Introduction To Bioinformatics

Jens Kleinjung
Division of Mathematical Biology
National Institute for Medical Research
London

Jülich 05.2010

Lectures

1. Evolutionary Processes and Phylogenetic Trees
2. Sequence Searching with Blast
3. Sequence Alignment
4. Principles of Information Theory and Thermodynamics
5. Secondary Structure Assignment and Prediction
6. Tertiary Structure Prediction and Modelling
7. Tertiary Structure: Folding
8. Protein Interactions: Complexes and Networks
9. Genome Analysis and Comparison
10. A Brief Guide to Algorithms and Programming

1.4 The Mechanism of Evolution

Mendel (1866)

Inheritance follows mathematical rules.

Darwin & Wallace (1859)

Natural variation and selection lead to evolutional adaptation. (mutations not yet known)

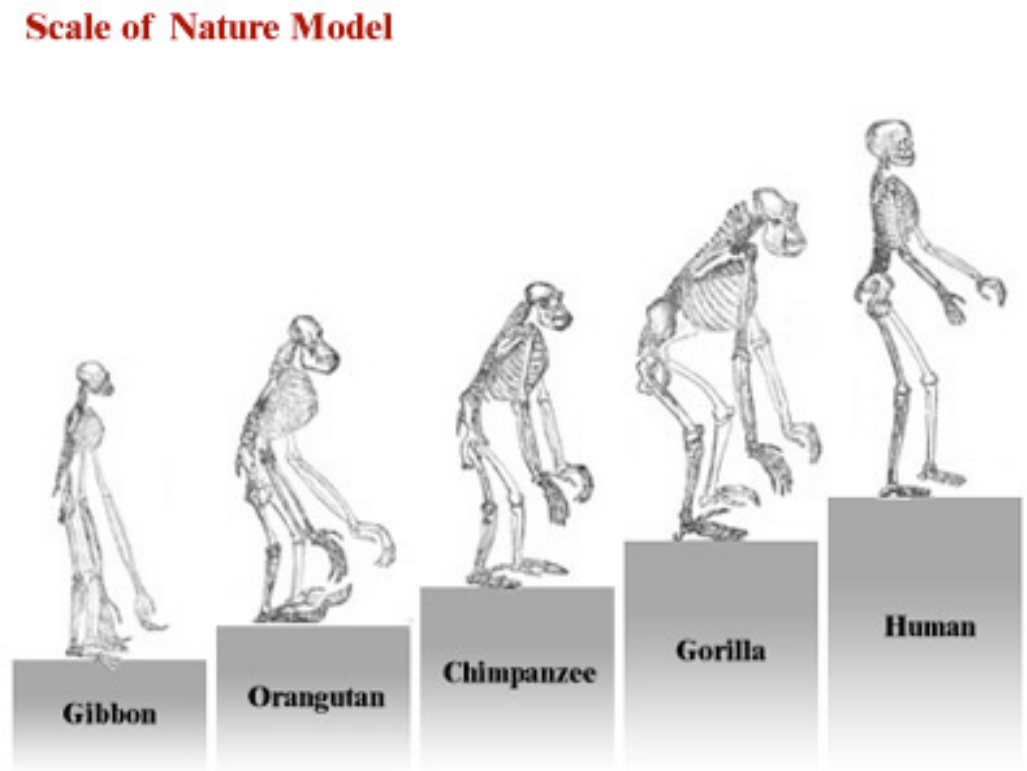
Watson & Crick (1953)

DNA double helix explains storage and replication of genetic information.

Human Genome Sequencing Consortium (2001)

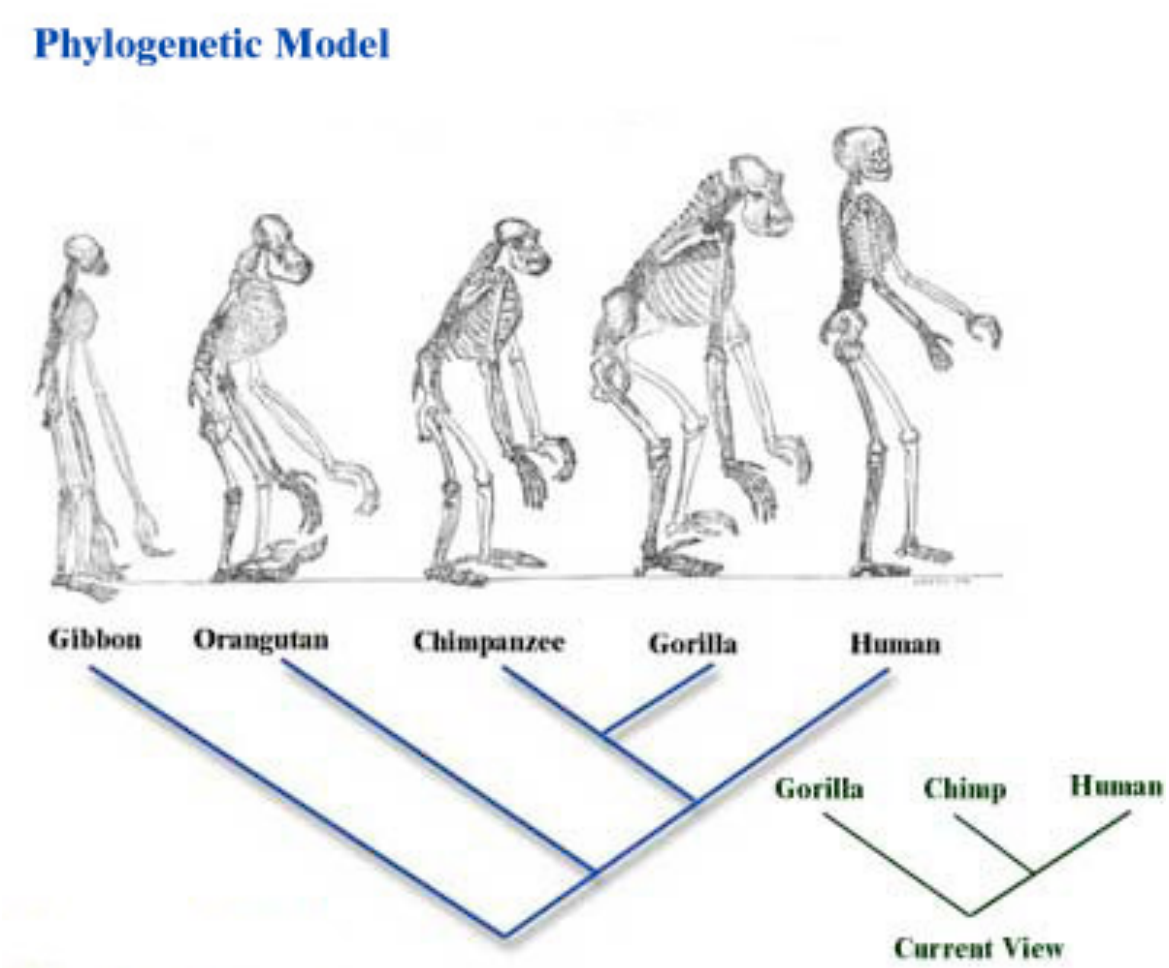
First draft of the human genome.

Wrong Model of Evolution



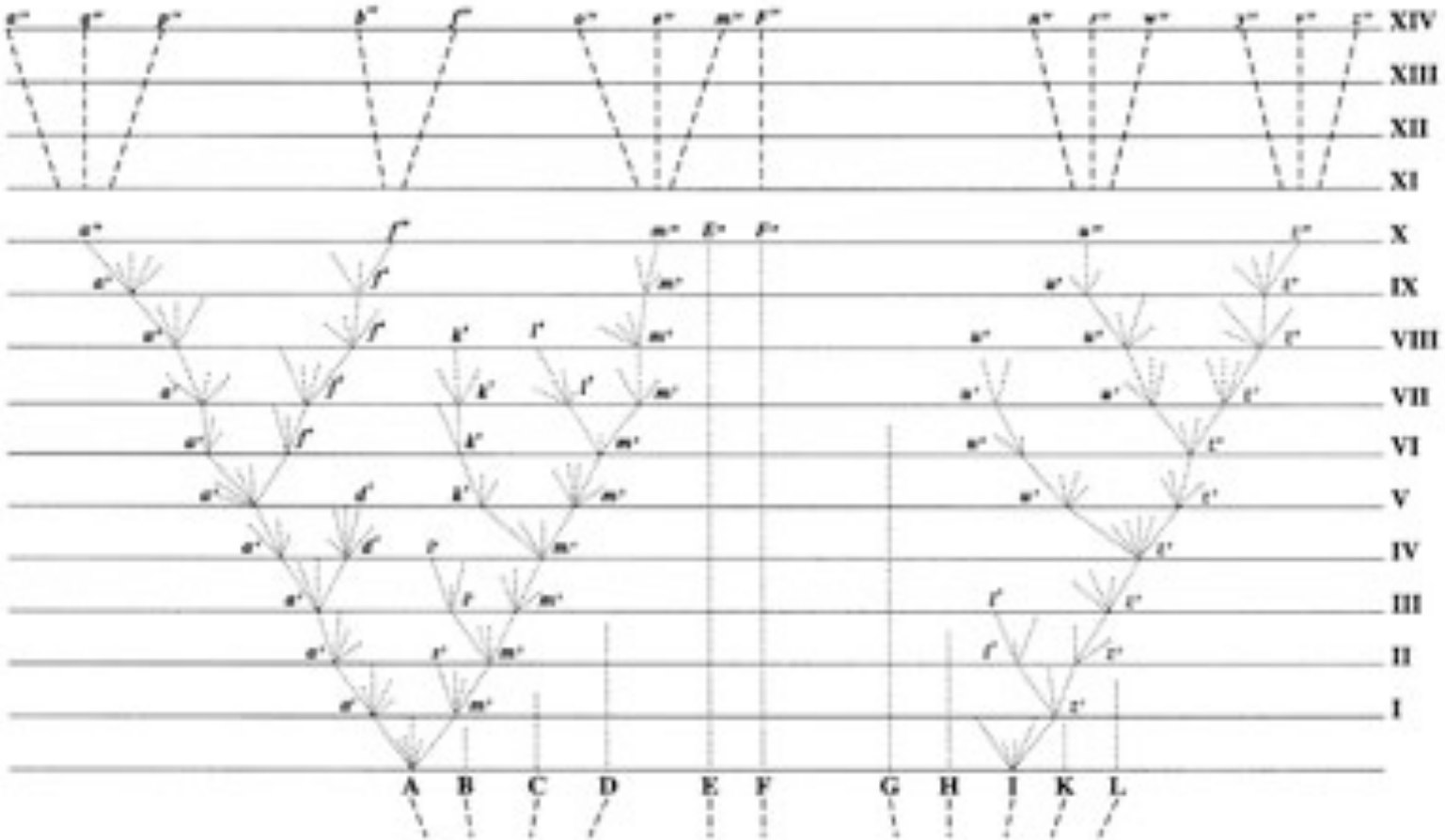
Before Charles Darwin, evolutionary history was thought to be a linear process.

