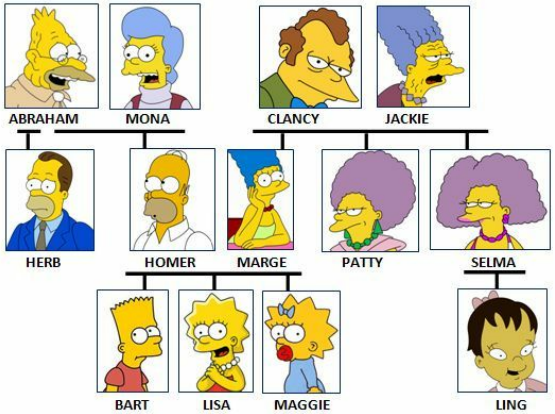
Familles de gènes, aspects phylogénétiques, problématiques bioinformatiques

Jean-François Dufayard
Équipe "Intégration des Données"

Famille de gènes ?
Ou "Il était une fois la phylogénie"...

?

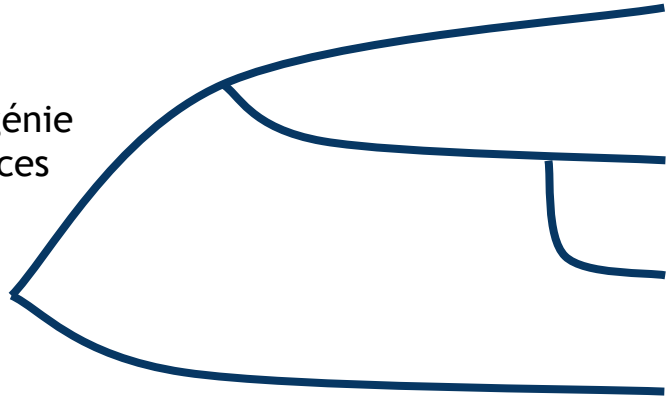


?

Il était une fois la phylogénie

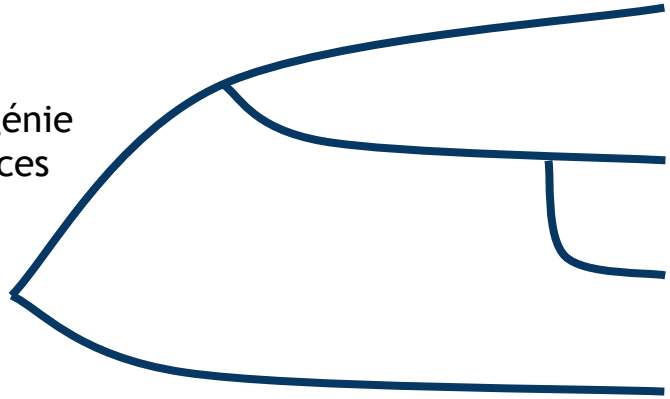
Histoire évolutive des espèces / classification des espèces

Phylogénie d'espèces

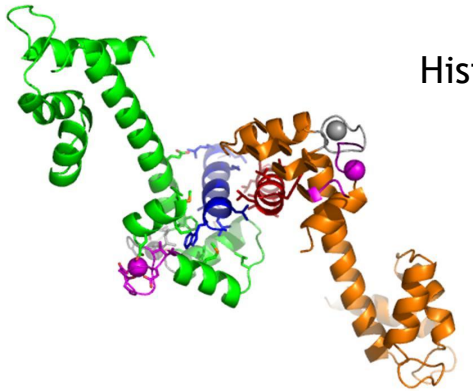


Il était une fois la phylogénie

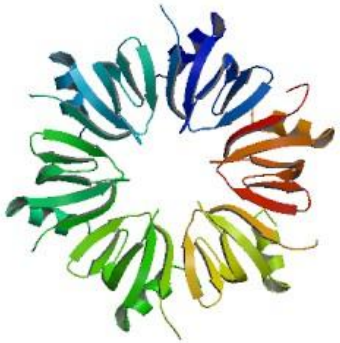
Histoire évolutive des espèces / classification des espèces

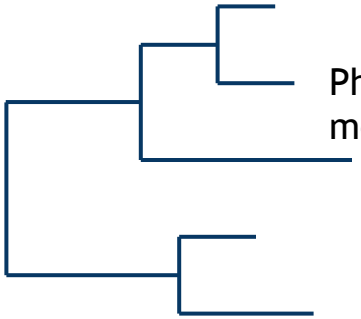


Phylogénie d'espèces



Histoire évolutive des molécules

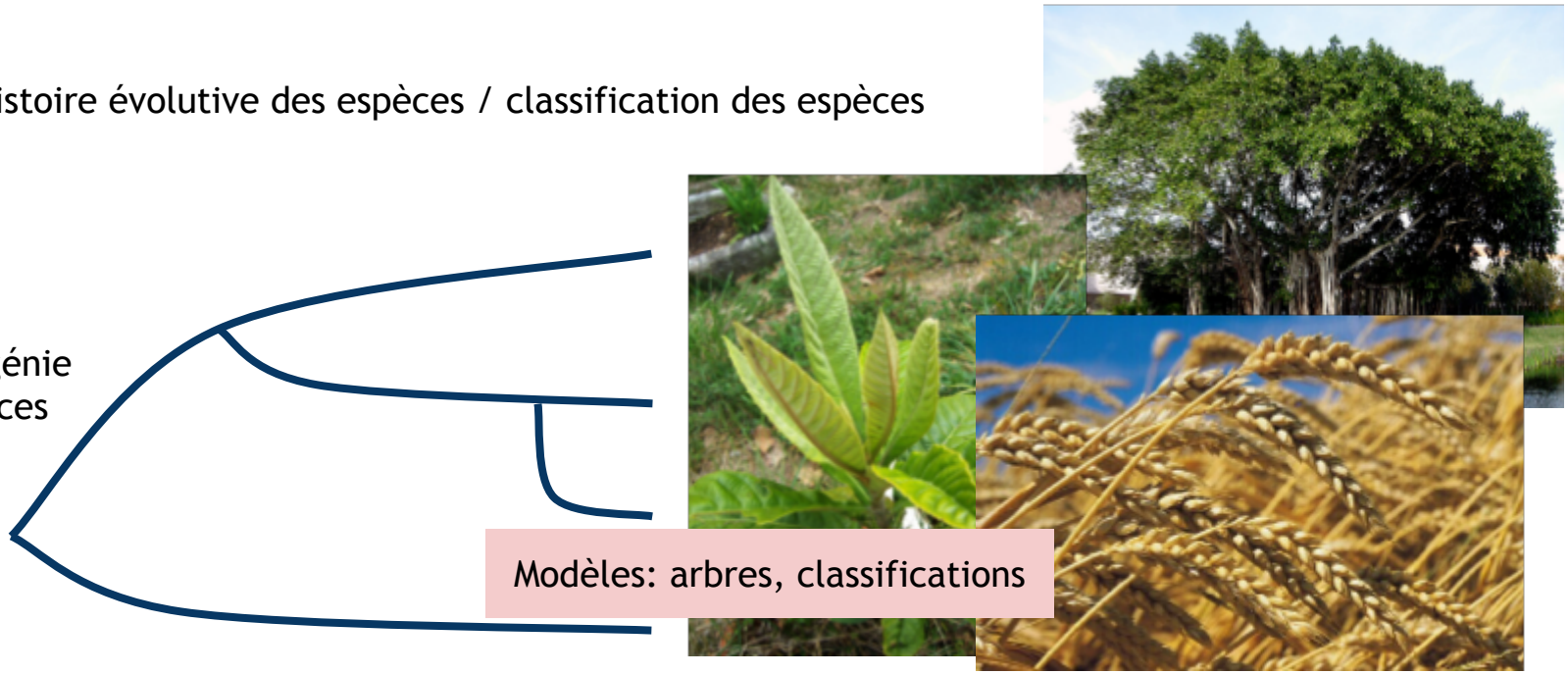


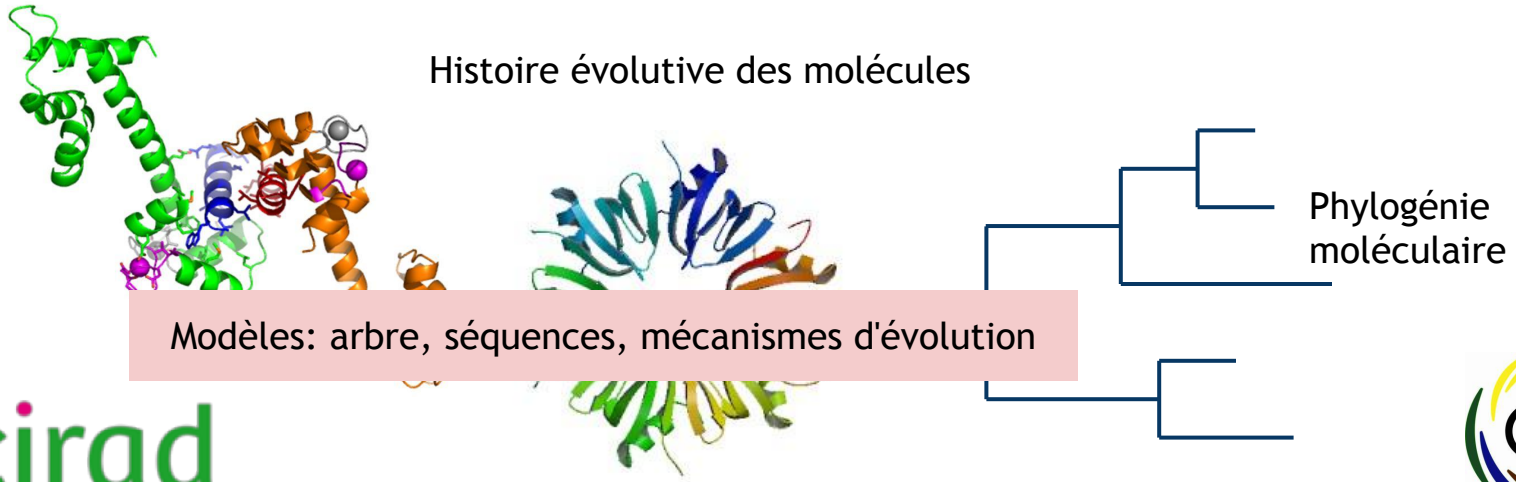


Phylogénie moléculaire

Il était une fois la phylogénie

Histoire évolutive des espèces / classification des espèces



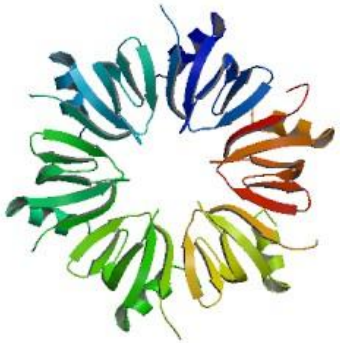


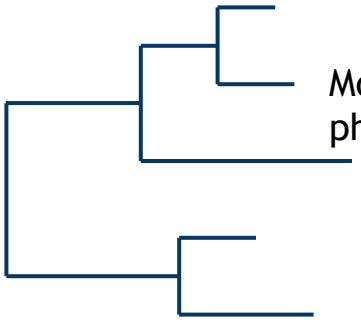
Famille de gènes ? Le point de vue biologique...

Famille de gènes = Famille de gènes homologues

- Une famille de gènes homologues sont un groupe de gènes dont on suppose l'existence d'un gène ancêstral commun.
- Deux gènes sont homologues s'ils ont un ancêtre commun.

Trivialement: on ne peut construire une phylogénie que pour une famille de gènes homologues..





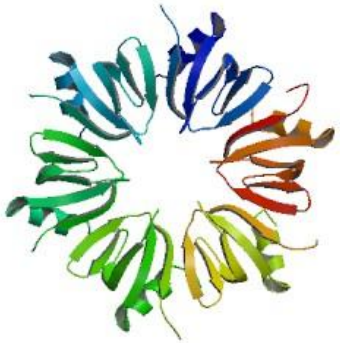
Famille de gènes ? Le point de vue bioinformatique...

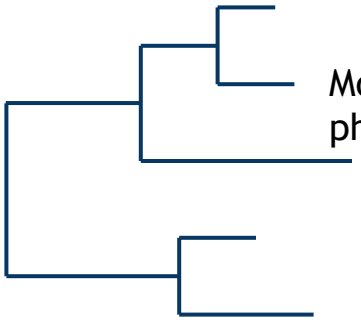
Famille de gènes ~ Un groupe de séquences similaires

- Similarité: mesure mathématique de ressemblance entre deux séquences comparables.
- Couverture: pourcentage de la longueur d'une séquence le long de laquelle elle est comparable à une seconde.

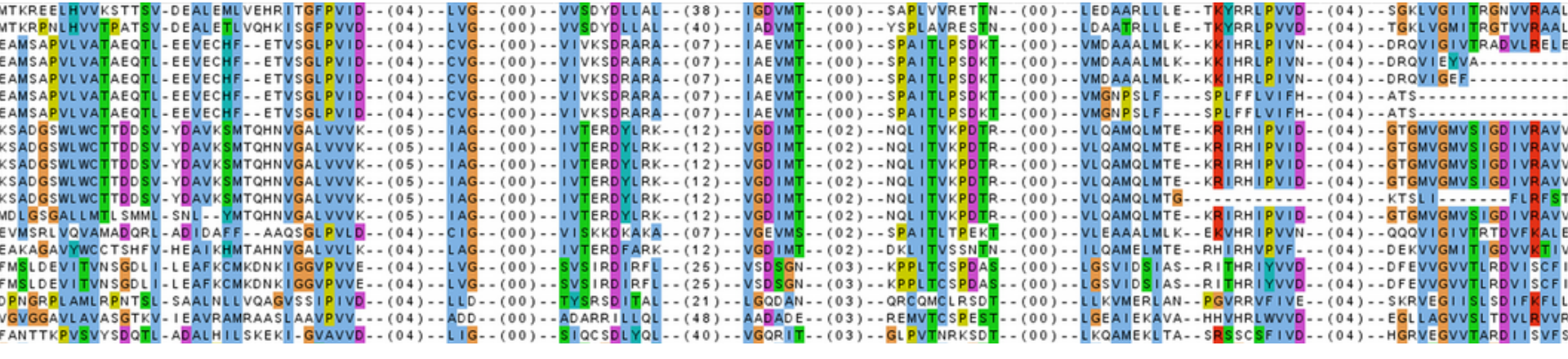
Un gène est modélisé par sa séquence.
L'hypothèse d'homologie entre deux gènes repose sur la similarité et la couverture des deux séquences.

```
---------------TCGAACTC---AAGTGAGATAACTTGTTATGAAAGGCAAAAGT
ATTCGTTATGCAAGTCCAAATCTGCAAGTGAAATAACTTGGTATGGAAGGCAAACGT
```





Comparaison de séquences, un problème trompeusement simple...



