

Introduction à la bioinformatique

La construction des arbres phylogénétiques

1

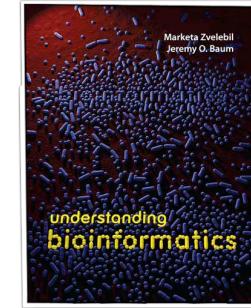
Objectifs

- Être capable d'expliquer les différents types d'arbres phylogénétiques
- Être capable d'expliquer les principes d'évolution moléculaire
- Comprendre quelles caractéristiques sont importantes pour la construction d'un modèle d'évolution moléculaire
- Être capable de calculer les distances entre des séquences
- Connaître les méthodes de construction
- Comprendre comment les méthodes UPGMA et Fitch-Margoliash fonctionnent

3

Bibliographie

- Zvelebil et Baum, Understanding bioinformatics



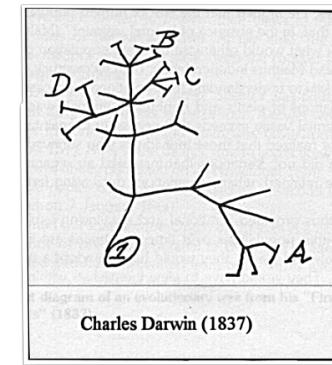
2

Quoi?

Comment représenter des relations (évolutives) entre des séquences ?

Comment montrer explicitement les distances évolutives entre ces séquences?

Arbre phylogénétique



4

Quoi?

Deux types de problèmes peuvent être étudiés

Les séquences orthologues sont des séquences qui ont une fonction commune et un ancêtre commun

1. En analysant des séquences orthologues appartenant à des espèces différentes, on peut examiner le rapport évolutif entre ces espèces

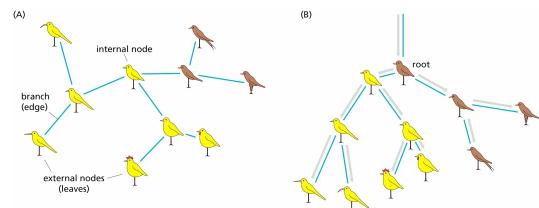
L'arbre qui résout ce problème s'appelle un arbre d'espèces

2. En étudiant un ensemble de séquences de la même classe protéique, nous pouvons examiner l'évolution de la fonctionnalité de cette classe de protéines.

5

Quoi? 3

Un arbre phylogénétique propose une hypothèse pour les rapports évolutifs entre des objets = taxons ou unités taxinomiques opérationnelles (UTO's).



Les noeuds internes = 1) événements de spéciation ou 2) des événements de duplication des gènes.

Chaque noeud interne prévoit d'avoir 3 branchements = une configuration de bifurcation

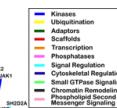
7

Quoi? 2

Cet arbre représente le rapport évolutif entre les séquences appartenant à la famille des domaines SH2

Chaque feuille est une protéine dans laquelle le domaine SH2 apparaît (les noms sont coloriés en rouge quand la structure est connue)

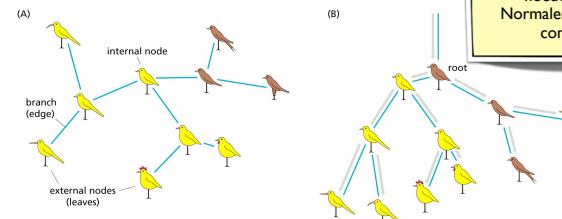
B.A.Liu et al. (2006) The Human and Mouse Complement of SH2 Domain Proteins—Establishing the Boundaries of Phosphotyrosine Signaling. Molecular Cell 22: 851-868



6

Quoi? 4

Attention, les ancêtres communs correspondent aux noeuds internes. Normalement on ne les connaît pas.



Si tous les noeuds internes ont 3 branchements (bifurcation) on dit que l'arbre est entièrement résolu

Un arbre qui est partiellement résolu a au moins un noeud interne avec quatre branchements ou plus

8

Quoi? 5

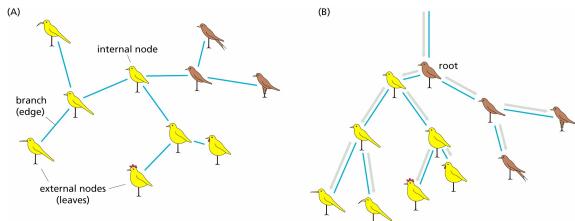


Figure A est un **arbre déraciné**

Il n'y a aucun ancêtre commun connu

Donc, il n'est pas clair quelle espèce est l'ancêtre commun. Il y a de l'ambiguité.

