Prédiction de gènes

Hélène Touzet

helene.touzet@univ-lille.fr

CNRS, Bonsai, CRIStAL

Prédiction de gènes, trois grandes approches

1. prédiction par homologie
   - alignement avec des génomes annotés
   - comparaison à des banques de données de protéines
     outil de localisation : BLASTX
     (traduction de l'ADN suivant les 6 cadres de lecture)

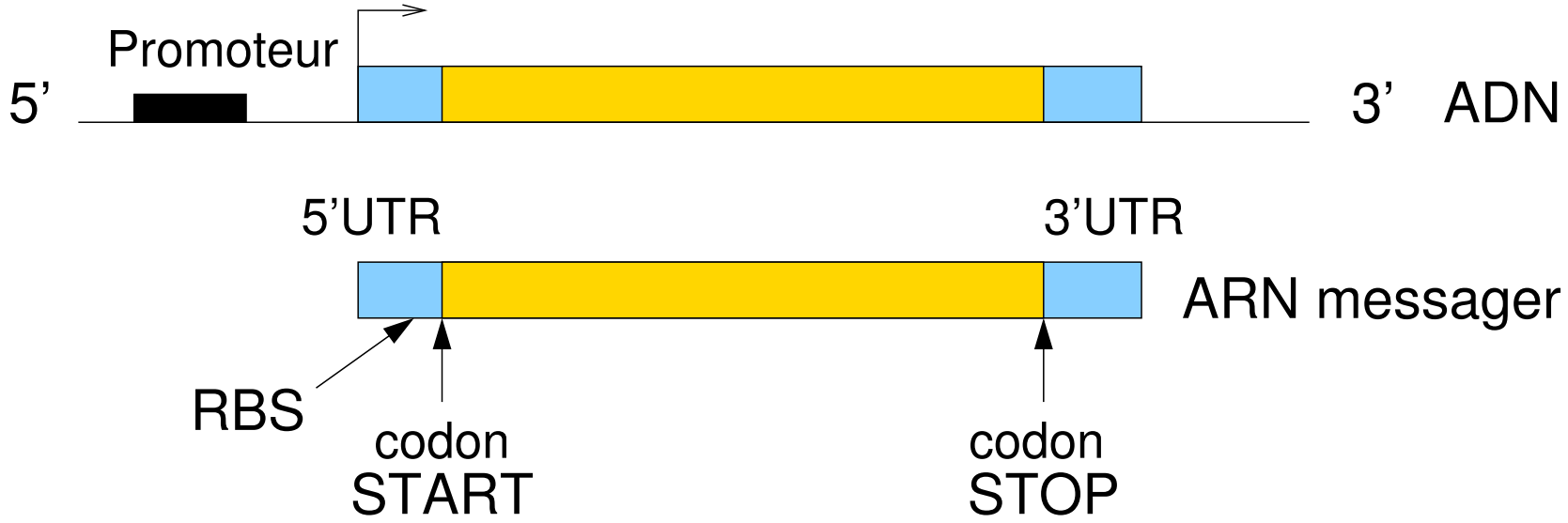
2. prédiction à partir de données de séquençage de transcriptome
   - RNA-seq
   - outils de mapping: STAR,

3. prédiction *de novo*, sans connaissance préalable

L'homme et la souris

- les deux premiers mammifères séquencés
- génome humain: 3 milliards de bases, environ 30 000 gènes
- génome de la souris : 2,5 milliards de bases, environ 30 000 gènes
- parenté génetique : 75 millions d'années
  disparition des derniers dinosaures: environ 60 millions d'années
- 99% de gènes similaires
  les plus grandes différences sont observées pour l'odorat, le système immunitaire et la détoxification
- souris : animal de laboratoire
  cycle de reproduction court (3 semaines de gestation), modèles de mutation