Aligner deux séquences: l'opération élémentaire

Sujet:
AAGTGAGATAAC
AAGAAATAAC

Algorithme: Needleman & Wunsch
Alignement global, programmation dynamique

Aligner deux séquences: l'opération élémentaire

Sujet:
AAGTGAGATAAC
AAGAAATAAC

Similarité (1 or -1)

Gap sur la séquence 1 (-1)

Gap sur la séquence 2 (-1)

	A	A	G	T	G	A	G	A	T	A	A	C
A												
A												
G												
A												
A												
A												
T												
A												
A												
C												

cirad

agap

Aligner deux séquences: l'opération élémentaire

Sujet:
AAGTGAGATAAC
AAGAAATAAC

Similarité (1 or -1)

Gap sur la séquence 1 (-1)

Gap sur la séquence 2 (-1)

	A	A	G	T	G	A	G	A	T	A	A	C
A	0											
A												
G												
A												
A												
A												
T												
A												
A												
C												

cirad

agap

Aligner deux séquences: l'opération élémentaire

Sujet:
AAGTGAGATAAC
AAGAAATAAC

Similarité (1 or -1)

Gap sur la séquence 1 (-1)

Gap sur la séquence 2 (-1)

	A	A	G	T	G	A	G	A	T	A	A	C
A	0	-1										
A												
G												
A												
A												
A												
T												
A												
A												
C												

cirad

agap

Aligner deux séquences: l'opération élémentaire

Sujet:
AAGTGAGATAAC
AAGAAATAAC

↖ Similarité (1 or -1)

↑ Gap sur la séquence 1 (-1)

← Gap sur la séquence 2 (-1)

	A	A	G	T	G	A	G	A	T	A	A	C
A	0 ←	-1 ←	-2 ←	-3 ←	-4 ←	-5 ←	-6 ←	-7 ←	-8 ←	-9 ←	-10 ←	-11
A												
G												
A												
A												
A												
T												
A												
A												
C												

cirad

agap

Aligner deux séquences: l'opération élémentaire

Sujet:
AAGTGAGATAAC
AAGAAATAAC

Similarité (1 or -1)

Gap sur la séquence 1 (-1)

Gap sur la séquence 2 (-1)

	A	A	G	T	G	A	G	A	T	A	A	C
A	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11
A	-1											
G												
A												
A												
A												
T												
A												
A												
C												

cirad

agap

Aligner deux séquences: l'opération élémentaire

Sujet:
AAGTGAGATAAC
AAGAAATAAC

Similarité (1 or -1)

Gap sur la séquence 1 (-1)

Gap sur la séquence 2 (-1)

	A	A	G	T	G	A	G	A	T	A	A	C
A	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11
A	-1	1										
G												
A												
A												
A												
T												
A												
A												
C												

cirad

agap

Aligner deux séquences: l'opération élémentaire

Sujet:
AAGTGAGATAAC
AAGAAATAAC

Similarité (1 or -1)

Gap sur la séquence 1 (-1)

Gap sur la séquence 2 (-1)

	A	A	G	T	G	A	G	A	T	A	A	C
A	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11
A	-1	1	0	-1	-2	-3	-4	-5	-6	-7	-8	-9
G	-2	0	2	1	0	-1	-2	-3	-4	-5	-6	-7
A	-3	-1	1	3	2	1	0	-1	-2	-3	-4	-5
A	-4	-2	0	2	2	1	2	1	0	-1	-2	-3
A	-5	-3	-1	1	1	1	1	1	0	-1	-2	-3
T	-6	-4	-2	0	0	0	2	1	2	1	0	-1
A	-7	-5	-3	-1	-1	-1	1	1	0	3	2	1
A	-8	-6	-4	-2	-2	-2	0	0	2	2	4	3
C	-9	-7	-5	-3	-3	-3	-1	-1	1	1	3	5

Aligner deux séquences: l'opération élémentaire

Sujet:
AAGTGAGATAAC
AAG--AAATAAC

Similarité (1 or -1)

Gap sur la séquence 1 (-1)

Gap sur la séquence 2 (-1)

	A	A	G	T	G	A	G	A	T	A	A	C
A	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11
A	-1	1	0	-1	-2	-3	-4	-5	-6	-7	-8	-9
G	-2	0	2	1	0	-1	-2	-3	-4	-5	-6	-7
A	-3	-1	1	3	2	1	0	-1	-2	-3	-4	-5
A	-4	-2	0	2	2	1	2	1	0	-1	-2	-3
A	-5	-3	-1	1	1	1	1	1	0	-1	-2	-3
T	-6	-4	-2	0	0	0	2	1	2	1	0	-1
A	-7	-5	-3	-1	-1	-1	1	1	0	3	2	1
A	-8	-6	-4	-2	-2	-2	0	0	2	2	4	3
C	-9	-7	-5	-3	-3	-3	-1	-1	1	1	3	5

Aligner deux séquences: l'opération élémentaire

Sujet:
AAGTGAGATAAC
AA--GAAATAAC

Similarité (1 or -1)

Gap sur la séquence 1 (-1)

Gap sur la séquence 2 (-1)

	A	A	G	T	G	A	G	A	T	A	A	C
A	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11
A	-1	1	0	-1	-2	-3	-4	-5	-6	-7	-8	-9
G	-2	0	2	1	0	-1	-2	-3	-4	-5	-6	-7
A	-3	-1	1	3	2	1	0	-1	-2	-3	-4	-5
A	-4	-2	0	2	2	1	2	1	0	-1	-2	-3
A	-5	-3	-1	1	1	1	1	1	0	-1	-2	-3
T	-6	-4	-2	0	0	0	2	1	2	1	0	-1
A	-7	-5	-3	-1	-1	-1	1	1	0	3	2	1
A	-8	-6	-4	-2	-2	-2	0	0	2	2	4	3
C	-9	-7	-5	-3	-3	-3	-1	-1	1	1	3	5

Aligner deux séquences: l'opération élémentaire

Sujet:
AAGTGAGATAAC
AA--GAAATAAC

Similarité (1 or -1)

Gap sur la séquence 1 (-1)

Gap sur la séquence 2 (-1)

	A	A	G	T	G	A	G	A	T	A	A	C
A	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11
A	-1	1	0	-1	-2	-3	-4	-5	-6	-7	-8	-9
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
A	-5	-3	-1	1	1	1	1	1	0	-1	-2	-3
T	-6	-4	-2	0	0	0	2	1	2	1	0	-1
A	-7	-5	-3	-1	-1	-1	1	1	0	3	2	1
A	-8	-6	-4	-2	-2	-2	0	0	2	2	4	3
C	-9	-7	-5	-3	-3	-3	-1	-1	1	1	3	5

Complexité: $O(n^2)$ ... Assez rapide, même sur de longues séquences.

Calcul l'alignement optimal, <u>d'un point de vue informatique</u>.

Aligner un ensemble de séquences (n>2) - en pratique

Étendre l'algorithme de NW à plus de 2 séquences
Méthode de l'alignement progressif

Calcul d'une matrice de distances par paires → Calcul d'un arbre approximatif → Alignement des séquences le long de cet arbre

Aligner un ensemble de séquences (n>2) - en pratique

Étendre l'algorithme de NW à plus de 2 séquences
Méthode de l'alignement progressif

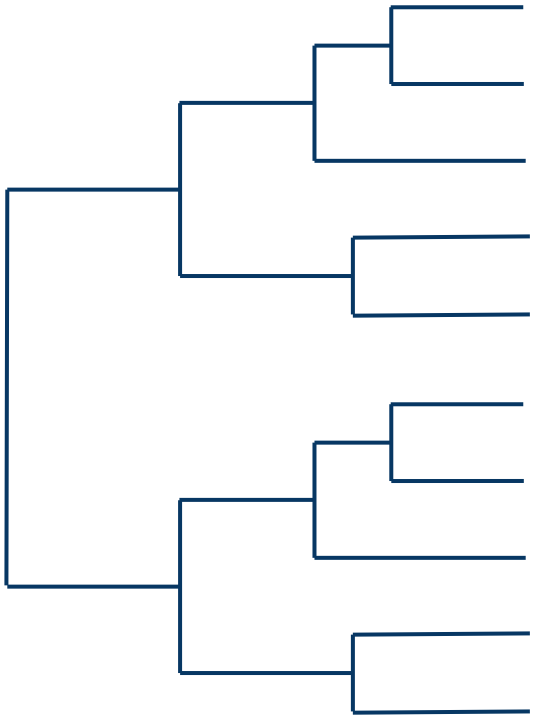
Calcul d'une matrice de distances par paires → Calcul d'un arbre approximatif → Alignement des séquences le long de cet arbre

SFM137	0.211	SFM137																				
SFW113	0.213	0.003	SFW113																			
SFW112	0.213	0.003	0.000	SFW112																		
SFW80	0.214	0.002	0.001	0.001	SFW80																	
SFW79	0.212	0.002	0.000	0.000	0.000	SFW79																
SFW78	0.214	0.003	0.002	0.000	0.001	0.001	SFW78															
SFW77	0.216	0.005	0.002	0.002	0.003	0.002	0.002	SFW77														
SMMY45	0.228	0.175	0.178	0.178	0.177	0.176	0.177	0.181	SMMY45													
SMMY46	0.226	0.180	0.182	0.182	0.182	0.181	0.182	0.186	0.003	SMMY46												
SL126	0.238	0.167	0.172	0.172	0.172	0.171	0.172	0.177	0.163	0.161	SL126											
SO125	0.239	0.168	0.166	0.166	0.165	0.165	0.166	0.170	0.156	0.158	0.144	SO125										
SBM131	0.238	0.162	0.168	0.168	0.166	0.166	0.168	0.170	0.160	0.161	0.154	0.112	SBM131									
SBM132	0.236	0.164	0.168	0.168	0.166	0.166	0.168	0.170	0.160	0.161	0.151	0.112	0.008	SBM132								
SBB117	0.231	0.168	0.172	0.172	0.170	0.171	0.172	0.174	0.167	0.168	0.161	0.112	0.017	0.018	SBB117							
SBB118	0.231	0.168	0.172	0.172	0.170	0.171	0.172	0.174	0.167	0.168	0.161	0.112	0.017	0.018	0.000	SBB118						
SMN37	0.228	0.150	0.155	0.155	0.153	0.153	0.154	0.157	0.161	0.163	0.161	0.117	0.066	0.067	0.066	0.066	SMN37					
SMN38	0.230	0.160	0.164	0.164	0.163	0.163	0.164	0.167	0.168	0.167	0.161	0.123	0.068	0.069	0.070	0.070	0.015	SMN38				
SMN39	0.228	0.159	0.163	0.163	0.161	0.162	0.163	0.165	0.167	0.166	0.157	0.122	0.068	0.069	0.070	0.070	0.015	0.002	SMN39			
SMN40	0.230	0.160	0.164	0.164	0.163	0.163	0.164	0.167	0.168	0.167	0.161	0.123	0.068	0.069	0.070	0.070	0.015	0.000	0.002	SMN40		
SMM133	0.220	0.153	0.158	0.158	0.156	0.156	0.157	0.160	0.163	0.161	0.153	0.120	0.070	0.072	0.073	0.073	0.017	0.021	0.021	0.021	SMM133	
SMM134	0.219	0.155	0.159	0.159	0.157	0.158	0.159	0.161	0.161	0.160	0.153	0.120	0.070	0.072	0.073	0.073	0.017	0.021	0.021	0.021	0.002	

Aligner un ensemble de séquences (n>2) - en pratique

Étendre l'algorithme de NW à plus de 2 séquences
Méthode de l'alignement progressif

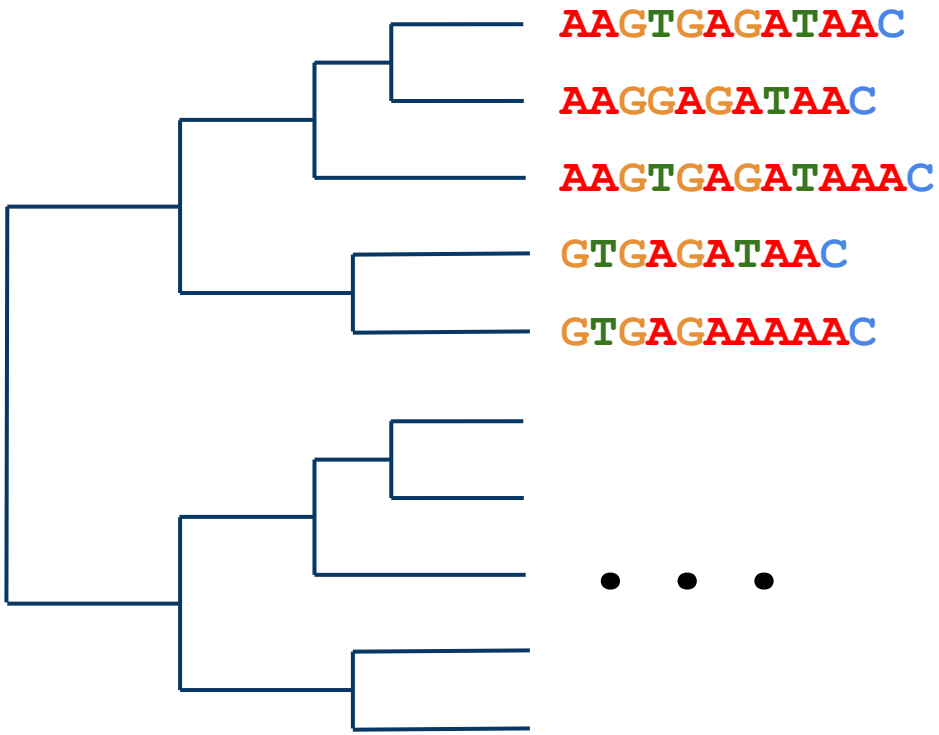
Calcul d'une matrice de distances par paires → **Calcul d'un arbre approximatif** → Alignement des séquences le long de cet arbre



Aligner un ensemble de séquences (n>2) - en pratique

Étendre l'algorithme de NW à plus de 2 séquences
Méthode de l'alignement progressif

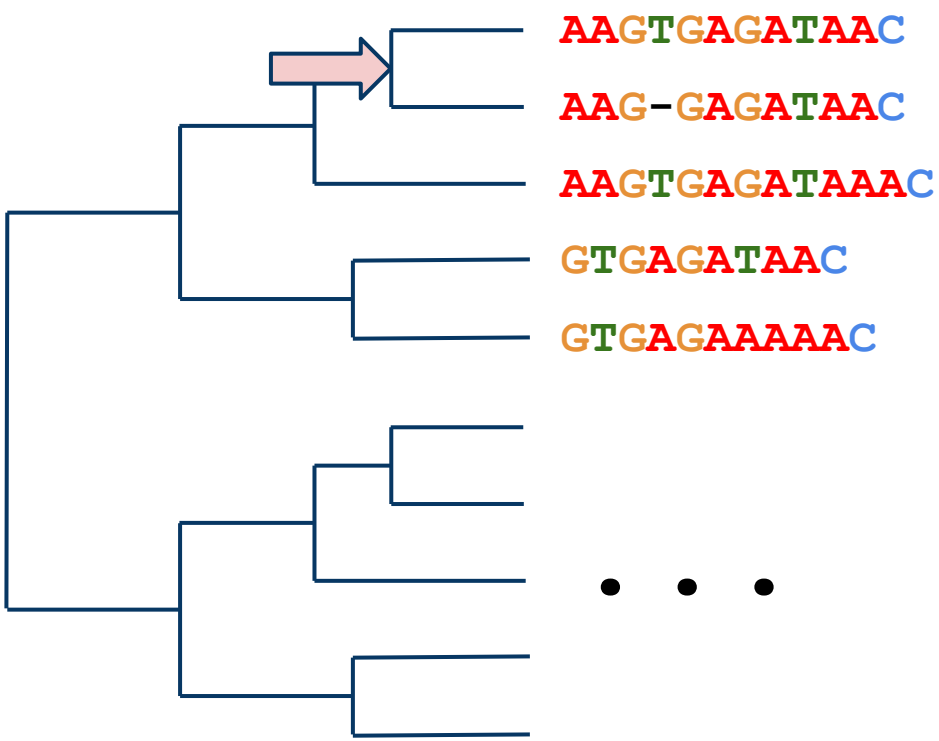
Calcul d'une matrice de distances par paires → Calcul d'un arbre approximatif → **Alignement des séquences le long de cet arbre**



