Introduction to Bioinformatics

Lecture 4:

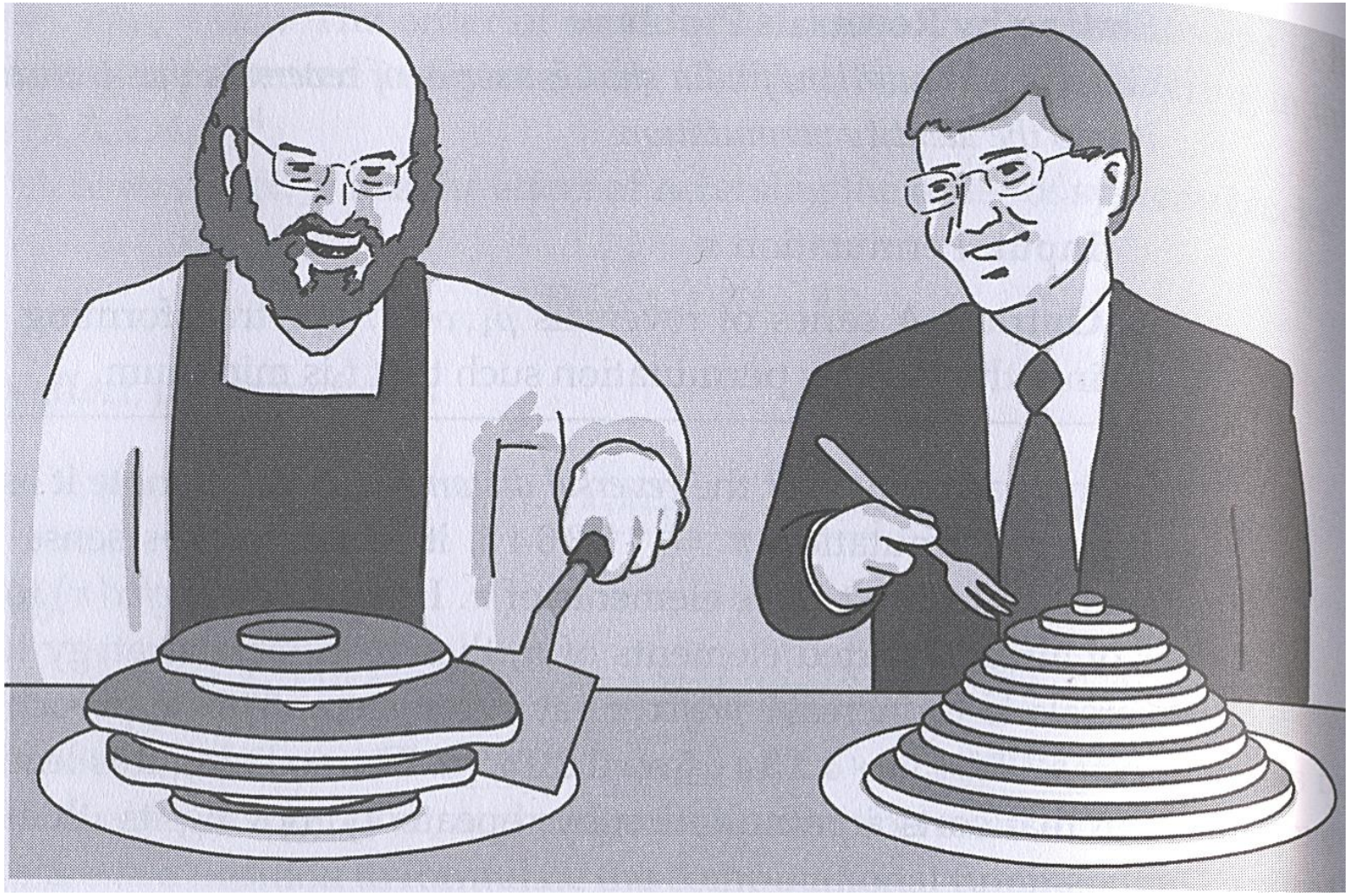
Genome rearrangements

Why study genome rearrangements?

- Provide insight into evolution of species
- Fun algorithmic problem!

- Structure of this lecture:
  - The biological phenomenon
  - How to computationally model it?
  - How to compute interesting things?
  - Studying the phenomenon using existing tools (continued in exercises)

