Reconstruction d’états ancestraux

- **Introduction**
  - Reconstruction d'états ancestraux: à partir des caractéristiques observées chez les individus, extrapoler ceux de leur ancêtre commun.

- Applications :
  - Evolution : individus, population, souches, espèces.
  - Génétique : reconstruction de séquences ancêtres (acides nucléiques, protéines)
  - Composition des génomes, en bases, gènes…
  - Ordre des gènes, reconstruction de génomes ancestraux
  - Evolution des caractères phénotypiques, des présences géographique (phylo-géographie)…
  - Repliement des protéines…

Modèles d'évolution réalistes

- Reconstruction d'états ancestraux:
  - doit reposer sur **des modèles d'évolution réalistes** pour inférer correctement les états ancestraux.

- Cependant, quelle que soit la qualité du modèle utilisée,
  - la qualité des prédictions se détériore avec l'augmentation des distances évolutives observées entre l'ancêtre et ses descendants.

- Remarque: les modèles les plus réalistes sont aussi les plus complexes et les plus difficiles à calculer.

- Les méthodes sont souvent appliquées sur un arbre qui a été lui-même inféré sur les mêmes données.
  - Elles sont donc dépendantes de la qualité de cet arbre.

- L'incertitude sur la topologie de l'arbre peut être prise en compte par certaines méthodes, en évaluant la reconstruction des états ancestraux sur plusieurs arbres concurrents (ex. approche Bayésienne)

A utiliser avec modération!

- Swofford et Maddison: ‘character state reconstructions can provide a powerful mechanism for studying many facets of the evolutionary process. However, the zeal with which these techniques are sometimes advocated belies the complexity of the problem’.

Historique

- Historique

- Le concept de reconstruction d'états ancestraux est souvent attribué à Emile Zuckerkandl et Linus Pauling (1963) dans le cadre de la reconstruction de séquences ancêtres.

- Ce concept avait déjà été introduit dans le champ de la cladistique (1900)

  Cladistique : inférer les relations évolutives entre les espèces sur la base de caractéristiques communes, qui sont supposées être présentes chez le dernier ancêtre commun des espèces étudiées.

- Ce concept a aussi été utilisé dans le cadre de la reconstruction d'inversions chromosomiques dans l'évolution de génomes de drosophiles (Dobzhansky et Sturtevant, 1938).

- => Concept apparu de façon indépendante dans différentes disciplines

Méthodes et algorithmes

- Méthodes et algorithmes

- Tout commence avec une **phylogénie** :
  - Modèle présentant sous la forme d'un arbre les liens de filiations entre un ancêtre commun et les feuilles de l'arbre (taxon).

- Les **nœuds internes** représentent les **ancêtres** des feuilles qui en dérivent

- La **racine** de l'arbre représente le dernier ancêtre commun à l'ensemble des feuilles

- Dans la majorité des méthodes (sauf…), cet arbre est une donnée connue (fournie par l'utilisateur).

Différentes approches

- Reconstruction d'états ancestraux :
  - appliquer un modèle évolutif sur une phylogénie connue

- L'objectif est d'estimer les paramètres de ce modèle évolutif à partir des caractéristiques des individus présents aux feuilles.

- Différentes approches:
  - Maximum de parcimonie
  - Maximum de vraisemblance
  - Inférence Bayésienne
  - méthode stochastique
  - …

Reconstruction d'états ancestraux

- **Maximum de parcimonie**

- Parcimonie : a pour principe de sélectionner l'hypothèse la plus simple

- Dans notre contexte :
  - trouver la distribution des états ancestraux qui **minimise le nombre de changements** le long des branches de l'arbre afin de rendre compte des états observés au niveau des feuilles (Fitch, 1971).

- Les méthodes MP sont intuitivement attirantes et très efficaces. De plus, elles peuvent être utilisées pour initialiser l'algorithme de maximisation de ML.

Algorithme

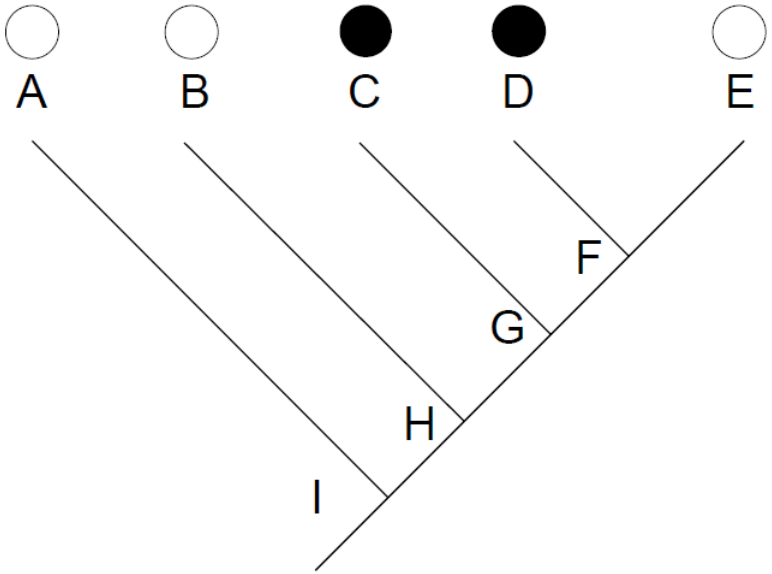
- Algorithme
- Exemple de MP non pondérée avec des caractères non ordonnées.

- L'optimisation de l'inférence des états ancestraux se fait par deux passages dans l'arbre:
  - Des feuilles à la racine 'downpass'
  - De la racine aux feuilles 'uppass'

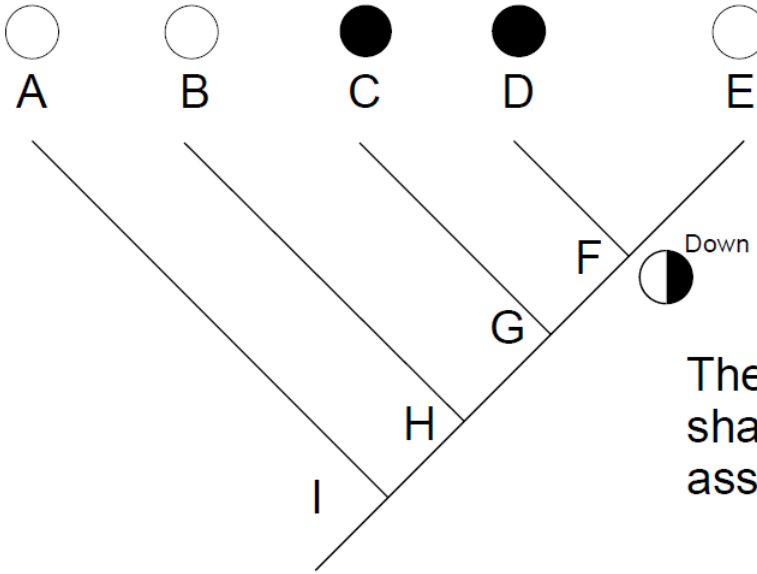
- Deux règles sont appliquées:
  - Règle 1: si deux descendants partage des états en commun, l'ensemble d'états commun est assigné à l'ancêtre (intersection).

  - Règle 2: si les descendant ne partagent aucun états, alors l'union des états observés est attribué à l'ancêtre (union).

Reconstruction d'états ancestraux



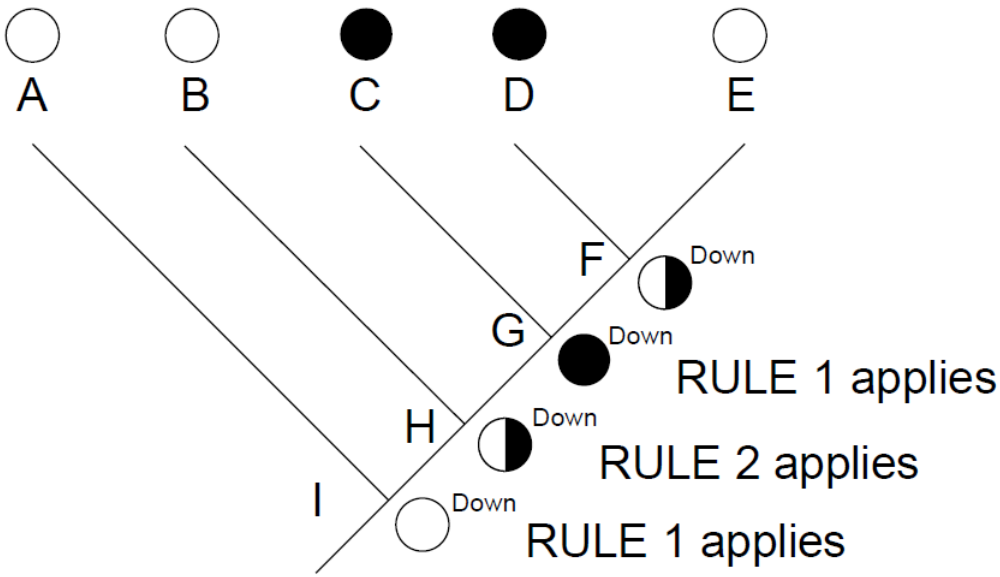
Reconstruction d'états ancestraux



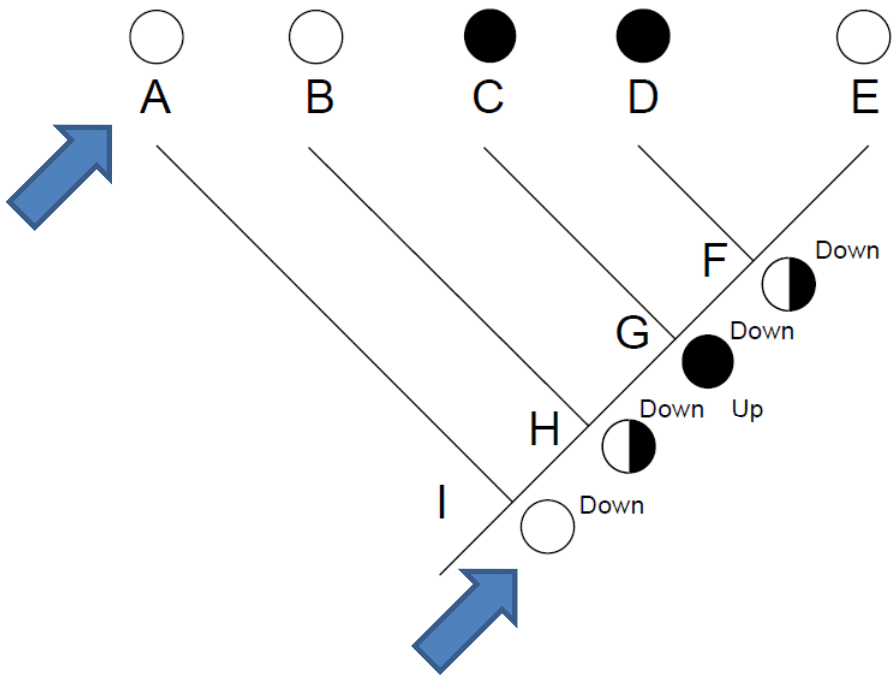
RULE 2 applies

The descendants of F (D and E) share no states so the union is assigned to F.

