Mar 10, 2025 Version 8

Guide to molecular databases, phylogenetics and molecular evolution V.8

PLOS One

Peer-reviewed method

DOI
https://dx.doi.org/10.17504/protocols.io.36wgq77e3vk5/v8

Florian G Jacques[1]

[1]Le Mans Université

Lunds University

Florian G Jacques
Masaryk University

OPEN ACCESS

DOI: https://dx.doi.org/10.17504/protocols.io.36wgq77e3vk5/v8

External link: https://doi.org/10.1371/journal.pone.0279597

Protocol Citation: Florian G Jacques 2025. Guide to molecular databases, phylogenetics and molecular evolution. **protocols.io** https://dx.doi.org/10.17504/protocols.io.36wgq77e3vk5/v8Version created by **Florian G Jacques**

Manuscript citation:
Jacques F, Bolivar P, Pietras K, Hammarlund EU (2023)Roadmap to the study of gene and protein phylogeny and evolution—A practical guide. PLOS ONE 18(2): e0279597. https://doi.org/10.1371/journal.pone.0279597

Protocol status: Working
We use this protocol and it's working

Created: February 07, 2025

Last Modified: March 10, 2025

Protocol Integer ID: 119822

Keywords: Evolution, bioinformatics, phylogenetic analysis, evolutionary studies, molecular evolution, Phylogenetic inference, molecular databases, molecular phylogenetic reconstruction, protocol for molecular phylogenetic reconstruction, phylogenetic analysis from molecular data, phylogenetic tree, phylogenetics, calibration of phylogenetic tree, phylogentic analysis, diverse aspects of molecular evolution, molecular database, numerous biological database, molecular evolution, protein database, proteomes of multiple species, annotations of genome, creation of numerous biological database, evolution at the molecular level, molecular evolution development, ancestral state reconstruction, such as ancestral state reconstruction, biodiversity, molecular data, species, multiple species, study of biodiversity, genome, compilation of dna, transcriptome, number of genome, studies into biodiversity, sequencing technology, sequencing, relevant database, store sequence, including maximum parsimony, evolution, maximum parsimony

Abstract

Developments in sequencing technologies and the sequencing of an ever-increasing number of genomes have revolutionised studies into biodiversity and molecular evolution. This accumulation of data has been accompanied by the creation of numerous biological databases that store sequences and annotations of genomes, transcriptomes, and proteomes of multiple species. However, to find relevant databases and perform phylogenetic analysis from molecular data can be challenging for non-bioinformatic users.

Here, we provide 1) a compilation of DNA and protein databases; 2) a protocol for molecular phylogenetic reconstruction using diverse methods, including Maximum Parsimony, Maximum Likelihood and Bayesian Inference, and 3) tools and protocols to reconstruct diverse aspects of molecular evolution from the phylogenetic tree, such as ancestral state reconstruction and time-calibration of phylogenetic trees.

This protocol aims at facilitating the study of biodiversity and evolution at the molecular level for the broad scientific community, especially students and researchers who are not familiar with informatic coding. We also provide tools and methods for more advanced users, especially in phylogentic analysis.

General introduction

1 This is a **step-by-step protocol** to reconstruct molecular (gene and protein) phylogeny and evolution. Based on a recently published up-to-date protocol for phylogenetic analysis (Jermiin *et al.*, 2020), we propose a guide to the study of molecular evolution, from molecular databases to phylogenetic analysis and diverse aspects of molecular evolution. This guide is especially intended to students and researchers who are not familiar with evolutionary biology and bioinformatics. We illustrate our protocol with two test-case studies on the evolution of human p53 (TP53) and Cyclins/CDKs protein families, and provide instructions and a few codes.

This guide includes:

- A brief introduction to phylogenetic analysis presenting concepts and definitions,
- A compilation of molecular (DNA and protein) databases,
- A description of every step of molecular phylogenetic reconstruction, with a list of relevant bioinformatic tools,
- A presentation of diverse evolutionary analyses that can be performed from a phylogenetic tree, including time calibration, ancestral state reconstruction, genome evolution etc., with a list of relevant bioinformatic tools.

Introduction to phylogenetic trees: concepts and definitions
Retracing the evolution of species, genes and proteins, usually requires reconstructing a phylogenetic tree. A **phylogenetic tree** or **phylogeny** is a graphical representation of the evolutionary relationship between taxa or genes/proteins. Mathematically, a phylogenetic tree is a **connected acyclic graph**. The shape of the tree is known as **topology**. The taxa or sequences are positioned at the extremity of the **branches** and are sometimes referred to as "leaves". Groups of related organisms are known as **OTU**s (operational taxonomic units). For instance, genera are OTUs. Internal **nodes**, also referred to as HTUs (hypothetical taxomonic units), indicate the most recent **common ancestors** between taxa. In phylogenetic analysis, ancestors are hypothetical, as nodes never indicate an identified extant or extinct species. A species (*e.g. Archaeopteryx lithographica* or*Tiktaalik rosaea*) is never considered the direct ancestor of a taxonomic group (*e.g.* birds or tetrapods). Furthermore, only one route is possible from an ancestor to a descendant, as horizontal transfers, although common in microorganisms, are usually ignored in phylogenetic analysis. Ideally, every node is followed by a **bipartition**. In this case, the tree is said to be **resolved**. If the evolutionary relationships between some taxa cannot be determined, the tree will include **polytomies**, meaning that the tree is not fully resolved.

The **root** of a phylogeny is the hypothetical common ancestor to all species or sequences in the tree. For instance, *LUCA* (last universal common ancestor), *LECA* (last

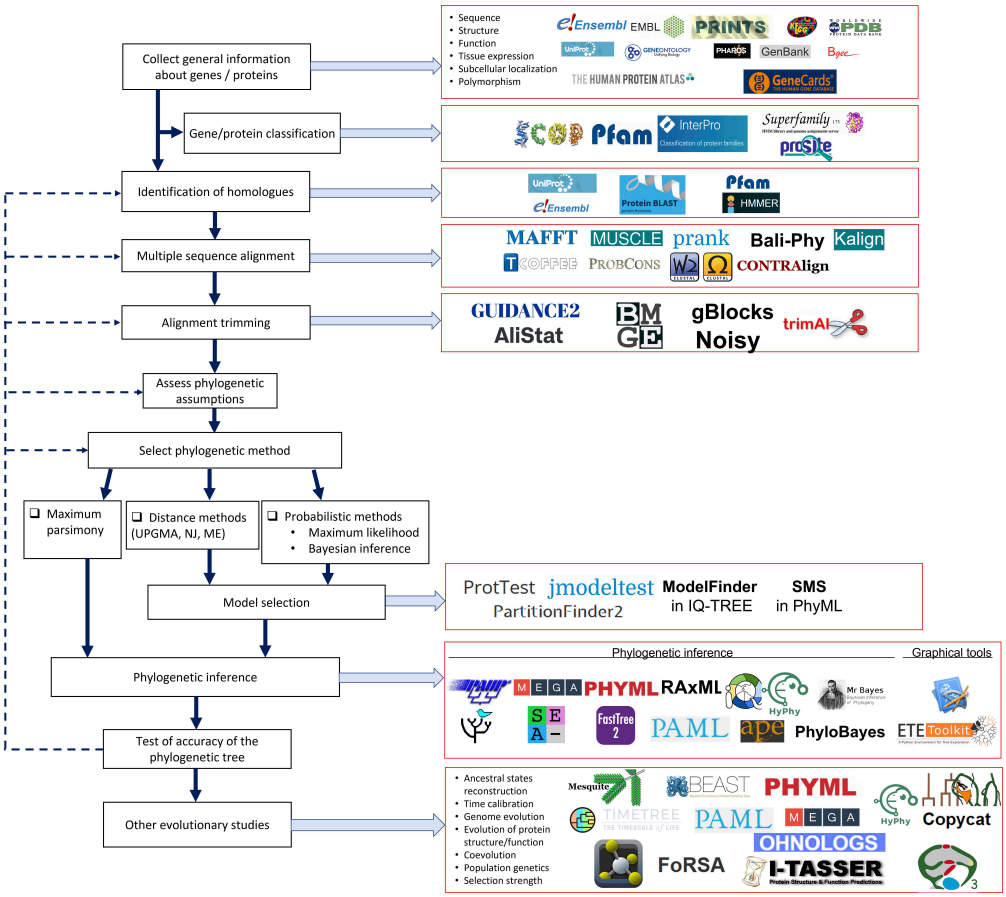
eukaryotic common ancestor), *Urmetazoa* and *Urbilateria* are the hypothetical common ancestors of all living organisms, all eukaryotes, all animals and all bilaterians, respectively. In molecular phylogenies, ancestors are hypothetical ancestral genes or peoteins. Phylogenetic trees can be rooted if the root is indicated, or unrooted if not.

