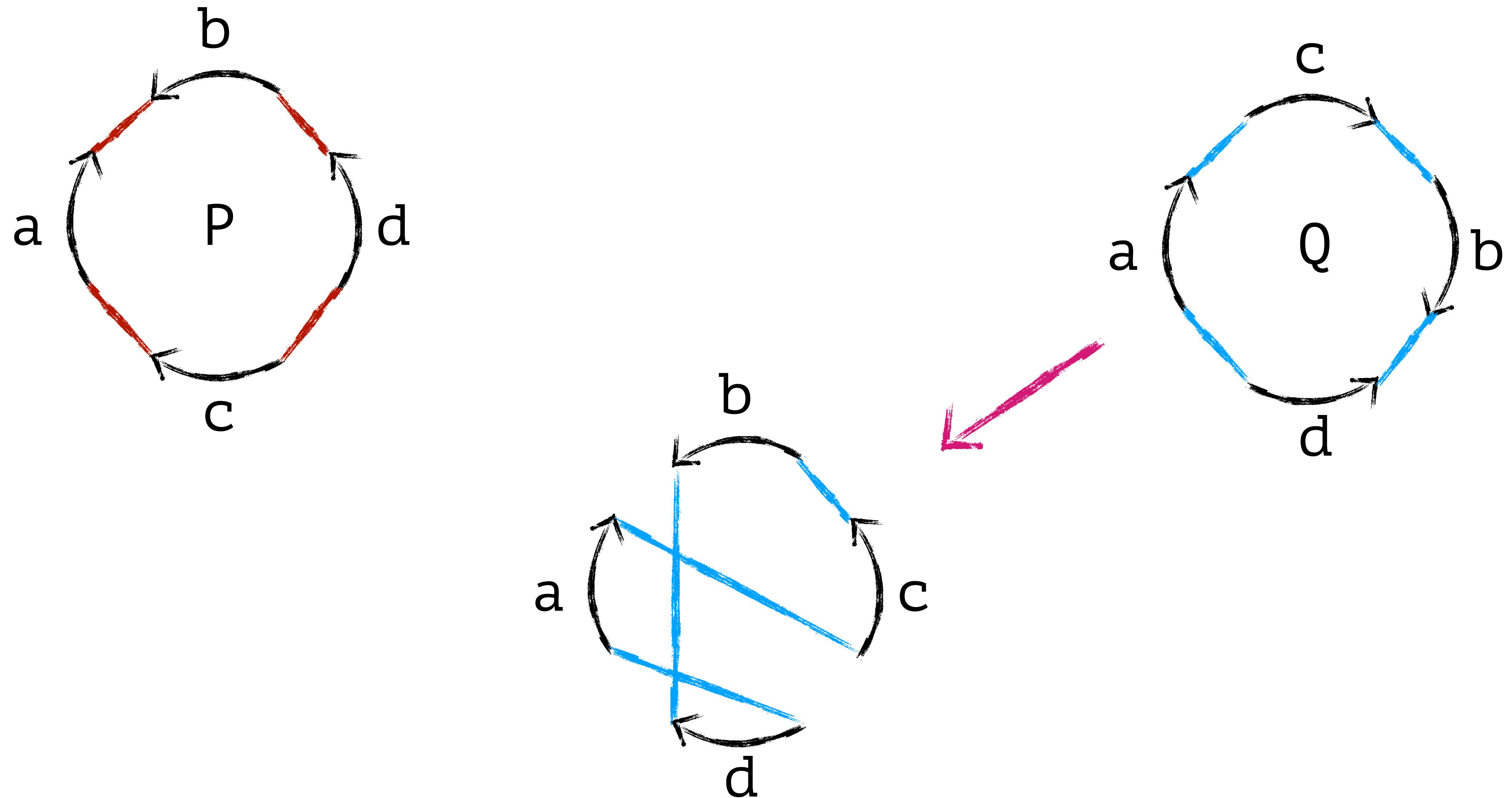
Что такое разворот участка генома в такой интерпретации?
Это удаление 2 красных ребер и добавление 2 новых! (2-breaks)

2-break distance – минимальное количество 2-breaks чтобы превратить один геном в другой.

