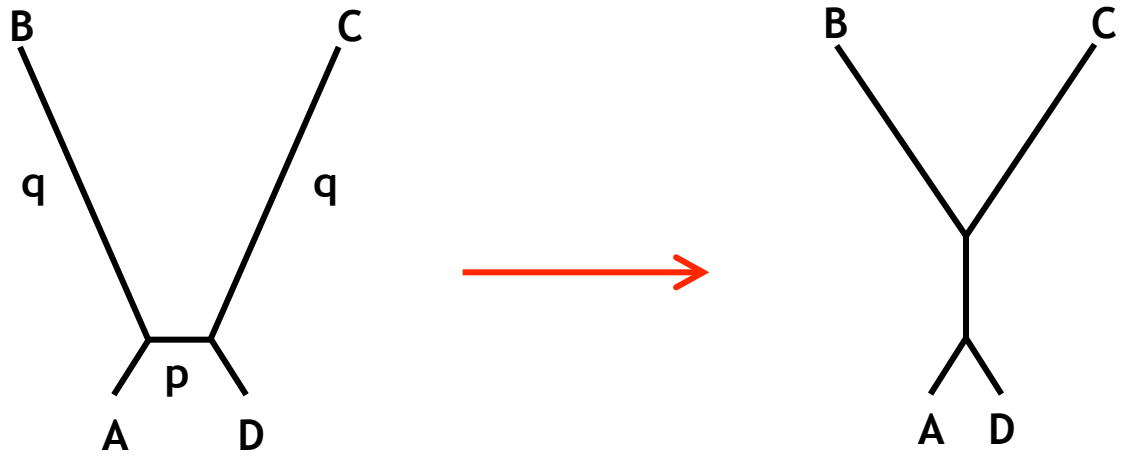
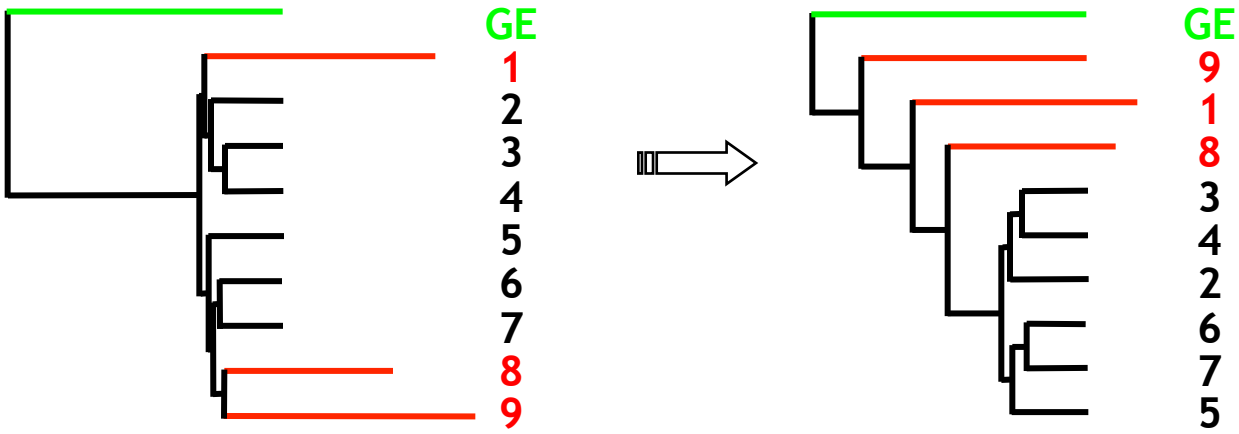
PARTIE IV

PARTIE IV

ATTRACTION DES LONGUES BRANCHES

Attraction des longues branches





GE = groupe extérieur distant

Base asymétrique

Artefact causé par les substitutions multiples

Attraction des longues branches

- Se produit lorsque les séquences sont très divergentes,

- Lié aux phénomènes de **substitutions multiples**. L'accumulation de **substitutions convergentes** est interprétée comme des synapomorphies.

- Les branches plus longues sont regroupées artificiellement car le nombre de **similitudes non homologues** entre les séquences est **supérieur** au nombre de similitudes que les séquences homologues ont conservé de leurs homologues véritables.

- L'attraction des longues branches affecte particulièrement les méthodes de parcimonie, notamment parce qu'elles n'autorisent pas de différence dans les taux de substitution.

Attraction des longues branches

Saturation – Cu-Zn Superoxyde-Dismutase

- Phylogénie intéressante car montre qu'un nœud bien soutenu (ici **99%**) peut-être totalement faux

⇨Un arbre robuste n'est pas forcément un arbre fiable !

⇨Trop grande différence sur certains caractères (sites de l'alignement) entre le nombre de mutations observées et le nombre de mutations inférées

Attraction des longues branches

Correction

1. Choix des sites : Gblock
2. Les méthodes de type **probabiliste** (ML, MB) permettent de réduire le phénomène d'attraction de branche longue en autorisant des différences de vitesse d'évolution (loi gamma, covarion) et en décrivant de manière réaliste les processus de substitution.

3. **L'ajout de taxons** dans une analyse phylogénétique permet de « casser les longues branches » en facilitant l'estimation des longueurs réelles de branches.

4. Le **modèle d'évolution moléculaire, CAT** (N. Lartillot, 2007), a permis de limiter certains phénomènes d'attraction.

Le modèle permet de rendre compte de l'hétérogénéité des sites de l'alignement en les affectant à des catégories possédant leurs propres modalités d'évolution moléculaire. Ces catégories reflètent théoriquement les propriétés biochimiques des domaines et des acides aminés concernés.

Lartillot N, Brinkmann H, Philippe H (2007) Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol Biol. 8;7 Suppl 1:S4

PARTIE IV

ROBUSTESSE DES ARBRES

Robustesse des arbres : Principe

- Méthodes statistiques pour l'estimation empirique de la variabilité d'un estimateur

En phylogénie, un arbre est un estimateur des données dont on dispose

- On va donc estimer la variabilité de l'arbre (ou d'une partie de l'arbre)
  Estimer nos données ⇔ Étude de la robustesse des arbres

=> Si un arbre est robuste i.e. fortement soutenu par les données, alors sa variabilité sera faible

=> Si un arbre est peu robuste alors il aura une grande variabilité

JackKnife
Bootstrap

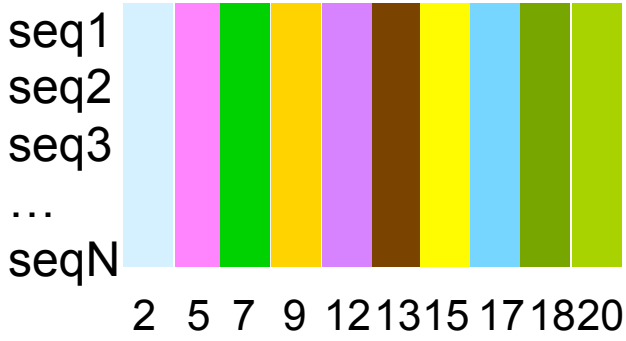
Test aLRT

Robustesse des arbres : Half-JackKnife

- Repose sur l'hypothèse de l'évolution indépendante des caractères au sein des séquences moléculaires
- On estime les phylogénies obtenues à partir d'un certain nombre d'échantillons de notre jeu de données initial
- Pour le half-jackknife on réalise X tirages sans remise de n/2 sites au sein du jeu de données initial

