



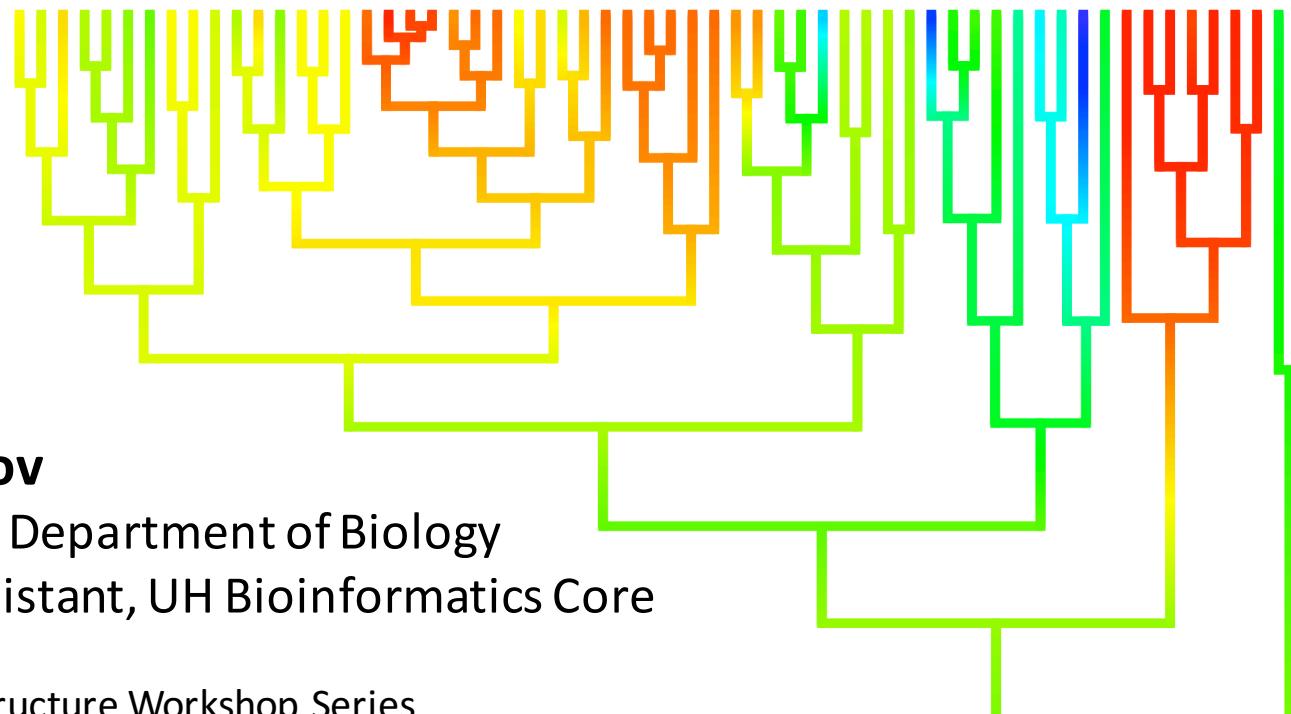
INFORMATION
TECHNOLOGY
SERVICES



University of Hawaii Bioinformatics Core



An Introduction to Molecular Phylogenetic Inference



Kirill Vinnikov

PhD Student, Department of Biology

Graduate Assistant, UH Bioinformatics Core

ITS Cyberinfrastructure Workshop Series

September 9, 2016



Goals for today

- How to read phylogenetic trees
- How to access sequence information in online databases
- How and when to use different sequence alignment methods
- How and when to use different phylogenetic methods
- How to visualize your phylogenetic trees



Workshop schedule

9:00 - 9:40 am Introduction and Basic Concepts

9:40 - 10:20 am Practical Lab 1

10:20 - 10:30 am **BREAK**

10:30 - 11:10 am Introduction to Phylogenetic Methods

11:10 - 12:00 pm Practical Lab 2

12:00 - 12:10 pm **BREAK**

12:10 – 1:00 pm Practical Lab 3

1:00 pm Q&A



1. Introduction and basic concepts

Outline

- The tree thinking challenge
- Understanding the basics:
homology, orthology, paralogy, xenology
- Sequence alignment approaches
- Introduction to online sequence databases
- Phylogenetics vs Phylogenomics: what's the difference?



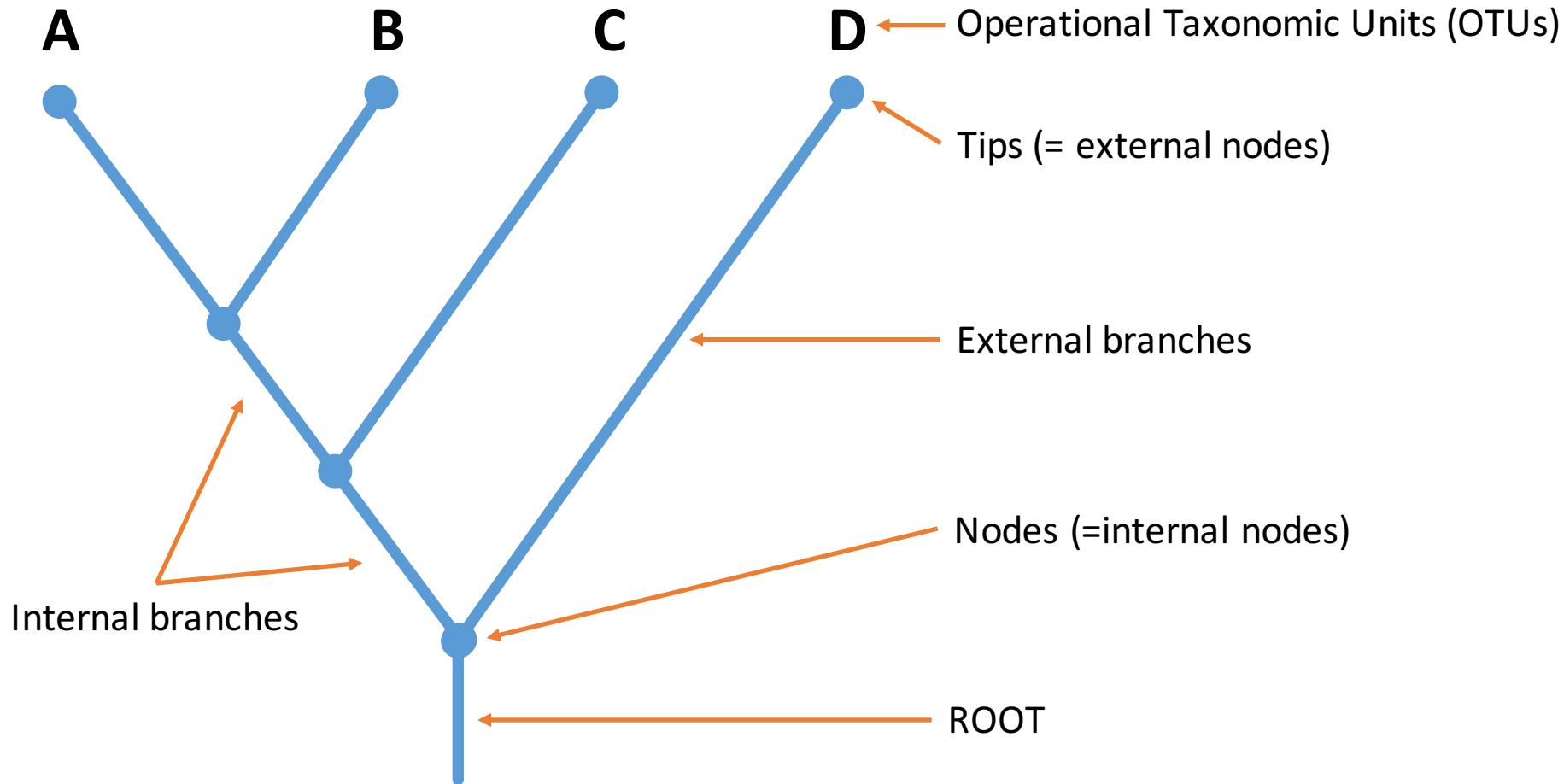
Why use phylogenies?

**“Nothing in biology makes sense
except in the light of evolution”**

Theodosius Dobzhansky (1900-1975)

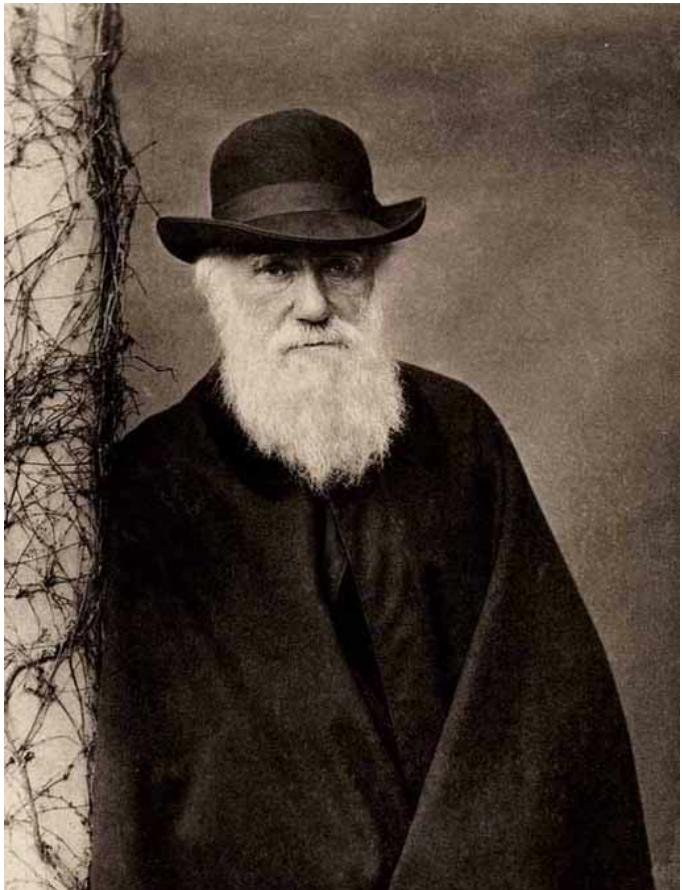
- Evolution gives us a mechanistic understanding of life
- It changed biology from a descriptive to predictive science
- Phylogenetic trees provide the easiest way to depict evolutionary relationships between organisms
- Phylogenies allow us to test different evolutionary hypotheses

Phylogenetic Tree Terminology

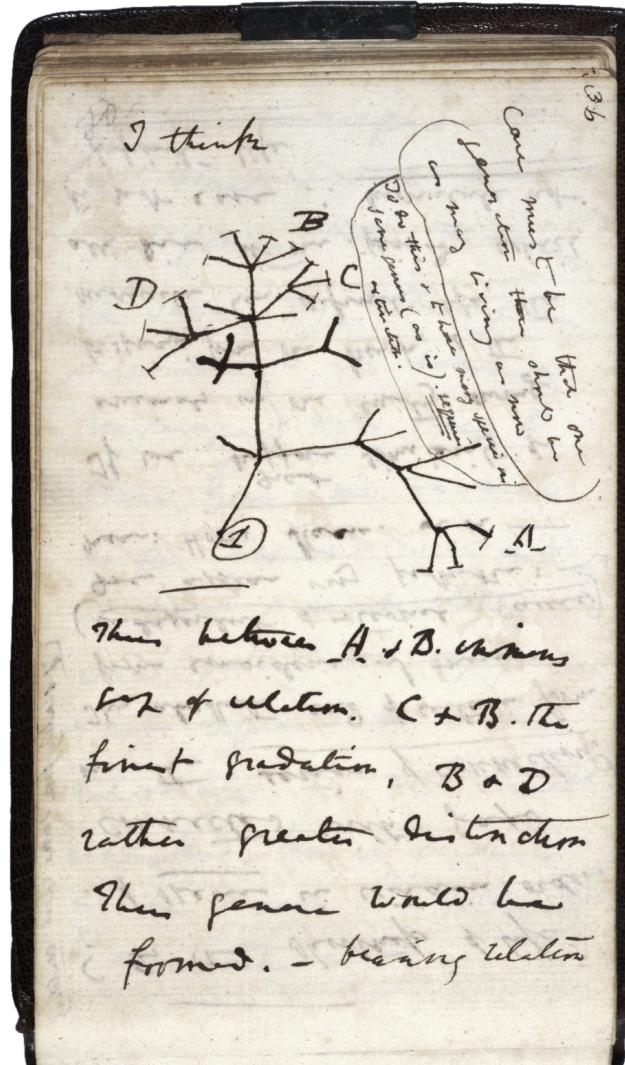




First phylogenetic trees



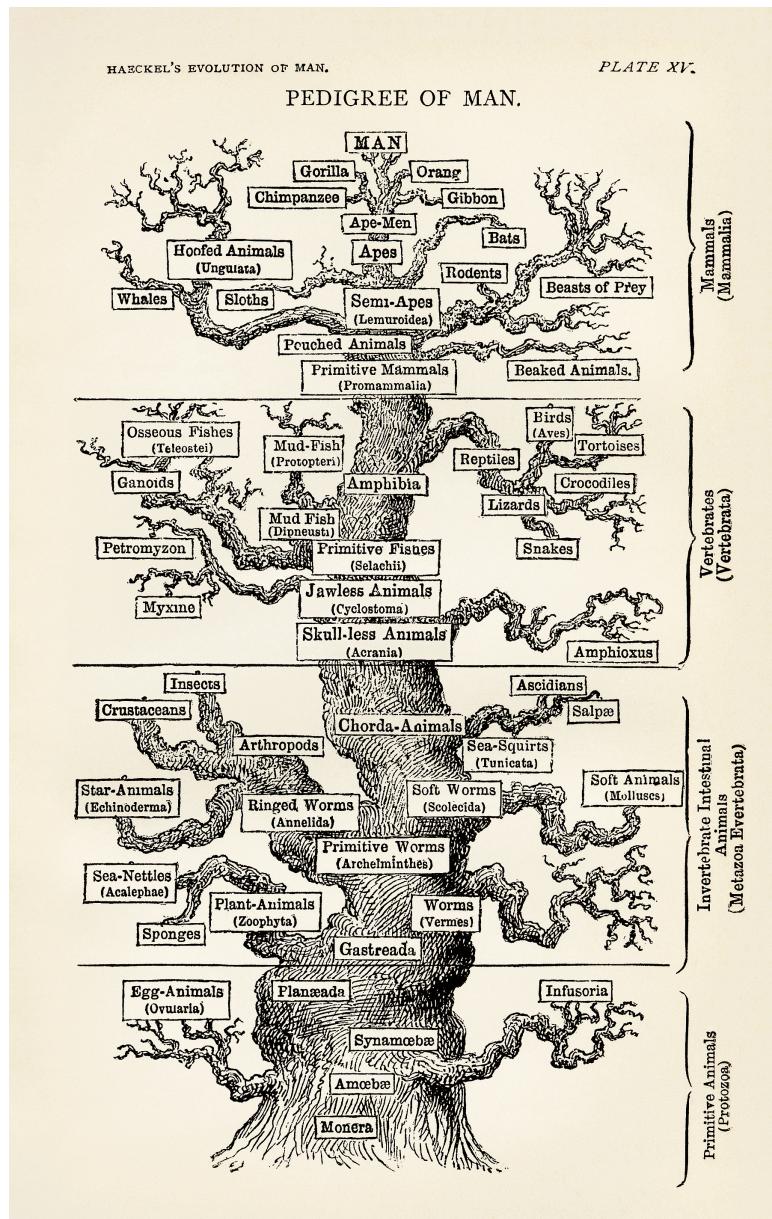
Charles Darwin (1809-1882)