Phylogenetic trees can consider branch lengths (**phylograms**) or not (**cladograms**). In the former, that take into account different evolution speeds, branch lengths can vary. In the latter, all branch lengths are equal.

Modern biological classification is based on phylogenetic trees and species evolutionary relatedness. Nowadays, only groups including an ancestor and all its descendants, known as **monophyletic groups** or **clades**, are considered. **Paraphyletic groups**, that include an ancestor and only one part of its desendants (e.g. protozoa or fish), and **polyphetic groups**, that include several taxa but not their last common ancestor, are excluded in modern classification.

Phylogenic analysis uses **characters** to retrace evolutionary relationships. They can be morphological or molecular. Molecular data are usually preferred on extant species. In molecular phylogenetics, the characters are **nucleotides** (A, T, G, C) or **aminoacids** (the 20 proteinogenic aminoacids) that constitute gene or protein sequences. Characters can be **ancestral** or **derived**, and only the phylogenetic analysis can determine the orientation. Only derived characters define monophyletic groups. Ancestral characters define paraphyletic groups, and they are not relevant for classifications because they are not phylogenetically informative. Characters can also result from convergent evolution, creating **homoplasy**.

Overview of phylogenetic analysis
We describe the protocol from sequence harvesting from databases to phylogenetic tree building and diverse evolutionary studies. Our protocol can be summerized as follows:



Protocol for the bioinformatic study of molecular phylogeny and evolution

I - Sequence collection and comparison

2 **Step1 : Collecting gene / protein sequences**

Most phylogenies are reconstructed based on genes and proteins, but phylogenetics and evolutionary analyses can be performed on virtually all kinds of molecular data, including:

- Genes
- Genomes
- Proteins
- mRNA
- Transposable elements
- Ribosomal RNA
- Other parts of the genome

The first step of phylogenetic analysis is to collect sequence data of genes or proteins. A wide range of public databases store sequences, usually in the Fasta format (*see tables*

below). The most commonly used are **Ensembl**, **EMBL** and **NCBI** for DNA sequences and **Uniprot**, **Interpro** and **Prosite** for protein sequences.

Other information, including protein structure, activity, biological function, tissue expression, sub-cellular localization and polymorphism can also prove relevant for evolutionary studies.

General nucleic acid databases:

- **NCBI**: Collection of databases for molecular biology and medicine, providing tools and services
- **Ensembl**: Genome browser of vertebrates, includes tools for identification of homology
- **Entrez**: Gene sequences and structures
- **GenBank**: Annotated DNA sequences
- **BAR**: plant genes and proteins
- **Bgee**: Gene expression patterns

Nucleic acid databases for human and model species

- **GeneCards**: Human gene function, genomics, transcription factor binding sites and protein products
- **FlyBase**: Genome and proteome of the model insect *D. melanogaster*
- **PomBase**: Genome and proteome of the model yeast *S. pombe*
- **TAIR**: Genome and proteome of the model plant *A. thaliana*
- **WormBase:**Genome and proteome of the model nematode *C. elegans*
- **Xenbase**: Genome and proteome of the model amphibian *X. laevis*

Protein databases

- **CATH**: Classification of protein domains based on their structure,
- functionality, and evolution
- **FSSP**: Classification of protein domains based on their structural similarity
- **Gene Ontology**: Unified annotation of molecular function, biological processes, and cellular components of proteins
- **InterPro**: Classification of proteins domains and functional sites
- **KEGG**: Protein function and biological pathways
- **PDB**: 3D structures of proteins
- **Pfam**: Information about protein families and domains
- **PHAROS**: Centralizes literature for human proteins
- **PRINTS**: Protein fingerprints classification database
- **PROSITE**: Protein family database
- **SCOP**: Classification of protein domains based on their structure, function and evolution
- **SUPERFAMILY**: Protein structure and functions

- **UniProt**: General information on proteins: sequence, structure, classification, function, subcellular localization and homology

→ To batch download a large number of sequences, users can use the **BioMart** tool of Ensembl (*https://www.ensembl.org/info/data/biomart/index.html*), the **Batch Entrez** tool of NCBI (https://www.ncbi.nlm.nih.gov/sites/batchentrez), and the **Retrieve/ID mapping** tool of UniProt (https://www.uniprot.org/id-mapping). They can also be used to convert and retrieve the identifiers of different databases such as NCBI, GenBank, Ensembl, Uniprot, Pfam or the PDB. **Pfam** also allows to batch download protein sequences in the same genome or different genomes.

Example:
*>Search for your gene or protein of interest (e.g. Human P53) in molecular databases (e.g. **NCBI** or **Uniprot**).*

>In Uniprot, HsTP53 is labelled as P04637 (https://www.uniprot.org/uniprot/P04637). click on "Sequence and Isoforms" *to display the sequence.*

>In NCBI, select "Nucleotide" *in the* "Database" *panel on the left, and type the name or GI number of the gene (https://www.ncbi.nlm.nih.gov/nuccore/?term=Homo+sapiens+P53). Click on* FASTA *to display the sequence.*

>To create a Fasta file with the sequences, paste them in the Fasta format, including the headlines, in Notepad. *Save the document using* fasta *as filename extention.*

The **Fasta format** includes a headline starting with ">", and the nucleic acid or amino-acid sequence. For example, in the case of the human P53 protein:

>sp|P04637|P53_HUMAN Cellular tumor antigen p53 OS=Homo sapiens OX=9606 GN=TP53 PE=1 SV=4
MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIEQWFTEDPGP
DEAPRMPEAAPPVAPAPAAPTPAAPAPAPSWPLSSSVPSQKTYQGSYGFRLGFLHSGTAK
SVTCTYSPALNKMFCQLAKTCPVQLWVDSTPPPGTRVRAMAIYKQSQHMTEVVRRCPHHE
RCSDSDGLAPPQHLIRVEGNLRVEYLDDRNTFRHSVVVPYEPPEVGSDCTTIHYNYMCNS
SCMGGMNRRPILTIITLEDSSGNLLGRNSFEVRVCACPGRDRRTEEENLRKKGEPHHELP
PGSTKRALPNNTSSSPQPKKKPLDGEYFTLQIRGRERFEMFRELNEALELKDAQAGKEPG
GSRAHSSHLKSKKGQSTSRHKKLMFKTEGPDSD

2.1 **Protein domain classification** (optional)

Protein classification can be relevant for evolutionary studies. Protein domains are classified into different categories based on 3-dimensional structure, function, and evolutionary relationship. Identifying the main domains of a protein can provide valuable

insight on its occurrence in living organisms and evolutionary origin. Several classification systems are published and listed in the table above.

*>Use a classification system (e.g. **Pfam** or **Interpro**) to identify the main domains of a protein and study their occurrence in living organisms. Pfam presents their occurrence as a sunburst plot.*

According to Pfam 35.0, HsTP53 contains four main protein domains:
- P53 TAD (transactivating domain)
- TAD2
- P53 DNA binding domain
- P53 tetramer.

P63 and p73 also contain the P53 DNA binding domain and the P53 tetramer domain. The P53 TAD and TAD2 domains are absent in P63 and P73, but both include a single SAM_2 domain instead.

P53 (PF00870 in Pfam) is the main domain of the p53 protein, covering the amino acids 99 to 289. Pfam contains 1765 P53-domain-contaning sequences from 382 species, all in choano-organisms (metazoans and choanoflagellates), including 5 sequences in choanoflagellates and 13 sequences in the genome of *Homo sapiens* (https://pfam.xfam.org/family/PF00870#tabview=tab7).

P53 TAD and TAD2 are two transcription scaffold domains. Pfam includes 253 sequences containing the P53 TAD domain, in bilaterians only. The domain TAD2 is present in 81 sequences, from primates only.
P53 tetramer serves for the oligomerization of the protein. Pfam includes 1392 sequences, in animals only, containing the domain p53 tetramer.
The SAM 2 (sterile alpha motif) domain is a putative protein interaction domain. More than 20000 sequences containing this domain, in more than 1400 species, are present in Pfam.