JDD initial

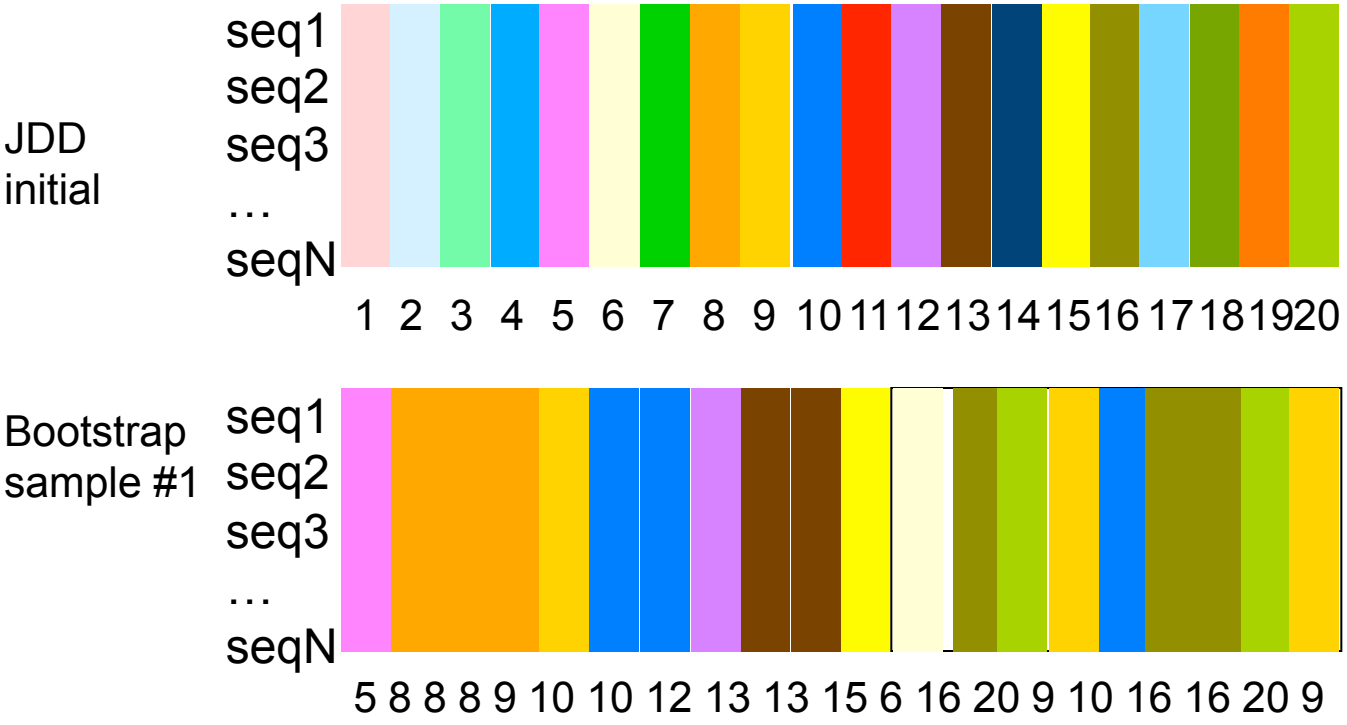
Jacknife sample #1



Utilisé pour valider l'ordre des gènes sur un génome :
Using jackknife to assess the quality of gene order phylogenies Jian Shi, Yiwei Zhang, Haiwei Luo and Jijun Tang. *BMC Bioinformatics* 2010, **11**:168

Robustesse des arbres : Bootstrap

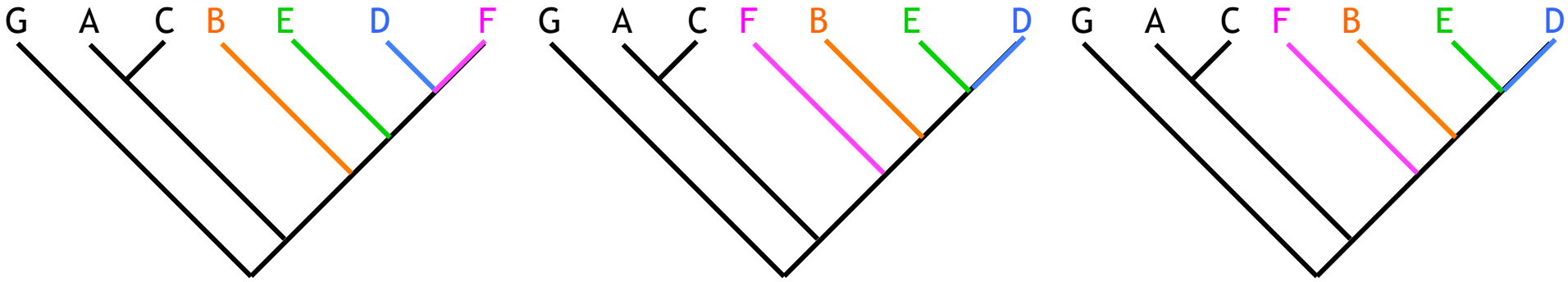
- Repose sur l'hypothèse de l'évolution indépendante des caractères au sein des séquences moléculaires
- On estime les phylogénies obtenues à partir d'un certain nombre de ré-échantillonnages de même taille que notre jeu de données initial
- On réalise X tirages avec remise de n sites parmi n sites au sein du JDD initial



Robustesse des arbres : Comparer

- Arbres consensus

Résumer sur un seul arbre l'information contenue dans plusieurs arbres obtenus par des méthodes ou d'après des données différentes pour le même jeu de données.



- Calcul de la distance entre arbres

En plus de s'intéresser aux structures communes entre des arbres, on peut essayer d'estimer comment ils sont différents

Arbre Consensus

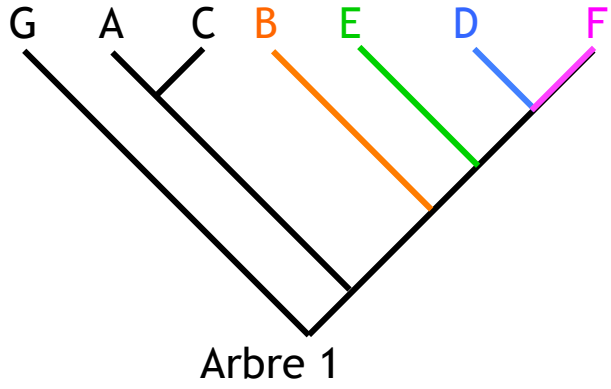
Consensus Strict (Rohlf 1982)

- Principe

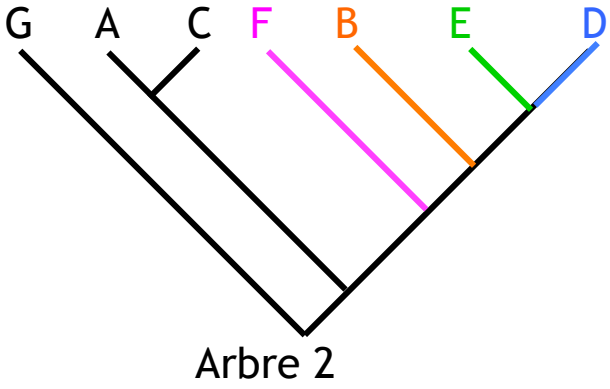
Construire un arbre qui montre tous **les groupes communs à tous les arbres**

