

Bioinformatics I

Sequence Analysis

Sequence Analysis

Prerequisites

- Comp Sci 1 and 2.

Can you program?

- Molecular biology.

Do you know the central dogma?

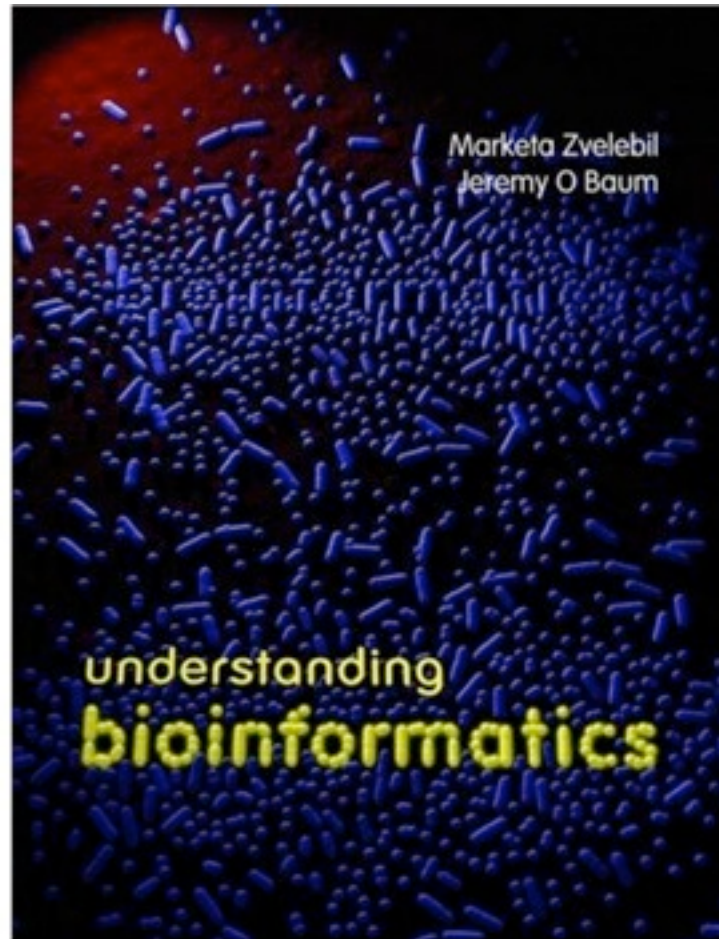
- Molecular biochemistry.

What molecular functions can you name?

What molecular functions can you name?

- Molecular biochemistry.

You need this:



Z&B

...and this.



<http://ugene.unipro.ru/>

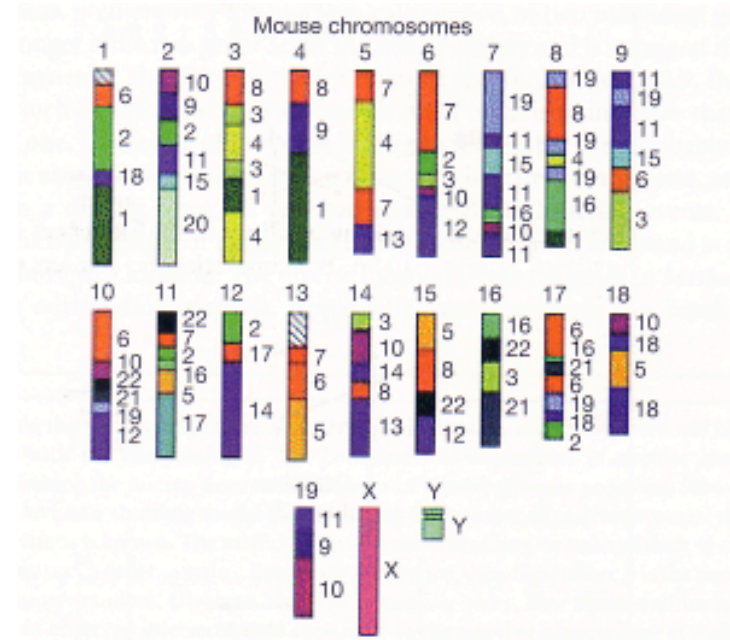
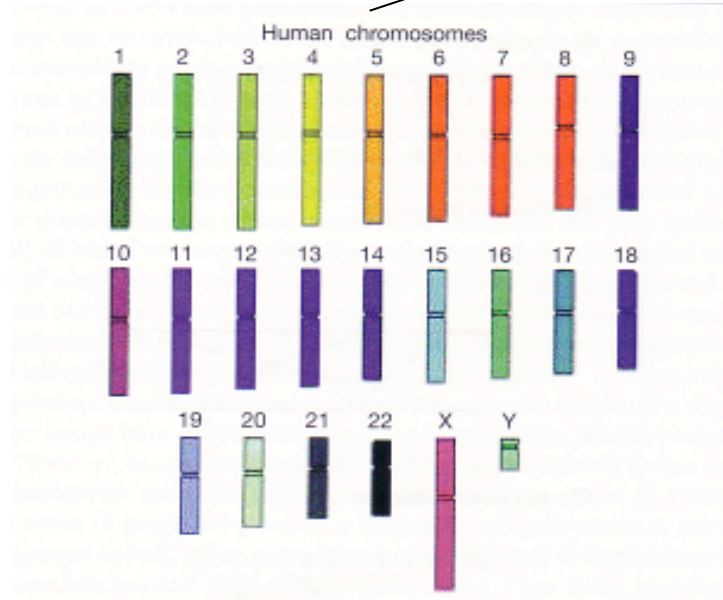
...and a 3-button mouse.

actual content starts here

*Consider how humans and mouse have diverged
since the last common ancestor...*

- Mouse and human had a common ancestor about 80 MYA.
- Evolution occurs by point **mutations, insertions, deletions** and *rearrangements*.
- Individual mouse genes and human genes are 80 to 95% identical.
- However, gene locations are **scrambled!** (Maybe gene location matters more than its sequence???)

Inter-chromosomal rearrangement



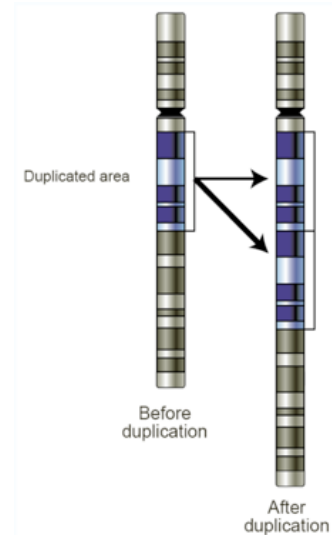
80 million years of scrambling...

NOTE: no rearrangements of X, Y chromosomes!