Correct Model of Evolution



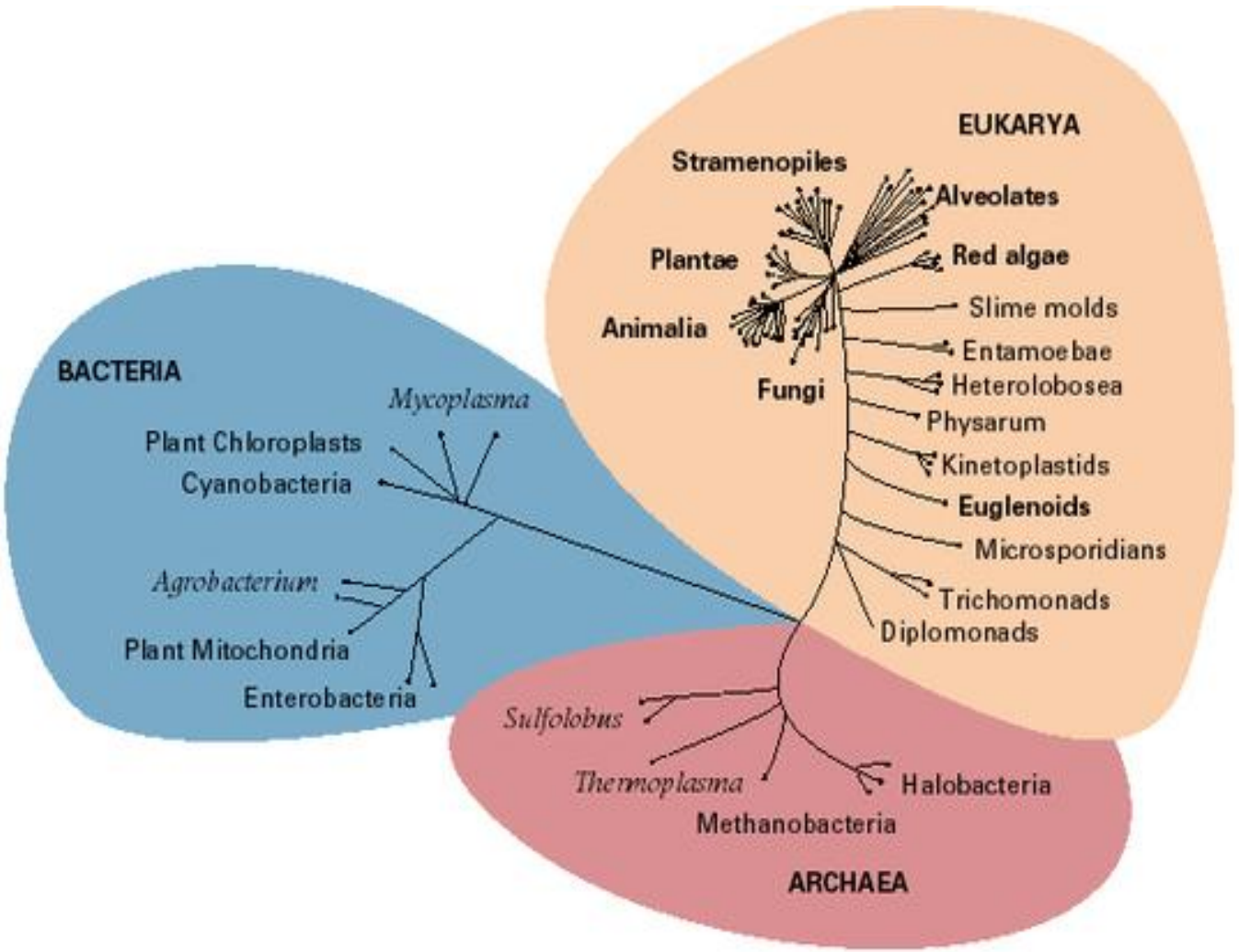
But in fact, it is a tree based on progressive divergence.

Darwin's Phylogenetic Tree



Phylogenetic tree by Charles Darwin in the 'Origin of Species'.

Tree of Life


EUKARYA
Stramenopiles
Alveolates
Plantae
Red algae
Animalia
Slime molds
Entamoebae
Fungi
Heterolobosea
Physarum
Kinetoplastids
Euglenoids
Microsporidians
Trichomonads
Diplomonads
BACTERIA
Mycoplasma
Plant Chloroplasts
Cyanobacteria
Agrobacterium
Plant Mitochondria
Enterobacteria
Sulfolobus
Thermoplasma
Halobacteria
Methanobacteria
ARCHAEA

Evolutionary Time Scales

Kingdom	When Evolved	Structure	Phothosynthesis
Prokaryotes:-			
Bacteria	3 to 4 billion years ago	Unicellular	Sometimes
Archaea	3 to 4 billion years ago	Unicellular	No
Eukaryotes:-			
Protista	1.5 billion years ago	Unicellular	Sometimes
Fungi	1 billion years ago	Unicellular or Multicellular	No
Animalia	700 million years ago	Multicellular	No
Plantae	500 million years ago	Multicellular	Yes

DNA and Culture

Texts are used in human culture for cultural inheritance since thousand of years. With the human genome sequencing the molecular inheritance became available in text form. This formalisation opens the possibility to decipher the 'book of life'.

DNA: .dat or .bin ?

The DNA stores the sequences of our genes.
In this respect the DNA behaves like data (.dat).

The DNA has regulatory areas that control the expression of genes. In this respect the DNA behaves like a program (.bin).

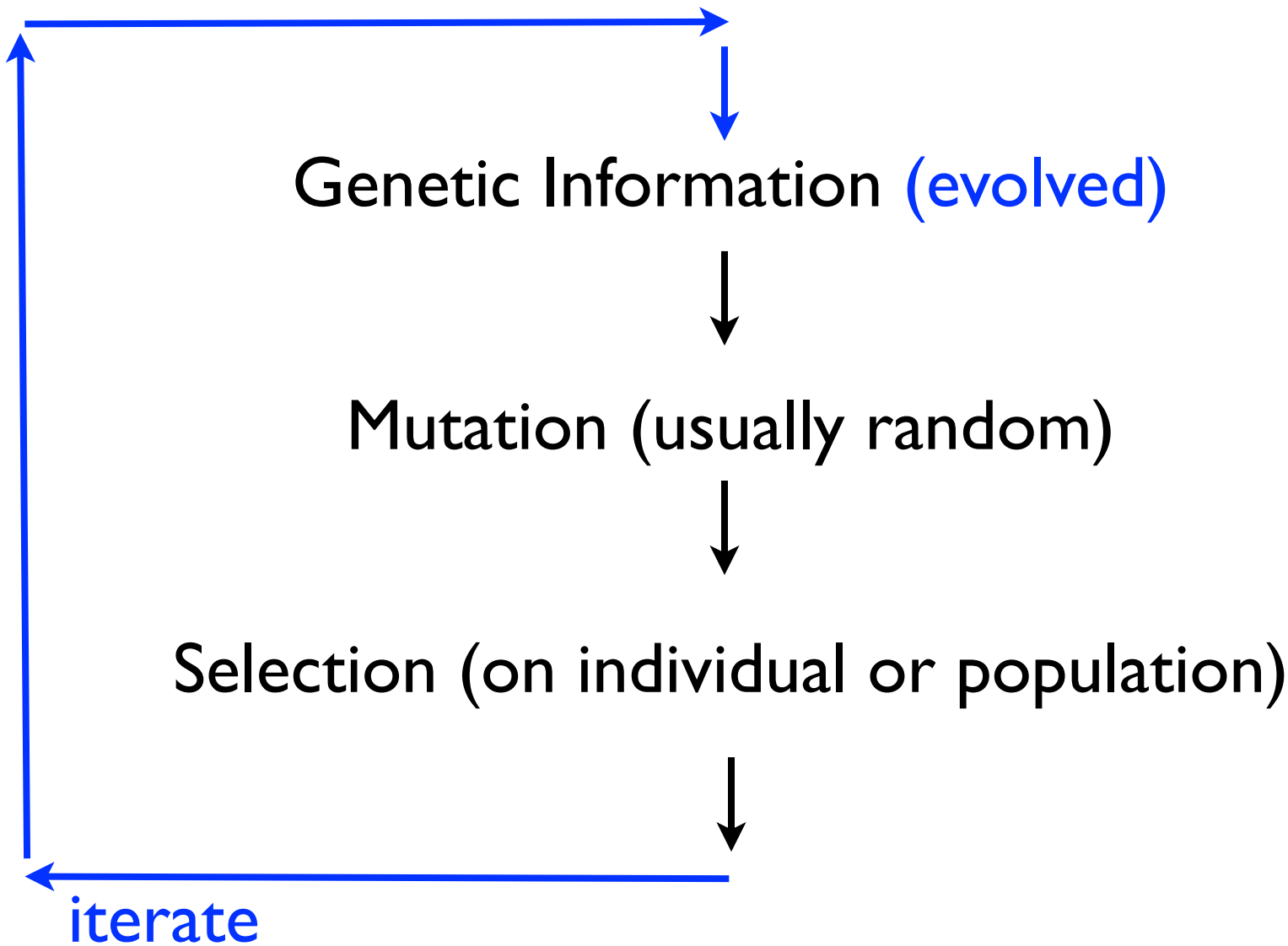
DNA has a temporal dimension: Consider that caterpillar and butterfly are the phenotype of the same genome.

Diseases

Each disease can be traced back to a molecular cause. To ultimately understand diseases, we need to know the underlying molecular mechanisms and ideally their effect at the level of the cell, organ and organism.

