M2 Mycologie

Approches phylogénétiques

Université Paris Cité

Phylogénie moléculaire

Principes

- arbre phylogénétique : représentation graphique des relations (de proximité) entre différents organismes, dans un contexte d'évolution temporelle à partir d'un ancêtre commun à tous les descendants
- relations de similarité et phénomène de divergence extraites à partir de l'alignement des séquences
- phylogény moléculaire : analyse des données génétiques (ADN) et des différences moléculaires héréditaires
- objectif : caractériser le processus évolutif ayant permis de générer la diversité observée [2, 4]

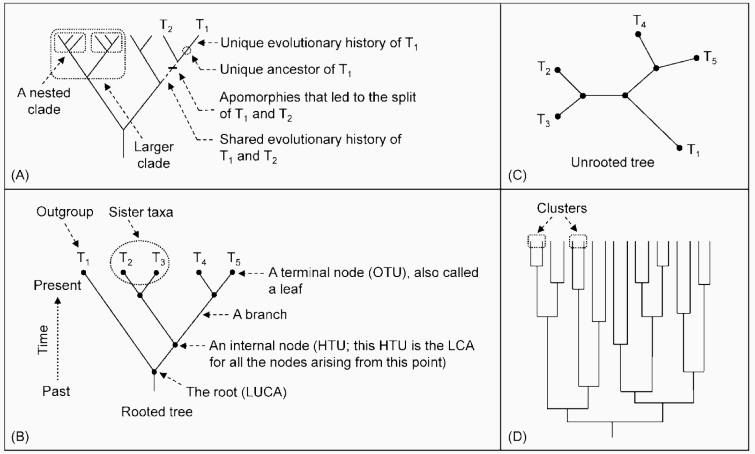
Classification de séquences

- séquences non redondantes, séquences consensus
- aggrégation/classification (EST, OTU, famille de protéines)
- mesure de distance, blast, autre
- logiciels : cd-hit, vsearch, OrthoFinder

Etapes de construction d'un arbre

- sélection des organismes ou famille de gènes
- choix des marqueurs moléculaires
- amplification, séquençage et assemblage
- alignement (ClustalW, MSA, MAFFT, T-Coffee)
- choix d'un modèle évolutionnaire
- analyse phylogénétique (Paup, PAML, PHYLIP, MEGA, Raxml, MrBayes)
- construction de l'arbre phylogénétique
- évaluation de l'arbre phylogénétique (bootstrap, parsimonie/distance, vraisemblance)

Nomenclature



Méthodes de classification

Classification automatique (en statistiques)

Approche non supervisée :

- statut inconnu
- identifier des groupes d'unités statistiques partageant des caractéristiques communes ou présentant un certain degré de similarité
- mesurer la co-occurence d'événements ou la fréquence de motifs plus ou moins réguliers
- Exemples : classification automatique ("clustering"), système de recommendation et filtrage collaboratif, text mining, etc.

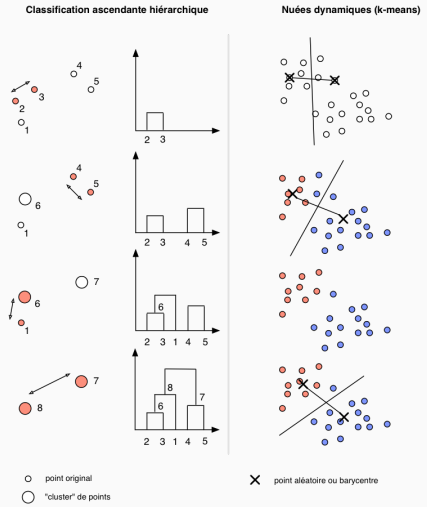
Classification hiérarchique

- mesure de dissimilarité ou de distance entre chaque paire d'observation : $||x - y||_2$ (euclidienne), $||x - y||_2^2$, $||x - y||_1$
- "aggrégation ascendante" : chaque observation définit son propre cluster, puis on regroupe les clusters par paires et on itère jusqu'à n'avoir plus qu'un seul cluster
- "aggrégation descendante" : on part d'un seul cluster et on divise
- critère d'agglomération : complete, single, average, centroid, diminution de la variance intra-classe (Ward), etc.

