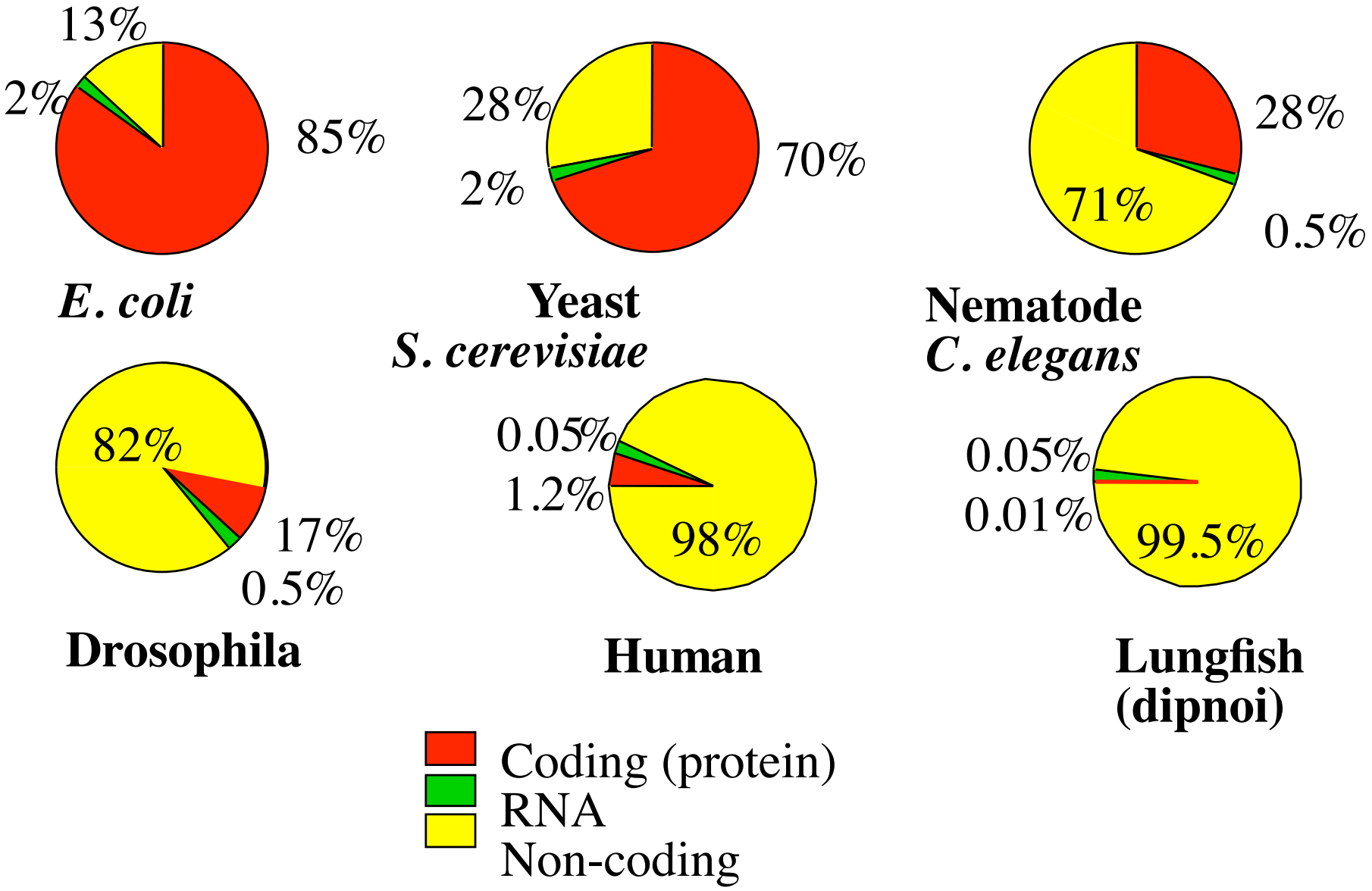
Bioinformatique: Annotation des génomes (eucaryotes)