Groupes ⇔ groupes monophylétiques si les arbres sont racinés

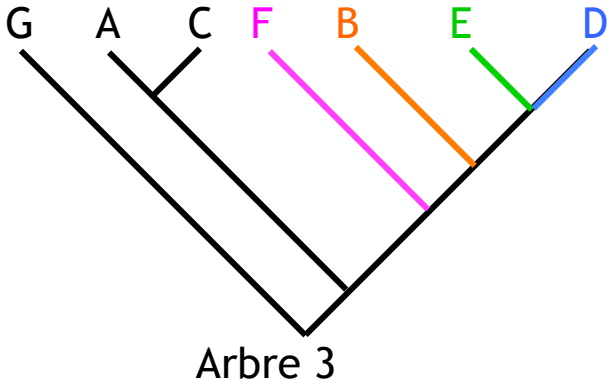
⇔ partitions si les arbres sont non racinés (Ex. la branche qui sépare BEDF du reste de l'arbre définit la partition {GAC|BEDF}



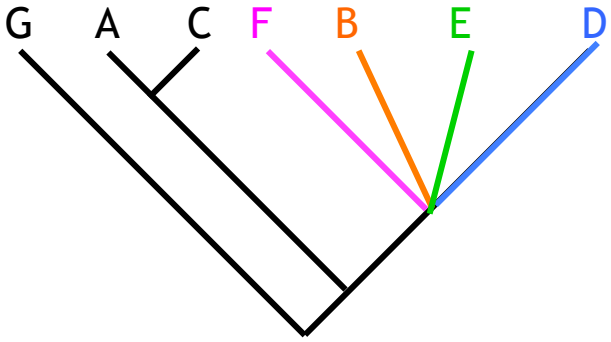
Arbre 1



Arbre 2



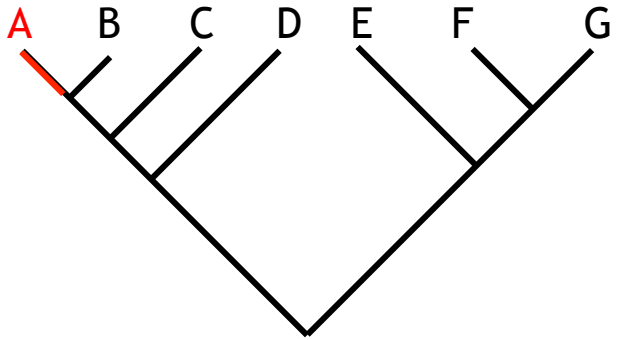
Arbre 3



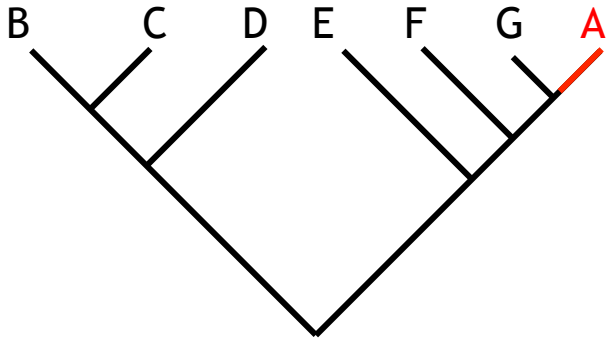
Arbre consensus

Les consensus stricts sont souvent trop stricts

Arbre Consensus : Consensus strict



Arbre 1

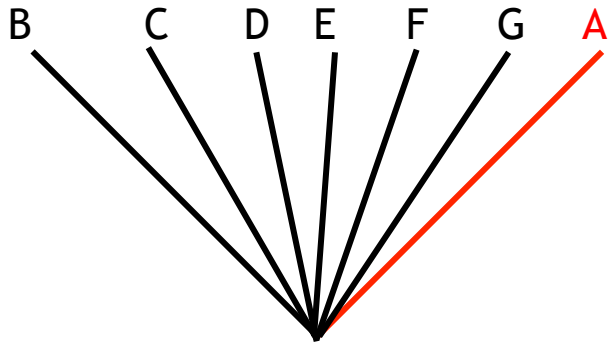


Arbre 2

Les arbres 1 et 2 sont très proches

=> Ils diffèrent uniquement par la position de l'espèce A

=> N'ont pas de groupes monophylétiques en commun



Arbre consensus

Arbre Consensus : consensus majoritaire

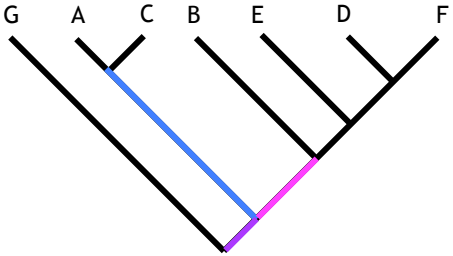
- Principe

Construire un arbre qui montre tous **les groupes communs à tous les arbres**

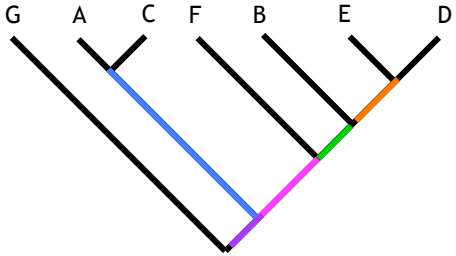
Groupes présents dans plus de X % des arbres

=> Arbres consensus (M_l)avec X = % variant de 50 à 100

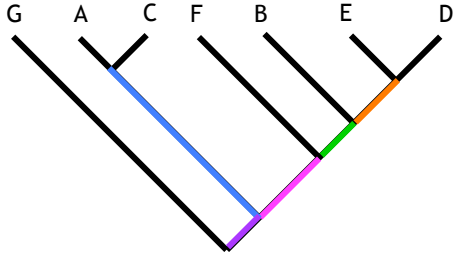
=> Si X = 100% l'arbre consensus construit est identique au consensus strict



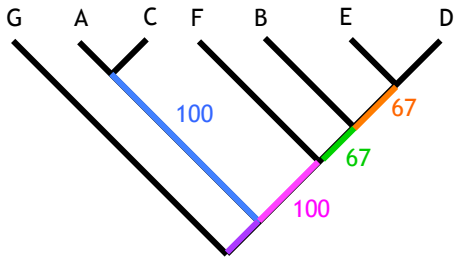
Arbre 1



Arbre 2



Arbre 3



Arbre consensus

Si on ajoute une copie d'un des arbres, l'arbre consensus majoritaire change

=>besoin d'une valeur de bootstraping importante

Robustesse des arbres: interprétation

- L'arbre consensus majoritaire quantifie de manière statistique les regroupements

  => Un nœud retrouvé dans 100% des arbres sera dit robuste

  => Un nœud retrouvé dans seulement 50% des arbres sera moins robuste…

  => Il n'existe pas de seuil ni de consensus universel (95%, 90%, 75%…)

  => On se contente de donner les nombres et d'interpréter

Une valeur de bootstrap de 100% ≠ un nœud vrai
ROBUSTESSE ≠ FIABILITE !