Aligner un ensemble de séquences (n>2) - en pratique

Étendre l'algorithme de NW à plus de 2 séquences
Méthode de l'alignement progressif

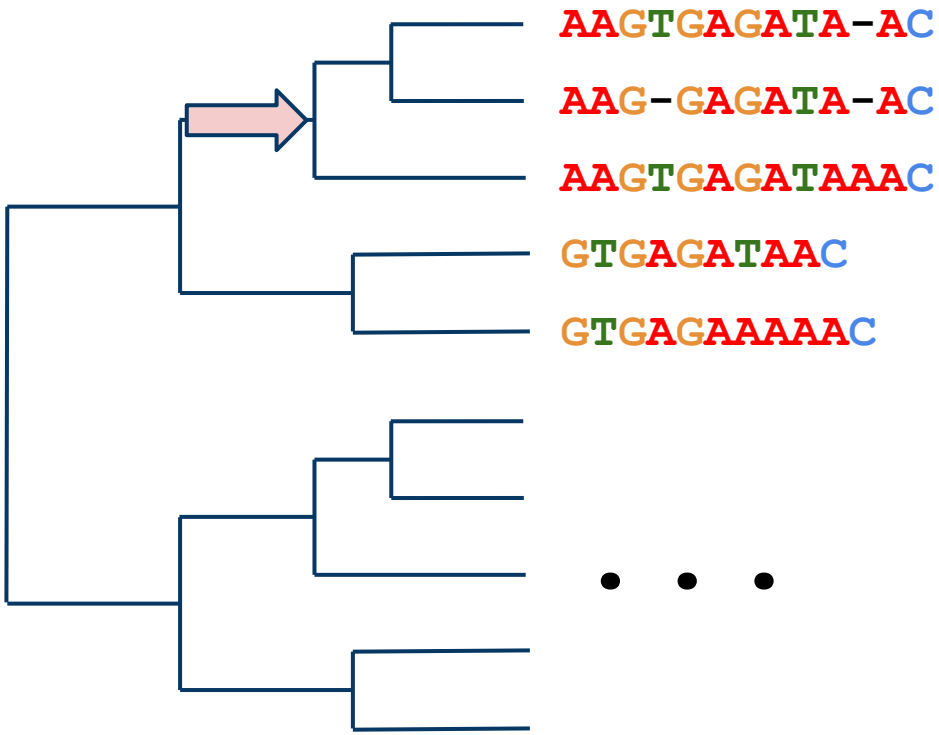
Calcul d'une matrice de distances par paires → Calcul d'un arbre approximatif → **Alignement des séquences le long de cet arbre**



Aligner un ensemble de séquences (n>2) - en pratique

Étendre l'algorithme de NW à plus de 2 séquences
Méthode de l'alignement progressif

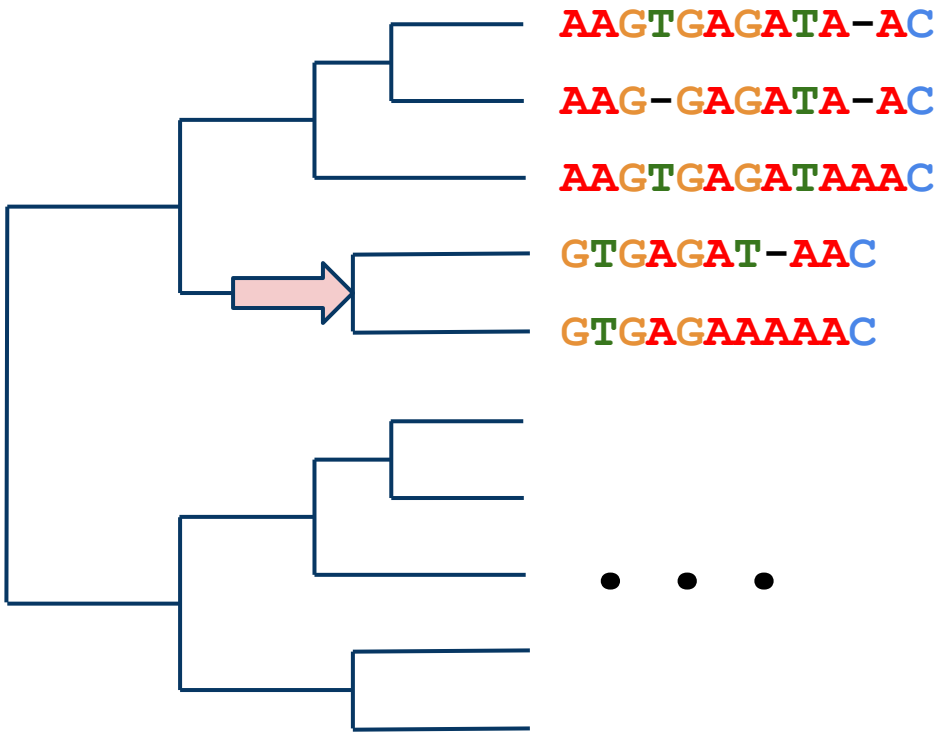
Calcul d'une matrice de distances par paires → Calcul d'un arbre approximatif → **Alignement des séquences le long de cet arbre**



Aligner un ensemble de séquences (n>2) - en pratique

Étendre l'algorithme de NW à plus de 2 séquences
Méthode de l'alignement progressif

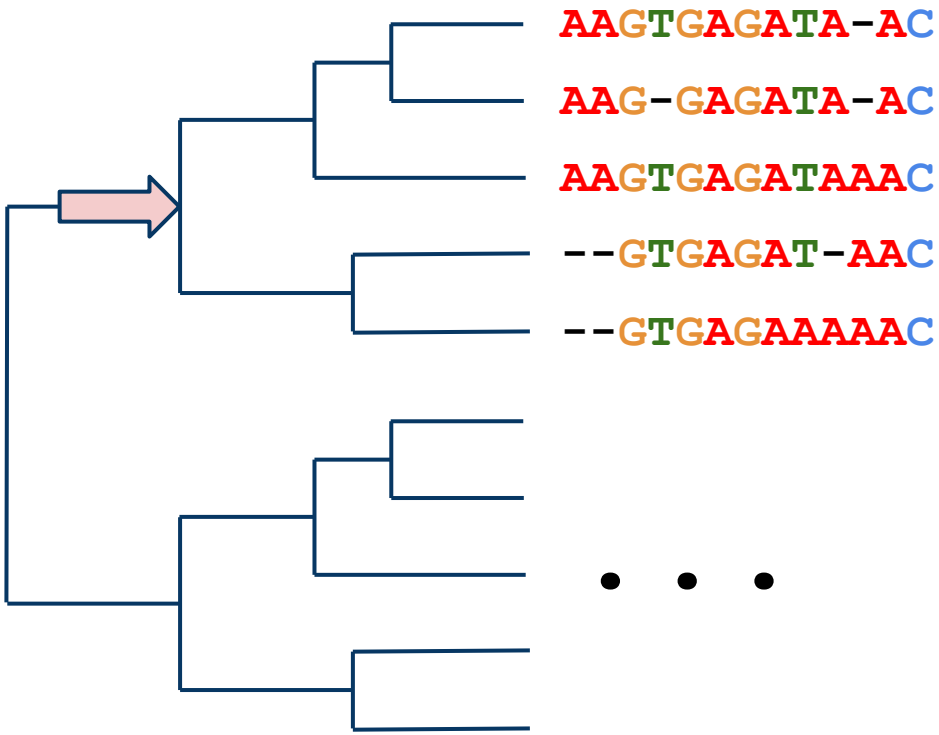
Calcul d'une matrice de distances par paires → Calcul d'un arbre approximatif → **Alignement des séquences le long de cet arbre**



Aligner un ensemble de séquences (n>2) - en pratique

Étendre l'algorithme de NW à plus de 2 séquences
Méthode de l'alignement progressif

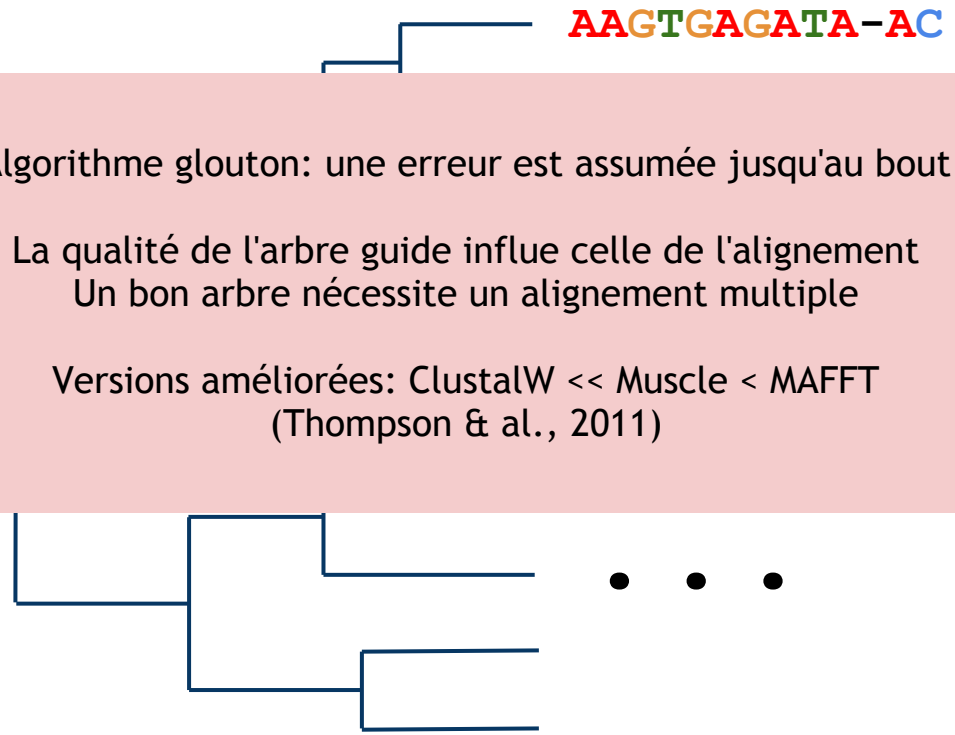
Calcul d'une matrice de distances par paires → Calcul d'un arbre approximatif → **Alignement des séquences le long de cet arbre**



Aligner un ensemble de séquences (n>2) - en pratique

Étendre l'algorithme de NW à plus de 2 séquences
Méthode de l'alignement progressif

Calcul d'une matrice de distances par paires → Calcul d'un arbre approximatif → Alignement des séquences le long de cet arbre



Algorithme glouton: une erreur est assumée jusqu'au bout

La qualité de l'arbre guide influe celle de l'alignement
Un bon arbre nécessite un alignement multiple

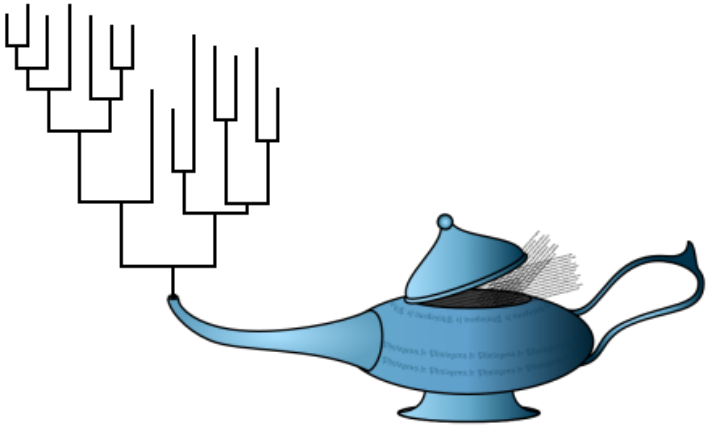
Versions améliorées: ClustalW << Muscle < MAFFT
(Thompson & al., 2011)

Aligner un ensemble de séquences (n>2) - en pratique

L'alignement global est central dans l'analyse des familles

Les alignements locaux sont une problématique très proche, et très utilisées en bioinformatique (BLAST par exemple)

Des séquences à la phylogénie



Reconstruction phylogénétique: 3 catégories de méthodes

3 types de méthodes,
pour 3 qualités de résultats et 3 vitesses d'exécution

Méthodes de distance	Méthodes par parcimonie	Méthodes par vraisemblance
Temps calculs		
***	**	*
Qualité de l'arbre obtenu		
*	**	***

cirad

agap

Reconstruction phylogénétiques: méthodes de distance

Méthodes de distance — Méthodes par parcimonie — Méthodes par vraisemblance

Hypothèse: l'histoire la plus courte est la meilleure

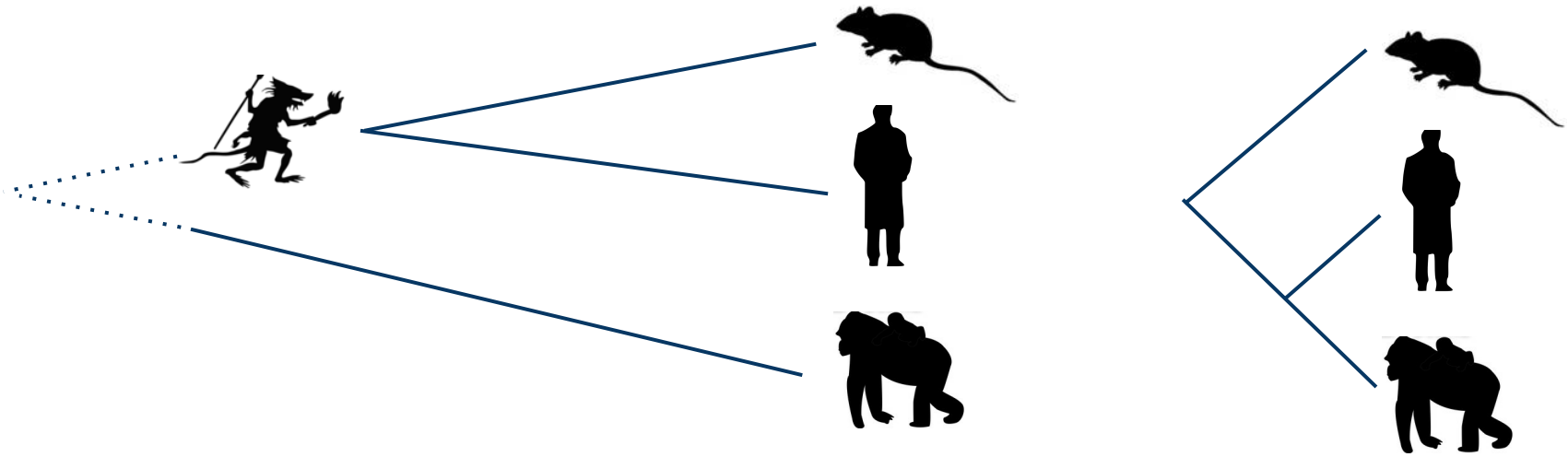
Alignement multiple → Matrice de distance → Arbre phylogénétique