Genome rearrangements as an algorithmic problem



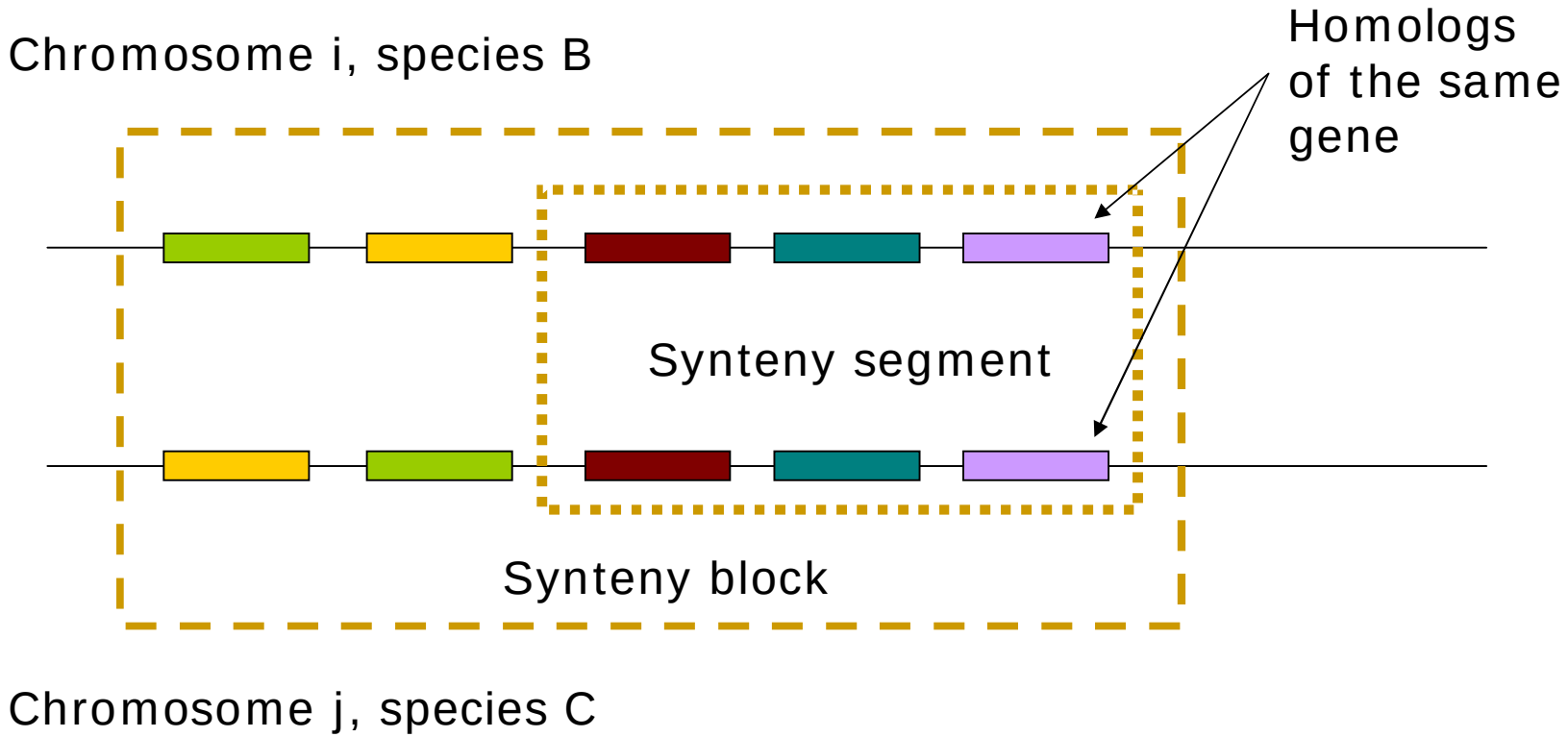
Background

- Genome sequencing enables us to compare genomes of two or more different species
  - -> Comparative genomics
- Basic observation:
  - Closely related species (such as human and mouse) can be almost identical in terms of genome contents...
  - ...but the order of genomic segments can be very different between species

Synteny blocks and segments

- Synteny – derived from Greek 'on the same ribbon' – means genomic segments located on the same chromosome
  - Genes, markers (any sequence)
- Synteny block (or syntenic block)
  - A set of genes or markers that co-occur together in two species
- Synteny segment (or syntenic segment)
  - Syntenic block where the *order* of genes or markers is preserved

Synteny blocks and segments

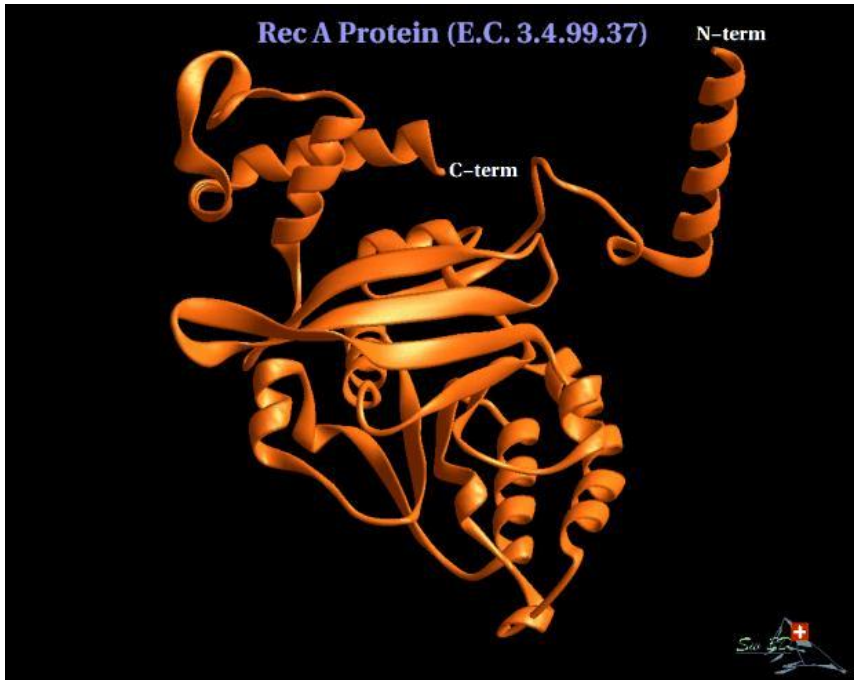


Observations from sequencing

1. Large chromosome *inversions* and *translocations* (we'll get to these shortly) are common
   - ...Even between closely related species
2. Chromosome inversions are usually symmetric around the *origin of DNA replication*
3. Inversions are less common *within species*...

What causes rearrangements?

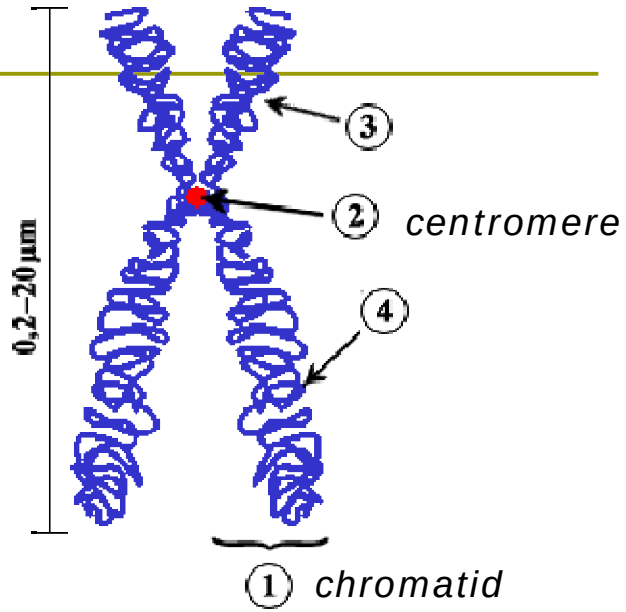
- RecA, Recombinase A, is a protein used to repair chromosomal damage
- It uses a duplicate copy of the damaged sequence as template
- Template is usually a homologous sequence on a *sister chromosome*



