Arbres phylogénétiques

et problèmes NP-complets

---

gilles.hunault@univ-angers.fr Avril 2003

ne prenez pas de notes !

Vous trouverez sur la page

`http://www.info.univ-angers.fr/pub/gh/Idas/Wphylog/tgamtr.htm`

ces transparents (postscript A4R)
les URL des conférences internationales
les URL des données biologiques

quelques articles de référence
des URL de logiciels et démonstrations en ligne

But des regroupements

A partir de données biologiques produire des arbres phylogénétiques qui représentent l'évolution des espèces selon des hypothèses biologiques de filiation, mutation, spéciation...

Les données sont des caractéristiques biophysiologiques, biochimiques ou des séquences génétiques (acides nucléiques, acides aminés...).

Données et méthodes

Character-based methods

les méthodes *discrètes* utilisent directement
les données biologiques.

Distance-based methods

les méthodes *comparatives* travaillent
sur des "distances" entre les données.

Familles de méthodes et problèmes associés

Méthodes discrètes

1. **parcimonie** : minimiser le nombre de mutations
2. **compatibilité** (dont phylogénie parfaite) :

   maximiser le nombre de de caractères compatibles

Méthodes comparatives

3. **taxonomie** : retrouver les distances

Importance du thème

Les regroupements phylogénétiques font partie de la Biologie moléculaire.

Ils sont directement liés aux alignements multiples.

Leur importance est croissante pour la connaissance du génome, des maladies pathogènes (dont le SIDA) et la prévention du bioterrorisme.

Quelques organismes payeurs
en millions d'euros

NSF	National Science Foundation
NIH	National Institutes of Health
CDC	Center for Disease Control
DOE	Department of Energy
DOD	Department of Defense

Quelques "conférences" internationales

RECOMB	Research in Computational Molecular Biology 10 - 13 avril, Berlin.
ISMB	Intelligent Systems for Molecular Biology 29 juin - 3 juillet, Brisbane.
BTC	O'Reilly Bioinformatics Technology Conference 3 - 6 février, San Diego.
ICSEB	Int. Congress for Systematic and Ev. Biology
IFCS	Int. Federation of Classification Societies

...

les Données utilisées

Données pour les méthodes "discrètes" (1/2)

```
Baboon      1 Y 0 2 1 1 0 1 Y 0 2 1 1 0 ...
Chimp       1 Y 0 1 2 1 1 0 N 1 0 1 1 1
Marmoset    1 N 0 1 0 0 1 0 Y 1 1 1 1 0
Gibbon      0 Y 1 1 1 1 0 1 Y 0 2 1 1 1
Lemur       0 N 1 0 1 1 1 1 N 0 1 0 0 1
Human       1 N 0 1 0 0 1 1 Y 0 2 1 1 0
...
```

Données pour les méthodes "discrètes" (2/2)

```
Baboon       CCCACCCTCTCTTGCT----TAGTCTATATACCGCCATC...
Chimp        CTCACCGCCTCTTGCT----CAGCCTATATACCGCCATC...
Gibbon       CTCACCATCTCTTGCT----CAGCCTATATACCGCCATC...
Gorilla      CTCACCACCTCTTGCT----CAGCCTATATACCGCCATC...
Human        CTCACCACCTCTTGCT----CAGCCTATATACCGCCATC...
Lemur        CTCACCACTTCTTGCTAATTCAACTTATATACCGCCATA...
Marmoset     CTCACCACGTCTAGCC-AT-CAGCCTGTATACCGCCATA...
Mouse        CTCACCATCTCTTGCTAATTCAGCCTATATACCGCCATC...
Orangutan    CTCACCACCCCTTGCT----CAGCCTATATACCGCCATC...
PygmyCh      CTCACCGCCTCTTGCT----CAGCCTATATACCGCCATC...
SakiMonkey   CTTACCACCTCTTGCC-AT-CAGCCTGTATACCGCCATG...
Tarsier      CTTACCACCTCTTGCTAATTCAGTCTATATACCGCCATC...
```

Données traitées par les méthodes comparatives (1/2)

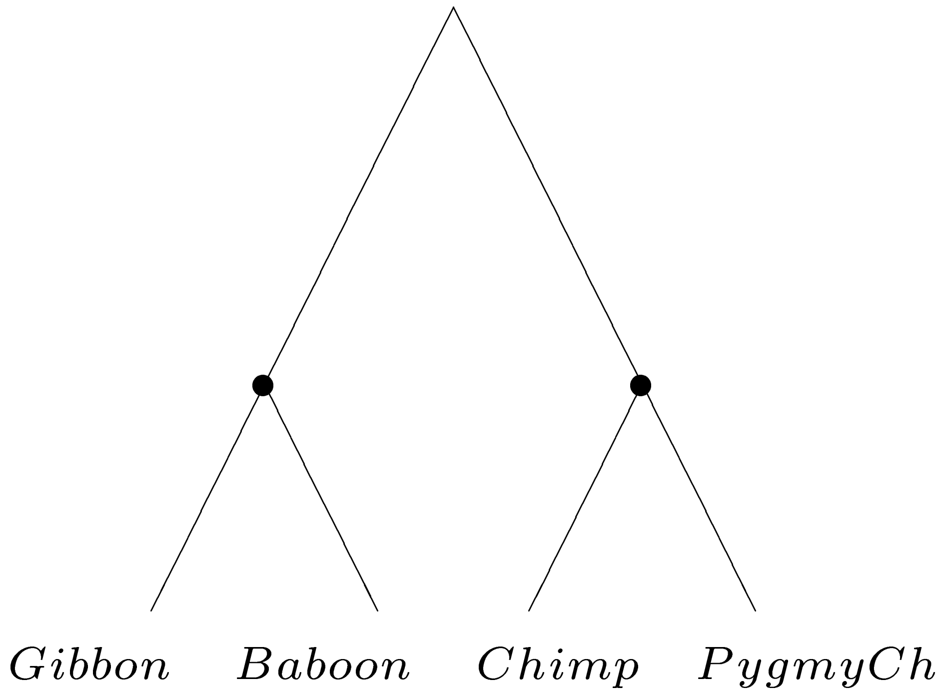
	Baboon	Gibbon	Orang	Gorilla	PygmyCh	...
Gibbon	0.184					
Orang	0.199	0.132				
Gorilla	0.184	0.116	0.096			
PygmyCh	0.178	0.119	0.097	0.041		
Chimp	0.182	0.113	0.099	0.046	0.017	
Human	0.176	0.114	0.096	0.041	0.032	...
...						

S'agit-il de "vraies" distances mathématiques
ou seulement de fonctions d'écarts ?

Données d'origine pour les méthodes comparatives (2/2)

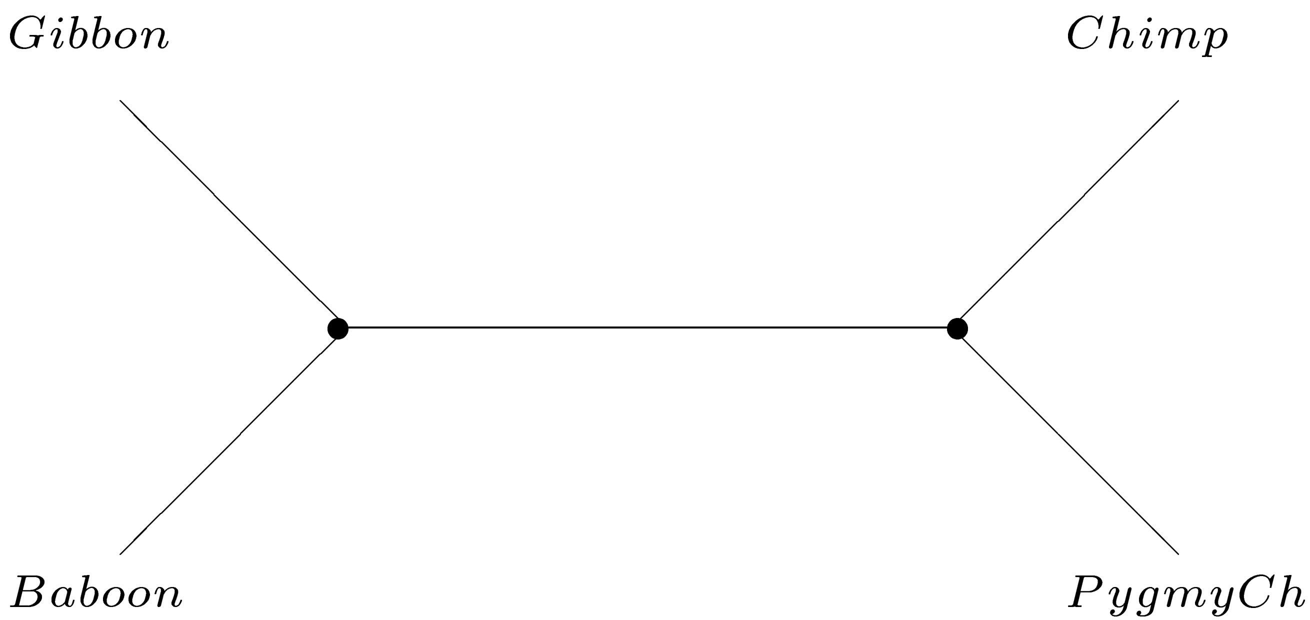
Baboon	CCCACCCUCUCUUGCUUAGUCUAUAUACCGCCAUC
Chimp	CUCACCGCCUCUUGCUCAGCCUAUAUACCGCCAUC
Gibbon	CUCACCAUCUCUUGCUCAGCCUAUAUACCGCCAUC
Gorilla	CUCACCACCUCUUGCUCAGCCUAUAUACCGCCAUC
Human	CUCACCACCUCUUGCUCAGCCUAUAUACCGCCAUC
Lemur	CUCACCACUUCUUGCUAAUUCAACUUAUAUACCGCCAUA
Marmoset	CUCACCACGUCUAGCCAUCAGCCUGUAUACCGCCAUA
Mouse	CUCACCAUCUCUUGCUAAUUCAGCCUAUAUACCGCCAUC
Orangutan	CUCACCACCCCUUGCUCAGCCUAUAUACCGCCAUC
PygmyChimp	CUCACCGCCUCUUGCUCAGCCUAUAUACCGCCAUC
SakiMonkey	CUUACCACCUCUUGCCAUCAGCCUGUAUACCGCCAUG
Tarsier	CUUACCACCUCUUGCUAAUUCAGUCUAUAUACCGCCAUC

Exemples d'arbre phylogénétique (1/2)



Arbre binaire enraciné dont les feuilles sont les espèces.

