Le 21 Mars 2003-03-22

Démonstration 7-8

La Phylogénie

1. L'horloge Moléculaire
2. Méthodes de construction
3. Récapitulatif

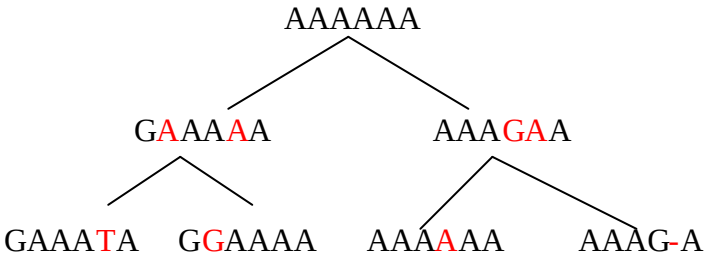
1. L'horloge moléculaire

La phylogénie consiste à comparer des gènes ayant un taux de mutation faible. La similarité des mécanismes moléculaires, suggère l'existence d'un ancêtre commun à tous les organismes. Ceci peut être représenté par un arbre phylogénétique dont la longueur des branches peut être vue comme une mesure de l'horloge moléculaire.
Pour dater la divergence entre les gènes, on peut la corréler avec une montre pour calibrer l'horloge.

- Trotteuse des secondes : taux de mutation élevé.
- Aiguille des minutes : taux de mutation moyen
- Aiguille des heures : taux de mutation faible.

Pour les protéines ayant la même fonction, la vitesse d'évolution est du même ordre de grandeur, ce qui n'est pas le cas pour les protéines de fonctions différentes.

Au cours du temps, les séquences accumulent des mutations : substitutions, délétions et insertions.



Malgré la remise en question de cette horloge moléculaire, il semblerait qu'elle fonctionne assez bien sur des longues périodes évolutives, pour des gènes ayant un taux de mutation faible.

2. Méthodes de reconstruction

Deux types de méthodes :

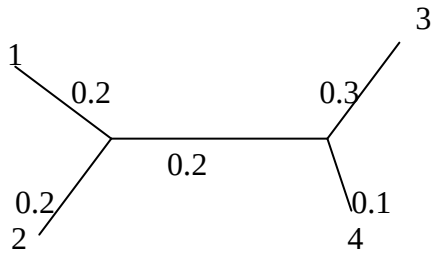
- ? Celles basées sur les distances entres séquences prises deux à deux.
- ? Celles basées sur les caractères (nombre de mutations).

Méthodes bases sur les distances :

Reconstruction d'arbre phylogénétique sans racine basée sur la recherche d'OTU (operationnal taxonomic units ? séquence) les plus proches. Ces méthodes sont rapides et donnent de bons résultats pour des séquences fortement similaires. Elles sont très utilisées.

- ? Calculer une matrice de distances entre séquences.
- ? Trouver l'arbre (topologie + longueurs de branches) qui respecte au mieux cette matrice.

Un arbre valué définit aussi une matrice de distance



d_{12} = 0.2+0.2=0.4
d_{13} = 0.2+0.2+0.3=0.7
……

UPGMA (Unweight Pair Group Method with Arithmetic mean)

Condition: si les séquences ne sont pas trop divergentes.

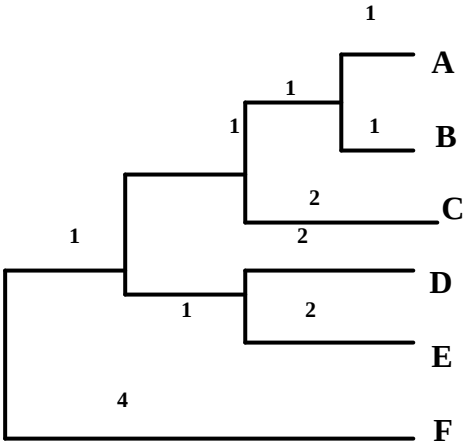
- ? L'algorithme de clusteristation séquentiel.
- ? Ordre de similarité.
- ? Reconstruction de l'arbre pas à pas grâce à cet ordre
- ? Identification de deux séquences très proches.
- ? Utilisation de ce groupe comme un tout.
- ? Condition d'arrêt : plus que deux groupes restants.

Exemple :

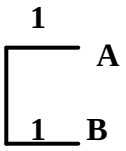
Soit la matrice suivante des distances entre 6 séquences .

	A	B	C	D	E	F
A						
B	2					
C	4	4				
D	6	6	6			
E	6	6	6	4		
F	8	8	8	8	8	

L'arbre associé qu'on aimerait obtenir



Les deux premières séquences avec la distance la plus faible sont A et B, le point de branchement est à une distance 2/2 = 1.



