Bioinformatique: prédiction de gènes

INSA - Février 2004

Laurent Duret

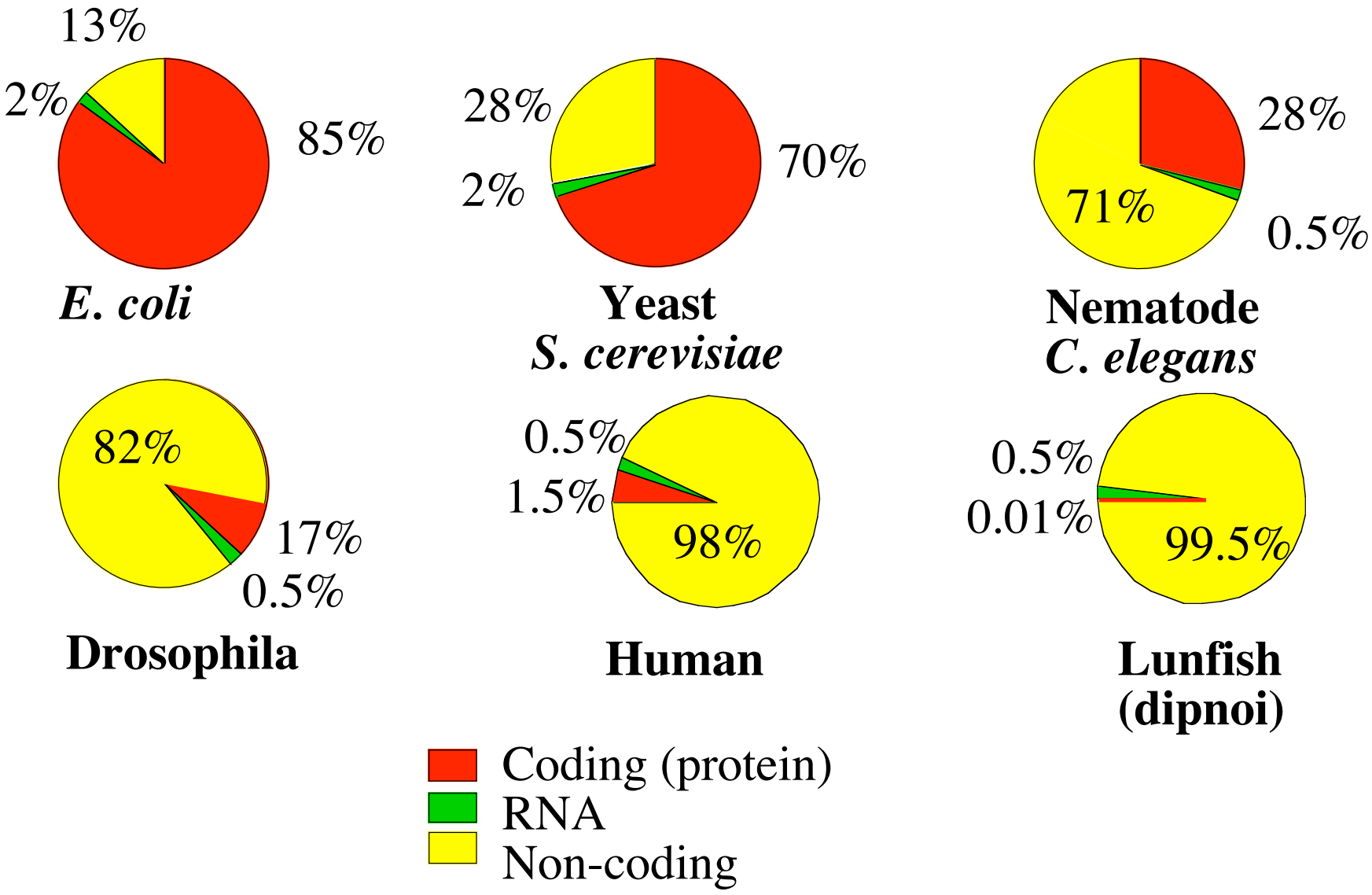
BBE – UMR CNRS n° 5558

Université Claude Bernard - Lyon 1

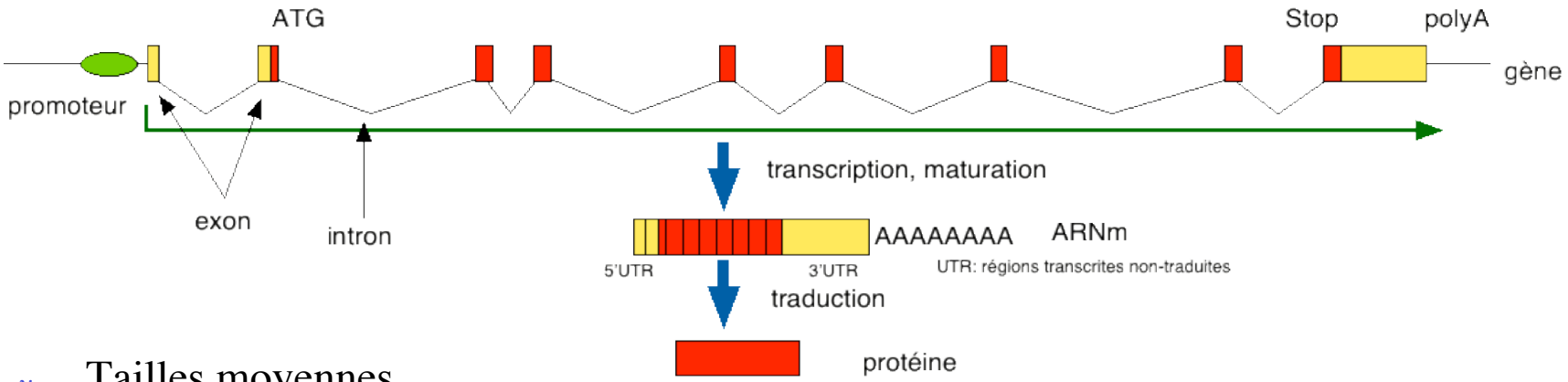
Genome annotation

- Identification of repeats (RepeatMasker, Reputer, …)
- Prediction of protein-coding genes
  - Intrinsic methods (GenScan, Genmark, Glimmer, ...)
  - Genomic/mRNA (EST) comparison (blastn, sim4, …)
  - Genomic/protein comparison (blastx, GeneWise, …)
- Prediction of RNA genes
  - Intrinsic methods (tRNA: tRNAScanSE, snoRNA …)
  - Genomic/RNA (EST) comparison (blastn, sim4, …)
- And more …
  - Replication origins (bacteria) (oriloc)
  - Pseudogenes (by similarity) (blastn, blastx)
  - Regulatory elements (CpG islands, promoters ??)

Proportion of functional elements within genomes



Structure des gènes humains



- Tailles moyennes
  - Gene 27 kb
  - CDS 1100 nt
  - Exon (interne) 145 nt
  - Intron 3400 nt
  - 5'UTR 210 nt
