

Genome Rearrangements

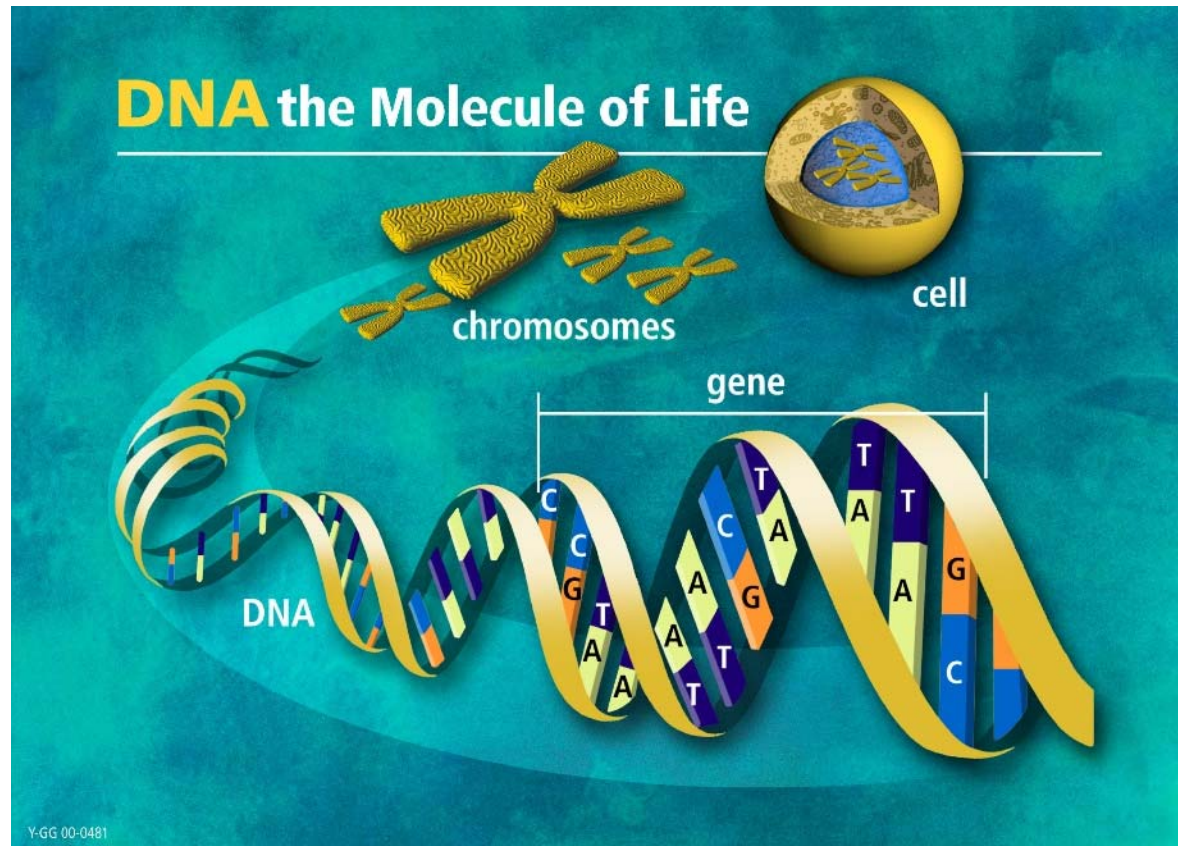
Guillaume Bourque
Genome Institute of Singapore

April 2009

Outline

- Sequencing large genomes
- Motivation
- Definitions and pairwise algorithms for genome rearrangements
- Genome rearrangement phylogenies
- Applications and latest developments

DNA sequence



Source: www.ornl.gov

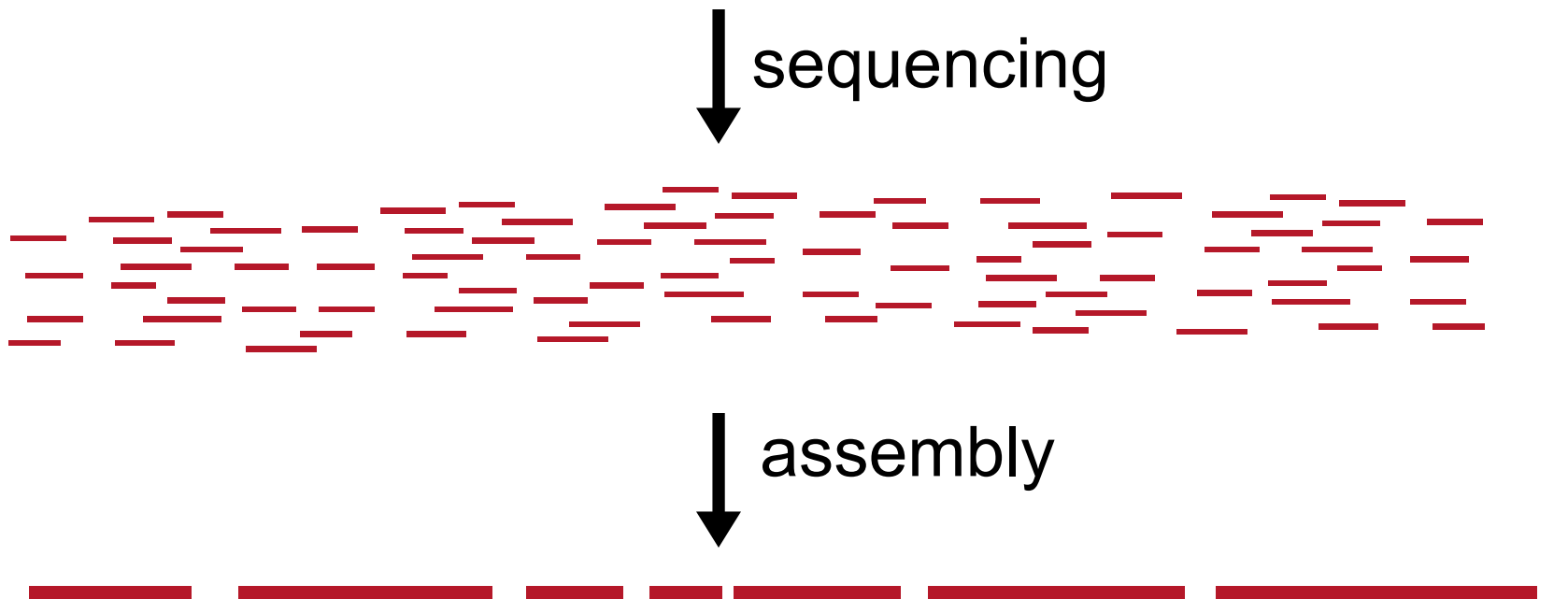
Longest contiguous DNA

1977	Bacteriophage	5,375
1981	Human Mitochondrion	16,569
1984	Epstein-Barr Virus	172,281
1992	S. cerevisiae chromosome 3	315,339
1997	Escherichia coli	4,639,221
1999	Human chromosome 22q	23,051,000
2003	Human chromosome 14	87,410,661

Source: www.tigr.org/tdb/contig_list.shtml

Sequencing and assembly

Genome



Assembling a genome

Assume you want to decode a sequence of length 10. You have read 4 fragments from the sequence:

T A G C C

A T G C

G C A T A

A T A G

Assembling a genome

Assume you want to decode a sequence of length 10. You have read 4 fragments from the sequence:

T A G C C

A T G C

A T A G

G C A T A

Assembling a genome

Assume you want to decode a sequence of length 10. You have read 4 fragments from the sequence:

T A G C C

A T G C

G C A T A

A T A G

Assembling a genome

Assume you want to decode a sequence of length 10. You have read 4 fragments from the sequence:

A T G C

G C A T A

A T A G

T A G C C

Assembling a genome

Assume you want to decode a sequence of length 10. You have read 4 fragments from the sequence:

```
      G  C  A  T  A
        A  T  A  G
          T  A  G  C  C
A  T  G  C
```

Assembling a genome

Assume you want to decode a sequence of length 10. You have read 4 fragments from the sequence:

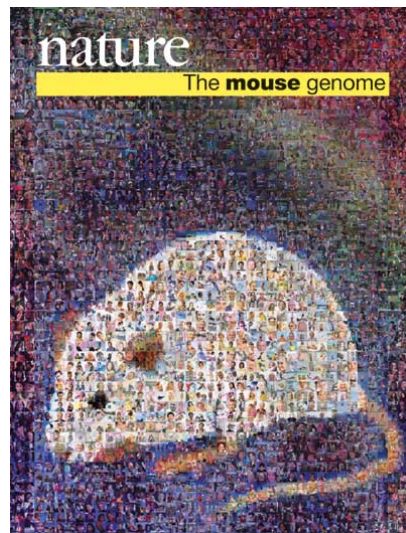
```
      G C A T A
        A T A G
          T A G C C
A T G C
A T G C A T A G C C
```

Big genomes

Human, 2001



Mouse, 2002



Rat, 2004



Chicken, 2004



Each of these genomes has over 1 billion base pairs.

