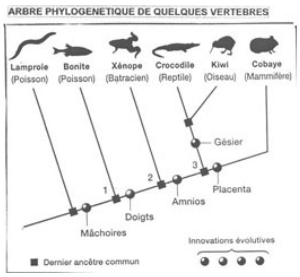
Reconstruction phylogénétique

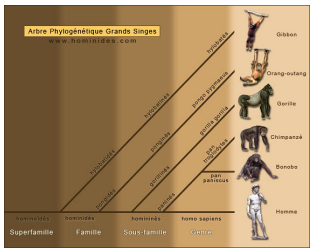
Équipe Bonsai

http://www.lifl.fr/bonsai

année 2011

Exemples





Problématique

- retracer l'historique des espèces ou d'un marqueur à partir des mutations observées
- données : gènes communs aux familles étudiées, pas trop divergents

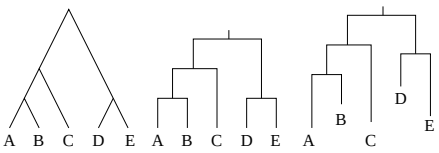
exemple : *cytochrome B de l'ADN mitochondrial*

- résultat : classification sous forme d'arbre phylogénétique

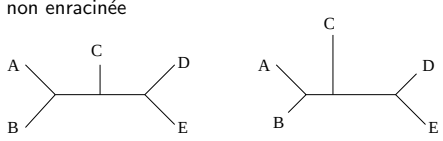
Les représentations

- enracinée

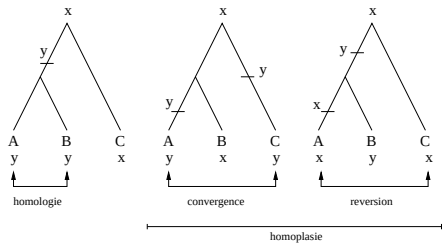
(((A,B),C),(D,E))



- non enracinée



Le concept de ressemblance



- homologie : héritage d'un ancêtre commun.
- homoplasie : apparition indépendante de caractères similaires.

Les écoles phylogénétiques

- systématique phénétique :
  - similitude globale basée sur le concept de distance
  - homologies et homoplasies confondues
  - arbre = dendogramme
- systématique cladistique :
  - seules les homologies sont regardées
  - similitude basée sur la recherche de passage d'un caractère d'un état primitif (plésiomorphe) → état dérivé (apomorphe)
  - arbre = cladogramme

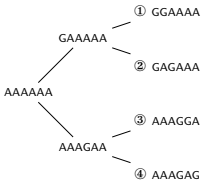
Quelles données ?

- données morphologiques :
  - forme des ailes,
  - présence/absence de dents,
  - nombre d'ongles ...
- données moléculaires :
  - ADN, ARN, protéines,
  - fréquence de gènes,
  - données d'expression ...
- 2 types de données
  - données discrètes : analyse de caractères
  - données globales : analyse par distance

Traitement des données moléculaires

- construction d'un alignement multiple
- quelle méthode d'alignement ?
  - pas de méthode a priori meilleure
  - attention aux hypothèses phylogénétiques (guide tree de CLUSTAL)
  - toujours revoir l'alignement à la main et l'ajuster
- quels sites conserver ?
  - supprimer tous les sites d'indels
  - supprimer les regions estimées trop variables ou modifier l'alignement pour ajouter des données manquantes
  - avantage : pas de modèle ad hoc à donner pour décrire les indels
  - désavantage : perte de l'information contenue dans ces régions

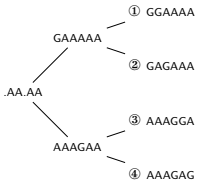
Un exemple idéal



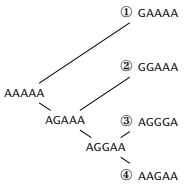
matrice de distances :

	①	②	③	④
①	0	...		
②	2	0	...	
③	4	4	0	...
④	4	4	2	0

arbre calculé :



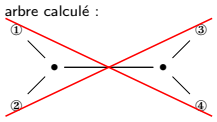
Un exemple plus réaliste



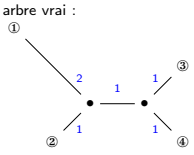
matrice de distances :

	①	②	③	④
①	0	...		
②	3	0	...	
③	4	3	0	...
④	4	3	2	0

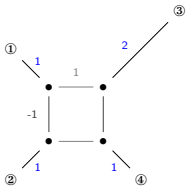
arbre calculé :



arbre vrai :



arbre représentant la matrice :