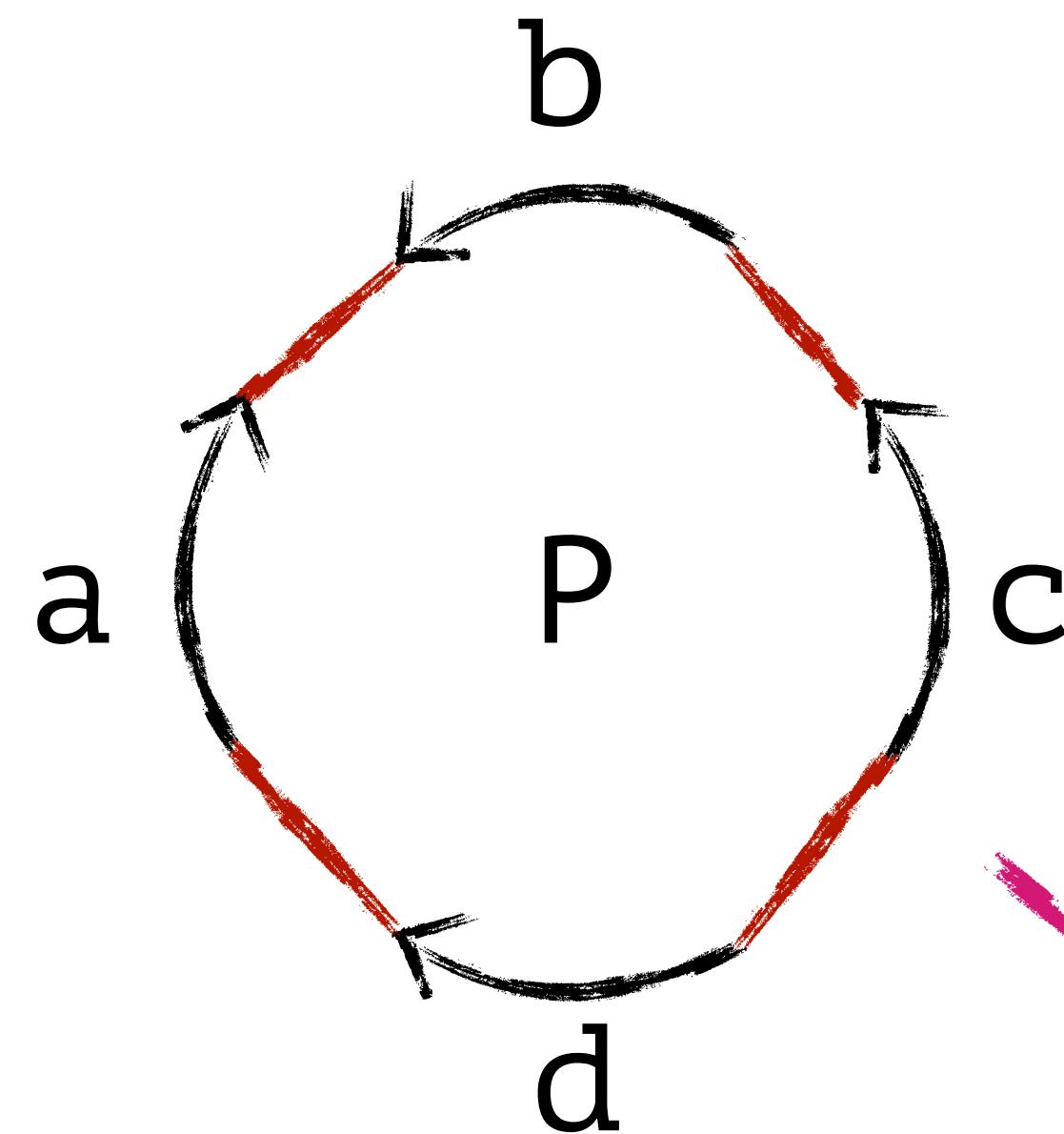
Breakpoint graph



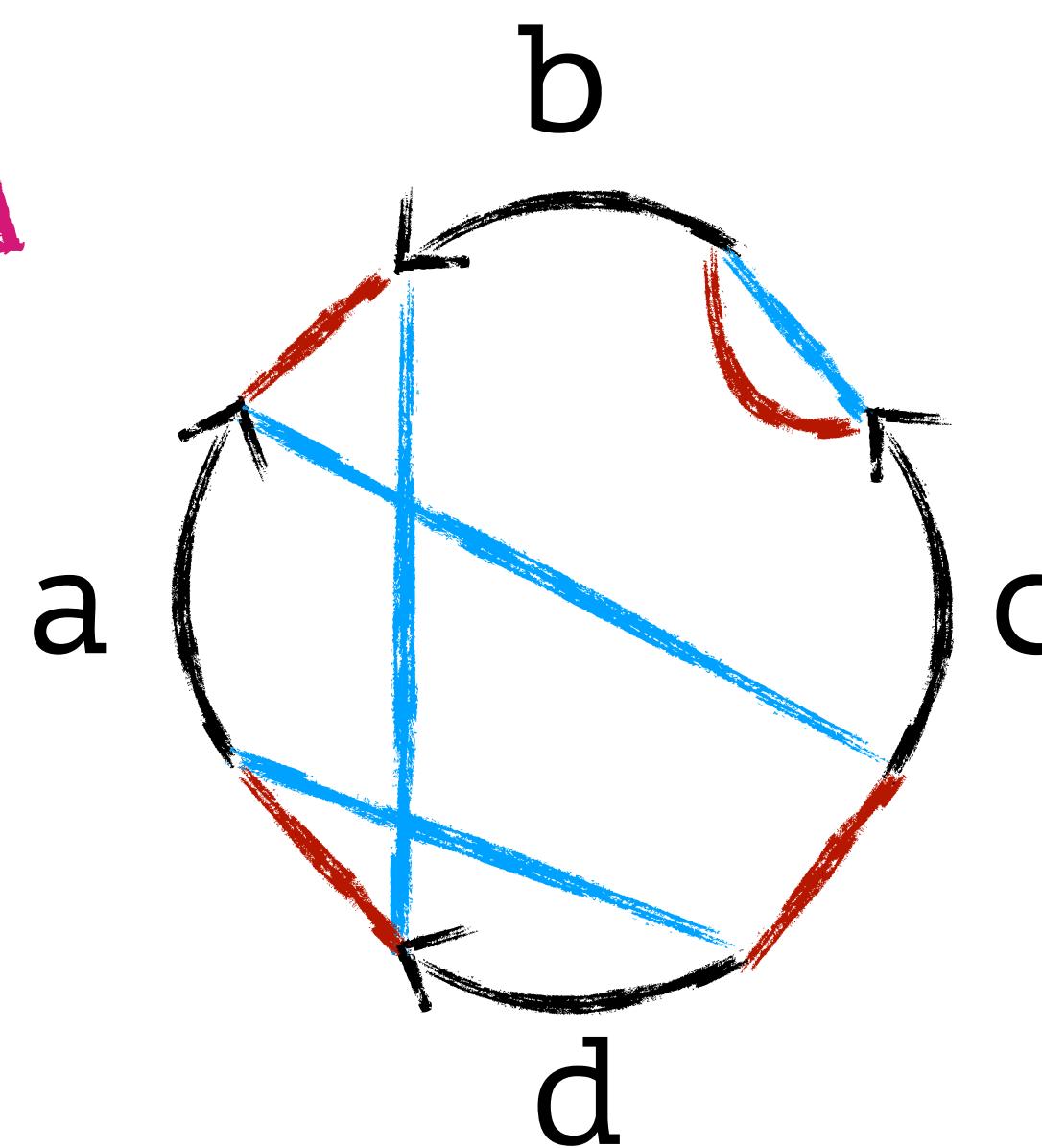
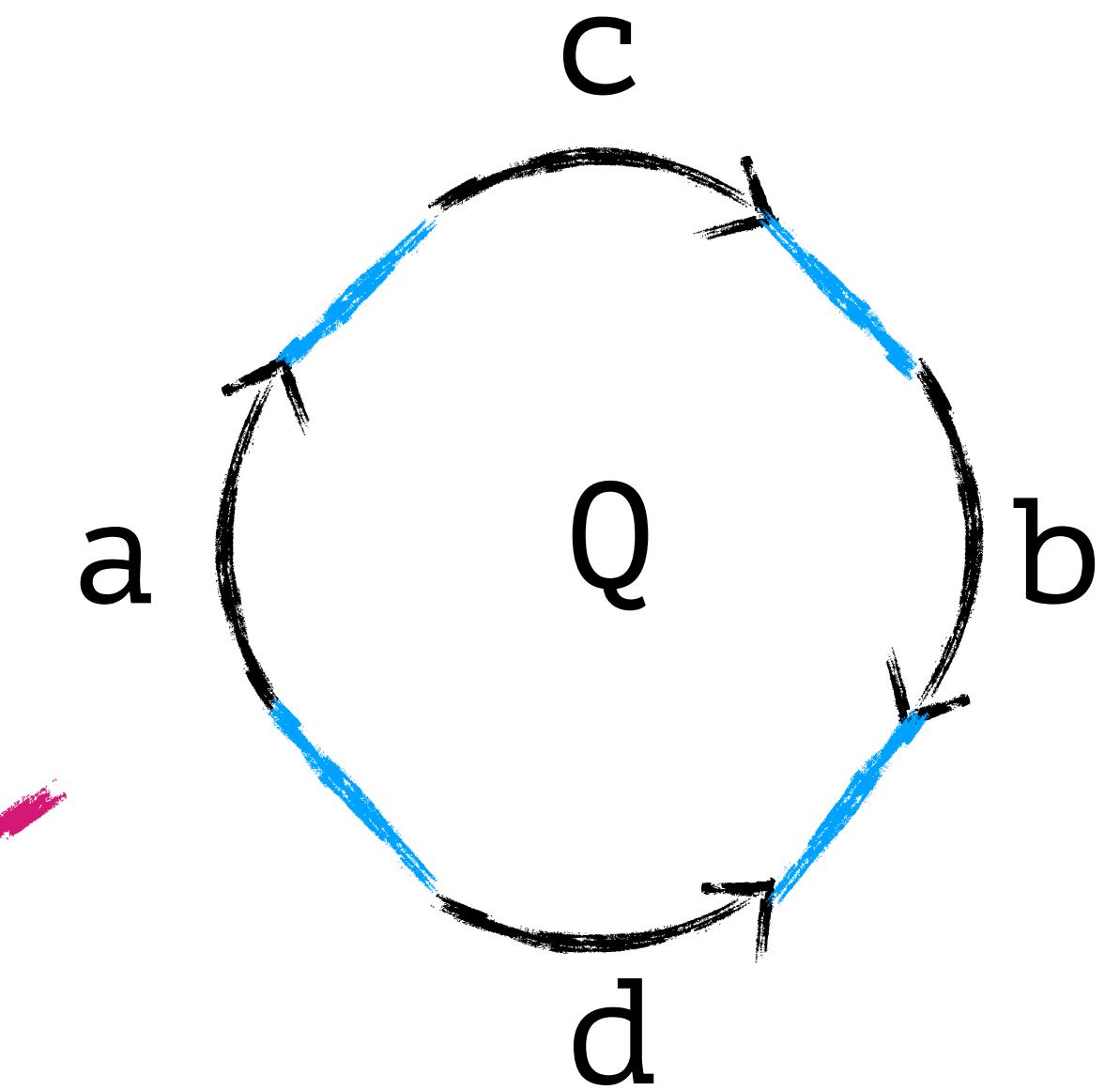
Breakpoint graph



