Introduction

Modèles d'évolution
- Deux approches
- Modèles pour l'ADN
- Modèles pour les AA
- Modèles plus complexes

Construction des arbres
- Maximum de vraisemblance
- Distances 2/2
- Maximum de parcimonie

Tests statistiques
- Sur les modèles
- Sur les arbres

Conclusion

Bibliographie

Molecular phylogenetics : state-of-the-art methods for looking into the past

Simon Whelan, Pietro Liò et Nick Goldman

[WLG01]

Emmanuel Cornet
École normale supérieure

Cours de Bioinformatique – MPRI

28 février 2005

Plan

1 Modèles d'évolution
- Deux approches
- Modèles pour l'ADN
- Modèles pour les AA
- Modèles plus complexes

2 Construction des arbres
- Maximum de vraisemblance
- Distances 2/2
- Maximum de parcimonie

3 Tests statistiques
- Sur les modèles
- Sur les arbres

Introduction : ADN $\leftrightarrow$ Mot sur $\{\mathtt{A}; \mathtt{T}; \mathtt{G}; \mathtt{C}\}$

Phylogenetics survey

Emmanuel Cornet

Introduction

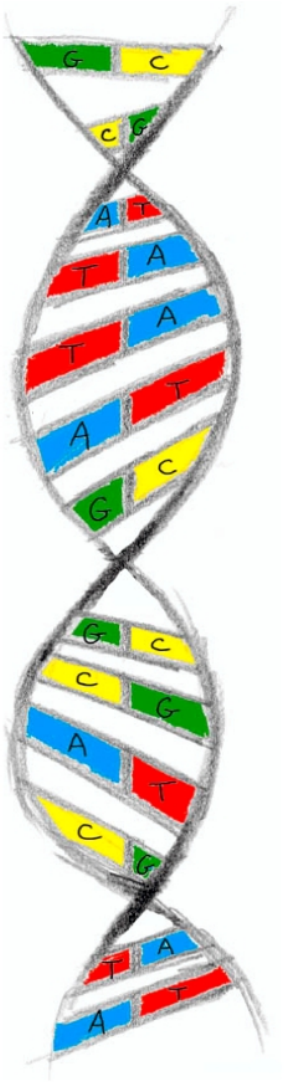
Modèles d'évolution
- Deux approches
- Modèles pour l'ADN
- Modèles pour les AA
- Modèles plus complexes

Construction des arbres
- Maximum de vraisemblance
- Distances 2/2
- Maximum de parcimonie

Tests statistiques
- Sur les modèles
- Sur les arbres

Conclusion

Bibliographie

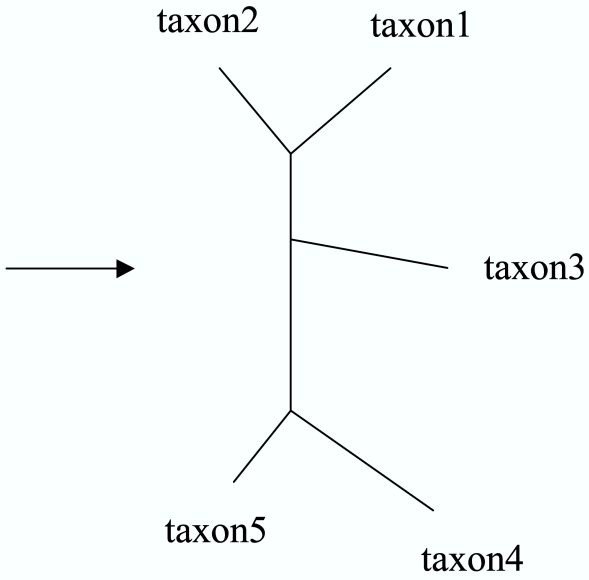


GCTAATCGCACAT

Introduction : Séquences ↔ Arbre phylogénétique

Phylogenetics survey
Emmanuel Cornet
Introduction
Modèles d'évolution
Deux approches
Modèles pour l'ADN
Modèles pour les AA
Modèles plus complexes
Construction des arbres
Maximum de vraisemblance
Distances 2/2
Maximum de parcimonie
Tests statistiques
Sur les modèles
Sur les arbres
Conclusion
Bibliographie

```
taxon1: AAGACATGTGGCACTGTGACTGA
taxon2: AGGACATGTGGCAGTGTGACTGA
taxon3: AGTACATGTGACAGAGTGACTGA
taxon4: AGCACATGTGACTGAGTGACCGA
taxon5: AGCACATGTGACAGAGTGACCGG
```



(séquences préalablement alignées)

Il nous faut :

- un modèle d'évaluation des « similarités » entre les séquences (ancêtre commun ?) ;
- une méthode de construction de l'arbre phylogénétique.