A partir de ce premier groupe on recalcule les distances :

D(A,B),C = (d(AC)+d(BC)) /2 = 4
D(A,B),D = (d(AD)+d(BD)) /2 = 6
D(A,B),E = (d(AE)+d(BE)) /2 = 6
D(A,B),F = (d(AF)+d(BF)) /2 = 8

	A,B	C	D	E	F
C	4				
D	6	6			
E	6	6	4		
F	8	8	8	8	

Regroupement de D et E

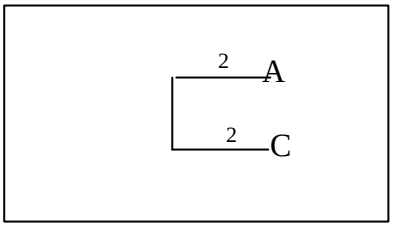
	A,B	C	D,E
C	4		
D,E	6	6	
F	8	8	8

Regroupement de (AB,C) et D,E

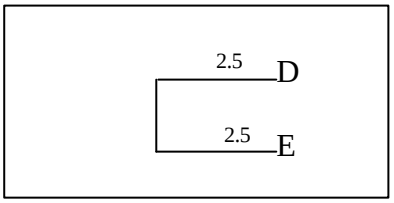
	ABC, DE
F	8

UPGMA donne un arbre non raciné, on applique la méthode du mid-point : la racine est équidistante à tous les OTUs soit (ABCDE),F/2 = 4.
Remarque : Quand la divergence entre est trop important, cette méthode ne s'applique très bien.
Exemple :

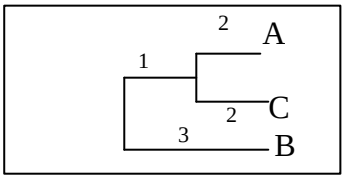
	A	B	C	D	E
B	5				
C	4	7			
D	7	10	7		
E	6	9	6	5	
F	8	11	8	8	9



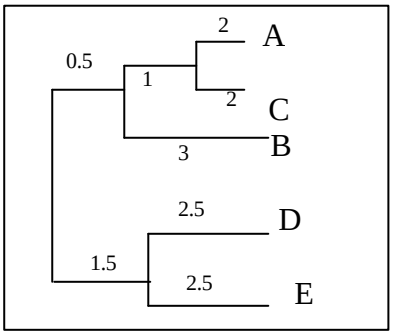
	A,C	B	D	E
B	4			
D	7	10		
E	6	9	5	
F	8	11	8	9



	A,C	B	D,E
B	6		
D,E	6.5	9.5	
F	8	11	8.5

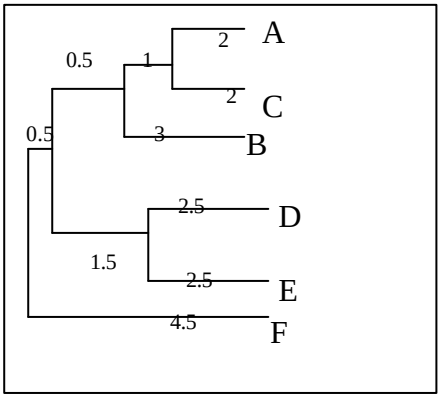


	AC,B	D,E
D,E	8	
F	9.5	9.5



	ABC,DE
F	9

Topologie Fausse



Il existe différentes façons de définir les distances, exemple une distance d qui donne la fraction f de sites ou x_u^i et x_u^j différent. Exemple du modèle de Jukes-Cantor pour l'ADN.
Soient les séquences suivantes :

ACCTTGATTGTTT
ACTTTAGTTGTTG

Différences =4, longueur = 13 dissimilarité = 4/13 = 0.307,
dissimilarité corrigée = 0.395 = -¾(log(1-4*f/3) avec f = 0.307.

rq : f <3/4 (sinon trop de différences par rapport à toute la séquence)

Cette correction vient de l'observation suivante de Jukes et Cantor en 1969 :
Si le temps de divergence entre deux séquences augmente, la probabilité d'avoir une seconde mutation a un site augmente également. Ceci implique que compter les différences entre les séquences ne reflète pas la réalité mais sou estime le nombre réel de substitutions.
L'hypothèse est que tous les sites sont équivalents (tous les changements ont une probabilité égale mais elle varie au cours du temps), qu'il n'y a pas de biais dans la direction du changement et qu'il n'y a eu ni insertions ni délétions. Hypothèse très simple mais pas forcément la plus correcte.