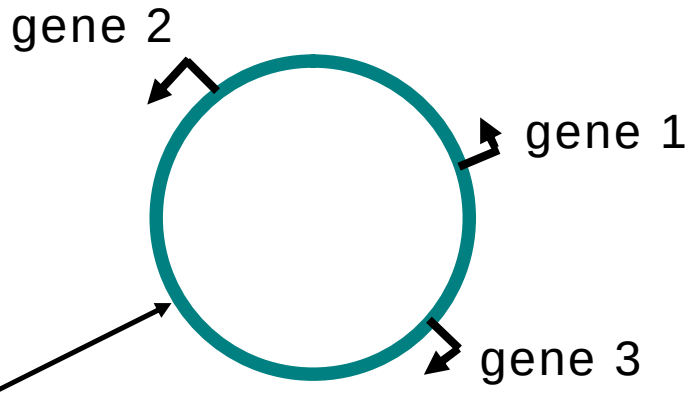
Diarmaid Hughes: Evaluating genome dynamics: the constraints on rearrangements within bacterial genomes, *Genome Biology* 2000, 1

Chromosomes: recap

- Linear chromosomes
  - Eukaryotes (mostly)



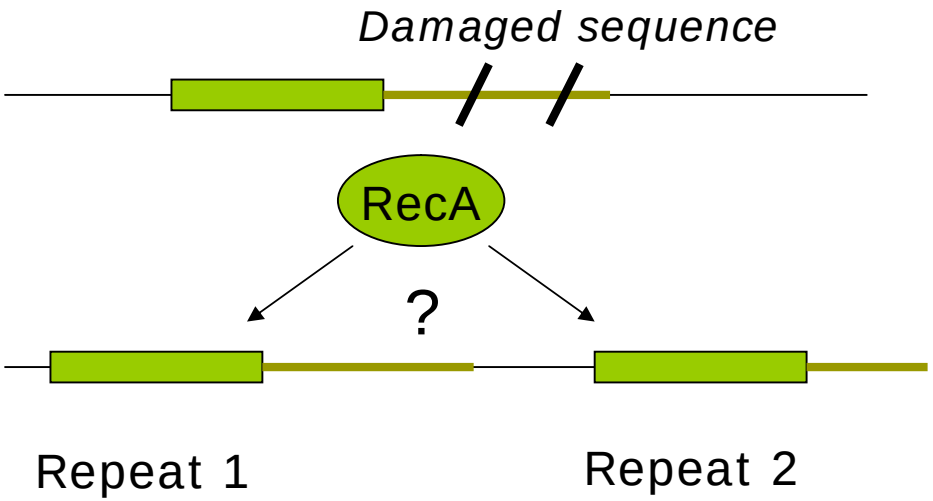
- Circular chromosomes
  - Prokaryotes (mostly)
  - Mitochondria

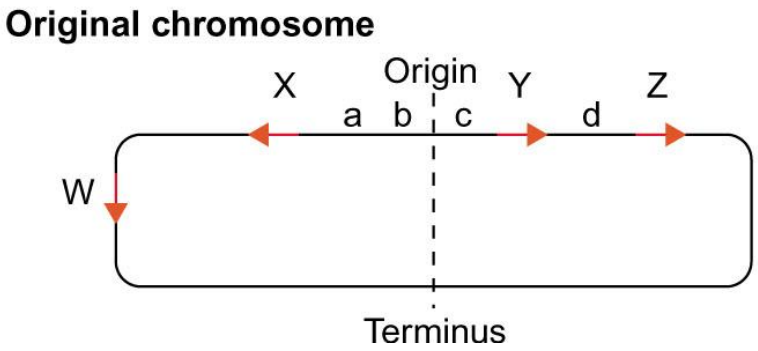


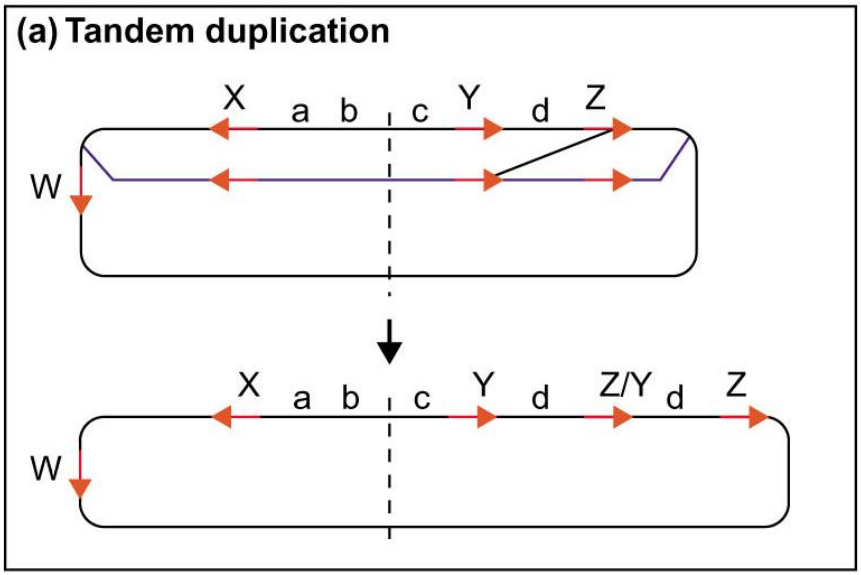
Also double-stranded: genes can be found on both strands (*orientations*)

What effects does RecA have on genome?