Large scale evolutionary changes in chromosomes

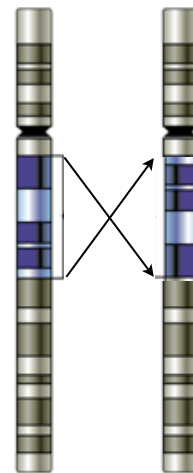
- Duplications

“...the most important evolutionary force since the emergence of the universal common ancestor.” --
Susumu Ohno



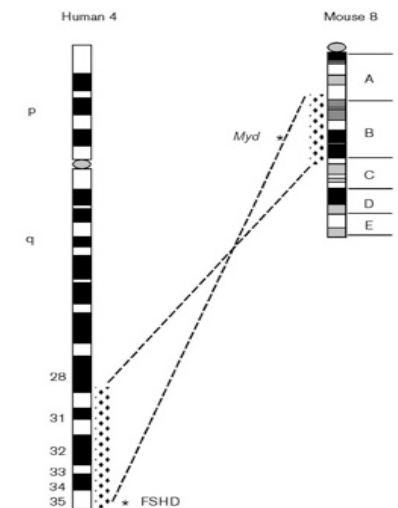
- Inversions

A syntenic group appears on the opposite strand, different location.

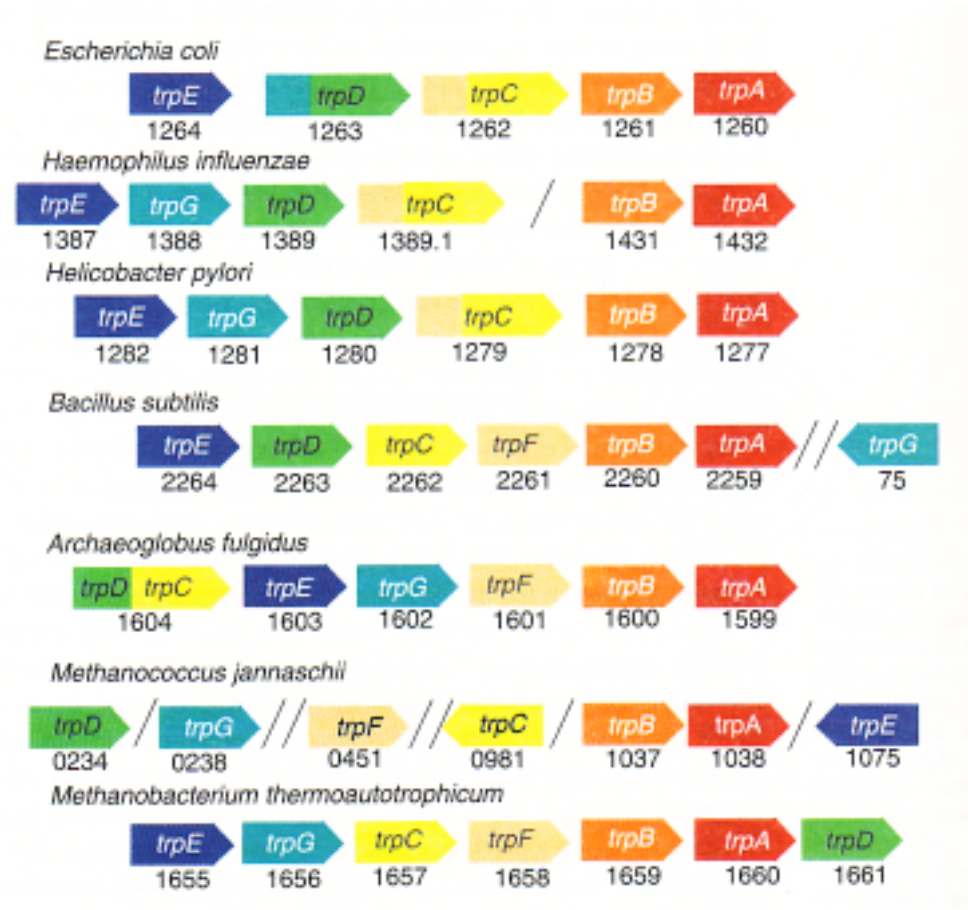


- Transpositions

A syntenic group appears on a different chromosome.