Prédiction de novo : structure d'un gène procaryote

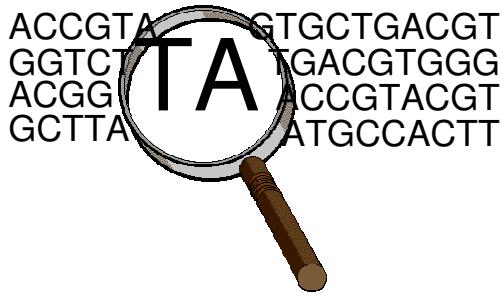


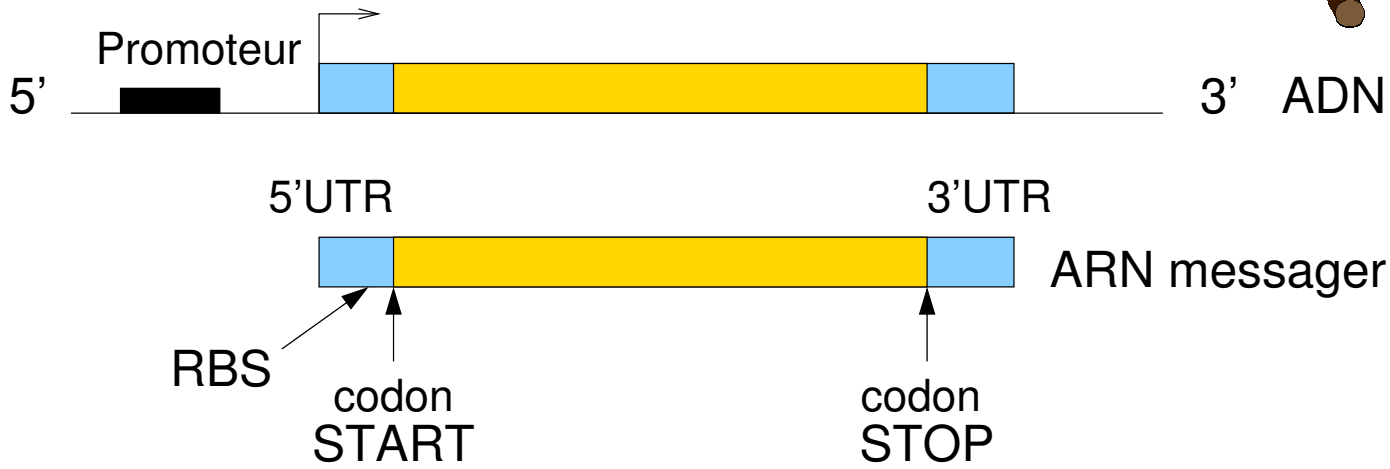
UTR : *UnTranslated Region*
région non traduite lors de la synthèse protéique

RBS : *Ribosome Binding Site*
site de fixation du ribosome à l'ARN messager lors de la traduction

Comment localiser les gènes ?

en regardant le génome de près





- signaux ADN
  promoteur, RBS, codons START et STOP
- composition en codons de la région codante
  table d'usage des codons

Etude du promoteur

- site d'initiation de la transcription
- reconnu par la sous-unité σ de l'ARN polymérase
  - σ^{70} : majorité des gènes (90%)
  - RpoH, SigS, RpoN, SigE, FliA
- séquences consensus pour σ^{70} chez *E.coli*

```
          -35        16-19 bp       -10      +1
------TTGACA------------------TATAAT---CAT
```