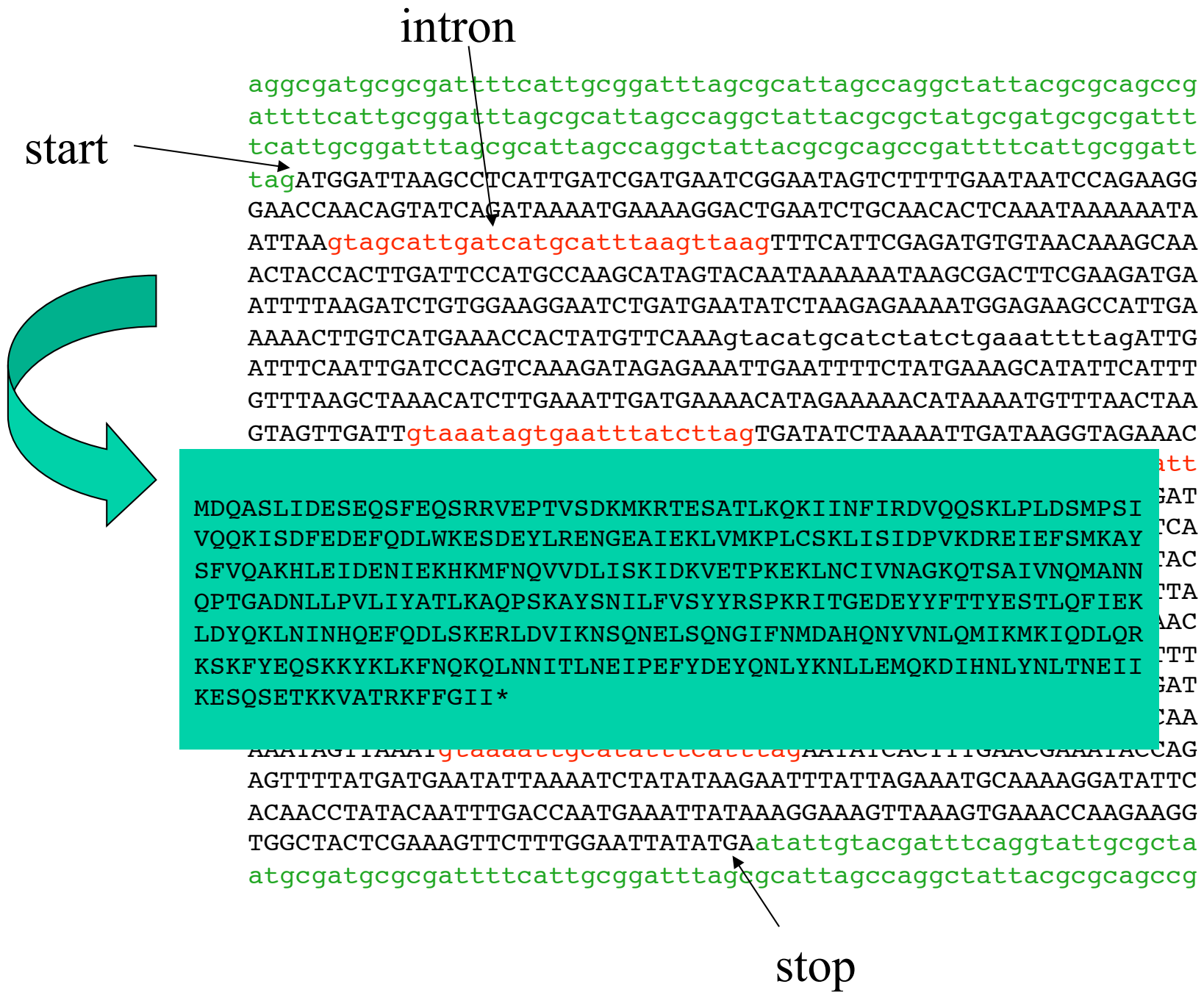
E2M2– Avril 2009

Laurent Duret
BBE – UMR CNRS n° 5558
Université Claude Bernard - Lyon 1

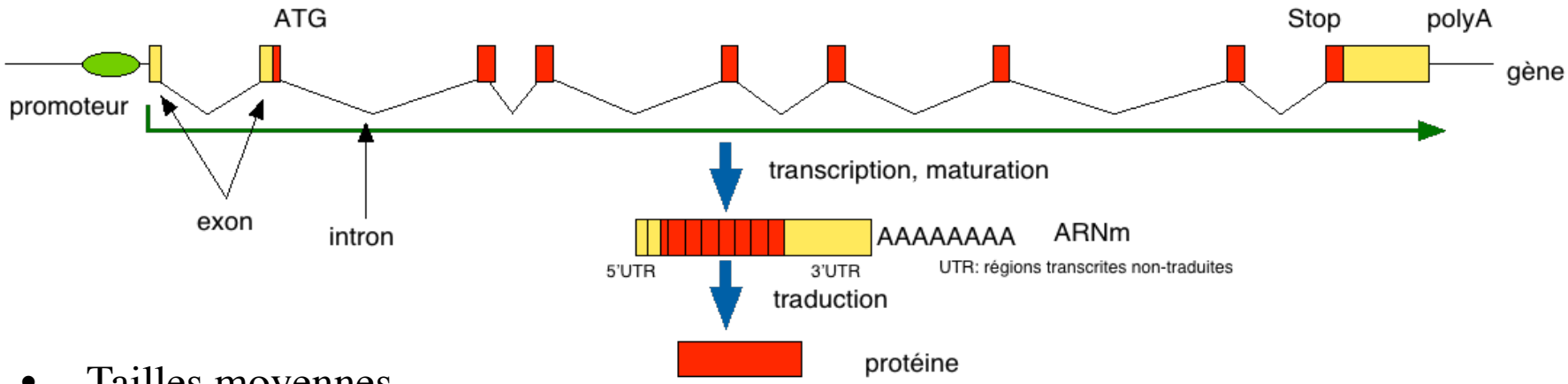
Proportion of functional elements within genomes