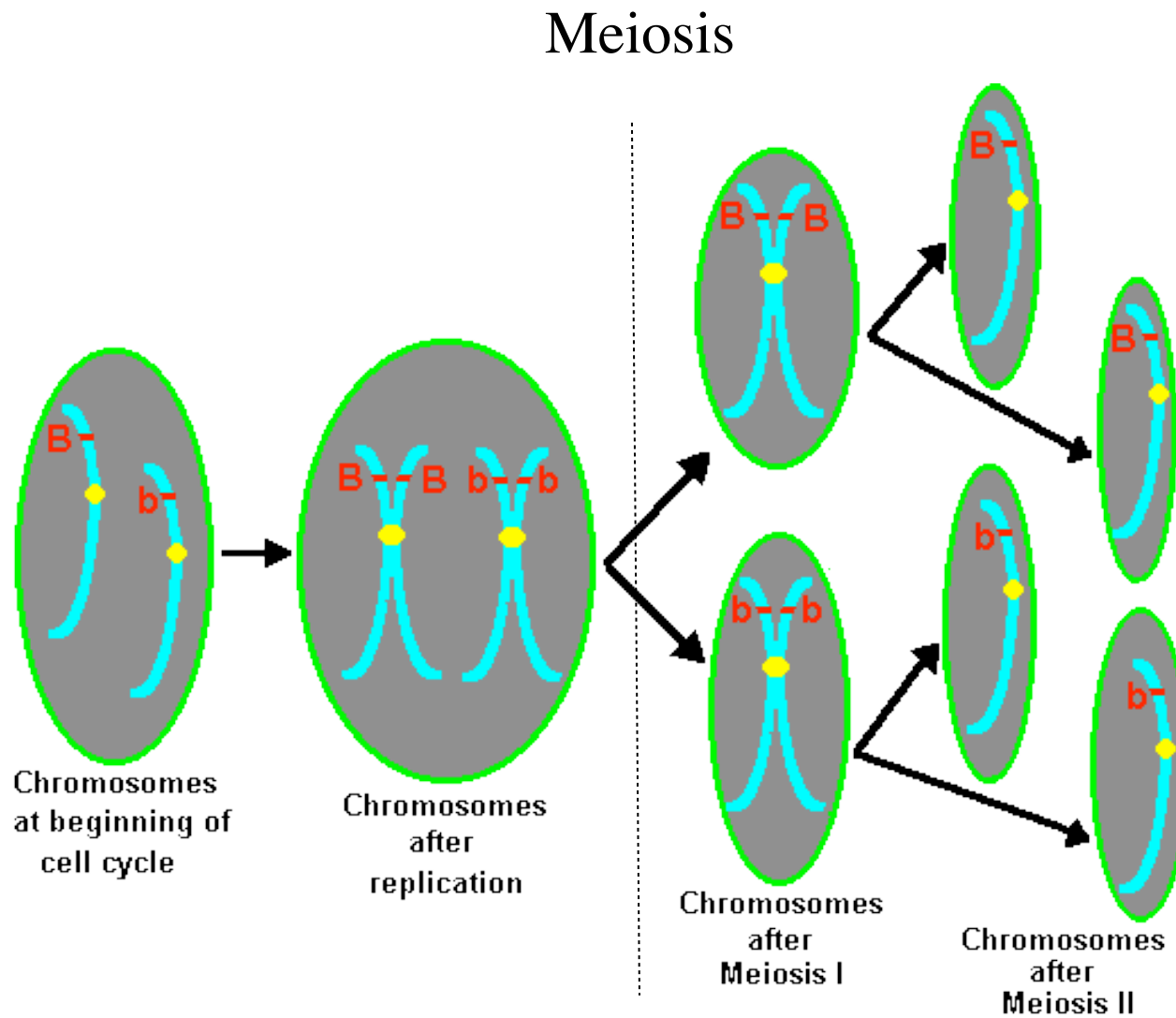


Syntenic groups in bacteria: Trp operon



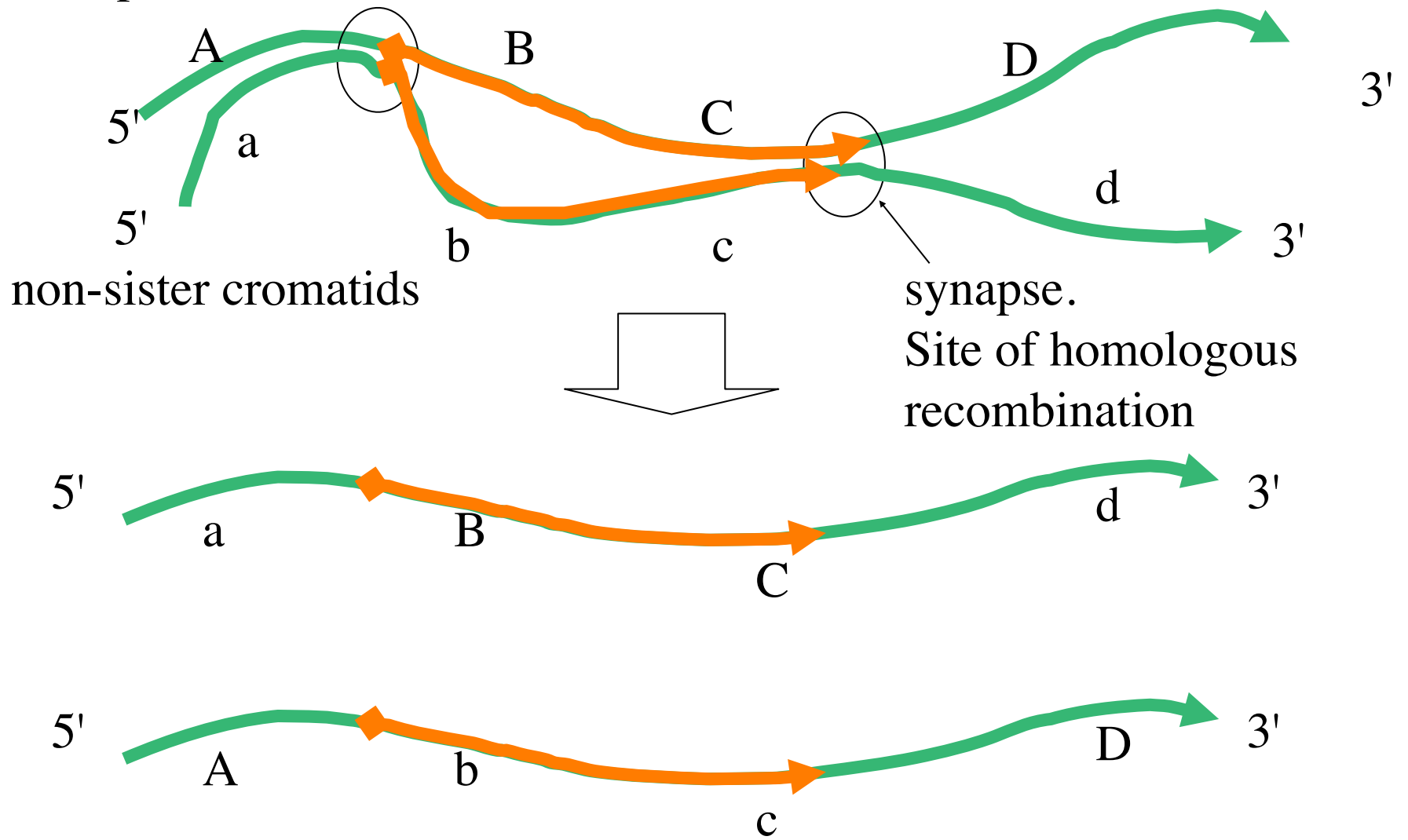
Defined as a set of genes that are always found together.

One way inversion can happen in Eukaryotes



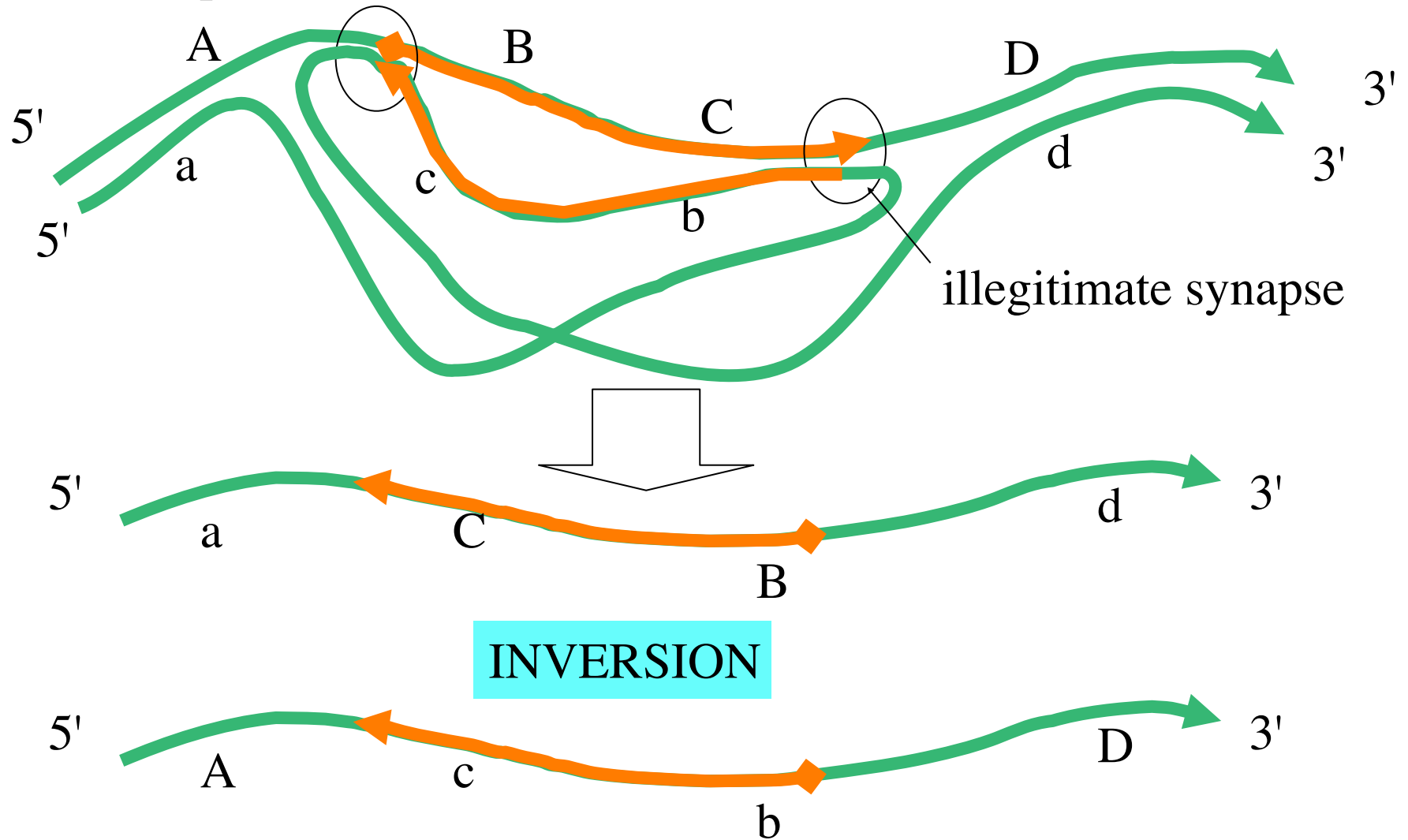
Normal prophase 1

Normal synapsis does not change the order of genes, just swaps alleles.