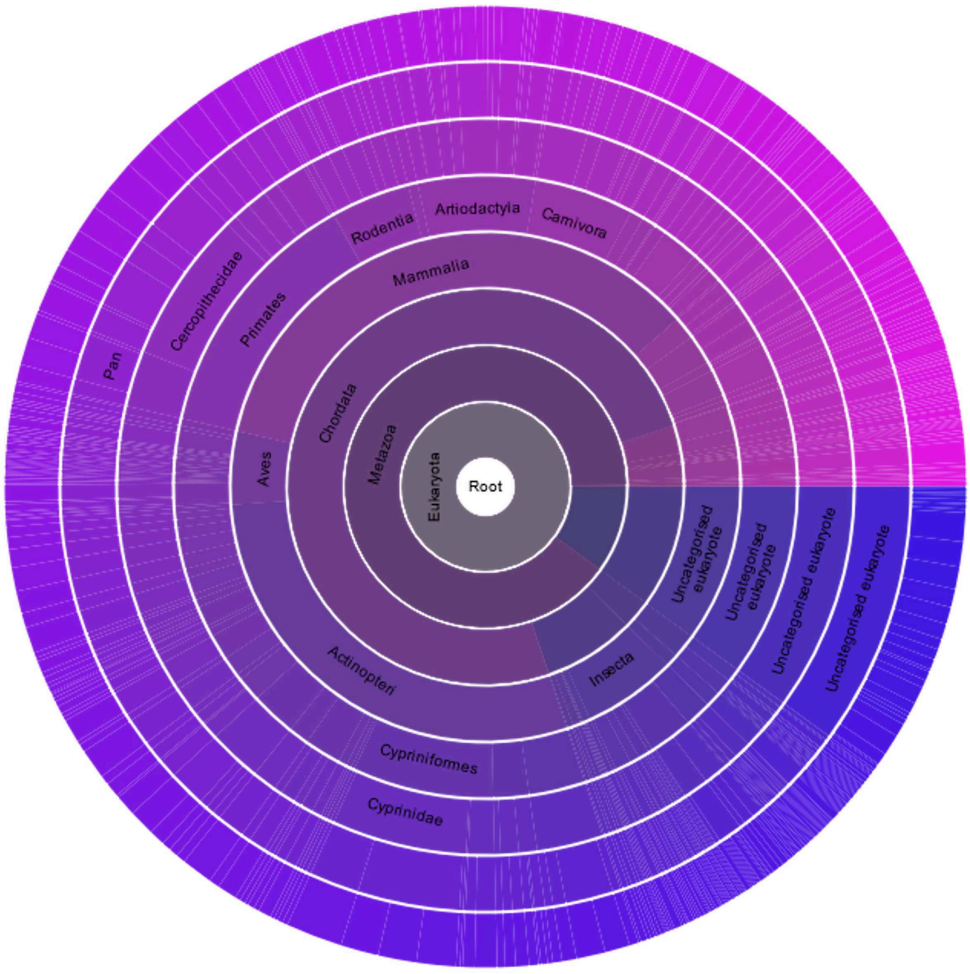
*>Retrieve the classification of the protein domains from SCOP (**https://scop.mrc-lmb.cam.ac.uk/**).*

The SCOP classification of the p53 DNA-binding domain is as follows (accessed September 06, 2021):
- **Class b**: all beta-proteins. This class contains 178 folds.
- **Fold: b.2 Common fold of diphtheria toxin/transcription factors/cytochrome f**. This fold contains 9 superfamilies.
- **Superfamily: b.2.5: p53-like transcription factors**. This superfamily contains 8 families.

- **Family: b.2.5.2: p53 DNA-binding domain-like**. 3 proteins belonging to this family are present in the database.
- **Protein p53 tumor suppressor**. The p53 DNA-binding domains proteins of *Homo sapiens* and *Mus musculus* are present in SCOP.

>Retrieve the sunburst plot of the occurrence of the P53 domains in living organisms from Pfam. Click "sequences" on the right to display the plot.

The plot shows the distribution of the 1,765 sequences containing the P53 binding domain across 382 species. Every bar on the periphery represents one single species, containing one or several p53 paralogues in their genome



Sunburst plot of the distribution of the P53 protein domain (PF00870) in living organisms according to Pfam. This domain is present in virtually all animals, and some of their close relatives, such as choanoflagellates, and suggests that it appeared before the divergence between animals and these protists.

3 Step 2: Identification of homologues

Studying the evolution of genes or proteins requires the identification of **homologues**, *i.e.*, genes or proteins with shared ancestry.

Homology is a central concept in evolutionary biology. It designates genes/proteins or organs that derived from the same ancestral gene or organ (*e.g.* in mammals, forelimbs of quadrupeds, hands of primates, bat wings and cetaceans fins are homologous organs). Homologous genes diverged from a common ancestral gene through an independant accumulation of mutations. The gene sequence and function of two homologues can greatly differ depending on the evolutionary age of the divergence. Genes belonging to the same multigenic superfamily/family (e.g. Hox genes of metazoans) are homologous, sharing the same ancestral hypothetical gene. Homologous genes/proteins usually have similar sequences. For example, proteins are classified in the same family if they share 30% sequence identity.

Homologues include:

- **Orthologues**, present in different species, resulting from speciation event (*e.g.*, human and murine hemoglobin)
- **Paralogues**, present in the same genome, resulting from gene duplication (*e.g.*, human hemoglobin and human myoglobin)
- **Xenologues** are homologous genes resulting from horizontal transfer

Bioinformatic tools such as **BLAST** (NCBI) and **BLAST/BLAT** (Ensembl) can be used to identify gene or protein homology based on sequence similarity, in the genomes of any species (*full list below*).

Bioinformatic tools for homology search

- **NCBI** provides a **BLAST** tool for protein or DNA homology search
- **Ensembl** includes a **BLASt/BLAT tool**
- **UniProt** includes tools for identification of similarity
- **BLAT**: Protein or DNA homology search in animal genomes
- **FASTA**: Protein or DNA homology search and sequence alignment
- **HMMER**: Gene and protein homology search
- **Pfam**: Protein families and domains, includes tools for homology identification
- **SSAHA**: DNA sequence search and alignment

In our example, we are studying the evolution of TP53 in animals. Vertebrates have three TP53 paralogues (TP53, TP63 and TP73). Our analysis is based on dos Santos *et al*, Plos One, 2016.

*>Using **BLAST*** (https://blast.ncbi.nlm.nih.gov/Blast.cgi), *paste the sequence of your gene of interest (in our example, human TP53) in the Fasta format to identify homologues in*

the genomes of a selection of other species covering the diversity of animals (e.g. all animals and choanoflagellates), for example.

In our example, we chose the cnidarian *Hydra vulgaris*, the fruit fly *Drosophila melanogaster*, three other insects *(Bombus terrestris, Apis mellifera* and *Aedes aegyptus)*, the urochordate *Ciona intestinalis*, and the teleost fish *Danio rerio*, the coelacanth *Latimeria chalumnae*, the amphibian *Xenopus tropicalis*, the lizard *Anolis carolinensis*, the bird *Gallus gallus*, and the mammals *Bos taurus* and *H. sapiens*.

>Select the homologous sequences based on E.value and significant homology (>30% identity). For vertebrate species, select one sequence for every paralogue (p53, p63 and p73).

>Download all the sequences in the Fasta format in a Fasta file.

>(Optional): Calculate the identity matrix of the sequences using alignment tools (e.g., CLUSTALW 2.1).