High priority list

- Mammals
 - Chimpanzee, Rhesus Macaque
 - Dog, Cat
 - Cow, Pig, Rabbit
 - Opossum, Elephant, Armadillo
 - ...
- Others
 - Honey Bee
 - Sea Urchin
 - ...

Motivation



Science, 1999

GENOME REVIEW

The Promise of Comparative Genomics in Mammals

Stephen J. O'Brien,¹ Marilyn Menotti-Raymond,¹ William J. Murphy,¹ William G. Nash,¹ Johannes Wienberg,¹ Roscoe Stanyon,¹ Neal G. Copeland,² Nancy A. Jenkins,² James E. Womack,³ Jennifer A. Marshall Graves⁴

Dense genetic maps of human, mouse, and rat genomes that are based on coding genes and on microsatellite and single-nucleotide polymorphism markers have been complemented by precise gene homolog alignment with moderate-resolution maps of livestock, companion animals, and additional mammal species. Comparative genetic assessment expands the utility of these maps in gene discovery, in functional genomics, and in tracking the evolutionary forces that sculpted the genome organization of modern mammalian species.

about 3.2 billion nucleotide pairs. Chromosome numbers range from a low of three pairs ($2N = 6$ in the Indian muntjac, *Muntiacus muntjak*) to a high of 67 pairs ($2N = 134$ in the black rhinoceros, *Diceros bicornis*). Gene maps have been constructed in human, mouse, and about 30 other mammal species for two general reasons: first, as a resource

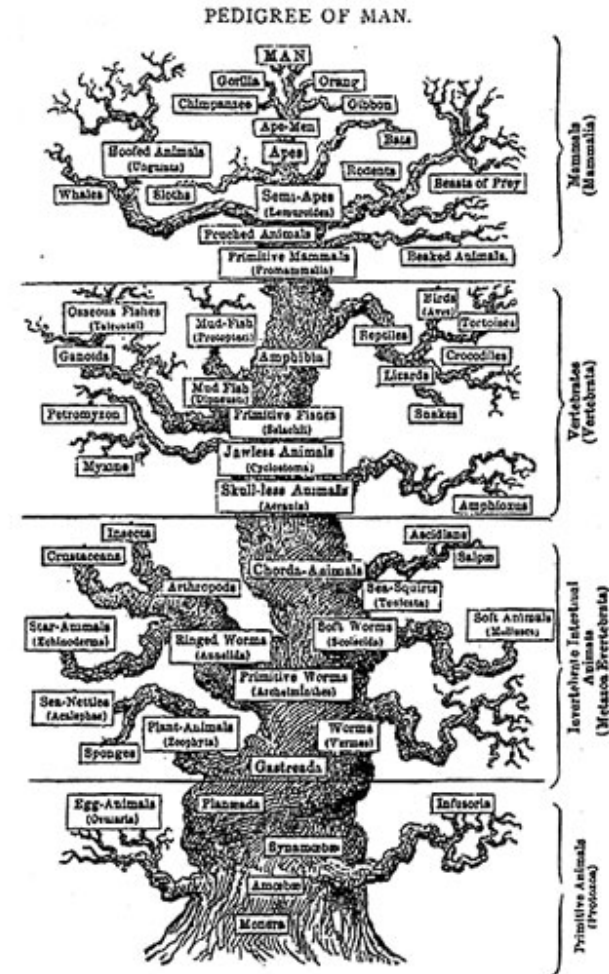
High hopes

- Explain the physical clustering of gene families (regulation, editing or retention).
- Understand whether even longer linkage associations were preserved by chance or by selection (developmental or functional).
- **Resolve the mammalian phylogeny using genomic segment exchanges as characters.**
- Discover molecular fossils of precipitous genomic events.
- Identify genetic determinants of reproductive isolation, adaptation, survival and species formation.

O'Brien et al, Science 1999

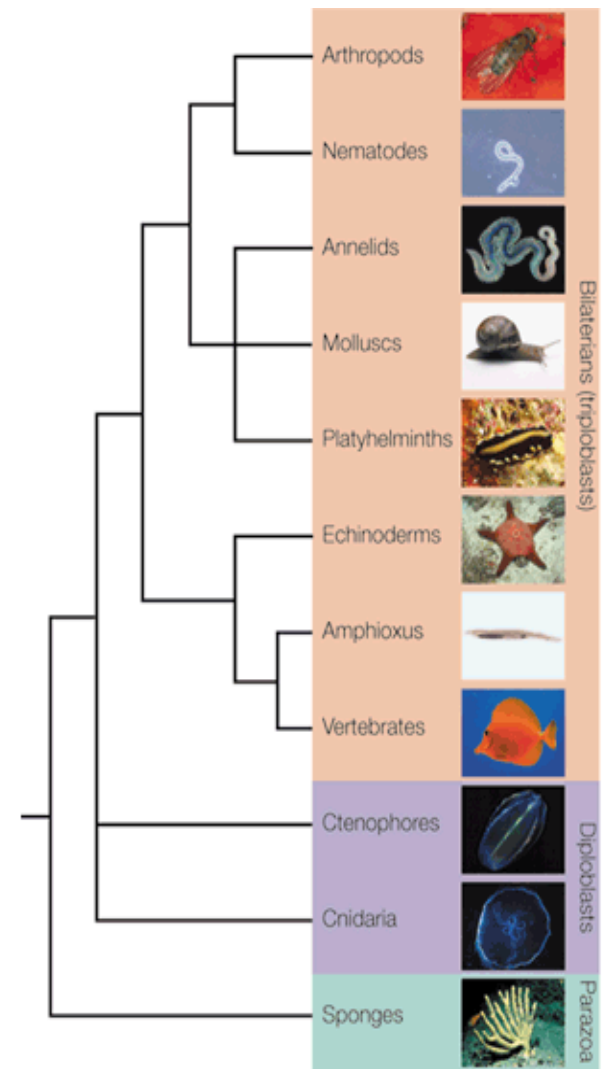
Early phylogenetic trees

- Evolution and genetic relationships between species have been represented by **phylogenetic trees** for over a century.
- Haeckel's "Pedigree of Man" Diagram, 1866.