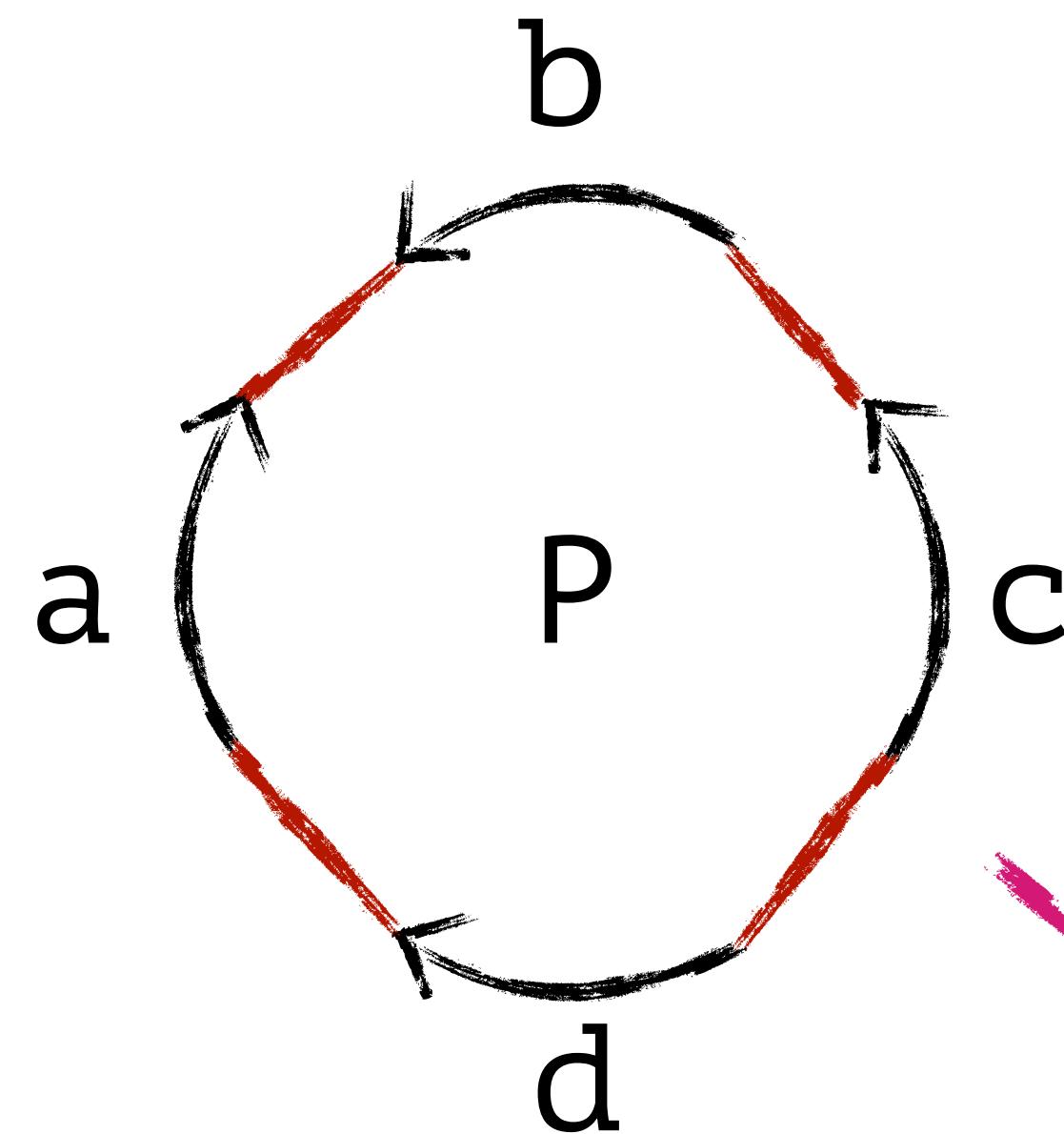
Breakpoint graph



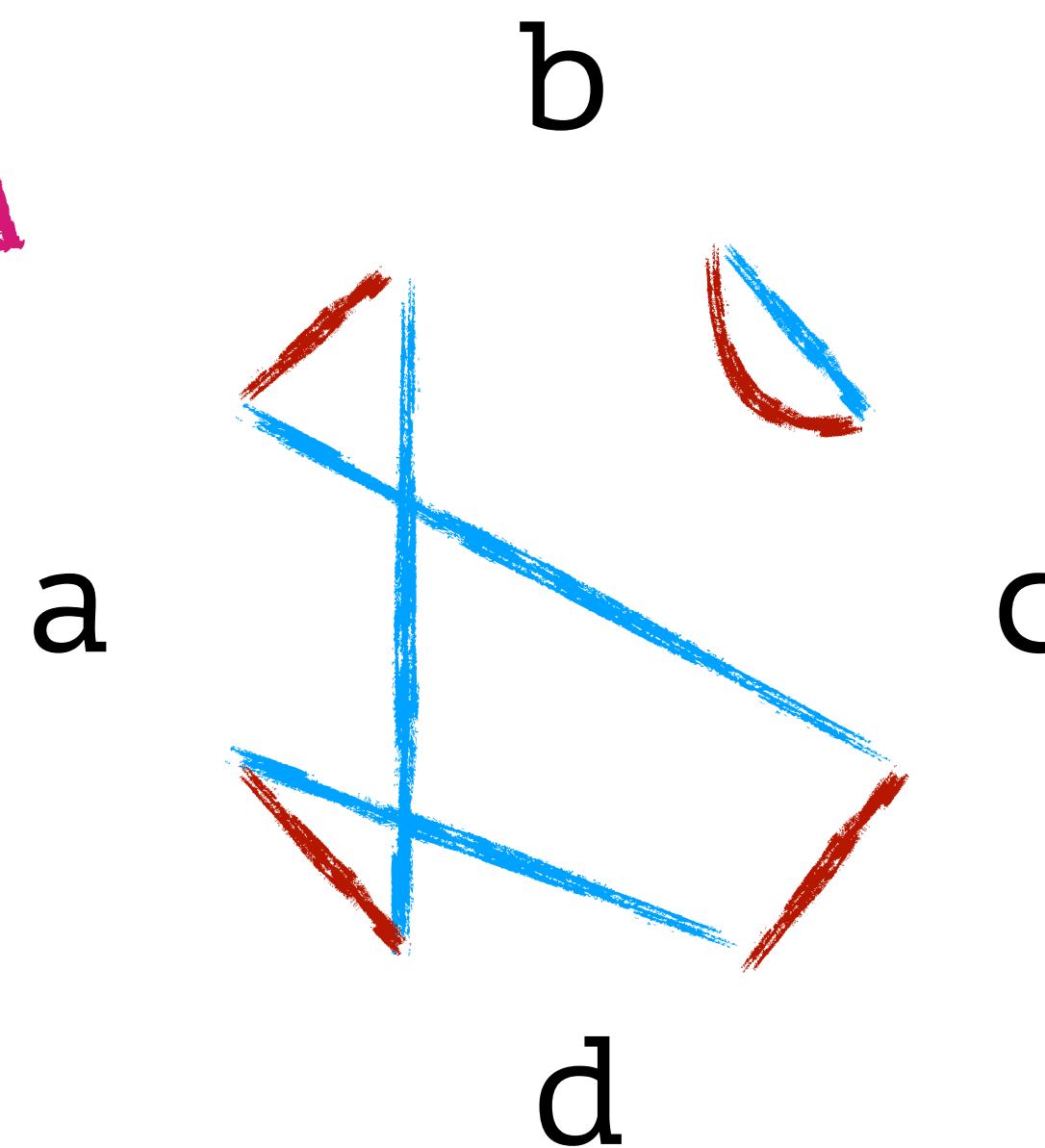
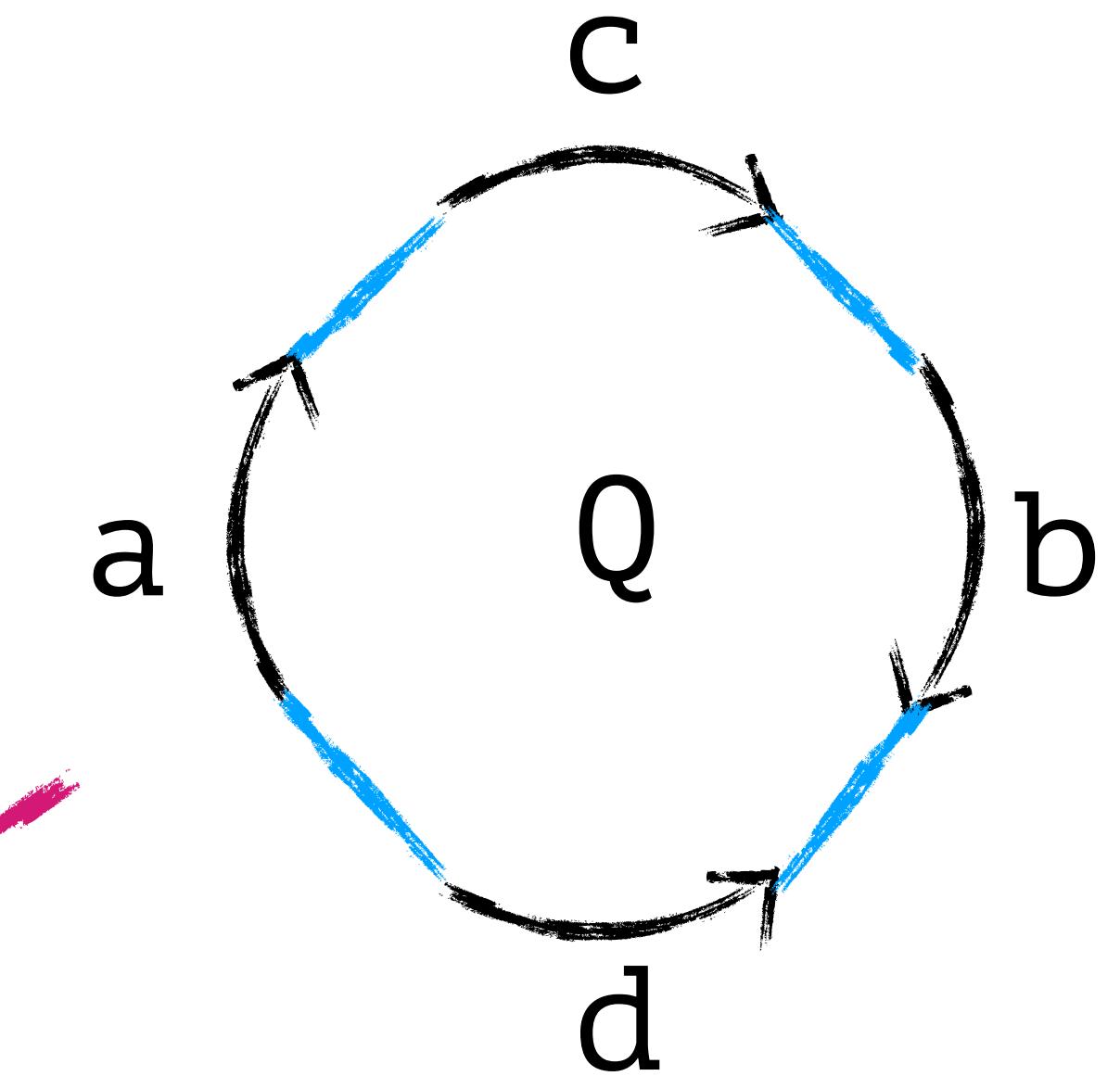
Breakpoint
Graph(**P**, **Q**)



Breakpoint graph



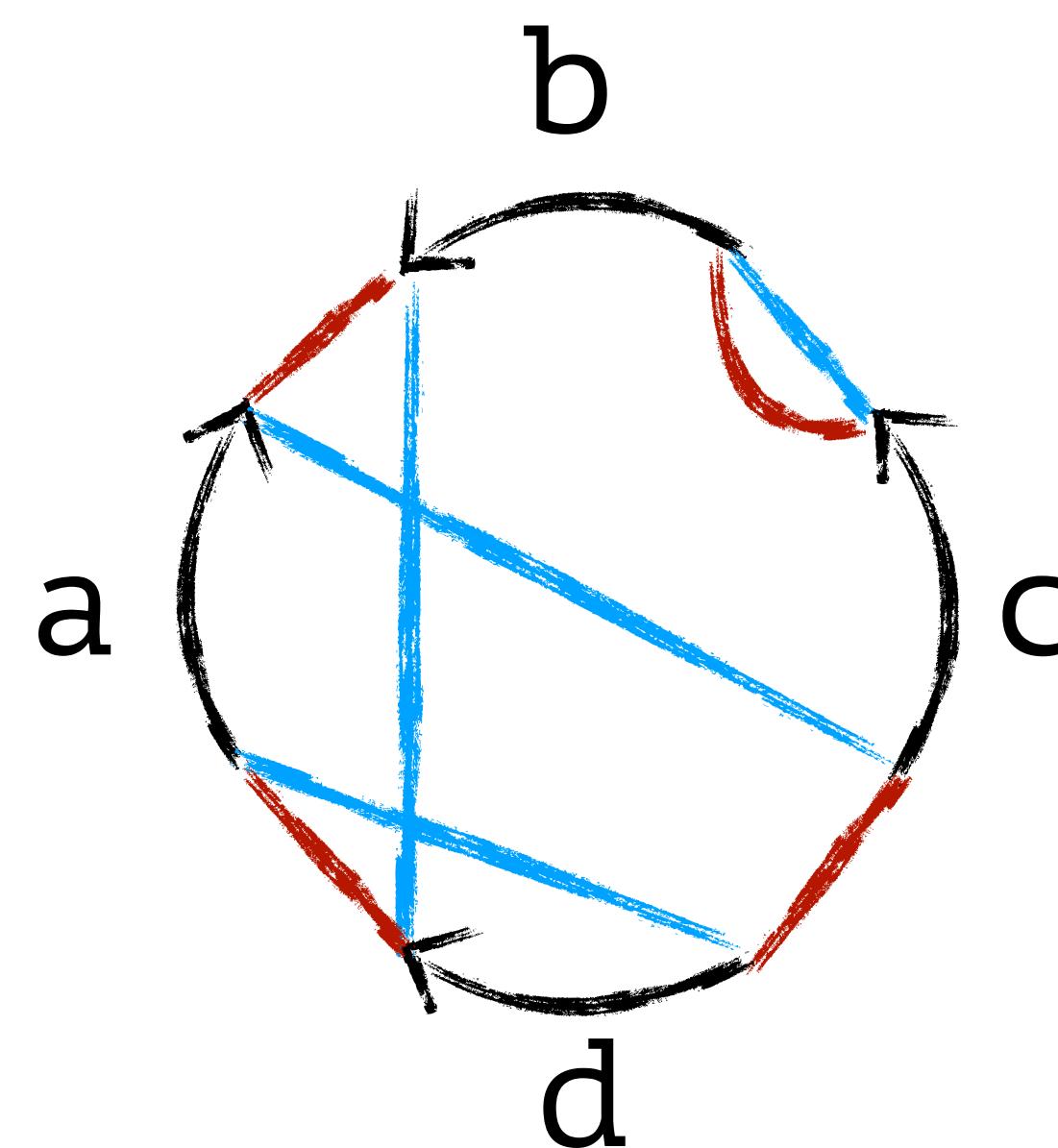
Breakpoint
Graph(**P**, **Q**)



