Prédiction de gènes

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], . . .)

Rappel : 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

Terminologie

similarité : notion algorithmique de relation entre séquences

homologue : relié par un ancêtre commun

→ orthologue : relié par événement de spéciation

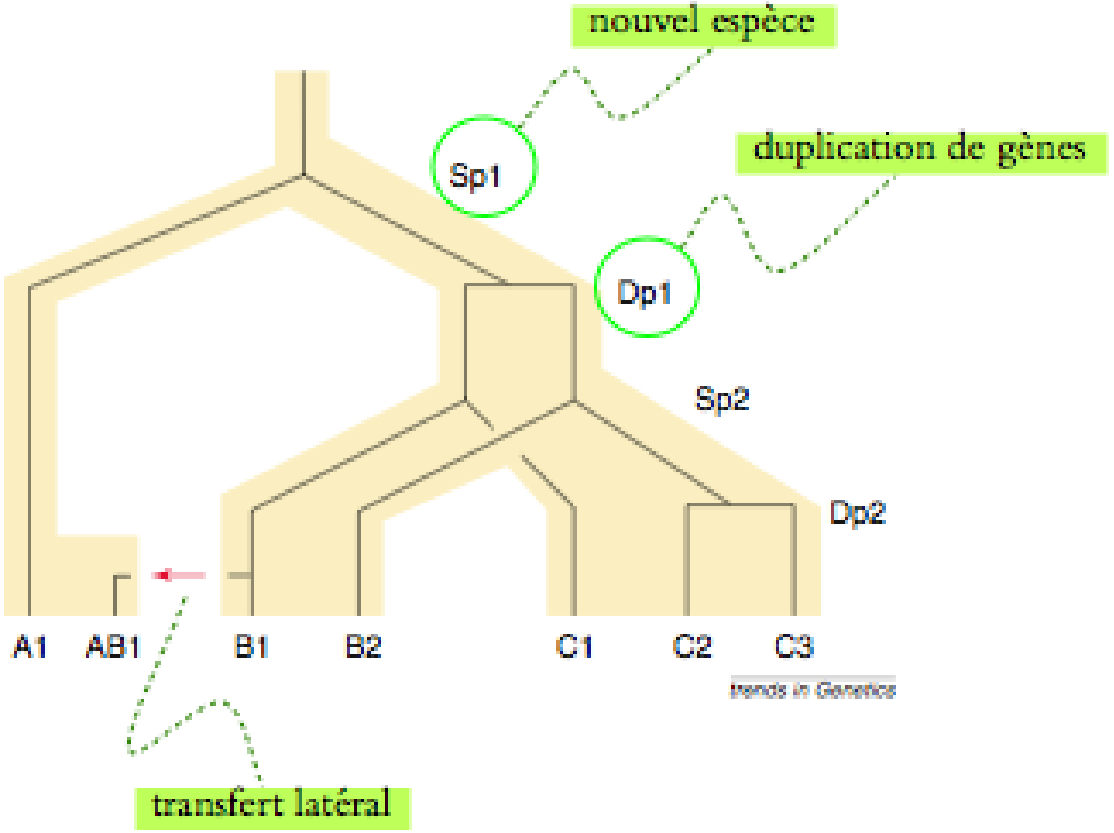
→ paralogue : relié par événement de duplication