Exemples d'arbre phylogénétique (2/2)



Arbre $bifurcant$ dont les sommets de degré 1 sont les espèces.

Quelques *datasets* de référence 1/2

Cyclophiline 106 x 85 ; 92 x 219 ; 13 x 190
http://www.mcb.ucdavis.edu/faculty-labs/gasser/Datasets.html

Angiospermes 500 x 1428
http://www.cis.upenn.edu/~krice/treezilla/

Données de PHYLM 218 x 4182 ; 500 x 1428
htttp://www.lirmm.fr/~guindon/phyml/data/

Base de données Treebase 28726 taxa, 2089 arbres
http://128.205.184.102/treebase/TreeBASE.acgi$treeDisplay.html
http://www.treebase.org/treebase/console.html

Quelques *datasets* de référence 2/2

Primates 9 x 888 ; 22 x 11538

http://www.biomath.ucla.edu/msuchard/datasets.html
http://www.genetics.wayne.edu/lgross/PapersData.html

Mammifères 195 x m=7096

http://www.cs.ubc.ca/∼rogic/evaluation/dataset.html

les Regroupements classiques

Techniques classiques de regroupements en Analyse des Données (Classifications)

non supervisées	*supervisées*
classifications hiérarchiques,	réseau de neurones,
k-moyennes, centres mobiles,	algorithmes génétiques,
nuées dynamiques,	algorithmes de colonies, de fourmis,
cartes autoorganisatrices...	algorithmes auto-adaptatifs...

Classification hiérarchique non supervisée
ou ”binary agglomerative clustering”

Principe : à partir d'une matrice de distance entre éléments [isolés]

- on choisit deux éléments qu'on fusionne en un nouvel élément,
- on calcule la distance entre les anciens éléments et le nouveau,
- on supprime les deux anciens éléments choisis,
- on recommence jusqu'à avoir regroupé tout le monde.

Difficultés : nombreux choix possibles dans l'application des méthodes

- *choix de la distance initiale, du cout, du score*
  binaire : jaccard, russel-rao, simpson, sokal...
  comptage : czekanowski, clark, kulczynski...
  info : hamming, wagner, levenstein, édition...

- *choix du critère de sélection* (ou "indice d'agrégation")
  "linkage" simple ou complet, "pair group"...
  divergence, inertie...

- *choix de la formule de recalcul des distances*
  min, max, moyenne, pondérée, ultramétrique...

Formule de recalcul

Si $D_{x,y}$ est la distance de x à y, la distance de l'ancien groupe k au nouveau groupe composé de i et de j est

$$D_{ij,k} \;=\; a_i\, D_{i,k} \,+\, b_j\, D_{j,k} \,+\, c\, D_{i,j} \,+\, d\, |D_{i,k} - D_{j,k}|$$

méthode	a_i	b_j	c	d	*interprétation*
single	1/2	1/2	0	-1/2	minimum
complete	1/2	1/2	0	+1/2	maximum
upgma	$n_i/(n_i+n_j)$	$n_j/(n_i+n_j)$	0	0	moy. pondérée
nj	1/2	1/2	$-1/2^*$	0	moy. réduite

Quelques méthodes utilisées

Méthodes comparatives

- UPGMA — Unweighted Pair Group Method with Arithmetic mean, $Sokal$ et $Michener$, 1958.

- NJ — Neighbor-Joining, $Saitou$ et Nei, 1987.

Méthodes discrètes

- MP — Maximum Parsimony, $Fitch$, 1971.

- ML — Maximum Likelihood, $Jukes$ et $Cantor$, 1969.

Détail des méthodes, utilisation de logiciels

http://www.infobiogen.fr/doc/tutoriel/PHYLO/phylogenie.html

http://evolution.genetics.washington.edu/phylip/software.html

http://www.info.univ-angers.fr/pub/gh/Idas/Wcabiq/cabiq.htm les matrices d'Etats

et les Arbres d'évolution

Terminologie

- Caractères, états, espèces, mutations
- Arbres binaires [enracinés], arbres bifurcants
- Noeuds, Ancêtres

Matrice d'états : caractères et espèces

Soit M une matrice avec k lignes et m colonnes contenant des entiers ou des lettres. Les *lignes* s_i de M sont les espèces. Les *colonnes* c_j de M sont les caractères. Chaque caractère a au plus r valeurs ou "états" possibles.

pour l'ADN $r = 4$ et pour les protéines $r = 20$

L'élément $M_{i,j}$ en ligne i colonne j est l'état de l'espèce i pour le caractère j. Si l'état de c_j change il y a mutation pour ce caractère. Une mutation *ponctuelle* de l'espèce s correspond au changement d'un seul état pour s.

par extension : $M_{i,j} = s_i(\ j\ ) = c_j(\ i\ )$

Exemples de matrice d'états

	c_1	c_2	c_3	c_4
s_1	A	C	G	T
s_2	A	C	C	T
s_3	A	C	C	G

	c_1	c_2	c_3
s_1	0	1	0
s_3	0	0	0
s_4	1	1	0
s_5	1	0	1
s_6	1	0	1
s_7	1	0	0

Arbres et ancêtres

Soit T un arbre binaire enraciné et S ses feuilles ($espèces$).

Les noeuds [internes] de T sont les sommets de T qui ne sont pas des feuilles.

Il existe alors un chemin unique entre deux sommets.

Si u et v sont les deux sommets successeurs du noeud a, on dit que a est un ancêtre [commun] de u et v.

Pour n espèces, T possède $n-1$ noeuds internes et $2n-2$ arcs.

Arbres bifurcants 1/3

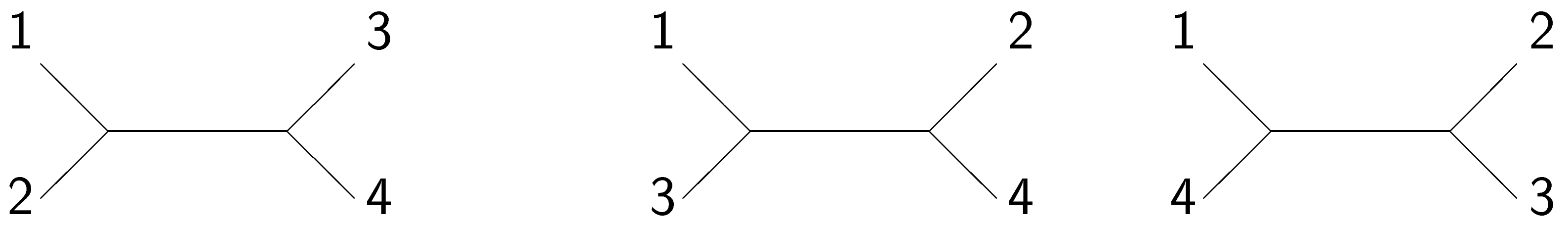
Un arbre T [non enraciné] est bifurcant sur n points *ssi* il possède n sommets de degré 1 (ou "feuilles"), étiquetés par $1, 2 \ldots n$ et si tous les autres sommets ("noeuds" [internes]) sont de degré 3.

Il y a bijection entre arbre bifurcant sur n points et arbre binaire enraciné sur $n-1$ feuilles.

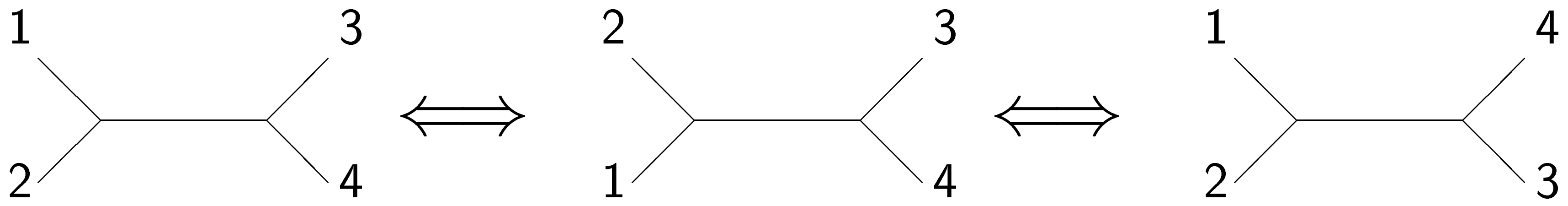
Si $B(n)$ désigne le nombre d'arbres bifurcants et $R(n)$ le nombre d'arbres binaires enracinés alors

