CHARLES UNIVERSITY

Bioinformatics and Microbiome Analysis MB140P94

Introduction & History of sequencing and phylogenetics

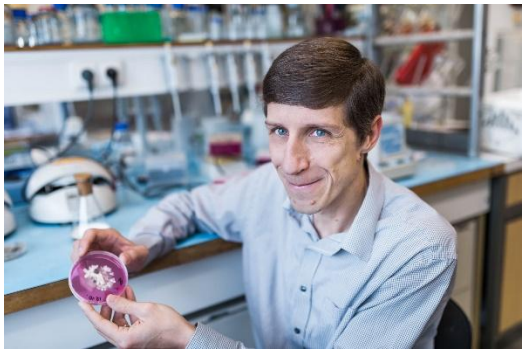
Tomáš Větrovský
Laboratory of Environmental Microbiology
Institute of Microbiology of the CAS

Introduction

Who are we?

Iñaki Odriozola

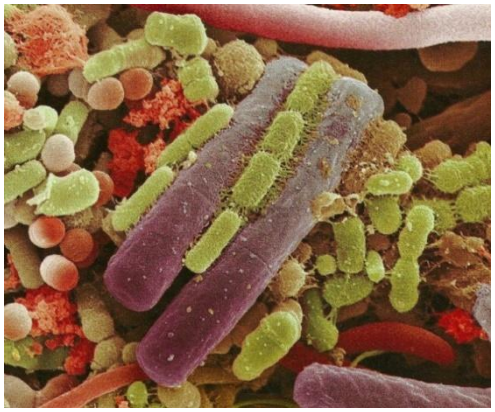
Petr Baldrian

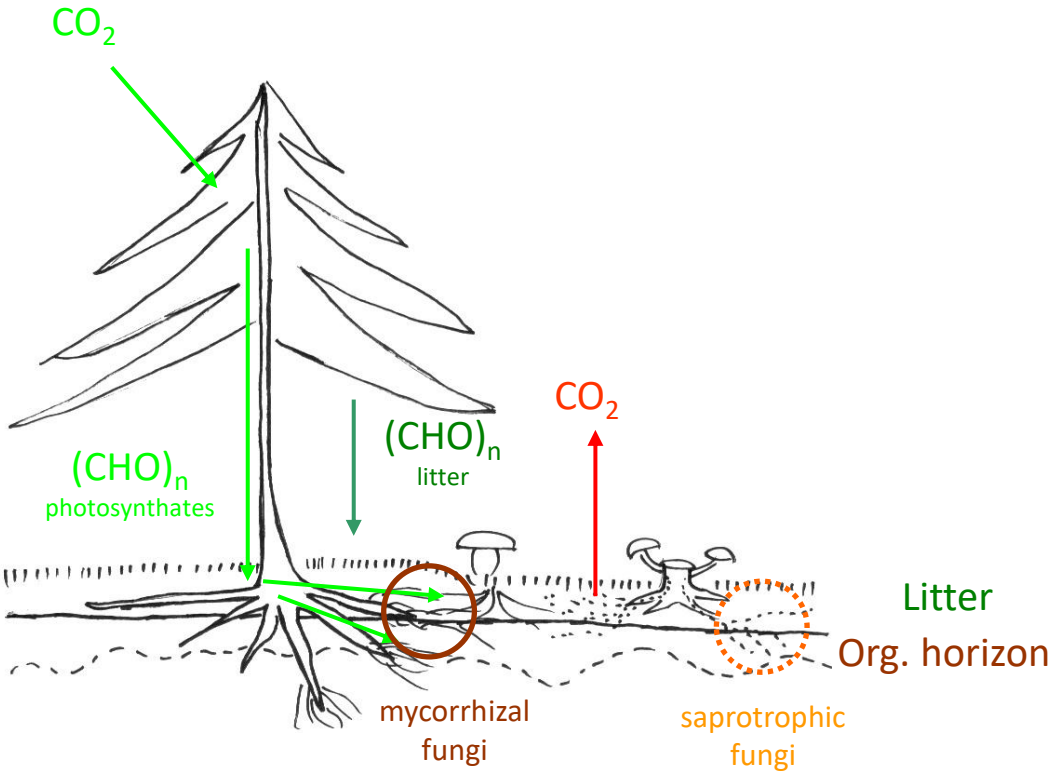


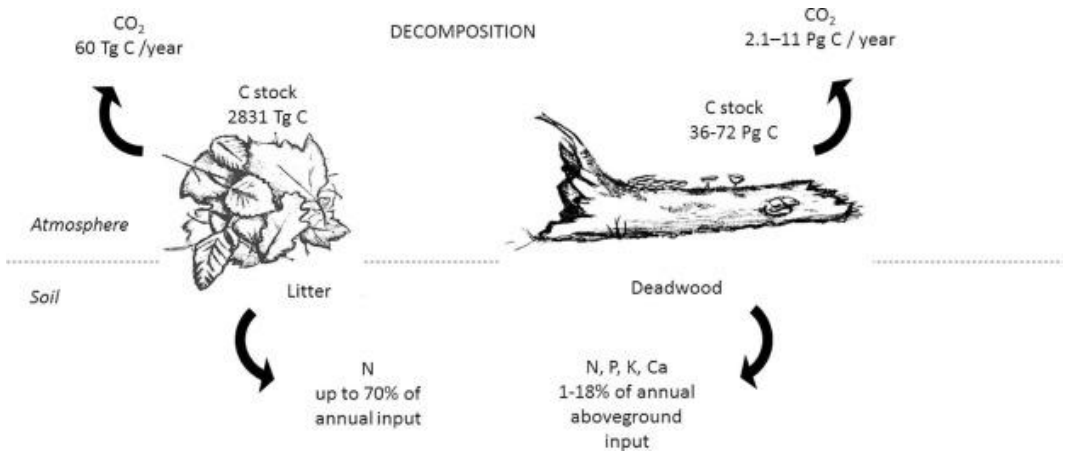
Priscila Thiago Dobbler

Algora Camelia

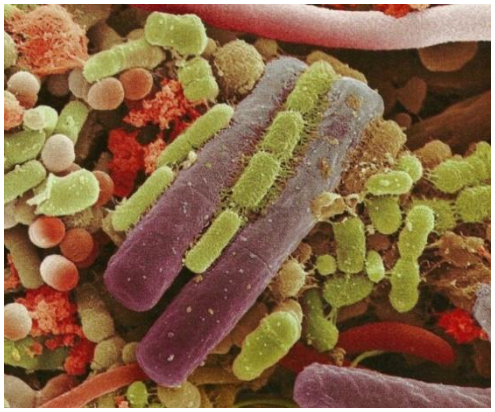
What we study...

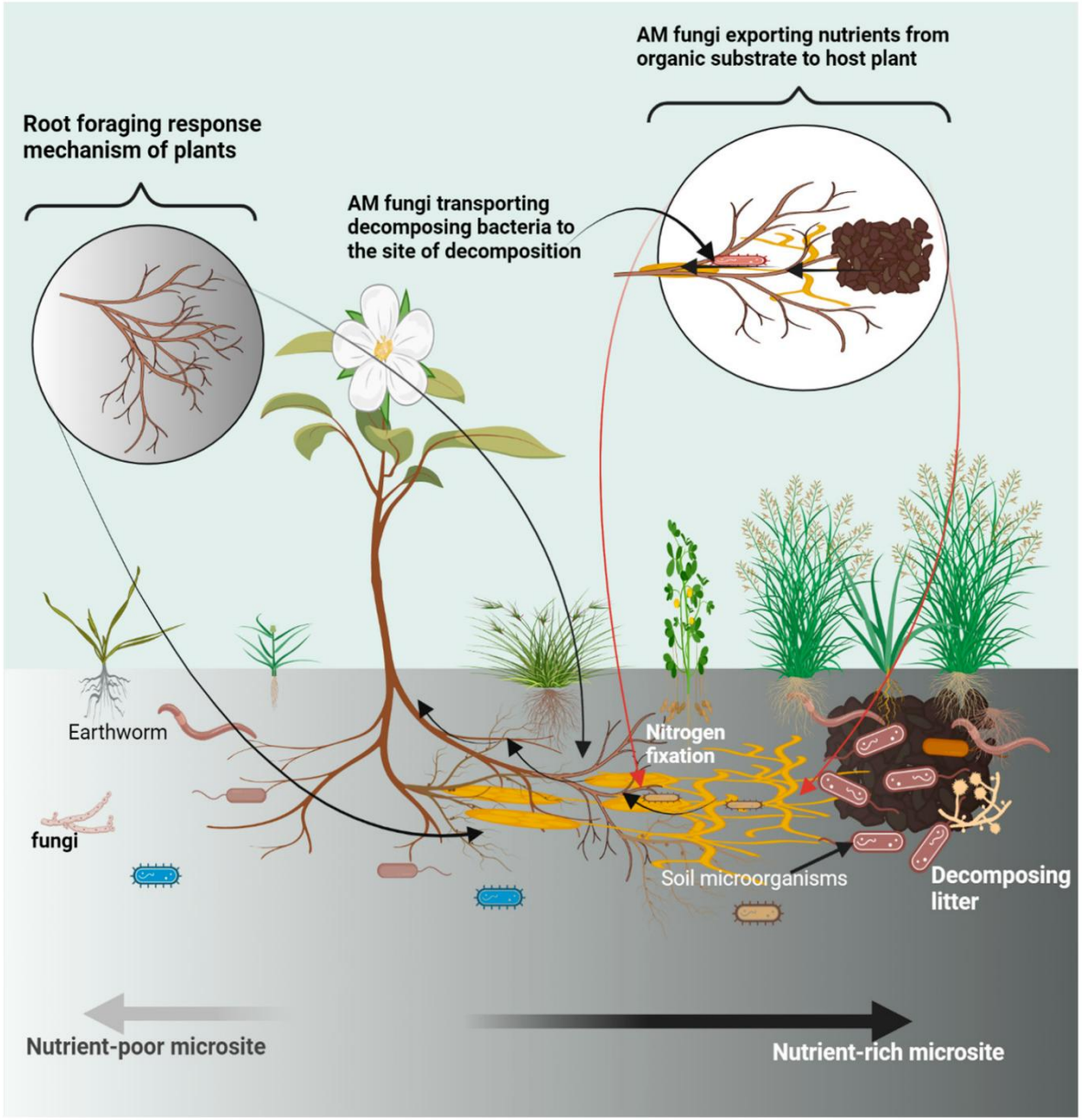






What we study...

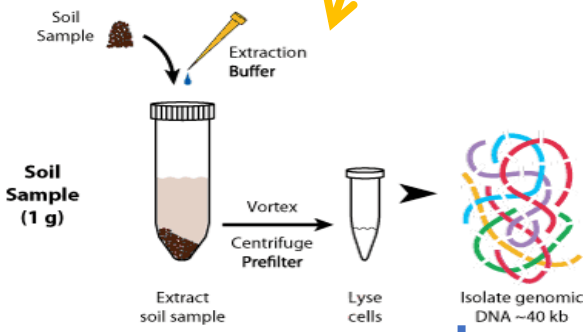




How do we study it...

Sampling → DNA extraction → PCR → Purification → Sequencing → Bioinformatics

PC



Unix server

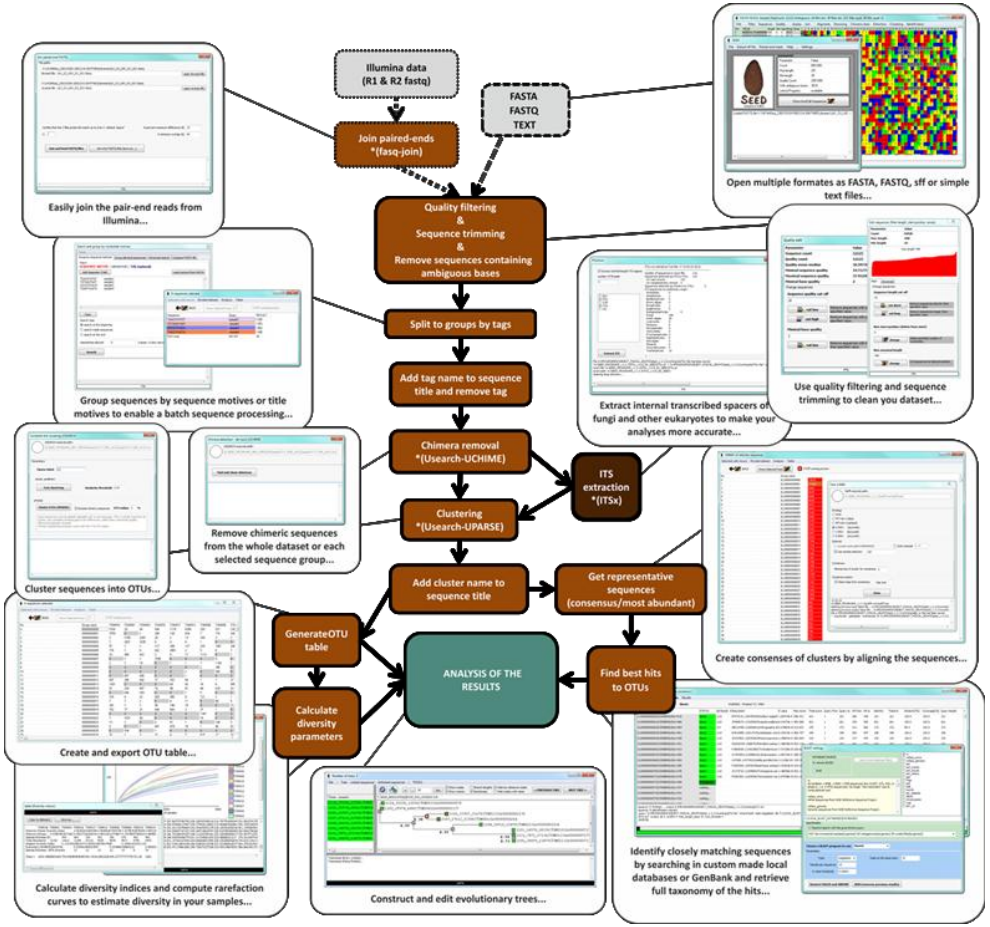
What we have done… (most relevant projects)

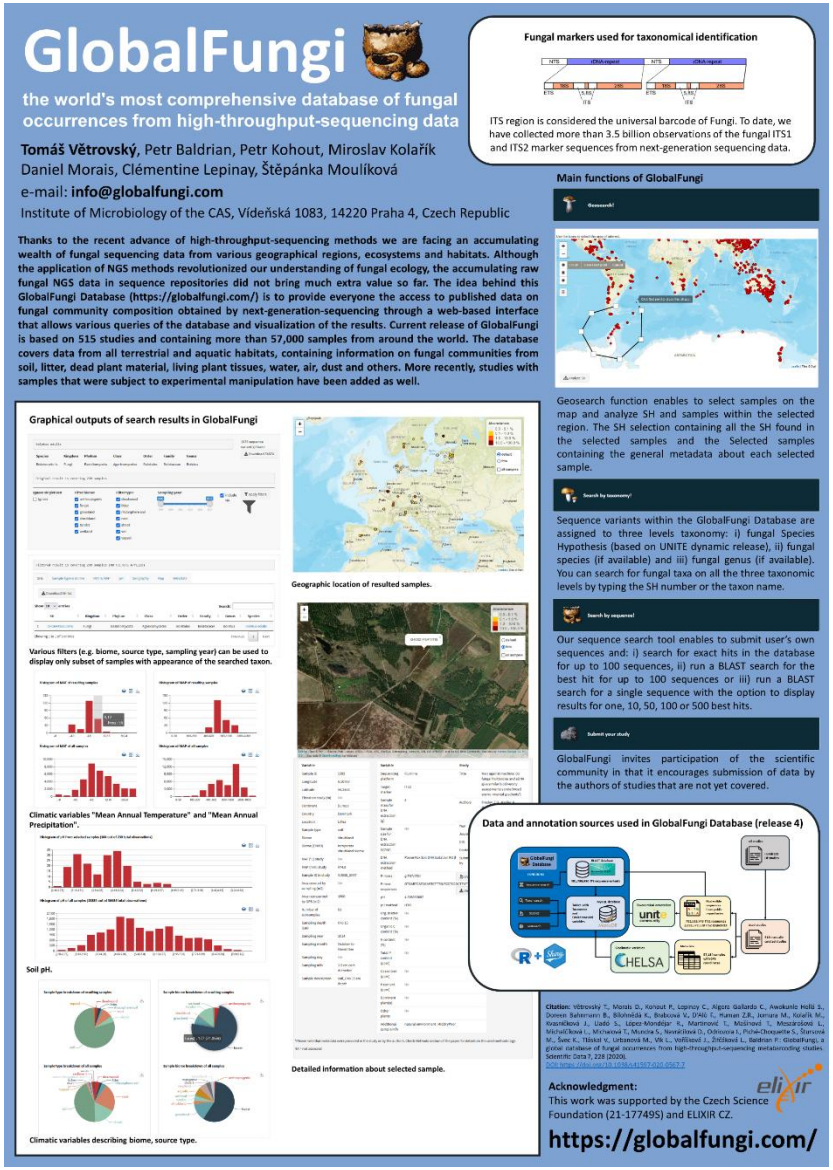




SEED 2: a user-friendly platform for amplicon high-throughput sequencing data analyses

GUI based alternative for Windows (http://www.biomed.cas.cz/mbu/lbwrf/seed/)





Content of the course

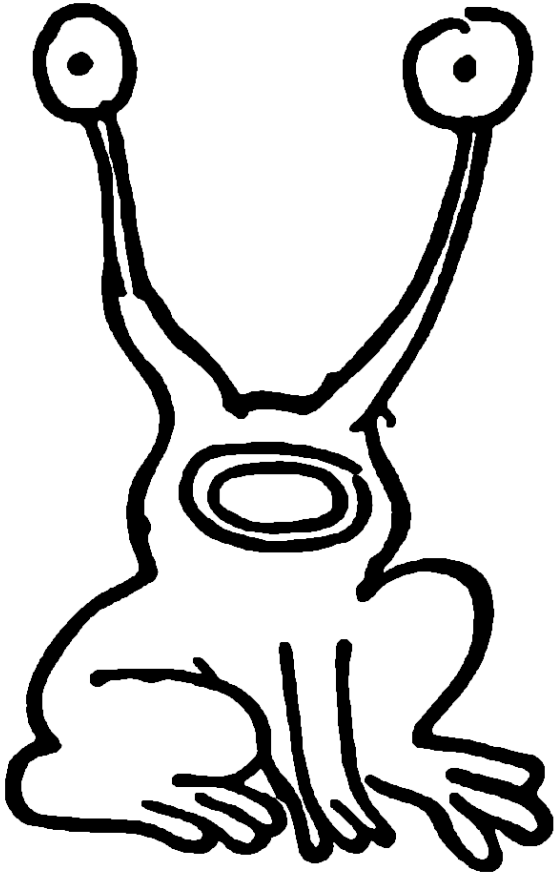
The idea is to understand the algorithm behind the most used programs and to acquire independence to explore other possibilities by yourselves.