- Repeated sequences cause RecA to fail to choose correct recombination start position
- This leads to
  - Tandem duplications
  - Translocations
  - Inversions







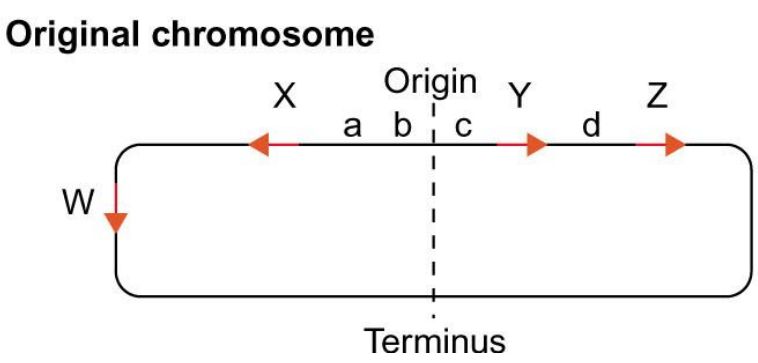
X, Y, Z and W are repeats of the same sequence.

a, b, c and d are sequences on genome bounded by repeats.

In a tandem duplication example, RecA recombines a sequence that starts from Y instead of Z after Z.

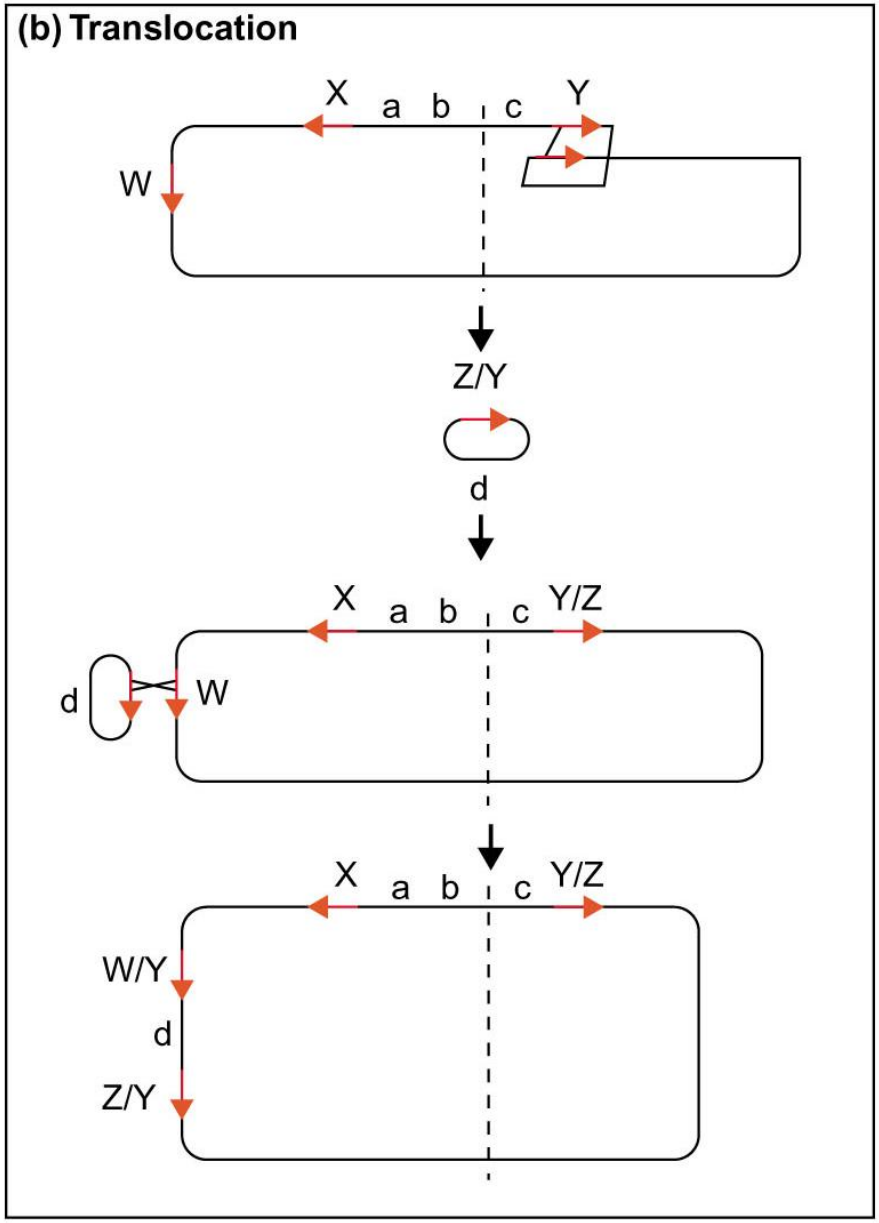
This leads to duplication of segment Y-Z.

Diarmaid Hughes: Evaluating genome dynamics: the constraints on rearrangements within bacterial genomes, *Genome Biology* 2000, 1