1: M.brevicolis_A9V4M3	100.00	21.40	23.18	21.54	17.61	18.75	24.16	20.00	25.26	24.86	22.17	20.45	21.38	20.81	20.84
2: M.brevicolis_A9UZX3	21.40	100.00	24.66	24.37	19.44	21.32	28.18	23.89	29.02	28.98	26.22	23.42	23.38	24.62	23.11
3: H.sapiens_E9PFT5.1/13-155	23.18	24.66	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	49.68	61.79	54.67
4: H.sapiens_J3KP33.1/99-289	21.54	24.37	100.00	100.00	100.00	92.28	94.02	96.38	100.00	94.51	94.02	96.79	39.88	43.77	42.55
5: H.sapiens_E7EMR6.1/99-165	17.61	19.44	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	30.91	36.97	35.76
6: H.sapiens_E7EQX7.1/99-270	18.75	21.32	100.00	92.28	100.00	100.00	82.54	91.06	85.62	82.54	82.54	92.28	38.87	43.93	42.44
7: H.sapiens_A0A087WT22.1/1-130	24.16	28.18	100.00	94.02	100.00	82.54	100.00	93.05	100.00	94.51	93.05	93.05	48.11	47.59	47.59
8: H.sapiens_A0A0U1RQC9.1/60-250	20.00	23.89	100.00	96.38	100.00	91.06	93.05	100.00	100.00	95.05	89.32	92.94	39.60	41.97	40.35
9: H.sapiens_E7ESS1.1/1-157	25.26	29.02	100.00	100.00	100.00	85.62	100.00	100.00	100.00	100.00	100.00	100.00	51.78	52.76	52.76
10: H.sapiens_A0A087WXZ1.1/1-130	24.86	28.98	100.00	94.51	100.00	82.54	94.51	95.05	100.00	100.00	95.05	95.05	49.44	49.45	49.45
11: H.sapiens_A0A087X1Q1.1/1-130	22.17	26.22	100.00	94.02	100.00	82.54	93.05	89.32	100.00	95.05	100.00	100.00	44.30	44.98	45.41
12: H.sapiens_P04637.4/99-289	20.45	23.42	100.00	96.79	100.00	92.28	93.05	92.94	100.00	95.05	100.00	100.00	38.50	42.11	41.44
13: H.sapiens_Q9H3D4.1/167-359	21.38	23.38	49.68	39.88	30.91	38.87	48.11	39.60	51.78	49.44	44.30	38.50	100.00	59.66	57.19
14: H.sapiens_A0A0C4DFW9.1/68-260	20.81	24.62	61.79	43.77	36.97	43.93	47.59	41.97	52.76	49.45	44.98	42.11	59.66	100.00	84.73
15: H.sapiens_O15350.1/117-309	20.84	23.11	54.67	42.55	35.76	42.44	47.59	40.35	52.76	49.45	45.41	41.44	57.19	84.73	100.00

Percent identity matrix of the seven p53 sequences of Homo sapiens and the two p53 sequences of the choanoflagellate *Monosiga brevicollis*. The matrix was realised using ClustalW 2.1.

According to the matrix, the 13 human p53 paralogues share 36% to 100% identity, and the two paralogues of *Monosiga brevicollis* share 21.4% identity. Human and *Monosiga* orthologues share 17% to 25% identity. Hence, all human paralogues are more similar to each other than to any of the *Monosiga* orthologues.

4 **Step 3: Multiple sequence alignment (MSA)**

Sequence evolution

Molecular sequences mostly evolve through the accumulation of mutations. Mutations create divergence and diversity in homologous genes/proteins. They are the main data that can be used to reconstruct molecular evolution. Two main kinds of mutations exist: **indels** and **substitutions**.

- Indels include **insertions** and **deletions** of one or several nucleotides.

- Substitution designates the replacement of one nucleotide by another one. Substitutions include **transitions** (purine to purine or pyminide to pyrimidine) and **transversions** (purine to pyrimidine or *vice versa*).

Phylogenetic analysis requires identifying homology between the sequences, which is inferred by a sequence alignment. A **sequence aligment** is a way of arranging molecular sequences (DNA or protein) as a table that identifies homologous nucleotides or aminoacids, supposedly derived from common ancestor, in the same column. The sequences are put in every row one after the other to arrange every homologous base or amino acid in the same column. All elements of a column are considered homologous, and the alignment can be used as a taxa/character matrix. Alignment of the homologous residues usually necessitates adding **gaps**, represented by the symbol "-", that indicate insertions or deletions (**indels**), into the sequences. A **mutliple sequence alignment** (**MSA**) is an alignment of a large number of molecular sequences used for phylogenetic analysis.

The first step of phylogenetic analysis is to infer homology between the sequences. Alignment tools reconstruct the best MSA. They require defining **gap opening penalty**. Increasing gap opening penalty increases the probability to align non-homologous positions, while decreasing the penalty increases the probability of creating gaps that do not correspond to indels. A good gap opening penalty should provide a good balance between the two, and can be arbitrary. For beginners, we recommend using the default settings of the programs.

Tools for alignment of molecular sequences (genes and/or proteins)
* indicates a web interface

- **BAli-Phy**: Multiple sequence alignment of nucleotide and amino acid sequences and phylogenetic analysis using BI
- **CLUSTAL Omega***: Speed-oriented multiple sequence alignment for nucleotide or aminoacid data
- **CLUSTALW***: Multiple sequence alignment for nucleotide or amino acid datac [39]
- **CONTRAlign** (ProbCons): Accuracy-oriented multiple sequence alignment for amino acid data
- **Kalign***: Speed-oriented multiple sequence alignment for nucleotide or amino acid data
- **MAFFT***: Multiple sequence alignment for nucleotide or amino acid data
- **MUSCLE***: Multiple sequence alignment for nucleotide or amino acid data
- **PASTA**: Speed-oriented multiple sequence alignment for nucleotide or aminoacid data, designed for very large datasets
- **PRANK/WebPRANK***: Multiple sequence alignment for nucleotide or amino acid data, should be preferred for close sequences
- **SATe**: Software package for multiple sequence alignments and phylogenetic inference

- **T-COFFEE***: Multiple sequence alignment of nucleotide and amino acid sequences
- **UPP**: Speed-oriented multiple sequence alignment of nucleotide and aminoacid sequences, designed for very large data sets

*>Use an alignment tool (e.g. MAFFT, **https://www.ebi.ac.uk/Tools/msa/mafft/**) to align the sequences. Paste the alignment in the Fasta format and submit. Retrieve the alignment and save it in a new Fasta file.*

5 **Step 4: Alignment trimming** (recommended)

It is recommended to check the alignment and, when necessary, to improve it manually or using alignment trimming tools. Trimming is the selection of phylogenetically informative sites in the alignment. Poorly aligned positions and highly variable regions are not phylogenetically informative, because these positions might not be homologous or subject to saturation. They should be excluded to maximize the phylogenetic signal of the alignment.

Here is a selection of tools to quantify the completeness of alignments and selection of the phylogenetic informative regions of the alignment. For beginners, we recommend **Guidance 2**, that includes a web version.

Alignment trimming tools
* indicates a web interface

- **AliStat**: Quantification of alignment completeness for alignment refinement
- **BMGE**: Selection of informative regions on MSA
- **GBlocks**: Selection of informative regions on MSA
- **Guidance 2***: Selection of informative regions on MSA
- **Noisy**: Selection of informative regions on MSA
- **TrimAl**: Selection of informative regions on MSA*>Use one of the tools below to compute the completeness of your alignment and exclude the poorly aligned regions (regions of the alignment with low scores).*

*>Alternatively, you can also directly download the sequences into the Guidance 2 server (**http://guidance.tau.ac.il/**) and proceed to the alignment using MAFFT. Open the color-coded MSA to identify poorly aligned and highly variable regions. You can delete them manually from the alignment or remove unreliable columns below a certain cutoff.*

>The new MSA, hereafter renamed sub-MSA, will be used for the phylogenetic analysis. Save the sub-MSA in the Fasta format.

6 **Assessing phylogenetic assumptions** (for more advanced users)

Phylogenetic models rely on simplifying assumptions stating for example that all sites in the alignment evolved under the same tree, that mutation rates have remained constant,

and that substitutions are reversible. If the phylogenetic data violate these assumptions, the phylogeny and evolutionary analyses can be biased. Once the alignment is performed and the sites selected for phylogenetic inference, it is recommended to assess those phylogenetic assumptions when possible. Tests for some of these assumptions have been included in IQ-TREE. More advanced coders can also use the package MOTMOT, written in the R language.

II - Phylogenetic analysis

7 **Phylogenetic methods**

The evolutionary history of genes, proteins or species is generally presented as a phylogenetic tree, a graphical illustration of the evolutionary relationships between the sequences or taxa.

Several methods for phylogenetic inference can be used:

- **Maximum Parsimony** (MP) (*section 7.2*)

- **Distance-based methods** (UPGMA, NJ, ME) (*section 7.3*)

- **Probabilistic methods** (ML, BI), nowadays the most widely used for molecular data (*section 7.4*).

Method 1: Maximum parsimony (MP)

Maximum parsimony is a classical and simple method, that calculates the minimum number of evolutionary steps, including nucleotide insertions, deletions or substitutions, between species.

