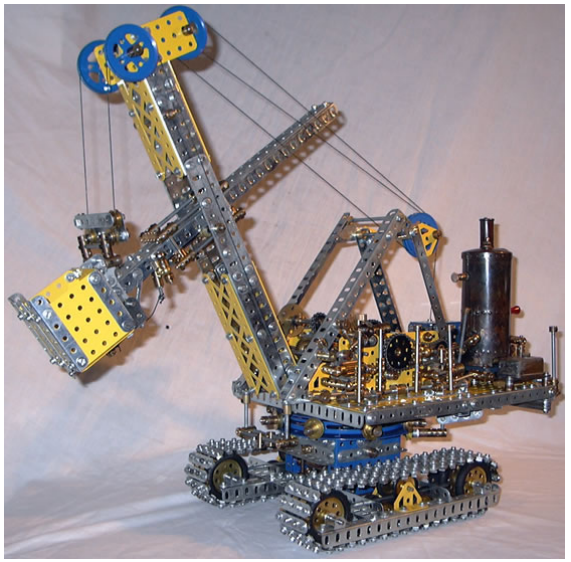
Profils phylétiques

François Jacob

prix Nobel en 1965, membre de l'Académie française depuis 1996

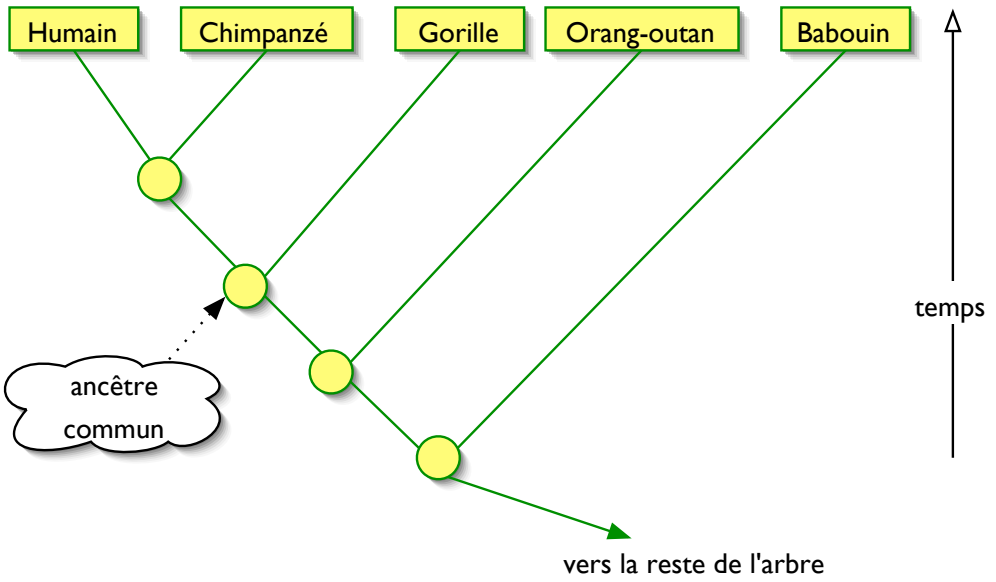
> Nous sommes faits d'un étrange mélange d'acides nucléiques et de souvenirs, de rêves et de protéines, de cellules et de mots.
>
> [...]
>
> Le monde vivant est fait de combinaisons d'éléments en nombres finis et ressemble aux produits d'un gigantesque Meccano résultant d'un bricolage incessant de l'évolution.



"Meccano", Wikipedia

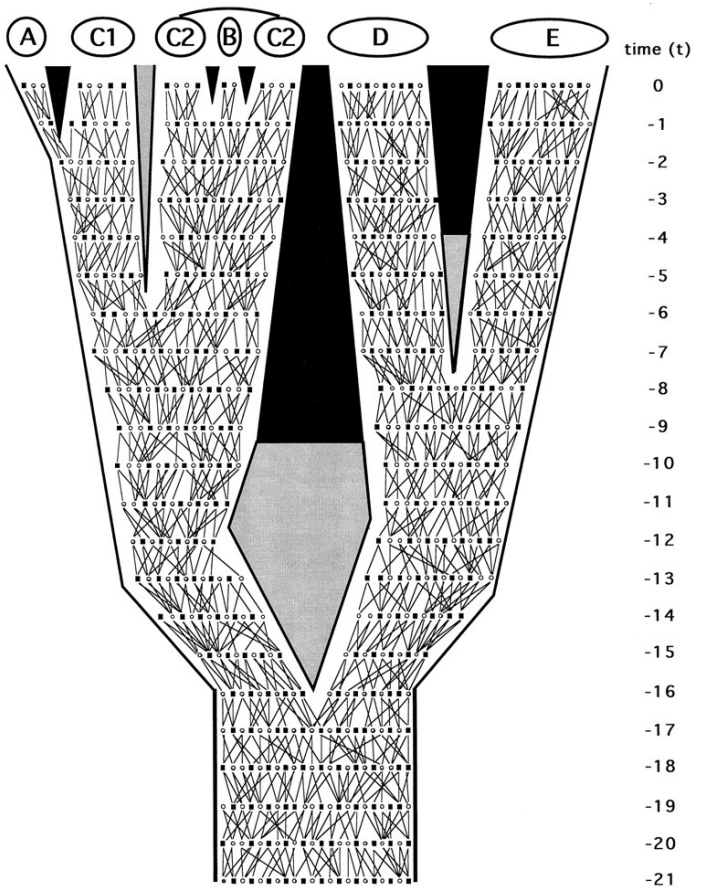
Phylogénie

phylogénie ou arbre évolutif



représente les relations évolutionnaires entre les organismes (relation : descendance d'un ancêtre commun)

