

mercredi 12 septembre 2007

programme de la matinée

LBA, de la parcimonie à la vraisemblance

La méthode de Maximum Likelihood

Comparaison entre les méthodes NJ, MP, ML

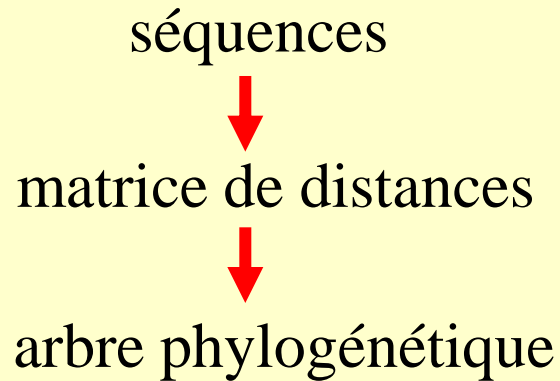
Introduction au programmes Treefinder

Phyml, Phylip, Raxml

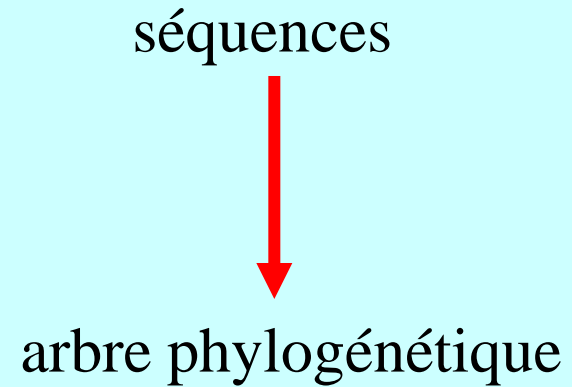
Exercices : analyses ML avec Treefinder

# Méthodes de reconstruction phylogénétique

## Méthodes de distances



## Méthodes de caractères



Maximum de Parcimonie

Méthodes Probabilistes

mercredi 12 septembre 2007

programme de la matinée

LBA, de la parcimonie à la vraisemblance

La méthode de Maximum Likelihood

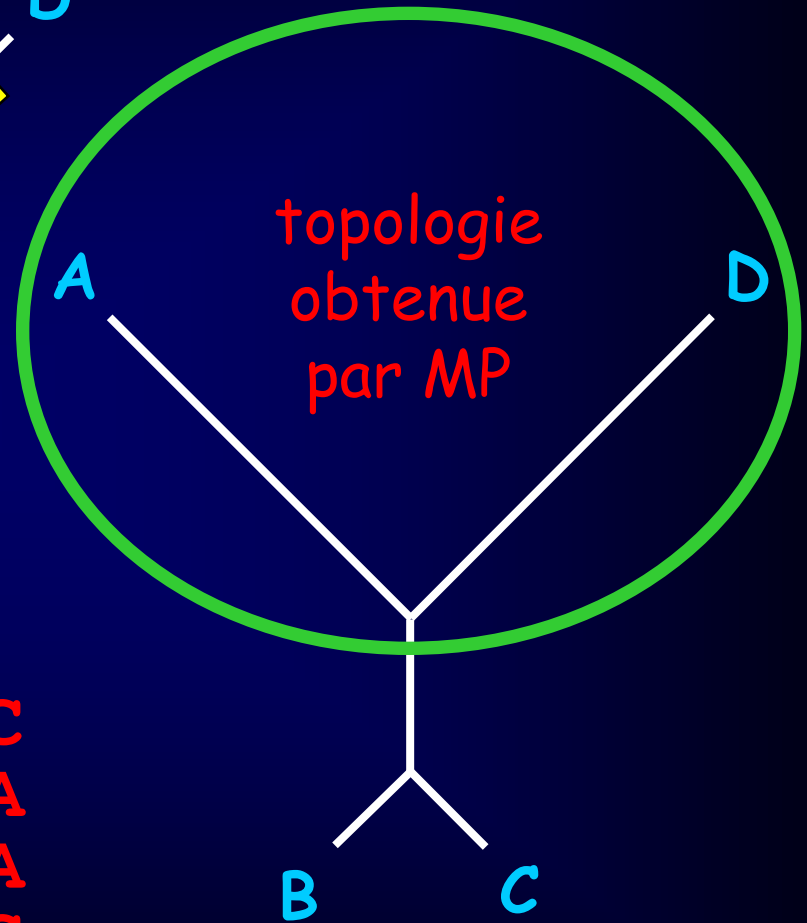
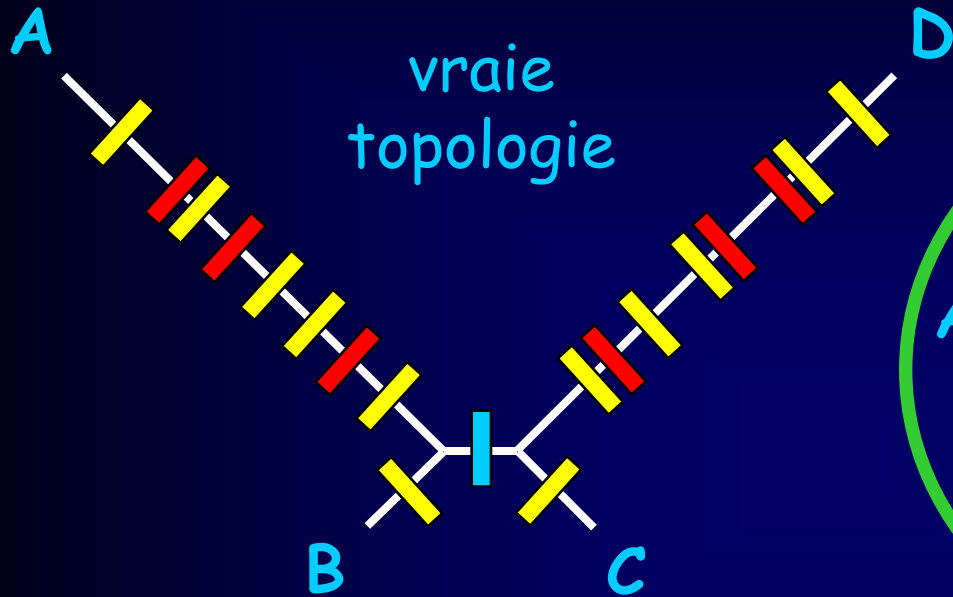
Comparaison entre les méthodes NJ, MP, ML

