

Introduction à la bioinformatique (UE SSV3U15)

TP5. Inférence phylogénétique

Emese Meglecz & Yvan Perez

Objectifs et notions

Objectifs

- Montrer la contribution de l'inférence moléculaire pour résoudre des questions concernant la phylogénie.

Exemples traités durant le TP

- Nous étudierons les relations phylogénétiques entre différentes espèces de reptiles (tortues, lézards, serpents...) et d'oiseaux.

Compétences acquises

A l'issue de ce TP, vous devrez avoir acquis les compétences suivantes.

- Décrire et interpréter des arbres phylogénétiques.
- Évaluer les valeurs de robustesse.
- Comparer des arbres et discuter différentes hypothèses.

- **Taxinomie** (ou **taxonomie**) : 1. Science de la classification, 2. Classification des éléments d'un domaine, en particulier les espèces biologiques
- **Arbre vrai / Arbre inféré** : Un arbre phylogénétique qui reflète exactement les relations de parenté entre groupes d'êtres vivants est qualifié d'arbre vrai. En réalité, l'arbre vrai n'est jamais connu. L'idée de l'inférence phylogénétique est de construire des arbres à partir des données à disposition (arbre inféré) qui s'approchent le plus possible de l'arbre vrai.
- **Arbre des espèces** : Arbre qui indique les relations de parenté entre des espèces d'êtres vivants – ou par extension d'autres niveaux taxonomiques.
- **Arbres des molécules** : Arbre phylogénétique inféré à partir des séquences biologiques, et qui reflète l'évolution vraisemblable des séquences.
- **Clade/Groupe monophylétique** : Un groupe d'organismes contenant un organisme ancestral et tous ses descendants, et n'incluant aucun autre organisme.
- **Groupe basal (ou lignée basale)** : groupe taxonomique qui se détache des autres à proximité de la racine d'un arbre phylogénétique. Le concept de groupe basal est questionnable car il dépend du choix des échantillons ayant servi à établir l'arbre phylogénétique.

N'oubliez pas que vous pouvez à tout moment consulter le [glossaire du cours](#) pour obtenir une définition sommaire des principaux termes utilisés.

- **Groupe frère** : Des groupes frères sont des groupes qui dérivent d'un ancêtre commun immédiat (les branches qui descendent d'un nœud).
- **Robustesse** : une estimation qui montre dans quelle mesure chaque nœud d'un arbre inféré est soutenu par le jeu de données. La méthode la plus connue est le bootstrap.
- **Méthode de bootstrap** : Estimation de la robustesse de chaque nœud d'un arbre. Cette méthode consiste à échantillonner les positions de l'alignement pour relancer la construction phylogénétique de façon itérative puis de comparer les arbres obtenus après de nombreuses répétitions. La valeur de bootstrap d'un nœud représente la proportion des arbres dans lequel le nœud a été retrouvé.
- **Phylogénomique** : Reconstruction phylogénétique sur base d'un grand nombre de gènes ou de génomes complets.

N'oubliez pas que vous pouvez à tout moment consulter le [glossaire du cours](#) pour obtenir une définition sommaire des principaux termes utilisés.

Etapes

- Exercice 1. **Analyse d'un arbre phylogénétique**
 - Dessin d'un arbre phylogénétique sur base de vos connaissances préalables – se familiariser avec le groupe taxinomique que nous allons étudier
 - Analyse de l'arbre phylogénétique – se familiariser avec le vocabulaire de la phylogénie et la description d'un dendrogramme (arbre)
- Exercice 2. **Arbre de référence – NCBI Common Tree**
 - Créer un arbre de référence avec la base de données du NCBI
 - Comprendre la structure phylogénétique de l'arbre et savoir le redessiner
- Exercice 3. **Phylogénie moléculaire de la protéine RAG2 chez les Sauria**
 - Aligner les séquences orthologues de RAG2
 - Effectuer une reconstruction phylogénétique avec le site NGphylogeny ou phylogeny.fr
- Exercice 4. **Phylogénie moléculaire basée sur la concaténation des séquences protéiques de 8 gènes différents chez les Sauria**

Complétion

- Tous les exercices doivent être réalisés par chaque étudiant.
- En principe, les 4 exercices devraient être faits en séance (avec explications par les enseignants).
- Si nécessaire, ils peuvent être terminés ultérieurement.

Rappel et précision des notions mises en oeuvre pour ce TP

Alignements multiples et distances génétiques

- **Caractères moléculaires**

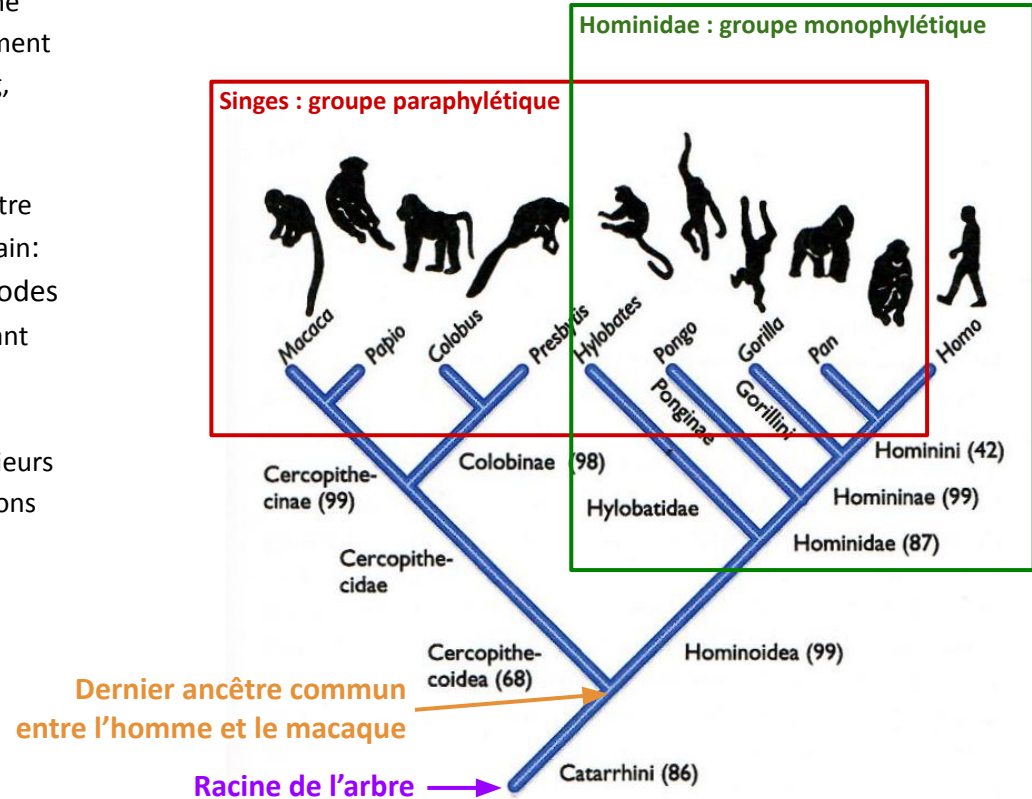
- La phylogénie “classique” repose sur des caractères phénotypiques (morphologiques, anatomiques, physiologiques, ...)
- En phylogénie moléculaire, on considère chaque position (colonne) d'un alignement multiple comme un caractère pour les constructions phylogénétiques.

- **Similarité et homologie (hypothèse primaire)**

- La présence de similarités entre caractères chez différents organismes peut a priori résulter d'un héritage commun (**homologie**) ou d'une convergence évolutive (**analogie**)
- En phylogénie moléculaire, on mesure un **taux de similarité** en calculant, à partir d'un alignement de séquences, le pourcentage de résidus alignés ayant un score positif (identités ou de substitutions conservatives).
- On peut également calculer des **scores probabilistes** (score en bits et **E-valeur,expect**) qui permettent d'estimer la **significativité statistique** des similarités observées entre deux séquences.
- Si la **significativité est élevée**, on en inférera qu'il s'agit d'une **homologie**.
- Attention, la similarité est un concept qu'on peut représenter sur une échelle quantitative (pourcentages de positifs, pourcentage d'identité) mais l'**homologie est un critère qualitatif, binaire** : deux séquences sont homologues, ou ne le sont pas.

Rappels des définitions

- **Groupe monophylétique** : groupe comportant un organisme ancestral et tous les organismes en descendant, et uniquement eux. Exemple : les **Hominidae** incluent gibbon, orang-outang, gorille, chimpanzé, bonobo (absent du dessin) et humain.
- **Groupe paraphylétique** : groupe qui inclut un organisme ancestral et ses descendant, mais en excluant certains d'entre eux. Exemples : les **singes** incluent les primates sauf l'humain : les **poissons** incluent les gnathostomes sauf les tétrapodes
- **Groupe polyphylétique** : assemblage d'organismes n'incluant pas leur ancêtre commun le plus récent. Exemples : mammifères marins, animaux cavernicoles.
- **Cénancêtre** = **dernier ancêtre commun** entre deux ou plusieurs groupes taxonomiques : espèce la plus récente que ces taxons ont pour ancêtre commun.
- Spéciation et duplication



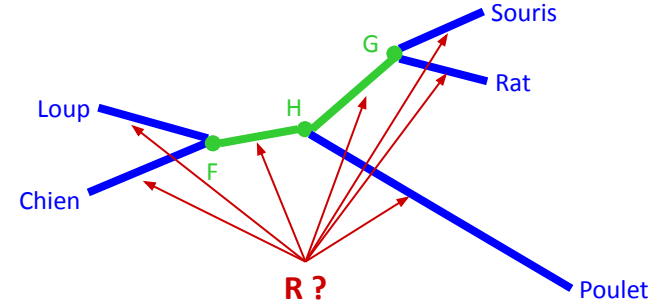
Arbres enracinés ou non enracinés

- Selon les méthodes utilisées, l'inférence phylogénétique produit soit un **arbre enraciné**, soit un **arbre non-enraciné**.
- Les arbres non-enracinés ne sont pas réellement des arbres phylogénétiques car ils n'ont pas de direction temporelle → les branches indiquent les étapes de séparation, avec une longueur proportionnelle distances évolutives, mais en absence de direction elles n'indiquent pas les relations de parenté (qui descend de qui).
- La **racine** définit une orientation de l'arbre, et donc un chemin évolutif unique vers chaque feuille. Elle symbolise le **dernier ancêtre commun** (*i.e.* le plus récent) de toutes les OTU.
- A priori, elle peut se situer à n'importe quelle position sur n'importe quelle branche de l'arbre.

