Sep 12, 2022 Version 2

Roadmap to the bioinformatic study of gene and protein phylogeny and evolution - a practical guide V.2

DOI
dx.doi.org/10.17504/protocols.io.36wgq77e3vk5/v2

florian.jacques[1], Paulina Bolivar[2]
[1]Masaryk University; [2]Lunds University

Protocol for studying ge...

Florian G Jacques
Masaryk University

DOI: https://dx.doi.org/10.17504/protocols.io.36wgq77e3vk5/v2

Protocol Citation: florian.jacques, Paulina Bolivar 2022. Roadmap to the bioinformatic study of gene and protein phylogeny and evolution - a practical guide . **protocols.io https://dx.doi.org/10.17504/protocols.io.36wgq77e3vk5/v2**Version created by **Florian G Jacques**

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

Created: August 24, 2022

Last Modified: September 12, 2022

Protocol Integer ID: 69101

Keywords: Evolution, bioinformatics, phylogenetic analysis, evolutionary studies, molecular evolution, Phylogenetic inference, bioinformatic tools for phylogenetic reconstruction, phylogenetic reconstruction, simple phylogenetic reconstruction, protein phylogeny, protein database, biological database, molecular evolution, bioinformatic tool, roadmap to the bioinformatic study, compilation of nucleic acid, protocol for information extraction, information extraction, population genetics, bioinformatic study, gene, evolution

Abstract

We present a compilation of nucleic acid and protein databases and bioinformatic tools for phylogenetic reconstructions and a wide range of studies on molecular evolution and population genetics. We provide a protocol for information extraction from biological databases and simple phylogenetic reconstruction.

Troubleshooting

Sequence collection and comparison

1 **Collecting sequence data and information on genes and proteins**

Evolutionary analyses on molecular data (genes, genomes or proteins), require retrieving information from public databases. Several dozens of databases store information about the state of the art on genes and proteins. Most of them provide the sequences in fasta format, which is necessary for evolutionary studies. Other information and annotations, including structure, activity, biological function, tissue expression, subcellular location and polymorphism can also prove relevant.

>Retrieve the sequence from one of the following databases (e.g. **NCBI** or **Uniprot**) and paste the sequence in a fasta file using fasta format, using .fasta as filename extension.

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

Here is a non-exhaustive list of nucleic acid databases, and a list of protein databases, with their main features.

	A	B	C
	Database	**Features**	**Link**
	BAR	Database of plant genes and proteins	http://bar.utoronto.ca/
	Bgee	Gene expression patterns	https://bgee.org/

	A	B	C
	Ensembl	Genome browser of vertebrates, includes tools for identification of homology	https://www.ensembl.org/index.html
	Entrez	Gene sequences and structures	https://www.ncbi.nlm.nih.gov/Web/Search/entrezfs.html
	FlyBase	Genome and proteins of the model insect *D. melanogaster*	https://flybase.org/
	GeneCards	Human gene function, genomics, transcription factor binding sites and protein products	https://www.genecards.org/
	GenBank	Annotated DNA sequences	https://www.ncbi.nlm.nih.gov/genbank/
	NCBI	Collection of databases for molecular biology and medicine providing many bioinformatics tools and services	https://www.ncbi.nlm.nih.gov/
	PomBase	Genes and proteins of the model yeast *S. pombe*	https://www.pombase.org/
	TAIR	Genome and proteins of the model plant *A. thaliana*	https://www.arabidopsis.org/
	WormBase	Genome and proteins of the model nematode *C. elegans*	https://wormbase.org//#012-34-5
	Xenbase	Genome and proteins of the model amphibian *X. laevis*	http://www.xenbase.org/entry/

List of nucleic acid databases

	A	B	C
	Database	**Features**	**Link**

	A	B	C
	Gene Ontology	Unified annotation of molecular function, biological processes, and cellular components of proteins	http://geneontology.org/
	Human Protein Atlas	Information on human protein and their link with diseases	https://www.proteinatlas.org/
	InterPro	Classification of proteins domains and functional sites	https://www.ebi.ac.uk/interpro/
	KEGG	Protein function and biological pathways	https://www.genome.jp/kegg/
	PDB	3-dimensional structures of proteins	https://www.rcsb.org/
	Pfam	Information about protein families and domains, includes tools for identification of homology	http://pfam.xfam.org/
	PHAROS	Centralizes literature for human proteins	https://pharos.nih.gov/
	PRINTS	Protein fingerprints classification database	http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/
	PROSITE	Protein family, domains and functional sites	https://prosite.expasy.org/
	SCOP	Structure-based classification of protein	https://scop.mrc-lmb.cam.ac.uk/
	SUPERFAMILY	Protein structure and functions	https://supfam.org/
	UniProt	General information on proteins, including sequences, structure, classification, function, subcellular localization and simple homology identification	https://www.uniprot.org/uniprot/