Une BV de 100% = un nœud ROBUSTE

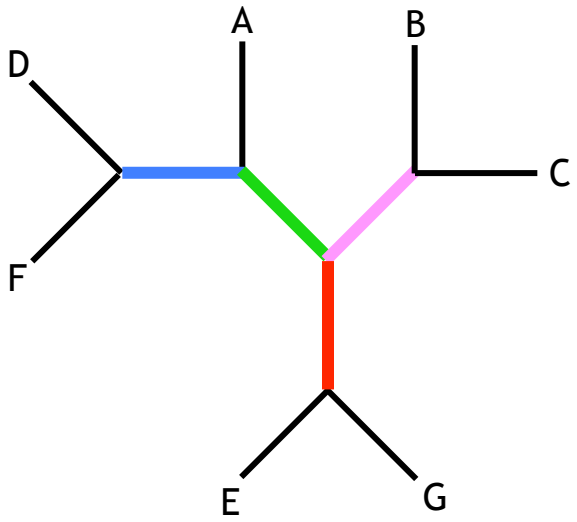
Distances entre arbres : Différence symétrique

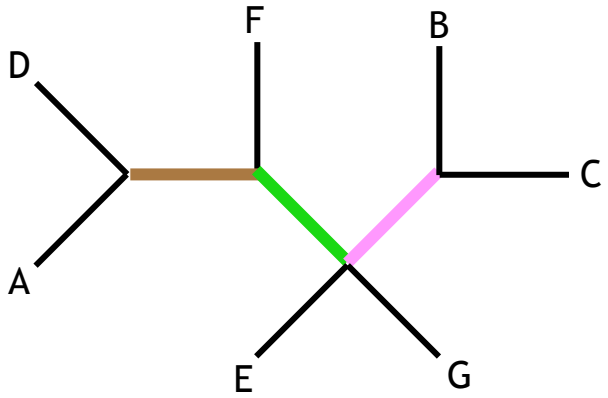
- **Principe**

Définition d'une distance entre une paire d'arbres basée sur le nombre de branches qui diffèrent entre les arbres

=> Pour chaque arbre, on établit la liste des partitions qu'il implique

=> La différence symétrique est le nombre de partitions, parmi cette liste de partitions, qui ne sont pas partagées par les autres arbres = nombre de partitions différentes



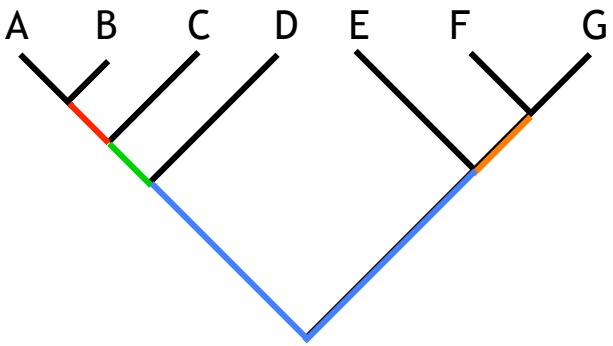


Partitions
{ADF | BCEG}
{DF | ABCEG}
{BC | ADEFG}
{EG | ABCDF}

Différence symétrique = 3

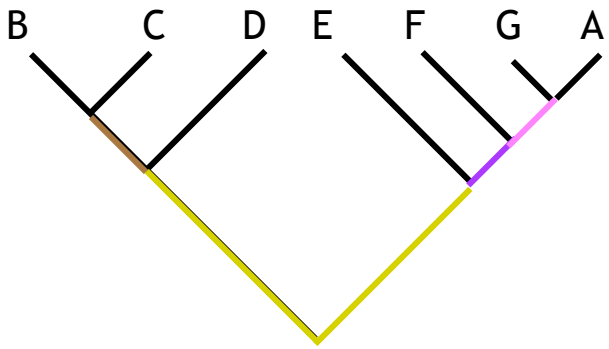
Partitions
{AD | FBCEG}
{ADF | BCEG}
{BC | ADEFG}

Distances entre arbres : Différence symétrique



Arbre 1

Partitions
{AB | CDEFG}
{ABC | DEFG}
{ABCD | EFG}
{FG | ABCDE}



Arbre 2

Partitions
{CB | DEFGA}
{BCD | AFGA}
{FGA | EDCB}
{GA | FEDCB}

Différence symétrique = 8

En dépit d'une structure proche, l'indice de différence symétrique est maximum (aucune partition n'est commune)

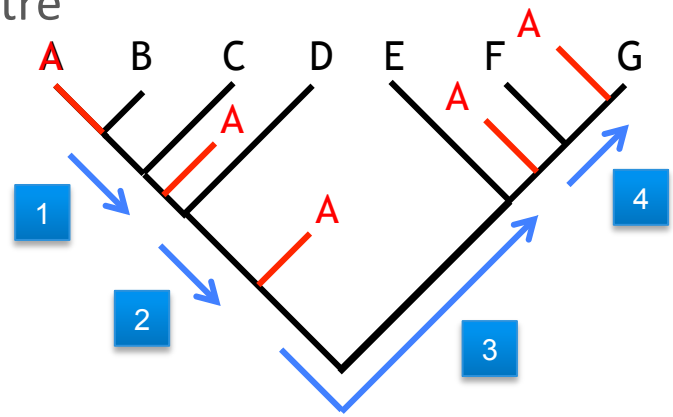
=> Méthode **insensible à des similarités partielles** qui peuvent exister entre les arbres

Distances entre arbres

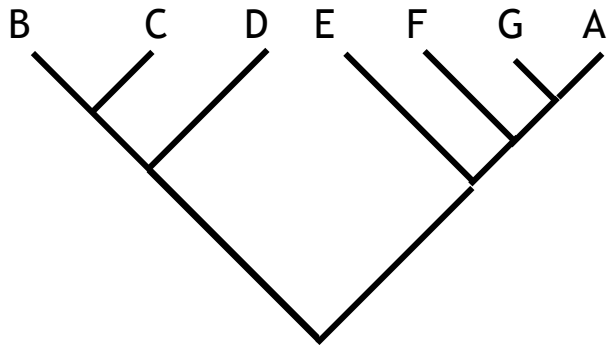
Distance de l'échange entre plus proche voisins (Waterman 1978)

- Principe

Utilise le nombre de réarrangements par échanges entre plus proches voisins (nearest-neighbor interchange NNI) nécessaires pour passer d'un arbre à l'autre



Arbre 1



Arbre 2

On peut passer de l'arbre 1 à l'arbre 2 en 4 réarrangements uniquement impliquant l'espèce A

NNI = 4

=> Devient rapidement incalculable pour des arbres impliquant un grand nombre de taxa

=> Utilisation de distances basées sur les réarrangements par SRP et TBR

Robustesse des arbres : Méthodes de vraisemblance

Likelihood Ratio Test :

- utilisé lorsque l'on souhaite comparer deux arbres ayant la même topologie, mais qui ont été obtenus avec des modèles d'évolution différents (ex. : ModelTest, test de l'horloge moléculaire, ...)