  - 3'UTR 740 nt
- Intron/exon
  - Nombres d'introns: 6 ±3 introns / kb CDS
  - Introns / (introns + CDS): 80%
- Epissage alternatif dans plus de 30% des gènes

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- utilisation de similarité ADN/protéines, ADN/ARNm, ADN/ADN

- méthodes intrinsèques (ab initio): utilisent 1 et 2
- méthodes extrinsèques (approche comparative): utilisent 3, et éventuellement 2

Prédiction de gènes : méthodes intrinsèques

- 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 intrinsèques (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

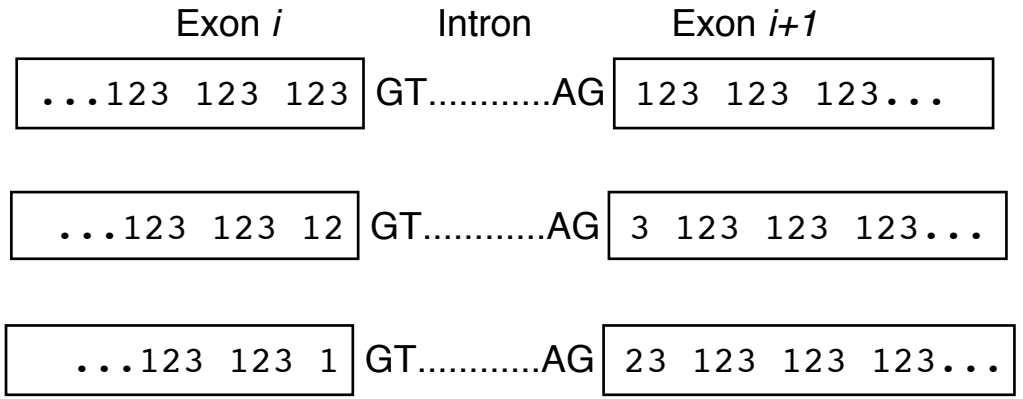
  - Utilisation de consensus (historique): e.g.

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

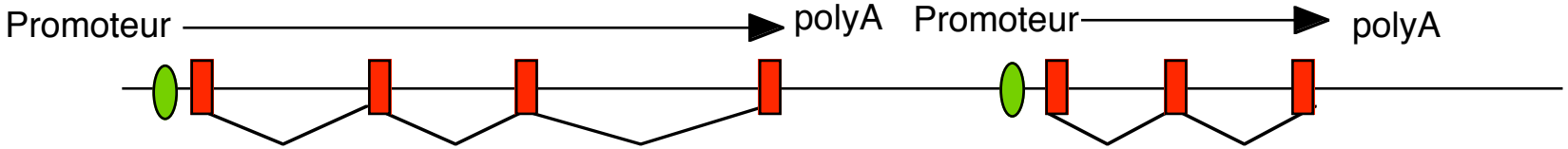
  - Utilisation de matrices de pondération position-dépendantes (profils): Cf suite du cours

Prédiction de gènes : méthodes intrinsèques (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