D'ou la matrice de taux de substitution uniformes

	A	T	G	C
A	1-f	f/3	f/3	f/3
T	f/3	1-f	f/3	f/3
G	f/3	f/3	1-f	f/3
C	f/3	f/3	f/3	1-f

Deuxième type de correction :

On fait la différence entre les substitutions en fonction des bases touchées par la mutation, autrement dit :

- Purine en Purine (transition) (A ____ G)
- Pyrimidine en Pyrimidine (transition) (C ——— T)
- Purine en Pyrimidine (transversions)(A_____C) ou (A_____T) ou (G_____ C) ou (G——— T)

Ceci correspond à la correction de Kimura ou 2 paramètres (1980) : modèle similaire à celui de Jukes-Cantor, mais avec une autre hypothèse : le taux de transition est différent de celui des transversions. Cette est venu suite à l'observation des transitions qui étaient plus fréquentes que les transversions.

Soient P la fréquence des transitions et Q celle des transversions.

$D_{K2p} = - (1/2) \ln(1-2P-Q) - 1/4 \ln(1-2Q)$

Soit la matrice de substitution Kimura à 2 paramètres

	A	T	G	C
A	1-P-Q	Q/2	p	Q/2
T	Q2	1-P-Q	Q/2	P
G	P	Q2	1-P-Q	Q/2
C	Q2	P	Q/2	1-P-Q

Le nombre total de substitutions estimé à travers une branche de longueur t est (2? +?)t égal à D_{K2p}

Construction d'un arbre avec Neighbor-Joining

Comme il a été montré au dessus, l'inconvénient majeur de UPGMA est sa sensibilité à des taux de mutations différents sur les différentes branches. La méthode NJ développée par Saitou et Nei (1987) tente de corriger l'inconvénient pour autoriser un taux de mutation différent sur les branches.

- ? On part d'un arbre étoile.
- ? On détermine la meilleure paire (i,j).
- ? On regroupe cette paire et on crée un nouveau nœud u.
- ? On calcule des distances dans l'arbre entre u et i,j.
- ? On calcule les distances dans la matrice entre u et les autres taxons.
- ? On oublie i et j et on recommence à l'étape 1.
- ? On termine quand il ne reste plus que 3 branches.

Exemple :

	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

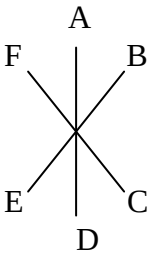
Etape 1: Calcul de la divergence de chacun des OTUs par rapport aux autres

R(A) = 5+4+7+6+8 = 30
R(B) = 42
R(C) = 32
R(D) = 38
R(E) = 34
R(F) = 44

Etape 2: calcul de la nouvelle matrice selon la formule : M(I,j) = d(i,j) –[r(i)+r(j)]/ (N-2)
Exple pour M(AB) = 5-[30+42]/4 = -13

	A	B	C	D	E
B	-13				
C	-11.5	-11.5			
D	-10	-10	-10.5		
E	-10	-10	-10.5	-13	
F	-10.5	-10.5	-11	-11.5	-11.5

L'arbre en étoile:



Etape 3: Choix des plus proches voisins, donc les OUT avec le M(i,j) le plus petit : A avec B et D avec E. On forme un nouveau nœud U avec A et B et on recalcule la longueur de la branche entre U et A ainsi que B. selon la formule suivante (acétate 26)

calcul de longueurs de branches : prenons la moyenne

$$L_{1X} = \frac{1}{n-2} \sum_{i=3}^{n} (d_{1i} + d_{12} - d_{2i})/2.$$

L_{2X} même chose

$$L_{iX} = (d_{1i} + d_{2i} - d_{12})/2.$$

avec 1 = A et 2 = B et X = U
L(AU) =1/4[(4+5-7)/2+(7+5-10)/2+6+5-9)/2+8+5-11)/2] = 8/8 = 1
L(BU) = d(AB) –L(AU) = 5-1 = 4

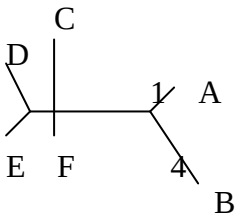
Etape 4: Les nouvelles distances entre U et les autres Out

d(CU) = d(AC) + d(BC) – D(AB)/2 = 3
d(DU) = d(AD)+d(BD)-d(AB)/2 =6
d(EU) = d(AE) + d(BE) –d(AB)/2 =5
d(DU) = d(AF) + d(BF) –d(AB)/2 = 7

Création d'une nouvelle matrice

	U	C	D	E
C	3			
D	6	7		
E	5	6	5	
F	7	8	9	8

Et d'un arbre en étoile



La procédure complète reprend à l'étape 1 avec N =N-1 = 5.