красно синих
циклов = 2

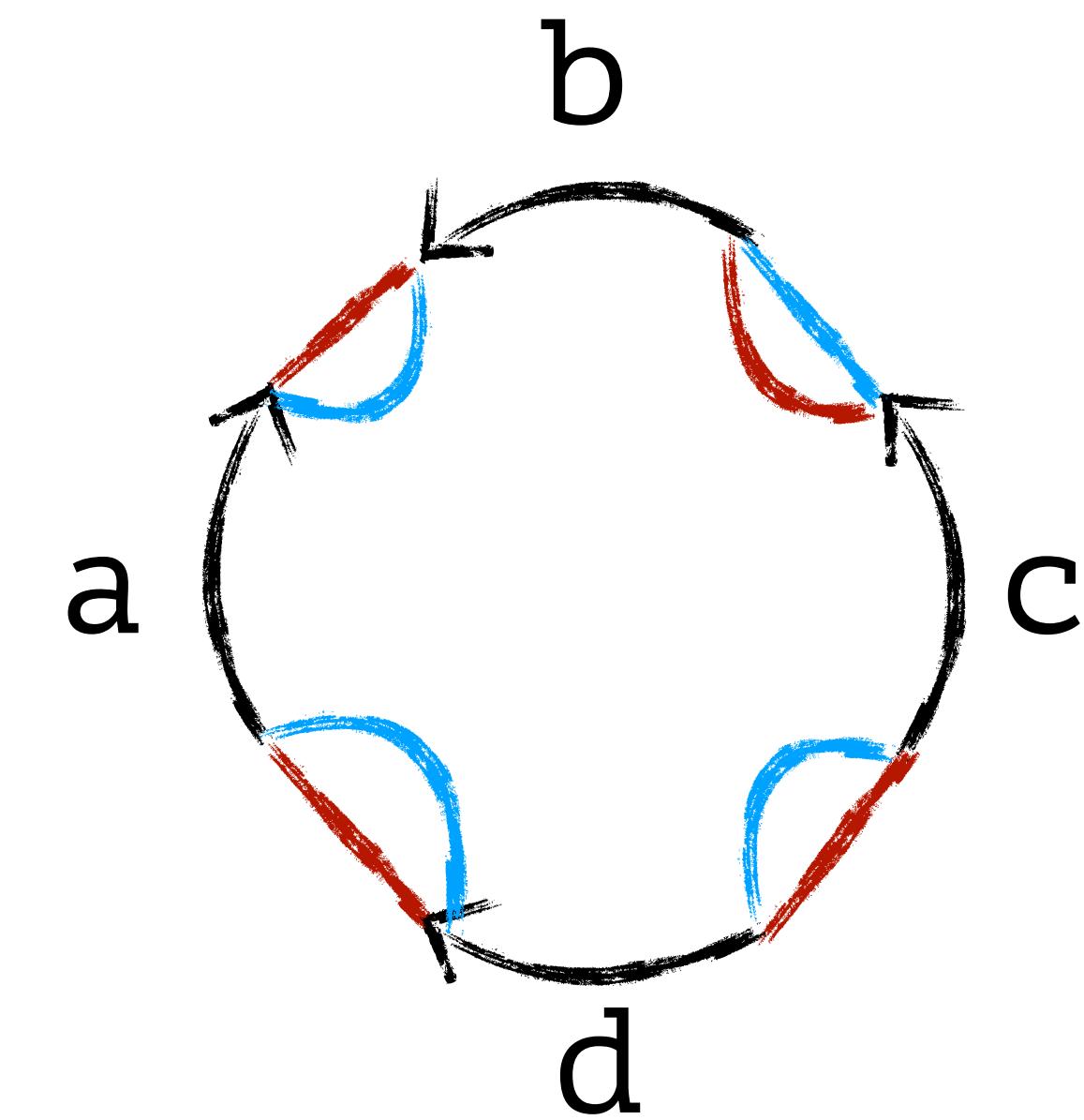
Breakpoint graph

Breakpoint
Graph(P, Q)



2-breaks

Breakpoint
Graph(P, P)



красно синих
циклов = 2

красно синих
циклов = 4

2-Break Distance Theorem

2-breaks

$$P \rightarrow \dots \rightarrow Q$$

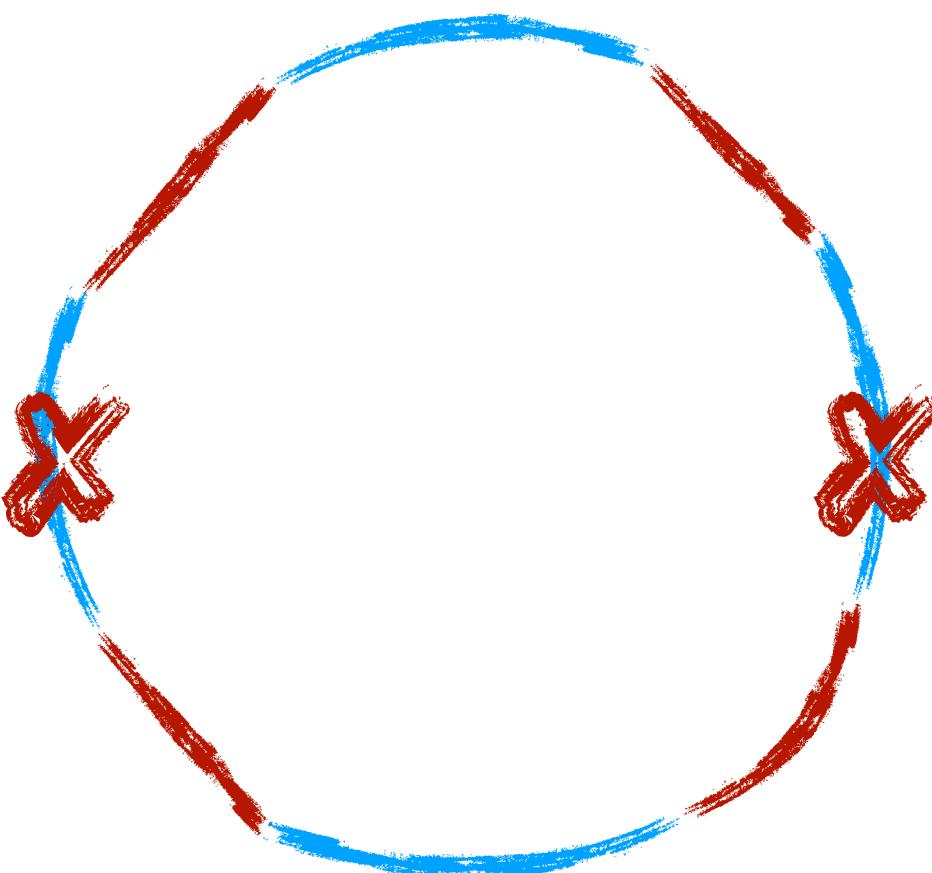
$$\text{BreakpointGraph}(P, Q) \rightarrow \dots \rightarrow \text{BreakpointGraph}(Q, Q)$$

$$\text{cycles}(P, Q) \rightarrow \dots \rightarrow \text{cycles}(Q, Q) = \text{blocks}(Q, Q)$$

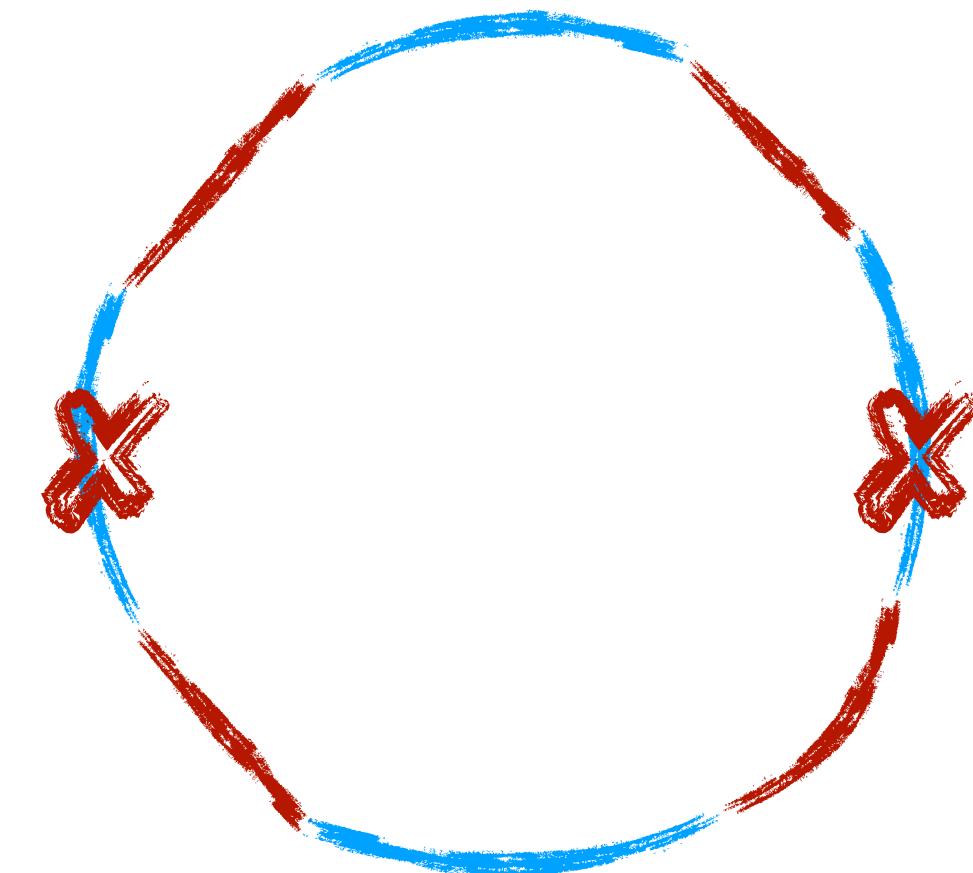
Увеличиваем число красно синих циклов на $\text{blocks}(Q, Q) - \text{cycles}(P, Q)$

Сколько нам понадобится операций 2-breaks?

