CS481: Bioinformatics Algorithms

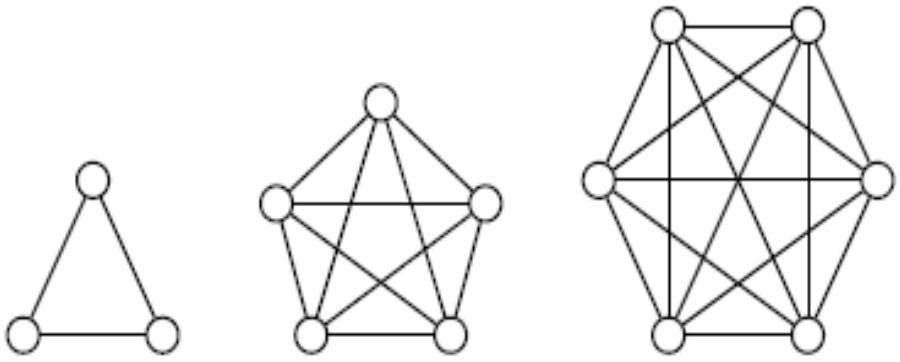
Can Alkan
EA224
calkan@cs.bilkent.edu.tr

http://www.cs.bilkent.edu.tr/~calkan/teaching/cs481/

CLUSTERING USING GRAPHS

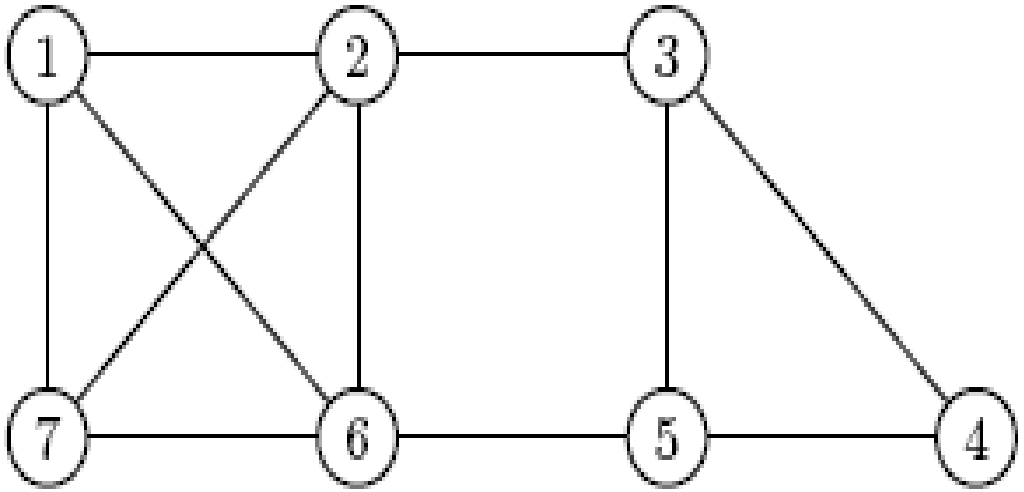
Clique Graphs

- A **clique** is a graph with every vertex connected to every other vertex
- A **clique graph** is a graph where each connected component is a clique



Transforming an Arbitrary Graph into a Clique Graphs

- **A graph can be transformed into a clique graph by adding or removing edges**



Corrupted Cliques Problem

Input: A graph *G*

Output: The smallest number of additions and removals of edges that will transform *G* into a clique graph

Distance Graphs

- Turn the distance matrix into a distance graph
  - Genes are represented as vertices in the graph
  - Choose a distance threshold θ
  - If the distance between two vertices is below θ, draw an edge between them
  - The resulting graph may contain cliques
  - These cliques represent clusters of closely located data points

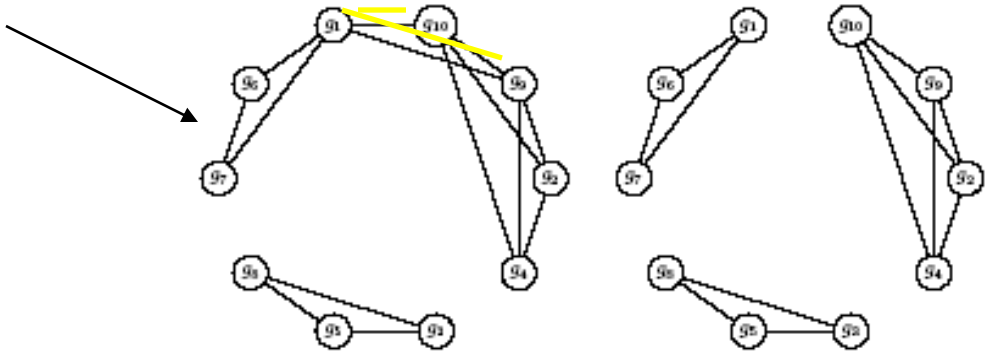
Transforming Distance Graph into Clique Graph

The distance graph (threshold θ=7) is transformed into a clique graph after removing the two highlighted edges

	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_{10}
g_1	0.0	8.1	9.2	7.7	9.3	2.3	5.1	10.2	6.1	7.0
g_2	8.1	0.0	12.0	0.9	12.0	9.5	10.1	12.8	2.0	1.0
g_3	9.2	12.0	0.0	11.2	0.7	11.1	8.1	1.1	10.5	11.5
g_4	7.7	0.9	11.2	0.0	11.2	9.2	9.5	12.0	1.6	1.1
g_5	9.3	12.0	0.7	11.2	0.0	11.2	8.5	1.0	10.6	11.6
g_6	2.3	9.5	11.1	9.2	11.2	0.0	5.6	12.1	7.7	8.5
g_7	5.1	10.1	8.1	9.5	8.5	5.6	0.0	9.1	8.3	9.3
g_8	10.2	12.8	1.1	12.0	1.0	12.1	9.1	0.0	11.4	12.4
g_9	6.1	2.0	10.5	1.6	10.6	7.7	8.3	11.4	0.0	1.1
g_{10}	7.0	1.0	11.5	1.1	11.6	8.5	9.3	12.4	1.1	0.0

(a) Distance matrix, d (distances shorter than 7 are shown in bold).