Nuées dynamiques (k-means)

- "aggrégation en centres mobiles" : regroupement itératif des observations par minimisation de la variance intra-cluster, après initialisation aléatoire des centres de classe
- choix d'une fonction distance : euclidienne, Manhattan (k-medians), et bien d'autres (k-medoids ou PAM)

Illustration



Tiré de B. Falissard, *Comprendre et utiliser les statistiques dans les sciences de la vie*, Masson 2005

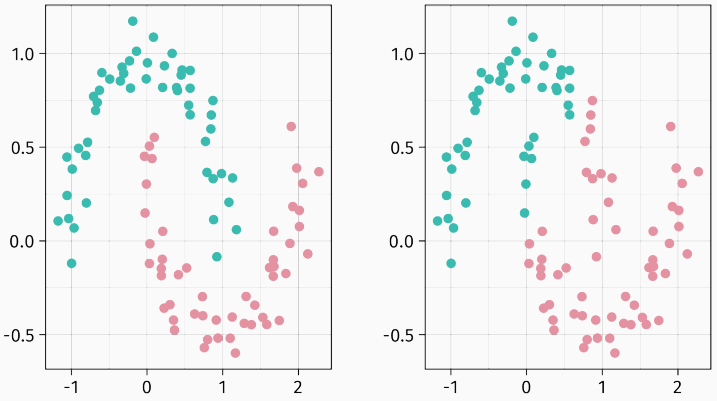
Application R

```
library(mclust)

hc.s ← hclust(dist(moon), method = "single")
hc.c ← hclust(dist(moon), method = "complete")

km ← kmeans(moon, centers = 2, nstart = 25)
mc ← Mclust(moon, G = 2)
```

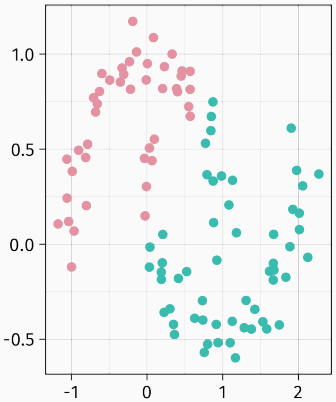
Application R (2)



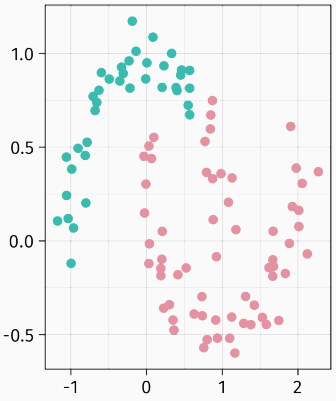
Single linkage

Complete linkage

Application R (3)



K-Means

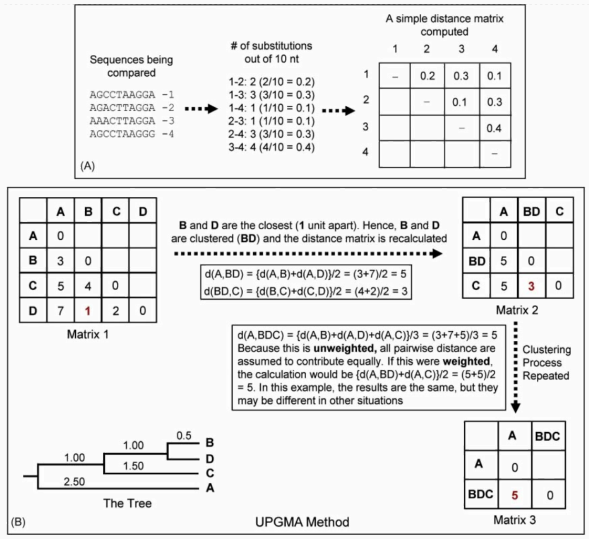


Model-based (Mclust 5.4)

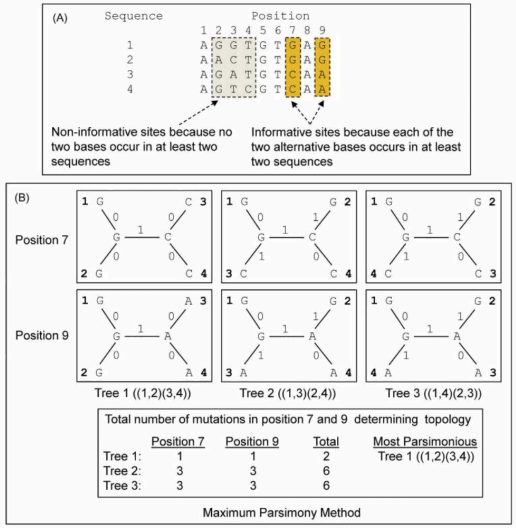
Classification automatique (en phylogénie)

- UPMGA : simple (proche de l'approche agglomérative) mais peu réaliste (suppose que la vitesse d'évolution est constante)
- NJ : tient compte de la vitesse d'évolution (additivité des distances) mais fournit un arbre non enraciné (et non ultramétrique) ; utile comme arbre initial pour les approches par vraisemblance
- parcimonie : minimiser le nombre "d'éditions" (mutations / substitutions) nécessaires pour passer d'une séquence à une autre (suppose les sites indépendants).