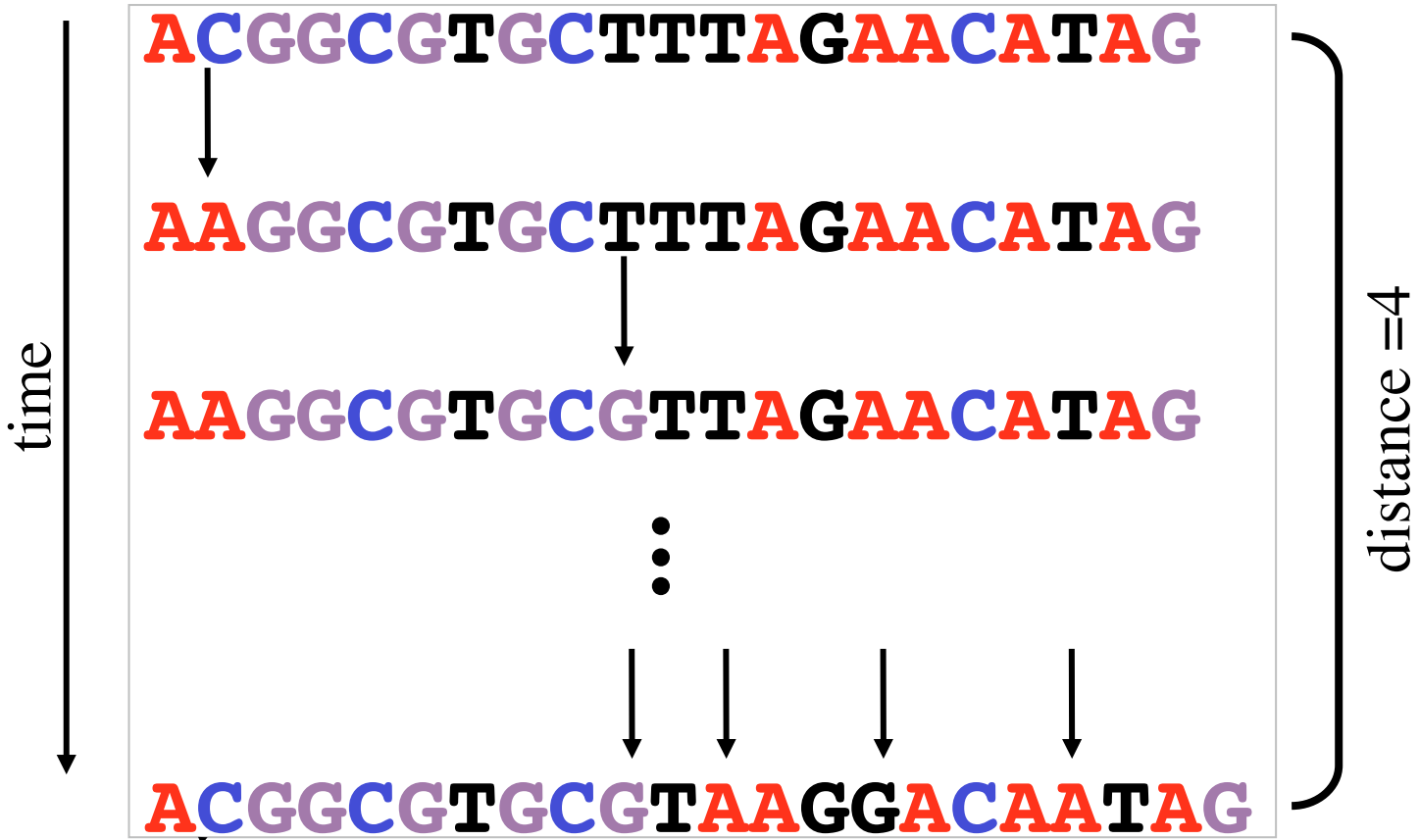


Abnormal prophase 1.

Illegitimate synapsis changes the order and direction of genes, and swaps alleles.