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

  - e.g. GenScan (mammifères): 78%

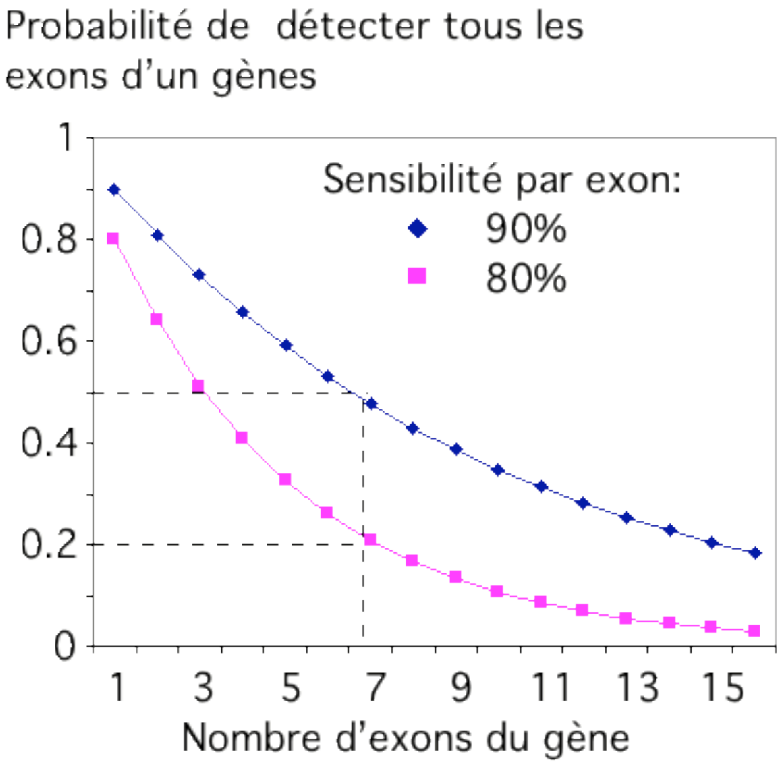
- Spécificité : fraction des vrais exons parmi tous ceux prédits

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

  - e.g. GenScan (mammifères): 81%

Prédiction de gènes protéiques complets

- Construction d ’un modèle de gène à partir de prédictions d ’exons de phases compatibles
- 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 intrinsèques (bilan)

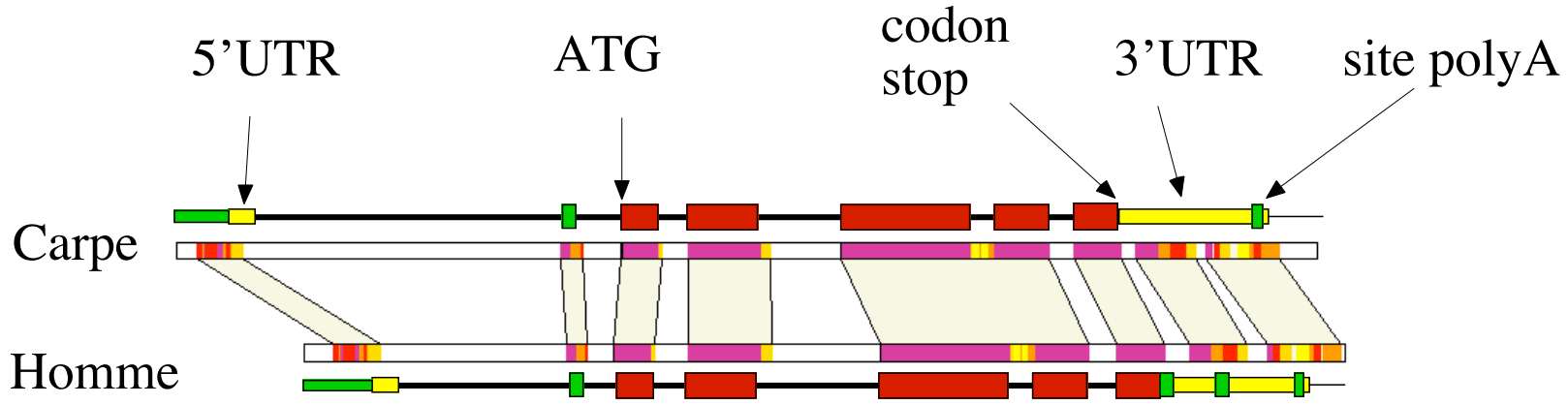
- Procaryotes (pas d'intron):
  - sensibilité et spécificité > 95% (dépend du taux de G+C du génome)
- Eucaryotes: efficacité variable (dépend du taux de G+C du génome et du nombre et de la taille des introns)
  - prédiction d'exons: sensibilité et spécificité 60-80%
  - prédiction de gènes complets: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

Prédiction de gènes : méthodes extrinsèques

- Utilisation des EST
  - comparaison séquence ADN génomique / mRNA : identification des exons (blastn, sim4)
  - informations sur épissage alternatif, expression
  - problème:
    - gènes faiblement exprimés ou à distribution tissulaire restreinte
    - artéfacts dans les EST