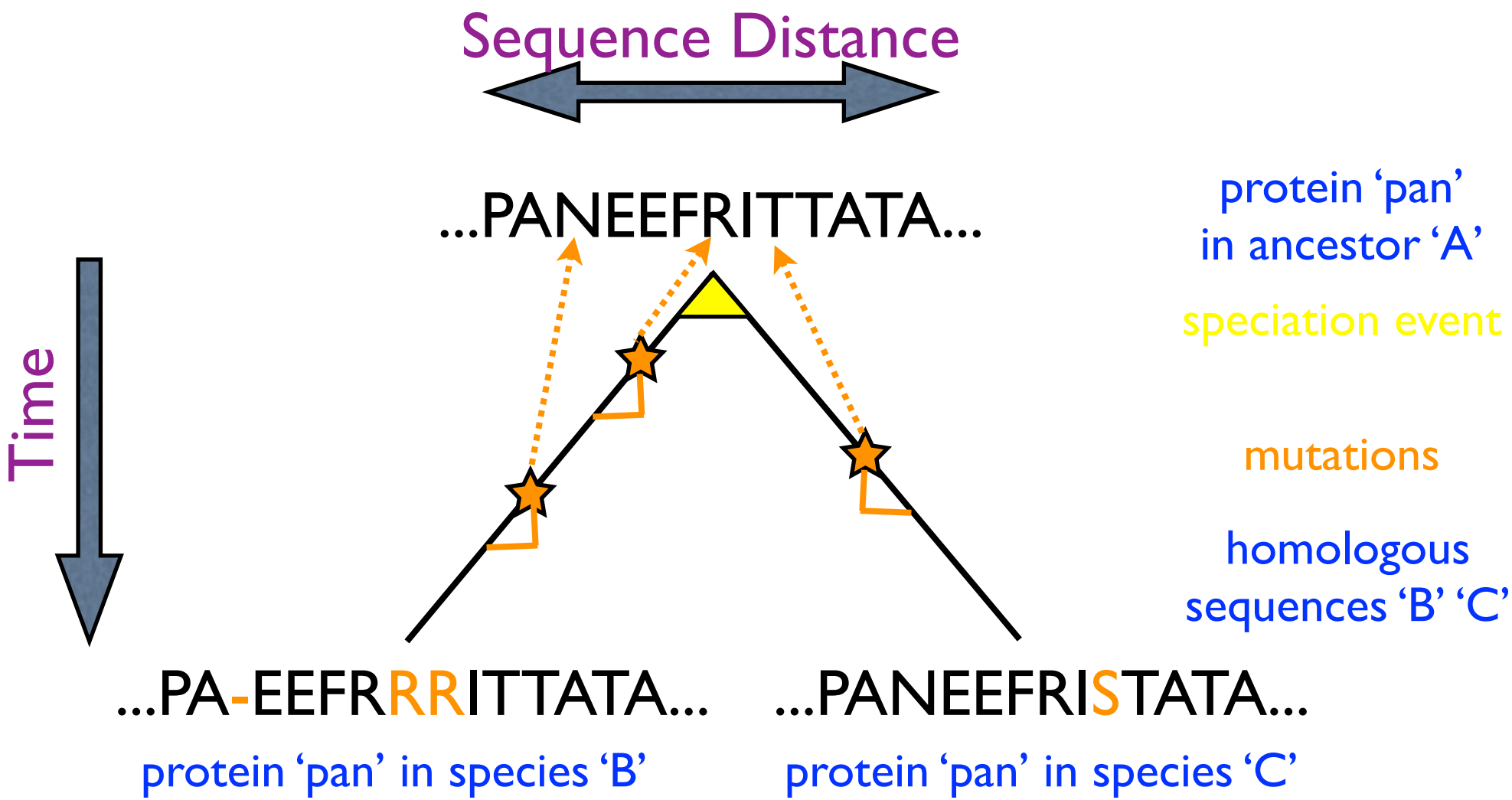
Bioinformatics is mostly concerned with the molecular level, Medicine mostly with organs and the organism. Systems Biology is a new field to bridge the different levels.

The Mechanism of Evolution



Divergence and Homology

Divergent evolution is the rule.



Comments to Previous Slides

Mutations of the DNA occur constantly. Mutations in somatic cells remain localised (although in the case of cancer they can spread), while mutations in the germline will be contained in the child generation and may be fixed in the population.

Within each species the exchange of genetic material keeps the average sequence distance short.
Example: Despite geographic separation there is only one human species.

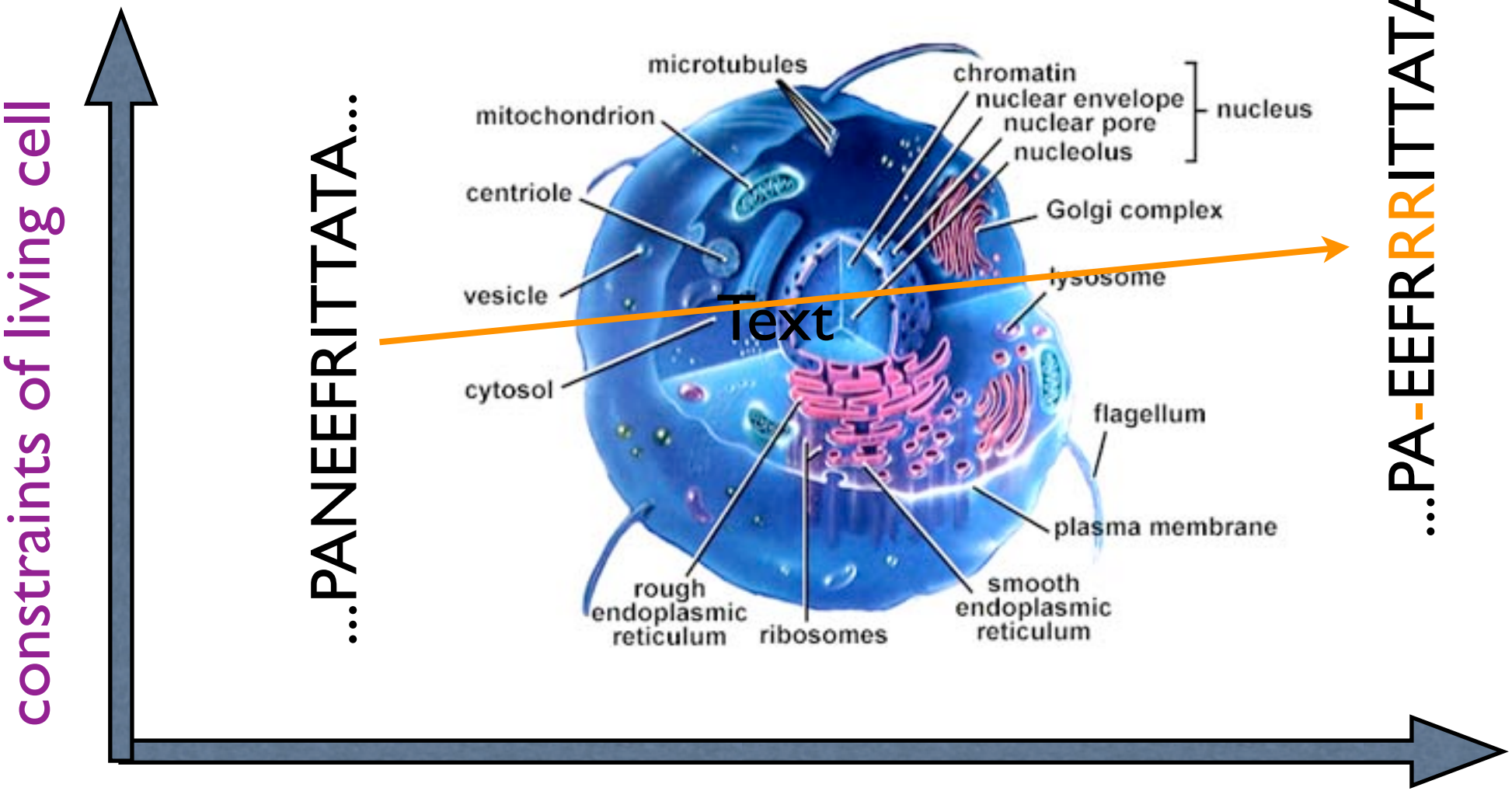
Comments to Previous Slides

If two groups of a species stop exchanging genes, their genomes will diverge up to a point where their genetic material becomes incompatible for reproduction.

Speciation event: The point in time when two groups of a species have diverged such that they cannot produce viable offspring together. Because species cannot mix genes between each other, mutations increase gradually in time the sequence distance.

Species definition: Two species cannot produce viable offspring together.

Variation versus Conservation in Biology



Sequence Alignment

Pairwise alignment of homologous sequences.
Insert gaps (-) to bridge INDEL events.

```
...PA-EEFRRRITTATA...
   ||  ||||   ||||||
...PANEEFR--ISTATA...
```

Sequence identity: 11/15 = 73%

One can assume homology if sequence identity > 30%.

Alignment and Homology

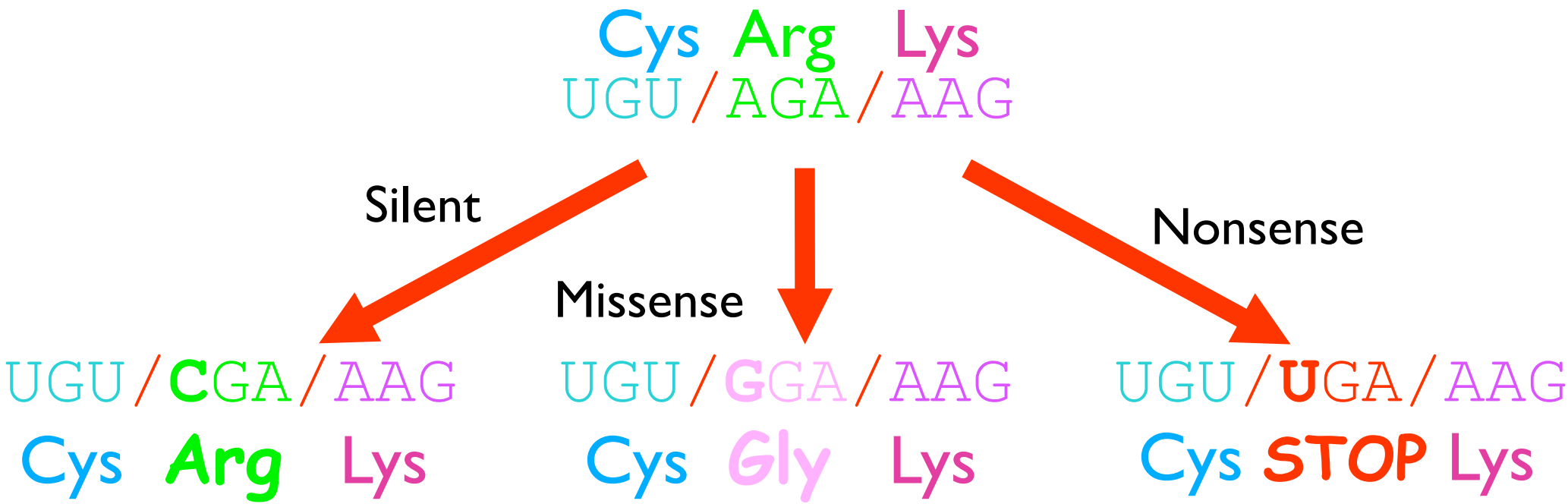
A sequence alignment matches amino acid or nucleotide residues that are evolutionary related.