Modèles basés sur les distances (UPMGA)



Modèles basés sur les distances (parcimonie)



Approches par vraisemblance

Modèle de Jukes-Cantor

Des mutations aléatoires intervenant sur le génome au cours de leur évolution, un alignement de deux séquences ne donnera pas 100 % d'identité.

Matrice de transition (ACGT) hypothétique :

$$\begin{pmatrix} -0.886 & 0.190 & 0.633 & 0.063 \\ 0.253 & -0.696 & 0.127 & 0.316 \\ 1.266 & 0.190 & -1.519 & 0.063 \\ 0.253 & 0.949 & 0.127 & -1.329 \end{pmatrix}$$

Si l'on se trouve dans l'état A, on y restera un temps exponentiel de paramètre $-q_{ii} = 0.886$. La probabilité d'observer la transition A→C est donnée par $-q_{ij}/q_{ii} = \frac{0.190}{0.886}$.

Modèle de Jukes-Cantor (2)

Considérons que les événements surviennent au cours du temps, t, avec une fréquence λ. Le temps d'attente avant d'observer le prochain événement est alors décrit par une loi exponentielle, $f(t) = \lambda e^{-\lambda t}$, tandis que le temps d'attente avant d'observer le k-ième événement suit une loi Gamma, $f(t) = \frac{\lambda^k}{\Gamma(k)} t^{k-1} e^{-\lambda t}$. Le nombre d'événements survenant dans l'intervalle T suit une loi de Poisson de paramètre λT, $\Pr(k) = \frac{e^{-\lambda t}(\lambda T)^k}{k!}$.

Si l'on part de A, on devra attendre $\exp(0.886) = 2.425$ unités de temps, and les probabilités de passer à un autre état sont 0.214 (pour C), 0.714 (G) and 0.071 (T). La transition la plus probbale, $A \to G$, suppose d'attendre $\exp(-0.633) = 0.531$ unité de temps.

Temps de divergence = fréquence des états finaux au bout d'un temps t (distribution stationnaire).

Modèle de Jukes-Cantor (3)

$$\begin{pmatrix} -3\mu & \mu & \mu & \mu \\ \mu & -3\mu & \mu & \mu \\ \mu & \mu & -3\mu & \mu \\ \mu & \mu & \mu & -3\mu \end{pmatrix}$$

Ceci signifie que chaque nucléotide a un taux constant de mutation μ. La probabilité que deux espèces ait le même nucléotide sur un site homologue vaut donc Pr(same) = $\frac{1}{4}(1 + 3e^{-8\mu t})$, d'où Pr(different) = 1 – Pr(same) = $\frac{3}{4}(1 - e^{-8\mu t})$.[1]

[1] Jukes, T. H. and Cantor, C. R., Evolution of protein molecules. In Mammalian Protein Metabolism, ed. Munro, H. N., pp. 21-132, New York: Academic Press, 1969

Modèles de substitution

- plusieurs modèles de substitution : GTR +I/+G[2]
- GTR+I+Γ = 10 paramètres
- compromis bias-variance, sur-ajustement, cadre de raisonnement hypothétique

[2] `https://www.hiv.lanl.gov/content/sequence/findmodel/findmodel.html` ; [1] ; [3]

Modèle GTR

AIC-SELECTED MODEL: **GTR: General Time Reversible plus Gamma** (model 55)
LnL = -1315.630093
AIC = 2649.260186
Parameter details

The matrix on the right illustrates the different models considered by Findmodel.
Rate Parameters with the same color are assumed to have the same value.
Base Frequencies are the same when they are shown as f_N, and different otherwise.

