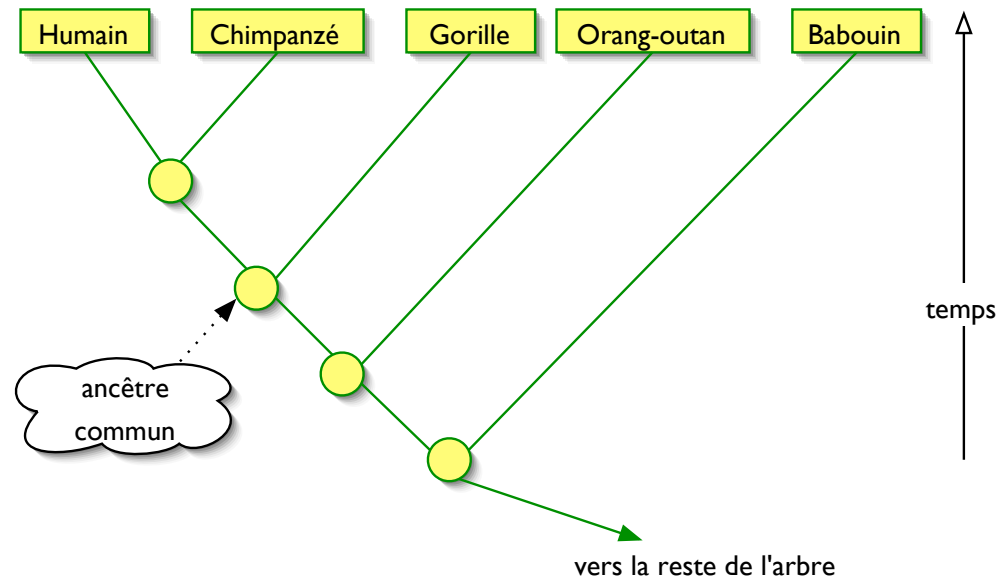


[RE]CONSTRUCTION DE PHYLOGÉNIES

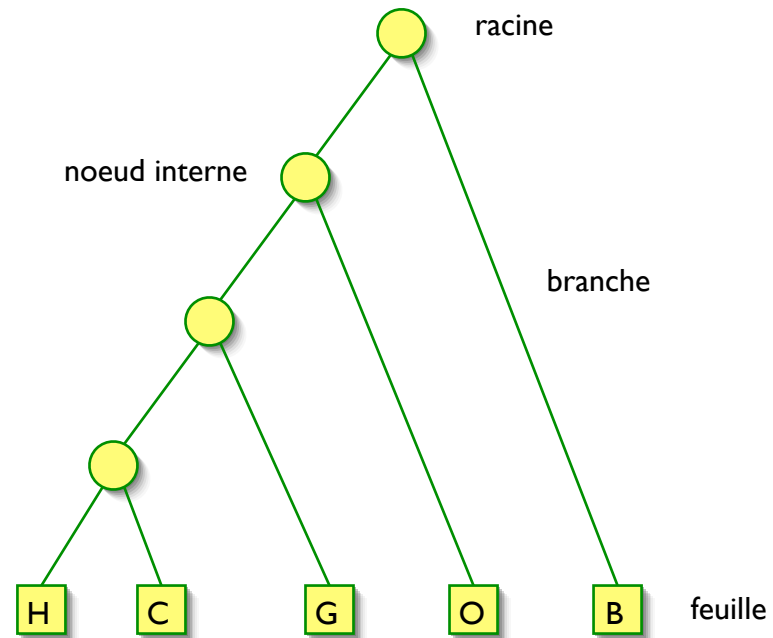
ou comment organiser notre collection de Livres

Phylogénies

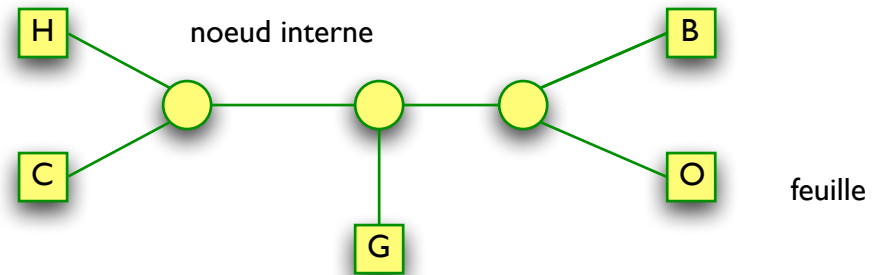
phylogénie ou arbre évolutif



Arbre raciné



Arbre non-raciné



Évolution de caractères/séquences

Séquence associée à chaque noeud : peut être des séquences moléculaires ou d'autres «caractères» (possède des ailes, présence de colonne vertébrale, présence d'intron, nombre de gènes dans la famille)