Therefore, sequence alignment makes only sense for homologous sequences!

The sequence space is so immense (20^{200} for a protein with 200 residues) that the probability of a significant similarity between unrelated (non-hololologous) sequences is close to zero.

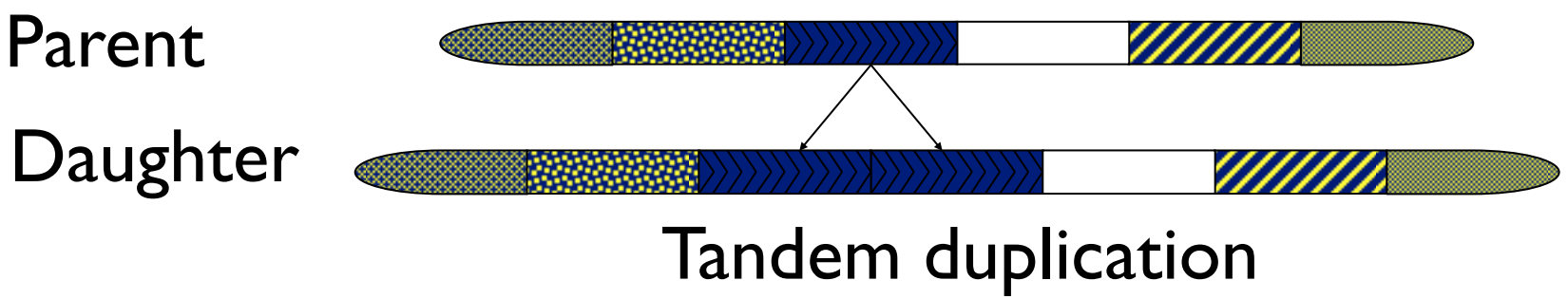
The only exceptions are relatively short segments of convergent evolution.

Substitutions in Coding Regions

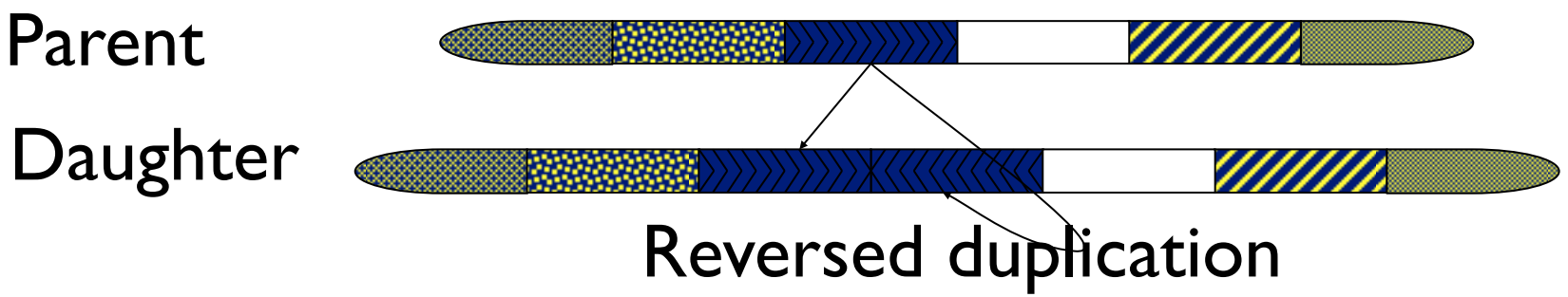


- First position: 4% of all changes silent
- Second position: no changes silent
- Third position: 70% of all changes silent (wobble position)

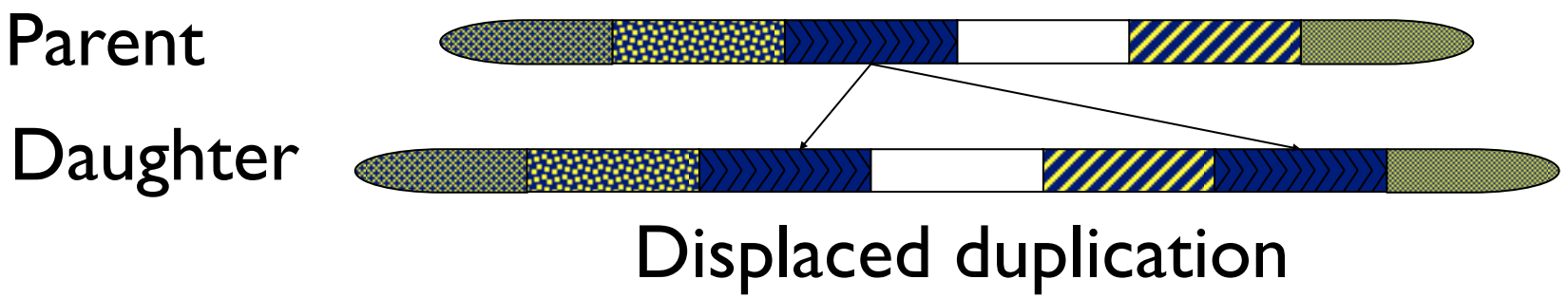
Chromosomal Mutations: Duplications



Tandem duplication

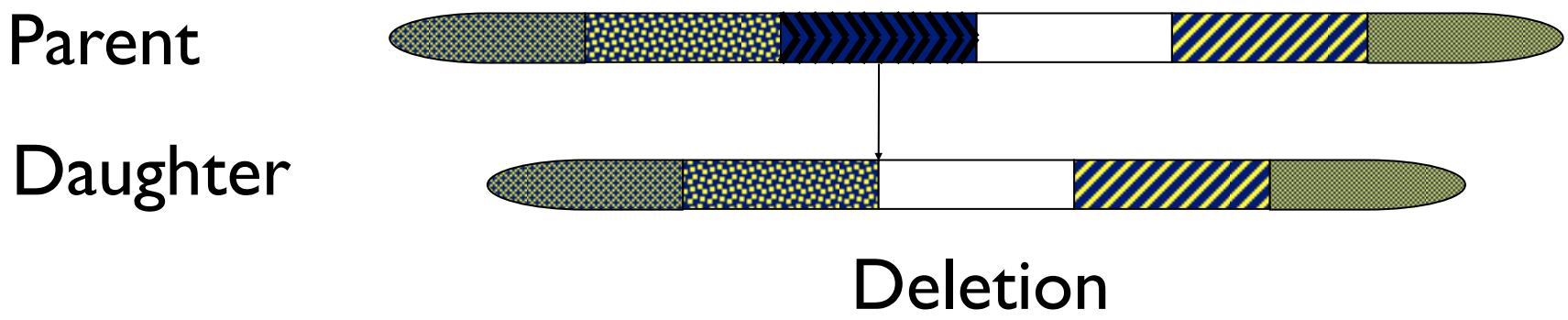


Reversed duplication

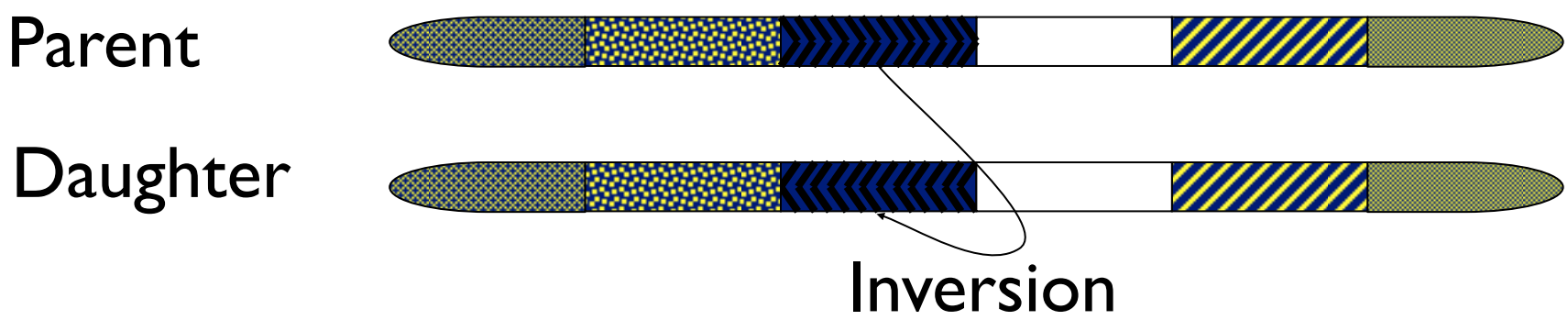


Displaced duplication

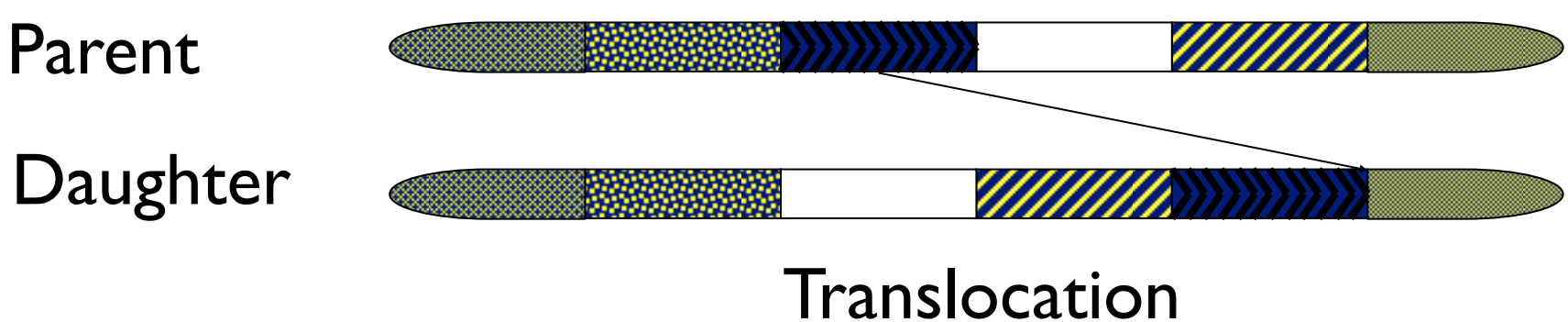
Chromosomal Mutations: Rearrangements



Deletion



Inversion



Translocation

Speciation: Horse and Donkey

Example: Horse and Donkey

Donkey: 62 chromosomes

Horse: 64 chromosomes

♂ donkey + ♀ horse → mule (63 chrom., sterile)

♀ donkey + ♂ horse → hinny (63 chrom., sterile)

Convergence

Convergent evolution is the exception.