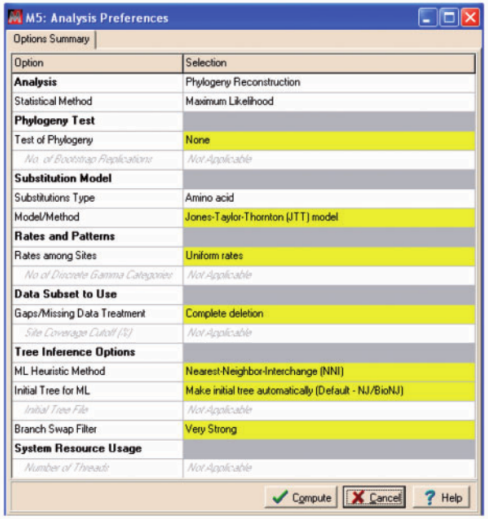
General Time Reversible

	T	C	A	G
T	f_T	a	b	c
C	a	f_C	d	e
A	b	d	f_A	f
G	c	e	f	f_G

Phyml

- utilisable directement depuis Seaview (sur séquences déjà alignées), ou en ligne
- utilisable également sous R (`plot` et `summary`)
- disponible en ligne sur NGPhylogeny

MEGA



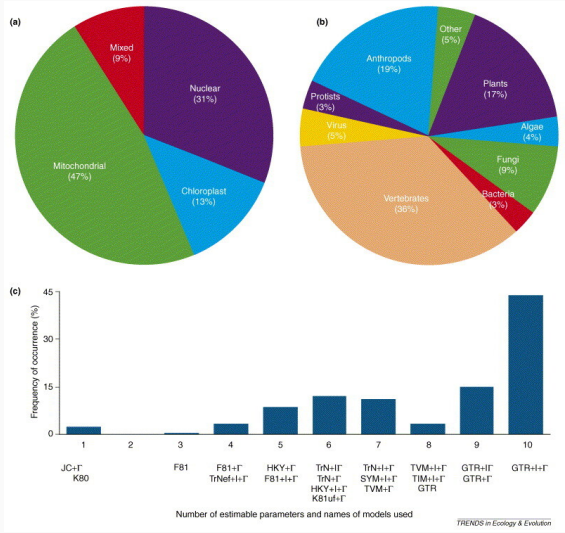
IQ-TREE : modèle de substitution NN

Model	df	Explanation	Code
JC or JC69	0	Equal substitution rates and equal base frequencies (Jukes and Cantor, 1969).	000000
F81	3	Equal rates but unequal base freq. (Felsenstein, 1981).	000000
K80 or K2P	1	Unequal transition/transversion rates and equal base freq. (Kimura, 1980).	010010
HKY or HKY85	4	Unequal transition/transversion rates and unequal base freq. (Hasegawa, Kishino and Yano, 1985).	010010
TN or TN93	5	Like `HKY` but unequal purine/pyrimidine rates (Tamura and Nei, 1993).	010020
TNe	2	Like `TN` but equal base freq.	010020
K81 or K3P	2	Three substitution types model and equal base freq. (Kimura, 1981).	012210
K81u	5	Like `K81` but unequal base freq.	012210
TPM2	2	AC=AT, AG=CT, CG=GT and equal base freq.	010212
TPM2u	5	Like `TPM2` but unequal base freq.	010212
TPM3	2	AC=CG, AG=CT, AT=GT and equal base freq.	012012
TPM3u	5	Like `TPM3` but unequal base freq.	012012
TIM	6	Transition model, AC=GT, AT=CG and unequal base freq.	012230
TIMe	3	Like `TIM` but equal base freq.	012230
TIM2	6	AC=AT, CG=GT and unequal base freq.	010232
TIM2e	3	Like `TIM2` but equal base freq.	010232
TIM3	6	AC=CG, AT=GT and unequal base freq.	012032
TIM3e	3	Like `TIM3` but equal base freq.	012032
TVM	7	Transversion model, AG=CT and unequal base freq.	012314
TVMe	4	Like `TVM` but equal base freq.	012314
SYM	5	Symmetric model with unequal rates but equal base freq. (Zharkikh, 1994).	012345
GTR	8	General time reversible model with unequal rates and unequal base freq. (Tavare, 1986).	012345