→ 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

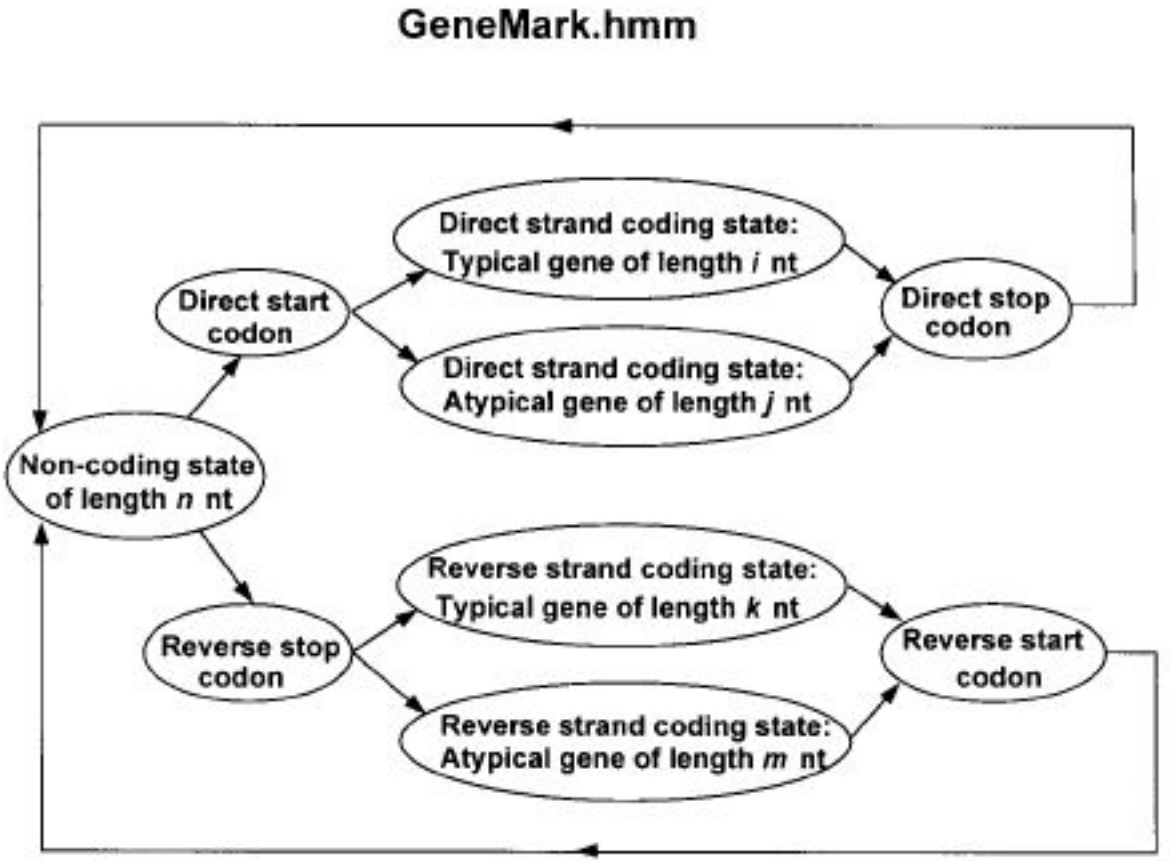
Gènes traduits — procaryotes

ORF : Open Reading Frame

`ATG`...Ter où Ter $\in \{$`TAA`, `TAG`, `TGA`$\}$

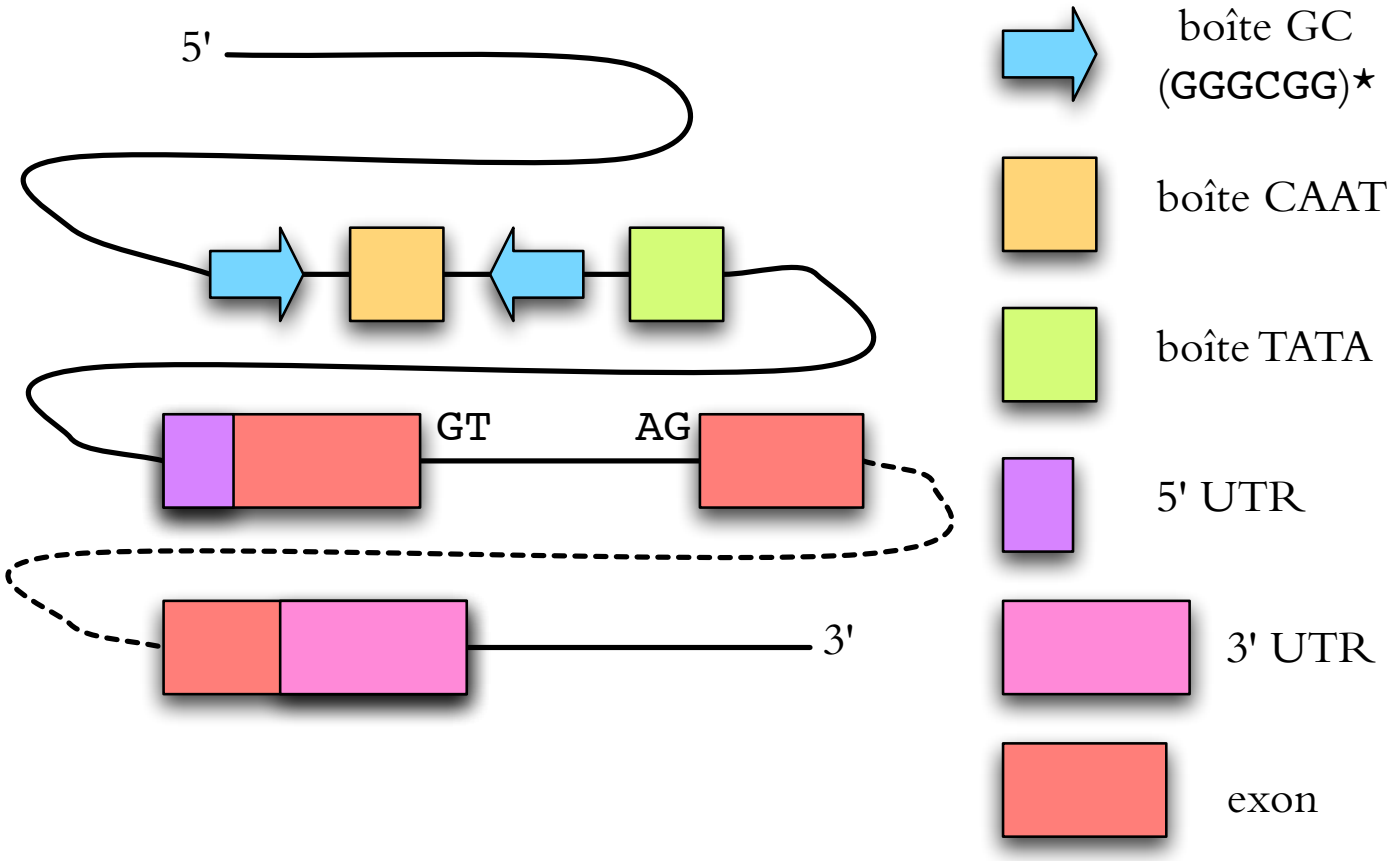
Gènes traduits — procaryotes