Évolution comme transmission d'une séquence ancestrale à partir de la racine vers les feuilles

Mutations introduites sur les branches

⇒ arbre sur les séquences (pas nécessairement le même que la phylogénie des espèces)

Mutations

Substitutions : $O(10^{-9})$ par site par an dans ADN nucléaire de mammifères
[beaucoup de variation !]

- transitions : $A \leftrightarrow G$ ou $T \leftrightarrow C$ et transversions
- peut changer le codon ou non [si région codante]

Insertions et suppressions

Recombinaisons

Inversions

Maximum de parcimonie (MP)

minimiser le nombre total de changements de caractères sur les branches (**valeur de parcimonie**)

Problème 1 : calculer la valeur de parcimonie pour un arbre donné — PD

Problème 2 : trouver l'arbre qui minimise la parcimonie — NP-difficile

Nombre d'arbres non-racinés avec n feuilles : $1 \cdot 3 \cdot 5 \cdots (2n - 5)$

Parcimonie 2

Problème 1 (la petite parcimonie)

Séquences [alignées] : $s_1, \dots, s_n$, de longueur ℓ

Arbre : chaque s_i assignée à une feuille ; déterminer les séquences ancestrales $s_j : j > n$ assignées aux noeuds internes

pénalités de substitution $\mathbf{C}$; $\mathbf{C} \begin{bmatrix} a \\ a \end{bmatrix} = 0$

Valeur de parcimonie sur une branche ij :

$$\text{VP}(ij) = \sum_{k=1}^{\ell} \mathbf{C} \begin{bmatrix} s_i[k] \\ s_j[k] \end{bmatrix}$$

On veut minimiser $\sum_{ij} \text{VP}(ij)$.

Parcimonie 3

Observation : $\sum_{ij} \text{VP}(ij) = \sum_{k=1}^{\ell} \sum_{ij} \mathbf{C} \begin{bmatrix} s_i[k] \\ s_j[k] \end{bmatrix}$ donc il suffit de trouver un algorithme pour des séquences de longueur 1 et l'utiliser ℓ fois.

Programmation dynamique : calculer pour chaque noeud u

- $V(u)$: VP pour le sous-arbre raciné à u
- $V(u, a)$: VP pour le sous-arbre si u est assigné caractère a

Feuilles : $V(u) = 0$; $V(u, a) = 0$ si u est assigné a , $V(u, a) = \infty$ sinon