Annotation of protein-coding genes


intron
start
aggcgatgcgcgattttcattgcggatttagcgcattagccaggctattacgcgcagccg
attttcattgcggatttagcgcattagccaggctattacgcgctatgcgatgcgcgattt
tcattgcggatttagcgcattagccaggctattacgcgcagccgattttcattgcggatt
tagATGGATTAAGCCTCATTGATCGATGAATCGGAATAGTCTTTTGAATAATCCAGAAGG
GAACCAACAGTATCAGATAAAATGAAAAGGACTGAATCTGCAACACTCAAATAAAAAATA
ATTAAgtagcattgatcatgcatttaagttaagTTTCATTCGAGATGTGTAACAAAGCAA
ACTACCACTTGATTCCATGCCAAGCATAGTACAATAAAAAATAAGCGACTTCGAAGATGA
ATTTTAAGATCTGTGGAAGGAATCTGATGAATATCTAAGAGAAAATGGAGAAGCCATTGA
AAAACTTGTCATGAAACCACTATGTTCAAAgtacatgcatctatctgaaattttagATTG
ATTTCAATTGATCCAGTCAAAGATAGAGAAATTGAATTTTCTATGAAAGCATATTCATTT
GTTTAAGCTAAACATCTTGAAATTGATGAAAACATAGAAAAACATAAAATGTTTAACTAA
GTAGTTGATTgtaaatagtgaatttatcttagTGATATCTAAAATTGATAAGGTAGAAAC
MDQASLIDESEQSFEQSRRVEPTVSDKMKRTESATLKQKIINFIRDVQQSKLPLDSMPSI
VQQKISDFEDEFQDLWKESDEYLRENGEAIEKLVMKPLCSKLISIDPVKDREIEFSMKAY
SFVQAKHLEIDENIEKHKMFNQVVDLISKIDKVETPKEKLNCIVNAGKQTSAIVNQMANN
QPTGADNLLPVLIYATLKAQPSKAYSNILFVSYYRSPKRITGEDEYYFTTYESTLQFIEK
LDYQKLNINHQEFQDLSKERLDVIKNSQNELSQNGIFNMDAHQNYVNLQMIKMKIQDLQR
KSKFYEQSKKYKLKFNQKQLNNITLNEIPEFYDEYQNLYKNLLEMQKDIHNLYNLTNEII
KESQSETKKVATRKFFGII*
AGTTTTATGATGAATATTAAAATCTATATAAGAATTTATTAGAAATGCAAAAGGATATTC
ACAACCTATACAATTTGACCAATGAAATTATAAAGGAAAGTTAAAGTGAAACCAAGAAGG
TGGCTACTCGAAAGTTCTTTGGAATTATATGAatattgtacgatttcaggtattgcgcta
atgcgatgcgcgattttcattgcggatttagcgcattagccaggctattacgcgcagccg
stop

Structure des gènes protéiques humains



- Tailles moyennes
  - Gene 45 kb
  - CDS 1500 nt
  - Exon (interne) 145 nt
  - Intron 5200 nt
  - 5'UTR 210 nt
  - 3'UTR 740 nt
- Intron/exon
  - Nombres d'introns: 6 ±3 introns / kb CDS
  - Introns / (introns + CDS): 92%
- Epissage alternatif dans plus de 30% des gènes

Quelles sont les approches envisageables pour identifier les gènes dans une séquence génomique ?

Prédiction de gènes: informations utilisées

- 1- caractérisation de la taille et du contenu des régions (codantes/non-codantes)
- 2- caractérisation des signaux au niveau de sites fonctionnels (e.g. signaux d'épissage, début et fin de traduction, ...)
- 3- données expérimentales: transcriptome (protéome)
- 4- conservation des régions fonctionnelles au cours de l'évolution

- Méthodes de prédiction de gènes
  - *ab initio* (méthodes intrinsèques): utilisent 1 et 2
  - Prédictions par analyse du transcriptome: utilisent 3 et éventuellement 2
  - Prédictions par approche comparative: utilisent 4, et éventuellement 2

Prédiction de gènes : méthodes *ab initio*

- Prédiction des régions codantes uniquement !
- Recherche de phases ouvertes de lecture (ORF: open reading frame) = série de codon sans STOP

```
Phase +0
Phase +1
Phase +2
       ATGTACCGTCGATCGTAGCTTGATCGATCG
       TACATGGCAGCTAGCATCGAACTAGCTAGC
Phase -0
Phase -1
Phase -2
```

  - Taille moyenne des ORF: ± 150 nt

- Distinction codant/non-codant : contenu et taille des séquences
  - usage des codons: utilisation non aléatoire des codons synonymes
  - fréquence des amino-acides (e.g. tryptophane est rare)
  - corrélations entre amino-acides (codons) successifs
  - taille des exons et introns

  - Apprentissage sur un ensemble de gènes connus
  - Fréquence d'oligomères (e.g. hexamères)
  - chaines de Markov

Prédiction de gènes : méthodes *ab initio* (suite)

- Recherche de signaux: sites fonctionnels conservés
  - signaux d'épissage: site donneur, accepteur d'épissage, point de branchement
  - codon d'initiation de la traduction
  - codon stop

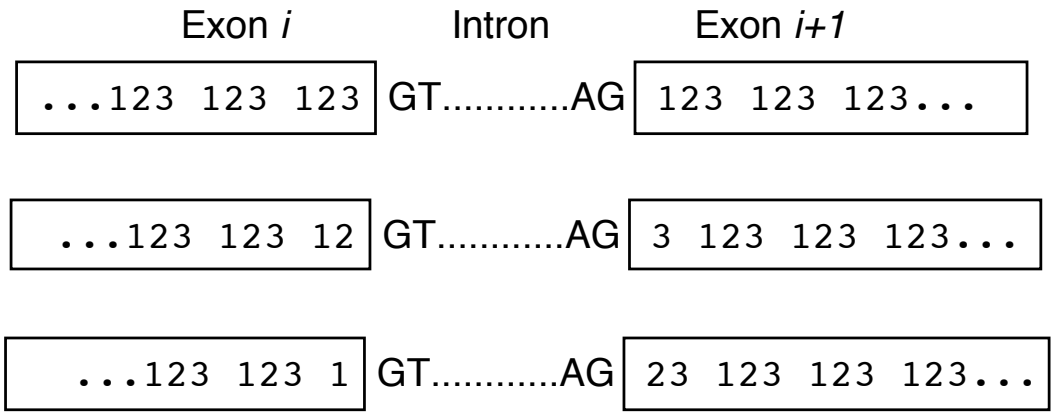
  - Utilisation de consensus (historique): e.g.