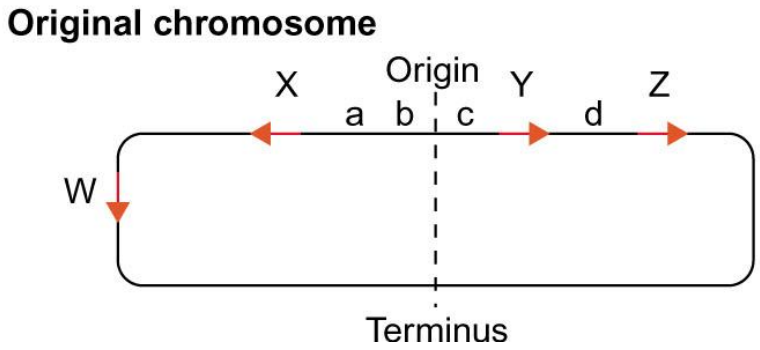


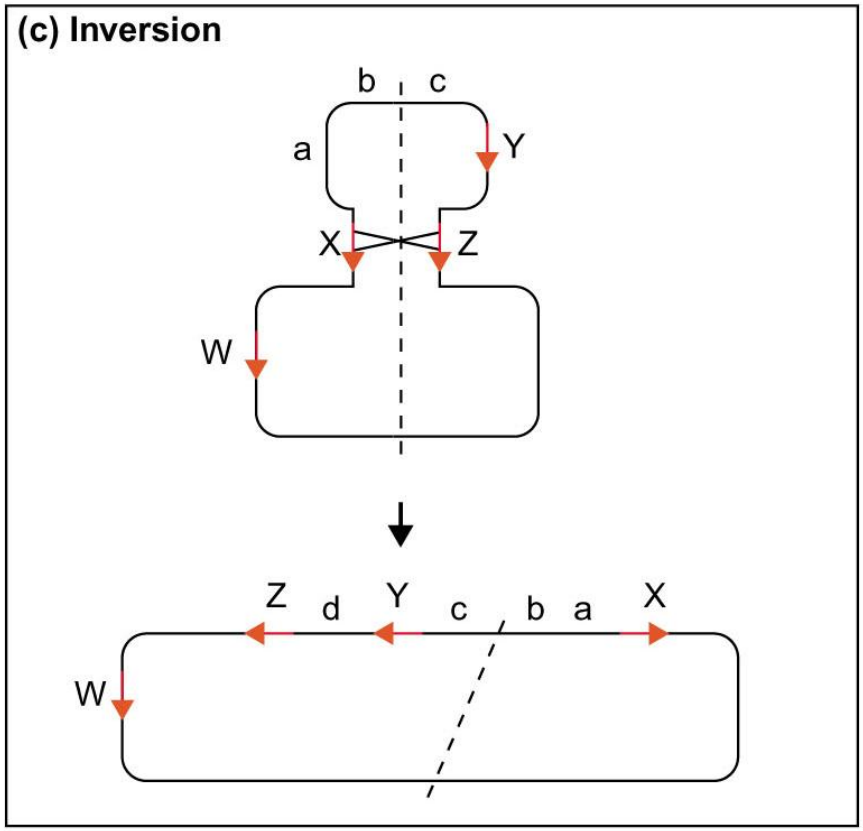
Recombination of two repeat sequences in the *same* chromosome can lead to a fragment translocation

Here sequence d is translocated



Diarmaid Hughes: Evaluating genome dynamics: the constraints on rearrangements within bacterial genomes, *Genome Biology* 2000, 1





Inversion happens when two sequences of *opposite* orientations are recombined.

Diarmaid Hughes: Evaluating genome dynamics: the constraints on rearrangements within bacterial genomes, *Genome Biology* 2000, 1

Example: human vs mouse genome

- Human and mouse genomes share thousands of homologous genes, but they are
  - Arranged in different order
  - Located in different chromosomes
- Examples
  - Human chromosome 6 contains elements from six different mouse chromosomes
  - Analysis of X chromosome indicates that rearrangements have happened primarily *within* chromosome