Introduction au programmes Treefinder

Phyml, Phylip, Raxml

Exercices : analyses ML avec Treefinder

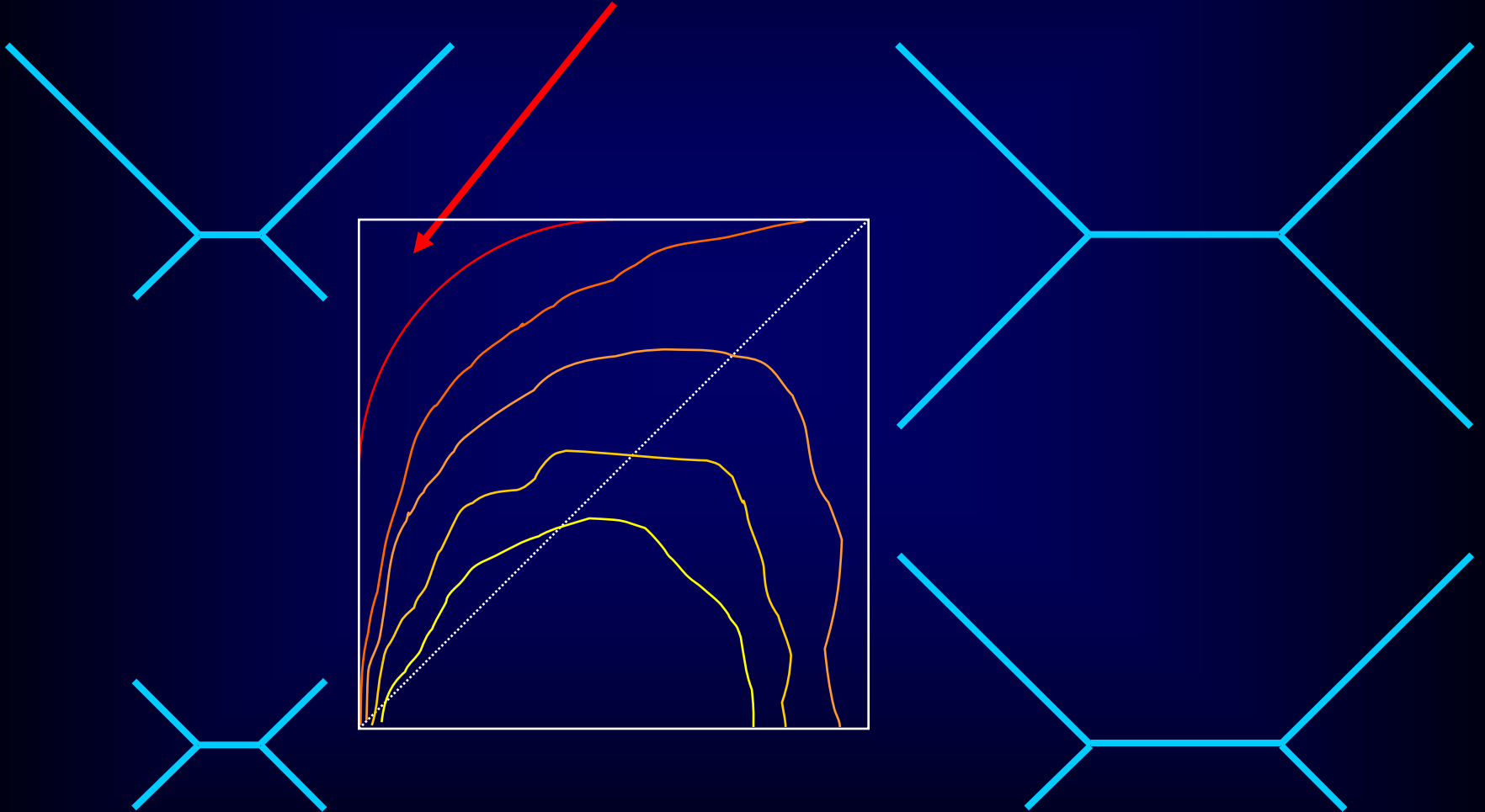
# LBA... c'est quoi?



A	A	A	A	G	A	T	C
B	A	C	A	A	A	A	A
C	A	T	A	T	A	A	A
D	A	T	A	A	A	C	C

# LBA, de la parcimonie à la vraisemblance

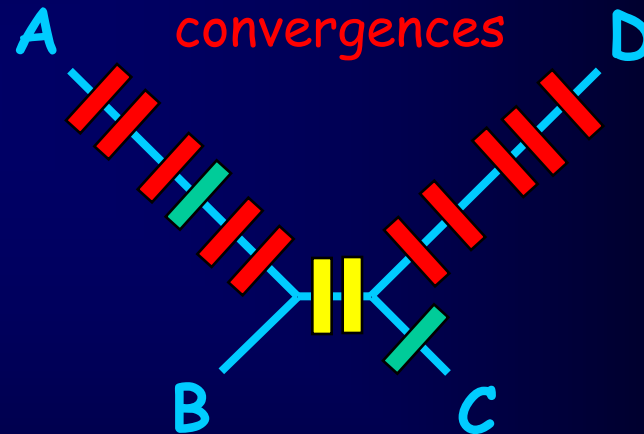
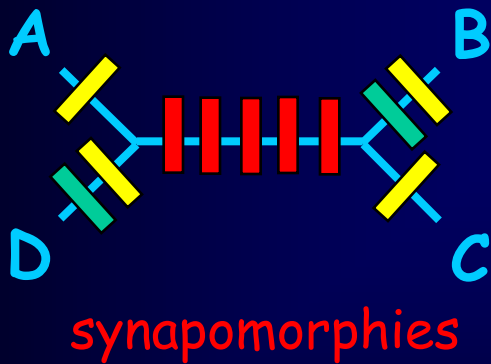
The « Felsenstein Zone »



# LBA, de la parcimonie à la vraisemblance

Lorsqu'il y a excès d'un des trois types de sites informatifs possibles (**rouge**), deux scénarios sont possibles :

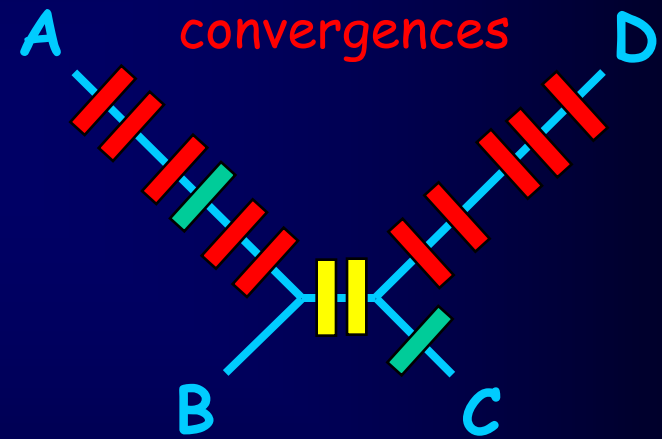
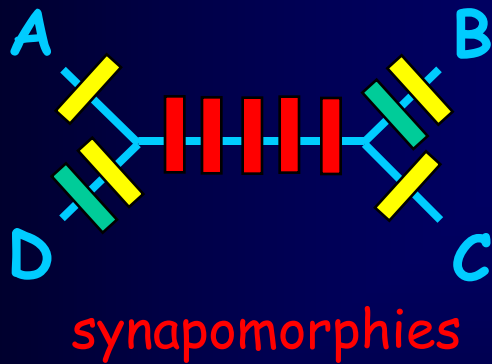
A	A	A	A	G	A	C	A	T
B	G	A	T	A	C	A	A	A
C	G	T	T	A	A	A	C	A
D	A	T	A	G	C	C	C	T



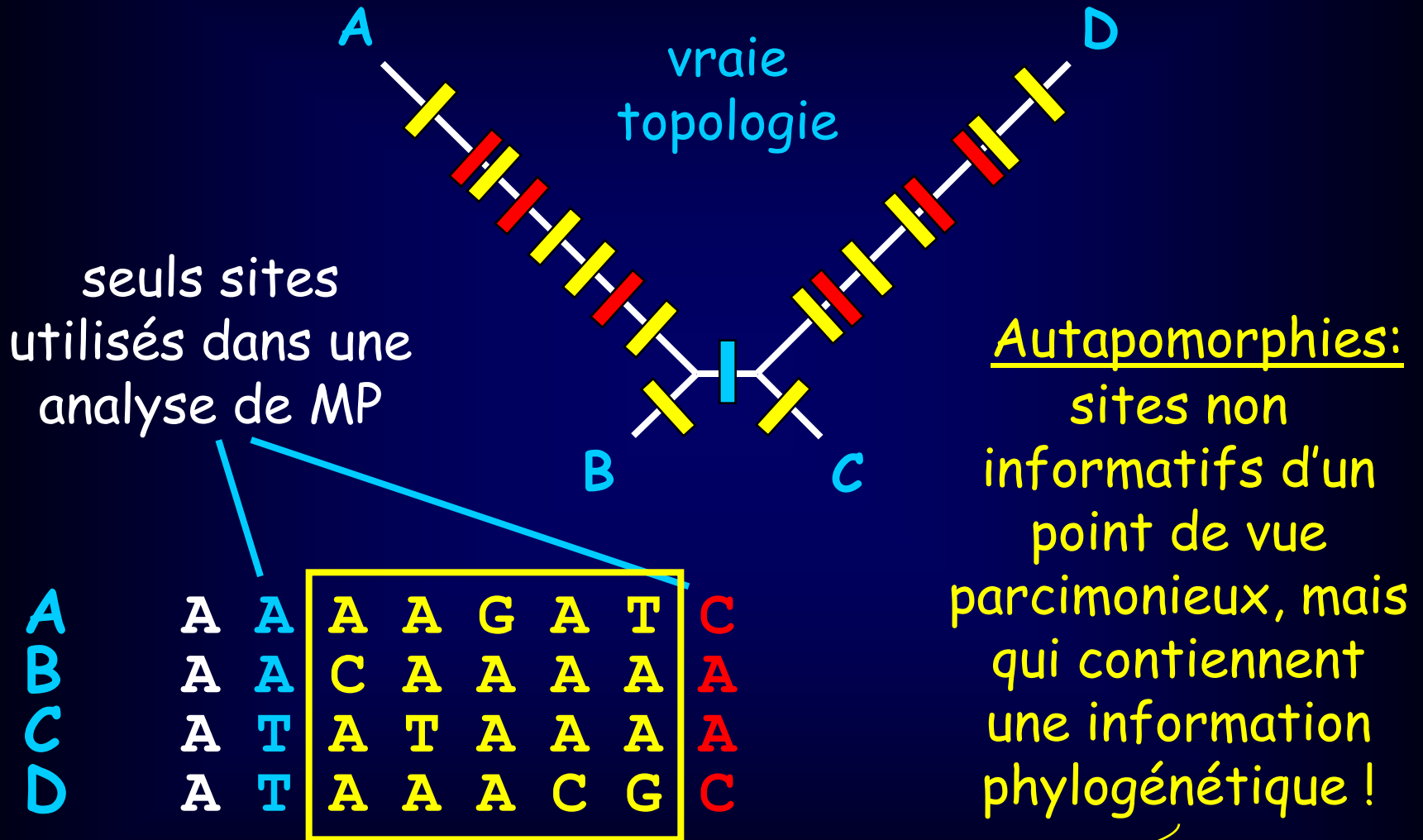
La méthode de MP retiendra automatiquement la topologie de gauche, qui implique moins de changements (synapomorphie = 1 changement alors que convergence = 2 changements)