Figure B est un **arbre enraciné**

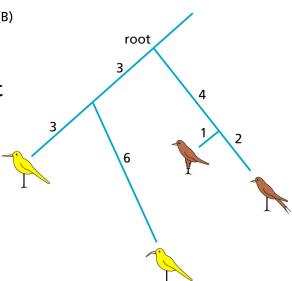
La racine de l'arbre est l'ancêtre commun. La direction d'évolution est définie sans ambiguïté

9

Quoi? 7

Un **arbre additif** montre les lignes de descente et aussi les longueurs des branches

(B)



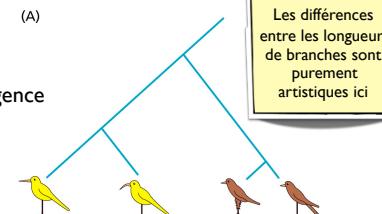
les **longueurs des branches** donnent de l'information évolutive quantitative, ce qui peut être proportionnelle au nombre de mutations entre les noeuds

La **distance évolutive totale** entre les taxa est donnée par la somme de toutes les longueurs des branches

11

Quoi? 6

En général il y a 3 types d'arbres phylogénétiques: 1) les **cladogram**, 2) les arbres **additifs** et 3) les arbres **ultramétriques**



Le **cladogram** montre la ligne de la descente du taxon, mais ne dit rien du chronométrage ou la mesure de divergence

La longueur des branches n'a aucune signification
Seule la topologie est importante

L'exemple montre que les quatre oiseaux ont un certain ancêtre commun et qu'un premier événement de spéciation a produit la différence entre les oiseaux de type jaune et marron.

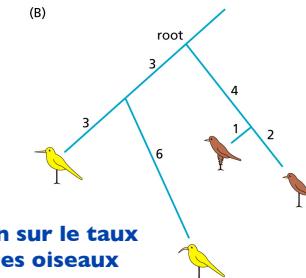
Sont rarement utilisés pour l'analyse des séquences

10

Quoi? 8

La distance évolutive entre l'ancêtre des oiseaux jaunes et la racine est plus petite (3) que la distance évolutive entre l'ancêtre des oiseaux bruns et la racine (4)

(B)



Cette relation nous donne de l'information concrète sur la divergence évolutive entre les espèces

Mais parce qu'il n'y a pas d'information sur le taux de mutations on ne peut pas dire que les oiseaux bruns sont divergés plus tard

Ces arbres peuvent être enracinés ou déracinés

12

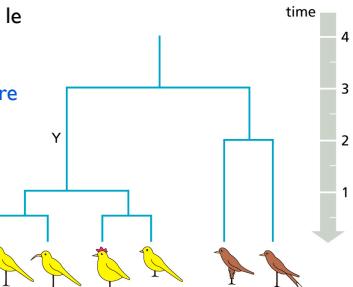
Quoi? ,

Un **arbre ultramétrique** est comme un arbre additif mais où le taux de mutation est constant le long chaque branche de l'arbre

Cette propriété est appelée **l'horloge moléculaire**

Les arbres ultramétriques ont toujours une racine et l'axe de l'arbre est proportionnel au temps

La distance évolutive entre la racine et les feuilles de l'arbre est la même pour tous les taxons



13

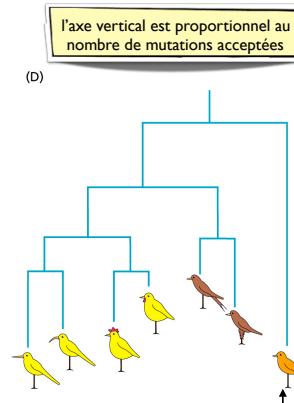
Quoi? 10

Les séquences ne conforment pas toujours à cette horloge moléculaire parce qu'on peut avoir des différences dans les taux de mutation à cause des différences dans la pression évolutionne.

Par conséquent **les arbres additifs sont utilisés le plus souvent**

Afin de situer la racine, on ajoute parfois un groupe de séquences qui ont un rapport évolutif distant avec l'ensemble des données initiales (= **outgroup**)

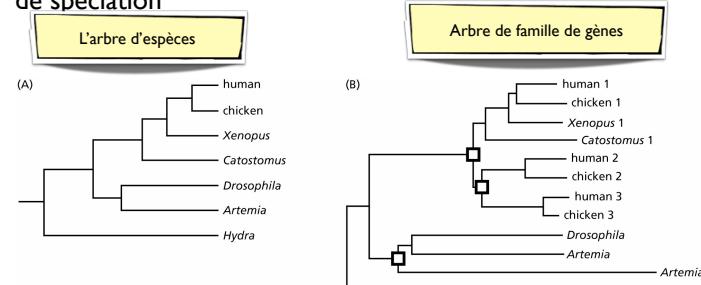
La racine est placée entre le **outgroup** et les autres taxons



14

Quoi? 11

(A) Des noeuds internes dans les arbres des espèces contiennent seulement de l'information sur les événements de spéciation