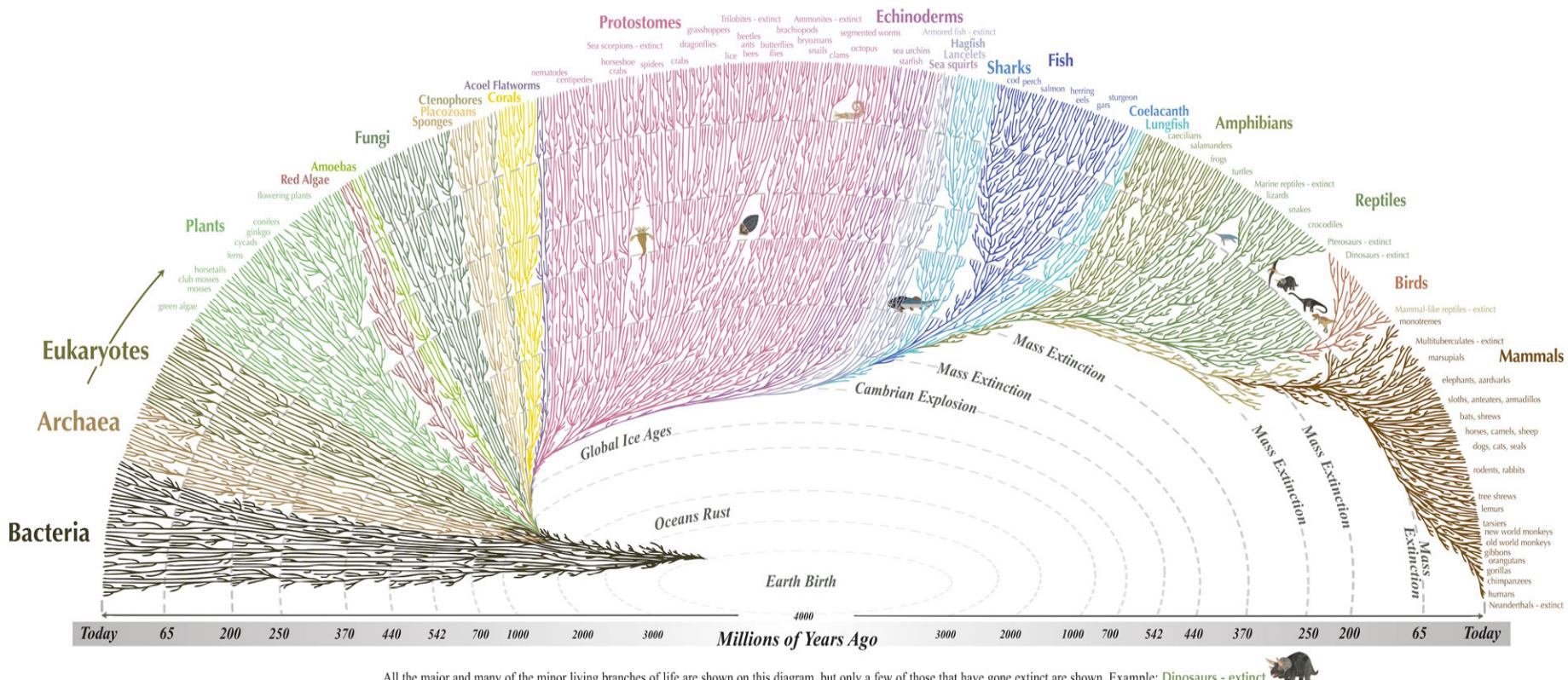
First phylogenetic trees



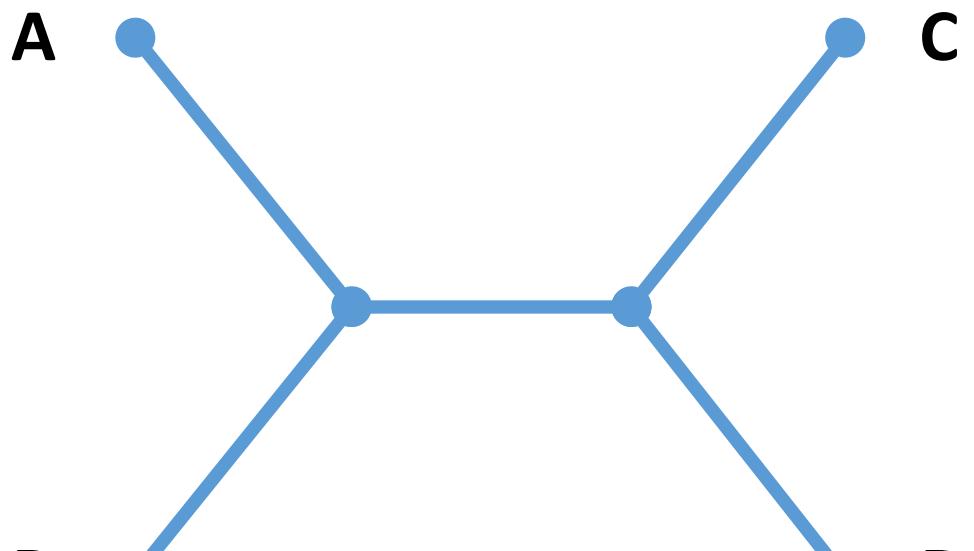
Ernst Haeckel (1834-1919)



Modern phylogenetic trees: the Tree of Life



Understanding Tree Topology



Unrooted tree

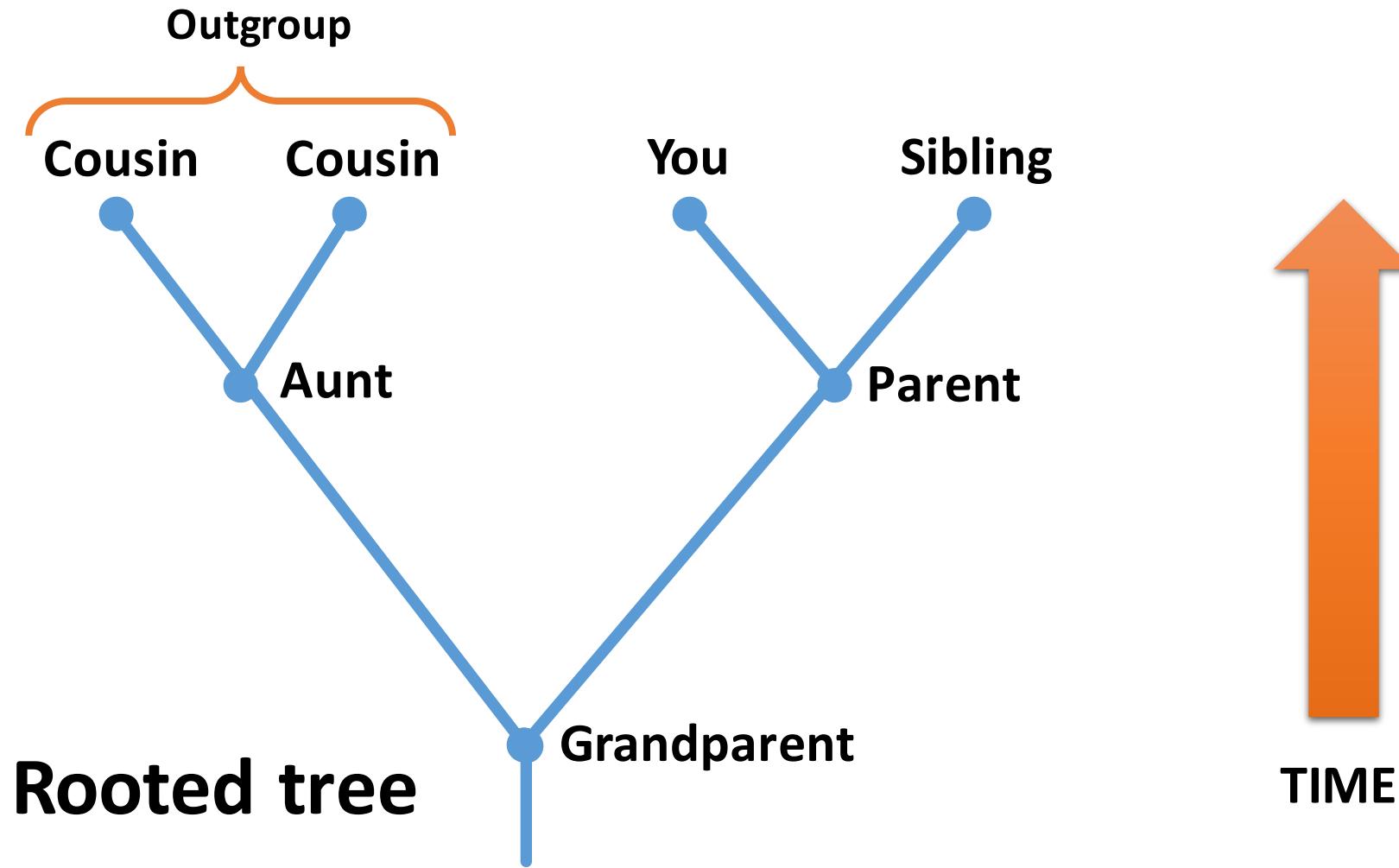
Ancestral Derived

?