donneur	accepteur
A/CAG GT RAGT	YYYYYYYYY*C AG G

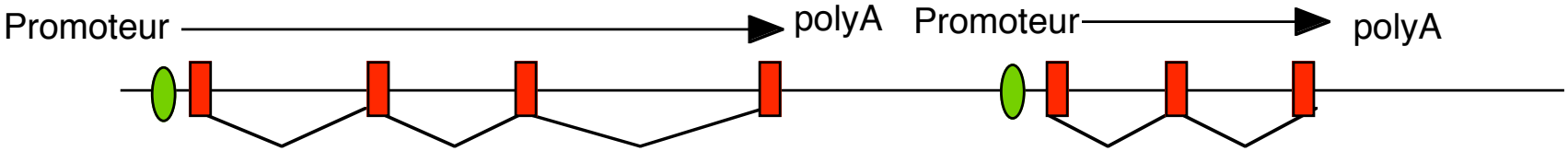
  - Utilisation de matrices de pondération position-dépendantes (profils)

Prédiction de gènes : méthodes *ab initio* (suite)

- Construction d ’un modèle de gène protéique
  - Combinaison d ’exons de phases compatibles (pondération en fonction des scores de chaque exon potentiel) - pas de codons stop en phase!



  - Recherche de limites de gènes
    - Exons terminaux (5 ’, 3 ’)
    - Promoteur
    - Signal de polyadénylation



Qualité de la prédiction par exon

- Évaluation de la fiabilité de la prédiction
  - essai des logiciels de prédiction sur un ensemble de séquences caractérisées expérimentalement (différentes de celles utilisées pour entrainer les logiciels)

- Sensibilité : fraction des exons présents dans la séquence qui sont retrouvés par le logiciel

$$sensibilité = \frac{VraisPositifs}{VraisPositifs + Faux\,Négatifs}$$

- Spécificité : fraction des vrais exons parmi tous ceux prédits

$$spécificité = \frac{Vrais\,Positifs}{VraisPositifs + Faux\,Positifs}$$

Prédiction de gènes eucaryotes: qualité de la prédiction

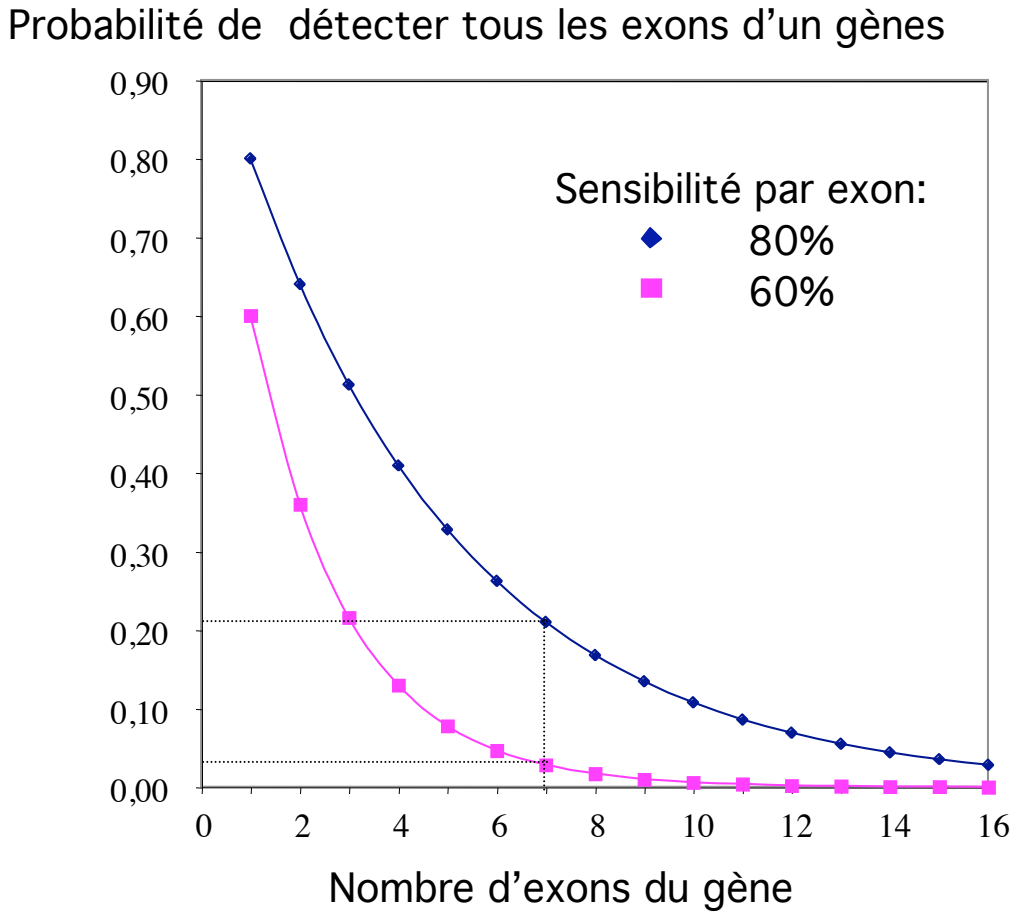
- Comparaison des différents logiciels: sensibilité/spécificité
  - Sn: sensibilité Sp: spécificité par exon (sn_e, sp_e) ou par nucéotide (sn_e, sp_e)

  - Locus BRCA2 (1.4 Mb, chrom. 13q) (Sanger Centre 1999): région "difficile" pour les logiciels de prédiction. 159 exons

	Sn_e	Sp_e	Sn_n	Sp_n
GenScan	0.66	0.36	0.81	0.44
FGENES 1.6	0.69	0.57	0.79	0.66
FGENES 1.6 masked	0.69	0.65	0.79	0.74
GenScan+FGENES	0.61	0.82	0.67	0.90

Prédiction de gènes protéiques complets

- Prédiction de gènes complets: sensibilité ?