$$R(n) = (2n-3)R(n-1) \quad \text{et} \quad R(n-1) = B(n).$$

Arbres bifurcants 2/3

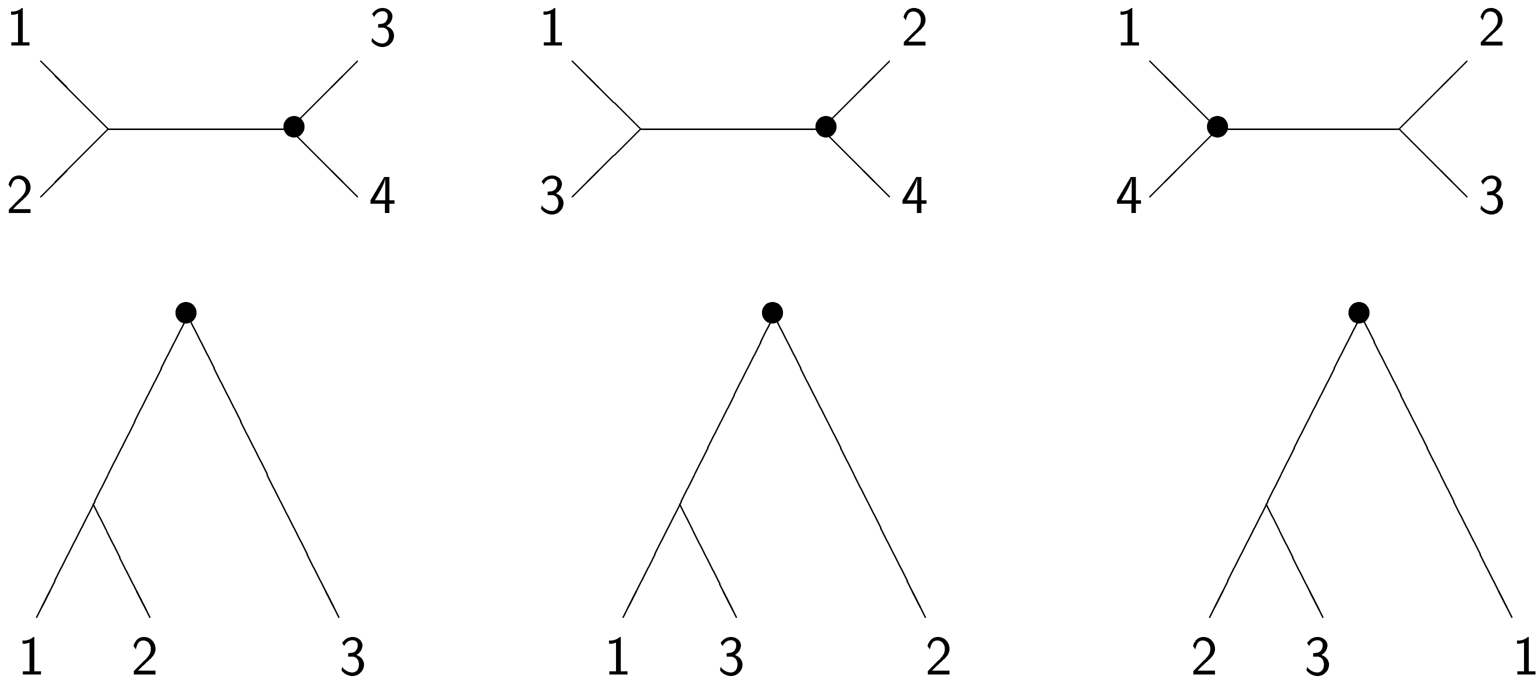


sachant :



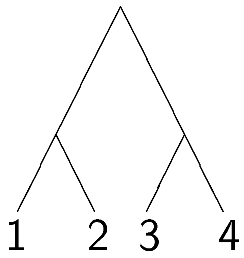
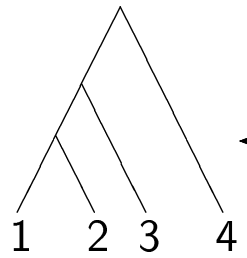
$B(4) = 3$: il y a 3 arbres $bifurcants$ ou "quartets" pour 4 espèces.

Arbres bifurcants 3/3



Bijection entre $B(4)$ et $R(3)$.

les 2 topologies pour les 15 arbres de $R(4)$

 $\Longleftrightarrow$	$12 \mid 34$	$13 \mid 24$	$14 \mid 23$	*3 arbres*
 $\Longleftrightarrow$	$12 \mid 3 \mid 4^*$	$13 \mid 2 \mid 4^*$	$14 \mid 2 \mid 3^*$	*6 arbres*
	$23 \mid 1 \mid 4^*$	$24 \mid 1 \mid 2^*$	$34 \mid 1 \mid 2^*$	*6 arbres*

nombre d'Arbres enracinés

n	$R(n)$	odg
5	105	1.1×10^2
10	34459425	3.4×10^7
15	213458046676875	2.1×10^{14}
20	8200794532637891559375	8.2×10^{21}
30	4951797690080198183901366117160891406250	5.0×10^{38}
35	48896013036866340154392278347307178464621367187 5	4.9×10^{47}
50	$\ldots$	2.8×10^{76}

à titre de comparaison : $2^{50} \simeq 1.1 \times 10^{16}$ et $50\,! \simeq 3.0 \times 10^{64}$
an 0 - an 2000 $\simeq 6.3 \times 10^{10}$ sec et 15 milliards d'années $\simeq 4.7 \times 10^{17}$ sec.

un peu de mathématiques

Depuis Schroder (1870) :

$$R(n) = (2n-3)!\,! = \prod_{i=1}^{n-1} (2i-1) = \frac{(2n-3)!}{2^{n-2}(n-2)!}$$

et grâce à la formule de Stirling et De Moivre (1730)

$$R(n) \sim \sqrt{2}\ \left(\frac{2}{e} \times (n-1) \right)^{n-1}$$

soit finalement l'encadrement : $n! < R(n+1) < n^n$ pour $n \geqslant 3$.

Arbres d'évolution

Soit C un ensemble de caractères, M une matrice d'états de C et soit E l'ensemble des espèces. On dit que T est un arbre d'évolution pour E (sous entendu via M et C) ssi

- T est un arbre binaire enraciné ou un arbre bifurcant,

- toutes les espèces de E étiquettent un sommet de T,

- toutes les feuilles de T sont étiquetées par E,

- tout noeud est étiqueté par une séquence de C.

Reconstructions phylogénétiques

Construire une phylogénie pour un ensemble d'espèces E connaissant les états des caractères de E, c'est trouver un arbre d'évolution pour E dont les feuilles sont les espèces avec des conditions données :

- arbre avec un nombre minimal de mutations,
- arbre avec un nombre maximal de caractères compatibles,
- arbre qui redonne au mieux les distances originales...

Comme $R(n) > (n-1)!$ la recherche de tels arbres ne se fait pas "naturellement" par un algorithme polynomial.

Trois Grands Types de Problèmes en bioInformatique...

- les problèmes de parcimonie
  - . restreinte
  - . large
- les problèmes de compatibilité
  - . totale (ou "phylogénie parfaite")
  - . maximale
- les problèmes de taxonomie
  - . ultramétrique
  - . additive

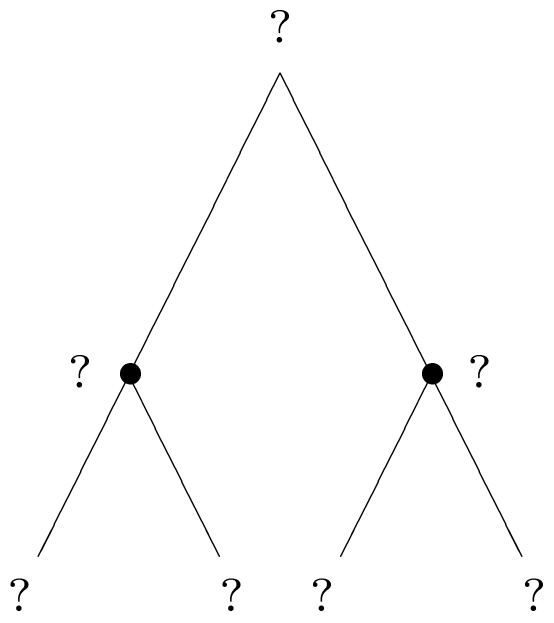
correspondant à des problèmes NP-complets classiques

- Parcimonie (*Steiner*)
- Phylogénie parfaite (*Ensembles indépendants*)
- Compatibilité (*Clique*)
- Taxonomie numérique (*Steiner*)

les problèmes de Parcimonie

la parcimonie en général

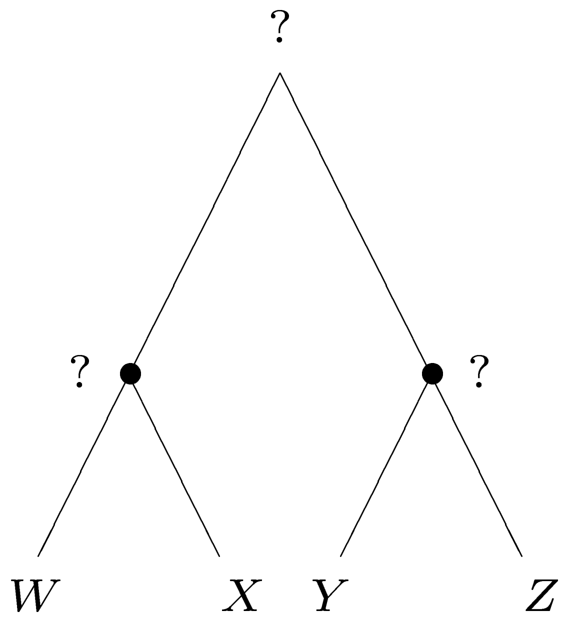
W	A	C	G	T
X	A	C	C	T
Y	A	C	C	G
Z	C	C	G	T



Trouver une topologie d'arbre d'évolution et les séquences des noeuds pour un score (longueur de parcimonie) minimale.