- Approche comparative
  - Comparaison d ’une séquence génomique avec des gènes déjà caractérisés dans d ’autres espèces (ADN/protéine) (blastx, genewise)
  - Comparaison de séquences génomiques homologues (ADN/ADN)

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

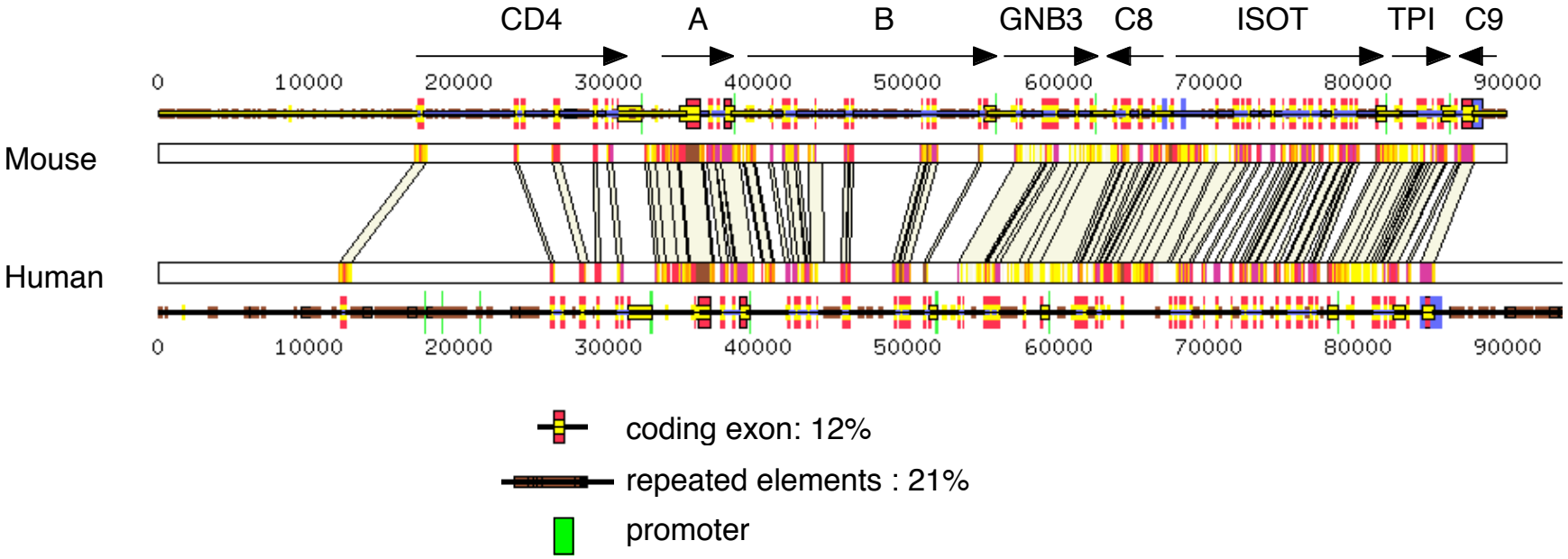
80 - 90% identité

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



Prédiction de gènes : démarche

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

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

Prédiction de régions régulatrices

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

Prédiction de promoteurs eucaryotes

- Combinaison de sites de fixation de facteur de transcription (ordre, orientation, distance)
- Motifs courts, dégénérés
  - Difficile de distinguer les vrais sites des faux positifs:
  - Motif à 4 bases: ≈1/256 pb (1/128 pb sur les deux brins)
- Boîtes TATA, CAAT , GC: absents dans beaucoup de promoteurs

- Banques de données de sites de fixation de facteurs de transcription (TRANSFAC), de promoteurs caractérisés expérimentalement (EPD)

- PromoterScan (Prestridge 1995): Mesure de la densité en sites potentiels de fixation de facteurs de transcription de long de la séquence (pondération en fonction de la fréquence des sites dans ou en dehors des vrais promoteurs)

Prédiction de promoteurs: sensibilité, spécificité

- Sensibilité: fraction des promoteurs qui sont trouvés par le logiciel

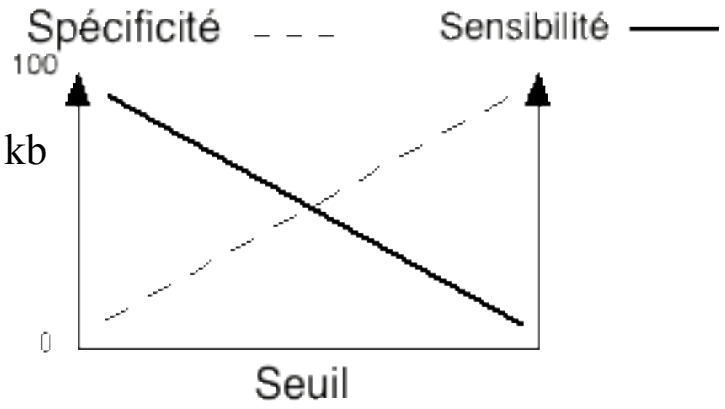
$$sensibilité = \frac{vrais\_positifs}{vrais\_positifs + faux\_négatifs}$$

