Phylogénies

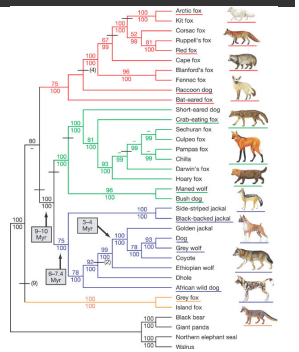
Marc Girondot

université PARIS-SACLAY | FACULTÉ DES SCIENCES D'ORSAY

1

Qu'est-ce qu'une phylogénie ?

Une phylogénie est une représentation de l'évolution des espèces incluant les phénomènes de cladogénèse et d'anagénèse (et autres phénomènes apparentés)



2

Caractères et états de caractères

- Caractères morphologiques
  - Les caractères sont des structures morphologiques et les états de caractère sont les modalités que prennent les structures
    - On peut rarement dénombrer plus d'une quarantaine de caractères qui ont chacun entre 2 et au maximum une dizaine d'états
- Caractères et états de caractère moléculaires
  - Une position sur une séquence est un caractère, la présence d'un A, T, G ou C ou d'un acide aminé est un état de caractère
    - Si on travaille en AN, pour une séquence de longueur *l*, il y a *l* caractères à 4 états
    - Si on travaille en AA, la même séquence a une longueur *l*/3 mais 20 états pour chaque caractère.

3

Caractères morphologiques

- Avantage
  - Peu cher à observer
  - Facile d'accès
  - Homoplasie peu fréquente
- Désavantages
  - On ne voit facilement que les différences
  - Non disponible entre espèces trop éloignées
  - Nombre de caractères disponibles très limités (souvent moins de 40)
  - Nécessite une grande expertise

4

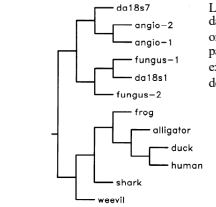
Caractères moléculaires

- Avantage
  - Grande quantité d'informations (>1000 bases, maintenant peut se compter en milliards)
  - Comparaison sur des gènes même entre espèces très éloignées
- Désavantage
  - Peu d'état pour chaque caractère d'où réversion et convergence fréquente (homoplasie)
  - Coût relativement élevé mais les prix baissent rapidement (4 000 € pour le séquençage d'un génome)

5

Moléculaire ou morphologie ?

- Si il s'agit de fossiles, la question peut ne pas se poser et on est obligé d'utiliser des caractères morphologiques.
- A moins qu'il s'agisse de sub-fossiles chez lesquels on peut extraire de l'ADN tout en gardant conscience que des faux positifs peuvent exister:
  - Wang HL, Yan ZY, Jin DY (1997) Reanalysis of published DNA sequence amplified from cretaceous dinosaur egg fossil. Molecular Biology and Evolution 14: 589-591



Les séquences da18s7 et da18s1 ont été clonées à partir d'ADN extrait d'un œuf de dinosaure.

FIG. 2.—Phylogenetic tree of partial 18S rDNAs from different species. The tree is based on a matrix of pairwise evolutionary distances expressed as substitutions per 100 bases and corrected by the Jin-Nei gamma distance method (Jin and Nei 1990) for multiple substitutions at a site. The tree reconstruction was performed using a neighbor-joining algorithm. A ciliate rDNA sequence was used for rooting. See legend to figure 1 for species names and GenBank accession numbers of the sequences in the tree; angio-1 = angiosperm-1; angio-2 = angiosperm-2.

6

Données moléculaires

- Les sequences en acides aminées sont plus courtes que les sequences en acides nucléiques mais l'homoplasie est moindre.

Le code génétique

Premier nucléotide	Deuxième nucléotide U		Deuxième nucléotide C		Deuxième nucléotide A		Deuxième nucléotide G		Troisième nucléotide
U	UUU	phényl-alanine	UCU	sérine	UAU	tyrosine	UGU	cystéine	U
	UUC		UCC		UAC		UGC		C
	UUA	leucine	UCA		UAA	STOP	UGA	STOP	A
	UUG		UCG		UAG		UGG	tryptophane	G
C	CUU	leucine	CCU	proline	CAU	histidine	CGU	arginine	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	glutamine	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	isoleucine	ACU	thréonine	AAU	asparagine	AGU	sérine	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	lysine	AGA	arginine	A
	AUG	méthionine	ACG		AAG		AGG		G
G	GUU	valine	GCU	alanine	GAU	acide aspartique	GGU	glycine	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	acide glutamique	GGA		A
	GUG		GCG		GAG		GGG		G

7

Données moléculaires

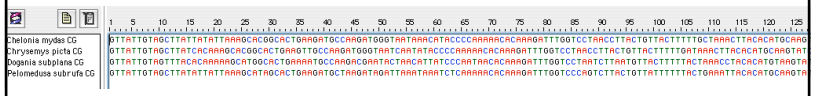
- Les données nucléiques sont maintenant devenues la norme pour étudier la phylogénie.
- En effet, les modèles actuels permettent de prendre en compte de façon différente les positions de codon 1, 2 et 3. Ainsi l'homoplasie qui est plus importante sur les séquences nucléiques, peut être contrôlée par des modèles d'évolution des séquences nucléiques.

8

Les séquences

Obtention des séquences par les méthodes classiques de la biologie moléculaire:

- Construction d'amorces en fonction des espèces
- PCR
- Séquençage

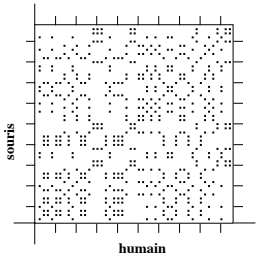


9

Etape 1: l'alignement

```
                1         2         3         4         4
      1         0         0         0         0         8
Humain:   ACCACCTCATCCTGGGCACCCTGGTTATATCAACTTCAGCTATGAGGT
Souris:   ACCACCTCATCCTGGAAGCCCTGGTTATATCAACTTAAGCTATGAGGT
                  ***                       *
```

