Gènes* dans le génome

∗ codant pour protéines

Gènes

Question principale : quel genre de gènes ?

Procaryotes [pas d'introns, opérons] et Eucaryotes [introns, signalisation compliquée]

Gènes traduits en protéines

Gènes non-traduits (gènes ARN)

Gènes non-transcrits (signaux de réplication, recombinaison, ségrégation [meiosis], . . .)

Code génétique

20 acides aminés encodés par 4 nucléotides ?

Encodages par triplets : 64 codons

AAA = Phe	AAG = Phe	AAT = Leu	AAC = Leu
AGA = Ser	AGG = Ser	AGT = Ser	AGC = Ser
ATA = Tyr	ATG = Tyr	ATT = FIN	ATC = FIN
		...	

3 triplets d'arrêt, 1 triplet de début (encode aussi Met)

De l'ADN à protéine

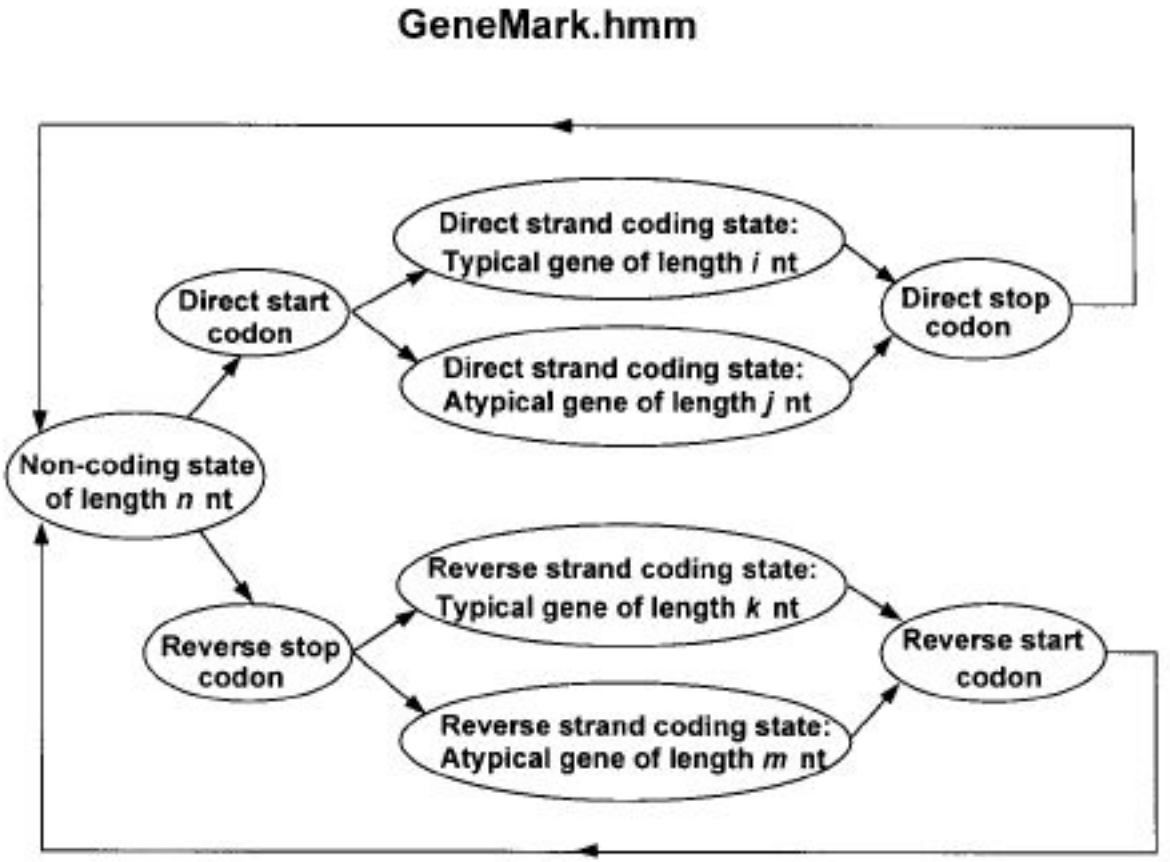
1. transcription : copie du brin informatif à ARN messager
(ARN : utilise Uracile au lieu de Thymine)

2. traduction : ARNm à protéine (par le ribosome) : acides aminés fournis par ARN de transfert

Mécanisme universelle !

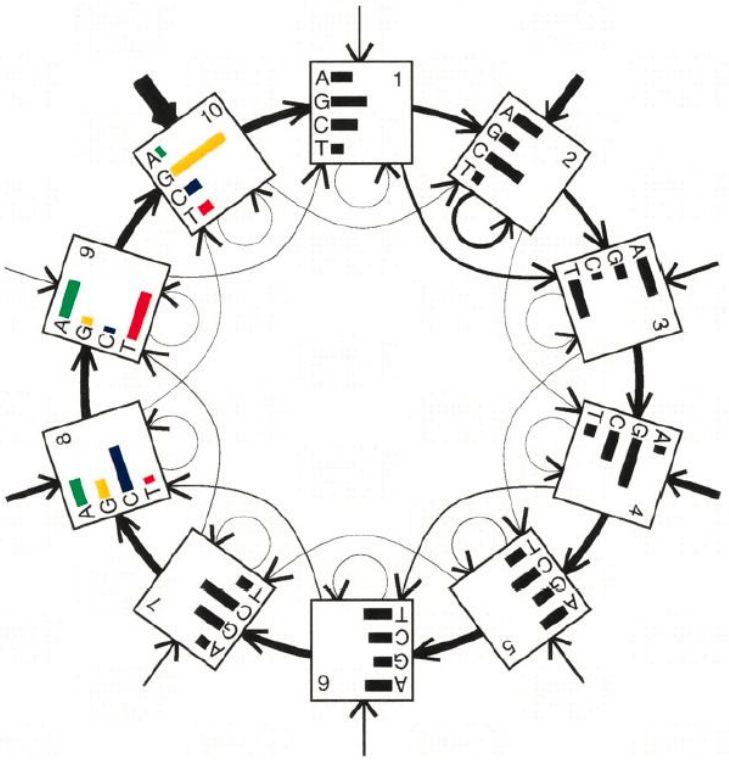
Gènes traduits — procaryotes

Procaryotes (pas d'exons !) — GeneMark



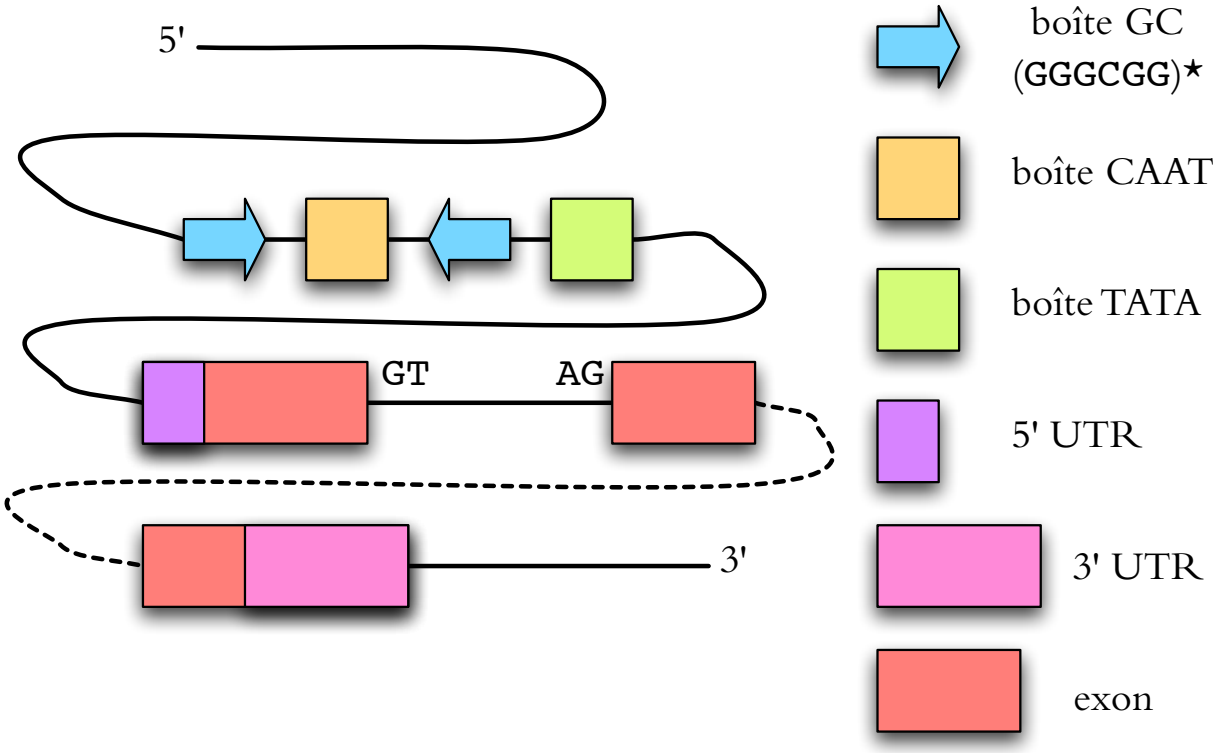
Lukashin & Borodovsky *Nucleic Acids Res* 26 : 1107 (1998)

