How to read and make phylogenetic trees

Zuzana Starostová

How to make phylogenetic trees?

Workflow:

- obtain DNA sequence
- quality check
- sequence alignment
- calculating genetic distances (optional)
- phylogeny estimation – topology and branch length
- reliability test (bootstrap)
- tree visualization

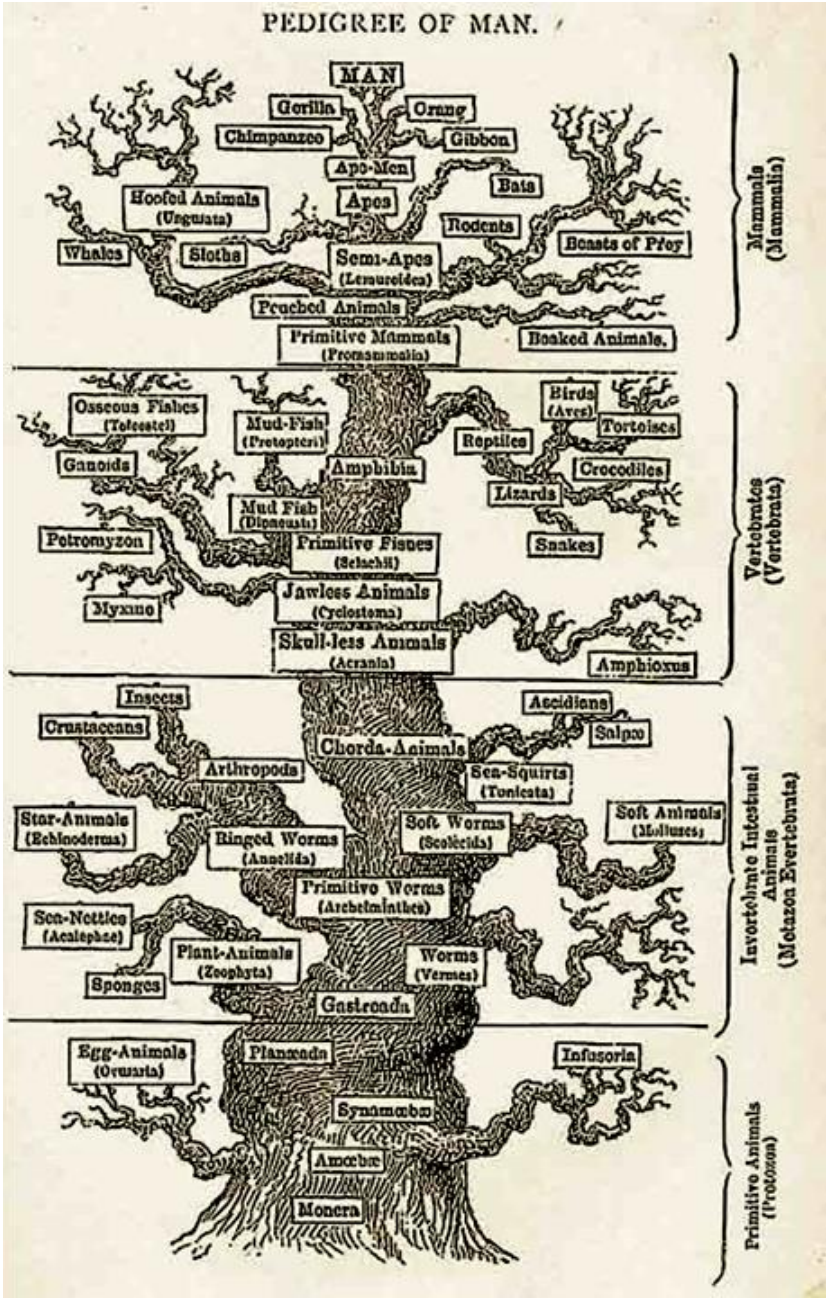
- the process of evolution produces a pattern of relationships between species - as lineages evolve and split and modifications are inherited, their evolutionary paths diverge

- this produces a branching pattern of evolutionary relationships – **phylogenetic tree**

- **phylogeny** – the evolutionary history of a species or group of related species

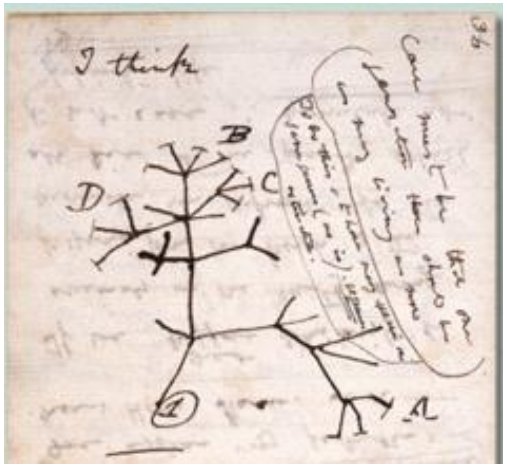
- phylogenies trace patterns of shared ancestry between lineages

- each lineage has a part of its history that is unique to it alone and parts that are shared with other lineages

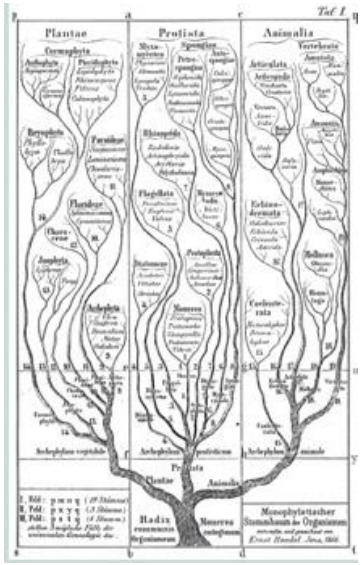


Tree of life

Charles Darwin
species share common ancestor
relationships among species compared to “the great Tree of Life“

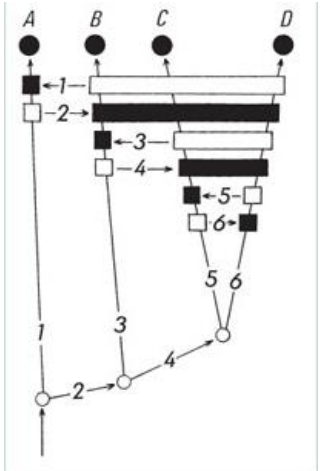


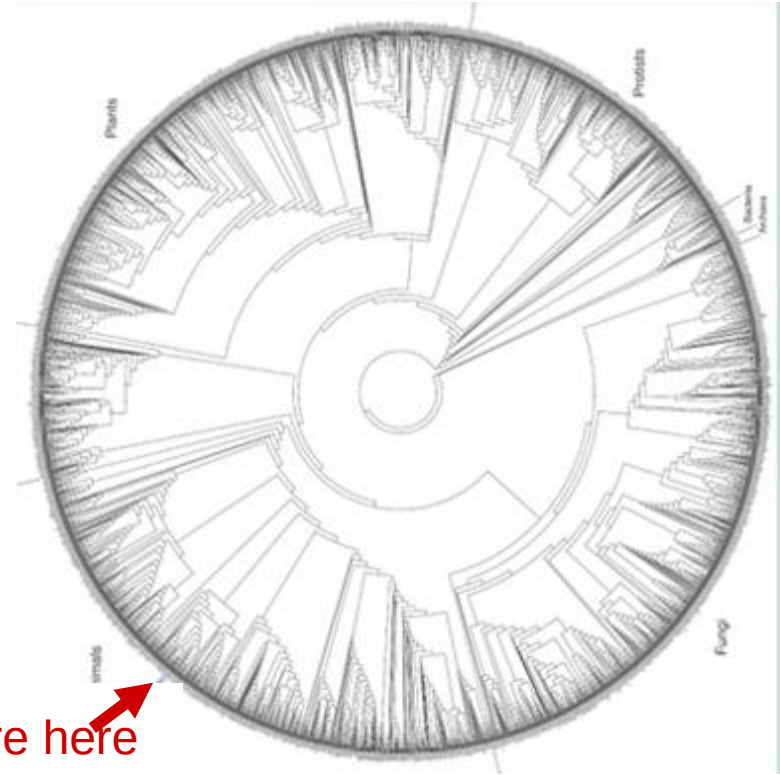
Darwin‘s diary, 1837



Ernst Haeckel – 1866
diagram of relationships among species based on general similarity

Willi Hennig – 1960s
foundation of modern phylogenetics





you are here

http://www.peabody.yale.edu/exhibits/treeoflife/index.html

Why is phylogeny important?

- understanding and classifying the diversity of life on Earth

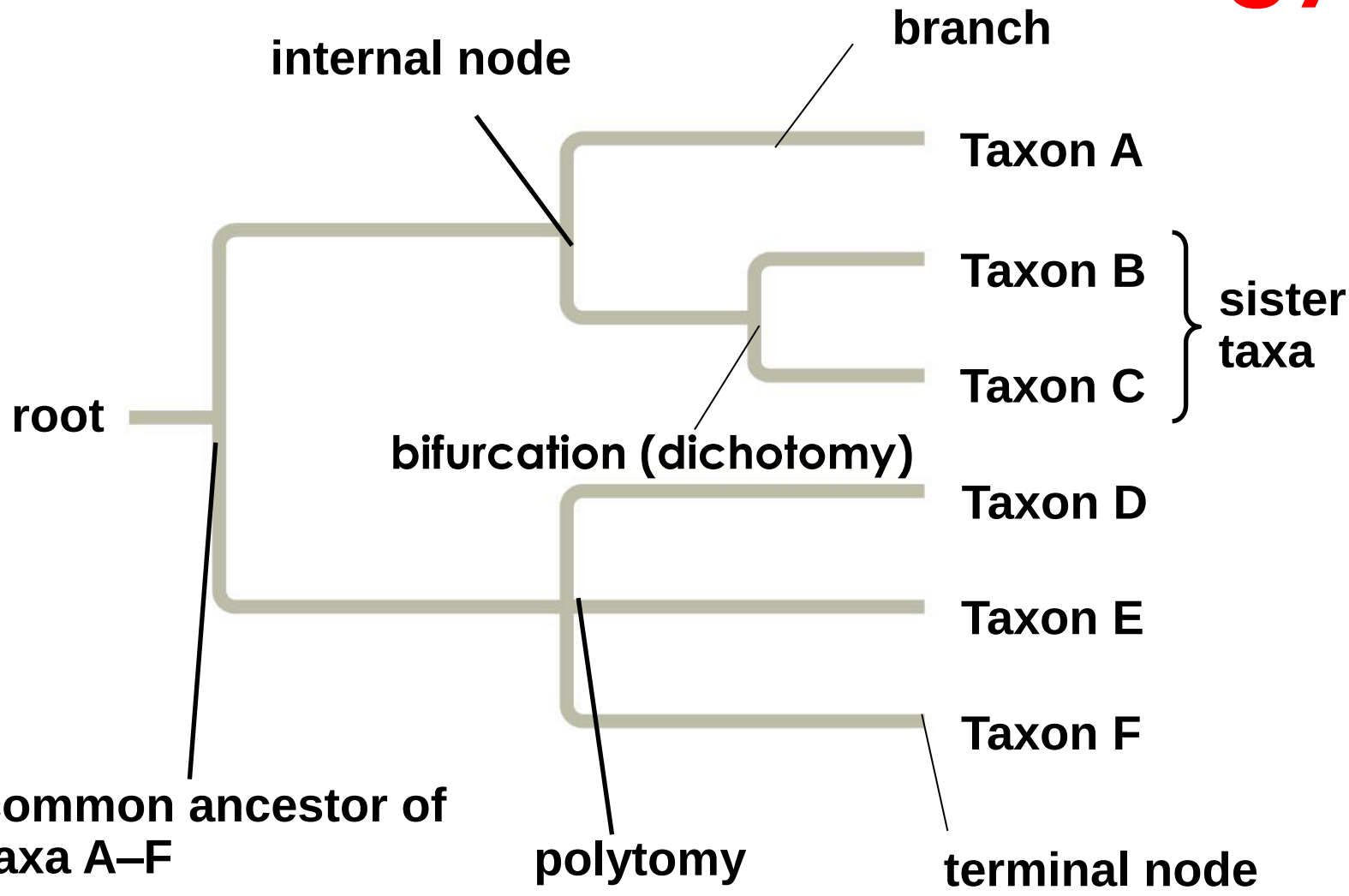
We use phylogenetic trees for:

- comparative analysis and character evolution
- biogeography
- dating – age of different taxa
- genetic ingeneering
- disease epidemiology
- conservation