# LBA, de la parcimonie à la vraisemblance

Comment faire la différence entre ces deux cas?



# LBA, de la parcimonie à la vraisemblance



# LBA, de la parcimonie à la vraisemblance

Comment se distribuent les sites autapomorphiques?

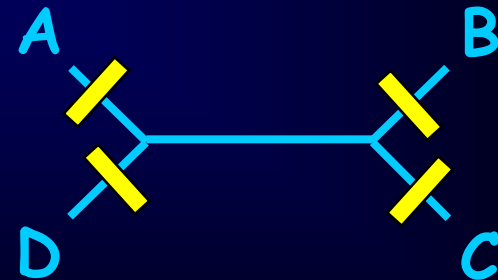
sites de type ((A,D),(B,C)) > sites de type ((A,B),(C,D))

A	A	A	A	A	G	A	T	C
B	A	A	C	A	A	A	A	A
C	A	T	A	T	A	A	A	A
D	A	T	A	A	A	C	G	C

└──────────────────┘

Cas 1 :

Les différents types de sites variables non informatifs ont plus ou moins tous la même fréquence



# LBA, de la parcimonie à la vraisemblance

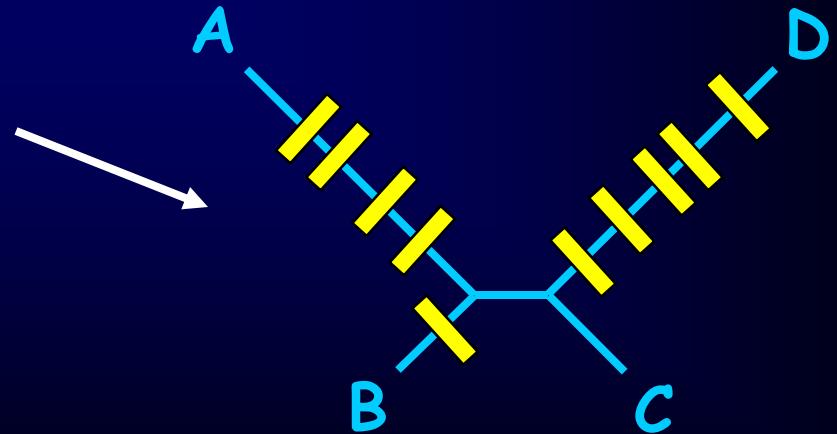
sites de type ((A,D),(B,C)) > sites de type ((A,B),(C,D))

A	A	A	A	A	G	A	T	C
B	A	A	C	A	A	A	A	A
C	A	T	A	T	A	A	A	A
D	A	T	A	A	A	C	G	C

}

Cas 2 :

Certains types de sites variables non informatifs sont nettement plus fréquents que les autres



## LBA, de la parcimonie à la vraisemblance

Les sites variables non informatifs d'un point de vue parcimonieux contiennent une information phylogénétique non négligeable, notamment sur la longueur des branches (et donc les taux de substitution).

### Hypothèse :

Un raisonnement **probabiliste** qui engloberait un modèle d'évolution, l'information phylogénétique détaillée de chaque site de l'alignement, les topologies possibles de l'arbre et la longueur de toutes ses branches devrait permettre de surpasser les résultats des analyses MP.

mercredi 14 septembre 2005

programme de la matinée

LBA, de la parcimonie à la vraisemblance

La méthode de Maximum Likelihood

Autres approches probabilistes

Comparaison entre les méthodes NJ, MP, ML

Introduction aux programmes Treefinder,  
PhyML, Phylip, Raxml

Exercices avec Treefinder

# La méthode de Maximum Likelihood

## Méthodes de reconstruction phylogénétique

### Méthodes de distances

séquences



matrice de distances



arbre phylogénétique

### Méthodes de caractères

séquences



arbre phylogénétique

Maximum de Parcimonie

Méthodes Probabilistes

# La méthode de Maximum Likelihood

La méthode de **Maximum Likelihood** (Felsenstein 1981) a été spécifiquement créée comme une amélioration directe de la méthode de MP, et vise à reconstituer une phylogénie en se basant sur une approche probabiliste.

Le critère à optimiser est la **vraisemblance (likelihood)** des données en fonction d'une hypothèse évolutive :

$$L_D = \Pr(D | H)$$

Dans le cas d'une phylogénie moléculaire, les données **D** sont **l'alignement de séquences**, et l'hypothèse **H** correspond à la **topologie** de l'arbre, à ses **longueurs de branches**, et au **modèle** d'évolution utilisé.

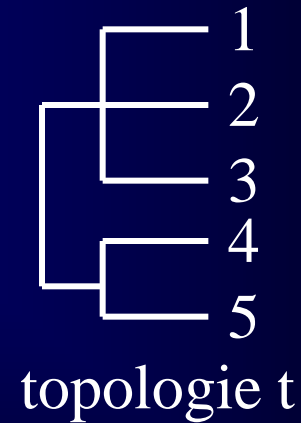
# La méthode de Maximum Likelihood

Vraisemblance d'une topologie étant donné l'alignement

On considère que chaque site est indépendant.

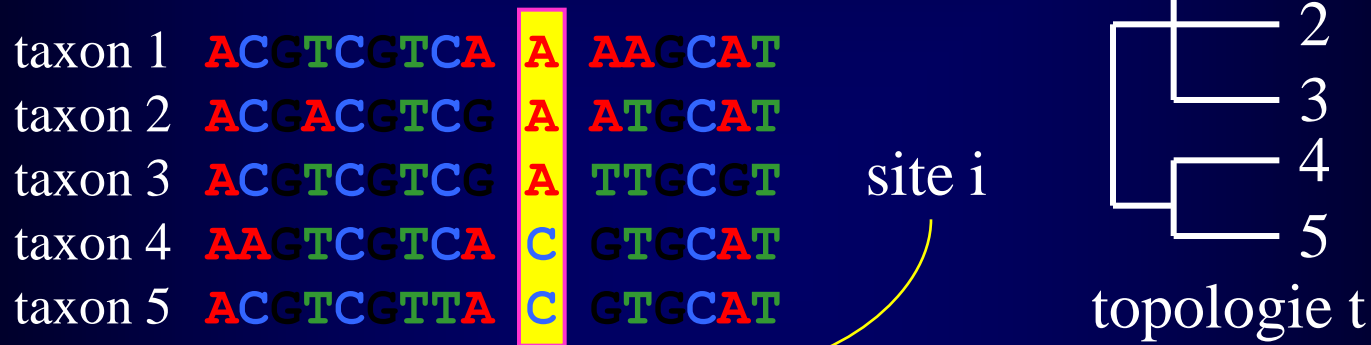
➔ La vraisemblance de l'alignement est alors le produit de la vraisemblance de chaque site.

taxon 1	ACGTCGTCA	A	AAGCAT
taxon 2	ACGACGTTCG	A	ATGCAT
taxon 3	ACGTCGTTCG	A	TTGCGT
taxon 4	AAGTCGTCA	C	GTGCAT
taxon 5	ACGTCGTTA	C	GTGCAT