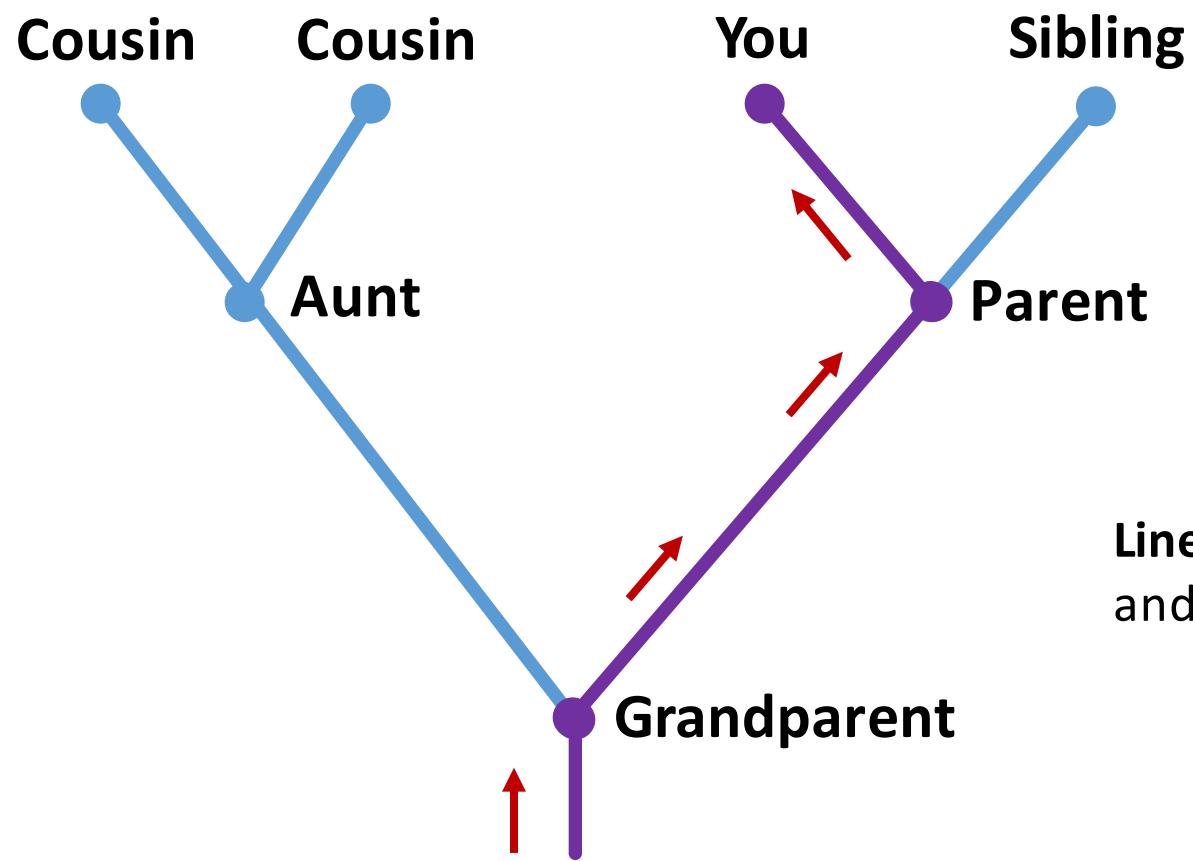


direction is unknown

Understanding Tree Topology



Understanding Tree Topology

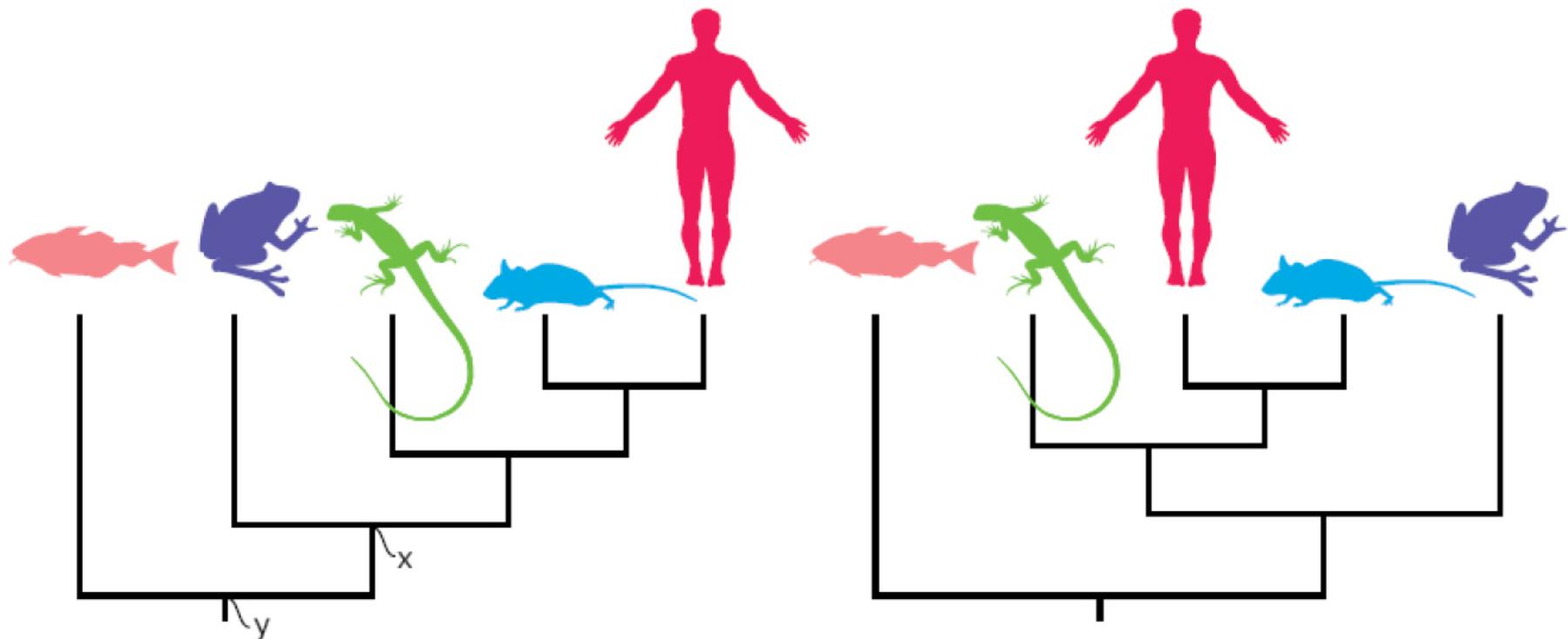


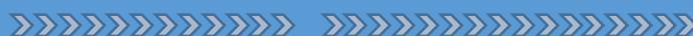
Lineage – a chain of ancestors and their descendants



The tree thinking challenge

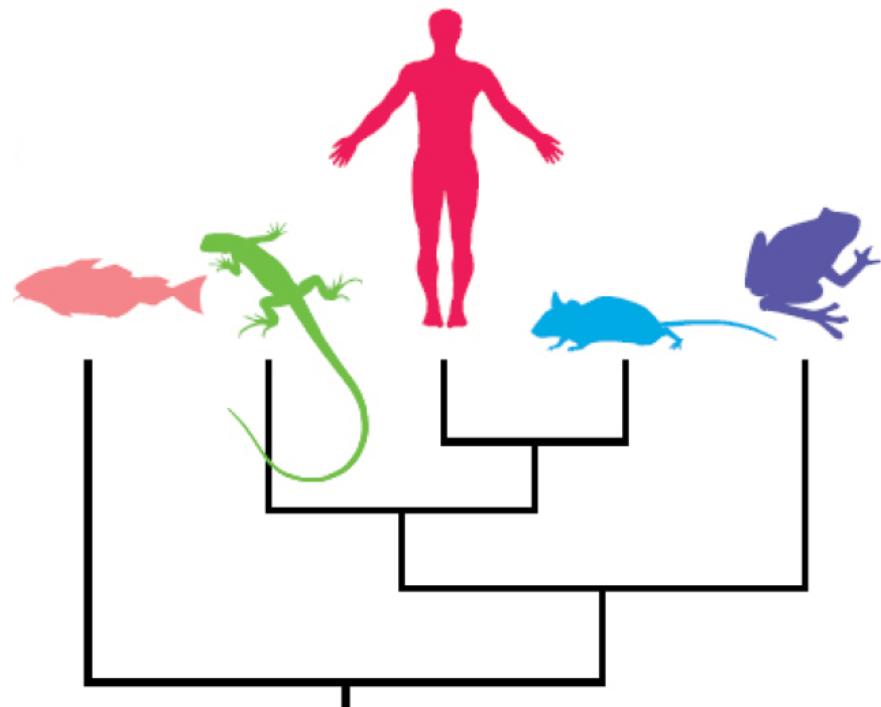
Which phylogenetic tree is more accurate?





The tree thinking challenge

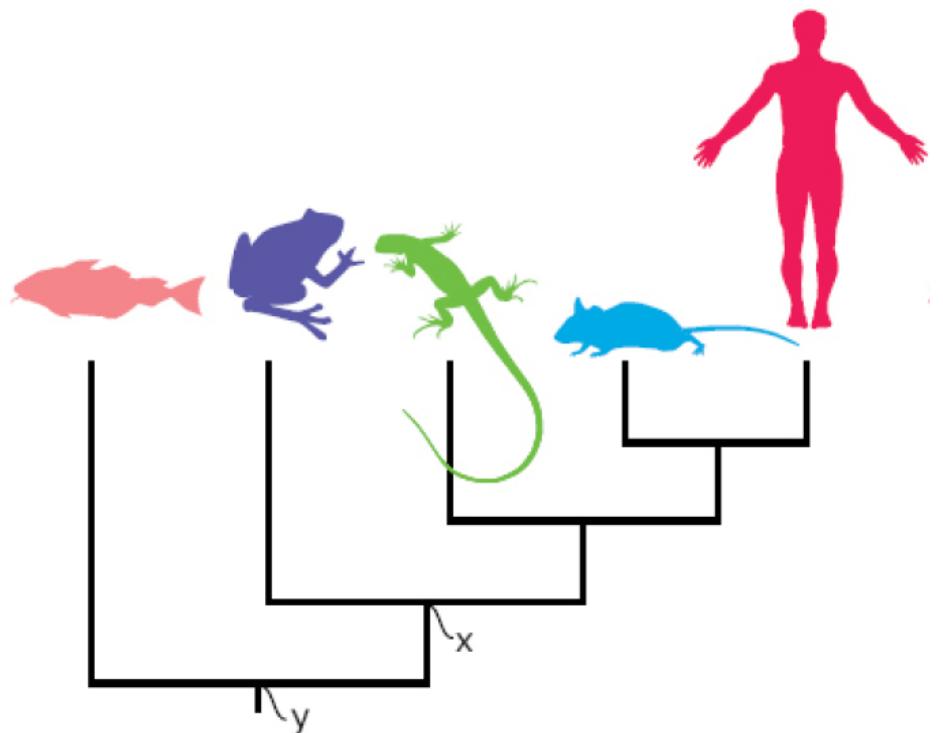
Is the frog more closely related
to the fish or to the human?





The tree thinking challenge

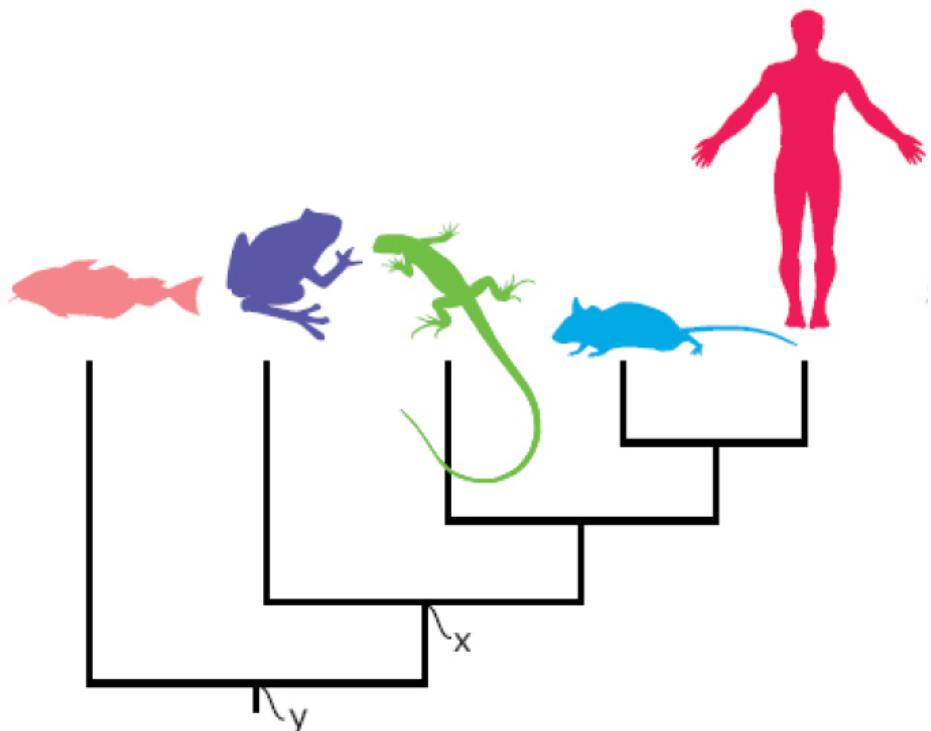
Is the frog more closely related
to the fish or to the human?





The tree thinking challenge

Is the frog more closely related
to the fish or to the human?



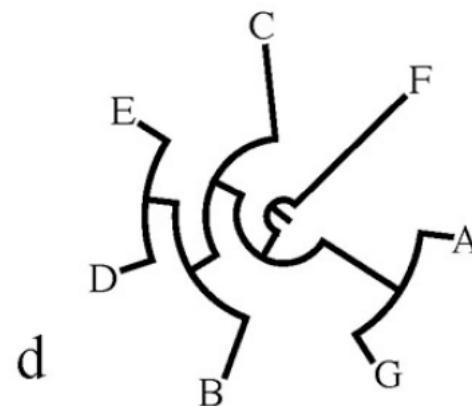
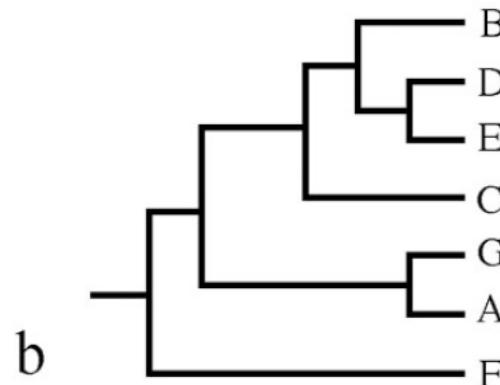
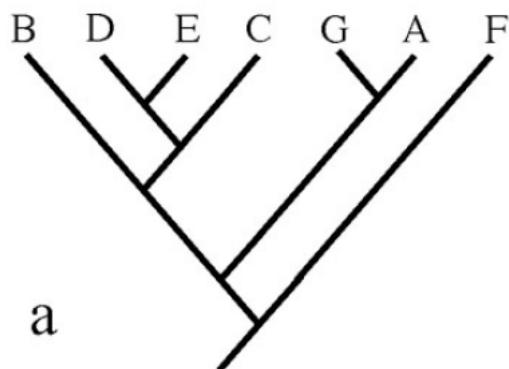
The more recently species share a common ancestor, the more closely related they are

X = the Most Recent Common Ancestor (MRCA)

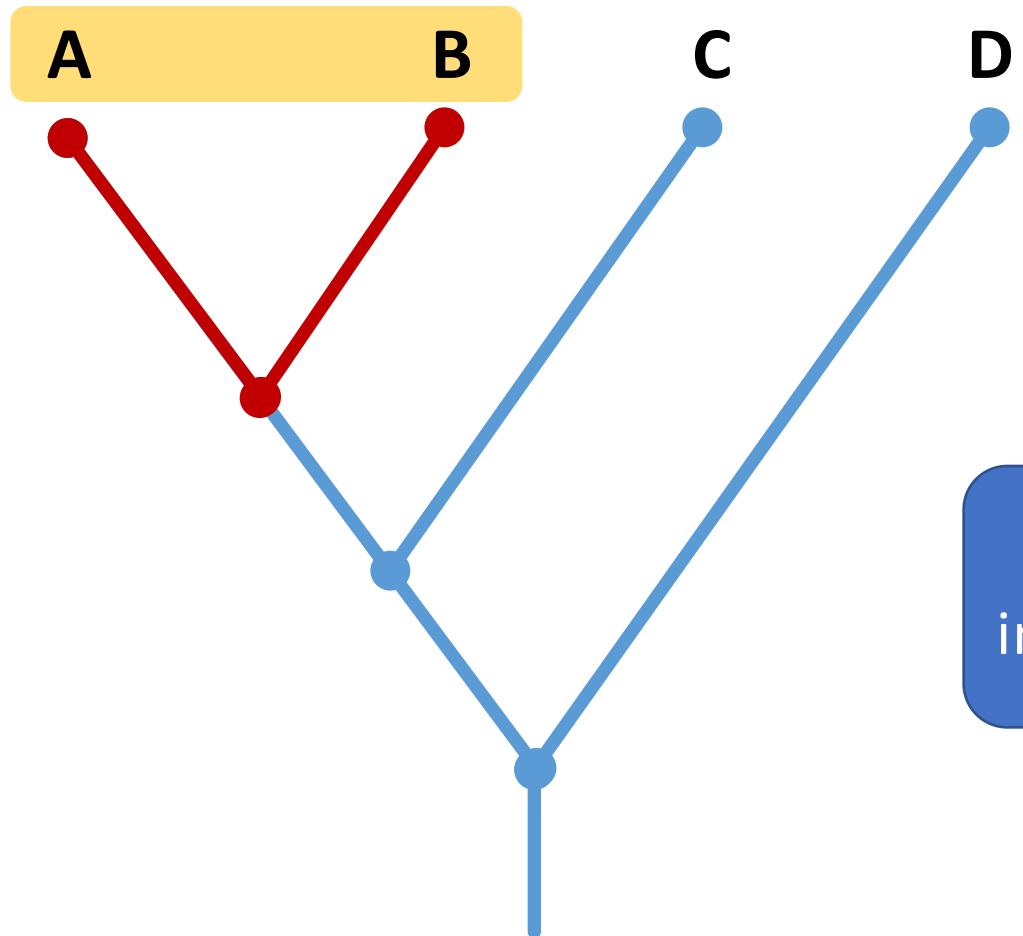


The tree thinking challenge

Which of the trees depicts a pattern different from the others?