- distance variable entre les deux boîtes
- le signal peut être très dégradé

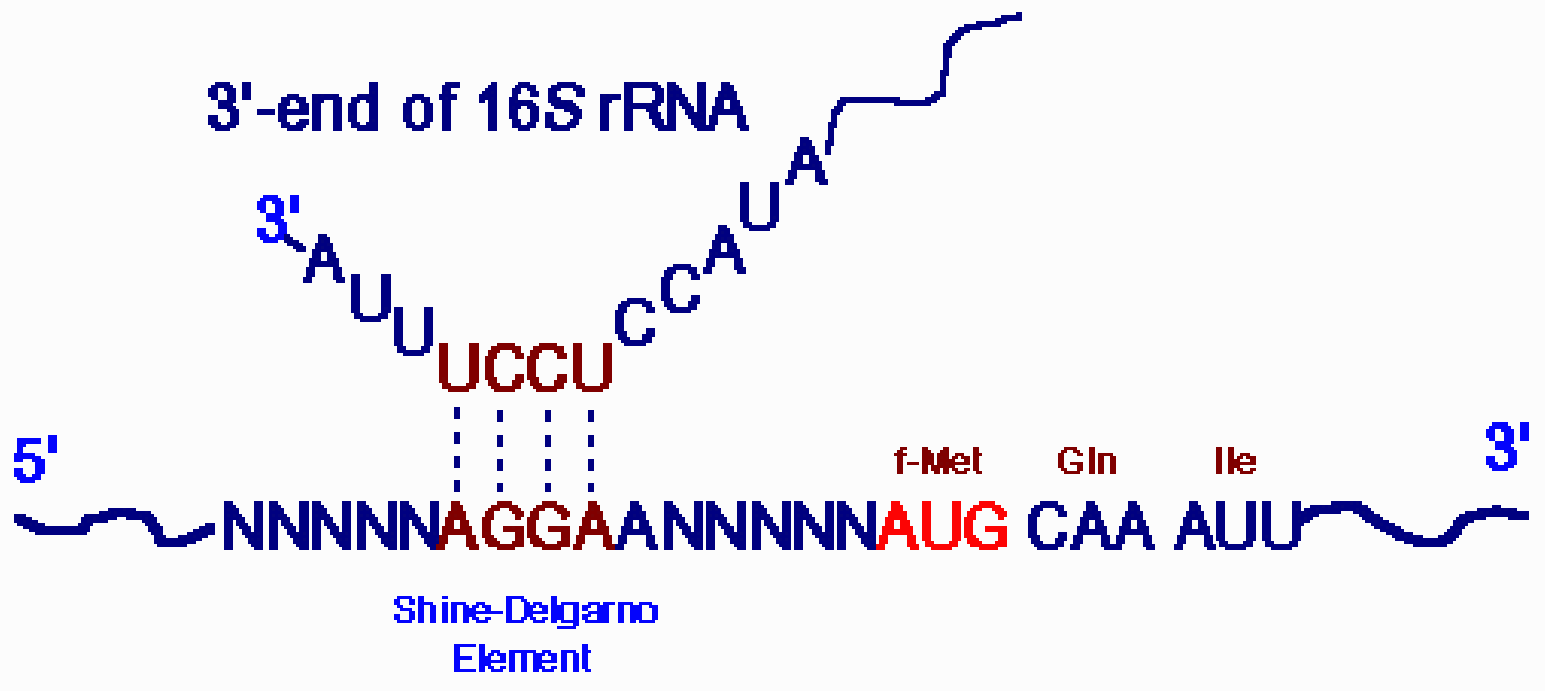
	position					
	1	2	3	4	5	6
A	0.04	0.88	0.26	0.59	0.49	0.03
C	0.09	0.03	0.11	0.13	0.22	0.05
G	0.07	0.01	0.12	0.16	0.12	0.02
T	0.80	0.08	0.51	0.13	0.18	0.89

matrice de fréquence pour `TATAAT` (pour 263 promoteurs connus)

- existence d'opérons
- pas de prise en compte de la structure de l'ADN : accessibilité du site
- pas suffisant

Etude du RBS – Ribosome Binding Site

- Séquence de Shine-Dalgarno: Site d'initiation de la traduction



- signal bref et dégradé
- distance entre le RBS et le codon START variable (≈ -10)

À la recherche des codons START et STOP

- ORF (*Open Reading Frame*) : région génomique
  - commençant par un codon START
    `ATG`, `CTG` ou `TTG`
  - terminant par un codon STOP dans la même phase
    `TAA`, `TGA` ou `TAG`
  - ne contenant pas de codon STOP dans la même phase entre les deux
- longueur moyenne d'un ORF ?