Procaryotes (pas d'exons !) — GeneMark



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

GÈNES TRADUITS — EUCARYOTES

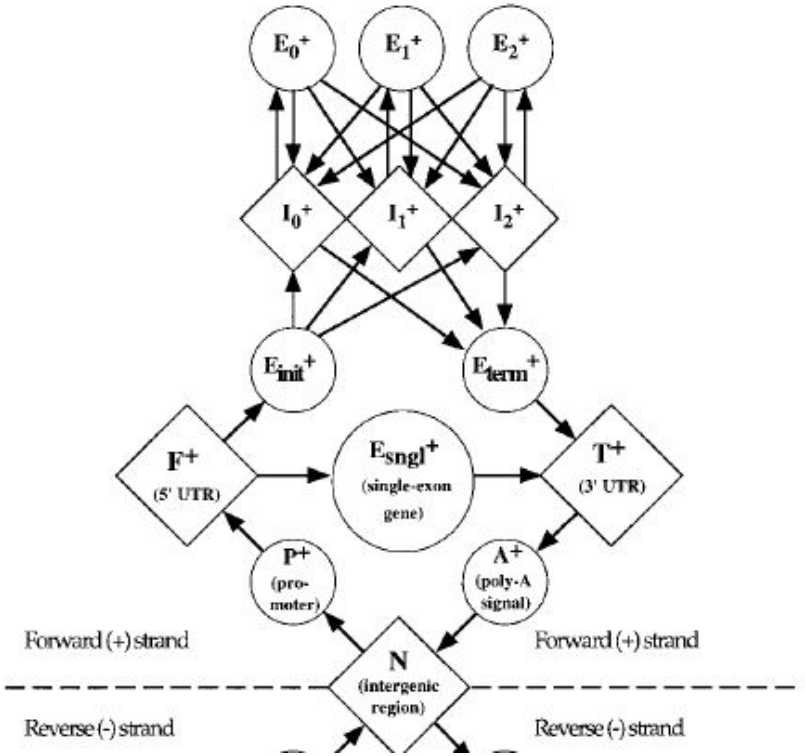


[processes : spliceosome, ribosome]

après Graur & Li *Fundamentals of Molecular Evolution* (2000)