Exam

- Oral
- Open schedule - end of the semester
- I would like to ask for suggestions for possible dates

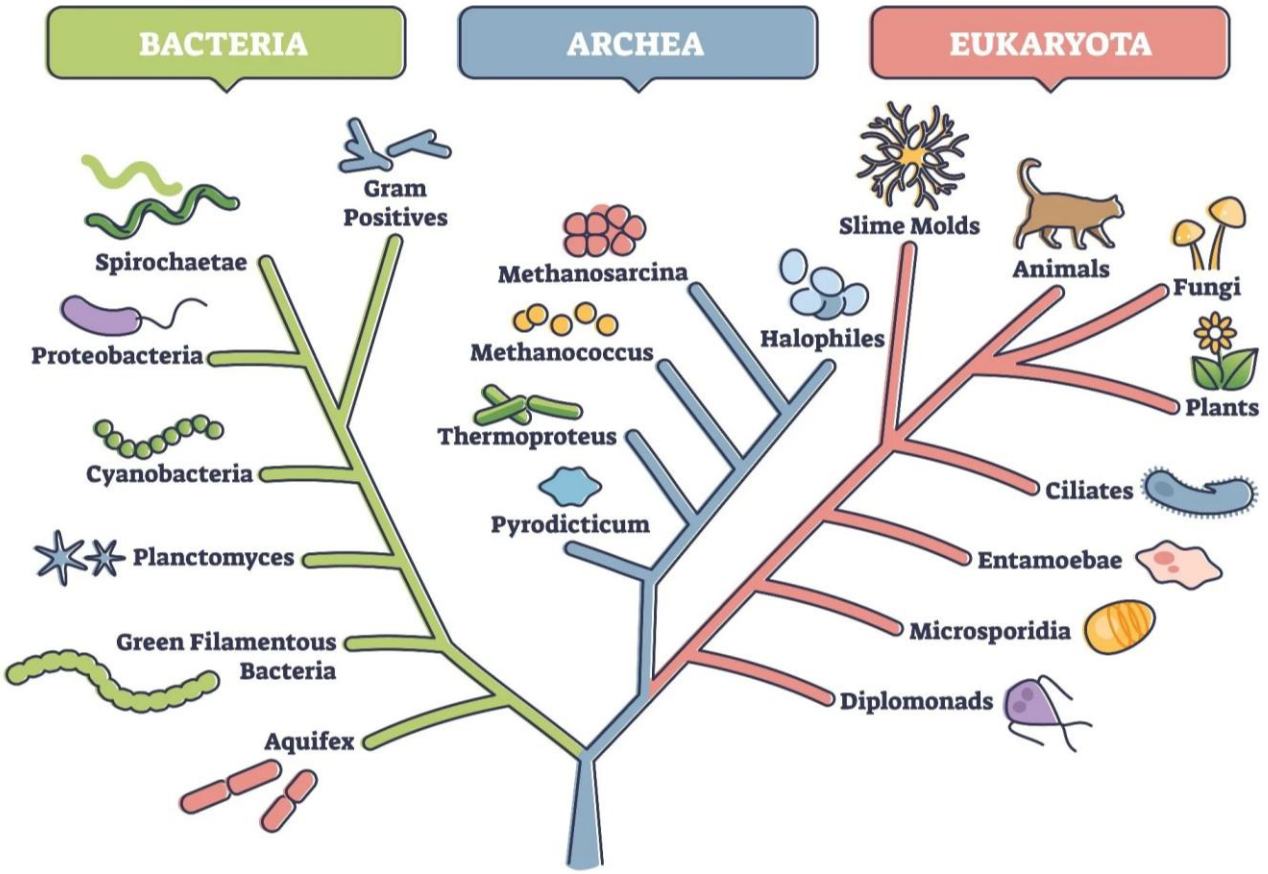
Who are you?
How much we already know?

- What is your computer operating system?
- How familiar are you with command line?
- Are you willing to spend some time to practice?



Phylogenetics

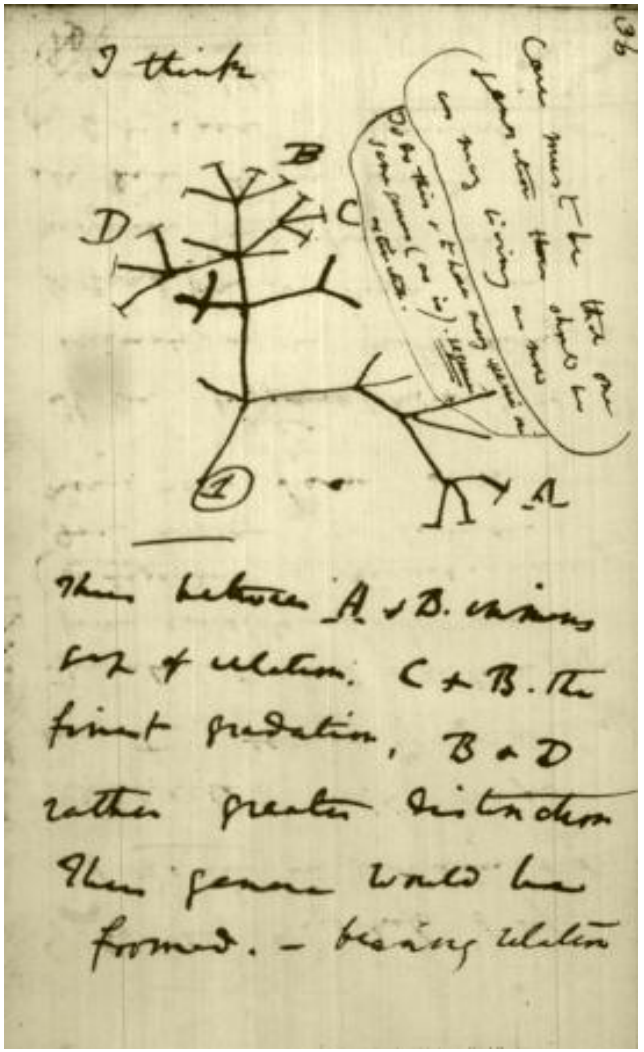
the **evolutionary history and relationships** among or within groups of organisms



Phylogenetics Tree of Life (TOL)

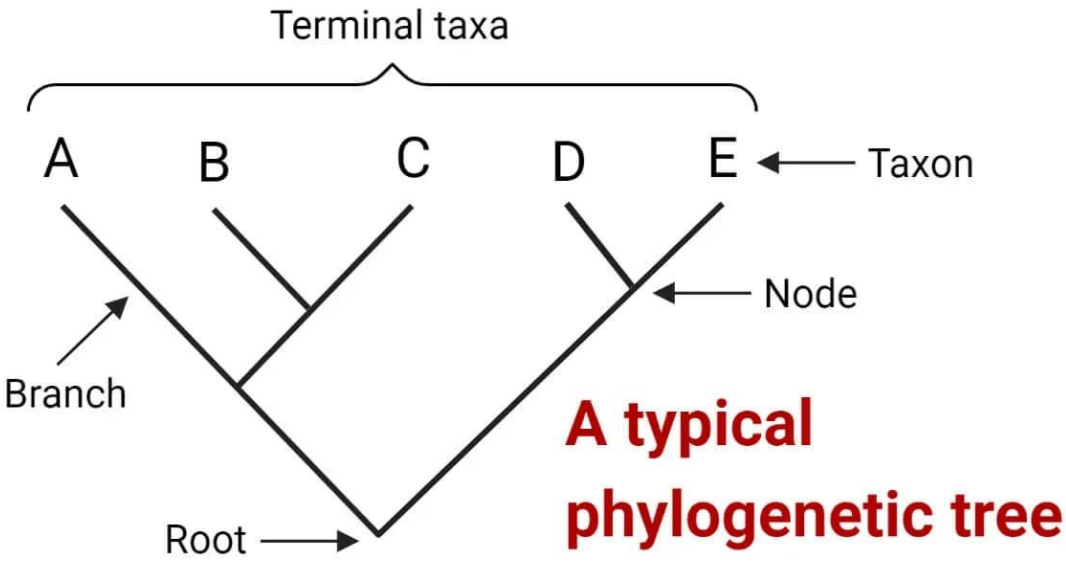
Phylogenetic trees

- **graphical representations** of the history of organisms
- shows **parental relationships** between them

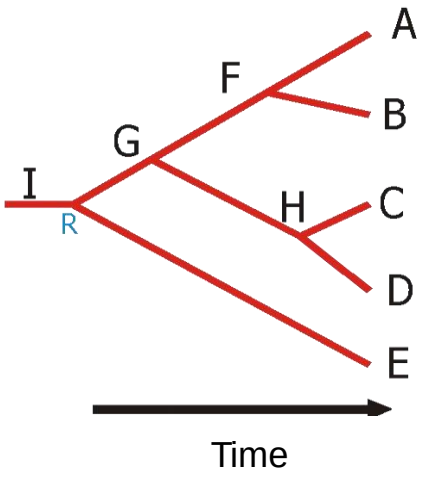


The first phylogenetic tree
Designed by Charles Darwin
(1837)

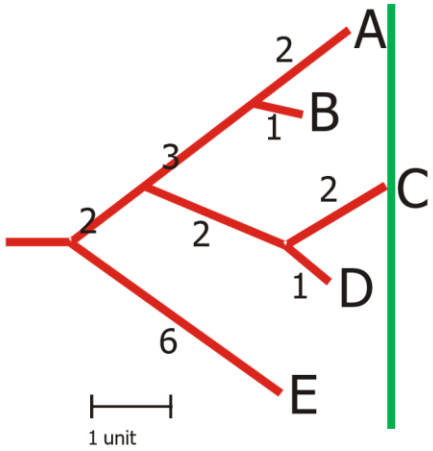
Phylogenetic trees



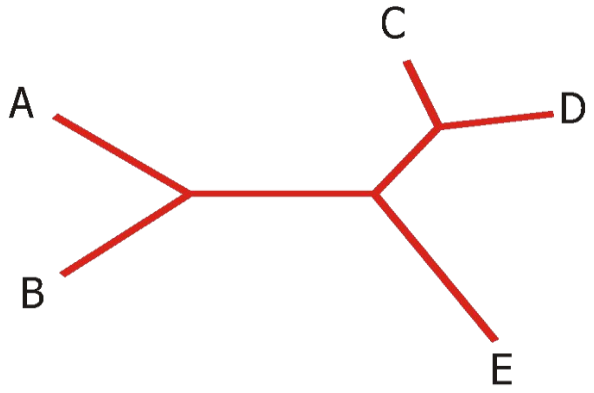
Roots and Outgroups...



Aligned terminal knots



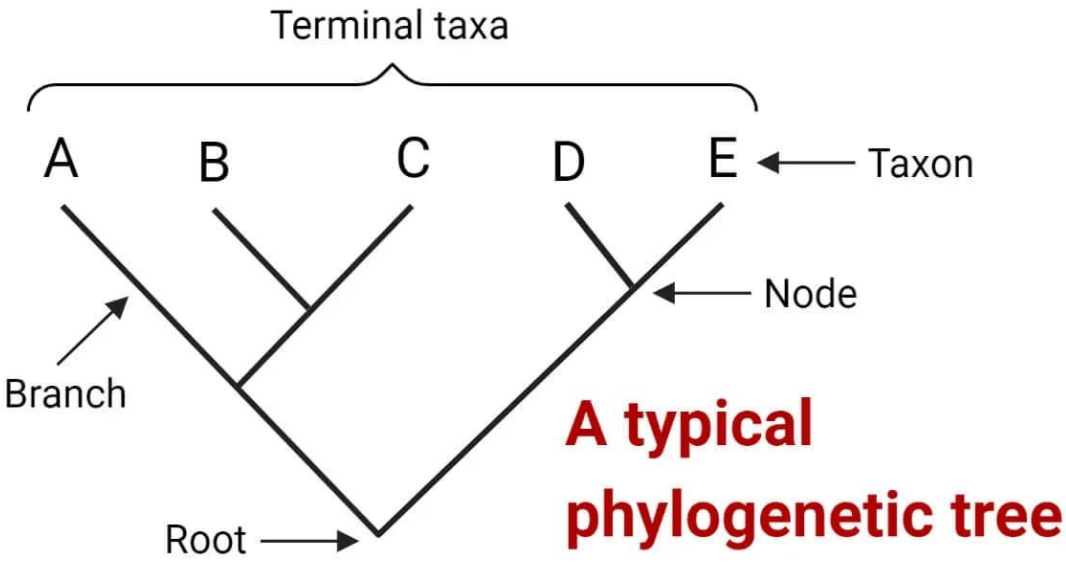
Unaligned terminal knots



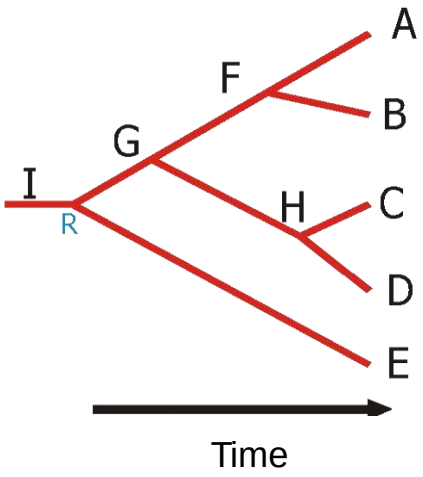
Is the root needed?

Phylogenetic trees

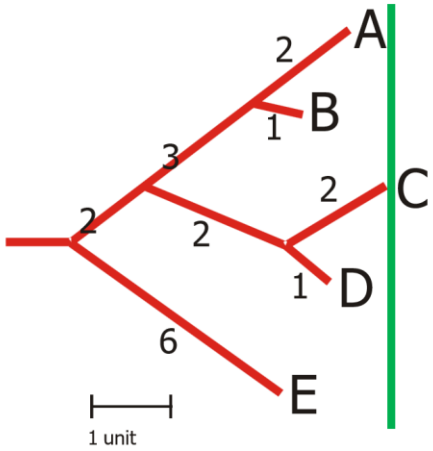
Branch sizes could represent the number of changes from each ancestral node



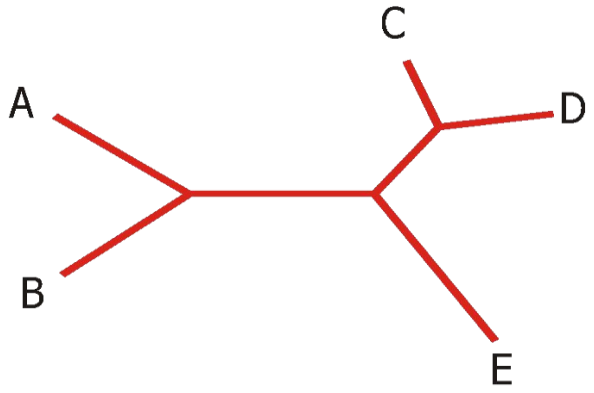
Roots and Outgroups...



Aligned terminal knots



Unaligned terminal knots



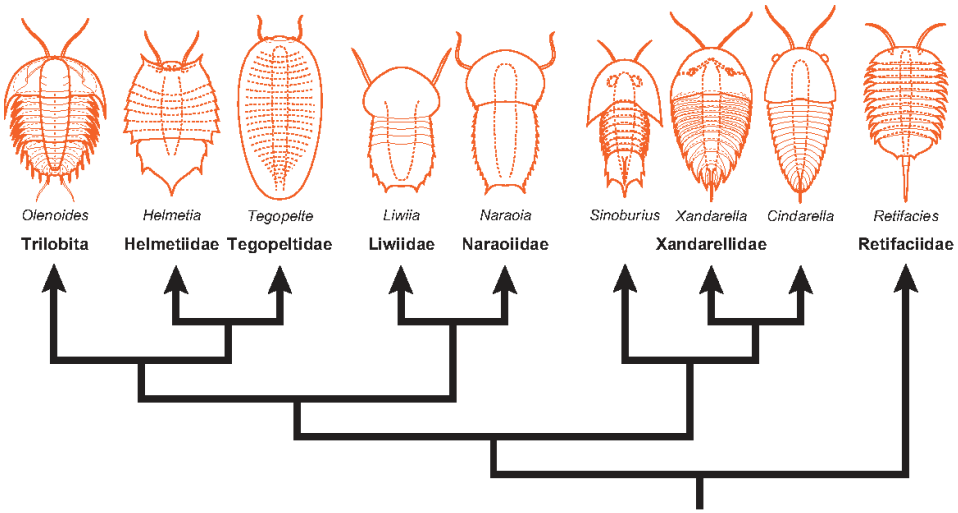
Is the root needed?

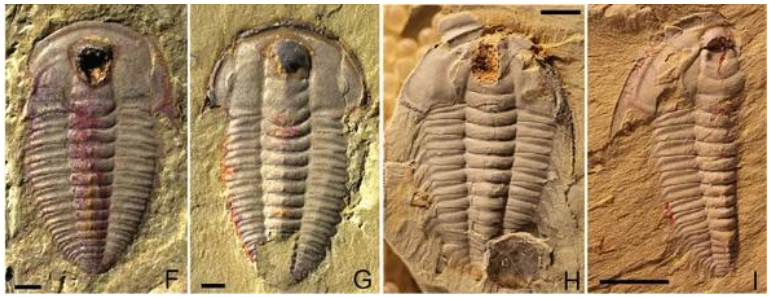
Evolution

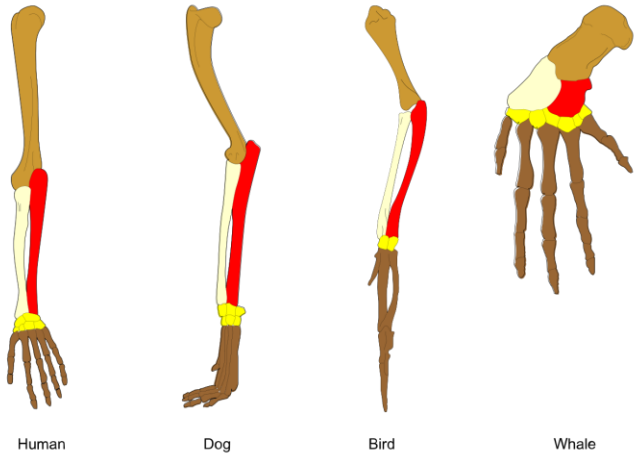
Changes in allelic frequencies (gene variants) over generations in a population.

- Historical science
  - Very complex
    - – it is based on very old events

- Documents
  - Fossil organisms
    - **Morphological characters**
    - **Dating**
  - Current organisms
    - **Morphological characters**
    - Behavior
    - Physiology
    - **Molecules***

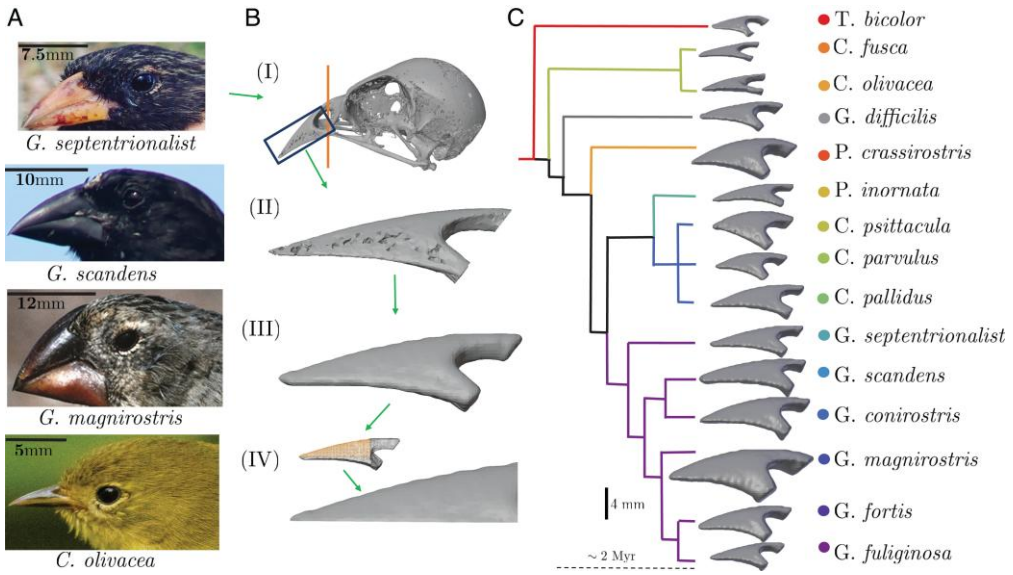






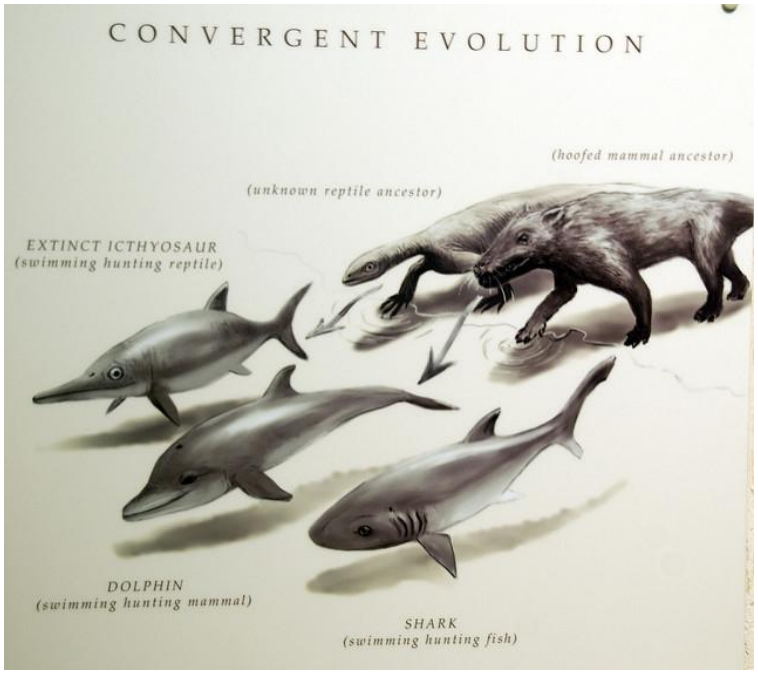
Morphology

- **First applied tool**
  - Allowed important findings (Darwin)
  - Allow fossil studies