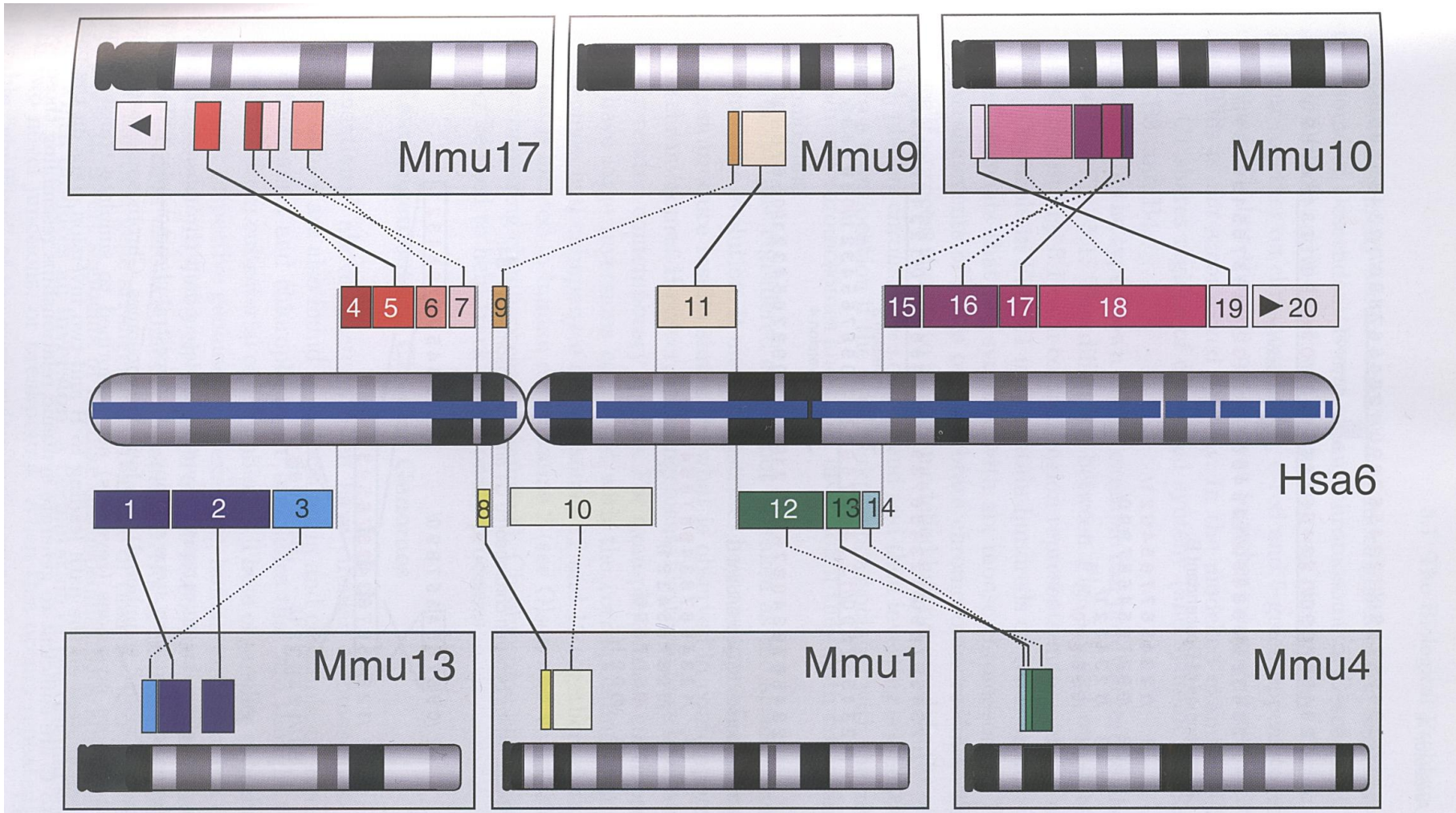


Fig. 5.1. Syntenic blocks conserved between human chromosome Hsa6 and mouse chromosomes. Broken lines indicate regions that appear in inverted orders in the two organisms. Reprinted, with permission, from Gregory SG et al. (2002) *Nature* 418:743–750. Copyright 2002 Nature Publishing Group.

Jones & Pevzner, 2004

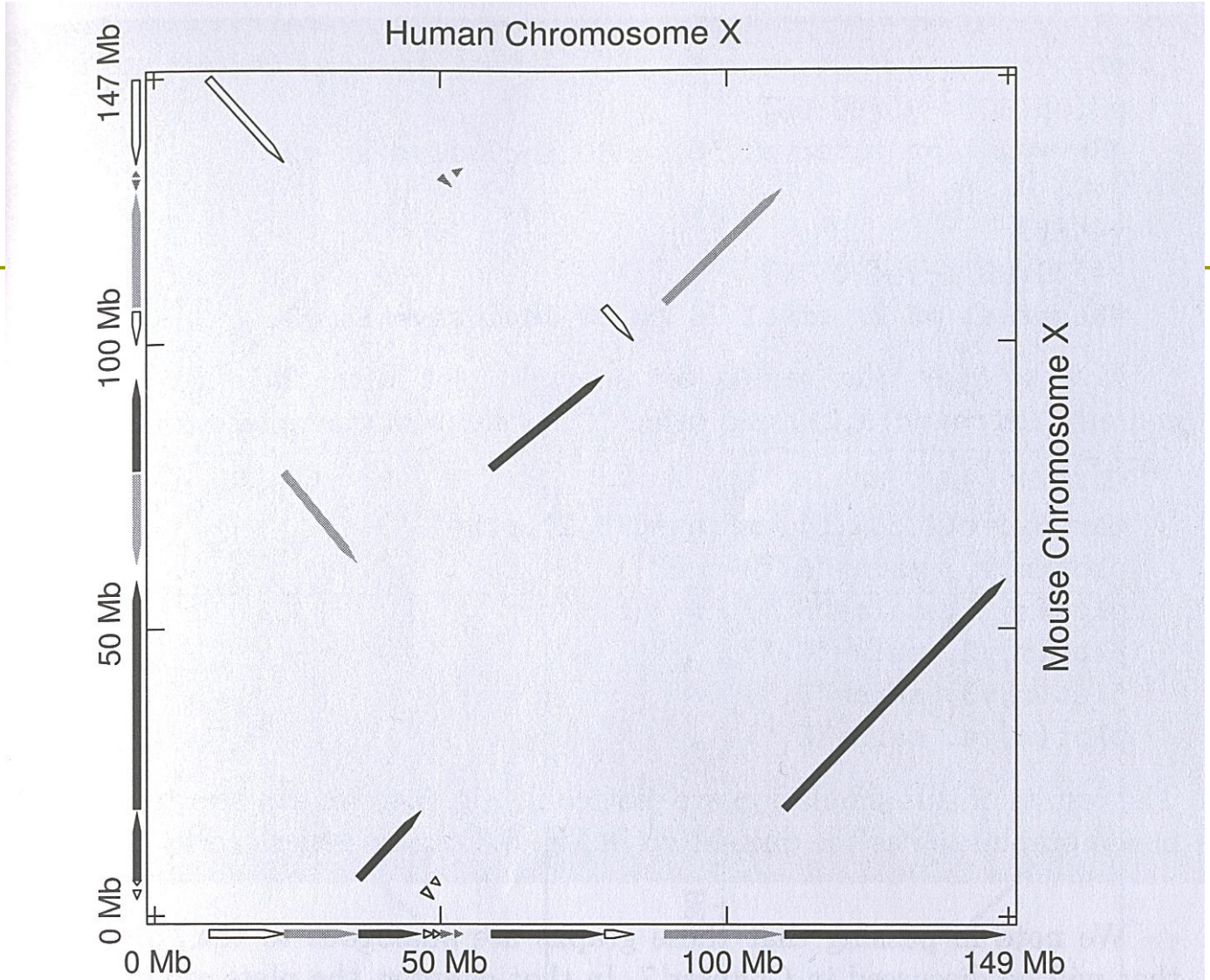


Fig. 5.3. Synteny blocks shared by human and mouse X chromosomes. The arrowhead for each block indicates the direction of increasing coordinate values for the human X chromosome. Reprinted, with permission, from Pevzner P and Tesler G (2003) *Genome Research* 13:37–45. Copyright 2003 Cold Spring Harbor Laboratory Press.