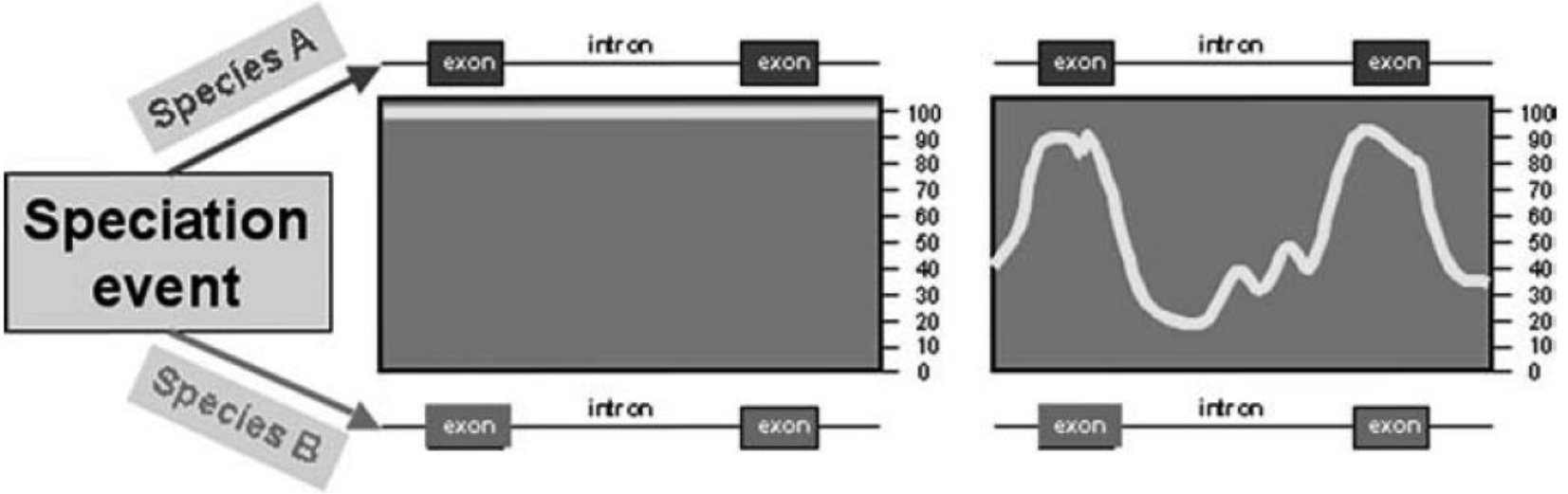
Prédiction de gènes (Genscan)