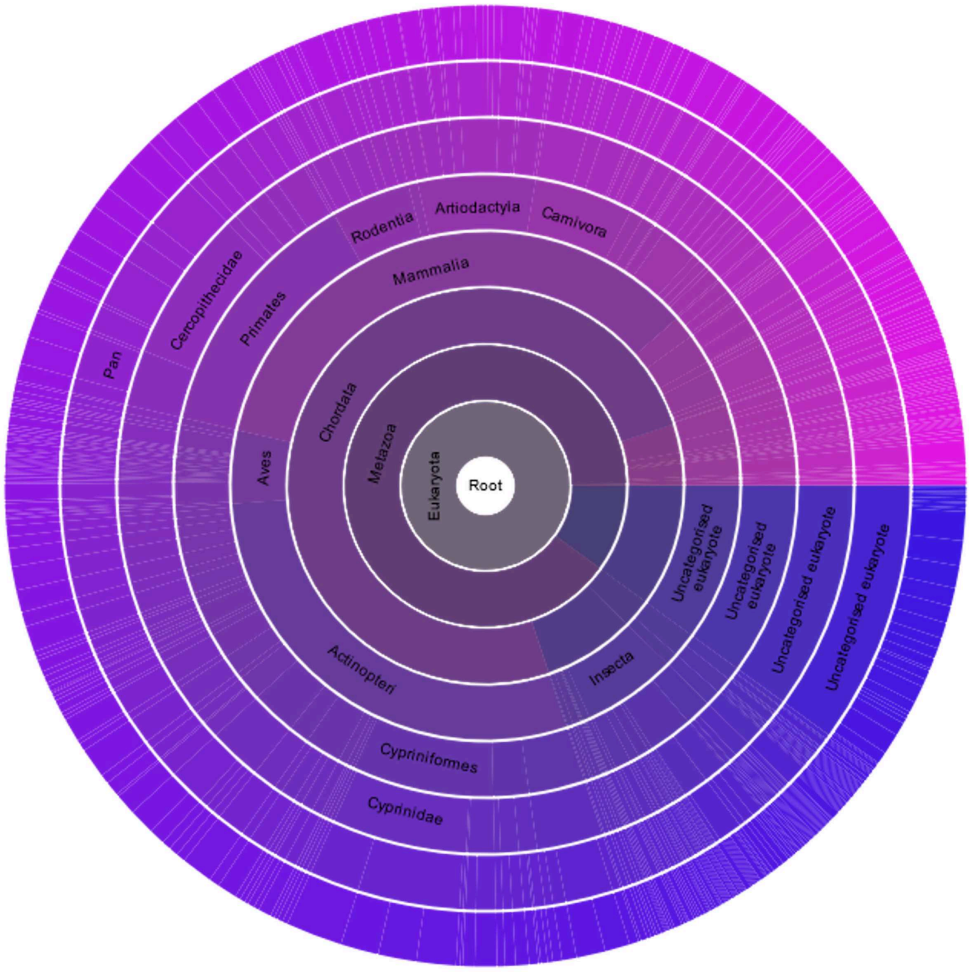
List of protein databases

1.1 **Protein domains classification** (optional)

Studying protein classification can also be useful for evolutionary studies. Proteins are classified into different categories based on structural and functional similarity, evolutionary relationship, or both. Retrieving the classification of a protein of interest and identifying the main protein domains often provides valuable insight on its diversity 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. Pfam presents also their occurrence in living organisms as a sunburst plot.

This figure displays the diversity of the domain P53 and can be retrieved from 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.



Sunburst plot of the distribution of the P53 protein domain (PF00870) in living organisms according to Pfam.

2 Identification of homologues

Studying the evolution of a family of genes or proteins requires the identification of homologues, *i.e.*, genes or protein with shared ancestry. Homologues include orthologues, that are present in different species, and paralogues, that are present in the same genome. Bioinformatic tools can be used to identify gene or protein homology based on sequence similarity, in the genomes of any species (see the list below). Here is a list of tools that can be used to identify sequence homology.

	A	B	C
	BLAST	Gene or protein homology search from NCBI	https://blast.ncbi.nlm.nih.gov/Blast.cgi
	BLAT	Sequence homology search in animal genomes	https://genome.ucsc.edu/cgi-bin/hgBlat
	Ensembl	Genome browser of vertebrates, includes tools for identification of homology	https://m.ensembl.org/index.html
	FASTA	Sequence search against protein databases and sequence alignment	https://www.ebi.ac.uk/Tools/sss/fasta/
	HMMER	Gene and protein homology search	http://hmmer.org/
	Pfam	Protein families and domains, includes tools for identification of homology	http://pfam.xfam.org/
	SSAHA	Gene sequence search and alignment	https://www.sanger.ac.uk/tool/ssaha/
	UniProt	General information on proteins, including simple homology identification	https://www.uniprot.org/uniprot/

List of bioinformatic tools for identification of gene and protein homologues

>Using **BLAST**, paste the sequence of your gene of interest (in our example, human TP53) under fasta format to identify homologues in the genomes of select other species covering the diversity of animals (e.g. all animals and choanoflagellates), for example.