Modèles d'évolution

Introduction
Modèles d'évolution
Deux approches
Modèles pour l'ADN
Modèles pour les AA
Modèles plus complexes
Construction des arbres
Maximum de vraisemblance
Distances 2/2
Maximum de parcimonie
Tests statistiques
Sur les modèles
Sur les arbres
Conclusion
Bibliographie

Un modèle de description de l'évolution de séquences est composé de deux éléments :

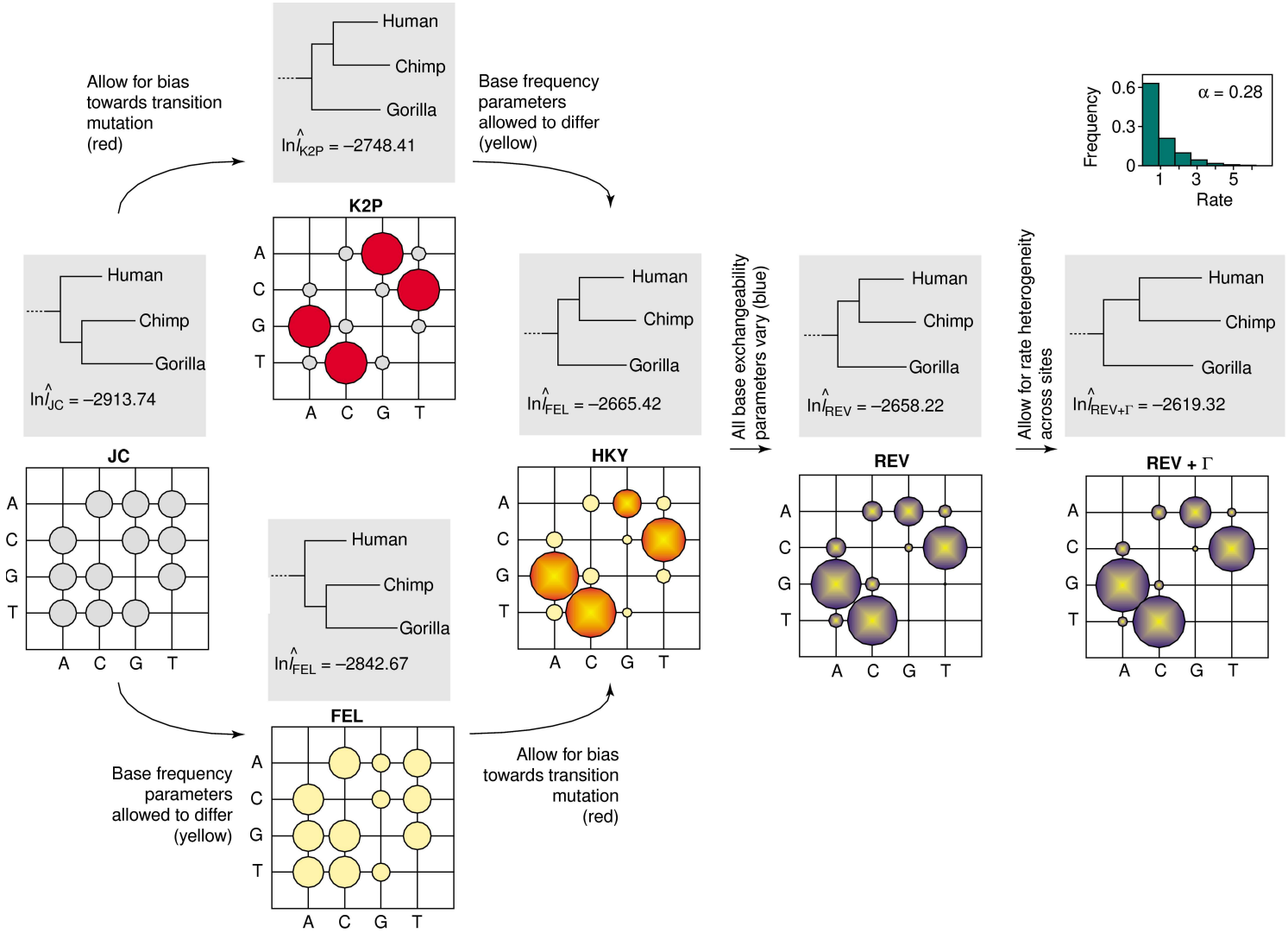
- Un arbre phylogénétique
- Une description de la manière dont les séquences évoluent