…

The Tree of Life Web Project (ToL) is a collaborative effort of biologists and nature enthusiasts from around the world. On more than 10,000 World Wide Web pages, the project provides information about biodiversity, the characteristics of different groups of organisms, and their evolutionary history.
http://tolweb.org/tree/

how to read trees - terminology

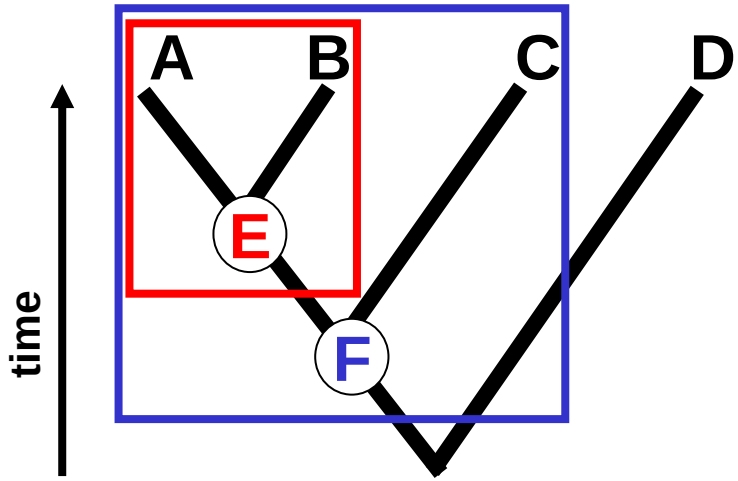


- root of the tree represents the ancestral lineage, and the tips of the branches represent the descendants of that ancestor
- as one moves from the root to the tips = moving forward in time

phylogenetic tree

branching diagram showing relationships between taxa based on their shared common ancestors

taxa (e.g.: species):



A and B are more closely related because they share a common ancestor (here "E") that C and D do not share

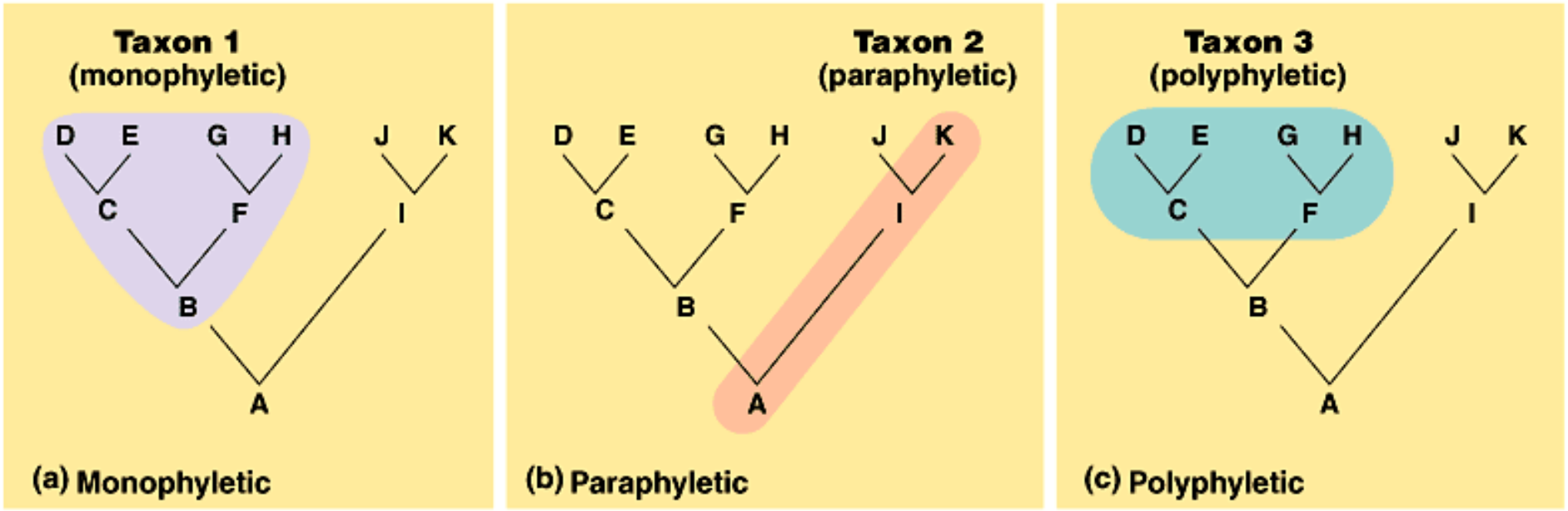
A+B+C are more closely related to each other than to D because they share a common ancestor ("F") that D does not share

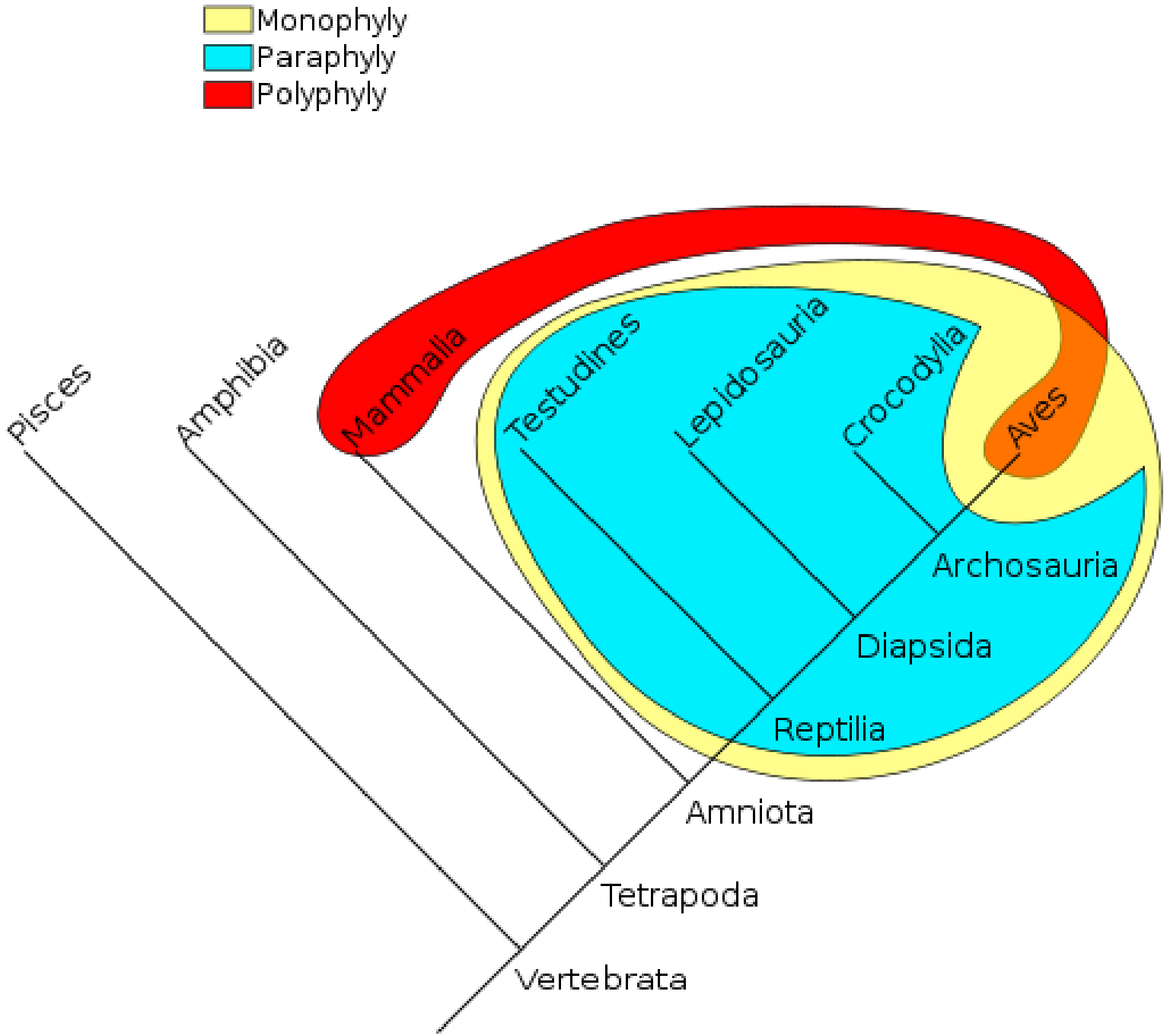
Phylogeny and classification

phylogenetic (cladistic) classification reflects evolutionary history

the only valid group for classification is **monophyletic**

group = clade = group that includes a common ancestor and all the descendants (living and extinct) of that ancestor.



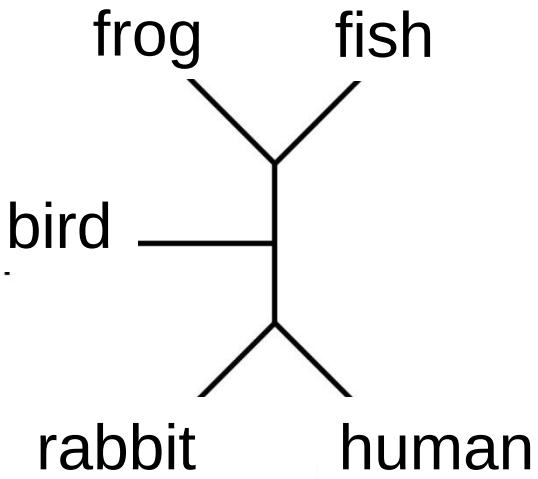

Monophyly
Paraphyly
Polyphyly
Pisces
Amphibia
Mammalia
Testudines
Lepidosauria
Crocodylia
Aves
Archosauria
Diapsida
Reptilia
Amniota
Tetrapoda
Vertebrata

rooted vs. unrooted phylogenetic trees

rooted tree – root leads to the common ancestor of all studied taxa (e.g.: species)
rooting the tree = indicating the direction of the evolutionary process
helps to determine what is plesiomorphic and apomorphic

How to root a tree?
- introducing outgroup

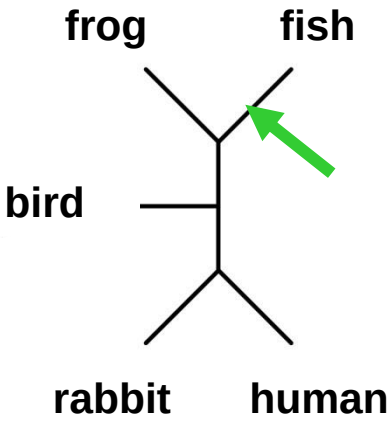
outgroup – a species or group of species that is clearly distant from all the species of interest (ingroup), but still closely related



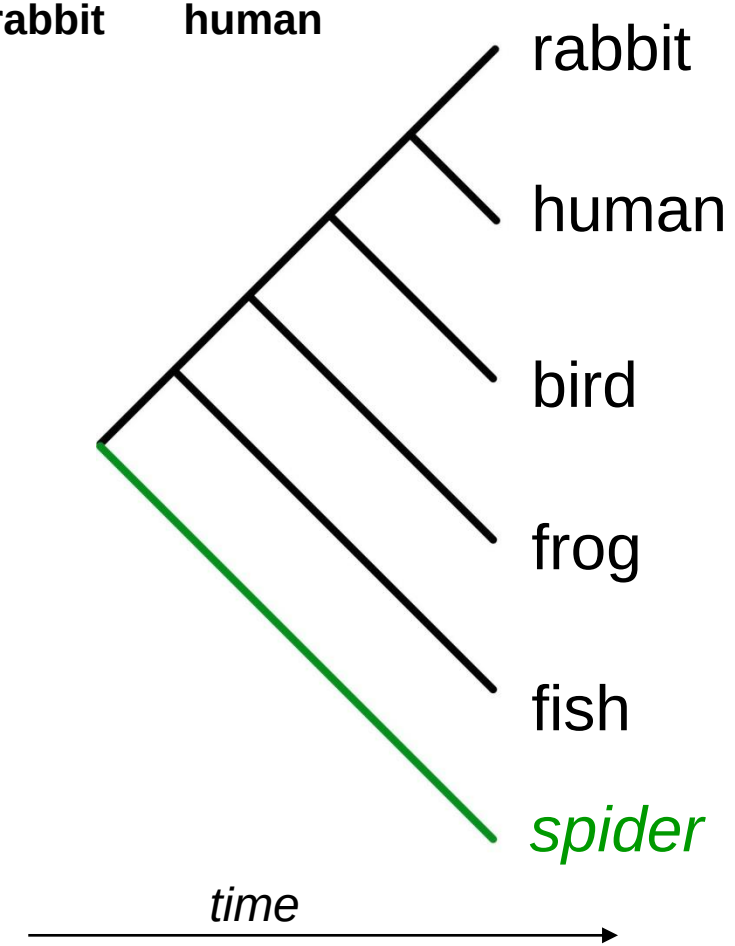
here we will use an invertebrate species to root the tree