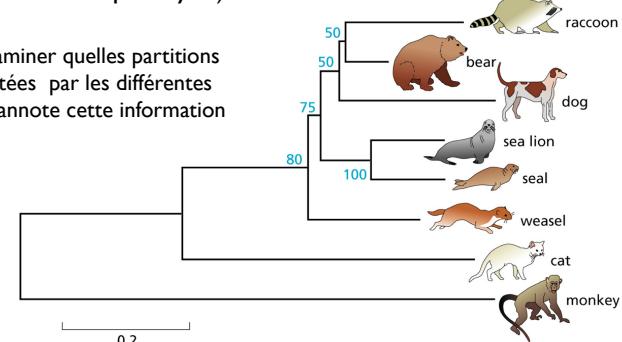


(B) Des noeuds internes peuvent correspondre soit à des événements de duplication des gènes soit à des événements de divergence au sein d'espèces ainsi qu'entre des espèces

l'arbre condensé et consensus

Les partitions peuvent être utilisées pour comparer des arbres qui sont produits par des méthodes différentes en utilisant les mêmes données ou par des échantillonnages différents d'une collection de données (= **bootstrap analysis**)

On peut examiner quelles partitions sont supportées par les différents arbres. On annote cette information sur l'arbre



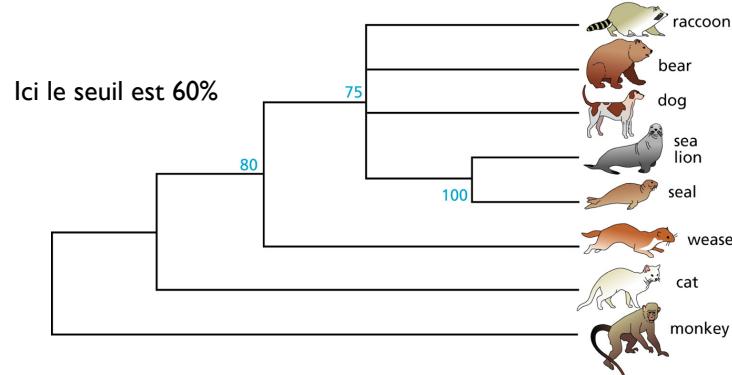
Soltis and Soltis (2003) Applying the bootstrap in phylogeny reconstruction. Statistical Science 18(2):256-267

15

16

l'arbre condensé et consensus 2

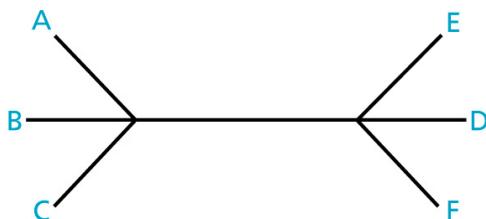
Quand on enlève toutes les partitions qui n'ont pas assez de support dans les différents arbres, on crée un arbre condensé



17

l'arbre condensé et consensus 4

On peut aussi récupérer les caractéristiques qui sont observées dans tous les arbres. Dans ce cas, L'arbre est appelé un arbre consensus

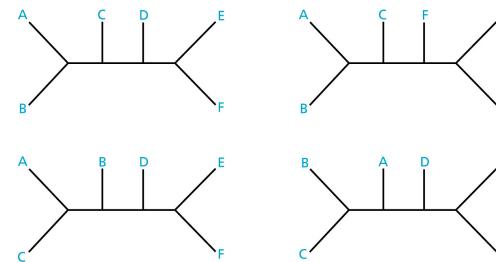


Ici, les partitions (A,B,C) et (D,E,F) sont observées dans tous les arbres

19

l'arbre condensé et consensus 3

On peut aussi récupérer les caractéristiques qui sont observées dans tous les arbres. Dans ce cas, L'arbre est appelé un arbre consensus



Ici, les partitions (A,B,C) et (D,E,F) sont observées dans tous les arbres

18

Evolution moléculaire

Construire des arbres phylogénétiques exige qu'on comprenne les principes d'évolution moléculaire, c.-a.-d. on a besoin d'un modèle d'évolution

L'évolution darwinienne par sélection naturelle se concentre sur les effets des changements sur le succès d'un organisme (survie de plus convenables) ⇒ l'ensemble de l'organisme

Les changements sont causés par des mutations et la qualité d'un modèle évolutif est liée à la qualité de la représentation des changements/préférences de chaque position dans les séquences

20

Evolution moléculaire 2

Quatre observations sont importantes pour la construction d'un modèle évolutif de grande qualité

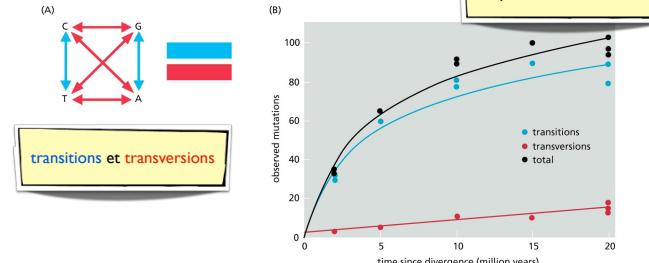
1. La plupart des séquences liées ont beaucoup de positions qui **ont subi à plusieurs mutations à la même position**
2. Le **taux de mutation accepté n'est pas le même** pour tous les types de substitutions dans les séquences.
3. Les **positions de codons ont des taux de mutation différents** et ces mutations peuvent produire des acides aminés identiques (synonyme) ou différents (non-synonyme)
4. Seuls des **gènes orthologues doivent être utilisés** pour la construction d'arbres phylogénétiques