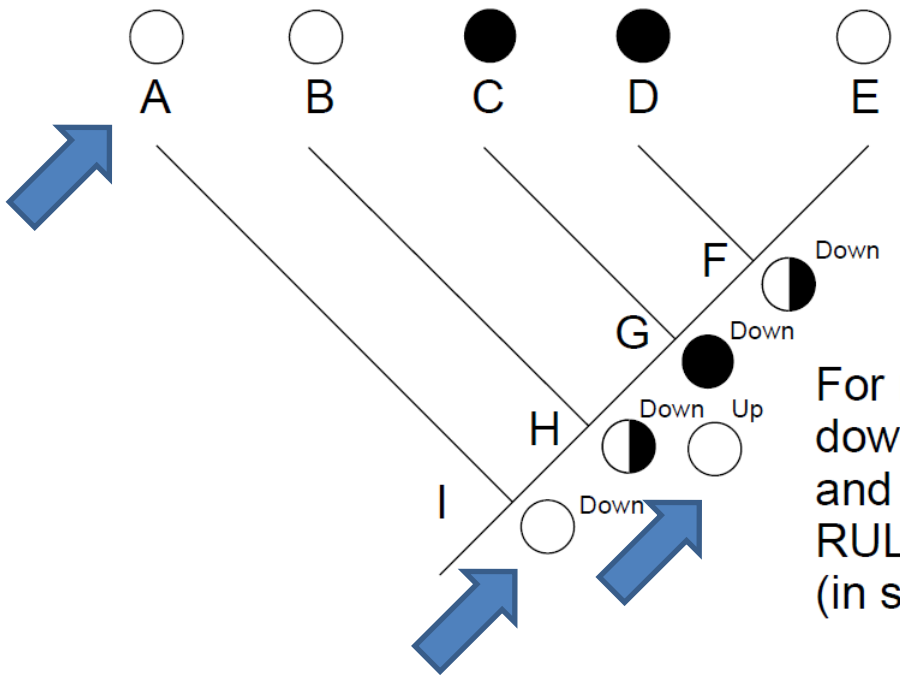
Reconstruction d'états ancestraux



Reconstruction d'états ancestraux

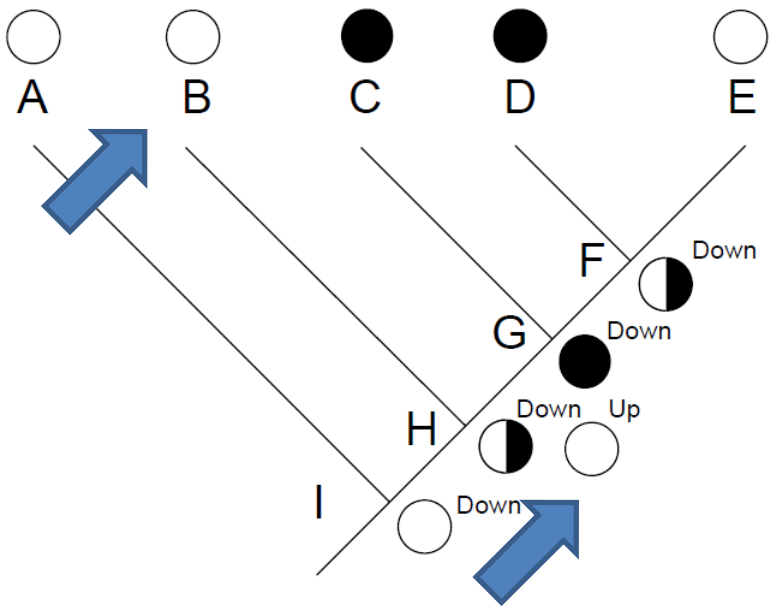


Reconstruction d'états ancestraux

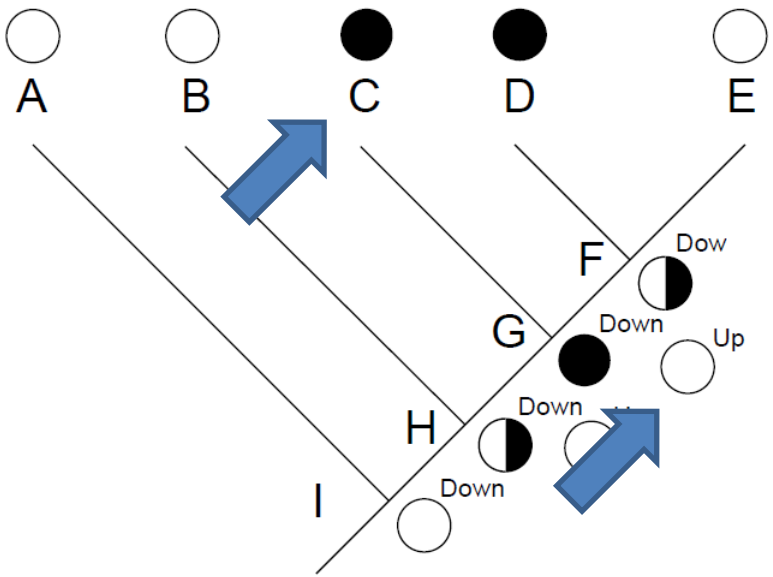


For node H, consider downpass state of ancestor I and ancestor's sister node (A) RULE 1 gives uppass state (in shaded box).

Reconstruction d'états ancestraux

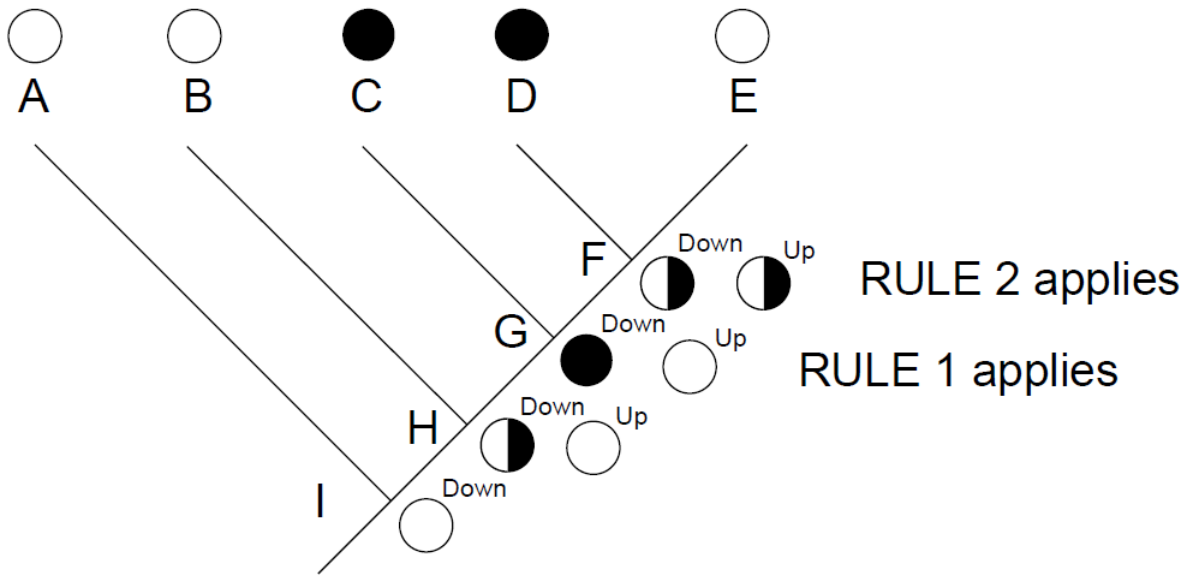


Reconstruction d'états ancestraux



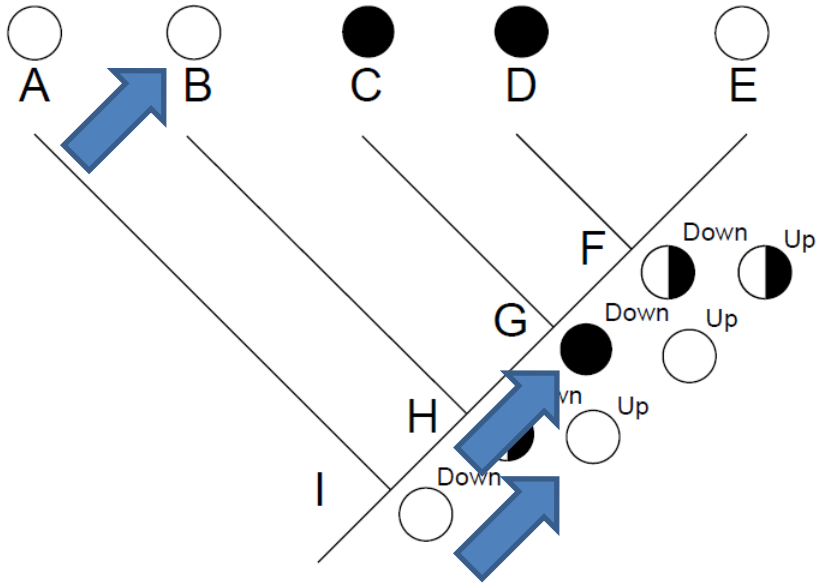
RULE 1 applies

Reconstruction d'états ancestraux



Reconstruction d'états ancestraux

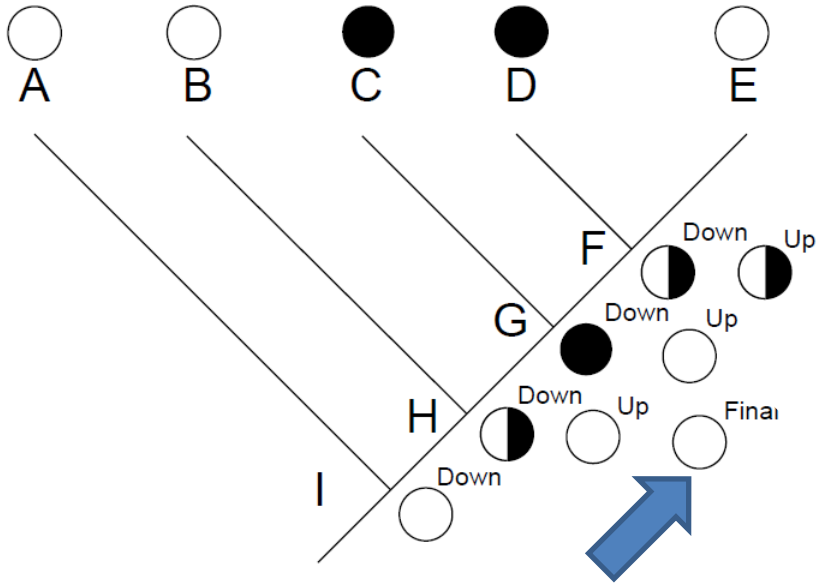
- Dans un dernier passage, nous allons considérer les affectations réalisées lors des deux premiers passages