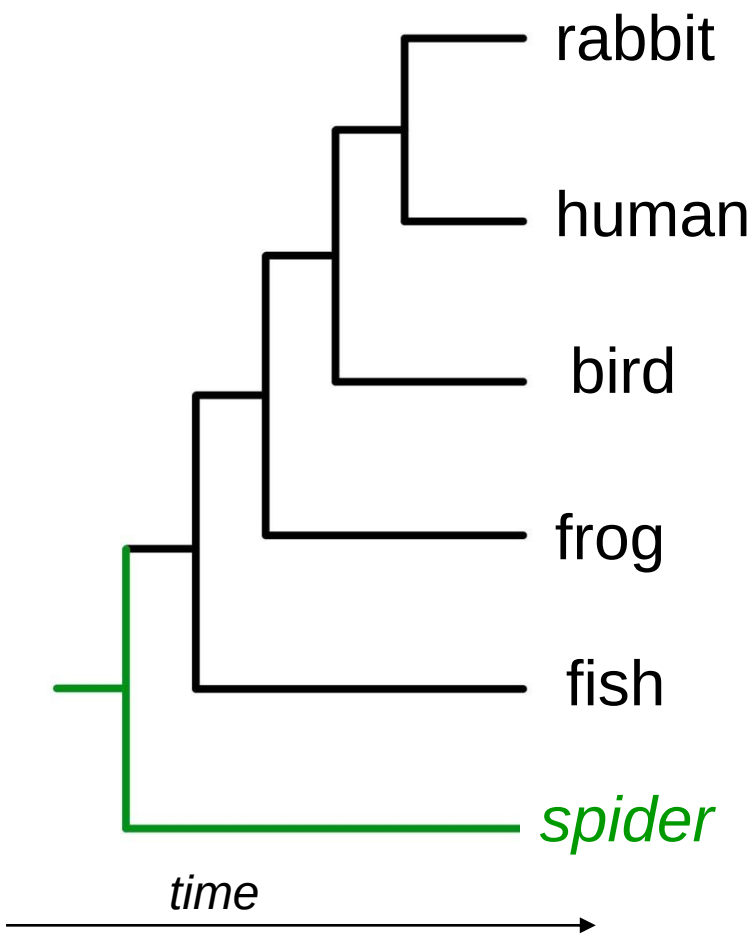
(since these species are all Craniata it would be better to use Urochordata (tunicates) to root the tree)

unrooted tree

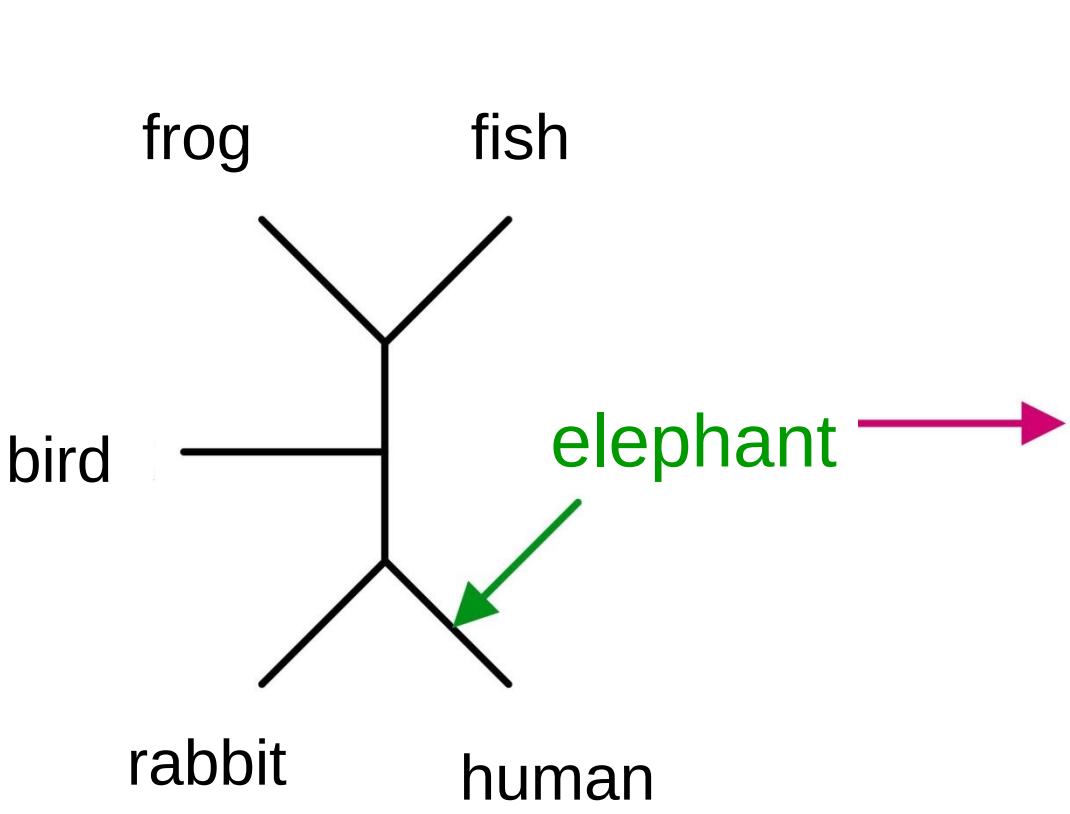


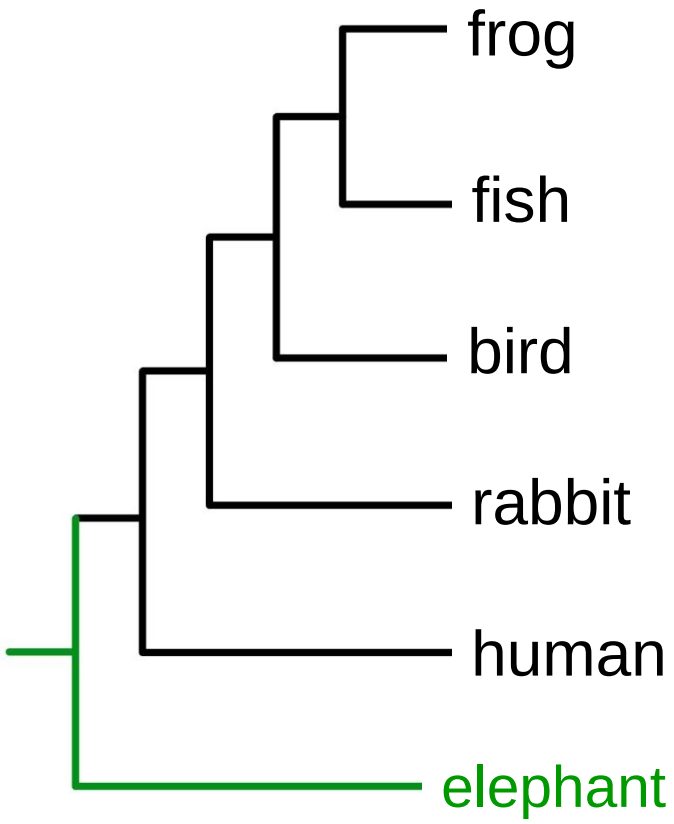
we place the root to the middle of the branch connecting outgroup with the rest of the tree



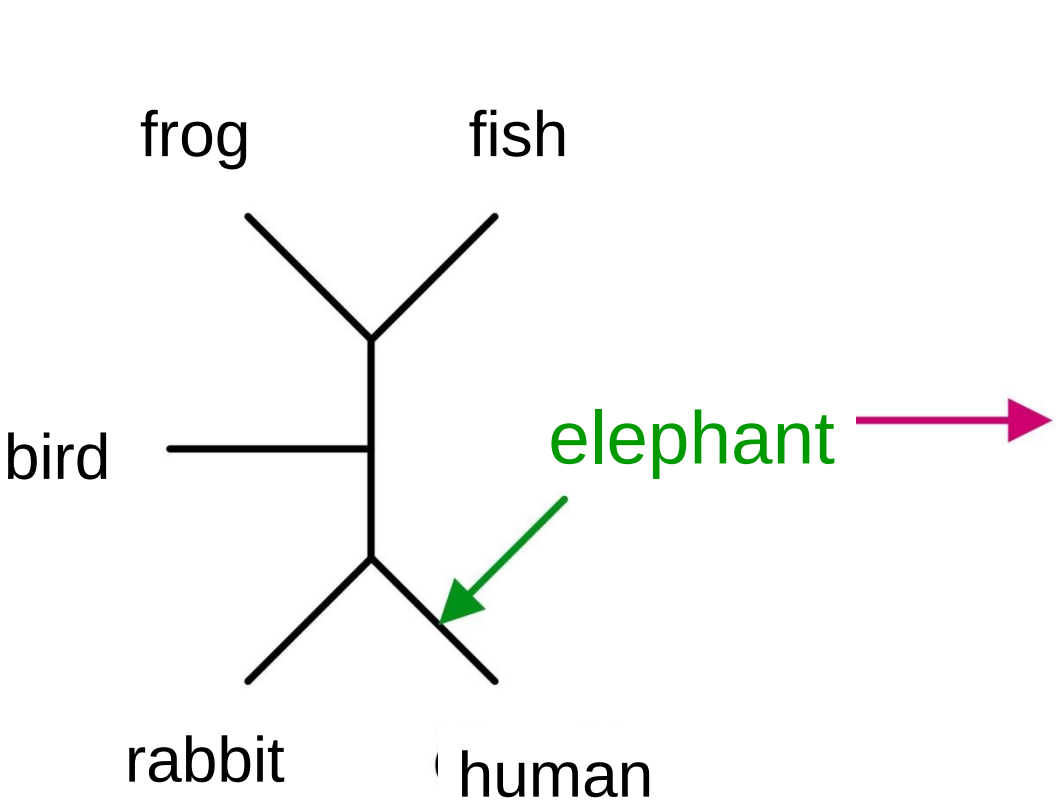


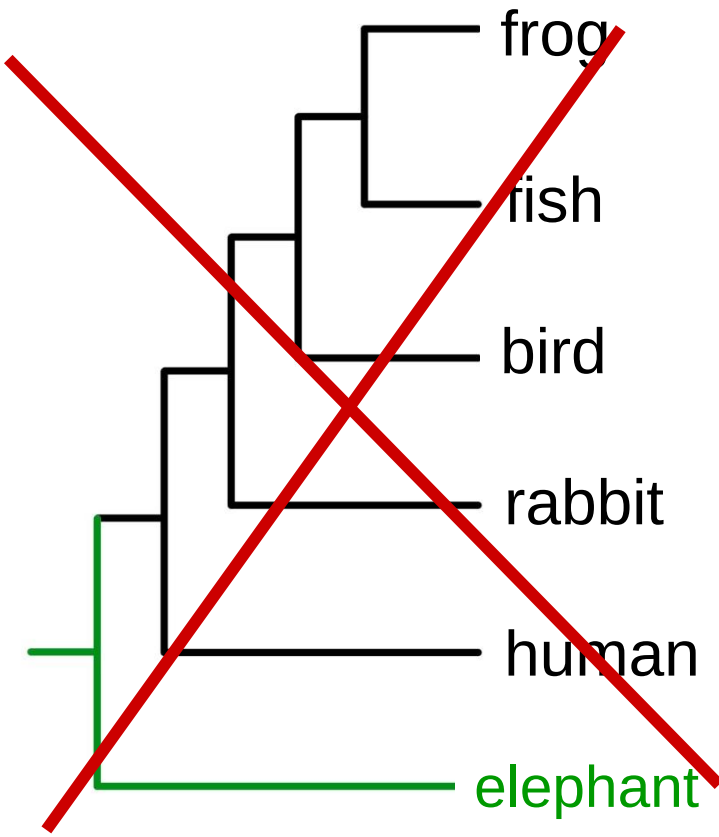
But what if our outgroup is wrong?