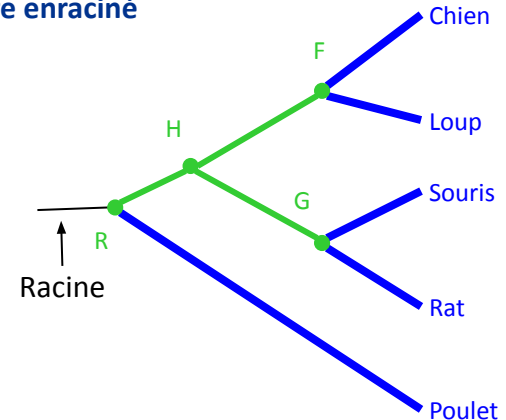
Comment enraciner un arbre ?

- Dans certains cas, on peut s'appuyer sur une connaissance *a priori* de la feuille la plus externe parmi les OTU étudiées, qualifiée de **groupe extérieur** (*outgroup* en anglais)
 - Exemple : si un arbre contient chien, loup, souris, rat et poulet → sur base des connaissances biologiques, on décide que le **groupe extérieur** est le poulet
- En absence de connaissance *a priori* du OTU les plus externes parmi les OTU étudiées, on peut envisager un **enracinement au poids moyen** : on enracine l'arbre sur la branche qui minimise la moyenne des distances aux feuilles.
 - **Note**: ceci implique une hypothèse d'**horloge moléculaire**: on considère que le taux de mutation est constant au cours de l'évolution, et égal entre les branches. Cette hypothèse n'est généralement pas très réaliste, il s'agit d'une approximation.

Arbre non-enraciné



Arbre enraciné



Méthode de bootstrap pour estimer la robustesse des arbres

En phylogénie moléculaire, on infère un arbre phylogénétique à partir d'un alignement multiple (après avoir supprimé les colonnes qui comportent des gaps).

On peut s'interroger sur la fiabilité de cette inférence, qui dépend des séquences particulières dont on dispose dans l'échantillon analysé.

Une méthode pour aborder cette question est le **bootstrap**.

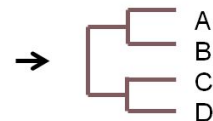
Pour évaluer la fiabilité de l'inférence, on peut appliquer la méthode du **bootstrapping**.

1. Etant donné un alignement de N séquences et M colonnes, on effectue une sélection aléatoire de M colonnes **avec remise**. Chaque colonne peut donc être tirée 0, 1 ou plusieurs fois.
2. On **calcule un arbre** avec ces colonnes ré-échantillonnées.
3. On **répète l'opération** un bon nombre de fois (ex: 1000)
4. On assigne à chaque branchement de l'arbre initial une **valeur de bootstrap** = le nombre de fois où ce branchement se retrouve à l'identique dans les N arbres produits.

La valeur de bootstrap est un **indice de la robustesse** de l'arbre phylogénétique par rapport aux fluctuations d'échantillonnage.

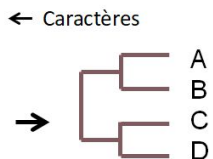
Alignement multiple initial

SeqA	A	T	T	C	A	T	G	A	T	T	C	T	G	G
SeqB	A	G	T	C	A	T	G	A	T	C	C	T	G	G
SeqC	A	C	T	C	A	T	G	A	G	T	C	T	G	G
SeqD	A	C	T	C	A	A	G	A	G	T	C	T	C	G
	1	2	3	4	5	6	7	8	9	10	11	12	13	14



Bootstrap1

SeqA	T	A	T	T	A	C	T	G	C	A	T	T	G	A
SeqB	T	A	T	T	A	C	C	G	C	A	T	T	G	A
SeqC	T	A	T	G	A	C	T	G	C	A	G	T	G	A
SeqD	T	A	T	G	A	C	T	G	C	A	G	T	C	A
	3	8	3	9	1	4	10	14	4	8	9	3	13	1



Bootstrap2

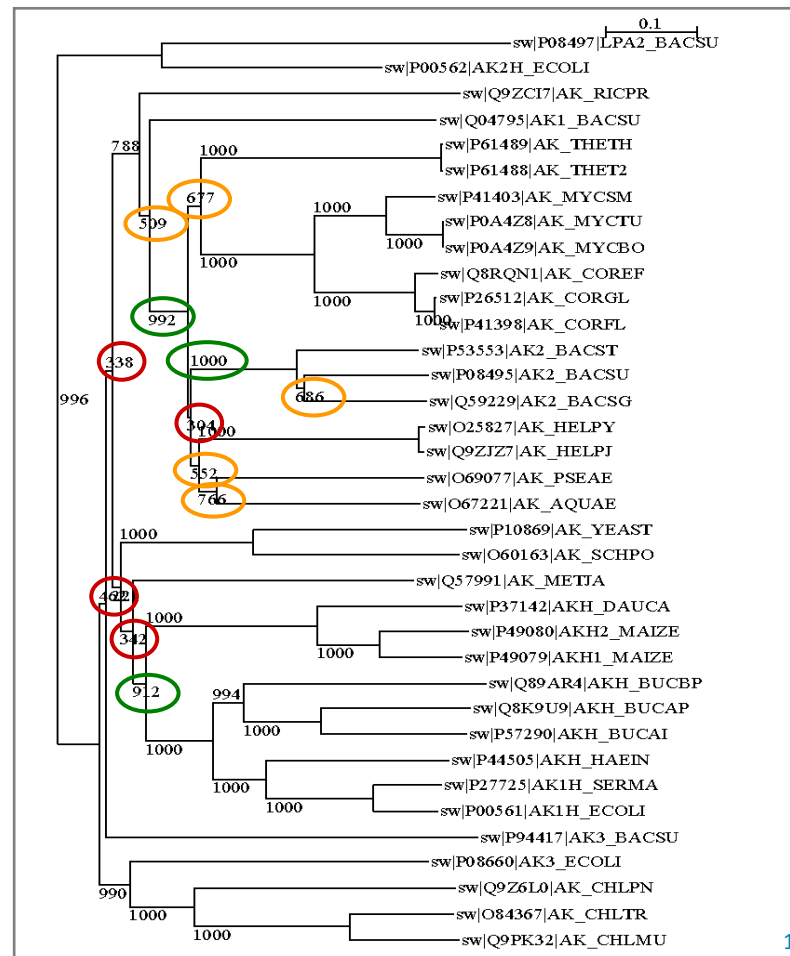
SeqA	T	T	G	C	A	A	T	T	T	A	A	T	A	A
SeqB	G	T	G	C	A	A	T	T	G	A	A	T	A	A
SeqC	C	G	G	C	A	A	G	T	C	A	A	G	A	A
SeqD	C	G	G	C	A	A	G	T	C	A	A	G	A	A
	2	9	7	4	1	8	9	3	2	1	8	9	5	1



JDD: Jeu de données; Figure adapté de cour de Céline Brochier

Bootstrapping

- Sur un arbre phylogénétique, une **valeur de bootstrap** est assignée à chaque branchement pour indiquer nombre de fois où ce branchement se retrouve à l'identique dans les N arbres de bootstrap.
- Une **valeur élevée** indique que le branchement est **robuste aux fluctuations d'échantillonnage**, et donc vraisemblablement fiable.
- Une **valeur faible** indique un branchement peu fiable.
 - Exemple : 338/1000 signifie que ce branchement n'est présent que dans ~1/3 des bootstraps ; il dépend donc fortement d'un sous-ensemble des colonnes plutôt que de représenter l'alignement complet.



Tutoriel et exercices

Ressources bioinformatiques utilisées

Nom	URL	Description
Base de données NCBI Taxonomy	https://www.ncbi.nlm.nih.gov/taxonomy/	Espèces et groupes taxinomiques avec leurs lignées
NCBI Taxonomy Common Tree	https://www.ncbi.nlm.nih.gov/Taxonomy/CommonTree/wwwcmt.cgi	Production d'un arbre pour une liste des espèces/groupes taxinomique
NGphylogeny	https://ngphylogeny.fr/	Inférence phylogénétique basée sur des séquences biologiques
phylogeny.fr	https://www.phylogeny.fr/ et http://phylogeny.lirmm.fr/	Inférence phylogénétique basée sur des séquences biologiques
ONEzoom	https://www.onezoom.org/	L'arbre du vivant interactif
Phylopic	https://www.phylopic.org/	Silhouettes d'organismes

Exercice 1. Arbre basé sur les connaissances préalables

Exercice 1. Arbre basé sur les connaissances préalables

- Au cours de cet exercice, vous allez vous familiariser avec le groupe que nous allons étudier, les Sauria. Le groupe des Sauria est un groupe monophylétique qui comprend les oiseaux et les reptiles.
- Nous allons utiliser les séquences de 10 espèces de Sauria, notre groupe d'étude et 2 espèces de mammifères. Ces derniers seront utilisés comme groupe extérieur (pour raciner l'arbre).

Nom d'espèce en latin	Nom français	Groupe taxinomique
<i>Serinus canaria</i>	Serin des Canaries	Aves
<i>Cygnus atratus</i>	Cygne noir	Aves
<i>Alligator mississippiensis</i>	Alligator d'Amérique	Crocodylia
<i>Crocodylus porosus</i>	Crocodile marin	Crocodylia
<i>Gavialis gangeticus</i>	Gavial du Gange	Crocodylia
<i>Chelonia mydas</i>	Tortue verte	Testudines
<i>Pelodiscus sinensis</i>	Tortue à carapace molle	Testudines
<i>Gekko japonicus</i>	Gecko	Squamata
<i>Lacerta agilis</i>	Lézard agile	Squamata
<i>Protobothrops mucrosquamatus</i>	Vipère à taches brunes	Squamata
<i>Gorilla gorilla</i>	Gorille de l'Ouest	Mammalia
<i>Orcinus orca</i>	Orque	Mammalia

Tableau 1. Liste des espèces étudiées.