21

Différences entre les taux de mutation

Le modèle évolutif le plus simple suppose qu'il n'y a aucune différence dans 1) les taux de mutation et 2) les préférences de substitution pour les différentes positions dans un alignement

En réalité, les deux peuvent varier



23

Mutations multiples

En réalité, les séquences génétiques évoluent à taux de mutation différents et à cause de ça, nombreuses positions ont muté plusieurs fois.

Même les nucléotides conservés peuvent avoir subi des mutations dans le passé ($A \rightarrow T \rightarrow A$)

L'évaluation de la distance évolutive en comptant simplement la fraction de positions alignées qui ne sont pas identiques, est **une sous-estimation** du vrai nombre de mutations

Construire un arbre phylogénétique correct exigera **qu'une correction de distance soit prise en considération**

22

Les mutations (non-)synonymes

Puisque le code génétique est dégénéré, les mutations dans un codon changent (**non-synonyme**) ou ne changent pas (**synonyme**) l'acide aminé encodé

	U	C	A	G	
U	UUU UCU UUA UUG	Phn UCU UCA UCG	Leu Ser	UAU UAC UAG UAA	Tyr Stop Stop Stop
C	CUU CUC CUA CUG	CCU CCC CCA CCG	Pro Leu Leu Leu	CAU CAC CAA CAG	His Gly Gln Gln
A	AUU AUC AUA AUG	ACU ACA ACA ACG	Ile Thr Thr Met	AAA AAC AAA AAG	Arg Asn Arg Lys
G	GUU GUU GUA GUG	GCU GCC GCA GGC	Ala Val	GAU GAC GAA GAG	Asp Glu

First position (5' end)		Third position (3' end)			
U	UUU	U	U	U	U
C	CUU	C	C	C	C
A	AAU	A	A	A	A
G	GAU	G	G	G	G

Amino acid names:

Ala = alanine Gln = glutamine Leu = leucine Ser = serine
Arg = arginine Glu = glutamate Lys = lysine Thr = threonine
Asn = asparagine Gly = glycine Met = methionine Tyr = tyrosine
Asp = aspartate His = histidine Phe = phenylalanine Val = valine
Cys = cysteine Ile = isoleucine Pro = proline

24

Les mutations (non-)synonymes 2

La plupart des mutations à la troisième position du codon produit des synonymes

Par conséquent, le taux de mutation à la troisième position sera plus élevé que pour les autres positions

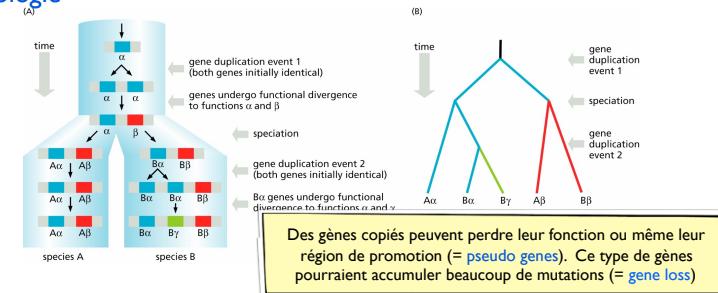
Des mutations non-synonymes changent le acide aminé et par conséquent ils peuvent changer la structure ou la fonction d'une protéine

Les mutations qui introduisent des changements dans les protéines peuvent être maintenues et devenir répandues ou sont rapidement perdues à cause de la sélection naturelle

25

Seulement les orthologues

La supposition clé en construisant un arbre phylogénétique est que toutes les séquences impliquées sont tirées d'un ancêtre commun = **homologie**



Les homologues peuvent être créés par spéciation (**orthologues**) ou par la duplication de gène (**paralogues**)

27

Les mutations (non-)synonymes 3

Quand il n'y a pas de pression sélective, des mutations s'étendent ou disparaîtront en conséquence d'événements aléatoires = **random drift**

S'il y a de la pression sélective et que la mutation fournit un avantage à l'organisme, la **sélection positive** augmentera la probabilité que cette mutation reste dans le génome

Quand la mutation fournit un désavantage à l'organisme, il disparaîtra à cause de la **sélection purifiante ou négative**

Kimura's **théorie d'évolution neutre** décrit qu'en réalité la plupart des mutations ne rencontrent pas la sélection forte (négative ou positive) et donc les mutations apparaissent à cause du *random drift*

