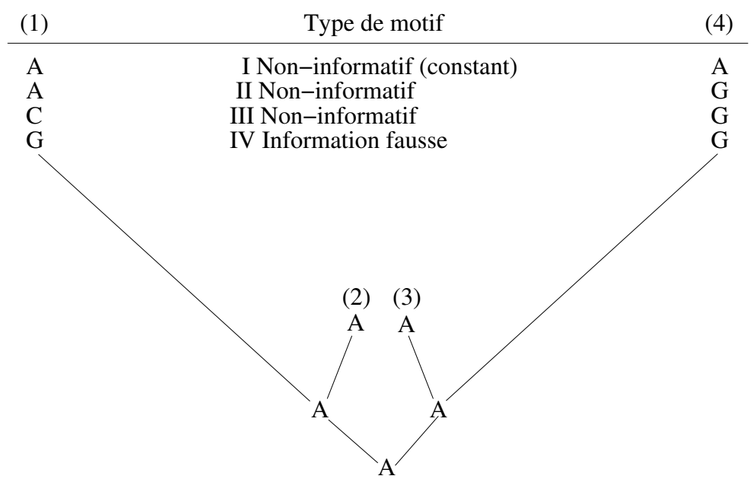
Inconsistance du modèle de parsimonie

Consistance d'une méthode d'estimation: Capacité à converger vers une bonne valeur (ici, le vrai arbre de phylogénie) avec l'augmentation des données.

En considérant un modèle d'évolution simple, la méthode de parsimonie peut entrainer une fausse estimation de l'arbre (*Felsenstein*)

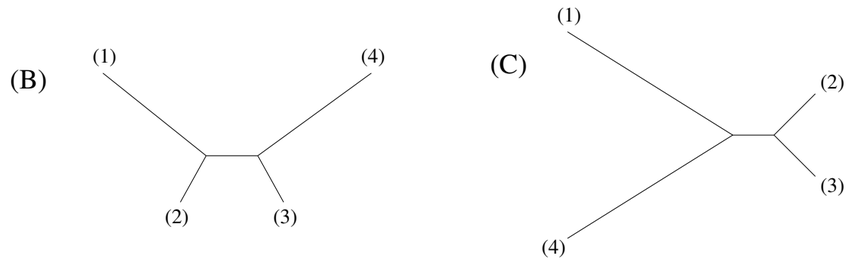
Supposons que la vraie phylogénie d'un groupe de 4 taxons soit:

(A)



Taille des branches reflète le taux d'évolution. Taux d'évolution accéléré pour les branches menant à (1) et (4). Les deux autres branches si courtes qu'il n'y a presque pas de difference entre (2) et (3).

4 classes possibles pour les nucléotides de (1) et (4). I, II, III ne fournissent aucune information permettant de clairement favoriser cet arbre par rapport à tous les autres. IV: seule classe permettant de favoriser un arbre particulier. Malheureusement, favorise le mauvais arbre:

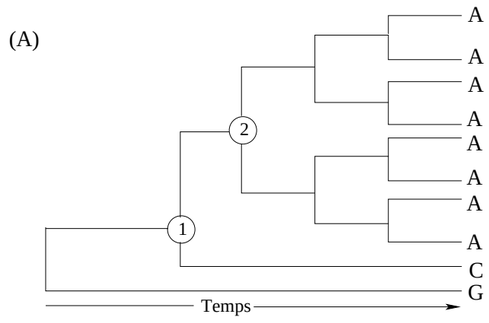


Felsenstein appelle une telle situation positively misleading car plus on a de caractères (plus les séquences sont longues), plus on est sur d'obtenir un arbre faux.

Lorsqu'on est dans la zone Felsenstein, le seul espoir d'obtenir un bon arbre est de séquencer suffisamment peu de caractères, de sorte à être induit en erreur le moins possible. Phénomène appelé attraction des longues branches.

Différence entre parsimonie et likelihood

Arbre non-enraciné:

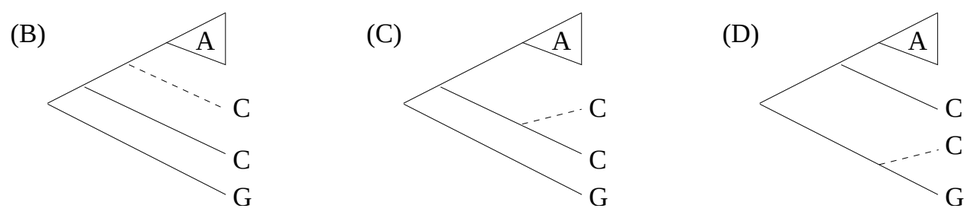


Comment deviner le nucléotide de l'ancêtre (1)?

Algorithme de Fitch: On peut attribuer à (1) n'importe lequel des nucléotides A, C ou G avec un poids de 2 pour l'arbre. T augmente ce poids de 1.

D'autre part, une nouvelle séquence avec C, A, ou G peut être insérée avec un poids de 2 à n'importe quelle branche. Également, nouvelle séquence contenant T peut être insérée avec un poids de 3

à n'importe quelle branche. Dans tous les cas, séquence non informative (ne favorise aucun arbre)



Maximum de vraisemblance: Choisir l'hypothèse qui maximise la probabilité d'observer le nucléotide obtenu.

Modèle d'évolution choisit: Taux de substitution identique pour tous les nucléotides; nombre moyen de substitutions le long d'une branche proportionnel á la longueur de la branche.

Observation: Tous les descendants de (2) ont des A. Donc, taux de mutation faible. D'où, phylogénie entrainant peu de mutations plus probable que phylogénie entrainant beaucoup de mutations. Donc,

présence d'un A à l'ancêtre (2) beaucoup plus probable que présence d'un C, G ou T (mais hypothèse d'un C,G,T non rejetée).

Nucléotide à l'ancêtre 1? A, C ou G? Supposons un A à (2). Plus probable que la substitution ait eu lieu sur la branche longue. Donc, plus probable d'avoir un A en (1). Plus généralement, ordre de probabilité: $A > C > G > T$.

Rajout d'une séquence avec un C: arbre (C) plus probable que les autres, car pour les arbres (B) et (D), deux substitutions $A \to C$ seraient nécessaires.

La taille des branches est une information importante pour la méthode de maximum likelihood, et donc pas de problème d'attraction des longues branches. Dans ce cas, arbre (8B) très probable.