>Select the sequences the low E.value and significant homology (typically >30% identity) for a wide range of animal species and download them under fasta format in another fasta file.

In our example, we are studying the evolution of TP53 in animals. It is relevant to select the 3 paralogues (TP53, TP63 and TP73) of a selection of species covering the diversity of animals and their relatives. In our example, we chose the cnidarian *Hydra vulgaris*, the

fruit fly *Drosophila melanogaster,* the urochordate *Ciona intestinalis,* and the teleost fish *Danio rerio,* the sarcopterygian *Latimeria chalumnae,* the amphibian *Xenopus tropicalis,* the reptile *Anolis carolinensis,* the bird *Gallus gallus,* and the mammals *Bos taurus* and *H. sapiens.* Only vertebrates possess 3 paralogues.

3 **Multiple sequence alignment**

Phylogenetic analysis requires identifying homologous bases or amino acid residues between the different sequences. Homology is inferred by a sequence alignment. The sequences are put in every row one after the other to arrange every homologous base or amino acid in the same column.

>Use an appropriate tool (e.g. ClustalW) to generate an alignment of the sequences. Insertions and deletions are indicated by gaps "-" added to the sequences. Once the alignment is completed, it can be exported using fasta format.

Here is a list of tools for nucleic acid or amino acid sequence alignment.

	A	B	C
	Software	**Features**	**Link**
	BAli-Phy	Multiple sequence alignment of nucleotide and amino acid sequences and phylogenetic analysis using a bayesian approach	http://www.bali-phy.org
	CLUSTAL Omega*	Speed-oriented multiple sequence alignment for nucleotide or amino acid data, suitable for large datasets	https://www.ebi.ac.uk/Tools/msa/clustalo/
	CLUSTALW*	Multiple sequence alignment for nucleotide or amino acid data	https://www.genome.jp/tools-bin/clustalw
	CONTRAlign (ProbCons)	Accuracy-oriented multiple sequence alignment for amino acid data	http://contra.stanford.edu/contralign/
	Kalign*	Multiple sequence alignment for nucleotide or amino acid data, suitable for large datasets	https://www.ebi.ac.uk/Tools/msa/kalign/

	A	B	C
	MAFFT*	Accuracy-oriented multiple sequence alignment for nucleotide or amino acid data	https://mafft.cbrc.jp/alignment/server/
	MUSCLE*	Multiple sequence alignment for nucleotide or amino acid data	https://www.ebi.ac.uk/Tools/msa/muscle/
	PASTA	Speed-oriented multiple sequence alignment for nucleotide or amino acid data, designed for very large datasets	https://bioinformaticshome.com/tools/msa/descriptions/PASTA.html
	PRANK/ WebPRANK*	Speed-oriented multiple sequence alignment for nucleotide or amino acid data, should be preferred for close sequences and large datasets	http://wasabiapp.org/software/prank/ https://www.ebi.ac.uk/goldman-srv/webprank/
	SATé	Software package for multiple sequence alignments and phylogenetic inference	https://phylo.bio.ku.edu/software/sate/sate.html
	T-COFFEE*	Accuracy-oriented multiple sequence alignment of nucleotide and amino acid sequences	http://tcoffee.crg.cat/
	UPP	Speed-oriented multiple sequence alignment of nucleotide and amino acid sequences, designed for very large data sets	https://github.com/smirarab/sepp.

List of programs for sequence alignment (* indicates a web interface)

4 **Alignment trimming**

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. These positions should be excluded prior to the phylogenetic analysis to maximize the phylogenetic signal of the alignment. Then, the alignment can be exported using fasta format.

>Use one of these tools (e.g. Guidance 2) to compute the completeness of your alignment and exclude the poorly aligned regions (regions of the alignment with low scores). Save the new alignment in fasta.

Here is a selection of tools to quantify the completeness of alignments and selection of the phylogenetic informative regions of the alignment.

	A	B	C
	Software	**Features**	**Link**
	AliStat	Selection of informative regions on multiple sequence alignments	https://github.com/thomaskf/AliStat
	BMGE	Selection of informative regions on multiple sequence alignments	https://gitlab.pasteur.fr/GIPhy/BMGE
	GBlocks	Selection of informative regions on multiple sequence alignments	http://molevol.cmima.csic.es/castresana/Gblocks.html
	Guidance 2*	Selection of informative regions on multiple sequence alignments	http://guidance.tau.ac.il/
	Noisy	Selection of informative regions on multiple sequence alignments	http://www.bioinf.uni-leipzig.de/Software/noisy/
	trimAl	Selection of informative regions on multiple sequence alignments	http://trimal.cgenomics.org/

List of programs for sequence alignment trimming (* indicates a web interface)

5 **Assessing phylogenetic assumptions** (optional)

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. Several statistical methods have been developped. Tests for all these assumptions have been included in IQ-TREE and R (package MOTMOT).