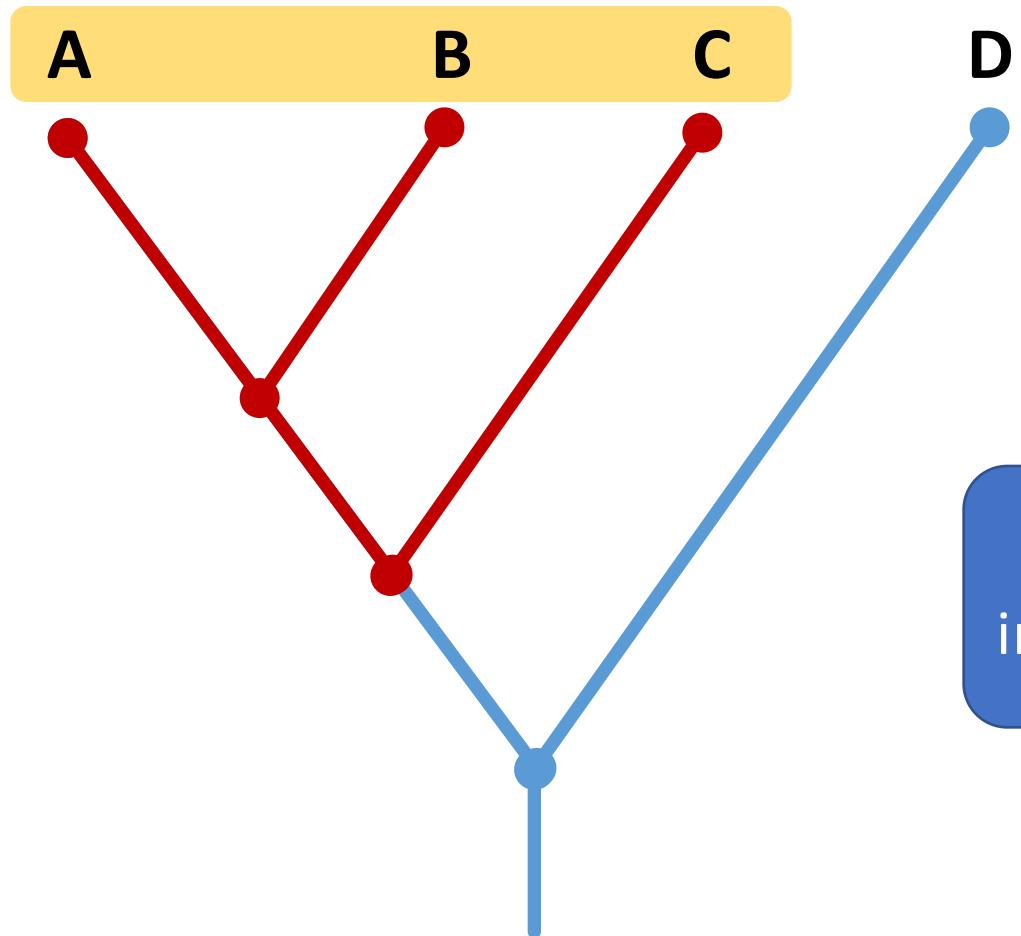
Understanding Tree Topology



Monophyletic group

OTUs share MRCA, which includes all of its descendants

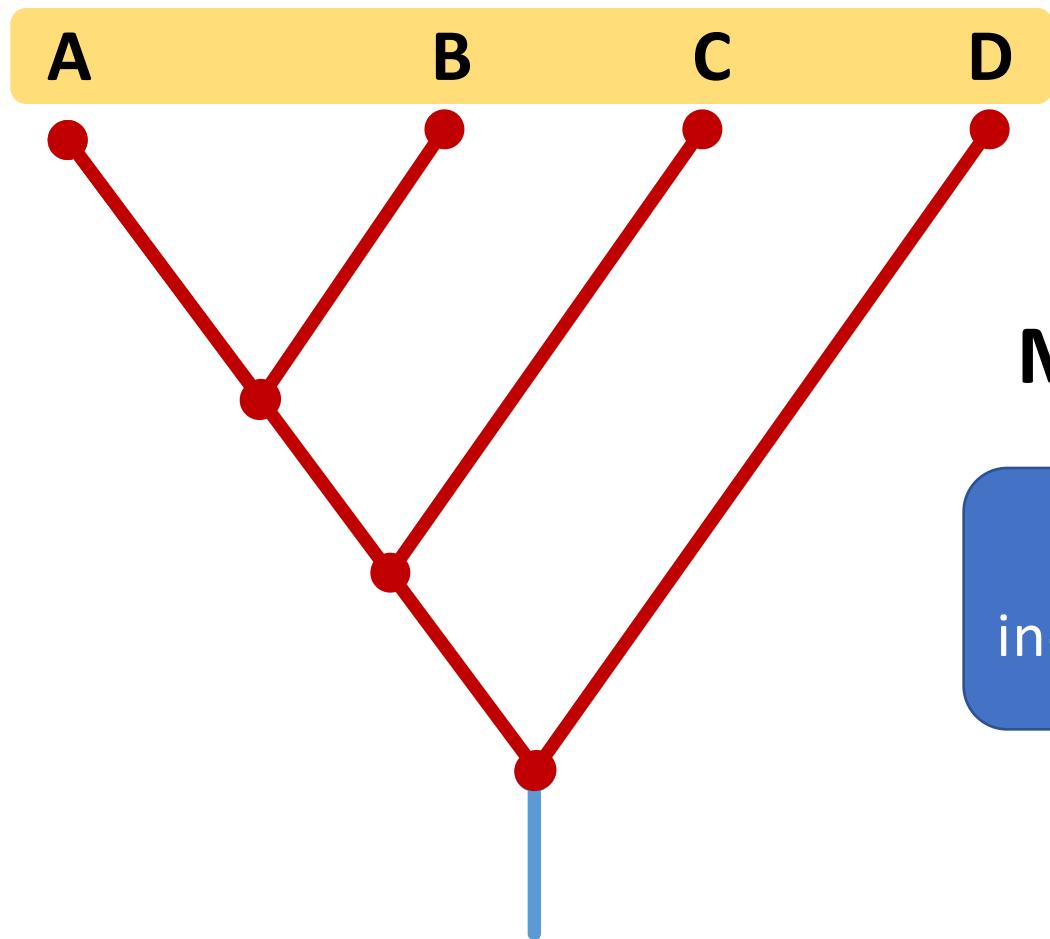
Understanding Tree Topology



Monophyletic group

OTUs share MRCA, which includes all of its descendants

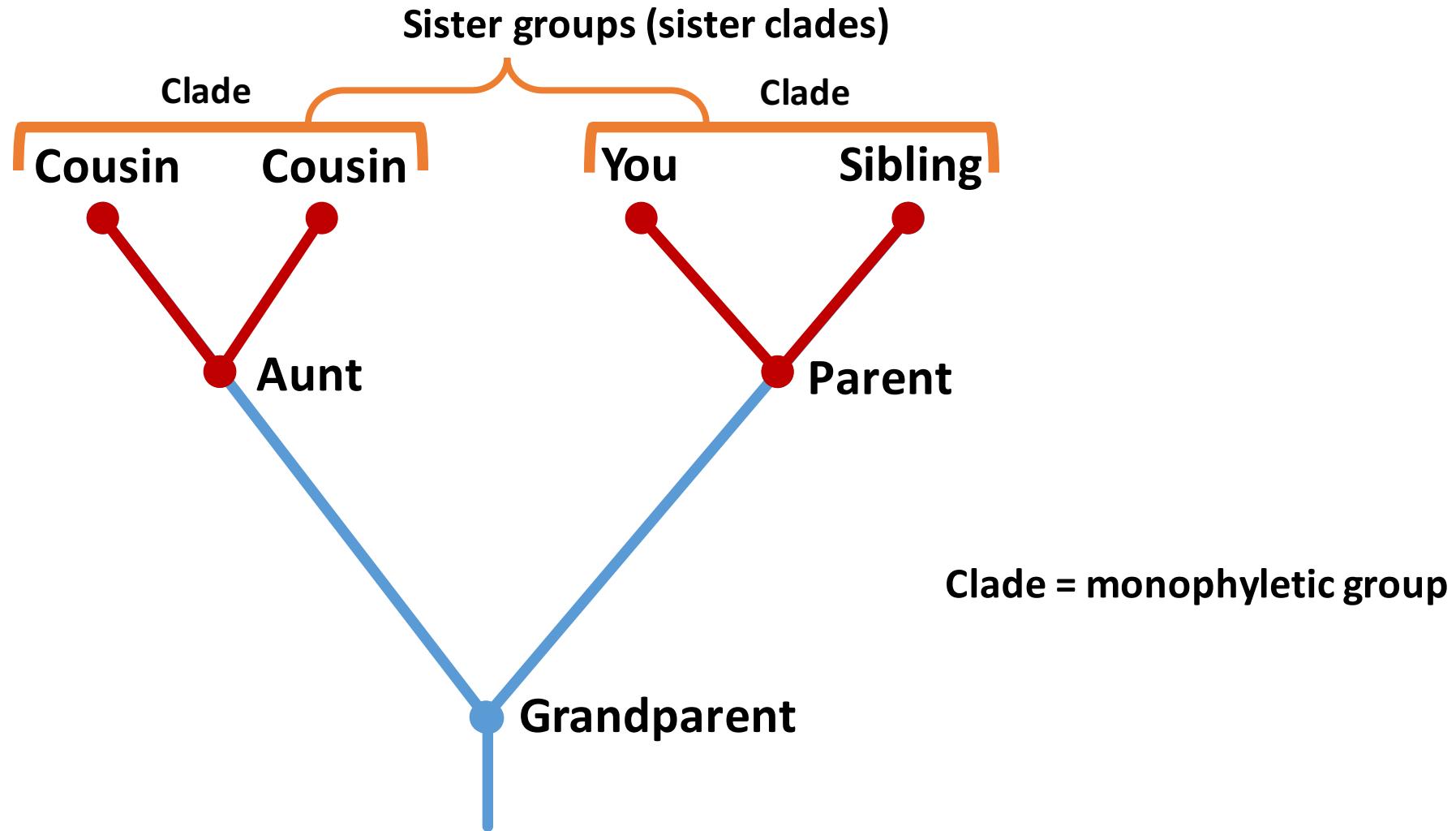
Understanding Tree Topology



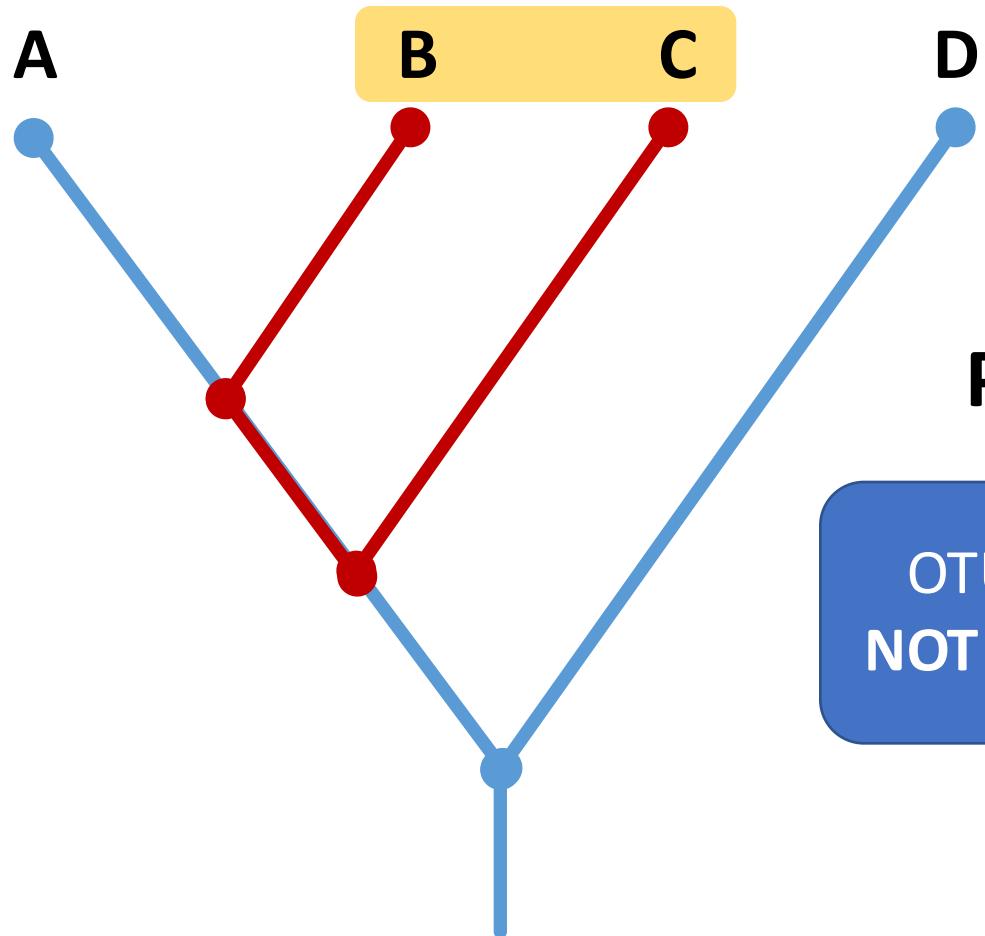
Monophyletic group

OTUs share MRCA, which includes all of its descendants

Understanding Tree Topology



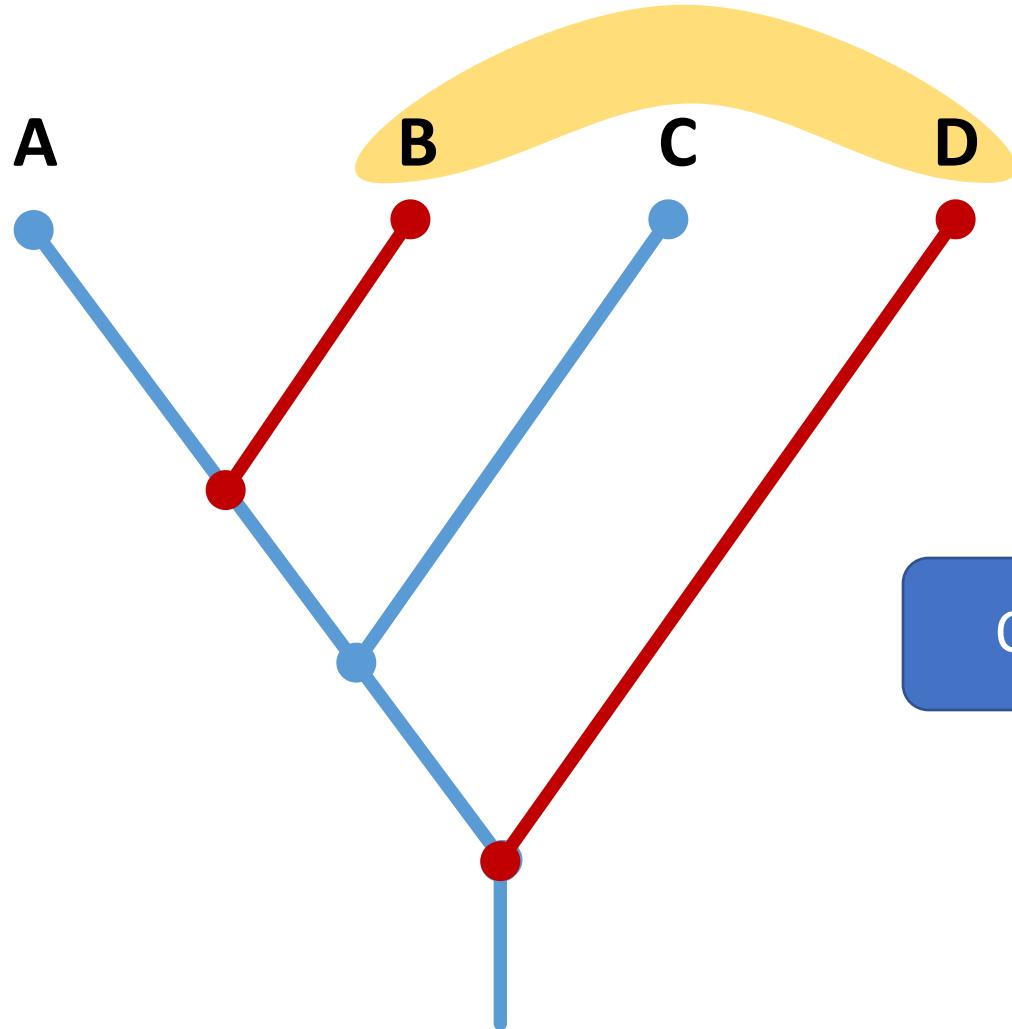
Understanding Tree Topology



Paraphyletic group

OTUs share MRCA, which does
NOT include all of its descendants

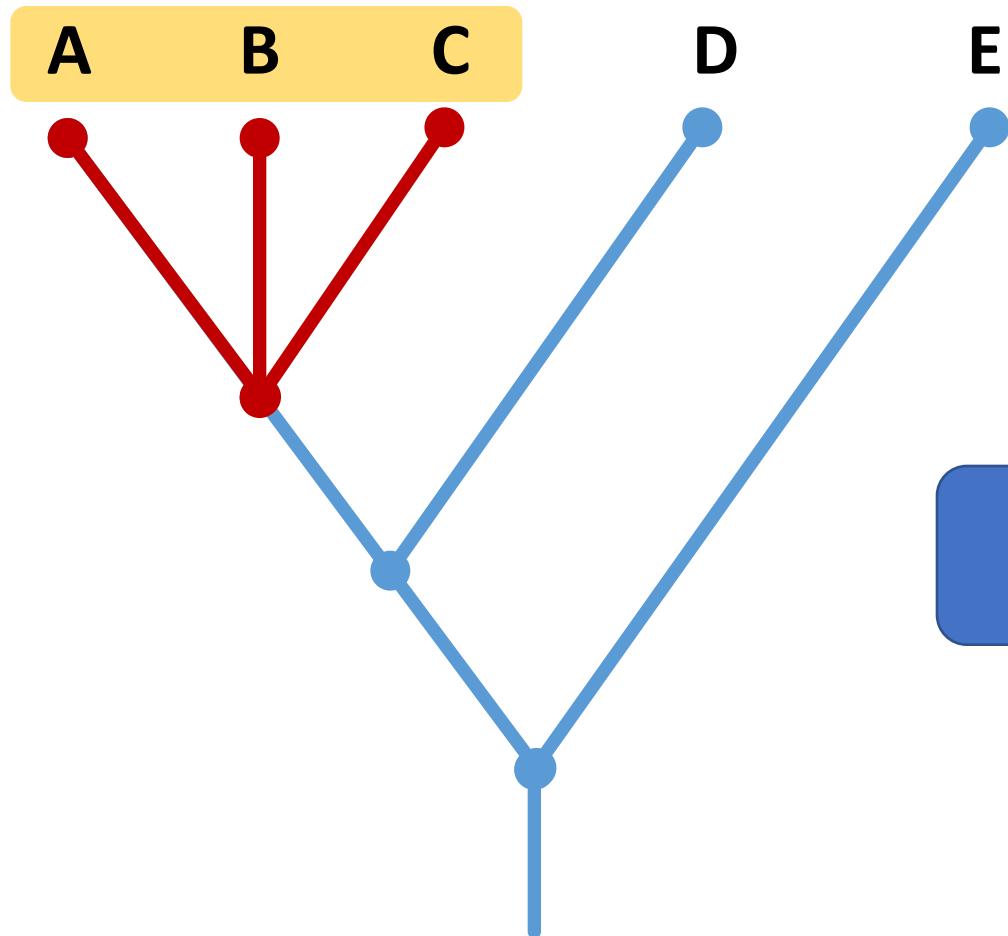
Understanding Tree Topology



Polyphyletic group

OTUs do NOT share MRCA

Understanding Tree Topology

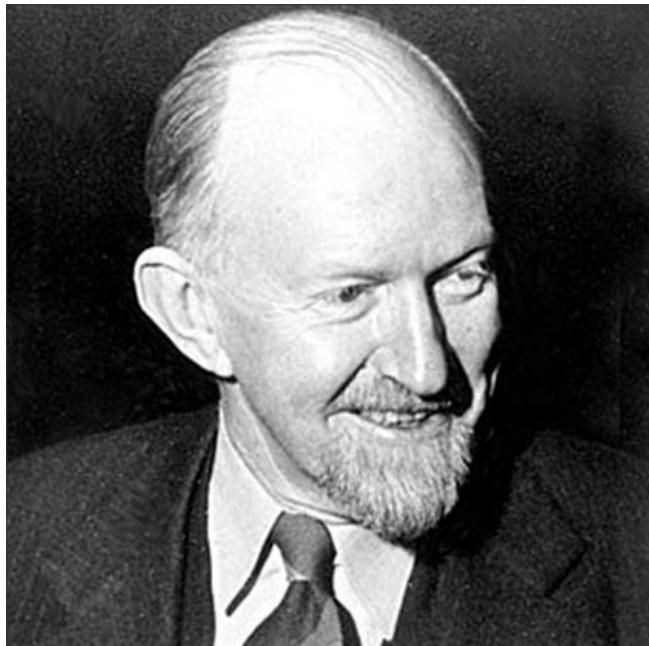


Polytomy

shows ambiguities in
relationships



Rise of Evolutionary Systematics (subjective approach)



George Simpson (1902-1984)



Ernst Mayr (1904-2005)

Organisms must be classified based on their evolutionary relationships rather than their morphological similarity... But how?