Pour cette dernière, deux approches sont possibles :

- **Empirique** : valeurs calculées sur des données préexistantes, une fois pour toutes. Facile à calculer et à utiliser, mais pas d'adaptation.
- **Paramétrique** : adaptative, mais plus complexe à gérer.

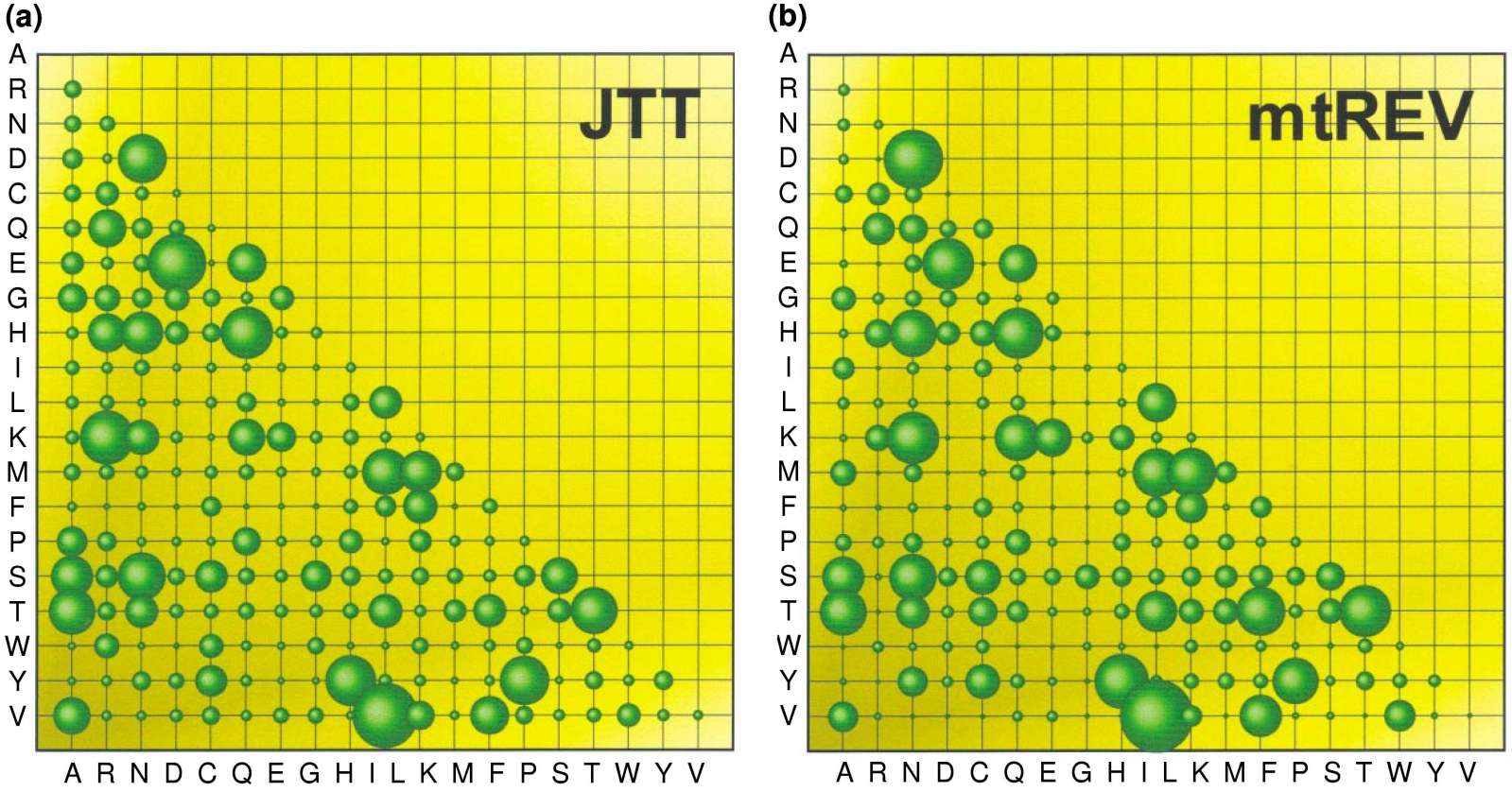
Modèles de substitution (ADN)