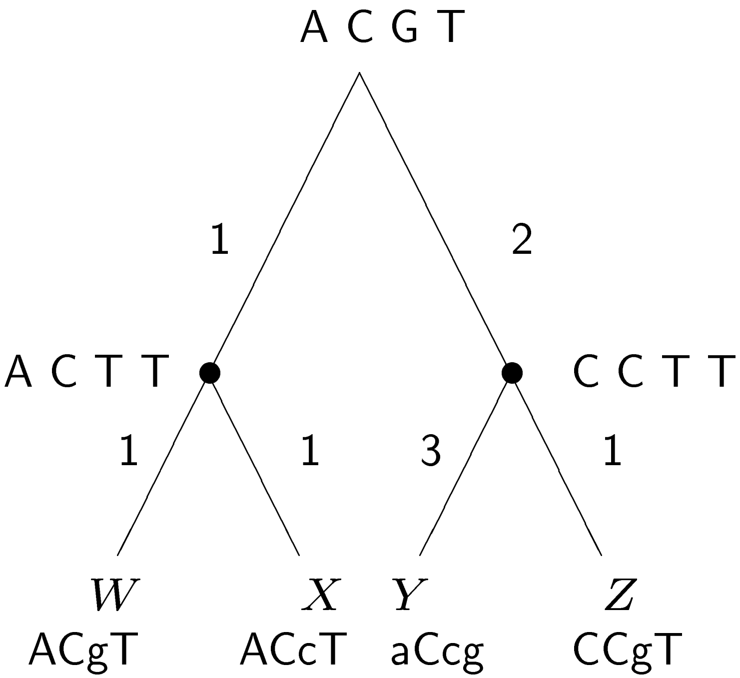
Problème de parcimonie restreinte (1/2)

W	A	C	G	T
X	A	C	C	T
Y	A	C	C	G
Z	C	C	G	T

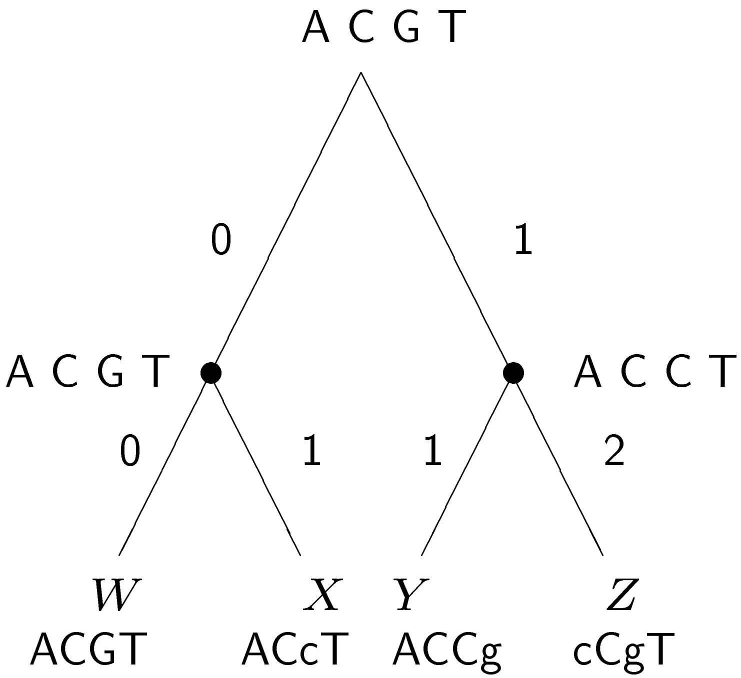


On connait la topologie d'arbre, il faut trouver les séquences des noeuds pour un score (longueur de parcimonie) minimale.

Problème de parcimonie restreinte (2/2)

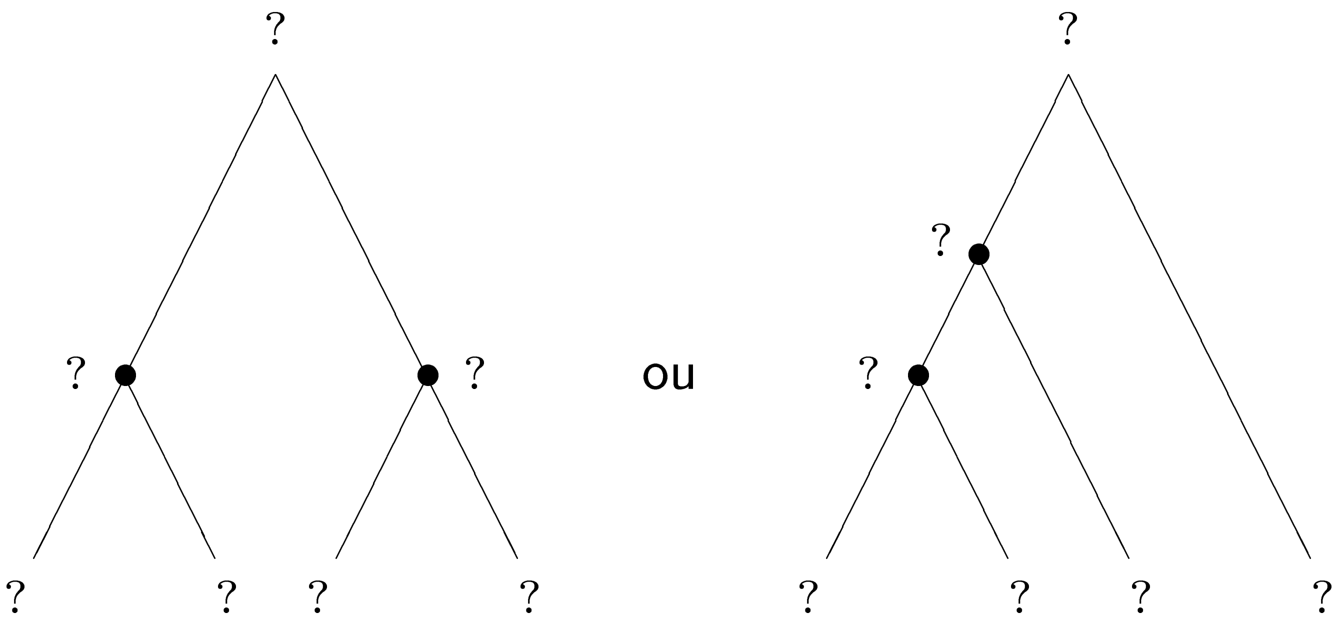


score 9



score 5

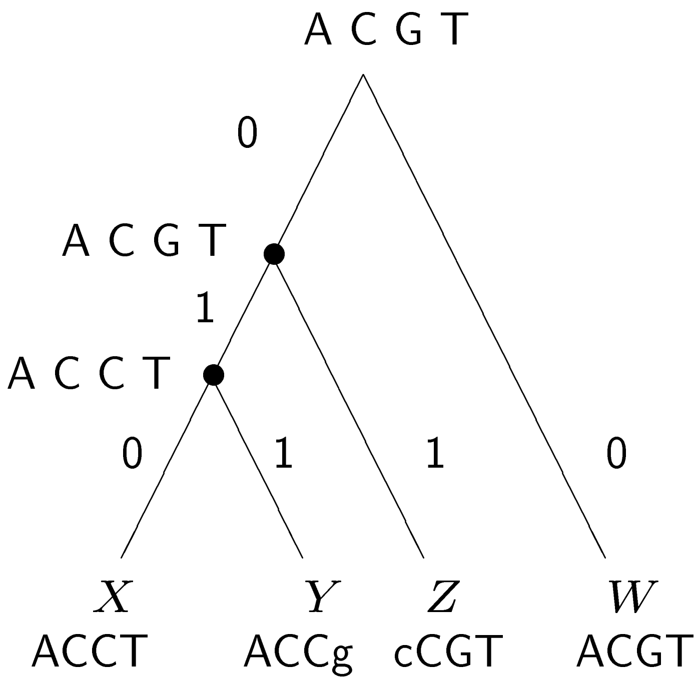
Problème de parcimonie large (1/2)



quelle topologie d'arbre ? quelles séquences pour les noeuds ?

Problème de parcimonie large (2/2)

W	A C G T
X	A C C T
Y	A C C G
Z	C C G T



$3 =$ meilleur score (longueur de parcimonie) toutes topologies, toutes séquences pour les noeuds confonfues : arbre de parcimonie [maximale].

Complexité des problèmes de parcimonie

- le problème de parcimonie restreinte est polynomial.
  (algorithme de Fitch en $O(mnk)$)

- le problème de parcimonie large est NP-complet
  même pour des séquences binaires.

  mais il en existe une 2-approximation : la longueur de parcimonie d'un arbre couvrant minimal sur le graphe complet sous-jacent est au plus le double de l'arbre de meilleure parcimonie.

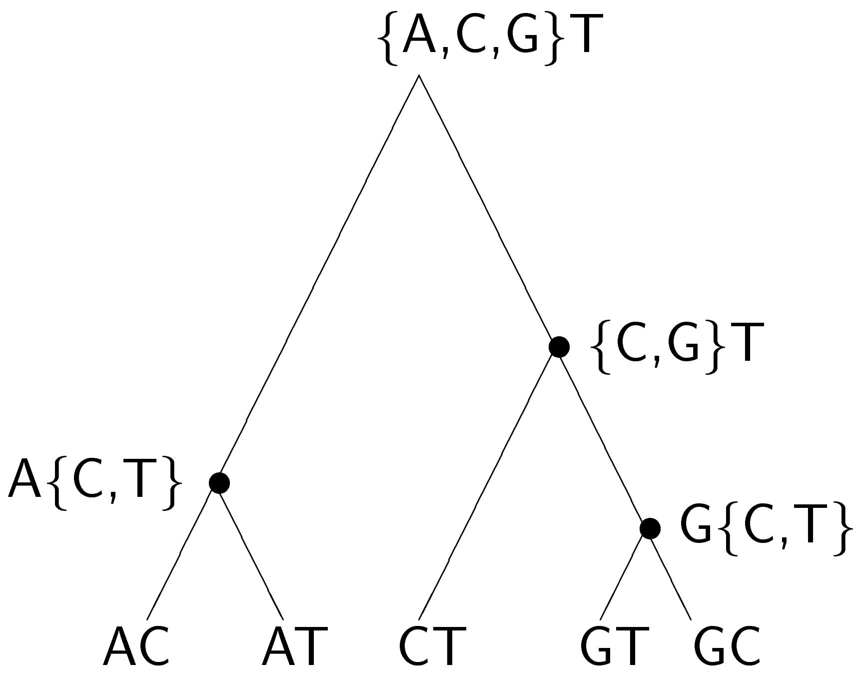
Algorithme de $Fitch$ 1/3 (polynomial)

Etape 1 / 2 : on va des feuilles à la racine pour déterminer les ensembles possibles d'états du caractère par $intersection$ $forcée$:

$$S_i \overline{\cap} S_j = S_i \cap S_j \text{ si } S_i \cap S_j \neq \emptyset,\ S_i \cup S_j \text{ sinon.}$$

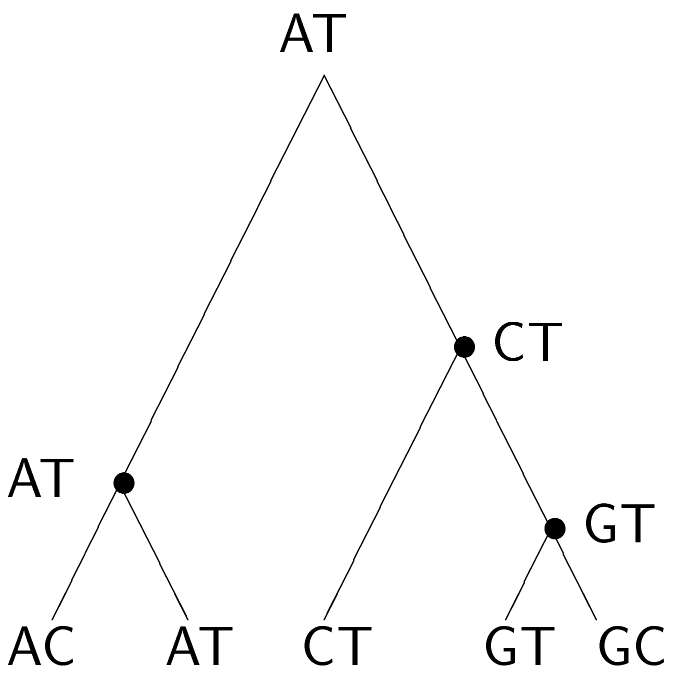
Etape 2 / 2 : on va de la racine aux feuilles à en sélectionnant dans les ensembles possibles les valeurs présentes dans les successeurs.