# La méthode de Maximum Likelihood

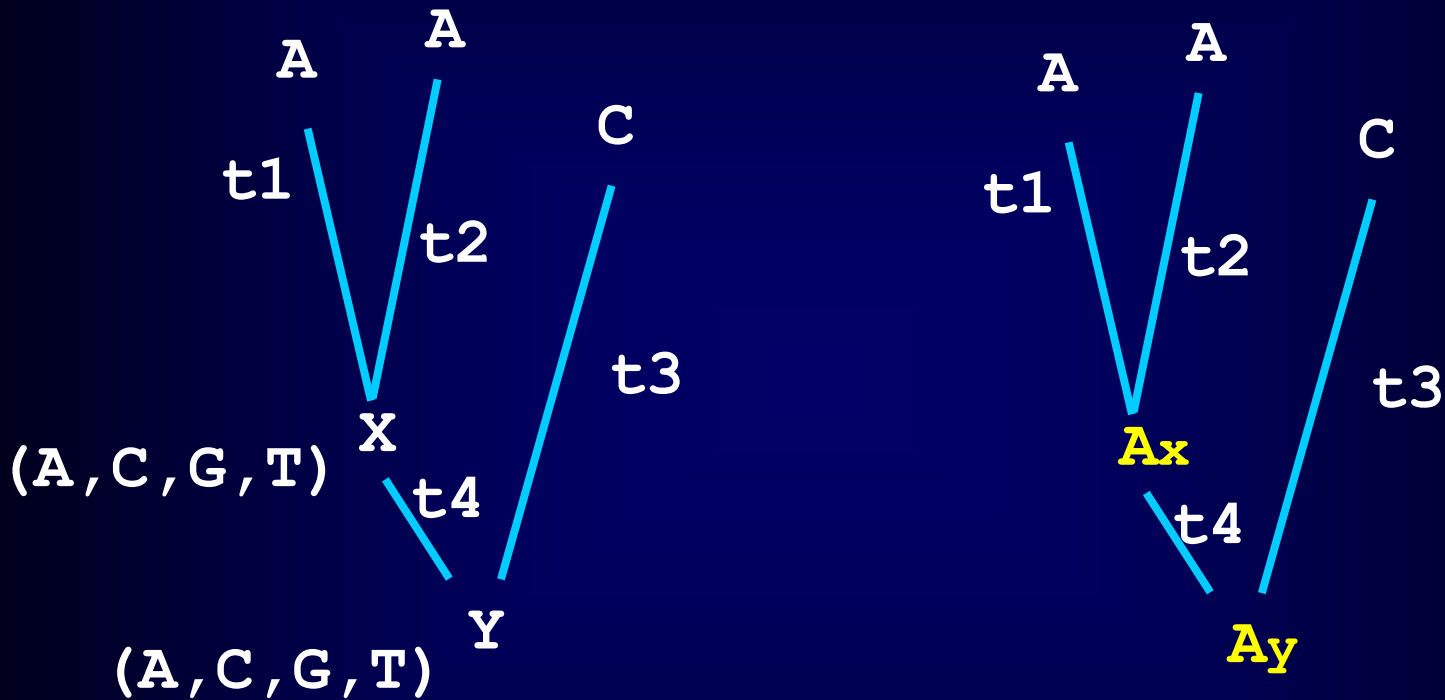
## Vraisemblance d'une topologie à un site donné



La vraisemblance  $L(t | i)$  d'une topologie  $t$  donnée pour le site  $i$  est la probabilité  $P(i | t)$  que le profil  $i = \{AAACC\}$  ait été généré chez les taxa (1,2,3,4,5) par la topologie  $t$ .

# La méthode de Maximum Likelihood

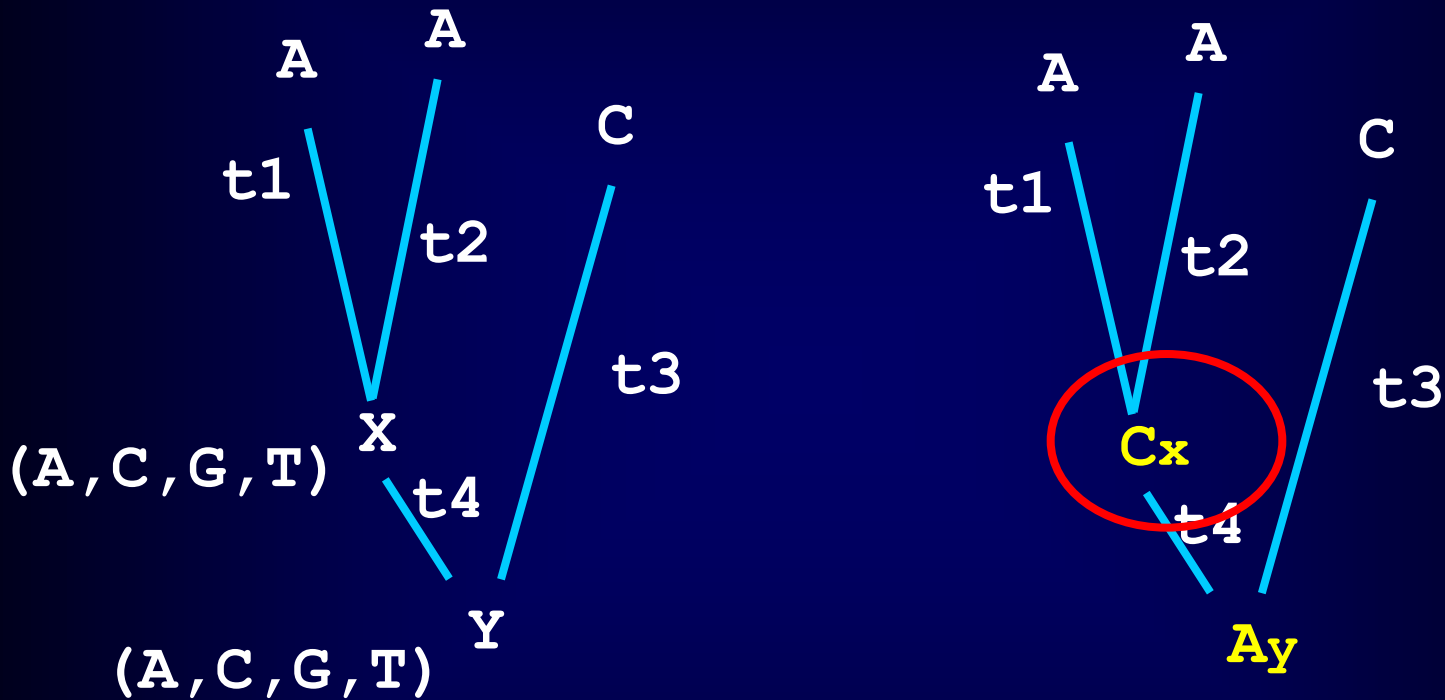
## Calcul de la vraisemblance pour un site



$$\begin{aligned} \text{Prob}(A, A, C, A_y, A_x | T) = & \\ & \text{Prob}(A_y) \text{Prob}(A_x | A_y, t_4) \\ & \text{Prob}(A | A_x, t_1) \text{Prob}(A | A_x, t_2) \\ & \text{Prob}(C | A_y, t_3) \end{aligned}$$