Approche paramétrique, en général.



Modèles de substitution (AA)

Approche empirique, en général.



Modèles plus complexes

Introduction
Modèles d'évolution
Deux approches
Modèles pour l'ADN
Modèles pour les AA
Modèles plus complexes
Construction des arbres
Maximum de vraisemblance
Distances 2/2
Maximum de parcimonie
Tests statistiques
Sur les modèles
Sur les arbres
Conclusion
Bibliographie

Il existe des modèles plus subtils, qui tiennent compte par exemple :

- à la fois des substitutions dans l'ADN et de la pression de **sélection** sur les **protéines** elles-mêmes ;
- des relations entre le taux de substitutions **synonymes** (code génétique) et **non synonymes** ;
- de l'aspect **structurel** des **protéines** et/ou de l'**ARN**.

L'étude simultanée de la phylogénie et de la structure des protéines permet de faire un va-et-vient entre les connaissances acquises sur l'une et sur l'autre.

Construction des arbres : trois méthodes

Introduction
Modèles d'évolution
Deux approches
Modèles pour l'ADN
Modèles pour les AA
Modèles plus complexes
Construction des arbres
Maximum de vraisemblance
Distances 2/2
Maximum de parcimonie
Tests statistiques
Sur les modèles
Sur les arbres
Conclusion
Bibliographie

Les trois méthodes principales pour la construction d'arbres phylogénétiques sont :

- La méthode du maximum de vraisemblance (*maximum likelihood*)
- La méthode du maximum de parcimonie (*maximum parsimony*)
- La méthode des distances deux à deux (*pairwise distances*)

Maximum de vraisemblance (*Maximum likelihood*)

Introduction
Modèles d'évolution
Deux approches
Modèles pour l'ADN
Modèles pour les AA
Modèles plus complexes
Construction des arbres
Maximum de vraisemblance
Distances 2/2
Maximum de parcimonie
Tests statistiques
Sur les modèles
Sur les arbres
Conclusion
Bibliographie

L'arbre "ML" est celui *qui rend les séquences observées les plus plausibles, étant donné le modèle choisi pour l'évolution des séquences*.

On renverse le problème habituel :

- **Probabilité** : paramètres connus $\longrightarrow$ prédiction des événements
- **Vraisemblance** : observation des données $\longrightarrow$ estimation des paramètres