Phylogenetic analysis

6 **Phylogenetic inference**

The evolutionary history of genes, proteins or species is generally presented as a phylogenetic tree, a graphical illustration of the evolutionary relationships between the studied taxa. Several methods for phylogenetic inference exist: Maximum Parsimony, the distance-based methods and the probabilistic methods. The probabilistic methods are nowadays the most widely used for molecular data, but other methods can be used in complement.

>Choose one or several phylogenetic methods to reconstruct the evolutionary history of your gene, protein or species of interest. See the sub-steps below for the specificities of each method. It can be interesting to use several approaches and compare the results (e.g. Maximum Likelihood, Neighbour Joining and Bayesian Inference).

Here is a list of tools for phylogenetic reconstruction using diverse methods, and visualization of phylogenetic trees that can be used in complement.

	A	B	C
	Software	**Features**	**Link**
	APE	R-written package for molecular phylogenetics	http://ape-package.ird.fr
	BAli-Phy	Sequence alignment and phylogenetic inference using a Bayesian approach	http://www.bali-phy.org
	BayesTraits	Phylogenetic inference and other evolutionary analyses using Bayesian inference	http://www.evolution.reading.ac.uk/BayesTraitsV4.0.0/BayesTraitsV4.0.0.html

	A	B	C
	BEAST	Diverse evolutionary analyses using Bayesian inference, including phylogenetic analysis, calibration and molecular clock	http://www.beast.community
	ETE Toolkit	Visualization and analysis of phylogenetic trees	http://etetoolkit.org/
	FastMe	Fast phylogenetic inference using distance methods.	http ://www.atgc-montpellier.fr/fastme/
	FastTree	Phylogenetic inference using ML for nucleotide (GTR and JC models) and amino acid (JTT and WAG models), and Shimodaira-Hasegawa test. Suitable for very large datasets.	http://www.microbesonline.org/fasttree/
	FigTree	Graphic software for phylogenetic trees	http://tree.bio.ed.ac.uk/software/figtree/
	GARLI	Phylogenetic inference using ML for nucleotide (GTR model), amino acid (most models) or codon data, with Gamma law and proportion of invariant sites.	http ://evomics.org/resources/software/molecular-evolution-software/garli/
	HYPHY*	Diverse evolutionary analyses including evolution model selection, phylogenetic inference using ML and distance methods and sequence evolution studies	https ://www.hyphy.org/
	IQ-TREE	ML phylogenetic inference, including model selection and ultrafast bootstrapping method. Includes the GHOST evolution model and tests for phylogenetic assumptions.	http ://www.iqtree.org/
	ITOL	Visualization and annotation of phylogenetic trees	https://itol.embl.de/
	MEGA	Sequence alignment, model selection, phylogenetic analysis (parsimony, distance methods). Includes all common nucleotide and amino acid evolution models, Gamma law and proportion of invariant sites, and Bootstrapping	https://www.megasoftware.net/

	A	B	C
		method.	
	MrBayes	Bayesian phylogenetic inference, ancestral states reconstruction, phylogenetic calibration and other evolutionary analyses	http://nbisweden.github.io/MrBayes/
	PAML	Maximum likelihood phylogenetic inference, estimation of selection strength, ancestral states reconstruction and other analyses	http://abacus.gene.ucl.ac.uk/software/paml.html
	PAUP	Phylogenetic inference using maximum parsimony and ML on nucleotide sequences (all ModelTest models), with Gamma law and proportion of invariant sites and bootstrapping method.	http://paup.phylosolutions.com/
	PHYLIP	Phylogenetic inference using parsimony, distance methods and ML	https://evolution.genetics.washington.edu/phylip.html
	PhyloBayes	Phylogenetic inference using Bayesian inference on proteins using a specific probabilistic model	http://www.atgc-montpellier.fr/phylobayes/
	PhyML*	Phylogenetic inference using ML, ancestral states reconstruction and various evolutionary analyses. Includes all common DNA and protein evolution models and diverse branch support methods (Bootstrap, Shimodaira-Hasegawa, aLTR...).	https://github.com/stephaneguindon/phyml Web interface : http://atgc.lirmm.fr/phyml/
	PyCogent	Phylogenetic inference and phylogeny drawing, various evolutionary analyses including partition models and ancestral states reconstruction	https://github.com/pycogent/pycogent
	RAxML	Phylogenetic inference using ML with nucleotide (GTR) or amino acid data (all common models) with Gamma law or CAT and proportion of invariant sites. Suitable for large datasets.	https://cme.h-its.org/exelixis/web/software/raxml/