  - PromoterScan: sensibilité = 70% (promoteurs à boîte TATA)
- Spécificité: fraction des vrais promoteurs parmi ceux qui ont été prédits

$$spécificité = \frac{vrais\_positifs}{vrais\_positifs + faux\_positifs}$$

  - PromoterScan: spécificité = 20%
  - Un faux positif / 10 kb

- Génome humain: ≈30 000 gènes, ≈1 promoteur/100 kb



Prédiction de promoteurs eucaryotes: recherches en cours

- Prise en compte de l'orientation relative et des distances entre sites de fixation de facteurs de transcription
  - COMPEL (Kolchanov 1998): banque de données d'éléments composites
  - FastM : recherche dans une séquence génomique d'une combinaison de deux sites de fixation de facteurs de transcription à une distance définie l'un de l'autre

- Recherche de corrélations entre sites
  - PromoterInspector (Werner 2000)
    - Sensibilité: 40%
    - Spécificité: 45%

http://www.gsf.de/biodv/index.html

- Combinaison recherche *ab initio* / approche comparative: recherche de sites potentiels parmi les régions conservées

Îlots CpG

- Génome de vertébrés :
  - méthylation des C dans les dinucléotides 5 ’-CG-3 ’(CpG)
- Me-C fortement mutable -> T
  - 5 ’-CG- 3 ’  5 ’-TG-3 ’ ou 5 ’-CA-3 ’
  - 3 ’-GC- 5 ’ 3 ’-AC-5 ’ 3 ’-GT-5 ’
- Génome des vertébrés: globalement dépourvu en CpG (excès de TG, CA)

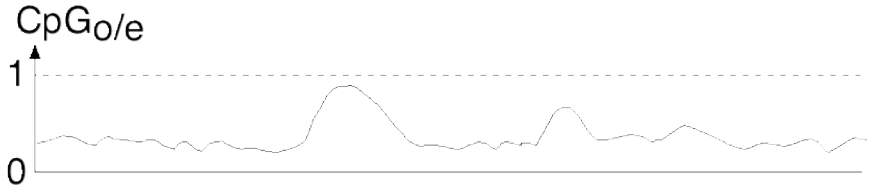
$$CpG_{o/e} = \frac{Nombre\_de\_CpG\_observé}{Nombre\_de\_CpG\_attendu} = 0.25$$

- Certaines régions (200 nt à plusieurs kb) échappent à la méthylation
  - Pas de déplétion en CpG: $CpG_{o/e}$ proche de 1
  - Riche en G+C
  - Îlot CpG:

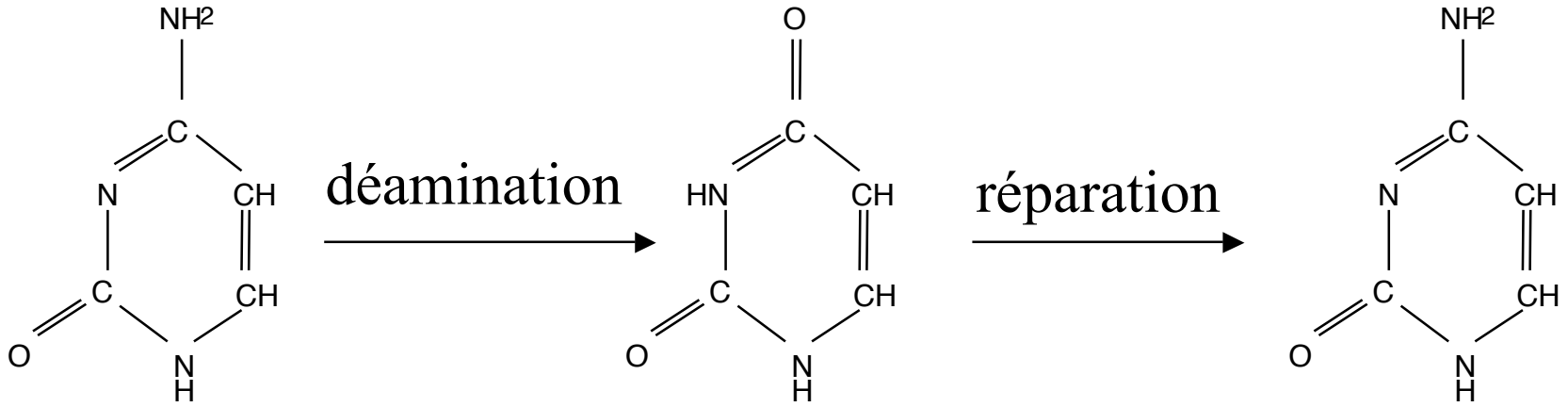
    Longueur > 500 nt

    CpGo/e > 0.6

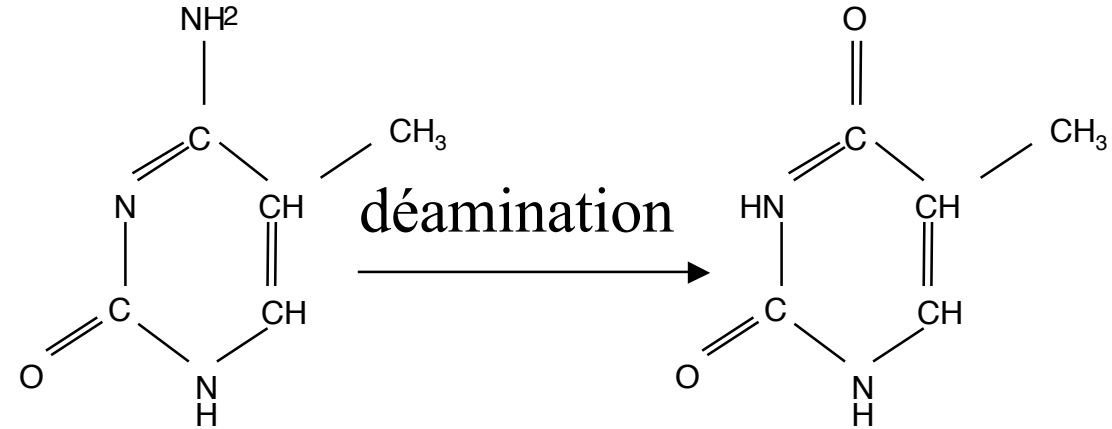
    G+C > 50%



La déamination des cytosines



Cytosine Uracile Cytosine



Cytosine méthylée Thymine

TpG
ou
CpA

Îlots CpG: associés aux régions promotrices ?

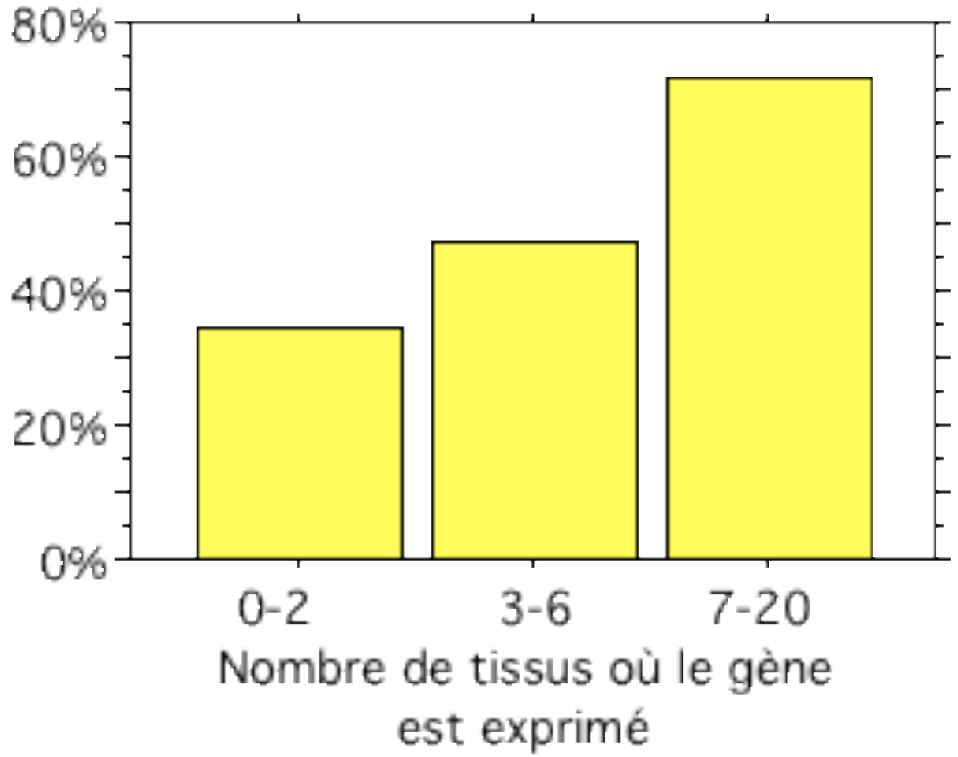
- Bird (1986), Gardiner-Garden (1987) Larsen (1992) ref
  - 40% des gènes tissu-spécifiques possèdent un îlot CpG en 5 ’
  - 100% des gènes ‘ housekeeping ’ possèdent un îlot CpG en 5 ’

- Rechercher des îlots CpG pour prédire des régions promotrices ?
  - Sensibilité: 40-100%
  - Spécificité ?? (Quelle fraction des îlots CpG correspond effectivement à des régions promotrices ?)