Algorithme de *Fitch* 2/3



Etape 1 / 2 : des feuilles à la racine

Algorithme de *Fitch* 3/3



Etape 2 / 2 : de la racine aux feuilles

Problèmes généraux de parcimonie (1/2)

On peut interdire certaines mutations comme

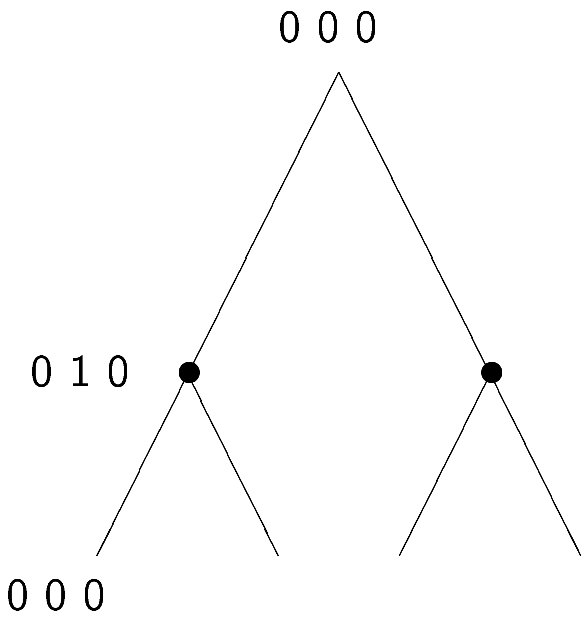
- la réversion vers un état antérieur
- la convergence ou mutation parallèle

L'arbre de parcimonie général (arbre de $Wagner$) associé peut alors être contraint :

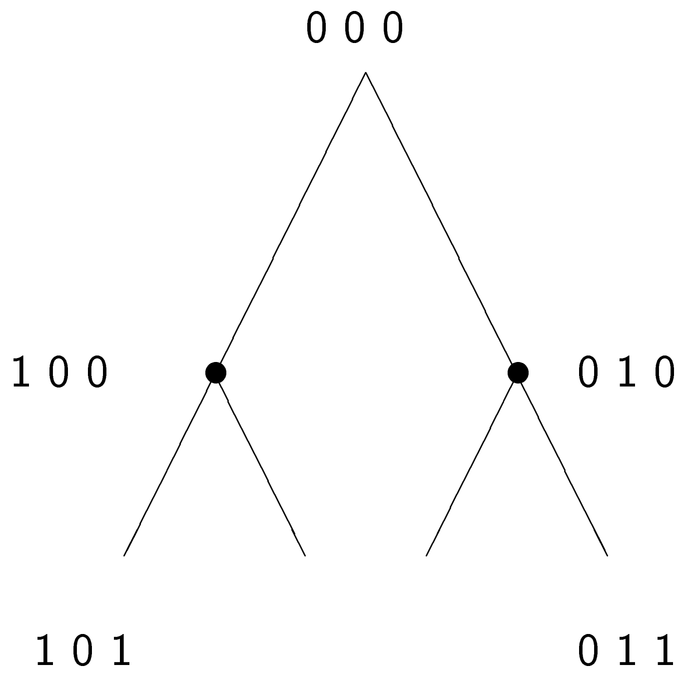
- à être un arbre sans réversion
  (parcimonie au sens de Camin-Sokal)

- à être un arbre sans convergence
  (parcimonie au sens de Dollo-Le Quesne)

Mutations interdites en parcimonie ordonnée



réversion du caractère 2



convergence pour le caractère 3

Problèmes généraux de parcimonie (2/3)

Il existe une version $pondérée$ du problème où on se donne des couts de transition entre états.

L'algorithme de $Sankoff$ qui généralise celui de $Fitch$ est alors utilisé.

Cet algorithme permet aussi de traiter, par le choix des poids, les restrictions apportées à la parcimonie.

Algorithme de Sankoff

On se donne une matrice de cout $c_{ij,}$ non nécessairement symétrique et des couts initiaux s_i^L pour chaque feuille L ; pour l'ADN on utilise à 0 si le nucléotide est présent et ont met ∞ s'il est absent.

Le cout s_i^k de l'état i pour le noeud k composé des successeurs a et b est défini par la formule

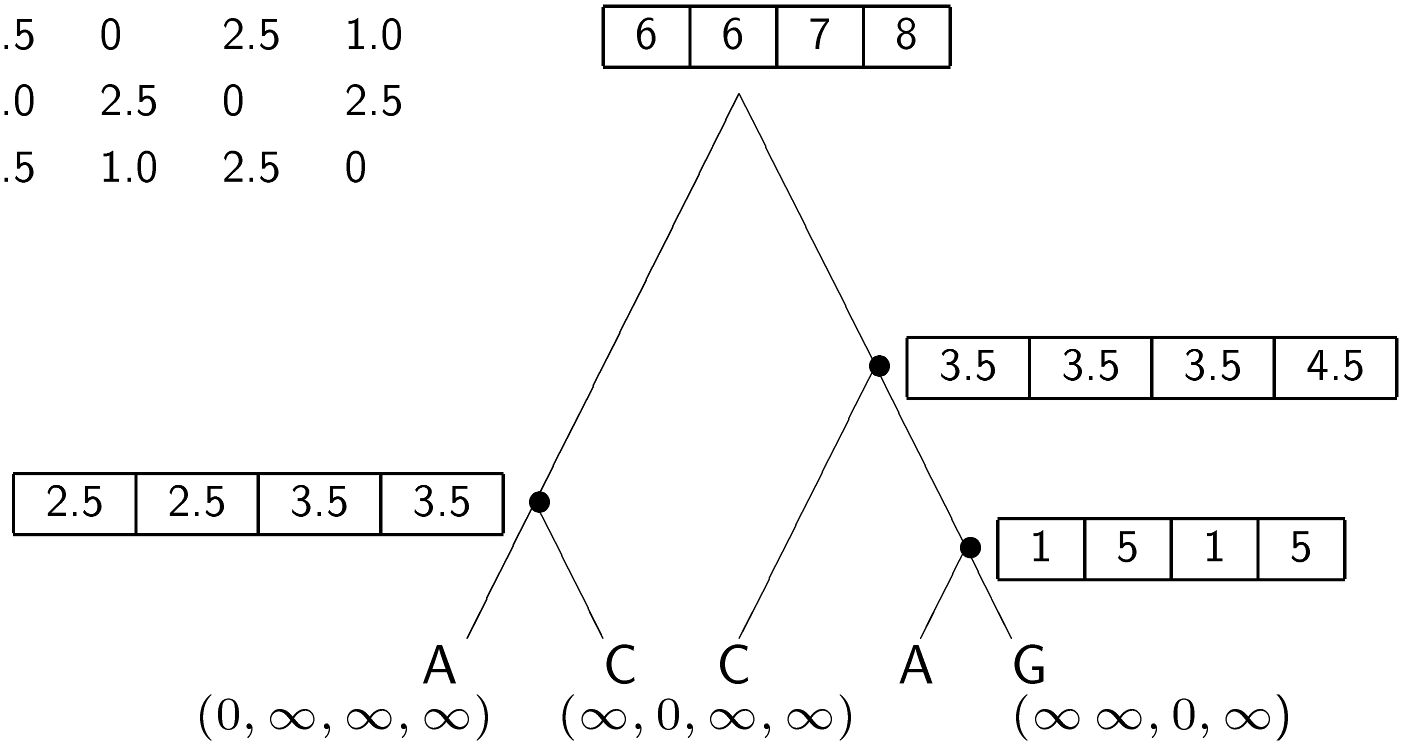
$$s_i^k = min(\ c_{i,m} + s_m^a\ ) + min(\ c_{i,n} + s_n^b)$$

où n et m parcourent tous les états possibles.

Le meilleur arbre est alors celui de cout minimal.

l'algorithme de Sankoff par l'exemple

	A	C	G	T
A	0	2.5	1.0	2.5
C	2.5	0	2.5	1.0
G	1.0	2.5	0	2.5
T	2.5	1.0	2.5	0



Les couts sont dans l'ordre

A	C	G	T

Problèmes généraux de parcimonie (3/3)

Le problème de parcimonie large est un cas particulier du problème de $Steiner$ pour les graphes où, pour n points donnés dans un espace métrique, il faut trouver un arbre nommé arbre de $Steiner$ sur ces points qui minimise la somme des distances entre les points.

Pour ce problème, l'ajout de $noeuds$ internes (ou "points de $Steiner$") est autorisé, ce qui le distingue du problème de l'arbre couvrant minimal.

quelques Heuristiques pour la parcimonie large

BB	Branch and Bound
NNI	Nearest Neighbour Interchange
SPR	Subtree Pruning and Regrafting
TBR	Tree Bisection and Reconnection
PHA	Phylip algorithms

quelques Articles de référence (1/2)

Première parution de l'algorithme de $Fitch$

W. Fitch. Toward defining the course of evolution:
Minimum change for specific tree topology.
Systematic Zoology, 20:406-416, 1971.