After transforming the distance graph into the clique graph, the dataset is partitioned into three clusters



(b) Distance graph for $\theta = 7$.

(c) Clique graph.

Figure 10.6 The distance graph (b) for $\theta = 7$ is not quite a clique graph. However, it can be transformed into a clique graph (c) by removing edges (g_1, g_{10}) and (g_1, g_9).

Heuristics for Corrupted Clique Problem

- Corrupted Cliques problem is NP-Hard, some heuristics exist to approximately solve it:
- **CAST** (Cluster Affinity Search Technique): a practical and fast algorithm:
  - **CAST** is based on the notion of genes *close* to cluster *C* or *distant* from cluster *C*
  - Distance between gene *i* and cluster *C*:

d(i,C) = average distance between gene *i* and all genes in *C*

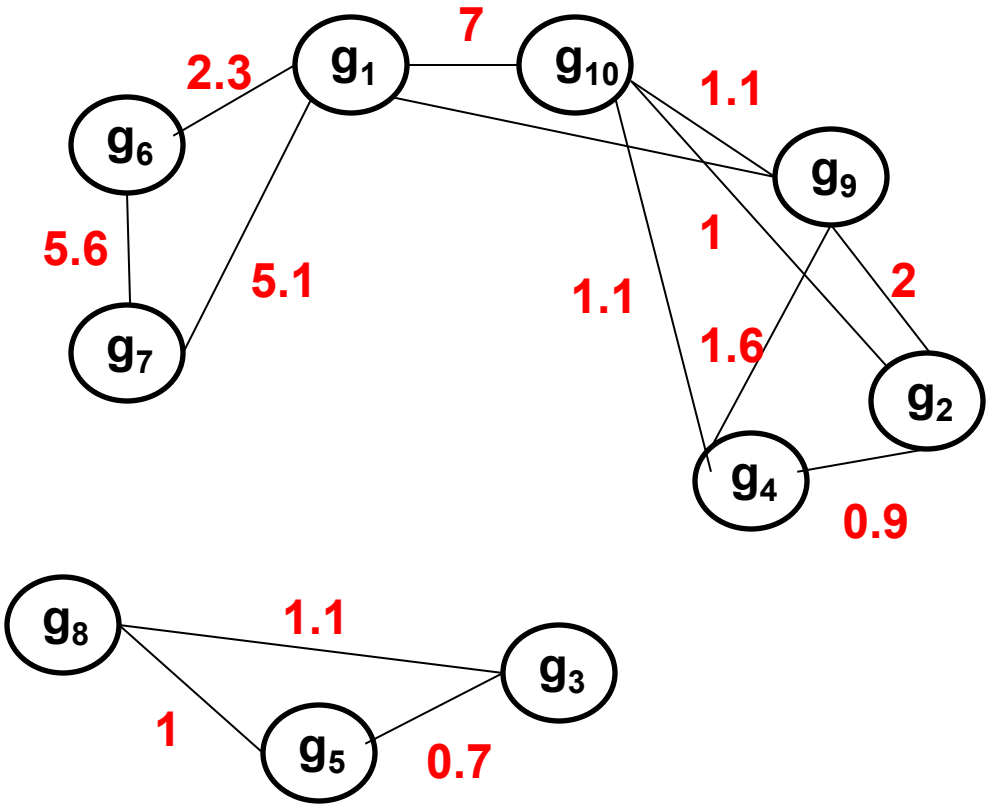
Gene *i* is ***close*** to cluster *C* if $d(i,C) < \theta$ and ***distant*** otherwise

CAST Algorithm

1. CAST(S, G, θ)
2. $P \leftarrow \emptyset$
3. **while** $S \neq \emptyset$
4. $V \leftarrow$ vertex of maximal degree in the distance graph G
5. $C \leftarrow \{v\}$
6. **while** a close gene i *not in* C or distant gene i *in* C exists
7. Find the nearest close gene i not in C and add it to C
8. Remove the farthest distant gene i in C
9. Add cluster C to partition P
10. $S \leftarrow S \setminus C$
11. Remove vertices of cluster C from the distance graph G
12. return P