However, this method ignores hidden mutations and does not consider branch lengths, potentially leading to long branch attraction, an incorrect clustering of unrelated taxa. Furthermore, it does not consider the possibility of hidden mutations, making it not relevant for distant taxa. While MP is still used with morphological data, it is rarely used with molecular data.

Method 2: Distance-based methods

Distance-based methods create a matrix of molecular distances based on the number of differences between the sequences, to reconstruct the phylogenetic tree. These methods ignore hidden mutations and are also subject to long branch attraction. Distance-based methods include the Unweighted Pair Group Method with Arithmetic mean (**UPGMA**), Neighbor Joining (**NJ**), and Minimum Evolution (**ME**).

Method 3: Probabilistic methods (requires selection of the molecular evolution model, see below)

The strength of probabilistic methods is the use of specified models of molecular evolution. Probabilistic methods consider different mutation rates between sites to avoid mutation saturation. Nowadays, most publications use probabilistic methods.

They include Maximum Likelihood (**ML**, described below in the section 7.4) and Bayesian Inference (**BI**, described below in the section 7.4). ML calculates the probability of observing the data (in this case, the sequence alignment) under different explicit models of molecular evolution. ML aims to identify the best fit model by exploring multiple combinations of model parameters. Inversely, BI evaluates the probability of each model of molecular evolution given the data.

>Choose one or several phylogenetic methods to reconstruct the evolutionary history of your gene, protein or species of interest.

It can be interesting to combine several approaches and compare the results (e.g. Maximum Likelihood, Neighbor Joining and Bayesian Inference). However, confirming phylogenetic relationships with several methods does not necessarily mean that the tree is biologically correct.

We encourage users to perform phylogenetic analysis on the same dataset using one distace-method such as NJ, and the probabilitic methods ML and BI. For beginners, we recommend **Mega** or **SeaView** for ML- and NJ-based phylogenies. More advanced users should prefer **IQ-TREE**, **PhyML** or **RaxML** for ML, and can also use **MrBayes** for BI-based phylogenies.

Tools for phylogenetic reconstruction using diverse methods:

Distance-based methods:

- **MEGA**: Sequence alignment, model selection, phylogenetic analysis using distance methods, MP and ML, and other evolutionary analyses. Complete and very user-friendly
- **APE**: R-written package for molecular phylogenetics
- **FastMe***: Phylogenetic inference
- **PHYLIP**: Phylogenetic inference using MP, distance methods and ML

Bayesian inference (BI)

- **MrBayes**: Phylogenetic inference using BI and diverse evolutionary analyses
- **BAli-Phy**: Phylogenetic inference
- **BayesTraits**: Phylogenetic inference and other evolutionary analyses

- **PhyloBayes**: Phylogenetic inference with protein data using BI using a specific probabilistic model

Maximum likelyhood (ML)

- **HYPHY***: Phylogenetic inference using ML and distance methods
- **IQ-TREE***: Phylogenetic inference using ML, including model selection and a very fast bootstrapping method
- **MEGA**: Sequence alignment, model selection, phylogenetic analysis using distance methods, MP and ML, and other evolutionary analyses. Complete and very user-friendly.
- **PAML**: phylogenetic inference using ML, and other evolutionary analyses
- **PAUP**: Phylogenetic inference using MP and ML
- **PHYLIP**: Phylogenetic inference using MP, distance methods and ML
- **PhyML***: Phylogenetic inference using ML and various evolutionary analyses
- **RAxML***: Phylogenetic inference using ML
- **SeaView**: Sequence alignment and phylogenetic inference using MP, NJ and ML
- **GARLI**: Phylogenetic inference using ML

Other:

- **PyCogent**: Phylogenetic inference and various evolutionary analyses
- **SplitsTree**: Phylogenetic inference for unrooted trees and phylogenetic networks

*> We propose to reconstruct the phylogeny of the TP53 family using a distance method: Neighbor Joining (NJ) with **MEGA 11**, and a probabilistic method: Maximum Likelihood (ML) using **IQ-TREE 2**.*

7.1 **Step 2: Molecular evolution model selection** (for probabilistic methods and distance methods)

Prior to phylogenetic analysis, probabilistic methods and distance methods require selection of the model of molecular evolution that best describes the data. A model of molecular evolution is a combination of a substitution model and a model of rate heterogeneity between sites. Nucleotide or amino acid substitution models exist. They differ in the number of parameters considered, like substitution rates and base/aminoacid frequencies.

Substitution models: The main nucleotide substitution models are, from the simplest to the most complex: JC69, K80, F81, HKY85, TN93 and GTR (see below for their specificities). The main amino acid substitution models include JTT, WAG, LG and Dayhoff. In addition, rate heterogeneity between sites are the Gamma distribution (G) and the proportion of invariant nucleotide or amino acid sites (I) can be included. The FreeRate model (R), a more complex model of rate heterogeneity is included in

ModelFinder, PhyML and IQ-TREE. The GHOST model, for alignments with variation in mutation rate, is also implemented in IQ-TREE.

The main models of nucleic acid evolution

- **JC69** (Jukes & Cantor 1969): equal substitution rates, equal base frequencies.
- **F81** (Felsenstein 1981): equal substitution rates, unequal base frequencies.
- **K80** (Kimura 1980): transversion rate lower than transition rate, equal base frequencies.
- **HKY** (Hasegawa, Kishino, Yano, 1985): transversion rate lower than transition rate, unequal base frequencies.
- **TN93** (Tamura Nei 1993): HKY with unequal purine/pyrimidine rates.
- **K81** (Kimura 1981): three substitution types, equal base frequencies.
- **GTR** (General time reversible): unequal rates for all substitutio types, unequal base frequencies.

The main models of amioacid evolution

- **Poisson** (simplest model): equal aminoacid substitution rates, equal frequencies.
- **JTT** (Jones, Taylor, Thornton)
- **Dayhoff**
- **WAG** (Wehlan, Goldman)
- **LG** (Le, Gascuel)

Selection of the best-fit model for the data

The likelihood of every model should be computed using appropriate program, such as **ModelTest** and its java version **jModelTest** for nucleotide sequences, and **ProtTest** for amino acid sequences. **ModelFinder**, implemented in IQ-TREE, is designed for alignments of nucleotides, codons or amino acid data. **PartitionFinder 2** can be used with nucleotide and amino acid data. Model test selectors are also included in programs such as MEGA and PhyML (**SMS**).

These tools calculate the Bayesian information criterion (**BIC**) and the Akaike information criterion (**AIC**) for every model of molecular evolution (combination of substitution model and rate-heterogeneity model). A model with lower AIC or BIC is considered more accurate. The model optimizing BIC or AIC (*i.e.,* with the lowest score) should be selected.

Molecular model selection tools

- ModelFinder, implemented in IQ-Tree: Fast model selection method with a model of rate heterogeneity between sites (nucleotides, amino acids, codons)
- **ModelTest / jModelTest**: Nucleotide substitution model selection
- **PartitionFinder 2**: Molecular evolution model selection
- **ProtTest**: Aminoacid substitution model selection (nucleotides, amino acids)

- SMS: Nucleotide or aminoacid model selection included in PhyML (nucleotides, amino acids)

>Here, we are studying protein sequences. Use ProtTest 3.4.2 to calculate the log-likelihood of a panel of 56 amino acid substitution models, and select the most relevant one based on the BIC or AIC score.

*>Alternatively, you can use the model selectors included in IQTREE or MEGA. For example, with the IQTREE web server (**http://iqtree.cibiv.univie.ac.at/**), open the* "Model Selection" *panel, download the sub-MSA, select "*protein sequences*", choose a selection criterion (AIC or BIC) and proceed to the analysis. With MEGA, download the sub-MSA, and select "*Find best DNA/protein models*" in the* "Model" *panel.*

>You will use this model to compute the phylogenetic tree of your protein and for further evolutionary analyses.

7.2 **Step 3: Phylogenetic analysis**

Method 1: Maximum parsimony (MP)

Maximum parsimony is rarely used in phylogenetic analysis but it is simple and very straightforward for beginners.
PAUP, **MEGA**, **SeaView** and **Phylip** can be used for phylogenetic analysis using MP.

Phylogenetic analysis using MP with MEGA 11:
>To reconstruct the phylogenetic tree of p53 sequences using MP, import the sub-MSA in MEGA. Then, in the "Phylogeny *panel", choose a phylogenetic analysis using parsimony. Select the bootstrap method with at least 1000 replicates and execute the analysis.*

7.3 **Method 2: Distance-based methods**

The most widely used distace-based methods include UPGMA and the more advanced Neighbour Joining (NJ) and Minimum Evolution (ME). For phylogenetic analysis, we recommand using NJ ad ME.
FastME, **PAUP**, **MEGA**, **FastTree** or **Phylip** can be used for all of them.