- Prenons le nœud H:
  - État de 'up' de H
  - Etats de 'down' de ses descendants B et G
  - Choisir l'état majoritaire dans ces trois ensembles

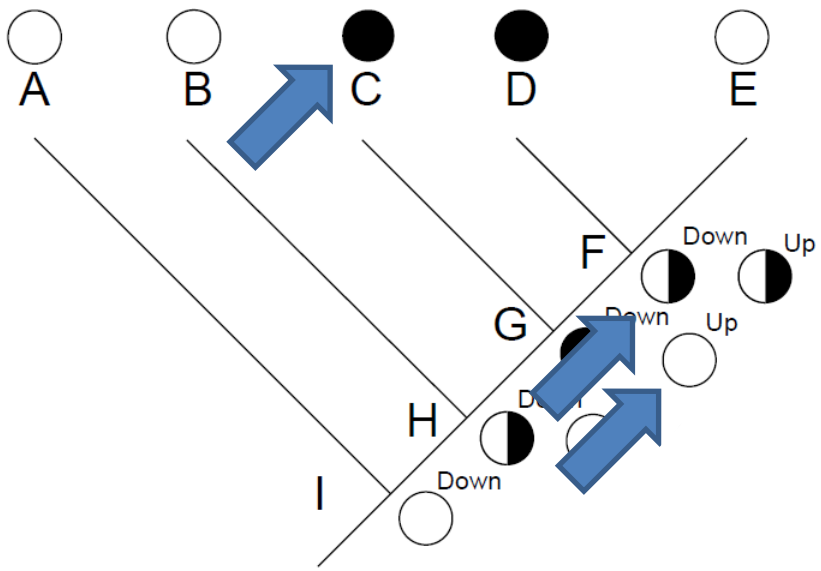
Reconstruction d'états ancestraux

- Dans un dernier passage, nous allons considérer les affectations réalisées lors des deux premiers passages



Reconstruction d'états ancestraux

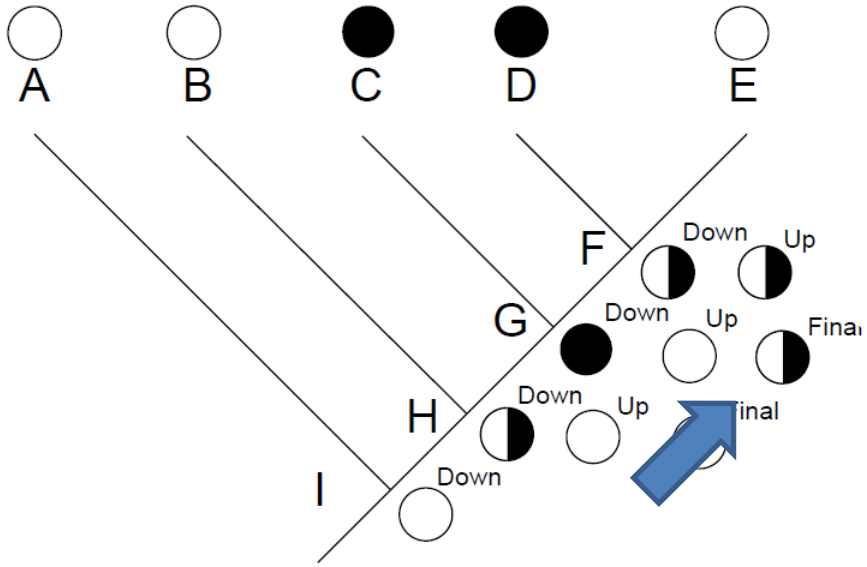
- Dans un dernier passage, nous allons considérer les affectations réalisées lors des deux premiers passages



- Prenons le nœud G:
  - État de 'up' de G
  - Etats de 'down' de ses descendants C et F
  - Choisir l'état majoritaire dans ces trois ensembles

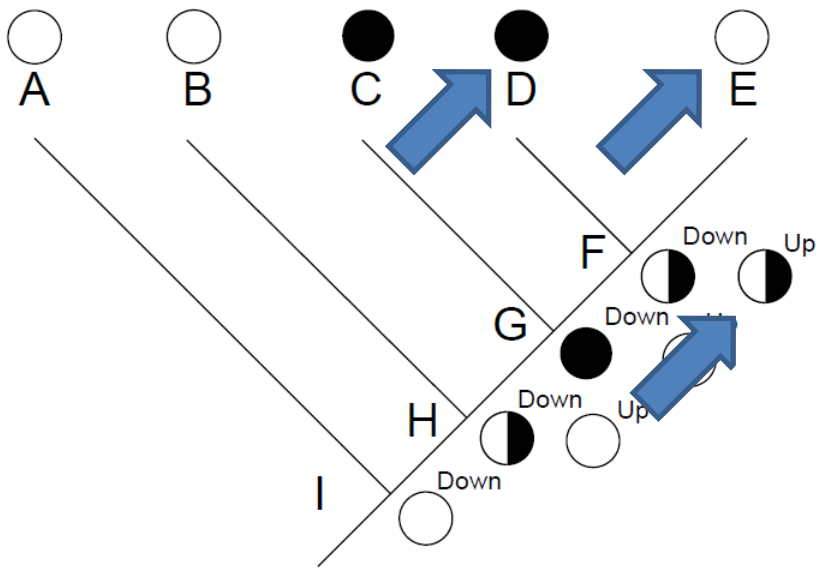
Reconstruction d'états ancestraux

- Dans un dernier passage, nous allons considérer les affectations réalisées lors des deux premiers passages



Reconstruction d'états ancestraux

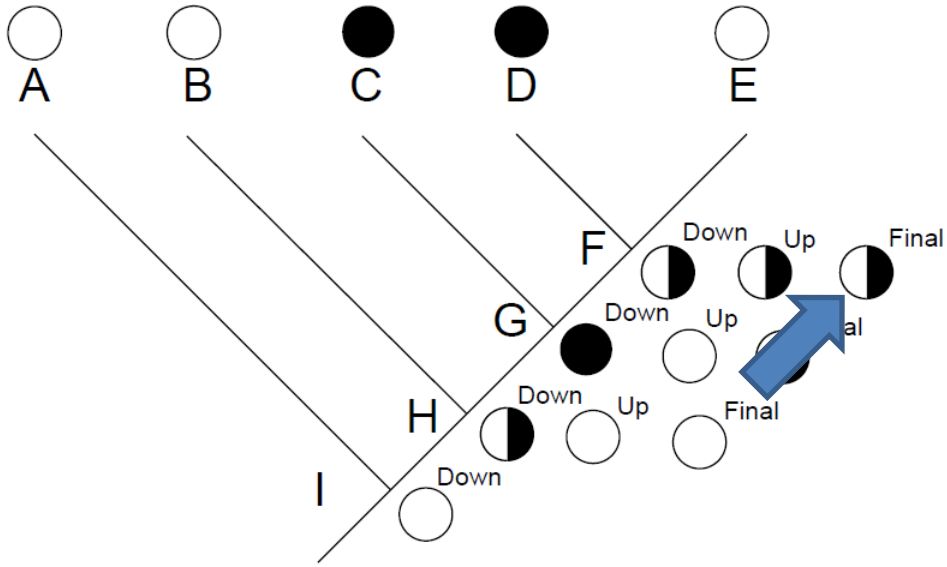
- Dans un dernier passage, nous allons considérer les affectations réalisées lors des deux premiers passages



- Prenons le nœud F:
  - État de 'up' de F
  - Etats de 'down' de ses descendants D et E
  - Choisir l'état majoritaire dans ces trois ensembles

Reconstruction d'états ancestraux

- Dans un dernier passage, nous allons considérer les affectations réalisées lors des deux premiers passages



Reconstruction d'états ancestraux

- Défauts et limites :

- **Variation des vitesses d'évolution** :
  - la méthode de Fitch suppose que les changements entre tous les états des caractères ont les mêmes chances de se produire
  - hypothèse souvent non réaliste (ex. transition versus transversions au niveau de l'ADN)
  - défaut « corrigé » par les méthodes de parcimonie pondérée (Sankoff, 1975)

- **Evolution rapide** :
  - l'hypothèse d'évolution minimum implique que les changements d'état sont rares. Contrainte forte peu réaliste en pratique