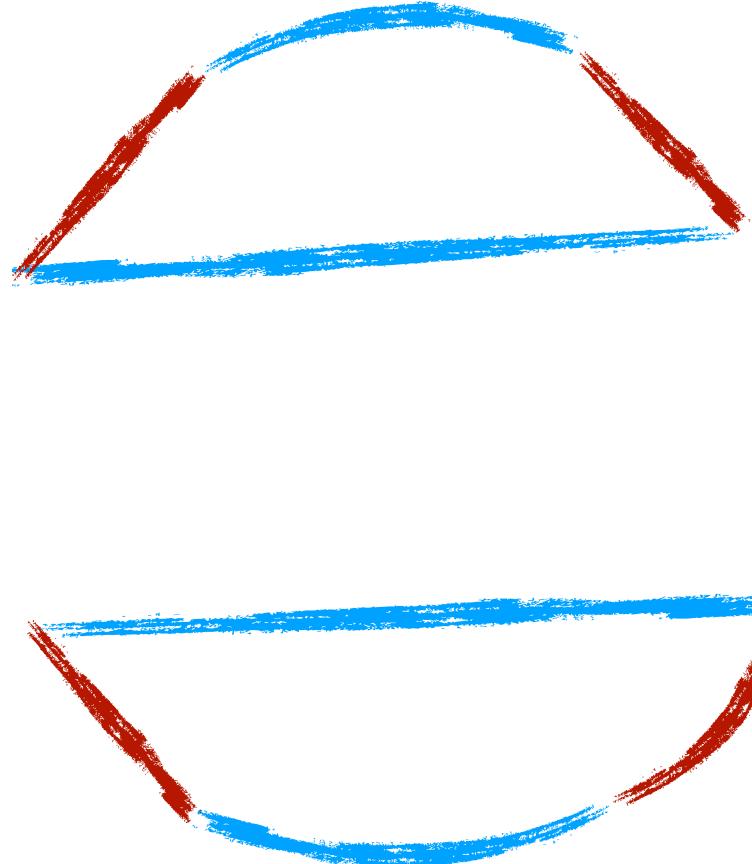
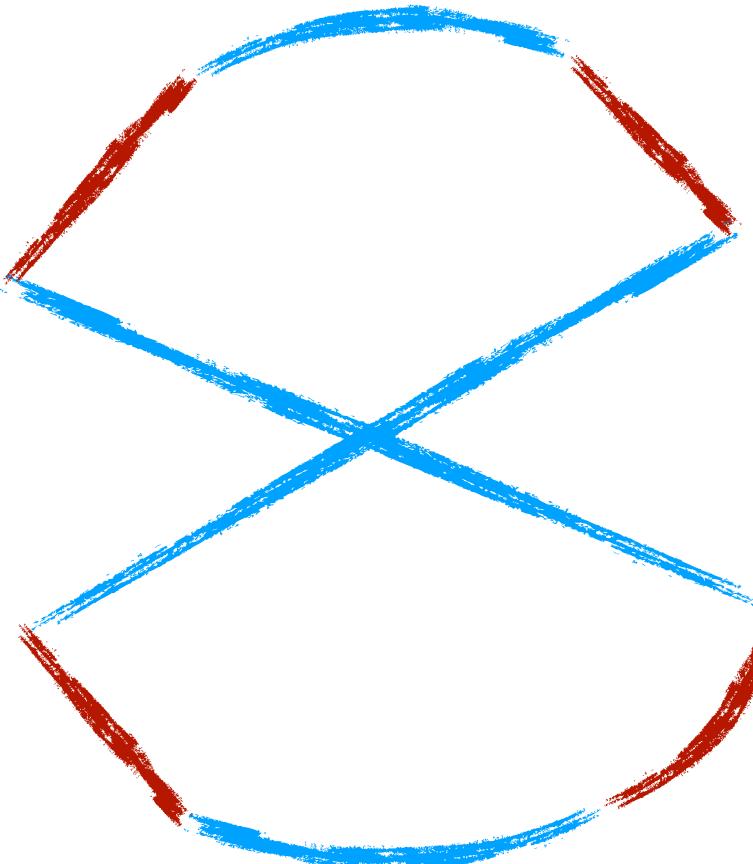
2-Break Distance Theorem