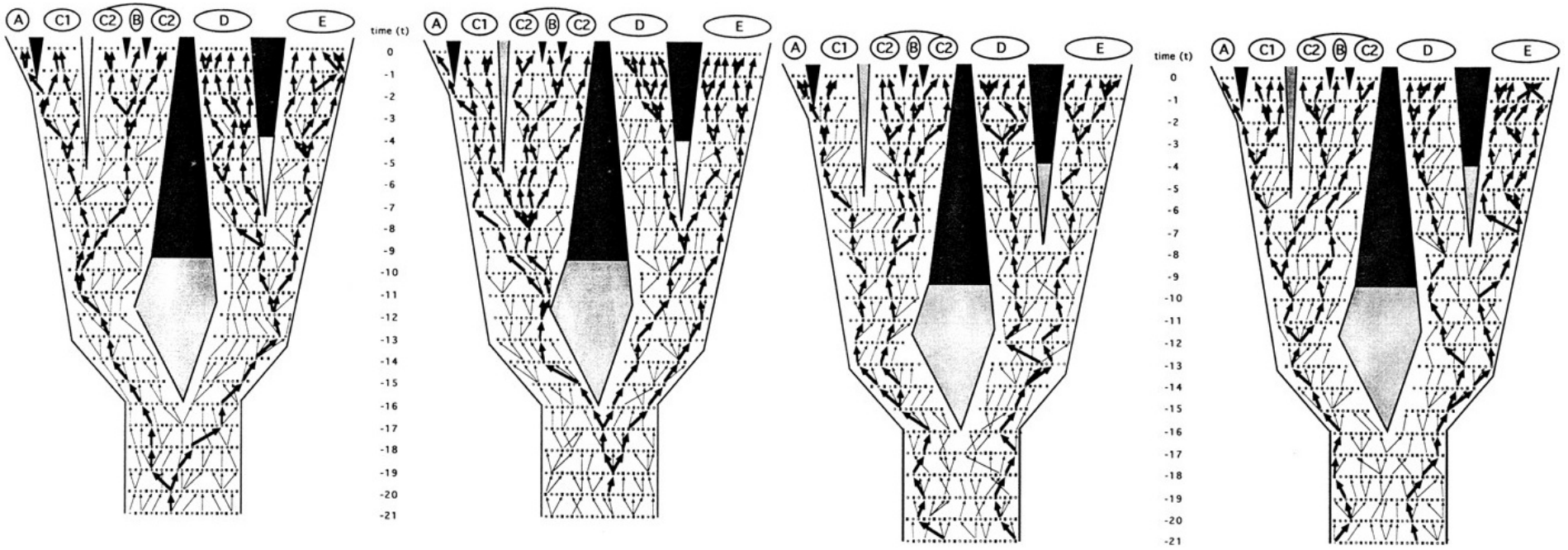
Espèces et populations



Avise & Wollenberg *PNAS* 94 :7748, 1997

Histoire des individus

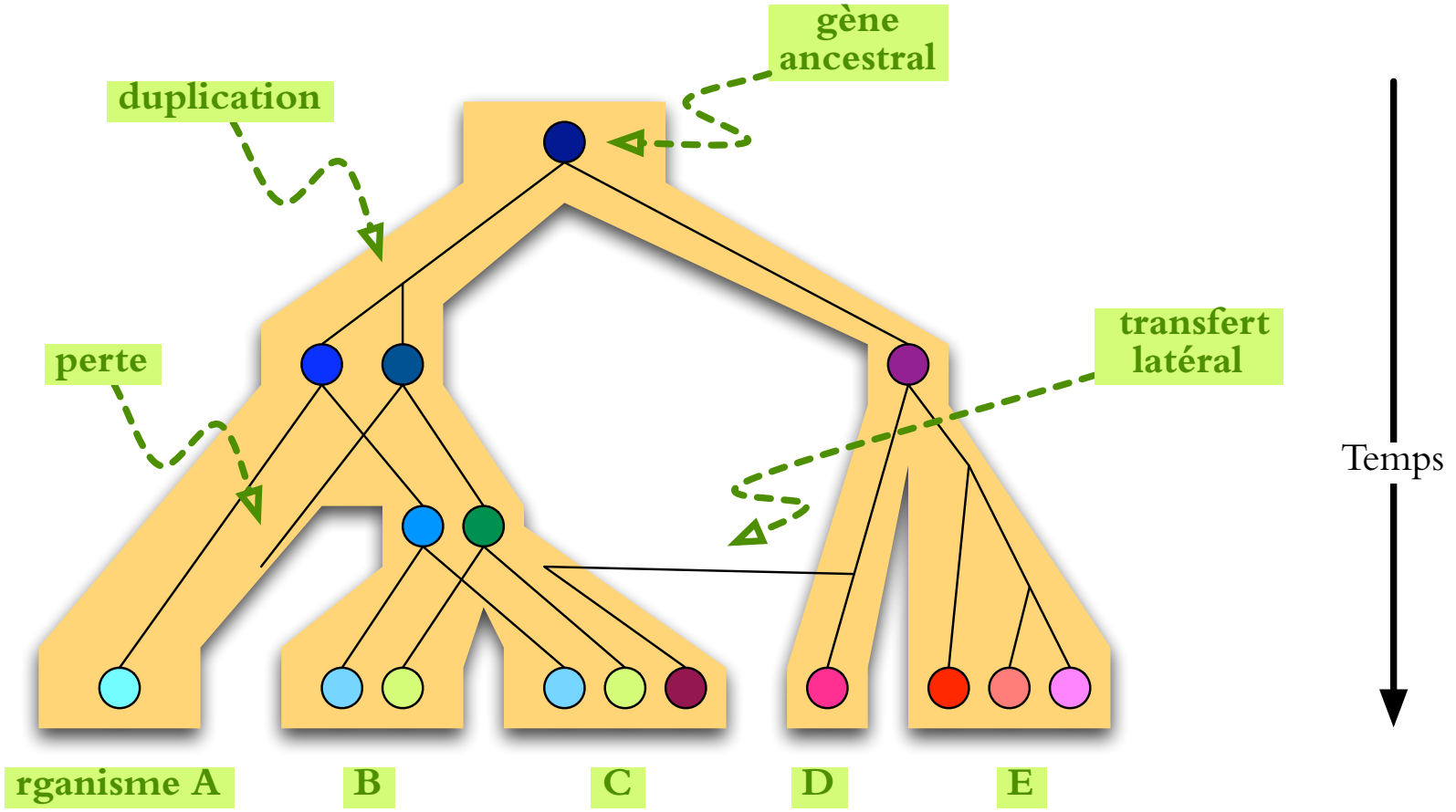
Correspondance entre phylogénie des espèces et l'histoire des individus (allèles, plutôt) n'est pas aussi simple . . . on va retourner à cette question