Reconstruction phylogénétiques: méthodes de distance

Méthodes de distance — Méthodes par parcimonie — Méthodes par vraisemblance

Hypothèse: l'histoire la plus courte est la meilleure

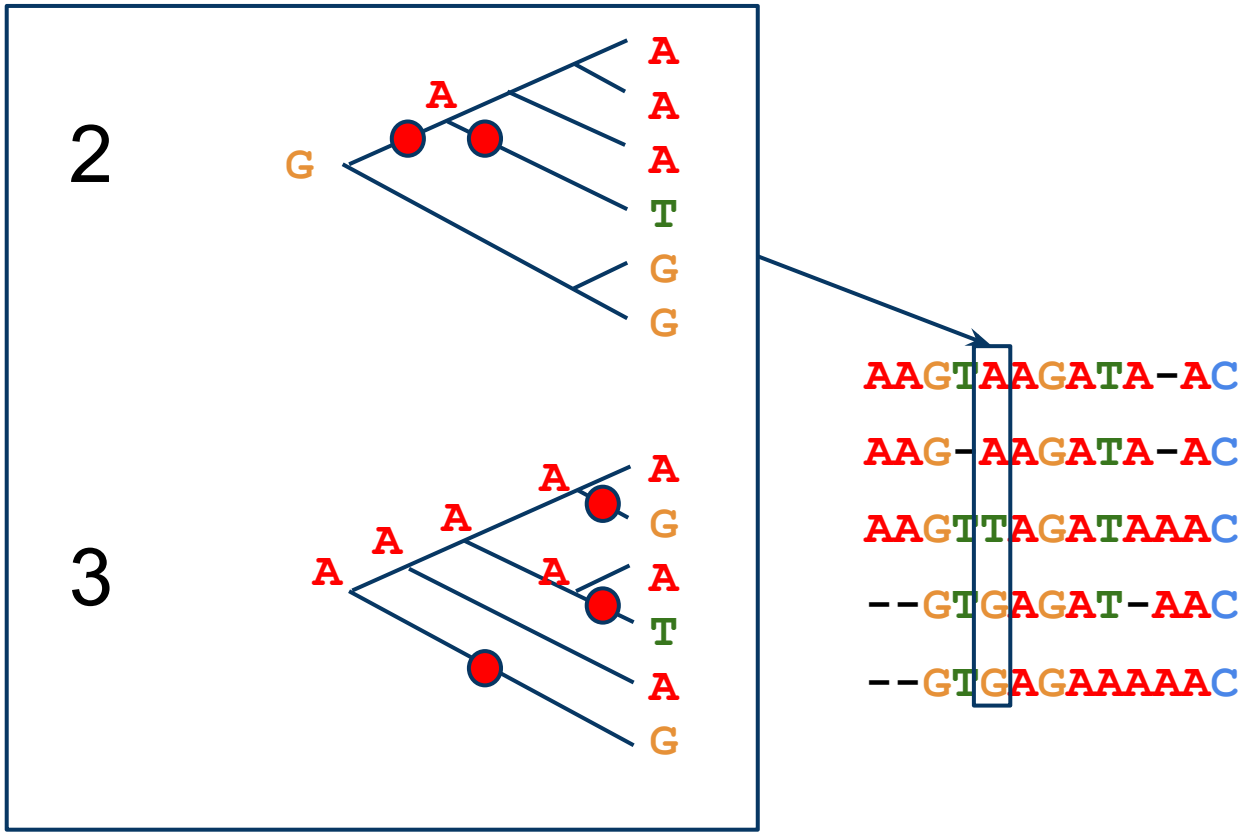
Alignement multiple → Matrice de distance → Arbre phylogénétique



Reconstruction phylogénétiques: méthodes par parcimonie

Méthodes de distance | **Méthodes par parcimonie** | Méthodes par vraisemblance

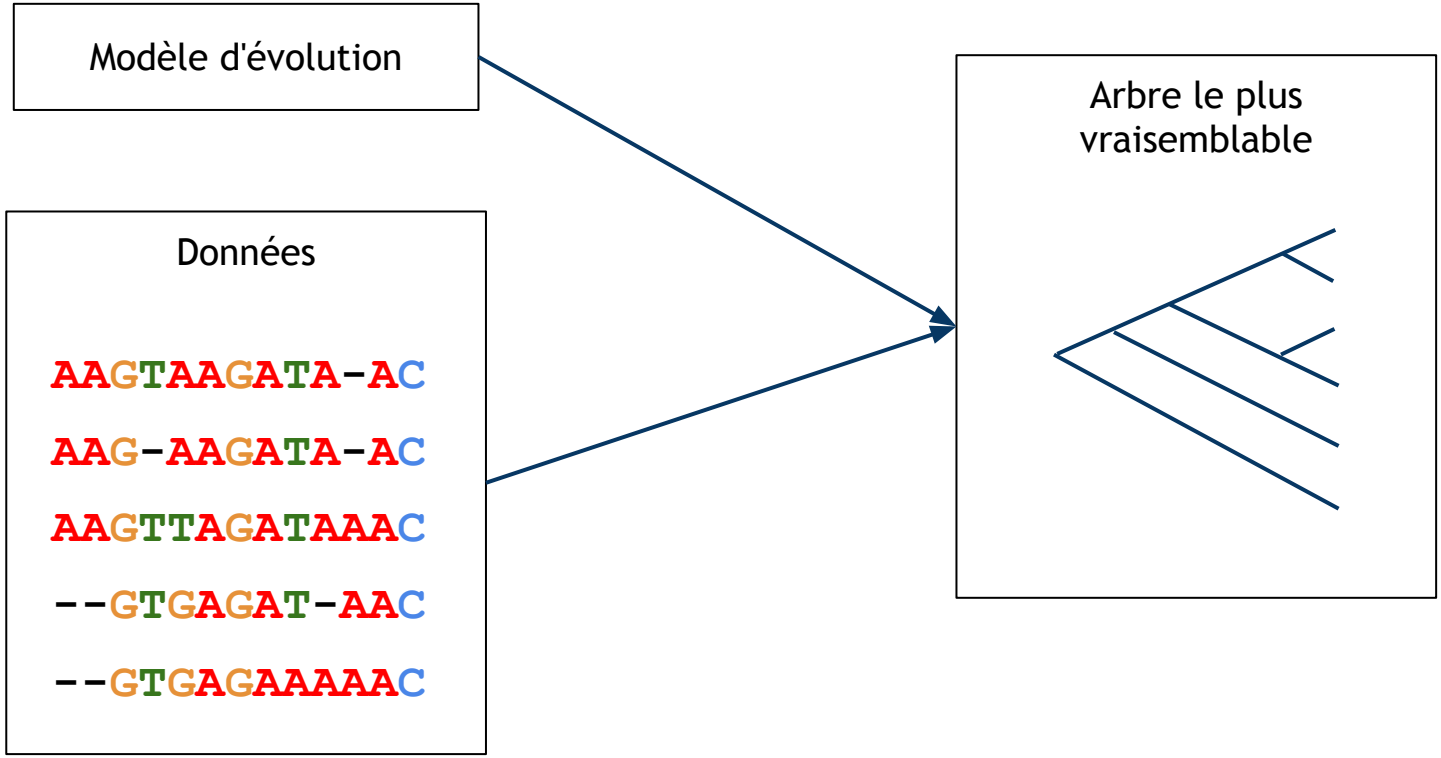
Hypothèse: l'arbre inférant le moins de mutation est le meilleur.



Reconstruction phylogénétiques: méthodes par maximum de vraisemblance

Méthodes de distance — Méthodes par parcimonie — **Méthodes par vraisemblance**

Hypothèse: l'arbre le plus vraisemblable selon les données et un modèle d'évolution.



cirad

agap

Reconstruction phylogénétique: trouver le meilleur arbre

Pour n séquences: il y a 1 x 3 x 5 x 7 x ... x (2n - 5) phylogénies résolues possibles...

- 3 arbres pour 4 séquences,
- 15 arbres pour 5 séquences,
- 2 millions d'arbres pour 10 séquences...

Reconstruction phylogénétique: trouver le meilleur arbre

Pour n séquences: il y a 1 x 3 x 5 x 7 x ... x (2n - 5) phylogénies résolues possibles...

- 3 arbres pour 4 séquences,
- 15 arbres pour 5 séquences,
- 2 millions d'arbres pour 10 séquences...

Au dessus de 10 séquences, évaluer tous des arbres est difficile

Quel que soit ce que l'on veut optimiser:

- Longueur de l'arbre
- Parcimonie
- Vraisemblance

Les logiciels doivent explorer les topologies possibles selon une euristique.

cirad

agap

Reconstruction phylogénétique: trouver le meilleur arbre

Pour n séquences: il y a 1 x 3 x 5 x 7 x ... x (2n - 5) phylogénies résolues possibles...

- 3 arbres pour 4 séquences,
- 15 arbres pour 5 séquences,
- 2 millions d'arbres pour 10 séquences...

Au dessus de 10 séquences, évaluer tous des arbres est difficile

Quel que soit ce que l'on veut optimiser:

- Longueur de l'arbre
- Parcimonie
- Vraisemblance

Les logiciels doivent explorer les topologies possibles selon une euristique.

Logiciels de distances: Phylip, BioNJ, FastME...
Logiciels de parcimonie: Phylip, PAUP, TNT...
Logiciels de maximum de vraisemblance : PhyML, RaxML...

cirad

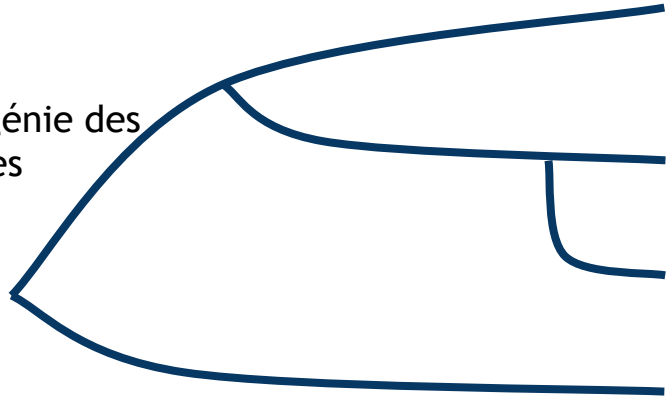
agap

Comparaison d'arbres: de la réconciliation d'arbres aux super-arbres

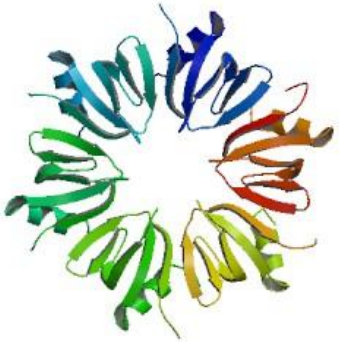
Phylogénie des espèces, et phylogénie moléculaire