Méthodes basées sur les caractères

Parcimonie : consiste à minimiser le nombre de mutations/substitutions pour passer d'une séquence à une autre dans la topologie de l'arbre.

Ses hypothèses :

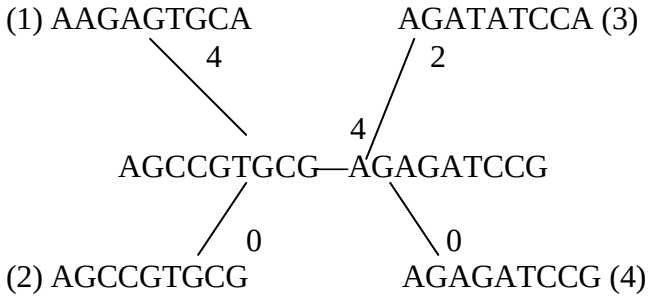
Les sites évoluent indépendamment les uns des autres.
Vitesse d'évolution lente et constante au cours du temps.

La méthode de maximum de parcimonie recherche toutes les topologies possibles afin de trouver l'arbre optimal (minimum) et le temps nécessaire pour cette exploration croît rapidement avec le nombre de séquences :
Avec n OTU nombre d'arbres enracinés possibles = (2n-3) ! / (2exp(n-2))(n-2) !
Et nombre non enracinés possibles = (2n-5) !/(2exp(n-3))(n-3) !

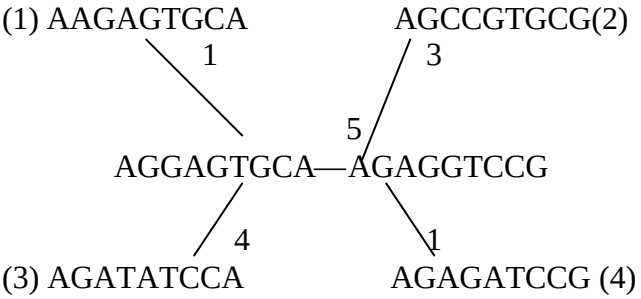
Exemple avec 4 séquences

	1	2	3	4	5	6	7	8	9
1	A	A	G	A	G	T	G	C	A
2	A	G	C	C	G	T	G	C	A
3	A	G	A	T	A	T	C	C	A
4	A	G	A	G	A	T	C	C	G

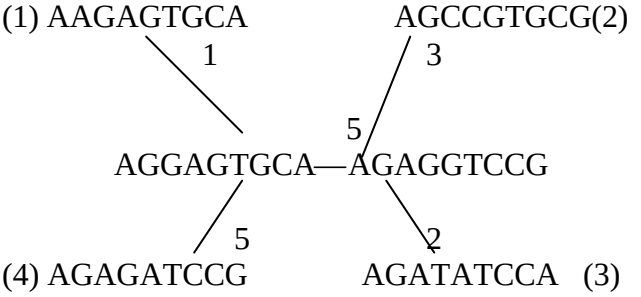
Il y a 3 arbres non enracinés possibles



Nombre de mutations = 10



Nombre de mutations = 14

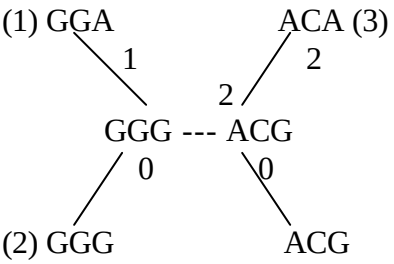


Nombre de mutations = 16

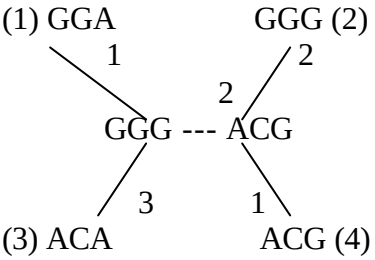
L'arbre le plus parcimonieux est le premier, avec le moins de mutations.

L'analyse peut se faire uniquement sur le sites informatifs, correspondant aux positions avec au moins deux nucléotides différents et chacun doit se retrouver au moins dans deux séquences. Donc les positions 5 , 7 et 9.

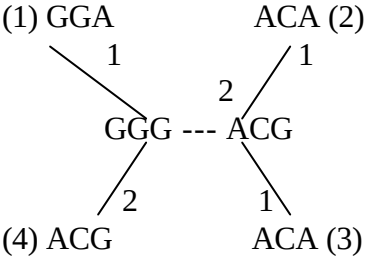
1	GGA
2	GGG
3	ACG
4	ACG



Nombre de mutations = 4



Nombre de mutations = 8



Nombre de mutations = 7

Le site 5 favorise l'arbre avec 4 mutations. L'arbre le plus parcimonieux est celui qui est supporté par le plus grand nombre de sites informatifs.