But what if our outgroup is wrong?





rooted vs. unrooted phylogenetic trees

$$\frac{(2n-3)!}{2^{n-2}(n-2)!}$$

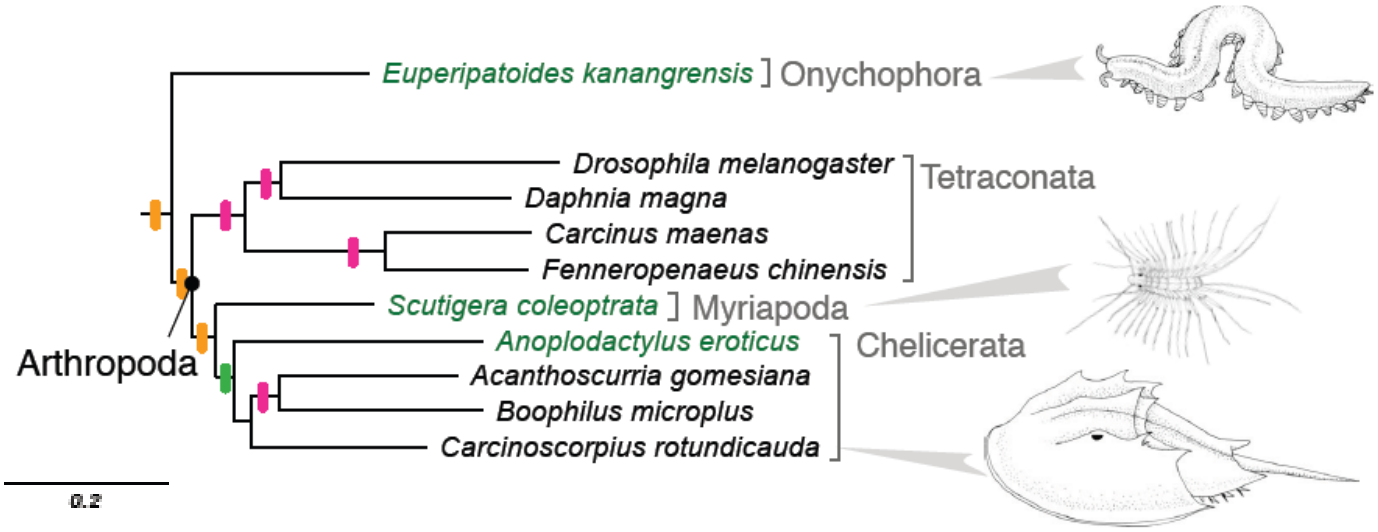
Number of bifurcating rooted trees for n taxa (OTUs)

OTU = operational taxonomical unit

Number of bifurcating trees is increasing with number of taxa. The number of unrooted trees for *n* taxa is equal for the number of rooted trees for (*n*-1) taxa.

No. of taxa	No. of unrooted trees	No. of rooted trees
3	1	3
4	3	15
5	15	105
10	2 027 025	34 459 425
11	34 459 425	654 729 075
12	654 729 075	13 749 310 575
50	1.00986×10^{57}	2.75292×10^{76}

From Zima a kol.
Genetické metody v zoologii



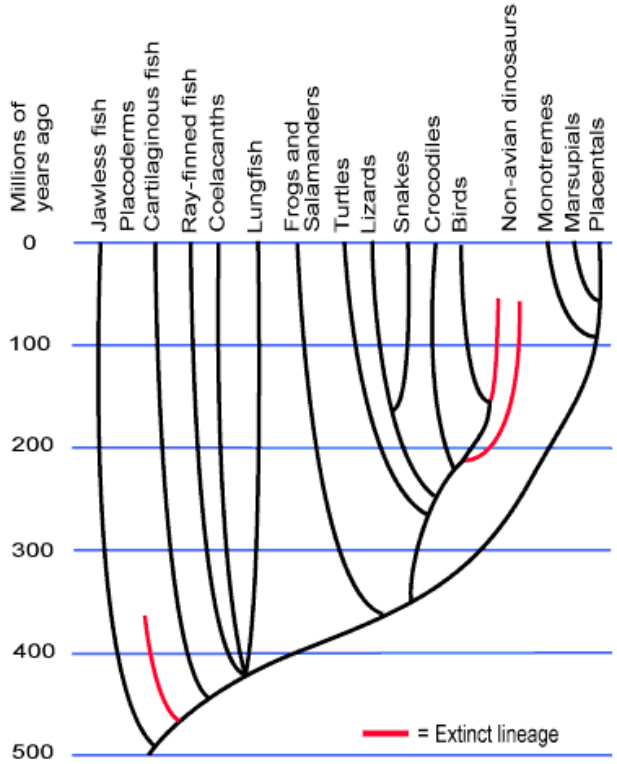
branch length = time of divergence from the common ancestor

molecular clock: sequence divergence increases over time linearly

when molecular clock holds (accumulation is linear over time) – all lineages in the tree have accumulated substitutions at the same rate

evolutionary rate dependent on metabolic rate, generation time, bottleneck events,…

we need calibration points (fossils, geological events)



How to make trees?

like family trees, phylogenetic trees represent patterns of ancestry

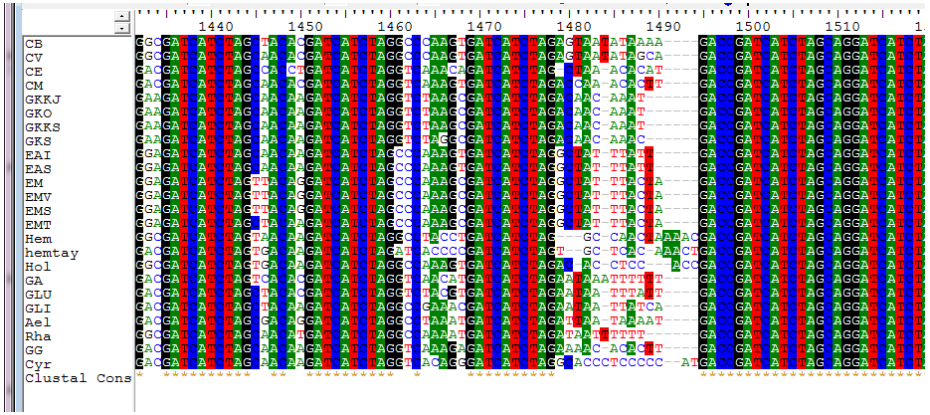
to reveal phylogenetic relationship, we have to compare characters which are **inherited from a common ancestor**

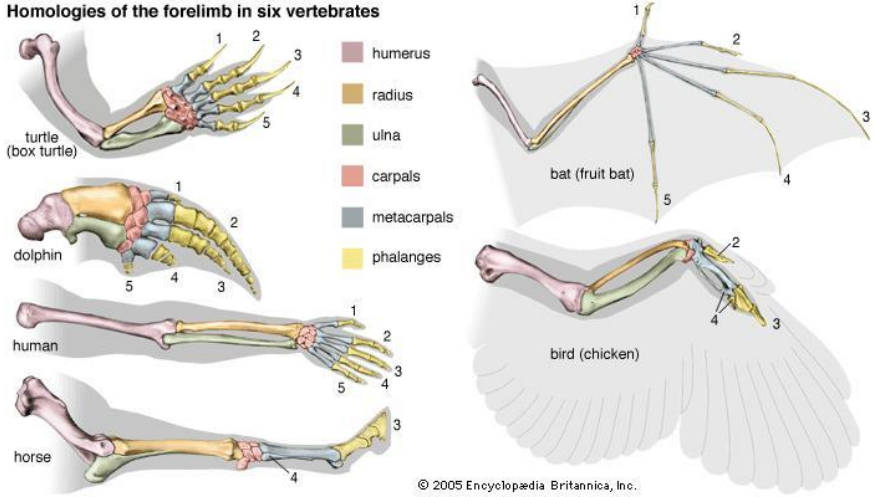
→ **homologous characters** (x analogous characters are result of convergent evolution)

character = heritable traits that can be compared across organisms

two or more forms = **character state**

types: physical characteristics (morphology), genetic sequences, and behavioral traits





Pros and cons of molecular characters (mainly sequence data)

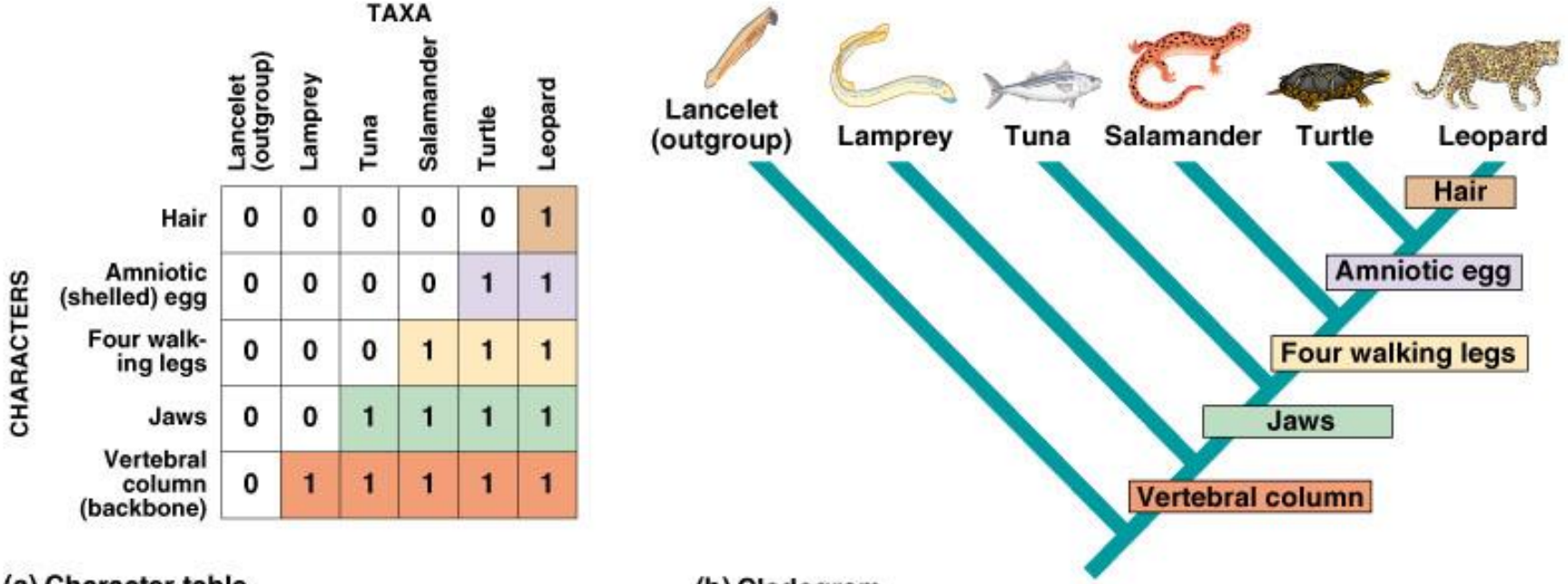
- molecular data - much more abundant (human genome 3,1 Gb, *E. coli* 4.6 Mb)
- independent (one position in the sequence of nucleotides x an eye – the eye is missing, but it means that also cornea, retina, etc. are missing)
- easy to describe (A, C, T, G at the position 175 in the cytochrome b gene) x some structure on the bone more pointed
- can resolve relationships at all different levels of organization, from species and populations to phyla and kingdoms
- less subjective
- neutral – number of shared characters mirrors phylogenetic relationship not just the same environmental selection pressures

Pros and cons of molecular characters (mainly sequence data)

- expensive
- technically demanding
- also non-experts on a given group can use them (lack of insight)
- no information about phenotype

How to make a tree - morphological example

-selected species have shared primitive and derived characters
-for reconstructing phylogeny derived shared characters = **synapomorphies** are important
-create a character table (matrix) with variable characters
-group the taxa based on synapomorphies - the more shared characters, the more closely related are the species
- for morphological data we use usually maximum parsimony method which prefers the simplest explanation of observed data



CHARACTERS	Lancelet (outgroup)	Lamprey	Tuna	Salamander	Turtle	Leopard
Hair	0	0	0	0	0	1
Amniotic (shelled) egg	0	0	0	0	1	1
Four walking legs	0	0	0	1	1	1
Jaws	0	0	1	1	1	1
Vertebral column (backbone)	0	1	1	1	1	1

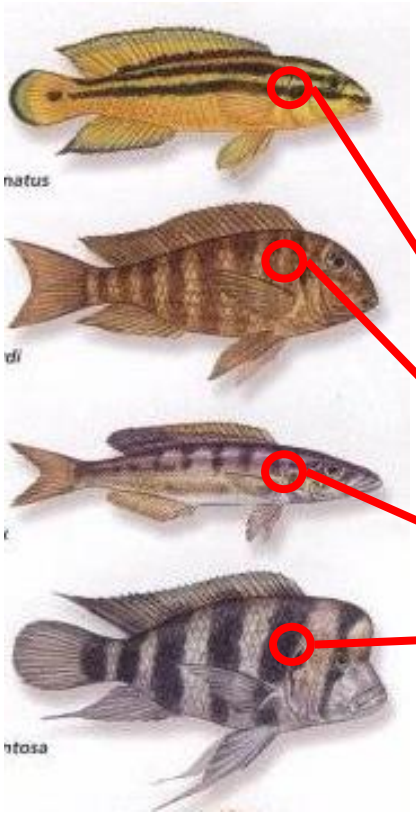
(a) Character table

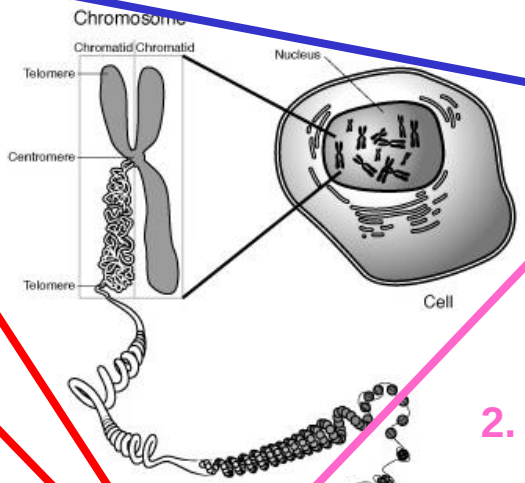
(b) Cladogram

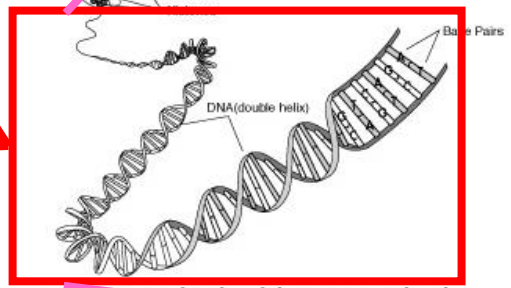
molecular characters

```
Outgroup   AAGCTTCATAGGAGCAACCATTCTAATAATAAGCCTCATAAAGCC
Taxon A    AAGCTTCACCGGCGCAGTTATCCTCATAATATGCCTCATAATGCC
Taxon B    GTGCTTCACCGACGCAGTTGTCCTCATAATGTGCCTCACTATGCC
Taxon C    GTGCTTCACCGACGCAGTTGCCCTCATGATGAGCCTCACTATGCA
```