# La méthode de Maximum Likelihood

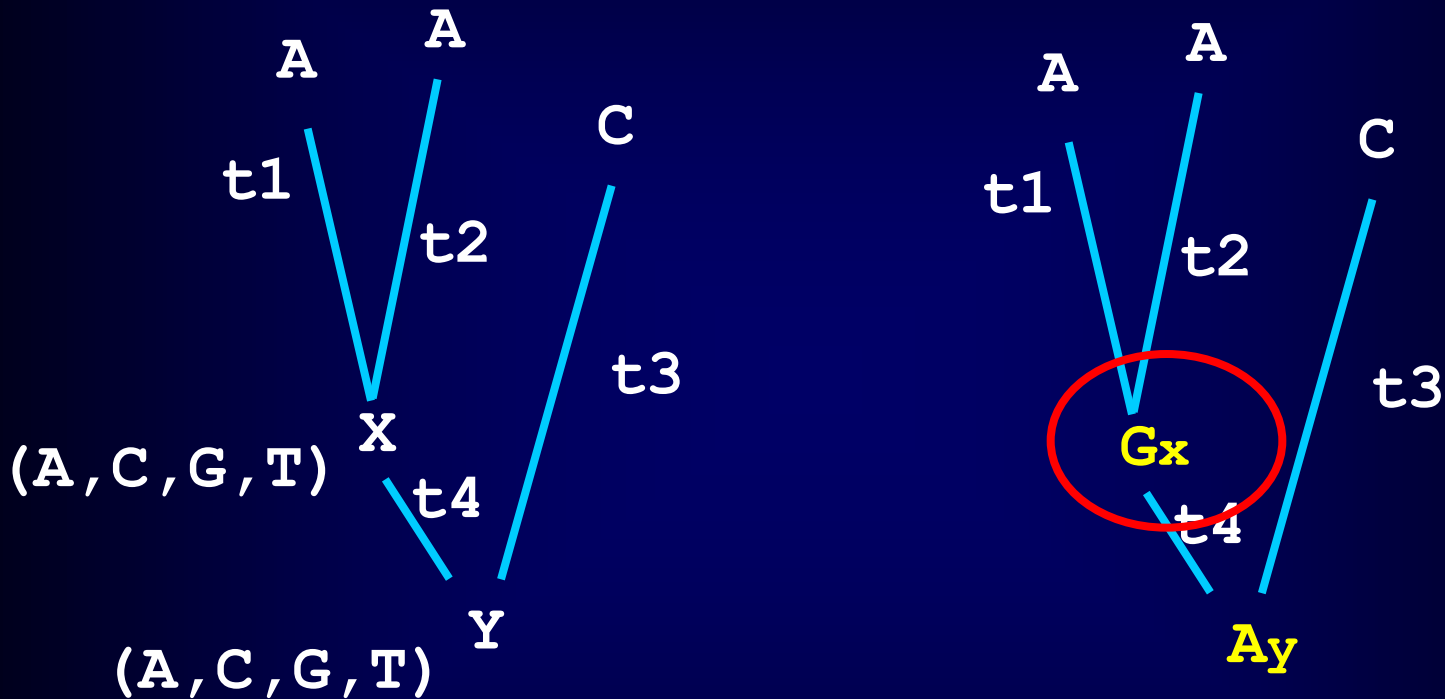
## Calcul de la vraisemblance pour un site



$$\begin{aligned} \text{Prob}(A, A, C, A_y, C_x | T) = & \\ & \text{Prob}(A_y) \text{Prob}(C_x | A_y, t_4) \\ & \text{Prob}(A | C_x, t_1) \text{Prob}(A | C_x, t_2) \\ & \text{Prob}(C | A_y, t_3) \end{aligned}$$

# La méthode de Maximum Likelihood

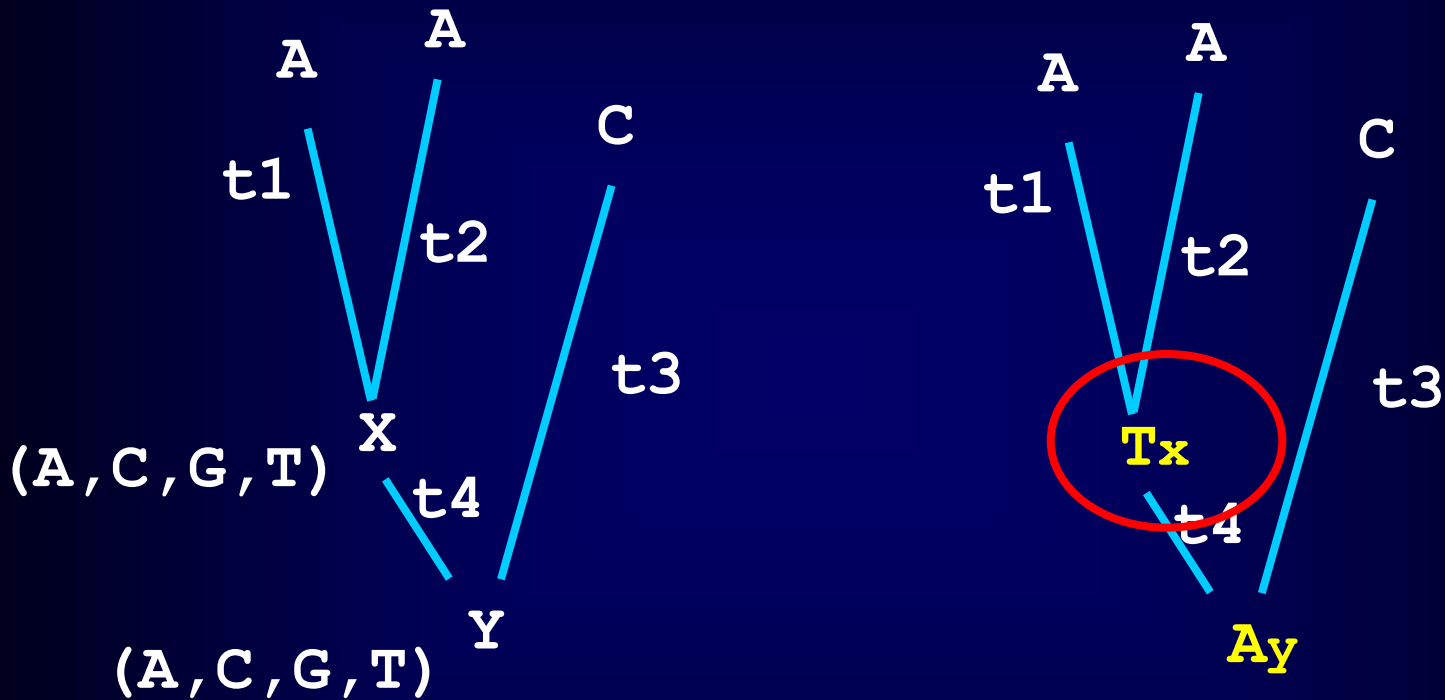
## Calcul de la vraisemblance pour un site



$$\begin{aligned} \text{Prob}(A, A, C, A_y, G_x | T) = & \\ & \text{Prob}(A_y) \text{Prob}(G_x | A_y, t_4) \\ & \text{Prob}(A | G_x, t_1) \text{Prob}(A | G_x, t_2) \\ & \text{Prob}(C | A_y, t_3) \end{aligned}$$

# La méthode de Maximum Likelihood

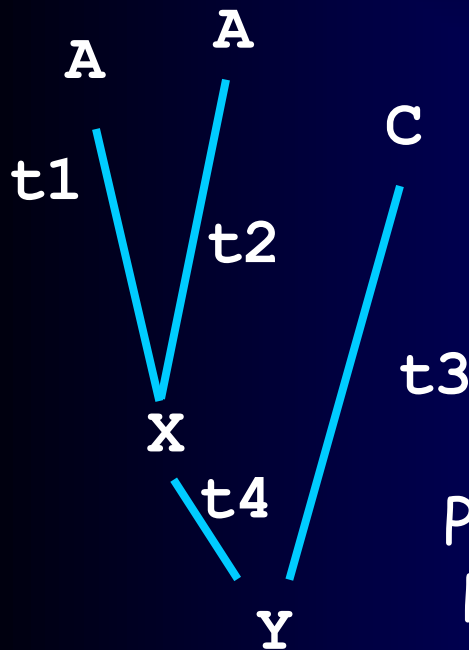
## Calcul de la vraisemblance pour un site



$$\begin{aligned} \text{Prob}(A, A, C, A_y, T_x | T) = & \\ & \text{Prob}(A_y) \text{Prob}(T_x | A_y, t_4) \\ & \text{Prob}(A | T_x, t_1) \text{Prob}(A | T_x, t_2) \\ & \text{Prob}(C | A_y, t_3) \end{aligned}$$