Reconstruction d'états ancestraux

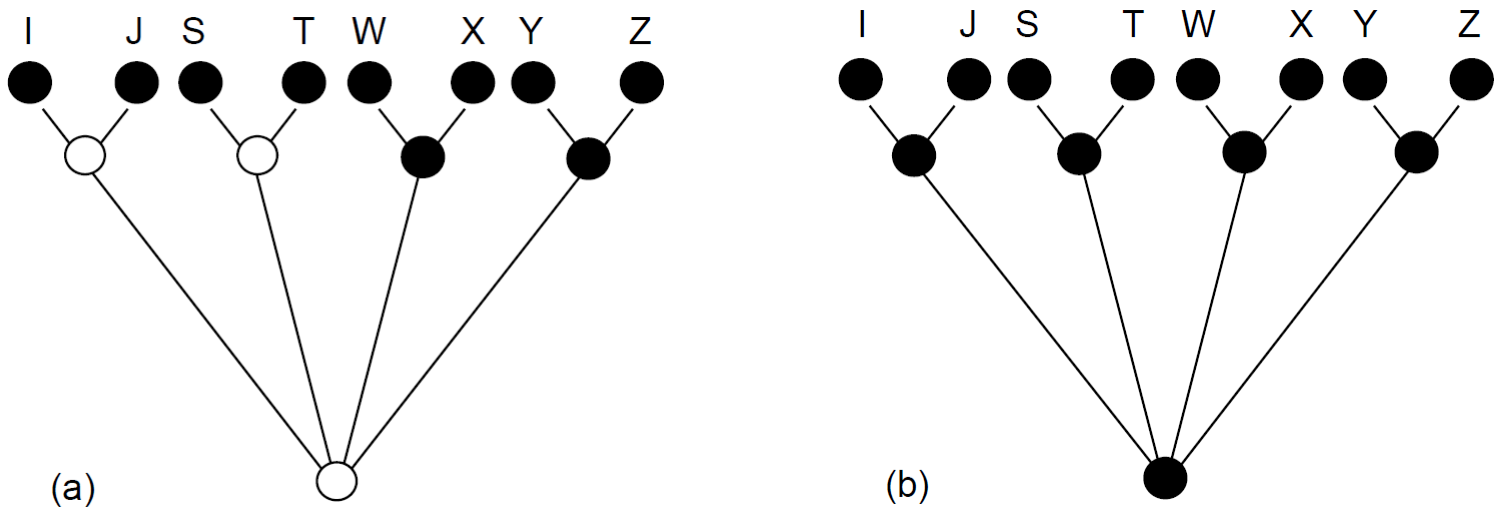
- Défauts et limites :

La figure montre un cas dans lequel une combinaison

d'évolution rapide

les probabilités de gains et de pertes inégales

conduisent à des erreurs dans la reconstruction des états ancestraux.



Dans cette phylogénie (T7) les vrais ancêtres sont connus (a phylogénie réelle, b estimation par MP),

=>évaluer la précision de la reconstruction par parcimonie.

Reconstruction d'états ancestraux

- Défauts et limites :

- **Variation des vitesses d'évolution** :
  - la méthode de Fitch suppose que les changements entre tous les états des caractères ont les mêmes chances de se produire
  - hypothèse souvent non réaliste (ex. transition versus transversions au niveau de l'ADN)
  - défaut « corrigé » par les méthodes de parcimonie pondérée (Sankoff, 1975)

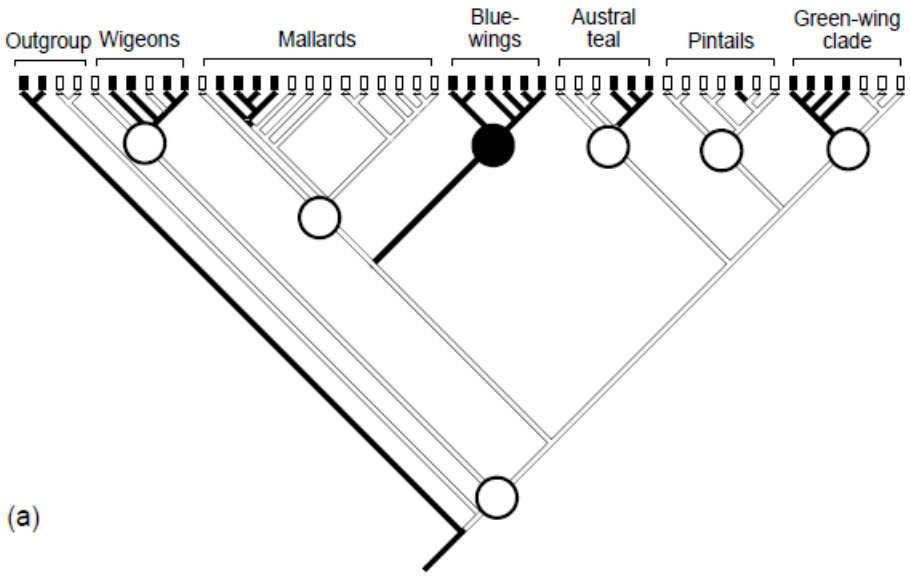
- **Evolution rapide** :
  - l'hypothèse d'évolution minimum implique que les changements d'état sont rares. Contrainte forte peu réaliste en pratique

- **Variation du temps écoulé le long des branches** :
  - ces méthodes ne prennent pas en compte les longueurs des branches de l'arbre, ce qui conduit à une sous-estimation du nombre de changements le long des branches longues

- **Pas de justification statistique** :
  - l'incertitude sur la reconstruction des états ancestraux n'est pas correctement estimée

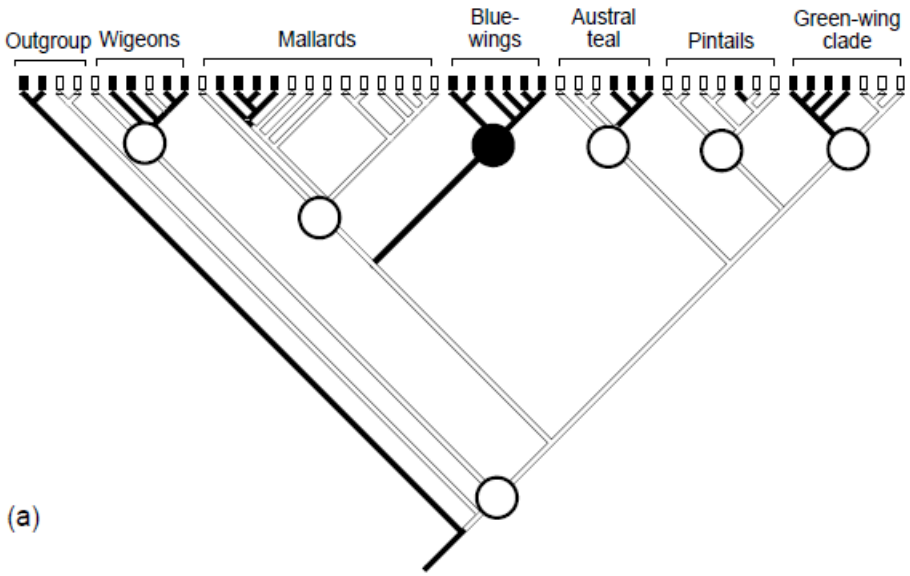
MP pondérée

- Evolution du dimorphisme du plumage chez les canards barboteurs (Anas).
  - blanc: espèces monochromatiques,
  - noir : espèces dichromatiques
  - Gris : ambiguës.
- Reconstruction états ancêtres par MP des ancêtres
  Etats ancestraux indiqués dans les cercles pour les six principaux groupes de canards et pour leur ancêtre commun.
  ombrage dans les branches : états ancestraux pour les autres nœuds



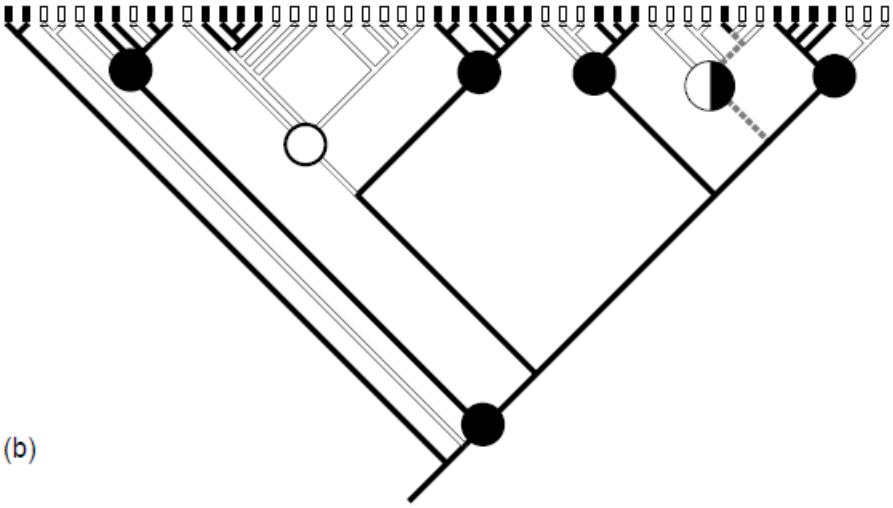
MP pondérée

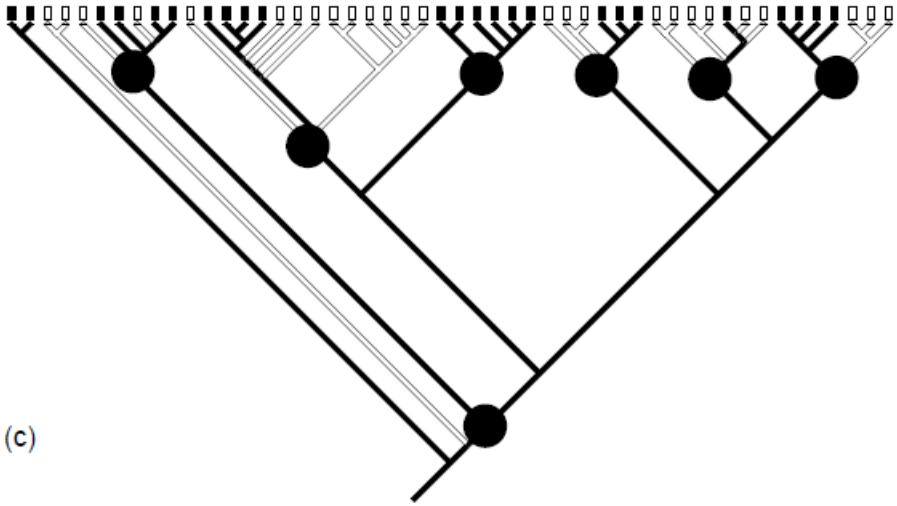
- MP avec des poids égaux pour les gains et les pertes.
- Cette reconstruction contredit l'hypothèse largement répandue que le dichromatisme a été perdu plusieurs fois.