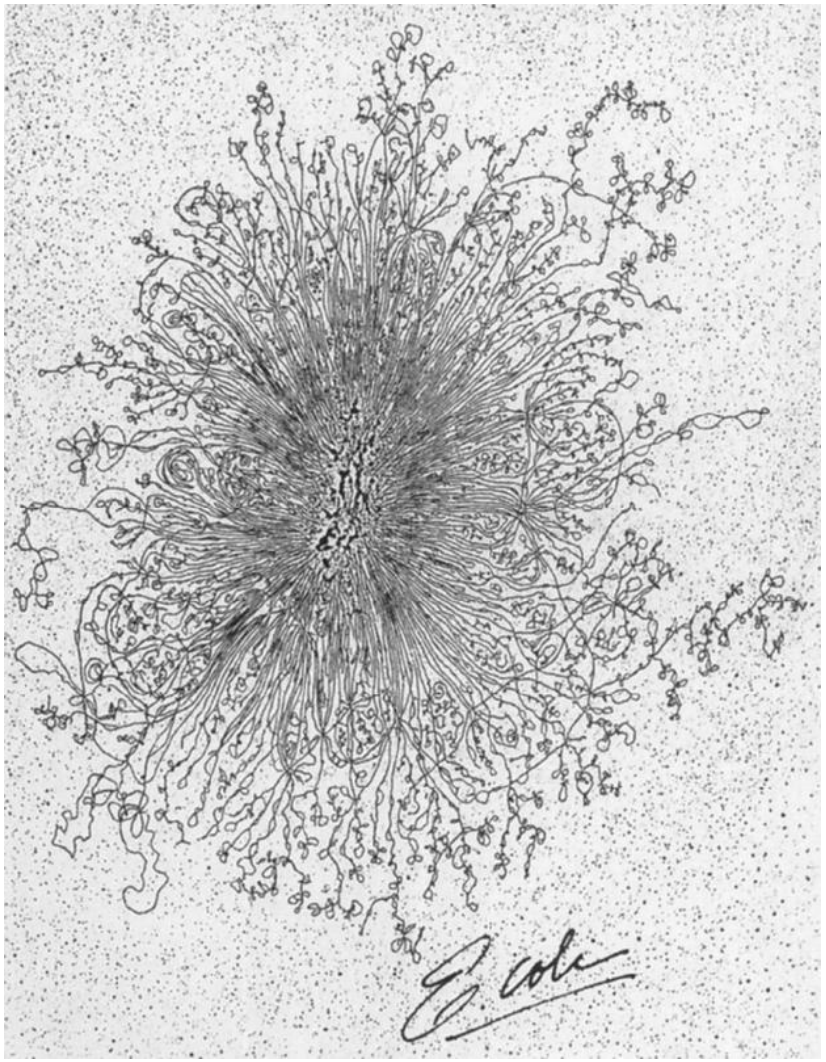


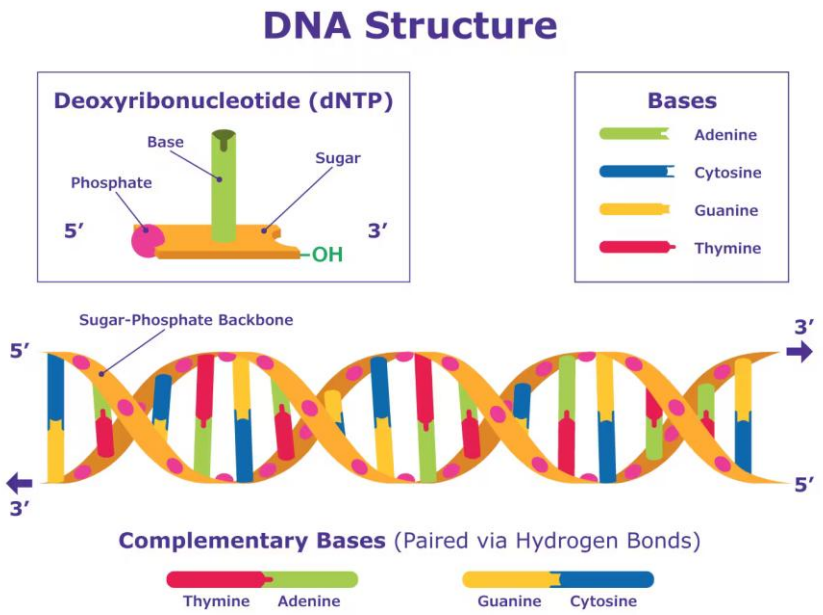
Beak morphology and phylogeny of Darwin's finches.

- **Limitations**
  - influenced by the environment
    - development (age stages)
    - life history (sexual dimorphism)
    - phenotypic plasticity
  - natural selection
    - convergent evolution

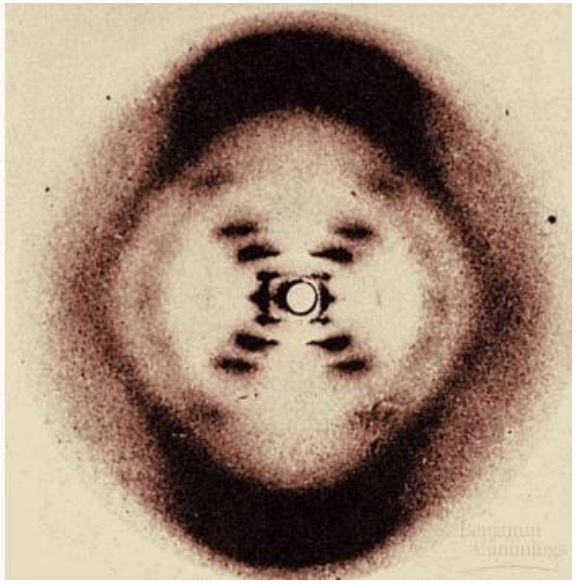


(bio)Molecules - Deoxyribonucleic acid

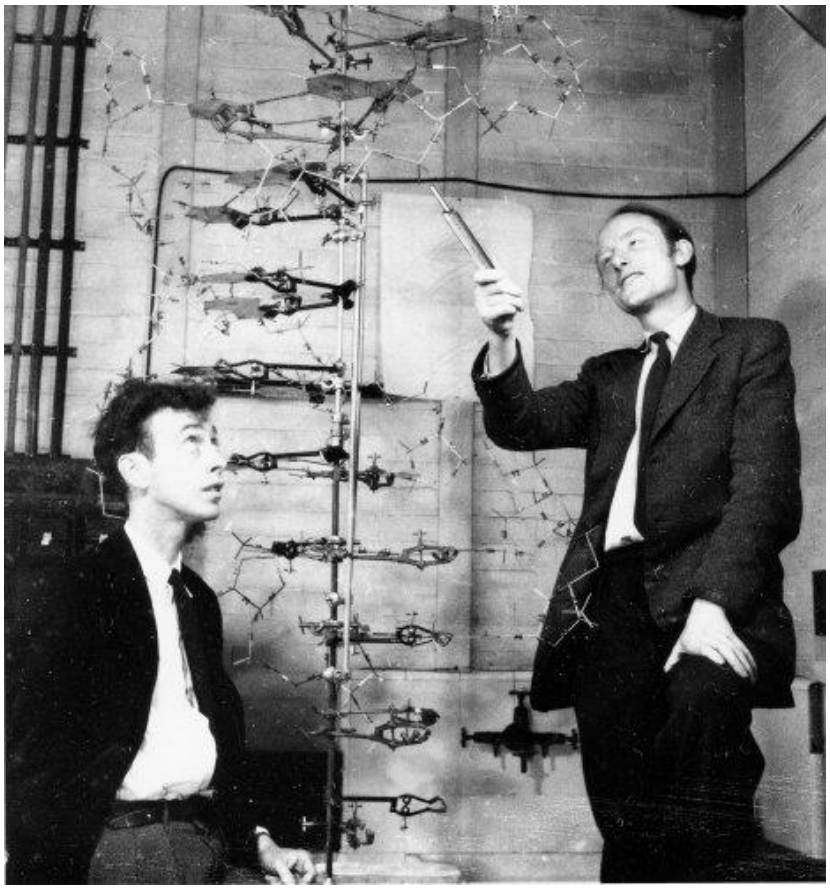




Discovery of DNA



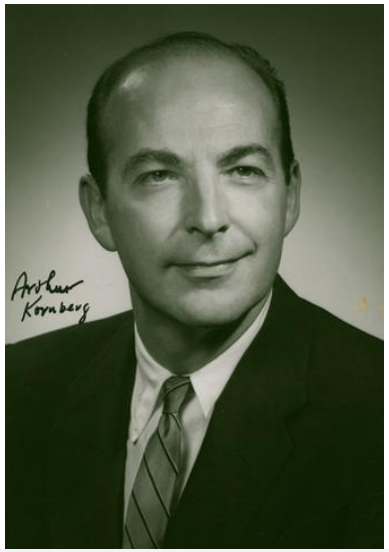
'Photo 51', taken by **Rosalind Franklin** and **Ray Gosling** at King's in **1952**, can claim to be one of the world's most important photographs. It demonstrated the helical structure of DNA and, with their own deductions, enabled James Watson and Francis Crick of the University of Cambridge to build the first correct model of the DNA molecule.



The discovery in **1953** of the double helix, the twisted-ladder structure of deoxyribonucleic acid (DNA), by **James Watson and Francis Crick**

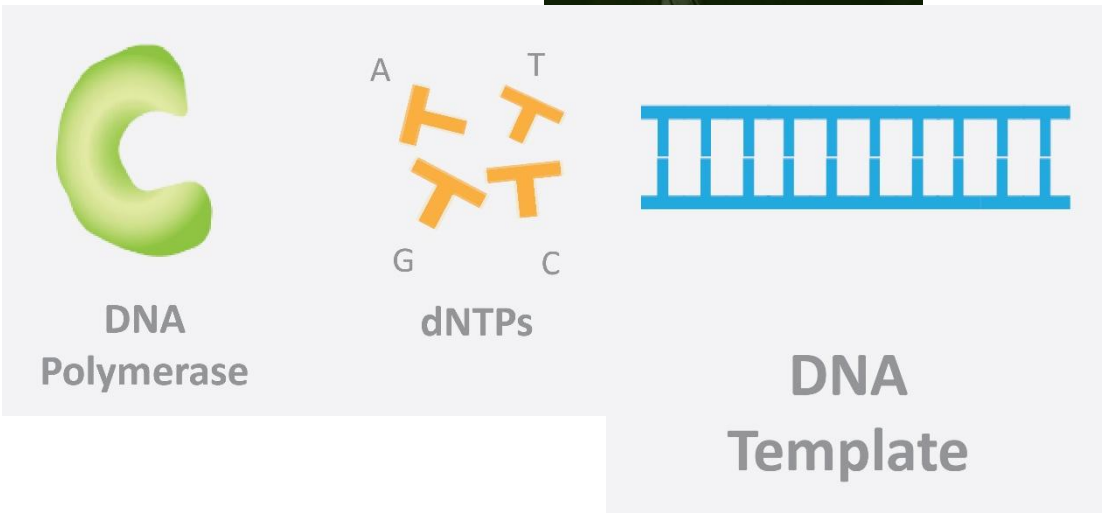
DNA sequencing – how to copy DNA?

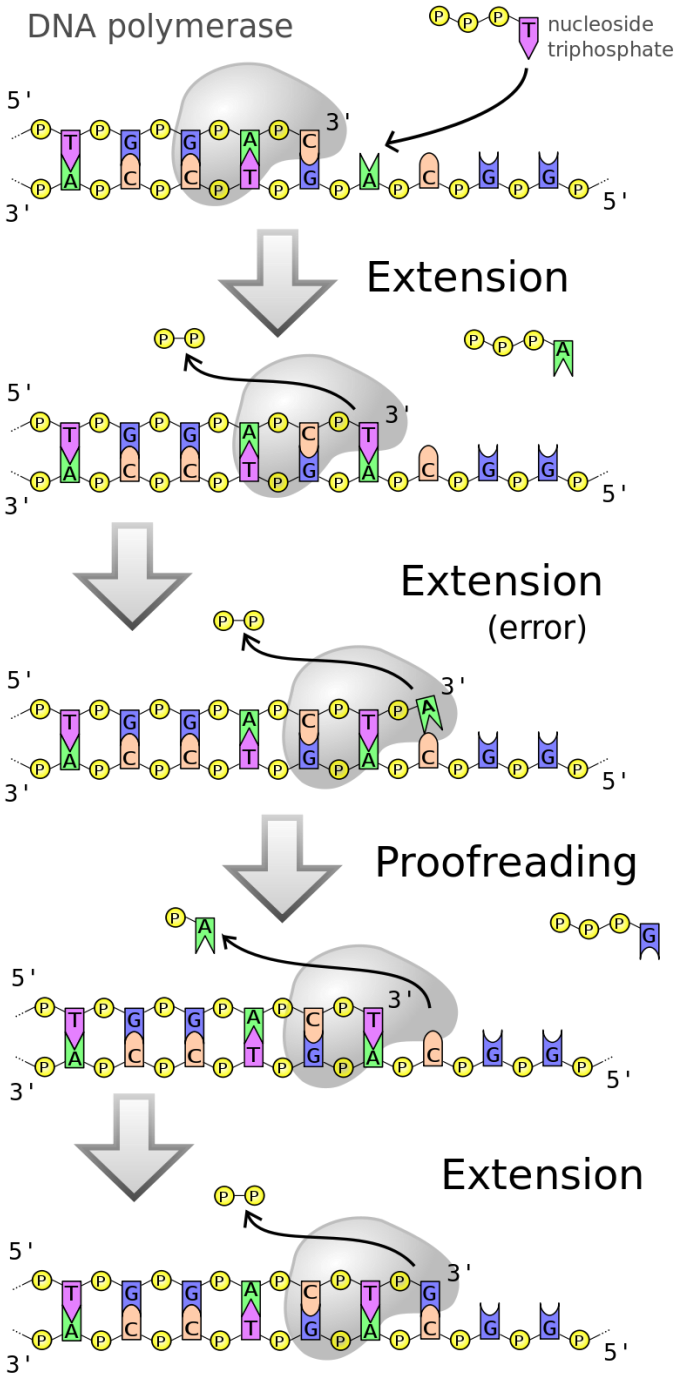
1956, **Arthur Kornberg** and his team of biochemists were the first to isolate and later characterize the enzyme which is now known as DNA polymerase I.



DNA polymerase

family of enzymes that catalyze the synthesis of DNA molecules from nucleoside triphosphates (dNTPs)

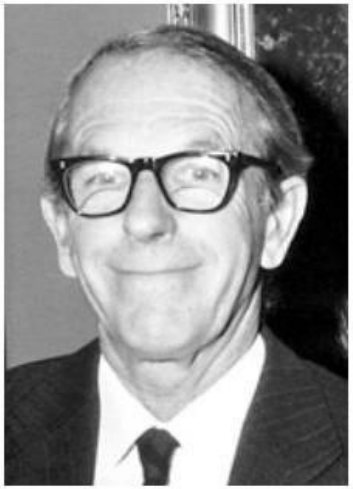




The Nobel Prize in Chemistry 1980

Paul Berg
Prize share: 1/2

Walter Gilbert
Prize share: 1/4

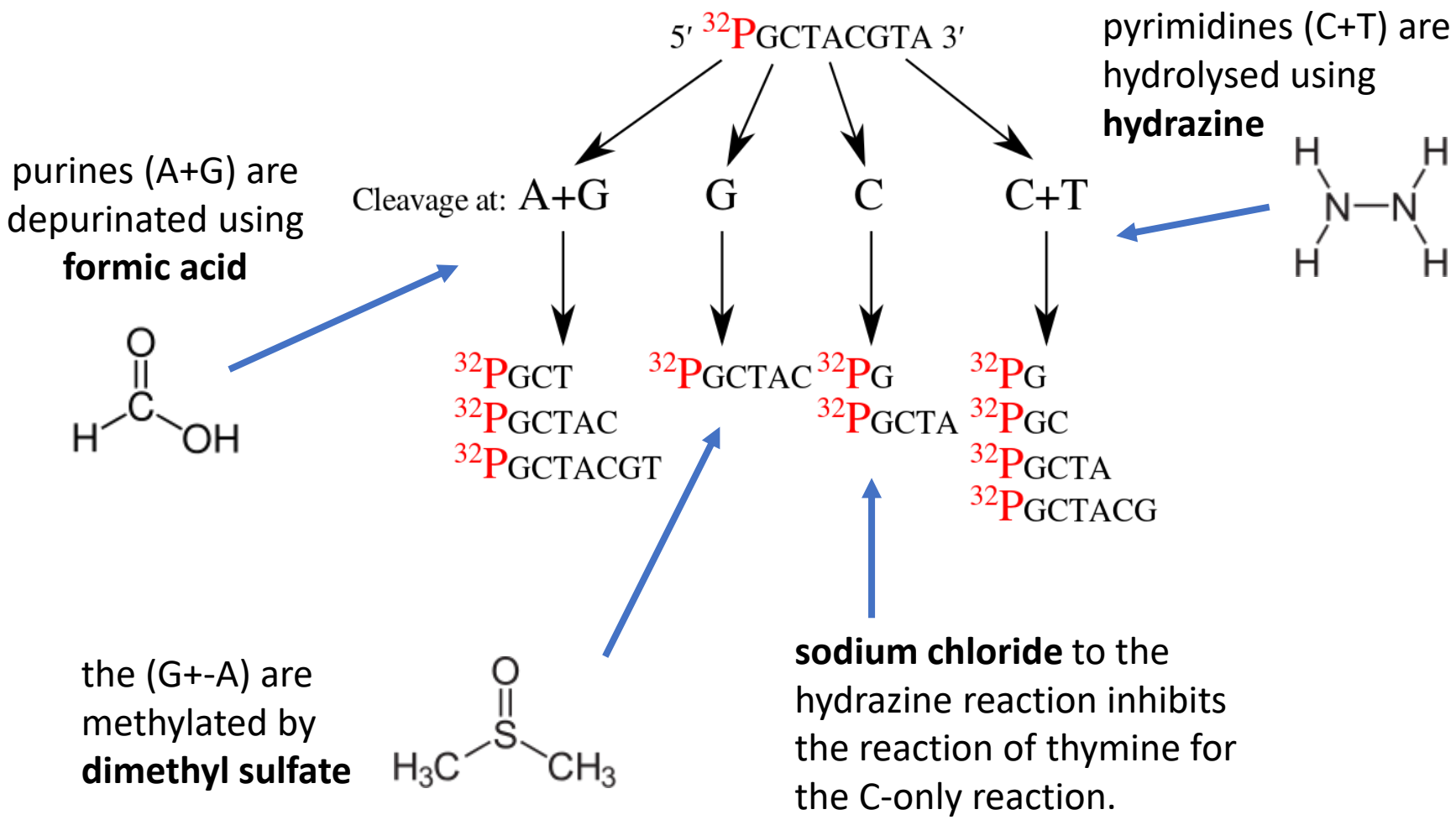


Frederick Sanger
Prize share: 1/4

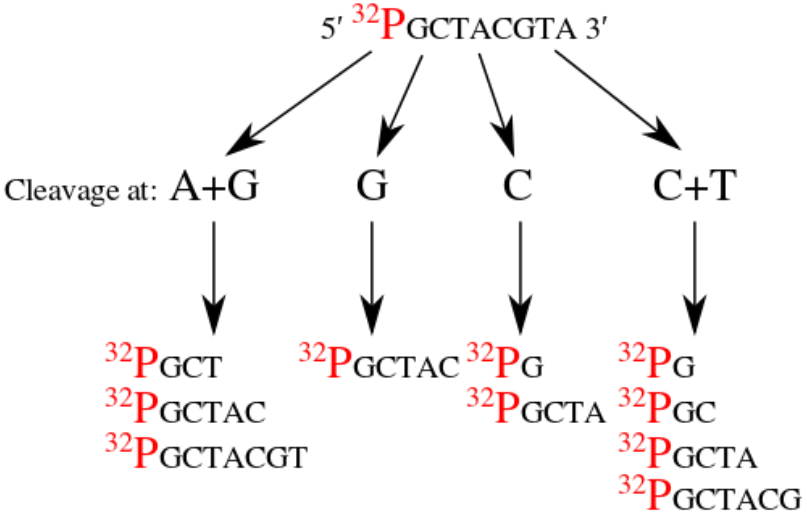
The Nobel Prize in Chemistry 1980 was divided, one half awarded to Paul Berg *"for his fundamental studies of the biochemistry of nucleic acids, with particular regard to recombinant-DNA"*, the other half jointly to Walter Gilbert and Frederick Sanger *"for their contributions concerning the determination of base sequences in nucleic acids"*.