Approche statistique : biais de composition

- code génétique: 20 acides aminés, $4 \times 4 \times 4 = 64$ codons
- redondance du code génétique
  plusieurs choix de codons sont possibles pour coder un acide aminé
- table d'usage des codons
  ce choix n'est pas équiprobable, et varie suivant les espèces

AAA	3.5	1.3	CAA	1.3	1.4	GAA	4.3	1.6	TAA	*	*
AAG	1.1	1.6	CAG	3.0	1.7	GAG	1.8	1.8	TAG	*	*
AAC	2.4	1.4	CAC	1.1	1.5	GAC	2.2	1.7	TAC	1.4	1.4
AAT	1.4	1.3	CAT	1.2	1.4	GAT	3.2	1.5	TAT	1.5	1.3
AGA	0.1	1.6	CGA	0.3	1.7	GGA	0.6	1.8	TGA	*	*
AGG	0.1	1.8	CGG	0.4	2.0	GGG	1.0	2.2	TGG	1.4	1.8
AGC	1.6	1.7	CGC	2.4	1.8	GGC	3.2	2.0	TGC	0.7	1.6
AGT	0.7	1.5	CGT	2.5	1.6	GGT	2.8	1.8	TGT	0.5	1.5
ACA	0.5	1.4	CCA	0.8	1.5	GCA	2.0	1.7	TCA	0.6	1.4
ACG	1.4	1.7	CCG	2.6	1.8	GCG	3.6	2.0	TCG	0.8	1.6
ACC	2.5	1.5	CCC	0.4	1.6	GCC	2.5	1.8	TCC	0.9	1.5
ACT	0.9	1.4	CCT	0.6	1.5	GCT	1.6	1.6	TCT	0.9	1.4
ATA	0.3	1.3	CTA	0.3	1.4	GTA	1.1	1.5	TTA	1.1	1.3
ATG	2.5	1.5	CTG	5.7	1.6	GTG	2.7	1.8	TTG	1.2	1.5
ATC	2.7	1.4	CTC	1.0	1.5	GTC	1.5	1.6	TTC	1.8	1.4
ATT	2.8	1.3	CTT	0.9	1.4	GTT	1.9	1.5	TTT	1.9	1.2

table d'usage des codons pour la bactérie *E. coli*

1ère colonne: codon

2ème colonne: fréquence observée (gènes connus)

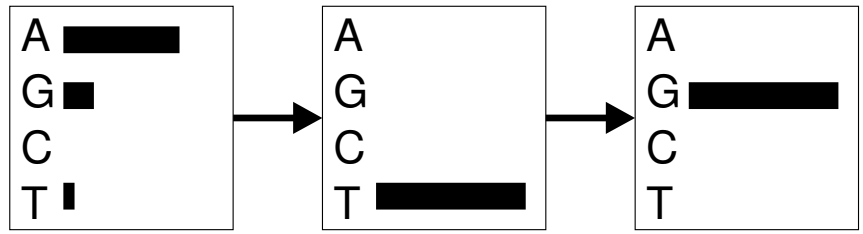
3ème colonne: fréquence théorique (modèle de base)

- régions codantes : modèle de Markov basé sur la table d'usage des codons (voir transparent suivant)
- régions intergéniques : en première approche, modèle de Markov avec indépendance des bases

$$\begin{array}{llll} \text{Proba(A)} = 0{,}237 & & \text{Proba(C)} = 0{,}253 \\ \text{Proba(G)} = 0{,}279 & & \text{Proba(T)} = 0{,}231 \end{array}$$

- codon start

$$\text{Proba(ATG)} = 0.905$$
$$\text{Proba(GTG)} = 0.090$$
$$\text{Proba(TTG)} = 0.005$$