	A	B	C
	SeaView	Sequence alignment and phylogenetic inference using maximum parsimony, NJ and ML	http://doua.prabi.fr/software/seaview
	SplitsTree	Phylogenetic inference, in particular unrooted trees, or phylogenetic networks	https ://uni-tuebingen.de/fakultaeten/mathematisch-naturwissenschaftliche-fakultaet/fachbereiche/informatik/lehrstuehle/algorithms-in-bioinformatics/software/splitstree/

List of programs and databases for phylogenetic analysis using diverse methods (* indicates a web interface)

6.1 Option 1: Maximum parsimony

Maximum parsimony is a classic 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 take into account branch lengths, potentially leading to long branch attraction, which is an incorrect clustering of unrelated taxa. Furthermore, it does not consider the possibility of hidden mutations, making it not relevant for distant taxa. While maximum parsimony is still used for morphological data, it is no longer considered relevant for molecular data.

PAUP, **MEGA**, **SeaView** and **Phylip** can be used for phylogenetic analysis using maximum parsimony.

6.2 Option 2: Distance-based methods

Distance-based methods create a matrix of molecular distance, defined by the number of differences between the sequences, to reconstruct the phylogenetic tree. Several distance-based methods exist. The Unweighted Pair Group Method with Arithmetic mean (UPGMA), Neighbor Joining (NJ), and Minimum Evolution (ME) are all based on the overall molecular distance, defined by the number of differences between the sequences. Distance-based methods also ignores hidden mutations and is also subject to long branch attraction.

FastME, **PAUP**, **MEGA**, **FastTree** or **Phylip** can be used for distance-based methods.

6.3 **Option 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 take into account different mutation rates between sites to avoid mutation saturation. Nowadays, almost all studies of phylogenetic reconstruction use probabilistic methods.

Probabilistic methods include Maximum Likelihood (ML) and Bayesian inference (BI). 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. BI evaluates the probability of each model of molecular evolution given the data. We recommend using different methods of phylogenetic reconstruction, including ML and BI.

Probabilistic methods require selection of the molecular evolution model.

6.4 **Selection of the molecular evolution model** (for probabilistic methods only)

Prior to phylogenetic analysis, probabilistic methods require selection of the model of molecular evolution that best describes the data. Nucleotide or amino acid substitution models exist. The nucleotide substitution models differ in the number of parameters considered, like mutation rates and base frequencies. The main nucleotide substitution models are, from the simplest to the most complex: JC69, K80, F81, HKY85, TN93, GTR. The main amino acid substitution models include JTT, WAG, LG and Dayhoff. Each model can be associated with the proportion of invariant nucleotide or amino acid sites (+I) or with the Gamma distribution (+G), which corresponds to a mutation rate heterogeneity between sites. More recently, the GHOST model for alignments with variation in mutation rate was introduced and implemented in **IQ-TREE**.

The likelihood of the different models should be computed by appropriate software. For every substitution model, these tools calculate the Bayesian information criterion (**BIC**) and the Akaike information criterion (**AIC**) from the log-likelihood scores. A model with lower AIC is considered more accurate. The model optimizing BIC or AIC (*i.e.,* with the lowest score) should be selected. Molecular evolution model selectors are also included in **MEGA** and **PhyML** (**SMS**).

>Using e.g. **jModelTest**, compute the BIC and AIC of the different models on the alignment. Select the model optimizing (minimizing) those criterions.

>You will use this model to compute the phylogenetic tree of your protein.

Here is a selection of tools that can be used for molecular evolution model selection.

	A	B	C
	Software	**Features**	**Link**
	ModelFinder	Fast model selection with a model of rate heterogeneity between sites (nucleotide, amino acids or codons)	Implemented in IQ-TREE
	ModelTest / jModelTest	Nucleotide substitution model selection	http://evomics.org/resources/software/molecular-evolution-software/modeltest/
	PartitionFinder 2	Molecular evolution model selection (nucleotide or amino acids)	http://www.robertlanfear.com/partitionfinder/
	ProtTest	Aminoacid substitution model selection	https://github.com/ddarriba/prottest3
	SMS	Molecular evolution model selection included in PhyML (nucleotide or aminoacid)	http://www.atgc-montpellier.fr/sms/

List of programs for molecular evolution model selection

6.5 Maximum Likelihood

Maximum Likelihood methods calculate the probability of observing the data under different explicit models of molecular evolution. Many programs for maximum-likelyhood-based phylogenetic analysis exist, that can be accuracy-oriented or speed-oriented.