Convergent evolution happens when the same selection pressure generates the same result in two unrelated organisms.

Example: The similarity of swift and swallow (rondone and rondine).

Convergence in Molecules

Chen et al, 97, PNAS, 94, 3811-16

Anti-freeze glycoprotein with $(ThrAlaAla)_n$
similar to Trypsinogen

Anti-freeze glycoprotein with $(ThrAlaAla)_n$

NOT

similar to Trypsinogen

SIMILAR sequences BUT DIFFERENT origin

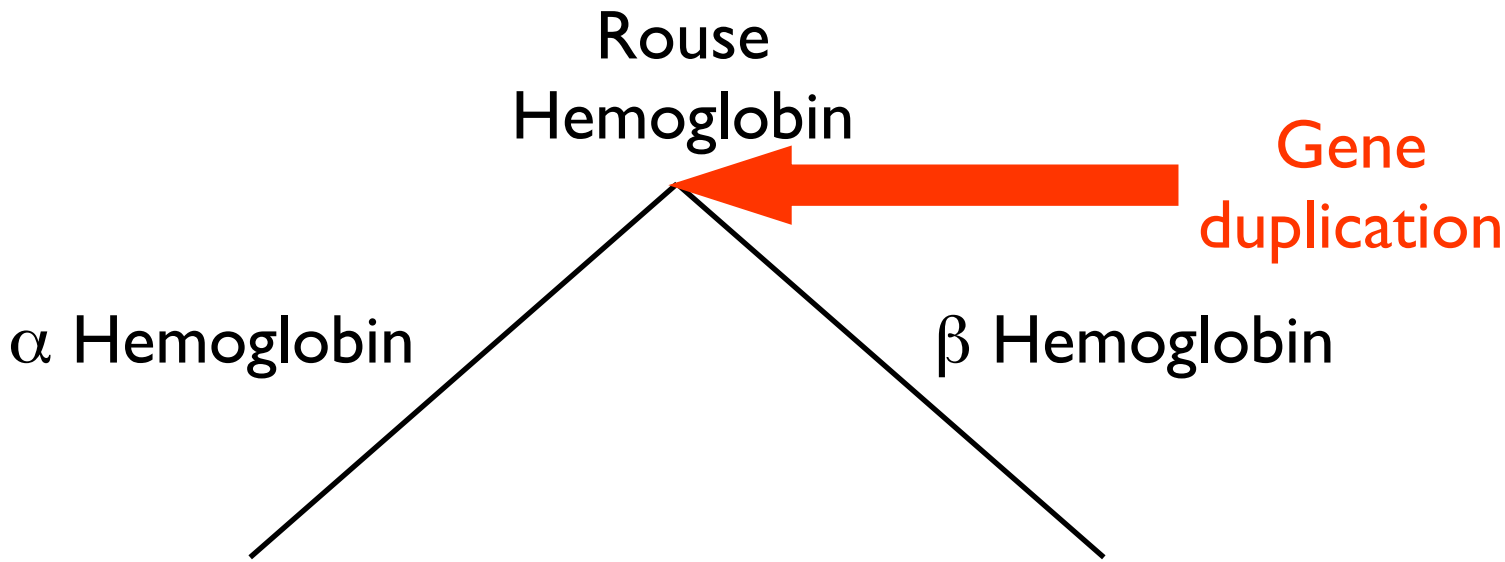
Slide by Cedric Notredame

Homologies

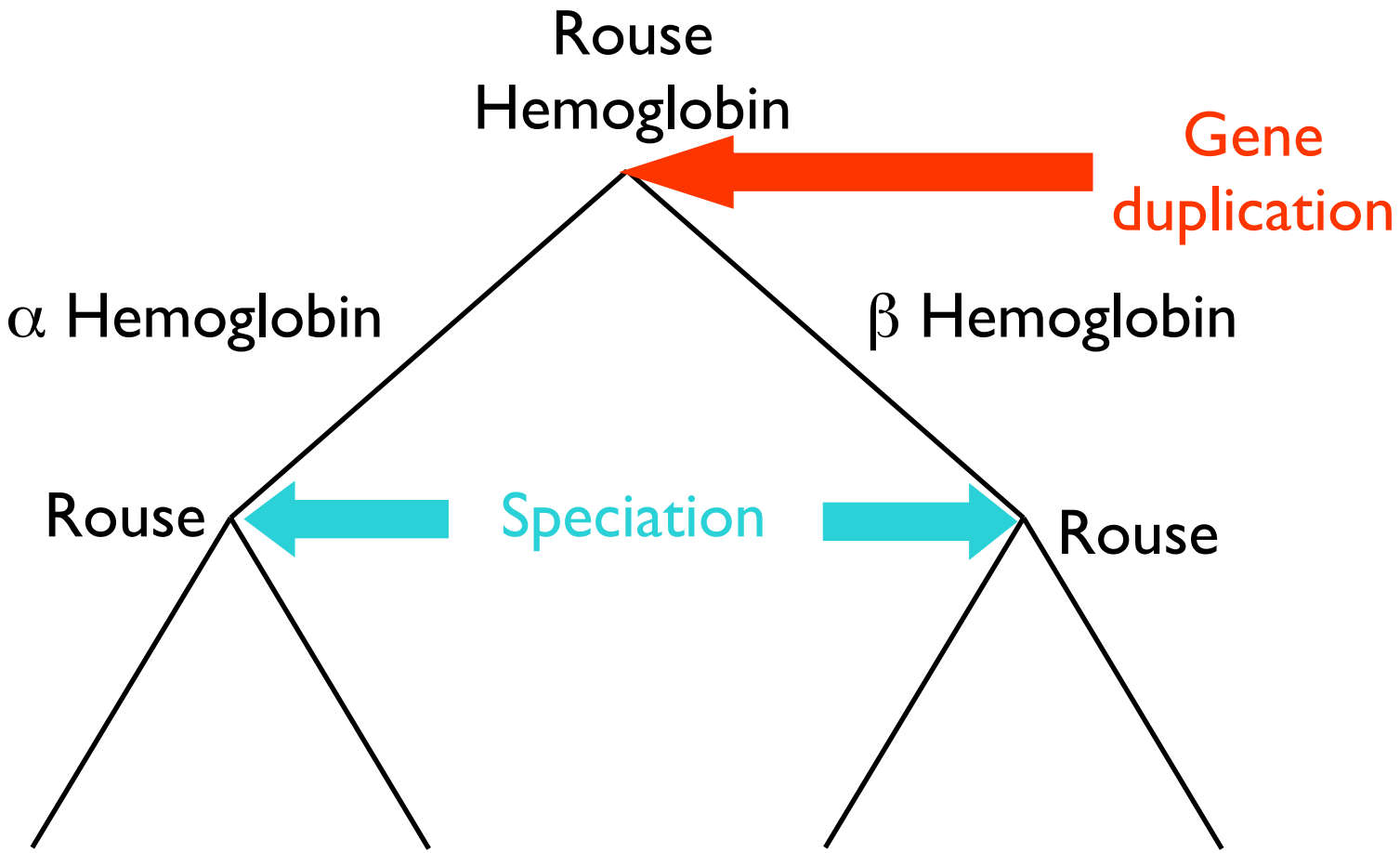
Homologies

Rouse
Hemoglobin