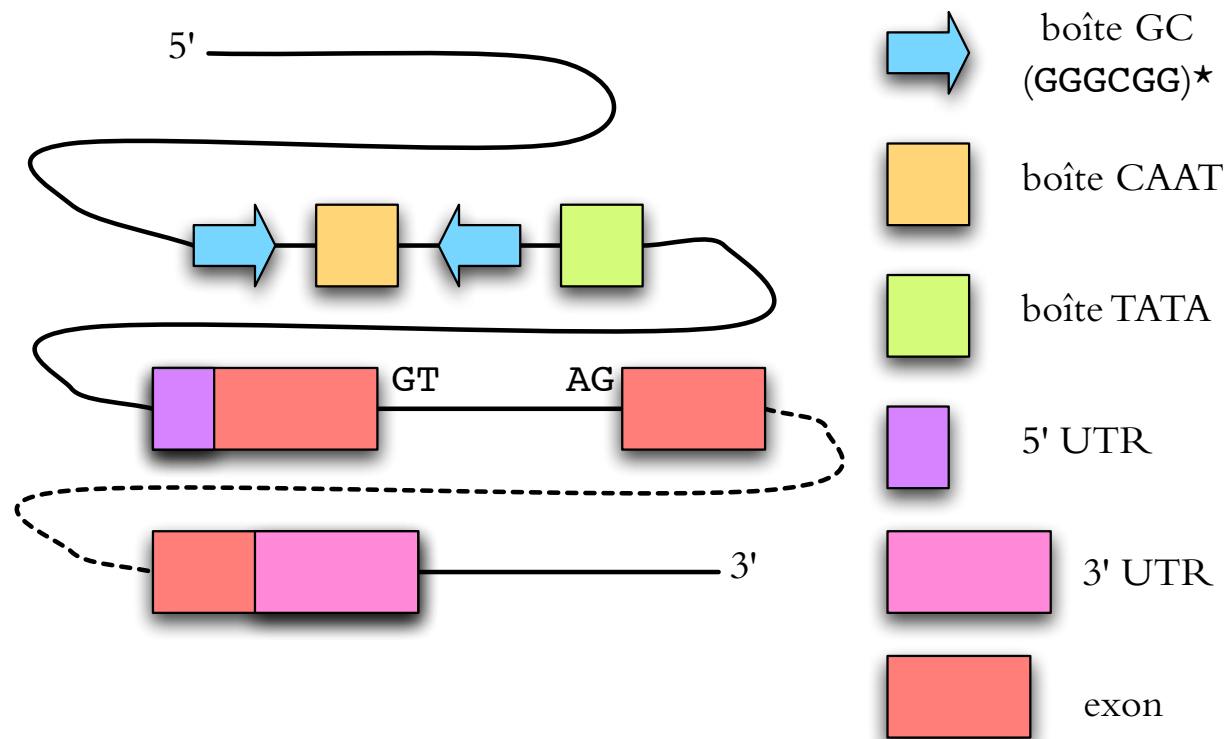
Réurrences pour noeud interne :

$$V(u, a) = \sum_{v \text{ est un descendant}} \min \{ V(v, b) + \mathbf{C} \begin{bmatrix} a \\ b \end{bmatrix} \};$$

$$V(u) = \min \{ V(u, a) \}$$

→ comme l'évaluation de scénarios gain-perte de gènes

Gènes en pièces



1977 : il y a des gènes qui sont «cousus» à partir des 'exons' (expressed)

Gilbert (1978)

Nature Vol. 271 9 February 1978

501

news and views

OUR picture of the organisation of genes in higher organisms has recently undergone a revolution. Analyses of eukaryotic genes in many laboratories¹⁻¹⁰, studies of globin, ovalbumin, immunoglobulin, SV40 and polyoma, suggest that in general the coding sequences on DNA, the regions that will ultimately be translated into amino acid sequence, are not continuous but are interrupted by 'silent' DNA. Even for

Why genes in pieces?

from Walter Gilbert

A gene, a contiguous region of DNA, now corresponds to one transcription unit, but that transcription unit can correspond to many polypeptide chains, of related or differing functions.

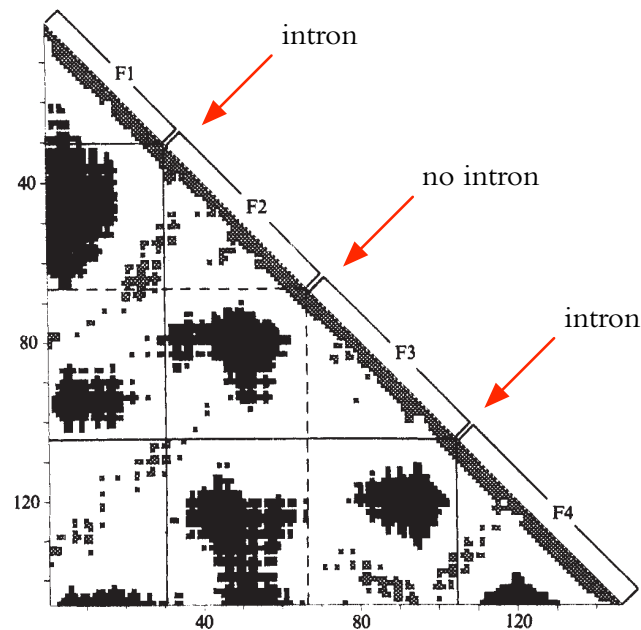
Recombination now becomes more rapid. Since the gene is spread out over a larger region of DNA, recombination, which should be hampered in higher cells by the inability of DNA molecules to get together, will be enhanced. Further-

... regions which will be lost from the mature messenger — which I suggest we call introns ...