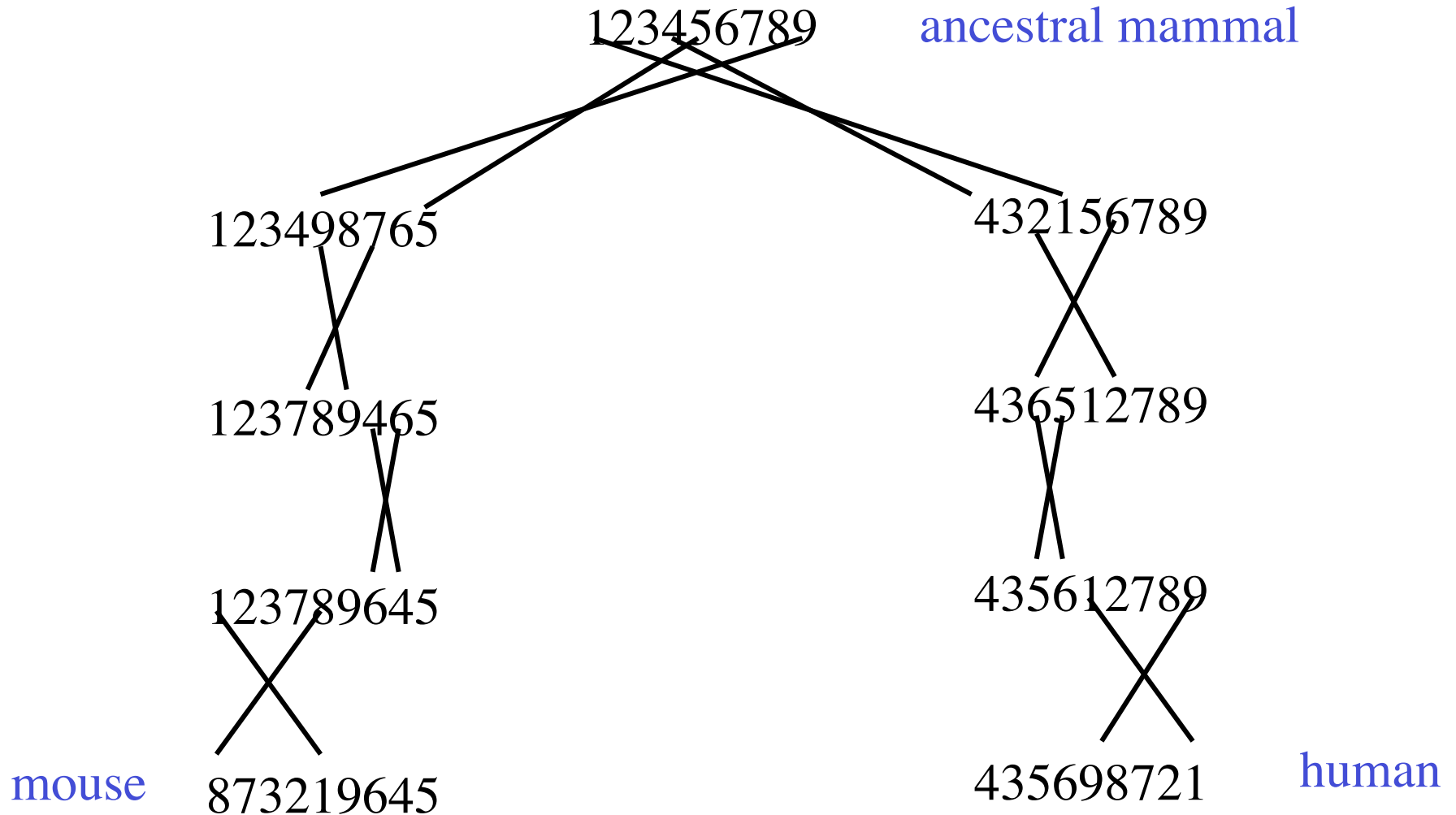


Edit distance as a series of mutations



homoplasy

Edit distance (genomic) as a series of reversals





Placental Ancestor

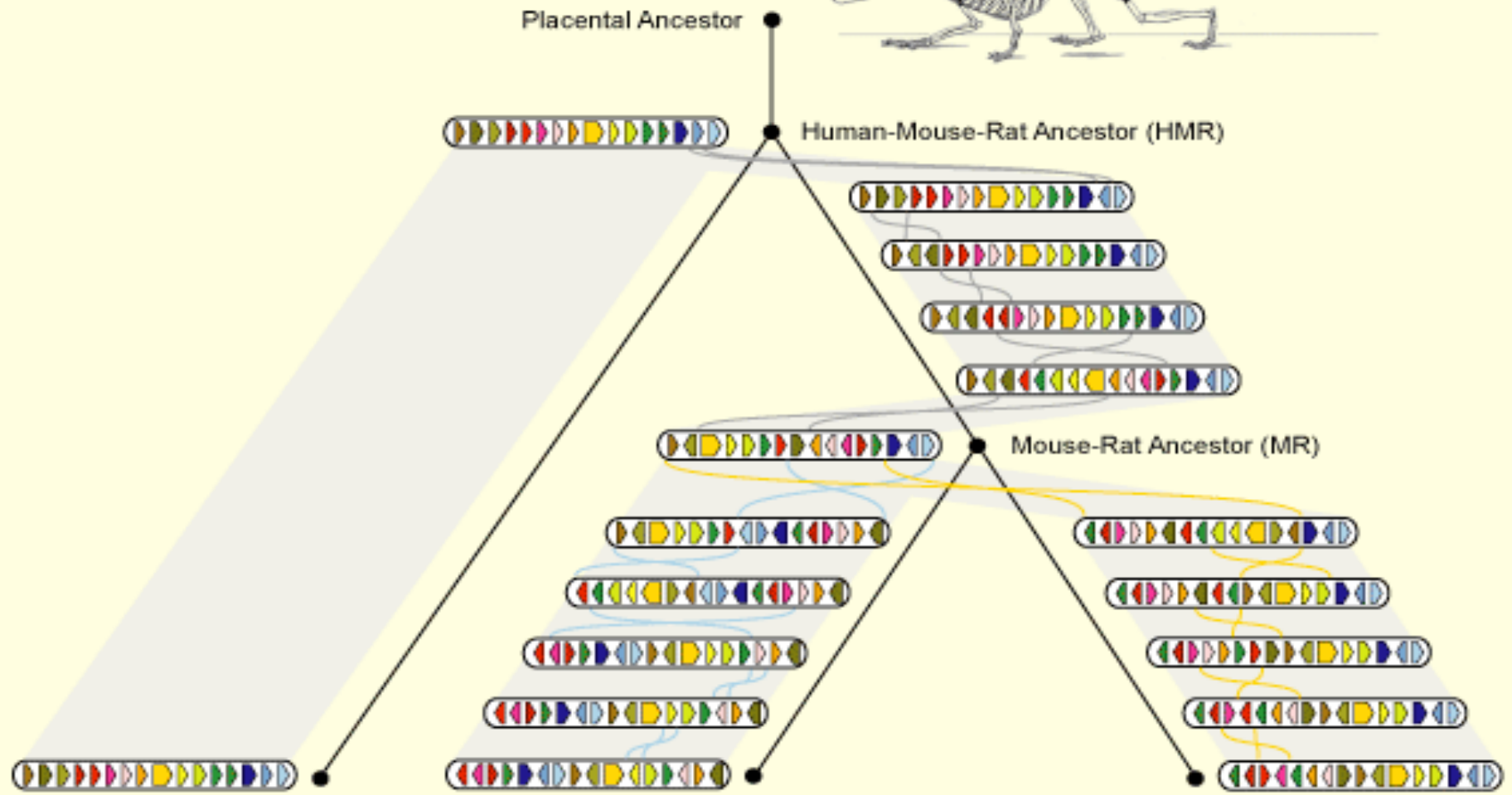
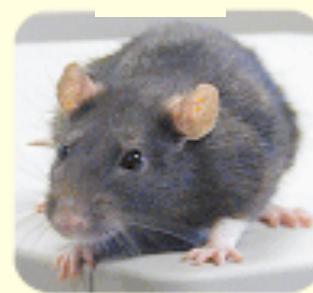
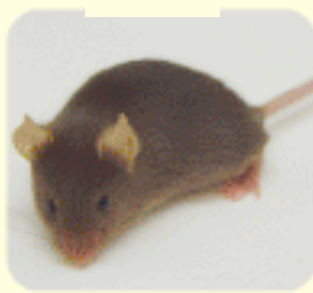
Human-Mouse-Rat Ancestor (HMR)

Mouse-Rat Ancestor (MR)

rug rat

mouse

rat



The Pancake Flipping Problem

A sloppy cook at a pancake diner makes pancakes of all different sizes and stacks them haphazardly.

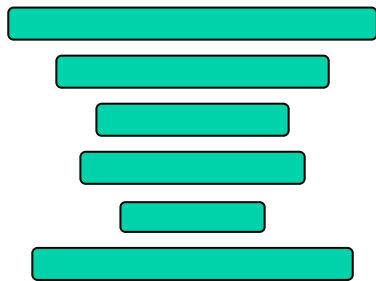
The waiter likes the pancakes to be stacked with the largest on the bottom and the smallest on top. On the way to the table, using only one hand with a spatula, he flips the pancakes until they are arranged by size, largest on bottom, smallest on top.