La notion de ressemblance

$d(i,j) = 1$
$p(i,j) = 1$

$d(i,j) = 0$
$p(i,j) = 2$
convergence

$d(i,j) = 0$
$p(i,j) = 2$
réversion

$d(i,j) = 1$
$p(i,j) = 2$
mutations mutiples

$d(i,j) = 1$
$p(i,j) = 4$
convergence + réversion + mutations mutiples

Les problèmes d'une analyse phylogénétique

- le nombre de substitutions réelles définit une distance d'arbre dont la représentation est l'arbre vrai (à l'horloge moléculaire près)
- à cause des substitutions multiples, le nombre de différences observées sous-estime le nombre de substitutions réelles
  - supposer que les subst. multiples sont rares : rechercher l'arbre le plus parcimonieux
  - poser un modèle stochastique d'évolution :
    - ★ estimer le nb réel de substitutions à partir de la distance observée et utiliser une méthode de distance,
    - ★ chercher l'arbre le plus vraisemblable dans le modèle

Correction des distances

- protéines : application d'une matrice de substitution
  PAM, BLOSUM, etc.
- ADN/ARN : application d'un modèle

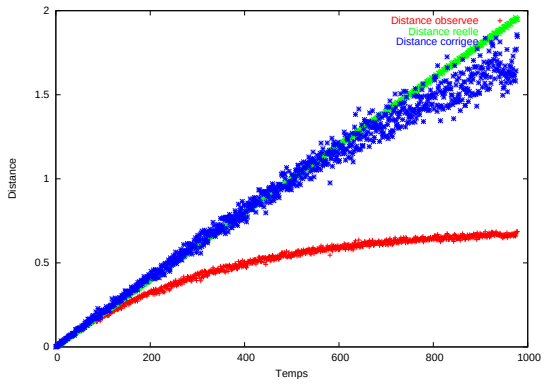
Le modèle de Jukes et Cantor

- soit u le taux de substitution par unité de temps
- pendant une courte période de temps dt, la probabilité d'un changement est $a = u \times dt$
- et donc la probabilité de conservation est $1 - 3 \times a$

$$\begin{pmatrix} 1-3a & a & a & a \\ a & 1-3a & a & a \\ a & a & 1-3a & a \\ a & a & a & 1-3a \end{pmatrix}$$

- probabilité de muter d'un site : $p = \frac{3}{4}(1 - e^{-\frac{4}{3}ut})$
- nouvelle distance : $d_{JC} = ut = -\frac{3}{4}\ln(1 - \frac{4}{3}p)$ où p sera la fréquence de mutation observée : $p = \frac{\text{nb mutations}}{\text{taille sequence}}$

Résultat de la correction JC



Le modèle de Kimura à 2 paramètres

- distingue transitions et transversions
- $a = u \times dt$ la probabilité de transition
- $b = v \times dt$ la probabilité de transversion (v est la demi-probabilité de transversion)
- on a alors :

$$\begin{pmatrix} 1-a-2b & a & b & b \\ a & 1-a-2b & b & b \\ b & b & 1-a-2b & a \\ b & b & a & 1-a-2b \end{pmatrix}$$

- $d_{K2P} = \frac{1}{2}\ln(\frac{1}{1-2P-Q}) + \frac{1}{4}\ln(\frac{1}{1-2Q})$ avec P et Q les fréquences de transition/transversion

Méthodes de parcimonie

- privilégier l'arbre qui minimise le nombre de mutations
- le nombre global de mutations est obtenu en faisant la somme des mutations le long de chaque branche

méthode :

- générer toutes les topologies
  $\Rightarrow$ nombre d'arbres non enracinés ?
- estimer le nombre de mutations
  $\Rightarrow$ nécessite de trouver les ancêtres : algo de Fitch appliqué indépendamment à chaque site
- problème NP-complet : utilisation d'heuristiques

Algorithme de Fitch

```
Entrée : un arbre dont les feuilles sont étiquetées
Sortie : un coût sur chaque noeud et un ensemble d'étiquettes possibles

si le noeud est une feuille
alors
  le coût est nul, l'étiquette n'est pas modifiée
sinon
  appeler récursivement la fonction sur les fils
  si l'intersection des étiquettes des fils est non vide
  alors
    étiqueter le noeud par l'intersection, ajouter les coûts
  sinon
    étiqueter le noeud par l'union, ajouter les coûts + 1
  fin si
fin si
```