ADN d'Eucaryotes



Baldi et al *J Mol Biol* 263 : 503 (1996)

Gènes traduits — eucaryotes



après Graur & Li (2000)

De l'ADN à protéine

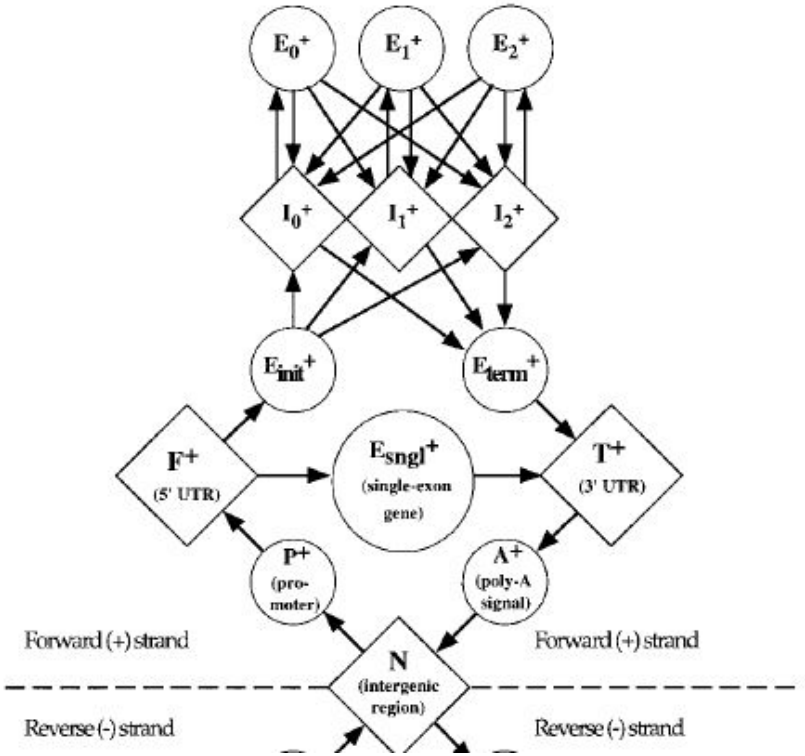
[animation]

[animation]

Lodish & al. *Molecular Biology of the Cell*, 2002

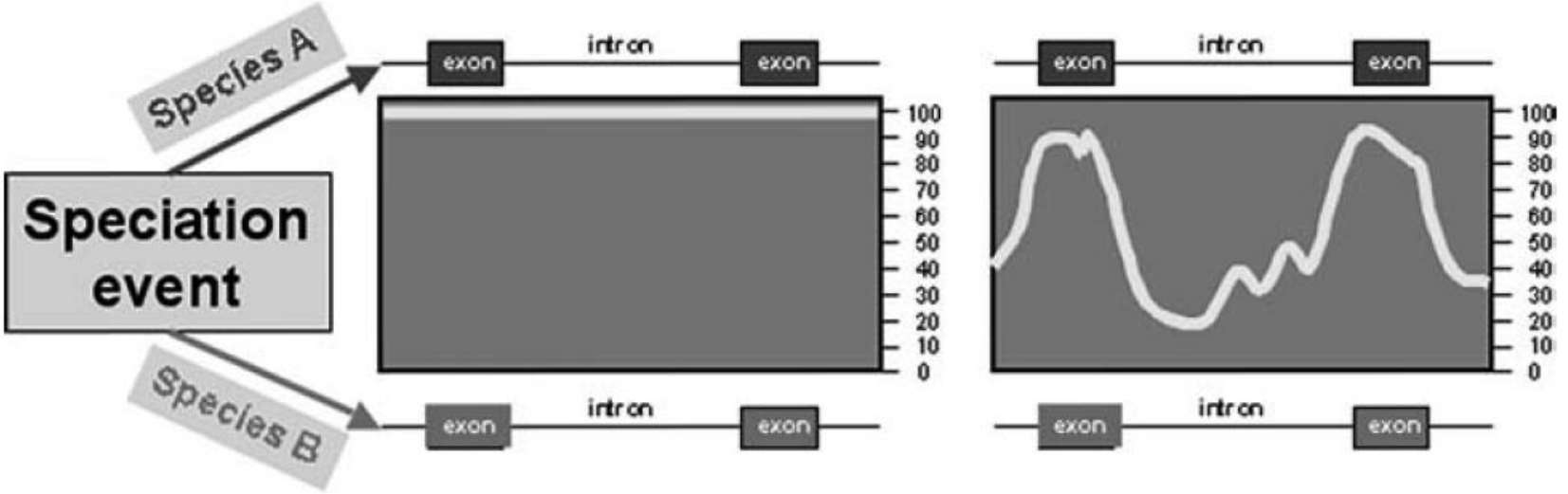
Prédiction de gènes (Genscan)

Eucaryotes — GENSCAN



Burge & Karlin *J Mol Biol* 268 : 78 (1997)