Maxam-Gilbert Chemical Sequencing (1976)

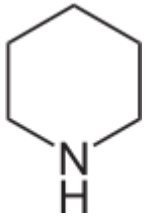
radioactive labeling at one 5′ end of the DNA fragment to be sequenced (typically by a kinase reaction using gamma-^{32}P ATP)



Maxam-Gilbert Chemical Sequencing (1976)

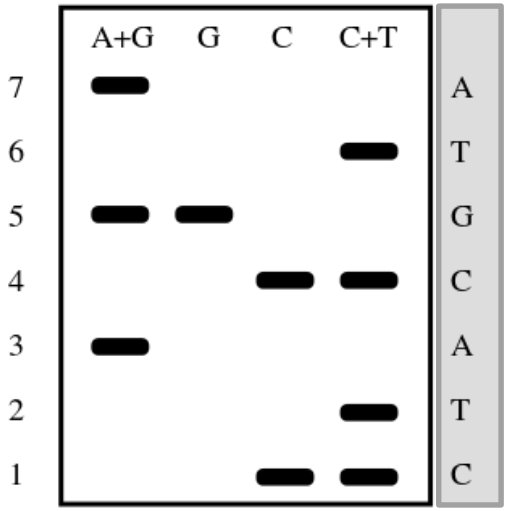


The modified DNAs may then be cleaved by hot **piperidine**



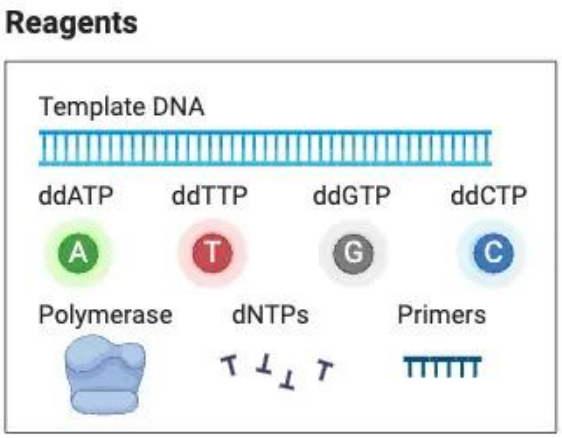
The fragments in the four reactions are **electrophoresed** side by side in denaturing acrylamide gels for size separation.

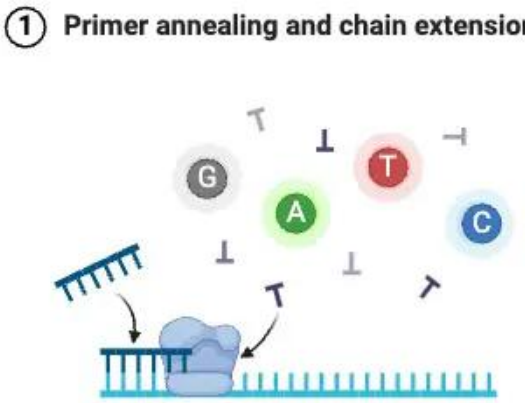
To visualize the fragments, the gel is exposed to X-ray film for autoradiography,

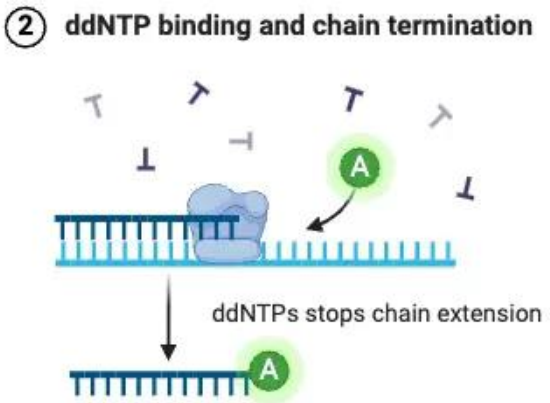


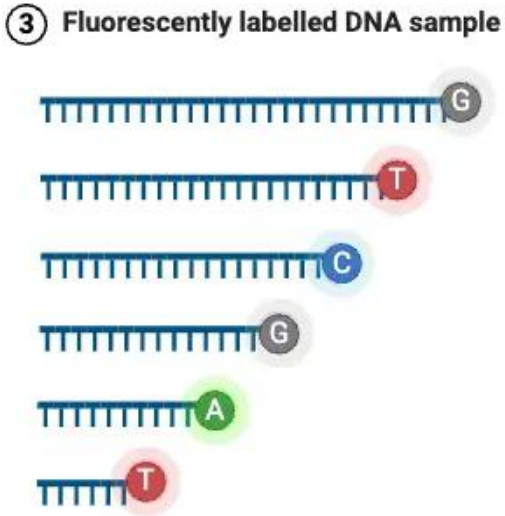
Sequencing Gel

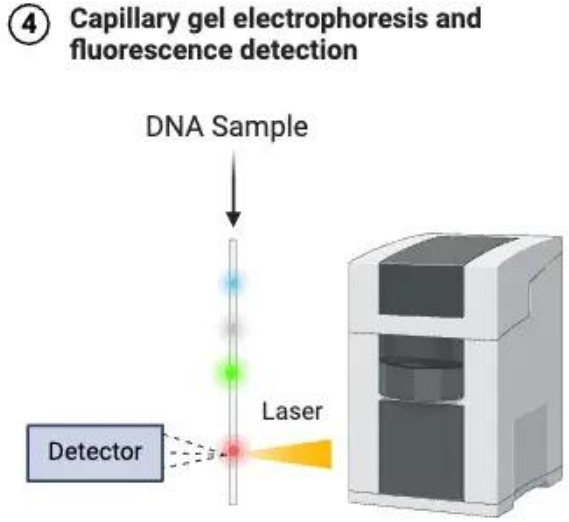
DNA sequencing – Sanger sequencing (1977)

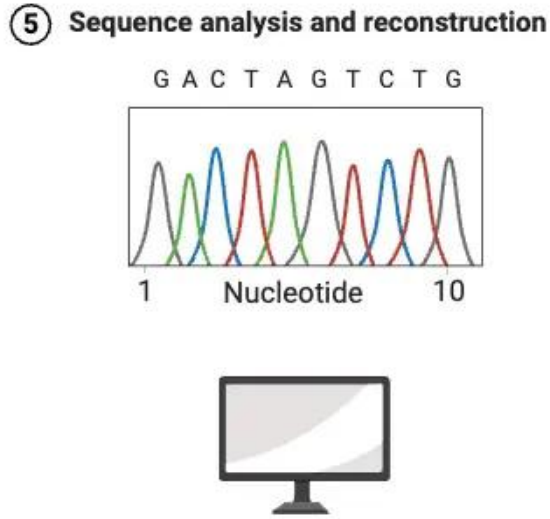








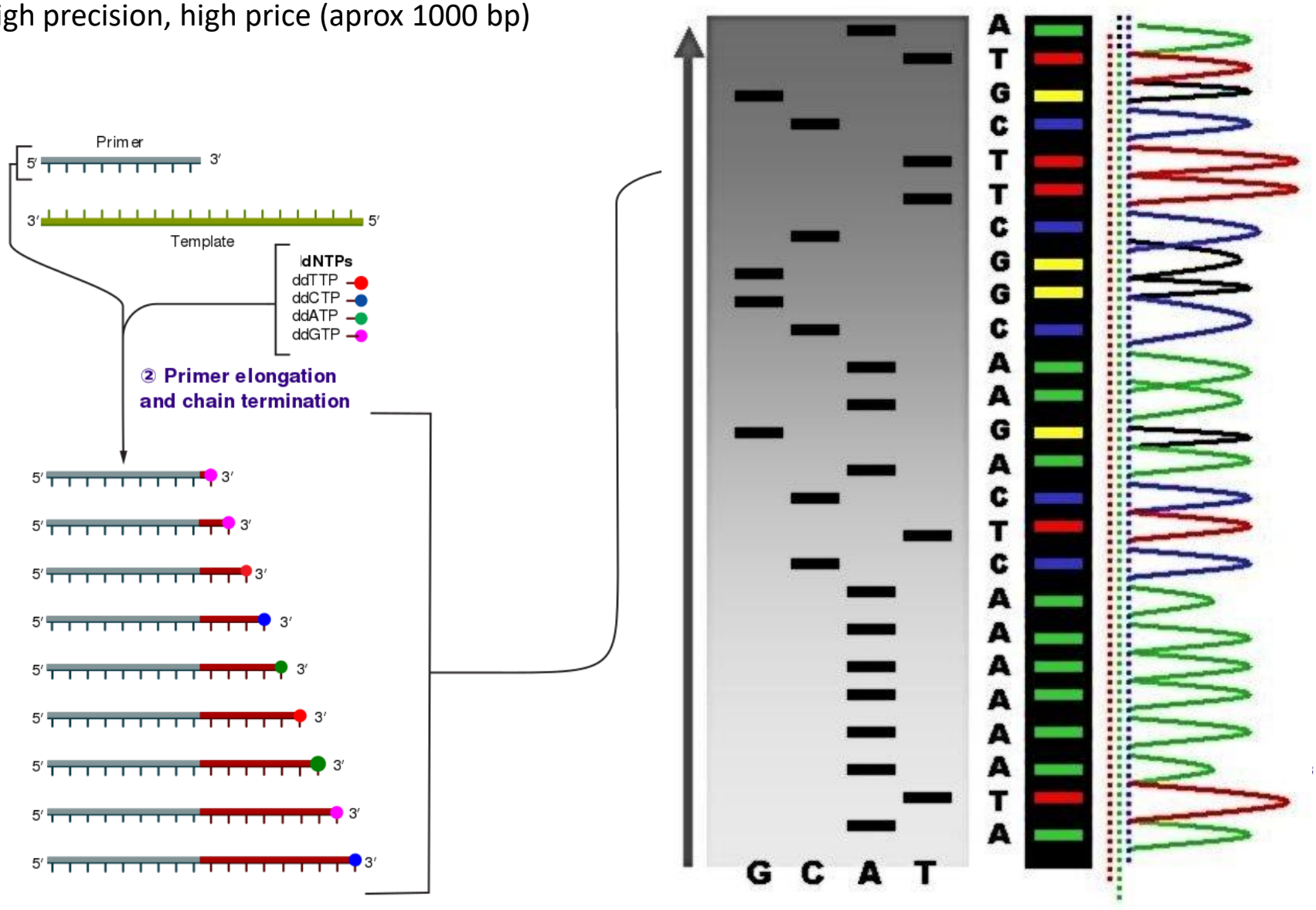




chain-termination DNA sequencing – using Fluorescent ddNTP molecules
high precision, high price (aprox 1000 bp)

DNA sequencing – Sanger sequencing (1977)

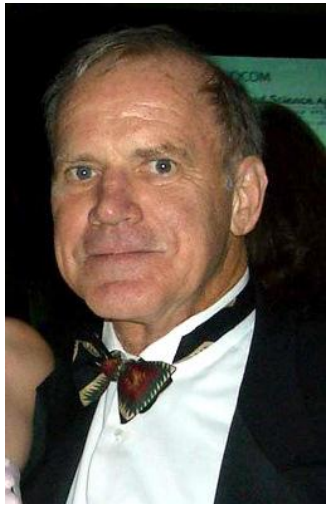
chain-termination DNA sequencing – using Fluorescent ddNTP molecules
high precision, high price (aprox 1000 bp)



DNA sequencing – how to copy DNA?

Polymerase chain reaction (PCR)

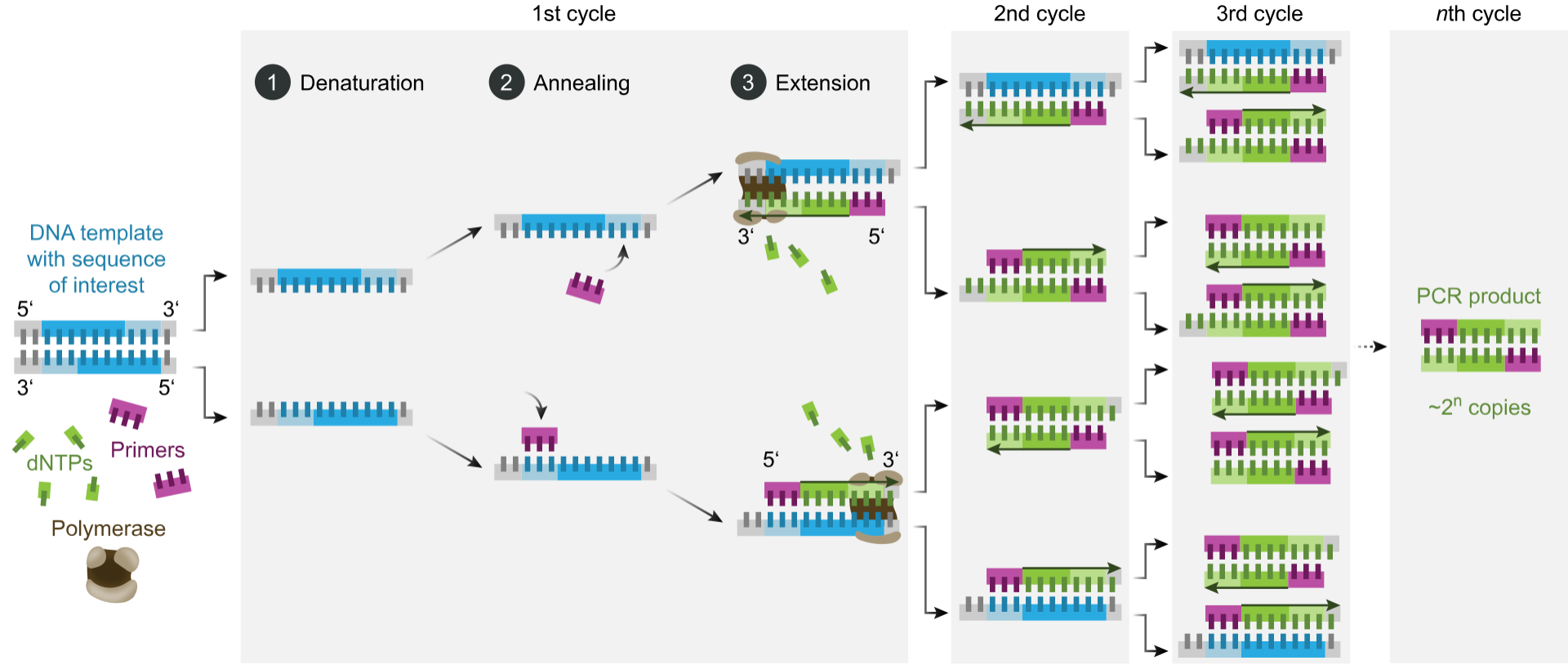
method widely used to rapidly
make millions to billions of copies
of a specific DNA sample



Kary Banks Mullis (1983)

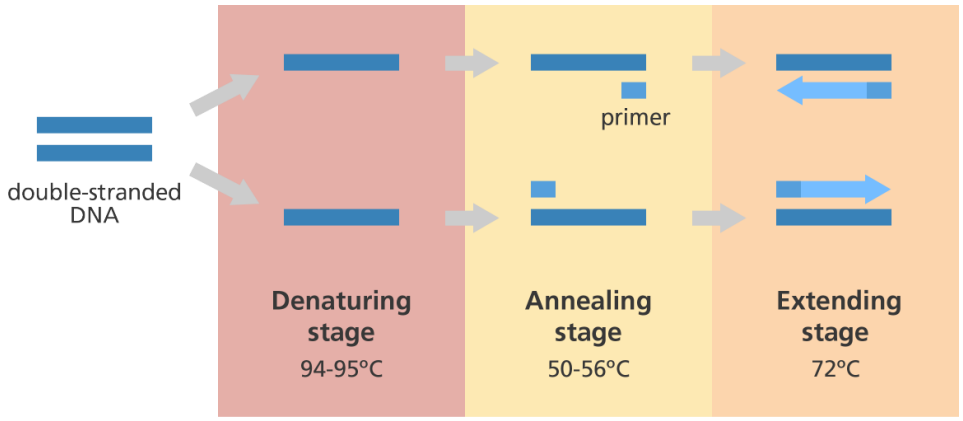
the idea to use a pair of primers

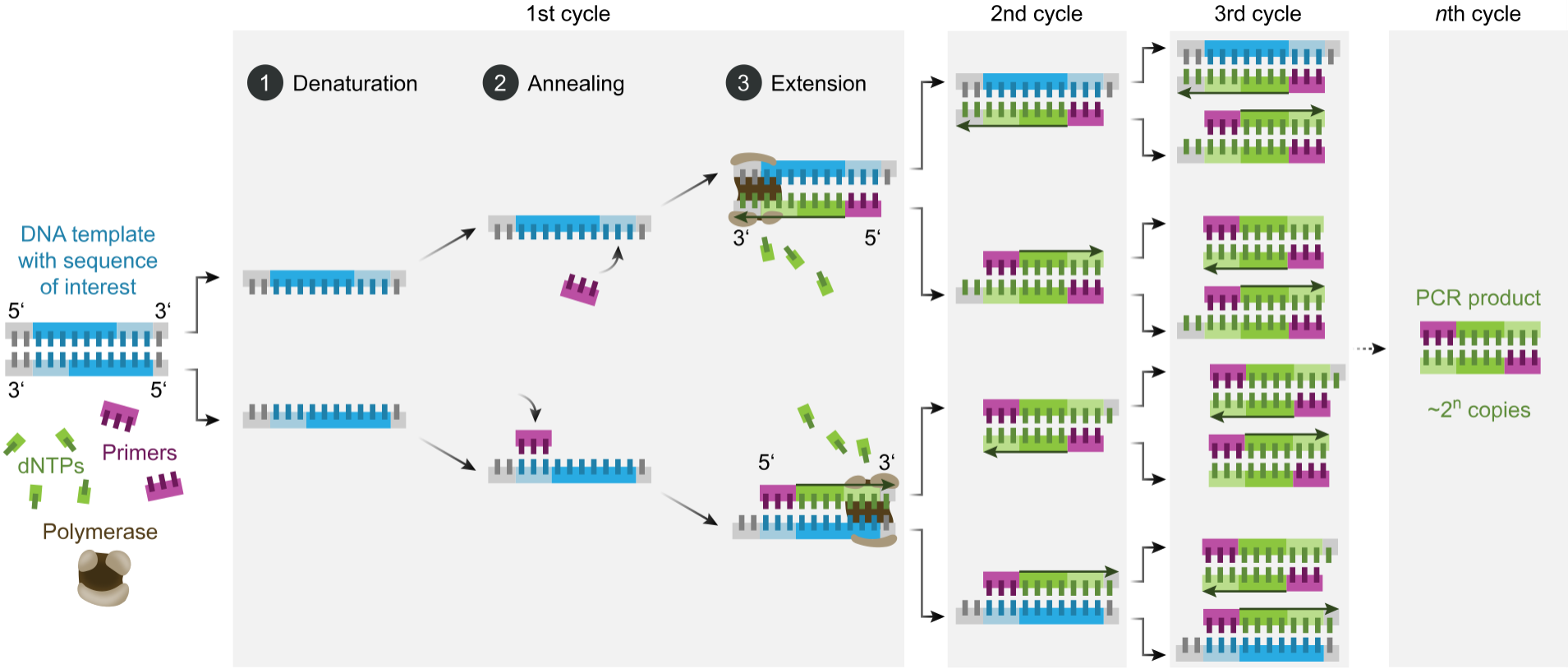
1993 Nobel Prize in Chemistry



DNA sequencing – how to copy DNA?

1986, **Randall Saiki** use *Thermophilus aquaticus* (Taq) DNA polymerase to amplify segments of DNA.