Principe et hypothèses

- les explications les plus courtes sont les meilleures (rasoir d'Occam)
- cherche à miniminer le nombre de changements le long des branches
- les sites évoluent de manière indépendante
- la vitesse d'évolution est lente et constante au cours du temps

Différentes parcimonies

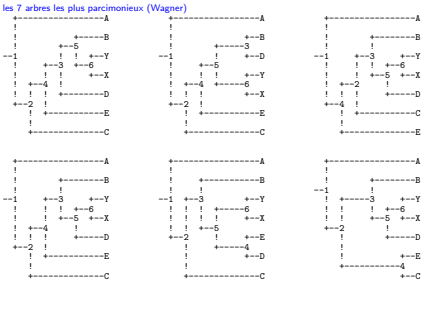
- Parcimonie de Wagner (1969, 1970)
  - convergences et réversions sont acceptées
  - un état peut passer de 0 à 1 puis de 1 à 0, et ceci plusieurs fois
  - un état peut passer de 0 à 1 sur des branches différentes
- Parcimonie de Camin-Sokal (1965)
  - les réversions sont exclues
  - il est donc nécessaire de connaître *a priori* l'état ancestral
- Parcimonie de Dollo (1972, 1977)
  - les convergences sont exclues
  - concept de caractère dérivé unique (Le Quesne, 1972)
  - plus facile de perdre un caractère, que de l'acquérir en parallèle

Consensus

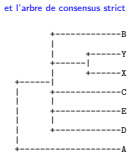
- idée : reconstruire un arbre à partir des noeuds les plus fréquents dans un ensemble d'arbres
- différents consensus :
  - strict : seuls les noeuds qui ont été trouvés dans tous les arbres sont inclus
  - majoritaire : seuls les noeuds présents dans au moins la moitié des arbres sont inclus

Résultat de penny + consense

les 7 arbres les plus parcimonieux (Wagner)

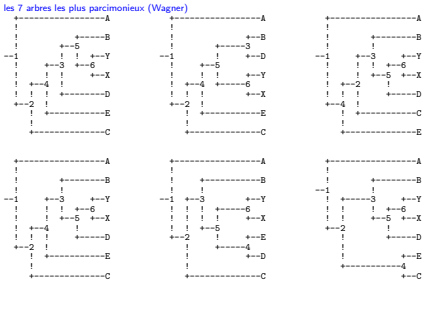


et l'arbre de consensus strict



Résultat de penny + consense

les 7 arbres les plus parcimonieux (Wagner)



et l'arbre de consensus majoritaire

Méthodes de distance

- point de départ : alignement multiple
- matrice de toutes les distances deux à deux
- classification hiérarchique

  On construit l'arbre à partir des feuilles en regroupant progressivement les noeuds 2 à 2 pour former des **clusters**.

- bootstrap : évaluation de la robustesse de l'arbre

UPGMA

Unweight Pair Group Method with Arithmetic mean

- algorithme de clusterisation séquentiel
- les relations sont envisagées dans l'ordre décroissant de leur similarité
- la reconstruction se fait pas à pas en suivant cet ordre
  - recherche des deux espèces les plus proches,
  - ces deux espèces forment un groupe considéré dans la suite comme une seule entité,
  - on recherche ensuite l'espèce (le groupe) la plus proche d'une autre espèce (un autre groupe)

$$d_{x,y} = \frac{1}{|x| \times |y|} \sum_{i \in x} \sum_{j \in y} d_{i,j}$$

  - on recommence jusqu'à avoir envisagé toutes les espèces

UPGMA

Exemple

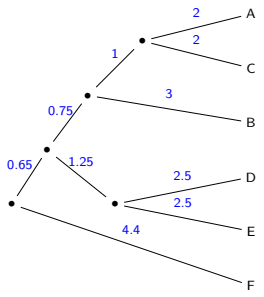
Etape 1	A	B	C	D	E
B	5				
C	4	7			
D	7	10	7		
E	6	9	6	5	
F	8	11	8	9	8