Méthode comparative pour la recherche de gènes

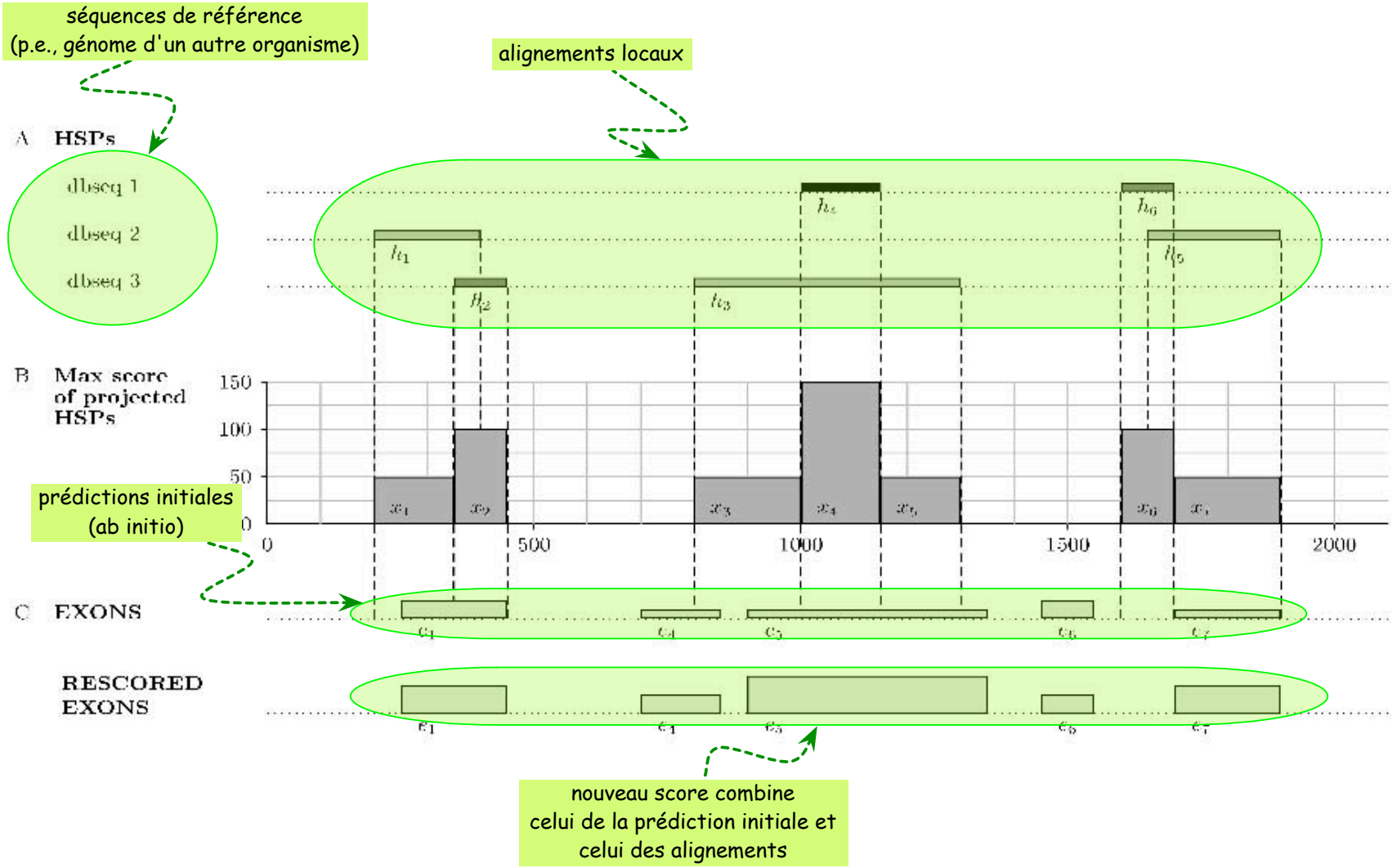


Alignement de deux régions aide à l'identification d'exons : les exons sont plus préservés (sélection négative)

Principe de génomique comparative : éléménts fonctionnels sont plus [séléction négative] ou moins [séléction positive] préservés que des éléments non-fonctionnels [évolution neutre]

Miller & al. *Annu Rev Genomics Hum Genet* 5 :15 (2004)

Prédiction de gènes (SGP2)



Parra & al. *Genome Res* 13 :108 (2003)

Terminologie

similarité : notion algorithmique de relation entre séquences

homologue : relié par un ancêtre commun

→ orthologue : relié par événement de spéciation

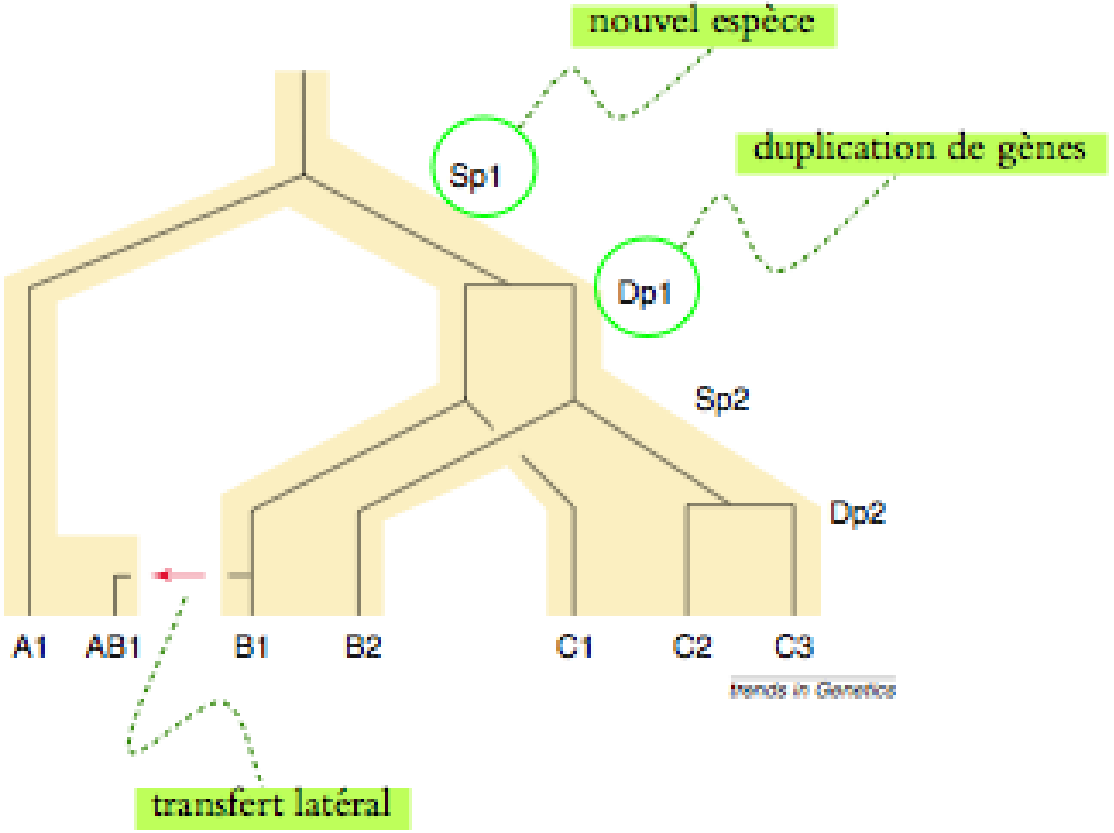
→ paralogue : relié par événement de duplication

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

similarité n'implique pas toujours la homologie : évolution convergente

homologie n'implique pas toujours la similarité non plus. . .

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)

Génomique comparative

Alignement de deux séquences :
- évidence de homologie
- conservation indique la fonctionnalité
- étudier les mécanismes de mutation
- étudier les forces d'évolution

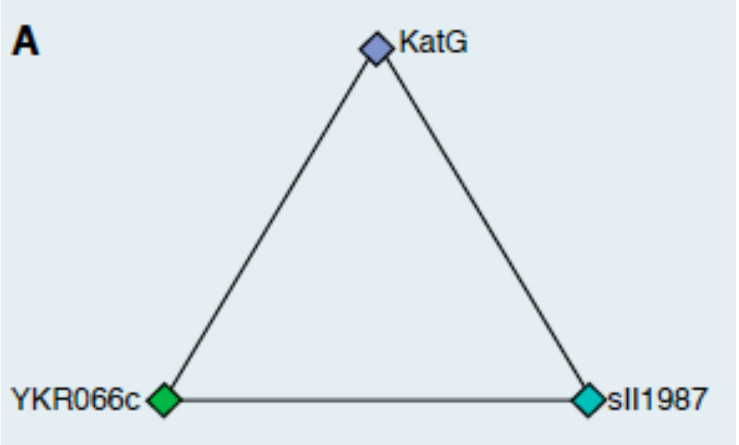
p.e. comparer le taux de mutations synonymes (entre codons encodant le même acide aminé) et celui de mutations non-synonymes

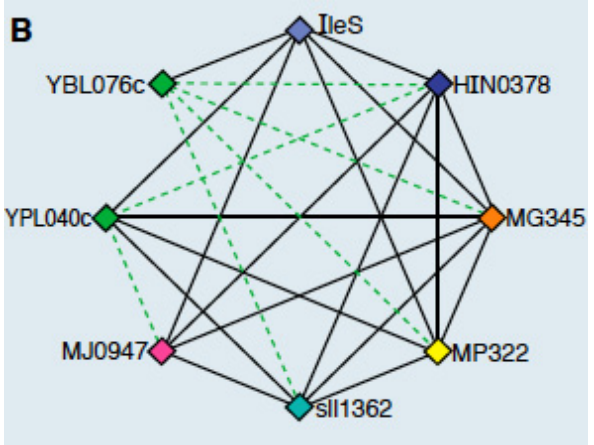
évolution neutre : aucune différence
évolution/sélection purificatrice/négative : synonyme plus fréquent
sélection positive [évolution Darwinien] : non-synonyme plus fréquent

Familles d'orthologues

BeT : [BLAST] *Best hit* (chaque "gène" de génome A comparé à ceux de génome B)