  - \+ les faux positifs ! + épissage alternatif ! + exons non-codants !

Un peu d ’optimisme

- Fraction de la longueur des gènes correctement prédits:

  70-80%

- Probabilité que deux exons potentiels consécutifs soient réels (et donc positifs en RT-PCR)

  0.5

Prédiction de gènes : méthodes *ab initio* (bilan)

- Procaryotes (pas d'intron):
  - sensibilité et spécificité > 95% (dépend du taux de G+C du génome)
- Eucaryotes: efficacité variable (dépend du taux de G+C du génome et du nombre et de la taille des introns)
  - prédiction d'exons: sensibilité et spécificité 60-80%
  - prédiction de gènes complets:
    - levure: >90% des gènes correctement prédits
    - nématode: 50% des gènes correctement prédits
    - homme: 20% (?) des gènes correctement prédits
- très utile pour guider les expérimentations

Prediction of protein genes from transcriptome data: mRNA / DNA comparison

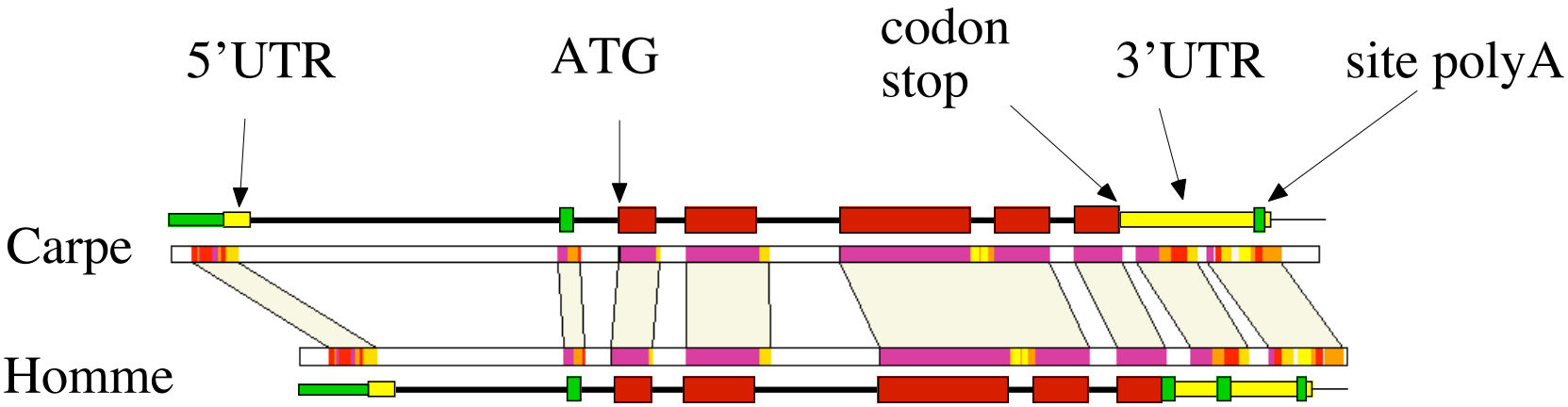
- Large scale transcriptome data
  - ESTs
  - full-length cDNA sequencing projects

- Alignment genomic DNA / mRNA : identification of exons (blastn, sim4, est2genome)
  - information on alternative splicing, gene expression pattern
  - not restricted to protein-coding regions (UTRs, non-coding RNAs)
- Problems:
  - weakly expressed genes; genes with a restricted tissue-distribution
  - artefacts in EST sequences (contamination with nuclear RNA, DNA)

Prediction of protein genes: comparative approach

- Comparison of a genomic sequence with genes that have been already characterized (e.g. in other species)
  - DNA/protein alignments: blastx, genewise
  - Warning! choice of the reference sequence:
    - Validated experimentally, or
    - Evolutionary distant
- Comparaison of homologous genomic sequences
  - DNA/DNA alignments

Analyse comparative des gènes de β-actine de l'homme et de la carpe



introns:

régions codantes:

éléments régulateurs:

échelle de similarité:

pas de similarité significative

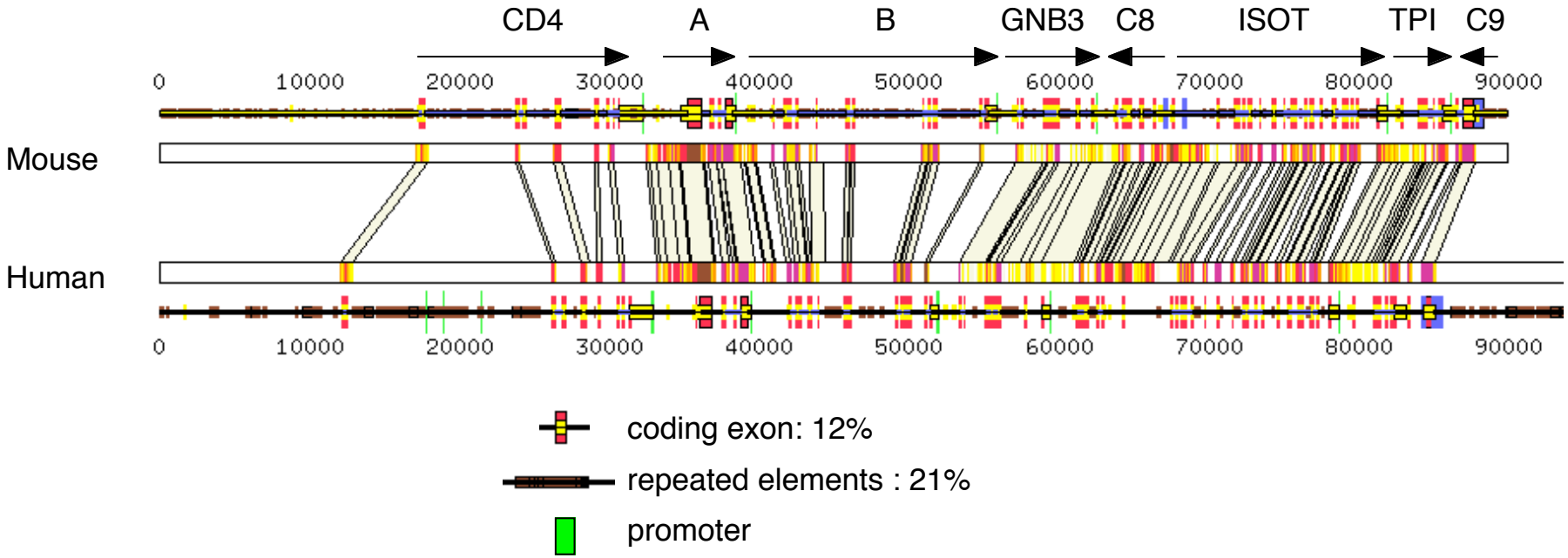
80 - 90% identité

70 - 80% identité

Comparison of human and mouse CD4-C9 locus: gene-rich, repeated-element poor, G+C-rich region (50.5%)

Human chromosome 12p13
Mouse chromosome 6

8 genes: CD4, A, B, GNB3, C8, ISOT, TPI, C9

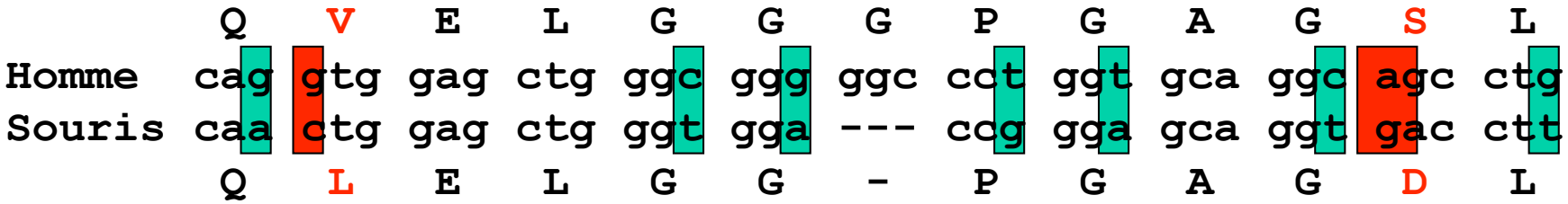


Prediction of protein genes: comparative approach

- Problems: sensitivity depends on the evolutionary distance
  - rapidly evolving genes
  - lineage-specific genes (orphans)

- How to distinguish protein-coding regions from conserved non-coding sequences ?

Distinction des régions conservées codantes *vs*. non-codantes



- Substitutions synonymes (Ks)
- Substitutions non-synonymes (Ka)
- \- Insertion ou délétion

Ratio Ka/Ks << 1 => région codante

Transposable elements: noise for gene prediction

- Transposable elements: ubiquitous in eukaryotic genomes (50% of mammalian genomes)
- TEs contain coding-regions (transposases, reverse-transcriptase) => recognized as “genes” by gene prediction software
- Domesticated (recruited) TEs are very rare
- Mask TEs before running gene prediction software (RepeatMasker)

Prédiction de gènes : démarche

- 1- recherche de séquences répétées (RepeatMasker)
- 2- méthodes intrinséques (consensus de différentes méthodes)
- 3- transcriptome: recherche de similarité ADN/mRNA (blastn/sim4)
- 4- approche comparative: recherche de similarité ADN/protéines (blastx/genewise), ADN/ADN
- COMBINER LES RESULTATS

- 6- prédiction de gènes RNA
  - tRNA: tRNAScanSE
  - rRNA: par similarité
  - snRNA ...

Annotation systématique du génome humain

- ENSEMBL project
  - http://www.ensembl.org/
- Human Genome Project Working Draft at UCSC
  - http://genome.ucsc.edu/