# La méthode de Maximum Likelihood

## Calcul de la vraisemblance pour un site



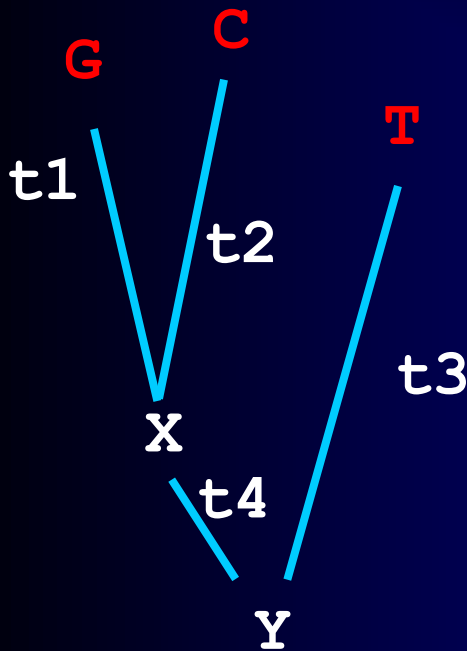
Parmi toutes les combinaisons  
laquelle possède la plus grande  
probabilité?

$$\text{Prob}(A, A, C, A_y, A_x | T) > \text{Prob}(A, A, C, C_y, T_x | T) > \\ \text{Prob}(A, A, C, A_y, T_x | T) > \dots\dots$$

Cette probabilité est la vraisemblance de cette  
topologie pour le site étudié

# La méthode de Maximum Likelihood

## Calcul de la vraisemblance pour un site



Cette même topologie est évaluée pour le site suivant (i+1) et ainsi de suite pour tout l'alignement

La vraisemblance de cette topologie pour tout l'alignement est produit de la vraisemblance des sites.

Site i = n

$$L(\tau) = \prod_{\text{Site } i = 1} L(\tau | i)$$

Site i = 1

# La méthode de Maximum Likelihood

La valeur de vraisemblance étant très petite, on l'exprime sous forme logarithmique. La somme des  $\ln L$  pour chaque site donne la vraisemblance de l'arbre.

$$\ln L(\tau) = \sum_{\text{Site } i = 1}^{\text{Site } i = n} \ln L(\tau | i)$$

Remarque :  $\ln L < 0$  car la probabilité calculée est  $< 1$

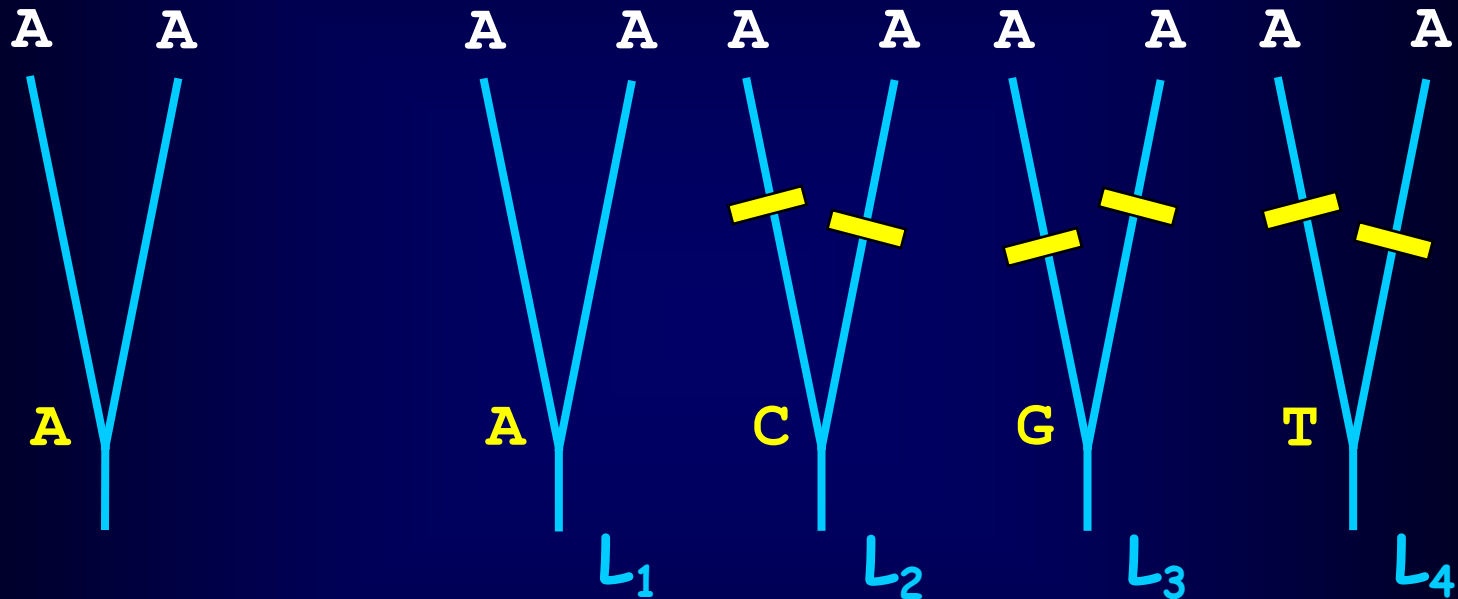
# La méthode de Maximum Likelihood

## Les étapes d'une analyse de likelihood (ML)

1. Pour une topologie donnée, calculer pour chaque site la probabilité d'observer les nucléotides présents, étant donné le modèle d'évolution.
2. Estimer les longueurs de branches qui maximisent le likelihood de la topologie, en tenant compte du fait que la vraisemblance de la topologie est égale au produit des vraisemblances de chaque site.
3. Refaire ce calcul pour toutes les topologies possibles.
4. Retenir la topologie la plus vraisemblable, c'est-à-dire celle qui possède la plus forte probabilité d'expliquer les données observées étant donné le modèle.

# La méthode de Maximum Likelihood

## Traitement des états ancestraux en ML



Raisonnement MP :  
seule l'explication la  
plus parcimonieuse  
est considérée

Raisonnement ML : tous les  
états possibles aux nœuds  
internes sont pris en compte

# La méthode de Maximum Likelihood

La somme de tous les likelihoods de toutes les données possibles étant donné tous les modèles et topologies possibles serait égale à 1, mais le likelihood des données observées pour un modèle donné est minuscule et s'exprime sous forme d'un logarithme.

1-10 séquences : **recherche exhaustive**

10-20 séquences : **recherche « branch and bound »**

plus de 20 séquences : **recherche heuristique**

L'arbre de départ peut être obtenu par « random sequence addition » (très long), mais on peut aussi utiliser un arbre de distance (très rapide).

# La méthode de Maximum Likelihood

Pourquoi une analyse de ML prend beaucoup de temps ?

Dans un raisonnement ML, trois choses doivent être optimisées en même temps : la **topologie** (valeur discrète), les **longueurs de branches** et les **paramètres du modèle** (valeurs continues).

Il est mathématiquement impossible d'optimiser ces trois valeurs en même temps, il faut donc passer par un cycle d'étapes intermédiaires où l'on fixe deux valeurs pour optimiser la troisième.

# La méthode de Maximum Likelihood

## Cycle complet d'un analyse ML :

- (1) estimation des paramètres du modèle à partir d'un arbre choisi au hasard (exemple : arbre NJ)
- (2) optimisation des longueurs de branches pour chaque topologie possible, en fonction des paramètres estimés
- (3) si la meilleure topologie est différente de l'arbre de départ, réestimation des paramètres
- (4) nouvelle optimisation des longueurs de branches pour chaque topologie possible

Tant qu'une amélioration du score de likelihood est possible, il faut continuer ce cycle...

# La méthode de Maximum Likelihood

## Avantages de la vraisemblance

Permet une utilisation optimale du modèle d'évolution en tenant compte de l'information contenue dans tous les sites ; théoriquement efficace même lorsque les taux de substitution varient entre taxa (LBA).

## Défauts de la vraisemblance

Méthode très coûteuse en temps de calcul.

mercredi 14 septembre 2005

programme de la matinée

LBA, de la parcimonie à la vraisemblance

La méthode de Maximum Likelihood

Autres approches probabilistes

Comparaison entre les méthodes NJ, MP, ML

Introduction aux programmes Treefinder,  
Phylip, Phym1, Raxml

Exercices avec Treefinder

## Autres approches probabilistes

Parce que les analyses de ML sont extrêmement coûteuses en temps de calcul, de grands efforts sont faits pour améliorer la rapidité et l'efficacité des programmes basés sur les méthodes probabilistes :

- Amélioration des algorithmes de likelihood

(cf. différence de temps de calcul PAUP\* / PhyML/Treefinder/...)