COGs (*Clusters of Orthologous Groups*) : déterminé par le graphe de BeTs





triangles formés par BeTs symétriques → squelette d'un COG

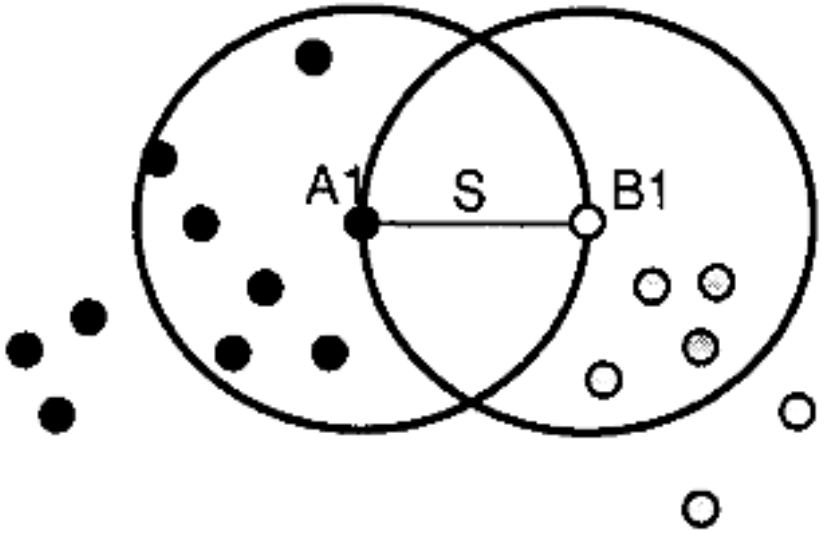
+ ajouter d'autres gènes avec des BeTs symétriques avec le groupe

+ inspection humaine

Tatusov & al. *Science* 278 :631 (1997)

Groupage — INPARANOID

Pour deux génomes



graphe BeTs symétriques — arêtes pondérées par score BLAST

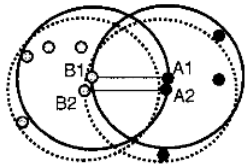
tout automatique

Remm & al *J Mol Biol* 314 :1041 (2001)

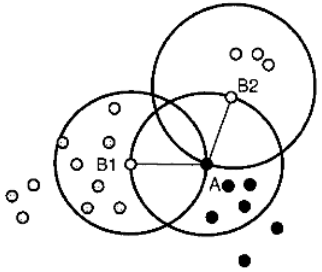
Règles — INPARANOID

Pour des groups chevauchant

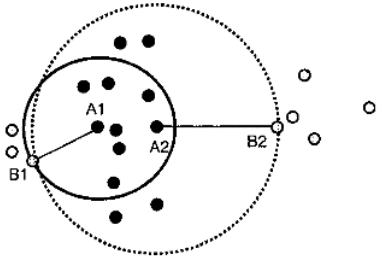
1) MERGE IF BOTH ORTHOLOGS ARE ALREADY CLUSTERED IN THE SAME GROUP



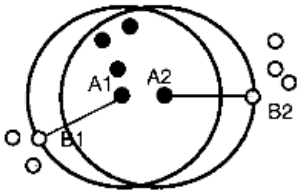
2) MERGE IF TWO EQUALLY GOOD BEST HITS FOUND



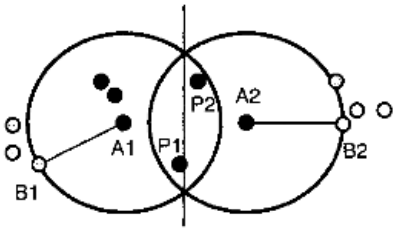
3) DELETE WEAKER GROUP IF (SCORE(A2-B2) - SCORE(A1-B1) > 50 bits)



4) MERGE IF (SCORE(A1-A2) < 0.5 * SCORE(A1-B1))



5) DIVIDE IN-PARALOGS IN OVERLAPPING AREAS

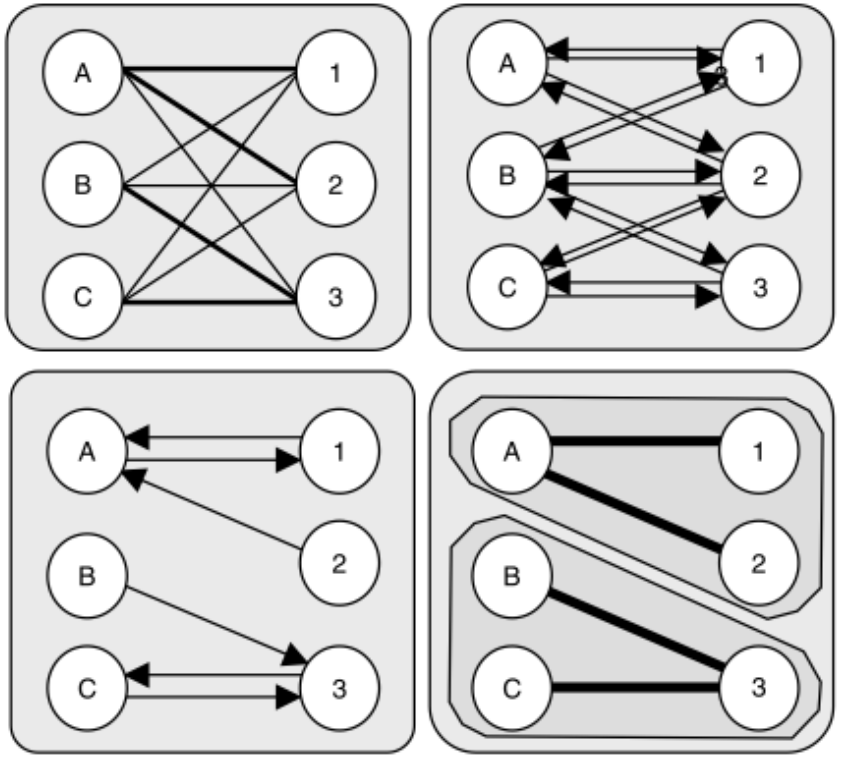


Remm & al *J Mol Biol* 314 :1041 (2001)

BUS — *best unambiguous subset*

Algorithme utilisé pour déterminer les orthologues parmi des levures

Graphe de BeTs — arêtes pondérées par identité aa et longueur



Kellis & al *J Comput Biol* 11 :319 (2004)

BUS

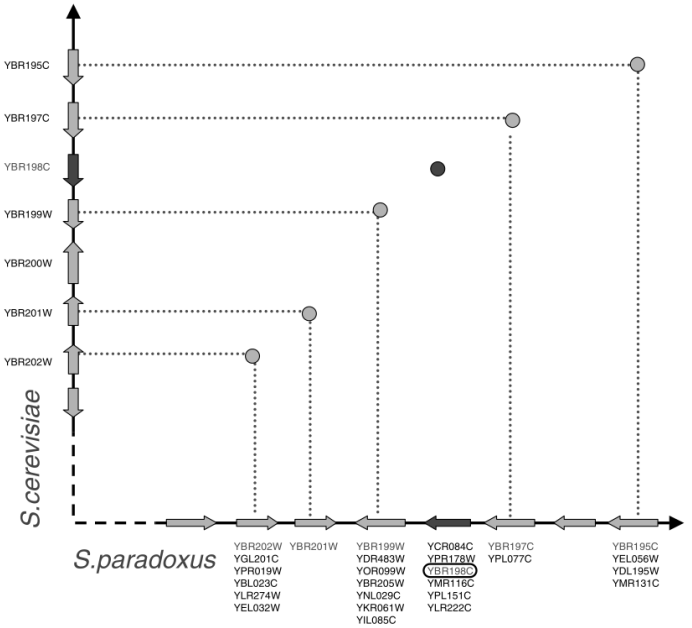
1. construction du graphe biparti

2. enlever les arêtes sous-optimales : score et longueur inférieurs à la 80% de la meilleure arête au sommet [! seuil relatif]

3. enlever des arêtes sortant de blocs de synténiques [! pas un «sac de gènes»]

4. composantes connexes dans le graphe de meilleures arêtes $\Rightarrow$ sous-ensembles non-ambigus