Avise & Wollenberg *PNAS* 94 :7748, 1997

Histoire d'une famille de gènes

L'histoire d'une famille de gènes dans la phylogénie



Terminologie

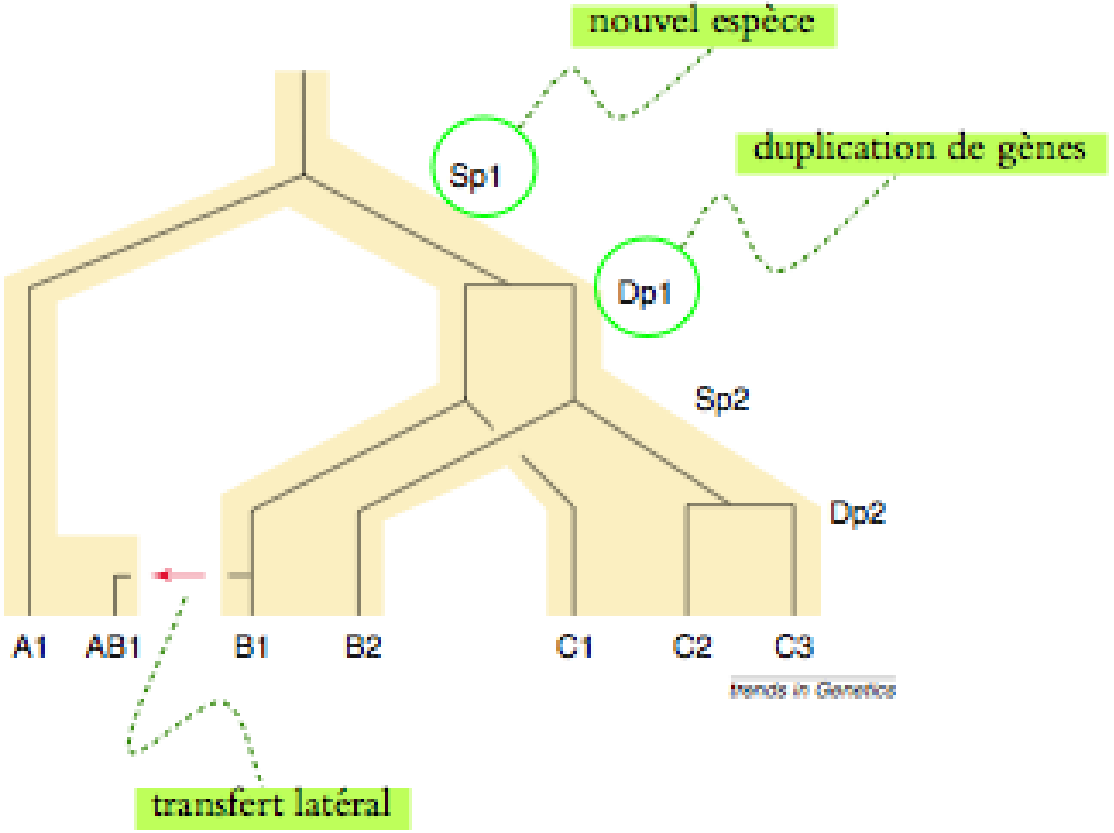
homologue : relié par un ancêtre commun

→ orthologue : relié par événement de spéciation

→ paralogue : relié par événement de duplication

→ xenologue : acquis par un autre mécanisme (transfert latéral)

Gènes homologues



orthologues : B1–A1, B1–C1

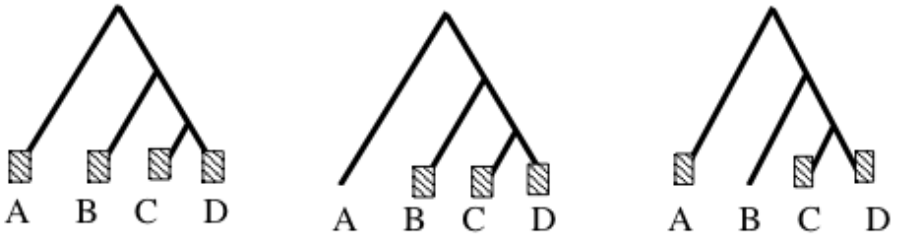
paralogues : B1–B2 (*in-paralog*), B1–C2 (*out-paralog*)

xenologues : A1–AB1 co-orthologues : {C1, C2, C3}–{B1, B2}

Fitch *Trends in Genetics* 16 :227 (2000)

Profils phylétiques

pour un gène — enregistrer dans quels espèces il y a au moins un homologue



`ABCD` `-BCD` `A-CD`

→ prédiction de fonction (profils pareils)