Le dot-plot

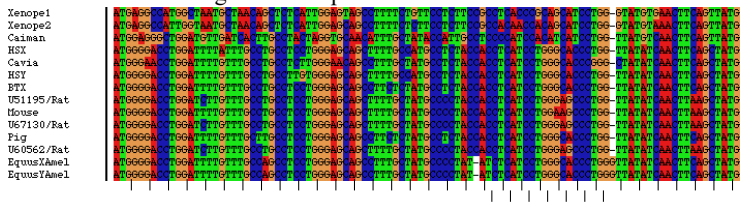


10

Alignement… suite

- L'alignement peut-être sans difficulté, par exemple c'est la cas avec la plupart des séquences codantes

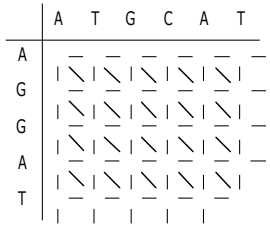
Mais…changement temporaire du cadre de lecture



11

L'algorithme de Needleman et Wunch (1970)

ATGCAT
AGGAT



Coût: Mauvais appariement=1
Indel= 2

Construction d'une matrice avec toutes les combinaisons possibles

A une case, arrivent 3 liens

Le diagonal: les deux bases sont mises en alignement

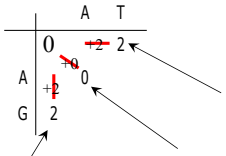
Le vertical: On ne change pas la base de la séquence horizontale mais on passe à la base suivante de la séquence verticale; ceci implique donc une insertion sur la séquence horizontale

L'horizontal: ceci implique donc une insertion sur la séquence verticale

12

Remplissage du tableau

- Le coût associé à une case est la somme de la case précédente et de l'événement que l'on envisage



T aligné avec A
Un indel; coût=2

G aligné avec A
Un indel; coût=2

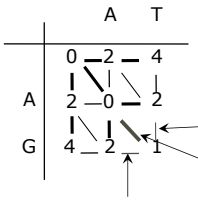
Signifie que le A est aligné avec le A
Une base alignée; coût=0

13

Remplissage du tableau

- On recherche le coût minimum pour une case en faisant la somme du coût de la case précédente et de l'événement que l'on envisage

Les lignes en gras désignent les chemins suivis.



AT-
A-G
Coût=2+2=4
Solution non retenue

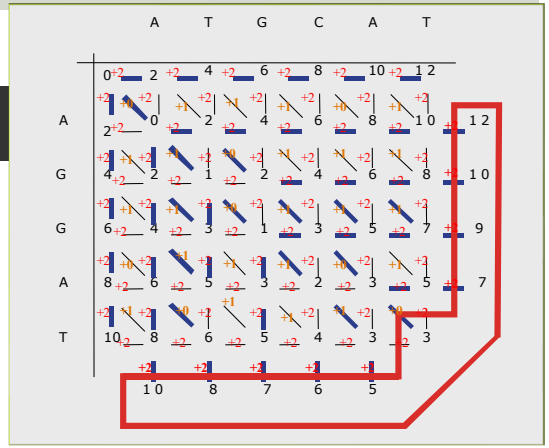
AT
AG
Coût=0+1=1
Solution retenue (trait bleu)

Coût=2+2=4
Solution non retenues
A-T
AG-

14

Résultat

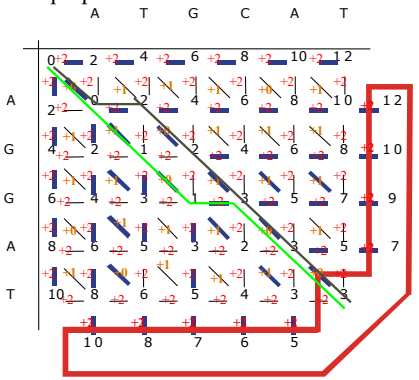
Les lignes bleues en gras sont les chemins choisis.



On recherche la valeur minimale sur les différentes cases en rouge qui représentent les alignements possibles.
Ici, c'est trois.

15

On remonte alors à partir du 3 pour voir les différents chemins bleus possibles.
Il y a seulement deux chemins possibles donc deux alignements qui présentent le même coût.



ATGCAT
A-GGAT

ATGCAT
AGG-AT

16

L'alignement…

- L'alignement des séquences peut être une étape cruciale dont dépend toute la suite de l'analyse.
- Or cette étape est déjà dépendante de la paramétrisation des coûts des changements.

17

Méthodes de reconstruction

- Méthodes phénétiques basées sur les distances
  - UPGMA: Unweighted pair group method with arithmetic mean
  - Fitch - Margoliash
  - Neighbor-joining
  - Bio-Neighbor-joining
- Méthodes basées sur les changements d'états de caractère
  - Méthode basée sur la parcimonie
  - Méthode basée sur la vraisemblance
  - Méthode bayésienne

18

Méthodes phénétiques

- Les méthodes phénétiques sont basées sur le calcul d'une matrice de distance et du calcul de la phylogénie qui minimise la distance entre les valeurs des distances et leur représentation sur la topologie.

19

Quelle distance utiliser ?

- Dans la matrice de distance, la valeur contenue dans chacune des cases mesure la distance entre les sequences correspondant au couple d'espèces.

20

Taux de mutation vs taux de substitution allélique

- Les taux de mutation se réfèrent à la vitesse à laquelle se produisent des mutations.
  - Le taux de mutation dans la lignée germinale humaine est de l'ordre de 0.5×10^{-9} par site par an. Il est communément noté μ.
  - Le taux de mutation est plus important chez l'homme que chez la femme et plus important lorsque les générations sont courtes que quand elles sont longues.
- Le taux de substitution se réfèrent à la vitesse à laquelle se produisent des substitutions d'allèles.
- Attention, une « substitution » est aussi un type particulier de mutation: les substitutions nucléotidiques, les insertions/délétions de quelques nucléotides et les remaniements géniques de grande taille. Ici on parle de substitution d'allèles=1 mutation fixée.