- Ponger (2001): comparaison des îlot CpG qui recouvre ou non le site d ’initiation de la transcription

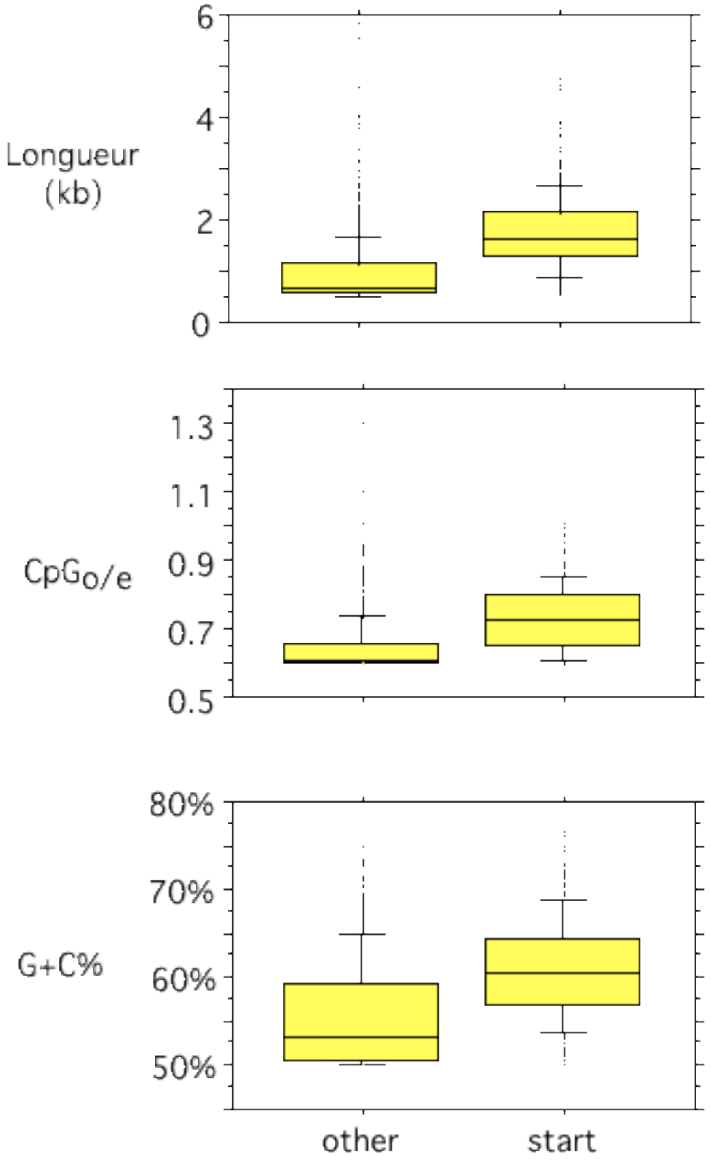
Fréquence des gènes humains avec un îlot CpG recouvrant le site d ’initiation de la transcription

- 800 gènes humains avec promoteur décrit
- Mesure de la distribution tissulaire à l ’aide d ’EST (20 tissus)



Comparaison des îlots CpG recouvrant ou non le site d ’initiation de la transcription

- 272 îlots start CpG recouvrant le site d ’initiation de la transcription (start)
- 1078 îlots CpG en dehors d ’un promoteur connu (other) (en excluant les séquences répétées)



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

- Enhancers de gènes HOX (Fugu/souris) (Aparicio et al. 1995)
- enhancer TCR α (homme/souris) (Luo, 1998)
- promoteur COX5B (11 primates) (Bachman, 1996)
- promoteur uPAR (homme/souris) (Soravia, 1995)

Prediction of gene function

- Analysis of expression pattern (ESTs, …)
- Prediction of the subcellular location of the protein : nucleus, membrane, excreted, etc.
  - SignalPep : http://www.cbs.dtu.dk/services/SignalP/
  - Psort: http://psort.nibb.ac.jp/
  - etc. (see http://www.expasy.org/tools/)
- Search for functional motifs (e.g. DNA binding domains, catalytic sites, …)

  http://hits.isb-sib.ch/cgi-bin/PFSCAN
- Prediction by homology

Function prediction by homology ?

- Similarity between proteins homology
- Homology conserved structure
- Conserved structure conserved function
- Yes, but …
  - Function: fuzzy concept
    - Identical biochemical activity ?
    - Identical expression pattern (tissu-specific isoforms) ?
    - Identical subcellular location (cytoplasm, mitochondria, etc.) ?
  - Homologous proteins with different function
    - e.g. homologous proteins binding a same receptor but opposite activity (activator/repressor)
    - homologous proteins with totally different functions: τ -cristalline / α-énolase
  - Orthology/paralogy
  - Modular evolution

Function prediction by homology ?