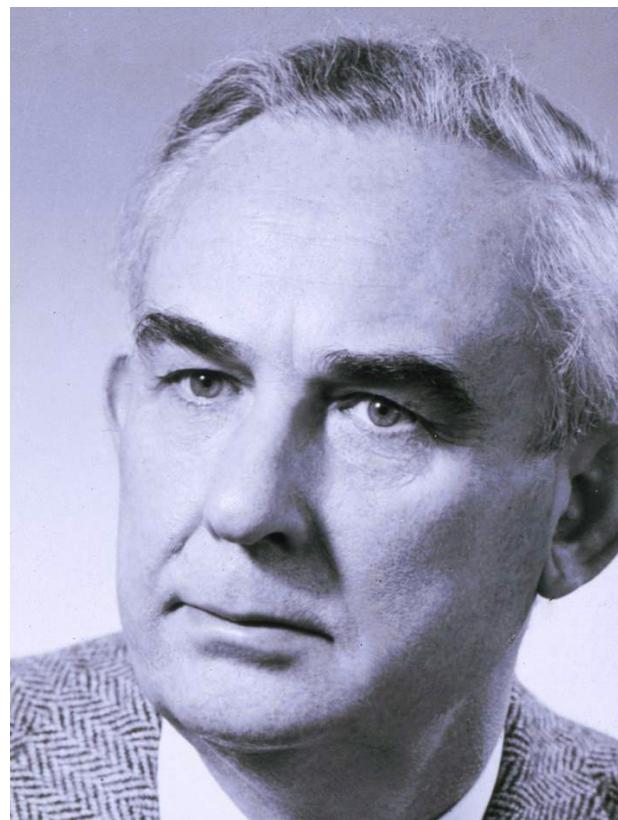


Phylogenetic Systematics (objective approach)

Hennig proposed the idea that evolutionary relationships between organisms must be revealed by using some objective analytical approach. He introduced the concepts of common ancestry and shared derived characters (synapomorphies) that were then implemented in **cladistics** analysis.

“Apomorphic” – derived or specialized character

“Plesiomorphic” – ancestral character



Willi Hennig (1913-1976)



Phylogenetic Systematics



Cladistics

Cladistics is an approach in biological classification in

which organisms are grouped in taxa based

on their **synapomorphies**; those are shared derived

characters within a monophyletic group



THE BIGGEST CHALLENGE:

How do we know which characters
to compare for shared ancestry?

= for reconstructing phylogenetic trees?

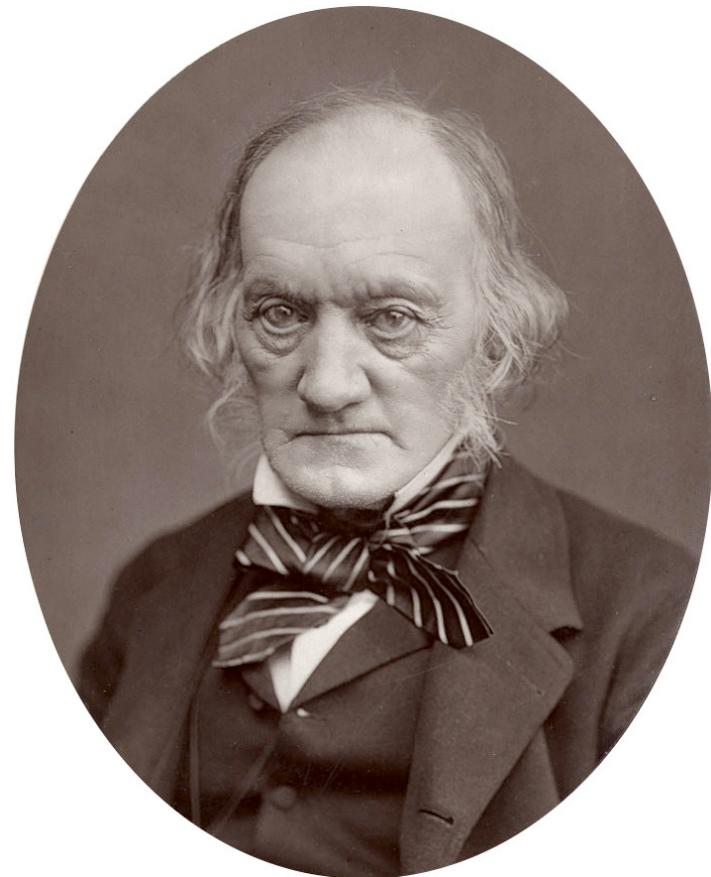


Homology

“the same organ in different animals under every variety of form and function”

Analogous characters or genes have similar structure or function but do NOT share a common ancestor.