→ évolution de répertoire de gènes

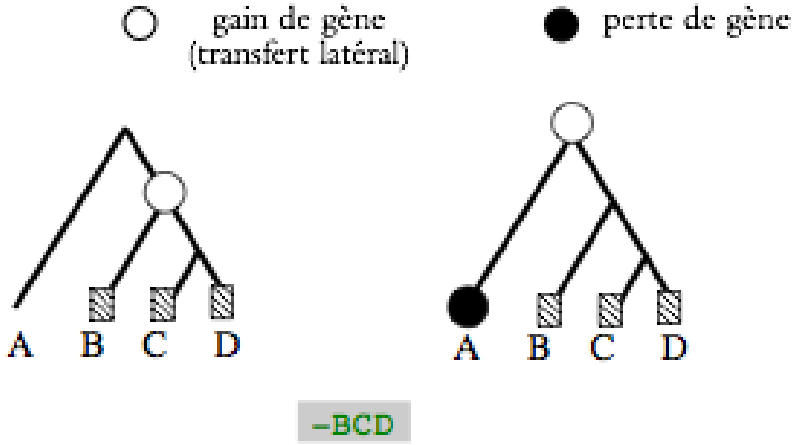
→ phylogénie d'espèces

→ validation d'annotation

Tatusov & al *BMC Bioinformatics* 4 :41 (2003)

Évolution du répertoire de gènes

Scénarios d'évolution



et d'autres scénarios avec plus d'événements

On peut calculer le meilleur scénario par parcimonie

Mirkin & al *BMC Evol Biol* 3 :2 (2003)

Parcimonie

Problème : on a un arbre phylogénétique et on connait les étiquettes $\xi[u]$ à chaque nœud *terminal* u

Objective : trouver des étiquettes $\xi[u]$ pour les nœuds internes

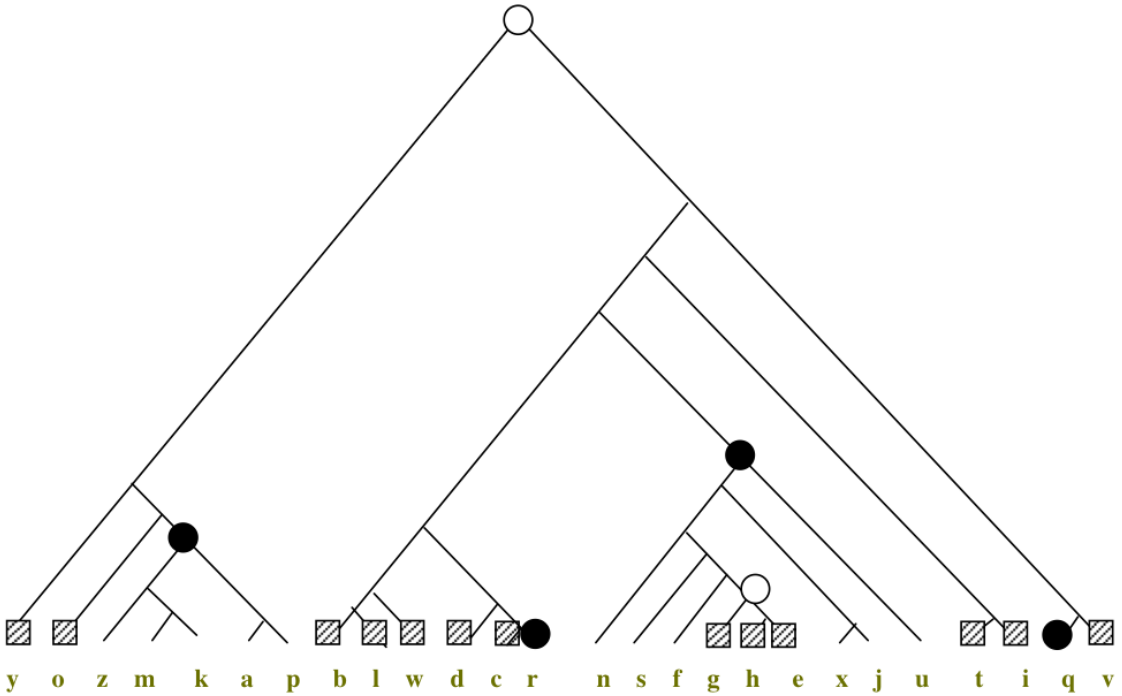
Parcimonie : minimiser les changements sur les arcs (c-à-d minimiser une somme de pénalités sur les arcs)

Variantes :
- simple : pénalité de 1 pour $\xi[u] \neq \xi[v]$ et 0 pour $\xi[u] = \xi[v]$
- poids : pénalisation de $\Delta(x \to y)$ si $\xi[u] = x$ et $\xi[v] = y$
- Dollo : étiquettes binaires, $0 \to 1$ seulement une fois dans l'arbre
- Wagner : étiquettes numériques, $\Delta(x \to y) = |x - y|$