S – set of elements, G – distance graph, θ – distance threshold

CAST Algorithm



$\Theta = 7$

$P = \emptyset$

$S=\{g_1,\ldots,g_{10}\}$

$degree(g_{10}) = 4$

$C_1 = \{g_{10}\}$

$C_1 = \{g_2, g_{10}\}$

$d(g_1, C_1) = (7+8.1) / 2 = 7.55$

$d(g_4, C_1) = (0.9+1.1) / 2 = 1$

$d(g_9, C_1) = (2+1.1) / 2 = 1.55$

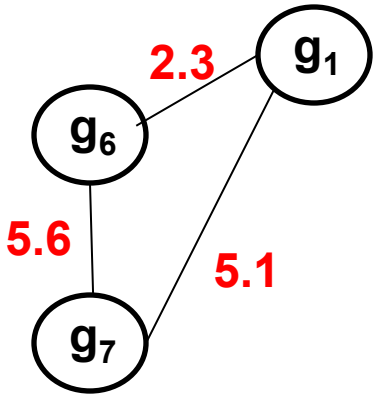
$C_1 = \{g_2, g_4, g_{10}\}$

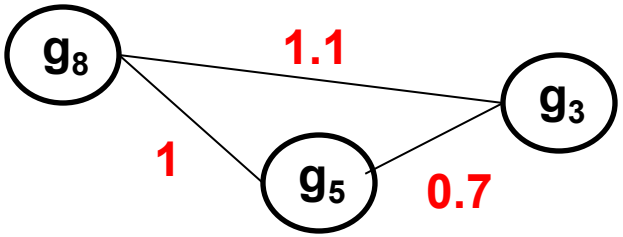
$d(g_9,C) = (2+1.6+1) / 3 = 1.53$

$C_1 = \{g_2, g_4, g_9, g_{10}\}$

$P = \{C_1\}$

CAST Algorithm





$\Theta = 7$

$P = \{C_1\}$

$C_1 = \{g_2, g_4, g_9, g_{10}\}$

$S=\{g_1,g_3,g_5, g_6,g_7, g_8\}$

$degree(g_1) = 2$

$C_2 = \{g_1\}$

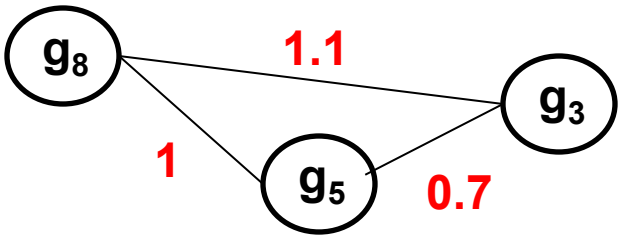
$C_2 = \{g_1, g_6\}$

$d(g_7, C_2) = (5.1+5.6) / 2 = 5.35$

$C_2 = \{g_1, g_6, g_7\}$

$P = \{C_1, C_2\}$

CAST Algorithm



$\Theta = 7$

$P = \{C_1, C_2\}$

$C_1 = \{g_2, g_4, g_9, g_{10}\}$

$C_2 = \{g_1, g_6, g_7\}$

$S=\{g_3, g_5, g_8\}$

$degree(g_3) = 2$

$C_3 = \{g_3\}$

$C_3 = \{g_3, g_5\}$

$d(g_8, C_3) = (1.1+1) / 2 = 1.05$

$C_3 = \{g_3, g_5, g_8\}$

$P = \{C_1, C_2, C_3\}$

CAST Algorithm

$\Theta = 7$

$P = \{C_1, C_2, C_3\}$

$C_1 = \{g_2, g_4, g_9, g_{10}\}$

$C_2 = \{g_1, g_6, g_7\}$

$C_3 = \{g_3, g_5, g_8\}$

$S = \emptyset$

… done

GENOME REARRANGEMENTS

Turnip vs Cabbage: Look and Taste Different

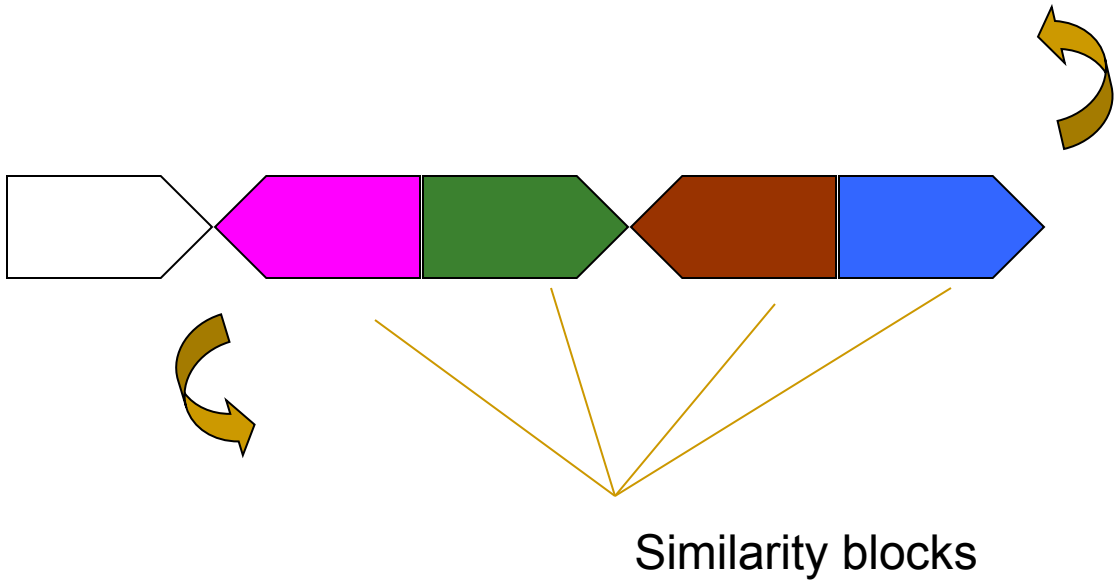
- Although cabbages and turnips share a recent common ancestor, they look and taste different