Si la probabilité de l'événement X, étant donnés les paramètres P du modèle, est

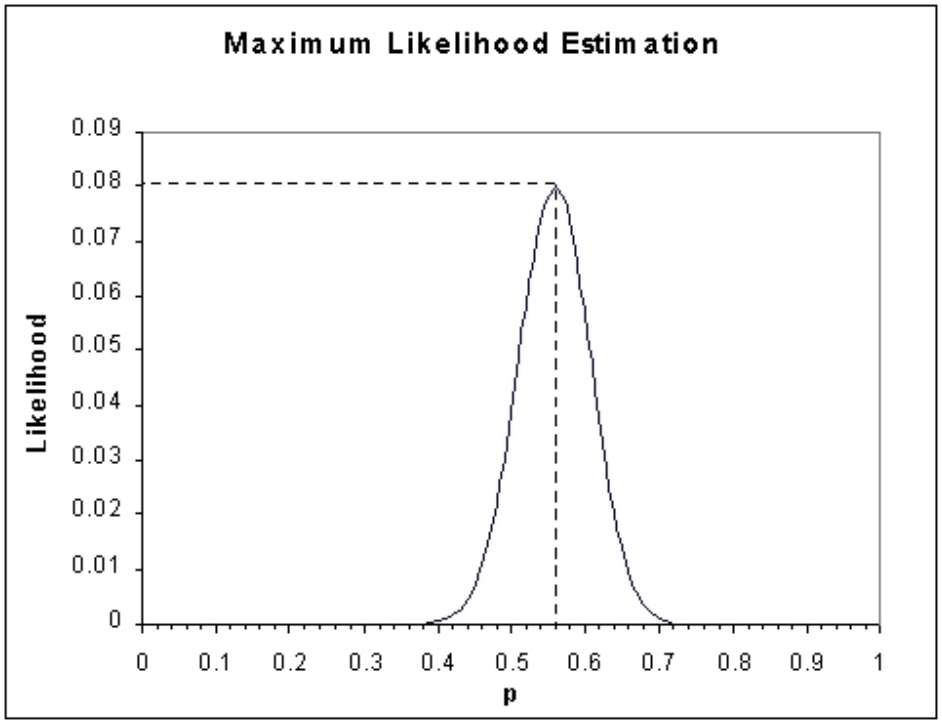
$$P(X|P)$$

alors la *vraisemblance* des paramètres, étant donné l'événement X, est

$$L(P|X) = \alpha\, P(X|P)$$

Maximum de vraisemblance : un exemple simple

100 tirages à pile ou face : 56 fois « face ». On représente $L(p|\text{données})$ en fonction de p et on cherche le maximum :



Maximum de vraisemblance : en pratique

Avantages :

- S'appuie sur des théories statistiques puissantes et bien explorées
- Par exemple : se prête bien à des tests statistiques
- Permet d'estimer plusieurs paramètres en même temps

Inconvénients :

- Nécessite d'explorer *a priori* tout l'espace des paramètres (*ie* toutes les topologies d'arbres)
- Il existe des heuristiques, mais le maximum global n'est plus garanti

Distances 2 à 2 (*pairwise distances*)

Introduction
Modèles d'évolution
Deux approches
Modèles pour l'ADN
Modèles pour les AA
Modèles plus complexes
Construction des arbres
Maximum de vraisemblance
Distances 2/2
Maximum de parcimonie
Tests statistiques
Sur les modèles
Sur les arbres
Conclusion
Bibliographie

Estimation de la distance d'évolution pour chaque paire de séquences, et construction d'un arbre à partir de ces distances.

- Plus rapide que ML
- Perte d'information inévitable
- Ne gère pas bien les modèles dont certains paramètres ne sont pas connus *a priori* (par exemple κ de transition/transversion)

Maximum de parcimonie (*maximum parsimony*)

Introduction
Modèles d'évolution
Deux approches
Modèles pour l'ADN
Modèles pour les AA
Modèles plus complexes
Construction des arbres
Maximum de vraisemblance
Distances 2/2
Maximum de parcimonie
Tests statistiques
Sur les modèles
Sur les arbres
Conclusion
Bibliographie

Sélectionne l'arbre nécessitant le minimum de changements évolutifs.

- Bonne approximation de ML dans les zones où les séquences sont proches
- Topologie erronée dans les zones où ce n'est pas le cas, dites zones de Felsenstein