- Recours à des approches dérivées

  - méthodes de quartets

  - approche bayésienne

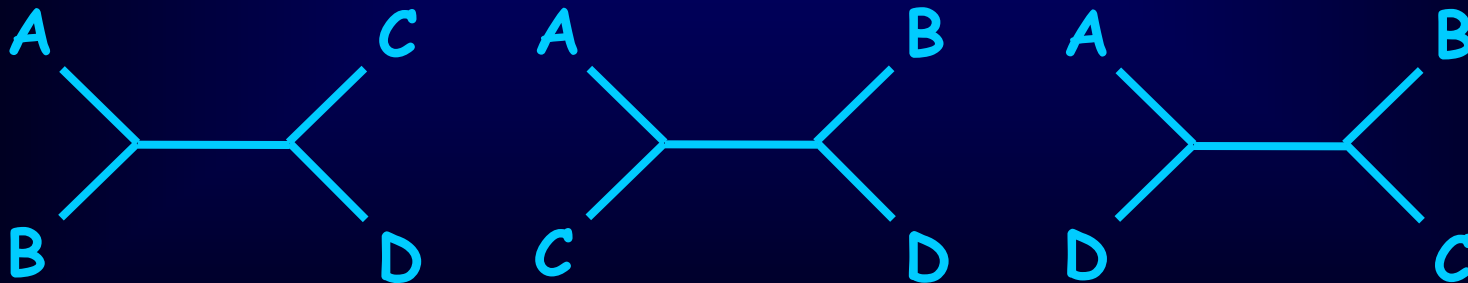
## Autres approches probabilistes

Exemple de méthode de quartets :

« **Quartet puzzling** » (Strimmer et von Haeseler 1996)

1. toutes les combinaisons de 4 séquences sont testées selon le critère de ML et la meilleure des 3 topologies possibles est déterminée pour chaque quartet

Rappel : pour 4 séquences, 3 topologies possibles



## Autres approches probabilistes

Exemple de méthode de quartets :

« **Quartet puzzling** » (Strimmer et von Haeseler 1996)

2. des arbres sont construits en ajoutant les séquences une à une selon le principe du random sequence addition ; les valeurs de ML calculées dans la première étape sont utilisées pour déterminer l'endroit où placer chaque séquence dans l'arbre naissant

3. on construit un consensus de tous les arbres trouvés lors des répétitions de la deuxième étape ; l'arbre final n'est donc pas forcément complètement résolu, mais contient une estimation du soutien des nœuds internes

mercredi 14 septembre 2005

programme de la matinée

LBA, de la parcimonie à la vraisemblance

La méthode de Maximum Likelihood

Autres approches probabilistes

Comparaison entre les méthodes NJ, MP, ML

Introduction aux programmes Treefinder,  
Phyml, Phylip, Raxml

Exercices avec Treefinder

# Comparaison entre les méthodes NJ, MP, ML

## méthodes d'analyse phylogénétique

caractères

distances

parcimonie

likelihood & Co

NJ,  
UPGMA

ME

stepwise  
addition

branch swapping

agglomératives

d'optimisation

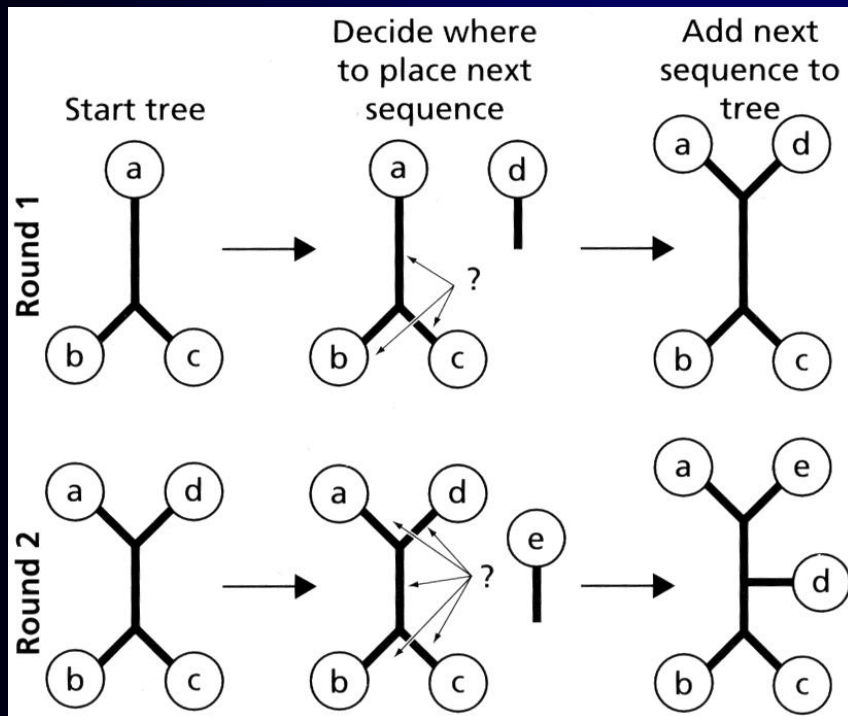
méthodes de reconstruction d'arbres

# Comparaison entre les méthodes NJ, MP, ML

## méthodes de reconstruction d'arbres

agglomératives

d'optimisation



Comparaison de toutes les topologies possibles

(ou, dans le cas d'une recherche heuristique, d'un certain nombre de topologies seulement)

## Comparaison entre les méthodes NJ, MP, ML

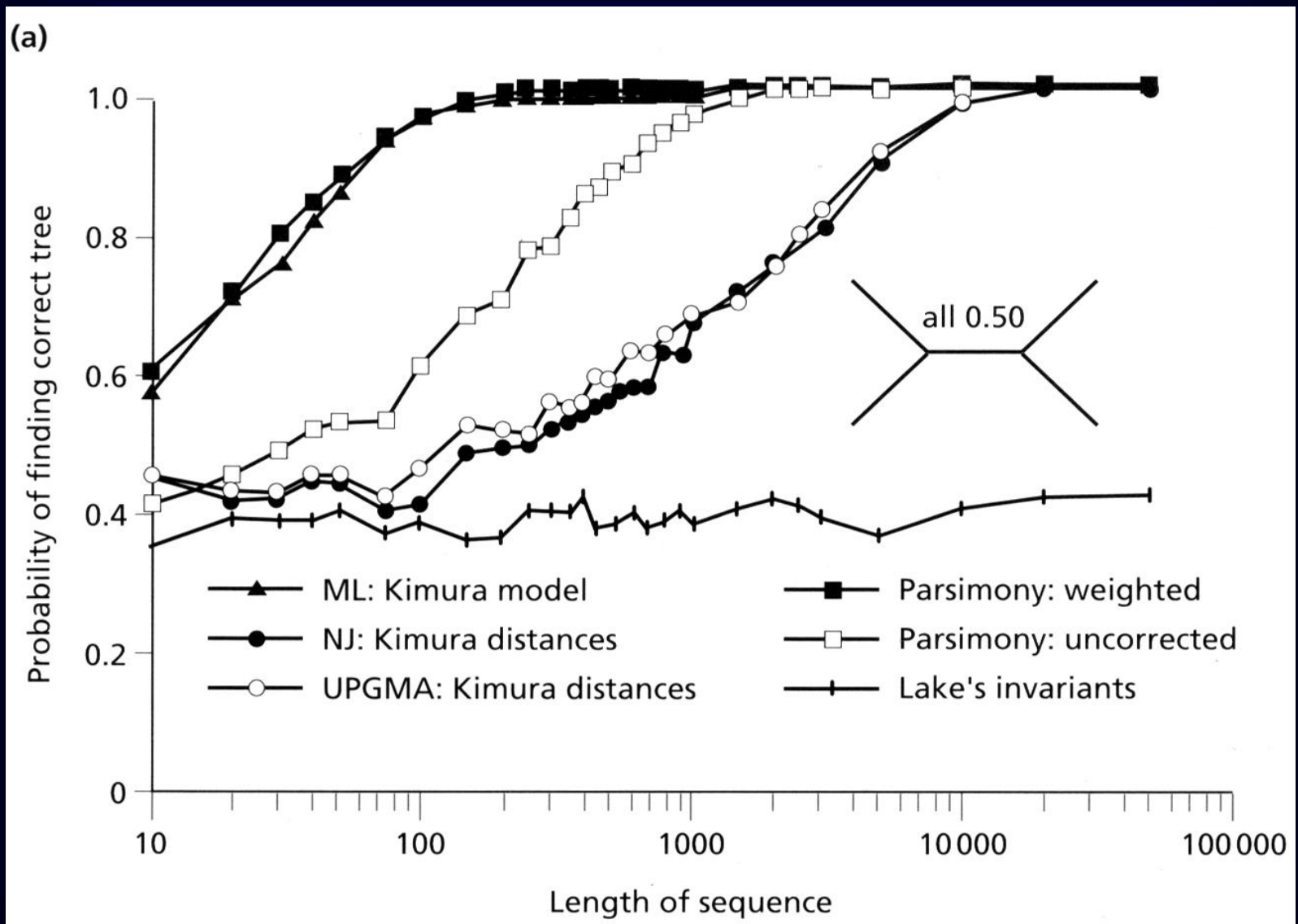
L'utilisation d'une **méthode agglomérative seule** est valable dans le cas de **NJ** (parfois meilleur que ME!), mais suffit rarement à trouver le meilleur arbre pour des analyses de MP et de ML.