Prédiction de régions régulatrices

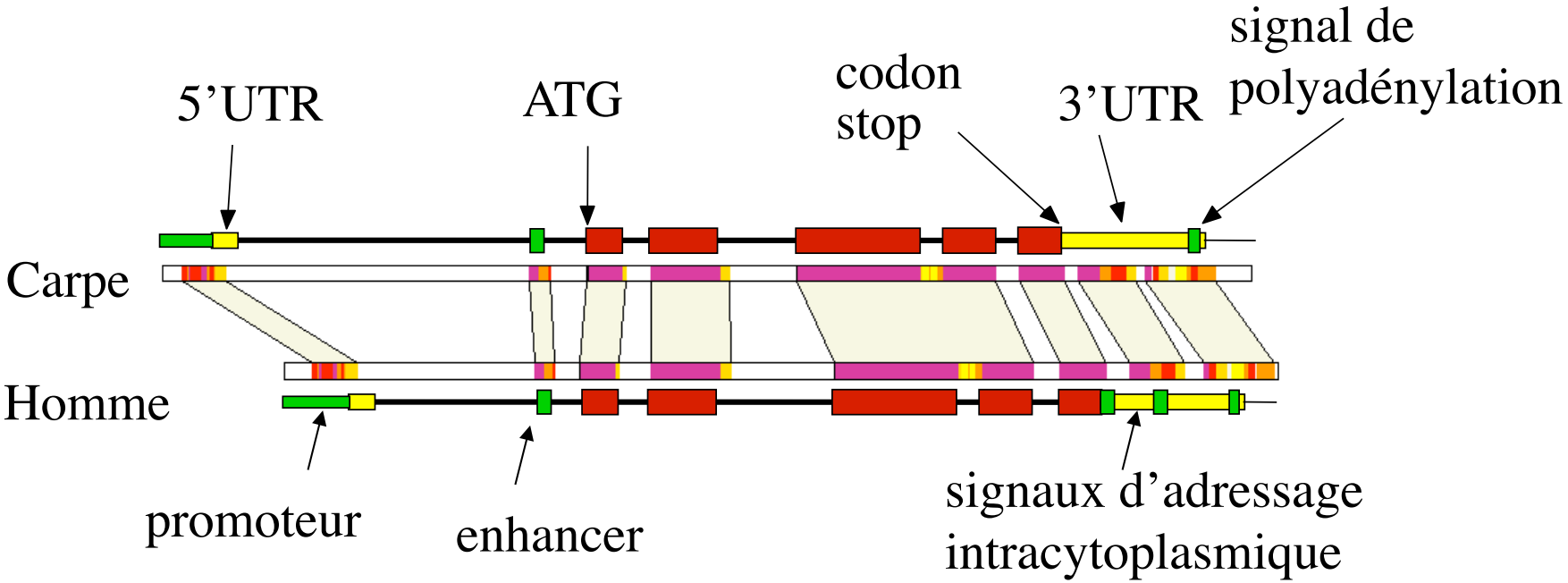
- Méthodes intrinsèques (*ab initio*)
  - Prédiction de promoteurs
  - Îlots CpG
- Approche comparative

Recherche de régions régulatrices par analyse comparative (empreintes phylogénétiques)

- Goodman et al. 1988: régulation de l'expression des gènes du cluster β-globine au cours du développement
  - Alignement de séquences orthologues de 6 mammifères (> 270 Ma d'évolution)
  - 13 empreintes phylogénétiques: ≥ 6 nt, conservation 100%
  - Analyse par retard de bande sur gel:
  - 12/13 (92%) correspondent à des sites de fixation de protéines

- 1996: 35 empreintes phylogénétiques avec protéines fixatrices identifiées

Analyse comparative des gènes de β-actine de l'homme et de la carpe



introns:

régions codantes:

éléments régulateurs:

échelle de similarité

pas de similarité significative

80 - 90% identité

70 - 80% identité

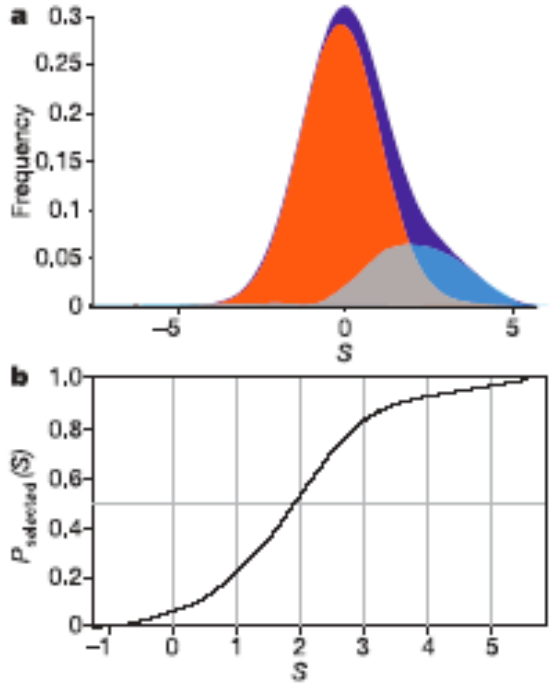
Recherche d'empreintes phylogénétiques à l'échelle du génome

- Empreintes phylogénétique = séquences qui évoluent plus lentement que des séquences non soumises à pression de sélection
- En absence de sélection, le taux d'évolution est égal au taux de mutation (= évolution neutre)
- Mammifères: taux de mutation environ 3.10^{-9} mutation/site/an
- Variation des taux de mutation le long du génomes (et entre espèces)
- Utilisation de marqueurs neutres pour mesurer le taux de mutation local : pseudogènes, éléments transposables défectifs

Comparaison des génomes de l'homme et de la souris

- Alignement des génomes de l'homme et de la souris: 40% du génome humain est alignable avec le génome de la souris
- Utilisation de marqueurs neutres pour mesurer le taux de mutation local : éléments transposables défectifs orthologues homme/souris (i.e. insérés dans le génome avant la divergence primates/ rongeurs)
- Comparaison des taux de substitution dans les séquences non-répétées et dans les marqueurs neutres

Comparaison des génomes de l'homme et de la souris



Distribution des taux de substitution

- Marqueurs neutres
- Séquences non-répétées

Probabilité d'être sous pression de sélection négative

- Plus de 5% du génome des mammifères est sous pression de sélection négative
- NB: seulement 1.2% du génome est codant !! 4 fois plus des regions non-codantes fonctionnelles que de régions codantes !!
- Problème: faible discrimination des séquences sous pression de sélection !!