(d'autres idées dans le même article : pistage alternatif, *exon shuffling*)

Gilbert Nature 271 :501, 1978

Exon shuffling



introns dans le gène hemoglobine β sont entre des unités structurales. . .
mais il en manque un [trouvé un peu plus tard dans la même année dans soy]

peut-être les modules des protéines correspond à des exons qui peuvent être permutés librement pour créer des gènes complexes — hélas la théorie n'est pas correcte

Gò *Nature* 291 :90, 1981

Origin d'introns

introns-early : introns sont plus anciens que l'ancêtre commun des eucaryotes et procaryotes — quelques uns sont perdus

introns-late : introns étaient acquis pendant l'évolution des eucaryotes

vérité : il y a des anciens, il y a des nouveaux, il y a des gains, il y a des pertes

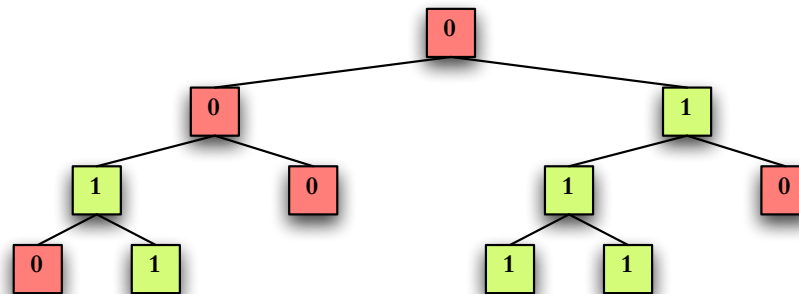
Introns homologues

On veut analyser les introns dans des positions homologues dans des différents organismes.

Il est difficile d'établir la homologie — les séquences des introns ont perdu leur similarité

