

# Prédiction de gènes

Hélène Touzet

`helene.touzet@univ-lille.fr`

CNRS, Bonsai, CRIStAL

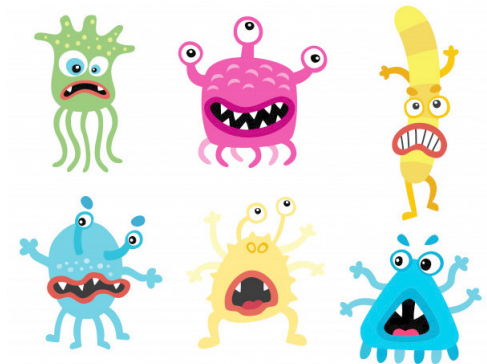
# Décoder le génome: prédiction des gènes

- première étape pour interpréter un génome nouvellement séquencé
  - distinction entre régions codantes et non codantes
  - identification des protéines
- réalisée par des programmes informatiques combinant différents types d'informations
  - possibilité de **faux négatifs**  
gènes qui échappent à la détection
  - possibilité de **faux positifs**  
certains gènes prédits ne correspondent pas à de vrais gènes
  - les bornes du gène sont parfois erronées

# Prédiction de gènes, trois grandes approches

- prédiction par homologie
- prédiction *ab initio*, sans connaissance préalable
- prédiction à partir de données de séquençage de transcriptome EST, RNA-Seq

# Génomes procaryotes



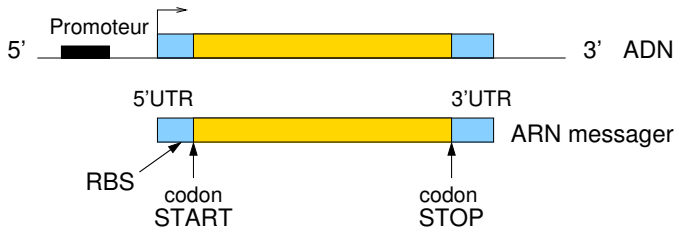
# Génomes procaryotes

- plus de 80% du génome est codant
  - séquences intergéniques courtes
  - en moyenne : un gène pour 1 000 nucléotides
- structure des gènes simple
  - régions transcrites mais non traduites (3' et 5' UTR) courtes
  - pas d'intron (sauf exception)

# Prédiction par homologie

- comparaison avec un génome proche
  - annotation d'une nouvelle souche pour une bactérie connue
  - conservation du contenu en gènes, avec possiblement des mutations ponctuelles au sein des gènes et des remaniements chromosomiques
  - BlastN, mauve, MUMmer
- recherche des régions pouvant coder pour des protéines orthologues
  - comparaison inter-espèces
  - traduction du génome suivant les 6 cadres de lectures puis comparaison à une base de données de protéines
  - BlastX

## Prédiction *ab initio* : structure d'un gène procaryote



UTR : *UnTranslated Region*

région non traduite lors de la synthèse protéique

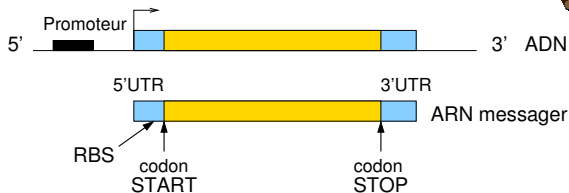
RBS : *Ribosome Binding Site*

site de fixation du ribosome à l'ARN messager lors de la traduction

# Comment localiser les gènes ?

en regardant le génome de près

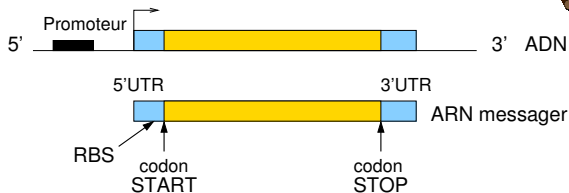
ACCGTAACTGCTGACGT  
GGTCTGACGTGGG  
ACGGCCGTACGT  
GCTTATGCCACTT



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

# Comment localiser les gènes ?

en regardant le génome de près



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

# 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, par hasard : environ 20 acides aminés (loi géométrique)  
Les ORF de grande longueur sont rarement dus au hasard.
- <https://www.ncbi.nlm.nih.gov/orffinder>

# Entre les codons START et STOP

- existence d'un biais statistique au sein des CDS (Coding Sequences)
- 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
  - biais d'usage du code  
ce choix n'est pas équiprobable au sein d'une espèce  
il varie suivant les espèces