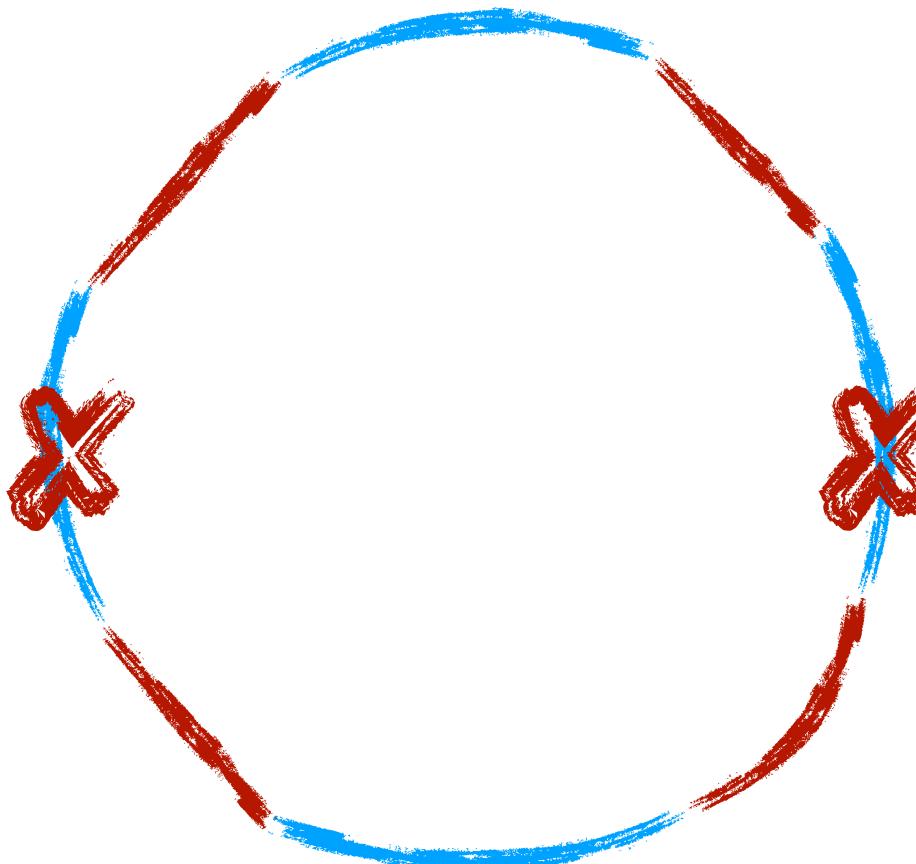
не меняется



увеличивается на 1



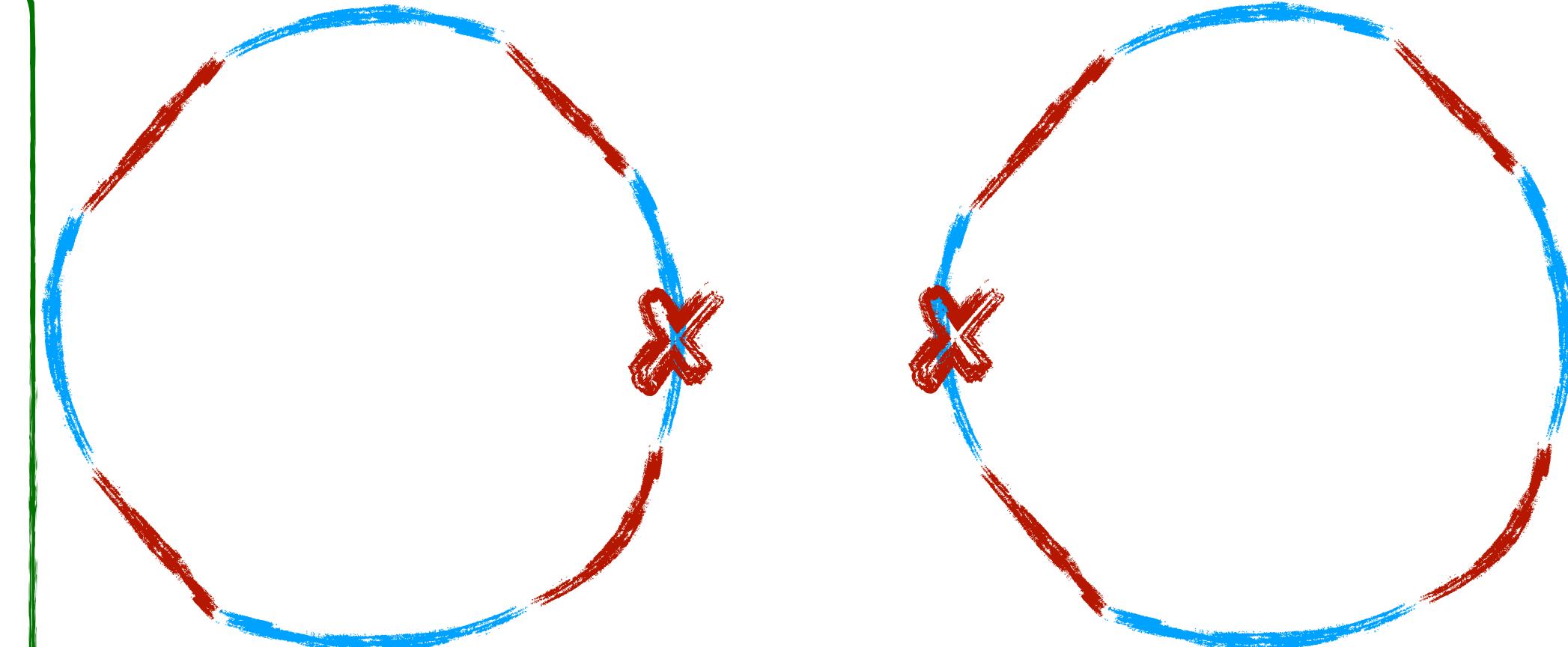
2-Break Distance Theorem



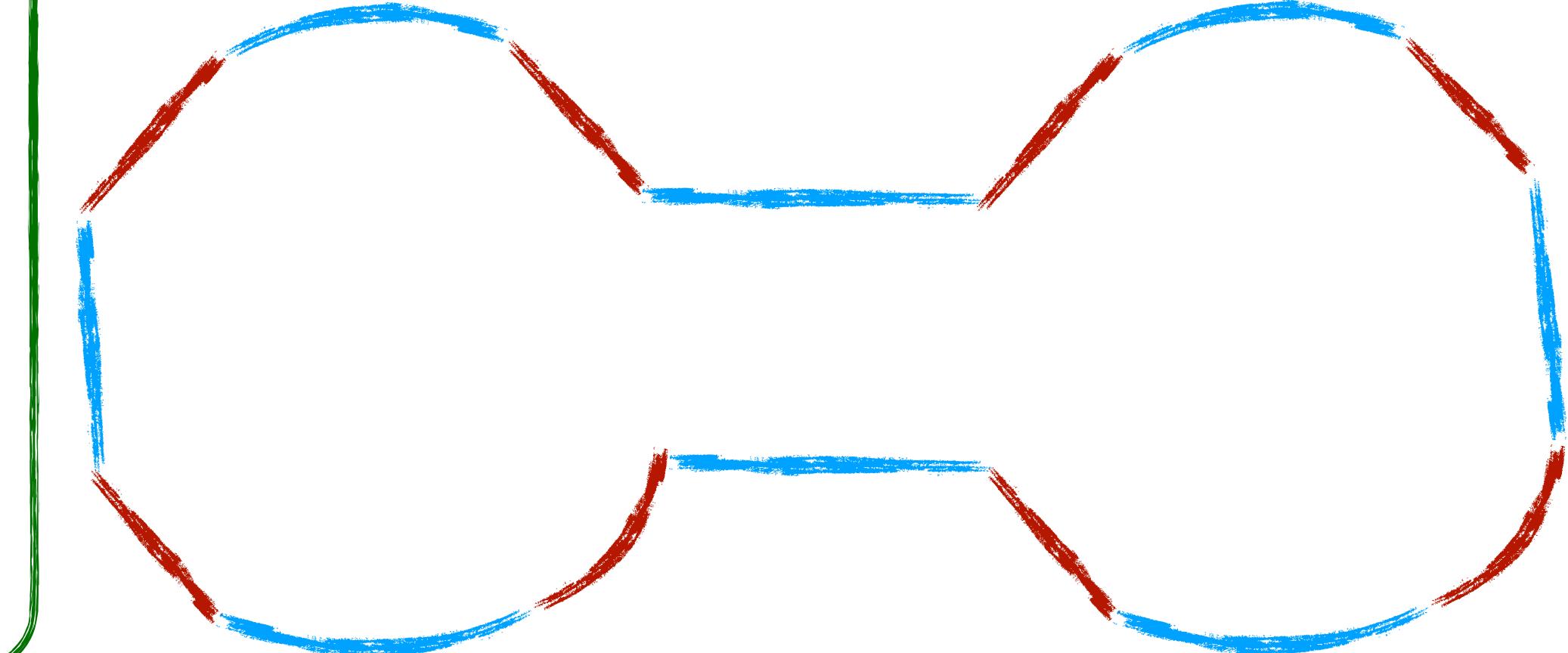
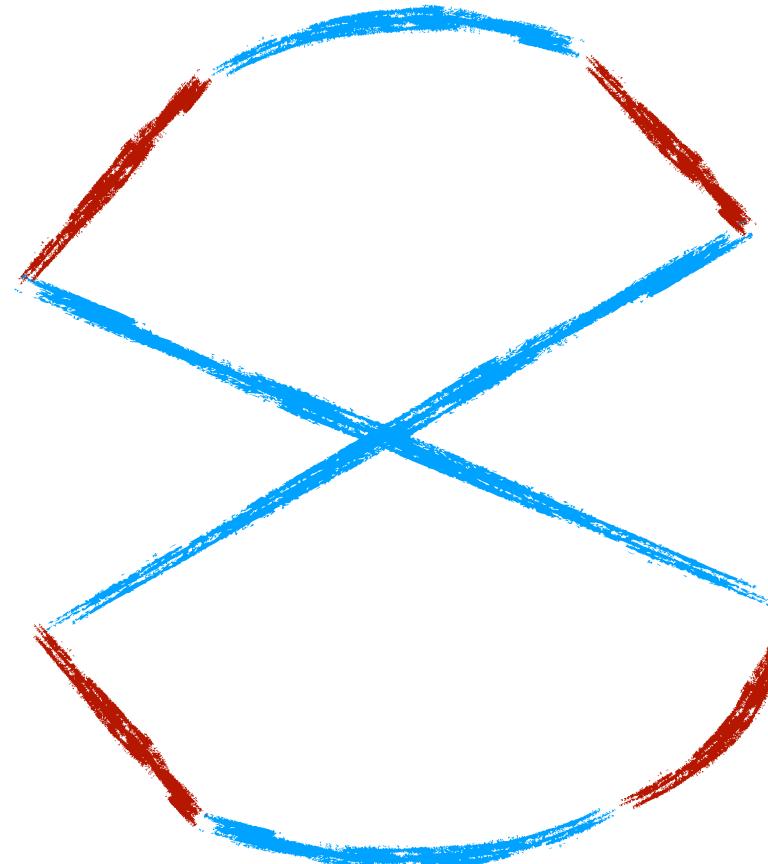
не меняется



увеличивается на 1



уменьшается на 1



2-Break Distance Theorem

Количество красно синих циклов изменяется на 1 или меньше!

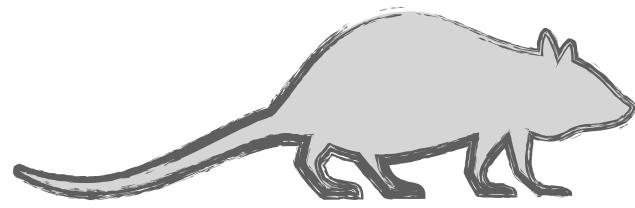
Причем существует такая операция, что увеличивается на 1

Так как при помощи операций 2-breaks нужно увеличить число циклов на $blocs(Q, Q) - cycles(P, Q)$, и мы можем совершив ровно столько операций, прервать P в Q, то:

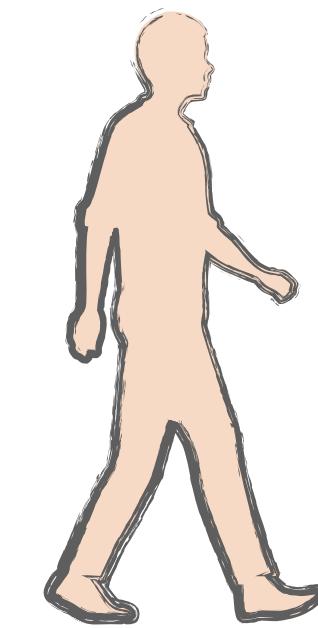
$$2BreakDistance(P, Q) = blocs(Q, Q) - cycles(P, Q)$$

2-Break Distance Theorem

M-геном



H-геном



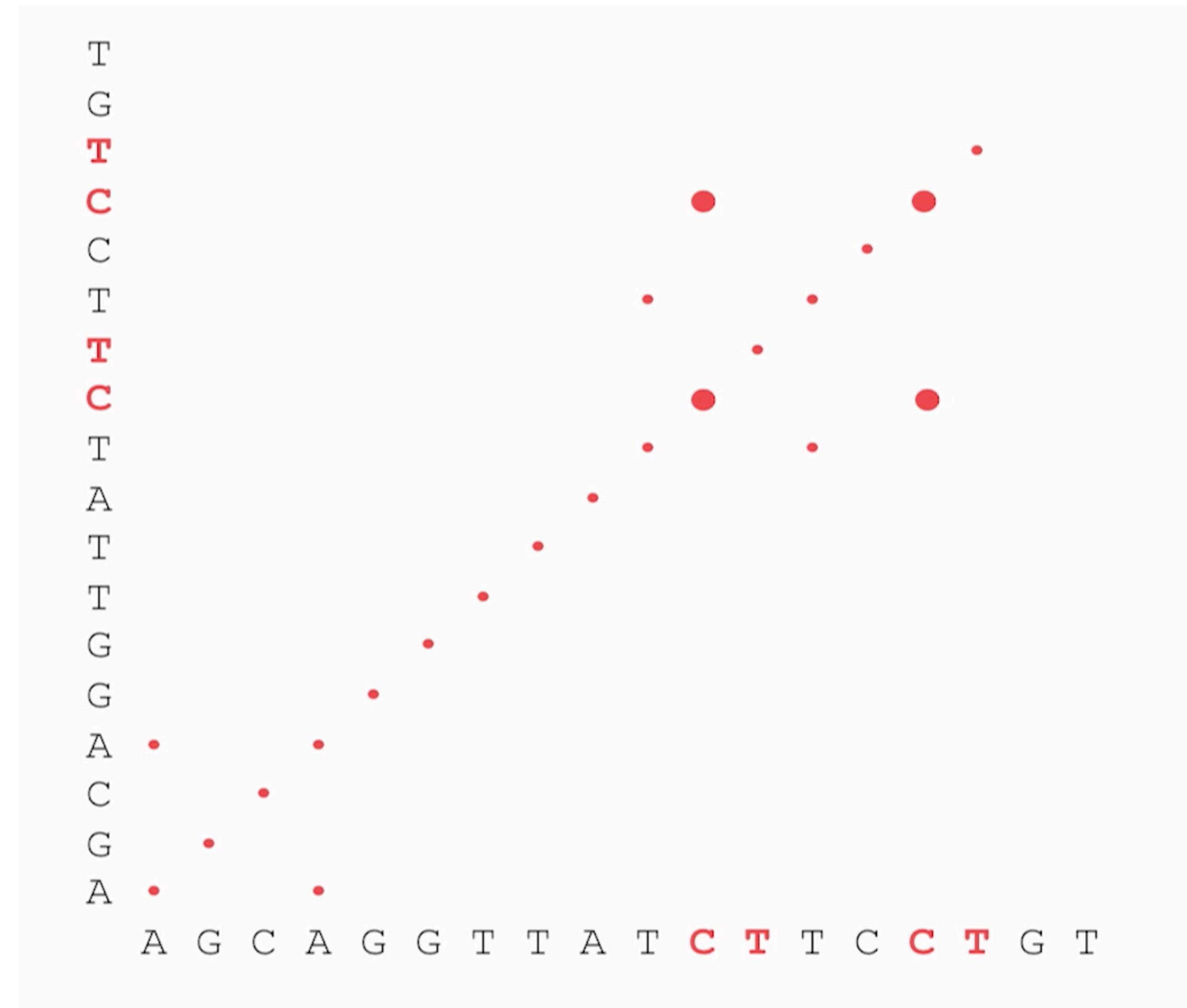
$\text{cycles}(M, H) = 35$, при том что $\text{blocks}(P, Q) = 280$