Exercice 1. Arbre intuitif

- En vous basant sur ces illustrations et sur vos connaissances, dessinez un arbre phylogénétique **de manière intuitive** en incluant les espèces ci-contre.
- Pendant le TP, vous allez comparer votre arbre avec l'arbre phylogénétique de ces espèces tel qu'on le connaît aujourd'hui.

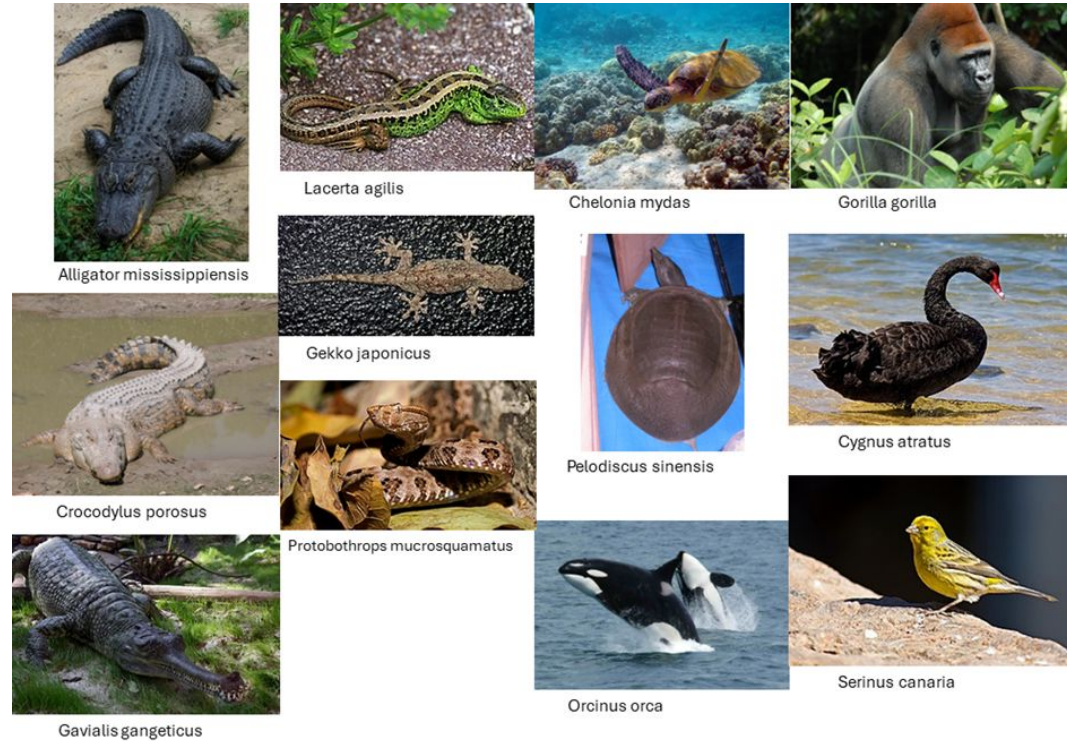


Figure 1. Photos des espèces étudiées.

Exercice 1. Analyse de l'arbre

- Familiarisez-vous avec le vocabulaire de la phylogénie utilisé pendant le TP.
- Comparez l'arbre ci- contre avec l'arbre que vous avez dessiné de manière intuitive.
- Décrivez les différences / ressemblances entre eux.

Sur Ametice, répondez au questionnaire 1

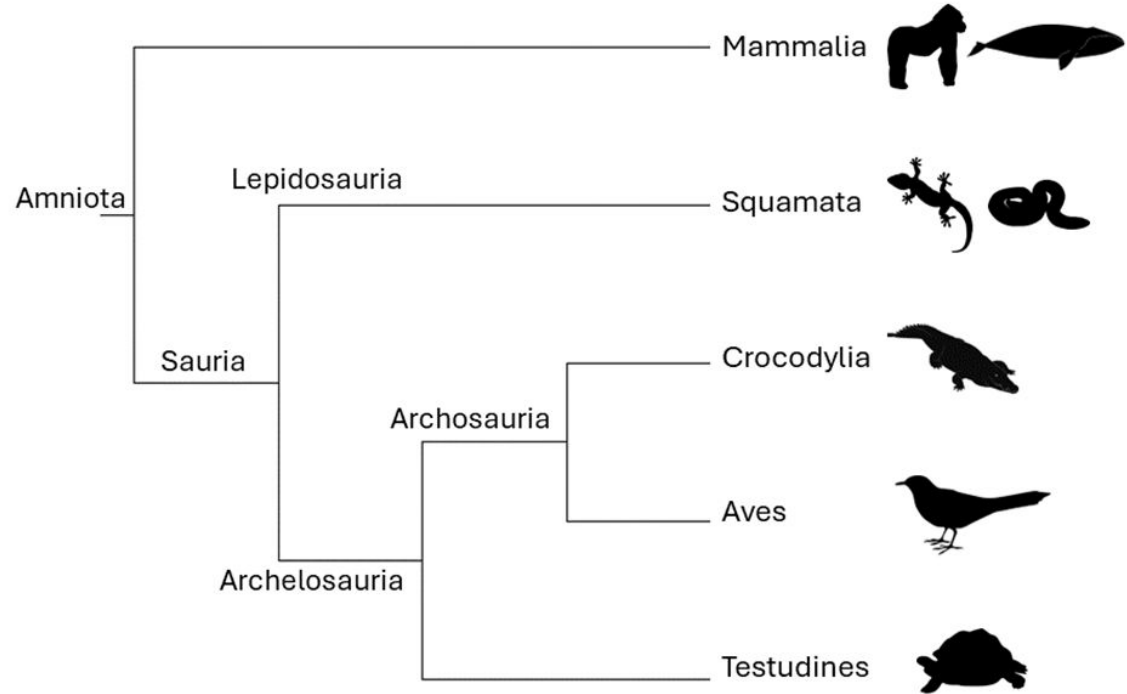


Figure 2. Arbre schématique des différents groupes étudiés pendant le TP

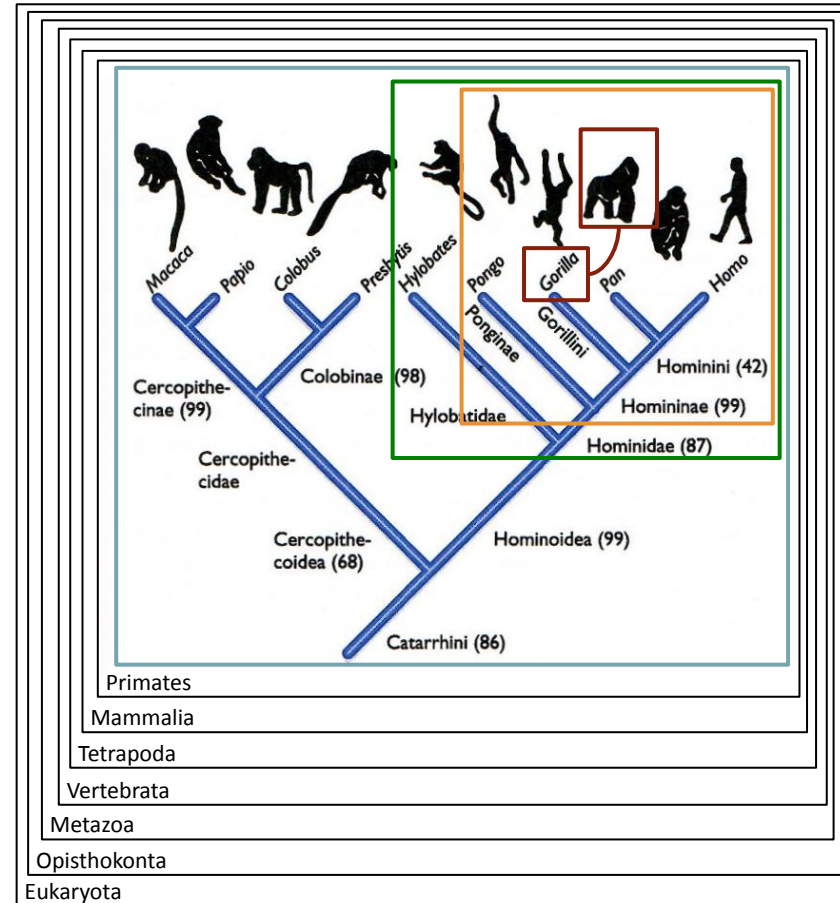
Exercice 2. Arbre de référence – NCBI Common Tree

Exercice 2. Arbre de référence – NCBI Common Tree

- Au cours de cet exercice, vous allez créer un **arbre de référence** contenant toutes les espèces choisies pour l'étude. Nous allons utiliser l'outil **Common Tree** de la base de données **NCBI Taxonomy**. Cette base de données contient toutes les espèces qui ont des séquences dans GenBank et leurs lignées taxinomiques.
- La **lignée** est une organisation hiérarchique. Le premier groupe est le plus large (Organismes cellulaires) et il contient le groupe suivant (Eukaryota) qui contient le groupe suivant (Opisthokonta), etc. Le gorille fait partie de chacun de ces groupes.