Comparaison à plus grande distance évolutive

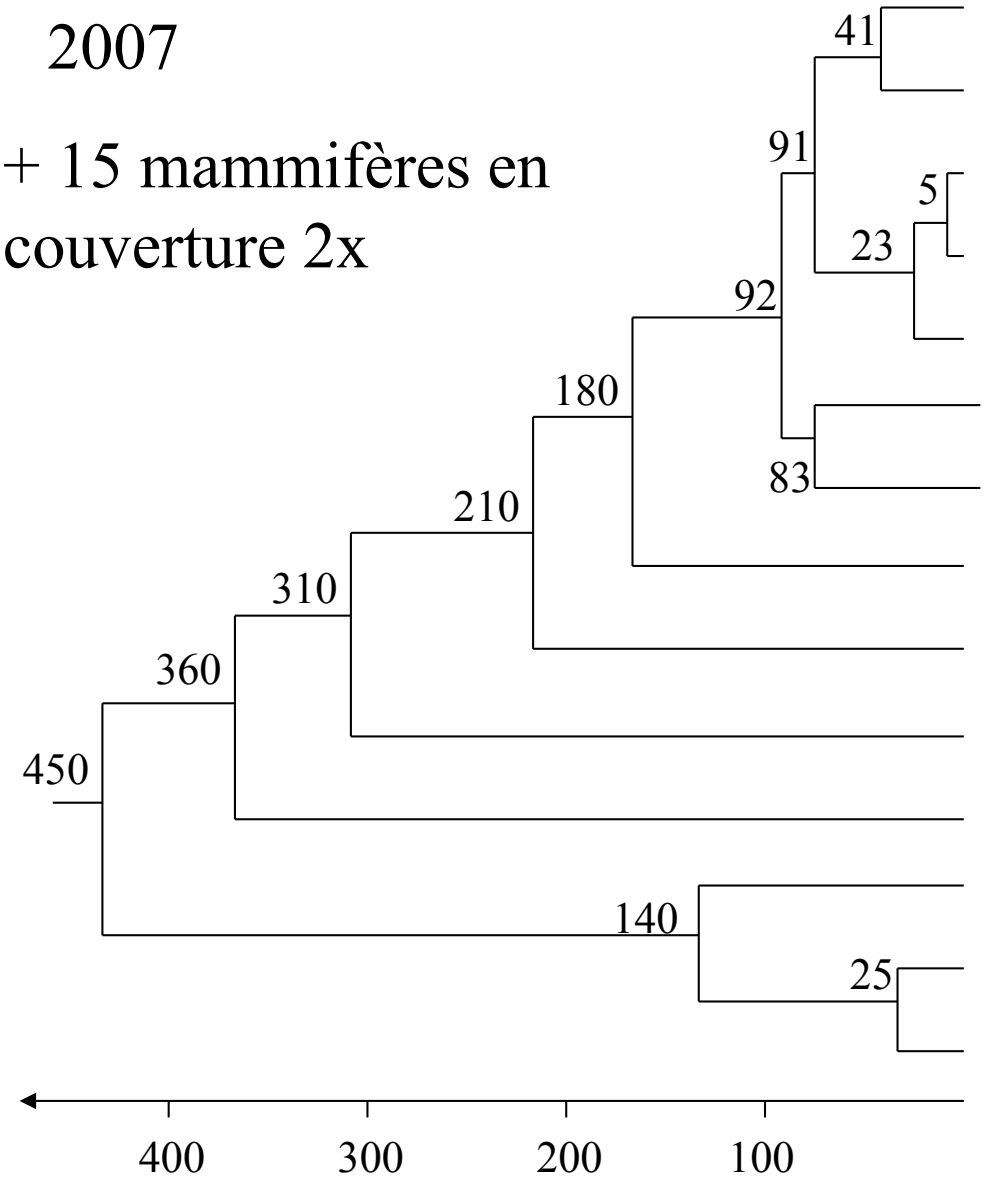
- Mammifères/Oiseaux
  - Empreinte phylogénétique couvrent 2,5% du génome humain
  - 56% des ces empreintes sont non-codantes
- Mammifères/Poissons:
  - Empreinte phylogénétique couvrent 1,8% du génome humain
  - 50% des ces empreintes sont non-codantes
  - Empreintes phylogénétiques non-codantes souvent à proximité des gènes impliqués dans le développement
  - 36 testées expérimentalement: 27 (75%) ont une activité enhancer (Nobrega 2003)

Empreintes phylogénétiques: compromis sensibilité/spécificité

- Grande distance évolutive: ne détecte pas les éléments fonctionnels spécifiques d'une lignée
  - Mammifères/poissons: 1.8% génome humain
  - Mammifères/oiseaux: 2.5% génome humain
  - Primates/rongeurs: 5% génome humain
  - Homme/grands singes: 10% génome humain ??
- Faible distance évolutive: ne discrimine pas les régions conservées contraintes vs. neutres
- Solution: augmenter le nombre d'espèces !

2007

+ 15 mammifères en couverture 2x



Mus musculus (souris) 2,6 Gb

Rattus norvegicus (rat) 2,7 Gb

Homo sapiens (homme) 3,1 Gb

Pan troglodytes (chimpanzé) 3,1 Gb

Macaca mulatta (macaque rhésus) 3,1 Gb

Canis familiaris (chien) ~ 3 Gb

Bos taurus (vache) ~ 3 Gb

Monodelphis domestica (opossum) ~ 3,6 Gb

Ornithorhynchus anatinus (ornithorynque) ~ 3 Gb

Gallus gallus (poulet) 1,2 Gb

Xenopus tropicalis (grenouille tropicale) 1,7 Gb

Danio rerio (poisson zèbre) 1,7 Gb

Fugu rubripes (poisson globe) 0,37 Gb

Tetraodon nigroviridis (poisson globe) 0,34 Gb

(millions d'années)

Séquençage complet, assemblage terminé

Séquençage presque complet; version préliminaire de l'assemblage disponible