$2\text{BreakDistance}(P, Q) = 245$

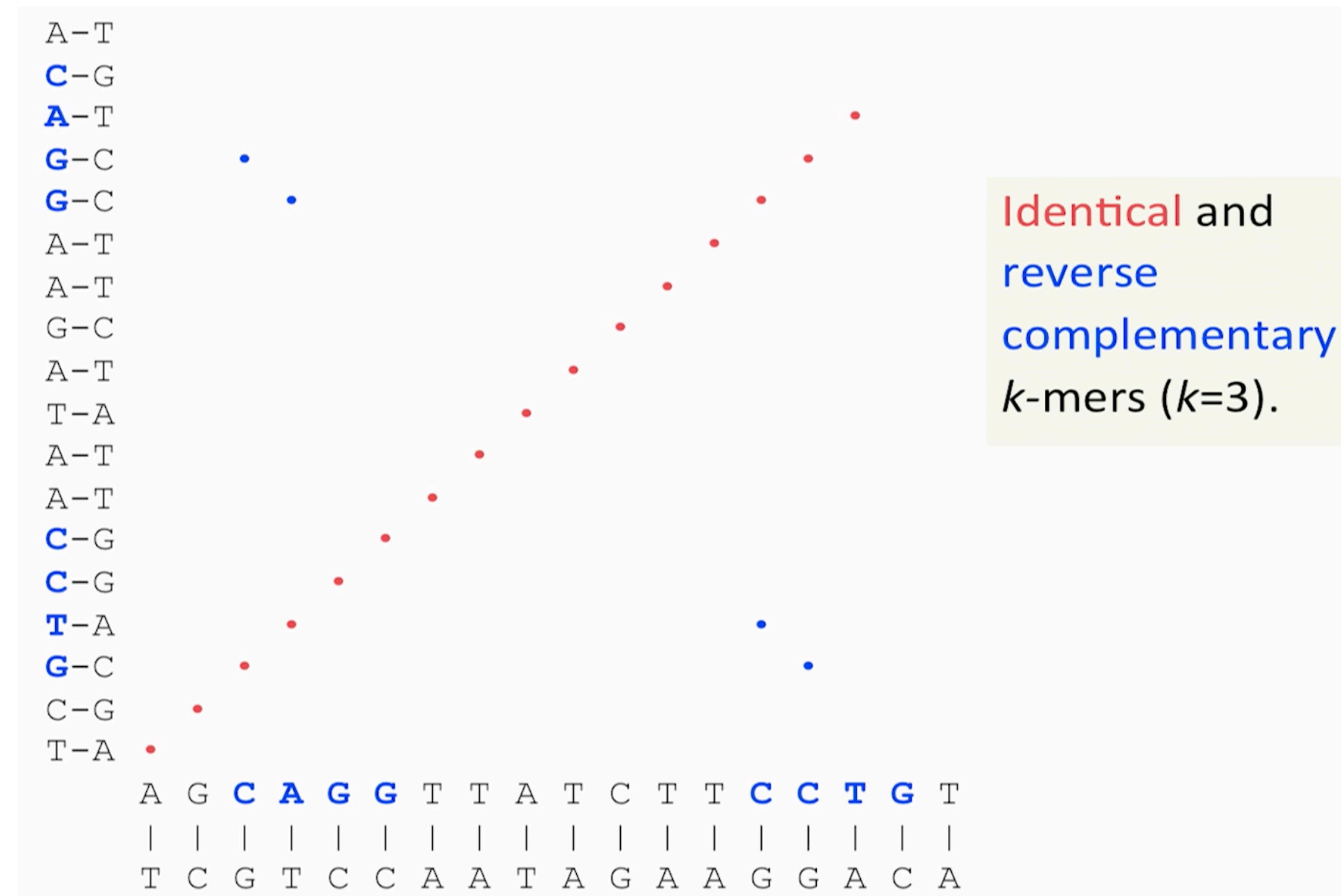
Способ искать Synteny blocks



Способ искать Synteny blocks



Способ искать Synteny blocks



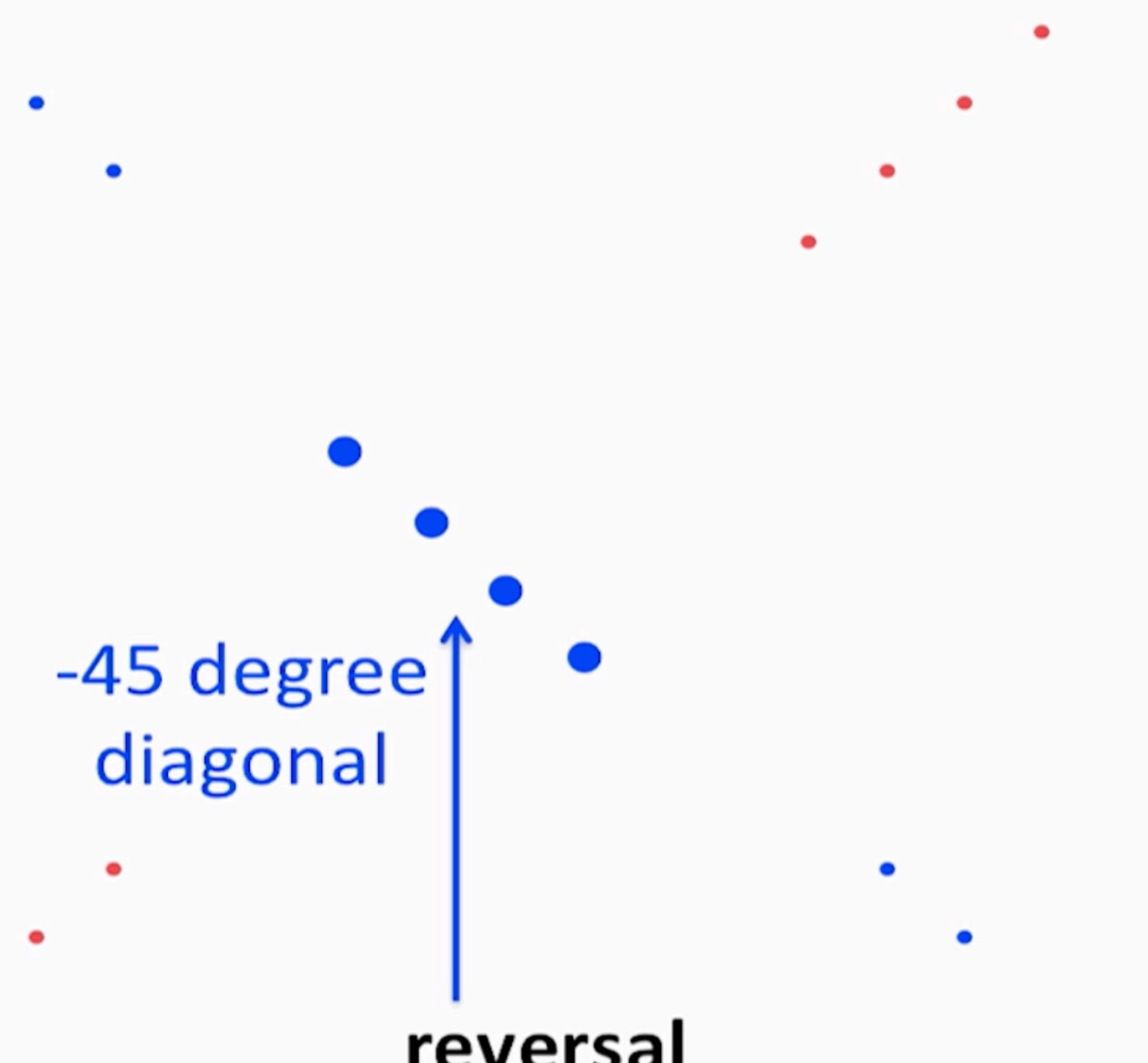
Способ искать Synteny blocks

*Genome*₂ A-T
 C-G

 differs A-T
 from G-C

*Genome*₁ T-A

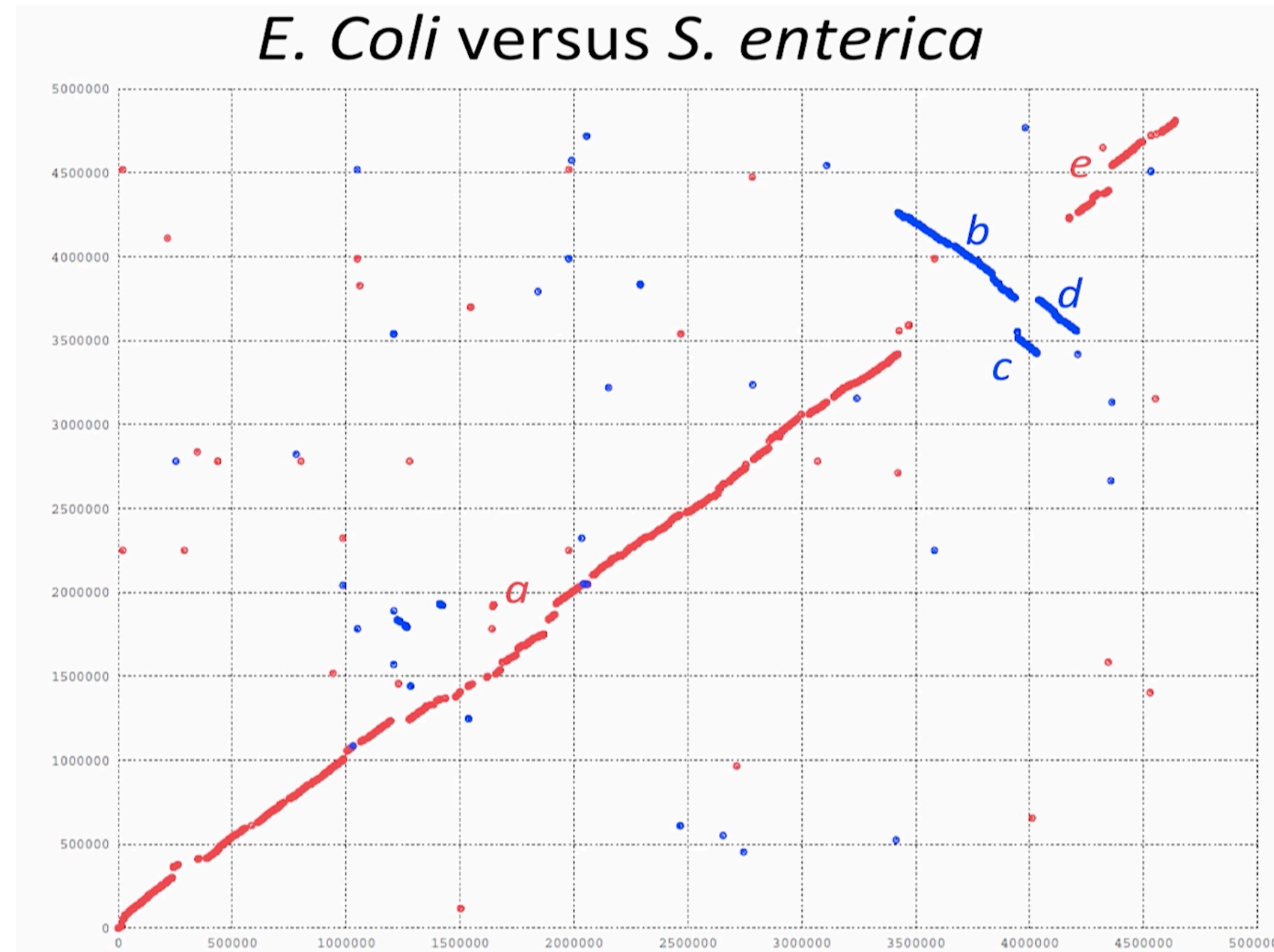
 by a T-**A**
 reversal T-**A**
 of **A**-T
 TTATCT T-**A**
 C-G



Identical and
reverse
complementary
 k -mers ($k=3$).