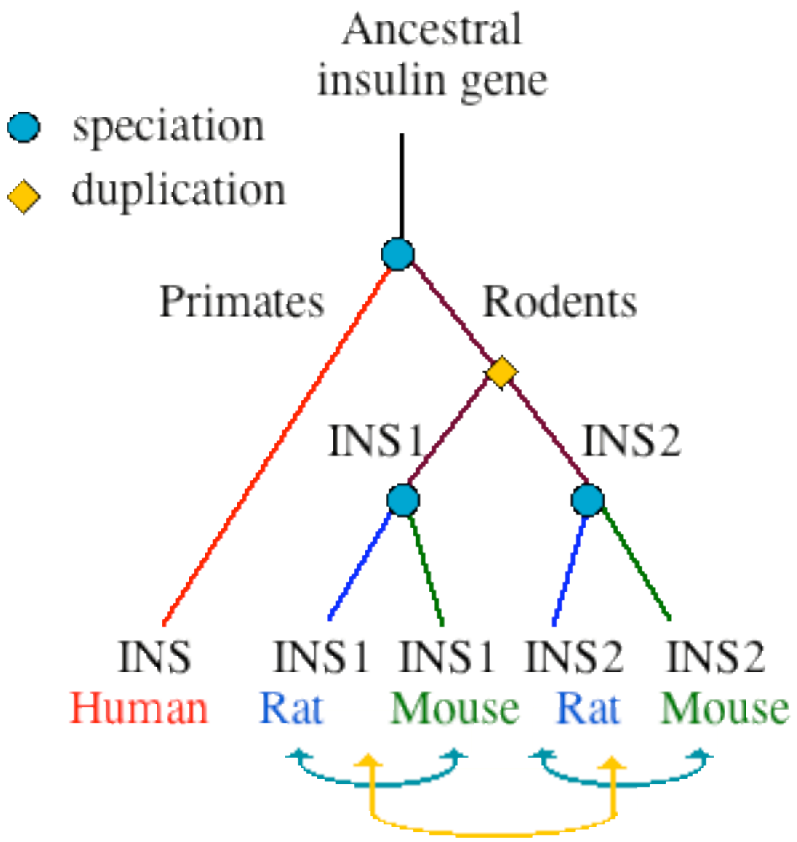
```
MZEORFG:  1   NSPDRACNLAKQAFDEAISELDSLGEESYKDSTLIMQLLDNLTLWTSDNE 59
              N+P++AC LAKQAFD+AI+ELD+L E+SYKDSTLIMQLLDNLTLWTSD E
BOV1433P: 186 NAPEQACLLAKQAFDDAIAELDTLNEDSYKDSTLIMQLLDNLTLWTSEGE 244

 Score = 87.4 bits (213), Expect = 1e-17
 Identities = 41/59 (69%), Positives = 50/59 (84%)

LOCUS       BOV1433P     1696 bp    mRNA            MAM   26-APR-1993
DEFINITION  Bovine brain-specific 14-3-3 protein eta chain mRNA
ACCESSION   J03868

LOCUS       MZEORFG       187 bp    mRNA            PLN   31-MAY-1994
DEFINITION  Zea mays putative brain specific 14-3-3 protein,
            tau protein homolog  mRNA, partial cds.
```

Orthology/paralogy



Homology: two genes are homologous if they share a common ancestor

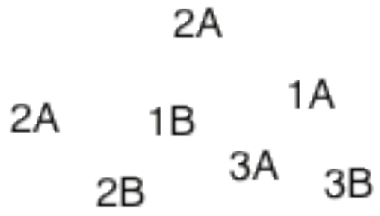
↔ Orthologues: homologous genes that have diverged after a speciation

↔ Paralogues: homologous genes that have diverged after a duplication

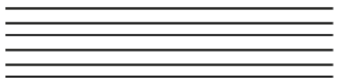
⚠ Orthology ≠ functional equivalence

Phylogenetic approach for function prediction

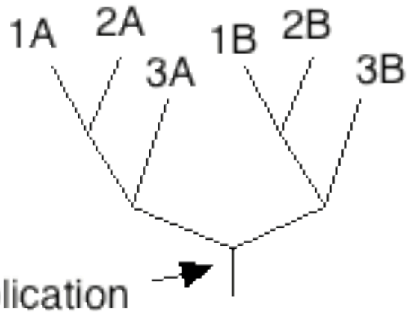
1) Identify homologues



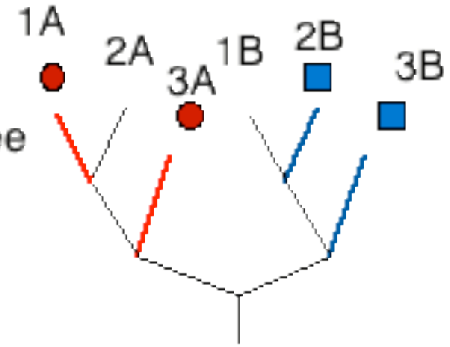
2) Align sequences



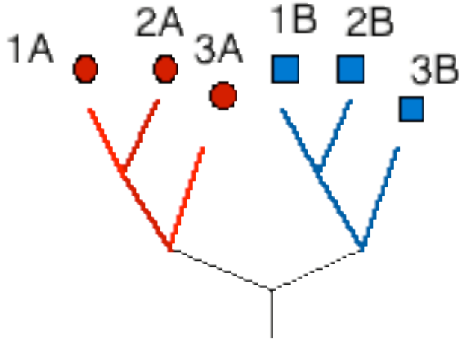
3) Compute the phylogenetic tree



4) Place known functions on the tree



5) Infer the likely function of other genes



Modular evolution