

# Prédiction de gènes

Hélène Touzet

Licence – master  
[www.lifl.fr/~touzet/M1/](http://www.lifl.fr/~touzet/M1/)

# Deux approches

## 1. Prédiction par similitude

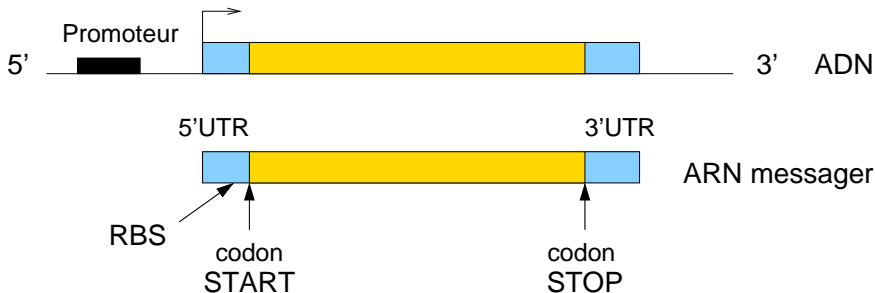
- ▶ comparaison à des banques d'**EST**  
EST : expressed sequence tags  
outil de localisation : BLAST
- ▶ comparaison à des banques de données de protéines  
outil de localisation : BLASTX  
(traduction de l'ADN suivant les 6 cadres de lecture)
- ▶ comparaison à un autre organisme  
homme/souris, bactéries, levures, etc.  
outil de localisation: BLAST

## 2. Prédiction *de novo*, sans connaissance préalable

# L'homme et la souris

- ▶ **L**es deux premiers mammifères séquencés
- ▶ **G**énome humain: 3 milliards de bases, environ 30 000 gènes
- ▶ **G**énome de la souris : 2,5 milliards de bases, environ 30 000 gènes
- ▶ **P**arenté génétique : 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
- ▶ **S**ouris : 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*

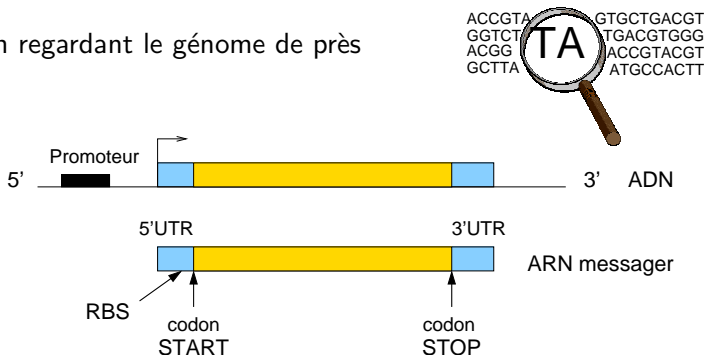
région non traduite lors de la synthèse protéique

**RBS** : *Ribosome Binding Site*

site de fixation du ribosome à l'ARN messenger 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



- ▶ Difficulté algorithmique (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

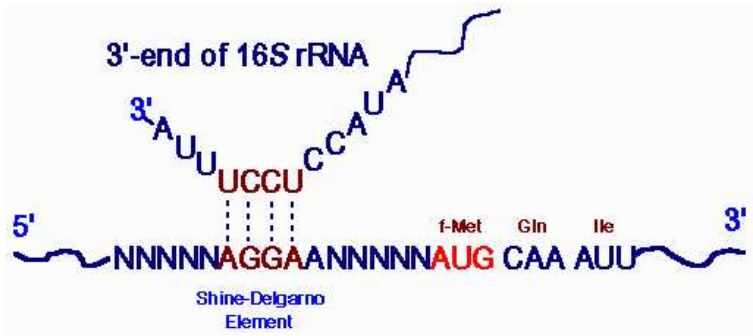
Matrice de scores pour TATAAT (pour 263 promoteurs connus)

? plus le promoteur est éloigné du consensus, moins le gène est exprimé ?

- ▶ Existence d'opérons
- ▶ Pas de prise en compte de la structure de l'ADN : accessibilité du site
- ▶ **Pas convaincant**

# 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 ( $\approx -10$ )

# À la recherche des codons START et STOP

- ▶ **ORF** (*Open Reading Frame*) : fragment d'ADN
  - ▶ 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 entre les deux, toujours dans la même phase
  
- ▶ Longueur moyenne d'un ORF ?
  
- ▶ Longueur moyenne d'une protéine ?