Turnip vs Cabbage: Almost Identical mtDNA gene sequences

- In 1980s Jeffrey Palmer studied evolution of plant organelles by comparing mitochondrial genomes of the cabbage and turnip
- 99% similarity between genes
- These surprisingly identical gene sequences differed in gene order
- This study helped pave the way to analyzing genome rearrangements in molecular evolution

Turnip vs Cabbage: Different mtDNA Gene Order

- Gene order comparison:



Turnip vs Cabbage: Different mtDNA Gene Order

- Gene order comparison:

Turnip vs Cabbage: Different mtDNA Gene Order

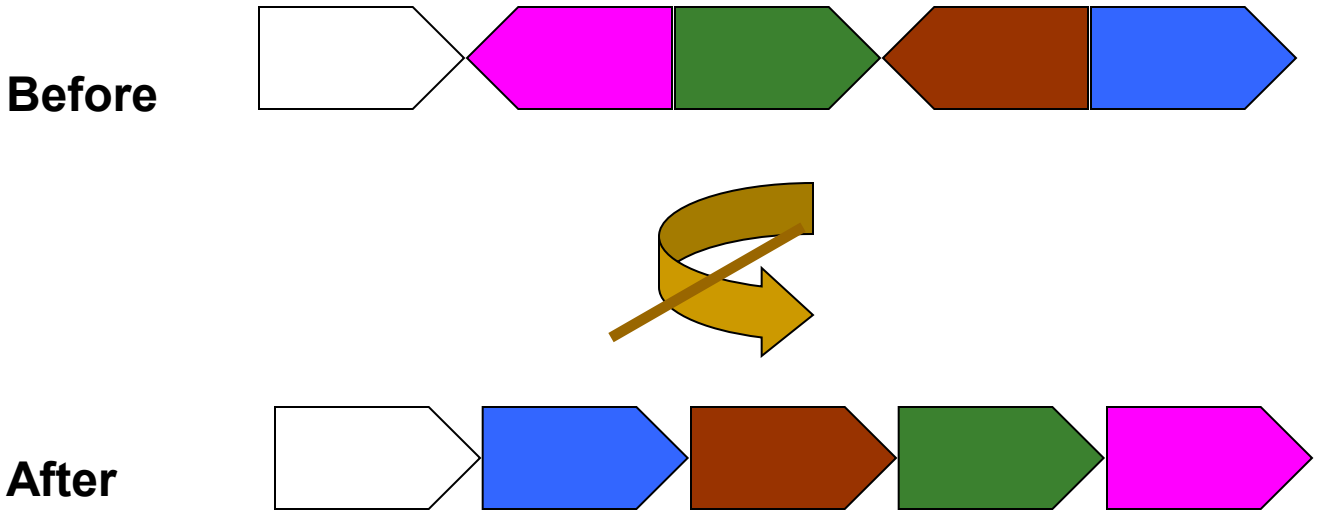
- Gene order comparison:

Turnip vs Cabbage: Different mtDNA Gene Order

- Gene order comparison:

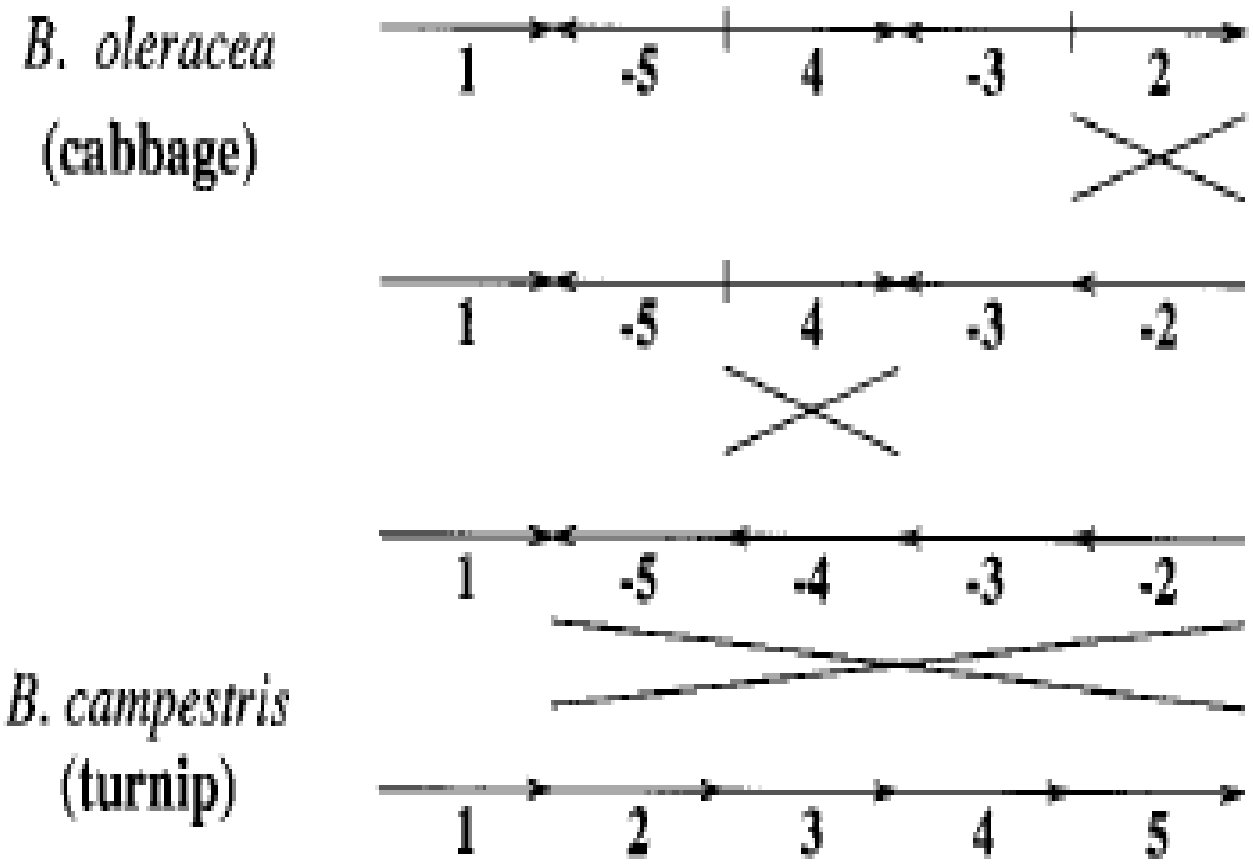
Turnip vs Cabbage: Different mtDNA Gene Order