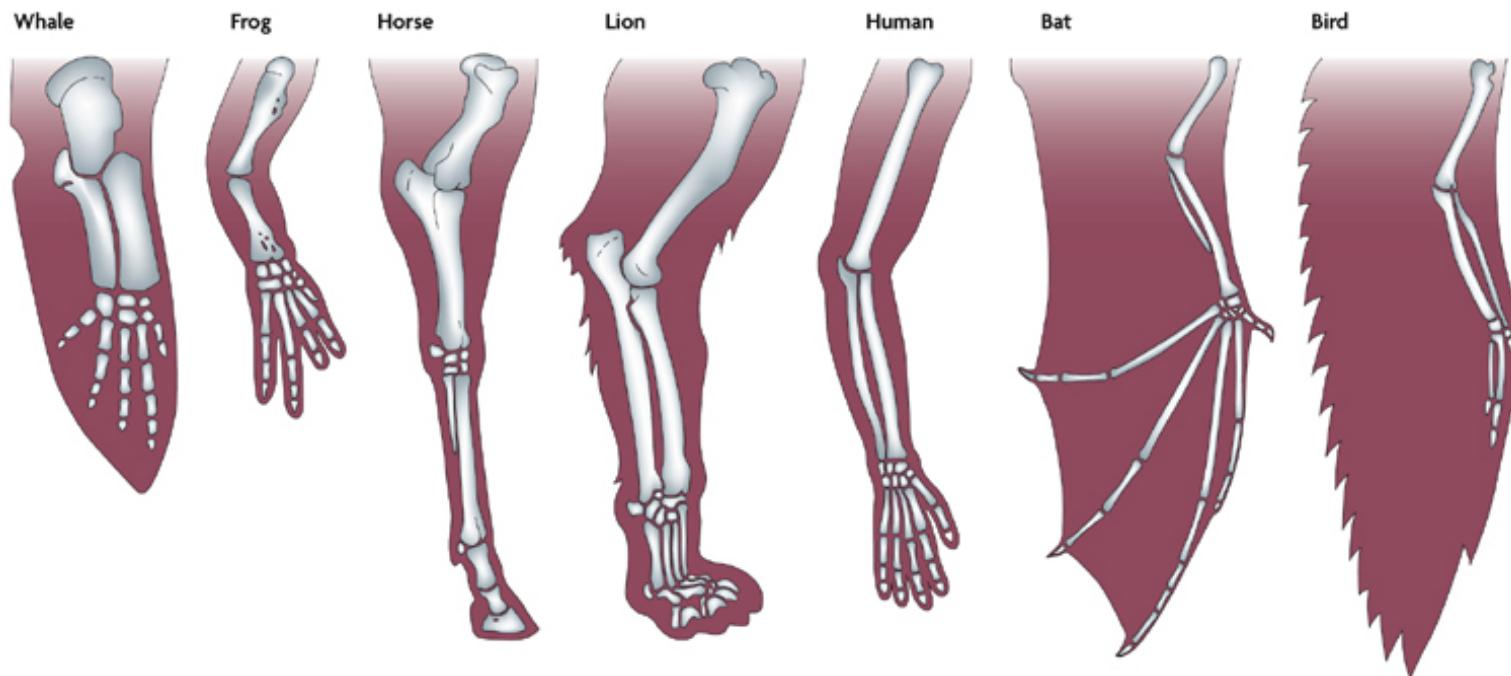
Homologous characters or genes are descending from a common ancestor.



Richard Owen (1804-1892)

Homology of forelimbs

Homologous characters can have different shapes and functions



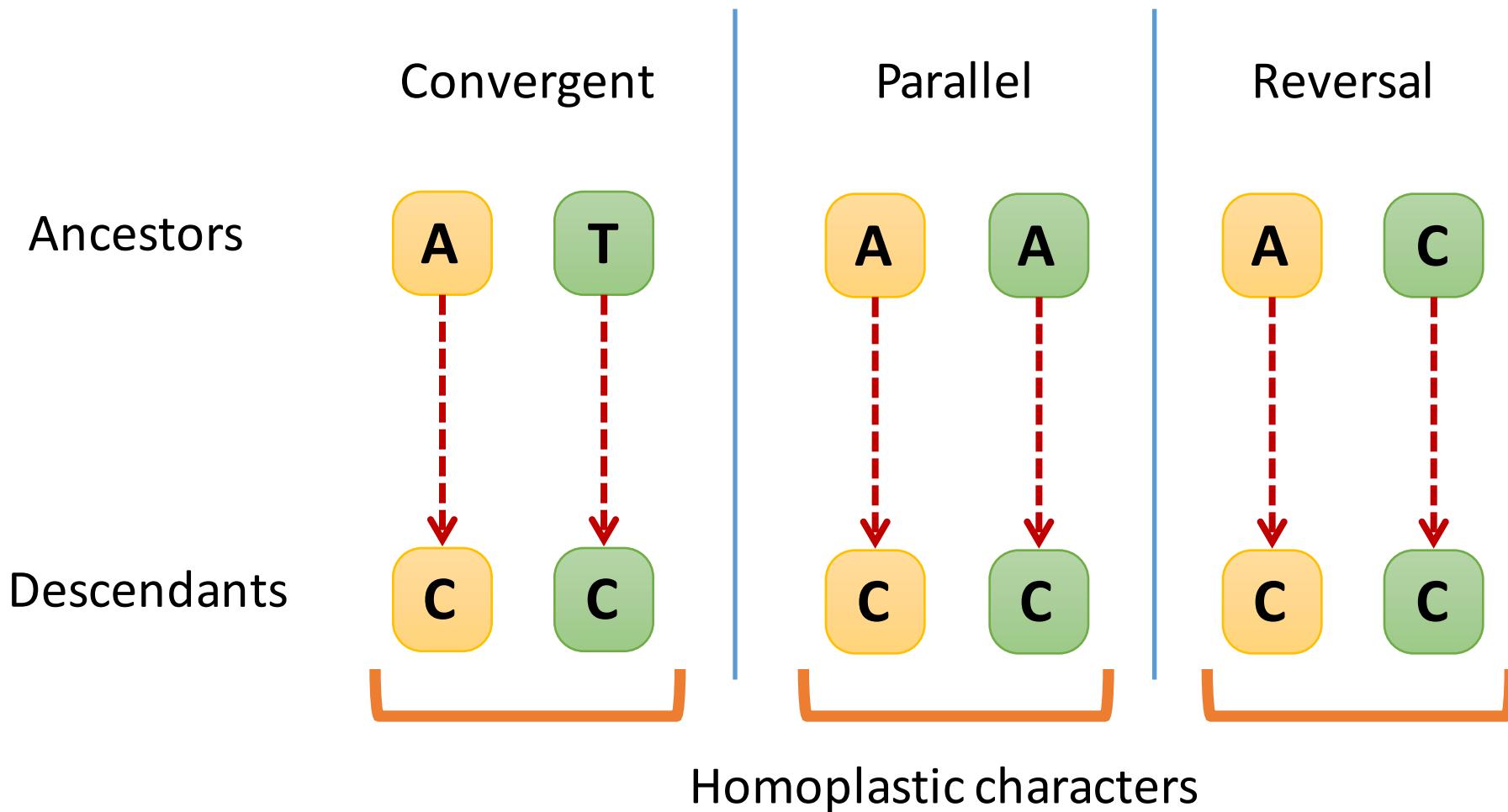


Analogous characters look or perform very similar but derived independently (do not share MRCA)



Convergent evolution –
independent evolution
of similar characters in
taxonomic groups from
different lineages

Homoplasy (= analogy) in nucleotide substitutions



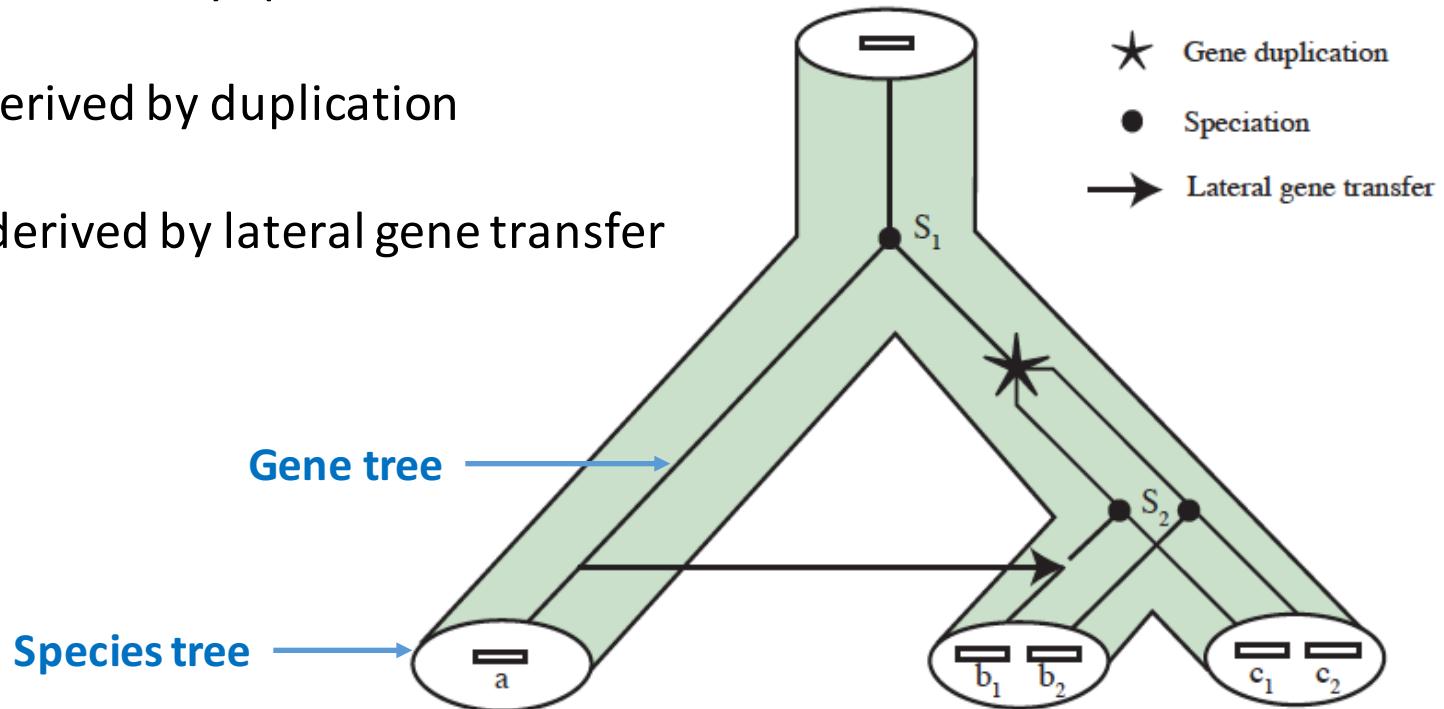


Homologous genes can be classified into:

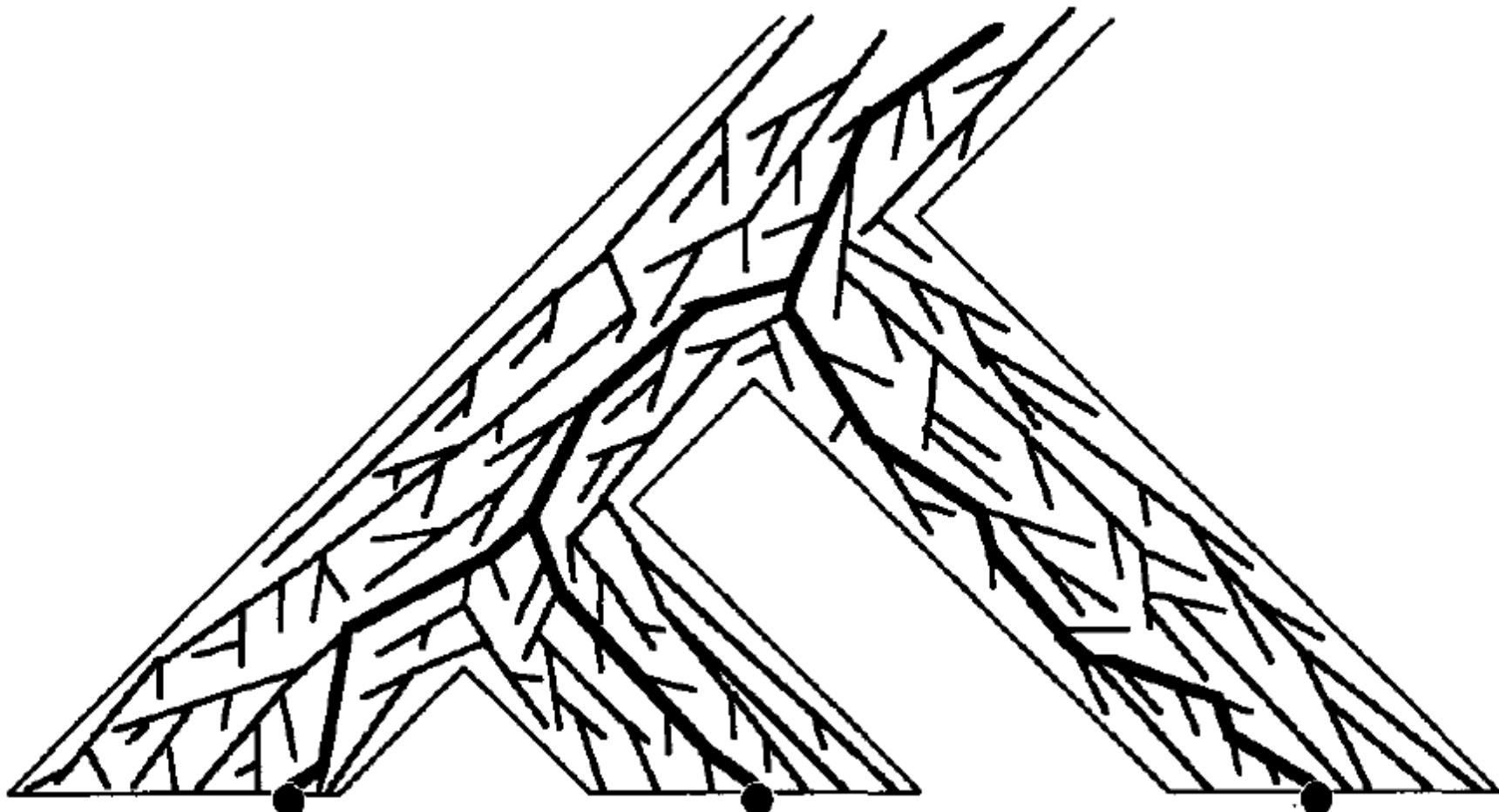
Orthologs – derived by speciation

Paralogs – derived by duplication

Xenologs – derived by lateral gene transfer



Genes trees contained within a species tree





Molecular markers for phylogenetic inference

- Must represent homologous sequences
- Orthologs can be inferred by sequence similarity (alignment score or a graph distance) and by gene vs. species tree reconciliation
- Paralogs are good for reconstructing gene trees but bad for reconstructing species trees, and therefore, must be avoided



THE BIGGEST CHALLENGE:

How do we know which ~~characters~~
~~to compare for shared ancestry?~~

are orthologous?

genes



How similar must two sequences be in order to be considered homologous?