Etape 2	A+C	B	D	E
B	6			
D	7	10		
E	6	9	5	
F	8	11	9	8

Etape 3	A+C	B	D+E
B	6		
D+E	6.5	9.5	
F	8	11	8.5

Etape 4	AC+B	D+E
D+E	7.5	
F	9	8.5

Etape 5	ACB+DE
F	8.8



Résultat de `neighbor` option `upgma`

```
   6 Populations

Neighbor-Joining/UPGMA method version 3.6a2.1

 UPGMA method

 Negative branch lengths allowed

Name                 Distances
----            ----------
sequence A    0.00000   5.00000   4.00000   7.00000   6.00000   8.00000
sequence B    5.00000   0.00000   7.00000  10.00000   9.00000  11.00000
sequence C    4.00000   7.00000   0.00000   7.00000   6.00000   8.00000
sequence D    7.00000  10.00000   7.00000   0.00000   5.00000   9.00000
sequence E    6.00000   9.00000   6.00000   5.00000   0.00000   8.00000
sequence F    8.00000  11.00000   8.00000   9.00000   8.00000   0.00000


                          +------------------------sequence A
                +---------1
          +-----3         +------------------------sequence C
          !     !
  +-------4     +--------------------------------sequence B
  !       !
  !       !               +--------------------------------sequence D
--5       +---------------2
  !                       +--------------------------------sequence E
  !
  +-----------------------------------------------------sequence F


From     To           Length          Height
----     --           ------          ------
   5      4          0.65000         0.65000
   4      3          0.75000         1.40000
   3      1          1.00000         2.40000
   1   sequence A    2.00000         4.40000
   1   sequence C    2.00000         4.40000
   3   sequence B    3.00000         4.40000
   4      2          1.25000         1.90000
   2   sequence D    2.50000         4.40000
   2   sequence E    2.50000         4.40000
   5   sequence F    4.40000         4.40000
```


Matrices de distance et UPGMA

- UPGMA ne fonctionne pas ici : d(E,F) = 8.8 sur l'arbre alors que c'était 8 dans la matrice
- pourquoi :

  il faut que la distance soit ultramétrique : pour tout triplet de séquences x, y, z deux des distances d_{xy}, d_{xz}, d_{yz} sont égales et plus grandes que la troisième
- tout arbre respectant l'horloge moléculaire implique une distance ultramétrique

Et si la distance n'est pas ultramétrique ?

- elle peut être additive : pour tout quadruplet de séquences x, y, z, t deux des sommes $d_{xy} + d_{zt}, d_{xz} + d_{yt}, d_{xt} + d_{yz}$ sont égales et plus grandes que la troisième (condition des quatre points)
- dans ce cas, on peut reconstruire un arbre
- l'horloge moléculaire n'est pas nécessaire

Neighbor-Joining

Satou et Nei, 1987

- autorise un taux de mutation différent sur les différentes branches,
- à partir des données initiales, calcule une matrice qui donne un arbre en étoile basé sur la divergence des taxons,
- la topologie de l'arbre est obtenu à partir de cette nouvelle matrice de distances,
- les longueurs des branches sont corrigées avec la divergence

Etapes du Neighbor-Joining

1. calcul des divergences
   $r_i = \sum_k d_{ik}$
2. calcul de la nouvelle matrice (représentation en étoile)
   $M_{ij} = d_{ij} - \frac{r_i + r_j}{n-2}$
3. choix des voisins les plus proches et calcul des branches de l'arbre
   $l_{iu} = \frac{d_{ij}}{2} + \frac{r_i - r_j}{2n-4}, \qquad l_{ju} = d_{ij} - l_{iu}$
4. calcul des nouvelles distances
   $d_{ku} = \frac{d_{ik} + d_{jk} - d_{ij}}{2}$
5. recommencer à partir de cette nouvelle matrice

Neighbor-Joining

Exemple

d_{ij}	A	B	C	D	E
B	5				
C	4	7			
D	7	10	7		
E	6	9	6	5	
F	8	11	8	9	8

1. $r_i = \sum_k d_{ik}$

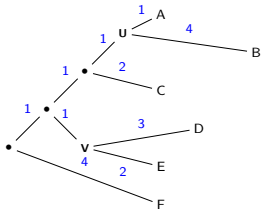
i	A	B	C	D	E	F
r_i	30	42	32	38	34	44

2. $M_{ij} = d_{ij} - \frac{r_i + r_j}{n-2}$

M_{ij}	A	B	C	D	E	F
B	-13					
C	-11.5	-11.5				
D	-10	-10	-10.5			
E	-10	-10	-10.5	-13		
F	-10.5	-10	-11	-11.5	-11.5	

3. $l_{AU} = \frac{d_{AB}}{2} + \frac{r_A - r_B}{2(n-2)} = \frac{5}{2} + \frac{30-42}{2(6-2)} = 1$