3. alignment

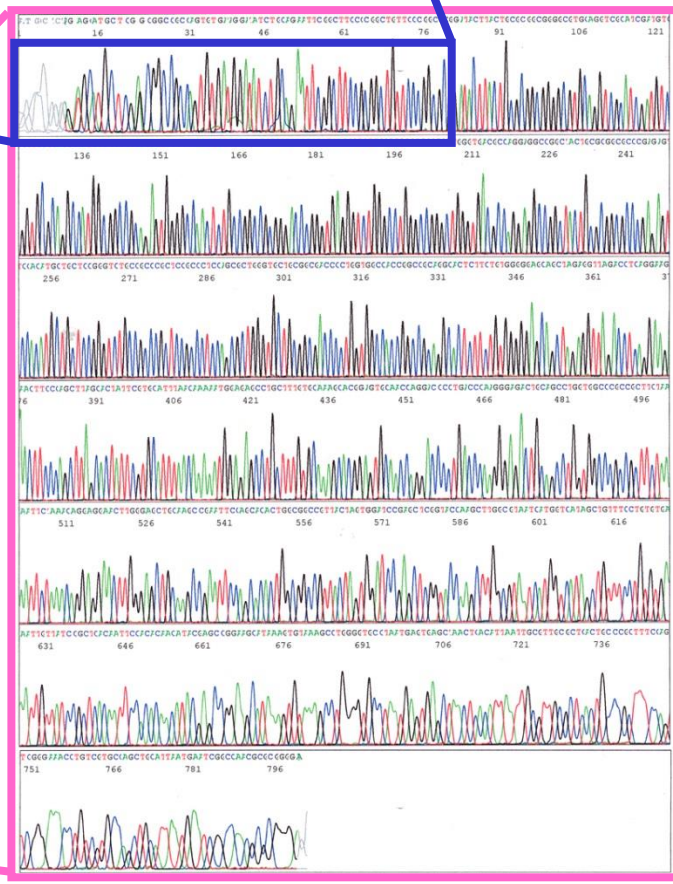






http://www.accessexcellence.org/AB/GG/chromosome.html

2. sequencing



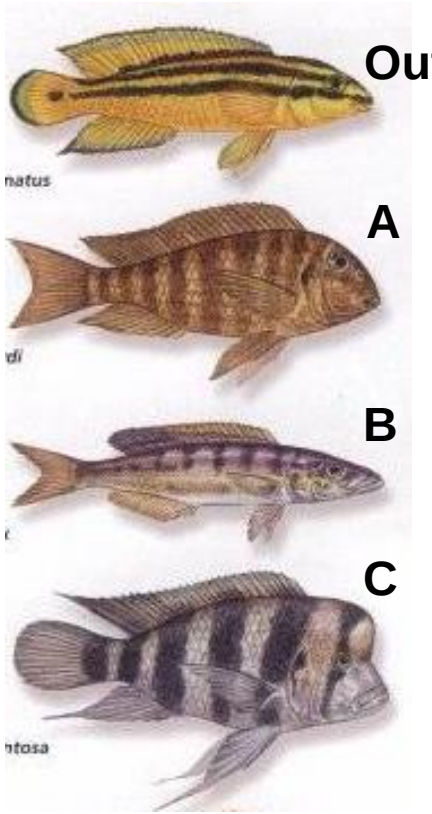
1.DNA isolation

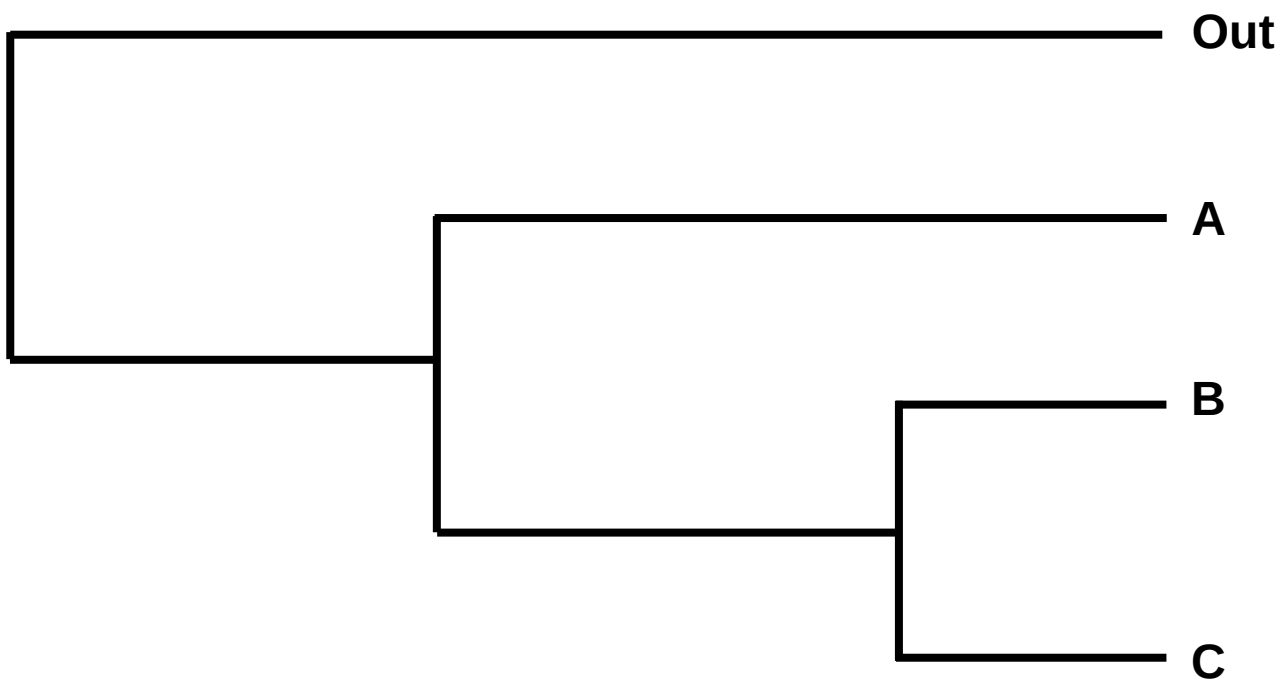
picture from R. Cox – Phylogeny. http://www.dartmouth.edu/~robertcox/Teaching.html

molecular characters

```
Outgroup   AAGCTTCATA
Taxon A    GAGCTTCACA
Taxon B    GTGCTTCACG
Taxon C    GTGCTTCACG
```

invariant sites





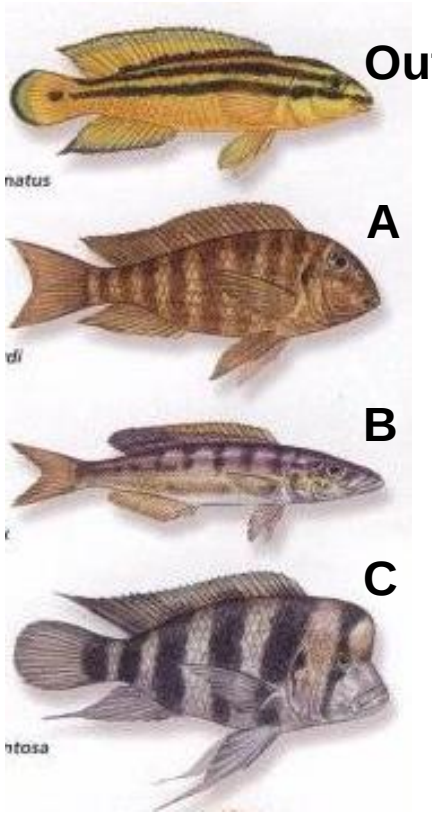
picture from R. Cox – Phylogeny. http://www.dartmouth.edu/~robertcox/Teaching.html

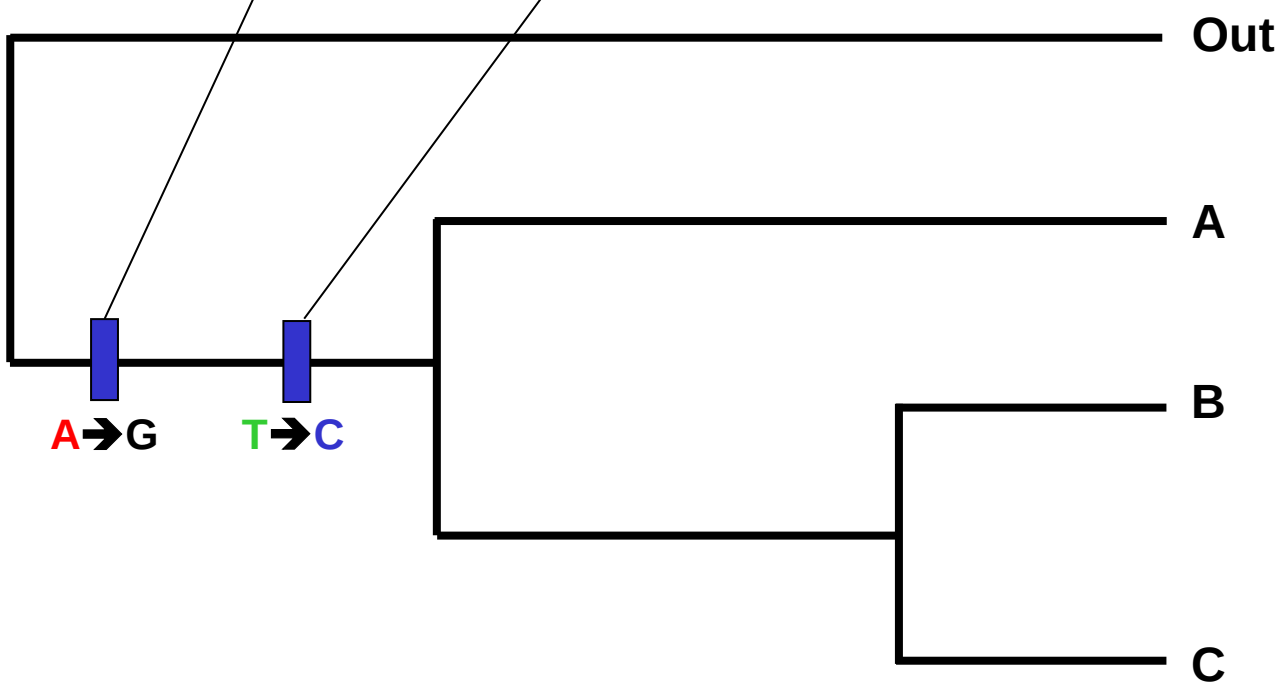
molecular characters

Outgroup
Taxon A
Taxon B
Taxon C

AAGCTTCATA
GAGCTTCACA
GTGCTTCACG
GTGCTTCACG

Synapomorphies supporting A+B+C





picture from R. Cox – Phylogeny. http://www.dartmouth.edu/~robertcox/Teaching.html