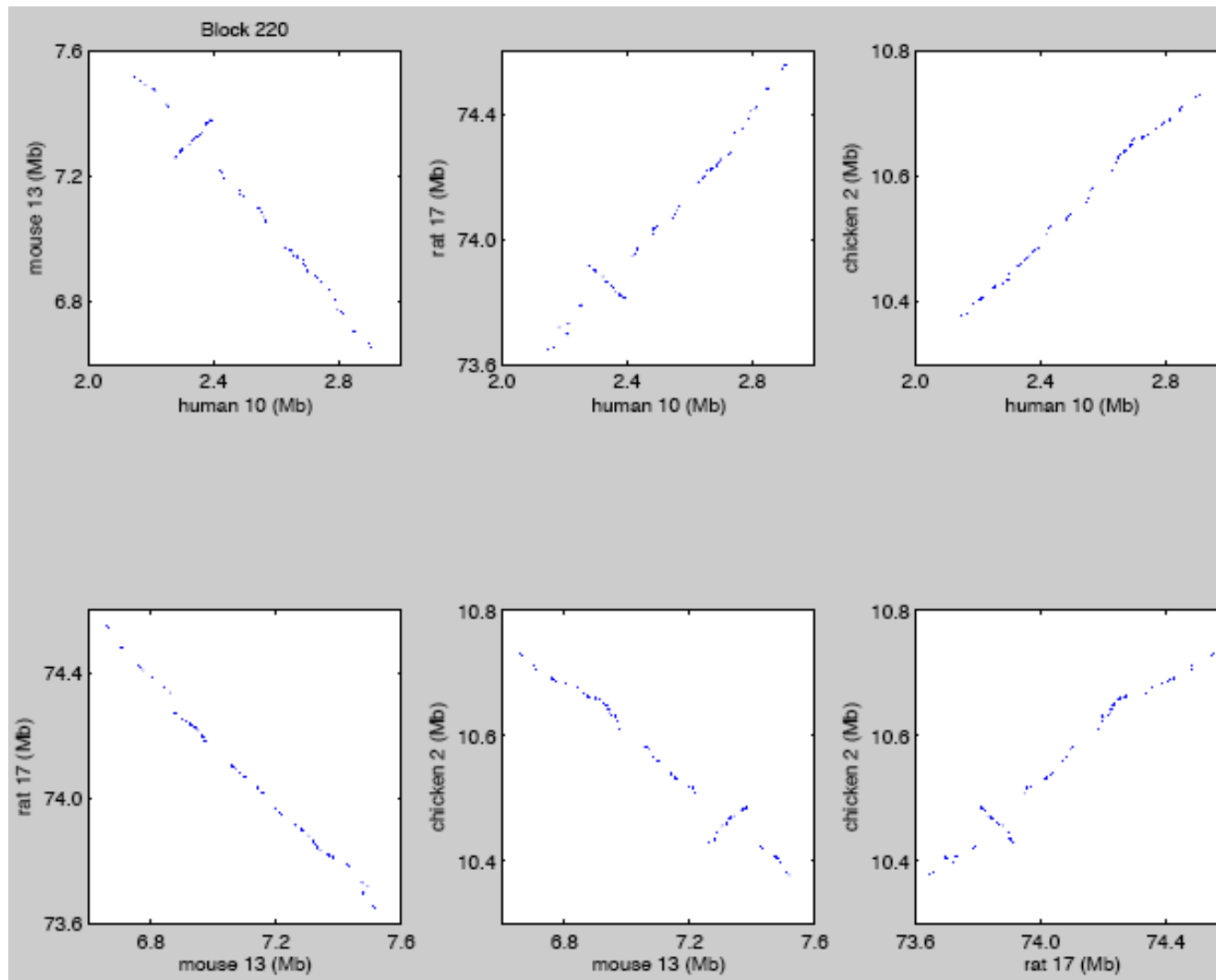


Gene trees

- With sequencing, people started constructing trees based on **DNA sequences**.
- E.g. this phylogeny of selected metazoans based on ribosomal DNA and Hox gene sequences. Ferrier and Holland, 2001.



Large-scale differences



Outline

- Sequencing large genomes
- Motivation
- Definitions and pairwise algorithms for genome rearrangements
- Genome rearrangement phylogenies
- Applications

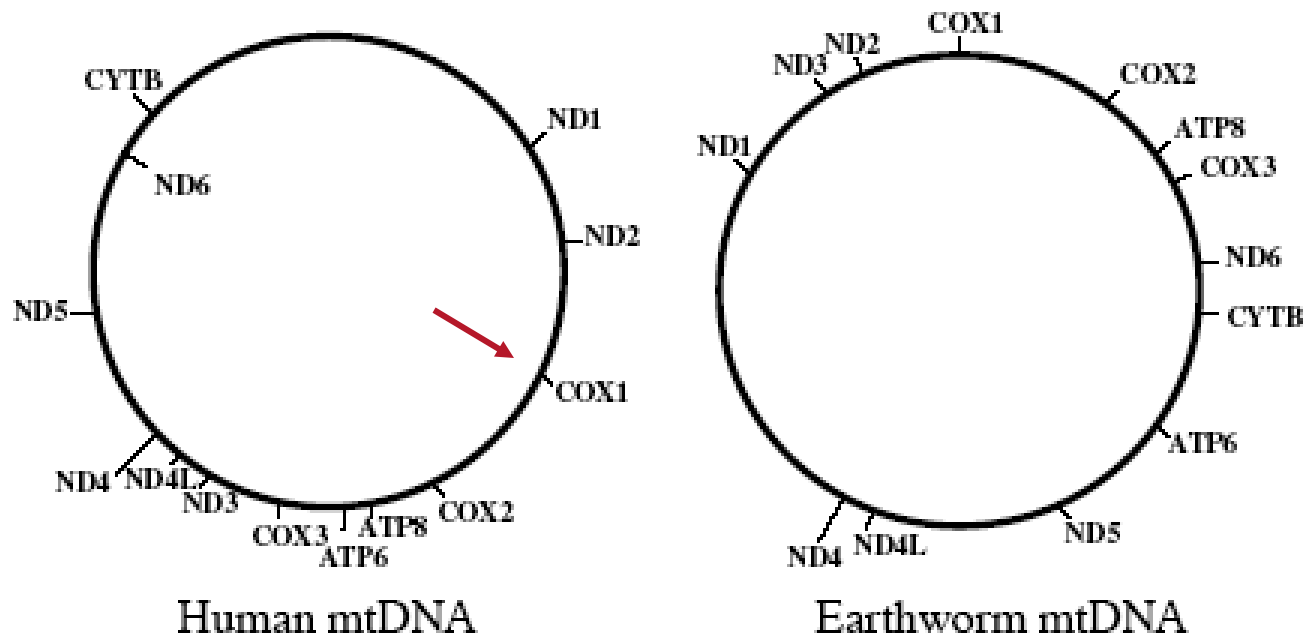
Genome representation

In the context of genome rearrangements, genomes are typically viewed as *signed permutations* where each integer corresponds to a unique gene and the sign corresponds to its orientation (strand). For instance, a segment:



Would be represented by a permutation: ... 1 2 -3

Two mtDNA



Human:	1	2	3	4	5	6	7	8	9	10	11	12	13
Earthworm:	1	2	3	5	-10	11	4	9	7	8	12	6	13

Chromosomal mutations

Mutation Type	Before	After	Impact
---------------	--------	-------	--------

Reversal	1 2 3 4 5 6	\Rightarrow 1 2 -5 -4 -3 6	gene order
----------	---	------------------------------	------------

Translocation	1 2 3 4 5 6 7 8	\Rightarrow 1 2 8 6 7 3 4 5	gene order
---------------	--	----------------------------------	------------

Fusion	1 2 3 4 5 6	\Rightarrow 1 2 3 4 5 6	gene order
--------	--	---------------------------	------------

Fission	1 2 3 4 5 6	\Rightarrow 1 2 3 4 5 6	gene order
---------	---	------------------------------	------------

Transposition	1 2 3 4 5 6	\Rightarrow 1 4 5 2 3 6	gene order
---------------	---	---------------------------	------------

Inverted Transpo.	1 2 3 4 5 6	\Rightarrow 1 4 5 -3 -2 6	gene order
-------------------	---	-----------------------------	------------

Insertion	1 2 3 4 5 6	\Rightarrow 1 2 3 4 7 5 6	gene content
-----------	---	-----------------------------	--------------

Deletion	1 2 3 4 5 6	\Rightarrow 1 4 5 6	gene content
----------	---	-----------------------	--------------

Duplication	1 2 3 4 5 6	\Rightarrow 1 2 3 4 3' 4' 5 6	gene content
-------------	---	---------------------------------	--------------

Distance between two genomes

- Breakpoint distance
 - Conservation distance
 - Common intervals
 - Conserved intervals
 - Rearrangement distance
- Model independent