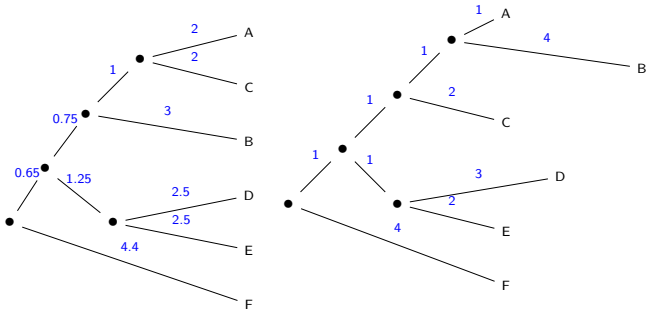
$l_{BU} = d_{AB} - l_{AU} = 5 - 1 = 4$



Résultat de `neighbor`

```
   6 Populations

Neighbor-Joining/UPGMA method version 3.6a2.1

 Neighbor-joining method

 Negative branch lengths allowed

Name                 Distances                                        | Between     And          Length
----                 ---------                                        | -------     ---          ------
sequence A    0.00000   5.00000   4.00000   7.00000   6.00000   8.00000 |       1     sequence B    4.00000
sequence B    5.00000   0.00000   7.00000  10.00000   9.00000  11.00000 |       1       2           1.00000
sequence C    4.00000   7.00000   0.00000   7.00000   6.00000   8.00000 |       2     sequence C    2.00000
sequence D    7.00000  10.00000   7.00000   0.00000   5.00000   9.00000 |       2       4           1.00000
sequence E    6.00000   9.00000   6.00000   5.00000   0.00000   8.00000 |       4       3           1.00000
sequence F    8.00000  11.00000   8.00000   9.00000   8.00000   0.00000 |       3     sequence D    3.00000
                                                                        |       3     sequence E    2.00000
                                                                        |       4     sequence F    5.00000
                                                                        |       1     sequence A    1.00000

  +-----------------------------sequence B
  !
  !       +--------------sequence C
  !       !
  1-------2              +----------------------sequence D
  !       !       +------3
  !       +-------4      +--------------sequence E
  !               !
  !               +---------------------------------------sequence F
  !
  +-------sequence A
```


UPGMA vs. NJ



Et si la distance n'est ni additive ni ultramétrique ?

- l'application de NJ donne un arbre avec des erreurs (des branches de longueur négative par exemple)
- l'application de UPGMA donne un arbre dont la topologie est fausse mais avec des longueurs de branches positives
- la matrice de distance déduite de ces arbres d^T est différente de la matrice de distance initiale d : estimation de l'écart entre la matrice et l'arbre

$$\text{ecart}(d, d^T) = \sum_{1 \le i < j \le N} (d_{ij} - d_{ij}^T)^2$$

Méthode des moindres carrés

- recherche un arbre (topologie et longueur des branches) qui estime le mieux la matrice de distances observées
- $$\text{SCE}(A) = \sum_{i=1}^{n} \sum_{j \neq i} \frac{(d_{ij} - e_{ij})^2}{d_{ij}^2}$$

  avec e_{ij} la distance estimée, *i.e.* la distance effective des longueurs de branches de l'arbre A

Méthode du maximum de vraisemblance

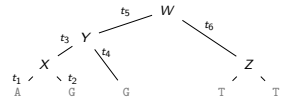
Maximum Likelihood

estime, sous un modèle évolutif donné, la probabilité d'un arbre (sa topologie et la longueur des branches)

- vraisemblance du site j : somme des probabilités de toutes les possibilités de reconstruction de l'état ancestral, sous le modèle choisi,
- vraisemblance de l'arbre : somme des log des vraisemblances pour chaque site

Vraisemblance d'une topologie donnée pour un site

- on se donne un modèle : $P_{ij}(t)$ donnant la probabilité de transition de i à j le long de la branche t
- on se donne un arbre et les valeurs des feuilles pour le site n :