Exemple : lignée taxinomique du gorille (*Gorilla gorilla*) :

Cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Sarcopterygii; Dipnotetrapodomorpha; Tetrapoda; Amniota; Mammalia; Theria; Eutheria; Boreoeutheria; Euarchontoglires; Primates; Haplorrhini; Simiiformes; Catarrhini; Hominoidea; Hominidae; Homininae; Gorilla



Exercice 2. Arbre de référence – NCBI Common Tree

Lignée taxinomique du gorille (*Gorilla gorilla*) :

Cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Sarcopterygii; Dipnotetrapodomorpha; Tetrapoda; Amniota; Mammalia; Theria; Eutheria; Boreoeutheria; Euarchontoglires; Primates; Haplorrhini; Simiiformes; Catarrhini; Hominoidea; Hominidae; Homininae; Gorilla

Lignée du crocodile marin (*Crocodylus porosus*)

Cellular organisms ; Eukaryota ; Opisthokonta ; Metazoa ; Eumetazoa ; Bilateria ; Deuterostomia ; Chordata ; Craniata ; Vertebrata ; Gnathostomata ; Teleostomi ; Euteleostomi ; Sarcopterygii ; Dipnotetrapodomorpha ; Tetrapoda ; Amniota ; Sauropsida ; Sauria ; Archelosauria ; Archosauria ; Crocodylia ; Longirostres ; Crocodylidae ; Crocodylus

- Le gorille et le crocodile font tous deux partie des **Amniota**, mais par la suite leurs lignées divergent.
- Nous allons utiliser cette structure pour créer un **arbre de référence**.
- En réalité, on est rarement certain de l'histoire évolutive d'un groupe, mais comme la taxinomie du NCBI est basée sur les études scientifiques utilisant à la fois les caractères morphologiques et moléculaires, on peut considérer que cet arbre sera une bonne approximation de l'évolution des groupes étudiés.

Exercice 2. Arbre de référence – NCBI Common Tree

- Connectez-vous à [NCBI Taxonomy](https://www.ncbi.nlm.nih.gov/taxonomy).
- Cliquez [Common Tree](#). Deux possibilités vous sont offertes:
 - Télécharger le fichier avec la liste des espèces ([species_list.txt](#)),
 - Téléchargez-le sur NCBI à l'aide du bouton 'Choisir un fichier / Choose File', puis cliquez sur 'Add from file.'
 - Alternativement, vous pouvez entrer un par un des noms d'espèce dans la boîte 'Enter name or ID' et en cliquant sur 'Add' après avoir ajouté chaque nom.

Serinus canaria

Cygnus atratus

Alligator mississippiensis

Crocodylus porosus

Gavialis gangeticus

Chelonia mydas

Pelodiscus sinensis

Gekko japonicus

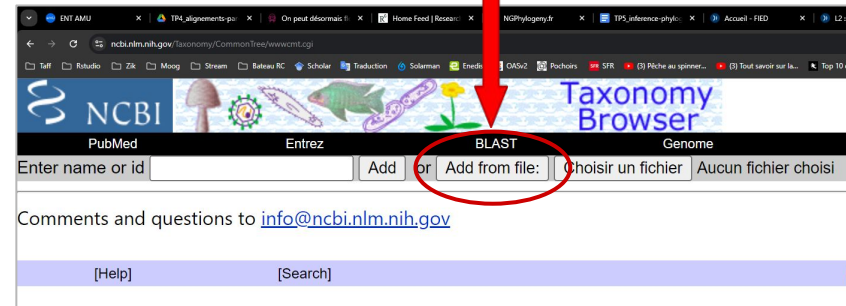
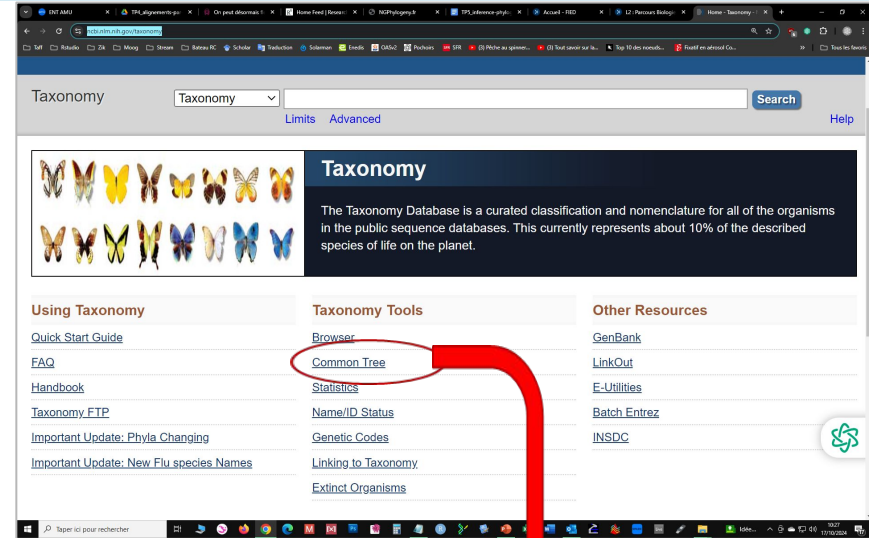
Lacerta agilis

Protobothrops

mucrosquamatus

Gorilla gorilla

Orcinus orca



Exercice 2. Arbre de référence – NCBI Common Tree

- Cochez la case ‘**include unranked (phylogenetic) taxa**’ pour afficher plus de niveaux taxinomiques intermédiaires.
- L’affichage qui apparaît sur l’écran n’est pas le plus simple à lire. Dessinez un arbre basé sur cette structure. Ajoutez après les noms d’espèces leurs groupes taxinomiques donnés dans le **Tableau 1**.
- Cet arbre sera utilisé par la suite de ce TP comme **arbre de référence** qui reflète l’évolution des Sauria tel qu’on les connaît maintenant.



Check Tax

Sur Ametice, répondez au questionnaire 2

Exercice 3.

Phylogénie moléculaire de la protéine RAG2 chez les Sauria

La suite logicielle la plus utilisée pour l'inférence logicielle, **Phylogeny.fr**, a été développée par le Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier (LIRMM) et est déployée à cette adresse : <https://phylogeny.lirmm.fr>.

Plus récemment, l'Institut Pasteur a développé une version "next generation" intitulée **NGPhylogeny.fr** : <https://ngphylogeny.fr/>. Cette plateforme présente des outils plus conviviaux pour l'affichage et l'analyse des alignements multiples et pour la visualisation des arbres phylogénétiques, mais en fonction de la charge du serveur, les temps d'attente sont parfois plus importants que sur la plateforme du LIRMM.

Ce diaporama contient les tutoriels pour les deux plateformes bioinformatiques.

Les deux plateformes donnent des résultats similaires et satisfaisants dans le cadre de ce TP.

En fonction de la disponibilité des serveurs durant la séance, les enseignants vous orienteront vers l'une ou l'autre.

- **Pour NGPhylogeny.fr, suivez les tuto des diapo "Exercice 3a"**
- **Pour Phylogeny.fr, suivez les tuto des diapo "Exercice 3b"**

En cas d'indisponibilité de NGPhylogeny.fr pendant le TP, les diapo "Exercice 3a" vous permettront, si vous le désirez de découvrir son interface pendant vos révisions, sans aucune obligation.

Au cours de cet exercice, vous utiliserez 12 séquences de la **protéine RAG2** pour inférer l'histoire évolutive des Sauria.

Vous allez travailler sur la plateforme NGphylogeny (ngphylogeny.fr) qui vous permet de construire un arbre phylogénétique à partir des séquences à l'aide d'un workflow qui associe différents logiciels.

Les 4 étapes de workflow sont les suivantes:

1. Alignement des séquences
2. Nettoyage (curation) de l'alignement
3. Inférence phylogénétique (construction de l'arbre)
4. Visualisation et édition de l'arbre

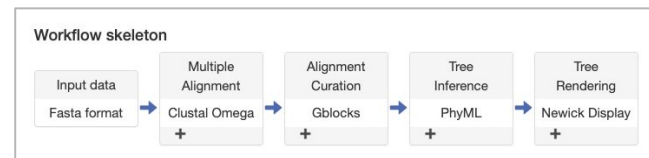
Pour chaque étape, vous avez le choix entre plusieurs logiciels ou algorithmes.

Exercice 3a. NGPhylogeny.fr. Phylogénie moléculaire de la protéine RAG2 chez les Sauria

- Téléchargez le fichier fasta [RAG2_protein.fas](#) contenant 12 séquences protéiques orthologues appartenant aux espèces de l'Exercice 2.
- Connectez-vous au site [NGphylogeny](#) pour effectuer une reconstruction phylogénétique basée sur ce jeu de données. La **phylogénie moléculaire** repose sur l'analyse des séquences biologiques (ADN, ARN, protéines) pour élaborer des arbres phylogénétiques en comparant les caractères moléculaires (ici les positions homologues de séquences).
- Cliquez sur l'option '**A la carte**'
- **Nommez** votre workflow (ex : 'RAG2prot').
Un workflow consiste à chaîner des outils logiciels modulaires qui réalisent les étapes successives d'une analyse.
- Créez un workflow en cochant les cases suivantes :
 - Multiple alignment: **MUSCLE**
 - Alignment curation: **Gblocks** (sélection des positions alignées de manière fiable)
 - Tree Inference: **PhyML**, qui construit un arbre phylogénétique sur base du maximum de vraisemblance (maximum likelihood)
 - Tree Rendering : **Newick** (visualisation et édition de l'arbre).
- Cliquez '**Create workflow**'.

The screenshot shows the NGPhylogeny.fr interface with the following elements:

- Name:** A text input field containing "RAG2_protein-Sauria", which is circled in red.
- Tools:** A section with four numbered icons representing the workflow steps: 1. Multiple Alignment (yellow circle with a grid), 2. Alignment Curation (blue circle with scissors), 3. Tree Inference (green circle with a tree), and 4. Tree Rendering (pink circle with an eye).
- Multiple Alignment:** A dropdown menu with "MUSCLE" selected and circled in red.
- Alignment Curation:** A dropdown menu with "Gblocks" selected and circled in red.
- Tree Inference:** A list of options including "FastME", "TNT", "PhyML", "MrBayes", and "trimAl". "PhyML" is selected and circled in red.
- Tree Rendering:** A dropdown menu with "Newick Display" selected and circled in red.



Reconstruction phylogénétique

- Sur la page suivante, vous pouvez soit charger le fichier à l'aide de bouton **Choose file**, soit copier le contenu du fichier RAG2_protein.fas dans la boîte de texte au bas de la section **Input data**.