BUS — blocs synténiques



bloc : au moins trois paires de gènes proches — paires sont formés par des cycles de taille 2 après étape 2

si un gène dans génome A est proche de ce bloc, alors on élimine les arêtes entre ce gène et des gènes dans génome B qui ne sont pas proches au bloc

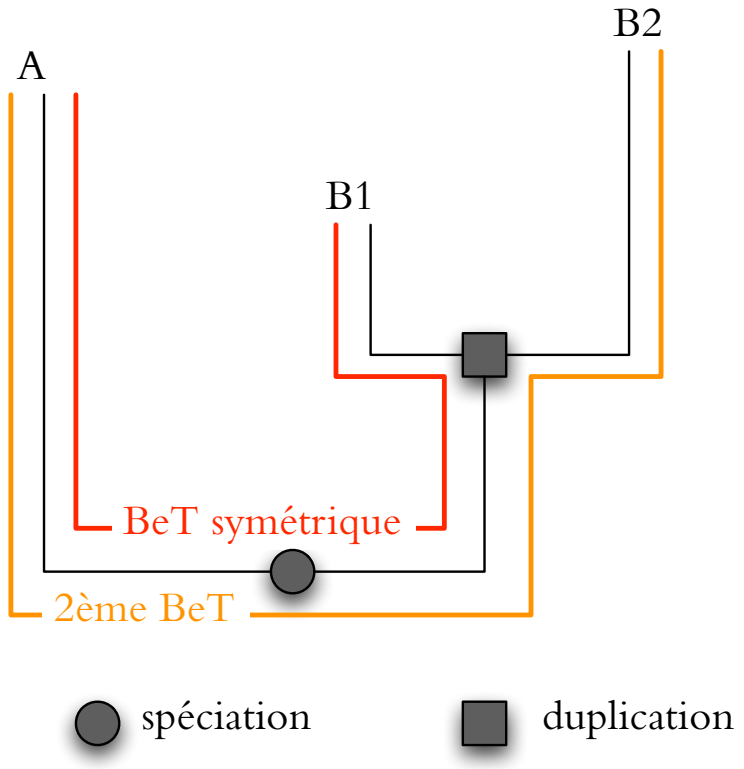
«proche» : distance $\leq$ 20kbp (moyenne entre gènes 2kbp)

Kellis & al *J Comput Biol* 11 :319 (2004)

BeTs ne suffissent pas

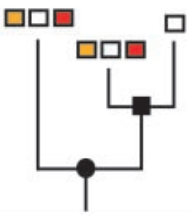
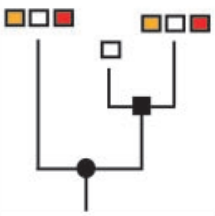
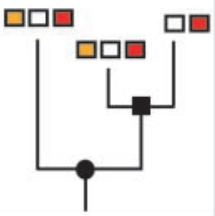
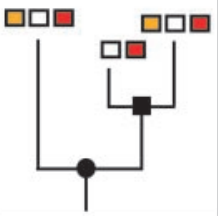
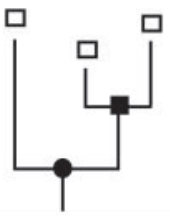
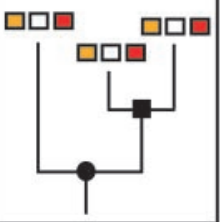
on considère des *in-paralogs*

est-ce que l'orthologue est toujours le *best hit*?



Notebaart & al *Nucleic Acids Res* 33 :6164 (2005)

BeTs et contexte

	1) 	2) 	3) 	4) 	5) 	6) 
G+	72	37	11	15	396	47
P	50	29	10	4	337	56
A	32	13	7	1	254	38
G+ and P	136	85	13	20	1219	103
G+ and A	52	31	13	14	822	68
P and A	44	25	19	8	687	55
Total	386	220	73	62	3715	367

■ Duplication ● Speciation

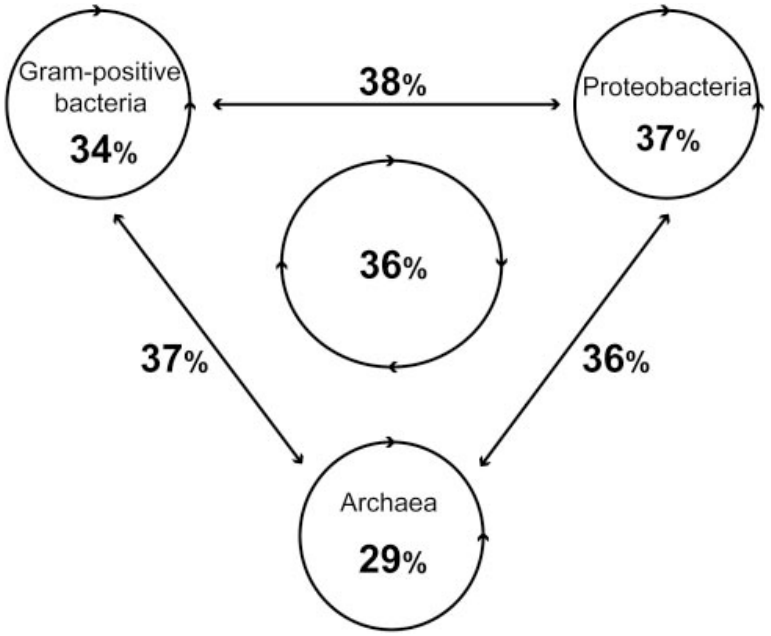
[groupes : (A) archebactéries (P) protéobactérie (G+) bactéries Gram+]

notion de contexte est beaucoup plus compliquée dans des eucaryotes : pas d'opérons

Notebaart & al *Nucleic Acids Res* 33 :6164 (2005)

BeTs et contexte

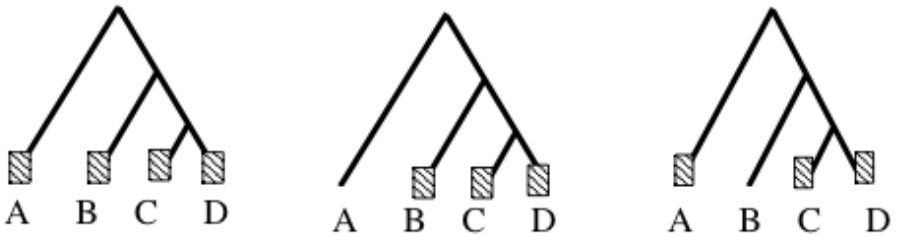
pourcentage de cas où il ya de la préservation de contexte pour un des co-orthologues mais le vrai orthologue est le deuxième *best hit* :



Notebaart & al *Nucleic Acids Res* 33 :6164 (2005)

Profiles phylétiques

pour un COG — enregistrer dans quels espèces il y a au moins un membre du groupe



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

→ prédiction de fonction (profiles pareils)