Phylogenetic analysis using NJ with MEGA 11:
>To reconstruct the phylogenetic tree of p53 sequences using NJ, import the sub-MSA in MEGA. Then, in the "Phylogeny" *panel, choose a phylogenetic analysis using NJ. Select the appropriate substitution model (e.g JTT+G) and the bootstrap method with at least 1000 replicates. Execute the analysis.*

7.4 **Method 3: Probabilistic methods** (requires selection of the molecular evolution model, see below)
Probabilistic methods are the most powerful and the most widely used for phylogenetic analysis. They include **Maximum likelihood** (ML) and **Bayesian inference** (BI). We recommend using those methods for publications, rather than or in combination with distance-based methods.

Option 1: Maximum Likelihood (ML)

ML-based phylogeetic trees are common in literature. Many programs for ML-based phylogenetic analysis exist.

- Beginners can use **SeaView** or **MEGA**, which include tools for sequence alignment, phylogenetic inference with probabilistic methods, and a tree editor.
- **IQ-TREE**, that includes ModelFinder and a very fast bootstrapping method (UFBOOT2), is reported to be both fast and accurate. IQ-TREE also includes a **web version**.
- **PhyML** is accurate, easy of use and, like **PAUP** and **MEGA**, includes many common models of molecular evolution. PhyML also includes a web interface.
- For very large datasets, we recommed **RAxML** and particularly **FastTree** are fast and well suited for large datasets. They use only a specific model of rate heterogeneity (CAT), in addition to the Gamma law and the proportion of invariant sites. Like **Garli**, their choice of nucleotide evolution model is limited to GTR.
- **PAUP** is slower than other programs, and uses nucleotide data only.

>Choose a program relevant with the type and size of your dataset (see Table in **Step 6**).

For ML-based phylogenetic analysis with IQ-TREE 2:
*>Download the sub-MSA, select the appropriate sequence type (DNA or protein) and the appropriate model of molecular evolution (e.g JTT+G). In the panel "*branch support analysis*", select the Ultrafast Bootstrap analysis with at least 1000 replicates. For single branch tests, you can also select the SH-aLRT test. Execute the analysis.*

Option 2: Bayesian Inference (BI)

The main programs used for BI-based phylogenetics are **MrBayes** and **BEAST**, that use the Markov Chain Monte Carlo (MCMC) algorithm. **PhyloBayes** is a Bayesian MCMC sampler for phylogenetic reconstruction with protein data using a specific probabilistic model, well adapted for large datasets and phylogenomics. **Bali-Phy** can also be used for phylogenetic analysis using BI.

8 **Tree rooting**

The root of a phylogenetic tree is the hypothetical last common ancestor of all the taxa in the tree. Phylogenetic trees can be unrooted or rooted. Rooting a phylogenetic tree consist in identifying ancestral and derived states, to study the direction of the evolution of the sequences.

Rooting methods
The most common requires icluding outgroups in the dataset. Outgroups are taxa that do not belong to the studied group but are closely related. Typically, two outgroups are selected, one being more closely related to the ingroup than the other, allowing for a proper identification of the states of characters.

Alternative methods include the Midpoint rooting, which places the root at the mid-point of the longest branches, and the molecular clock rooting, which assumes that the evolution speed is constant between the sequences.

>In our example, we include the P53 homologues from the choanoflagellate *Monosiga brevicollis*. *Designate the P53 of* M. brevicollis *as the outgroup when drawing the phylogenetic tree with graphical programs.*

>Alternatively, select "Midpoint rooting" *when drawing the phylogenetic tree with graphical programs.*

II - Phylogenetic analysis

9 **Test of reliability of the tree**

It is recommended to estimate the reliability of the clades of the inferred phylogenetic tree. Most
programs of phylogenetic analysis use the **non-parametric bootstrapping method**.

Bootstrapping is a resampling technique used to assess the repeatability of the clade and estimate how consistently it is supported by the data. The nucleotides/aminoacids in the alignment are randomly
resampled with replacement, and a new phylogeny is calculated for each replicate. A **bootstrap value** is calculated for every clade, corresponding to the proportion of replicate phylogenies that recovered the clade. For beginners, we recommend using at least 1000 replicates to obtain good bootstrap accuracy.

Ex: A bootstrap value of 100% means that the branch is supported by all resampled datasets, while low values mean that only few of these datasets support the branch. Values lower than 70% are usually considered a poor support for the clade. Users

should keep in mind that bootstrapping gives a measure of the consistency of the estimate, but it is not a measure of the accuracy of the tree.

Bootstrapping programs
-Beginners can use the bootstrapping tool that is included in **Mega** or **SeaView**.
-In ML-based phylogenetic analysis, the **Shimodaira-Hasegawa** (**SH**) test and its improved version, **Approximately unbiased** (**AU**), can be used with PAUP, PhyML, FastTree and IQ-TREE.
-IQ-TREE includes the fast and accurate bootstrapping tools **UFBoot** and **UFBoot2**.
-PhyML includes the **approximate likelihood ratio test** (**aLRT**), an alternative to the non-parametric
bootstrap implemented.
-Bayesian inference methods use **posterior probabilities** (**PP**) to measure branch support.

With **Mega**
*>Before launching the analysis, select "**Boostrap method**" and the **number of replicates** (e.g. 1000). With large datasets, bootstrapping can be very time-consuming.*

With the web version of **IQ-TREE**
*>Select the **standard** or the **ultrafast bootstrap method**, the latter is less accurate but should be preferred with large datasets. Select the number of replicates. You can also select the single-branch SH-aLRT test.*

With PhyML
*>Select the **standard bootstrap analysis** and the umber of replicates, or one of the fast likelihood based methods (aLRT or aBayes)*

II - Phylogenetic analysis

10 **Tree drawing**

Once the phylogenetic tree is computed, it can be exported using *e.g.* Newick file format and visualized using a graphical software such as **FigTree**, **ETE Toolkit** or **ITOL**. **MEGA** and **SeaView** also include visualization tools. Using different sets of options, several types of phylogenetic trees can be drawn (rooted or not, cladogram or phylogram), and branch support values (bootstrap values or posterior probabilities) can be displayed.

Tools for graphical visualization and annotation of phylogenetic trees
- ETE Toolkit: Visualization and analysis of phylogenetic trees
- FigTree: Graphic software for phylogenetic trees

- ITOL*: Visualization and annotation of phylogenetic trees
- MEGA and Seaview include tree visualization and annotation tools

With Mega 11:
>Click "file" > "export current tree (Newick)", select "Bootstrap" and "branch length", retrieve the phylogenetic tree in the Newick format and save it with nwk as filename extention.

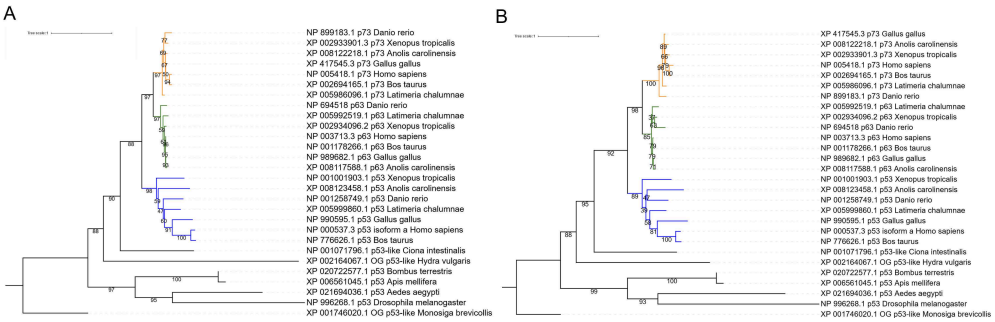
With IQ-TREE 2:
>Paste the phylogenetic tree in the Newick format to Notepad, and save it using nwk as filename extension.

>Use a program (e.g. FigTree) in the list below and open the nwk files. You can also directly paste the phylogeny in the Newick format in the graphical tools.

>Many options of tree drawing are available. For example, you can display the bootstrap values, the posterior probabilities, or the SH-aLRT values, collapse clades below a certain bootstrap threshold (e.g. 50), and highlight or add color to the clades.