- Gene order comparison:

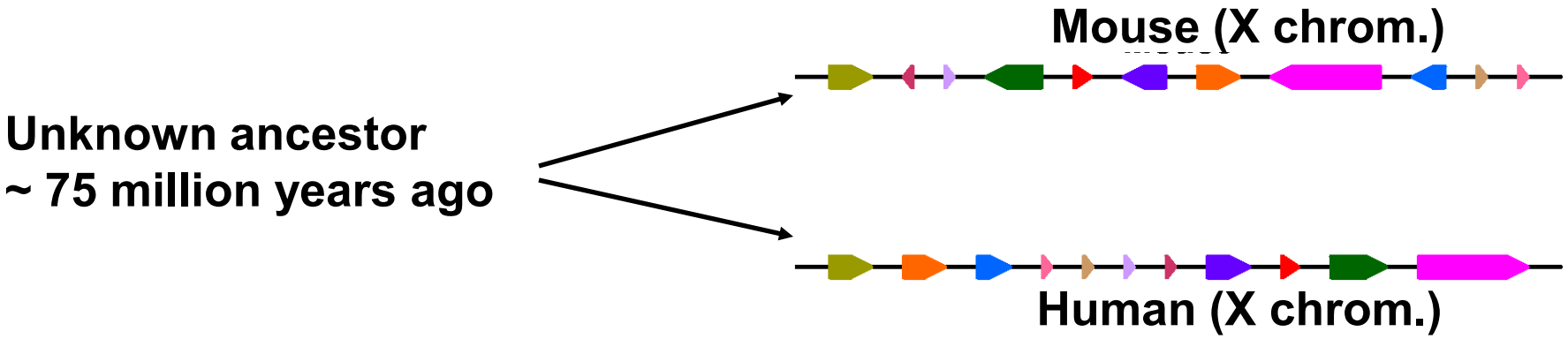


Evolution is manifested as the divergence in gene order

Transforming Cabbage into Turnip

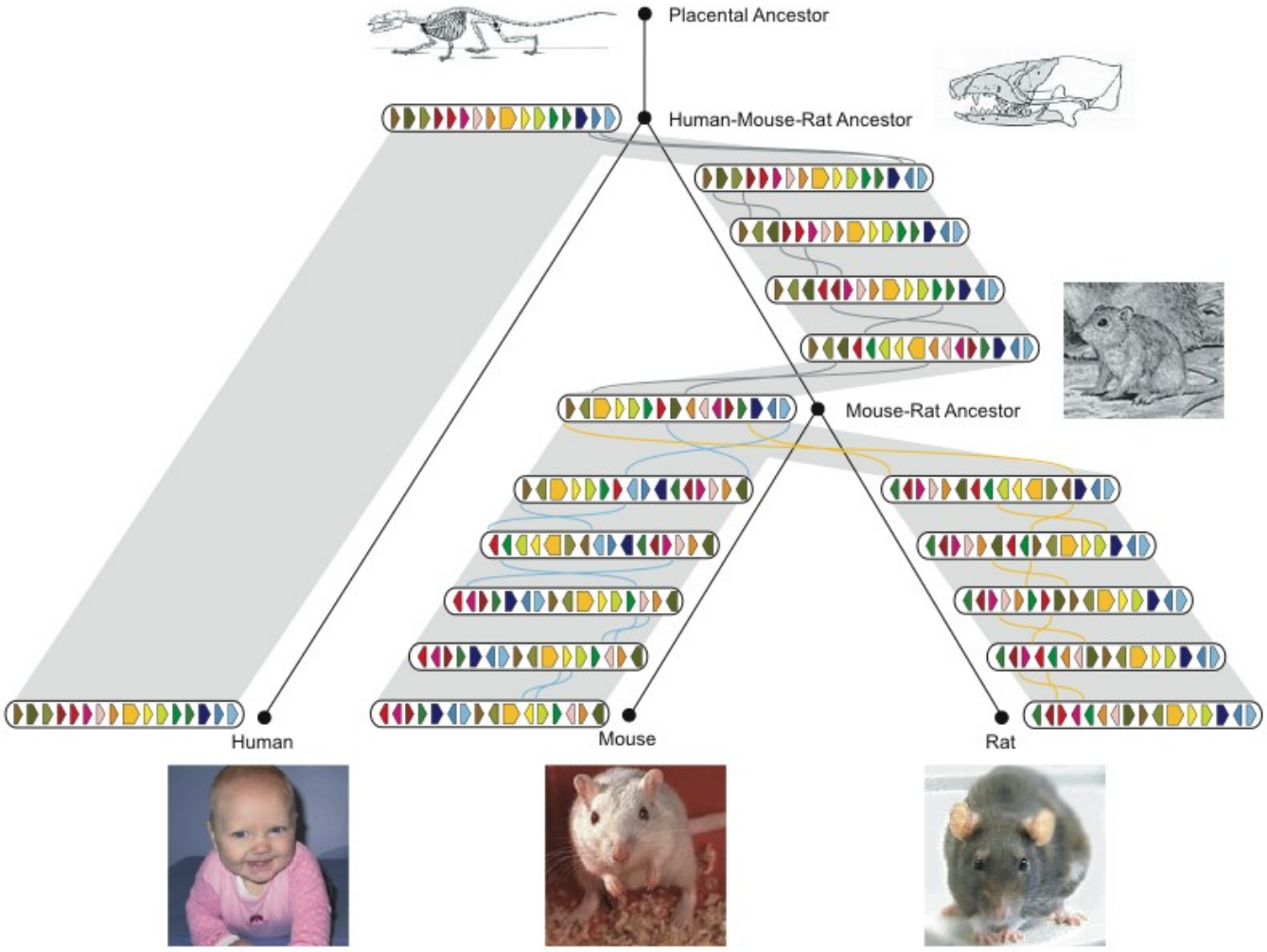


Genome rearrangements