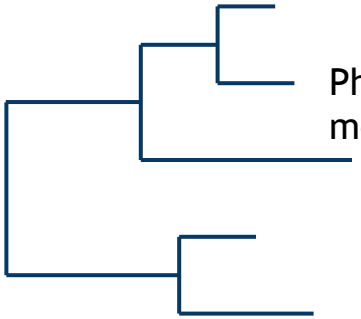
Histoire évolutive des espèces

Phylogénie des espèces



Histoire évolutive des molécules





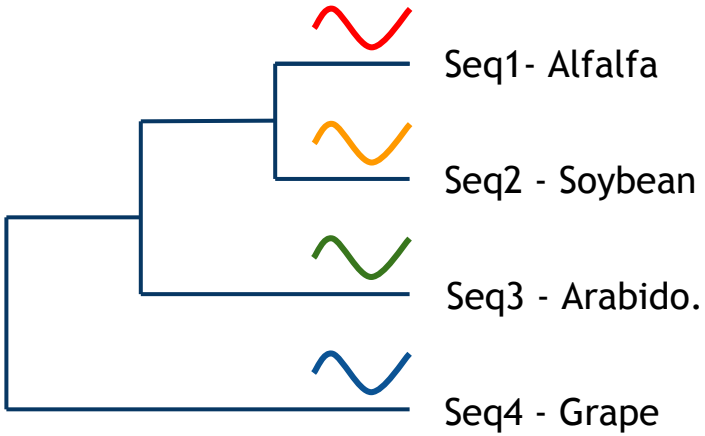
Phylogénie moléculaire

Réconciliation d'arbres

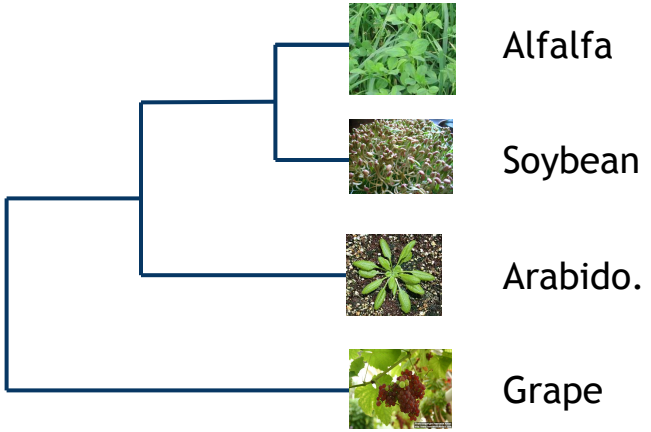
Question ouverte

Inférer l'histoire des gènes
= ?
Inférer l'histoire des espèces

Phylogénie moléculaire

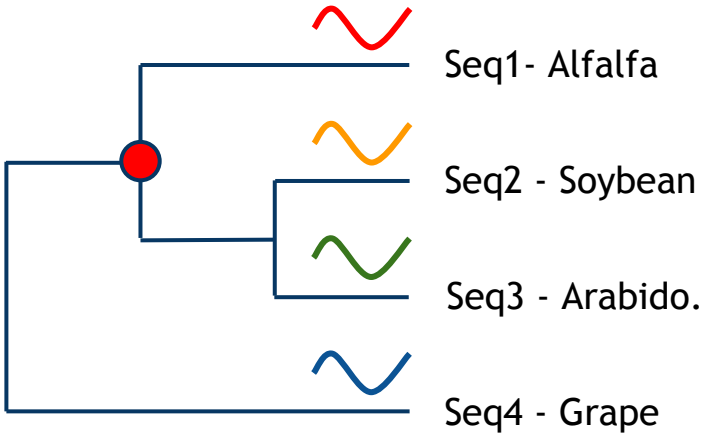


Phylogénie des espèces

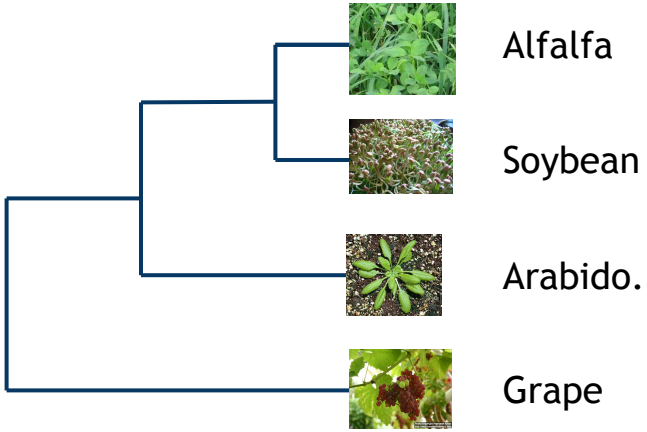


Réconciliation d'arbres

Phylogénie moléculaire

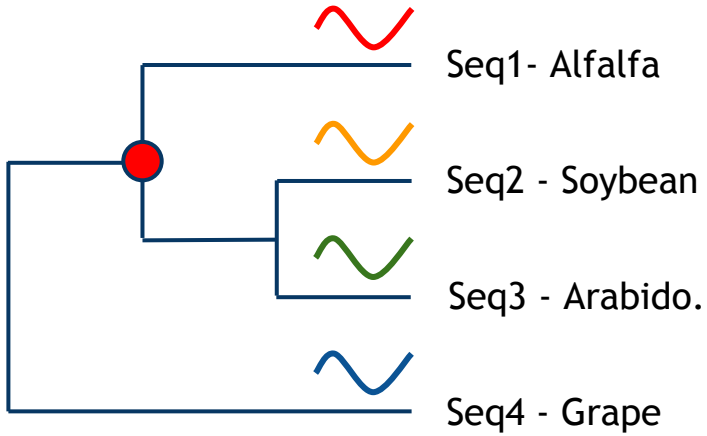


Phylogénie des espèces

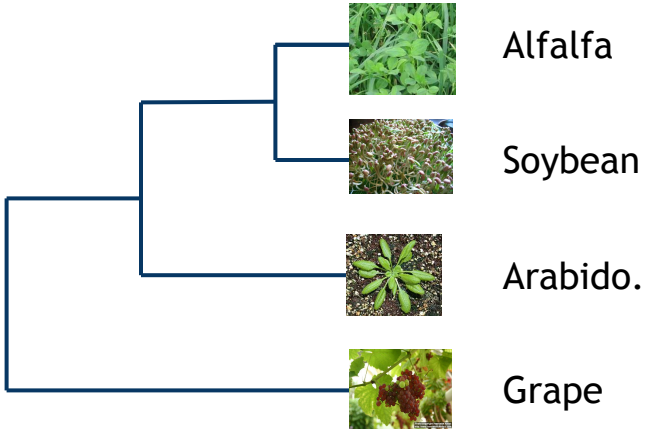


Réconciliation d'arbres

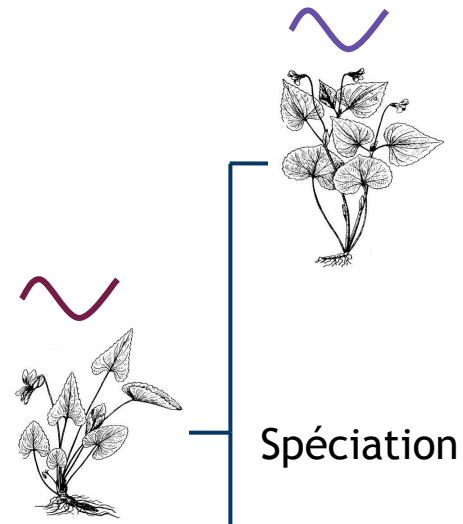
Phylogénie moléculaire



Phylogénie des espèces

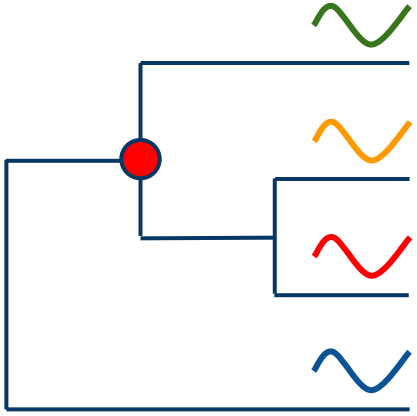


Réconciliation d'arbres



Seq4 - grape

Phylogénie moléculaire



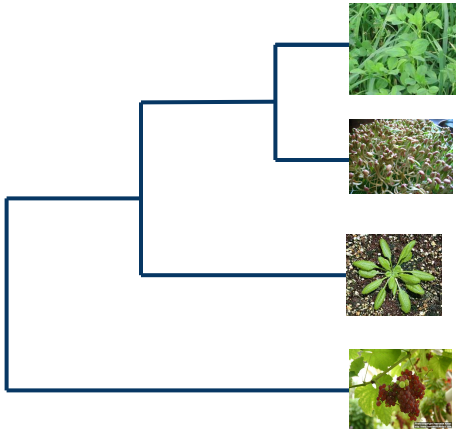
Seq1- Alfalfa

Seq2 - Soybean

Seq3 - Arabido.

Seq4 - Grape

Phylogénie des espèces



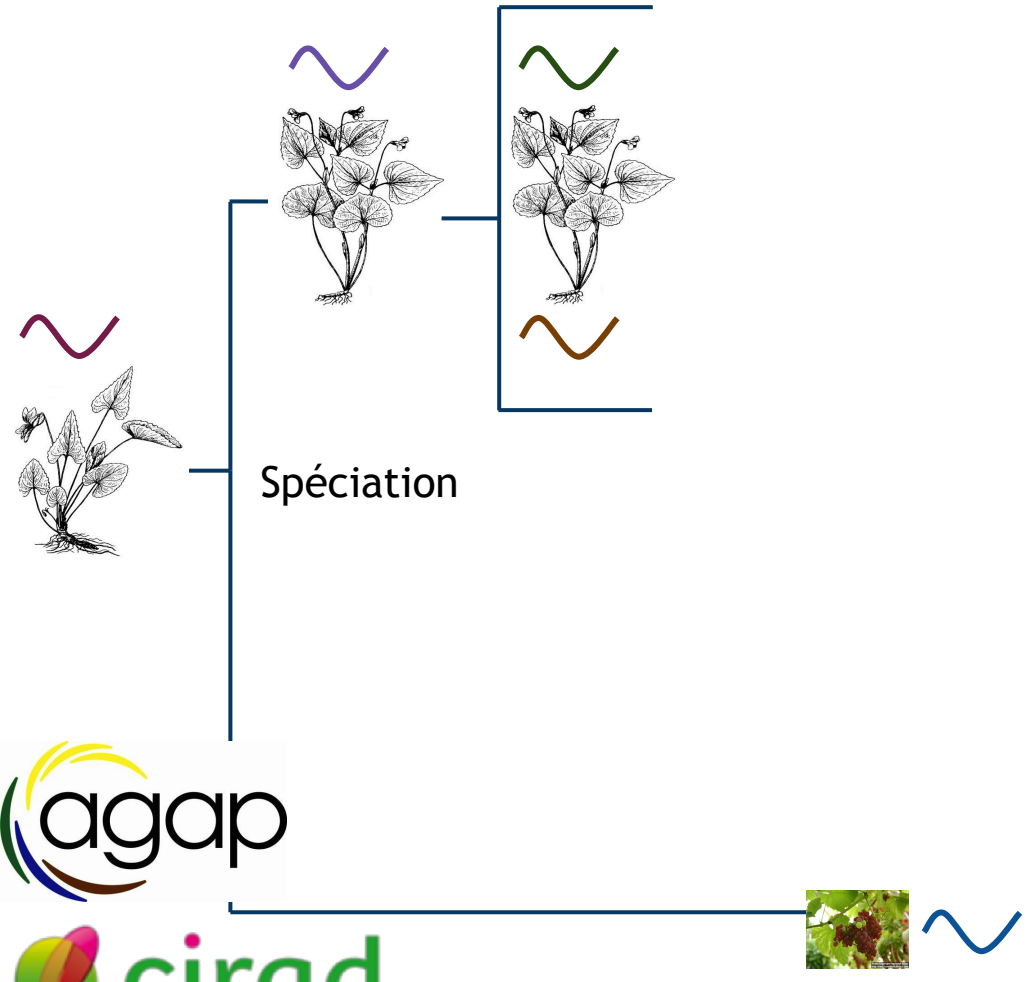
Alfalfa

Soybean

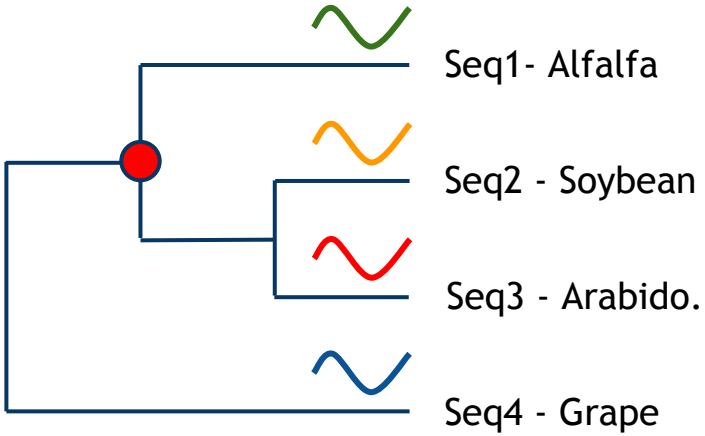
Arabido.