Representing genome rearrangments

- When comparing two genomes, we can find homologous sequences in both using BLAST, for example
- This gives us a map between sequences in both genomes

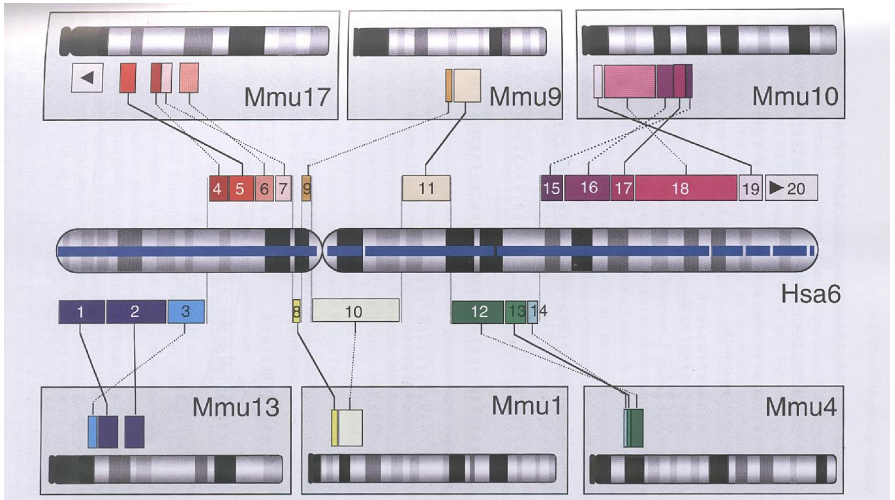
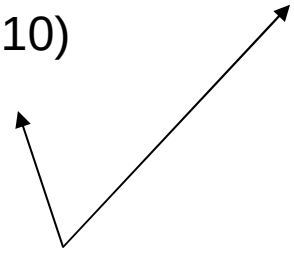


Fig. 5.1. Syntenic blocks conserved between human chromosome Hsa6 and mouse chromosomes. Broken lines indicate regions that appear in inverted orders in the two organisms. Reprinted, with permission, from Gregory SG et al. (2002) *Nature* 418:743–750. Copyright 2002 Nature Publishing Group.

Representing genome rearrangments

- We assign numbers 1,…,n to the found homologous sequences
- By convention, we number the sequences in the first genome by their order of appearance in chromosomes
- If the homolog of *i* is in reverse orientation, it receives number *–i* (*signed data*)
- For example, consider human vs mouse gene numbering on the right

Human		**Mouse**	
1	(gnat2)	12	(inpp1)
2	(nras)	13	(cd28)
3	(ngfb)	14	(fn1)
4	(gba)	15	(pax3)
5	(pklr)	-9	(il10)
6	(at3)	-8	(pdc)
7	(lamc1)	-7	(lamc1)
8	(pdc)	-6	(at3)
9	(il10)		



List order corresponds to *physical order* on chromosomes!

Permutations

- The basic data structure in the study of genome rearrangements is *permutation*
- A permutation of a sequence of *n* numbers is a reordering of the sequence
- For example, 4 1 3 2 5 is a permutation of 1 2 3 4 5

Genome rearrangement problem

- Given two genomes (set of markers), how many
  - duplications,
  - inversions and
  - translocations

  do we need to do to transform the first genome to the second?

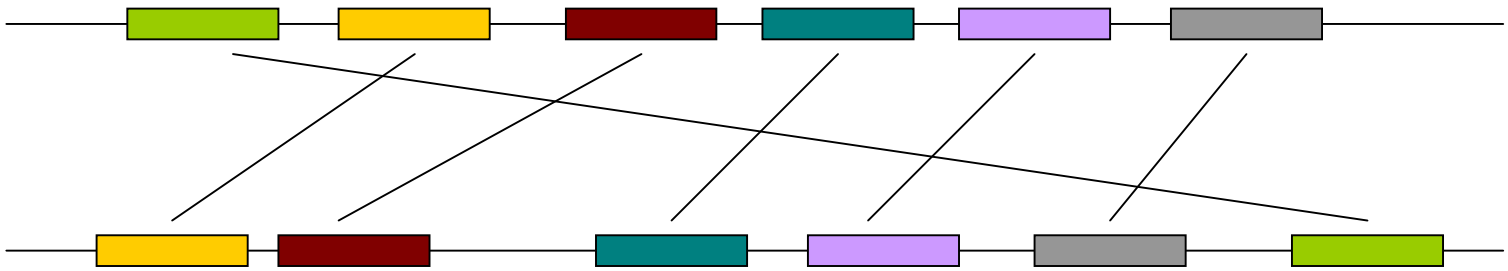
  Minimum number of operations?

  What operations? Which order?

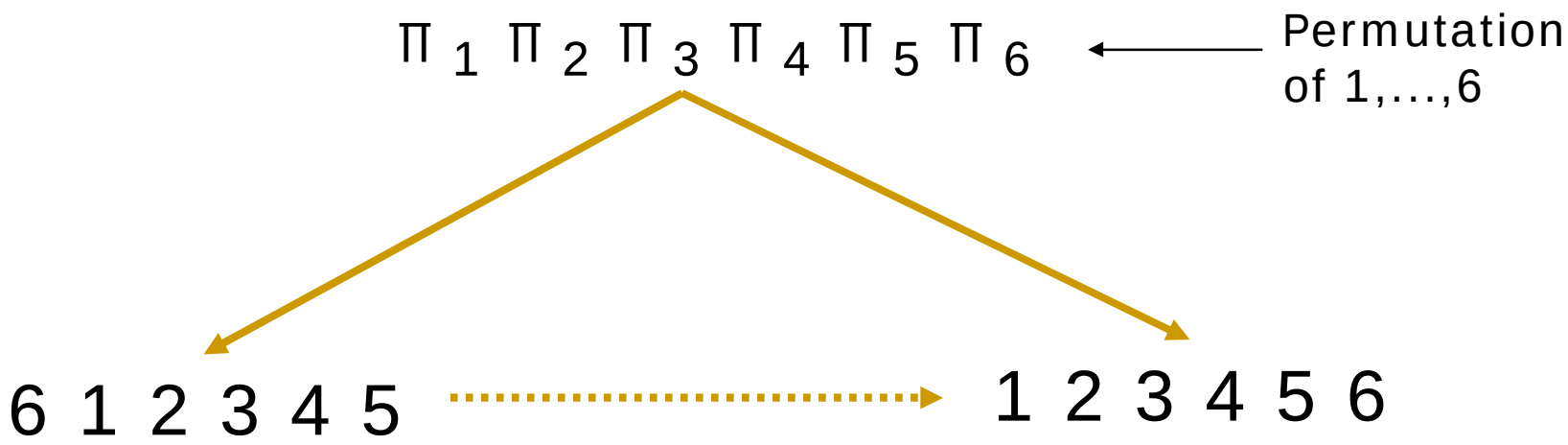
Genome rearrangement problem

#duplications?
#inversions?
#translocations?