On peut aligner des alignements des séquences protéiques

Présence/absence d'introns encodés par 1/0



Positions homologues

Intron positions 33 55 144 169 233

Pf MSRRTKKVGLTGKYGTRYGSSLRQIKKIELMQHAKYLCF CGKTATKRTC VGIWKCK--KCKRKVCGGAWSLTPAAVAAKSTIIRLRKQKEEAQKS

At MTKRTHKARIVGKYGTRYGASLRQIKKMEVSQHNKYFCEFCGKYSVKRKVVGIWKCK--DCGKVKAGGAYTMNTASAVTVRSTIRRLREQTES

Sc MAKRTKKVGITGKYGVRYGSSLRQVKKLEIQQHARYDCSFCGKKT VKRGAAGIWTCS--CCKKT VAGGAYTVSTAAAATVRSTIRRLREMVEA

Sp MTKRTHKVGITGKYGVRYGASLRDVRKIEVQQHSRYQCF CGRLTVKR TAAGIWKCSGKGC SKTLAGGAWTVTTAAATSARSTIRRLREMVEV

Ce MAKRTKKVGIVGKYGTRYGASLRKMAKKLEVAQHSRYTCSFCGKEAMKRKATGIWNCA--KCHKVVAGGAYVYGTVTAAATVRSTIRRLRLDKE

Dm MAKRTKKVGIVGKYGTRYGASLRKMVKKMEITQHSKYTCSFCGKDSMKRAVVG IWSCK--RCKRTVAGGAWVYSTTAAASVRS AVRRLRETKEQ

Ag YLPKMAKRTRKVGIVGKYGTRYGASLRKMVKKMEITQHAKYTCTFCGKDAMKRSCVGIWSCK--RCNRVVAGGAWVYSTTAAASVRS AVRRLREM

Hs MAKRTKKVGIVGKYGTRYGASLRKMVKKIEISQHAKYTCSFCGKTKMKRRAVGIWHCG--SCMKT VAGGAWTYNTTS AVTVKSAIRRLKELKDQ

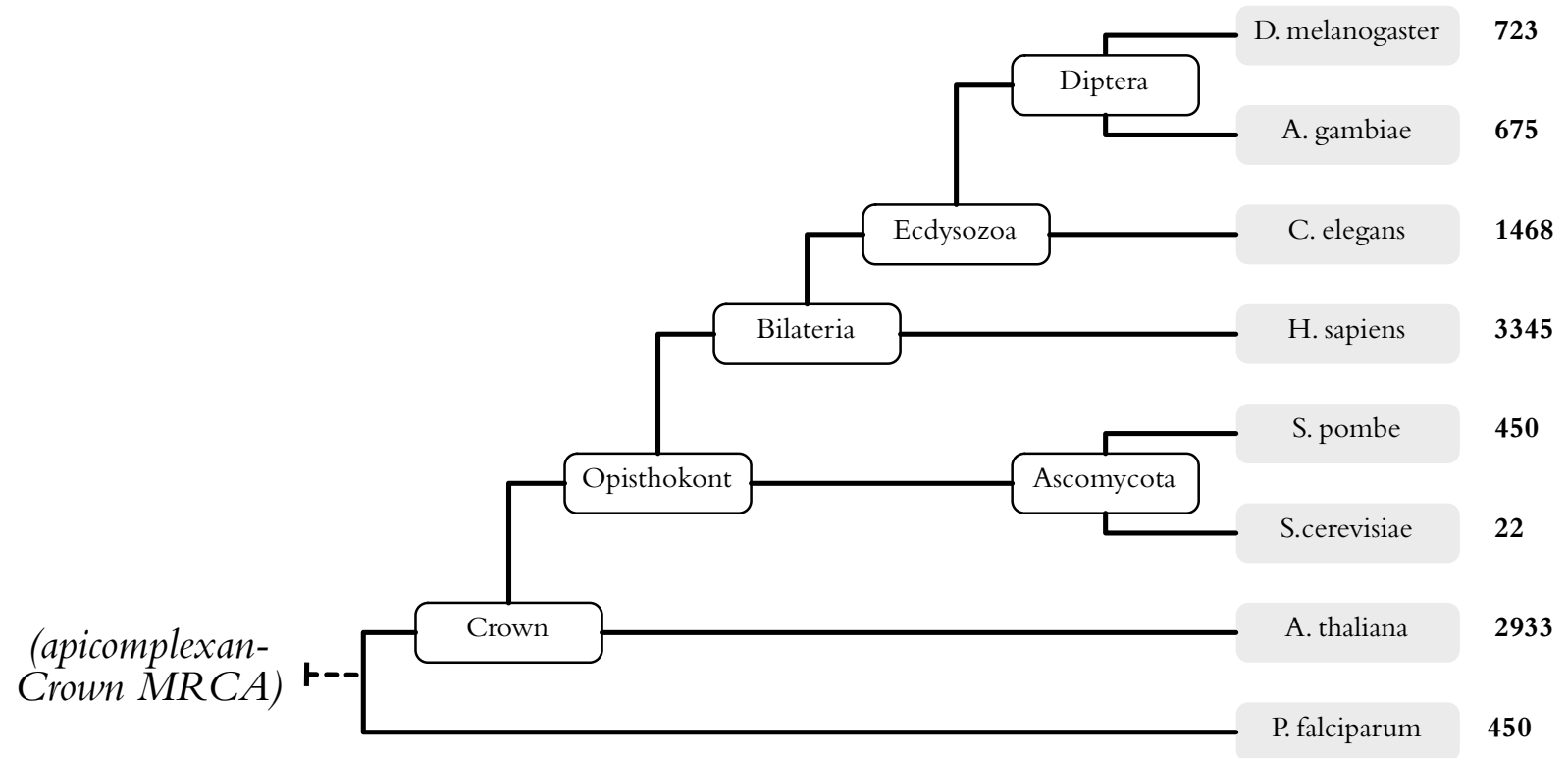


	33	55	144	169	233
Pf	1	0	1	0	0
At	0	1	1	0	0
Sc	0	0	0	0	0
Sp	0	0	0	1	0
Ce	0	0	0	0	0
Dm	0	0	1	0	0
Ag	0	0	1	0	0
Hs	0	0	1	0	1

Rogozin++ *Current Biology* 13 :1512 (2003)