21

Calcul du taux de substitutions

Lorsque l'on veut calculer ce taux, il convient d'avoir au moins deux séquences alignées. Par exemple ici l'exon 3 de l'amélogénine (gène majeur impliqué dans la formation de l'émail) chez la souris et l'homme:

```
                  1         2         3         4         4
        1         0         0         0         0         8
Humain: ACCACCTCATCCTGGGCACCCTGGTTATATCAACTTCAGCTATGAGGT
Souris: ACCACCTCATCCTGGAAGCCCTGGTTATATCAACTTAAGCTATGAGGT
                     ***                     *
```

Sur les 48 positions analysées, on dénombre 4 différences (*). Pour calculer le taux de substitution, il convient de prendre en compte le fait qu'une même position peut muter plusieurs fois au cours de l'évolution et donc le taux observé est une sous-estimation du taux réel.
Pour calculer le taux réel, il faut définir un modèle d'évolution des nucléotides. On étudiera ici le modèle de Jukes et Cantor (1969), les modèles plus généraux seront cités simplement à la fin.

22

Modèles d'évolution des bases

1 JC69 model (Jukes and Cantor 1969)
2 K80 model (Kimura 1980)
3 K81 model (Kimura 1981)
4 F81 model (Felsenstein 1981)
5 HKY85 model (Hasegawa, Kishino and Yano 1985)
6 T92 model (Tamura 1992)
7 TN93 model (Tamura and Nei 1993)
8 GTR model (Tavaré 1986)

23

1. JC69

JC69, le modèle Jukes et Cantor 1969, est le modèle de substitution le plus simple. Il suppose des fréquences de base et des taux de mutation égaux. Le seul paramètre de ce modèle est donc μ, le taux de substitution global.

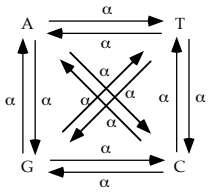
Le modèle JC69 suppose que toutes les bases sont également fréquentes ($\pi A = \pi G = \pi C = \pi T = 0{,}25$).

24

Modèle d'évolution des bases

Soit α la probabilité qu'une base X_1 mute en une autre base X_2 en un temps t (figure 2). Comme il y a 4 bases au total, la probabilité que X_1 change pendant le temps t est 3α et la probabilité qu'il ne change pas est simplement 1-3α.

Figure 2: Les 12 types de mutations possibles dans le modèle de Jukes et Cantor (1969) à un seul paramètre.



Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Munro HN (ed) Mammalian protein metabolism. Academic press, New York, pp 21-23

On peut donc écrire $P_{AA}(t=1)$ la probabilité qu'une base A présente à un site au temps t=0 soit toujours A au temps t=1 et $P_{AX}(t=1)$ la probabilité qu'une base A présente à un site au temps t=0 soit G, C ou T au temps t=1:

$P_{AA}(t=1)=1-3\alpha$

$P_{AX}(t=1)=3\alpha$

25

Suite…

La probabilité qu'au temps t=2 il y ait un A est calculée en prenant en compte qu'il est possible qu'au temps t=1 il y ait un A ($P_A(t=1)$) ou une autre base ($1-P_A(t=1)$).
S'il y a un A, il faut qu'il y reste, soit une probabilité de $(1-3\alpha)$ et s'il n'y a pas de A, il faut que la base en question le devienne soit une probabilité de α.

D'où:

$(P_A(t=2))=(1-3\alpha).(P_A(t=1))+\alpha.(1-P_A(t=1))$

Et en généralisant:

$(P_A(t+1))=(1-3\alpha).(P_A(t))+\alpha.(1-P_A(t))$

D'où

$\Delta P_A(t)=P_A(t+1)-P_A(t)=-3\alpha P_A(t)+\alpha.(1-P_A(t))=-4\alpha P_A(t)+\alpha$

En rendant continu le modèle, $\Delta P_A(t)$ est le taux de changement au temps t,

$$\frac{dP_A(t)}{dt} = -4\alpha P_A(t) + \alpha$$

26

Solution…

Ce qui est une équation différentielle du premier ordre avec comme solution:

$$P_A(t) = \frac{1}{4} + \left(P_A(0) - \frac{1}{4}\right) e^{-4\alpha t}$$

Si le nucléotide est A au temps 0, alors $P_A(t=0)=1$ d'où:

$$P_A(t) = \frac{1}{4} + \frac{3}{4} e^{-4\alpha t}$$

Si le nucléotide n'est pas A au temps 0, alors $P_A(t=0)=0$ d'où:

$$P_A(t) = \frac{1}{4} - \frac{1}{4} e^{-4\alpha t}$$

Comme dans le modèle de Juke et Cantor tous les nucléotides sont traités identiquement, on peut généraliser l'équation avec i et j des nucléotides différents:

$$P_{ii}(t) = \frac{1}{4} + \frac{3}{4} e^{-4\alpha t}$$

$$P_{ij}(t) = \frac{1}{4} - \frac{1}{4} e^{-4\alpha t}$$

27

Le taux de substitutions

A partir du modèle de Juke et Cantor, la proportion de nucléotides identiques entre deux séquences qui ont divergées depuis un temps t est:

$$I(t) = \frac{1}{4} + \frac{3}{4} e^{-8\alpha t}$$

Notez le 8 en exposant plutôt que le 4, car il y a un temps t pour aller de la séquence ancêtre à une des deux séquences et encore un temps t pour aller de la séquence ancêtre à l'autre séquence soit une divergence totale de 2t entre deux séquences qui ont divergé depuis un temps t.

La probabilité que deux séquences soient différentes à un site après ce même temps t est donc p=1-I(t):

$$p = \frac{3}{4}(1 - e^{-8\alpha t})$$

D'où

$$8\alpha t = -\ln(1 - \frac{4}{3} p)$$ Equation 5

28

Suite et fin…

La valeur t est inconnue, par contre on peut estimer K, le nombre réel de substitution par site depuis la divergence entre les deux séquences (qui prend donc en compte les évènements multiples).
Sous le modèle de Juke et Cantor, on a $K = 2.(3.\alpha.t)$ où $3.\alpha.t$ est le nombre moyen réel de substitutions par site sur une des lignées.