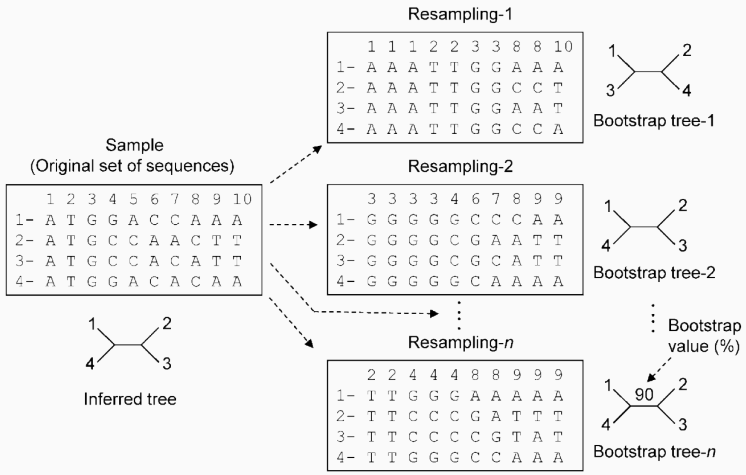
IQ-TREE : modèle de substitution AA

Model	Region	Explanation
Blosum62	nuclear	BLOcks SUbstitution Matrix (Henikoff and Henikoff, 1992). Note that `BLOSUM62` is not recommended for phylogenetic analysis as it was designed mainly for sequence alignments.
cpREV	chloroplast	chloroplast matrix (Adachi et al., 2000).
Dayhoff	nuclear	General matrix (Dayhoff et al., 1978).
DCMut	nuclear	Revised `Dayhoff` matrix (Kosiol and Goldman, 2005).
FLAVI	viral	Flavivirus (Le and Vinh, 2020).
FLU	viral	Influenza virus (Dang et al., 2010).
GTR20	general	General time reversible models with 190 rate parameters. *WARNING: Be careful when using this parameter-rich model as parameter estimates might not be stable, especially when not having enough phylogenetic information (e.g. not long enough alignments).*
HIVb	viral	HIV between-patient matrix HIV-B_{bt} (Nickle et al., 2007).
HIVw	viral	HIV within-patient matrix HIV-W_{in} (Nickle et al., 2007).
JTT	nuclear	General matrix (Jones et al., 1992).
JTTDCMut	nuclear	Revised `JTT` matrix (Kosiol and Goldman, 2005).
LG	nuclear	General matrix (Le and Gascuel, 2008).
mtART	mitochondrial	Mitochondrial Arthropoda (Abascal et al., 2007).
mtMAM	mitochondrial	Mitochondrial Mammalia (Yang et al., 1998).
mtREV	mitochondrial	Mitochondrial Vertebrate (Adachi and Hasegawa, 1996).
mtZOA	mitochondrial	Mitochondrial Metazoa (Animals) (Rota-Stabelli et al., 2009).
mtMet	mitochondrial	Mitochondrial Metazoa (Vinh et al., 2017).
mtVer	mitochondrial	Mitochondrial Vertebrate (Vinh et al., 2017).
mtInv	mitochondrial	Mitochondrial Inverterbrate (Vinh et al., 2017).
NQ.bird	nuclear	Non-reversible Q matrix (Dang et al., 2022) estimated for birds (Jarvis et al., 2015).
NQ.insect	nuclear	Non-reversible Q matrix (Dang et al., 2022) estimated for insects (Misof et al., 2014).
NQ.mammal	nuclear	Non-reversible Q matrix (Dang et al., 2022) estimated for mammals (Wu et al., 2018).
NQ.pfam	nuclear	General non-reversible Q matrix (Dang et al., 2022) estimated from Pfam version 31 database (El-Gebali et al., 2018).
NQ.plant	nuclear	Non-reversible Q matrix (Dang et al., 2022) estimated for plants (Ran et al., 2018).
NQ.yeast	nuclear	Non-reversible Q matrix (Dang et al., 2022) estimated for yeasts (Shen et al., 2018).
Poisson	none	Equal amino-acid exchange rates and frequencies.
PMB	nuclear	Probability Matrix from Blocks, revised `BLOSUM` matrix (Veerassamy et al., 2004).

Modèles d'évolution [2]



Stabilité de l'arbre (bootstrap)



Applications

- Evaluer la qualité du modèle
- Comparaison d'arbres
- Extraction de clades
- Combinaison d'arbres

Références i

[1] Subha Kalyaanamoorthy et al. "ModelFinder: fast model selection for accurate phylogenetic estimates". In: *Nature Methods* 14.6 (1958), pp. 587–589.

[2] Scot A. Kelchner and Michael A. Thomas. "Model Use in Phylogenetics: Nine Key Questions". In: *TRENDS in Ecology and Evolution* 22.2 (2006), pp. 87–94.

[3] Posada, D. and K. A. Crandall. "Modeltest: testing the model of DNA substitution". In: *Bioinformatics* 14.9 (1998), pp. 817–818.

[4] Ziheng Yang and Bruce Rannala. "Molecular phylogenetics: principles and practice". In: *Nature Reviews Genetics* 13 (2012), pp. 303–314.