- la vraisemblance que l'ancêtre de A et G soit X est donné par :

$$L_n(X) = P_{X\mathtt{A}}(t_1)P_{X\mathtt{G}}(t_2)$$

- la vraisemblance que l'ancêtre de X et G soit Y est donné par :

$$L_n(Y) = P_{Y\mathtt{G}}(t_4)\sum_X L_n(X)P_{YX}(t3)$$

- et ainsi :

$$L_n(W) = \sum_Y \sum_Z L_n(Y)P_{WY}(t_5)L_n(Z)P_{WZ}(t_6)$$

Vraisemblance d'une topologie donnée pour l'arbre

- la vraisemblance de l'arbre pour le site n est :

$$L_n = \sum_W \pi_W L(W)$$

avec π_W la fréquence ancestrale de la base W

- à partir de la vraisemblance calculée pour chaque site :

$$\ln L_{\text{arbre}} = \sum_n \ln L_n$$

- produit une valeur souvent largement négative
- donc une vraisemblance bien inférieure à 1

Méthode des quartets

Tree-Puzzling

- pour tout les quadruplets d'espèces, évalue l'arbre le plus vraisemblable (assez rapide)
- effectue un assemblage des quartets ainsi obtenus
- permet de traiter par le ML des problèmes plus grands

Les problèmes de la reconstruction

- quelle confiance avoir en un arbre ?
- quels sont les noeuds dont je suis sûr ?
- l'ordre des espèces a-t-il une importance ?
- et si je supprime une espèce ?
- pourquoi ai-je un peigne ?
- et ces deux branches, elles sont bien longues !

Bootstrap

- basée sur l'hypothèse que les caractères évoluent de manière indépendante
- protocole en 3 étapes (réalisé au moins 100 fois) :
  1. réalisation d'un pseudo alignement à partir de l'alignement initial en prenant arbitrairement n colonnes (avec remise),
  2. estimation de l'arbre obtenu T',
  3. comparaison des arbres T et T' (décompte du nombre de fois où un noeud est trouvé commun entre ces deux arbres)
- on assigne ensuite aux noeuds de l'arbre T les fréquences obtenues

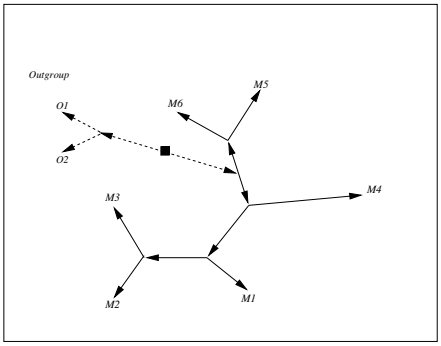
Enracinement

principe

- ajout d'un groupe externe (outgroup)
  - la racine se situe sur la branche menant à l'outgroup,
  - si l'on utilise une seule espèce, la monophylie du groupe étudié est considérée comme acquise
  - l'outgroup est une séquence ancienne,
  - ni trop éloignée des autres données (biais de reconstruction, attraction des longues branches)
  - ni trop proche (certitude quant au fait que ce soit un outgroup ?)
  - l'ajout d'un groupe plutôt que d'une seule espèce renforce la confiance en la topologie

Enracinement

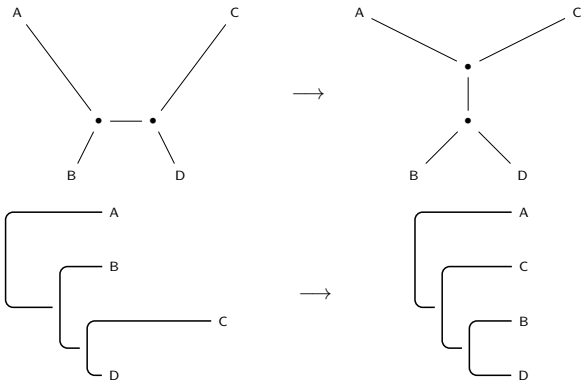
exemple



Mêmes espèces mais pas même phylogénie

- tous les gènes n'ont pas la même vitesse d'évolution,
- les phènomènes de recombinaison,
- les transferts de gènes,
- la confusion entre gènes paralogues et orthologues

Attention aux longues branches



Récapitulatif

Méthodes	Espèces	Avantages	Inconvénients
Distances	Très proches	Rapides, faciles à mettre en œuvre	Tous les sites sont traités de manière équivalente, pas applicables à des séquences éloignées
Parcimonie	Relativement éloignées	Evaluation de plusieurs arbres, informations sur des séquences ancestrales	Lente, très vite limitée en terme de nombre d'espèces
ML	Eloignées	Robuste, modélisation des différentes substitutions, estime la longueur des branches	Lente, très vite limitée en terme de nombre d'espèces