>Choose a program according to the size of your dataset (see Table in Step 6). For beginners, we recommend **SeaView** or **MEGA**, which include several tools for sequence alignment, phylogenetic inference including probabilistic methods and others, and a tree editor. **PhyML** is accurate, easy of use and, like **PAUP** and **MEGA**, includes all common models of molecular evolution. PhyML also includes a web interface. **RAxML** and particularly **FastTree** are fast and well suited for large datasets (up to 1 million sequences

with FastTree). They use only a restricted a specific model of rate heterogeneity, in addition to Gamma law and proportion of invariant sites. Like **Garli**, their choice of nucleotide evolution model is limited to GTR. **IQ-TREE**, that includes ModelFinder and a very fast bootstrapping method, is reported to be both fast and accurate. **PAUP** is slower than other programs, and uses nucleotide data only.

>In our example, we used Maximum likelihood to build the phylogenetic tree of the P53 family (including the p73, p63 and p53 proteins of different animal species). Use an appropriate programme (e.g. **MEGA**) to reconstruct the phylogeny of P53 family using Maximum Likelihood with the appropriate model of evolution.

6.6 Bayesian inference

The most recent method for phylogenetic reconstruction uses Bayesian inference, which calculates the probability of the molecular evolution model given the data. The main software 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 Bayesian inference.

7 Tree rooting

The root of a phylogenetic tree is the hypothetical last common ancestor of all the taxa present in the tree. Depending on the addressed question, phylogenetic trees can be unrooted or rooted. The latter corresponds to the identification of ancestral and derived states, aiming at studying the direction of the evolution and interpreting the evolution of the studied taxa. Diverse methods have been developed to root phylogenetic trees. The most common require including outgroups in the analysis. Outgroups are taxa that do not belong to the studied ingroup but are closely related. Typically, two outgroups are selected, one being more closely related to the ingroup than to the other outgroup, allowing for a proper identification of the states of characters (ancestral or derived).

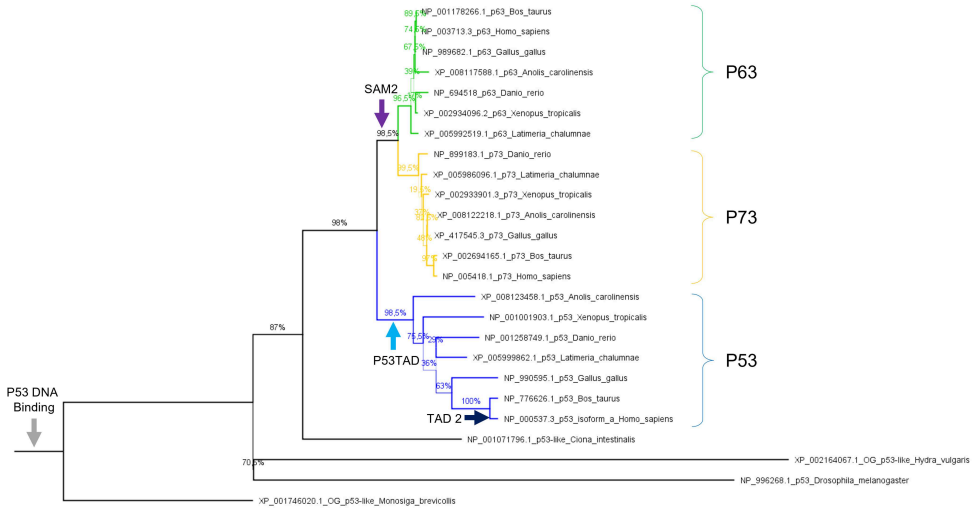
>In our example, we include the P53 homologues from the choanoflagellate *Monosiga brevicollis* and the Cnidarian *Hydra vulgaris*.

When outgroups are not identified, like in the case of viruses, alternative methods can be used. For example, midpoint rooting places the root at the mid-point of the longest branches, and molecular clock rooting assumes that evolution speed is constant between the sequences.

8 Tree drawing

Once the phylogenetic tree has been 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.

>In our example, we used **FigTree** for the graphical representation of the phylogenetic tree of the P53 family.



Phylogenetic tree of p53 domain-containing proteins of metazoans using maximum likelihood. The tree was realized according to the model JTT+G with Gamma law, as calculated by ProtTest 3.4.2 using AIC. The percentages indicate the bootstrap values. The phylogenetic tree was inferred using MEGA 11 and the figure was generated using FigTree 1.4.4. Green branches represent the p63 family, yellow branches represent the p73 family and blue branches represent the p53 family. The appearence of the different protein domains of the P53 family are indicated by arrows.

Working with phylogenies

9 Reconstruct the evolution of the gene or protein

Sequence alignments and phylogenetic trees can be used to reconstruct diverse aspects of the evolutionary history of genes, proteins and species, as well as the study the genetic structures within populations. In this last section, we provide a brief and non-exhaustive overview of evolutionary studies that can be performed using bioinformatic tools.