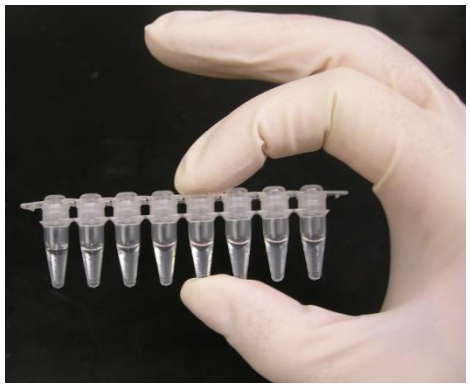




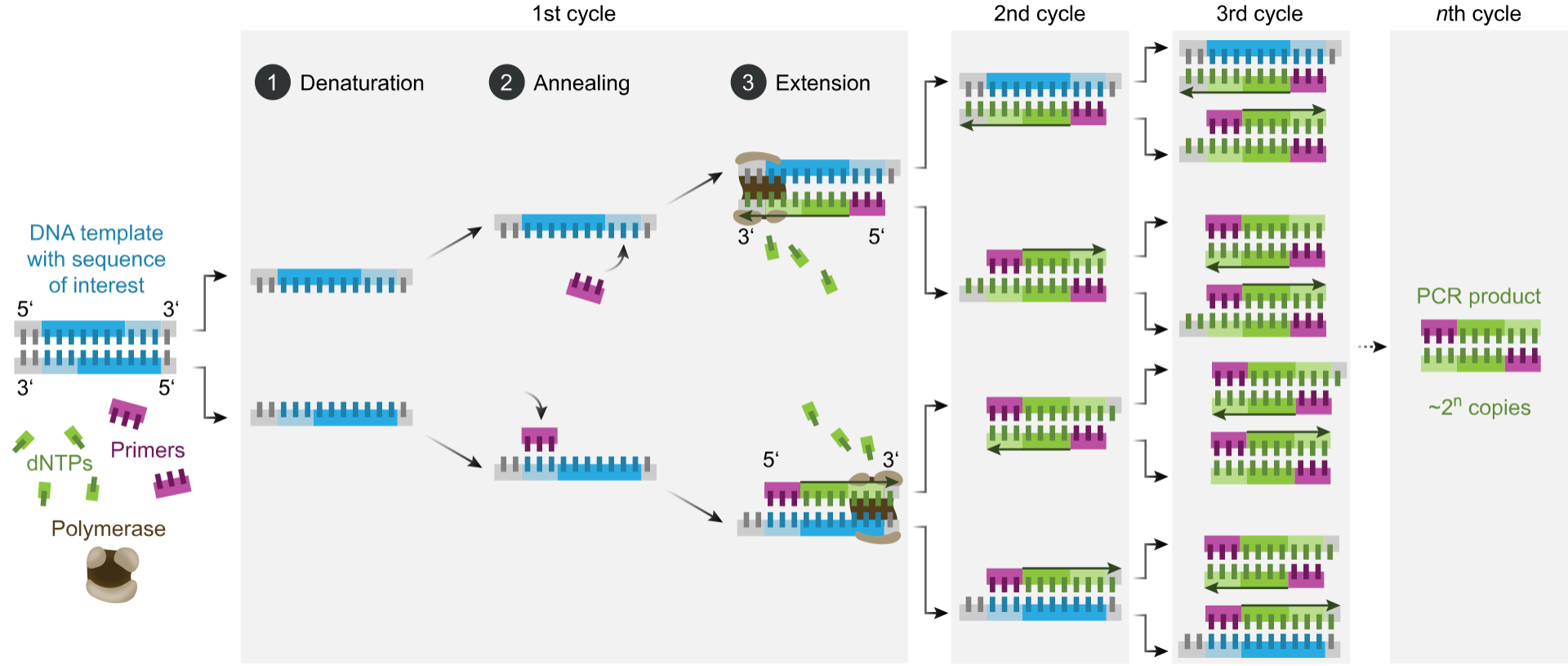
DNA sequencing – how to copy DNA?

Polymerase chain reaction (PCR)

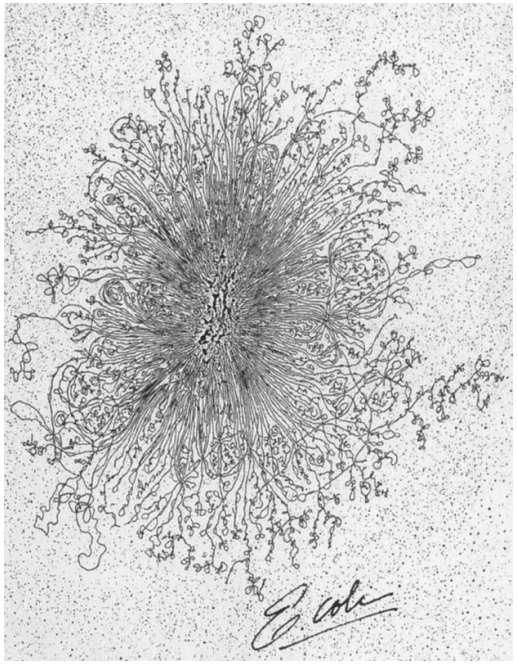
method widely used to rapidly
make millions to billions of copies
of a specific DNA sample



thermal cycler



Molecules - advantages

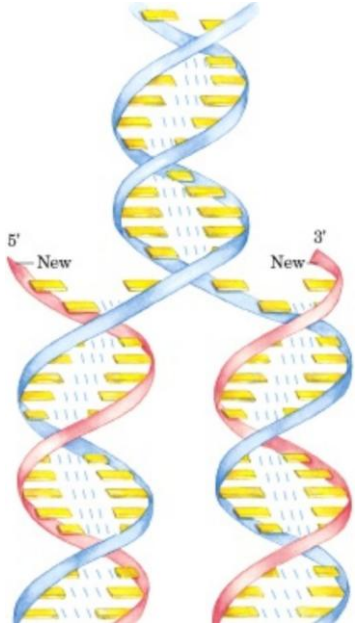


Are genetic

We know how they are inherited, it does not depend on the environment. This is exactly the level where evolutionary innovations arise - **mutations in DNA.**

There's a lot of information

The size of genomes ranges from $0.5 * 10^6$ - $600 * 10^9$. The human genome contains over 3 billion base pairs. It is estimated that people differ from each other in 0.1%, i.e. 3 million bases.



Molecules - advantages

Are selectively neutral

We can better distinguish **homology** and **homoplasy**

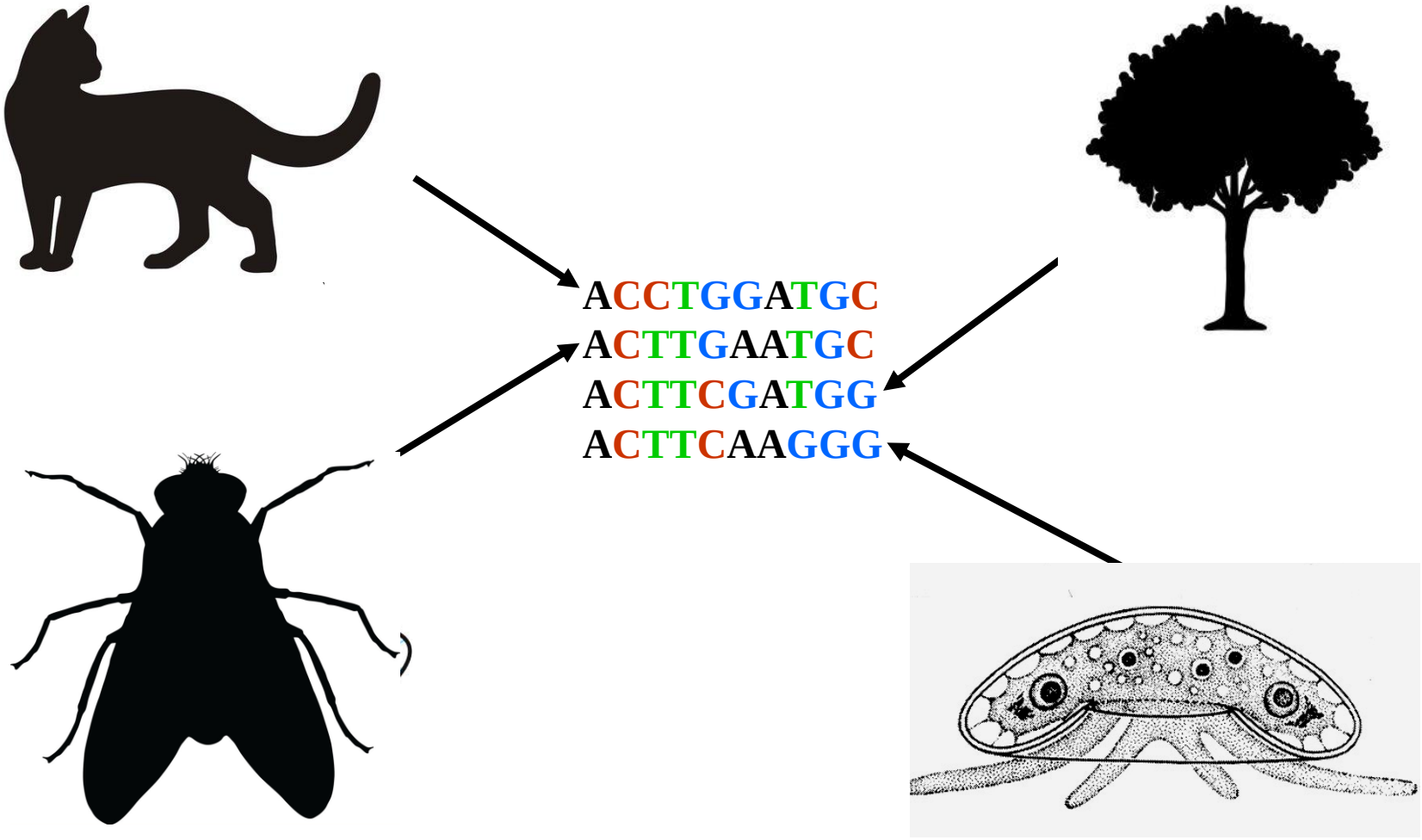
Bat

Bird

Butterfly

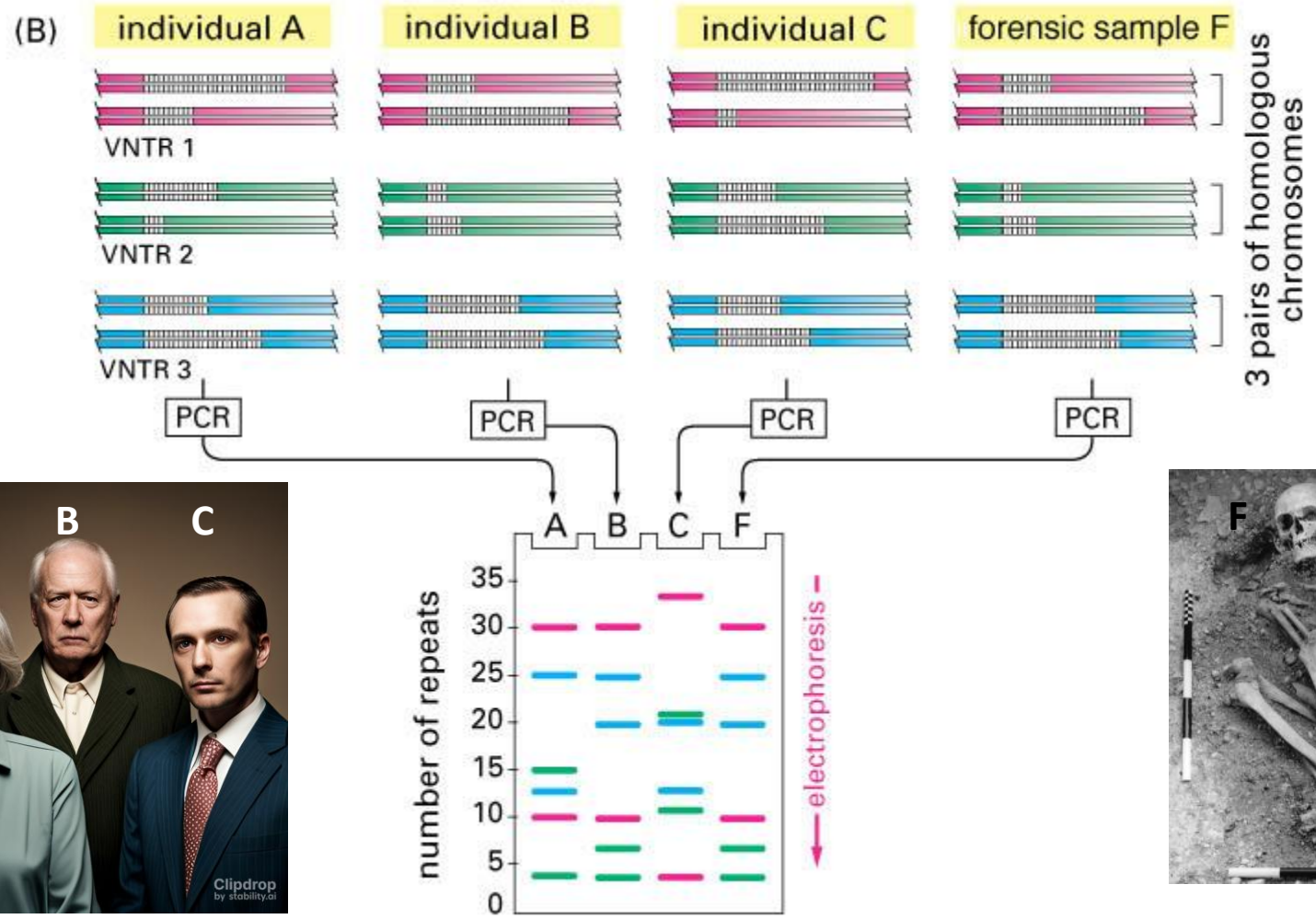
Molecules - advantages

Are useful for the most distant comparisons…



Molecules - advantages

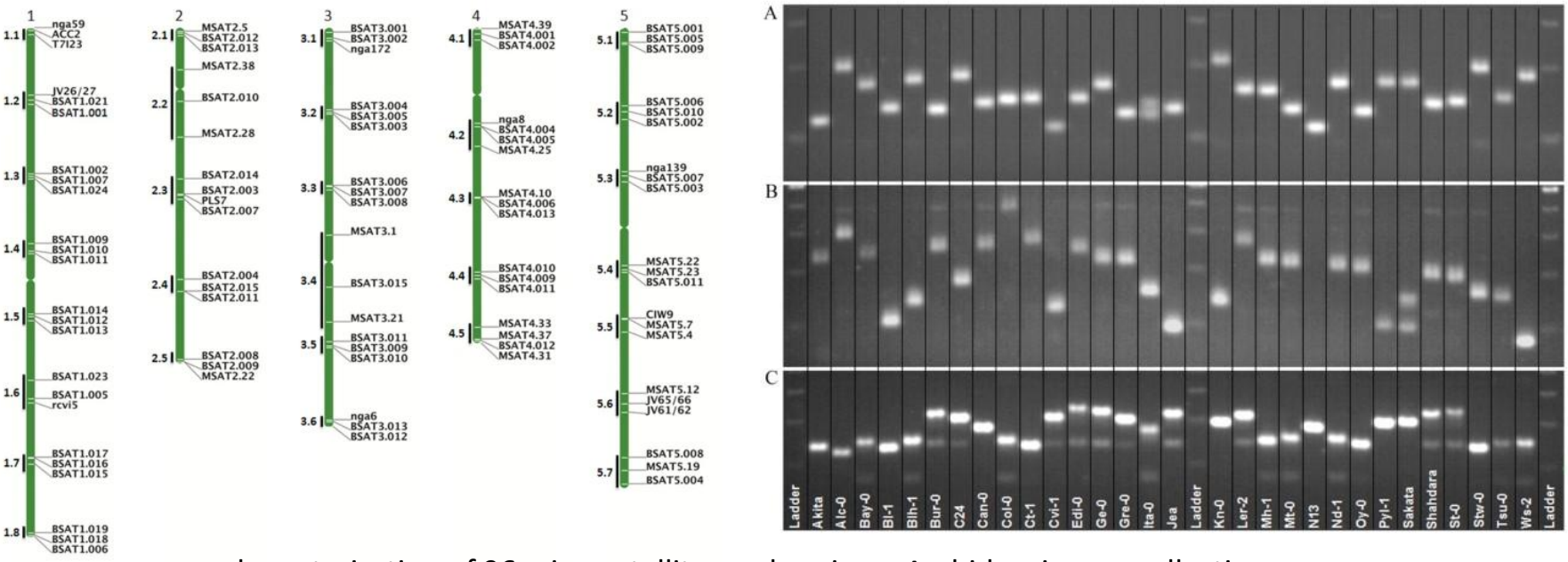
… to comparisons of individuals within species.



(VNTR - variable number tandem repeat)

Gene choice

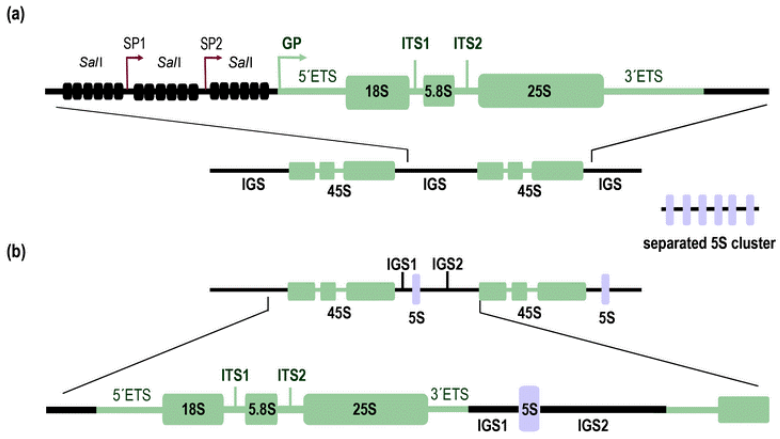
- According to the rate of nucleotide substitutions, considering the estimated time of divergence of the organisms being compared
- **Pseudogenes, introns and intergenic regions** are suitable for closely related species or populations (**variable**)
- **Histones** are indicated for phylogenies between **kingdoms** (**conserved**)
- The most appropriate is to test several genes for the same problem, check the phylogenetic signal and evaluate similar results.



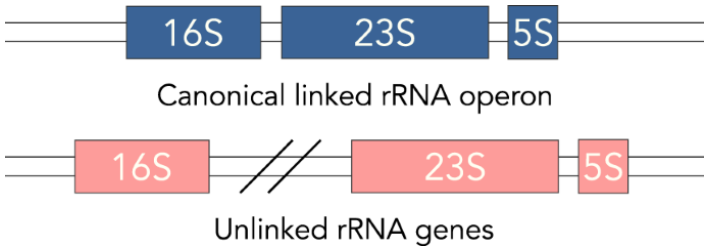
characterization of 96 microsatellite markers in an Arabidopsis core collection

Why ribosomal genes such as 16S/18S/ITS are good choices?