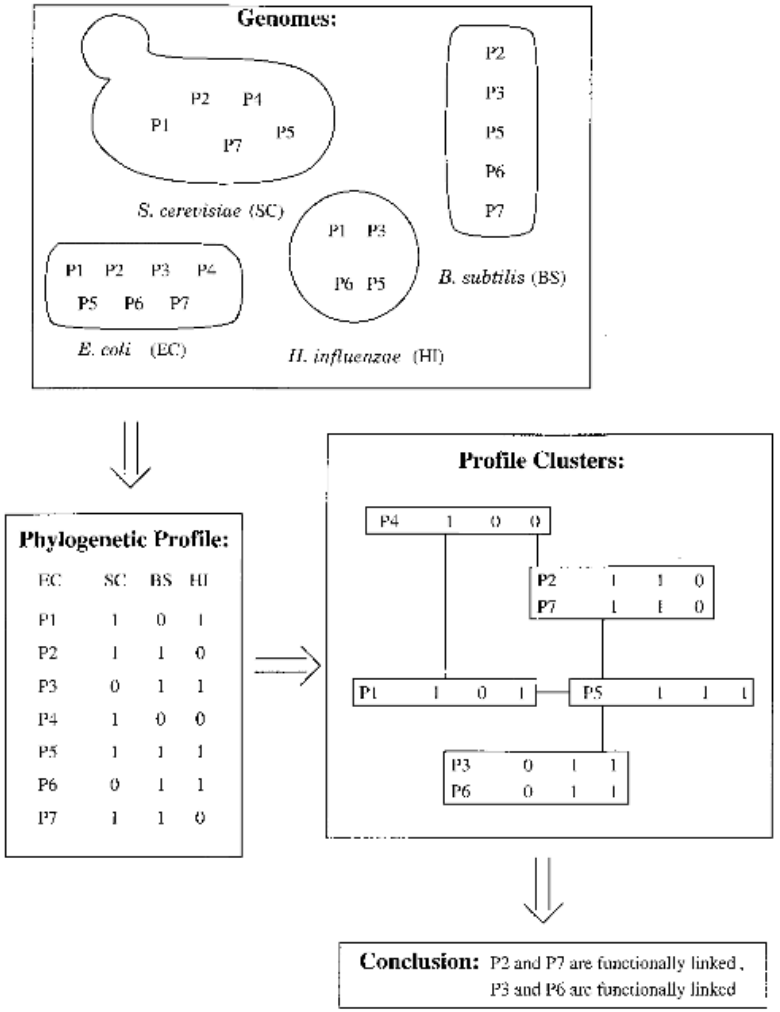
→ évolution de répertoire de gènes

→ phylogénie d'espèces

→ validation d'annotation

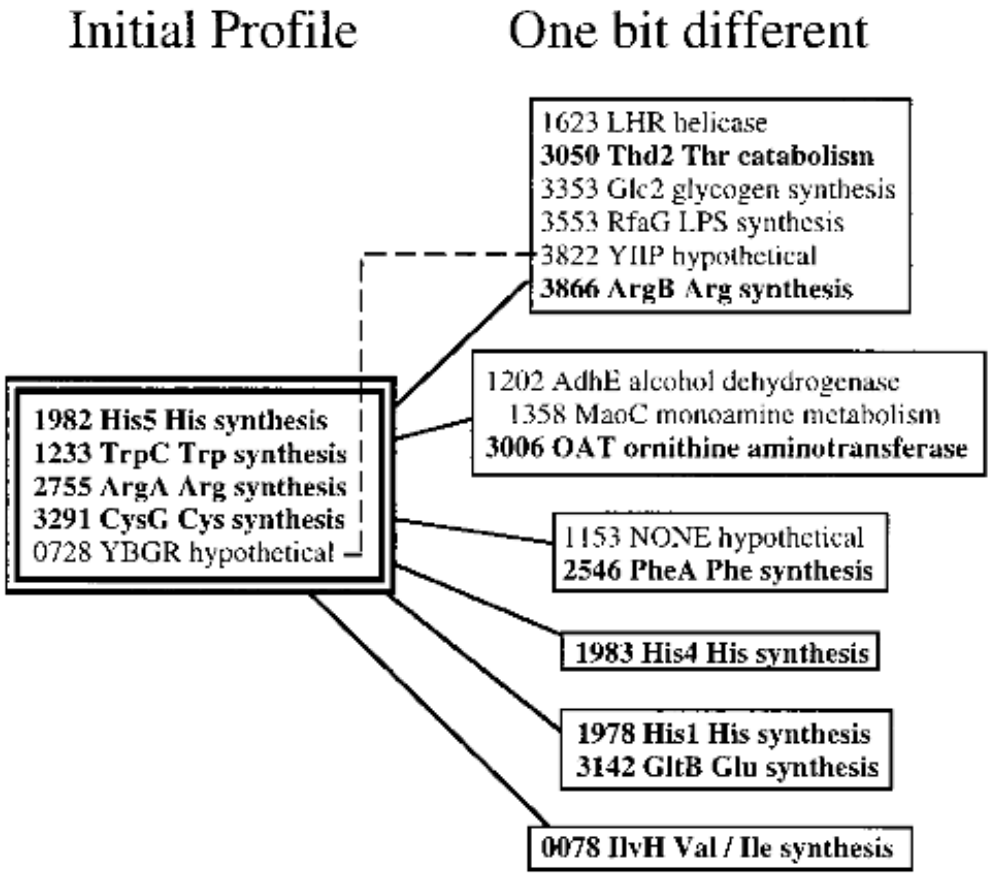
Tatusov & al *BMC Bioinformatics* 4 :41 (2003)

Prédiction de fonction



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

Fonction et profile phylétique



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

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

Profiles complémentaires

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

Exemple : synthase de thymidilate

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

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

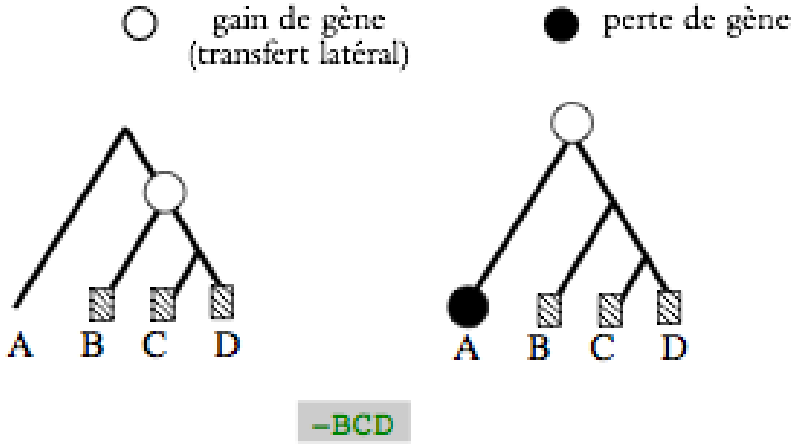
COG1351 (fonction inconnue) : `-o-pkz-qv-r--c------u-xit-`

NOGD — *Non-orthologous gene displacement*

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

Évolution du répertoire de gènes

Scénarios d'évolution



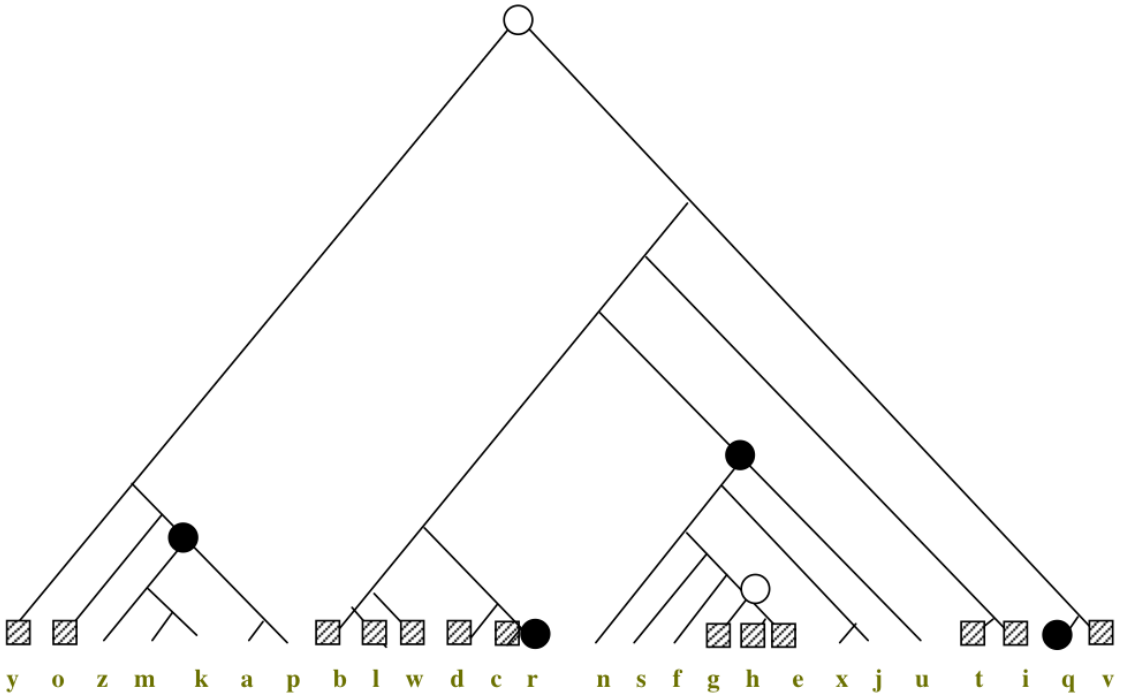
et d'autres scénarios avec plus d'événements