Kishino-Hasegawa (KH) et Shimodaira-Hasegawa (SH) tests :

- utilisés pour comparer des arbres dont les topologies diffèrent (ex. : tester la monophylie d'un groupe, déterminer si deux topologies distinctes sont significativement différentes ou non, ...)
- Les KH et SH tests sont implémentés dans PAUP*

Robustesse des arbres : Méthodes de vraisemblance

Test aLRT SH-like

- Permet d'étudier la significativité de la vraisemblance d'une branche (support de branche)
- Alternative au bootstrap, plus rapide qui ne nécessite pas de générer X arbres
- Implémenté dans PhyML

Vérifier la vraisemblance de l'arbre sans la branche étudiée :

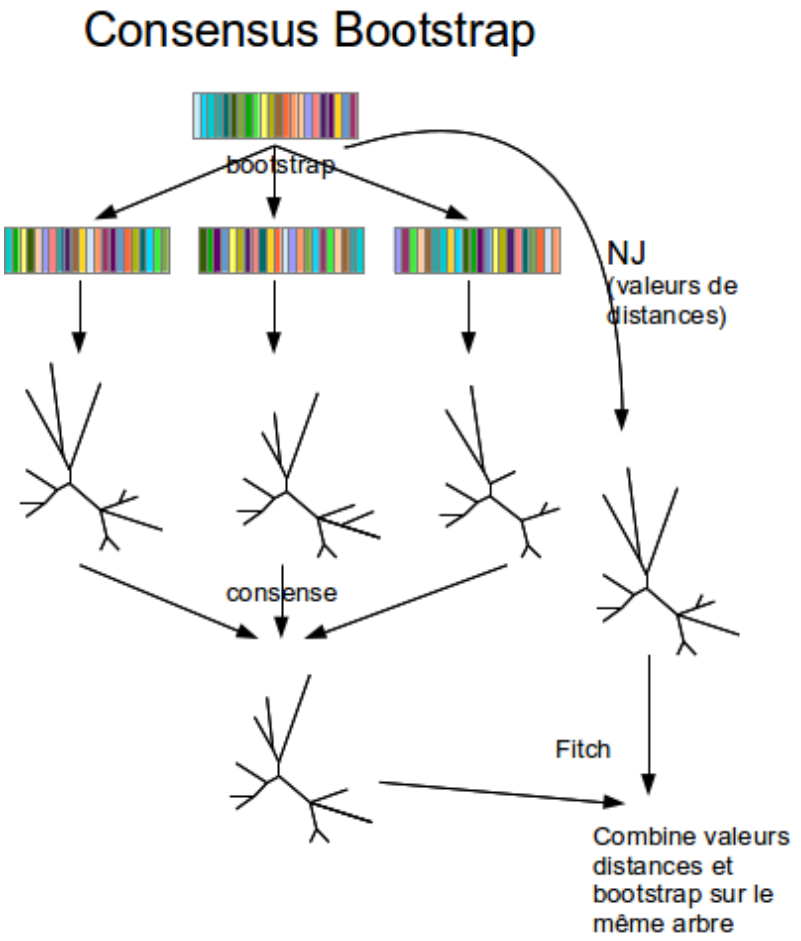
Test LRT standard : 2(L1 – L0) avec

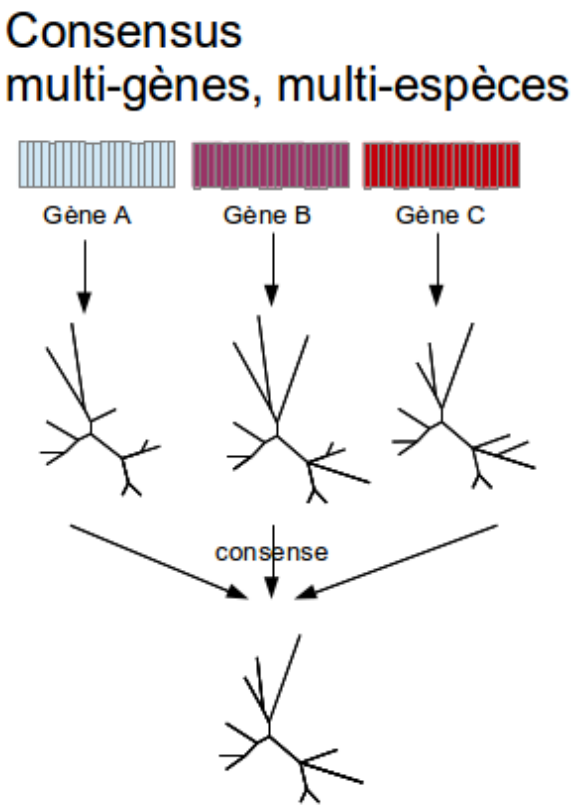
- L1 log vraisemblance de l'arbre
- L0 log vraisemblance de l'arbre sans la branche étudiée

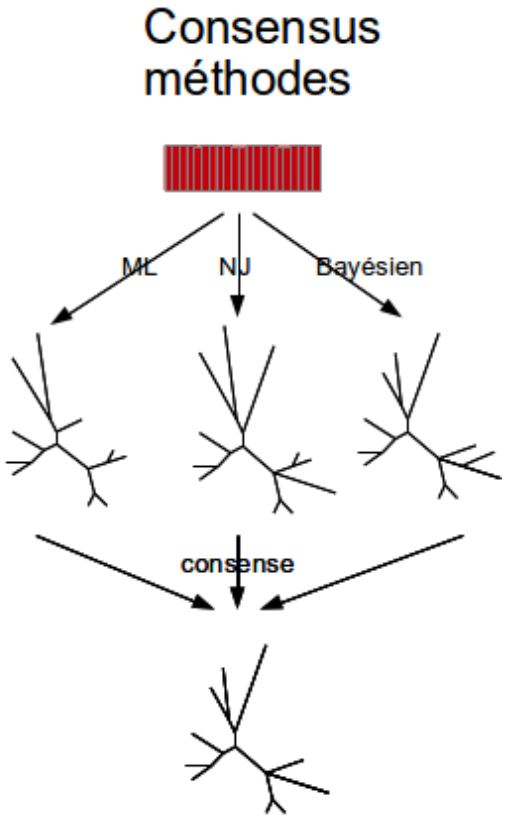
aLRT : 2(L1 – L2) avec

- L2 log vraisemblance avec la deuxième meilleure configuration NNIs autour de cette branche

Arbre consensus







PARTIE IV

ANALYSE MULTIGÉNIQUE ET SUPERARBRES

Analyses multigéniques et Superarbres

- Analyses multigéniques

Lorsque l'on souhaite analyser plusieurs marqueurs pour résoudre un problème phylogénétique, deux possibilités existent :

➢ Mettre bout à bout les séquences des différents gènes pour chaque espèce et faire les analyses sur ce "superalignement"

Conditions : Il faut si possible avoir le même échantillonnage d'espèces pour chaque gène, et il faut que les différents gènes puissent être considérés comme évoluant selon le même modèle.

Analyses multigéniques et Superarbres

- Analyses multigéniques

Lorsque l'on souhaite analyser plusieurs marqueurs pour résoudre un problème phylogénétique, deux possibilités existent :

➢ Analyser chaque gène séparément, constituer des arbres et comparer les résultats, en construisant par exemple un superarbre