In our example (figure below), we used **ITOL** for the graphical representation of the phylogenetic tree of the P53 family.

Interpretation of a phylogenetic tree
Both methods reveal four major clades containing respectively the p53 of insects and the p53, p63, and p73 of all vertebrates. The p53, p63, and p73 of vertebrates are more closely related to each other than to any other p53. Furthermore, the p63 and p73 of vertebrates are more closely related to each other than to vertebrate p53. This indicates that two duplication events in the p53 family preceded the origin of vertebrates. First, the p53 family and the p63/p73 cluster diverged. The second one caused the p63 and p73 families to diverge. The p53 of insects are clustered together. This indicates that insects diverged from the other bilaterians before these two duplications.



Phylogenetic trees of p53 domain-containing proteins of metazoans using Neighbour Joining (A) and Maximum Likelihood (B). The trees were realized according to the model JTT+G, as calculated by ModelFinder using AIC. The numbers indicate the bootstrap values as calculated by the standard bootstrapping method and UFBoot2, respectively. The phylogenetic trees were inferred using MEGA 11 and IQ-TREE 2, respectively, and the figures were generated using ITOL. Green branches represent the p63 family, orange branches represent the p73 family and blue branches represent the p53 family.

III - Evolutionary analyses using phylogenetic trees

11 Sequence alignments and phylogenetic trees can be used to reconstruct diverse aspects of the evolutionary history of genes, proteins and species. In this last section, we provide a brief and non-exhaustive overview of evolutionary studies that can be performed using bioinformatics.

11.1 1 - Time-calibration of phylogenetic trees

Phylogenetic calibration consists in estimating the age of speciation or duplication events (the nodes in the phylogenetic tree), using events with a known age, such as fossil and other geological data (that can only give minimal ages) as calibration points. Alternatively, mutation rates can be used to calculate the divergence time between two sequences.

Tools for time-calibration of phylogenetic trees

- MEGA provides tools for time calibration
- **BayesTraits**: Evolutionary analyses using Bayesian inference
- **BEAST**: Diverse evolutionary analyses using BI, including time-calibration of phylogenetic trees
- **Mesquite**: Comparative analyses and statistics
- **Ohnologs**: Database of vertebrate ohnologues, resulting from whole genome duplications. Can be used to estimate the evolutionary age of gene paralogues
- **TimeTree**: database of evolutionary age of speciation events

Databases such as **TimeTree** compute the estimated divergence time between all species and the relevant literature. **Mesquite** also provides tools to calibrate phylogenetic trees in geological times using fossil data. **Ohnologs** can be used to estimate the divergence time between homologues resulting from whole genome duplications in vertebrates.

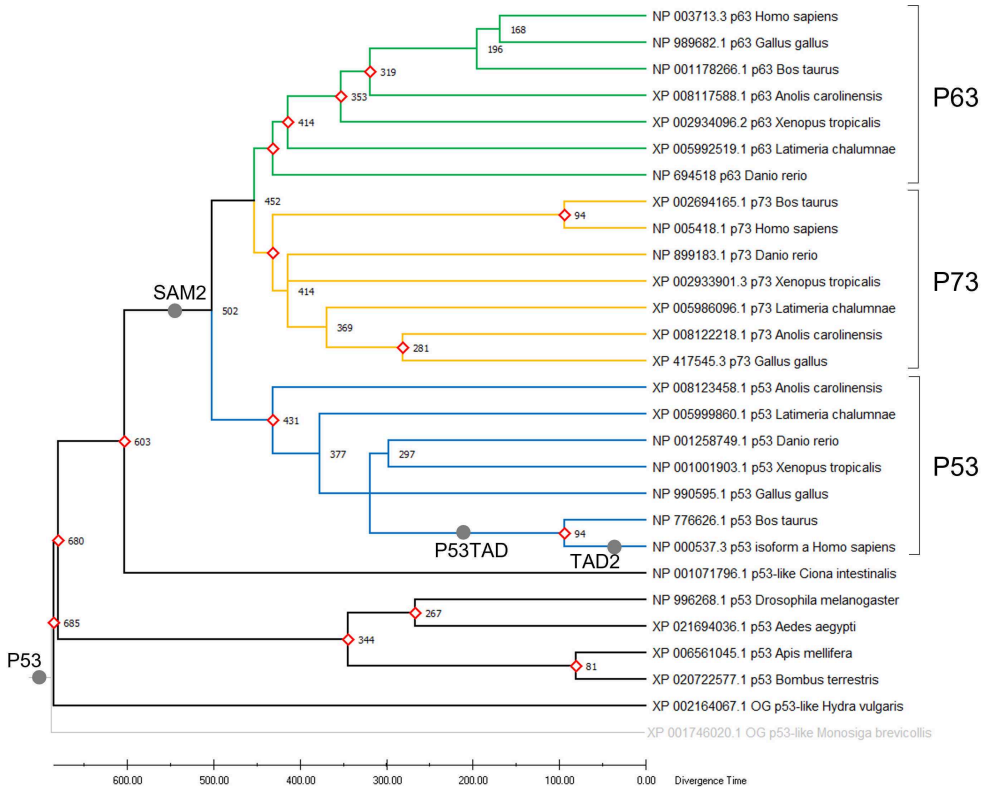
To calibrate the phylogenetic tree of P53 with Mega 11 and TimeTree:
>Download the alignment file in the Fasta format and the tree file in the Newick format in MEGA. In the Compute *panel, select* "Compute Timetree"*.*

>In the Specify outgroup *section, define* Monosiga brevicollis *as the outgroup by moving it to the* Outgroups *panel. In the section* Calibrate nodes, *select* "internal nodes constraints"*.*

*>In TimeTree (**http://www.timetree.org/**), enter the names of two species to retrieve their estimated divergence time. For example,* Homo *and* Drosophila *diverged between 630 and 830 million years ago, with 694 million years as median time.*

>In MEGA 11, click "add new calibration point" *and select the node in the phylogenetic tree, or enter the names of the two taxa, and define the speciation age with a minimum, maximum or fixed time (for example, 694 million years between* Homo *and* Drosophila*). Use TimeTree to define several calibration points between different species in the tree before and after the duplication events.*

>Click "Launch the analysis", *and retrieve and save the calibrated tree. You can export it with the divergence times. Use the graphical programs to add color, highlight clades, etc...*



Time-calibrated phylogenetic tree of p53 domain-containing proteins of metazoans. The tree was realized according to the model JTT+G as calculated by ModelFinder using AIC. The phylogenetic tree and the figure were realized using MEGA 11. Time calibration was performed using TimeTree. The values on the branches and the scale indicate the divergence time in million years. Green branches represent the p63 family, orange branches represent the p73 family and blue branches represent the p53 family. Grey spots on the branches indicate the origin of the different protein domains during the evolution of the TP53 family.

11.2 2 - Reconstruction of ancestral states

Retracing the functional evolution of genes, proteins, or biological traits often requires the reconstitution of ancestral states. They can be inferred from a phylogenetic tree using MP, ML, or BI; and requires the aligned sequences and the model of molecular evolution that has been used for the phylogenetic analysis when using probabilistic and distance methods.

Tools for ancestral states reconstruction

BayesTraits: Evolutionary analyses using Bayesian inference
BEAST: Diverse evolutionary analyses using BI
IQ-TREE, IQ-TREE2: Ancestral states reconstruction
Mesquite: Comparative analyses and statistics
PAML: Evolutionary analyses using ML including ancestral states reconstruction

PyCogent: Numerous evolutionary analyses, including ancestral states reconstruction
RASP: Ancestral states reconstruction

Reconstruction of ancestral states with Mega 11:
>Import the alignment file in the Fasta format, and the tree file in the Newick format in MEGA. In the Ancestors *panel click* "infer ancestral sequences" *and select the method (MP or ML). In ML, select the appropriate substitution model. Launch the analysis to display the reconstructed ancestral state of each site of the sequence at every node.*

11.3 **5 - Genome evolution**

Genome evolution studies the appearence and divergence of genes belonging to the same family, or multigenic families, such as the well-known Hox genes in metazoans. Genome databases (see section 2) are very useful for gene and genome comparison between species. Evolutionary events such as single nucleotide polymorphism (SNP), insertions and deletions (indels), create diversity in paralogues/orthologues (genes that share a common origin) and can be identified using sequence alignment (see section 4).