PROTEINS: > 25 % of the amino acids

MUST BE SIMILAR*

DNA & RNA: > 70 % of the nucleotides

* Assumption usually applies when sequence length is more than 100 amino acids or nucleotides



Large scale orthology inference software

Phylogeny based methods:

- NOTUNG (Chen et al., 2000)
- Orthostrapper (Storm & Sonnhammer, 2002)
- LOFT (Heijden et al., 2007)
- Ensemble (Flicek et al., 2008)

Pairwise based methods:

- OrthoMCL (Li et al., 2003)
- Roundup (Deluca et al., 2006)
- MultiParanoid (Alexeyenko et al., 2006)
- OMA (Roth et al., 2008)
- InParanoid 7 (Ostlund et al., 2010)
- Ortholog-Finder (Horiike et al., 2016)

Graph based methods:

- OrthoMCL-DB (Chen et al., 2006)
- COG (Tatusov et al., 2000)
- eggNOG (Jensen et al., 2008)



Online databases for identifying orthologous genes



[Orthologous Group Search](#)
[Query Orthology Classification](#)
[About OrthologID](#)

Welcome to OrthologID Online

OrthologID automates gene orthology determination within a character-based phylogenetic framework.

OrthologID Online identifies orthologous groups for complete genomes compiled in our [database](#) (*Orthologous Group Search*), and classifies user-input query sequences into orthologous groups generated from complete genomes (*Query Orthology Classification*). It identifies diagnostic characters that define each orthologous group, as well as diagnostic characters responsible for classifying query sequences. The output is presented in phylogenetic tree format.

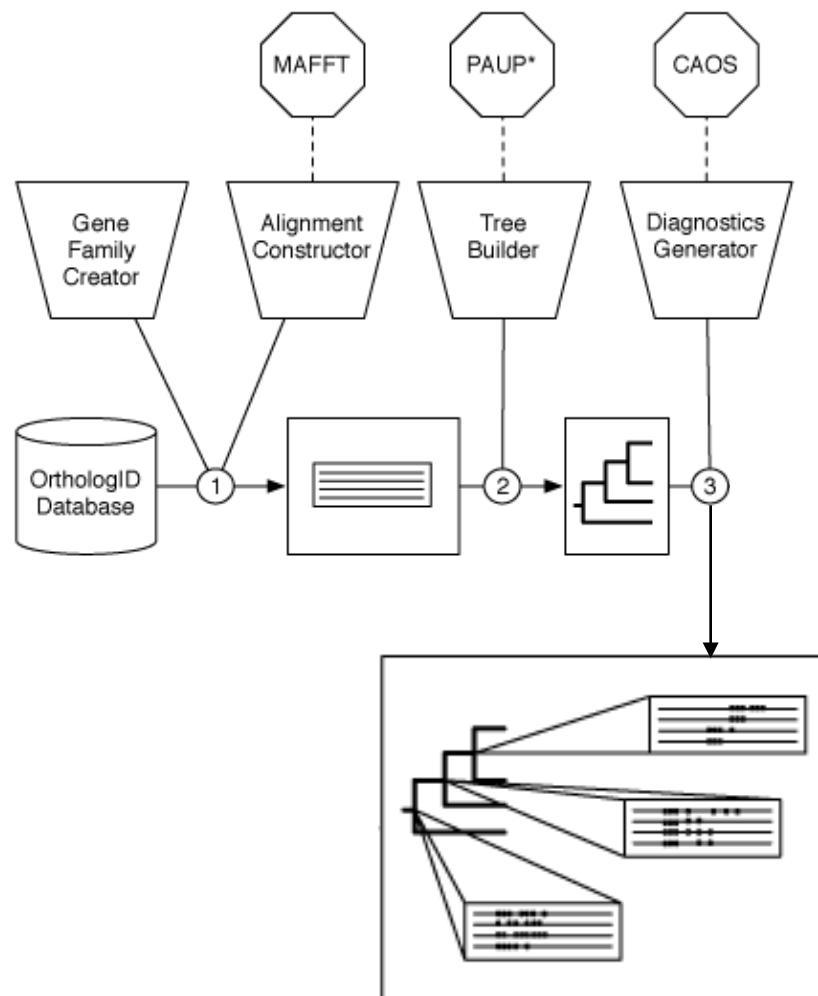
Joanna C. Chiu, Ernest K. Lee, Mary G. Egan, Indra Neil Sarkar, Gloria M. Coruzzi, and Rob DeSalle. OrthologID: automation of genome-scale ortholog identification within a parsimony framework. *Bioinformatics* (2006) 22(6): 699-707 first published online January 12, 2006 doi:10.1093/bioinformatics/btk040

This site uses an early version of the OrthologID Plant database and pipeline. To install the latest version of the OrthologID pipeline on your own cluster, follow the instructions [here](#) to download the source code from our Mercurial repository.

<http://nypg.bio.nyu.edu/orthologid>

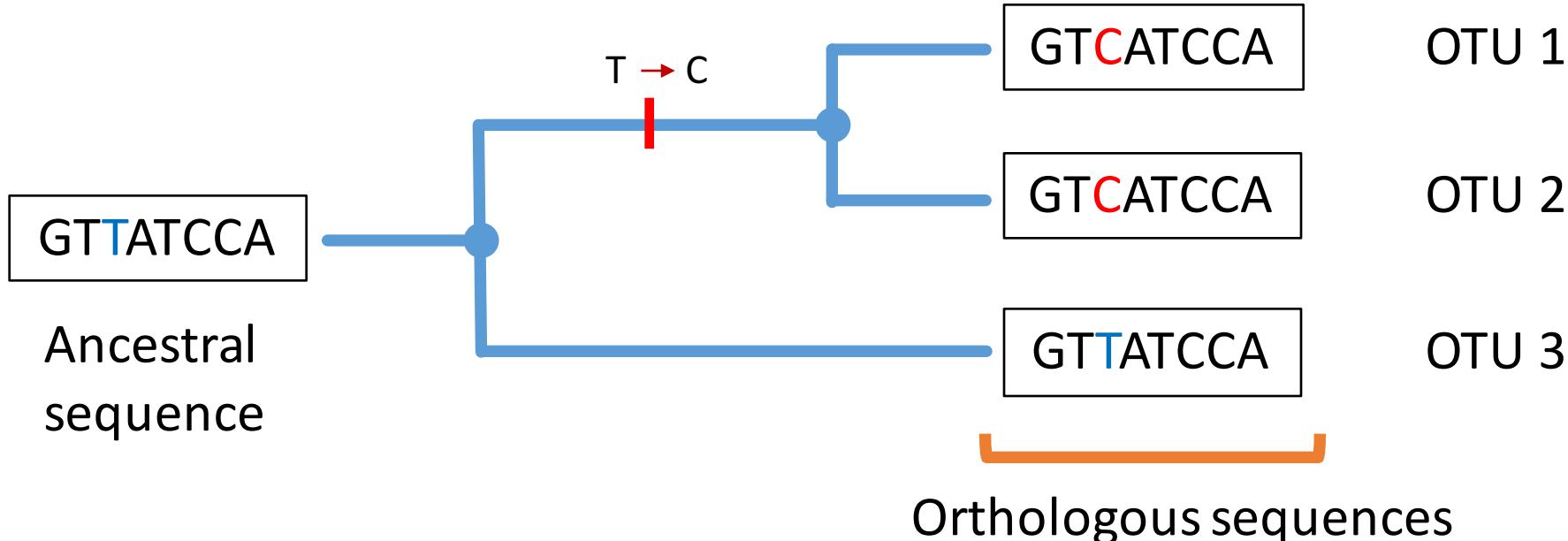
<http://www.orthodb.org>

<http://pantherdb.org/genes/>





Sequence alignment is an important step in phylogenetic analysis

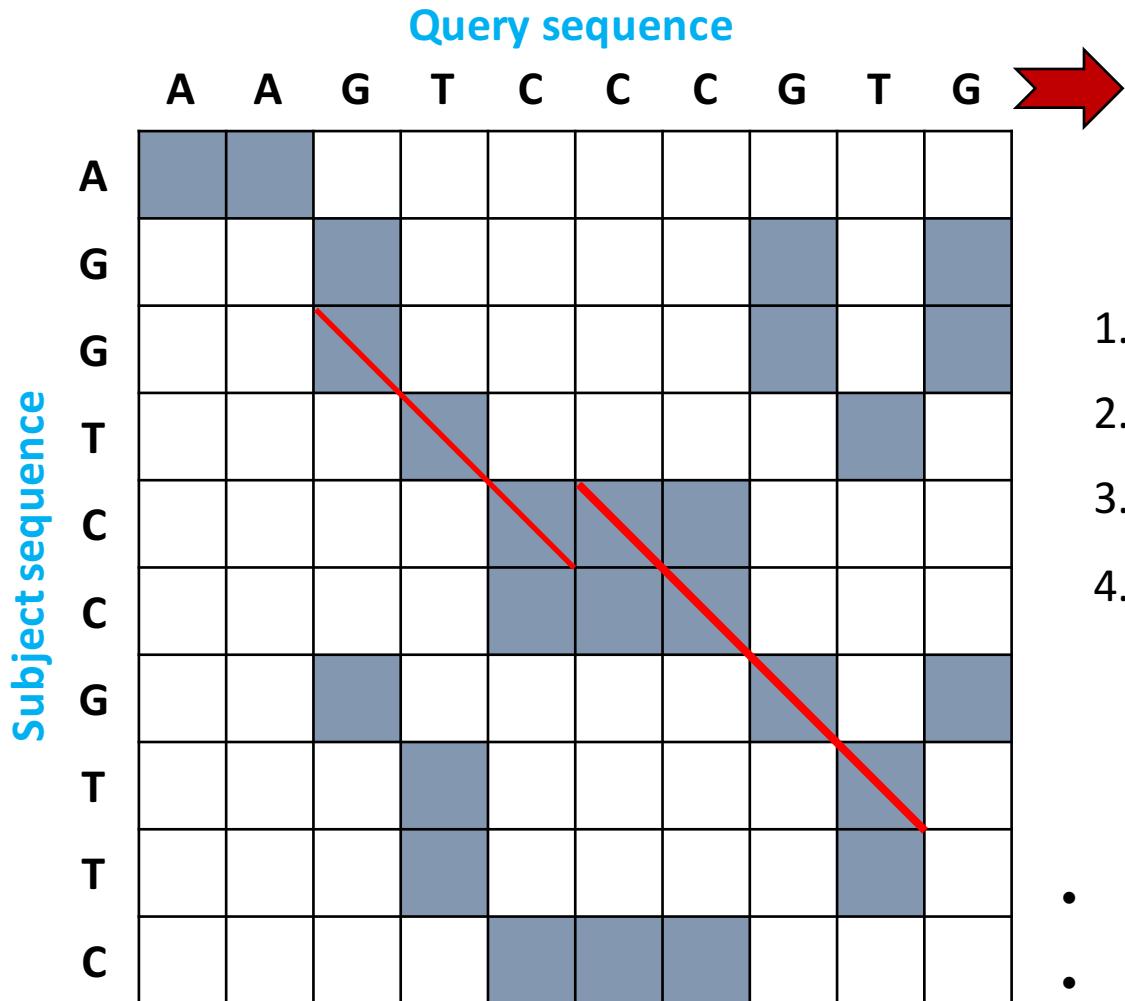




Important applications of sequence alignment:

- Prediction of function (e.g., protein annotation)
- Database searching (e.g., barcode gene id)
- Gene finding (e.g., finding gene within a genome)
- Sequence divergence (e.g., homology test)
- Sequence assembly (e.g., Illumina reads)