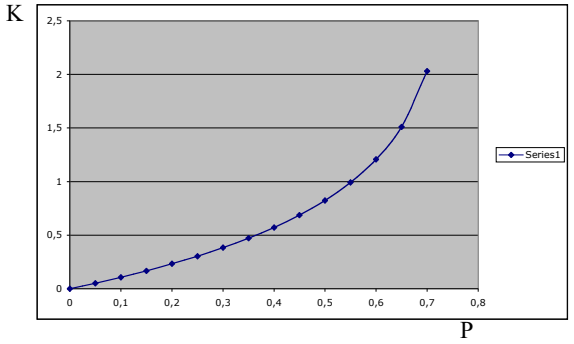
En remplacement $8.\alpha.t$ par son équivalent en fonction de K dans l'équation 5, on a:

$$K = -\frac{3}{4} ln\left(1 - \frac{4}{3}p\right)$$

Pour résumer, deux séquences qui ont une proportion p de bases différentes ont subi K substitutions par site. Par exemple, sur 48 bases, les deux séquences précédentes ont 4 différences, soit p=4/48=1/12=0,083 d'où K=0,088.
On peut montrer que $K>p$, ce qui est logique car dans l'estimation de p on perd les évènements multiples de substitutions sur une même base.

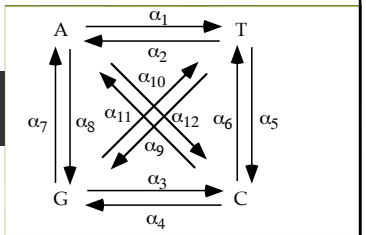
29

P versus *K*



30

Généralisation



Ce modèle ne prend en compte qu'un seul paramètre pour décrire les mutations, α, alors que l'on sait, par exemple, que les transitions (A<->G) et (C<->T) sont plus fréquentes que les transversions (C<->A), (C<->G), (T<->A), (T<->G). Par ailleurs, le C des dinucléotides CpG est souvent méthylé et a tendance à muter en T lors de la réplication. Chez les procaryotes, c'est le premier T des dinucléotides TpT.

31

2. K80

K80, le modèle de Kimura 1980, souvent appelé modèle à deux paramètres de Kimura (ou modèle K2P), distingue les transitions (purine vers purine[AG] ou pyrimidine vers pyrimidine[CT]) et les transversions (de la purine à la pyrimidine ou vice versa).

Le modèle K80 suppose que toutes les bases sont à égale fréquence ($\pi A = \pi G = \pi C = \pi T = 0{,}25$).

32

4. F81

F81, le modèle de Felsenstein de 1981, est une extension du modèle JC69 dans lequel les fréquences de base peuvent être différentes de 0,25 ($\pi A \neq \pi G \neq \pi C \neq \pi T \neq 0{,}25$).

33

5. HKY85 model

HKY85, le modèle Hasegawa, Kishino et Yano 1985, peut être considéré comme combinant les extensions des modèles Kimura80 et Felsenstein81. À savoir, il distingue le taux de transitions et de transversions (en utilisant le paramètre κ), et il permet des fréquences de base inégales ($\pi A \neq \pi G \neq \pi C \neq \pi T \neq 0{,}25$).

34

8. GTR86

Le GTR, le modèle généralisé réversible dans le temps de Tavaré 1986, est le modèle le plus général possible: neutre, indépendant, à sites finis et réversible dans le temps. Il a été décrit pour la première fois sous une forme générale par Simon Tavaré en 1986.
Les paramètres GTR consistent en un vecteur de fréquence de base d'équilibre $\prod = (\pi A, \pi G, \pi C, \pi T)$, donnant la fréquence de chaque base, et la matrice Q des taux

$$Q = \begin{pmatrix} -(\alpha\pi_G + \beta\pi_C + \gamma\pi_T) & \alpha\pi_G & \beta\pi_C & \gamma\pi_T \\ \alpha\pi_A & -(\alpha\pi_A + \delta\pi_C + \epsilon\pi_T) & \delta\pi_C & \epsilon\pi_T \\ \beta\pi_A & \delta\pi_G & -(\beta\pi_A + \delta\pi_G + \eta\pi_T) & \eta\pi_T \\ \gamma\pi_A & \epsilon\pi_G & \eta\pi_C & -(\gamma\pi_A + \epsilon\pi_G + \eta\pi_C) \end{pmatrix}$$

avec
$$\begin{aligned} \alpha &= r(A \to G) = r(G \to A) \\ \beta &= r(A \to C) = r(C \to A) \\ \gamma &= r(A \to T) = r(T \to A) \\ \delta &= r(G \to C) = r(C \to G) \\ \epsilon &= r(G \to T) = r(T \to G) \\ \eta &= r(C \to T) = r(T \to C) \end{aligned}$$

35

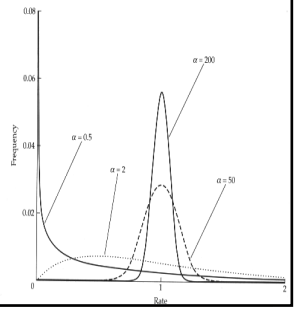
8. General time-reversible (GTR+I+Γ)

Substitution rates: A-C, A-G, A-T, C-G, C-T, G-T
(5 free parameters)

Base frequencies: π_A π_G π_C π_T
(3 free parameters)

Proportion of invariant sites: I
(1 free parameter)

Shape of the Γ distribution: α
(1 free parameter)



Le modèle GTR86+I+Γ utilise 10 paramètres.

36

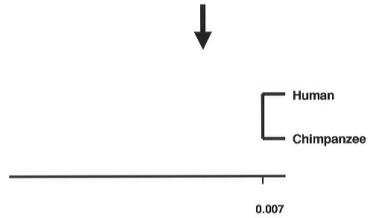
Reconstruction phylogénétique

- Méthodes phénétiques basées sur les distances
  - UPGMA: Unweighted pair group method with arithmetic mean
  - Fitch - Margoliash
  - Neighbor-joining
  - Bio-Neighbor-joining
- Méthodes basées sur les changements d'états de caractère
  - Méthode basée sur la parcimonie
  - Méthode basée sur la vraisemblance
  - Méthode bayésienne

37

UPGMA Example

	Human	Chimp.	Gorilla	Orangutan	Gibbon
Human	—	0.015	0.045	0.143	0.198
Chimpanzee	1	—	0.030	0.126	0.179
Gorilla	3	2	—	0.092	0.179
Orangutan	9	8	6	—	0.179
Gibbon	12	11	11	11	—