Sur cette même page, vous pouvez modifier les paramètres de chacun des logiciels choisis. Nous allons garder les paramètres par défaut pour toutes les étapes sauf pour celle de reconstruction phylogénétique:

- Cliquez sur la boîte grisée **PhyML**
- Dans le menu '**Statistical test for branch support**' sélectionnez l'option 'Bootstrap', et indiquez une valeur de 100 pour l'option **Number of bootstrap replicates**.
- Cliquez sur 'Submit'



Configure your workflow

Clustal Omega
Gblocks
PhyML
Newick Display

Input data

Choose a file or Paste content
(Fasta format with more than 3 sequences)

Choose file RAG2_protein.fas

Blast range
--

Files in session
--



PhyML

Statistical test for branch support

Bootstrap

Use aLRT or aBayes to save computing time.

Number of bootstrap replicates

100

Must be a positive integer

Exécution de l'analyse

Au bout de quelques secondes, NGPhylogeny affiche une page avec le statut de réalisation des différentes tâches.

Les tâches s'affichent de bas en haut, par ordre d'exécution.

La colonne **Status** indique le statut de chaque tâche:

- accomplie
- en cours d'exécution
- en attente des résultats de l'étape précédente

Dès qu'une étape est terminée, ses résultats peuvent être consultés sans attendre la finalisation des étapes suivantes.

Astuce: cliquez droit (Ctrl-clic) sur les boutons des étapes accomplies pour ouvrir les résultats dans un autre onglet. Ceci vous permettra de revenir ultérieurement sur la page de suivi des tâches.

The screenshot shows the NGPhylogeny.fr History page. The table lists tasks with columns: Tool, Step, File Name, and Status. The tasks are grouped by tool: Newick Display, PhyML, Gblocks, Clustal Omega, and Upload File. The Status column shows the progress of each task.

Tool	Step	File Name	Status
Newick Display	15.	All tree images	En attente
	14.	Tree image	En attente
PhyML	13.	Mapping between short sequence to and names (used to interpret some bootstrap log files if any)	Travail en cours
	12.	PhyML Newick tree	Travail en cours
	11.	PhyML Statistics	Travail en cours
	10.	PhyML log	Travail en cours
	9.	PhyML bootstrap trees: align.phy_phyml_boot_trees.txt	Travail en cours
	8.	Booster: Tree with [id]avg transfer distances[dep] as branch labels: tbe_raw_tree.nhx	Travail en cours
	7.	Booster: Tree with normalized supports: tbe_norm_tree.nhx	Travail en cours
	6.	Booster: tbe_log.txt	Travail en cours
Gblocks	5.	Gblocks Sequences information	Tâches accomplies
	4.	Gblocks Cleaned sequences	Tâches accomplies
	3.	Gblocks log	Tâches accomplies
Clustal Omega	2.	alignment	Tâches accomplies
	1.	RAG2_protein.fas	Tâches accomplies

References of tools to cite

bioRxiv | text

Visualisation de l'alignement

Une fois la page de résultats affichée, vous pouvez visualiser le résultat de chaque étape.

- Observez les alignements en cliquant sur le bouton **'MSAViewer'** en fin de ligne
 - Résultat de MUSCLE (2. Muscle alignment)
 - Alignement “nettoyé” par Gblocks (5. Gblocks Cleaned sequences).
- **Gardez ces fenêtres ouvertes**, pour pouvoir répondre au questionnaire un peu plus tard.

Astuce: cliquez droit (Ctrl-clic) sur les boutons des étapes accomplies pour ouvrir les résultats dans un autre onglet. Ceci vous permettra de revenir ultérieurement sur la page de suivi des tâches et sur chaque page de résultat.

Tool	Step	File Name	Status
Newick Display	15.	All tree images	✓
	14.	Tree image	✓
	13.	Mapping between short sequence id and names (useful to interpret some bootstrap log files if any)	✓
	12.	PhyML Newick tree	✓
	11.	PhyML Statistics	✓
PhyML	10.	PhyML log	✓
	9.	PhyML bootstrap trees: align.phy_phyml_boot_trees.txt	✓
	8.	Booster: Tree with [ldjavg transfer distances[depth] as branch labels: tbe_raw_tree.nhx	✓
	7.	Booster: Tree with normalized supports: tbe_norm_tree.nhx	✓
	6.	Booster: tbe_log.txt	✓
Gblocks	5.	Gblocks Sequences information	✓
	4.	Gblocks Cleaned sequences	✓
	3.	Gblocks log	✓
Clustal Omega	2.	alignment	✓
	1.	RAG2_protein.fas	✓

Alignement nettoyé par Gblocks

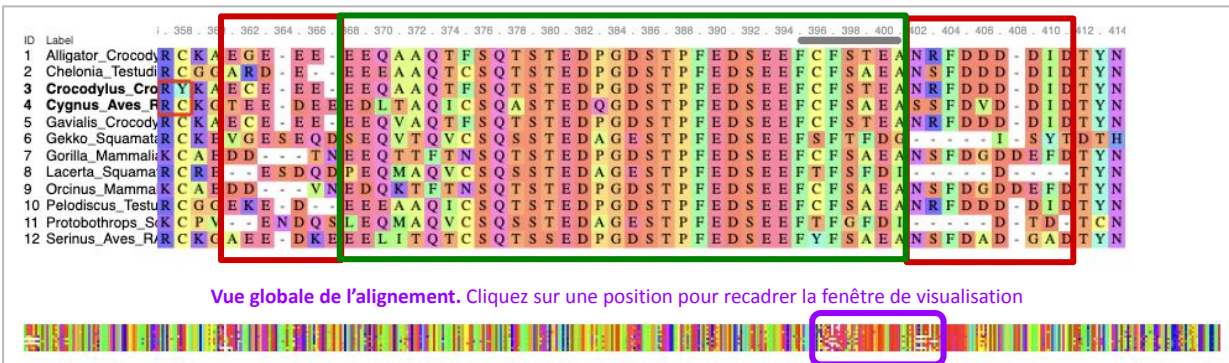
Alignement produit par MUSCLE

Visualisation de l'alignement

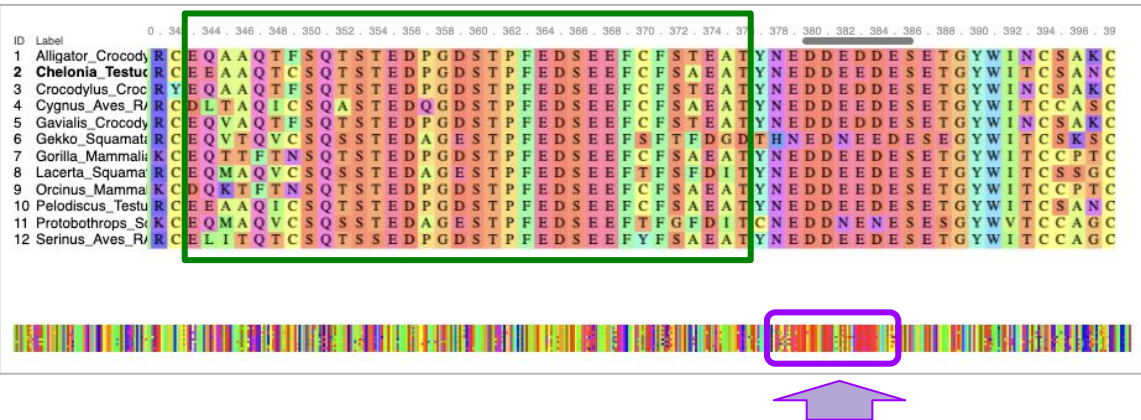
Comparez les deux alignements (longueur de l'alignement, distribution des gaps, des régions conservées...).

- Quelles différences voyez-vous entre les deux alignements ?
- Quel est l'intérêt de l'étape de curation ?

Alignement multiple produit par MUSCLE



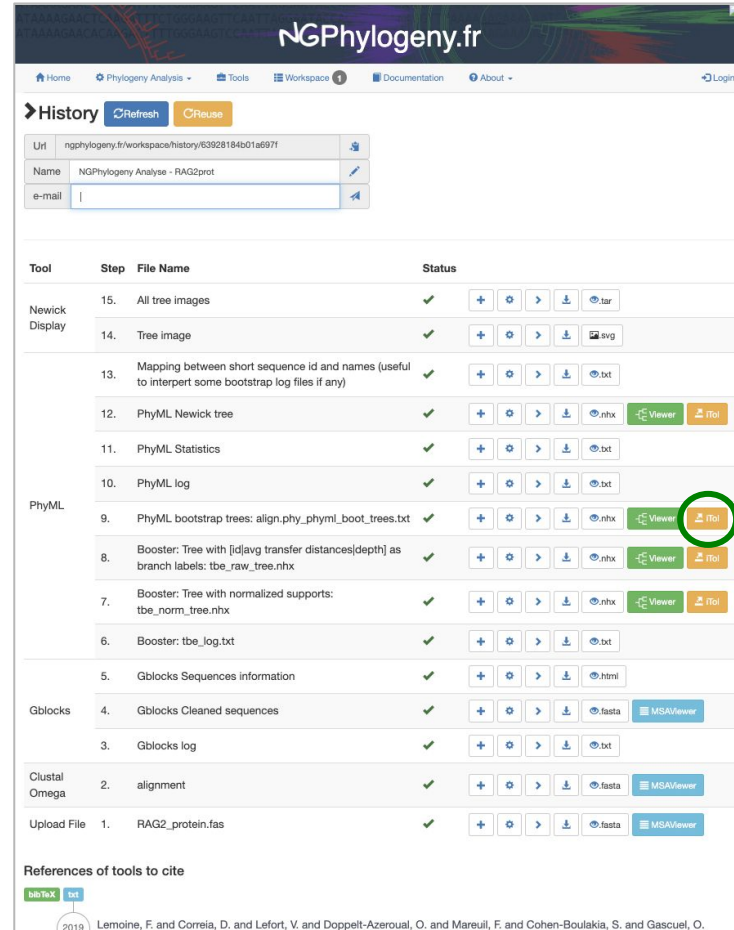
Alignement nettoyé par Gblocks



Exercice 3a. NGPhylogeny.fr. Phylogénie moléculaire de la protéine RAG2 chez les Sauria

Visualisation et édition de l'arbre

- Revenez à la **fenêtre de statut**
 - Cliquez sur le lien du viewer **iTol**.
- Vous pouvez à présent éditer votre arbre.