Solution : par programmation dynamique

Évolution du répertoire de gènes

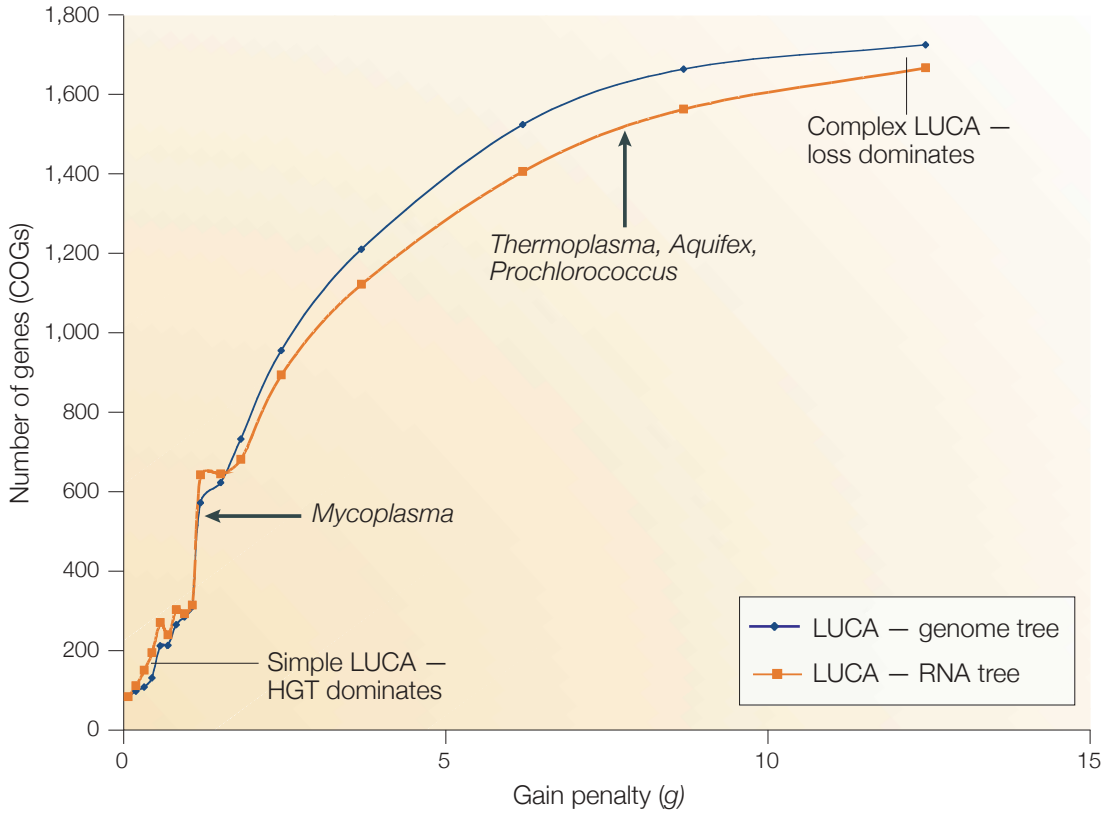
questions : (1) est-ce que le gène à la racine est pénalisé ? ; (2) quel est le score rélatif des gains et des pertes ?



Mirkin & al *BMC Evol Biol* 3 :2 (2003)

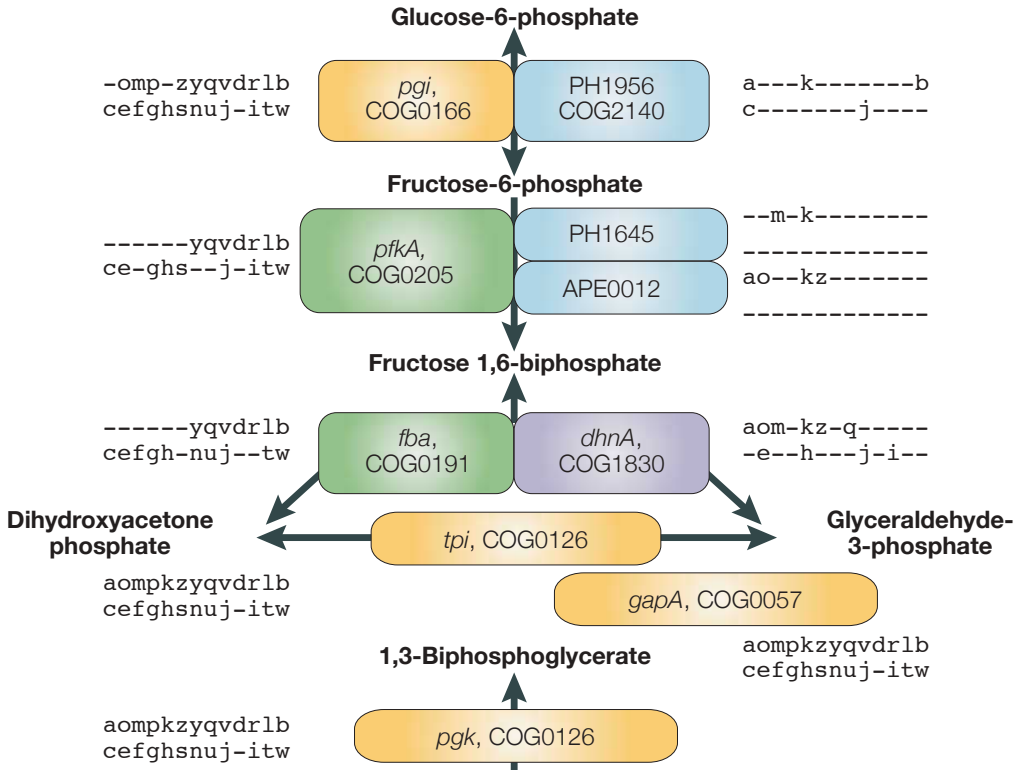
Gènes de LUCA

LUCA (*Last Universal Common Ancestor*) : le plus récent ancêtre commun de toutes les organismes vivantes



Koonin *Nat Rev Microbiol* 1 :127 (2003)