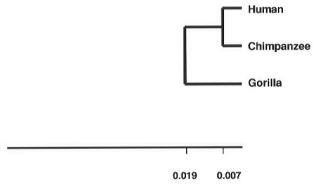
38

d(human-chimp) – gorilla= ½ [d(human-gorilla) + d(chimp – gorilla)]
= ½ [0.045 + 0.030]
= 0.037

d(human-chimp) – oran. = ½ [d(human-oran.) + d(chimp – oran.)]
= 0.135

d(human-chimp) – gibbon = ½ [d(human-gibbon) + d(chimp – gibbon)]
= 0.189

	Human-chimp	Gorilla	Orangutan	Gibbon
Human-chimp	—	0.037	0.135	0.189
Gorilla		—	0.092	0.179
Orangutan			—	0.179
Gibbon				—



39

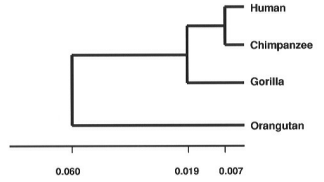
d(human-chimp-gorilla) – oran. = 1/3 [d(human-oran.) + d(chimp – oran.) + d(gorilla – oran.)]

= 0.121

d(human-chimp-gorilla) – gibbon = 1/3 [d(human-gibbon) + d(chimp – gibbon) + d(gorill – gibbon)]

= 0.185

	Human-chimp-gorilla	Orangutan	Gibbon
Human-chimp-gorilla	—	0.121	0.185
Orangutan		—	0.179
Gibbon			—

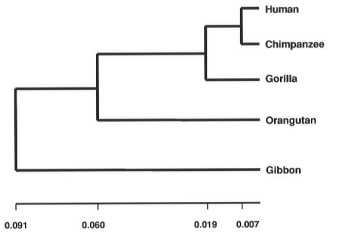


40

d(human-chimp-gorilla– oran.) - Gibbon = 1/4 [d(human-gibbon) + d(chimp – gibbon) + d(gorilla – gibbon) + d(oran. - gibbon]

= 0.183

	Human-chimp-gorilla-oran.	Gibbon
Human-chimp-gorilla-oran.	—	0.183
Gibbon		—



41

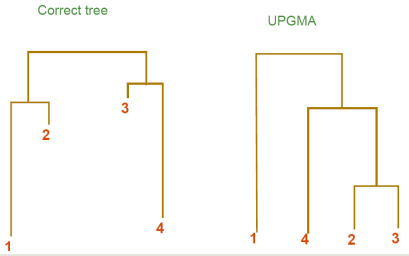
Problème lié à l'UPGMA

- L'algorithme produit un arbre ultramétrique: la distance de la racine à n'importe quelle feuille est la même
- L'UPGMA suppose une horloge moléculaire constante: toutes les espèces accumulent des mutations (évoluent) au même rythme.

42

Problème lié à l'UPGMA

UPGMA échoue à proposer un arbre correct lorsque l'hypothèse d'horloge moléculaire n'est pas validée.



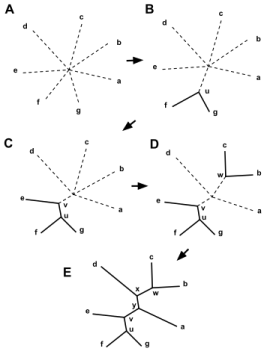
43

Neighbor-joining (NJ)

- Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 4: 406-425
- Studier JA, Keppler KJ (1988) A note on the neighbor-joining method of Saitou and Nei. Molecular Biology and Evolution 5: 729-731

44

Exemple de NJ



En commençant par un arbre en étoile (A), on choisit une paire de nœuds à joindre, dans ce cas f et g. Ceux-ci sont joints à un nœud nouvellement créé, u, comme indiqué dans (B). La partie de l'arborescence représentée en traits pleins est désormais fixe et ne sera pas modifiée lors des étapes de jonction suivantes. Les distances du nœud u aux nœuds a-e sont recalculées.
Ce processus est ensuite répété, en utilisant une matrice composée uniquement des distances entre les nœuds, a, b, c, d, e et u. Dans ce cas, u et e sont joints au v nouvellement créé, comme indiqué en (C). Deux autres itérations mènent d'abord à (D), puis à (E), moment auquel l'algorithme est terminé, car l'arbre est entièrement résolu.

45

Olivier Gascuel

Bio-NJ

- Nouvel algorithme qui améliore la précision du NJ en utilisant le fait que les distances plus courtes dans la matrice de distances sont généralement mieux connues que les distances plus longues.
- Gascuel O (1997) BIONJ: An improved version of the NJ algorithm based on a simple model of sequence data. Molecular Biology and Evolution 14: 685-695