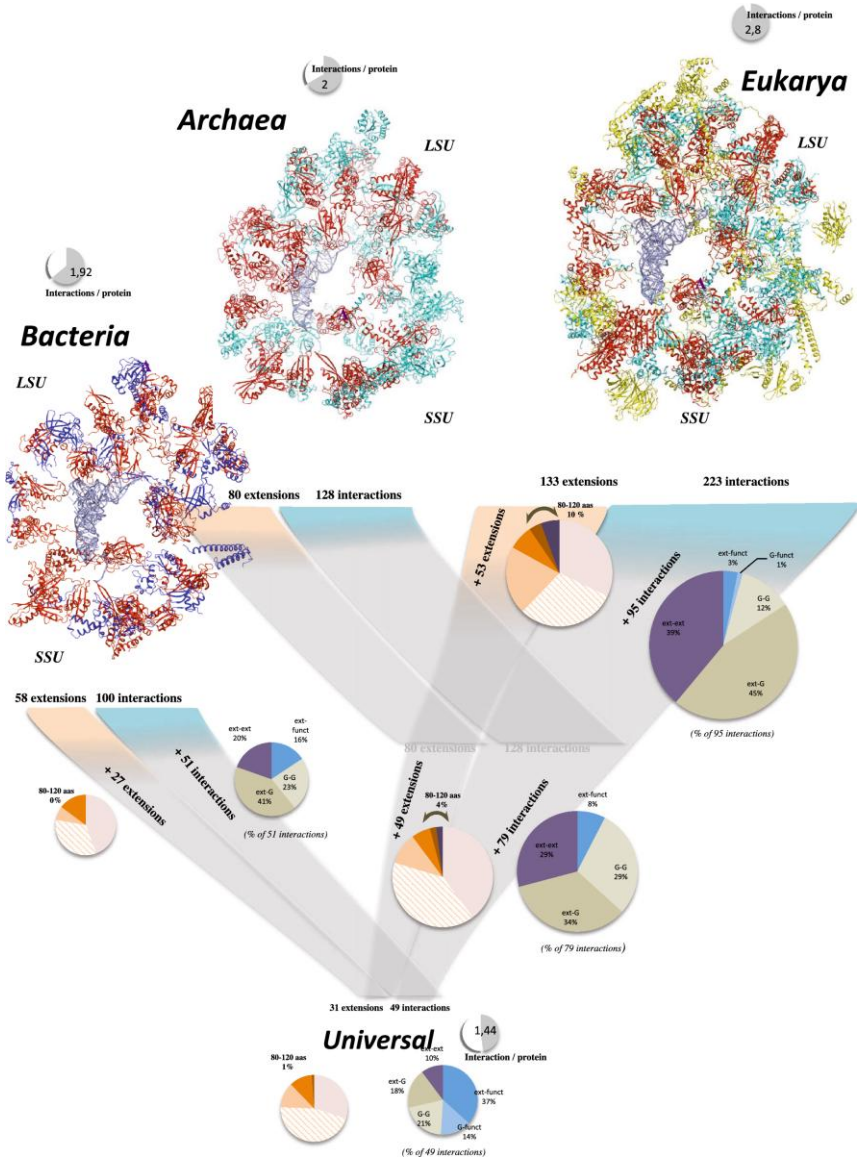
- They are found in all organisms (all?)
- Low mutation rate, but enough to allow comparative studies among all organisms



Organization of nuclear rRNA genes in plants

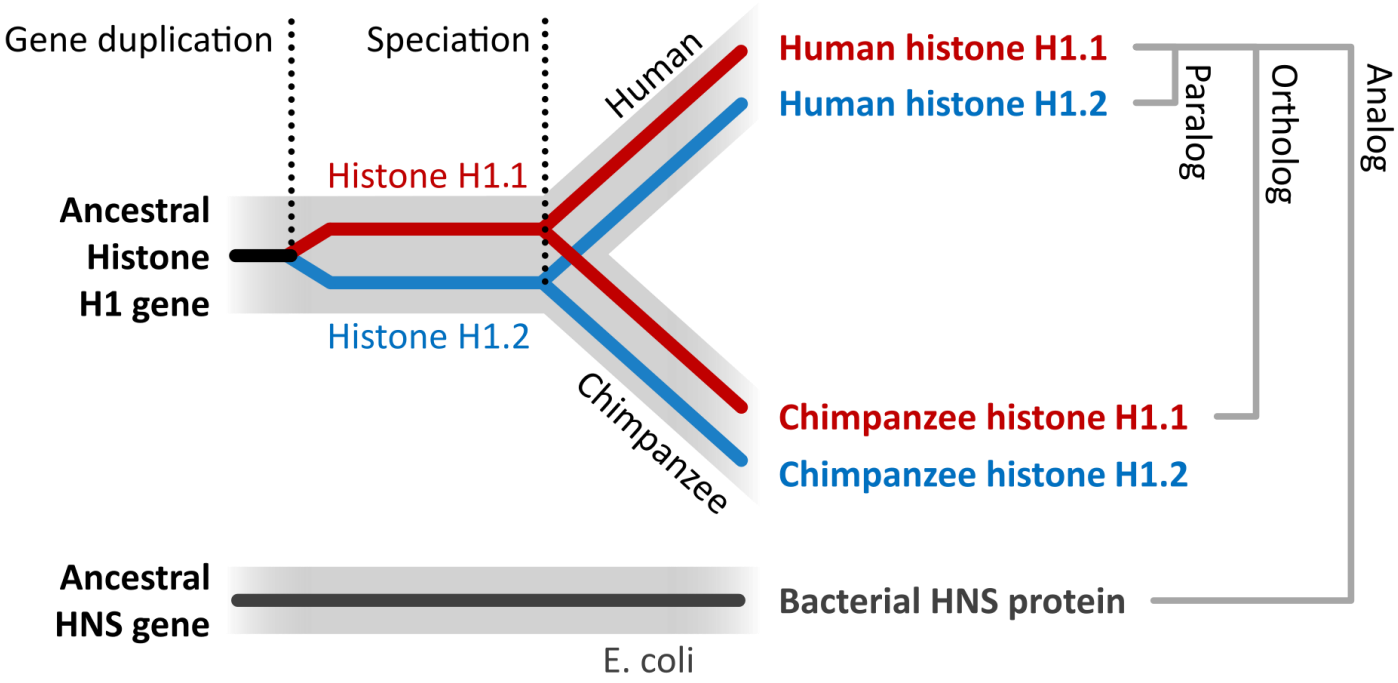


Organization of nuclear rRNA genes in bacteria and archaea



Evolution of the ribosomal protein interactomes

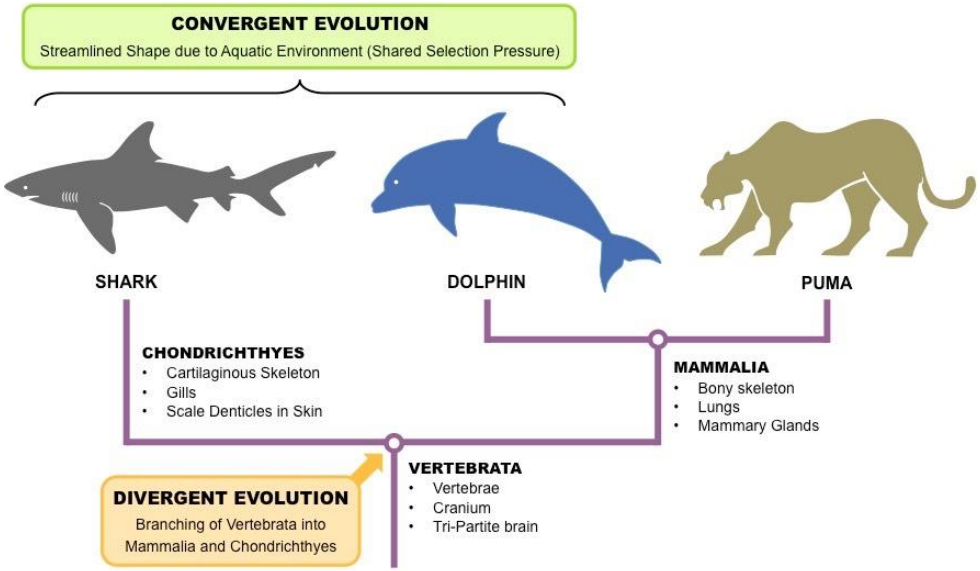
Sequence homology



Gene phylogeny as red and blue branches within grey species phylogeny. Top: An ancestral **gene duplication produces two paralogs** (histone H1.1 and 1.2). A **speciation event produces orthologs** in the two daughter species (human and chimpanzee). Bottom: in a separate species (E. coli), a gene has a similar function (histone-like nucleoid-structuring protein) but has a **separate evolutionary origin and so is an analog**.

Homology vs. analogy

It is important to compare states of the homologous characters



Alignments of multiple sequences are used to indicate which regions of each sequence are homologous.

Histone H1 (residues 120-180)

```
HUMAN KKASKPKKAASKAPTKKPKATPVKKAKKKLAATPKKAKKPKTVKAKPVKASKPKKAKPVK
CHIMP KKASKPKKAASKAPTKKPKATPVKKAKKKLAATPKKAKKPKTVKAKPVKASKPKKAKPVK
MOUSE KKAAKPKKAASKAPSKKPKATPVKKAKKKPAATPKKAKKPKVVKVKPVKASKPKKAKTVK
RAT   KKAAKPKKAASKAPSKKPKATPVKKAKKKPAATPKKAKKPKIVKVKPVKASKPKKAKPVK
COW   KKAAKPKKAASKAPSKKPKATPVKKAKKKPAATPKKTKKPKTVKAKPVKASKPKKTKPVK
      ***:**********:*************** ******:****  **.**********:*  **
```

Homology of sequence characters

- A character is homologous in two organisms, **if inherited by both from their common ancestor**

For sequence analyses:

- **There is no % of homology:** **A sequence is homologous or not.**
- Higher similarities between sequences, increase their likelihood of being homologous;
- However, 2 sequences can be homologous but showing low (or null) similarity (depends on the time of divergence between them)

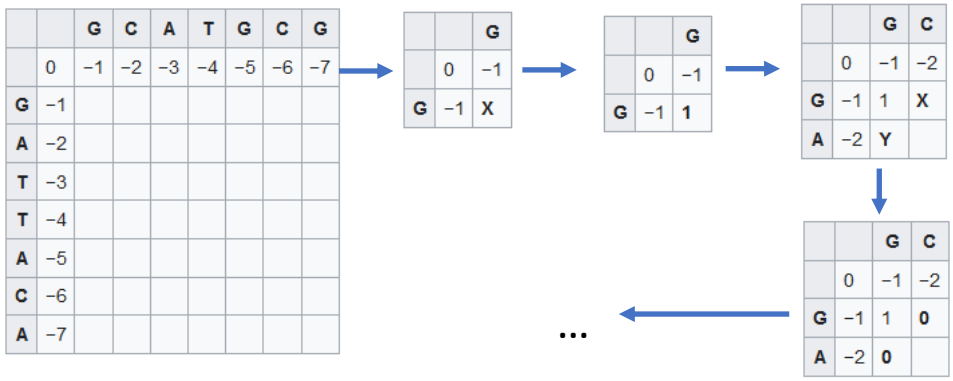
Phylogeny reconstruction using nucleotide or protein sequences

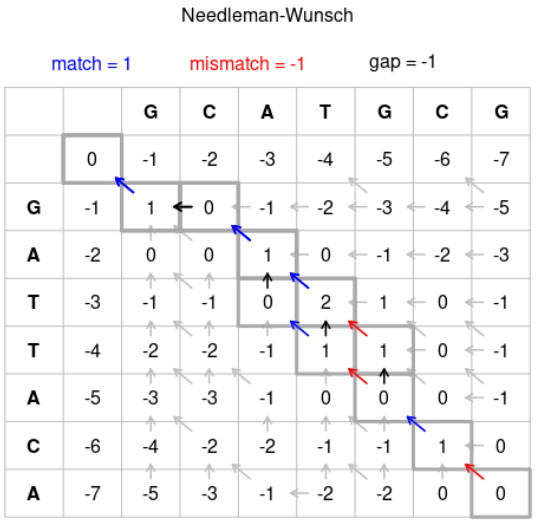
Computational Analysis

1) Compute distance matrix
   - Alignment based (**Multiple sequence alignment**)
   - Alignment-Free Methods (***k*-mer frequency, Composition vector**)

2) Generate phylogenetic tree based on this matrix
   - Minimum evolution
   - Neighbor joining
   - Maximum Parsimony
   - Maximum Likelihood
   - Bayesian inference
   - Using independent information

Compute distance matrix



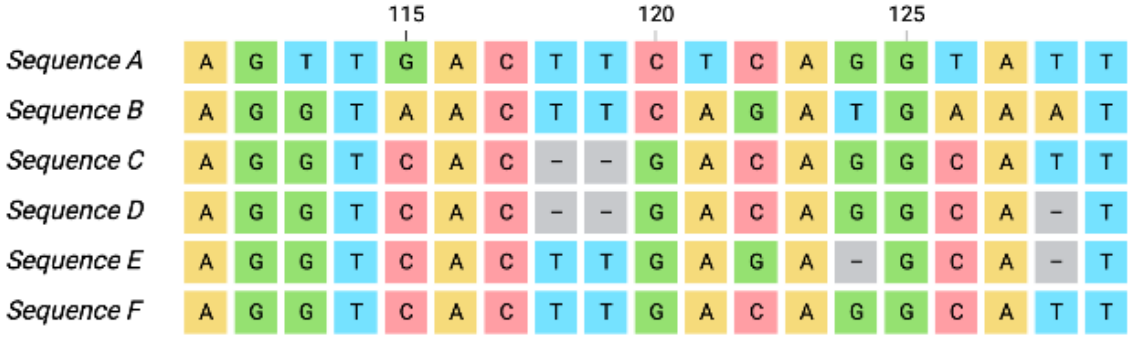


Needleman-Wunsch pairwise sequence alignment (1970)

```
G → CG → GCG → -GCG → T-GCG → AT-GCG → CAT-GCG → GCAT-GCG
A → CA → ACA → TACA → TTACA → ATTACA → -ATTACA → G-ATTACA
                ↓
    (branch) → TGCG → -TGCG → ...
             → TACA → TTACA → ...
```

Multiple sequence alignment

MAFFT
MUSCLE

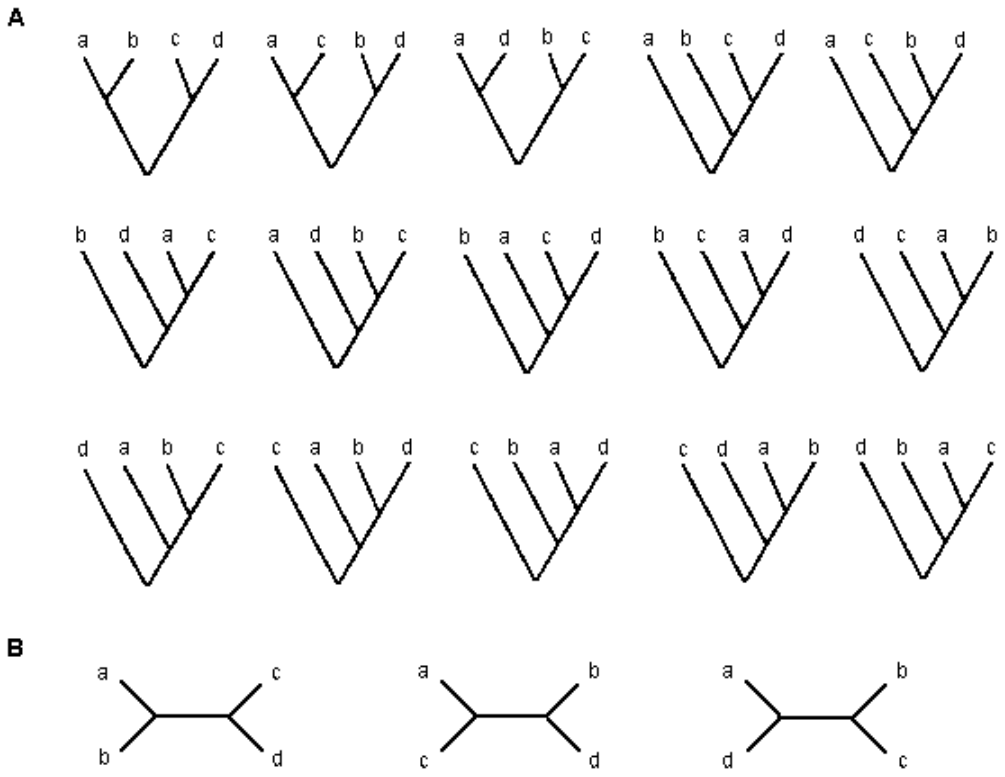


https://mafft.cbrc.jp/alignment/server/index.html

distance matrix

Phylogeny reconstruction

Reconstruction of phylogenetic trees is a statistical problem



Fifteen possible rooted trees (A) and three possible unrooted trees (B) for four species.

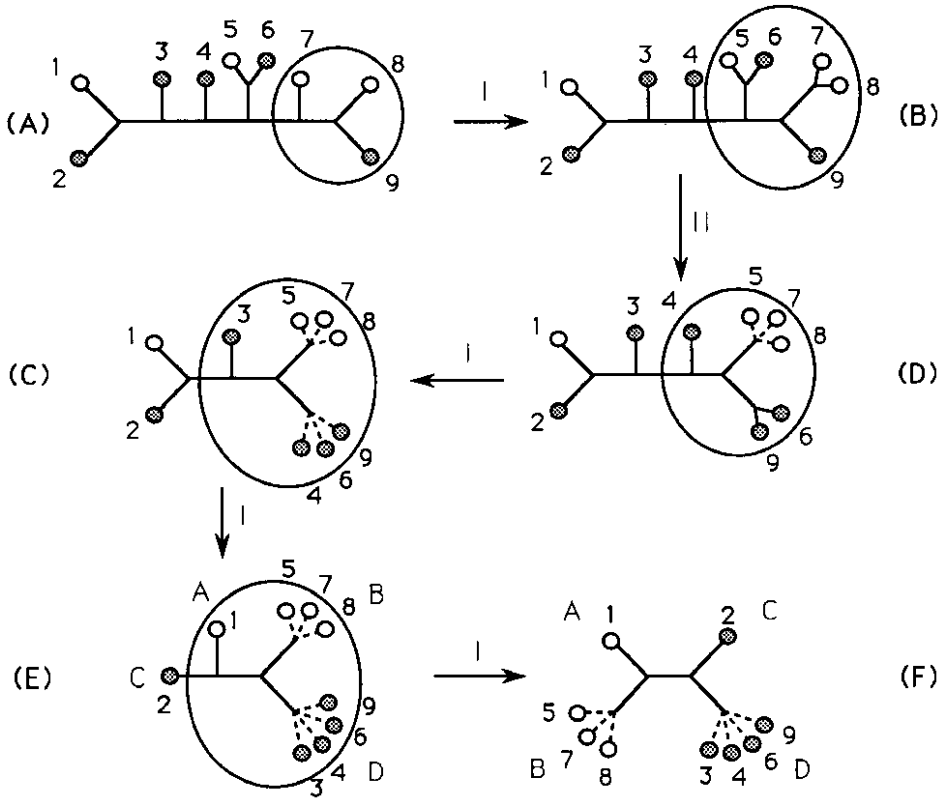
Phylogeny reconstruction

Methods that seek, among all possible trees, the one that best represents the evolutionary history of a given organism:

- **Minimum evolution**
  - Choice of topology that display the smaller branches - geometric methods
- **Neighbor-joining**
  - It allows to obtain the minimum evolution tree without building all the trees.
- **Maximum Parsimony**
  - Choice of the topology with fewer substitutions.
- **Maximum Likelihood**
  - Choice of the topology presenting the highest degree of fitness for a given substitution model.
- **Bayesian inference**
  - Similar to the Maximum likelihood, but instead of calculating the probability of the data under a given model, the probability of the model is calculated from *a priori* information.