Tests statistiques sur les modèles

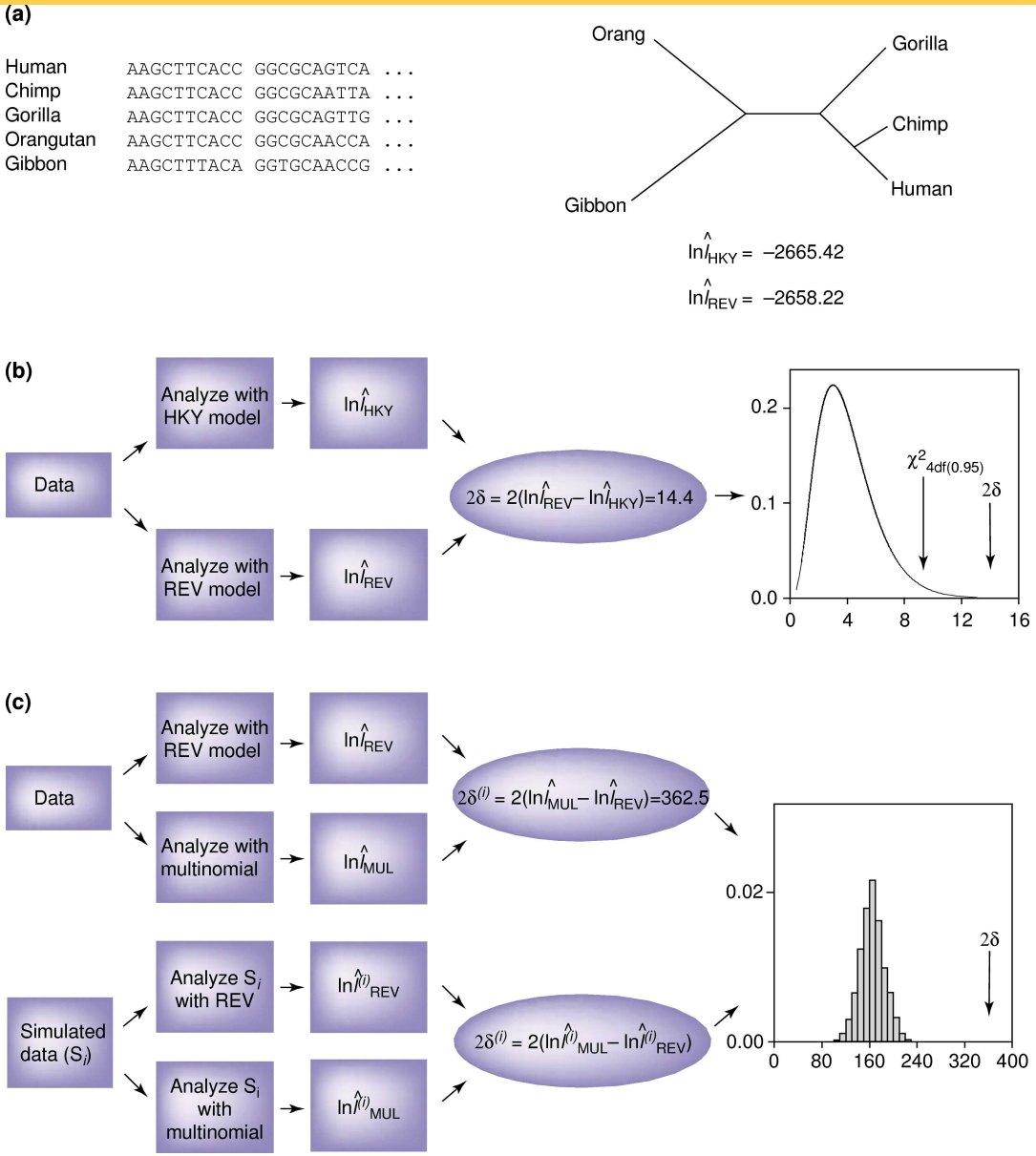
Introduction
Modèles d'évolution
Deux approches
Modèles pour l'ADN
Modèles pour les AA
Modèles plus complexes
Construction des arbres
Maximum de vraisemblance
Distances 2/2
Maximum de parcimonie
Tests statistiques
Sur les modèles
Sur les arbres
Conclusion
Bibliographie

Disponibles presque exclusivement pour l'approche ML : *likelihood ratio test*.

Comparaison entre deux modèles
On utilise la valeur du maximum de vraisemblance (différence des logarithmes, d'où le nom) de deux modèles pour savoir à quel point un modèle plus complexe traduit mieux la réalité des observations qu'un modèle plus simple. On prend également en compte le nombre de degrés de liberté supplémentaires dans le modèle complexe.

Test d'un seul modèle
On compare un modèle donné avec une valeur hypothétique et optimale du maximum de vraisemblance.

Tests statistiques sur les modèles



Quelques observations

Introduction
Modèles d'évolution
Deux approches
Modèles pour l'ADN
Modèles pour les AA
Modèles plus complexes
Construction des arbres
Maximum de vraisemblance
Distances 2/2
Maximum de parcimonie
Tests statistiques
Sur les modèles
Sur les arbres
Conclusion
Bibliographie

- Les modèles actuels n'expliquent pas tous les phénomènes évolutifs présents dans les séquences analysées.
- Les modèles les plus complexes sont généralement les plus proches de la réalité.
- Les rares cas où le test a montré qu'un modèle était adéquat, il s'agissait d'un modèle relativement complexe, appliqué à des séquences sous une faible pression de sélection.