Trouver l'arbre le plus parcimonieux est NP-complet :

L. R. Foulds, R. Graham.
The Steiner problem in phylogeny is NP-complete.
Adv. Appl. Math., 3:43-49, 1982.

quelques Articles de référence (2/2)

Premiers algorithmes d'approximation de l'arbre le plus parcimonieux (basés sur les heuristiques de l'arbre de Steiner couvrant minimal)

L. R. Foulds, M. D. Hendy, D. Penny.
A graph theoretic approach to the development of minimal phylogenetic trees.
J. Mol. Evol., 13:127-149, 1979.

Premiers algorithmes en "branch and bound"

M. D. Hendy, D. Penny.
Branch and bound algorithms to determine minimal evolutionary trees.
Math. Biosci., 59:277-290, 1982.

les problèmes de Compatibilité

(dont la phylogénie parfaite)

Compatibilité entre des caractères et un arbre

l'état k de c_j est compatible avec T si toutes les espèces s telles que $c_j(s) = k$ sont dans un même sous-arbre.

c_j est compatible (ou "convexe") avec T si tous ses états sont compatibles avec T.

C est compatible avec T ssi tous ses caractères sont compatibles avec T.

Remarque : dans le cas binaire, si C est compatible avec T alors les c_j sont des étiquettes pour les arcs de T qui sont adjacents à la valeur 1 du caractère.

problèmes généraux de compatibilité

Pour une matrice d'états $M(S, C)$

- le problème de compatibilité totale ou *phylogénie parfaite* est de trouver un arbre d'évolution avec lequel tous les caractères de C sont compatibles,

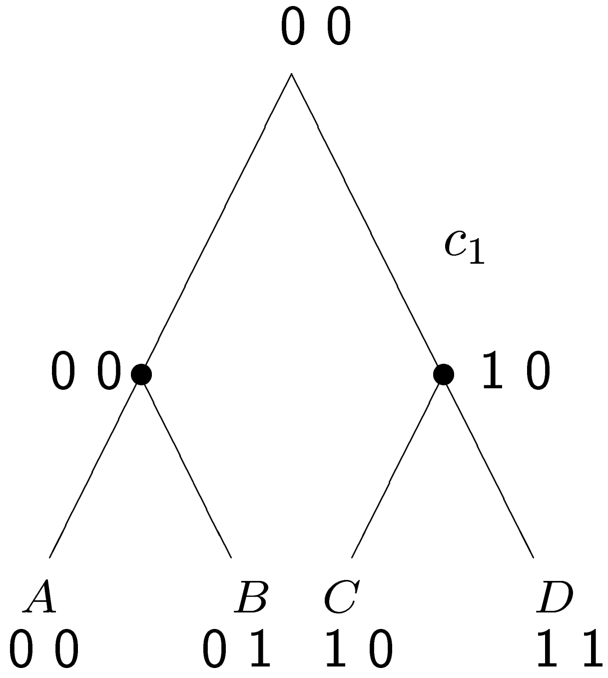
- le problème de *compatibilité large* (ou "maximale") est de trouver un sous-ensemble maximal de caractères de C pour lesquels il existe un arbre de phylogénie parfaite.

On distingue les caractères binaires du cas général.

On distingue les caractères ordonnés du cas général.

Exemples de compatibilité (1/3)

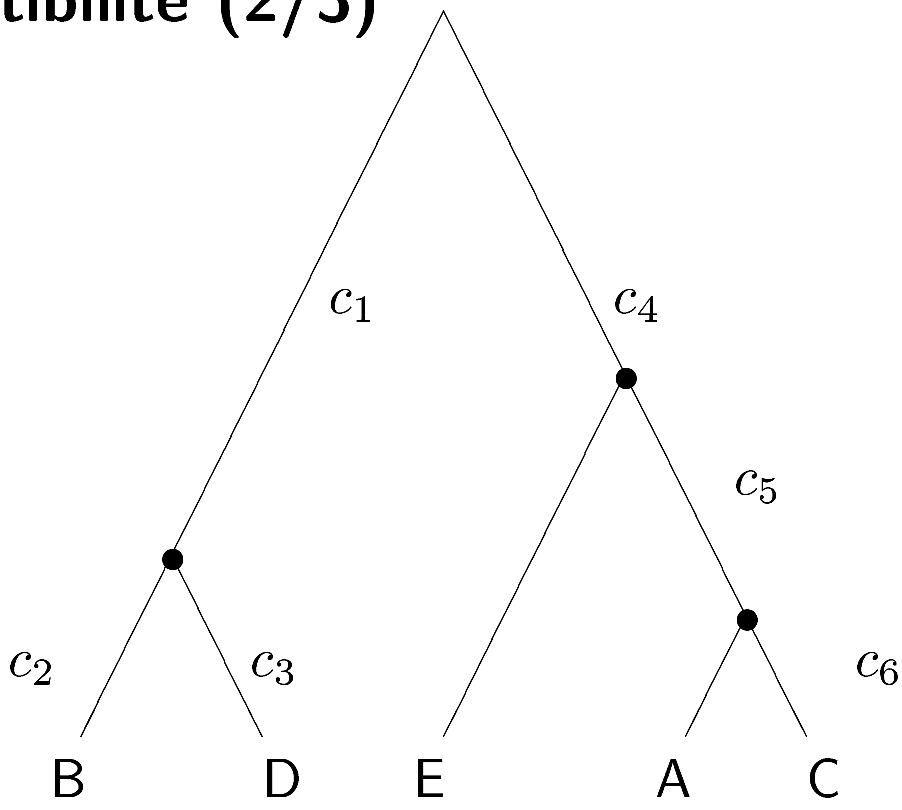
	c_1	c_2
A	0	0
B	0	1
C	1	0
D	1	1



c_1 est compatible avec T mais c_2 n'est pas compatible avec T.

Exemples de compatibilité (2/3)

	c_1	c_2	c_3	c_4	c_5	c_6
A	0	0	0	1	1	0
B	1	1	0	0	0	0
C	0	0	0	1	1	1
D	1	0	1	0	0	0
E	0	0	0	1	0	0

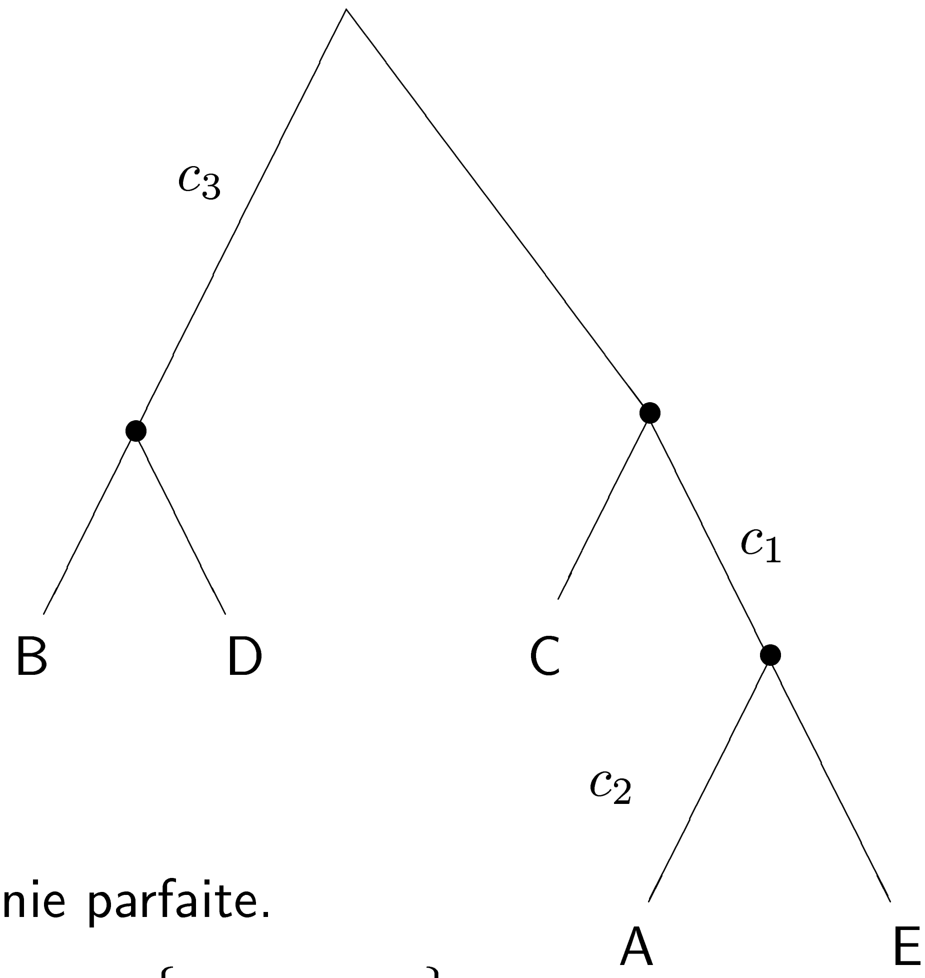


La compatibilité est totale : la phylogénie est parfaite.

L'arc afférent à E peut être étiqueté $c_4 \wedge \sim c_5$.

Exemples de compatibilité (3/3)

	c_1	c_2	c_3	c_4
A	1	1	0	1
B	0	0	1	0
C	0	0	0	1
D	0	0	1	1
E	1	0	0	1



Il ne peut pas y avoir phylogénie parfaite.

La compatibilité est maximale pour $\{ c_1, c_2, c_3 \}$.

Phylogénie parfaite binaire (1/4)

Soit M une matrice de n espèces et m caractères binaires.

Il existe un algorithme polynomial intuitif en $O(nm^2)$ optimisé en $O(nm)$ pour tester si M admet un arbre phylogénétique parfait.

Lorsqu'il existe un tel arbre, l'algorithme de construction de cet arbre s'exécute en $O(nm)$ par "*radix sort*" et "*last index method*".