Métabolisme de LUCA

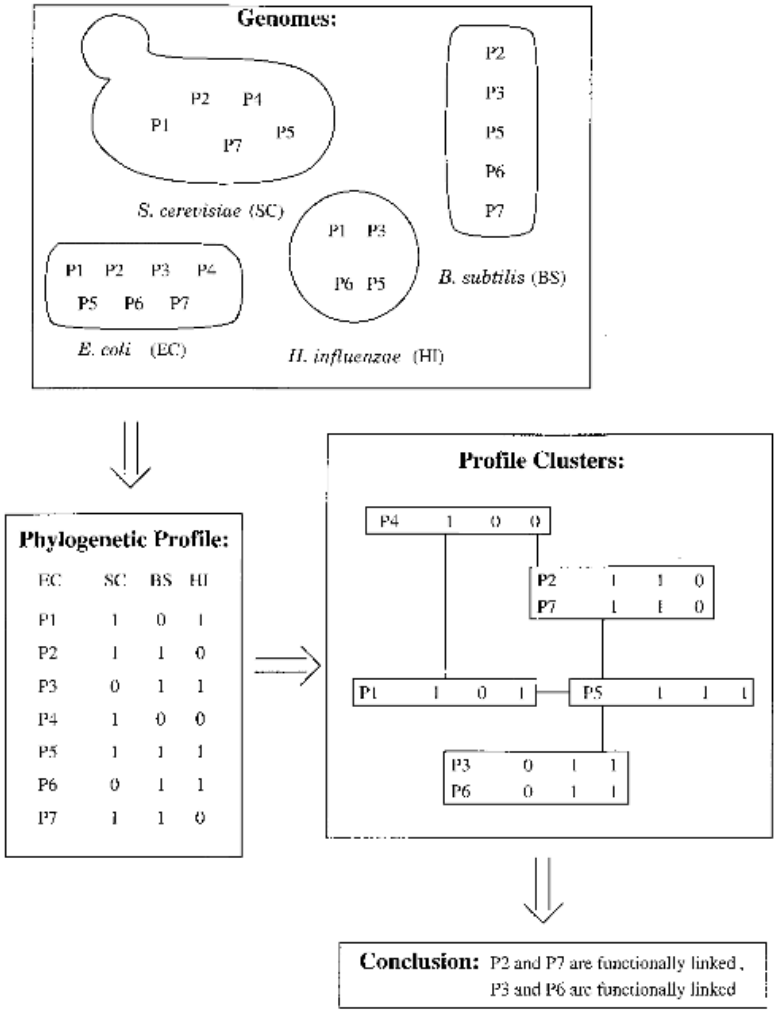


couleur	coût de gain (g)
jaune	0.9
vert	1
violet	2
bleu	non-LUCA

avec $g = 1$, presque tous les chemins essentiels sont présents dans LUCA : ≈ 600 gènes

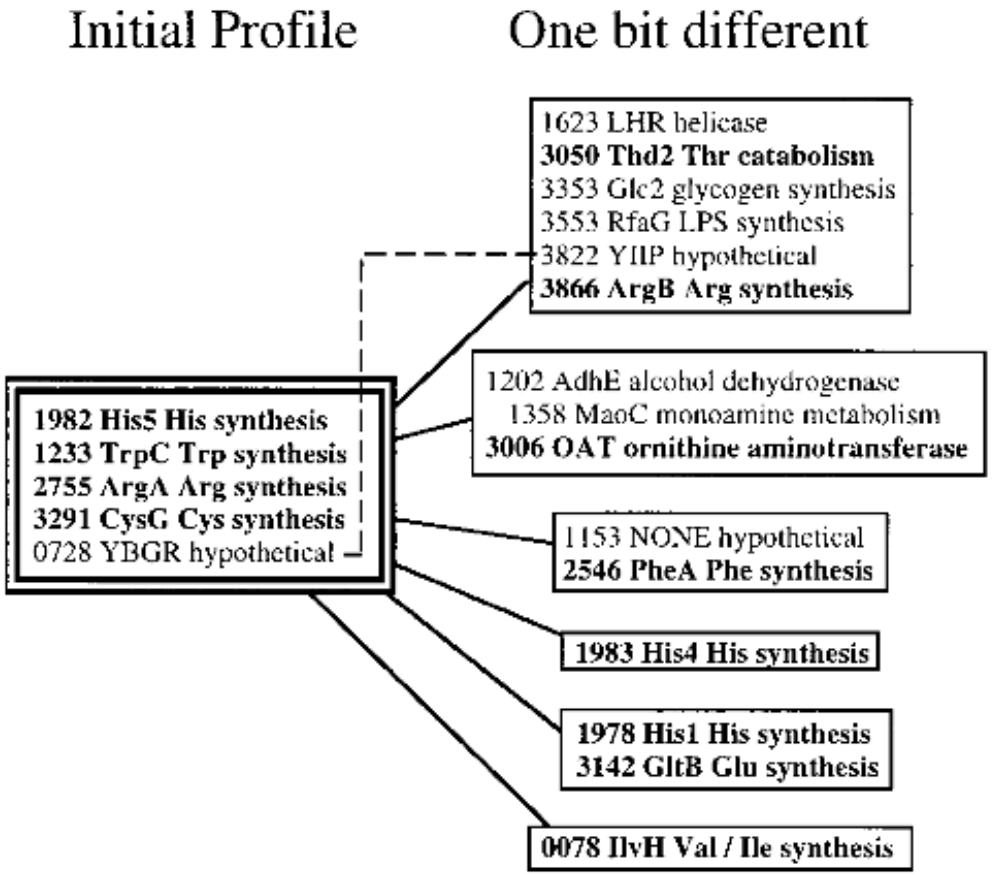
Koonin *Nat Rev Microbiol* 1 :127 (2003)

Prédiction de fonction par profils



Pellegrini & al *PNAS* 96 :4285 (1999)

Fonction et profil phylétique



(synthèse de histidine — les profils proches incluent aussi d'autres protéines de métabolisme d'acides aminés)

Pellegrini & al *PNAS* 96 :4285 (1999)

Profils complémentaires

Il y a des exemples ou la fonction d'un COG était déterminé par *complémentarité* à un COG de fonction connue

Exemple : synthase de thymidilate

COG0207 (fonction connue) : `a-m---y--drlb-efghsn-j---w`

Recherche d'un profil complémentaire (c'est un enzyme essentiel)