MP pondérée

- b) gains trois fois plus fréquents que les pertes,
  - le nombre d'origines indépendantes du dichromatisme diminuent.
- c) gains cinq fois plus probable que les pertes,
  - le nombre d'origines indépendantes correspond a l'hypothèse privilégiée par les biologistes.

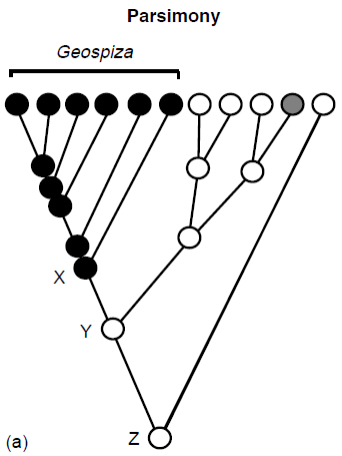




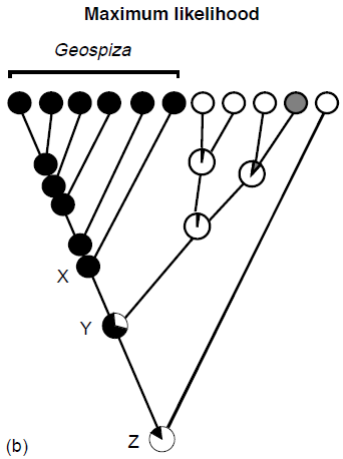
MP et ambiguitées

- Exemple : reconstruire le comportement alimentaire des pinsons (Geospiza).
  - noir : granivores,
  - blancs : insectivores,
  - gris : folivores.

(a) Reconstruction par MP. Toutes les reconstructions sont sans ambiguïté et impliquent que granivory est née dans l'ancêtre commun du clade Geospiza.



b) Reconstitution par ML. Suggère que granivory est née avant la diversification de Geospiza.
Notez que la reconstruction ML est moins parcimonieuse!)



Maximum de vraisemblance

Les méthodes de ML de reconstruction d'état ancestral

- traitent les états de caractère aux noeuds internes du arbre comme **paramètres**
- tentent de trouver les **valeurs de paramètres** qui maximisent la probabilité de les données (les états de caractère observés) étant donné l'hypothèse
  - un modèle d'évolution
  - une phylogénie reliant les séquences observées

Premières approches ont étés développées dans le contexte

- de l'évolution de la séquence génétique
- de l'évolution des caractères discrets.

Ces approches utilisent le même cadre probabiliste.

Maximum de vraisemblance

En bref,

l'évolution d'une séquence génétique est modélisée par un processus de Markov en temps continus.

Dans le plus simple, tous les caractères subissent des transitions d'état

- indépendantes
- à un taux constant au cours du temps.

Souvent étendu pour permettre

- des taux différents sur chaque branche de l'arbre
- des taux variant au cours du temps (changements environnementaux)

Un modèle définit les probabilités de transition des états i à j le long d'une branche de longueur t (temps en unités d'évolution).

La probabilité d'une phylogénie est calculée comme la somme emboitée des probabilités de transition

$\Rightarrow$ correspond à la structure hiérarchique de l'arbre proposé.

Maximum de vraisemblance

A chaque nœud, la probabilité de ses descendants est la somme sur tous les états de caractères ancestraux possibles à ce nœud:

La vraisemblance du sous arbre enraciné au **noeud *x*** avec des descendants directs *y* et *z*,

$$L_x = \sum_{S_x \in \Omega} P(S_x) \left(\sum_{S_y \in \Omega} P(S_y | S_x, t_{xy})\, L_y \sum_{S_z \in \Omega} P(S_z | S_x, t_{xz})\, L_z \right)$$

Notation :

S_i indique l'état de caractère du *i*-ième nœud,

t_{ij} est la longueur de la branche (temps évolutif) entre les nœuds *i* et *j*,

Ω est l'ensemble de tous les états de caractères possibles (ex, A, C, G et T).

- Ainsi, l'objectif de la reconstruction ancestrale est trouver les états S_x pour tous les *x* nœuds internes qui maximisent la probabilité des données observées pour un arbre donné.

Rappels sur le maximum de vraisemblance

Spécification du modèle
Fonction de densité de probabilité

En statistique, le vecteur de données $y = (y_1, \ldots, y_m)$ est un échantillon aléatoire d'une population inconnue.

L'objectif de l'analyse des données est d'identifier la population qui est la plus susceptible d'avoir généré cet échantillon.

En statistique, chaque population est identifiée par un **distribution de probabilité** correspondante. Associé à chaque distribution de probabilité est une valeur unique de la paramètre du modèle. Lorsque le paramètre change de valeur, différentes distributions de probabilité sont générées. Officiellement, un modèle est défini comme la famille de probabilité distributions indexées par les paramètres du modèle.

Fonction de densité de probabilité

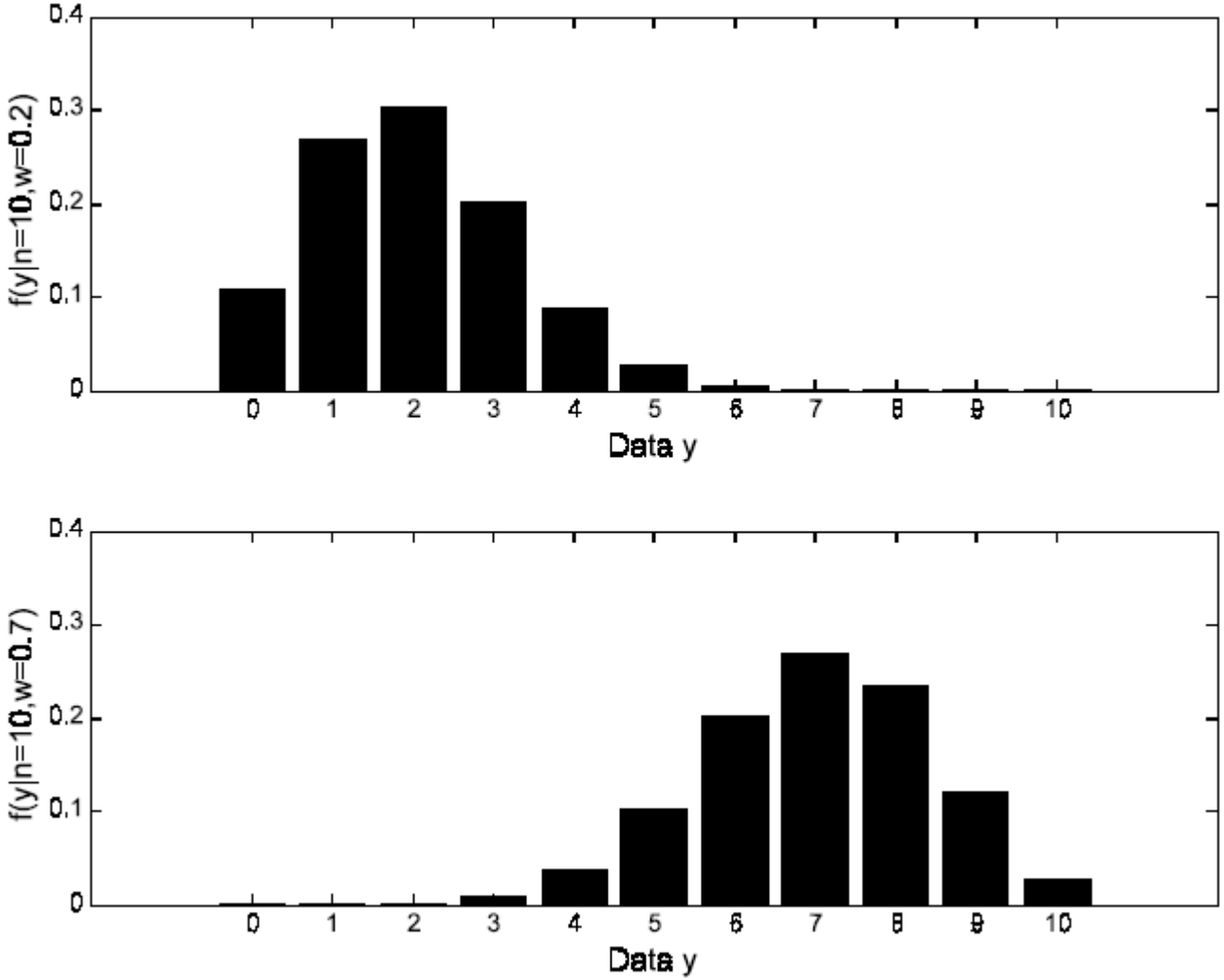
Spécification du modèle
Fonction de densité de probabilité (FDP)

- Pour illustrer l'idée d'une FDP,
  - cas avec une observation et un paramètre, $m = k = 1$:

  Supposons
  - que les données y représentent le nombre de succès dans une séquence de 10 lancer une pièce de monnaie (Bernoulli)
  - que la probabilité de un succès sur un essai, représenté par le paramètre w, est 0,2.
  - La PDF dans ce cas est donné par :

$$f(y \mid n = 10, w = 0.2) = \frac{10!}{y!(10-y)!}(0.2)^y(0.8)^{10-y}$$
$$(y = 0, 1, \ldots, 10)$$

Fonction de densité de probabilité



Les distributions binomiales des probabilités pour un échantillon de taille n = 10

Paramètre de probabilité w = 0: 2 (en haut) et w = 0: 7 (en bas).

Fonction de densité de probabilité

Généralisation pour des valeures arbitraires de *w* et *y*:

$$f(y|n,w) = \frac{n!}{y!(n-y)!} w^y (1-w)^{n-y}$$
$$(0 \leqslant w \leqslant 1; \quad y = 0, 1, \ldots, n)$$

Spécifie la probabilité des données *y* pour des valeurs données des paramètres *n* et *w*.

La collection de tous ces FDP générées en faisant varier le paramètre à travers son intervalle (0-1 pour w et n>=1) **définit un modèle**.

Fonction de vraisemblance

Étant donné un ensemble de valeurs de paramètres, les FDP vont montrer que certaines données sont plus probables que autres.

Dans l'exemple précédent, le PDF avec w = 0: 2; y = 2 est plus susceptible de se produire que y = 5 (0.302 vs.0,026).

En réalité, nous avons déjà observé les données.