Eucaryotes — GENSCAN



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

Méthode comparative pour la recherche de gènes

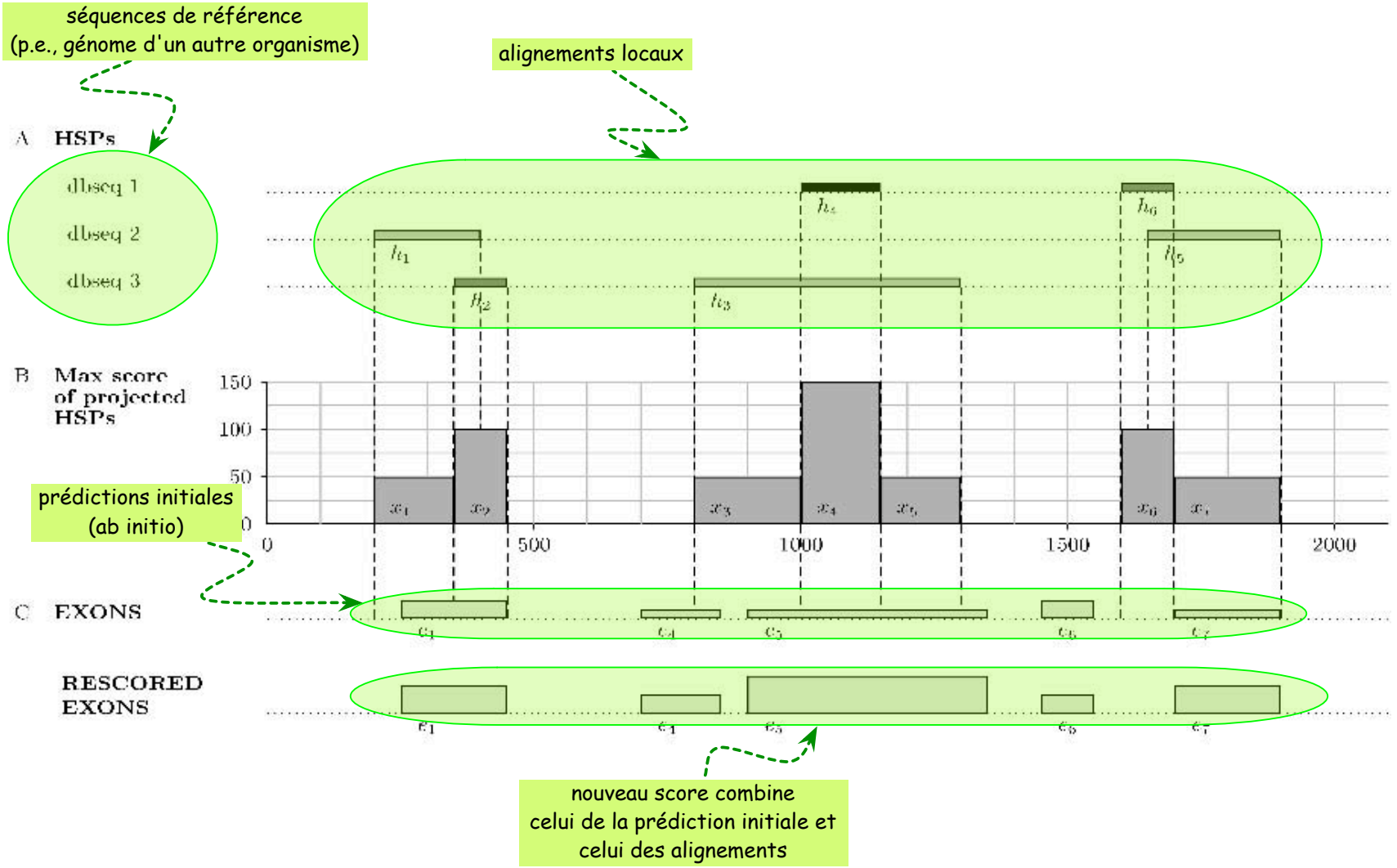


Alignement de deux régions aide à l'identification d'exons : les exons sont plus préservés (sélection négative)

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)

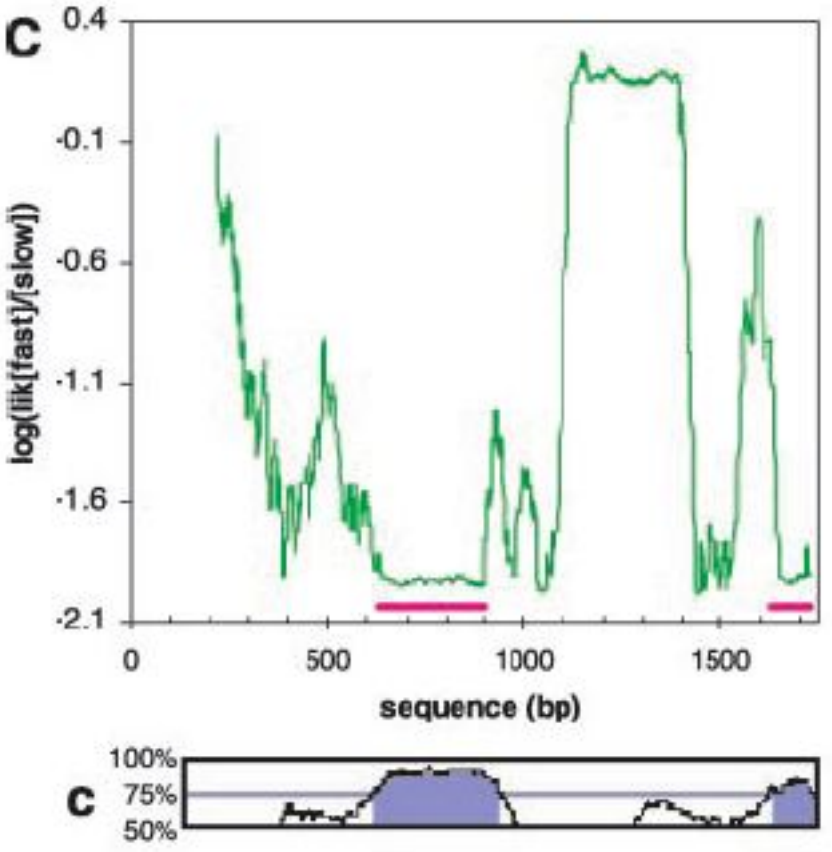
Prédiction de gènes (SGP2)