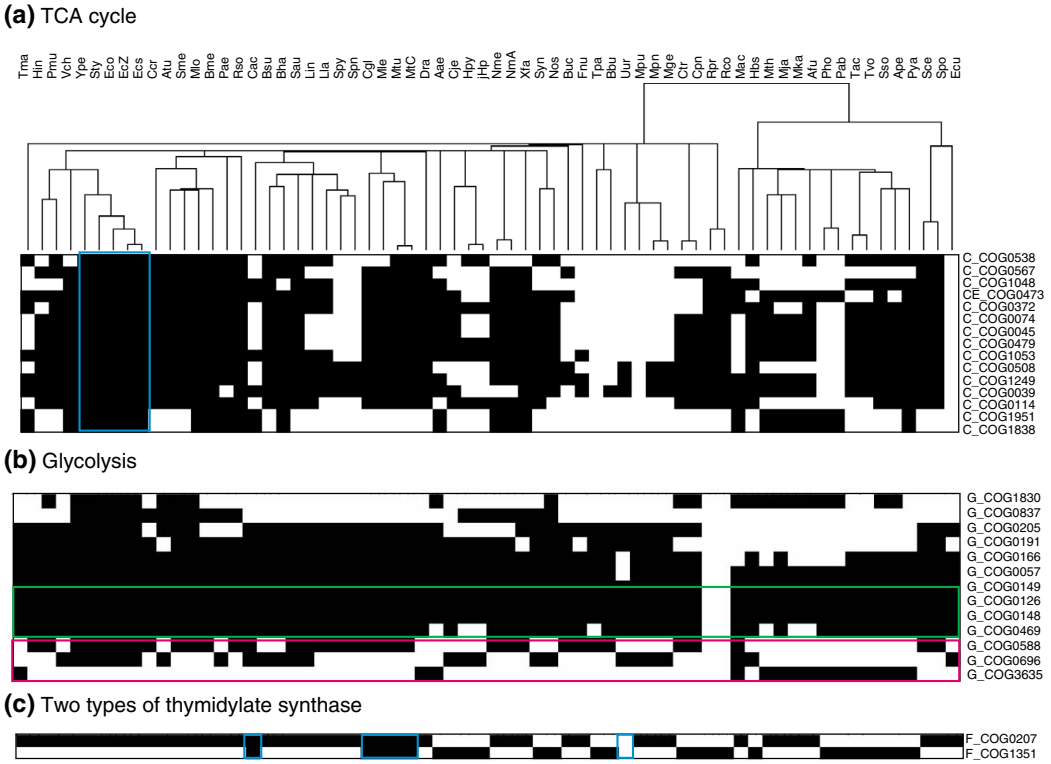
COG1351 (fonction inconnue) : `-o-pkz-qv-r--c------u-xit-`

NOGD — *Non-orthologous gene displacement*

exemple de Koonin & Galperin *Sequence-Evolution-Function* (2003)

Pas toujours aussi propre. . .

profils phylétiques pour groupes de gènes reliés par fonction

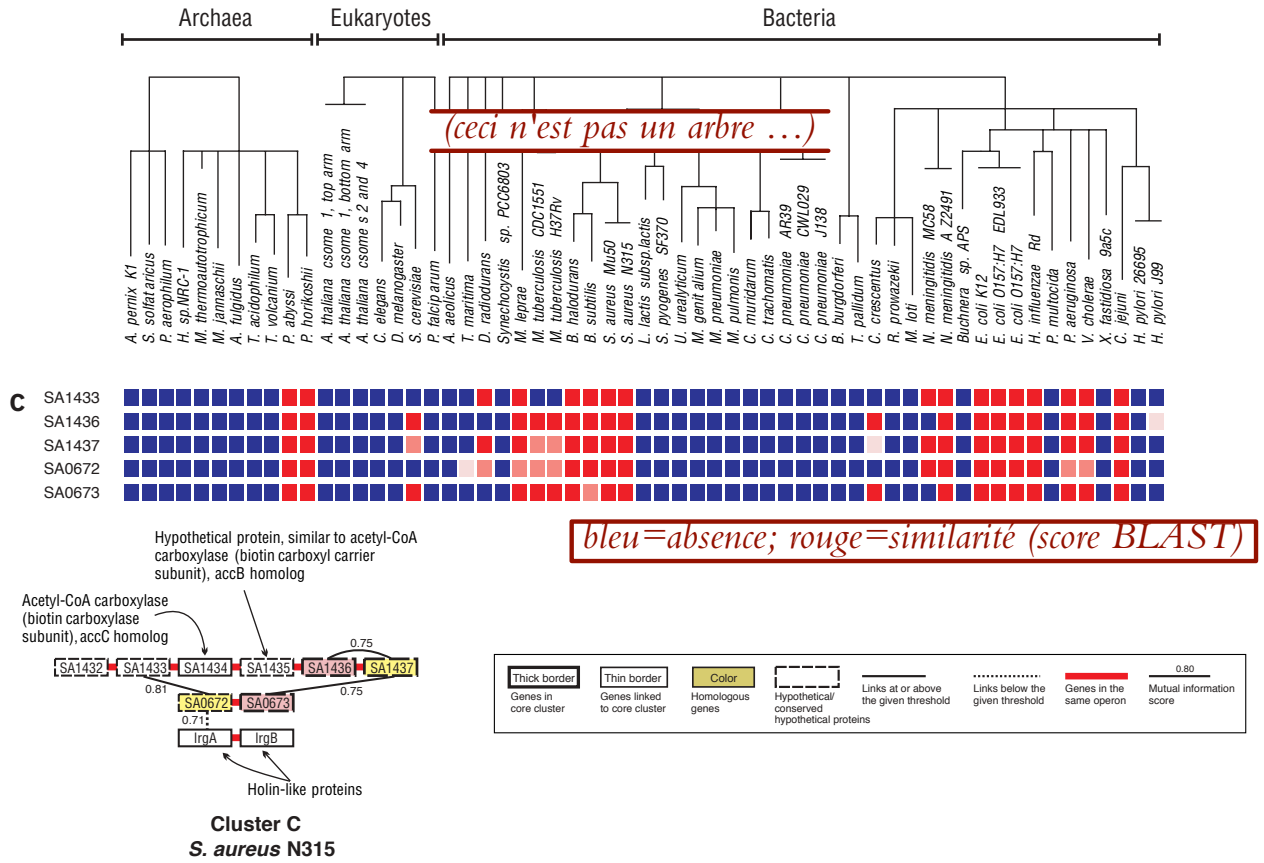


→ pertes et déplacements fréquents ; c'est des fragments de chemins qui sont préservés