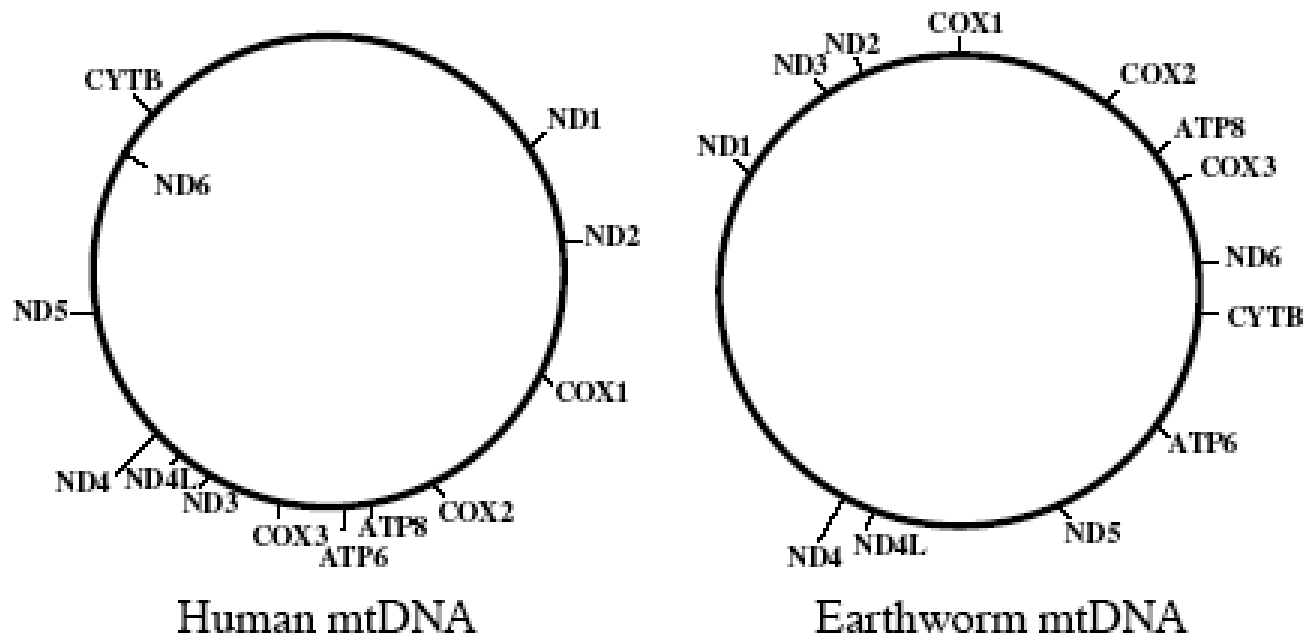
Breakpoint distance

- The breakpoint distance compares two permutations by directly counting the number of gene order disruptions between two genomes.
- Given two sequences of size n , π and γ , extend both permutations so that they start with 0 and end with $n+1$.
- The *breakpoint distance* is defined as the number of pairs $(\gamma_i \gamma_{i+1})$, $0 \leq i \leq n$, such that neither $(\gamma_i \gamma_{i+1})$ nor $(-\gamma_{i+1} - \gamma_i)$ appears in π .

1 -4 -3 -5 2 6 Vs. 1 2 3 4 5 6
 ↑ ↑ ↑ ↑
 ↑ ↑

Breakpoint distance = 4

Two mtDNA



Breakpoint distance = 9

Breakpoint distance

- Easily computable in linear time.
- Does not require any assumptions about the underlying rearrangement mechanisms.
- Nadeau and Taylor (1984) successfully used this criterion, on a very limited set of markers, to make a prophetic prediction for the approximate number of conserved segments between human and mouse .

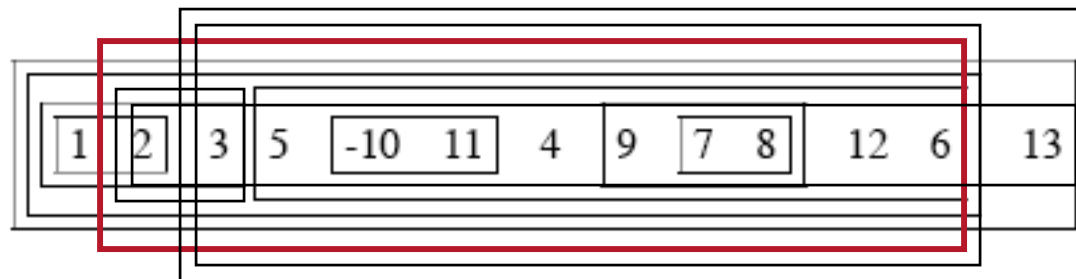
Conservation distance

- Generalization of the breakpoint distance but considers intervals instead of just adjacencies.
- Two different criterias:
 - Common intervals
 - Conserved intervals
- Two important properties
 - It can be directly defined on a set of more than two genomes and allows the identification of shared features in a family of organisms.
 - It does not rely on an a priori model of rearrangements.

Common intervals

Given two signed permutations, a *common interval* is a set of two or more integers that is an interval in both permutations.

Human:	1	2	3	4	5	6	7	8	9	10	11	12	13
Earthworm:	1	2	3	5	-10	11	4	9	7	8	12	6	13

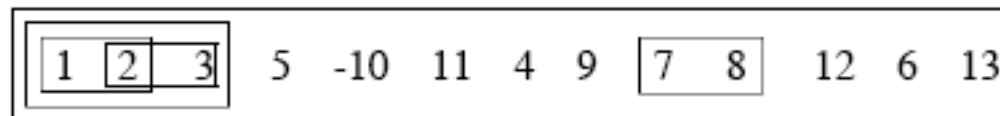


14 common intervals

Conserved intervals

Given two signed permutations, π and γ , a *conserved interval* is an interval $[a, b]$ such that a precedes b or $-b$ precedes $-a$ in π and γ , and the set of elements, without signs, between a and b is the same in π and γ .

Human:	1	2	3	4	5	6	7	8	9	10	11	12	13
Earthworm:	1	2	3	5	-10	11	4	9	7	8	12	6	13



5 conserved intervals

Rearrangement distance

The types of rearrangements events that we consider are:

- Reversals (inversions)
- Translocations
- Fusions
- Fissions



Multichromosomal
genomes

Examples of rearrangements



Reversal

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

Translocation

1 2 3 4
5 6  1 2 6
5 3 4

Fusion

1 2 3 4
5 6  1 2 3 4 5 6


Fission

Sorting by reversals

Polynomial algorithm for computing the rearrangement distance and the most parsimonious scenario between 2 unichromosomal genomes (Hannenhalli and Pevzner 1995). For example:

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



1 2 3 4 5 6 7 8

Sorting by reversals

Polynomial algorithm for computing the rearrangement distance and the most parsimonious scenario between 2 unichromosomal genomes (Hannenhalli and Pevzner 1995). For example:

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

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

1 2 3 6 7 -4 -5 8

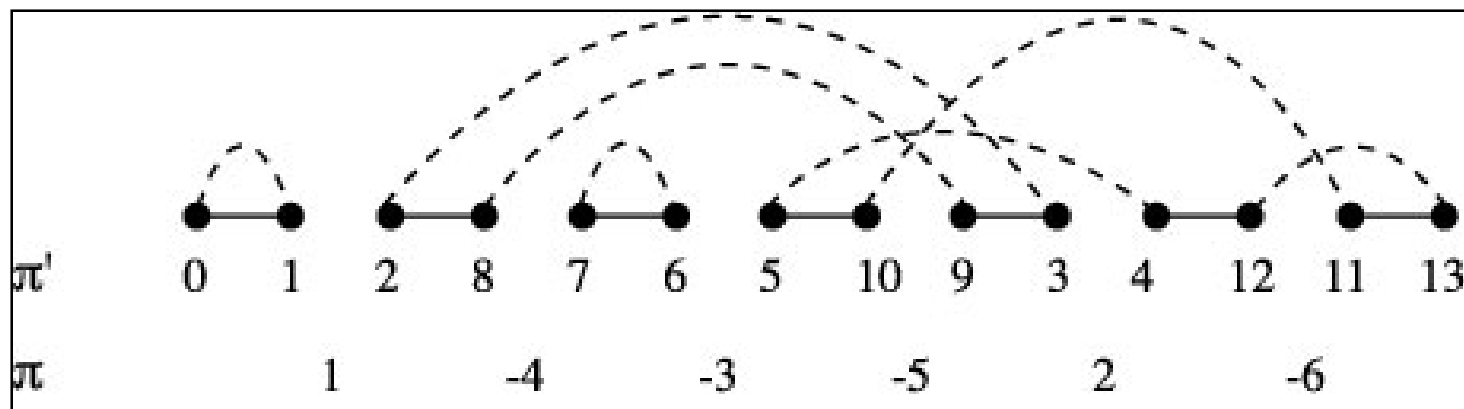
1 2 3 4 -7 -6 -5 8

1 2 3 4 5 6 7 8

Breakpoint graph

To construct the breakpoint graph of π which we wish to sort with respect to the identity permutation:

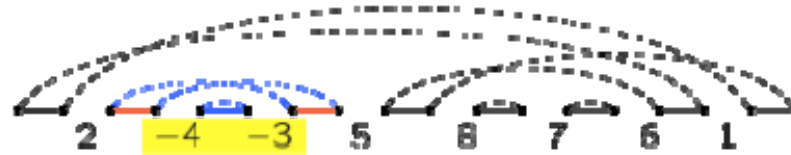
- Replace each positive element $+i$ by $2i-1$ and $2i$ and each negative element $-i$ by $2i$ and $2i-1$ in a new permutation π' .
- Add 0 and $2n + 1$ at the end of the permutation π' .
- Add a *black* edge between π'_{2i} and π'_{2i} . Add a *gray* edge between $2i$ and $2i + 1$.