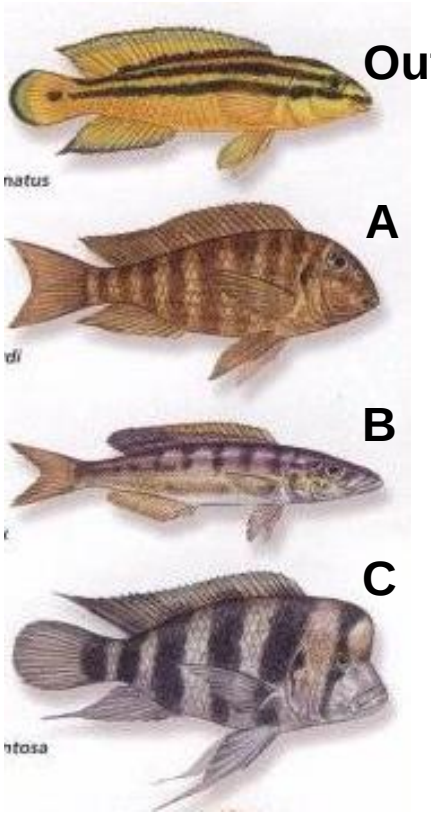
molecular characters

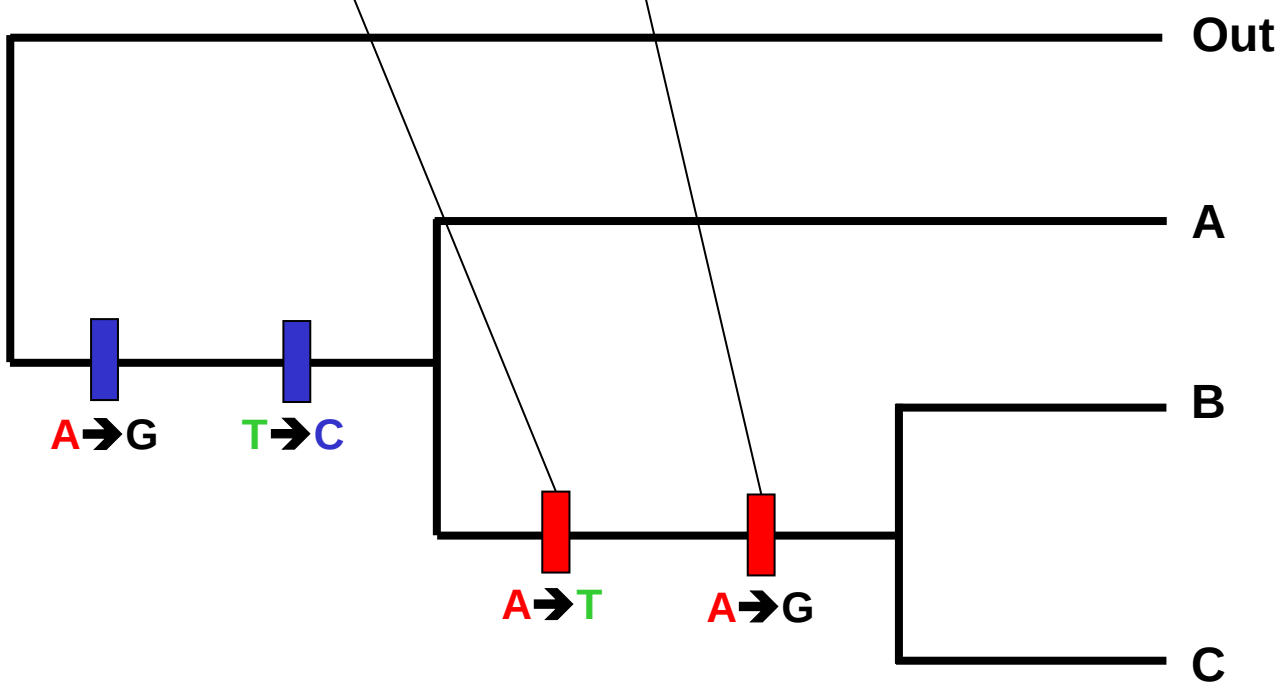
Outgroup
Taxon A
Taxon B
Taxon C

AAGCTTCATA
GAGCTTCACA
GTGCTTCACG
GTGCCTCACG

Synapomorphies supporting A+B+C

Synapomorphies supporting B+C





picture from R. Cox – Phylogeny. http://www.dartmouth.edu/~robertcox/Teaching.html

molecular characters

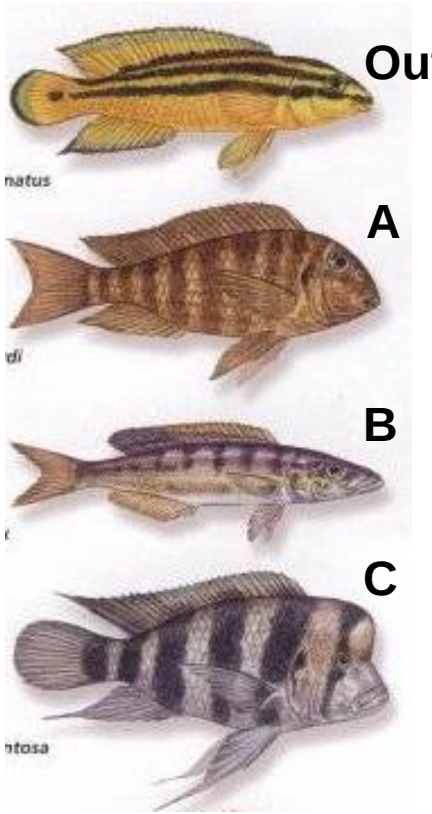
Outgroup
Taxon A
Taxon B
Taxon C

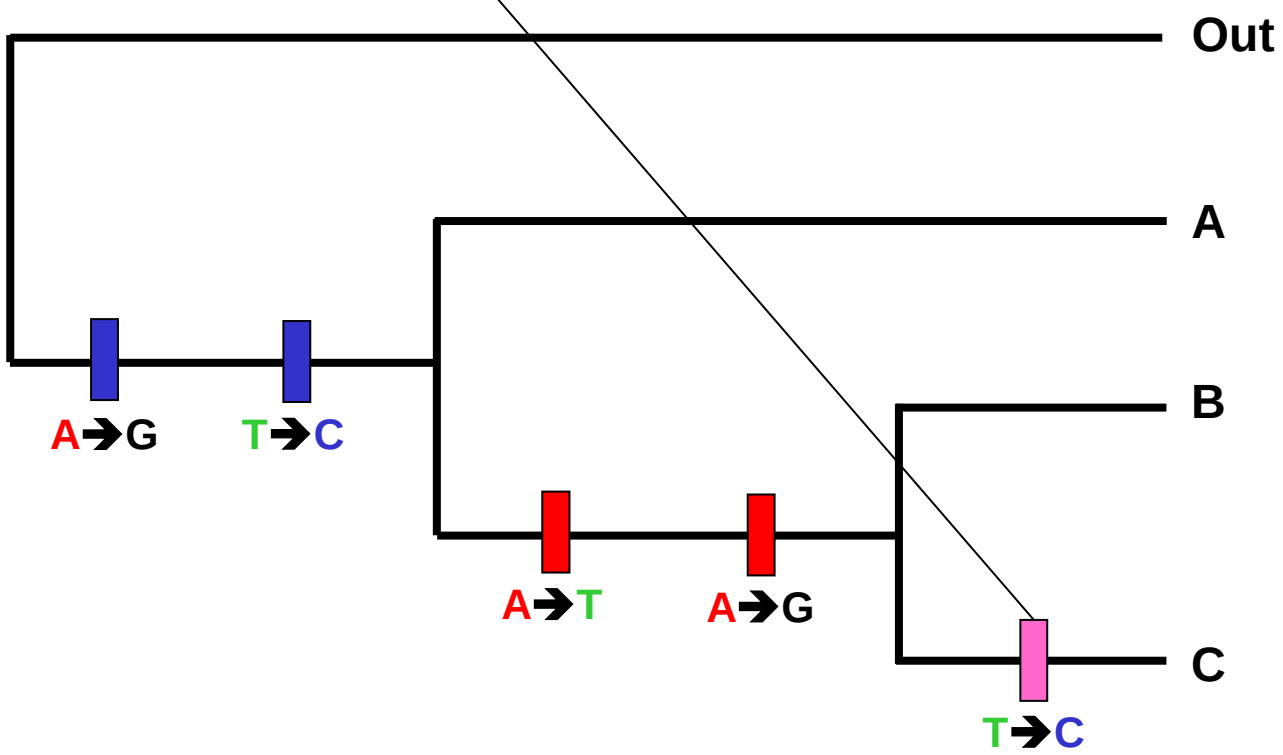
AAGCTTCATA
GAGCTTCACA
GTGCTTCACG
GTGCCTCACG

Synapomorphies supporting A+B+C

Synapomorphies supporting B+C

Apomorphy for C





picture from R. Cox – Phylogeny. http://www.dartmouth.edu/~robertcox/Teaching.html

How to make phylogenetic trees?

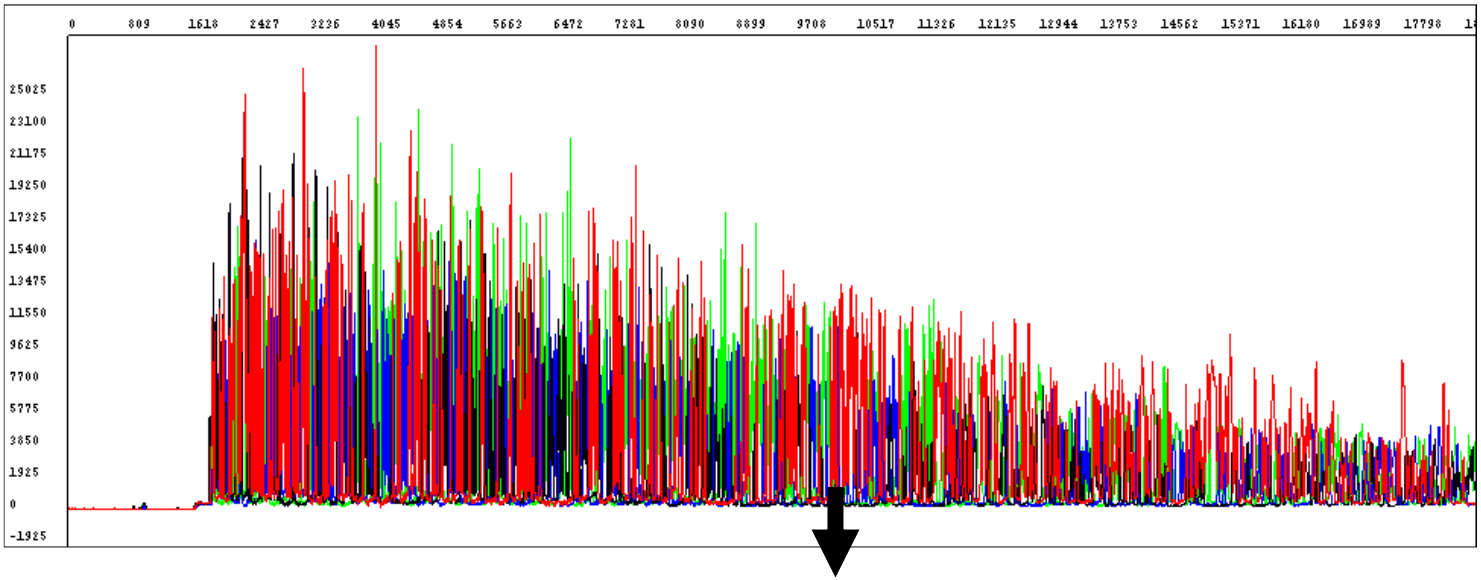
Workflow:

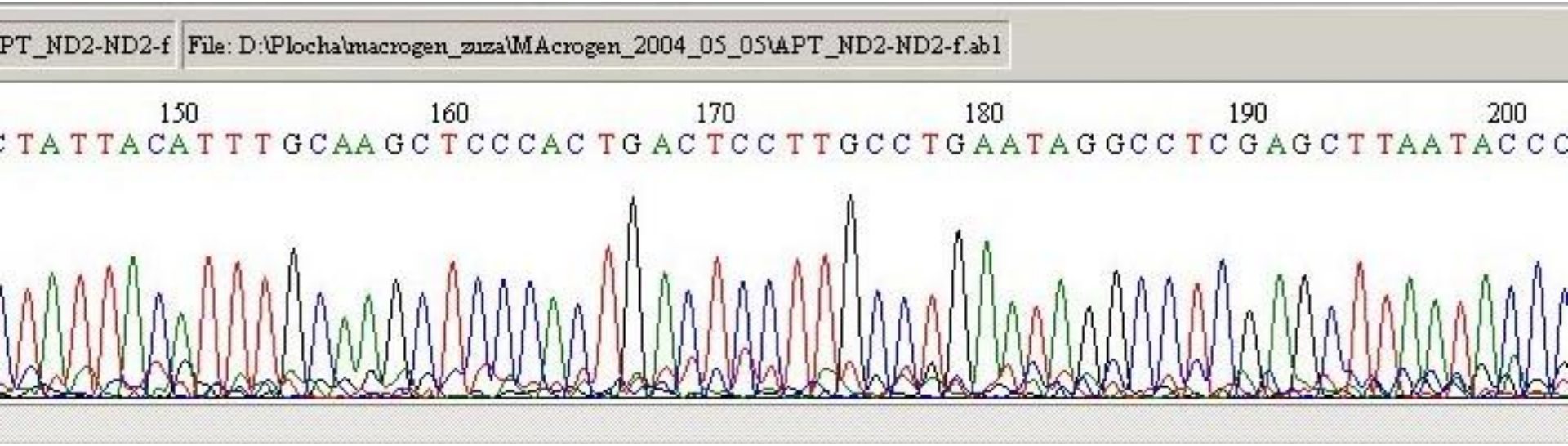
obtain DNA sequence

- quality check
- sequence alignment
- calculating genetic distances
- phylogeny estimation – topology and branch length
- reliability test (bootstrap)
- tree visualization

Chromatogram files:

Can be opened in different software- Chromas, BioEdit, DNASTAR , …





Checking sequences and chromatograms programs: Chromas, BioEdit, Geneious

Quality control:

Protein coding sequences: translation into amino acids

Checking for stop codons within the sequence of the gene (UAG, UAA, UGA)

Is start codon AUG (codes methionine (Met)) present?