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

Phylogenetic shadowing

Comparaison de séquences entre des espèces proches :
modèles d'évolution rapide/lente (HMM)

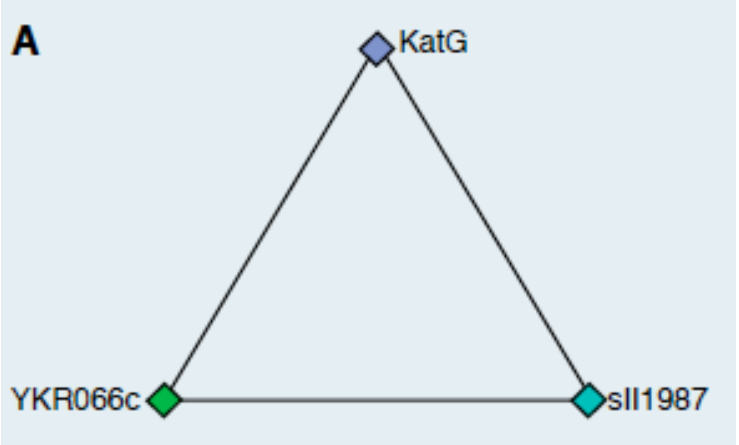


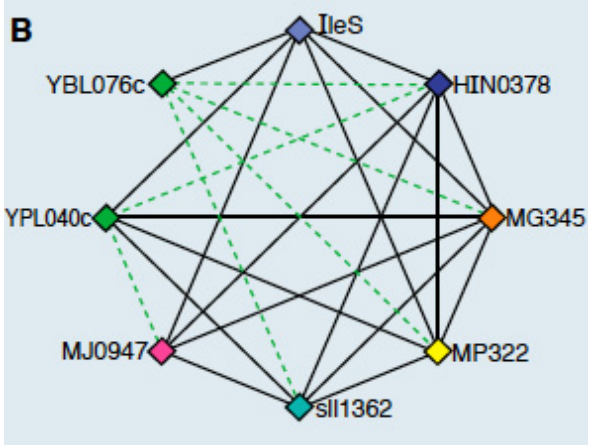
Boffelli et al. *Science* 299 :1391 (2003)

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

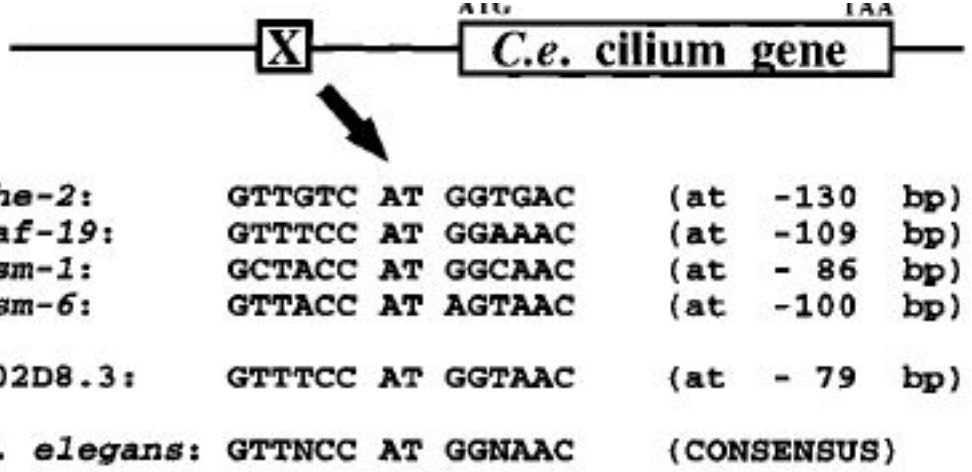
+ ajouter d'autres gènes avec des BeTs symétriques avec le groupe