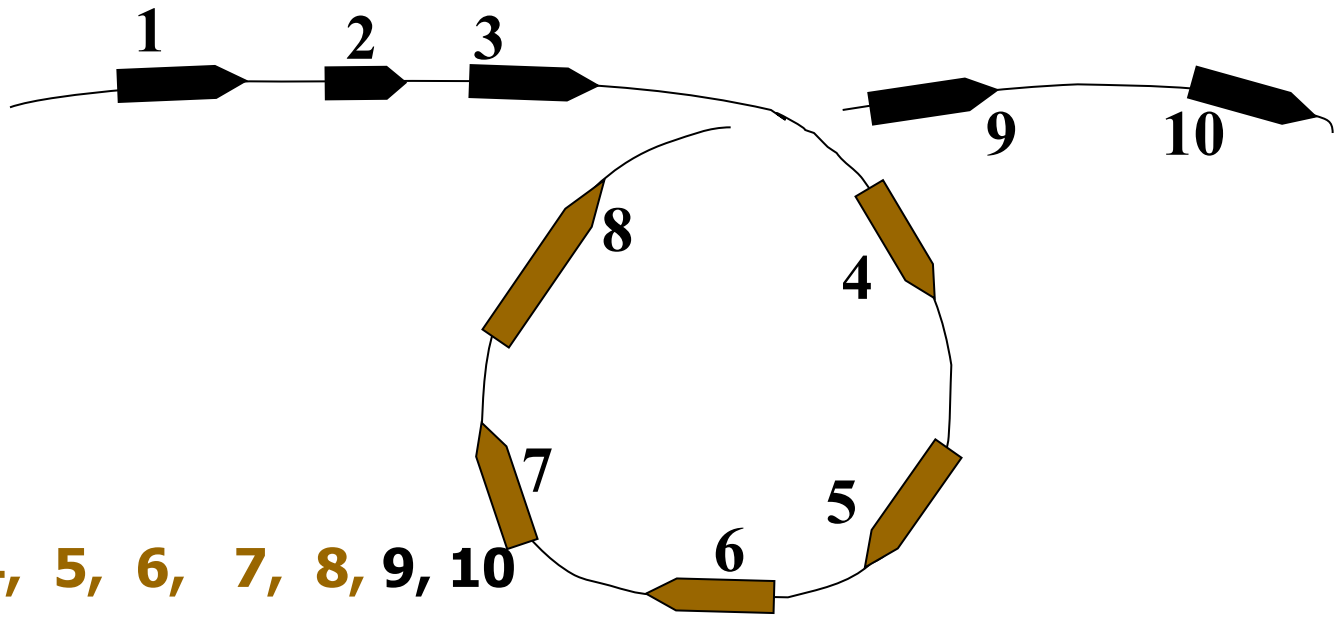
- What are the similarity blocks and how to find them?
- What is the architecture of the ancestral genome?
- What is the evolutionary scenario for transforming one genome into the other?