Minimum evolution

Minimum evolution is searching for the phylogeny that has the **shortest total sum of branch lengths** – all topologies need to be generated



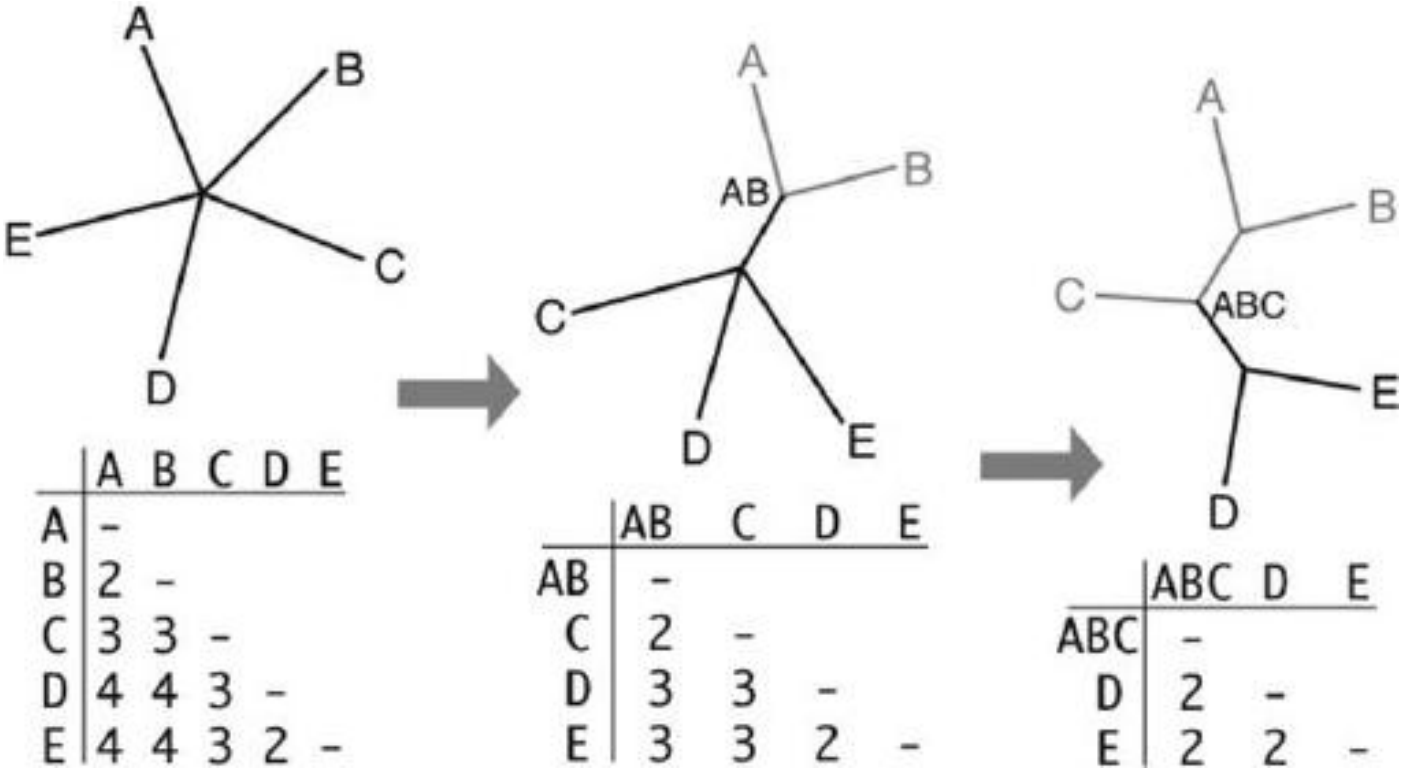
LIMITATION

The number of topologies exponentially increases with the number of OTUs - Very computer consuming

Number (*n*) of OTUs (taxa)	Number of bifurcating unrooted tree topologies (N_u)
4	3
5	15
10	2,027,025
20	2.2164×10^{20}
50	2.8381×10^{74}
60	5.01×10^{94}

Neighbor Joining (NJ)

bottom-up (agglomerative) clustering method for the creation of phylogenetic trees, created by Naruya Saitou and Masatoshi Nei (1987)



	A	B	C	D	E
A	–				
B	2	–			
C	3	3	–		
D	4	4	3	–	
E	4	4	3	2	–

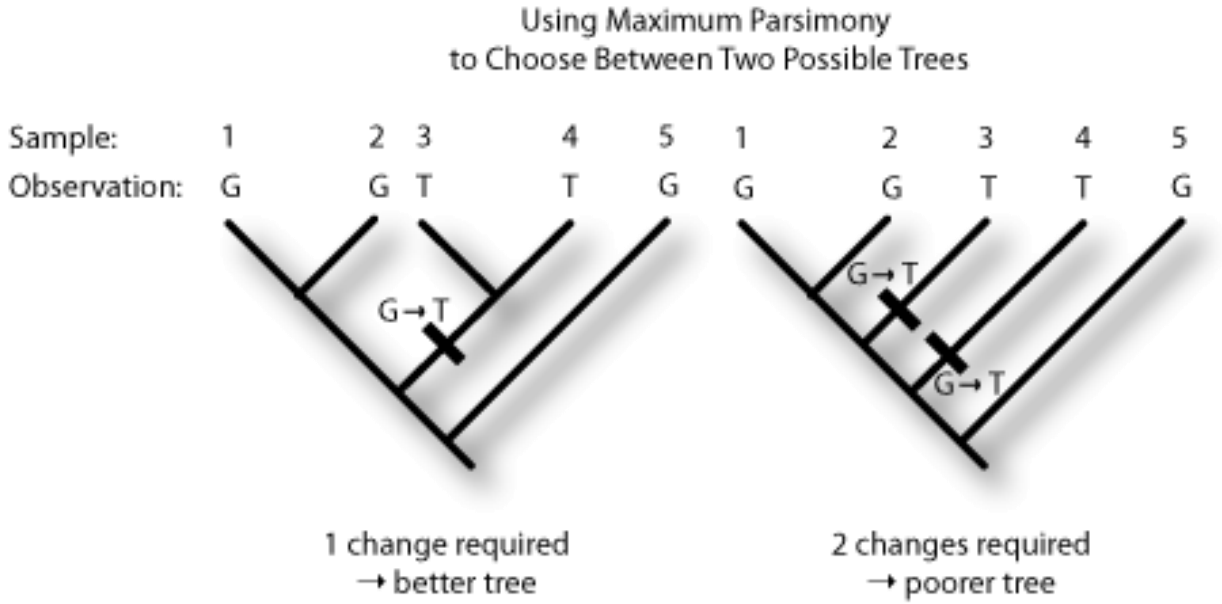
	AB	C	D	E
AB	–			
C	2	–		
D	3	3	–	
E	3	3	2	–

	ABC	D	E
ABC	–		
D	2	–	
E	2	2	–

It allows to obtain the minimum evolution tree without building all the trees

Maximum parsimony

Maximum parsimony is used to build an evolutionary tree by choosing the "simplest" one



The "starting point", or ancestral state, is chosen so as to obtain the most parsimonious tree.

Parsimony is used both for choosing the topology and for choosing the ancestral states for a given topology:

1) The parsimony score for a given topology is computed for the set of ancestral states minimizing the number of transformations.

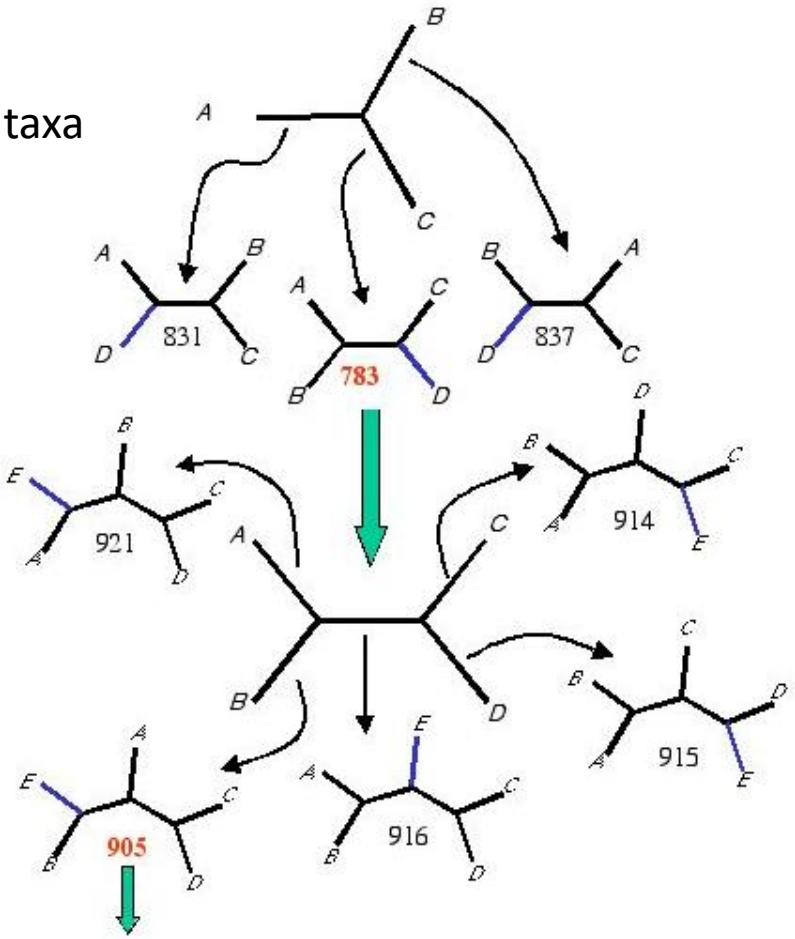
2) Then, the topology minimizing this minimized number of transformations is chosen as the most parsimonious tree.

Maximum Likelihood

determine the tree topology, branch lengths, and parameters of the evolutionary model **that maximize the probability of observing the sequences at hand** (or minimizes the cost of differentially weighted character-state changes)

Stepwise addition method

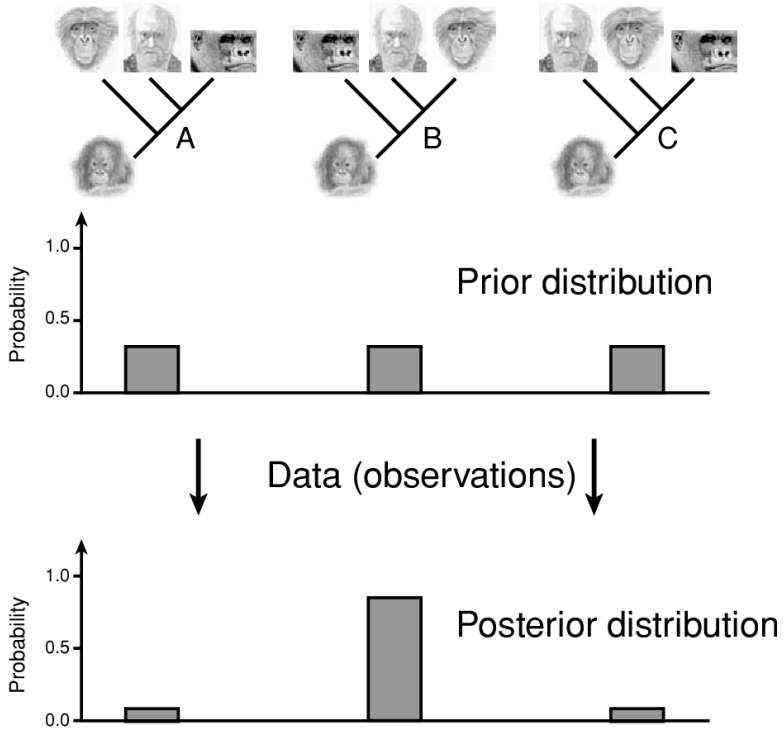
- select three random taxa from n terminal taxa
- find the most likely tree
- add another taxon
- find the most likely tree
- repeat n-3 times



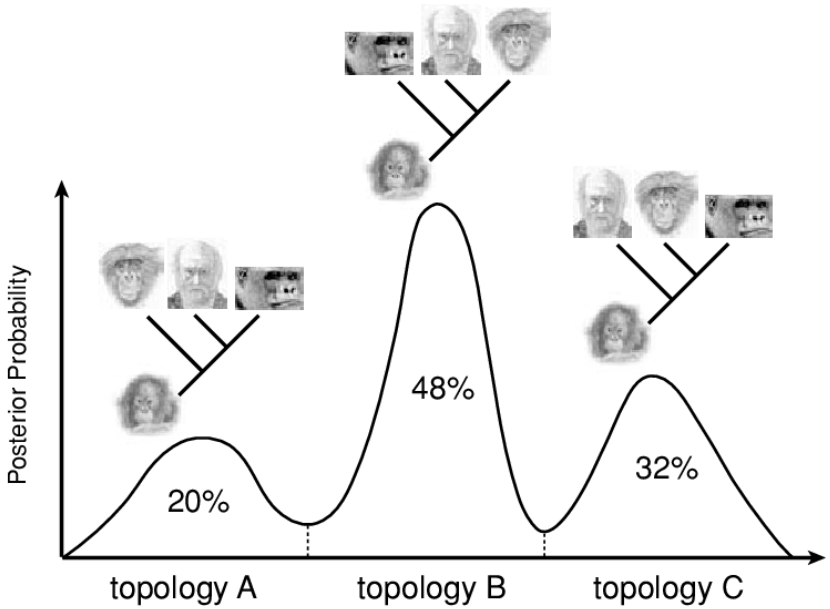
- will find a locally optimal tree
- other addition orders may give a more optimal tree
- perform tree rearrangements to search other optimal trees

Bayesian inference

Similar to the Maximum likelihood, but instead of calculating the probability of the data under a given model, the probability of the model is calculated from a priori information.



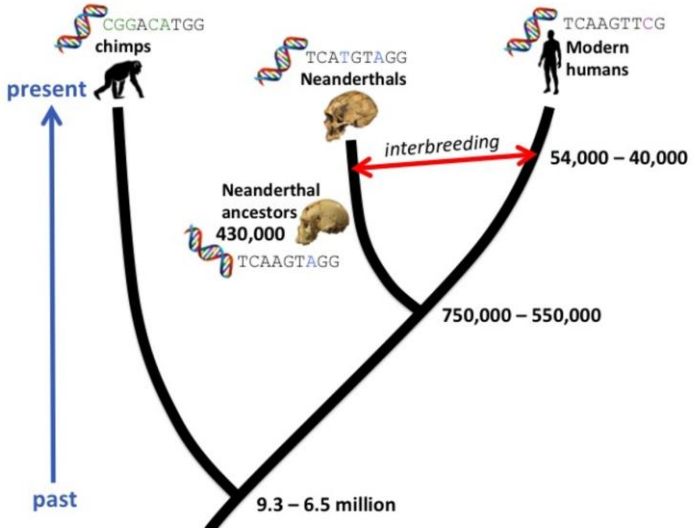
We start the analysis by specifying our prior beliefs about the tree. In the absence of background knowledge, we might associate the same probability to each tree topology. We then collect data and use a **stochastic evolutionary model and Bayes' theorem** to update the prior to a posterior probability distribution. **If the data are informative, most of the posterior probability will be focused on one tree** (or a small subset of trees in a large tree space).



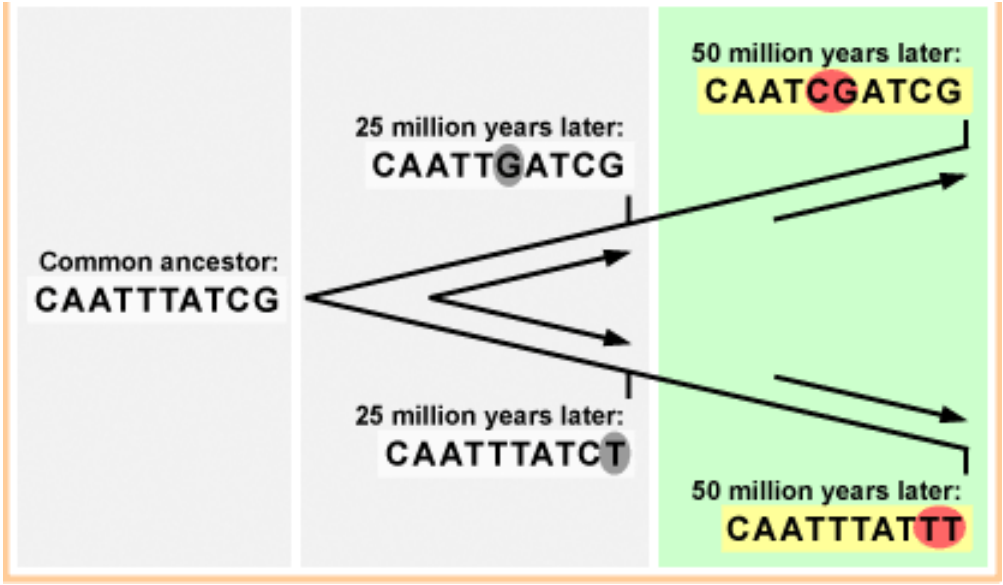
Posterior probability distribution for our phylogenetic analysis. The x-axis is an imaginary one-dimensional representation of the parameter space. It falls into three different regions corresponding to the three different topologies. Within each region, a point along the axis corresponds to a particular set of branch lengths on that topology. It is difficult to arrange the space such that optimal branch length combinations for different topologies are close to each other. Therefore, the posterior distribution is multimodal. The area under the curve falling in each tree topology region is the posterior probability of that tree topology

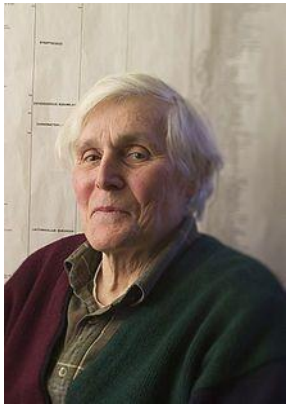
Molecular clock

a figurative term for a technique that uses the **mutation rate of biomolecules to deduce the time** in prehistory when two or more life forms diverged



2 different species acumulate mutations as they diverge, in a regular rate, becoming most different along the time...





Proc. Natl. Acad. Sci. USA
Vol. 74, No. 11, pp. 5088–5090, November 1977
Evolution

Phylogenetic structure of the prokaryotic domain: The primary kingdoms

(archaebacteria/eubacteria/urkaryote/16S ribosomal RNA/molecular phylogeny)

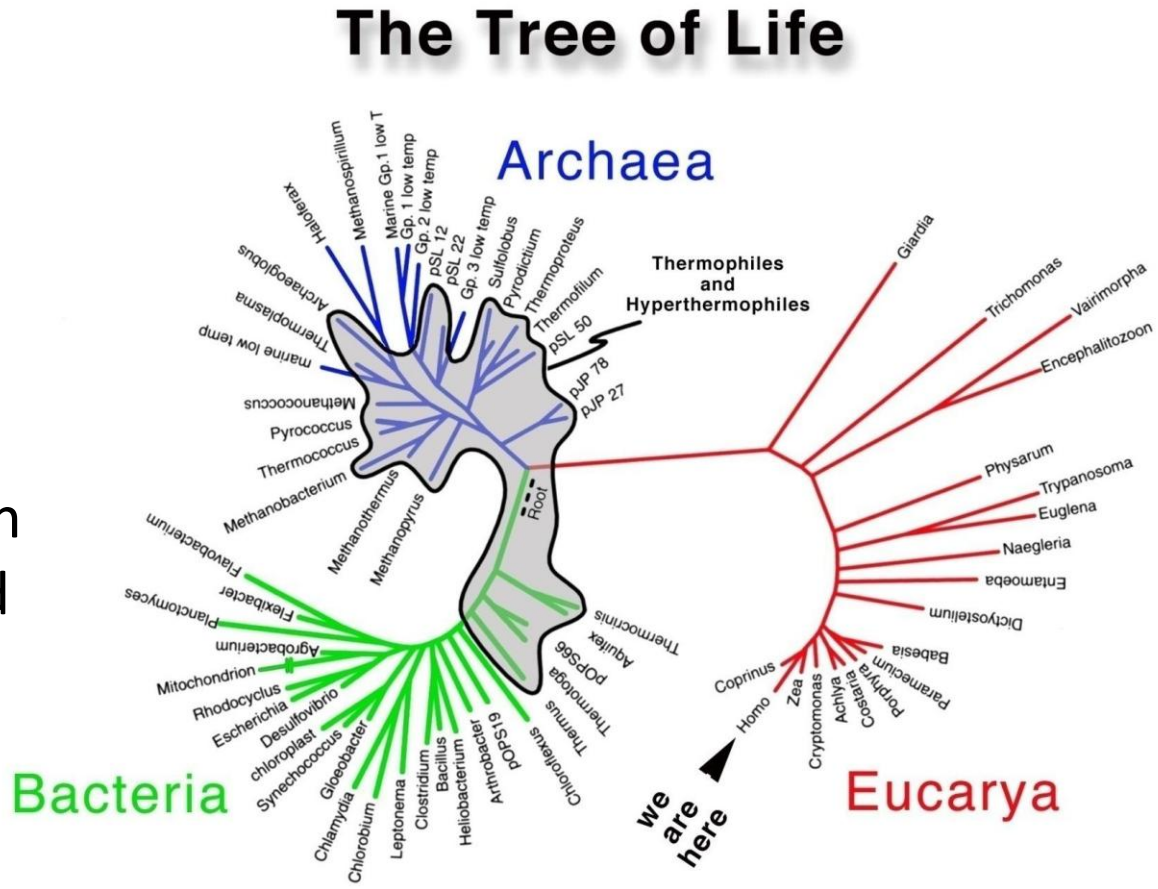
CARL R. WOESE AND GEORGE E. FOX*

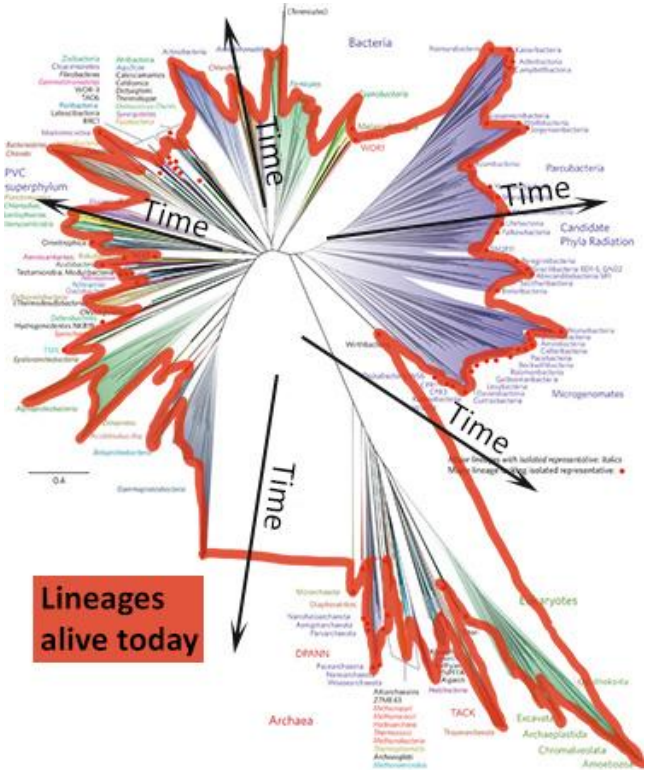
Department of Genetics and Development, University of Illinois, Urbana, Illinois 61801