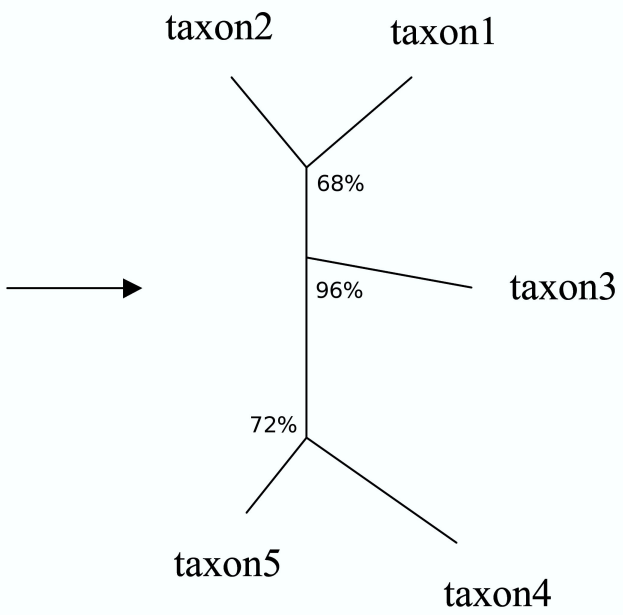
Conclusion : il semble que les effets de sélection soient parmi les principaux facteurs pas encore bien intégrés aux modèles d'évolution actuels.

Bootstrap non paramétrique

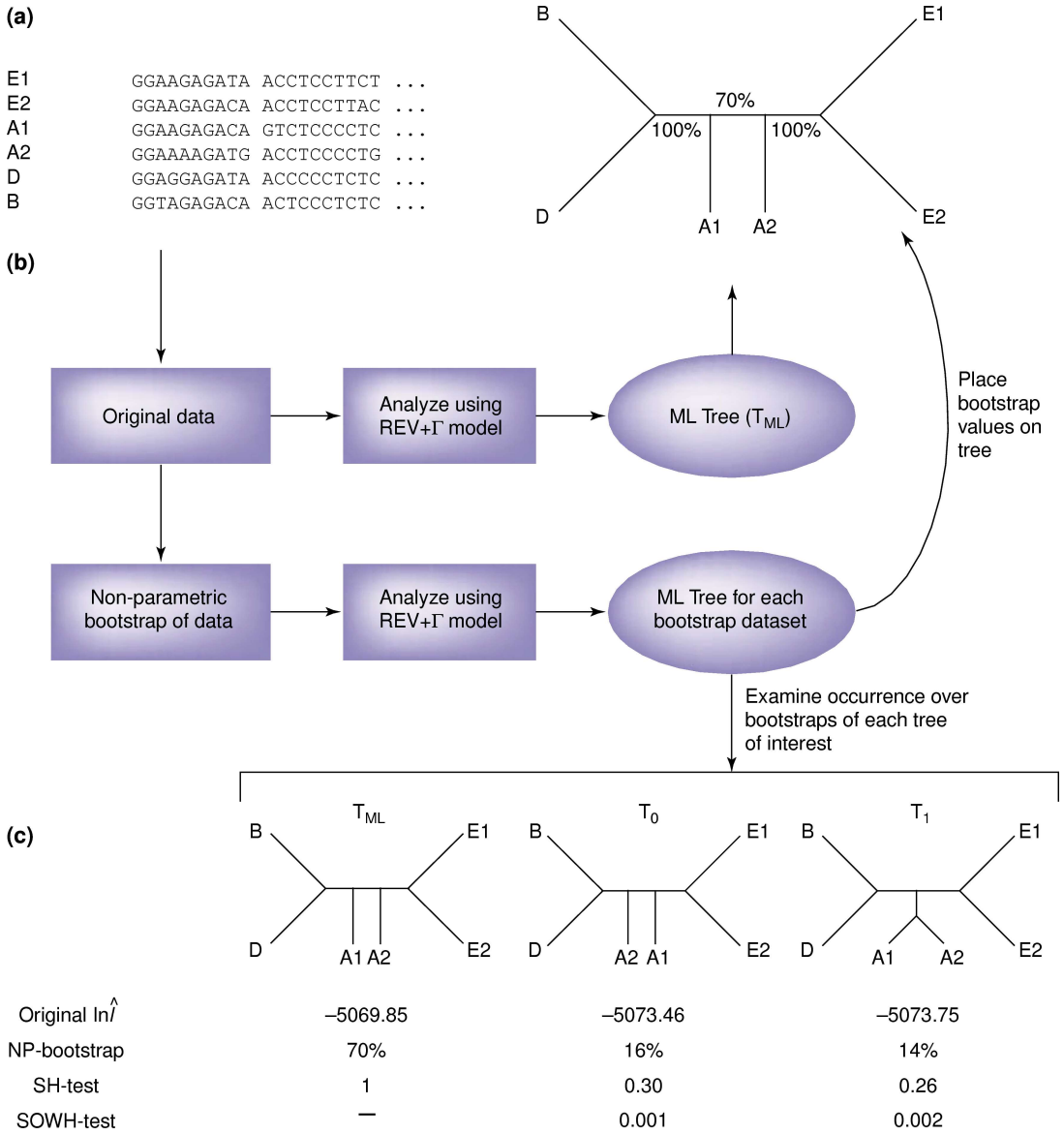
Principe du *bootstrap* :