Dot-plot alignment



Algorithm

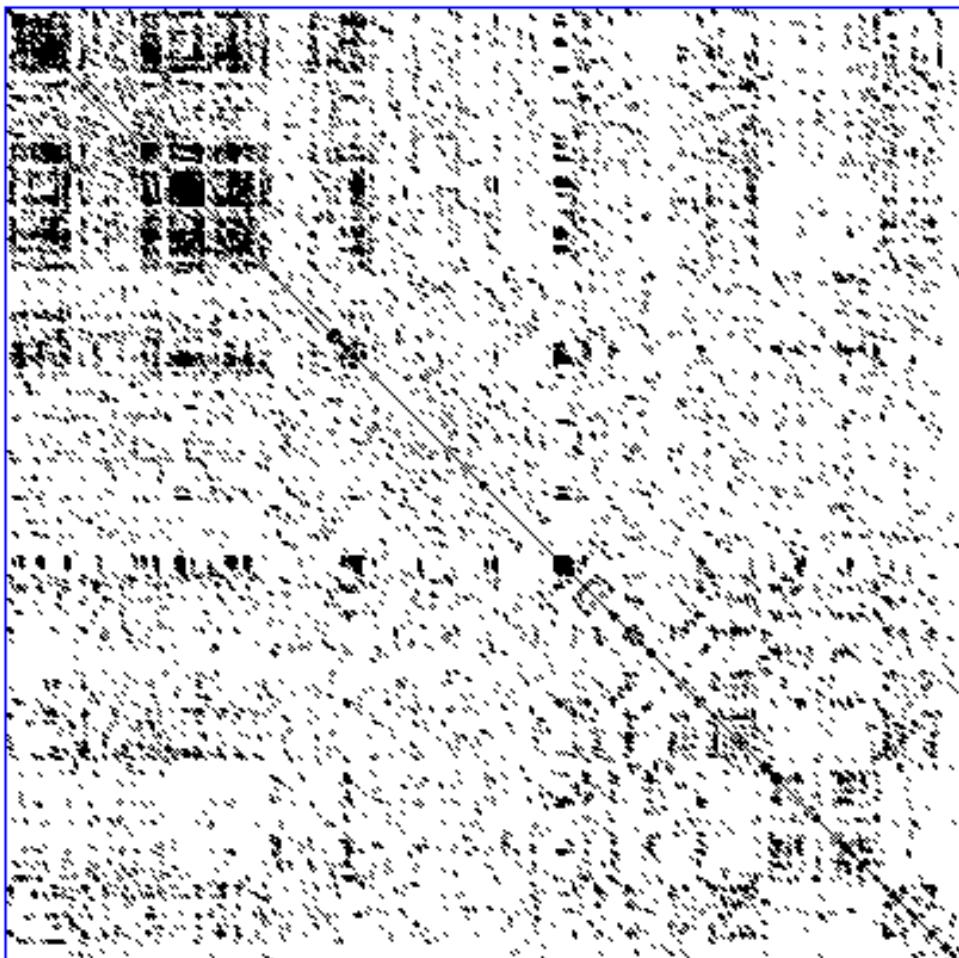
1. Query sequence moves right
2. Mark down all residual matches
3. Count matches on a diagonal
4. Choose the alignment with the largest score

Parameters

- Minimal diagonal length (window)
- Minimal # of matches (stringency)



Dot-plot alignment



Dot-plots are **NOT** predictive of homology. They are solely used to highlight similarity between a pair of sequences.

Self-similarity of a human zinc finger transcription factor



Dynamic alignment methods: Global vs. Local

Global (Needleman-Wunsch algorithm):
sequences are aligned over their entire lengths

```
--T--CC-C-AGT--TATGT-CAGGGGACACG-A-GCATGCAGA-GAC
 | | | | | | | | | | | | | | | | | | | | | | | | | |
AATTGCCGCC-GTCGT-T-TTCAG---CA-GTTATG-T-CAGAT--C
```

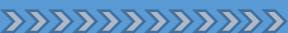
Local (Smith-Waterman algorithm):
only the most similar portions are aligned

```
TCCCAGTTATGTCAGGGACACGAGCATGCAGAGAC
 | | | | | | | | | |
AATTGCCGCCGTCGTTTCAGCAGTTATGTCAGATC
```



When to use pairwise alignments?

Method	Usage
Dot Plot	<ul style="list-style-type: none">General sequence explorationSearch for repeatsFinding indels
Local	<ul style="list-style-type: none">Comparing sequences with partial homologyMaking high-quality alignmentsResidue-per-residue analysis
Global	<ul style="list-style-type: none">Comparing sequences over their entire lengthIdentifying long indels and intronsCommonly used in phylogenetics/phylogenomics



Some popular sequence alignment software

Program	Alignment	Algorithm	Accuracy	Speed	Link
BLAST	local	heuristic	medium	slow	blast.ncbi.nlm.nih.gov
BLAT	local	heuristic	medium	medium	www.kentinformatics.com
Needle	global	heuristic	medium	fast	emboss.toulouse.inra.fr
FastA	local	heuristic	high	slow	fasta.bioch.virginia.edu
HMMER	both	Consistency-based	high	slow	hmmer.org
MAFFT	global	Iterative	medium	fast	mafft.cbrc.jp
MUSCLE	global	Iterative	medium	fast	www.drive5.com/muscle
SSEARCH	local	heuristic	high	very slow	www.biology.wustl.edu/gcg/search.html
T-coffee	both	Structure-based Consistency-based	high	slow	www.tcoffee.org
USEARCH	global	heuristic	low	very fast	www.drive5.com/usearch

Online DATABASES for sequence data:

NCBI: National Center for Biotechnology Information (aka 'GeneBank')

NCBI Resources How To Sign in to NCBI

All Databases Search

NCBI Home

Resource List (A-Z)

All Resources

Chemicals & Bioassays

Data & Software

DNA & RNA

Domains & Structures

Genes & Expression

Genetics & Medicine

Genomes & Maps

Homology

Literature

Proteins

Sequence Analysis

Taxonomy

Training & Tutorials

Variation

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[About the NCBI](#) | [Mission](#) | [Organization](#) | [NCBI News](#) | [Blog](#)

Submit
Deposit data or manuscripts into NCBI databases

Download
Transfer NCBI data to your computer

Learn
Find help documents, attend a class or watch a tutorial

Develop
Use NCBI APIs and code libraries to build applications

Analyze
Identify an NCBI tool for your data analysis task

Research
Explore NCBI research and collaborative projects

Popular Resources

PubMed

Bookshelf

PubMed Central

PubMed Health

BLAST

Nucleotide

Genome

SNP

Gene

Protein

PubChem

NCBI Announcements

VAST+ update provides refined alignments 23 Aug 2016

The new version of VAST+ provides a refined structure-based alignment of

September 12th class at NLM: EDirect - Command Line Access to NCBI's Biomolecular Databases 22 Aug 2016



Online DATABASES for sequence data:

ENSEMBL: vertebrate genome browser

Ensembl BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors

Search: for Go

e.g. BRCA2 or rat 5:62797383-63627669 or rs699 or coronary heart disease

Browse a Genome

Ensembl is a genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation. Ensembl annotate genes, computes multiple alignments, predicts regulatory function and collects disease data. Ensembl tools include BLAST, BLAT, BioMart and the Variant Effect Predictor (VEP) for all supported species.

Popular genomes

 Human GRCh38.p7	 Human GRCh37
 Mouse GRCm38.p4	 Zebrafish GRCz10

[★ Log in to customize this list](#)

All genomes

-- Select a species --

[View full list of all Ensembl species](#)

Other species are available in [Ensembl Pre!](#) and [EnsemblGenomes](#)

Did you know...?

Ensembl can deliver free training at your institute. [Find out more](#).

Still using Human GRCh37? [Go to GRCh37](#)

Variant Effect Predictor 

Gene expression in different tissues 

Find SNPs and other variants for my gene

```
GTRATACATTG
CCTAAAGTCTT
CTTCAATTTGT
GAAACATTTC
```

Retrieve gene sequence

```
GCTCTACCTTCGGGTTGC
GGCGTTTGCGGGCGACG
GGCGCTCTGGCGGCGT
AGGGCACAGATTTTGTAG
ACCTCTGGAGCCGGTTT
CCCTTCAGCGGTGGCG
```

Compare genes across species 

Use my own data in Ensembl 

ENCODE data in Ensembl 

Ensembl supports data from external projects through [Track hubs](#) 

What's New in Ensembl Release 85 (July 2016)

- Update to Ensembl-Havana human GENCODE gene set (release 25)
- 30 new epigenomes from the Roadmap Epigenomics Project
- New zebrafish rnaseq
- Update to Rat Ensembl-Havana gene set
- Mouse: update to Ensembl-Havana GENCODE gene set

[Full details](#) | [All web updates, by release](#) | [More news on our blog](#)

- 19 Aug 2016: [Ensembl website maintenance August 23rd](#)
- 19 Aug 2016: [What's coming in Ensembl release 86](#)
- 15 Aug 2016: [GTEX eQTL data now in Ensembl](#)

[Go to Ensembl blog](#)

Tweets by @ensembl

 **Ensembl** @ensembl IGF2BP2, LPL alleles associated w/ lower obesity, #diabetes risk more frequent in #EnduranceAthletes: buff.ly/2aKdt9Q

 **WGC engage** @WGCengage Happy #InternautDay! The worldwide web is crucial in furthering science research and sharing data like @ensembl. Thank you @timberners_lee!

 **Ensembl** @ensembl Embed [View on Twitter](#)



Online DATABASES for sequence data:

EMBL-EBI: The European Bioinformatics Institute

The European Bioinformatics Institute

The home for big data in biology

At EMBL-EBI, we use bioinformatics — the science of storing, sharing and analysing biological data — to help people everywhere understand how living systems work, and what makes them change.

EMBL-EBI
Other EMBL locations >

26-30 August, 2016

Due to planned essential maintenance at one of our datacentres, some EMBL-EBI services may be unavailable or experience degraded performance. We apologise for any inconvenience.

Find a gene, protein or chemical:

Examples: blast, keratin, bfl1...

Explore EMBL-EBI

Services >

Research >

Training >

Industry >

ELIXIR >

Online DATABASES for sequence data:

UniProt (Swiss-Prot & TrEMBL): protein functional information

The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

UniProtKB
UniProt Knowledgebase

- Swiss-Prot (551,705)
Manually annotated and reviewed.
- TrEMBL (65,378,749)
Automatically annotated and not reviewed.

UniRef
Sequence clusters

UniParc
Sequence archive

Proteomes

Supporting data

Literature citations

Cross-ref. databases

Taxonomy

Diseases

Subcellular locations

Keywords

News

Forthcoming changes
Planned changes for UniProt

UniProt release 2016_07
(Bacterial) immigration under control

UniProt release 2016_06
Strength through unity | Removal of the cross-references to NextBio | Change of URIs for neXtProt

UniProt release 2016_05

News archive

Getting started

UniProt data

- Text search
Our basic text search allows you to search all the resources available
- BLAST
Find regions of similarity between your sequences
- Sequence alignments
Align two or more protein sequences using the Clustal Omega program
- Retrieve/ID mapping
This tool merges the "Retrieve" and "ID Mapping" tools
- Download latest release
Get the UniProt data
- Statistics
View Swiss-Prot and TrEMBL statistics
- How to cite us
The UniProt Consortium
- Submit your data
Submit your sequences and annotation updates
- SPARQL
Query UniProt data using a SQL like graph query language

Protein spotlight

Of Plastic And Men
July 2016

Nature has extraordinary resources. Here we are trashing her land, sea and atmosphere - and have been for over a century now - with all sorts of chemistry she didn't ask for and which, sooner or later, will prove to be harmful to those who are putting it there. Despite this, sometimes she manages to find ways of twisting something bad into something good. Polyethylene terephthalate, also known as PET, is one...

www.uniprot.org

Online DATABASES for sequence data:

ExPaSy: Bioinformatics Resource Portal

  ExPaSy Bioinformatics Resource Portal

Home About Contact

Query all databases help

Visual Guidance

Categories

- proteomics
- genomics
- structure analysis
- systems biology
- evolutionary biology
- population genetics
- transcriptomics
- biophysics
- imaging
- IT infrastructure
- medicinal chemistry
- glycomics

Resources A..Z

Links/Documentation

ExPaSy is the SIB Bioinformatics Resource Portal which provides access to scientific databases and software tools (i.e., *resources*) in different areas of life sciences including proteomics, genomics, phylogeny, systems biology, population genetics, transcriptomics etc. (see [Categories](#) in the left menu). On this portal you find resources from many different SIB groups as well as external institutions.

Featuring today

Selectome
Database of positive selection
[\[details\]](#)

NV₁TETTA
NV₁NEVTA
Selectome
NADTEMIA
NOBTRETA

• • • • •

How to use this portal?

- Features and updates
- New to ExPaSy
- Experienced ExPaSy users: what is different

Popular resources

-  UniProtKB
-  SWISS-MODEL
-  STRING
-  PROSITE

Latest News

Protein Spotlight: Of plastic and men - 2016-07-28
Nature has extraordinary resources. Here we are trashing her land, sea and atmosphere - and have been for over a century now - with all sorts of chemistry she didn't ask for and which, sooner or later, will prove to be harmful [More..](#)

UniProtKB Knowledgebase release 2016_07 - 2016-07-06
Release notes
551,705 UniProtKB/Swiss-Prot entries ([More..](#))
65,378,749 UniProtKB/TrEMBL entries ([More..](#))

[More news] [SIB news]

www.expasy.org

Phylogenetics vs Phylogenomics: what's the difference?

	Phylogenetics	Phylogenomics
DATA	<ul style="list-style-type: none">• One or a few genes with known function• Long sequences	<ul style="list-style-type: none">• Genomic data which often includes non-coding regions• Short sequences
Alignment	<ul style="list-style-type: none">• Accurate and relatively fast• Easy to visualize and analyze	<ul style="list-style-type: none">• Not accurate and slow• Hard to visualize and analyze
Phylogenetic inference	<ul style="list-style-type: none">• Partition test• Model test• Model based methods	<ul style="list-style-type: none">• No partition test• No model test• Distance based methods• Simple model based with uniform prior for all genes
Tree visualization	<ul style="list-style-type: none">• Conventional	<ul style="list-style-type: none">• Super-trees• Phylogenetic networks• Cladogram