On peut calculer le meilleur scénario par parcimonie

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

Évolution du répertoire de gènes

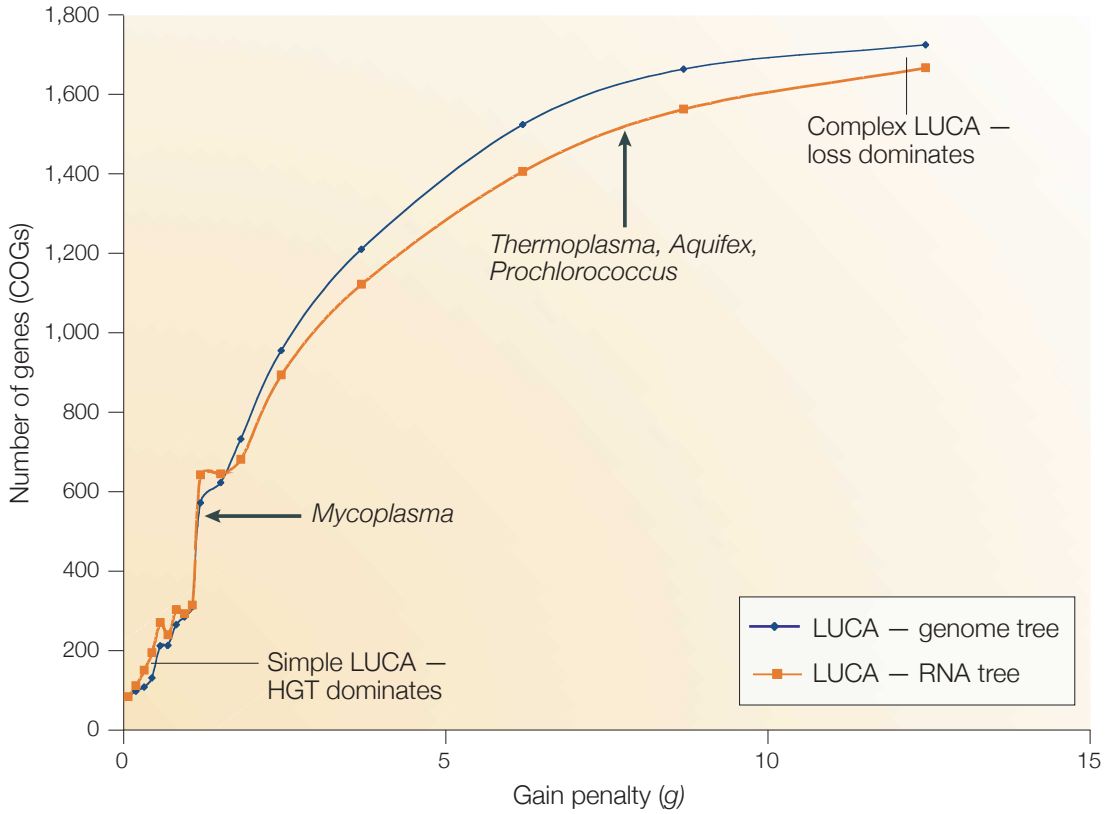
questions : (1) est-ce que le gène à la racine est pénalisé ? ; (2) quel est le score rélatif des gains et des pertes ?