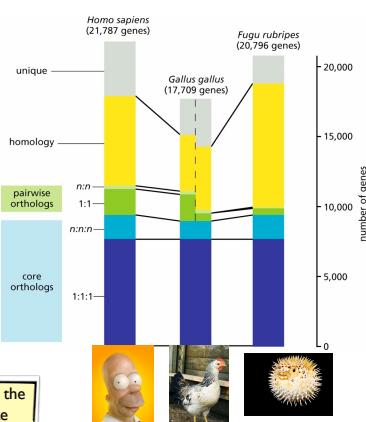
26

Seulement les orthologues 2

Une comparaison du nombre d'orthologues, d'homologues et des gènes identiques entre trois espèces : homme, poulet et une certaine poisson (*puffer fish*)

Une majorité des gènes orthologues ont les mêmes fonctions dans la cellule

LW Hillier et al (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 432:695-716



28

Seulement les orthologues 3

Parce que **seules les séquences orthologues ont la capacité d'identifier des événements de spéciation**, on préfère les utiliser pour la construction d'un arbre d'espèces.

Ajouter des séquences paralogues rend l'interprétabilité confuse puisque certains noeuds correspondent aux événements de duplication et pas aux événements de spéciation

Le transfert horizontal ou latéral peut également embrouiller les résultats.

Pour ces raisons, on préfère utiliser des séquences orthologues à moins que le but ne soit pas d'étudier les relations entre des protéines avec la même fonction.

29

Quel modèle?

Model name	Base composition	Different transition and transversion rates	All transition rates identical	All transversion rates identical	Reference
JC (JC69)	1:1:1:1	No	Yes	Yes	Jukes and Cantor (1969)
Felsenstein 81 (F81)	Variable	No	Yes	Yes	Felsenstein (1981)
K2P (K80)	1:1:1:1	Yes	Yes	Yes	Kimura (1980)
HKY85	Variable	Yes	No	No	Hasegawa et al. (1985)
Tamura-Nei (TN)	Variable	Yes	No	Yes	Tamura and Nei (1993)
K3P (K81)	Variable	Yes	No	Yes	Kimura (1981)
SYM	1:1:1:1	Yes	No	No	Zharkikh (1994)
REV (GTR)	Variable	Yes	No	No	Rodriguez et al. (1990)

Plusieurs modèles évolutifs sont proposés prenant en compte des éléments spécifiques comme la préférence pour certaines bases

Ces modèles sont utilisés pour corriger des fonctions de distance ou pour le calcul de la qualité des arbres pendant le processus-même

31

Construction d'arbre

Pour construire un arbre qui représente l'histoire évolutive entre des espèces, la première étape est **de bien choisir les données**, c.-à.-d. les séquences d'ARN ou les gènes qui représentent les espèces

Après, **trois décisions supplémentaires** doivent être faites:

1. Quelle méthode est utiliser pour construire l'arbre?
2. Quel modèle évolutif?
3. Quel test sera exécuté pour évaluer la robustesse des prédictions qui résultent de l'arbre?

30

Quel modèle? 2

La chose la plus logique à faire est d'utiliser plusieurs modèles évolutifs avec la même méthode de construction et d'accepter la méthode qui s'adapte le mieux aux données

Pour l'instant, il n'est pas clair quelle approche de comparaison est la meilleure.

En plus, même si on a créé des méthodes statistiques pour certains mécanismes de construction, il n'y a pas de consensus sur la validité des méthodes

par exemple: hierarchical likelihood ratio test, Akaike Information criterion et Bayesian information criterion

32

Calcul de distance

Le plus simple (mais imprécis) qui mesure la distance évolutive en estimant le nombre des mutations entre l'ancêtre commun et les séquences, est la *p-distance* (ou la *différence d'alignement fractionnaire*)

$$p = \frac{D}{L}$$

Nombre de substitutions dans les positions alignées
←
Nombre de positions alignées entre les deux séquences

Pour améliorer l'exactitude, cette méthode est modifiée en utilisant des modèles d'évolution plus sophistiqués

33

Calcul de distance 3

Les deux corrections précédentes ne prennent pas en compte les caractéristiques biochimiques des éléments qui construisent la séquence

elles peuvent être utilisées aussi bien avec des séquences de nucléotides qu'avec des séquences d'acides aminés

Le *modèle de Jukes-Cantor (JC)* est un des modèles les plus simples qui se focalise sur les séquences de nucléotides

Supposition de base: toutes les positions sont indépendantes et les taux de mutation ($= \alpha$) sont identiques pour chaque base (A,C,G ou T).