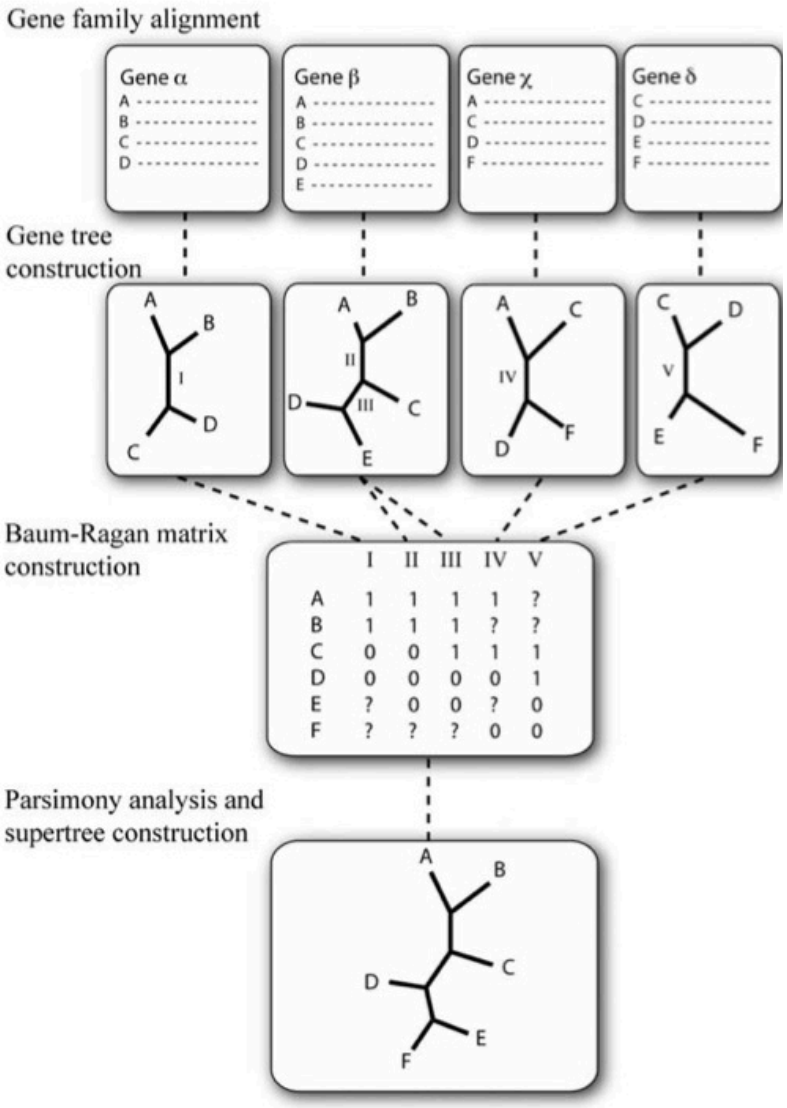
Principe : Le programme compare les arbres, code les noeuds internes de chaque arbre en fonction de la totalité des taxa présents dans tous les arbres, et crée une nouvelle matrice, qui permettra d'obtenir le superarbre.

Avantage : Il n'est pas nécessaire d'avoir le même échantillonnage d'espèces pour chaque gène.

Programmes : *RadCon*, *Clann, ….*

Methodes : Matrix representation with parsimony (MRP) method, Calculation of average consensus, Most Similar Supertree Algorithm (MSSA), Quartet, …

Analyses multigéniques et Superarbres



Matrix representation with parsimony (MRP) method.

La procédure MRP est la suivante:

- Une fois que l'alignement des familles de gènes est terminée, les arbres sont construits pour chacun des gènes séparément.

- Dans chacun de ces arbres, les branches internes (ou splits) sont identifiées (I-V ci-dessus).

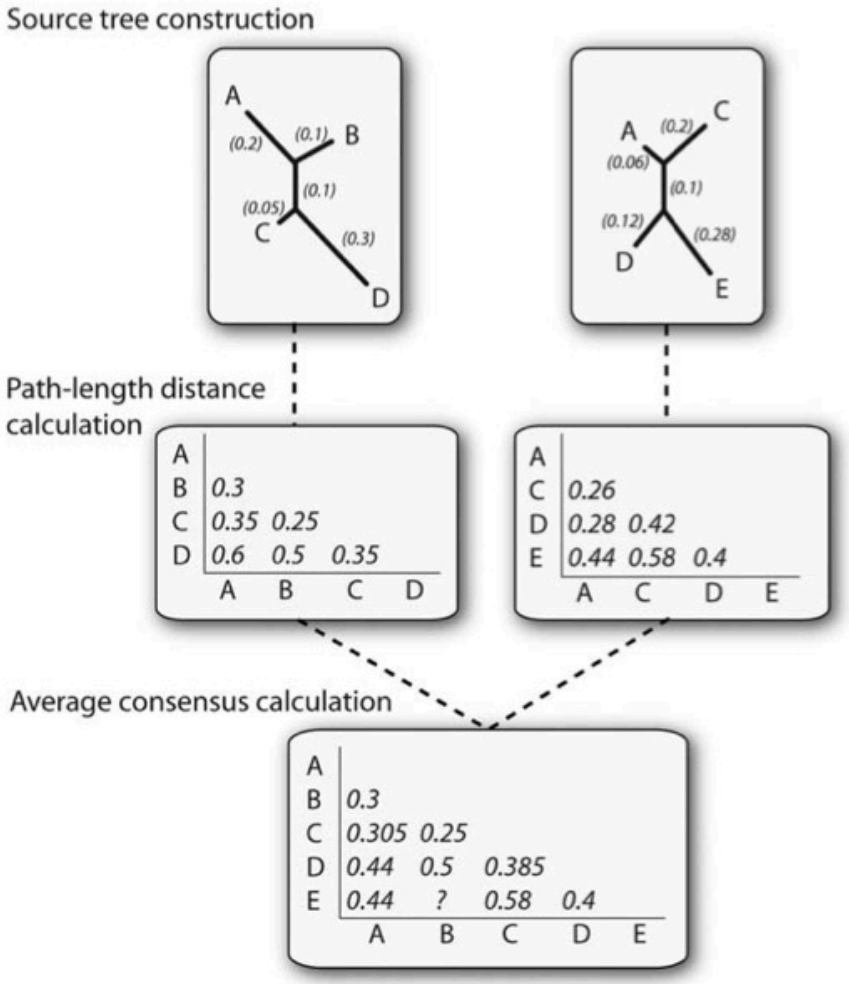
Un schéma de codage de Baum-Ragan est construit, contenant une colonne pour chacune des branches internes.

Remplissage de la matrice

Par exemple pour la branche interne I, les taxa A et B sont d'un côté et les taxa C et D sont de l'autre. Dans la matrice, les taxons A et B sont tous deux marqués d'un ’’1'' et taxons C et D sont tous deux marqués d'un'' 0.'' Comme les taxons E et F ne sont pas dans cet arbre, ils sont marqués par un ''?'' dans la colonne I.

Lorsque la matrice est terminée, une approche maximale de parcimonie est généralement utilisée pour reconstruire le superarbre.

Analyses multigéniques et Superarbres



Calcul du consensus moyen

Dans cette approche, les longueurs des branches des arbres sources sont utilisées pour calculer les distances de chaque taxon à tous les autres taxons.

Dans ces arbres, les chiffres entre parenthèses indiquent les longueurs des branches connexes.