History of Chromosome X



Rat Consortium, *Nature*, 2004

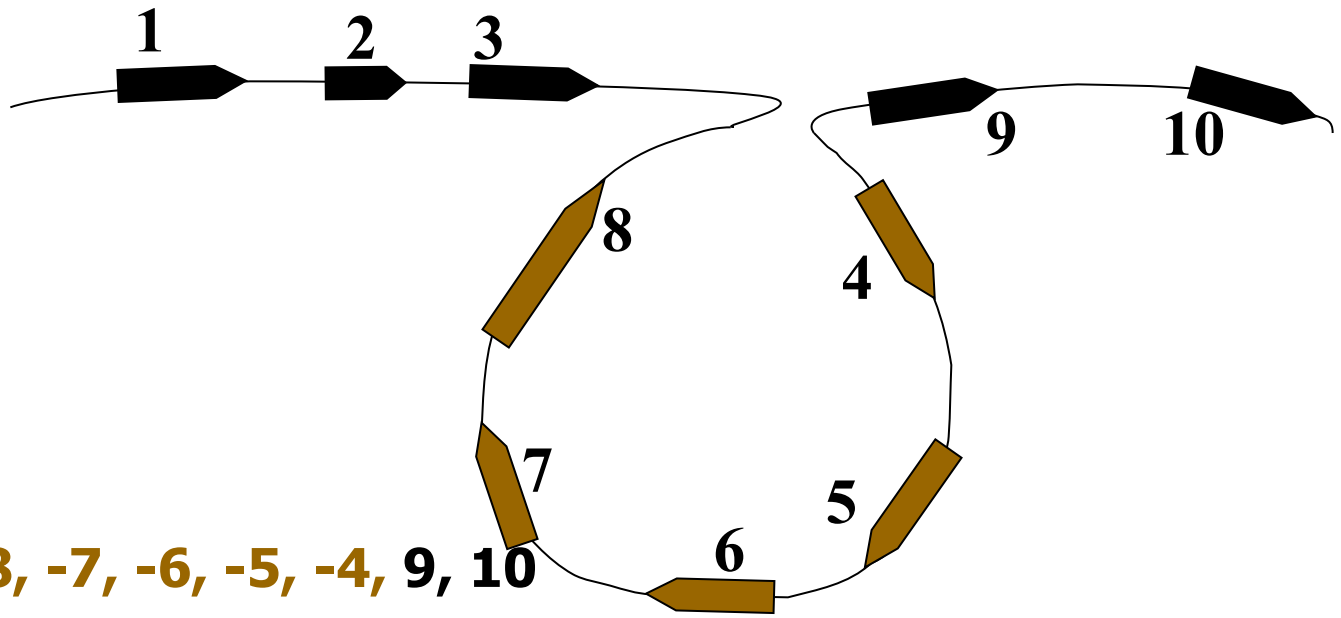
Reversals



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

- Blocks represent conserved genes.

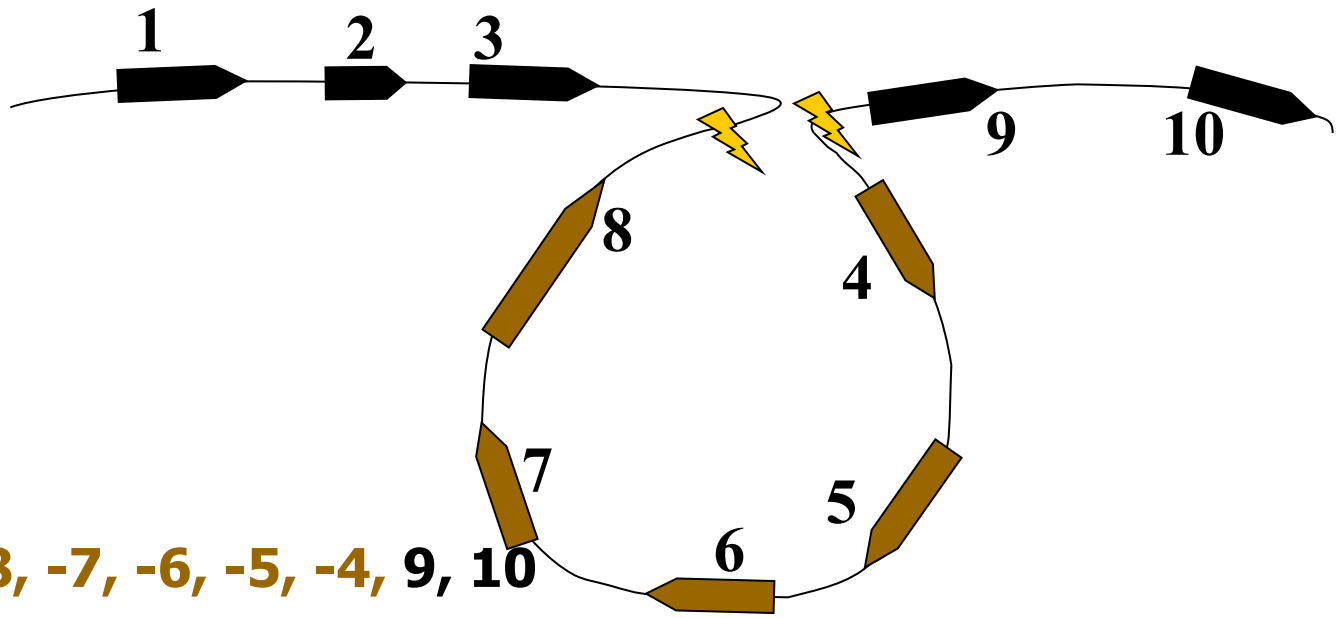
Reversals



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

- **Blocks represent conserved genes.**
- **In the course of evolution or in a clinical context, blocks 1,…,10 could be misread as 1, 2, 3, -8, -7, -6, -5, -4, 9, 10.**