Nous sommes confrontés au **problème inverse**:

- Compte tenu des **données** observées et d'un **modèle**, trouver la FDP qui parmi toutes les celles du modèle est le plus susceptible d'avoir généré les données.

Problème inverse

Pour résoudre ce problème inverse,
Définition d'une fonction de vraisemblance en inversant les rôles de le vecteur de données *y* et le vecteur de paramètres *w* :

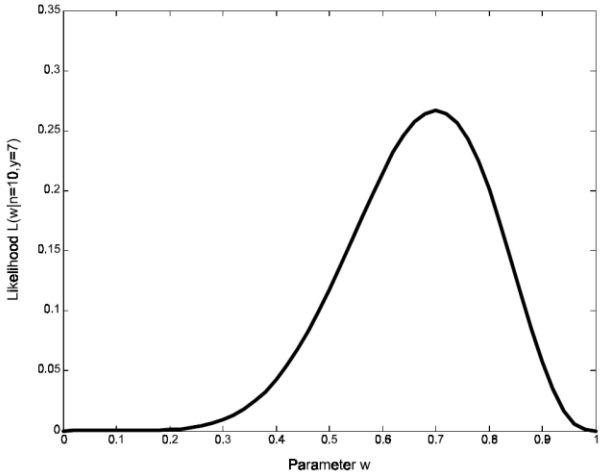
$$L(w|y) = f(y|w).$$

$L(w|y)$ est la vraisemblance du paramètre w connaissant les données observées y

$L(w|y)$ est donc une fonction de w.

Pour $y = 7$ et $n=10$:

$$L(w \mid n = 10, y = 7) = f(y = 7 \mid n = 10, w)$$
$$= \frac{10!}{7!3!} w^7 (1 - w)^3 \quad (0 \leqslant w \leqslant 1).$$



Problème inverse

FDP et FV sont définies sur différents axes, et ne sont donc pas directement comparables.

- FDP est en fonction des données (*y*) donnant un ensemble particulier de valeurs de paramètres (*w*), définies sur l'échelle de données.
- FV est une fonction des paramètre (*w*) connaissant un ensemble de données observées (*y*), définies sur échelle des paramètres.

- FDP : probabilité d'observer une valeur de donnée particulière pour un paramètre fixe.
- FV : vraisemblance d'une valeur de paramètre particulière pour un ensemble de données fixe.

Notez que la fonction de vraisemblance dans cette figure est une courbe.

Si le modèle a deux paramètres, le FV sera une surface au-dessus de l'espace des paramètres.

En général, pour un modèle avec k paramètres, la fonction de vraisemblance prend la forme d'une «surface» géométrique à k-dim au-dessus d'un hyperplan k-dim engendré par le vecteur de paramètres.

Estimation du maximum de vraisemblance

- But : trouver la valeur des paramètres qui correspondent à la distribution de probabilité souhaitée.

- Le principe de l'estimation du maximum de vraisemblance (MLE) indique que la distribution de probabilité souhaitée est celui qui rend les données observées «**les plus vraisemblable**»

=> on doit trouver la valeur du vecteur de paramètres qui maximise la fonction de vraisemblance.

Ce vecteur est appelé l'estimation MLE, et est noté $w_{MLE} = (w_{1,MLE}, \ldots, w_{k,MLE})$

Sur l'exemple, l'estimation MLE est $w_{MLE} = 0{:}\ 7$

Pour résumer, l'estimation du maximum de vraisemblance est une méthode pour chercher la distribution de probabilité qui rend les données observées les plus vraisemblable.

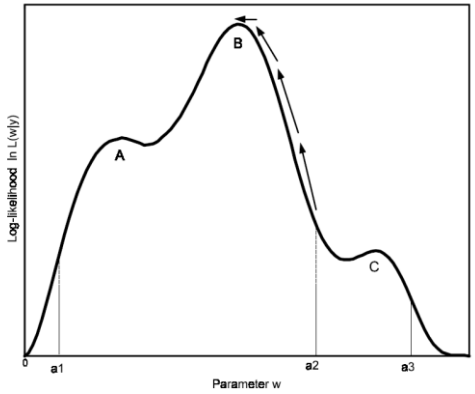
Estimation du maximum de vraisemblance

- En pratique, il n'est généralement pas possible de obtenir une solution de **forme analytique** pour l'estimation MLE, surtout quand le modèle implique de nombreux paramètres et son PDF est généralement non linéaire.

- L'estimation MLE doit être recherchée numériquement en utilisant des algorithmes d'optimisation non linéaire.

- L'idée de base est de trouver rapidement des paramètres optimaux qui maximise la log-vraisemblance.
  - recherche de sous-ensembles beaucoup plus petits plutôt que de faire une cherche exhaustive.
  - procède itérativement par essais/erreurs.

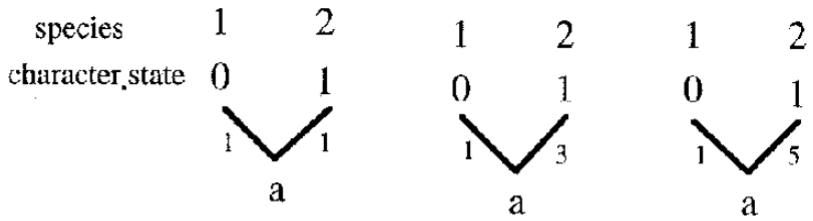
  A chaque itération un nouvel ensemble de valeurs de paramètres est obtenu en ajoutant de petites modifications aux paramètres précédant de telle sorte que les nouveaux paramètres sont susceptible d'entraîner une amélioration des performances.



Minimum locaux!

Estimer l'état ancestral de deux espèces

- Cas le plus simple : estimer l'état ancestral de deux espèces.



- Observations aux feuilles : $d = \{d1, d2\}$ ici $d = \{0, 1\}$
- Etat ancestral au nœud : $n = \{a\}$

- Hypothèses :
  - l'arbre et la longueur de ses branches sont connues.
  - L'état du caractère au nœud n'est pas connu. Il constitue un des paramètres à estimer.

Estimer l'état ancestral de deux espèces

- Soit ***m*** un modèle d'évolution d'un caractère qui peut adopter deux états (0, 1) avec des transitions réversibles de 1-> 0 et 0 -> 1.
- Il peut être modéliser par un **processus de Markov en temps continu** qui représente le probabilité d'une transition de caractère en fonction de deux paramètres : le taux de transition et le temps.

- Soit $P_{ij}(t) = m(\alpha, \beta, t)$ représente la probabilité d'une transition du caractère de l'état *i* à l'état *j* le long d'une branche de longueur *t*.
- Le paramètre α est le taux de transition de l'état 0 à l'état 1, et β est la transition de état 1 à l'état 0.
- => quatre probabilités possibles correspondant aux états de début et de fin de chaque branche de la phylogénie.

	State at end of branch	
State at beginning of branch	0	1
0	$P_{00}(t) = 1 - P_{01}(t)$	$P_{01}(t)$
1	$P_{10}(t) = 1 - P_{11}(t)$	$P_{11}(t)$

Estimer l'état ancestral de deux espèces

- Soit $P(d|m)$ (noté f($y|w$) plus haut) la probabilité d'observées les données étant donné le modèle d'évolution.
  Calculer $P(d|m)$ nécessite une estimation les deux paramètres de taux de transition (α, β) du modèle m, les longueurs de branche étant connues.

- L (m; d) (noté $L(w|y)$ plus haut) est la probabilité d'observer les données.

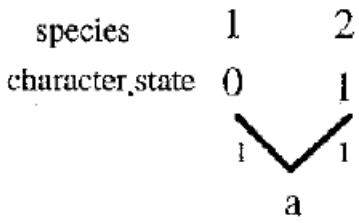
$$L(m; d) \propto P(d|m)$$

- Note : $\propto$ signifie "est proportionnel à«

Par commodité, L (m; d) peut être abrégé en L (m) .

Estimer l'état ancestral de deux espèces

- La probabilité des données de la figure 1 est trouvé à partir de



$$
\begin{aligned}
L(m) &\propto P(d|m) \\
&= \sum_{a=0}^{1} w(a)P(d|m, a) \\
&= \sum_{a=0}^{1} w(a)(P_{a0}(t) \,.\, P_{a1}(t)) \qquad (2) \\
&= w(0)(P_{00}(t) \,.\, P_{01}(t)) + w(1)(P_{10}(t) \,.\, P_{11}(t))
\end{aligned}
$$

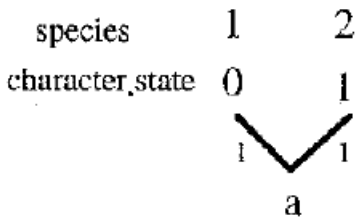
La probabilité d'observer les caractères aux feuilles sont la somme de deux termes qui sont le produit des deux probabilités

- nœud *a* est 0,
  - pas de transition pour passer à feuille 1
  - mais transition 0-> 1 pour passer à feuille 2
- nœud *a* est 1,
  - transition 1-> 0 pour passer à feuille 1
  - pas de transition pour passer à feuille 2

Ces deux alternatives pour le nœud *a* sont pondérées par leur probabilité d'occurrence antérieure $w(a)$.
En l'absence de toute autre information, les états alternatifs sont également probables et $w(a) = 0{,}5$.

Estimer l'état ancestral de deux espèces

- La probabilité des données de la figure 1 est trouvé à partir de



$$
\begin{aligned}
L(m) &\propto P(d|m) \\
&= \sum_{a=0}^{1} w(a) P(d|m, a) \\
&= \sum_{a=0}^{1} w(a)(P_{a0}(t) \,.\, P_{a1}(t)) \qquad (2) \\
&= w(0)(P_{00}(t) \,.\, P_{01}(t)) + w(1)(P_{10}(t) \,.\, P_{11}(t))
\end{aligned}
$$

La probabilité d'observer les caractères aux feuilles sont la somme de deux termes qui sont le produit des deux probabilités

- nœud *a* est 0,
  - pas de transition pour passer à feuille 1
  - mais transition 0-> 1 pour passer à feuille 2
- nœud *a* est 1,
  - transition 1-> 0 pour passer à feuille 1
  - pas de transition pour passer à feuille 2

Ces deux alternatives pour le nœud *a* sont pondérées par leur probabilité d'occurrence antérieure $w(a)$.
En l'absence de toute autre information, les états alternatifs sont également probables et $w(a) = 0{,}5$.