The screenshot shows the NGPhylogeny.fr interface. The 'History' tab is active, displaying a table of analysis steps. A green circle highlights the 'iTol' viewer link for step 9, 'PhyML bootstrap trees: align.phy_phyml_boot_trees.txt'. A green arrow points from the text 'Arbre phylogénétique' to this link.

Tool	Step	File Name	Status	Actions
Newick Display	15.	All tree images	✓	+ ⚙️ ▶️ ⬅️ ⬇️ ⬆️
	14.	Tree image	✓	+ ⚙️ ▶️ ⬅️ ⬇️ ⬆️
PhyML	13.	Mapping between short sequence id and names (useful to interpret some bootstrap log files if any)	✓	+ ⚙️ ▶️ ⬅️ ⬇️ ⬆️
	12.	PhyML Newick tree	✓	+ ⚙️ ▶️ ⬅️ ⬇️ ⬆️
	11.	PhyML Statistics	✓	+ ⚙️ ▶️ ⬅️ ⬇️ ⬆️
	10.	PhyML log	✓	+ ⚙️ ▶️ ⬅️ ⬇️ ⬆️
	9.	PhyML bootstrap trees: align.phy_phyml_boot_trees.txt	✓	+ ⚙️ ▶️ ⬅️ ⬇️ ⬆️
	8.	Booster: Tree with [ldjavg transfer distances[depth] as branch labels: tbe_raw_tree.nhx	✓	+ ⚙️ ▶️ ⬅️ ⬇️ ⬆️
	7.	Booster: Tree with normalized supports: tbe_norm_tree.nhx	✓	+ ⚙️ ▶️ ⬅️ ⬇️ ⬆️
	6.	Booster: tbe_log.txt	✓	+ ⚙️ ▶️ ⬅️ ⬇️ ⬆️
Gblocks	5.	Gblocks Sequences information	✓	+ ⚙️ ▶️ ⬅️ ⬇️ ⬆️
	4.	Gblocks Cleaned sequences	✓	+ ⚙️ ▶️ ⬅️ ⬇️ ⬆️
	3.	Gblocks log	✓	+ ⚙️ ▶️ ⬅️ ⬇️ ⬆️
Clustal Omega	2.	alignment	✓	+ ⚙️ ▶️ ⬅️ ⬇️ ⬆️
Upload File	1.	RAG2_protein.fas	✓	+ ⚙️ ▶️ ⬅️ ⬇️ ⬆️

References of tools to cite

bioRxiv 2019 Lemoine, F. and Correia, D. and Lefort, V. and Doppelt-Azeroual, O. and Mareuil, F. and Cohen-Boulakia, S. and Gascuel, O.

Enracinement de l'arbre

Cette étape est la plus importante.

- Notre groupe d'étude est le clade des **Sauria**.
- Les deux séquences de mammifères (gorille et orque) nous servent de **groupe extérieur** (outgroup), car nous savons que les mammifères ne font pas partie des Sauria, mais sont phylogénétiquement proches.
- Ceci nous indique que la racine de l'arbre devrait être placée entre les Sauria et ces mammifères.

Pour raciner votre arbre correctement,

- Cliquez sur la **branche** qui sépare *Gorilla* et *Orcinus* des autres espèces.
- Sans le menu '**Tree structure**', sélectionnez la fonction '**Re-root the tree here**'.

The screenshot shows the ITOL (Interactive Tree Of Life) web interface. At the top, there is a navigation bar with 'ITOL INTERACTIVE TREE OF LIFE', 'Tree of Life', 'Upload', 'Data sharing', and 'Help'. Below this, a phylogenetic tree is displayed with a 'Tree scale: 0.1' indicator. A red arrow points to a specific branch on the tree, which is the branch separating the mammalian clade (Gorilla and Orcinus) from the Sauria clade. A context menu is open over this branch, showing various 'Clade functions' and 'Editing' options. The 'Tree structure' option is selected, and a sub-menu is visible with the following options: 'Re-root the tree here' (highlighted with a red circle), 'Delete only this node', 'Delete the whole clade', 'Move the clade', and 'Root the tree at midpoint'. To the right of the tree, a list of species names and their corresponding RAG2-XP sequence identifiers is shown, including Gorilla, Orcinus, Cygnus, Serinus, Alligator, Crocodylus, Gavialis, Pelodiscus, Chelonia, Gekko, Protobothrops, and Lacerta.

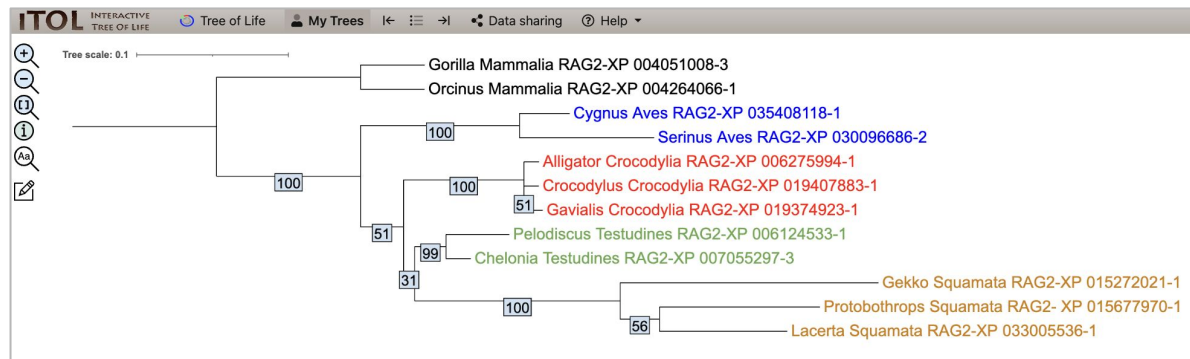
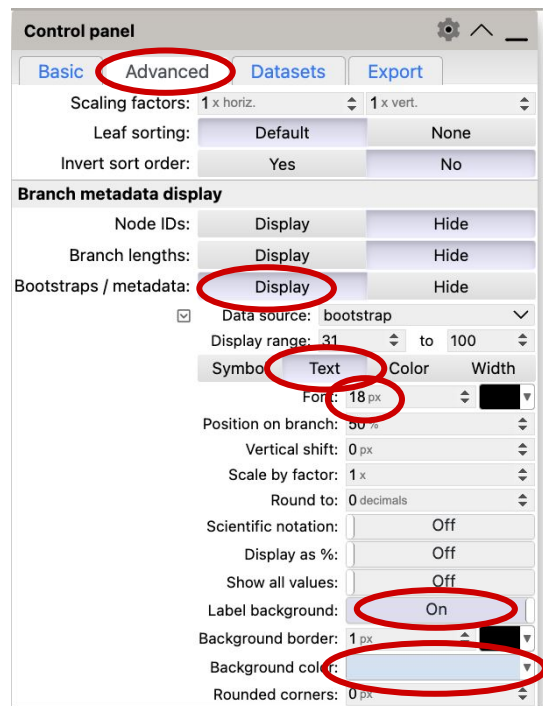
Visualisation et édition de l'arbre

- On peut également modifier la couleur des branches et des noms d'espèce: dans l'onglet '**Basics**', sélectionnez '**at tips**' dans '**Label options**'. Ceci repositionne les étiquettes des feuilles (noms d'espèces) sans changer la structure de l'arbre.
- En cliquant sur les branches adéquates, coloriez les différents groupes avec les couleurs suivantes (Labels → Whole clade → Color):
 - Aves > bleu
 - Crocodylia > rouge
 - Testudines > vert
 - Squamata > orange

The screenshot displays the ITOL web interface. At the top, there's a navigation bar with 'Tree of Life', 'Upload', 'Data sharing', and 'Help'. The main area shows a phylogenetic tree with species names like 'Gorilla Mammalia RAG2-XP 004051008-3' and 'Ornithus Aves RAG2-XP 03009686-2'. A red arrow points to the 'Clade functions' menu, which is open, showing options like 'Collapse clade', 'Rotate children', and 'Add clade to pruned tree'. The 'Labels' option is highlighted with a red circle. Below the tree, a 'Control panel' is visible, with tabs for 'Basic', 'Advanced', 'Datasets', and 'Export'. The 'Basic' tab is active, showing various settings. The 'Branch metadata display' section has 'Bootstraps / metadata' set to 'Display'. The 'Tree scales' section has 'Internal tree scale' and 'Tree scale box' set to 'Display'. The 'Node options' section has 'Leaf node symbols' and 'Internal node symbols' set to 'Display'. The 'Other functions' section has 'Label functions' set to 'Multi-style' and 'Auto assign taxonomy' set to 'NCBI'. A color selection dialog is open, showing a color palette and a 'Whole clade' selection. The 'Color' dropdown is set to 'blue'.

Visualisation et édition de l'arbre

- Affichez également les valeurs statistiques pour chaque nœud de l'arbre:
 - Advanced > Bootstraps/metadata > text**
 - Ajustez l'affichage des valeurs statistiques (taille de police, décimal, position...).



Sur Ametice, répondez au questionnaire 3

Au cours de cet exercice, vous utiliserez 12 séquences de la **protéine RAG2** pour inférer l'histoire évolutive des Sauria.

Vous allez travailler sur la plateforme phylogeny.fr du LIRMM (phylogeny.lirmm.fr/phylo.cgi/alacarte.cgi) qui vous permet de construire un arbre phylogénétique à partir des séquences à l'aide d'un workflow qui associe différents logiciels.

Les 4 étapes de workflow sont les suivantes:

1. Alignement des séquences
2. Nettoyage (curation) de l'alignement
3. Inférence phylogénétique (construction de l'arbre)
4. Visualisation et édition de l'arbre

Pour chaque étape, vous avez le choix entre plusieurs logiciels ou algorithmes.

Exercice 3b. **Phylogeny.fr**. Phylogénie moléculaire de la protéine RAG2 chez les Sauria

- Téléchargez le fichier fasta [RAG2_protein.fas](#) contenant 12 séquences protéiques orthologues appartenant aux espèces de l'Exercice 2.
- Allez sur la page [Phylogénie 'A la carte'](#) du site phylogeny.fr pour effectuer une reconstruction phylogénétique basée sur ce jeu de données.