L'utilisation d'une **méthode d'optimisation seule** n'est possible qu'avec un petit nombre de séquences (cela correspond à une **recherche exhaustive**).

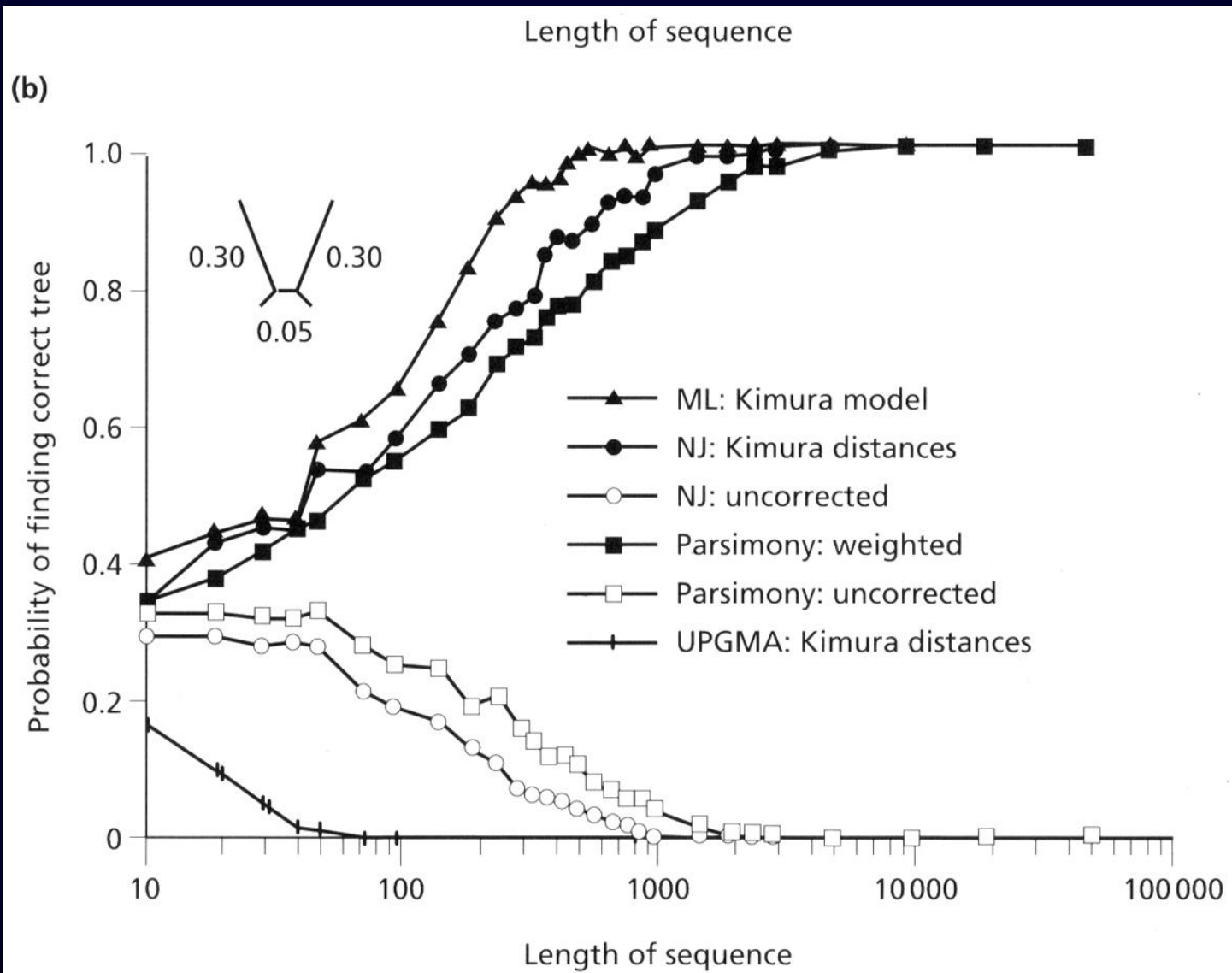
Dans le cas d'une **recherche heuristique**, les deux approches sont utilisées successivement :

l'arbre de départ est généralement obtenu par une approche agglomérative / les réarrangements sont effectués dans une optique d'optimisation.

# Comparaison entre les méthodes NJ, MP, ML



# Comparaison entre les méthodes NJ, MP, ML



## Comparaison entre les méthodes NJ, MP, ML

Actuellement, la méthode de parcimonie n'est plus utilisée pour analyser des séquences (sauf par quelques irréductibles cladistes n'ont pas compris la différence entre données moléculaires et données morphologiques)

Les méthodes de distances sont extrêmement pratiques pour des analyses préliminaires ou pour des recherches simples (nombre de phylotypes distincts, etc.).

Les méthodes probabilistes ont l'avantage de rendre possible la résolution de problématiques annexes à la simple phylogénie : choix du modèle d'évolution, tests statistiques (topologies, horloge moléculaire), analyses biogéographiques, datation, ...

mercredi 14 septembre 2005

programme de la matinée

LBA, de la parcimonie à la vraisemblance

La méthode de Maximum Likelihood

Autres approches probabilistes

Comparaison entre les méthodes NJ, MP, ML

Introduction au programme Treefinder, PhymI,  
Phylip, Raxml

Exercices avec Treefinder

# Introduction au programme PhyML

## PhyML (Guindon et Gascuel, 2003)

Programme basé sur la méthode de ML qui a été proposé récemment pour estimer des grandes phylogénies.

L'algorithme utilise un arbre de départ (obtenu par BioNJ ou fourni par l'utilisateur) pour estimer les paramètres du modèle. Cet arbre est ensuite modifié pour augmenter sa vraisemblance. Les paramètres sont réestimés après chaque amélioration de la topologie. L'optimisation simultanée des paramètres, de la topologie et des longueurs des branches permet d'atteindre plus rapidement la vraisemblance optimale.

PhyML est accessible sur

<http://atgc.lirmm.fr/phyml/>

# Introduction au programme PhyML

## Comparaison entre PhyML et PAUP\*

	PhyML	PAUP*
estimation directe des paramètres	rapide	très lente
branch swapping	rapide mais pas exhaustif	long mais efficace

Pour des données complexes et un grand nombre de séquences, l'idéal est d'utiliser PhyML pour estimer les paramètres et obtenir une bonne topologie de départ, puis d'utiliser PAUP\* pour vérifier que la topologie obtenue avec PhyML est vraiment la meilleure.

# Introduction au programme PhyML

- PHYML v2.4.1 -

Settings for this run:

D	Data type (DNA/AA)	DNA
I	Input sequences interleaved (or sequential)	interleaved
S	Analyze multiple data sets	no
B	Non parametric bootstrap analysis	no
M	Model of nucleotide substitution	HKY
E	Base frequency estimates (empirical/ML)	empirical
T	Ts/tv ratio (fixed/estimated)	fixed (ts/tv = 4.00)
V	Proportion of invariable sites (fixed/estimated)	fixed (p-invar = 0.00)
R	One category of substitution rate (yes/no)	yes
U	Input tree (BIONJ/user tree)	BIONJ
O	Optimise tree topology	yes

Are these settings correct? (type Y or letter for one to change)

Introduction au programmes PhymI,Phylip, Raxml

Liens internet

**Introduction au programme Treefinder**  
**démonstration**