Sorting with the breakpoint graph

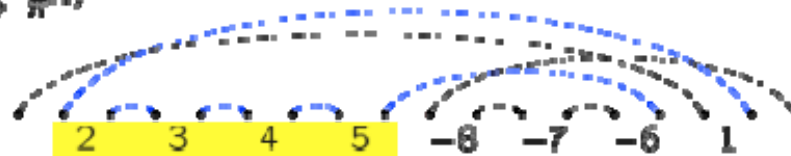
Step 0: $\pi^{(0)} = \pi$

$G(\pi, \gamma)$



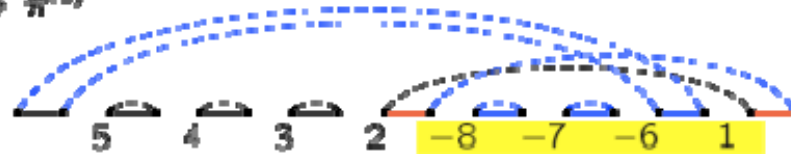
Step 1: Reversal $\rightarrow \pi^{(1)}$

$G(\pi^{(1)}, \gamma)$



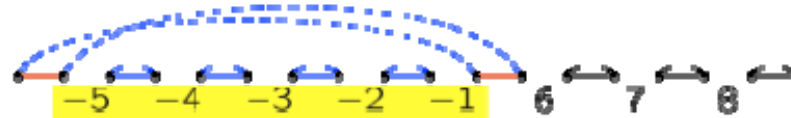
Step 2: Reversal $\rightarrow \pi^{(2)}$

$G(\pi^{(2)}, \gamma)$



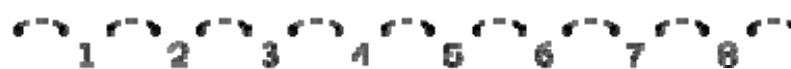
Step 3: Reversal $\rightarrow \pi^{(3)}$

$G(\pi^{(3)}, \gamma)$



Step 4: Reversal $\rightarrow \pi^{(4)} = \gamma$

$G(\gamma, \gamma)$



(slide by Glenn Tesler)

Reversal distance

The reversal distance:

$$d(\pi) = n + 1 - c(\pi) + h(\pi) + f(\pi)$$

Where $c(\square)$, $h(\square)$ and $f(\square)$ are parameters of the breakpoint graph:

- $c(\square)$ is the number of « cycles »
- $h(\square)$ is the number of « hurdles »
- $f(\square)$ is 1 if the permutation satisfies the properties of a « fortress » and 0 otherwise.

Sorting by rearrangements

-3 7 -1 4 5 6 -2



1 2 3 4 5 6 7

Sorting by rearrangements

-3 7 -1 4 5 6 -2 translocation

-3 -2 -1 4 5 6 7 fusion

-3 -2 -1 4 5 6 7 fission

-3 -2 -1 4 5 6 7 reversal

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

Multichromosomal rearrangements



By *concatenating* chromosomes, this may be mimicked by a single reversal:

$$\begin{array}{cc}
 (\textcolor{red}{5} \textcolor{red}{9} \textcolor{orange}{4} \textcolor{orange}{10}) & (\textcolor{blue}{2} \textcolor{blue}{-7} \textcolor{blue}{-11} \textcolor{green}{1} \textcolor{green}{6}) \\
 (\textcolor{red}{5} \textcolor{red}{9} \textcolor{blue}{11} \textcolor{blue}{7} \textcolor{blue}{-2}) & (\textcolor{orange}{-10} \textcolor{orange}{-4} \textcolor{green}{1} \textcolor{green}{6})
 \end{array}$$

Clinical: A specific translocation (BCR/ABL in chr. 9/22) is observed in 15—20% of leukemia patients.

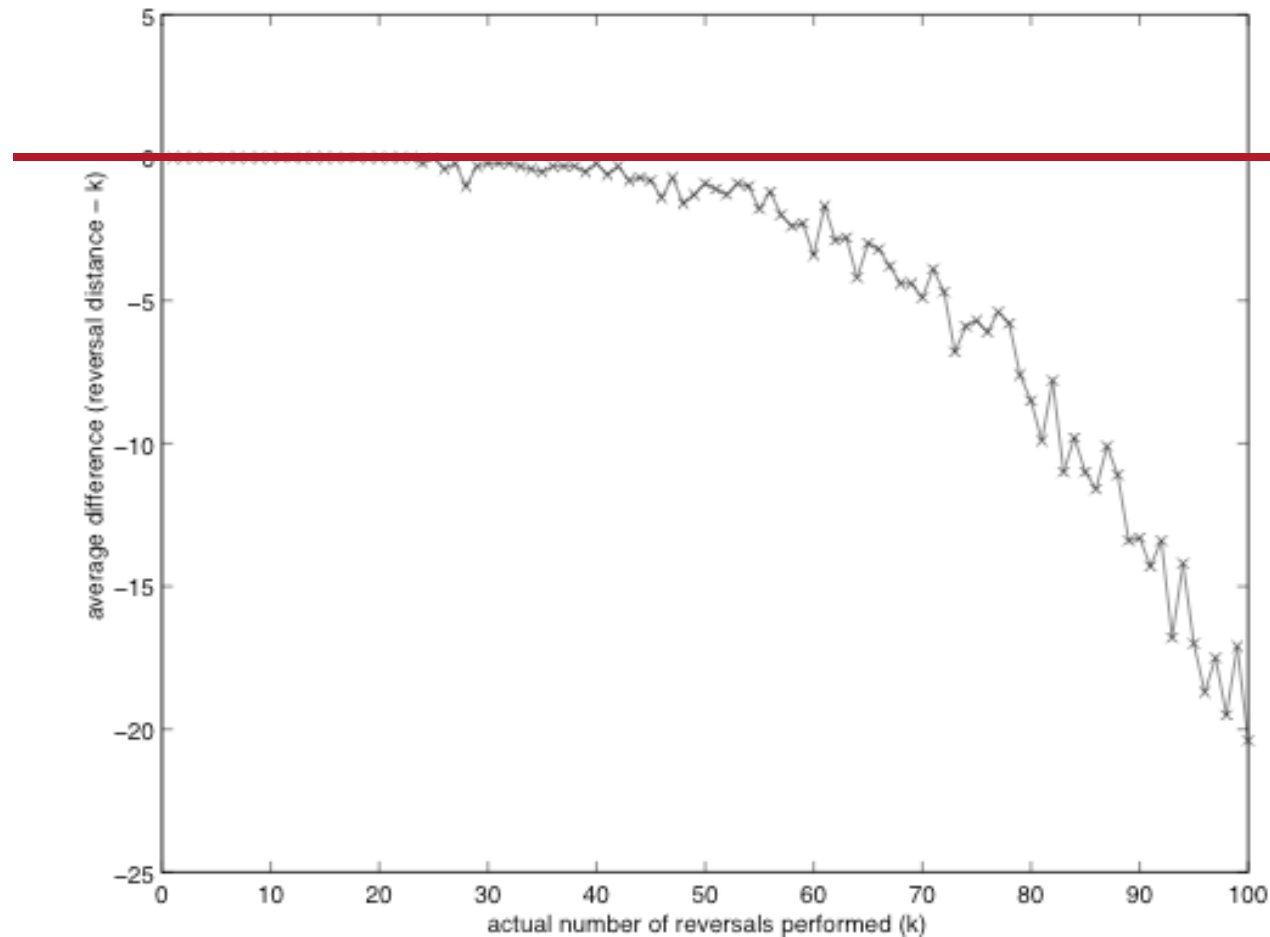
Multichromosomal rearrangements

Most concatenates don't work!