Même si le modèle est vraiment simple (et connu pour être incorrect) il a prouvé être très utile

35

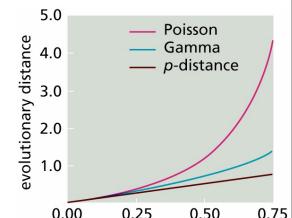
Calcul de distance 2

La *correction de Poisson* prend en compte le fait que plusieurs mutations peuvent se produire à la même position

$$d_p = -\ln(1-p)$$

La *correction de Gamma* prend en compte le fait que le taux de mutation peut dépendre de la position dans la séquence

$$d_\Gamma = a [(1-p)^{-1/a} - 1]$$



34

Calcul de distance 4

Les taux de substitution dans le modèle de Jukes-Cantor (JC) peuvent être décrits dans une matrice:

	A	C	G	T
A	-3α	α	α	α
C	α	-3α	α	α
G	α	α	-3α	α
T	α	α	α	-3α

la somme des éléments d'une ligne est égale à zéro, ce qui signifie que **la taille de la séquence ne peut pas changer**

la somme des éléments d'une colonne est aussi égale à zéro, ce qui signifie que **la composition des séquences reste constante**

36

Calcul de distance 5

En utilisant ce modèle on a dérivé une nouvelle fonction de distance (pour la dérivation, regardez le livre)

	A	C	G	T
A	-3α	α	α	α
C	α	-3α	α	α
G	α	α	-3α	α
T	α	α	α	-3α

$$d_{JC} = - (3/4) \ln [1 - (4/3)p]$$

cette fonction est [similaire à la correction de Poisson](#), signifiant qu'elle prend aussi en compte que plusieurs mutations peuvent se produire à la même position

37

Calcul de distance 7

Notez que tous ces modèles supposent toujours que toutes les positions dans les séquences ont le même taux de mutation

Les différences de taux de mutation pour les positions de séquences peuvent être incorporées en ajoutant la correction gamma

Donc si nous **combinons la correction avec le modèle de JC** (par exemple), on obtient la fonction suivante

$$d_{JC+\Gamma} = (3/4)\alpha [(1-(4/3)p)^{-1/\alpha} - 1]$$

39

Calcul de distance 5

Des modèles plus complexes distinguent entre les fréquences relatives de différentes types de mutation

Le [modèle de Kimura](#) fait la distinction entre les transitions (α) et les transversions (β)

	A	C	G	T
A	-2β-α	β	α	β
C	β	-2β-α	β	α
G	α	β	-2β-α	β
T	β	α	β	-2β-α

On produit une nouvelle fonction de distance, dans laquelle $P(Q)$ est la fraction observée de mutations de transition (transversion) entre des bases

Observée = information extraite des séquences alignées

$$d_{K2P} = - (1/2) \ln [1-2P-Q] - (1/4) \ln [1-2Q]$$

38

Calcul de distance 8

Jusqu'à maintenant la discussion s'est concentrée surtout sur des séquences de nucléotides (ARN ou gènes)

On peut appliquer les mêmes idées pour la construction des modèles évolutifs de protéines, sachant qu'on a besoin d'une matrice de dimension 20x20 qui exprime la probabilité de transition

$$d_{JC} = - (19/20) \ln [1 - (20/19)p]$$

40

Quelle méthode?

Deux méthodes générales

1. Les méthodes de groupements construisent les arbres graduellement, en commençant avec un nombre limité de séquences

Force: vitesse, robustesse et elles fonctionnent pour des grandes collections de données

Faiblesse: aucune mesure de qualité pour comparer l'arbre aux données

2. Les Méthodes de recherche topologique produisent une collection d'arbres en utilisant des mesures de qualité pour comparer les arbres aux données

Forces et faiblesses sont l'inverse de celles de la méthode précédente

Le choix de la méthode dépend de la taille de la collection de données et leur qualité

41

Chercher des topologies

Les algorithmes de construction d'arbre qui appartiennent à cette classe sont composés de deux processus:

1. Un processus qui produit plusieurs arbres avec des topologies différentes, souvent en utilisant les méthodes de groupement

2. Un processus pour évaluer l'optimalité des arbres et calculer les longueurs des branches

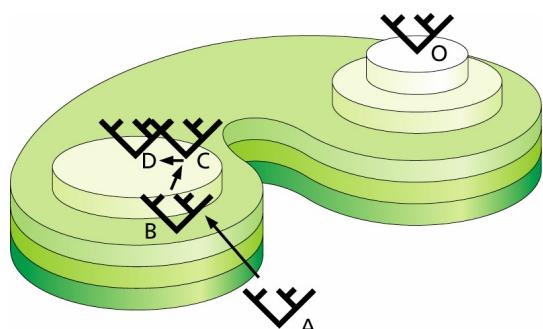
Parcimonie maximale: arbre avec la moindre mutation (problème de minimisation)

Probabilité maximale: arbre qui représente les données de la meilleure façon (problème de maximisation)

42

Chercher des topologies 2

Un espace de recherche imaginaire pour des topologies



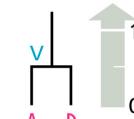
43

Groupement

Le *unweighted pair-group method using arithmetic averages* (**UPGMA**) est une des plus vieilles méthodes de groupement qui produit des **arbres ultramétriques enracinés** (tous les noeuds externes ont la même distance avec la racine)

(A)

d_{ij}	A	B	C	D	E	F
A	-	6	8	1	2	6
B		-	8	6	6	4
C			-	8	8	8
D				-	2	6
E					-	6

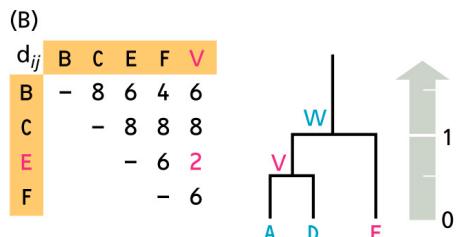


étape 1: les deux séquences avec la plus petite distance d_{ij} sont assumés être les dernières à avoir divergé. Leurs branches ont la même longueur, c.a.d la moitié de la distance entre elles.