Other events, such as gene duplication, chromosome duplication and fusion, chromosome reorganization and whole genome duplication (WGD) events shape the constitution of the genomes, as well as pseudogenization (loss of function on one gene) and horizontal gene transfer. These evolutionary events can be identified and studied using genomics tools and databases in complement with phylogenetic trees.

Tools for genome evolution studies
- HGT-Finder: Identification of horizontal gene transfer
- Ohnologs: Database of vertebrate ohnologues, resulting from whole genome duplications
- CAFE: Gene family evolution
- CoGE: Comparative genomics analyses

General molecular databases:
- NCBI: Collection of databases for molecular biology and medicine, providing tools and services
- Ensembl: Genome browser of vertebrates, includes tools for identification of homology
- Entrez: Gene sequences and structures
- GenBank: Annotated DNA sequences

11.4 **4 - Study of co-evolution**

Co-evolution refers to the genetic and/or morphological changes between different species in interaction. It is widely used in evolutionary ecology and parasitology to study the evolution of hosts and parasites. Co-evolutionary events include co-speciation, host change, duplication and loss of interaction. The evolution of the parasite is partly driven by the evolution of the host, which is considered independent from the evolution of the parasite. The co-evolutionary history can be presented as a co-phylogeny with the two entities.

Some programs co-evolution study, including **Jane**, **CoRe-PA** and **TreeMap** [170] (**Table 8**), are based on the hypothesis that the evolution of the parasite is driven by the evolution of the host. Others, such as **Copycat** [162], reconciliate the two phylogenies under the hypothesis that the situation is symmetric and evaluate the significance of co-evolution under a statistical framework. Co-evolution of genes or proteins can also be studied using these tools.

In our second test-case study, we used ML to reconstruct the evolution of cyclins and CDKs, two families of proteins involved in cell cycle control and closely interacting. We used **Jane** and **TreeMap** to reconstruct cophylogenies between the two families.

>With Jane and TreeMap, a single nexus file containing the phylogenies of cyclins and CDKs, and their associations is needed. Create a nexus file (starting with #NEXUS). This file should contain the two trees in Newick format, in the sections BEGIN HOST and BEGIN PARASITE, and the associations in the section BEGIN DISTRIBUTION. This section should mention every association between Cyclins and CDK following the pattern "Host: Parasite,". All three sections should end with "ENDBLOCK;". The names of the taxa in the three files should be identical. Cyclins interacting with several CDKs and vice versa should be repeated.

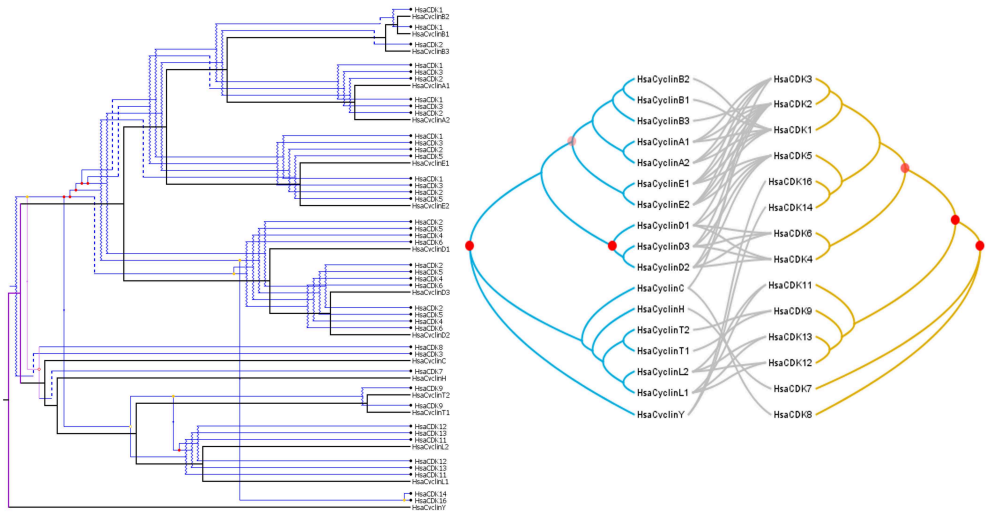
>Import this file to Jane and launch the analysis in the Solve Mode. The costs of coevolutionary events can be set. The stats mode can be used to compute the cost range of the solutions. With TreeMap, import the nexus file and launch the analysis in "Solve the tanglegram". We optain a coevolutionary scenario that represents the best way to associate the two trees. You can test the significance of the reconstruction in "estimate significance" or perform a heuristic test.

In both figures, the clustering of cyclins and CDKs indicate an interaction (the cyclin can bind the CDK and activate it). Red spots indicate significant events of coevolution between the two families of proteins. Co-speciation (hollow red circle), duplications (solid red circle), duplications with host switch (yellow circle), loss of interaction (dashed lines), failures to diverge (jagged lines) are indicated on the figure.

List of programs for coevolution studies

- Copycat

- CoRe-PA
- Jane
- TreeMap



Two co-evolutionary scenarios of the associations between, and co-evolution of human cyclins and CDKs

11.5 6 - Phylogenetic comparative analysis

Evolutionary biology often employs the so-called phylogenetic comparative methods to study the adaptive significance of biological traits. These methods aim at identifying biological characters, in terms of morphology, physiology, or ecology, that result from a shared ancestry. Comparative analyses can be done for quantitative or qualitative variables. **Mesquite** is a very appropriate tool for comparative analysis and to compute statistics on phylogenetic trees. **BayesTraits** can also be used.

11.6 3 - Measure of selection strength

The type and strength of selection on protein coding genes may be of interest. It is calculated by evaluating the ratio of the number of non-synonymous substitutions (substitutions changing the protein sequence) per non-synonymous site (dN), and the number of synonymous substitutions (substitutions with no effect on the protein sequence due to the redundancy of the genetic code) per synonymous site (dS). If dN/dS > 1, then the non-synonymous substitutions are higher than expected and the gene is under positive selection. If dN/dS<1, the gene is under purifying selection and if dN/dS=1, the selection is neutral. It is recommended not to use the dN/dS ratio for closely related species. The ratio can be calculated using **PAML**, **MEGA**, **Bio++** and **HyPhy**.

Measure of selection strength with Mega 11:
>Import the alignment file in the Fasta format, and the tree file in the Newick format in MEGA. In the Selection *panel click* "infer ancestral sequences".

11.7 **7 - Evolution of populations**

Genetic diversity can be explored at the population level by analyzing polymorphism between members of the same species. Bioinformatic tools are designed to study allele diversity within a population, including single nucleotide polymorphisms (SNPs), indels, microsatellites or transposable elements. Mathematical models have been developed to describe polymorphism. Several programs are suitable for population genetics studies.

Tools for population genetics studies

- **Arlequin**: Population genetics analyses
- **DNAsp**: Analysis of DNA polymorphism
- **Genepop**: Population genetics analyses
- **SNiplay** SNP detection and other population genetics analyses

11.8 **8 - Study of protein structure and function evolution**

Studying the functional evolution of proteins can require structure alignments, that can be realized using **PyMol**, and the mean distance in ångström between homologous residues can be calculated**.**

Protein structures are stored in the Protein Data Bank (**PDB**), that provides the 3-dimensional structures of proteins and their interacting ligands established by X-ray crystallography, electron microscopy, or NMR spectroscopy, which can be retrieved as pdb files. The PDB also displays a 3D visualization tool, programs for 3D analyses such as pairwise structure alignment and pairwise symmetry, and cross links to other protein databases. Annotation for protein families based on fingerprints, *i.e.*, conserved 3-dimensional motifs specific for a protein family, are gathered in the database **PRINTS**. PRINTS includes a 3D visualization software and search tools for protein sequence homology and pairwise or multiple sequence alignments.

I-TASSER, **HHPred** of the HH suite and **Alpha fold** can be used to predict the 3-dimensional structure of proteins from their amino-acid sequences. **FoRSA** is able to identify a protein fold from its amino acid sequence or a protein sequence in the proteome of a species from a crystal structure.

Tools protein structures analyses

- **Alpha**: fold Protein structure prediction from amino-acid sequence
- **FoRSA**: Protein structure prediction from amino-acid sequence

- **HHPred**: Protein structure prediction from amino-acid sequence
- **I-TASSER**: Protein structure prediction from amino-acid sequence
- **PyMOL**: 3D visualization of molecules, including proteins
- **PDB**: Database that provides all published 3D structures of proteins
- **PRINTS**: Protein fingerprints classification database