Grape

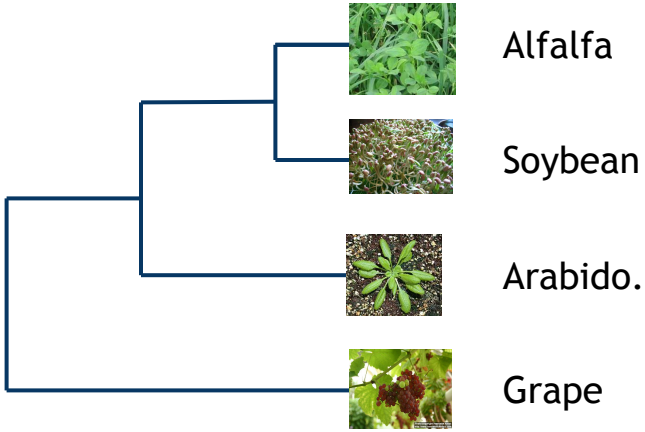
Réconciliation d'arbres



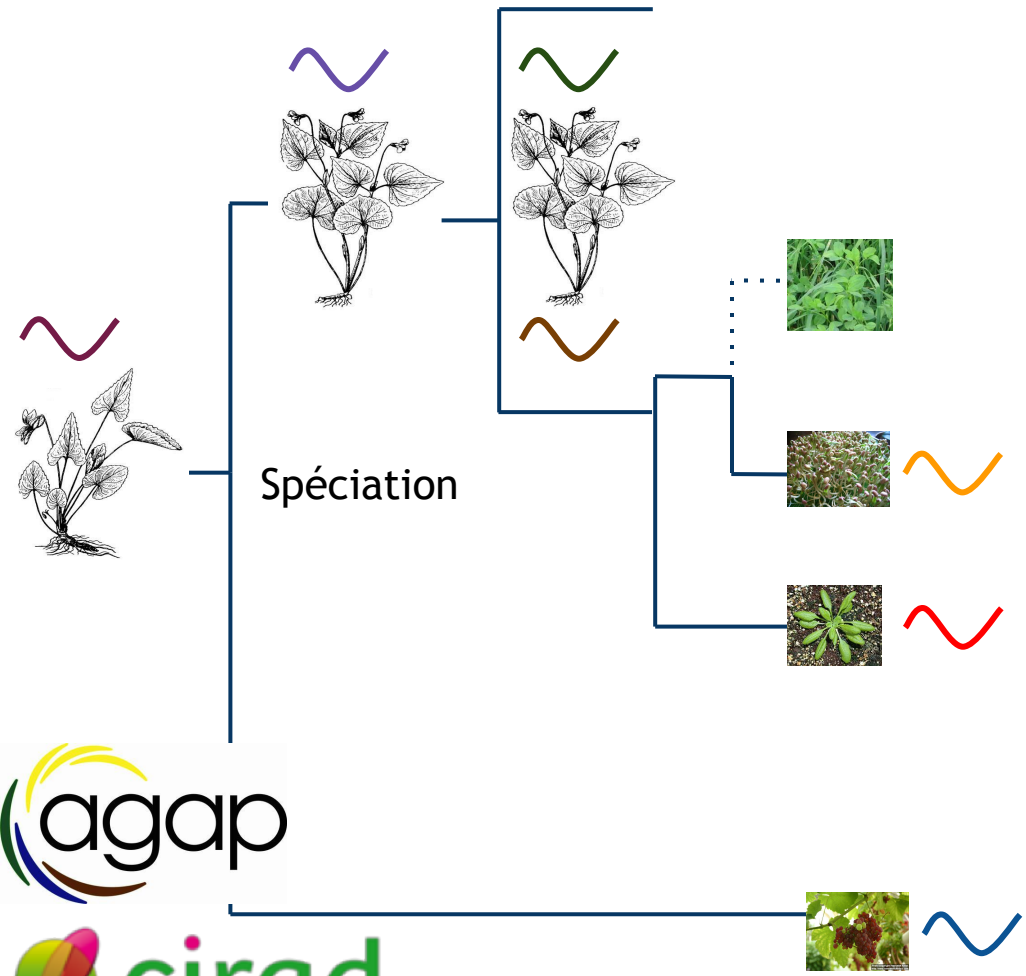
Phylogénie moléculaire



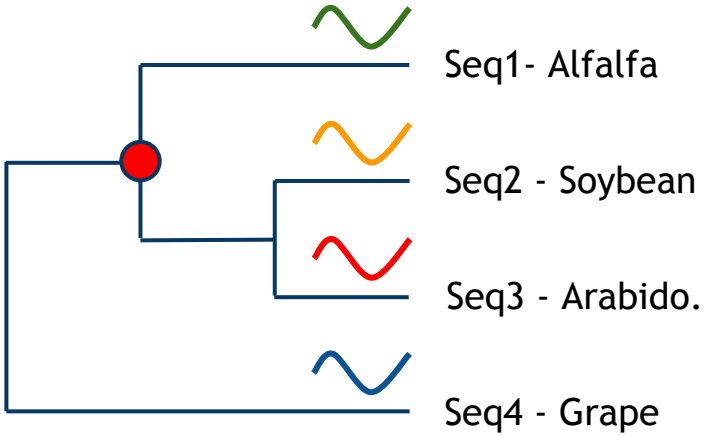
Phylogénie des espèces



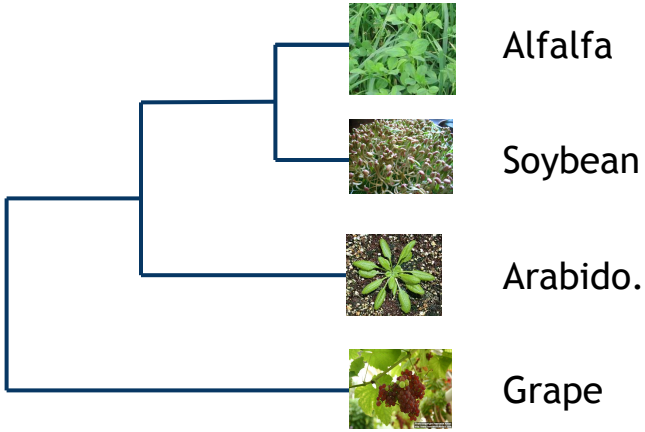
Réconciliation d'arbres



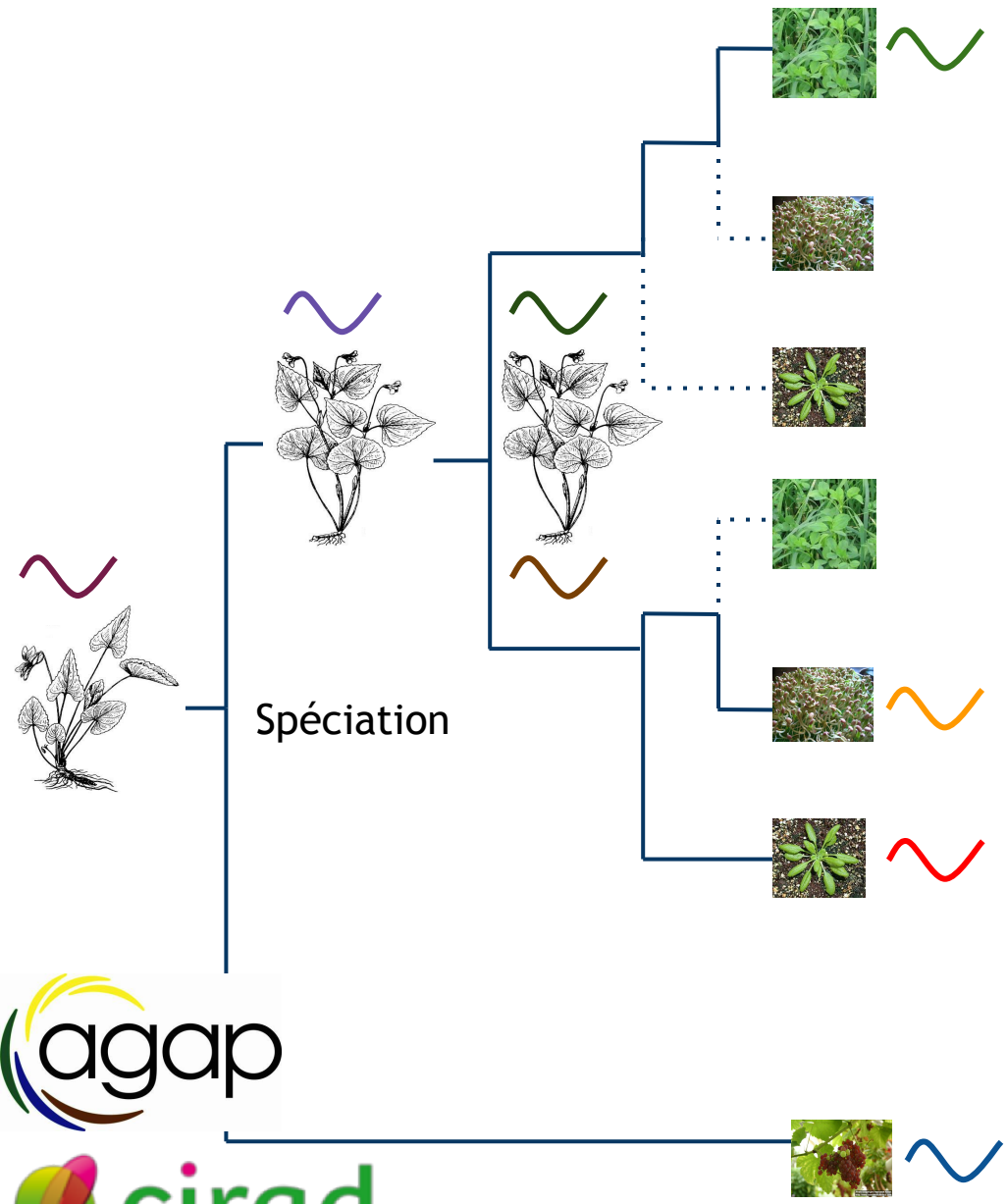
Phylogénie moléculaire



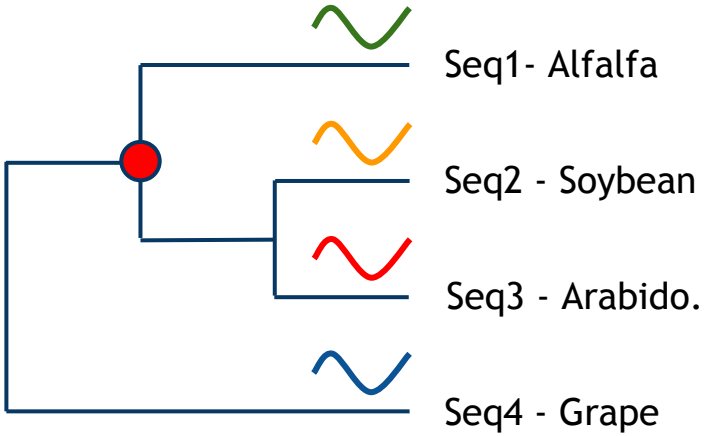
Phylogénie des espèces



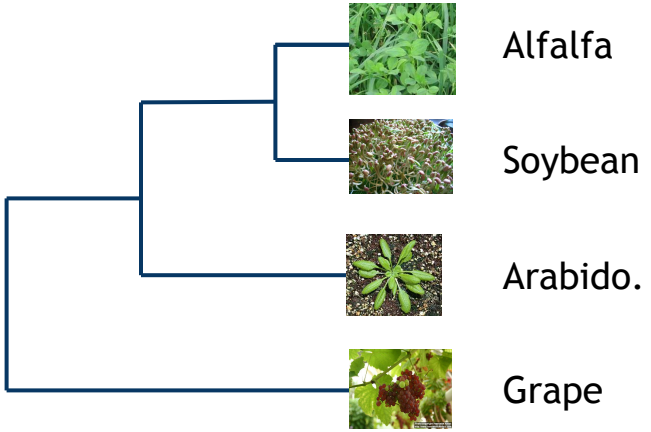
Réconciliation d'arbres



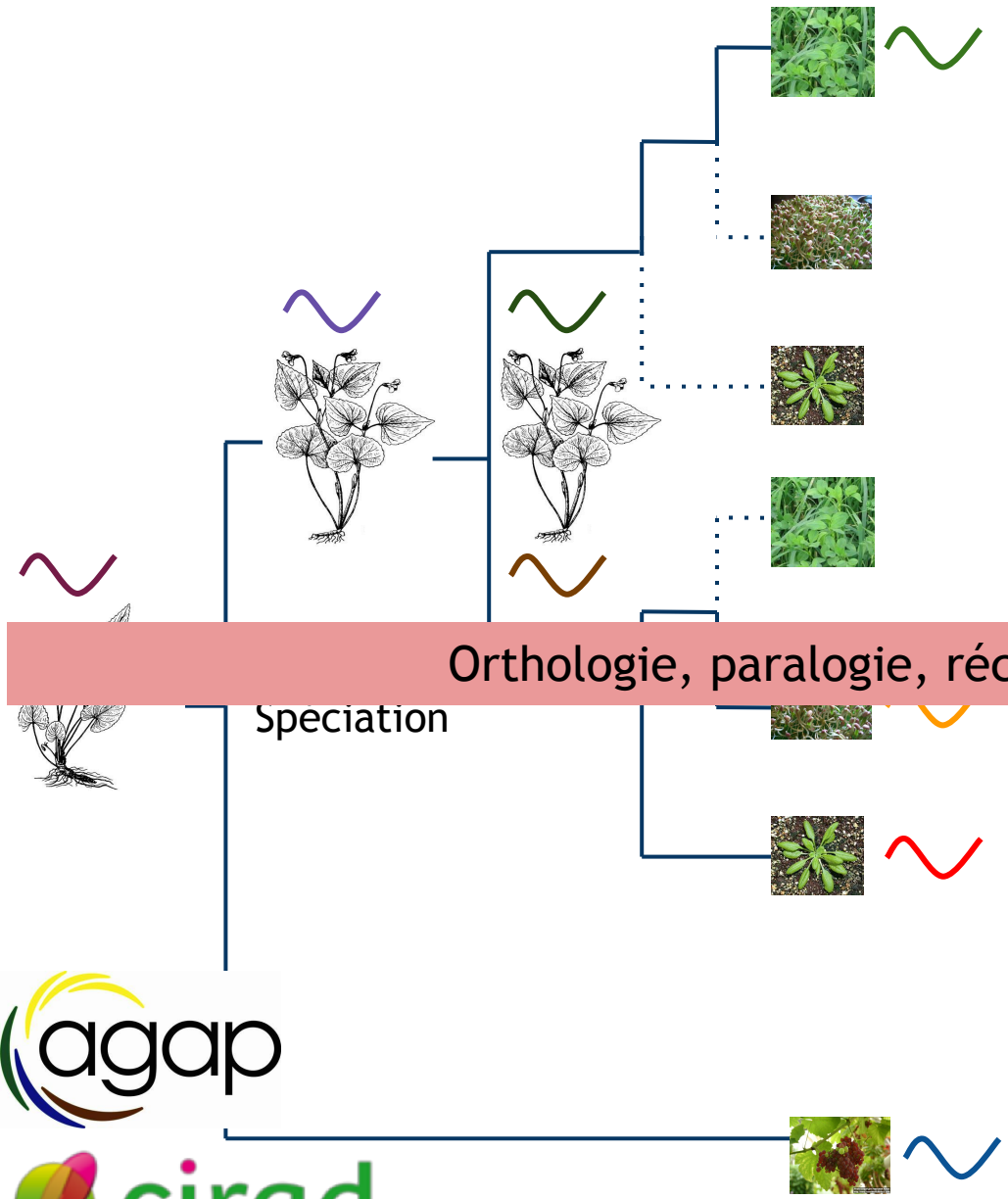
Phylogénie moléculaire



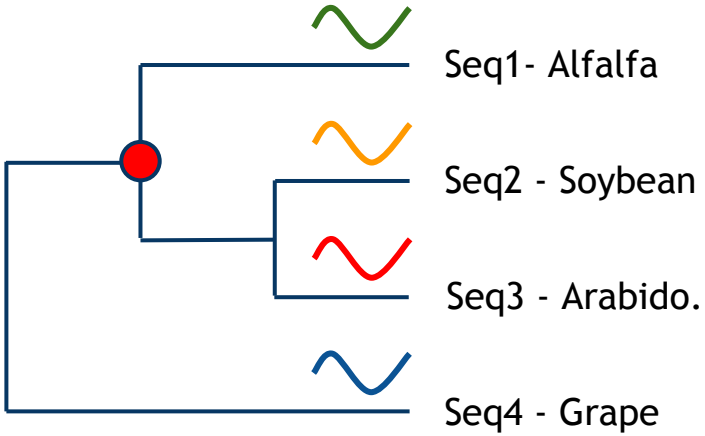
Phylogénie des espèces



Réconciliation d'arbres

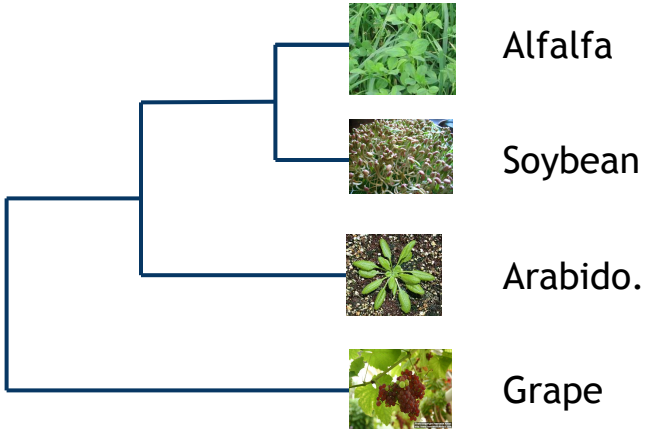


Phylogénie moléculaire

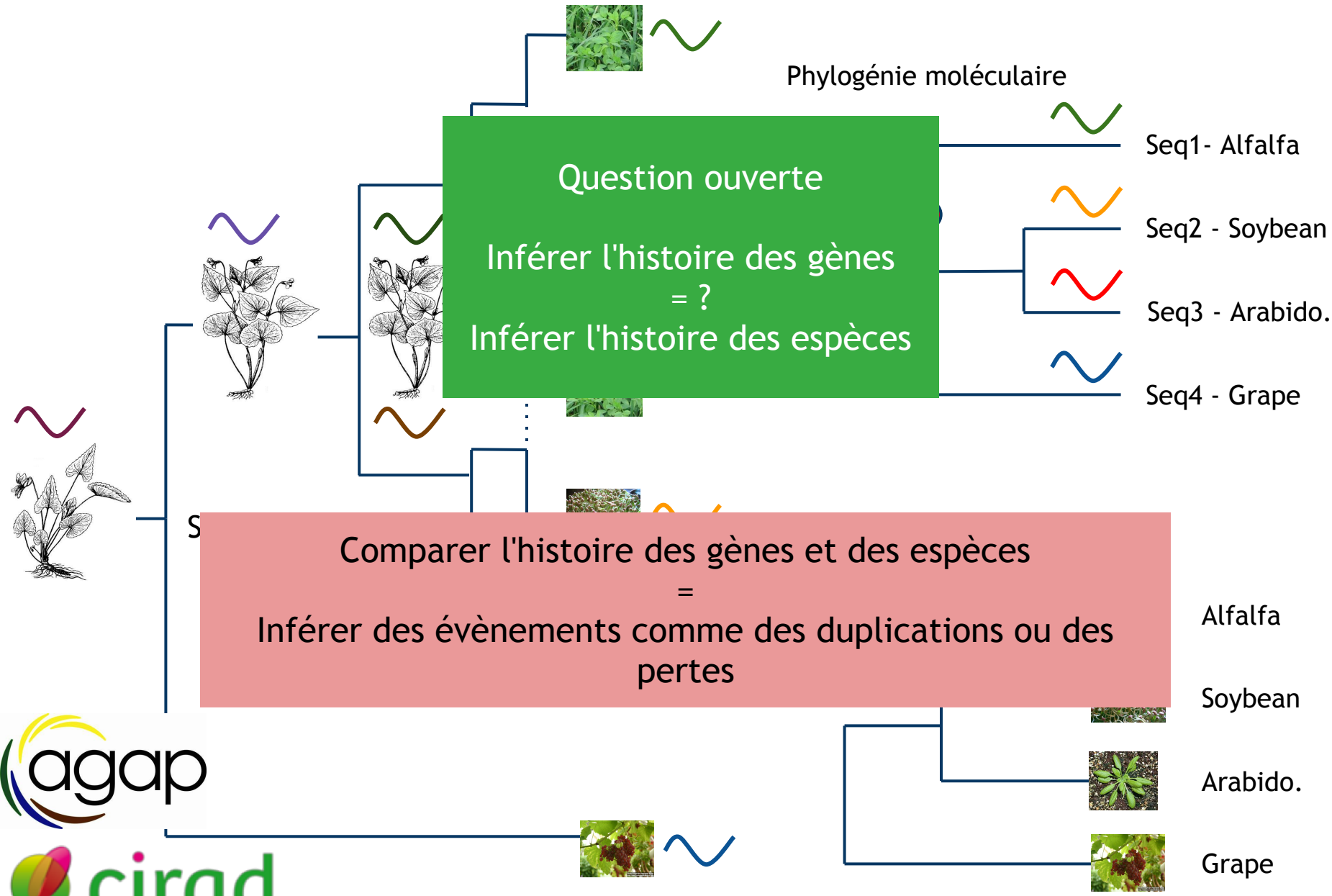


Orthologie, paralogie, réconciliation d'arbres