9.1 Reconstitution of ancestral states

Retracing the functional evolution of genes, proteins, or biological traits often requires the reconstitution of ancestral states. Ancestral states can be inferred from a phylogenetic tree using maximum of parsimony, maximum likelihood, or Bayesian inference; and requires the aligned sequences and the model of molecular evolution that has been used for the phylogenetic analysis (when using probabilistic methods only).

	A	B	C
	Software	**Features**	**Link**
	BayesTraits	Evolutionary analyses using Bayesian inference	http://www.evolution.rdg.ac.uk/BayesTraitsV3.0.5/BayesTraitsV3.0.5.html
	BEAST	Diverse evolutionary analyses using Bayesian inference, including phylogenetic analysis, calibration and molecular clock	http://www.beast.community
	Mega	Sequence alignment, model selection, phylogenetic analysis (parsimony, distance methods)	https://www.megasoftware.net/
	Mesquite	Comparative analyses and statistics	http://www.mesquiteproject.org/
	MrBayes	Bayesian phylogenetic inference, ancestral states reconstruction, phylogenetic calibration...	http://nbisweden.github.io/MrBayes/
	PAML	Maximum likelihood phylogenetic inference, estimation of selection strength, ancestral states reconstruction and other analyses	http://abacus.gene.ucl.ac.uk/software/paml.html
	RASP	Ancestral states reconstruction	http://mnh.scu.edu.cn/soft/blog/RASP/index.html

List of programs and databases for ancestral states reconstruction

9.2 Measure of selection strength

The strength of selection on protein coding genes can be calculated by evaluating the ration of the number of non-synonymous mutations (mutations changing the protein sequence) per non-synonymous site (dN) and the number of synonymous mutations (mutations with no effect on the protein sequence due to the redundancy of the genetic code) per synonymous site (dS). The ratio dN/dS reveals, if >1, that non-

synonymous mutations 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. Selection strength can be calculated by sequence alignment programs such as **MEGA**.

9.3 Phylogenetic calibration

Phylogenetic calibration consists in estimating the age of speciation events (the nodes in the phylogenetic tree) with events with a known age, for example fossil data 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.

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

	A	B	C
	Software	**Features**	**Link**
	BayesTraits	Evolutionary analyses using Bayesian inference	http://www.evolution.rdg.ac.uk/BayesTraitsV3.0.5/BayesTraitsV3.0.5.html
	BEAST	Diverse evolutionary analyses using Bayesian inference, including phylogenetic analysis, calibration and molecular clock	http://www.beast.community
	Mega	Sequence alignment, model selection, phylogenetic analysis (parsimony, distance methods)	https://www.megasoftware.net/
	Mesquite	Comparative analyses and statistics	http://www.mesquiteproject.org/
	MrBayes	Bayesian phylogenetic inference, ancestral states reconstruction, phylogenetic calibration...	http://nbisweden.github.io/MrBayes/
	Ohnologs	Database of vertebrate ohnologues, resulting from whole genome duplications	http://ohnologs.curie.fr/

	A	B	C
	Timetree	Tree calibration	http://www.timetree.org/

List of programs and databases that can be used for phylogenetic calibration using diverse methods

9.4 Study of coevolution

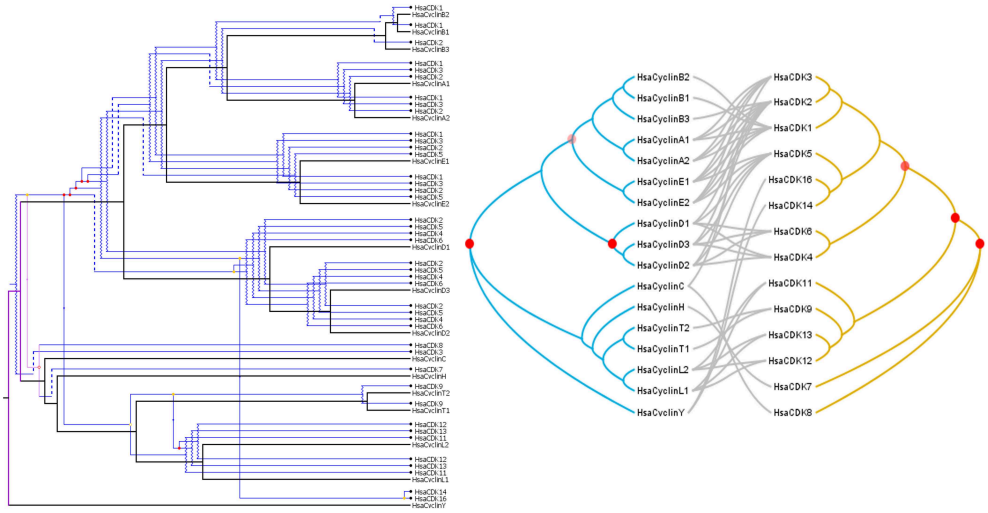
Evolutionary ecology and parasitology often study the evolution of hosts and parasites association through an approach of coevolution, or reciprocal genetic change between different species. Coevolution can also be used to study associations of genes or proteins. Studying coevolution consists in identifying evolutionary events such as co-speciation, host change, and duplication or loss of interaction from the phylogenetic trees of the hosts and the parasites.