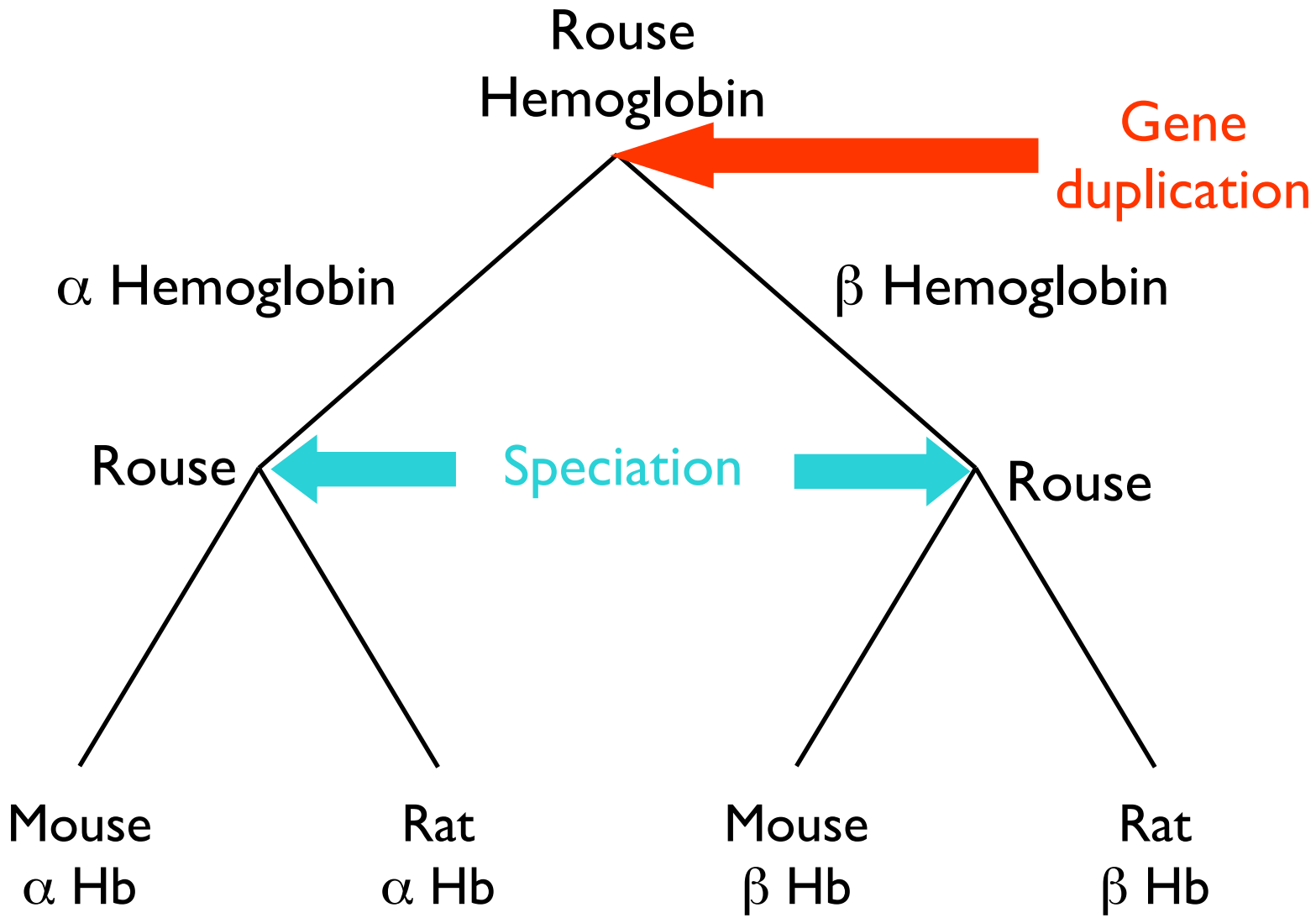
Homologies


Rouse
Hemoglobin
Gene
duplication
α Hemoglobin
β Hemoglobin

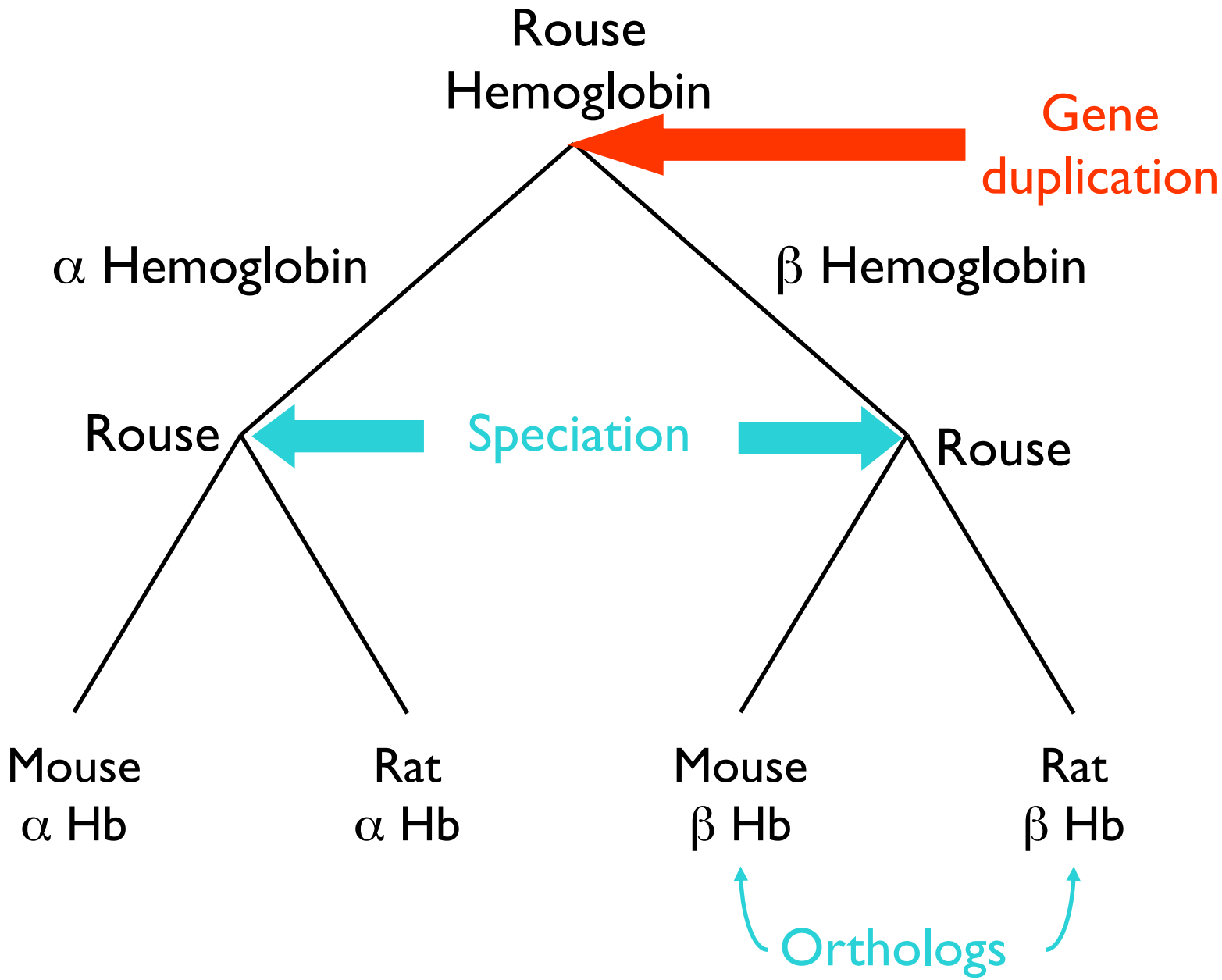
Homologies


Rouse
Hemoglobin
Gene
duplication
α Hemoglobin
β Hemoglobin
Rouse
Speciation
Rouse

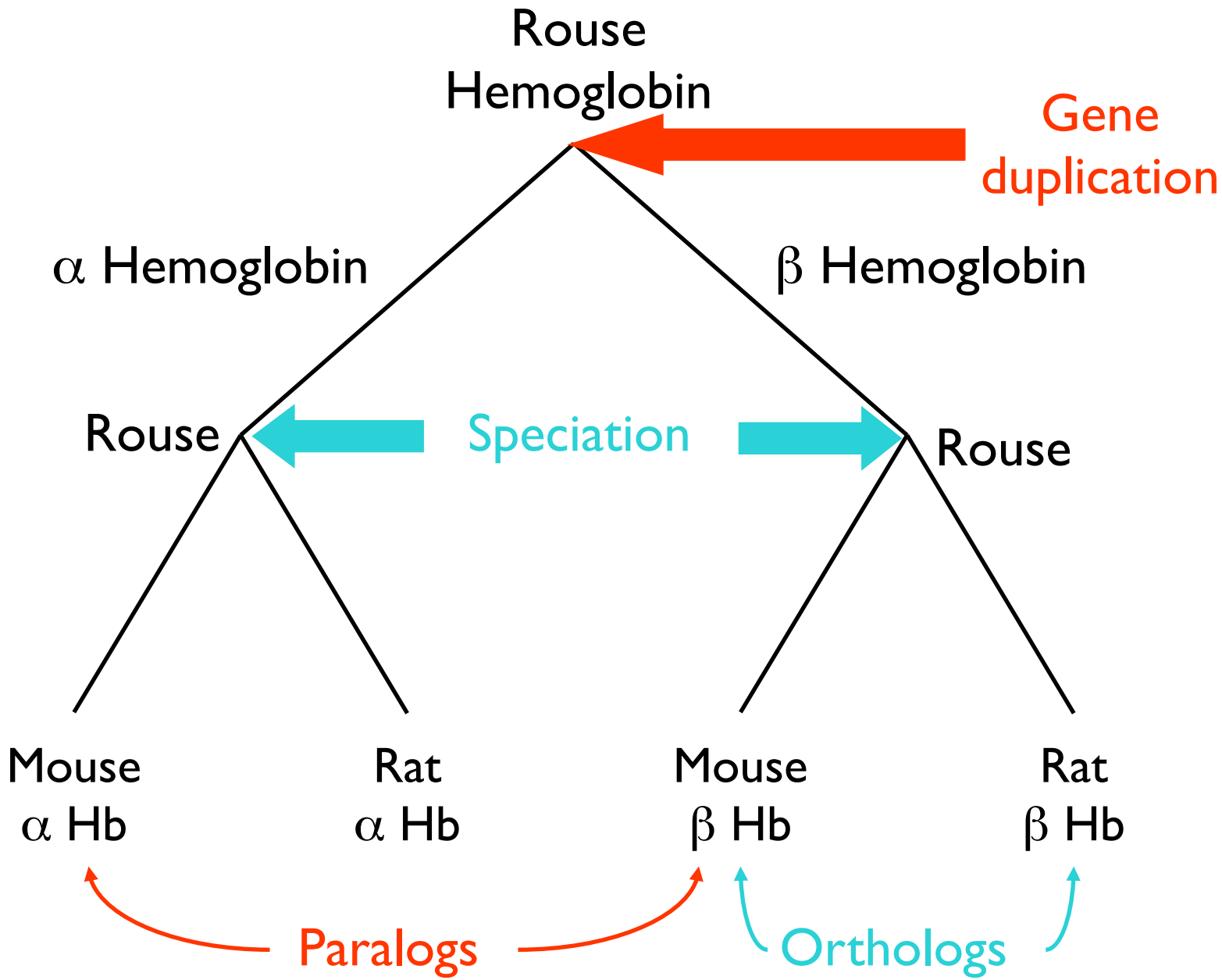
Homologies


Rouse
Hemoglobin
Gene
duplication
α Hemoglobin
β Hemoglobin
Rouse
Speciation
Rouse
Mouse
α Hb
Rat
α Hb
Mouse
β Hb
Rat
β Hb

Homologies


Rouse
Hemoglobin
Gene
duplication
α Hemoglobin
β Hemoglobin
Rouse
Speciation
Rouse
Mouse
α Hb
Rat
α Hb
Mouse
β Hb
Rat
β Hb
Orthologs