Be aware of using correct translation table:

– nuclear vs. mitochondrial DNA, mitochondrial DNA (vertebrates, invertebrates, …)

http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi?mode=c#SG2

```
LOCUS       AF248660                1036 bp    DNA     linear   INV 14-NOV-2000
DEFINITION  Gryllus bimaculatus haplotype 2 cytochrome b gene, partial cds;
            mitochondrial gene for mitochondrial product.
ACCESSION   AF248660
VERSION     AF248660.1  GI:11141460
KEYWORDS    .
SOURCE      mitochondrion Gryllus bimaculatus (two-spotted cricket)
  ORGANISM  Gryllus bimaculatus
            Eukaryota; Metazoa; Arthropoda; Hexapoda; Insecta; Pterygota;
            Neoptera; Orthopteroidea; Orthoptera; Ensifera; Grylloidea;
            Gryllidae; Gryllinae; Gryllus.
REFERENCE   1  (bases 1 to 1036)
  AUTHORS   Huang,Y., Orti,G., Sutherlin,M., Duhachek,A. and Zera,A.
  TITLE     Phylogenetic relationships of north American field crickets
            inferred from mitochondrial DNA data
  JOURNAL   Mol. Phylogenet. Evol. 17 (1), 48-57 (2000)
   PUBMED   11020304
REFERENCE   2  (bases 1 to 1036)
  AUTHORS   Huang,Y., Sutherlin,M. and Duhachek,A.
  TITLE     Direct Submission
  JOURNAL   Submitted (24-MAR-2000) School of Biological Sciences, University
            of Nebraska, 314 Manter Hall, Lincoln, NE 68588, USA
FEATURES             Location/Qualifiers
     source          1..1036
                     /organism="Gryllus bimaculatus"
                     /organelle="mitochondrion"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:6999"
                     /haplotype="2"
                     /country="Cyprus"
     CDS             <1..>1036
                     /codon_start=2
                     /transl_table=5
                     /product="cytochrome b"
                     /protein_id="AAG31952.1"
                     /db_xref="GI:11141461"
                     /translation="FGSLLGLCLVIQIMTGIFLAMHYTADINMAFNSVSHICRDVNFG
                     WLLRTMHANGASFFFICLYTHVGRGMYYGSYNLMFTWMTGTLILFLVMAAAFMGYVLP
```

```
LOCUS       AF020255                902 bp    DNA     linear   VRT 30-SEP-1997
DEFINITION  Cyclura nubila cytochrome b (cytb) gene, mitochondrial gene
            encoding mitochondrial protein, partial cds.
ACCESSION   AF020255
VERSION     AF020255.1  GI:2444366
KEYWORDS    .
SOURCE      mitochondrion Cyclura nubila
  ORGANISM  Cyclura nubila
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Lepidosauria; Squamata; Iguania; Iguanidae; Iguaninae; Cyclura.
REFERENCE   1  (bases 1 to 902)
  AUTHORS   Petren,K. and Case,T.J.
  TITLE     A Phylogenetic analysis of body size evolution and biogeography in
            chuckwallas (Sauromalus) and other iguanines
  JOURNAL   Evolution 51, 206-219 (1997)
REFERENCE   2  (bases 1 to 902)
  AUTHORS   Petren,K. and Case,T.J.
  TITLE     Direct Submission
  JOURNAL   Submitted (20-AUG-1997) EEB, Princeton University, Guyot Hall,
            Washington Rd., Princeton, NJ 08544-1003, USA
FEATURES             Location/Qualifiers
     source          1..902
                     /organism="Cyclura nubila"
                     /organelle="mitochondrion"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:51214"
                     /note="specimen from Cuba, (pop. #27)"
     gene            <1..>902
                     /gene="cytb"
     CDS             <1..>902
                     /gene="cytb"
                     /codon_start=2
                     /transl_table=2
                     /product="cytochrome b"
                     /protein_id="AAB71582.1"
                     /db_xref="GI:2444367"
                     /translation="FIDLPTPSNISSWWNFGSLLGLCLIIQILTGLFLAMHYTANISH
```

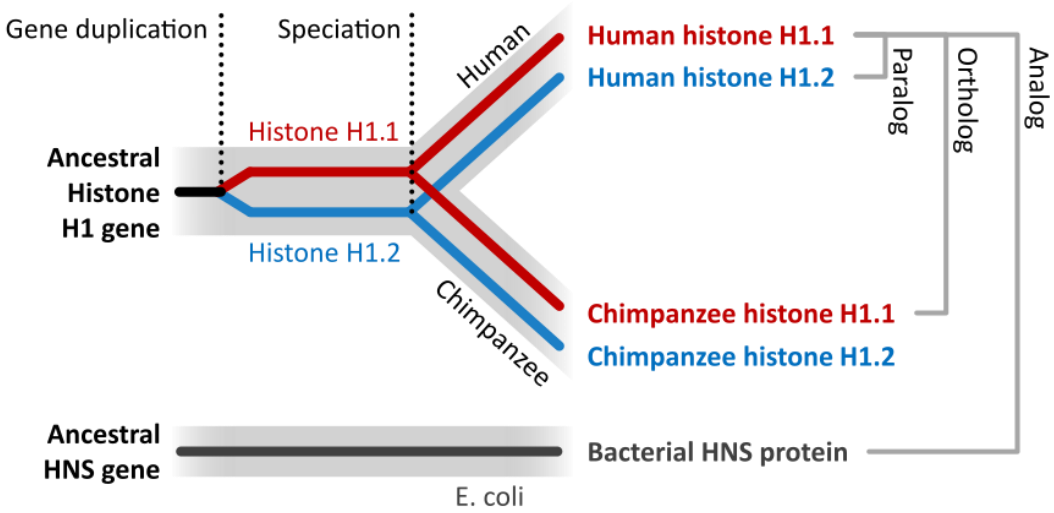
homology of molecular characters

it is necessary to compare sequences (of e.g. genes) which are **orthologous =** inferred to be descended from the same ancestral sequence separated by a speciation event

X

xenolog – sequence of gene incorporated from other species by horizontal transfer

pseudogene – sequence copied from the mitochondrial genome to the nuclear DNA



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**.

source: Wikipedia

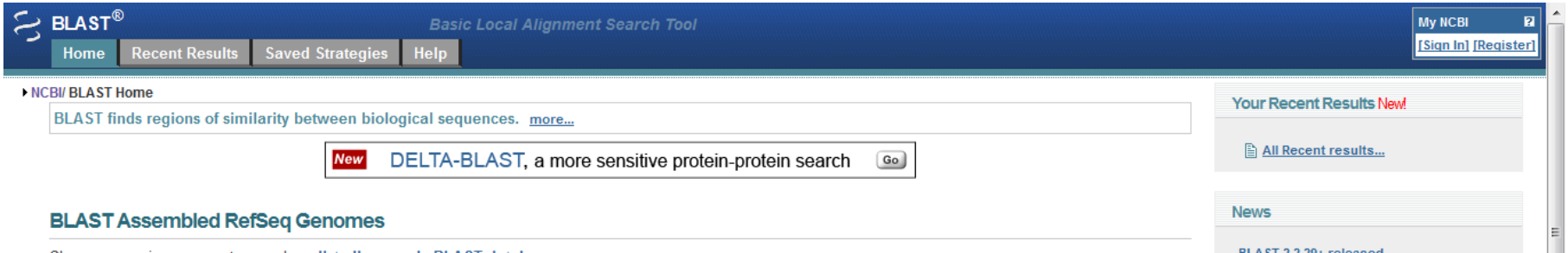
Additional sources of DNA sequences:

- public databases

http://www.ncbi.nlm.nih.gov/

BLAST - **Basic Local Alignment Search Tool**

- algorithm for searching in databases for similar sequences

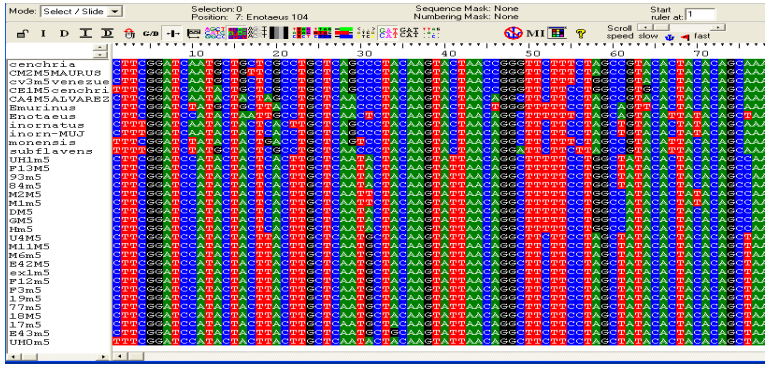


How to make phylogenetic trees?

Workflow:

- [x] obtain DNA sequence
- [x] quality check
- sequence alignment
- calculating genetic distances
- phylogeny estimation – topology and branch length
- reliability test (bootstrap)
- tree visualization

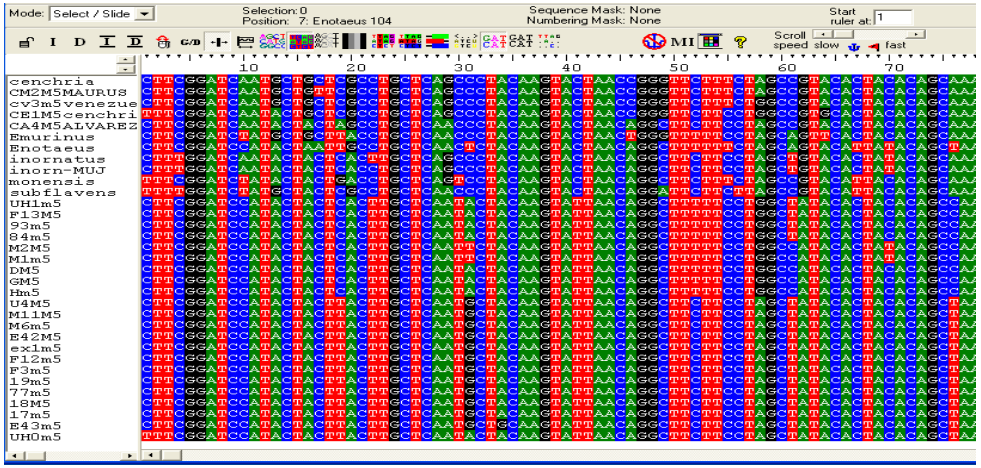
Where sequences differ and where are the same?