Estimer l'état ancestral de deux espèces

- Utilisation du logiciel Discrete (Pagel, 1994) pour effectuer les calculs.

- Si les deux branches sont de longueur égale à 1,0
- Les taux de transitions sont estimés à :

$$\hat{\alpha} = \hat{\beta} = 8.0.$$

- Avec ces estimations toutes les probabilités (cf tableau) sont égales à 0,50

- => la probabilité globale est (0,5 0,5) + (0,5 0,5), qui est multiplié par la probabilité à priori de 0,5.

Etat ancestral le plus probable

- L'état ancestral le plus probable *L*(*m*) a été trouvé en maximisant sur les deux états à la "racine".
- Cela supprime la racine (ou plus généralement tout nœud ancestral) de la probabilité et rend la probabilité indépendante de toute valeur particulière.

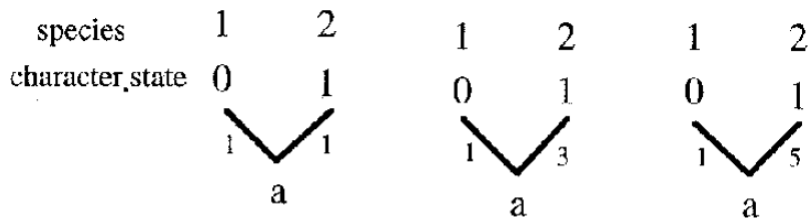
- Pour estimer la vraisemblance des deux états ancestraux possibles, deux vraisemblances :

$$L(m, a = i) \propto w(a = i)P(d|m, a = i)$$

- où i = 0, 1 correspond à chacun des deux états possibles au nœud *a*.
- Les deux vraisemblances sont estimées séparément, ayant fixé la racine à *a* = 0 ou *a* = 1 et ré-estimer les paramètres de *m*;

- La somme des deux vraisemblances n'est pas nécessairement égales à *L*(*m*).
- Les probabilités les états ancestraux sont estimés de cette façon car l'attribution d'une valeur à la racine (plus généralement à n'importe quel nœud) implique un ensemble différent de transition (les *α* et *β* de *m*).

Etat ancestral le plus probable



$\hat{L}(d)$	0.250	0.256	0.296
$\hat{L}(a = 0 / d)$	0.125	0.236	0.291
$\hat{L}(a = 1 / d)$	0.125	0.125	0.125

- Lorsque les longueurs de branches sont égales, la vraisemblance que l'ancêtre soit dans l'état 0 ou 1 est la même (bon sens!).

- Si les branches ont des longueurs différentes, à la fois l'intuition et la vraisemblance suggèrent une préférence pour l'état du caractère à la racine sur

Méthode marginale *versus* méthode jointe

- Phylogénie : calculer la probabilité globale pour les arbres alternatifs.
- Reconstruction des états ancestraux : trouver la combinaison des états des caractères à chaque nœud ancestral avec la vraisemblance maximum (ML) marginale la plus élevé.

- Deux approches pour résoudre ce problème.
  1. Partir des descendants d'un arbre pour assigner progressivement les états les plus probables au caractère de chaque ancêtre, en ne prenant en considération que ses descendants immédiats.

     - Approche appelée **reconstruction marginale**.

     Algorithme glouton qui fait des choix localement optimaux à chaque étape du problème d'optimisation.

     Bien qu'il puisse être très efficace, il n'est pas garanti d'atteindre une solution globalement optimale.

  2. Trouver la combinaison d'états de caractères ancestraux dans l'arbre qui maximise conjointement la vraisemblance des données.

     - Approche st appelée **reconstruction jointe**.

     Elle n'est pas aussi rapide que la reconstruction marginale, mais elle moins de chance d'être prise dans un optimum local.

Une reconstruction marginale peut affecter un état au caractère de l'ancêtre immédiat qui est localement optimal, mais qui écarte la distribution conjointe des états de caractères ancestraux loin de l'optimum global.

Sans surprise, la reconstruction jointe est plus complexe en termes de calcul que la reconstruction marginale.

Mais les algorithmes pour la reconstruction jointe ont été développés avec une complexité de temps qui est généralement linéaire avec le nombre de taxa ou de séquences observées.

Forces et faiblesses de ML

- Forces
- Les méthodes ML tendent à fournir une **plus grande précision** que les méthode MP en présence de variation des taux d'évolution entre les caractères (ou entre les sites dans un génome).

- Faiblesses
- Ces méthodes ne sont pas encore en mesure de s'adapter à la variation des taux de l'évolution au fil du temps (hétérotachie).
  - Si le taux d'évolution pour un caractère spécifique s'accélère sur une branche de la phylogénie, avec un modèle de taux constant d'évolution pour ce caractère, la vitesse d'évolution sera sous-estimée sur cette branche.

- Avec ML (contrairement à MP) l'utilisateur doit spécifire un modèle d'évolution. Un modèle incorrect à toute les chances d'affecter la qualité des inférences!
- ML ne peut fournir qu'une seule reconstruction des états de caractères («estimation ponctuelle»)
  - Problème lorsque la surface de vraisemblance est fortement non convexe, comprenant plusieurs pics (optima local). Dans ce cas une approche bayésienne peut être plus appropriée.

Inférence bayésienne

- Inférence bayésienne
  L'approche consiste a actualiser les croyances *a priori* sur un évènement à la lumière des nouvelles informations disponibles pour obtenir une description quantitative des connaissances actuelles.

- Comment l'expérience doit changer notre opinion sur un paramètre θ? Cela suppose :
  - Une opinion *a priori* sur les différentes valeurs plausibles du paramètre avant l'expérience,
  - L'information issue de l'expérience (-> Vraisemblance),
- L'intégration de ces deux informations nous donne :
  - L'opinion a posteriori sur θ.

La transformation de l'opinion *a priori* à partir des données (la fonction de vraisemblance) en une opinion *a posteriori* se fait à l'aide du **théorème de Bayes** :

$$P(\theta|y) \propto P(y|\theta)P(\theta)$$

$$\text{A Posteriori} \propto \text{Vraisemblance} * \text{A Priori}$$

Théorème de Bayes

- **Théorème de Bayes**

$$P(B|A) = \frac{1}{P(A)} P(A|B)P(B)$$

- En appliquant à un paramètre θ et des données y :

$$P(\theta|y) = \frac{1}{P(y)} P(y|\theta)P(\theta)$$

- En appliquant a une hypothèse nulle H0 sur θ :

$$P(H_0|y) = \frac{1}{P(y)} P(y|H_0)P(H_0)$$

- P(y) = Constante

$$P(\theta|y) \propto P(y|\theta)P(\theta)$$

$$\text{A Posteriori} \propto \text{Vraisemblance} * \text{A Priori}$$

Inférence bayésienne

- Inférence bayésienne
- Dans le contexte de la reconstruction des états ancestraux, l'objectif est de déduire les probabilités *a posteriori* des états de caractères ancestraux à chaque nœud d'un arbre donné.

- De plus, on peut intégrer ces probabilités sur les distributions postérieures, sur les paramètres du modèle d'évolution et dans l'espace de tous les arbres possibles.
- Ceci peut être exprimé comme une application du théorème de Bayes:

$$P(S|D,\theta) = \frac{P(D|S,\theta)\ P(S|\theta)}{P(D|\theta)}$$
$$\propto P(D|S,\theta)\ P(S|\theta)\ P(\theta),$$

- où S représente les états ancestraux, D correspond aux données observées, et θ représente à la fois le modèle évolutif et l'arbre phylogénétique.
- $P(D|S,\theta)$ est la vraisemblance des données observées.
- $P(S|\theta)$ est la probabilité *a priori* des états ancestraux pour un modèle et un arbre donnés.
- $P(D|\theta)$ est la probabilité des données pour un modèle et un arbre donnés, intégrés sur tous les états ancestraux possibles.

Inférence bayésienne

- Bayes empirique et hiérarchique
  Développé par Yang et ses collègues, où les estimations du modèle évolutif et de l'arbre obtenus par ML ont été utilisés pour définir les distributions *a priori*.

- La méthode de Bayes empirique est donc proche des reconstructions ancestrales basées ML sauf qu'a la place d'assigner les états en fonction de leur probabilité, ce sont les distributions de probabilité qui sont rapportées directement.

- Faiblesse :
  - La méthode requière que les paramètres du modèle évolutif et que l'arbre soient connus sans erreur.

- Quand la complexité des données rend cette hypothèse irréaliste, il peut être plus prudent d'adopter l'approche bayésienne entièrement hiérarchique et déduire la distribution postérieure conjointe à partir des états de caractères ancestraux, du modèle et de l'arbre .

- **L'approche empirique** calcule les probabilités de diverses états ancestraux pour un arbre spécifique et un modèle d'évolution. En exprimant la reconstruction des ancêtres comme un ensemble de probabilités, on peut directement quantifier l'incertitude pour l'attribution d'un état à un ancêtre.

- **L'approche hiérarchique** moyenne ces probabilités sur tous les arbres possibles et les modèles d'évolution, proportionnellement à la probabilité de ces arbres et de ces modèles sur les données observées. En pratique, cette approche se limite à l'analyse d'un petit nombre de séquences ou de taxons, car l'espace de tous les arbres possibles devient rapidement trop vaste.

Méthode stochastique

- La méthode stochastique permet d'échantillonner les états des nœuds et les histoires des caractères à partir de leur distribution bayésienne à postériori (empirique ou hiérarchique) (Huelsenbeck *et al.*, 2003 ; Bollback, 2006)

- L'algorithme de reconstruction par la méthode stochastique est le suivant :
  1. Calcul de la vraisemblance conditionnelle pour chaque état de caractère à chaque nœud interne de l'arbre.
  2. Simulation des états ancestraux à chaque nœud interne de l'arbre en échantillonnant à partir des distributions à posteriori des états (à partir des résultats de l'étape 1).
  3. Simulation d'une histoire de substitution des états le long de chaque branche en échantillonnant à partir de la distribution à posteriori conditionnée par les reconstructions à l'étape 2 et les états observés aux feuilles de l'arbre.

  L'estimation des paramètres et la reconstruction des états ancestraux peuvent être réalisées en utilisant le maximum de vraisemblance comme décrit dans l'algorithme, ou en utilisant la méthode MCMC. Cependant, cette dernière requiert un temps de calcul excessivement long!.