44

Groupement 2

étape 2: à chaque étape la distance est recalculée et un autre noeud est ajouté à l'arbre

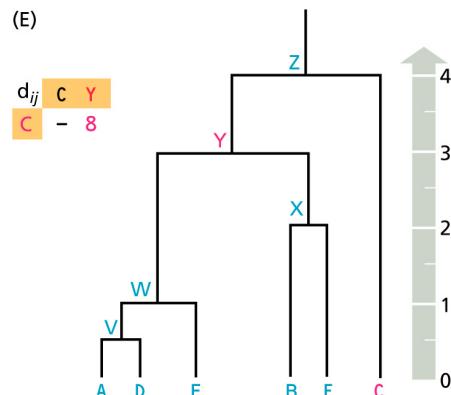


La distance entre deux groupes est : $d_{XY} = \frac{l}{N_X N_Y} \sum_{i \in X} \sum_{j \in Y} d_{ij}$

45

Groupement 4

à la fin, on obtient



47

Groupement 3

(C)

d_{ij}	B	C	F	W
B	- 8	4	6	
C	- 8	8		
F	- 6			

(D)

d_{ij}	c	W	X
c	- 8	8	
W	- 6		

Il existe un moyen de calculer plus vite la distance entre des groupes(p.e. W et X)

$$d_{XW} = \frac{N_B d_{BW} + N_F d_{FW}}{N_B + N_F}$$

46

Groupement 5

Donc en général, chaque élément appartient initialement à son propre groupe et à chaque étape les deux groupes avec la plus petite distance sont fusionnés.

Le processus se termine quand on a seulement un groupe

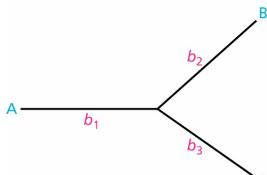
La [structure de données Disjoint-set](#) fournit une structure de données idéale pour mettre en oeuvre ces algorithmes de groupement

48

Groupement 6

L'algorithme de Fitch-Margoliash produit des arbres additifs déracinés

L'approche est basée sur une analyse d'un arbre avec trois branchements



La distance entre les paires de séquences est déterminée par la longueur de branches b_1 , b_2 et b_3

$$b_1 = \frac{1}{2} (d_{AB} + d_{AC} - d_{BC})$$

$$b_2 = \frac{1}{2} (d_{AB} + d_{BC} - d_{AC})$$

$$b_3 = \frac{1}{2} (d_{AC} + d_{BC} - d_{AB})$$

ce qui montre l'additivité de cette arbre: $d_{AB} = b_1 + b_2$, $d_{AC} = b_1 + b_3$ et $d_{BC} = b_2 + b_3$

49

Groupement 8

A) STEP 1 ($N = 5$)

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

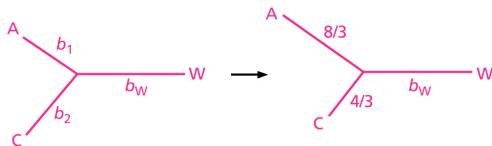
recalculer la distance

$$d_{AC} = 4$$

$$d_{AW} = \frac{5+9+8}{3} = \frac{22}{3}$$

$$d_{CW} = \frac{5+7+6}{3} = 6$$

$B, D, E \in W$
 $A, C \in X$



calculer les longueurs des branches

$$b_1 = \frac{1}{2} \left(4 + \frac{22}{3} - 6 \right) = \frac{8}{3}$$

$$b_2 = \frac{1}{2} \left(4 + 6 - \frac{22}{3} \right) = \frac{4}{3}$$

Les trois groupes initiaux sont $\{A\}$, $\{C\}$ et $W=\{B,D,E\}$, dans lesquels W est composé des éléments restants.