La **phylogénie moléculaire** repose sur l'analyse des séquences biologiques (ADN, ARN, protéines) des organismes vivants pour élaborer des arbres phylogénétiques en comparant les caractères moléculaires (ici les positions homologues de séquences).

- Nommez votre '**workflow**' (ex: 'RAG2_prot_sauria').

Un workflow consiste à chaîner des outils logiciels modulaires qui réalisent les étapes successives d'une analyse.

- Créez un workflow en cochant les cases suivants:
 - Multiple alignment: **MUSCLE**
 - Alignment curation: **Gblocks** (sélection des positions alignées de manière fiable)
 - Construction of phylogenetic tree : **PhyML** (Maximum Likelihood: méthode du maximum de vraisemblance)
 - Visualisation of phylogenetic tree : **Treedyn** : visualisation et édition de l'arbre.
- Cliquez '**Create workflow**'

"A la Carte" Mode

Alignment MUSCLE → Curation Gblocks → Phylogeny PhyML → Tree Rendering TreeDyn

1. Workflow Setup

Workflow Settings

Name of the analysis (optional):

Choose processing steps to run and select software to use:

☒ Multiple Alignment:

- ☒ MUSCLE
- ☐ 3DCoffee
- ☐ ClustalW

☒ Alignment curation:

- ☒ Gblocks
- ☐ positions with gaps

☒ Construction of phylogenetic tree:

- ☒ Maximum Likelihood
- ☐ TNT
- ☐ Distances
- ☐ ProtDist/FastDist + BioNJ
- ☐ ProtDist/FastDist + Neighbor

☒ Visualisation of phylogenetic tree:

- ☒ TreeDyn
- ☐ Dendrogram
- ☐ Drawtree

Run workflow:

- ☒ all at once
- ☐ step by step

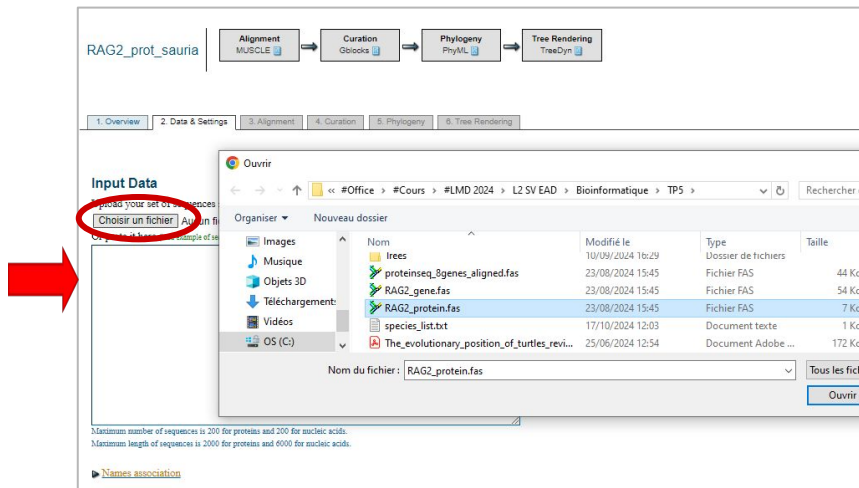
Reconstruction phylogénétique

- Sur la page suivante, vous pouvez soit charger le fichier à l'aide de bouton '**Choose file**, soit **coller le contenu du fichier RAG2_protein.fas** dans la fenêtre de la section '**Input data**'.

Sur cette même page, vous pouvez modifier les paramètres de chacun des logiciels choisis. Nous allons garder les paramètres par défaut par tout sauf pour l'étape de reconstruction phylogénétique:

- Dans la section '**Phylogeny: PhyML**', au bas du formulaire sélectionnez l'option '**Approximate Likelihood-Ratio (aLRT)**', et sélectionnez '**SH-like**' dans le menu déroulant
- Remplissez la case '**email address**'
- Cliquez sur '**Submit**'
- Cliquez sur le lien (**link**) qui apparaît au bas du formulaire.

Astuce: le lien vous permettra de suivre en direct la réalisation des étapes successives du workflow, et d'afficher les résultats des premières étapes pendant que les dernières sont en cours d'exécution.



Phylogeny: PhyML

Settings

Statistical tests for branch support:

☒ Approximate Likelihood-Ratio Test (aLRT): **SH-like**

☐ Bootstrapping procedure: Number of bootstraps: **100**

Substitution model: **Default**

The job is processing, you will be sent an email when the job will be completed.

You can also bookmark **this link** for results.

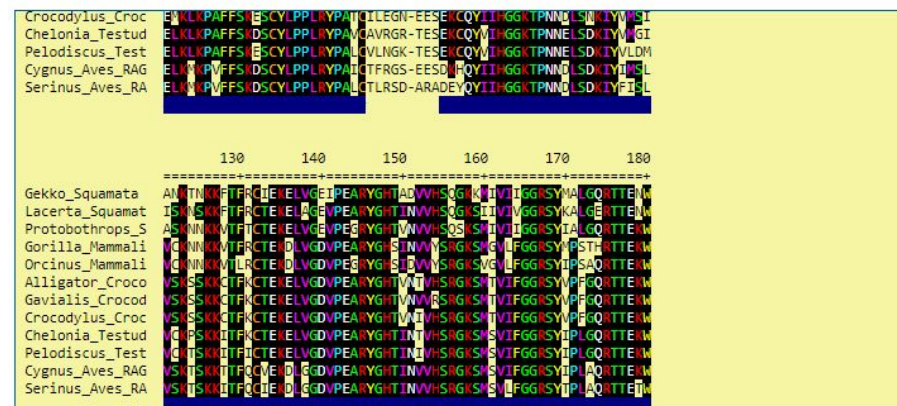
Visualisation de l'alignement

Une fois la page de résultats affichée, vous pouvez visualiser le résultat de chaque étape grâce aux différents onglets.

- Observez l'alignement produit par MUSCLE (**Alignment**) et l'alignement “nettoyé” par Gblocks (**Curation**) en cliquant sur le bouton '**MSAviewer**' en fin de ligne.
- **Gardez ces fenêtres ouvertes**, pour pouvoir répondre au questionnaire un peu plus tard.
- Comparez les deux alignements (longueur de l'alignement, distribution des gaps, des régions conservées...).
- Quelles différences voyez-vous entre les deux alignements ?
- Quel est l'intérêt de l'étape de curation ?



Curation results



Visualisation et édition de l'arbre

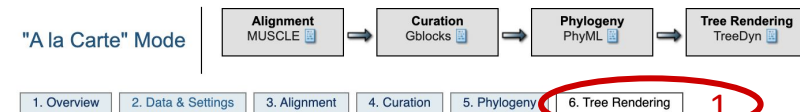
Vous pouvez à présent éditer l'arbre phylogénétique dans l'onglet '**Tree rendering**'

Enracinement de l'arbre

Cette étape est la plus importante. Notre groupe d'étude est le clade des **Sauria**. Les deux séquences de mammifères (gorille et orque) nous servent de **groupe extérieur** (outgroup), car nous savons que les mammifères ne font pas partie des Sauria, mais sont phylogénétiquement proches. Ceci nous indique que la racine de l'arbre devrait être placée entre les Sauria et ces mammifères.

Pour enraciner votre arbre

1. Cliquez sur l'onglet '**6. Tree rendering**'
2. Cliquez sur l'icône **Reroot (outgroup)**
3. Cliquez sur la **branche** qui sépare Gorilla et Orcinus des autres espèces.



Tree Rendering results

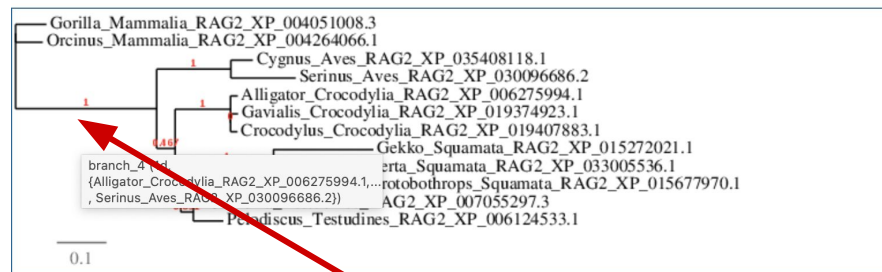


Figure 1: Phylogenetic tree (the branch length is proportional to the number of substitutions per site).

3. Cliquez ici

Dynamic Tree Edition

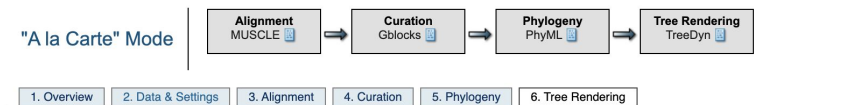
<input checked="" type="checkbox"/> Color	<input checked="" type="checkbox"/> leaf using color <input type="text" value="blue"/>	Tree manipulation:
<input checked="" type="checkbox"/> branch	and assign the group name <input type="text"/>	
<input type="checkbox"/> Flip subtree		<input type="button" value="Reset to original tree"/>
<input type="checkbox"/> Change leaf name		<input type="button" value="Reroot using mid-point rooting"/>
		<input checked="" type="button" value="Reroot (outgroup) 2"/>
		<input type="button" value="Swap subtrees"/>
		<input type="text" value="red"/>

Visualisation et édition de l'arbre

On peut également modifier la couleur des branches et des noms d'espèce.

Astuce : il est parfois difficile de cliquer sur les branches très courtes. Pour contourner cela, cochez l'option "ignore branch length", qui remplace le phylogramme par un cladogramme. Le cladogramme permet également de mieux afficher les valeurs de bootstrap.

- Cochez la case **Ignore branch length**
- Dans le menu 'Color', sélectionnez 'Leaf' et 'Branch'.
- Choisissez la couleur bleue
- Cliquez sur la branche qui sépare les oiseaux des autres groupes.
- Faites de même avec les autres clades pour obtenir les couleurs suivantes.
 - Aves > bleu
 - Crocodylia > rouge
 - Testudines > vert
 - Squamata > orange
- Ajustez l'affichage des valeurs statistiques (taille de police, décimales, position...).
- Expérimentez pour découvrir les autres modifications d'affichage.