6 1 2 3 4 5 → 1 2 3 4 5 6



Genome rearrangement problem

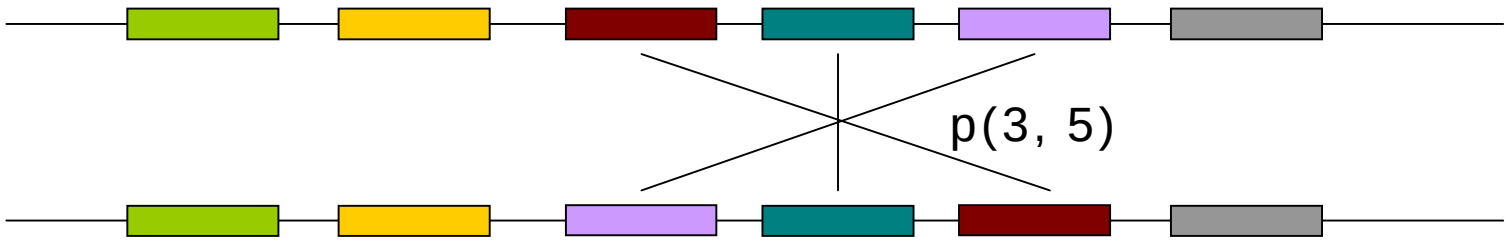


Keep in mind, that the two genomes have been evolved from a common ancestor genome!

Genome rearrangements using reversals (=inversions) only

- Lets consider a simpler problem where we just study reversals with *unsigned data*
- A reversal p(i, j) reverses the order of the segment $\Pi_i \Pi_{i+1} \ldots \Pi_{j-1} \Pi_j$ (indexing starts from 1)
- For example, given permutation

  6 1 2 3 4 5 and reversal p(3, 5) we get permutation 6 1 4 3 2 5



...note that we do not care about exact *positions* on the genome

Reversal distance problem

- Find the shortest series of reversals that, given a permutation Π, transforms it to the *identity* permutation (1, 2, …, n)
- This quantity is denoted by $d(\Pi)$

- Reversal distance for a pair of chromosomes:
  - Find synteny blocks in both
  - Number blocks in the first chromosome to identity
  - Set Π to correspond matching of second chromosome's blocks against the first
  - Find reversal distance

Reversal distance problem: discussion

- If we can find the minimal series of reversals for some pair of genomes
  - Is that what happened during evolution?
  - If not, is it the correct number of reversals?

- In any case, reversal distance gives us a measure of evolutionary distance between the two genomes and species

Solving the problem by sorting

- Our first approach to solve the reversal distance problem:
  - Examine each position i of the permutation
  - At each position, if $\Pi_i \neq i$, do a reversal such that $\Pi_i = i$
- This is a *greedy* approach: we try to choose the best option at each step

Simple reversal sort: example

6 1 2 3 4 5 -> 1 6 2 3 4 5 -> 1 2 6 3 4 5 -> 1 2 3 4 6 5

-> 1 2 3 4 5 6

Reversal series: p(1,2), p(2,3), p(3,4), p(5,6)

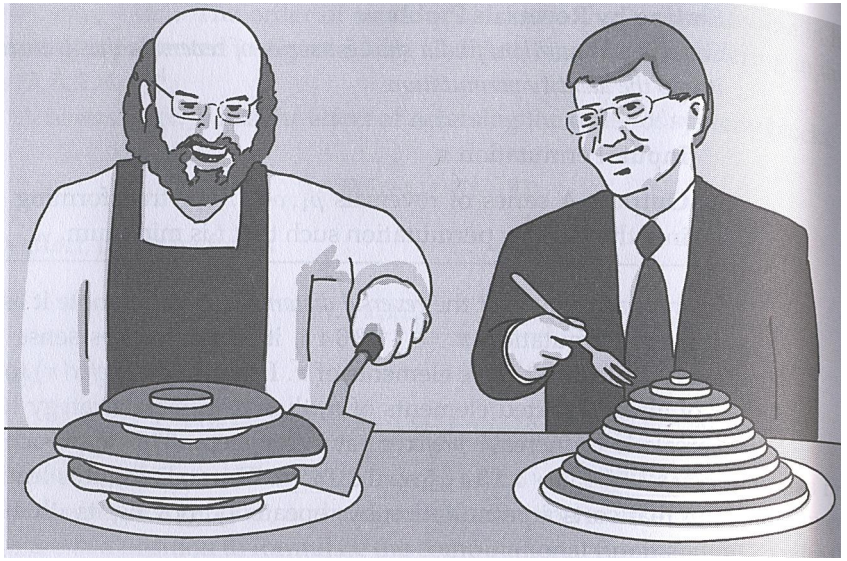
Is d(6 1 2 3 4 5) then 4?

6 1 2 3 4 5 -> 5 4 3 2 1 6 -> 1 2 3 4 5 6

D(6 1 2 3 4 5) = 2

Pancake flipping problem

- No pancake made by the chef is of the same size
- Pancakes need to be rearranged before delivery
- Flipping operation: take some from the top and flip them over
- This corresponds to always reversing the sequence *prefix*



1 2 3 6 4 5 -> 6 3 2 1 4 5 ->

5 4 1 2 3 6 -> 3 2 1 4 5 6 ->

1 2 3 4 5 6