Phylogénie des espèces



Réconciliation d'arbres

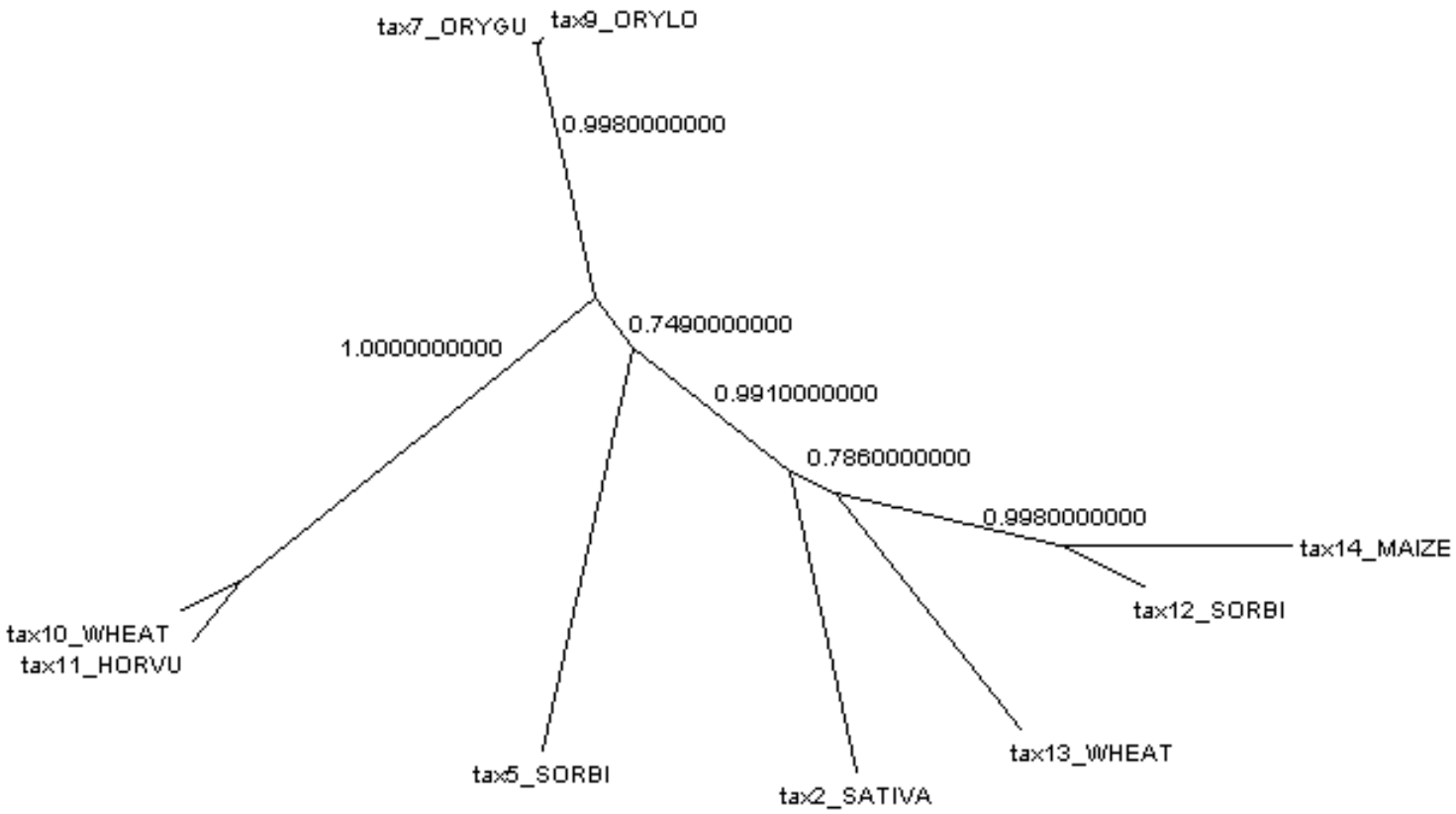


Question ouverte

Inférer l'histoire des gènes
= ?
Inférer l'histoire des espèces

Comparer l'histoire des gènes et des espèces
=
Inférer des évènements comme des duplications ou des pertes

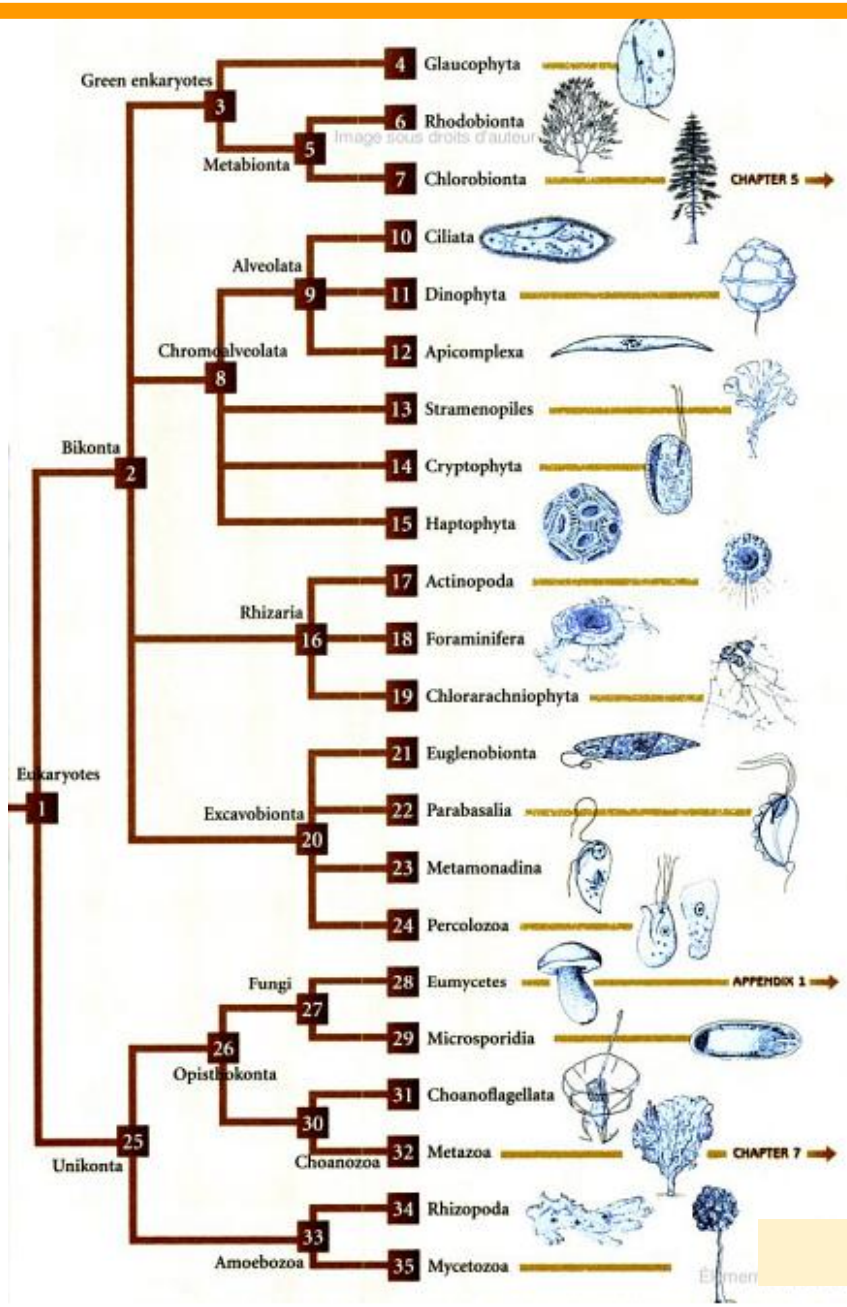
Réconciliation d'arbres



La topologie d'un arbre n'est pas toujours bien résolue

Un arbre phylogénétique n'a pas de racine

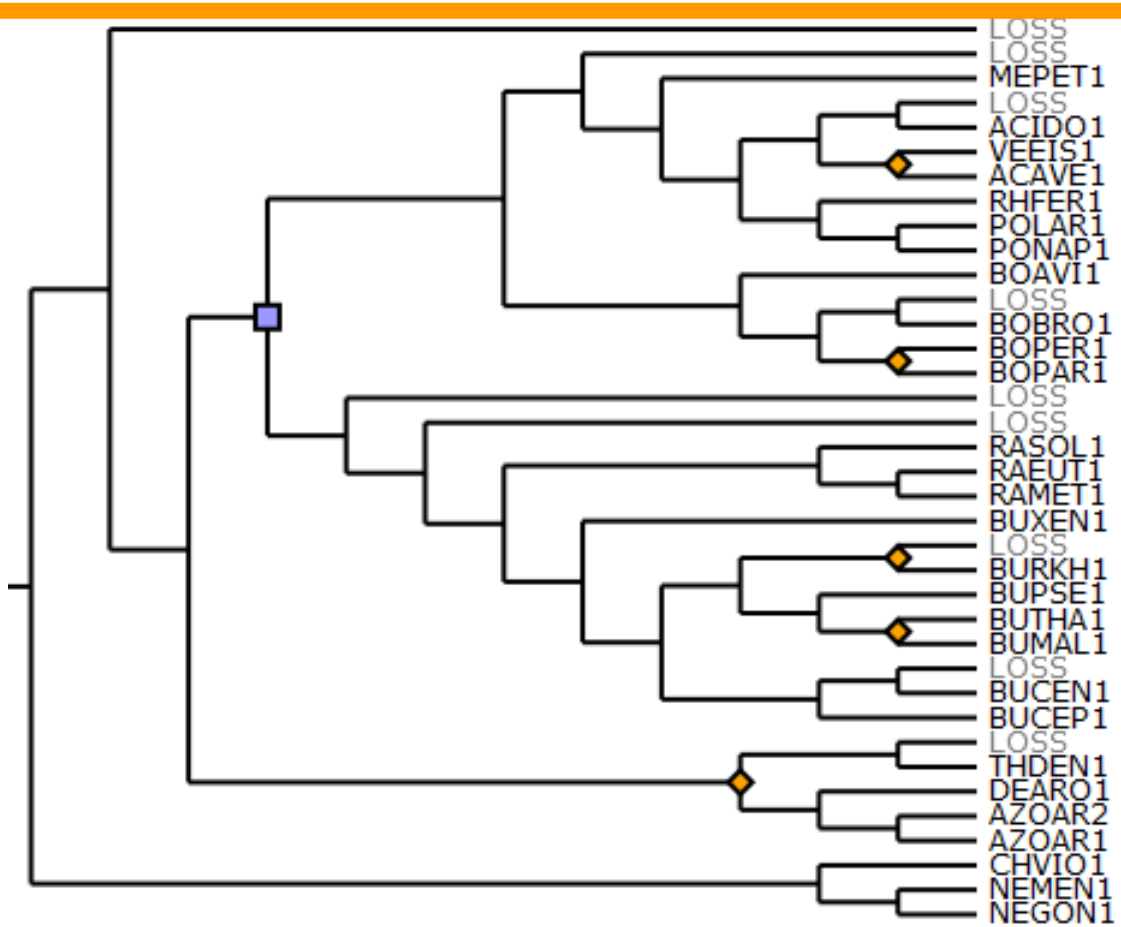
Réconciliation d'arbres



La phylogénie des espèces n'est pas complètement résolue

Lecointre, Le Guyader

Réconciliation d'arbres

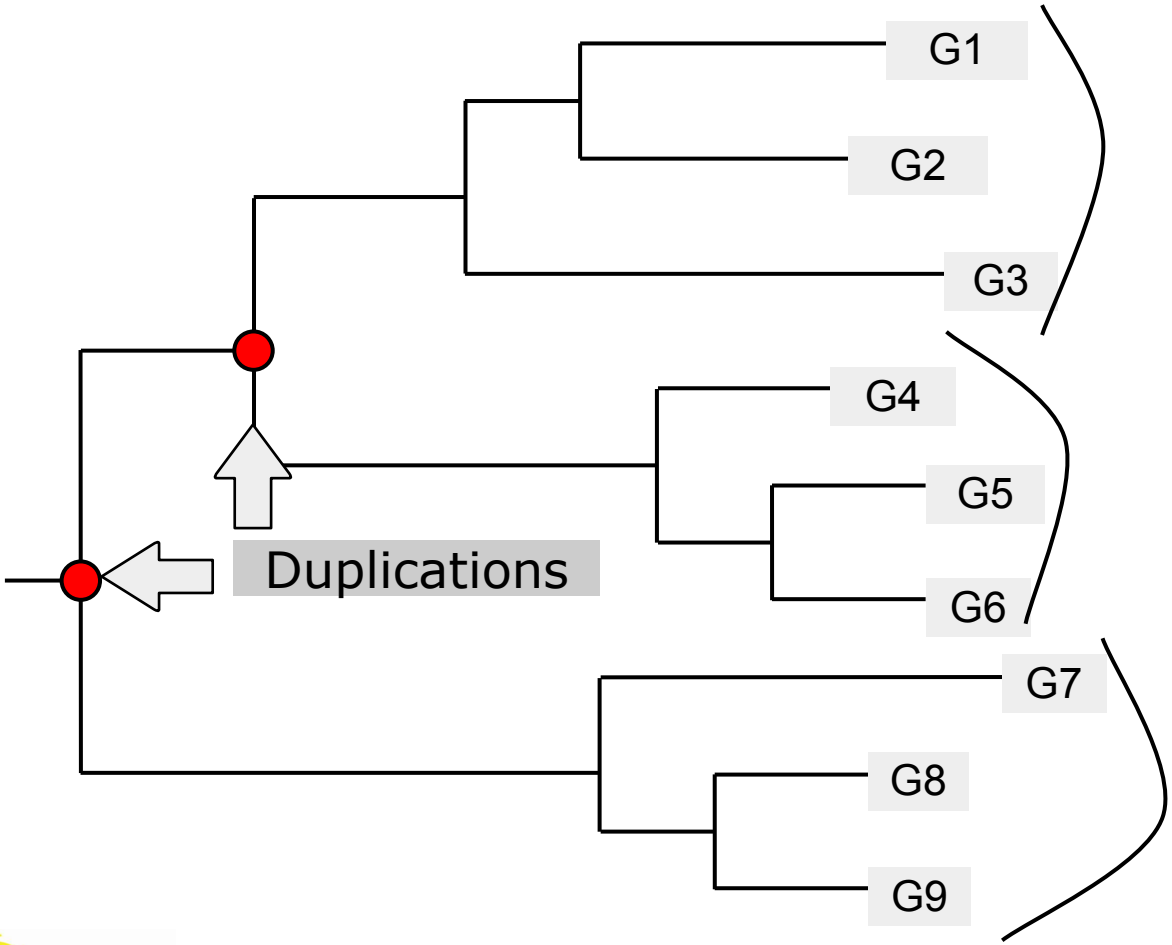


Annoter des évènements de l'histoire évolutive

en gérant:

1. les topologies mal résolues,
2. les supports,
3. les racinements,
4. ...