- On rééchantillonne les séquences sur les colonnes (tirage de plusieurs fois la même colonne : généralement autorisé)
- On réitère le processus sur chaque nouveau jeu de données
- On reporte la robustesse des branchements sur l'arbre
- Interprétation des pourcentages !

```
taxon1: AAGACATGTGGCACTGTGACTGA
taxon2: AGGACATGTGGCAGTGTGACTGA
taxon3: AGTACATGTGACAGAGTGACTGA
taxon4: AGCACATGTGACTGAGTGACCGA
taxon5: AGCACATGTGACAGAGTGACCGG
```



Tests statistiques de la topologie d'un arbre



Possibilités d'amélioration

Introduction
Modèles d'évolution
Deux approches
Modèles pour l'ADN
Modèles pour les AA
Modèles plus complexes
Construction des arbres
Maximum de vraisemblance
Distances 2/2
Maximum de parcimonie
Tests statistiques
Sur les modèles
Sur les arbres
Conclusion
Bibliographie

Quelques moyens pour améliorer la fiabilité d'un arbre :

- Inclure plus de séquences
- Prendre des séquences plus longues
- Concaténer plusieurs gènes

Conclusion

Introduction

Modèles d'évolution
- Deux approches
- Modèles pour l'ADN
- Modèles pour les AA
- Modèles plus complexes

Construction des arbres
- Maximum de vraisemblance
- Distances 2/2
- Maximum de parcimonie

Tests statistiques
- Sur les modèles
- Sur les arbres

Conclusion

Bibliographie

- De plus en plus de séquences dans le domaine public : importance de la phylogénie moléculaire
- Méthodes d'estimation des paramètres : meilleure compréhension des phénomènes de l'évolution
- Tycho Brahe et Kepler

Bibliographie

- L'article étudié :
  [WLG] Simon Whelan, Pietro Liò and Nick Goldman, Molecular phylogenetics : state-of-the-art methods for looking into the past, *TRENDS in Genetics* Vol. 17 No. 5, May 2001

- Sur le maximum de vraisemblance :
  `http://statgen.iop.kcl.ac.uk/bgim/mle/sslike.html`
- Sur le *likelihood ratio test* :
  `http://workshop.molecularevolution.org/resources/lrt.php`
- Sur la phylogénie moléculaire en général :
  `http ://www.isem.univ-montp2.fr/PPP/PM/TEACH/MBPE/PHYLOGENY/M1W_C09-Galtier_ML.pdf`

Retrouvez les documents sur le Web

Introduction

Modèles d'évolution
- Deux approches
- Modèles pour l'ADN
- Modèles pour les AA
- Modèles plus complexes

Construction des arbres
- Maximum de vraisemblance
- Distances 2/2
- Maximum de parcimonie

Tests statistiques
- Sur les modèles
- Sur les arbres

Conclusion

Bibliographie

Cette présentation est disponible au format PDF à l'adresse

`www.eleves.ens.fr/home/cornet/Survey_presentation.pdf`