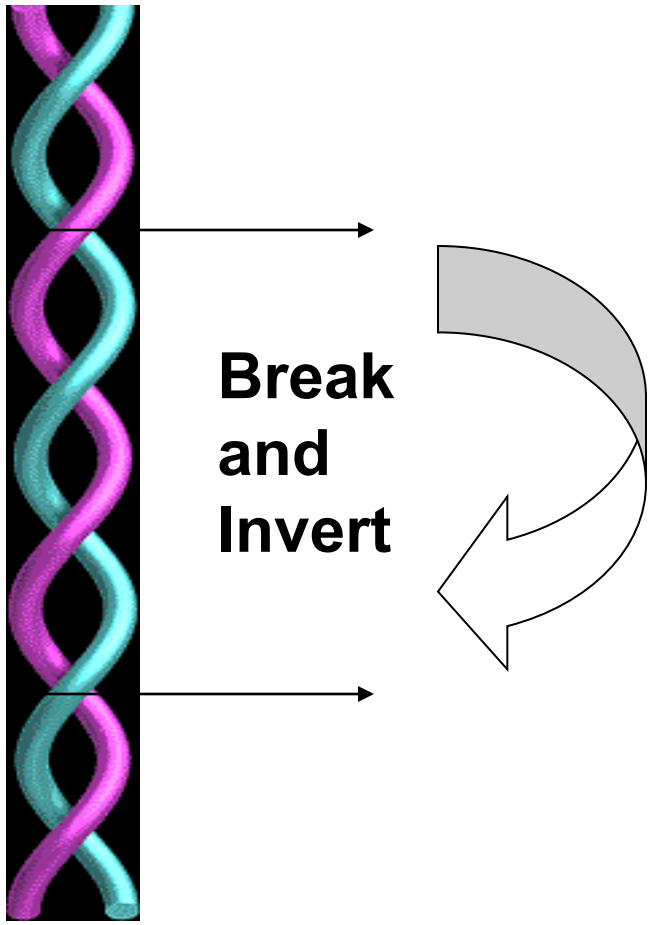
Reversals and Breakpoints



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

The reversion introduced two *breakpoints* (disruptions in order).

Reversals: Example



5’ ATGCCTGTACTA 3’

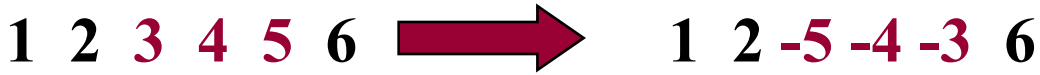
3’ TACGGACATGAT 5’

5’ ATGTACAGGCTA 3’

3’ TACATGTCCGAT 5’

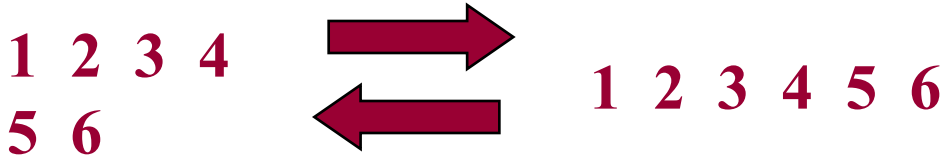
Types of Rearrangements

Reversal



Translocation

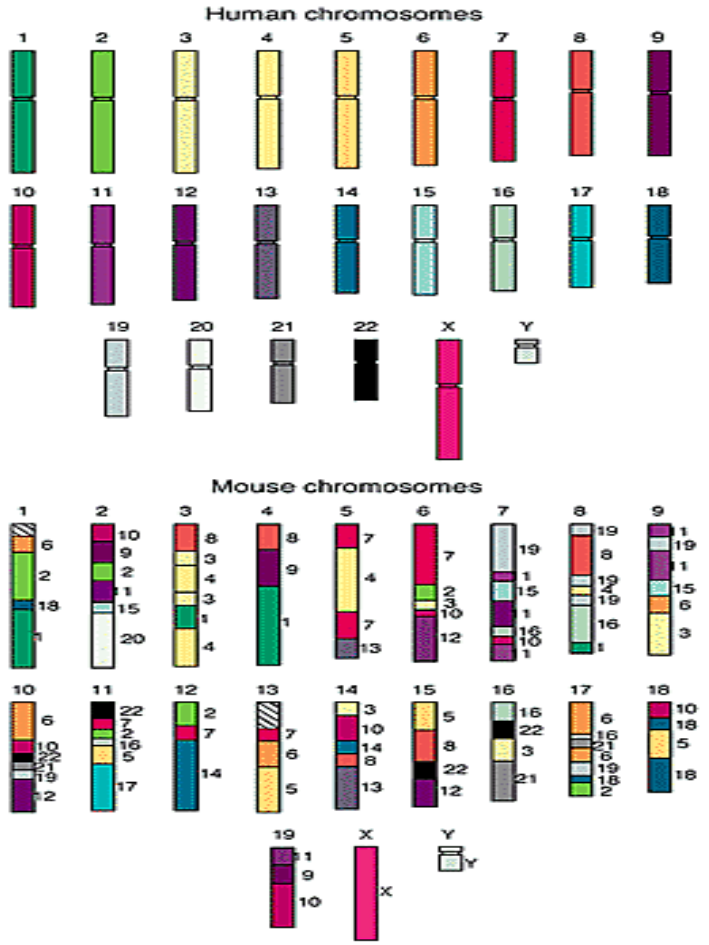
Fusion



Fission

Comparative Genomic Architectures: Mouse vs Human Genome

- Humans and mice have similar genomes, but their genes are ordered differently
- ~245 rearrangements
  - Reversals
  - Fusions
  - Fissions
  - Translocation



Human chromosome 2