Les données

7236 positions

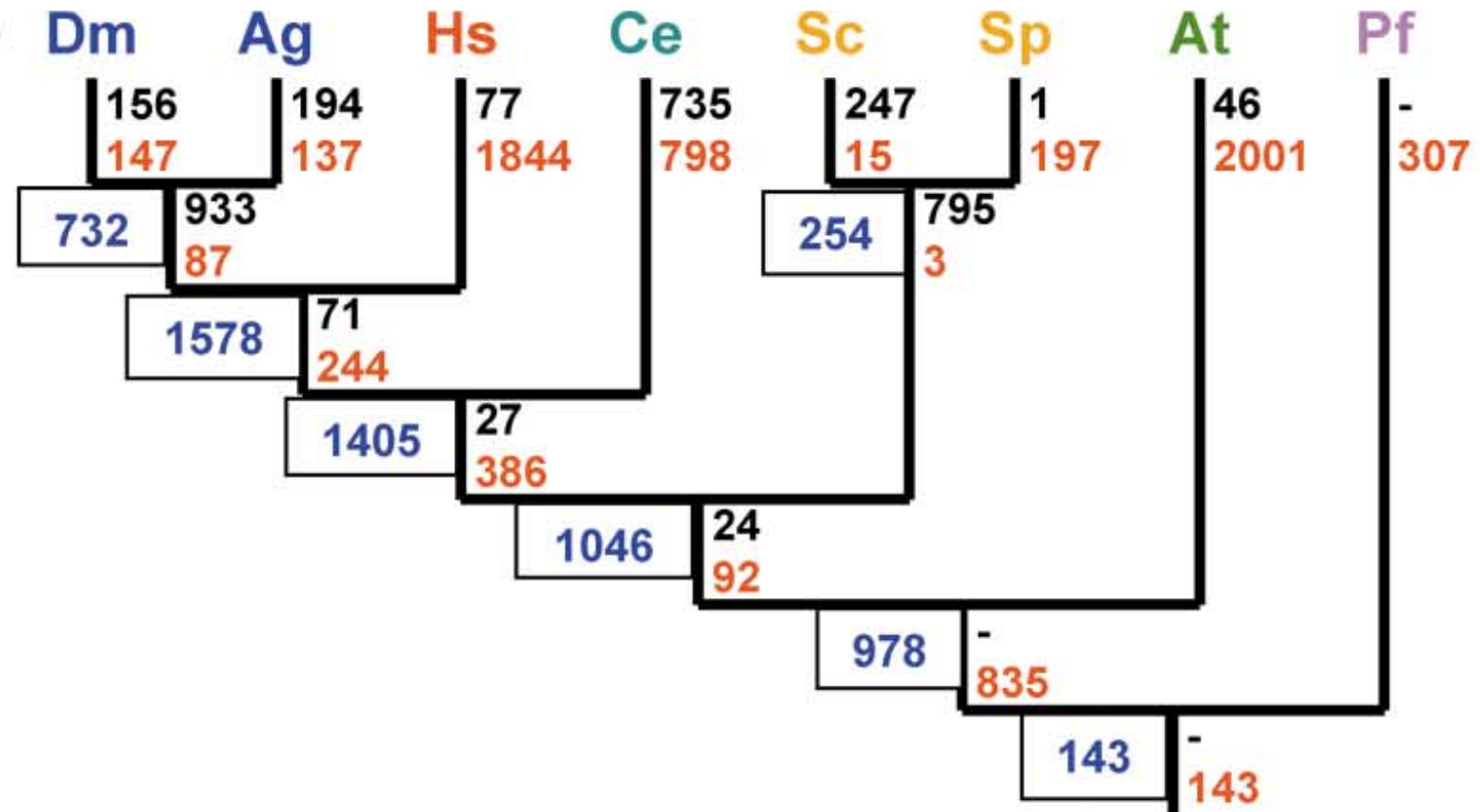


Parcimonie Dollo

Il est peu probable qu'un intron est apparu plus qu'une fois

Parcimonie : $0 \rightarrow 1$ sur une branche dans l'arbre, $1 \rightarrow 0$ sur autant de branches que nécessaire

Résultats — Dollo



Rogozin++ *Current Biology* 13 :1512 (2003)

Modèles probabilistes

Modèle d'évolution de séquences sur un arbre :

relai d'un message de la racine vers les feuilles

transmission avec du «bruit» sur chaque branche

bruit = mutations

modèle probabiliste pour les mutations

les séquences correspondent à un «échantillon» aléatoire

Cavender-Farris

modèle de Cavender-Farris : séquences binaires, mutations symmetriques

Exemple : noeud x est le parent de y , transmission d'un bit $X \rightarrow Y$.