- What is the algorithm for flipping?
- What is the algorithm for finding the fewest flips?
- Same problem, pancakes burned on one side.

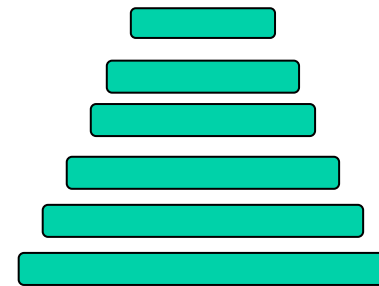
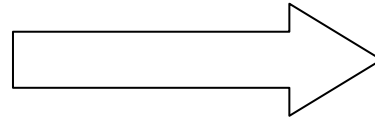


In class exercise:

Given the arrangement below, flip the pancakes until they are in order. How many flips? (You can order the numbers instead of the pancakes.)



642315



123456

In class exercise: *work in pairs*

- Write detailed instructions on how to stack six pancakes in order by flipping. *The instructions should not depend on the starting order.*

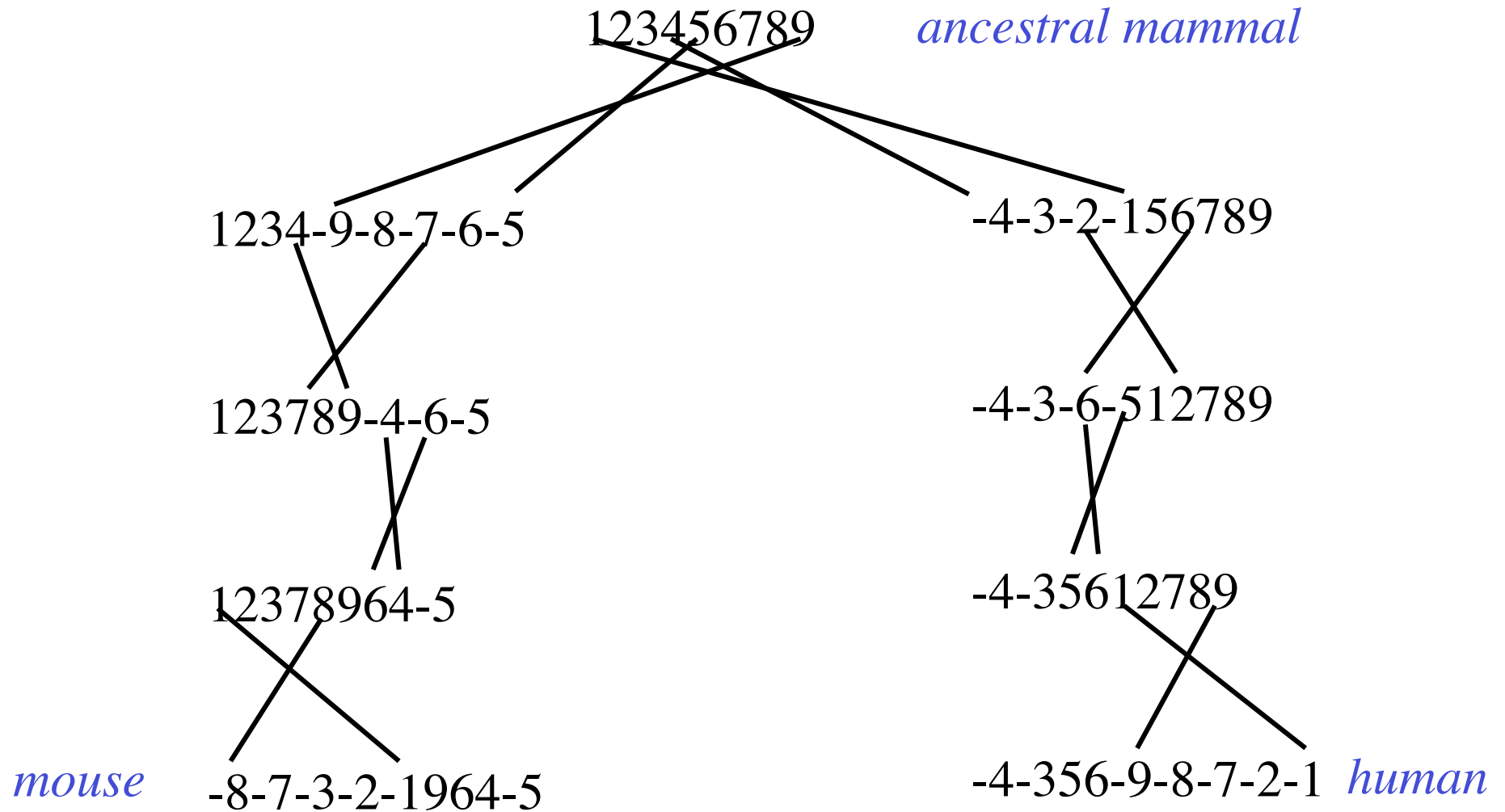
- Give your instructions to your partner. Follow your partner's written instructions to stack the following six "pancakes" in order:

125436 ---> ... ---> ... ---> 123456

- On the board: Convert these instructions to pseudocode.

“pancakes burned on one side” problem

Reversals put genes on the opposite strand.



“-” indicates that the gene is on the *reverse complement* strand.

Finding the minimum number of inversions by graph decomposition.

01234567 \implies 03152647

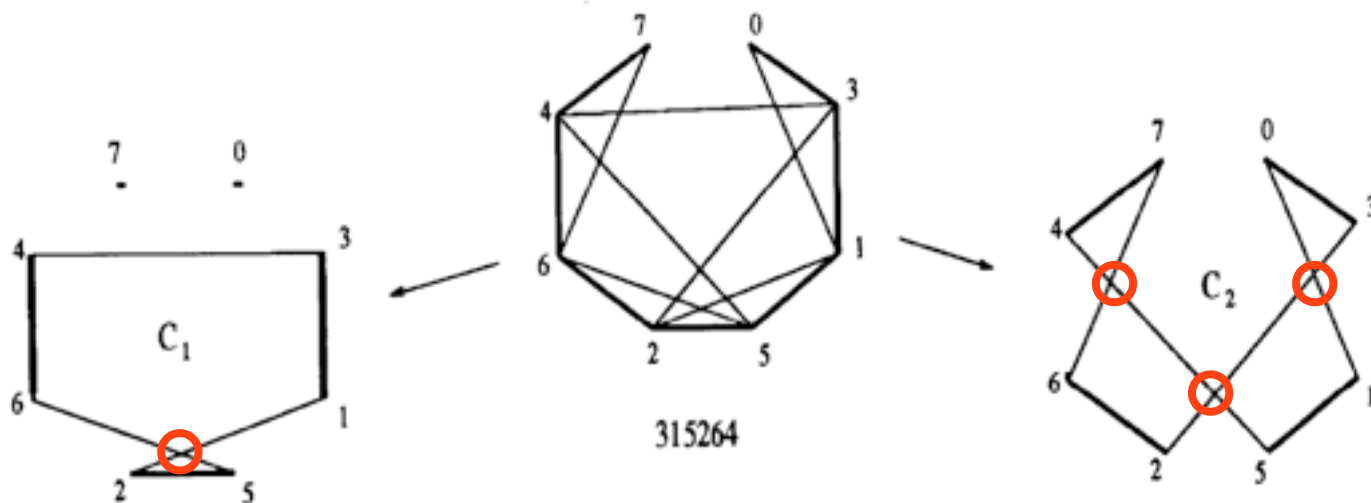
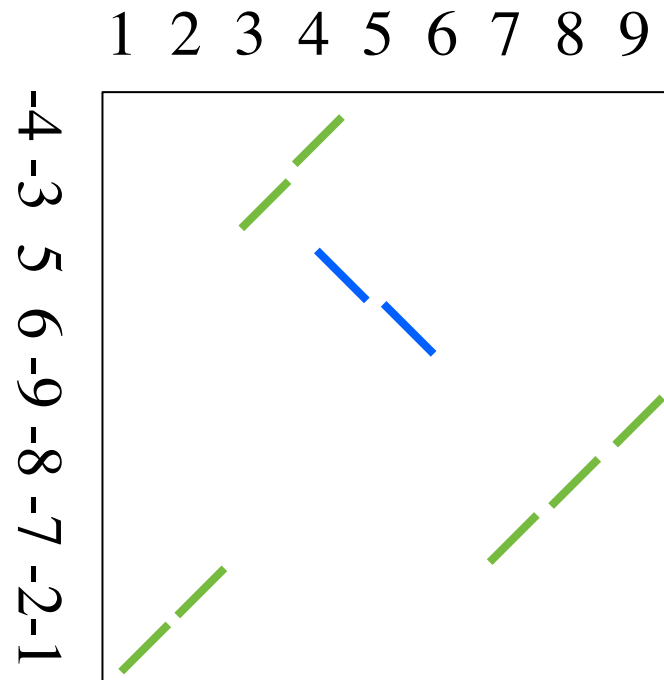