⇒ on veut un modèle de coévolution (avec arbre) ou de corrélation (sans arbre)

Glazko & Mushegian *Genome Biol* 4 :R32 (2004)

Découverte de modules fonctionnels



Date & Marcotte *Nat Biotech* 9 :1062 (2003)

Corrélation entre profils

sans arbre évolutif : mesurer la corrélation entre les profils
vecteurs binaires (ou réelles — coordonnées expriment similarité)

profils $x[1..n]$ et $y[1..n]$ (sur n génomes)

$\longrightarrow$ corrélation linéaire (Pearson) n'est pas une bonne idée

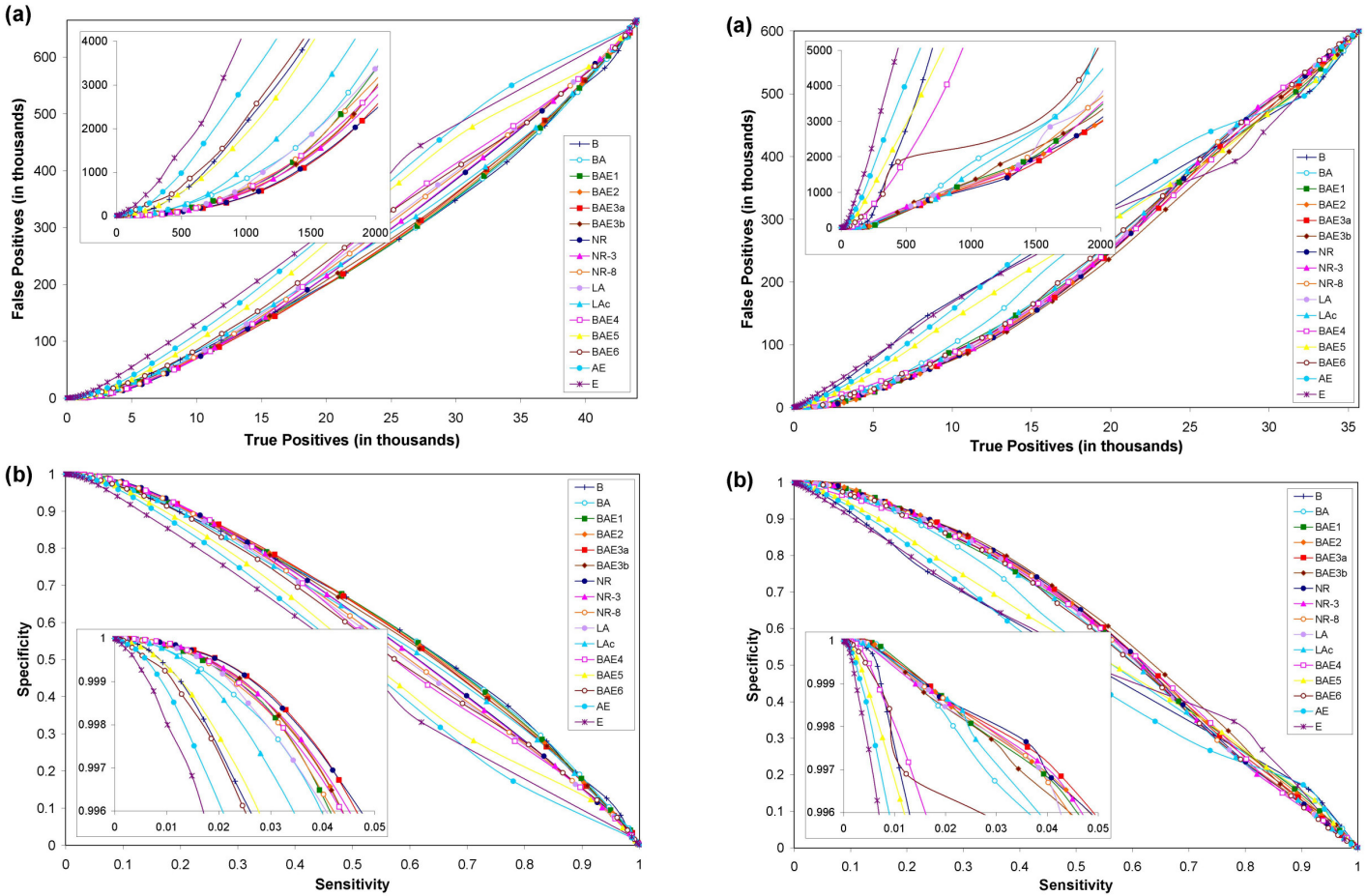
$\longrightarrow$ information mutuelle marche mieux

estimer les distributions de $x[i] \sim p_x$, $y[i] \sim p_y$ et $\Big(x[i], y[i]\Big) \sim p_{xy}$ (communes pour $i = 1, 2, \ldots, n$), et calculer

$$\mathsf{I}(x, y) = \sum_{\alpha,\beta} p_{xy}(\alpha, \beta) \log \frac{p_{xy}(\alpha, \beta)}{p_x(\alpha) \cdot p_y(\beta)}$$

comparer avec I pour profils aléatoires

Quels génomes ?



BAE3 = bactérie+archée+quelques eucaryotes ; NR = sans génomes proches

Jothi, Przytycka & Aravind *BMC Bioinformatics* 8 :173 (2007)