Modèles

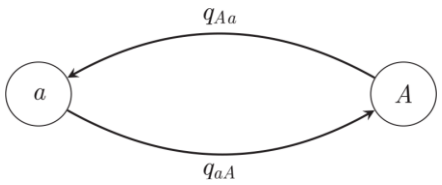
- Modèles
  De nombreux modèles ont été développés pour estimer les états ancestraux de caractères discrets et continus.

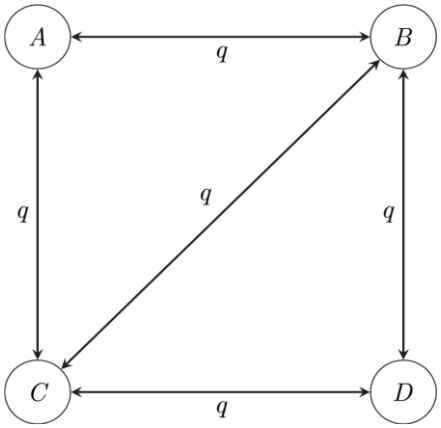
- Ces modèles supposent que l'évolution d'un trait à travers le temps peut être modélisé comme un processus stochastique.
- Pour les caractères à **valeur discrète** , ce processus est généralement considéré comme une chaîne de Markov,
- Pour les caractères à **valeur continue** (taille), le processus est souvent considéré comme un mouvement brownien ou un processus Ornstein-Uhlenbeck.

Modèles discret

- Modèles à états discrets
  Supposons que le trait en question puisse tomber dans l'un des k états, étiquetés 1 ,. . ., k.

- Modélisation via une chaîne de Markov en temps continu :
  - Chaque état est associé à des taux de transition vers les autres états.
  - Chaque état peut passer d'un état à l'autre au cours du temps.



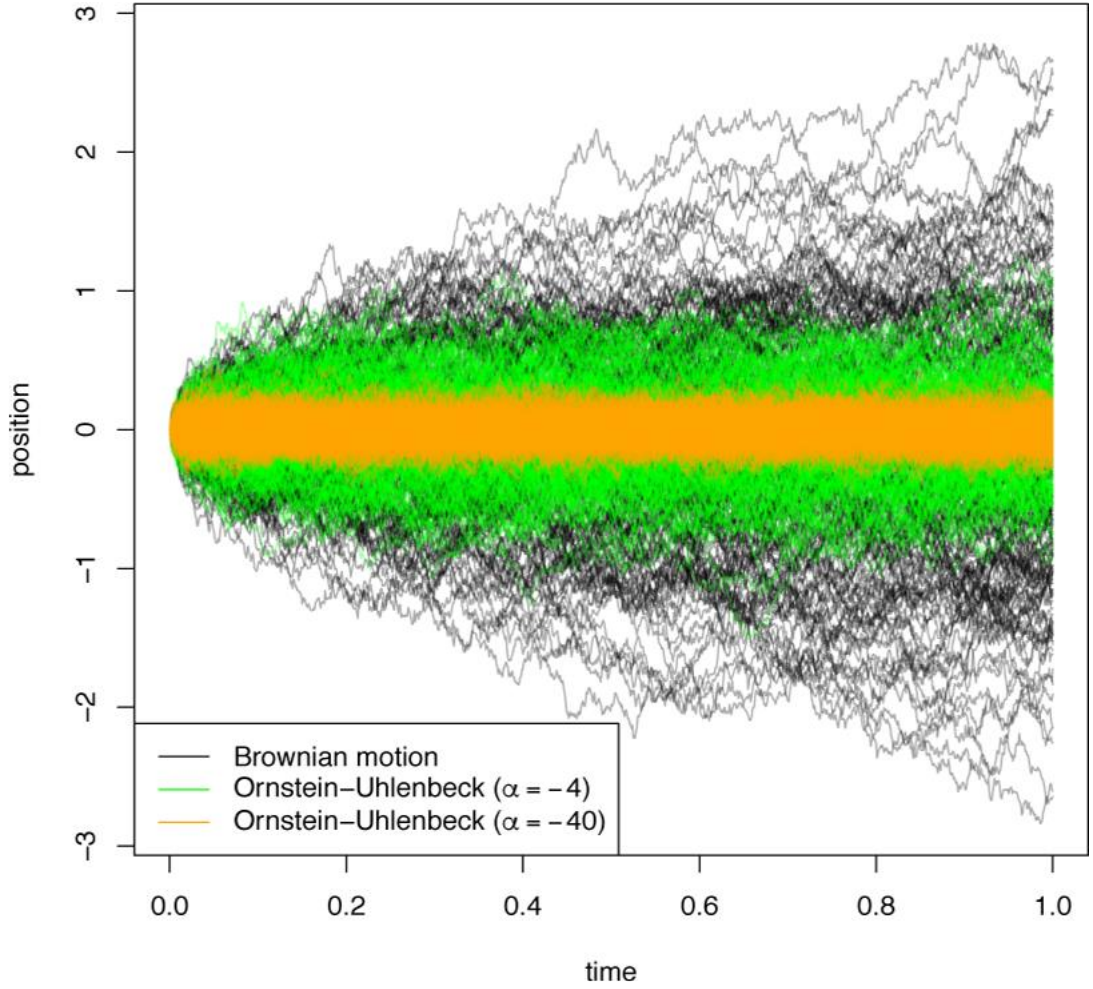


  - Les taux de transition *q* peuvent être estimée en utilisant, par exemple, des méthodes de ML.

Modèles à état continu

- Modèles à état continu
  L'inférence des états ancestraux par ML (ou par MB) procède comme pour les états discrets mais avec les probabilités de transitions d'état entre nœuds adjacents donné par une autre distribution de probabilité continue.
-
  - **Mouvement brownien**:
    - Si les nœuds U et V sont adjacents dans la phylogénie et séparés par une branche de longueur *t*, la probabilité d'une transition de U étant dans l'état *x* à V étant dans l'état *y* est donné par une densité gaussienne avec la moyenne 0 et la variance σ2t.
      - un seul paramètre (σ2),
      - le modèle suppose que le trait évolue librement sans un biais vers l'augmentation ou la diminution,
      - le taux de changement est constant dans toutes les branches de l'arbre phylogénétique.
-
  - **Processus d'Ornstein-Uhlenbeck**: se comporte comme un mouvement brownien, mais attiré vers une valeur centrale, où la force de l'attraction augmente avec la distance de cette valeur.
    - utile pour modéliser des scénarios où le trait est soumis à une sélection stabilisante autour d'une certaine valeur (ex 0).
    - deux paramètres: σ2 et α, qui décrit la force de l'attraction à 0.
    - Si α tend vers 0, le processus est de moins en moins contraint par son attraction à 0 et le processus devient un Mouvement brownien.

Modèles à état continu



Tracés de 200 trajectoires des
mouvement brownien avec σ2 = 1 (noir);
Ornstein-Uhlenbeck avec σ2 = 1 et α = -4 (vert);
et Ornstein-Uhlenbeck avec σ2 = 1 et α = -40 (orange).

Name	Methods	Platform	Supported Input Formats	Character Types	Continuous (C) or Discrete (D) Characters	Software Licence
PAML	ML	Unix, Mac, Win	PHYLIP, NEXUS, FASTA	Nucleotide, Protein	D	Proprietary
BEAST2	Bayesian	Unix, Mac, Win	NEXUS, BEAST XML	Nucleotide, Protein, Geographic	C, D	GNU Lesser General Public License
APE	ML	Unix, Mac, Win	NEXUS, FASTA, CLUSTAL	Nucleotide, Protein	C, D	GNU General Public License
Diversitree	ML	Unix, Mac, Win	NEXUS	Qualitative and quantitative traits, Geographic	C, D	GNU General Public License, version 2
HyPhy	ML	Unix, Mac, Win	MEGA, NEXUS, FASTA, PHYLIP	Nucleotide, Protein (customizable)	D	GNU Free Documentation License 1.3
BayesTraits	Bayesian	Unix, Mac, Win	TSV or space delimited table. Rows are species, columns are traits.	Qualitative and quantitative traits	C, D	Creative Commons Attribution License
Lagrange	ML	Linux, Mac, Win	TSV/CSV of species regions. Rows are species and columns are geographic regions	Geographic	-	GNU General Public License, version 2
Mesquite	Parsimony, ML	Unix, Mac, Win	Fasta, NBRF, Genbank, PHYLIP, CLUSTAL, TSV	Nucleotide, Protein, Geographic	C, D	Creative Commons Attribution 3.0 License
Phylomapper	ML, Bayesian (version 2)	Unix, Mac, Win	NEXUS	Geographic, Ecological niche	C, D	-
Ancestors	ML	Web	Fasta	Nucleotide (indels)	D	-
Phyrex	Maximum Parsimony	Linux	Fasta	Gene expression	C, D	Proprietary
SIMMAP	SM	Mac	XML-like format	Nucleotide, qualitative traits	D	Proprietary
MrBayes	Bayesian	Unix, Mac, Win	NEXUS	Nucleotide, Protein	D	GNU General Public License
PARANA	Maximum Parsimony	Unix, Mac, Win	Newick	Biological networks	D	Apache License
PHAST (PREQUEL)	ML	Unix, Mac, Win	Multiple Alignment	Nucleotide	D	BSD License
RASP	ML, Bayesian	Unix, Mac, Win	Newick	Geographic	D	-
VIP	Maximum Parsimony	Linux, Win	Newick	Geographic	D (grid)	GPL Creative Commons
FastML	ML	Web, Unix	Fasta	Nucleotide, Protein	D	Copyright
MLGO	ML	Web	Custom	Gene order permutation	D	GNU
BADGER	Bayesian	Unix, Mac, Win	Custom	Gene order permutation	D	GNU GPL version 2
COUNT	Maximum Parsimony	Unix, Mac, Win	Tab-delimited text file of rows for taxa and count data in columns	Count data (homolog family size)	D	BSD
MEGA	Maximum parsimony, ML	Mac, Win	MEGA	Nucleotide, Protein	D	Proprietary
ANGES	Local Parsimony	Unix	Custom	Genome maps	D	GNU General Public License, version 3
EREM	ML	Win, Unix, Matlab module	Custom text format for model parameters, tree, observed character values	Binary	D	None specified, although site indicates software is freely available