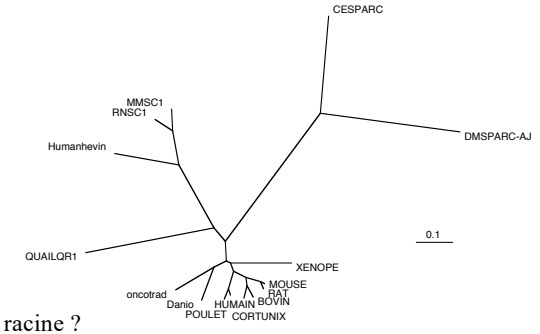
46

Exemple d'une matrice de taux de substitutions

	15														
MOUSE	0.000	0.120	0.111	0.036	0.171	0.185	0.293	0.257	0.267	0.607	0.607	0.623	0.615	0.929	0.954
BOVIN	0.120	0.000	0.075	0.136	0.174	0.188	0.300	0.282	0.274	0.592	0.600	0.604	0.614	0.910	0.952
HUMAIN	0.111	0.075	0.000	0.120	0.161	0.170	0.301	0.281	0.286	0.609	0.612	0.616	0.636	0.902	0.954
RAT	0.036	0.136	0.120	0.000	0.183	0.195	0.308	0.264	0.272	0.609	0.595	0.606	0.628	0.929	0.980
CORTUNIX	0.171	0.174	0.161	0.183	0.000	0.046	0.254	0.234	0.263	0.599	0.550	0.561	0.587	0.876	0.940
POULET	0.185	0.188	0.170	0.195	0.046	0.000	0.272	0.244	0.278	0.619	0.580	0.583	0.608	0.861	0.957
XENOPE	0.293	0.300	0.301	0.308	0.254	0.272	0.000	0.311	0.339	0.630	0.596	0.606	0.624	0.962	1.066
Danio	0.257	0.282	0.281	0.264	0.234	0.244	0.311	0.000	0.216	0.551	0.581	0.568	0.575	0.928	0.997
oncotrad	0.267	0.274	0.286	0.272	0.263	0.278	0.339	0.216	0.000	0.602	0.614	0.565	0.617	0.939	1.041
QUAILQR1	0.607	0.592	0.609	0.609	0.599	0.619	0.630	0.551	0.602	0.000	0.639	0.678	0.781	1.060	1.287
RNSC1	0.607	0.600	0.612	0.595	0.550	0.580	0.596	0.581	0.614	0.639	0.000	0.109	0.328	1.072	1.313
MMSC1	0.623	0.604	0.616	0.606	0.561	0.583	0.606	0.568	0.565	0.678	0.109	0.000	0.327	1.055	1.289
Humanhevin	0.615	0.614	0.636	0.628	0.587	0.608	0.624	0.575	0.617	0.781	0.328	0.327	0.000	1.049	1.307
CESPARC	0.929	0.910	0.902	0.929	0.876	0.861	0.962	0.928	0.939	1.060	1.072	1.055	1.049	0.000	0.655
DMSPARC-AJ	0.954	0.952	0.954	0.980	0.940	0.957	1.066	0.997	1.041	1.287	1.313	1.289	1.307	0.655	0.000

47

L'arbre non-enraciné



Où mettre la racine ?

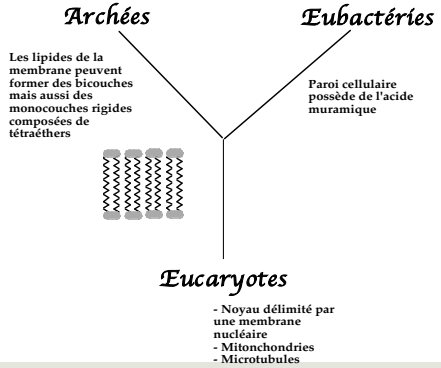
48

Comment enraciner…

- Centre de gravité de l'arbre
  - Implique une notion d'horloge moléculaire
- En utilisant une séquence pour laquelle on a de forts arguments qu'elle soit extérieure… l'*outgroup*
  - Dans l'activité scientifique, on ne « sait » rien car tout est réfutable ! Il y a toujours le risque de se tromper !

49

L'enracinement du vivant



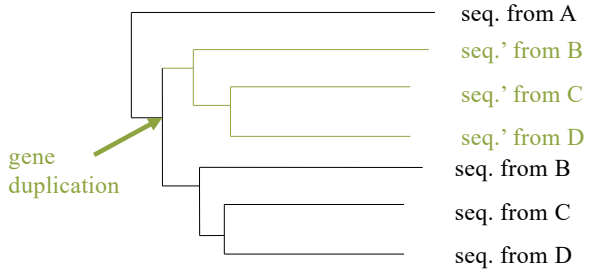
50

Phylogénie de gènes

- Une phylogénie peut aussi représenter l'évolution des gènes incluant les évènements de duplication et de transferts horizontaux.
  - Orthologues: Séquences dérivant d'un évènement de spéciation
  - Paralogues: Séquences dérivant d'un évènement de duplication
  - Ohnologues: Séquence dérivant d'un évènement de duplication complète du génome
  - Xénologues: Séquences dérivant d'un évènement de transfert horizontal

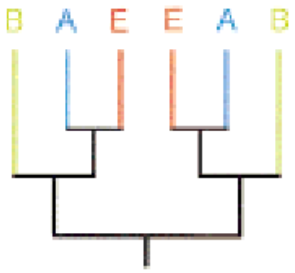
51

Exemple d'orthologues et paralogues

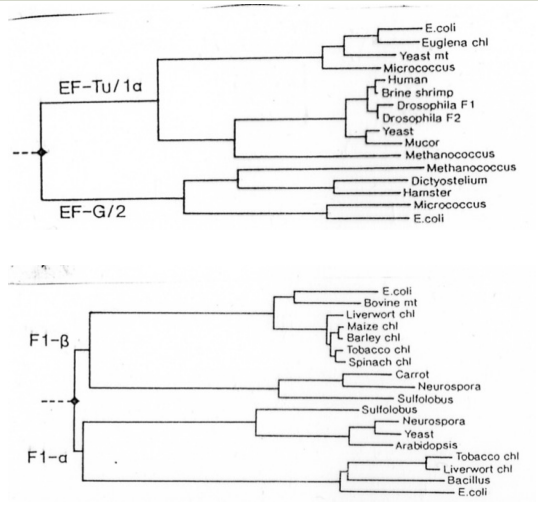


52

Les duplications de gènes chez Luca ?



53



54

Méthodes de reconstruction

- Méthodes phénétiques basées sur les distances
- Méthodes basées sur les changements d'états de caractère
  - Méthode basée sur la parcimonie
  - Méthode basée sur la vraisemblance
  - Méthode bayésienne
- **On cherchera la topologie de l'arbre qui minimise le nombre de changements requis pour expliquer l'alignement (parcimonie) ou qui maximise la vraisemblance de l'alignement (maximum de vraisemblance) ou qui maximise la crédibilité de l'arbre (méthode bayésienne).**

55

Nombre d'arbres possibles

- Nombre d'arbres non enracinnés pour n taxons
  Nu=(2n-5)*(2n-7)*...*3*1=(2n-5)!/[2n-3*(n-3)!]
- Nombre d'arbres enracinnés pour n taxons
  Nr=(2n-3)*(2n-5)*(2n-7)*...*3*1=(2n-3)!/[2n-2*(n-2)!]
- Ainsi, pour 50 espèces, il existe $2{,}8.10^{74}$ arbres possibles. Un ordinateur extrêmement performant qui analyserait un milliard d'arbres par seconde aurait ainsi besoin de 10^{58} années pour mener à bien le calcul exhaustif de tous les arbres.