Probabilité de mutation sur la branche : p_{xy} .

Alors $\mathbb{P}\{X \neq Y\} = p_{xy}$.

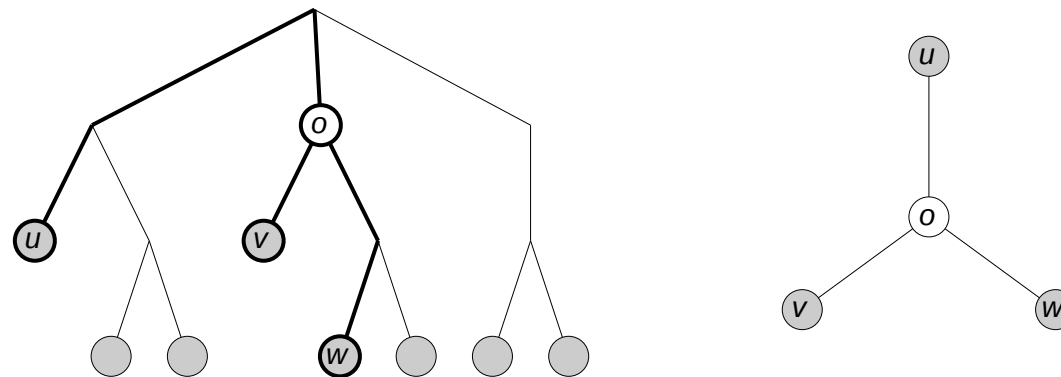
Similarité de deux noeuds : $S(x, y) = \mathbb{P}\{X = Y\} - \mathbb{P}\{X \neq Y\}$.

Thm. $S(x, y)$ se multiplie sur chaque chemin.

Donc $D(x, y) = -\log S(x, y)$ est une distance additive sur l'arbre.

Distances - 2

Algorithme pour la construction d'un arbre à partir des distances : triples



$$D(u, o) = \left(D(u, v) + D(u, w) - D(v, w) \right) / 2.$$

Algorithme pour construire l'arbre

[Esquissé...]

Feuilles : $1, \dots, n$; matrice de distances entre feuilles de taille $n \times n$

1. Commencer avec le triple $1, 2, 3$ et son centre
2. Pour $i = 4, \dots, n$:
 - 2.1 choisir★ un triple j_1, j_2, i avec $j_1, j_2 < i$;
 - 2.2 insérer son centre c sur le chemin entre j_1 et j_2 dans l'arbre ;
 - 2.3 joindre i à c et calculer les nouvelles longueurs de branches

★ le centre doit être différent d'autres vertices déjà dans l'arbre

Jukes-Cantor

Jukes-Cantor-Neyman : évolution de séquences sur un alphabet $\Sigma = \{a_1, \dots, a_m\}$

mutations de caractères indépendantes en chaque position des séquences

mutation sur une branche XY : symétrique, spécifiée par la probabilité de mutation p

$$\mathbb{P}\left\{Y = a' \mid X = a\right\} = \begin{cases} 1 - p & \text{si } a = a'; \\ p/(m - 1) & \text{si } a \neq a'. \end{cases}$$

Distance : $D(X, Y) = -\log\left(1 - \frac{m}{m-1}\mathbb{P}\{X \neq Y\}\right)$

Modèle Markovien

Mutations indépendantes ; probabilités de substitutions spécifiées par une matrice de transition

$$\mathbf{M}_{X \rightarrow Y}[a, a'] = \mathbb{P}\{Y = a' \mid X = a\}$$

On a que $\mathbf{M}_{X \rightarrow Z} = \mathbf{M}_{X \rightarrow Y} \mathbf{M}_{Y \rightarrow Z}$.

Donc $D(X, Y) = -\log \left(\det \mathbf{M}_{X \rightarrow Y} \det \mathbf{M}_{Y \rightarrow X} \right)$ est une distance sur l'arbre (additive et symétrique) — appelée distance paralinéaire