# 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*

1<sup>ère</sup> colonne: codon

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

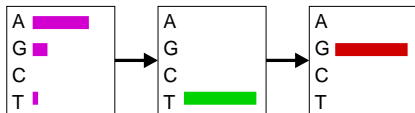
3<sup>ème</sup> colonne: fréquence théorique (modèle de base)

- ▶ **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

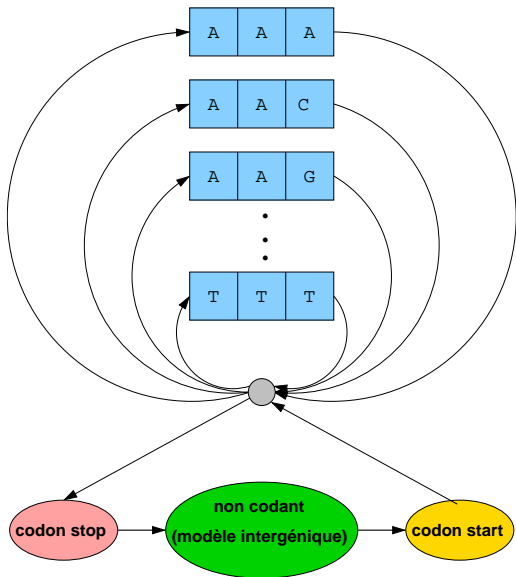
$$\begin{array}{ll} \text{Proba(A)} & = 0,237 & \text{Proba(C)} & = 0,253 \\ \text{Proba(G)} & = 0,279 & \text{Proba(T)} & = 0,231 \end{array}$$

- ▶ **Codon start**

$$\begin{array}{ll} \text{Proba(ATG)} & = 0.905 \\ \text{Proba(GTG)} & = 0.090 \\ \text{Proba(TTG)} & = 0.005 \end{array}$$



- ▶ **Codon stop** : idem



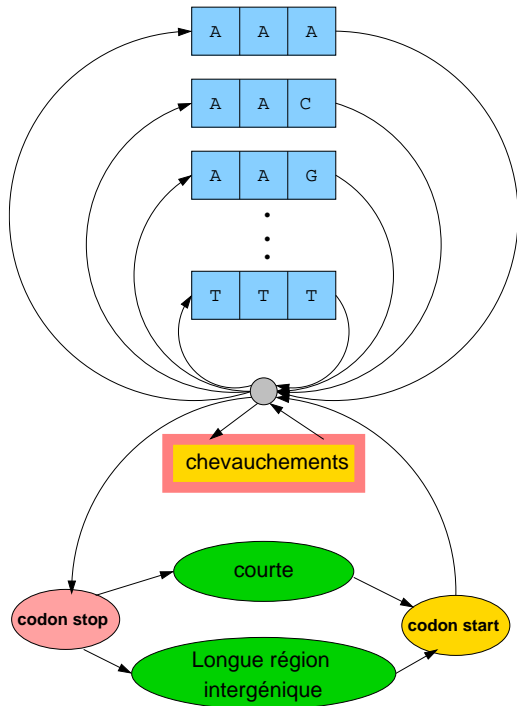
Modèle de Markov basé sur la table d'usage des codons

## Ecoparse – 1994

- ▶ **M**odèle plus évolué
- ▶ **C**ourtes régions intergéniques ( $< 10$ )
- ▶ **L**ongues régions intergéniques ( $> 10$ )  
Apparition de motifs connus
  - ▶ après le codon STOP (*Repetitive Extragenic Palindromic sequences*)
  - ▶ avant le codon START (RBS)
- ▶ **T**raitement des chevauchements de taille 1 et 4

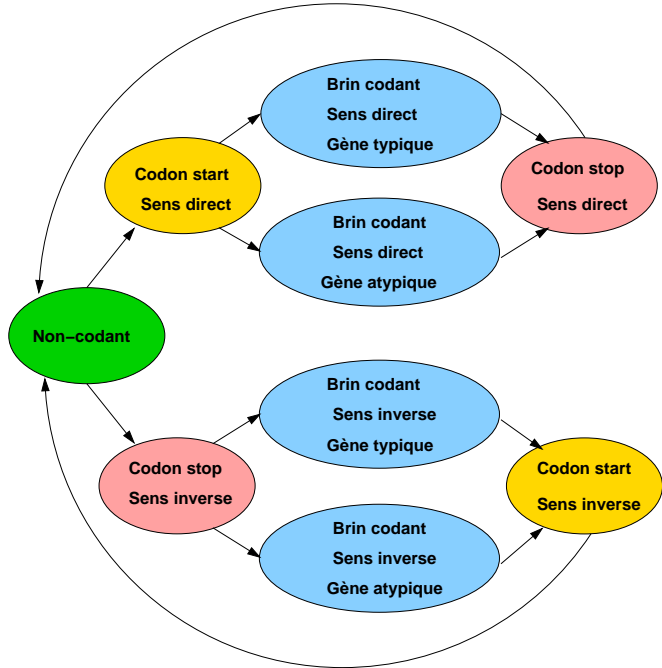
NN[AG]TGANN

N : n'importe quel nucléotide



# 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.



GeneMark.hmm

# Méthodologie

- ▶ **A**pprentissage à partir des gènes détectés pour la prédiction de nouveaux gènes
  - ▶ Signaux (transcription, traduction)
  - ▶ Composition en codons
- ▶ **C**onstitution de l'ensemble d'apprentissage
  - ▶ Homologie
  - ▶ Expériences
- ▶ ? Problème de biais ? (On prédit ce qu'on connaît)