56

Algorithmes

- Dans la pratique, il est donc hors de question de parcourir tout l'espace de recherche et il faut donc accepter que le programme ne produise pas nécessairement le meilleur arbre mais, au moins, un arbre s'en approchant. On dit de tels programmes qu'ils sont des heuristiques.
- En pratique, on est obligé de faire un compromis entre temps de calcul et efficacité de la recherche, ce qui fait que l'on est rarement certain d'avoir trouvé le meilleur arbre, même si l'on est toujours sûr d'avoir trouvé un très bon arbre (c'est-à-dire proche du meilleur).

57

Maximum de parcimonie

- On cherchera la topologie qui nécessite le moins de changement évolutif.

58

Cladistique

Willi Hennig
1913-1976

- La parcimonie
  - L'hypothèse parcimonieuse est celle qui nécessite le moins d'événements évolutifs.
  - La parcimonie est basée sur la règle qu'entre deux solutions identiques, il convient de préférer celle se présentant le plus simplement (nécessitant le moins de paramètres).

Cette méthodologie a été aussi proposée dans:

Edwards A.W.F., Cavalli-Sforza L.L. 1963. The reconstruction of evolution. Ann. Hum. Genet. 27:104-105.

59

Le rasoir d'Ockam

- Le rasoir d'Ockham est un principe de raisonnement philosophique entrant dans les concepts de rationalisme. Son nom vient du philosophe franciscain Guillaume d'Ockham (1285, Ockham, Royaume-Uni - 1347, Munich, Allemagne).
- On le trouve également appelé principe de simplicité, principe d'économie ou principe de parcimonie (en latin *lex parsimoniae*).

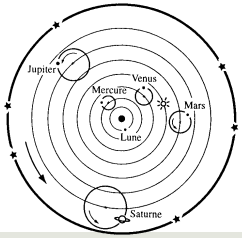
Pluralitas non est ponenda sine necessitate

« Les multiples ne doivent pas être utilisés sans nécessité. »

60

Exemple de modèle

- Le mouvement des planètes est expliqué par Ptolémée (90 - 168) par une série de mouvements circulaires avec la terre au centre du système soit un modèle géocentrique.



61

Exemple de modèle

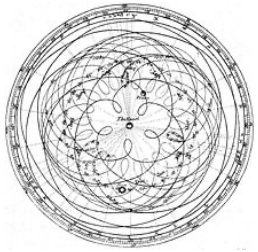
- Ce modèle est utilisé pendant très longtemps et encore au 16ème siècle.
- Il est matérialisé par une sphère armillaire qui représente le mouvement de la lune, des planètes et du soleil autour de la terre (armille = cercles métalliques).

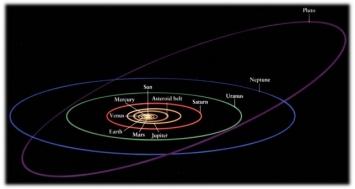
Jan Gossaert (ca 1478-1532)
Portrait de jeune princesse portant une sphère armillaire, vers 1530,
The National Gallery, Londres

62

Comparaison de modèles

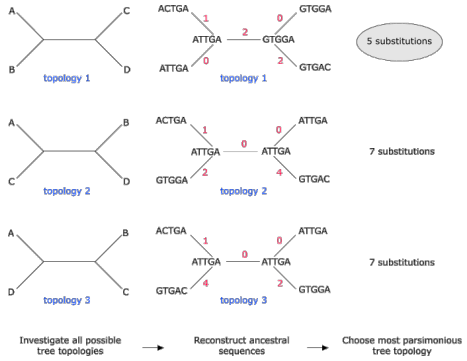
- On va donc chercher à utiliser un critère de meilleur ajustement mais en pénalisant par le nombre de paramètres utilisés dans le modèle pour arriver à la proposition de Galilée du modèle héliocentrique.





63

Exemple de parcimonie



64

Limites de la parcimonie

- Ne permet pas d'implémenter des modèles complexes d'évolution de caractères.
  - Il est possible de pondérer le poids des différents changements, mais c'est tout.
- Ne prend pas en compte la longueur des branches
- Par contre, cette méthode est assez intuitive.

65

Maximum de vraisemblance

$$L = \Pr(Data|Model)$$

- On recherchera l'arbre qui maximise la probabilité d'observer les données (séquences) conditionnellement à un modèle d'évolution, une topologie et des longueurs de branche.

Première utilisation en phylogénie: Cavalli-Sforza and Edwards (1967) pour des données de fréquence génique et Felsenstein (1981) des sequences d'ADN

Joe Felsenstein

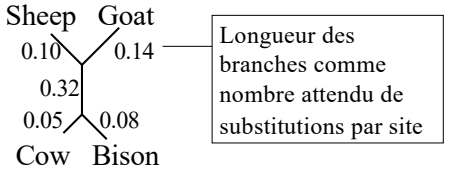
66

Probabilité, fréquence, vraisemblance

- Il existe de nombreuses définitions de ces termes qui ne recouvrent donc pas exactement la même chose chez différents auteurs.
- En général, le terme probabilité est utilisé pour décrire la chance (ou le risque !) qu'un événement particulier se produise.
- Le terme fréquence (ou plus précisément fréquence empirique) désigne une réalisation particulière de la probabilité qu'un événement particulier se produise.
- Le terme vraisemblance désigne une probabilité conditionnelle, c'est à dire la probabilité d'observations en fonction de paramètres supposés connus.
- En anglais, *frequency* est utilisé pour nommer des dénombrements (par exemple sur la commande hist() de R, freq=TRUE permet d'afficher les nombres d'observations); on distingue:
  - *Frequency (or absolut frequency)*, des nombres
  - *Relative frequency (or empirical probability)*, des fréquences !

67

En phylogénétique, l'hypothèse est une topologie, les longueurs de branches et un modèle d'évolution sous lequel les données ont évolué:



La parcimonie cherche à minimiser le nombre de substitutions

La vraisemblance cherche à estimer le nombre de substitutions dans un modèle donné pour qu'elles soient le plus proche possible des observations.

68

Relations entre parcimonie et vraisemblance

- Il a été montré que la parcimonie peut être considérée comme un sous-ensemble simplifié de la vraisemblance.
- **Parsimony, Likelihood, and the Role of Models in Molecular Phylogenetics**
- *Mike Steel and David Penny*
- Mol. Biol. Evol 2000 17(6): 839-850.