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

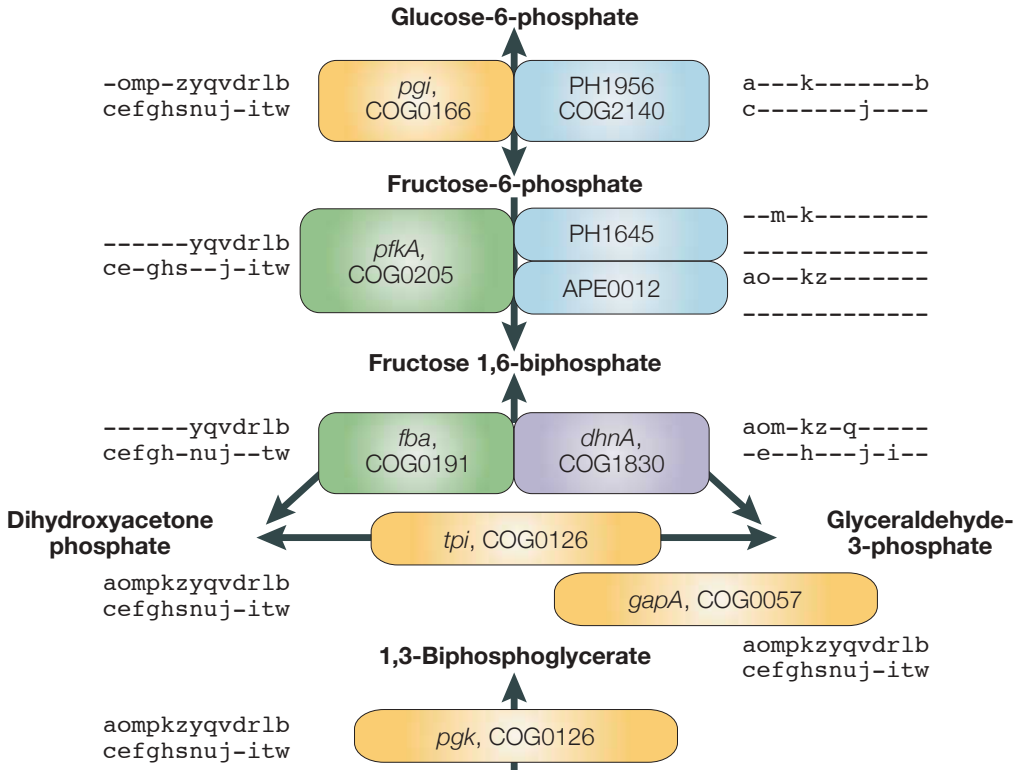
Gènes de LUCA

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

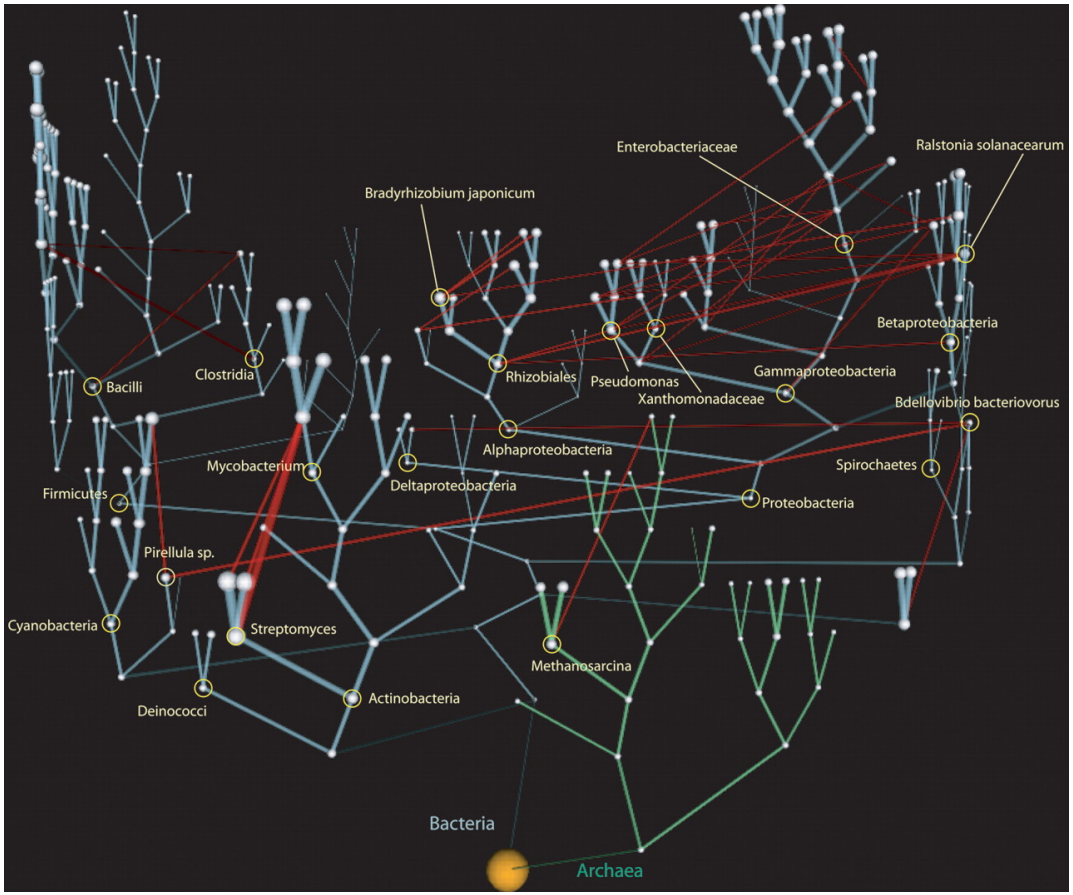


couleur	coût de gain (g)
jaune	0.9
vert	1
violet	2
bleu	non-LUCA

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

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

Fréquence de transfert latéral

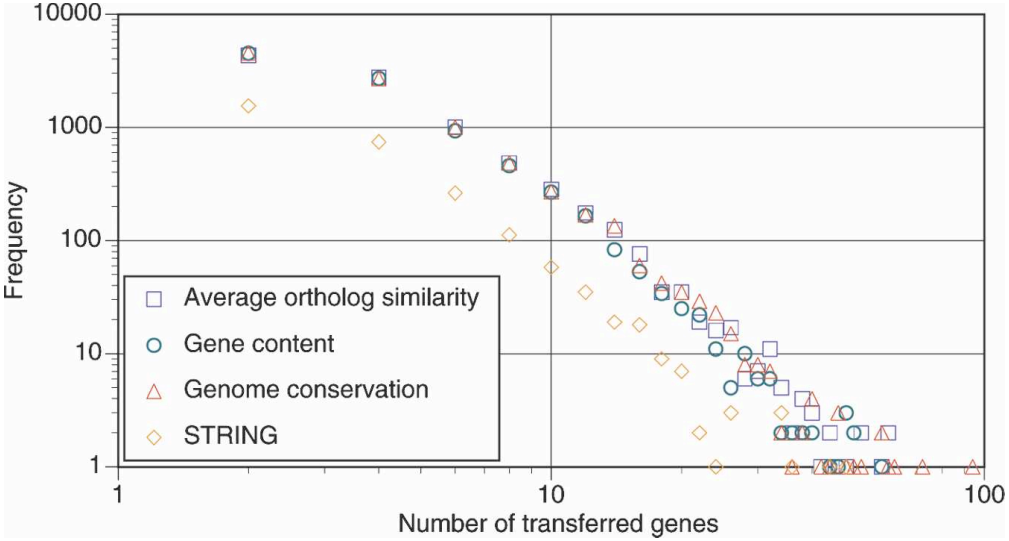


Kunin & al *Genome Res* 15 :954 (2005)

Fréquence de transfert latéral

Transfert latéral est rélativement fréquent
(0–7% de gènes dans un génome bactérien)

? *power law* ?



[symboles dénotent des méthodes différentes de reconstruction phylogénétique et des données d'orthologie]

Kunin & al *Genome Res* 15 :954 (2005)

Duplication

le mécanisme classique d'invention

quand un gène est dupliqué, une des copies peut acquérir une nouvelle fonction sans pression purificatrice (Ohno 1970)

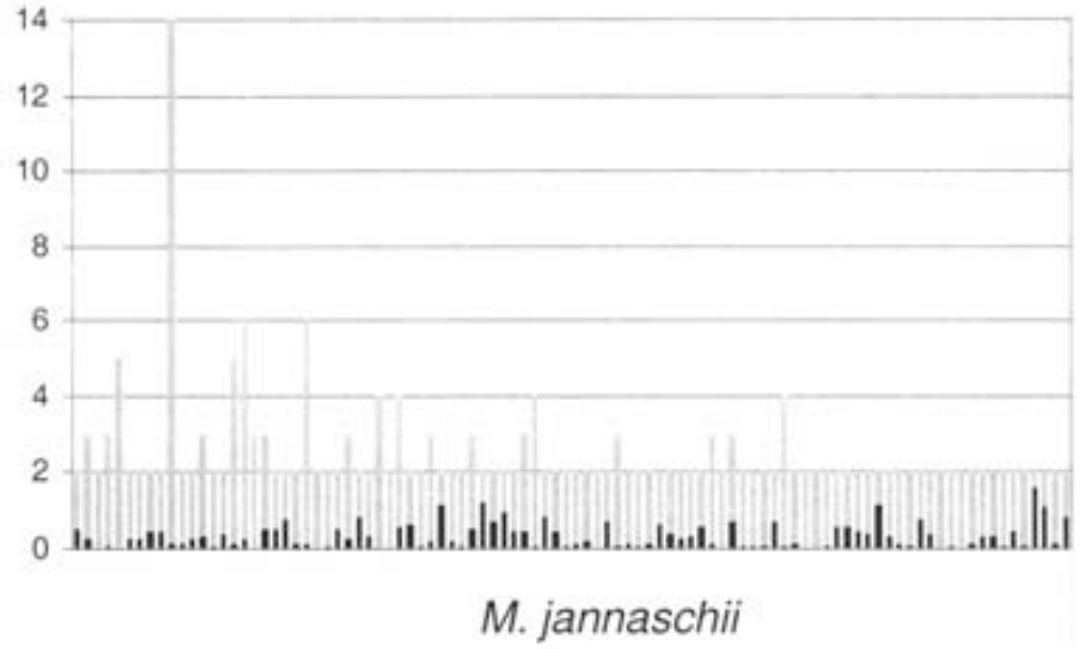
cas très intéressant : grande expansion d'une famille de gènes
(LSE - *Lineage Specific Expansion*)

→ fonction très avantageuse pour l'organisme (sélection positive)

typiquement 10–20% de gènes d'un procaryote

Identification de LSEs

méthode : (1) groupage de gènes dans un génome, (2) BeTs contre un génome de référence — un groupe de LSE devrait avoir le même BeT dans l'autre génome ou aucun *hit*



[gris : taille du groupe, noir : nombre moyen de BeTs dans un autre génome]

Jordan & al *Genome Res* 11 :555 (2001)

Exemples de LSE

[similarité moyenne entre deux membres dans une position alignée]

pathogènes

invasion des cellules de l'hôte

Species	Cluster size	Average score density	Domain organization[b]	Function
M. tuberculosis	90	0.38	Multitransmembrane proteins; PPE family	Predicted surface protein, interaction with host cells
M. tuberculosis	67	0.37	Signal-peptide-containing, non-globular proteins, consist mostly of glycine-rich repeats; PE family	Predicted surface protein, interaction with host cells
H. pylori	34	0.34	Outer membrane protein	Predicted surface protein, interaction with host cells
E. coli	31	0.30	Helix-turn-helix DNA-binding domain (LysR family), solute-binding domain	Transcription regulation of various metabolic operons
Synechocystis sp.	30	0.26	Histidine kinase	Signal transduction, sensing of environmental stimuli
M. pneumoniae	25	0.62	Predicted non-globular domain	Unknown
M. tuberculosis	24	0.21	Signal-peptide-containing protein	Predicted surface protein (mce1), interaction with host cells
A. fulgidus	24	0.23	Histidine kinase	Signal transduction, sensing of environmental stimuli
Synechocystis sp.	22	0.39	Diguanylate cyclase/phosphodiesterase (GGDEF and EAL domains)	Signal transduction, sensing of environmental stimuli
M. tuberculosis	21	0.29	Short chain dehydrogenase	Dehydrogenases with different specificities (related to short-chain alcohol dehydrogenases)
M. tuberculosis	20	0.45	Beta-ketoacyl synthase, acyl transferase, thioesterase	Polyketide synthase

Jordan & al *Genome Res* 11 :555 (2001)