FIG. 2.—Breakpoint graph $G(315264)$ and its maximal cycle decomposition

1. Draw a graph: Nodes are genes. Edges are sequential orders.
2. Decompose graph into cycles.
3. Count crossing edges.

Plotting rearranged genomes

1 2 3 4 5 6 7 8 9 → -4 -3 5 6 -9 -8 -7 -2 -1



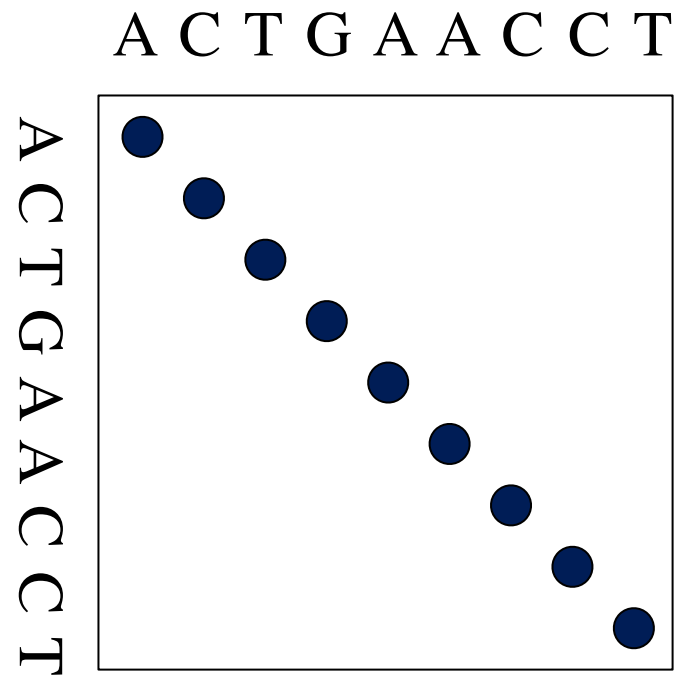
Alignment matrix

	A	C	T	G	A	A	C	C	T
A	1	0	0	0	0	0	0	0	0
C	0	1	0	0	0	0	0	0	0
T	0	0	1	0	0	0	0	0	0
G	0	0	0	1	0	0	0	0	0
A	0	0	0	0	1	0	0	0	0
A	0	0	0	0	0	1	0	0	0
C	0	0	0	0	0	0	1	0	0
C	0	0	0	0	0	0	0	1	0
T	0	0	0	0	0	0	0	0	1

1=aligned (associated)
0=not aligned

- Boolean matrix.
- Only one “1” per row.
- Only one “1” per column.
- If $A(i,j)=1$, then $A(m,n)=0$ for all $(m < i \ \&\& \ n > j)$
- If $A(i,j)=1$, then $A(m,n)=0$ for all $(m > i \ \&\& \ n < j)$