$$\begin{array}{rcl}
 (\textcolor{red}{5} \textcolor{red}{9} \textcolor{brown}{4} \textcolor{brown}{10}) & (\textcolor{green}{-6} \textcolor{green}{-1} \textcolor{blue}{11} \textcolor{blue}{7} \textcolor{blue}{-2}) \\
 \hline
 (\textcolor{red}{5} \textcolor{red}{9} \textcolor{green}{1} \textcolor{green}{6}) & (\textcolor{brown}{-10} \textcolor{brown}{-4} \textcolor{blue}{11} \textcolor{blue}{7} \textcolor{blue}{-2}) \\
 \hline
 (\textcolor{red}{5} \textcolor{red}{9} \textcolor{green}{1} \textcolor{green}{6}) & (\textcolor{blue}{2} \textcolor{blue}{-7} \textcolor{blue}{-11} \textcolor{brown}{4} \textcolor{brown}{10}) \\
 \hline
 (\textcolor{red}{5} \textcolor{red}{9} \textcolor{blue}{11} \textcolor{blue}{7} \textcolor{blue}{-2}) & (\textcolor{green}{-6} \textcolor{green}{-1} \textcolor{brown}{4} \textcolor{brown}{10})
 \end{array}$$

- These concatenates required 3 reversals instead of 1!
- The second reversal just flipped a whole chromosome to position it correctly;
*this is an artifact of our genome representation,
not a biological event.*
- We want to avoid such extra steps and artifacts.

Most parsimonious assumption



Outline

- Sequencing large genomes
- Motivation
- Definitions and pairwise algorithms for genome rearrangements
- Genome rearrangement phylogenies
- Applications

Genome rearrangement phylogenies

- Distance-based methods
- Maximum likelihood methods
- Maximum parsimony methods

Distance-based methods

- Generic approach to construct trees from any pairwise distance matrix.
- Neighbor-Joining, Weighbor, etc.
- Polynomial time inference of phylogenies.
- But...
 - No labeling of internal nodes.
 - No rearrangement scenario associated with solution.
 - May lead to unrealistic solutions...

Maximum likelihood methods

- Requires detailed assumptions of mechanisms and rate of evolution.
- Computationally intensive.
- Provide global picture of solution space.
- So far has been limited to few applications and only to unichromosomal genomes.
- Promising developments by Larget et al., 2005, who developed a program called BADGER.

Maximum parsimony methods

Given m genomes, the *Multiple Genome Rearrangement Problem* is to recover the phylogenetic tree T and the ancestral gene orders that minimize $D(T)$, the total rearrangement distances over all the edges of the tree.

$$D(T) = \sum_{(\pi, \gamma) \in T} d(\pi, \gamma)$$

Similar for breakpoint distance or conservation distance...

Breakpoint analysis

- Blanchette and Sankoff, 1997, developed a method to recover the ancestor of the 3 genomes by minimizing the **breakpoint distance**.
- Their approach reduce this problem to an instance of the Traveling Salesman Problem.
- Blanchette and Sankoff, 1999, generalized their method for m genomes and called it **Breakpoint Analysis**.
- Moret et al. 2000, improved and reimplemented the method in a program called **GRAPPA**.

MGR algorithm

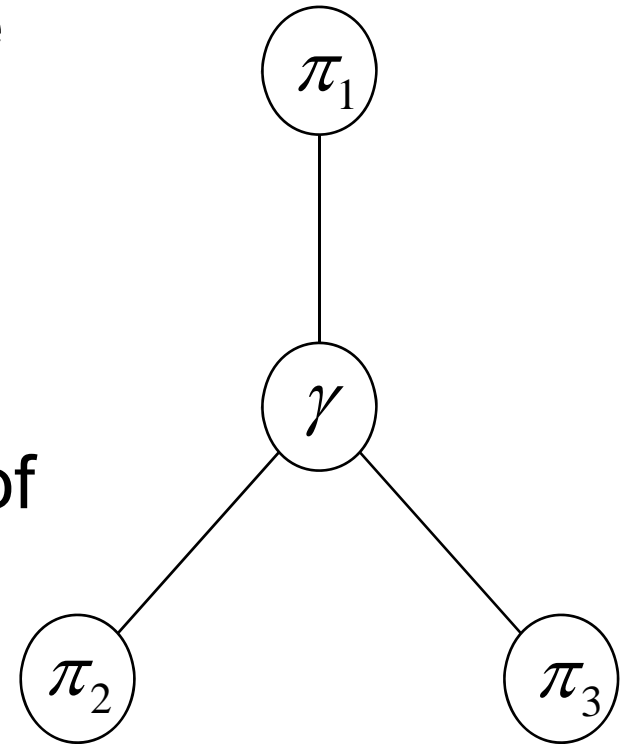
- Multiple Genome Rearrangements (MGR)
 - extends on the Hannenhalli-Pevzner theory
 - applicable to 3 or more genomes
 - works for uni- and multichromosomal genomes
 - can be used for phylogenetic tree reconstruction
- Bourque and Pevzner, *Genome Research*, 2002.

Median problem

The special case of the Multiple Genome Rearrangement Problem with 3 genomes is known as the **Median Problem**.

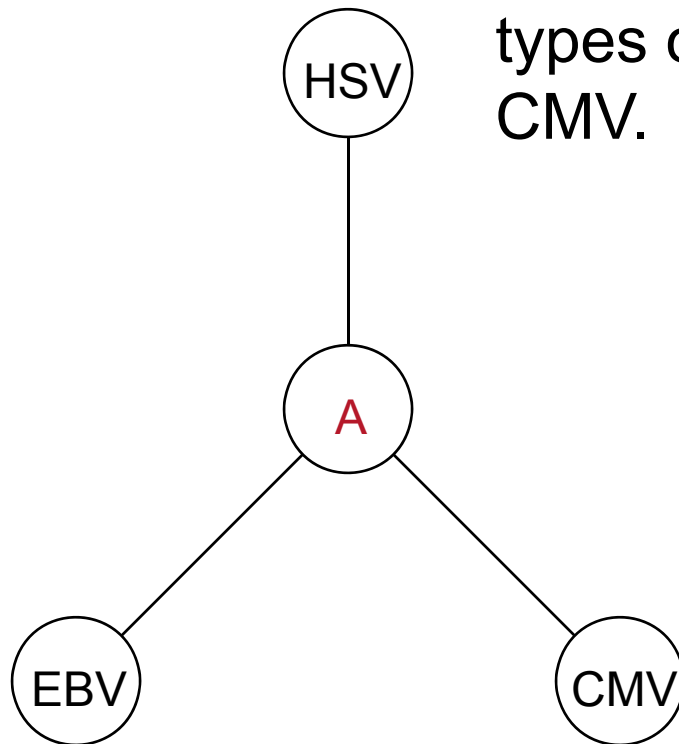
It is equivalent to the recovery of the ancestral gene order, γ , minimizing:

$$D(T) = d(\pi_1, \gamma) + d(\pi_2, \gamma) + d(\pi_3, \gamma)$$



MGR: main idea of algorithm

We have the gene order of 3 different types of herpes viruses: HSV, EBV and CMV.

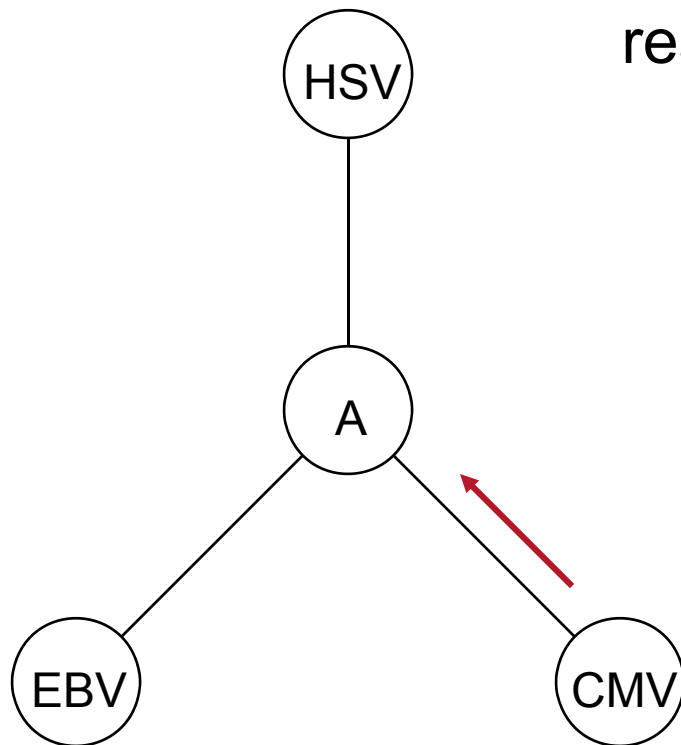


The problem is find the gene order of A, their ancestor.