In our second example, we used Maximum likelihood 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 coevolutionary scenarios (cophylogenies) between the two families.

In both figures, the clustering of cyclins and CDKs indicate an interaction (in this case, that the cyclin can bind the CDK). 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.

	A	B	C
	Software	**Features**	**Link**
	Copycat	Software for studying coevolution	http://www.cophylogenetics.com/
	Core-PA	Software for studying coevolution	http://pacosy.informatik.uni-leipzig.de/49-1-CoRe-PA.html
	Jane	Software for studying coevolution	https://www.cs.hmc.edu/~hadas/jane/
	TreeMap	Software for studying co-evolution	https://sites.google.com/site/cophylogeny/treemap

List of programs to study coevolution



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

9.5 Genome evolution

Evolutionary events, such as mutations, insertions, deletions, gene or whole genome duplications, genome reorganization, and genetic exchanges can be identified using phylogenetic trees in complement with genomics tools and databases. Here is a list of databases tools to study diverse aspects of genome evolution, including genome browsers of diverse lineages and tools for comparative genomics and evolutionary genomics:

	A	B	C
	Software	**Features**	**Link**
	BAR	Database of plant genes and proteins	http://bar.utoronto.ca/
	CAFE	Gene family evolution	https://github.com/hahnlab/CAFE5
	CoGE	Comparative genomics analyses	https://genomevolution.org/coge/

	A	B	C
	Ensembl	Genome browser of vertebrates, includes tools for identification of homology	https://www.ensembl.org/index.html
	Entrez	Gene sequences and structures	https://www.ncbi.nlm.nih.gov/Web/Search/entrezfs.html
	FlyBase	Genome and proteins of the model insect *D. melanogaster*	https://flybase.org/
	GenBank	Annotated DNA sequences	https://www.ncbi.nlm.nih.gov/genbank/
	HGT-Finder	Horizontal gene transfer finding	http://cys.bios.niu.edu/HGTFinder/HGTFinder.tar.gz
	Ohnologs	Database of vertebrate ohnologues, resulting from whole genome duplications	http://ohnologs.curie.fr/
	PomBase	Genes and proteins of the model yeast *S. pombe*	https://www.pombase.org/
	TAIR	Genome and proteins of the model plant *A. thaliana*	https://www.arabidopsis.org/
	WormBase	Genome and proteins of the model nematode *C. elegans*	https://wormbase.org//#012-34-5
	Xenbase	Genome and proteins of the model amphibian *X. laevis*	http://www.xenbase.org/entry/

List of programs and databases to study genome evolution

9.6 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.

9.7 Population genetics

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

	A	B	C
	Software	**Features**	**Link**
	DNAsp	Analysis of DNA polymorphism	http://www.ub.edu/dnasp/
	Genepop	Population genetics analyses	https://genepop.curtin.edu.au/
	SNiplay	SNP detection and other population genetics analyses	https://sniplay.southgreen.fr/cgi-bin/home.cgi
	Arlequin	Population genetics analyses	http://cmpg.unibe.ch/software/arlequin35/

List of programs and databases for population genetics

9.8 Study of protein structure and function evolution

Studying the functional evolution of proteins can require structure alignments, that can be realized by appropriate programs such as **PyMol**, and the mean distance in Å between homologous residues can be calculated**.**

Protein structures are described by databases such as the Protein Data Bank (**PDB**). The PDB 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. 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 sequence alignment or multiple sequence alignment.

Other programs allow to identify and study structures conservation between proteins and infer a structures from a protein sequence. These analyses and a phylogenetic tree of the protein families, together with the reconstruction of ancestral states, can facilitate the study of the evolution of protein functions within the family.

Here is a list of tools that can be used for analyses on protein structures in an evolutionary framework:

	A	B	C
	Software	**Features**	**Link**
	PyMol	3D visualization of molecules	http://www.mesquiteproject.org/
	PDB	3-dimensional structures of proteins	https://www.rcsb.org/
	Prints	Protein fingerprints classification	http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/
	PyMol	3D visualization of molecules and diverse analyses on protein structures	https://pymol.org/2/
	I-Tasser	Protein structure prediction	https://zhanglab.dcmb.med.umich.edu/I-TASSER/
	Forsa	Protein structure prediction	http://www.bo-protscience.fr/forsa/
	HHPred	Protein structure prediction	https://toolkit.tuebingen.mpg.de/tools/hhpred

List of programs for protein structure analyses