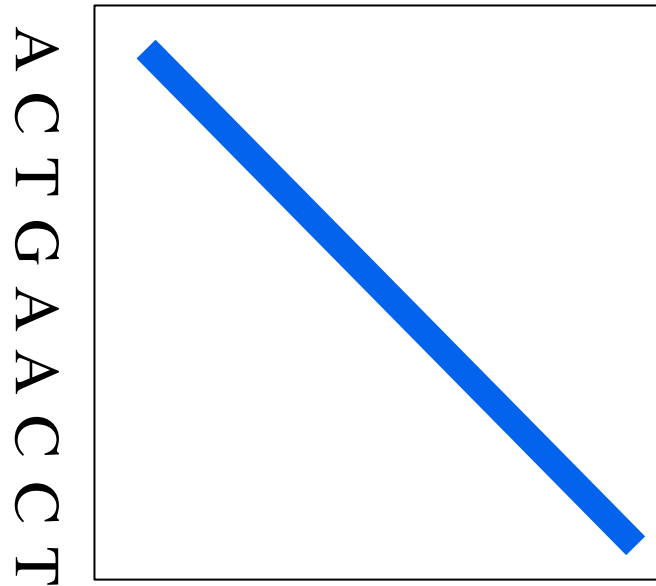
Alignment matrix



easier to draw this way

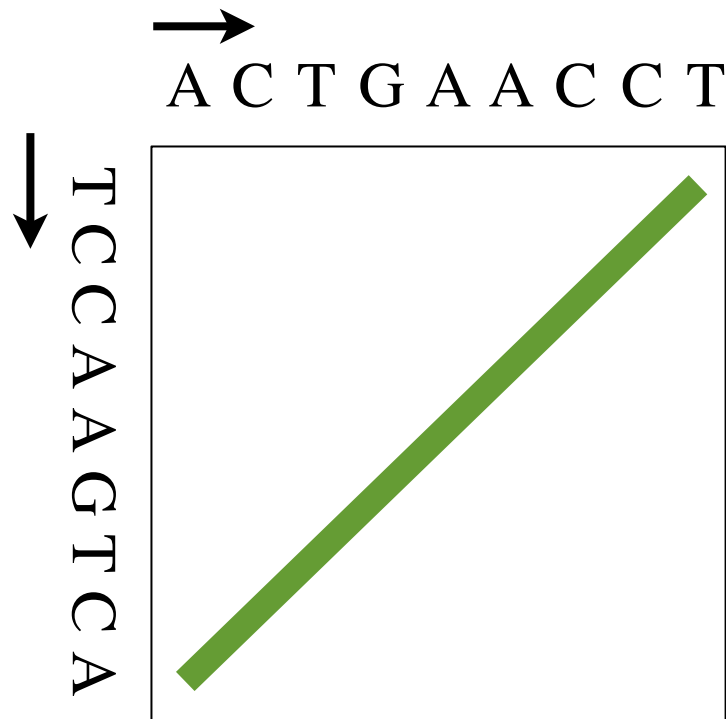
Alignment matrix

A C T G A A C C T



easier still

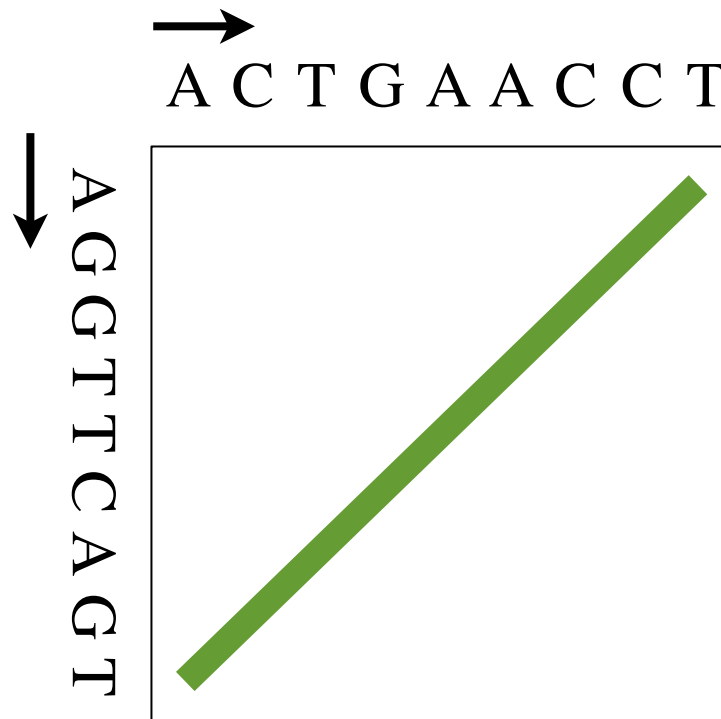
Alignment matrix



reverse alignment

(NOT biologically relevant!)

Alignment matrix



reverse complement alignment

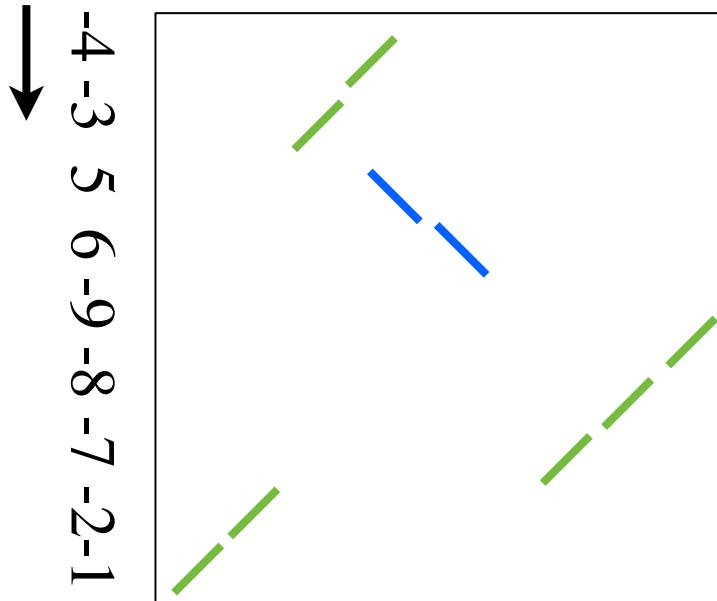
(yes, biologically relevant, for nucleotides!)

Plotting rearranged genomes

1 2 3 4 5 6 7 8 9 → -4 -3 5 6 -9 -8 -7 -2 -1

Each number represents a long sequence

→
1 2 3 4 5 6 7 8 9



Each negative number represents its reverse complement

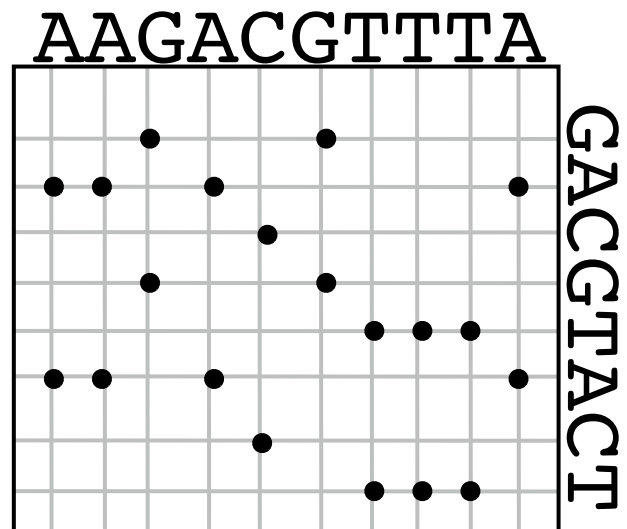
Z&B p. 158

Dot Plot

Each position in the matrix $D[i,j]$ is either

→ **dot**, if $A[i] == B[j]$

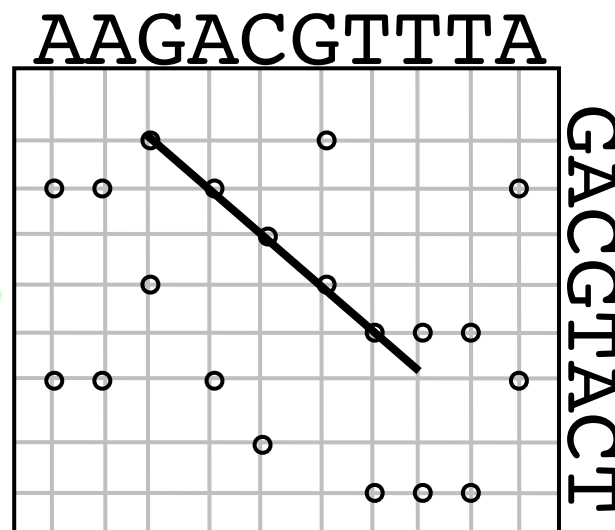
→ **blank**, otherwise.



A more advanced dot plot

With thousands of bases, it is impossible to plot all dots in the matrix. Instead we look for stretches of sequence with few mismatches. If the number of mismatches is less than the cutoff, plot a dot or line.

Show all diagonals with at least 4 out of 5 matches.



"**window size**" is the length of a diagonal, "**stringency**" is minimum number of matches in the window.

Install UGENE

- <http://ugene.unipro.ru/>
- Download UGENE manual.
- Bookmark podcasts

Explore UGENE

- Open data/samples/FASTA/human_T1.fa
- Right-click in sequence window. Select Analyze...Build Dotplot... Set x=y=human_T1. Direct repeats. Set minimum length = 50, 100%identity. Click OK.
- Navigate the dotplot window by zooming and scrolling. Find the longest repeat. Select it. Make an annotation “Longest direct repeat”.

-