To recover A, we want to recover the history of rearrangements.

MGR: finding rearrangements

How can we find the most “recent” rearrangement that happened to CMV?

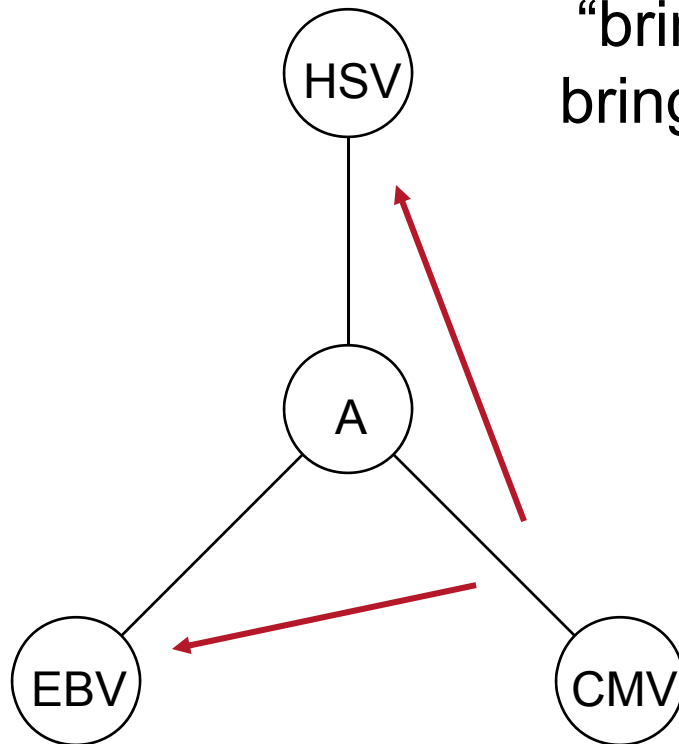


The hope is that it is a rearrangement that “brings” CMV closer to A, a rearrangement that reduces $d(\text{CMV}, A)$.

The problem is that the ancestor A is unknown!

MGR: finding rearrangements

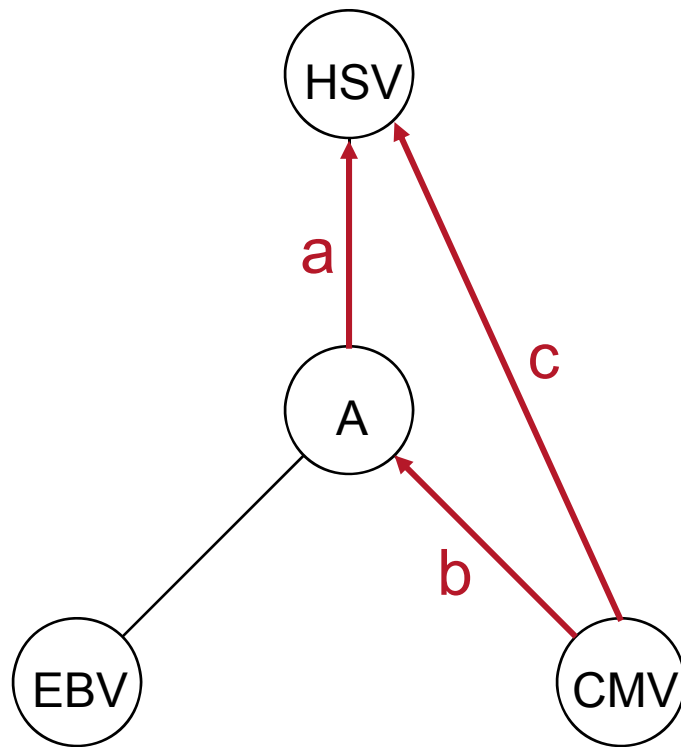
Fortunately, if the correct rearrangement “brings” CMV closer to A, it probably also brings it closer to HSV and EBV.



In other words, it probably also reduces $d(\text{CMV}, \text{HSV})$ and $d(\text{CMV}, \text{EBV})$.

These rearrangements are what we call: **good rearrangements**

Additive trees



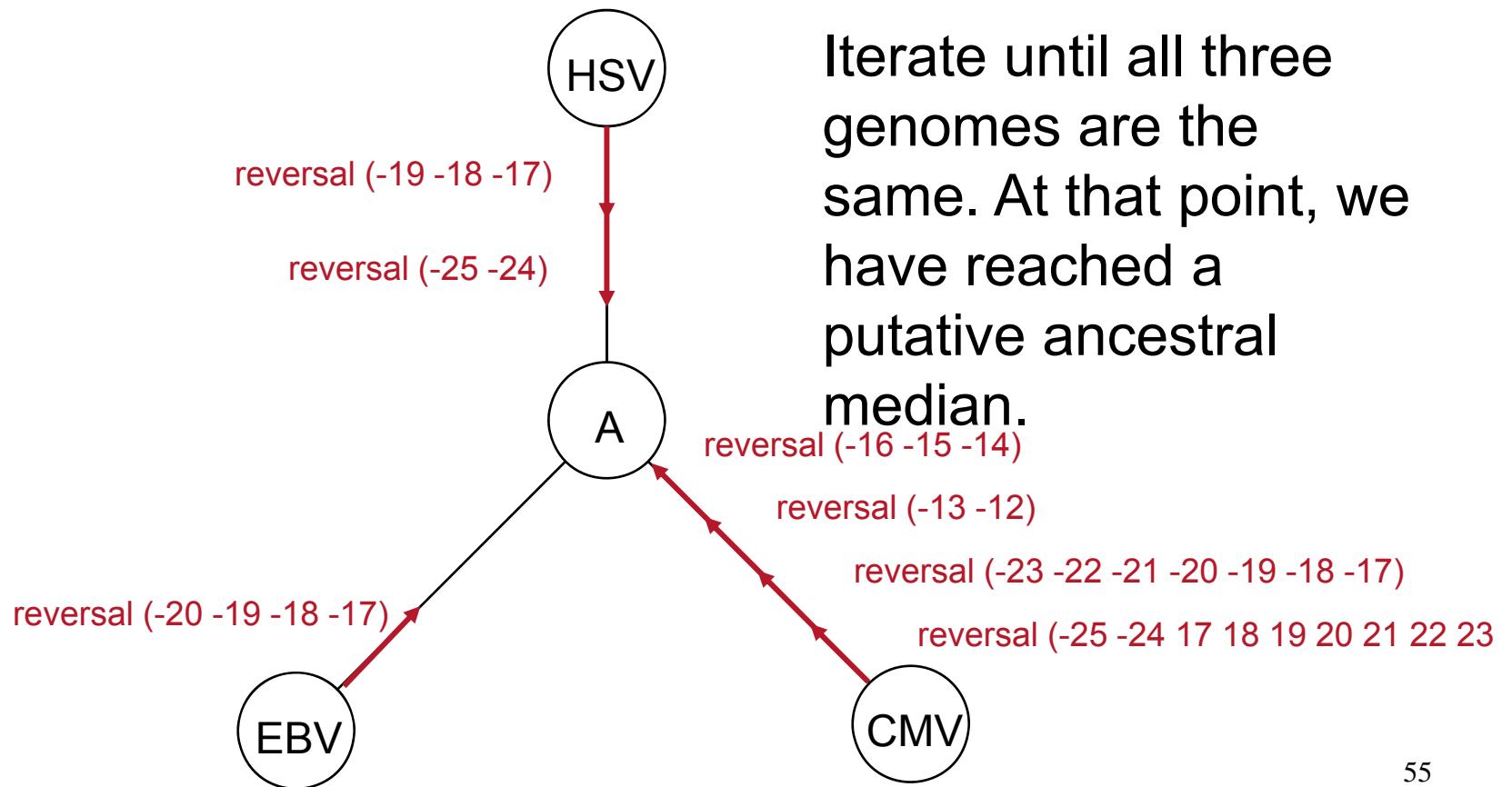
Method works best for **additive** or nearly additive trees.