Phylogénie parfaite binaire (2/4)

Des ensembles H_i des ensembles sont hiérarchiquement compatibles ssi $H_i \cap H_j \in \{ \emptyset, H_i, H_j \}$.

L'orbite O_i du $i-$ème caractère [binaire] est l'ensemble des espèces qui valent 1 pour ce caractère.

Si les orbites des caractères sont hiérarchiquement compatibles alors il existe un arbre phylogénétique parfait.

Phylogénie parfaite binaire (3/4)

	c_1	c_2	c_3	c_4
A	1	1	0	1
B	0	0	1	0
C	0	0	0	1
D	0	0	1	1
E	1	0	0	1

$$\begin{array}{lcl} O_1 & = & \{\ A,\ E\ \} \\ O_2 & = & \{\ B\ \} \\ O_3 & = & \{\ B,\ D\ \} \\ O_4 & = & \{\ A,\ C,\ D,\ E\ \} \end{array}$$

O_3 et O_4 ne sont pas hiérarchiquement compatibles.

$\{c_1, c_2, c_3, c_4\}$ n'admet pas d'arbre phylogénétique parfait.

$\{c_1, c_2, c_3\}$ admet un arbre phylogénétique parfait.

Phylogénie parfaite à deux caractères non-binaires

Une matrice d'états dont les m caractères définissent un coloriage c admet un un arbre phylogénétique parfait ssi le graphe d'intersection d'états peut être $c-$triangulé.

Une matrice d'états pour n espèces dont les deux caractères définissent un coloriage admet un arbre phylogénétique parfait ssi le graphe d'intersection d'états est acyclique.

Il existe un algorithme en O(n) pour décider de l'acyclicité.

L'algorithme de construction de l'arbre parfait s'exécute alors en O(n).

Phylogénie parfaite générale

Le problème de phylogénie parfaite générale $P(k, m, r)$ avec k espèces, m caractères et r états au plus :

- polynomial en km pour $P(k, m, 2)$

- polynomial en k pour $P(k, 2, r)$

- NP-complet pour $m > 2$ et $r > 2$

Heuristiques pour la phylogénie parfaite

Réorganisation de Voisinage

`NNI`	Nearest Neighbour Interchange
`SSB`	Single Step branching
`SPR`	Subtree Pruning and Regrafting

Recherche locale

`II`	Iterative Improvement (`NNI`, `SSB`, `SPR`)
`VND`	Variable Neighborhood Descent

Métaheuristiques

`GRASP`	Greedy Randomized Adaptive Search Procedure
`VNS`	Variable Neighborhood Search
`SPRTS`	Tabu Search for `SPR`

quelques Articles de référence (1/3)

Premier algorithme polynomial pour la phylogénie parfaite binaire ($r = 2$)

D. Gusfield. Efficient algorithms for inferring evolutionary trees. Networks, 21:19-28, 1991.

La phylogénie parfaite générale ($r > 2$) est NP-complète

H. Bodlaender, M. Fellows, T. Warnow.
Two strikes against perfect phyloeny.
Proc. 19-th Int. Coll. Automata, Lang. and Program.
pages 273-283. Lecture Notes Comp. Sci., 1992.

M. A. Steel. The complexity of reconstructing trees from qualitative characters and subtrees.
J. Classification, 9:91-116, 1992.

quelques Articles de référence (2/3)

Un algorithme en $O(km^2)$ pour $r = 3$

A. Dress, M. Steel. Convex tree realizations. AML, 5:3-6, 1992.

La phylogénie parfaite est équivalente à la coloration de graphes triangulés d'où un algorithme en $O(\ (rm)^{m+1} + km^2)$

F. R. McMorris, T. Warnow, T. Wimer.
Triangulating vertex-colored graphs.
SIAM J. Discr. Math., 7:296-306, 1993.

Un algorithme en $O(k^2m)$ pour $r \leqslant 4$ donc pour l'ADN

S. K. Kannan, T. Warnow.
Inferring evolutionary history from DNA sequences.
SIAM J. Comput., 23(4):713-737, 1994.

quelques Articles de référence (3/3)

Programmation dynamique et algorithme en $O(2^{2r}km^2)$

R. Agarwala, D. Fernandez-Baca.
A polynomial time algorithm for the perfect
phylogeny problem when the number of character states is fixed.
Proc. 34-th Annual IEEE Symp. Found. Comput. Sci., 1993.

S. Kannan, T. Warnow.
A fast algorithm for the computation and enumeration
of perfect phylogenies when the number of character states is fixed.
pages 595-603, 1995.

les problèmes de Taxonomie

Propriétés métriques

Soit f une fonction définie sur $E \times E$ et $q_f(a,b,c,d) = f(a,b) + f(c,d)$.

	f est	si et seulement si f vérife
C1	d-nulle	$f(x,x) = 0$
C2	symétrique	$f(x,y) = f(y,x)$
C3	d-triangulaire	$f(x,y) \leqslant f(x,z) + f(z,y)$
C4	séparée	$f(x,y) = 0 \Rightarrow y = x$
C5	sous-maximale	$f(x,y) \leqslant max(\ f(x,z), f(y,z)\ )$
C6	buneman	$q_f(x,y,z,t) \leqslant max(\ q_f(x,z,y,t), q_f(x,t,y,z)\ )$

La propriété de $Buneman$ est également nommée "condition des 4 points".

Dissimilarités, écarts, distances, ultramétriques

nom	conditions	remarques
indice de dissimilarité	C1 C2	
pseudo-métrique (ou écart)	C1 C2 C3	
métrique (ou distance)	C1 C2 C3 C4	
ultramétrique	C1 C2 C3 C5	ce qui implique C4
métrique d'arbres	C1 C2 C3 C4 C6	ou métrique additive

On dit que f vérifie la "condition des 3 points" ssi

$$f(x,z) = f(y,z) = c \Rightarrow f(x,y) \leqslant c$$

Matrices de distances

Une "matrice de distances" M est une matrice de réels positifs ou nuls.

Une "vraie matrice de distances" M est une matrice telle que f_M : (i,j) $\rightarrow M_{i,j}$ est une distance.

Définition 1 : M est additive ssi f_M est additive.

Définition 2 : M est ultramétrique ssi f_M est ultramétrique.

Propriétés élémentaires

Théorème 1 :

f est additive
ssi f vérifie la condition de $Buneman$.

Théorème 2 :

f est ultramétrique
ssi f vérifie la condition des trois points.

Théorème 3 :

si f est ultramétrique alors f est additive.

Exemples de matrices

Matrice 1	*Matrice 2*	*Matrice 3*	*Matrice 4*	*Matrice 5*

`  A B C D`
`A 0`
`B 8 0`
`C 8 2 0`
`D 0 2 3 0`
	`  A B C D`
	`A 0`
	`B 8 0`
	`C 8 2 0`
	`D 1 2 3 0`
		`  A B C D`
		`A 0`
		`B 2 0`
		`C 4 2 0`
		`D 1 2 3 0`
			`  A B C D`
			`A 0`
			`B 11 0`
			`C 10 3 0`
			`D 9 12 11 0`
				`  A B C D`
				`A 0`
				`B 8 0`
				`C 8 2 0`
				`D 14 14 14 0`

n'induit pas une distance	n'induit pas une distance	distance ni A ni U	distance A non U	distance U donc A

0 $= M_1(4,1)$

8>2+1:$M_2(2,1) > M_2(2,4) + M_2(4,1)$

6>max(5,3):$M_3(2,4)+M_3(3,1) > max(M_3(4,3)+M_3(2,1), M_3(4,1)+M_3(2,3)$

11 $>$ max(9, 10) : $M_4(3,4) > max(\ M_4(3,1)\ ,\ M_3(4,1)\ )$

Distance d'arbre (1/2)

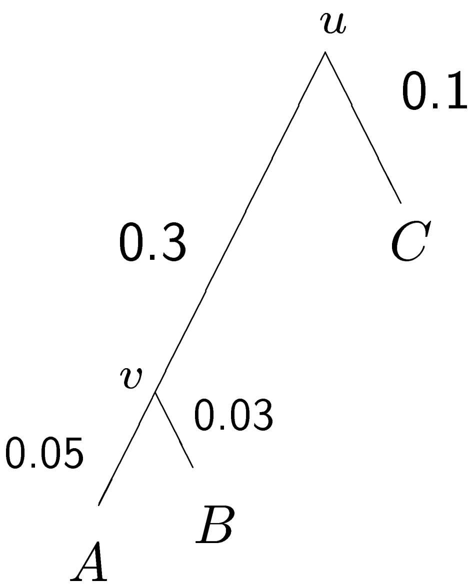
Rappel :

si T est un arbre binaire enraciné ou bifurcant, il existe un unique chemin qui relie deux sommets donnés.

Définition :

Si T est un arbre pondéré, la fonction d_T qui associe à deux sommets la somme des poids des arcs pour l'unique chemin qui les relie est nommée "distance d'arbre" induite par T.

Distance d'arbre (2/2)



d_T	A	B
B	0.08	
C	0.45	0.43

$$d_T(A,C) = d_T(A,v) + d_T(v,u) + d_T(u,C) = 0.05 + 0.3 + 0.1 = 0.45$$

le problème de taxonomie

Soit M une matrice de distances M pour des séquences S. Trouver un arbre T d'évolution de S tel que $d_T = f_M$.

Si ce n'est pas possible, trouver T "au mieux" c'est à dire tel que $||M - d_T||$ soit minimale par exemple pour la norme $\mathcal{L}_\infty$.

$Variantes$

- trouver U, matrice ultramétrique avec $||M - U||_{\mathcal{L}_\infty}$ minimale.

- trouver A, matrice additive avec $||M - A||_{\mathcal{L}_\infty}$ minimale.

Reconstruction ultramétrique

Théorème

si M est ultramétrique, il existe un arbre unique T tel que $d_T = f_M$.

T est alors fourni par l'algorithme UPGMA qui s'exécute en $O(n^2)$ c'est à dire en prenant comme critère d'agrégation la distance minimale et comme formule de recalcul entre k et $\{i, j\}$ la moyenne

$$d(k, ij) = \frac{n_i\, d(k, i) + n_j\, d(k, j)}{n_i + n_j}$$

Les successeurs de a sont alors équidistants de a et la distance de la racine aux feuilles est constante.