Après $\{A\}$ et $\{C\}$ sont combinés dans le groupe $X=\{A,C\}$

51

Groupement 7

Comme dans les arbres produits par la méthode UPGMA, l'arbre ici est aussi construit étape par étape

À chaque étape on a trois groupes : 2 pour les groupes qui seront combinés et un pour le reste de séquences

Les longueurs des branches sont déterminées en utilisant les équations des slides précédentes

En même temps on calcule la position du noeud interne

Les deux groupes seront combinés dans un nouveau groupe, pour lequel on doit recalculer la distance avec les autres groupes

50

Groupement 8

B) STEP 2 ($N = 4$)

d_{ij}	D	E	X
B	10	9	5
D	7	8	
E	7		

A, C $\in X$
D, E $\in Y$
B, X $\in Z$

$$d_{XB} = 5$$

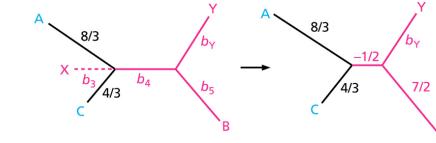
$$d_{BY} = \frac{10+9}{2} = \frac{19}{2}$$

$$b_3 = \frac{1}{2} \left(\frac{8}{3} + \frac{4}{3} \right) = 2$$

$$(b_3 + b_4) = \frac{1}{2} \left(5 + \frac{15}{2} - \frac{19}{2} \right) = \frac{3}{2}$$

$$b_4 = \frac{3}{2} - b_3 = \frac{3}{2} - 2 = -\frac{1}{2}$$

$$b_5 = \frac{1}{2} \left(5 + \frac{19}{2} - \frac{15}{2} \right) = \frac{7}{2}$$



Le trois groupes sont maintenant X, {B} et Y={D,E}

Après X et {B} sont combinés dans le groupe Z={B,{A,C}}

52

Groupement 9

C) STEP 3 ($N = 3$)

d_{ij} E Z
D 7 26/3
E 23/3

A,B,C $\in \mathbb{Z}$

$$d_{DE} = 7$$

$$d_{DZ} = \frac{26}{3}$$

$$d_{EZ} = \frac{23}{3}$$

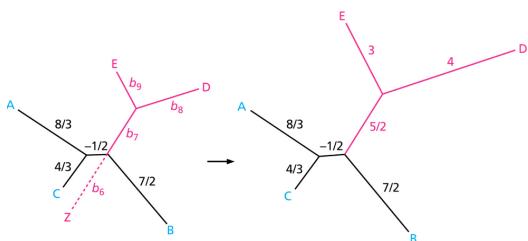
$$(b_6 + b_7) = \frac{1}{2} \left(\frac{26}{3} + \frac{23}{3} - 7 \right) = \frac{14}{3}$$

$$b_6 = \frac{1}{3} \left(\left[\frac{8}{3}, \frac{1}{2} \right] + \frac{7}{2} \left[\frac{4}{3}, \frac{1}{2} \right] \right) = \frac{13}{6}$$

$$b_7 = \frac{14}{3} - b_6 = \frac{14}{3} - \frac{13}{6} = \frac{5}{2}$$

$$b_8 = \frac{1}{2} \left(7 + \frac{26}{3} - \frac{23}{3} \right) = 4$$

$$b_9 = \frac{1}{2} \left(7 + \frac{23}{3} - \frac{26}{3} \right) = 3$$



53

Groupement 11

(D) patristic distance matrix Δ_{ij} from the tree and errors e_{ij}

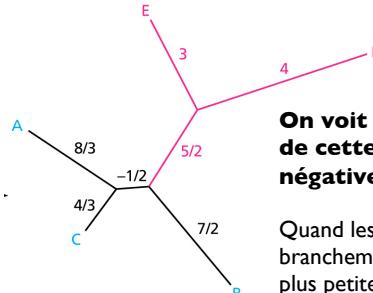
Δ_{ij}	B	C	D	E
A	5.7	4.0	8.7	7.7
B		5.3	10.0	9.0
C			7.3	6.3
D				7.0

e_{ij}	B	C	D	E
A	2/3	0	-1/3	-1/3
B	1/3	0	0	
C		1/3	1/3	
D				0

Le même problème devient visible quand on calcule les distances entre les noeuds externes (les distances patristiques) et qu'on les compare avec les distances initiales (regarder les erreurs)

55

Groupement 10



On voit immédiatement une faiblesse de cette approche : une longueur négative!

Quand les taux d'évolution entre les branchements sont différents, les éléments avec la plus petite distance ne sont pas nécessairement les éléments les plus proches.

Comme résultat on peut avoir des longueurs négatives.

54

Groupement 12

L'algorithme de groupement des voisins (*Neighbor-joining algorithm*) produit aussi des arbres additifs déracinés

L'idée fondamentale est que l'arbre réel sera l'arbre pour lequel la somme de toutes les longueurs de branches S est la plus petite.

Comme UPGMA et Fitch-Margoliash, des couples de noeuds sont identifiés à chaque étape et graduellement l'arbre se forme

La différence est dans la façon dont les noeuds sont identifiés

56

Résumé

Il devrait être clair qu'une grande collection de méthodes de construction d'arbre et de modèles d'évolution existe

Cela signifie qu'il n'y pas une approche optimale pour résoudre ce problème

Finalement, pour vérifier la qualité des arbres on a besoin de méthodes d'évaluation qui peuvent être utilisées pour comparer les arbres aux données

De cette façon on peut quantifier la qualité des arbres

Mais cela est en dehors le contenu de ce cours.