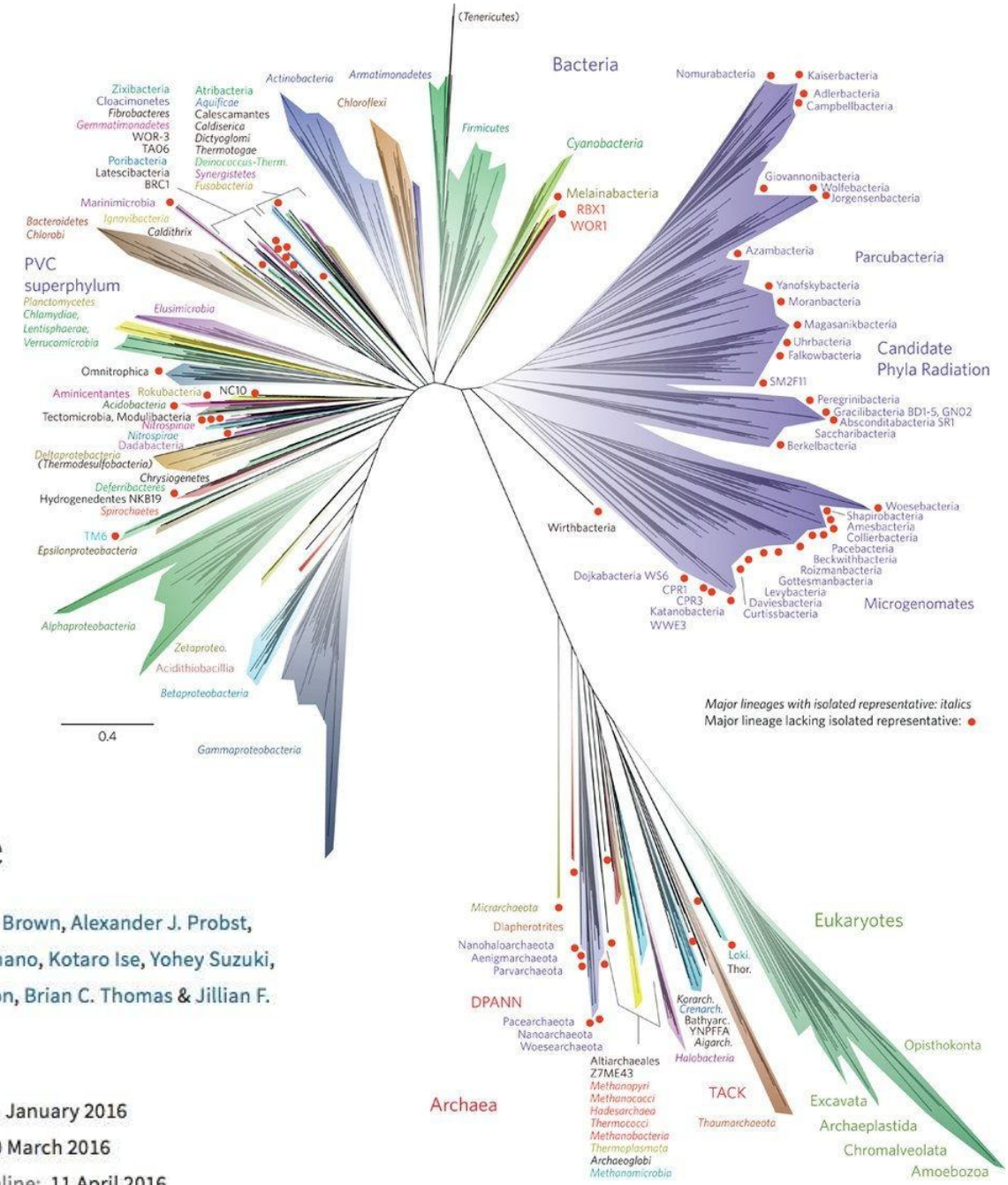
Communicated by T. M. Sonneborn, August 18, 1977

Carl Woese
(1928-2012)

- defined the new realm or domain of Archaea
- results were based on rRNA sequencing and alignment







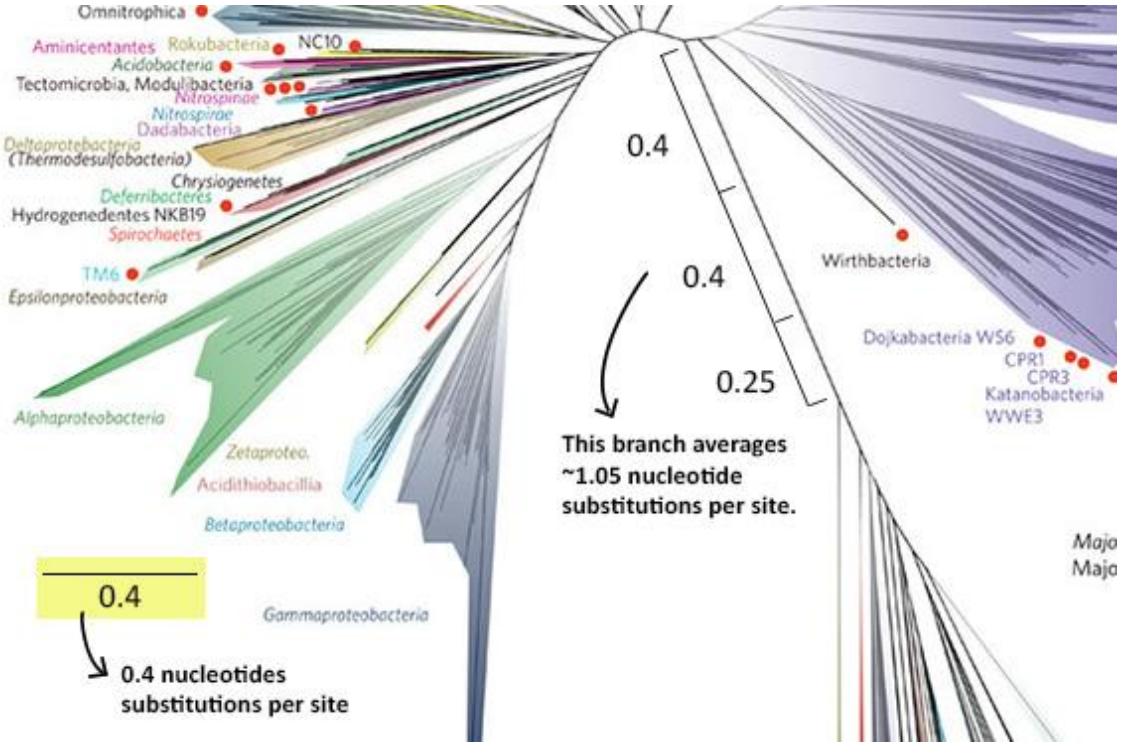
Letter | OPEN

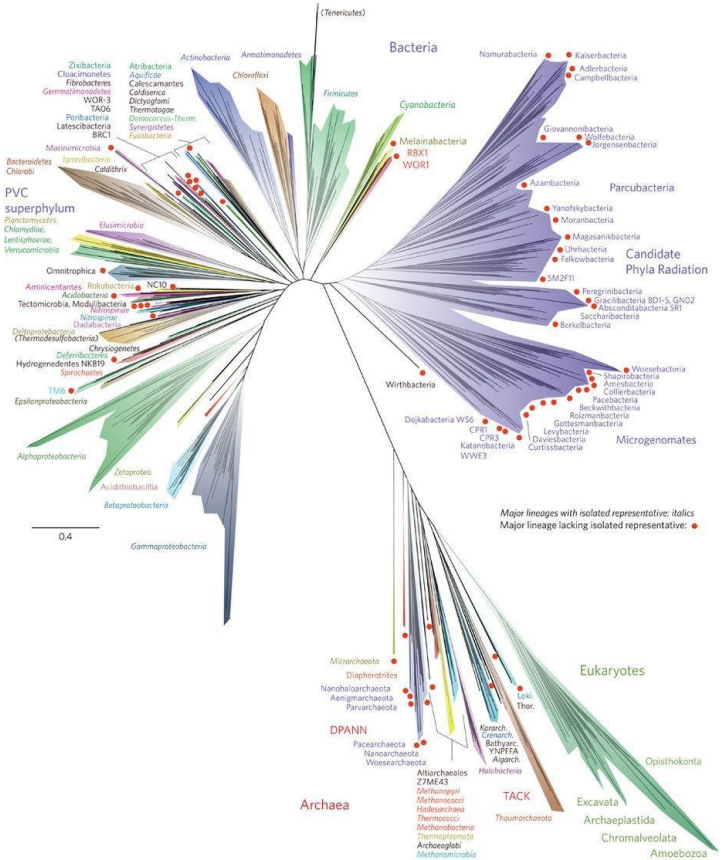
A new view of the tree of life

Laura A. Hug, Brett J. Baker, Karthik Anantharaman, Christopher T. Brown, Alexander J. Probst, Cindy J. Castelle, Cristina N. Butterfield, Alex W. Hernsdorf, Yuki Amano, Kotaro Ise, Yohey Suzuki, Natasha Dudek, David A. Relman, Kari M. Finstad, Ronald Amundson, Brian C. Thomas & Jillian F. Banfield

Nature Microbiology **1**, Article number: 16048 (2016)
doi:10.1038/nmicrobiol.2016.48

Received: 25 January 2016
Accepted: 10 March 2016
Published online: 11 April 2016





Letter | OPEN

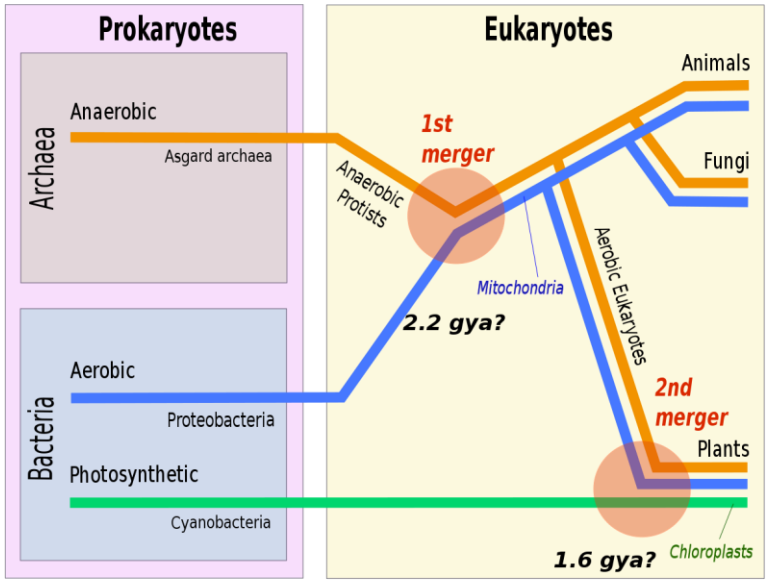
A new view of the tree of life

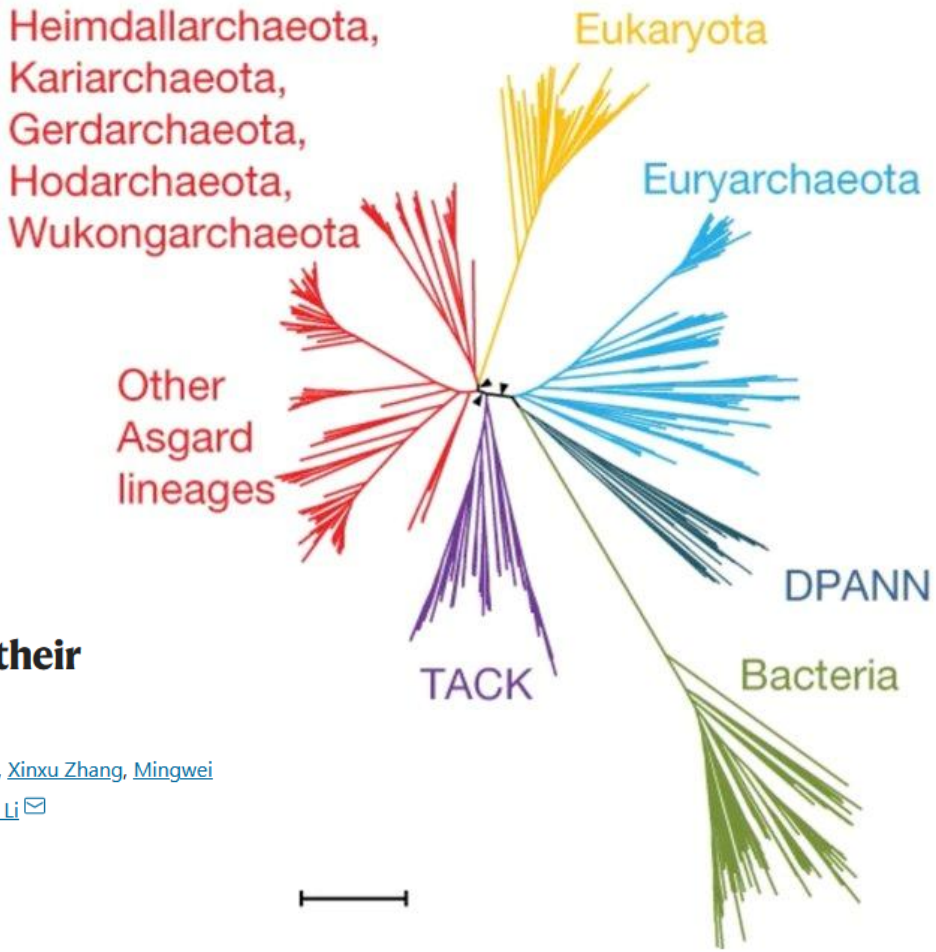
Laura A. Hug, Brett J. Baker, Karthik Anantharaman, Christopher T. Brown, Alexander J. Probst, Cindy J. Castelle, Cristina N. Butterfield, Alex W. Hernsdorf, Yuki Amano, Kotaro Ise, Yohey Suzuki, Natasha Dudek, David A. Relman, Kari M. Finstad, Ronald Amundson, Brian C. Thomas & Jillian F. Banfield

Nature Microbiology **1**, Article number: 16048 (2016)
doi:10.1038/nmicrobiol.2016.48

Received: 25 January 2016
Accepted: 10 March 2016
Published online: 11 April 2016

Asgard is a recently discovered superphylum of archaea that appears to contain the closest archaeal relatives of eukaryotes.





Article | Published: 28 April 2021

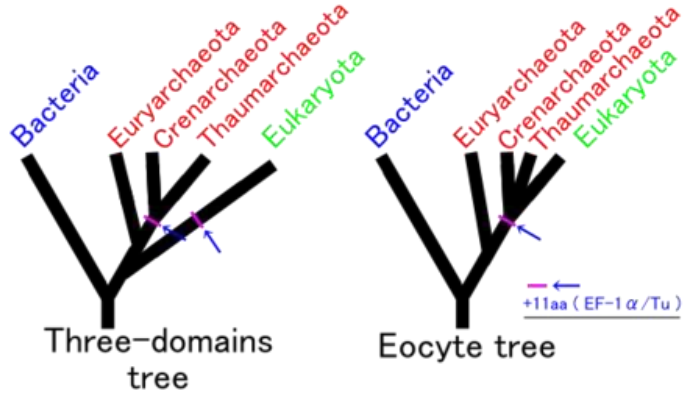
Expanded diversity of Asgard archaea and their relationships with eukaryotes

Yang Liu, Kira S. Makarova, Wen-Cong Huang, Yuri I. Wolf, Anastasia N. Nikolskaya, Xinxu Zhang, Mingwei Cai, Cui-Jing Zhang, Wei Xu, Zhuhua Luo, Lei Cheng, Eugene V. Koonin & Meng Li

Nature **593**, 553–557 (2021) | Cite this article

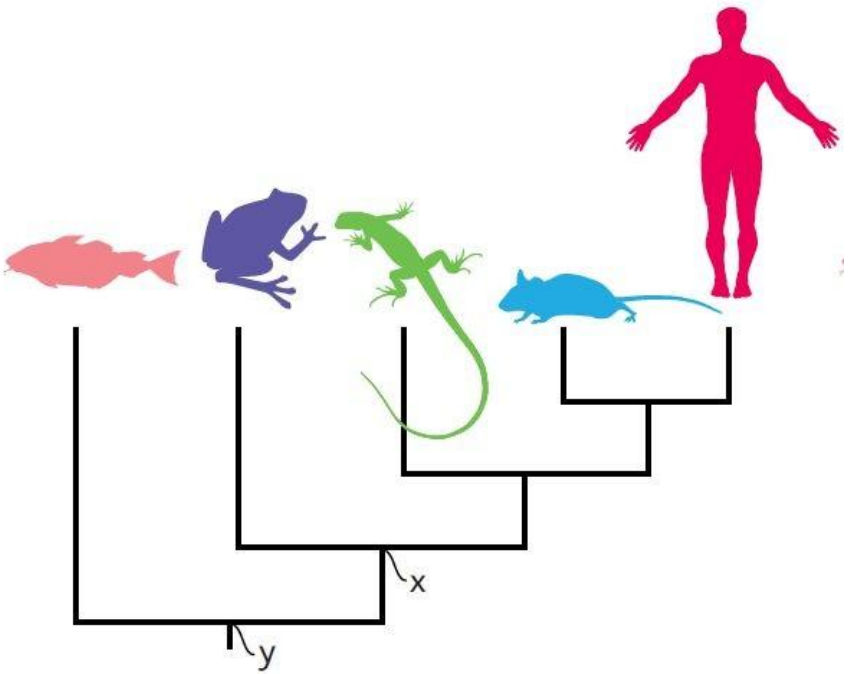
Possible flaws...

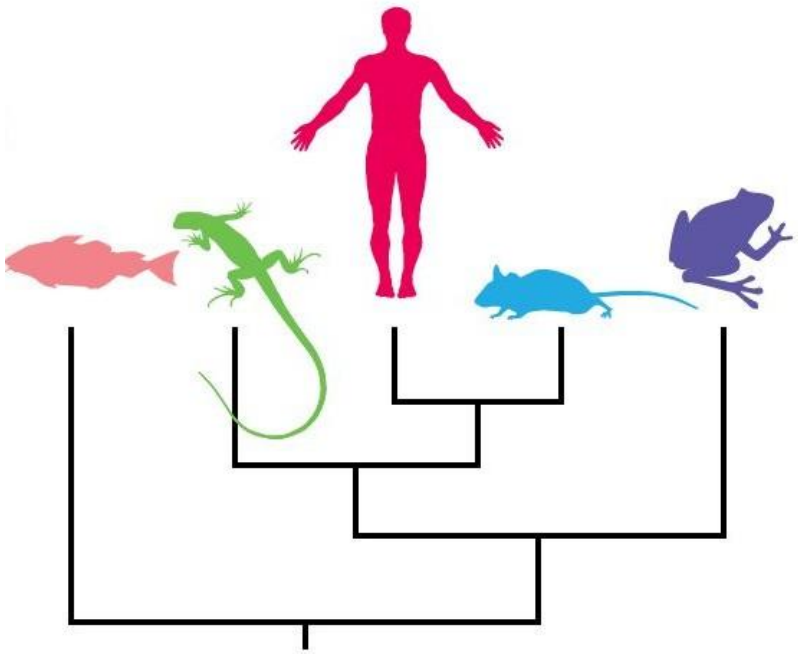
- **Heterotachy**:
  - The nucleotide substitution rate in a gene may change during time.
    - You should not use concatenated sequences.



- Unbalanced sequence distribution
  - Bacteria (too many)
  - Archaea / Eukaria (too few)
    - The "*outgroup*" effect – long branch attraction ("*LBA*")
      - random convergence or parallel evolution
    - To balance the number of sequences in all nodes.
- Alignment problems

Is “Frog” more closely related to “Fish” or “Humans”?





END