|                           |   | Deuxième lettre    |     |     |     |                      |      |     |      |   |                            |   |
|---------------------------|---|--------------------|-----|-----|-----|----------------------|------|-----|------|---|----------------------------|---|
|                           |   | U                  |     | C   |     | A                    |      | G   |      |   |                            |   |
| Première lettre (côté 5') | U | UUU                | Phe | UCU | Ser | UAU                  | Tyr  | UGU | Cys  | U | Troisième lettre (côté 3') |   |
|                           |   | UUC                | Phe | UCC | Ser | UAC                  | Tyr  | UGC | Cys  |   |                            | C |
|                           |   | UUA                | Leu | UCA | Ser | UAA                  | Stop | UGA | Stop |   |                            | A |
|                           |   | UUG                | Leu | UCG | Ser | UAG                  | Stop | UGG | Trp  |   |                            | G |
|                           | C | CUU                | Leu | CCU | Pro | CAU                  | His  | CGU | Arg  | U |                            |   |
|                           |   | CUC                | Leu | CCC | Pro | CAC                  | His  | CGC | Arg  | C |                            |   |
|                           |   | CUA                | Leu | CCA | Pro | CAA                  | Gln  | CGA | Arg  | A |                            |   |
|                           |   | CUG                | Leu | CCG | Pro | CAG                  | Gln  | CGG | Arg  | G |                            |   |
|                           | A | AUU                | Ile | ACU | Thr | AAU                  | Asn  | AGU | Ser  | U |                            |   |
|                           |   | AUC                | Ile | ACC | Thr | AAC                  | Asn  | AGC | Ser  | C |                            |   |
|                           |   | AUA                | Ile | ACA | Thr | AAA                  | Lys  | AGA | Arg  | A |                            |   |
|                           |   | AUG                | Met | ACG | Thr | AAG                  | Lys  | AGG | Arg  | G |                            |   |
|                           | G | GUU                | Val | GCU | Ala | GAU                  | Asp  | GGU | Gly  | U |                            |   |
|                           |   | GUC                | Val | GCC | Ala | GAC                  | Asp  | GGC | Gly  | C |                            |   |
|                           |   | GUA                | Val | GCA | Ala | GAA                  | Glu  | GGA | Gly  | A |                            |   |
|                           |   | GUG                | Val | GCG | Ala | GAG                  | Glu  | GGG | Gly  | G |                            |   |
|                           |   | codon d'initiation |     |     |     | codon de terminaison |      |     |      |   |                            |   |

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

|     |     |     |     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 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 |

1<sup>ère</sup> colonne: codon

2<sup>ème</sup> colonne: fréquence observée (gènes connus)

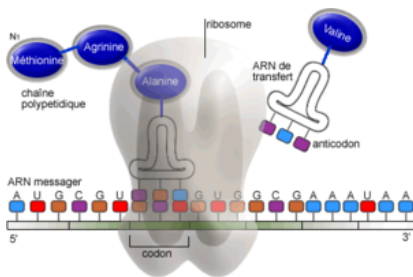
3<sup>ème</sup> colonne: fréquence théorique (modèle de base)

## Exemples d'usage des codons Sérine dans différents organismes

| Codon | <i>E.coli</i> | <i>D.melanogaster</i> | <i>H.sapiens</i> | <i>S.cerevisiae</i> |
|-------|---------------|-----------------------|------------------|---------------------|
| AGT   | 3             | 1                     | 10               | 5                   |
| AGC   | 20            | 23                    | 34               | 4                   |
| TCG   | 4             | 17                    | 9                | 1                   |
| TCA   | 2             | 2                     | 5                | 6                   |
| TCT   | 34            | 9                     | 13               | 52                  |
| TCC   | 37            | 48                    | 28               | 33                  |

(fréquences arrondies – source: D. Gautheret)

- hypothèse neutraliste : résulte d'un biais mutationnel. La probabilité de mutation d'une base nucléique vers une autre varie en fonction des bases considérées, et cela se reflète dans les codons
- hypothèse sélectionniste : optimisation de l'efficacité de la traduction. Les codons préférés sont ceux pour lesquels les ARN de transfert sont les plus fréquents dans le génome



## Exemple de logiciel: GeneMark.hmm

- suite d'outils développée depuis 1998 par le Georgia Institute of Technology
  - version procaryote
  - version eucaryote
- utilisée pour annoter le génome d'*Haemophilus influenzae*
- hmm: modélisation statistique à base de Modèles de Markov cachés (Hidden Markov models)
- <https://exon.gatech.edu/GeneMark>

# Prédiction chez les procaryotes : quelques pièges

- plusieurs codons START possibles
  - présence de RBS
  - comparaison de séquence
- événements rares ou difficiles à identifier
  - transferts horizontaux (sauf si homologie)
  - décalage de phase de lecture
  - gènes chevauchants ou en anti-sens
  - (gènes non codants)

# Génomes eucaryotes

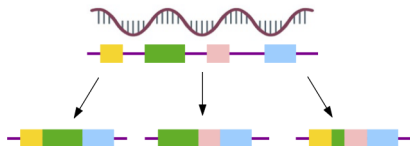


# Complexité des génomes eucaryotes

- structure complexe des génomes
  - faible pourcentage de séquences codant pour des protéines (environ 2% du génome humain)
  - présence de répétitions
  - présence de pseudo-gènes
- structure complexe des gènes
  - longues régions 3' et 5' non traduites (exons non codants)
  - alternance d'introns et d'exons

# Les introns et les exons

- taille des introns non multiple de 3  
un intron peut couper un codon en deux → changement de phase d'un exon à l'autre
- existence d'exons courts ( $\sim 10\text{nt}$ )
- existence d'introns très longs (plus longs que les exons)
- signal d'épissage faible (sites donneurs et accepteurs)
- épissage alternatif (plus de 70% des gènes humain)



# Analyse en deux temps

- localisation des ARN messagers
- étude des bornes des gènes: site d'initiation de la transcription, codon START, structure exonique

# Exercice

- annotation du génome miniature de *Staphylococcus aureus*
- recherche d'ORF avec ORFfinder  
<https://www.ncbi.nlm.nih.gov/orffinder>
- Comparaison des ORF trouvées avec des banques de protéines avec BlastX  
<https://blast.ncbi.nlm.nih.gov>
- Prédiction de gènes avec Genemark.hmm  
<https://exon.gatech.edu/GeneMark>