Homologies

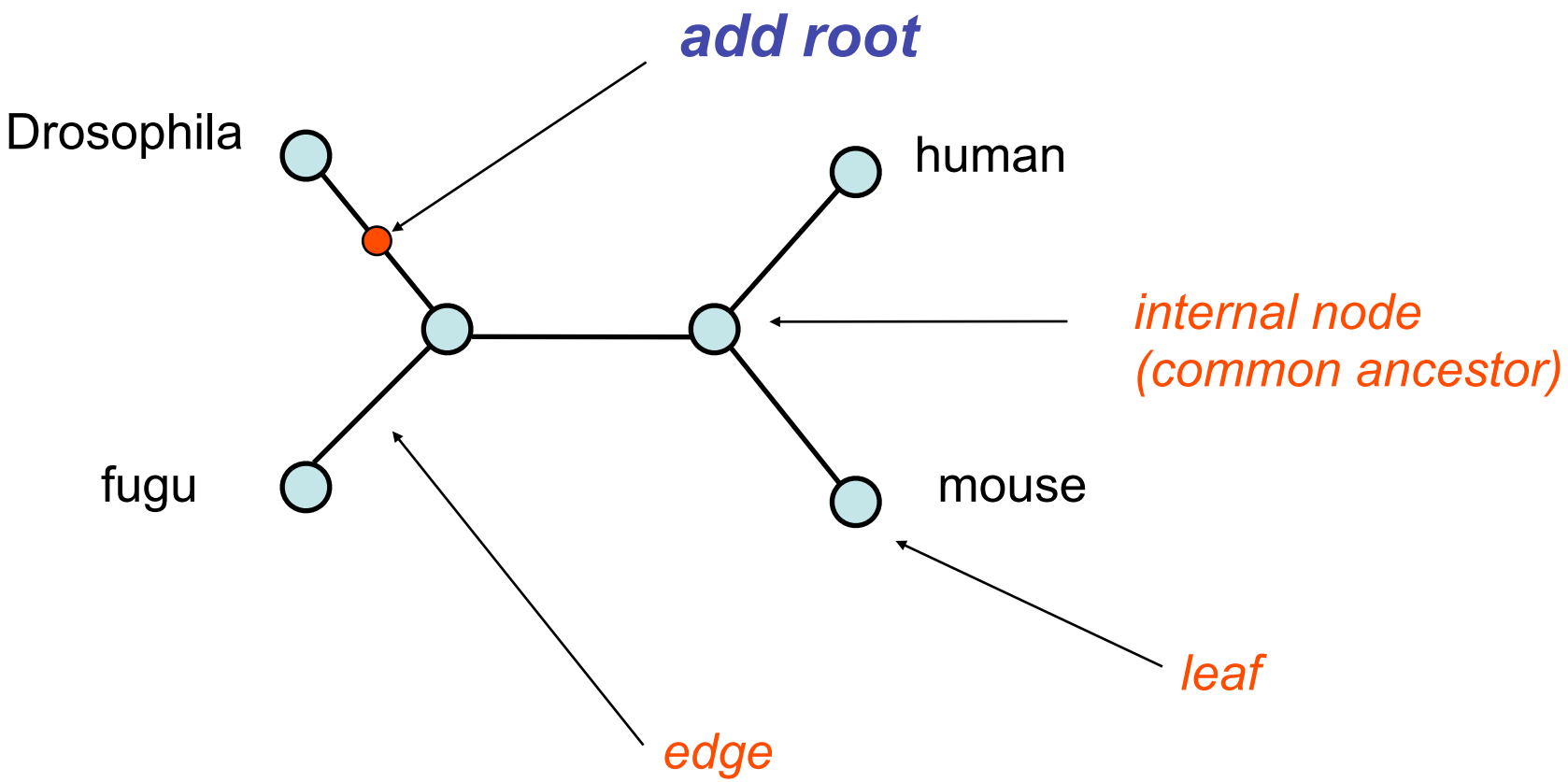

Rouse
Hemoglobin
Gene
duplication
α Hemoglobin
β Hemoglobin
Rouse
Speciation
Rouse
Mouse
α Hb
Rat
α Hb
Mouse
β Hb
Rat
β Hb
Paralogs
Orthologs

Definitions of Homologies

- Homologues are similar sequences in two different organisms that have been derived from a common ancestor sequence. Homologues can be described as either orthologues or paralogues.
- Orthologues are similar sequences in two different organisms that have arisen due to a speciation event. Orthologs typically retain identical or similar functionality throughout evolution.
- Paralogues are similar sequences within a single organism that have arisen due to a gene duplication event.

1.5 Phylogenetic trees

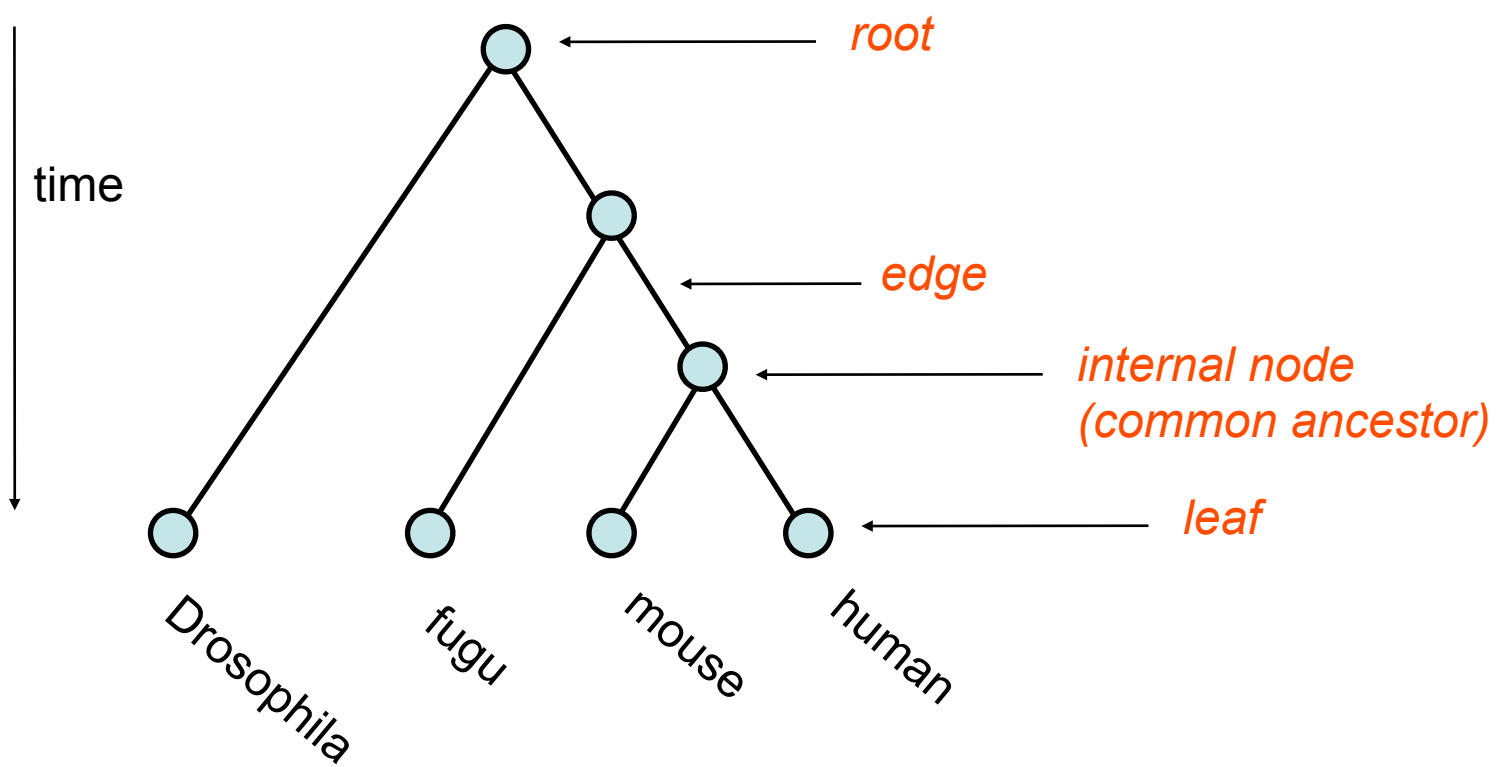
unrooted tree



Trees are binary.
Leaves are observed taxonomical units.
Edge (branch) lengths are proportional to evolutionary distance.