Pour garantir de trouver le meilleur arbre possible, il faut évaluer toute s les topologies, impossible avec plus de 12 séquences.

Le nombre de topologies croît avec le nombre des séquences, il est possible de rechercher l'arbre d'une façon plus efficace sans passer par l'énumération de tus les arbres.

Avec la parcimonie, il y a une alternative pour trouver le meilleur arbre, elle exploite le fait que le nombre de substitutions dans un arbre ne peut croître qu'en rajoutant des branches.
L'idée derrière la méthode «**Branch and Bound**» est de construire l'arbre en incrémentant le nombre de feuilles, mais on abandonne la direction de construction des que l'arbre courant incomplet a un coût supérieur à celui obtenu par un arbre complet construit précédemment

Une méthode de recherche heuristique réarrange les branches à chaque étape, cette méthode ne garantit pas de trouver l'arbre optimal.

La méthode de l'arbre consensus, comme la méthode du maximum de parcimonie peut conduire à trouver plusieurs arbres équivalents, on peut créer un arbre consensus (avec utilisation du bootstraping). Cet arbre consensus est construit à partir des nœuds les plus fréquemment rencontrés sur l'ensemble des arbres possibles.

Maximum de vraisemblance :

Reconstruction de l'arbre en terme de probabilités, l'ordre des branchements et de la longueur des branches d'un arbre sous un modèle évolutif donné.
Le meilleur candidat est celui qui maximise la probabilité de vraisemblance. La stratégie est de chercher à travers tous les arbres, et pour chaque topologie T de trouver les longueur t, qui maximisent la vraisemblance $P(x_u|T, t)$

Exemple :

1 **j**
CGAGA**C**
AGCGA**C**
AGATT**A**
GGATA**G**

La vraisemblance au site j est la somme des probabilités de toutes les possibilités de reconstruction de l'état ancestral sous le modèle choisi.
La vraisemblance de l'arbre A est en général évaluée en sommant les logs des vraisemblances pour chaque site (la somme des probabilités étant trop faible.
L'arbre du maximum de vraisemblance est celui avec la vraisemblance la plus élevée.

Les modèles évolutifs :
Les probabilités à chaque site dépendent du modèle choisi et dans le modèle le plus simple il est supposé :

La probabilité de chaque changement est indépendante des précédents (Modèle de Markov)
Les probabilités des substitutions ne changent pas au cours du temps (le long de l'arbre).
Les changements sont réversibles P(a/u) = P(u/a).

On peut inclure d'autres paramètres dans le modèle pour accroître son réalisme.

Des taux de substitutions différents pour chaque remplacement.
Une correction pour le nombre de sites susceptibles de muter et des taux de substitutions variables pour ces sites.
Un taux de variation différents pour chaque site : on peut par exemple utiliser une distribution statistique (distribution gamma).

Méthodes d'évaluation :

Il faut évaluer la confiance de cet arbre.
La méthode la plus utilisée est celle de bootstrap.Elle part du postulat que les caractères évoluent de manière indépendante.
Etapes de la méthode :

1. Réalisation de pseudo alignement A' à partir des séquences d'origine en prenant arbitrairement n colonnes de l'alignement original.
2. Estimation de l'arbre obtenu T'

3. Comparaison des arbres T et T', pour chaque sous arbre de T, on regarde s'il est présent dans T'.
4. On compte ensuite pour chaque sous arbre le nombre de fois où il est présent dans les T'.
5. Cette fréquence avec laquelle on retrouve un sous arbre est la valeur de bootstrap (plus elle est élevée, plus la fiabilité de la branche est importante) .

3. Récapitulatif

Méthodes	Séquences	Avantages	Inconvénients	Programmes
Distances	Trés proches	Rapides Faciles à mettre en œuvre	Tous les sites sont de manière équivalente d'où une perte d'informations. Non applicable à des séquences éloignées	DNAdist Protdist FITCH KITSCH
Parcimonie	Relativement éloignées	Évaluation de différents arbres. Essaye de donner les infos sur les séquences ancestrales	Lente Inutilisable lorsque l'on a un grand nombre de séquences	DNApars PROTpars
ML	Éloignées	Robuste Taux de transition/transversions différents. Estimation de la longueur des branches de l'arbre final	Lente Inutilisable avec un grand nombre de séquences.	FastDnaml