An **additive tree** is a tree such that distance between each pair of leaves is equal to sum of the lengths of the branches on the tree.

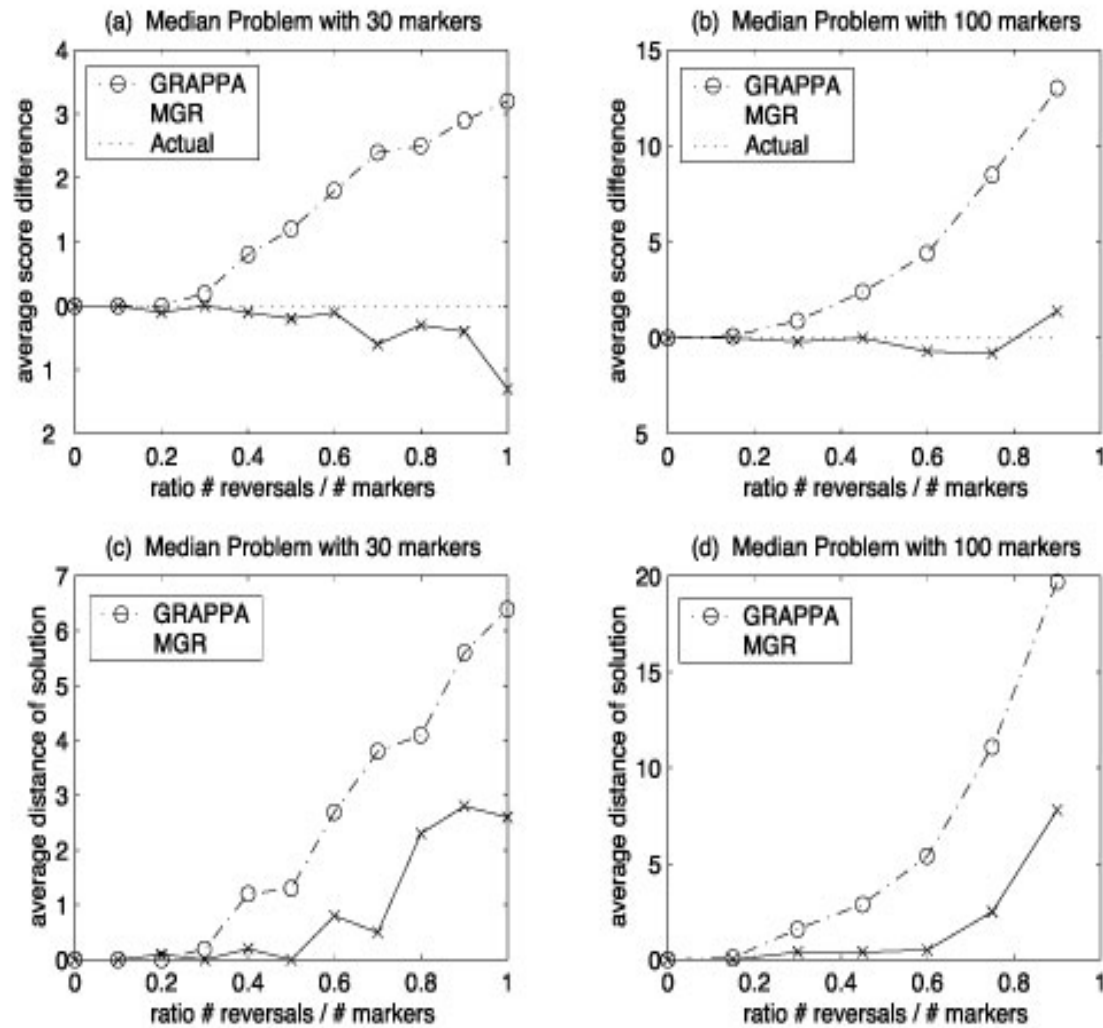
$$c = a + b$$

Median problem example

HSV: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 -19 -18 -17 20 21 22 23 -25 -24
EBV: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 -20 -19 -18 -17 21 22 23 24 25
CMV: 1 2 3 4 5 6 7 8 9 10 11 -13 -12 -16 -15 -14 -23 -22 -21 -20 -19 -18 -17 24 25
A: ???

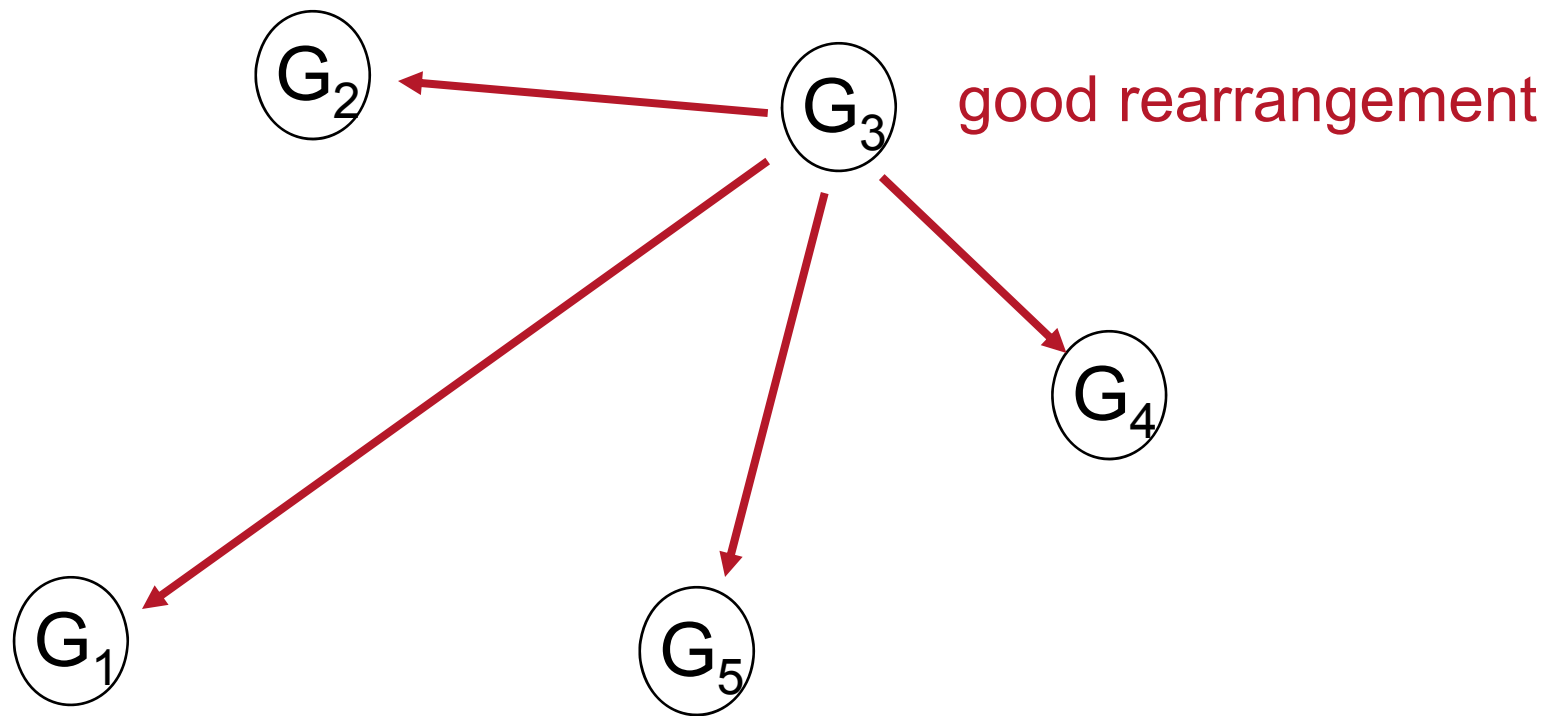


Tests (3 genomes)



MGR Algorithm

The idea is the same, a rearrangement which brings (G₃) closer to all other genomes is probably good.



Outline

- Sequencing large genomes
- Motivation
- Definitions and pairwise algorithms for genome rearrangements
- Genome rearrangement phylogenies
- Applications and latest developments