Phylogenetic trees

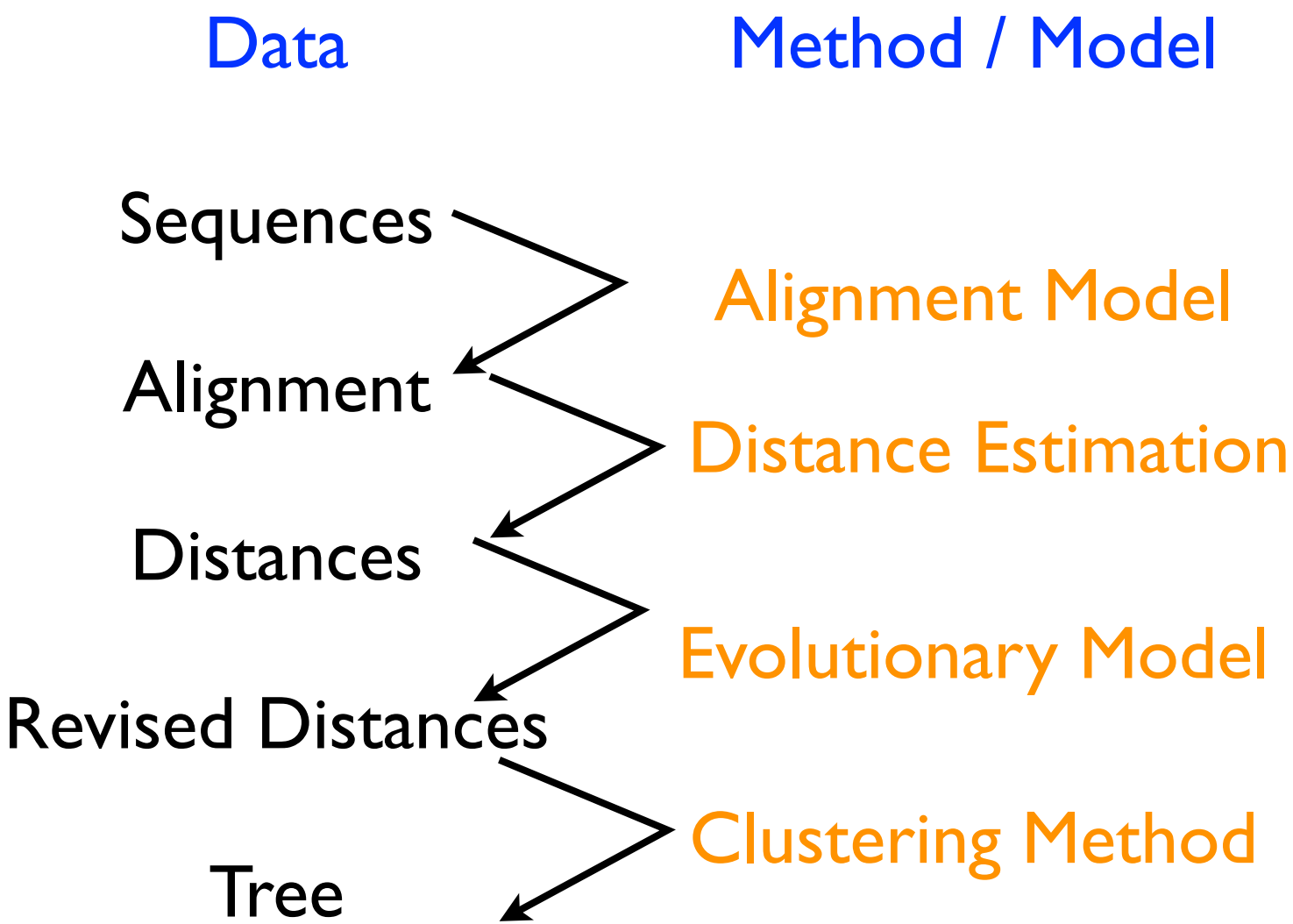
rooted tree



Types of Phylogenetic Trees

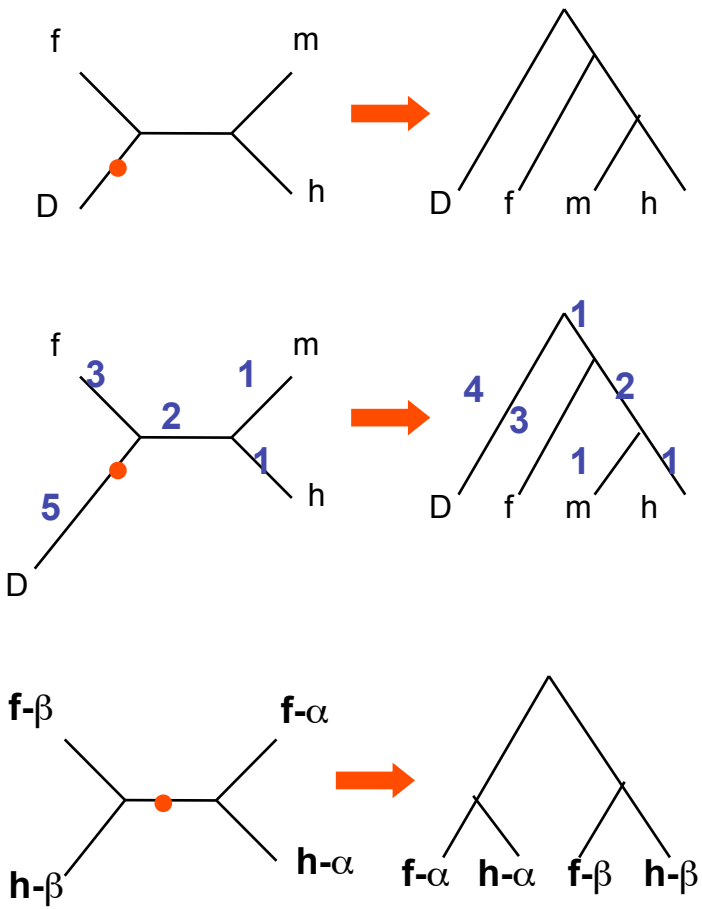
- Cladogram : Branch lengths are arbitrary
- Additive Tree : Branch lengths are proportional to the number of mutations per site.
- Ultrametric Tree : Like additive tree, in addition the mutation rate of all branches is the same.
  All leaves have the same distance from the root.

How to Build a Tree



How to Root a Tree

- Outgroup : place root between distant sequence and rest group.
- Midpoint : place root at midpoint of longest path (sum of branches between any two observed taxonomical units (OTUs)).
- Gene duplication : place root between paralogous gene copies.



Phylogenetic Distance Computation

Parsimony : fewest number of evolutionary events (mutations) – relatively often fails to reconstruct correct phylogeny.

Distance-based : pairwise sequence distances.

Maximum Likelihood :

L(ikelihood) = P(robability)[Tree | Data]

Bayesian Statistics :

P[Tree | Data] = P[Data | Tree] * P[Tree] / P[Data]

Algorithm: Markov Chain Monte Carlo (MCMC).

Tree Construction Methods

Single Tree Methods

- UPGMA
- Neighbour Joining

+ fast
+ robust
- no model comparison
- no fit quality assessment

Multiple Tree Methods

- Maximum Likelihood
- Bayesian Statistics

+ model comparison
+ fit quality assessment
- slow
- many parameters

Tree Confidence Intervals

Bayesian method : Compute probability of tree given the data and compare to probability of other well-fitting trees.

Distance method – bootstrap : Select multiple alignment columns with replacement and recalculate tree.

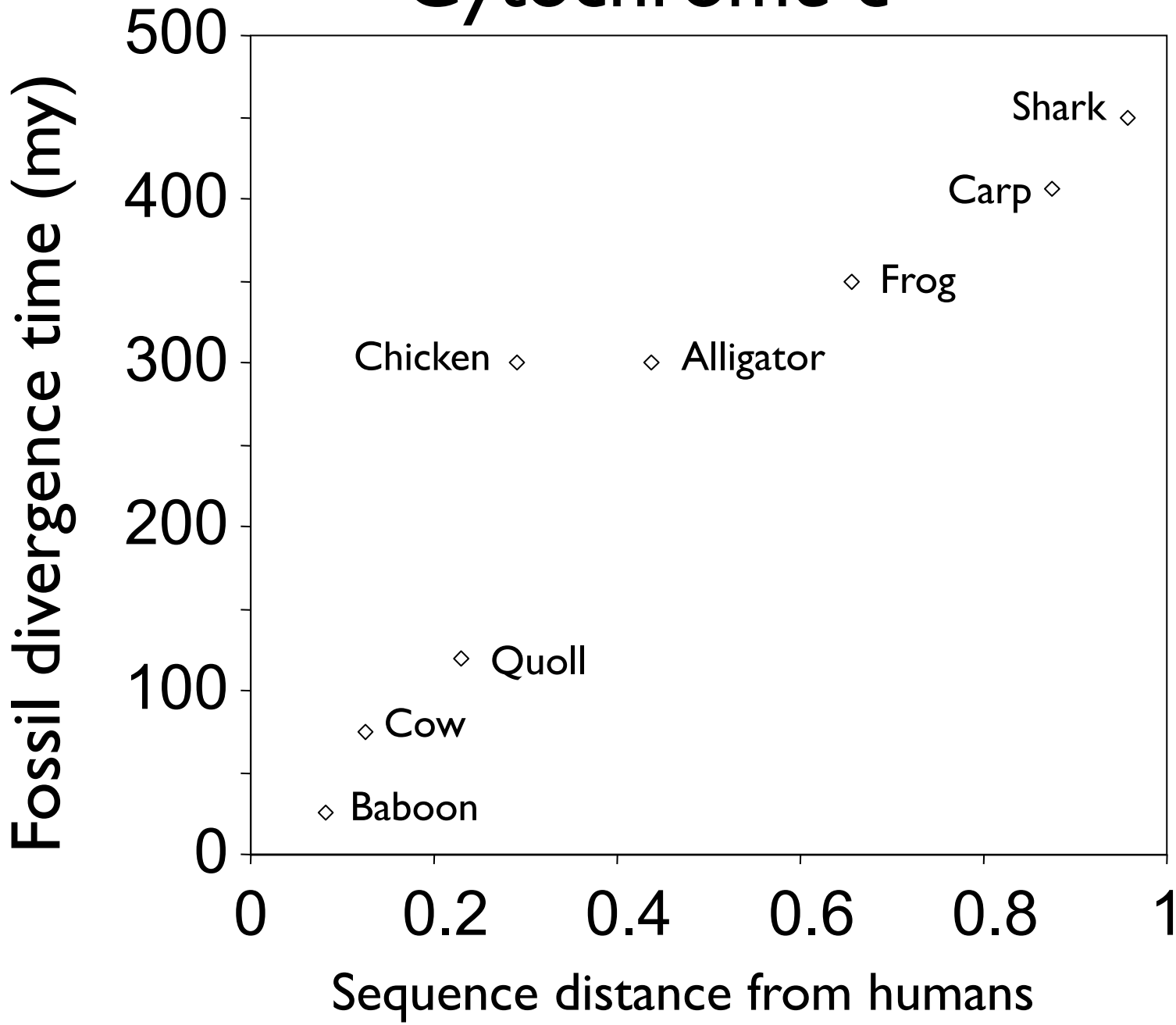
Compare branches with original (target) tree.

Repeat 100-1000 times, so calculate 100-1000 different trees, and derive confidence intervals for internal nodes.

Combinatorics of Trees

Nr (species)	Nr (unrooted trees)	Nr (rooted trees)
2	1	1
3	1	3
4	3	15
5	15	105
6	105	945
7	945	10395
8	10395	135135
9	135135	2027025
10	2027025	34459425

Evidence for the Molecular Clock: Cytochrome c



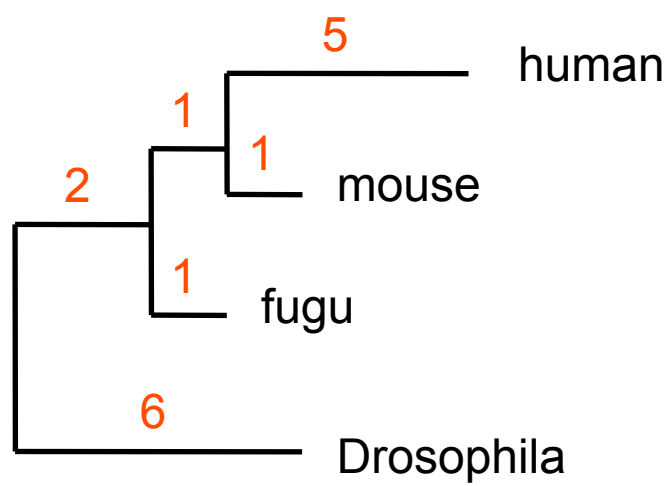
Distances on a Tree

One assumes that the species distance is proportional to the sequence distance.

The distance (in evolutionary time) is represented by the horizontal branch length.

human	x			
mouse	6	x		
fugu	7	3	x	
Drosophila	14	10	9	x

distance matrix



phylogenetic tree

Learning Outcomes

- Principles of inheritance and evolution
- Mutation and Selection
- Homologues and Paralogues
- Phylogenetic tree construction