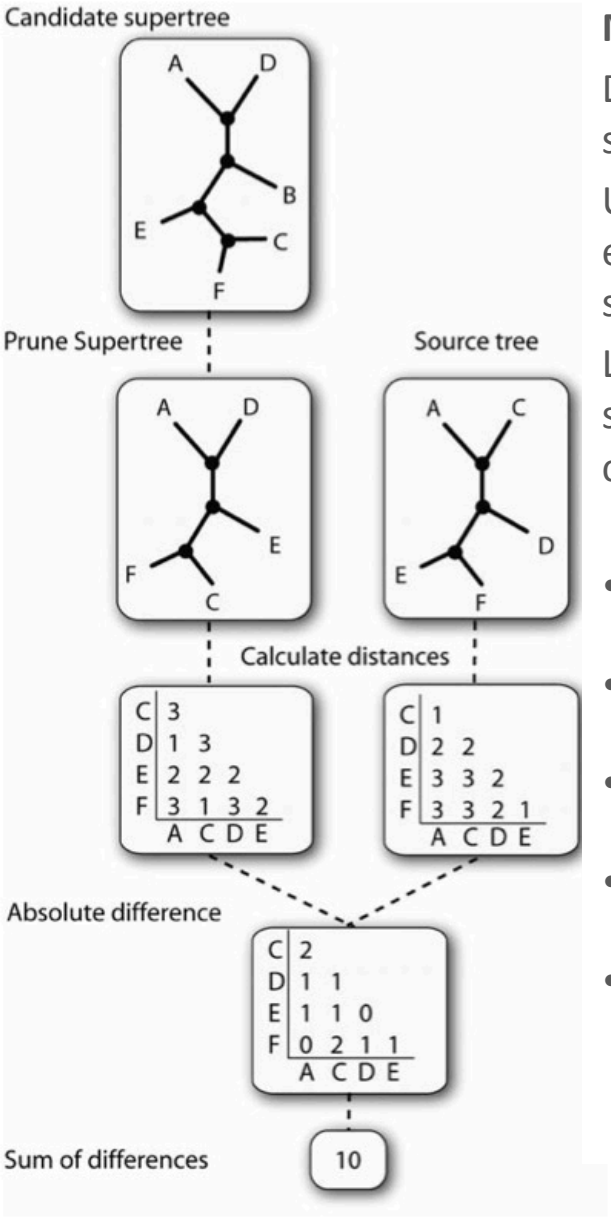
Par exemple, l'arbre sur la gauche a une distance du taxon A au taxon D de 0,6 (0,2 + 0,1 + 0,3).

La distance moyenne de chaque taxon à tout autre taxon est ensuite calculée pour déterminer le consensus moyen.

Par exemple, la distance de A à C dans les deux arbres sont 0,35 et 0,26. La moyenne (0,305) est le résultat mis dans la matrice de consensus moyenne.

Quelques distances ne sont pas calculables car les taxons n'apparaissent pas dans tous les arbres (comme avec les taxons B et E); dans ce cas, la valeur est estimée à partir des valeurs qui l'entourent dans la matrice de consensus moyen.

Analyses multigéniques et Superarbres



Most Similar Supertree Algorithm (MSSA).

Dans cette approche, une fonction est utilisée pour évaluer les superarbres candidats .

Une recherche heuristique ou exhaustive de l'espace des supers arbres est effectuée et le super arbre qui minimise la fonction est le plus semblable à l'ensemble des arbres sources .

La différence entre le SA candidat et chaque arbre source est calculée séparément et la somme de ces scores est la note globale pour le SA candidat.

- Pour chaque comparaison à un arbre source le SA est d'abord « réduit » au même nombre de taxa que l'arbre source.
- Un score de distance représentant les différences entre les deux arbres est calculé.
- Les distances sont le nombre de noeuds internes (cercles pleins dans les arbres ci-dessus) qui sont dans le chemin entre deux taxa sur l'arbre.
- La somme des différences absolues entre les matrices est le score représentant la différence entre le SA et cet arbre source.
- Cette valeur est généralement divisée par soit le nombre de comparaisons dans la matrice soit par le nombre d'espèces partagées entre le SA et l'arbre source (afin de pondérer le biais lié à l'étude de grands arbres sources).

PARTIE IV

MESURE DE LA PRESSION ÉVOLUTIVE

Pression de sélection

Pression de sélection

Evolution : concepts modernes

La théorie sélectionniste

- La plupart des nouveaux allèles apparus par mutations se fixent dans les populations parce qu'ils sont avantageux pour les porteurs dans le milieu où ils vivent (sélection darwinienne).
- La plupart des mutations sont délétères, elles réduisent la fitness de l'organisme
- Elles sont éliminées de la population par sélection purificatrice.
- Quelques mutations sont avantageuses, elles augmentent la fitness de l'organisme
- Elles sont gardées par sélection positive darwinienne (sélection adaptative)
- La fitness est fonction du taux de survie et de la fécondité

La théorie neutraliste (Kimura)

- La plupart des mutations restent neutres, se fixent au hasard (seules les mutations très défavorisantes ou létales pour l'individu sont éliminées) et le milieu n'a pas de rôle sélectif.
- Dans cette théorie, la sélection naturelle perd son caractère de **facteur évolutif** prépondérant et devient **un facteur parmi d'autres** au nombre desquels on compte les facteurs stochastiques tel que la dérive génétique. Le rôle de la sélection naturelle reste néanmoins très important pour l'évolution.

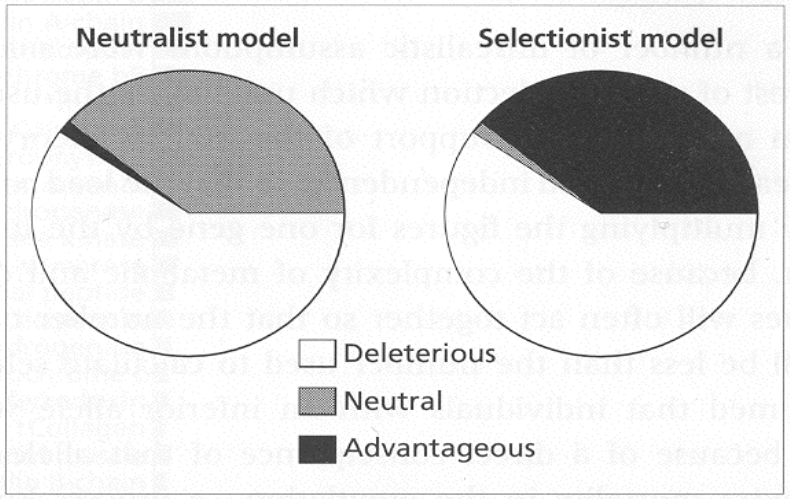
-> corollaire horloge moléculaire

Motoo Kimura
En 1968 Kimura analyse les changements dans l'hémoglobine, le cytochrome c et le triose-phosphate dehydrogenase.
Kimura (1968) Nature 217 624-626

Pression de sélection

Evolution : concepts modernes

Une différence de vue de l'évolution entre "Selectionistes" and "Neutralistes"



Fraction des mutations aléatoires supposées être délétères, neutres, et avantageuses

Point de vue Selectioniste :
La plupart des mutations observées sont des innovations fonctionnelles

Point de vue Neutraliste :
La plupart des mutations observées représentent des changements conservatifs cad des changements dans des régions non importantes.

Pression de sélection

Evolution : concepts modernes

L'HORLOGE MOLECULAIRE.

Le concept d'horloge moléculaire est basé sur l'hypothèse que l'accumulation progressive des mutations au cours du temps ne dépend que du taux d'erreur de la DNA polymérase et de la pression de sélection.

1. pour des régions non codantes, l'horloge tourne à son rythme maximal
2. Pour des régions codantes, la vitesse de l'horloge dépend des contraintes de la pression de sélection sur cette protéine.

Dans la réalité les taux d'accumulation des mutations :

1. Peuvent être différents d'un organisme à un autre (f° de la capacité de reproduction)
2. Peuvent varier au cours du temps dans une lignée
3. Ne sont pas identiques d'un résidu à un autre

Ces problèmes peuvent :

1. Affecter l'efficacité des méthodes de reconstruction.
2. Empêcher de dater précisément les divergences.

-> Désormais : Horloge moléculaire relaxée

Pression de sélection

Différents type de sélection

$$\frac{\text{Substitution non synonymes}}{\text{Substitution synonymes}} = K_A / K_S = d_N / d_S = \omega$$

Pathway 1: CTA (Leu) --> GTA (Val) --> GTT (Ile) 2 non-synonymous changes

Pathway 2: CTA (Leu) --> CTT (Leu) --> GTT (Ile) 1 synonymous, 1 non-synonymous change

Les substitutions synonymes sont peu ou pas soumises à la sélection.