Exemple de Reconstruction ultramétrique

	A	B	C	D	E
A	0				
B	8	0			
C	4	8	0		
D	6	8	6	0	
E	8	4	8	8	0

	A	BE	C	D
A	0			
BE	8	0		
C	4	8	0	
D	6	8	6	0

	AC	BE	D
AC	0		
BE	8	0	
D	6	8	0

	ACD	BE
ACD	0	
BE	8	0

1. on regroupe B et E au niveau 4
   donc $d(B, BE) = d(E, BE) = d(B, E)/2 = 2$.
   et $d(A, BE) = (1.d(A, B) + 1.d(A, E))/2 = 8$
2. on regroupe A et C au niveau 4.
3. on regroupe AC et D au niveau 6.
4. on regroupe ACD et BE au niveau 8.

Reconstruction additive

Théorème

si M est additive, il existe un arbre unique T
tel que $d_T = f_M$.

Rem. T est fourni par l'algorithme NJ qui s'exécute en $O(n^2)$ c'est à dire en prenant comme critère d'agrégation la distance moyenne recentrée minimale et comme formule de recalcul entre k et $\{i,j\}$ l'expression

$$d(k,ij) = \frac{d(k,i) + d(k,j) - d(i,j)}{2}$$

La distance entre a et ses successeurs est alors proportionnelle à la somme des distances entre le successeur et les autres noeuds.

Reconstruction générale

Les "matrices de distances" n'en sont pas en général !

Par exemple $p = d(a, b) =$ pourcentage de différents n'est ni additif, ni ultramétrique, pas plus que sa correction de Poisson

$$p^* = -ln(1 - p)$$

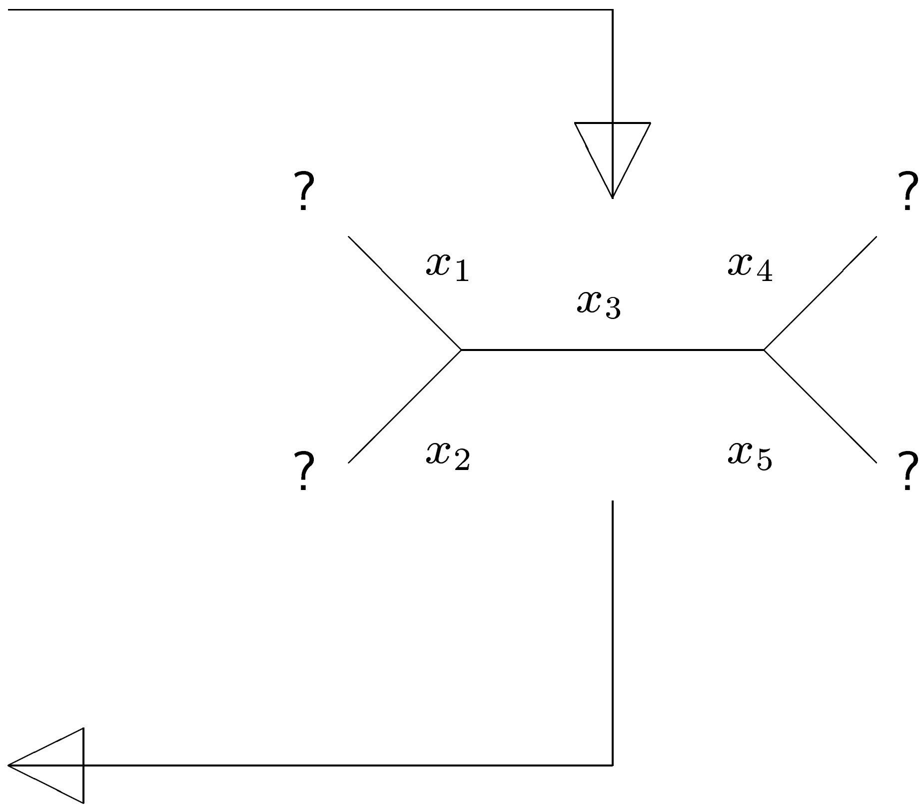
On utilise alors des méthodes comme

NJ (en priant le Seigneur !)

LSM (moindre carrés) et autres méthodes pour la métrique $\mathcal{L}_\infty$

exemple de Reconstruction générale (1/2)

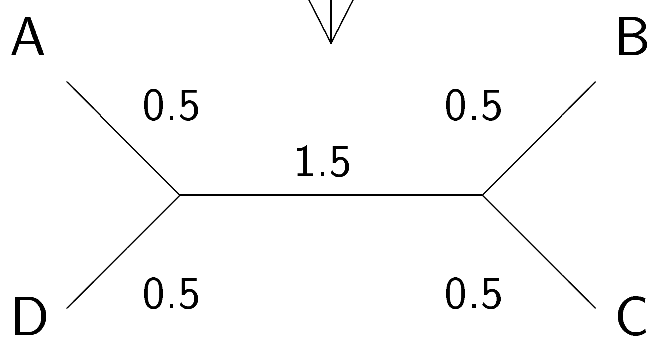
	A	B	C	D
A	0			
B	3	0		
C	2	1	0	
D	1	2	3	0



	A	B	C	D
A	0			
B	?	0		
C	?	?	0	
D	?	?	?	0

exemple de Reconstruction générale (2/2)

	A	B	C	D
A	0			
B	3	0		
C	2	1	0	
D	1	2	3	0



	A	B	C	D
A	0			
B	2.5	0		
C	2.5	1.0	0	
D	1.0	2.5	2.5	0

quelques Articles de référence (1/3)

Méthodes classiques de classification dont UPGMA

P. H. Sneath, R. R. Sokal.
Numerical Taxonomy.
an Francisco: Freeman, 1963.

L'article original sur NJ

N. Saitou, M. Nei. The neighbor-joining method:
A new method for reconstructing phylogenetic trees.
Mol. Biol. Evol., 4:406-425, 1987.

quelques Articles de référence (2/3)

Caractérisation des matrices additives

P. Bunemann. The recovery of trees from measures of dissimilarity. Mathematics in the archaeological and historical sciences, F. Hodson, D. Kendall, and P. Tautu, editors, pages 387-395. Edinburgh University Press, 1971.

Construction algorithmique des matrices additives pour un arbre donné et unicité de telles matrices.

M. S. Waterman, T. F. Smith, M. Singh, W. A. Beyer. Additive evolutionary trees. J. Theoret. Biol., 64:199-213, 1977.

quelques Articles de référence (3/3)

Trouver une matrice [u.m., add.] proche d'une matrice donnée est NP-complet

M. Krivanek, J. Moravek.
NP-hard problems in hierarchical-tree clustering.
Acta Inform., 23:311-323, 1986.

W. H. Day.
Computational complexity of inferring phylogenies from dissimilarity matrices.
Bull. Math. Biol., 49:461-467, 1987.

Validation d'arbres

On utilise principalement deux techniques statistiques classiques : le *bootstrap* (Efron, 1979) et le *jackknife* (Quenouille, Tukey, 1977) dont le principe est

- on modifie un peu les séquences
  . rééchantillonnage
  . pseudoréplication

- on recalcule l'arbre

- on compare ou fusionne les arbres

Comparaison d'arbres

- Similarité : MAST Maximum Agreement Subtree
- Dissimilarité :
  . Robinson-Foulds Distance (+ algorithme de Day)

  . NNID Nearest Neighbor Interchange Distance

  . STTD Subtree Transfer Distance

Fusion d'arbres

- techniques de consensus

  (si ce sont les mêmes espèces initiales)

  . strict ; majoritaire ; relaché / semi-strict / compatible
  . moyen ; local ; glouton ; Nelson-Page / médiane asymétrique
  . Adam ; "prune and graft" ; MRP ; Buneman

- techniques de "super abres"

  (pour des espèces différentes avec des arbres emboitants)

Non évoqués

- les hypothèses biologiques sous-jacentes,

- les indices d'adéquation entre arbre et séquences,

- l'aspect des paysages de recherche,

- les modèles probabilistes,

- les algorithmes récents.

Articles de synthèse 1 / 2

J. Kim, T. Warnow.
Tutorial on Phylogenetic Tree Estimation.
Intelligent Systems for Molecular Biology, Heidelberg 1999.

S. Holmes.
Phylogenies: an overview.
In Statistics and Genetics, IMA (112), Springer 1999.

T. Warnow.
Some combinatorial optimization problems in phylogenetics,
in Graph Theory and Combinatorial Biology, L. Lovasz et al. (eds.),
Bolyai Society Mathematical Studies, Volume 7, Budapest, 1999, 363-413.

Articles de synthèse 2 / 2

L. Salter.
Algorithms for Phylogenetic Tree Reconstruction.
ICMETMBS Proc.,
Vol. 2, pgs. 459-465. 2000.

A. Andreatta, C. Ribeiro.
Heuristics for the phylogeny problem.
Journal of Heuristics 8, 2002, 429-447.

D. Bryant.
A classification of Consensus Methods for Phylogenetics.
Bioconsensus . DIMACS-AMS, 2003.

Vous avez échappé à... (1/2)

homologie	orthologie	paralogie	polytomie
apomorphie symplésiomorpie	autapomorphie synapomorphie	plésiomorphie homplasie	
dendrogramme	cladogramme	phylogramme	phénogramme
taxon	otus	clades	
outgroup	midpoint rooting	outlier	
newick	nexus	phylip	

mais aussi à... (2/2)

kishno-hasegawa test, shimodaira-hasegawa test

ML, Jukes-Cantor, Kimura, Delete-half Jackknife, Permutation Analysis

upgma/wpgma, sattah-tversky, split decomposition

ADN, ARN [16s], ADN mitochondrial

gene, codon, intron, exon

quartet cleaning, Q* de Buneman, Maximum Quartet Compatibility

en guise de Conclusion

Après avoir atteint un niveau **suffisant** d'efficacité grâce à l'informatique,il faudra **beaucoup** de biologie pour valider puis interpréter les résultats.