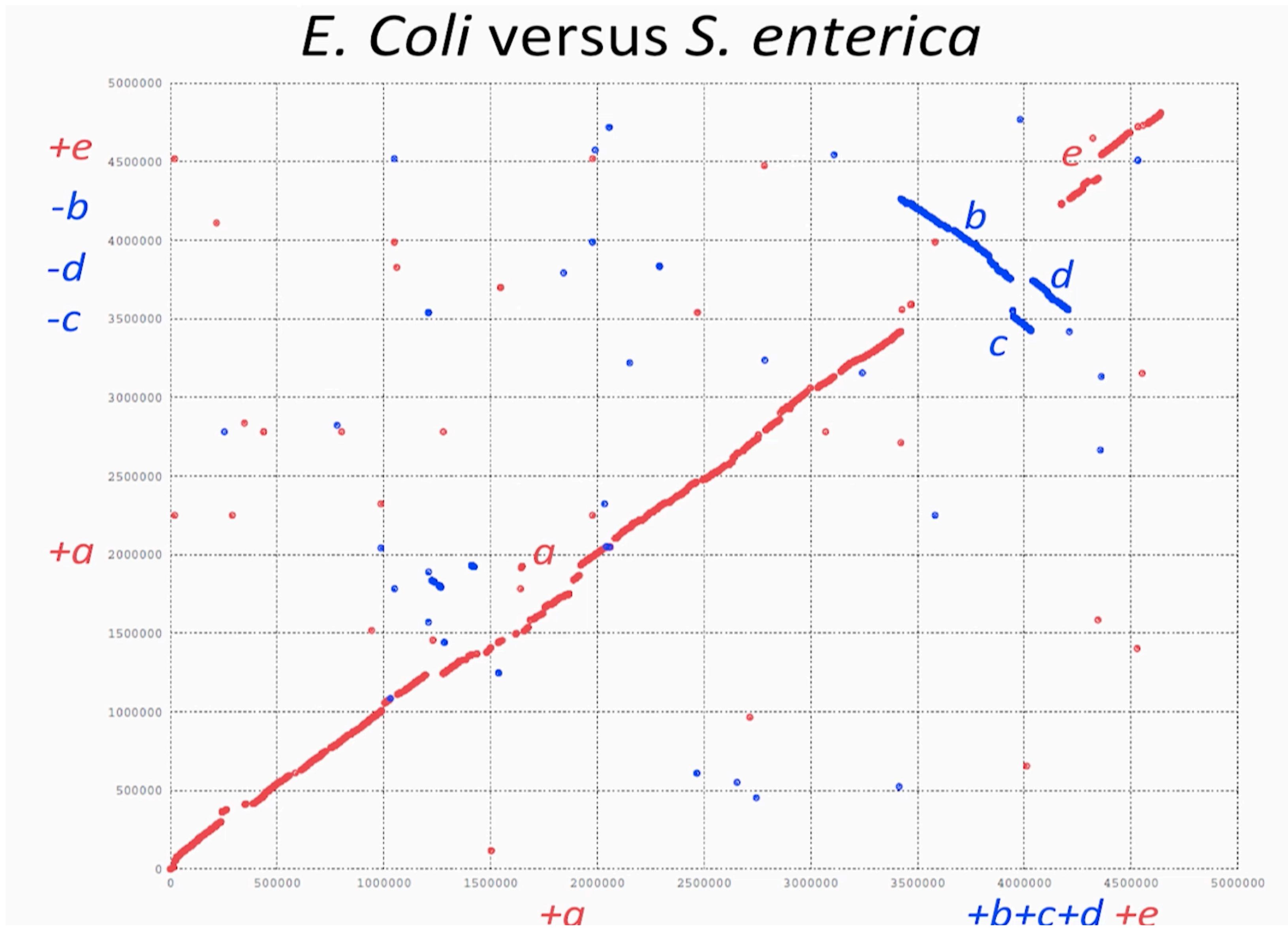
Способ искать Synteny blocks

E. Coli versus *S. enterica*

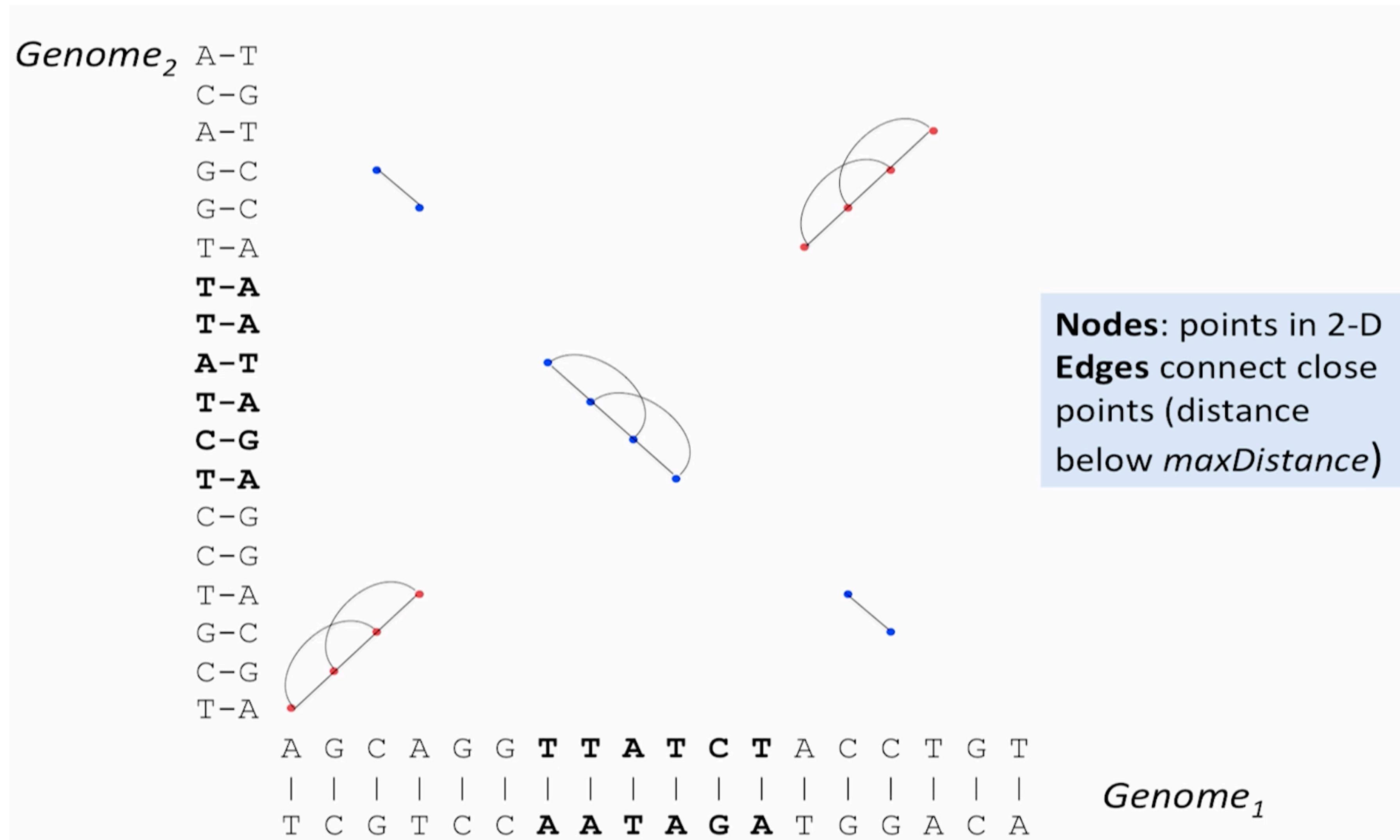


Способ искать Synteny blocks

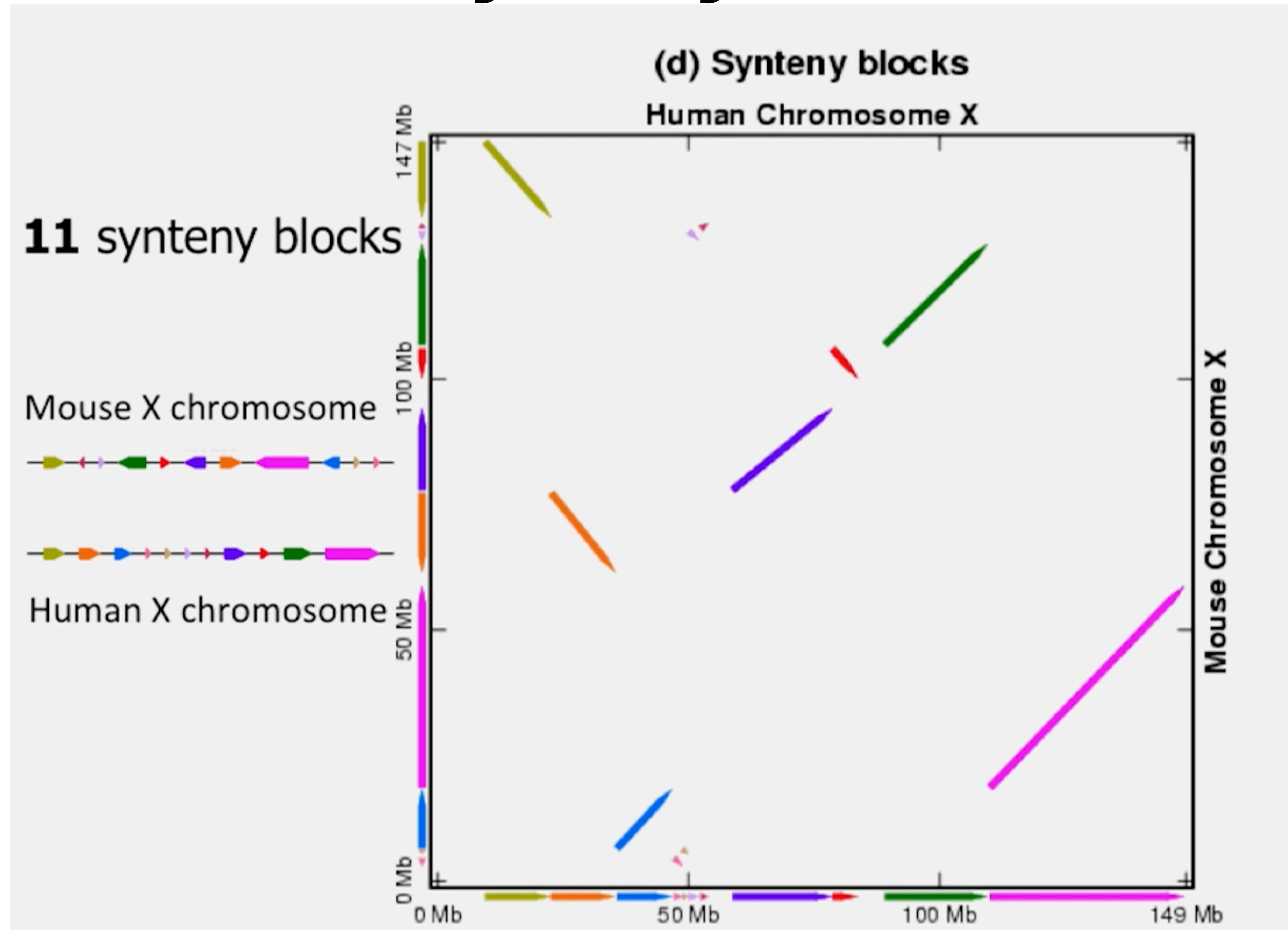
E. Coli versus *S. enterica*



Способ искать Synteny blocks



Способ искать Synteny blocks



Резюмируем

- Геномы состоят из синтентных блоков, которые получились в результате эволюционного процесса
- Минимальное количество перестроек помогает точнее понимать общую эволюционную картину
- Строя breakpoint граф, можно измерять расстояние как 2BreakDistance
- Есть эвристические способы искать синенные блоки сравнивая 2 генома