Tree Rendering results

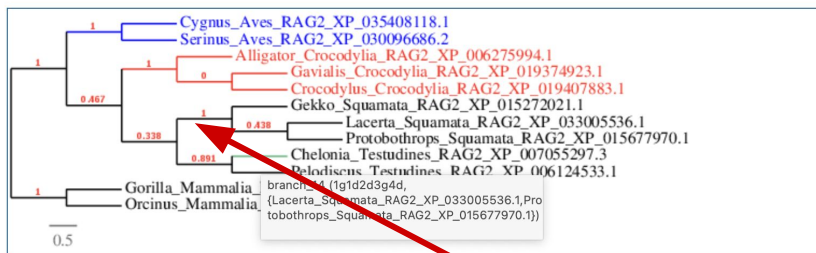


Figure 1: Phylogenetic tree (the branch length is proportional to the number of substitutions per site).

4. Cliquez ici

Dynamic Tree Edition

2 **Color** leaf using color orange 3 **Tree manipulation**

and assign the group name

Flip subtree

Change leaf name

Reset to original tree

Reroot using mid-point rooting

Reroot (outgroup)

Swap subtrees

Add annotations using color red

Display:

Branch annotation: Branch support values Branch length values Use color: red

Legend at position 25 130 Update Use color: dimgray

Ignore branch length 1

Exercice 4. Concaténation des séquences protéiques de 8 gènes différents chez les Sauria

Exercice 4. Concaténation des séquences protéiques de 8 gènes différents chez les Sauria

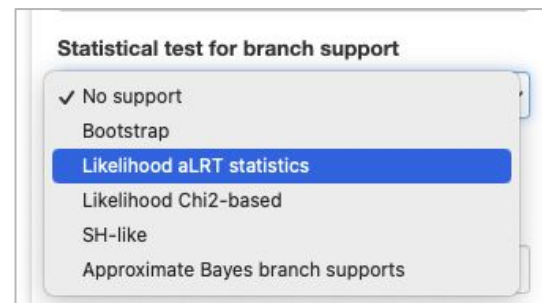
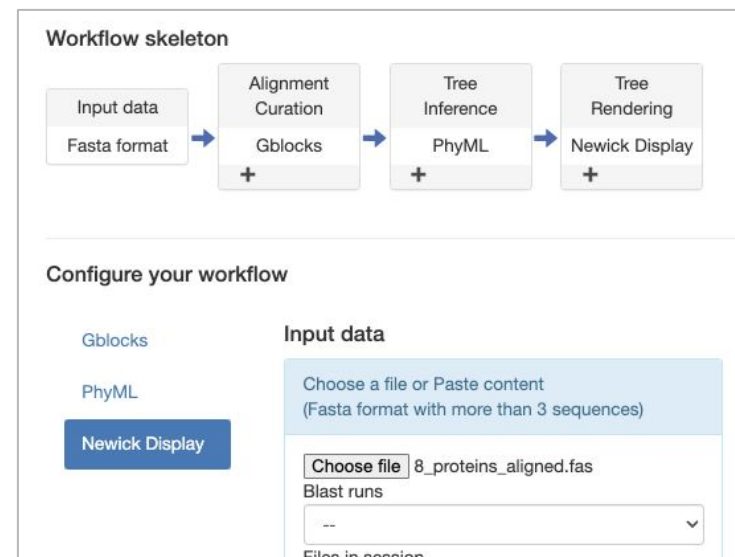
En phylogénie moléculaire, la concaténation des séquences consiste à assembler plusieurs séquences macromoléculaire provenant de différents gènes ou protéines en une seule séquence continue pour chaque organisme étudié. Cette approche est généralement utilisée pour inférer un arbre des espèces à partir de données moléculaires, en augmentant la quantité d'information disponible. L'intérêt principal de la concaténation est de fournir une meilleure résolution et une plus grande robustesse aux analyses phylogénétiques, en réduisant les biais et les erreurs qui pourraient survenir si l'on utilisait des gènes individuels séparément.

- Téléchargez le fichier fasta [8 proteins aligned.fas](#) construit à partir de la concaténation de 8 protéines différentes (BDNF, RAG-1, RAG-2, PDC, ND1, COX1, CYTB, ATP6) appartenant aux espèces traitées dans les questions 2 et 3. Pour créer ce fichier nous avons aligné des séquences protéine par protéine, puis nous avons concaténé les séquences alignées.

Exercice 4a. NGPhylogeny.fr. Concaténation des séquences protéiques de 8 gènes différents chez les Sauria

Créez un arbre PhyML à partir de ce nouveau jeu de données

- Ouvrez une connexion au site [NGphylogeny](https://ngphylogeny.fr),
- Cliquez l'option '**A la carte**'
- Créez un workflow nommé 'Concatprot'.
 - Multiple Alignment: **attention, ne sélectionnez aucun logiciel d'alignement à l'étape de création de workflow, car les séquences sont déjà alignées**
 - Alignment curation: **Gblocks** (sélection des positions informatives dans l'alignement)
 - Tree Inference: **PhyML** (construction phylogénétique en maximum de vraisemblance, *Maximum Likelihood*)
 - Tree rendering: **Newick Display** (visualisation+édition de l'arbre)
- Cliquez sur '**Create workflow**'.
- Sur la page suivante, vous pouvez soit charger le fichier à l'aide de bouton '**Choisir un fichier**', soit copier le contenu du fichier [8_proteins_aligned.fas](#) dans la fenêtre de la section '**Input data**'.
- Sur cette même page, déroulez le menu '**PhyML**' et dans le champ '**Statistical test for branch support**' sélectionnez l'option '**Approximate Likelihood-Ratio (aLRT) - SH-like**'.
- Cliquez sur '**Submit**'
- Visualisez l'arbre sur iTol. Racinez l'arbre sur les séquences de mammifères comme décrit dans l'exercice précédent.
- **Sur Ametice, répondez au questionnaire 4**



Exercice 4b. **Phylogeny.fr**. Concaténation des séquences protéiques de 8 gènes différents chez les Sauria

- Créez un arbre PhyML à partir de ce nouveau jeu de données
- Ouvrez une connexion sur la **plateforme phylogeny.fr** du LIRMM (phylogeny.lirmm.fr/phylo_cgi/alacarte.cgi).
- Créez un **workflow** nommé 'Concatprot_sauria' en sélectionnant
 - Multiple Alignment: **attention, désactivez la case "Multiple alignment", car les séquences fournies sont déjà alignées**
 - Alignment curation: **Gblocks** (sélection des positions informatives dans l'alignement)
 - Tree Inference: **PhyML** (construction phylogénétique en maximum de vraisemblance, Maximum Likelihood)
 - Tree rendering: **Treedyn** (visualisation+édition de l'arbre)
- Cliquez sur '**Create workflow**'.
- Sur la page suivante, vous pouvez soit copier le contenu du fichier [8_proteins_aligned.fas](#) dans la fenêtre de la section '**Input data**' soit charger le fichier à l'aide de bouton '**Choisir un fichier**'.
- Sur cette même page, dans le menu '**Settings**' de '**Phylogeny: PhyML**', assurez-vous que sélectionnez l'option '**bootstrapping procedure**', et indiquez une valeur de 100.
- Cliquez sur '**Submit**'
- **Racinez l'arbre** sur les séquences de mammifères comme décrit dans l'exercice précédent.
- **Sur Ametice, répondez au questionnaire 4**

"A la Carte" Mode

Curation Gblocks → Phylogeny PhyML → Tree Rendering TreeDyn

1. Workflow Setup

Workflow Settings

Name of the analysis (optional): Concatprot_sauria

Choose processing steps to run and select software to use:

☐ Multiple Alignment:

- ☐ MUSCLE
- ☐ T-Coffee
- ☐ 3DCoffee
- ☐ ClustalW

☒ Alignment curation:

- ☒ Gblocks
- ☐ Remove positions with gaps

☒ Construction of phylogenetic tree:

- ☒ Maximum likelihood
- ☐ Parsimony
- ☐ TNT

Distances

- ☐ ProtDist/FastDist + BioNJ
- ☐ ProtDist/FastDist + Neighbor

☒ Visualization of phylogenetic tree:

- ☒ TreeDyn
- ☐ Cytoscape
- ☐ Drawtree

Run workflow:

☒ all at once ☐ step by step

Create workflow

Debriefing post-TP

Au terme de ce TP, vous devriez avoir acquis les connaissances suivantes

- Concepts de base concernant la taxinomie des organismes
- Représentations arborées des relations phylogénétiques
- Typologie des groupes d'organismes par rapport un arbre phylogénétique (groupes monophylétique, paraphylétique, polyphylétique)
- Processus de construction d'un arbre de référence à partir d'une base de données taxinomique (NCBI).
- Importance de la curation (nettoyage) en phylogénie moléculaire pour améliorer la qualité de l'alignement et la robustesse des arbres phylogénétiques.
- Approche de concaténation de séquences pour inférer un arbre des espèces sur base d'analyses plus robustes.

Vous aurez également acquis les compétences suivantes

- Utiliser une plateforme web bioinformatique pour inférer un arbre phylogénique sur base de données de séquence.
- Modifier les paramètres d'affichage d'un arbre phylogénétique pour obtenir une représentation pertinente
- Interpréter, décrire et comparer des arbres phylogénétiques.

Dans ce cours d'introduction, nous n'avons pas traité les aspects suivants, qui pourront être couverts dans des UE de L3 ou de Master.

Connaissances

- Approches algorithmiques pour inférer un arbre à partir d'un alignement multiple : UPGMA, Neighbour Joining, maximum de vraisemblance (Maximum Likelihood, ML), parcimonie
- Modèles probabilistes d'évolution des séquences nucléiques

Compétences

- Choisir un algorithme et un modèle probabiliste en fonction des données utilisées
- Evaluer l'impact des choix algorithmiques sur la topologie des arbres phylogénétiques