$d_N / d_S < 1$ ➔ sélection « purifiante » (majorité des gènes : une mutation délétère a peu de chance de se fixer)

$d_N / d_S = 1$ ➔ neutralité (typique des pseudogènes)

$d_N / d_S > 1$ ➔ sélection « positive » (ou « Darwinienne») (exemple relation hote-pathogène où adaptation permanente aux modifications des protéines du partenaire)

Mesure possible sur un gène complet ou sur un site

Voir document additionnel pdf Ka/Ks

Pression de sélection

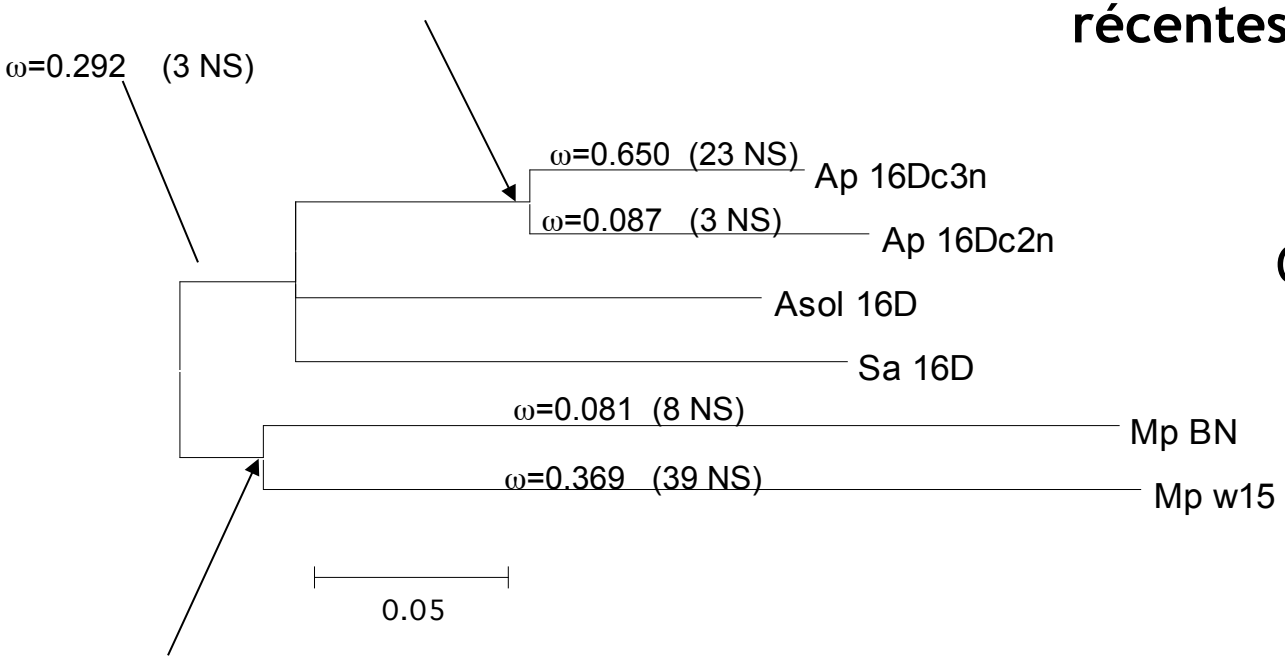
Mesure de pression

cysteine protease genes of the family cathepsin B
Acyrthosiphon pisum and *Myzus persicae*

Pression de sélection

Mesure de pression

Hétérogénité des taux d'évolution (entre copies récentes) ?



Clade « 16D »

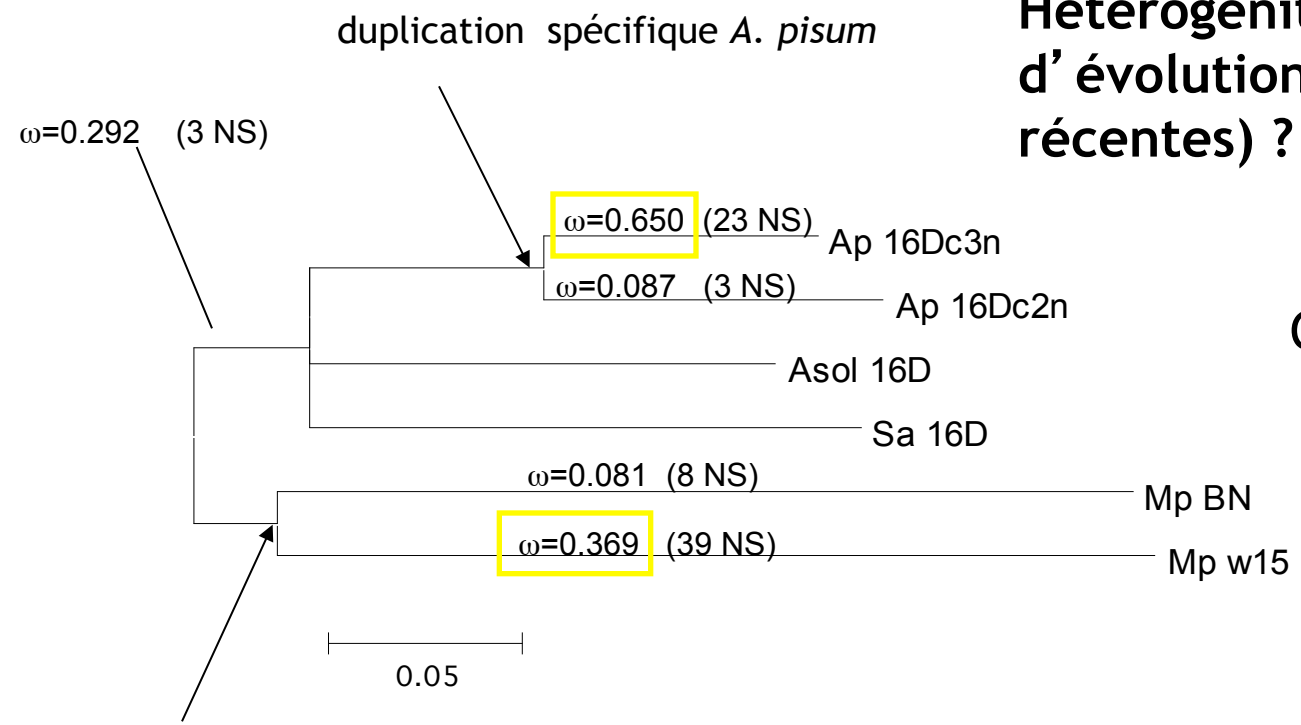
Analyse PAML, modèle « several ratios »

Topologie fixée, longueurs de branches / dS

Significativité / modèle M0: P=0.00003

Pression de sélection

Mesure de pression



Hétérogénité des taux d'évolution (entre copies récentes) ?

Clade « 16D »

Analyse PAML, modèle « several ratios »

Topologie fixée, longueurs de branches / dS

Significativité / modèle M0: P=0.00003

Une des deux copies dupliquées (chez *Acyrthosiphon pisum* et *Myzus persicae*) connaît une accélération évolutive – sélection relâchée ?