- codon stop : idem

Données pour *E.coli*

GeneMark.hmm – 1998

- analyse des deux brins simultanément

  sens direct et inverse

- gènes typiques et atypiques

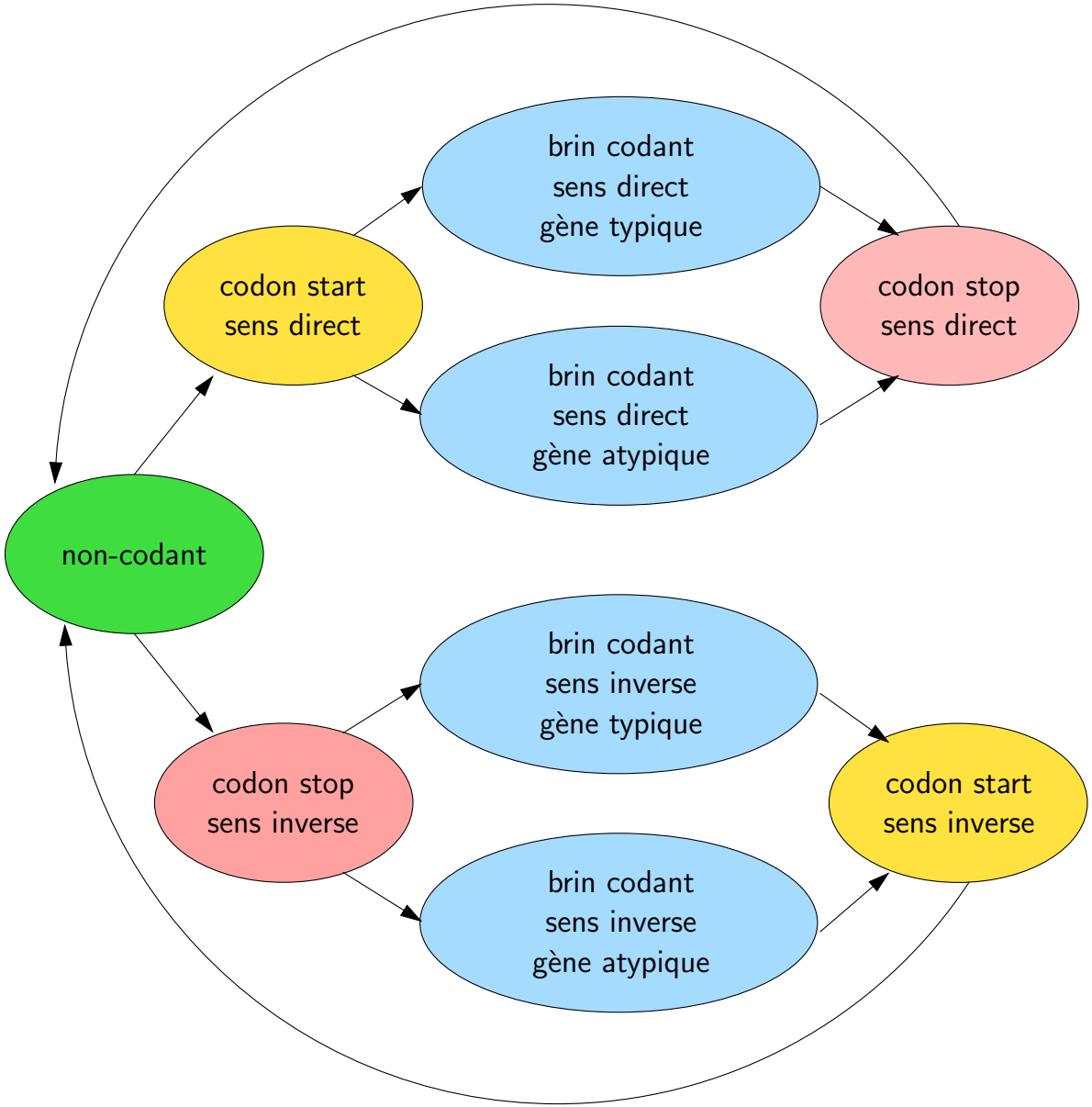
  typique : 90% des gènes connus

  atypique : transfert horizontal entre espèces

- post-traitement pour limiter les problèmes des gènes chevauchants

  à partir du codon START prédit par l'algorithme de Viterbi,

  recherche du premier codon START préservant l'ORF et précédé par un un RBS



HMM pour GeneMark.hmm

Méthodologie

- apprentissage à partir de génes connus pour la prédiction de nouveaux gènes
  - signaux (transcription, traduction)
  - composition en codons
- constitution de l'ensemble d'apprentissage
  - homologie
  - séquençage de transcriptome
- risque de biais : on prédit ce qu'on connait