De l'usage des concepts d'orthologie et paralogie



Groupes d'orthologues

HOGENOM

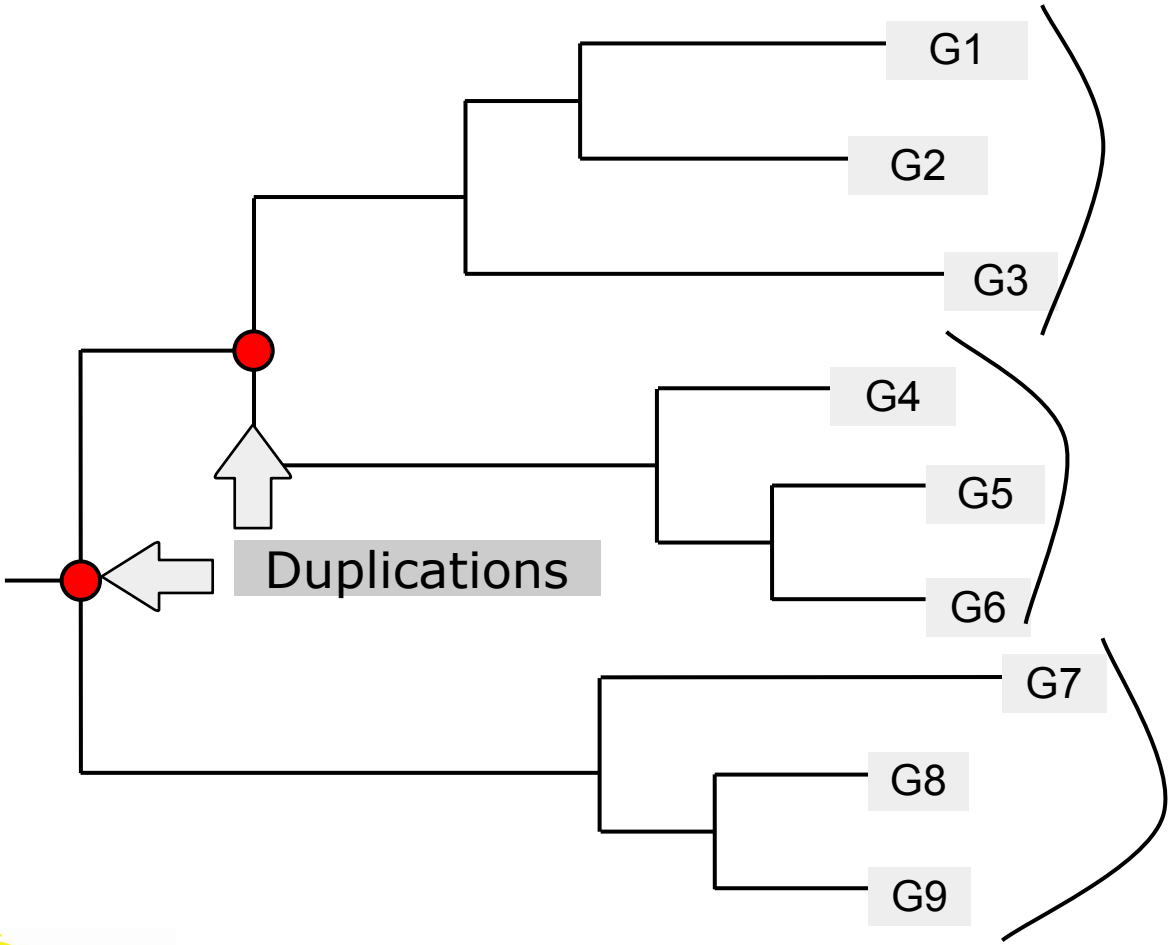
Penel et al. 2011

GreenPhyl

Rouard et al. 2010
Conte et al. 2008

Hypothèse implicite: "Deux gènes orthologues ont des fonctions plus proches entre eux que deux gènes paralogues."

De l'usage des concepts d'orthologie et paralogie



Groupes d'orthologues

HOGENOM

Penel et al. 2011

GreenPhyl

Rouard et al. 2010
Conte et al. 2008

Hypothèse implicite: "Deux gènes orthologues ont des fonctions plus proches entre eux que deux gènes paralogues"

Question ouverte

Inférer l'histoire des gènes
= ?
Inférer l'histoire des espèces

De la phylogénétique à la phylogénomique

Question ouverte

Inférer l'histoire des gènes
= ?
Inférer l'histoire des espèces

Analyse phylogénétique: un gène, aligné avec soin

Impact des évènements évolutifs

De la phylogénétique à la phylogénomique

Question ouverte

Inférer l'histoire des gènes
= ?
Inférer l'histoire des espèces

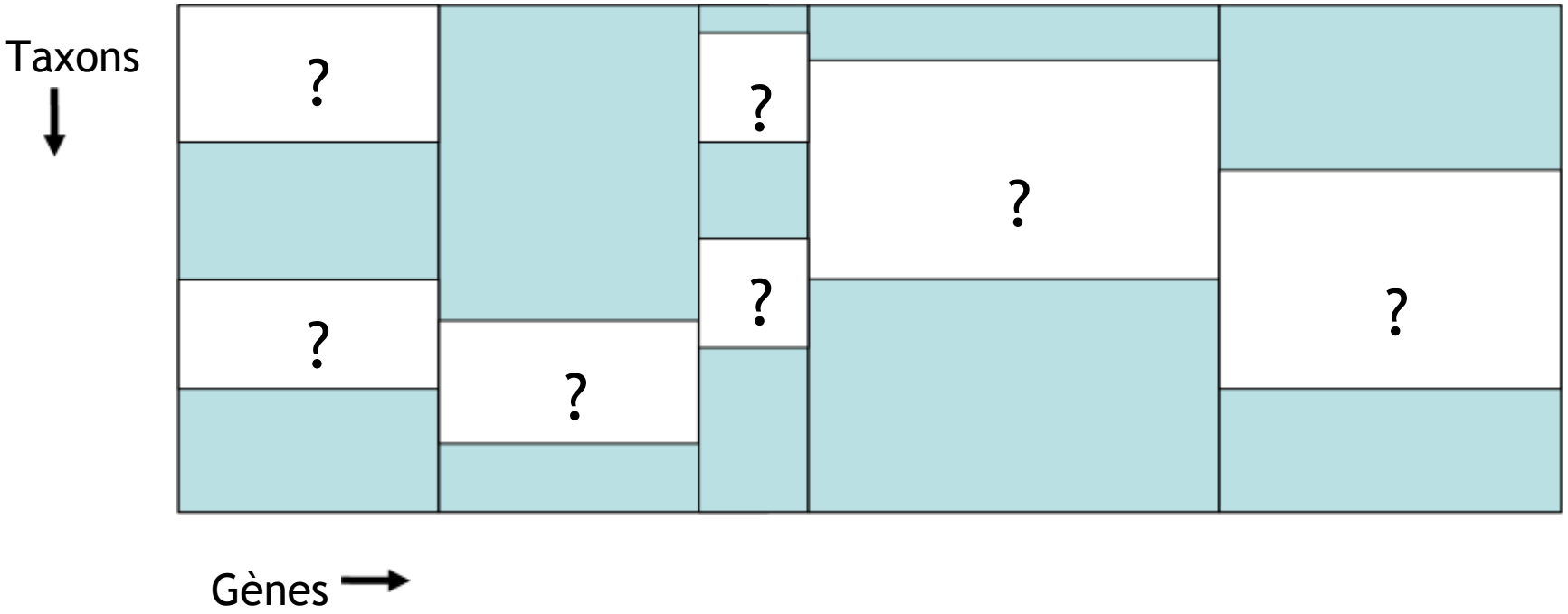
Analyse phylogénétique: un gène, aligné avec soin

Impact des évènements évolutifs

Analyse phylogénomique: nombreux gènes, analyse automatique

Trouver le signal majoritaire

Données (super-matrice) :

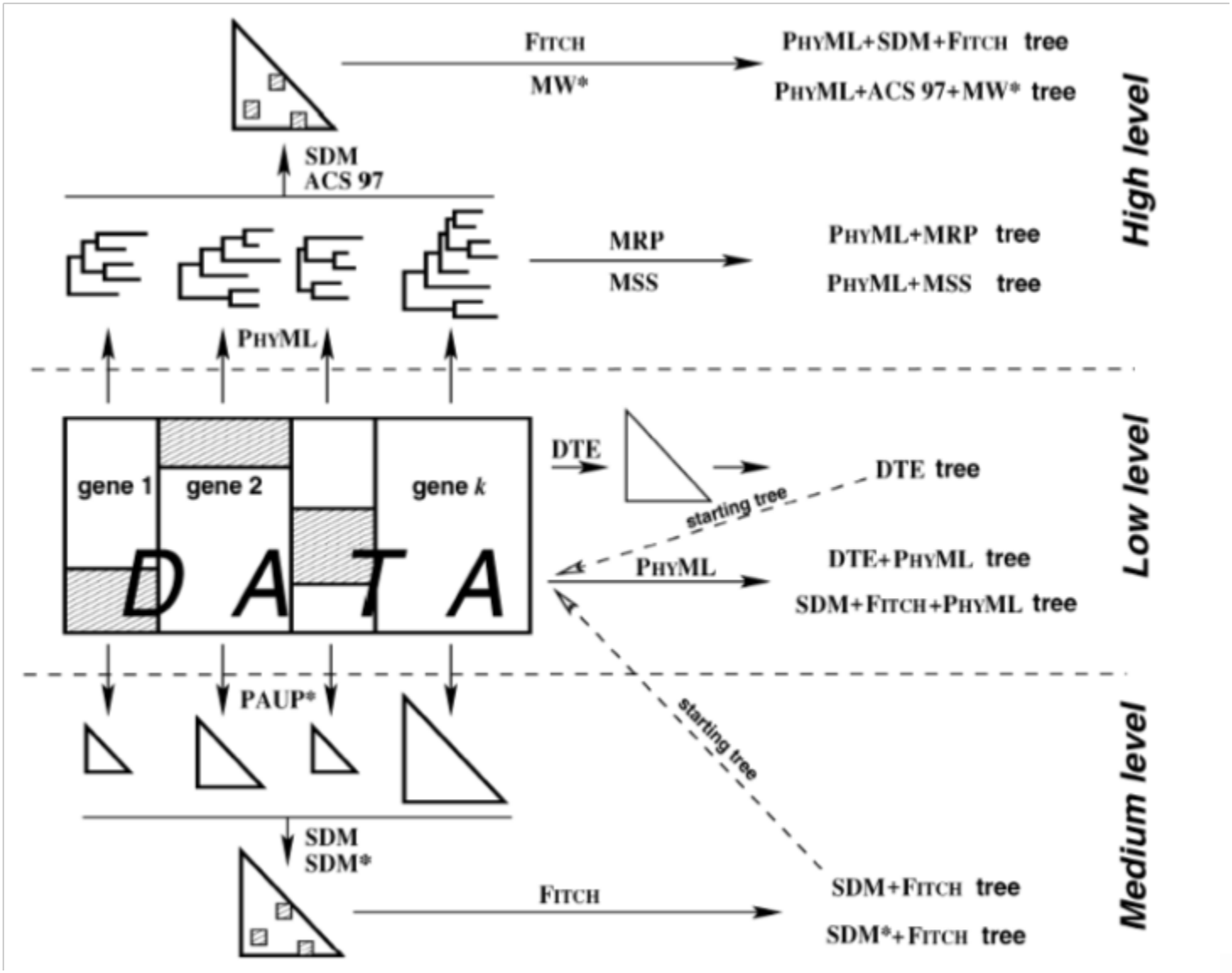


De la phylogénétique à la phylogénomique

4 principales difficultés:

- Quantité de données manquantes (Driskell et al).
- Hétérogénéité des processus évolutifs (Gatesy et al 2002; Criscuolo et al 2006).
- Duplications, transferts.
- Grande complexité informatique.

De la phylogénétique à la phylogénomique



Bininda-Emonds O., ed., (2004). *Phylogenetic supertrees: Combining information to reveal the tree of life*. Kluwer Academic Publishers, Dordrecht.

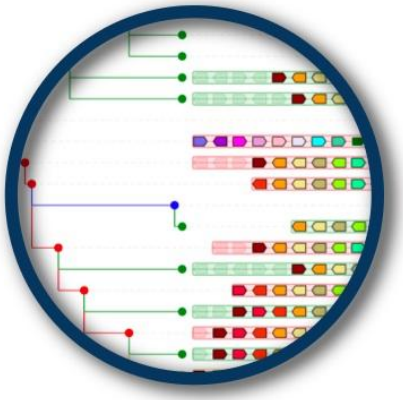
Criscuolo A., Berry V., Douzery E., Gascuel O. (2006). SDM: A Fast Distance-Based Approach for (Super)Tree Building in Phylogenomics. *Systematic Biology* 55(5):740–755.

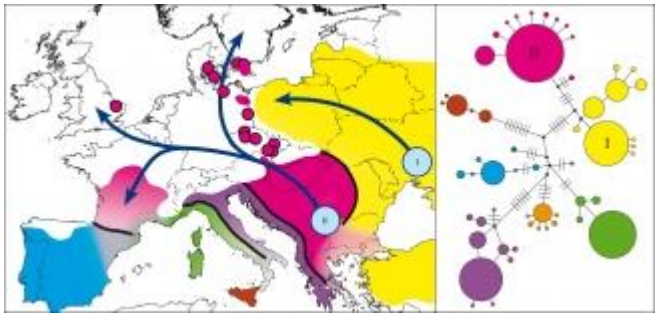
Sanderson et al. (2007). Fragmentation of large datasets in phylogenetic analyses. In *Reconstructing Evolution: New Mathematical and Computational Advances*, Gascuel & Steel, eds., Oxford University Press.

PhySIC_IST : healing source trees to infer healthy supertrees (2008) Celine Scornavacca, Vincent Berry, Emmanuel J. P. Douzery and Vincent Ranwez. *BMC Bioinformatics* 9:413

En conclusion...

Intégration de données, car la phylogénie n'est qu'une information parcéllaire

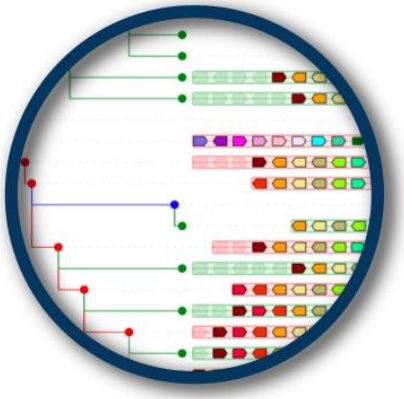




Des projets d'intégration de données, en France

PhyloSpace: projet ANR de phylogéographie, appliqué aux insectes et aux plantes

Genomicus: genome browser intégrant phylogénie et contexte génomique



Ancestrome: projet intégratif d'inférence d'entités ancestrales (génomes, interactomes...)

cirad

agap

Un sujet de thèse dans l'équipe Intégration de Données (CIRAD)

Collecte et intégration de données pour des familles de gènes homologues, appliqué aux stress environnementaux

Représentation des connaissances

Interopérabilité

Fouille de données

...