69

Bayésien

- On recherchera l'arbre qui maximise la vraisemblance des données dans un modèle en prenant en compte des *priors*.
- Les *priors* sont des informations que l'on connait, par exemple d'études antérieures ou en raison de la biochimie des mutations.
  - Si les *priors* sont tirés dans une loi uniforme large, la méthode s'apparente à du maximum de vraisemblance.

70

Le bootstrap

- En statistiques, les techniques de bootstrap sont des méthodes d'inférence statistique basées sur la réplication multiple des données à partir du jeu de données étudié, selon les techniques de rééchantillonnage.
- Plus précisément, et c'est le sens du terme « rééchantillonnage », un bootstrap consiste à créer des « nouveaux échantillons » statistiques, mais uniquement par tirage avec remise, à partir de l'échantillon initial.

71

Exemple de bootstrap

- Soit une série de 200 nombres, $k=k_i$ tirés dans une loi quelconque. Je veux connaître les valeurs des quantiles 0.025 et 0.975.
- Connaître la valeur du quantile est relativement simple: k[order(k)][5] et k[order(k)][195] me les donne.

```
k <- rgamma(n=200, shape=2, rate=10)*runif(n=200, min=0, max=3)
hist(k, ylim = c(0, 120), las=1)
points(k[order(k)][5], 120, pch="|", col="red")
points(k[order(k)][195], 120, pch="|", col="red")
```

Hyndman, R. J. and Fan, Y. (1996) Sample quantiles in statistical packages, American Statistician 50, 361–365. doi: 10.2307/2684934.

72

Exemple de bootstrap

- Soit une série de 200 nombres, $k=k_i$ tirés dans une loi quelconque. Je veux connaître les valeurs des quantiles 0.025 et 0.975.
- Connaître la valeur du quantile est relativement simple: k[order(k)][5] et k[order(k)][195] me les donne… mais quelle est l'incertitude sur ces valeurs ?
- Si j'en avais plusieurs, ce serait simple, je calculerai simplement la moyenne et l'écart-type des estimations obtenues sur les différentes séries.
  - Cela rappelle le problème des réplicats en écologie quand on a une seule terre.

73

Exemple de bootstrap

- Si j'en avais plusieurs, ce serait simple, je calculerai simplement la moyenne et l'écart-type des différentes séries.

```
q0.025 <- NULL
q0.975 <- NULL
for (i in 1:1000) {
  kp <- rgamma(n=200, shape=2, rate=10)*runif(n=200, min=0, max=3)
  q0.025 <- c(q0.025, kp[order(kp)][5])
  q0.975 <- c(q0.975, kp[order(kp)][195])
}
mean(q0.025); sd(q0.025)
mean(q0.975); sd(q0.975)

hist(k, ylim = c(0, 120), las=1)
points(mean(q0.025), 120, pch="|")
segments(x0=mean(q0.025)-1.96*sd(q0.025), y0=120, x1=mean(q0.025)+1.96*sd(q0.025), y1=120)
points(mean(q0.975), 120, pch="|")
segments(x0=mean(q0.975)-1.96*sd(q0.975), y0=120, x1=mean(q0.975)+1.96*sd(q0.975), y1=120)
```

74

Exemple de bootstrap

- Mais là, j'en ai qu'un… alors je peux en créer des pseudo-nouvelles séries par rééchantillonnage.

```
k <- rgamma(n=200, shape=2, rate=10)*runif(n=200, min=0, max=3)
kp <- sample(x=k, size = length(k), replace = TRUE)
```

75

Exemple de bootstrap

```
q0.025_bootstrap <- NULL
q0.975_bootstrap <- NULL
for (i in 1:1000) {
  kp <- sample(x=k, size = length(k), replace = TRUE)
  q0.025_bootstrap <- c(q0.025_bootstrap, kp[order(kp)][5])
  q0.975_bootstrap <- c(q0.975_bootstrap, kp[order(kp)][195])
}
mean(q0.025_bootstrap); sd(q0.025_bootstrap)
mean(q0.975_bootstrap); sd(q0.975_bootstrap)

hist(k, ylim = c(0, 120))
points(mean(q0.025_bootstrap), 120, pch="|", col="red")
segments(x0=mean(q0.025_bootstrap)-1.96*sd(q0.025_bootstrap), y0=120,
        x1=mean(q0.025_bootstrap)+1.96*sd(q0.025_bootstrap), y1=120, col="red")
points(mean(q0.975_bootstrap), 120, pch="|", col="red")
segments(x0=mean(q0.975_bootstrap)-1.96*sd(q0.975_bootstrap), y0=120,
        x1=mean(q0.975_bootstrap)+1.96*sd(q0.975_bootstrap), y1=120, col="red")
```

76

Exemple de bootstrap

mean(q0.025_bootstrap); sd(q0.025_bootstrap)
mean(q0.025); sd(q0.025)

mean(q0.975_bootstrap); sd(q0.975_bootstrap)
mean(q0.975); sd(q0.975)

77

Applications à la phylogénie

- Pour connaître le support statistique des nœuds d'une phylogénie, on utilisera la même méthode mais ce que l'on ré-échantillonnera, ce sont des séquences.

```
00000000011…
12345678901…
ATGGTCGTAGT…
ATGGTCGTAGT…
ATGGTCGTAGT…
ATGGTCGTAGT…
ATGGTCGTAGT…
```

Ré-échantillonnage avec remise, donc même taille de séquences

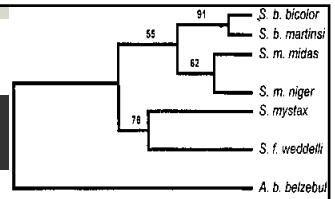
```
00000101010…
73752150911…

00100101000…
95253150972…
```

n fois, avec n=100 ou 1000 ou 5000

78

Résultats



- Évaluation de la fiabilité de la topologie de l'arbre
- 100, 1000 ou 5000 arbres construits suite au ré-échantillonnage des séquences
- Combien de fois une certaine branche est reproduite dans l'échantillon des arbres obtenu avec les séquences ré-échantillonnées
  - Valeurs comprises entre 1 et 100 (%)
- La fréquence à laquelle chacun des nœuds de votre phylogénie dans l'ensemble ré-échantillonné donne une certaine mesure de la confiance que vous pouvez avoir pour ces nœuds

79