+ inspection humaine

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

Recherche de motifs

Site de liaison de facteur de transcription : signal de régulation de l'expression génique



Placement inconnu, copies non-identiques

Swoboda & Adler & Thomas *Mol Cell* 5 :411 (2000)

Recherche de motifs 2

Problème abstrait : séquences $S_1, \ldots, S_n$, avec des copies d'un motif M

Si on a une représentation du motif (signature, HMM, PSSM,. . .) on sait comment trouver les copies

Qu'est-ce qu'on fait si on connait pas M ?

Problème difficile. . .

Solution 1 : force brut — essayer tous les m-mers comme candidats de M, trouver les m-mers les plus proches à M en chaque séquence, et calculer un score (p.e., moyenne de distance de M, ou entropie)

Branch & bound

Recherche de motifs 3

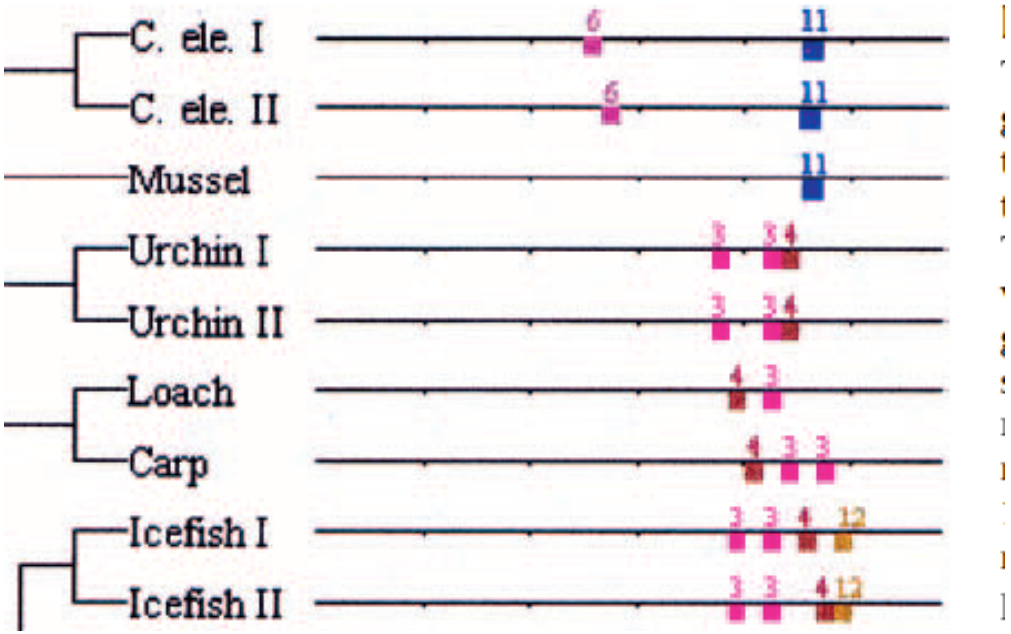
Score : étant donné des copies $M_1, \ldots, M_n$, former profile et calculer le score comme LODS en utilisant un modèle de background

Solution 2 : Gibbs sampler

1. copies $M_1, \ldots, M_n$
2. enlever M_j et le remplacer par le meilleur m-mer en S_j (comparaison au profile défini par $M_1, \ldots, M_{j-1}, M_{j+1}, \ldots, M_n$)
3. répéter jusqu'à convergence

Phylogenetic foorptinting

Régions non-codantes conservées parmi des espèces distants : éléments de régulation



Blanchette & Tompa, *Genome Res.* 12 : 739 (2002)