Alignment

- a way of arranging the sequences of DNA (also e.g. amino acid in the protein sequence) to identify regions of similarity
- start of every phylogenetic analysis
- assessing of position homology of each base in the sequence
- each position (column in the alignment) in the sequence represents character potentially useful for the phylogenetic analysis
- different programs for calculating and editing alignments
  - manual: BioEdit, Macaw
  - automatic – different algorithms

Clustal X, PileUp, Multalin, Mafft – often online



Alignment - pairwise alignment (two sequences)

- multiple alignment (more sequences)

```
AATGCCCTAAA
AATGCGGCTAAA
AACGCGCTAAA
ATGCTAA
```

```
AATGCC-CTAAA
AATGCGGCTAAA
AACGCG-CTAAA
-ATG---CTAA-
```

gaps are inserted between the bases so that identical or similar characters are aligned in successive columns

during alignment: algorithm is searching for the best position for each base and **receives points**:
Plus points for every match,
Minus points for inserting or extension of the gap,

The goal is to obtain the best score

(A)

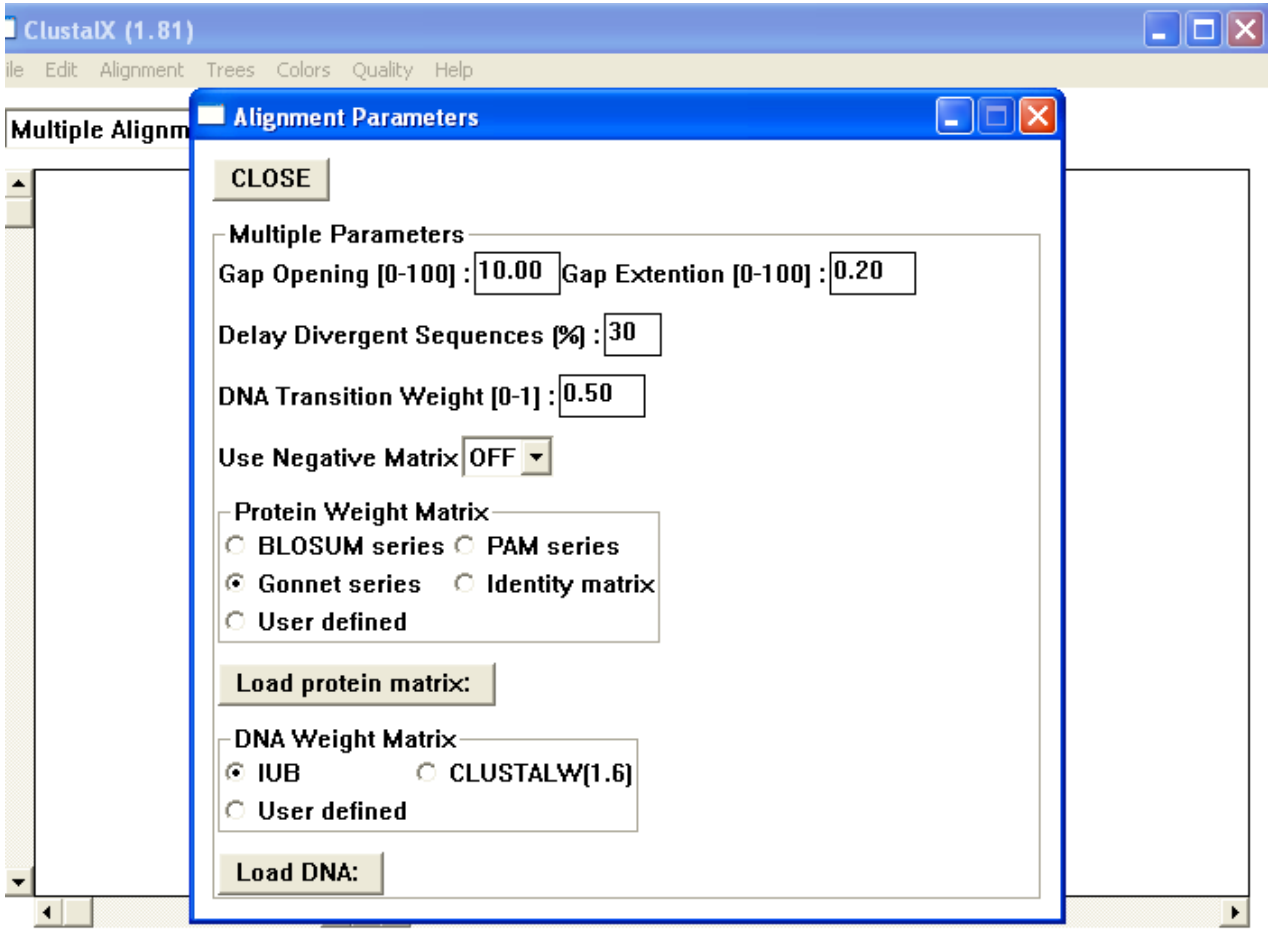
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	9																			
S	−1	4																		
T	−1	1	5																	
P	−3	−1	−1	7																
A	0	1	0	−1	4															
G	−3	0	−2	−2	0	6														
N	−3	1	0	−2	−2	0	6													
D	−3	0	−1	−1	−2	−1	1	6												
E	−4	0	−1	−1	−1	−2	0	2	5											
Q	−3	0	−1	−1	−1	−2	0	0	2	5										
H	−3	−1	−2	−2	−2	−2	1	−1	0	0	8									
R	−3	−1	−1	−2	−1	−2	0	−2	0	1	0	5								
K	−3	0	−1	−1	−1	−2	0	−1	1	1	−1	2	5							
M	−1	−1	−1	−2	−1	−3	−2	−3	−2	0	−2	−1	−1	5						
I	−1	−2	−1	−3	−1	−4	−3	−3	−3	−3	−3	−3	−3	1	4					
L	−1	−2	−1	−3	−1	−4	−3	−4	−3	−2	−3	−2	−2	2	2	4				
V	−1	−2	0	−2	0	−3	−3	−3	−2	−2	−3	3	2	1	3	1	4			
F	−2	−2	−2	−4	−2	−3	−3	−3	−3	−3	−1	−3	−3	0	0	0	−1	6		
Y	−2	−2	−2	−3	−2	−3	−2	−3	−2	−1	2	−2	−2	−1	−1	−1	−1	3	7	
W	−2	−3	−2	−4	−3	−2	−4	−4	−3	−2	−2	−3	−3	−1	−3	−2	−3	1	2	11

Figure 4.4
Amino acid substitution scoring matrices. (A) The BLOSUM-62 matrix and (B) the PAM120 substitution matrix. Each cell represents the score given to a residue paired with another residue (row × column). The values are given in half-bits, as discussed in Section 5.1. The colored shading indicates different physicochemical properties of the residues (see Figure 2.3): small and polar, yellow; small and nonpolar, white; polar or acidic, red; basic, blue; large and hydrophobic, green; aromatic, orange.

scoring matrices

Alignment:

Clustal W (Clustal X)
www.clustal.org – frequently used software (Thompson et al. 1994 a 1997)



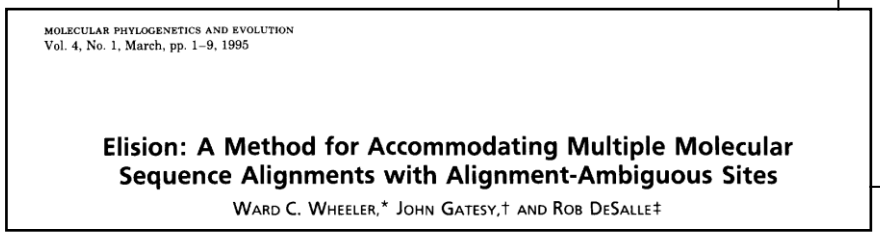
What can help:

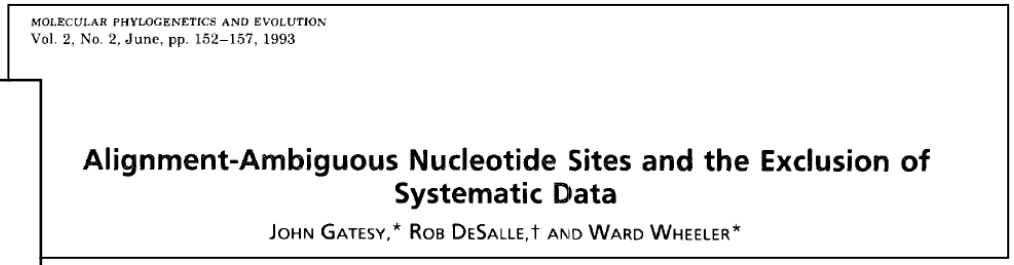
- first use default parameters, later increase penalty for gap opening and decrease penalty for gap extension

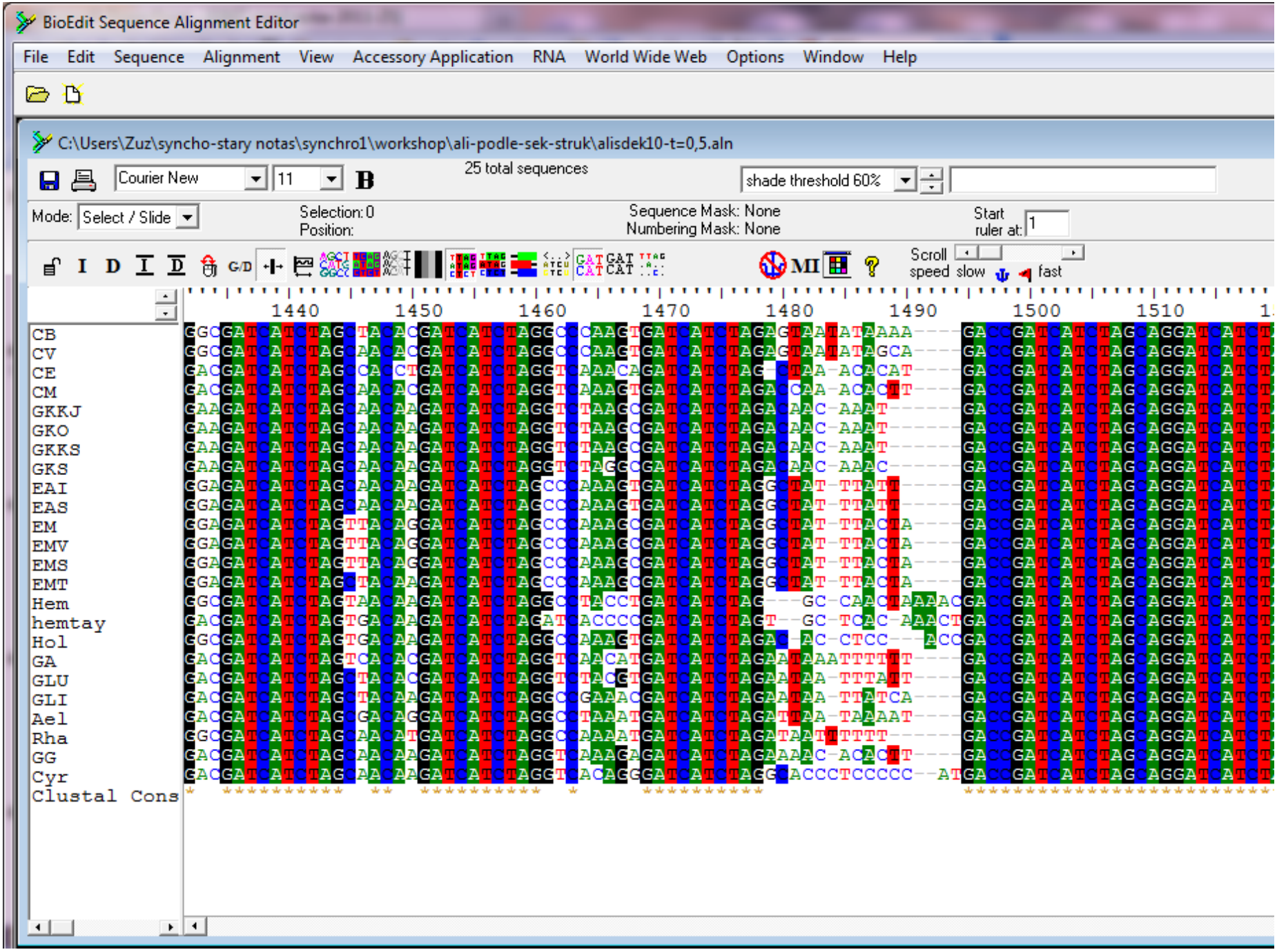
- coding genes align as amino acids

- use information about sequence secondary structure (genes for 12S and 16S RNA) – database with alignments with secondary structure in mind (http://www.arb-silva.de/)

- delete ambiguously aligned positions






BioEdit Sequence Alignment Editor
File Edit Sequence Alignment View Accessory Application RNA World Wide Web Options Window Help
C:\Users\Zuz\syncho-stary notas\synchro1\workshop\ali-podle-sek-struk\alisdek10-t=0,5.aln
Courier New
11
25 total sequences
shade threshold 60%
Mode: Select / Slide
Selection: 0
Position:
Sequence Mask: None
Numbering Mask: None
Start ruler at: 1
Scroll speed slow fast
1440 1450 1460 1470 1480 1490 1500 1510
CB
CV
CE
CM
GKKJ
GKO
GKKS
GKS
EAI
EAS
EM
EMV
EMS
EMT
Hem
hemtay
Hol
GA
GLU
GLI
Ael
Rha
GG
Cyr
Clustal Cons

Phylogeny estimation

character based (maximum parsimony, maximum likelihood, Bayesian analysis)

two types of methods

distance based (Neighbour-joining, UPGMA)

Two different approaches:
algorithm – number of specific steps resulting in one best tree
methods: UPGMA, Neighbour-joining

optimality criterion – consider and compare all theoretically possible trees based on selected criteria- number of evolutionary steps, likelihood value