

Prédiction des gènes chez les procaryotes

Hélène TOUZET

touzet@lifl.fr

Deux approches

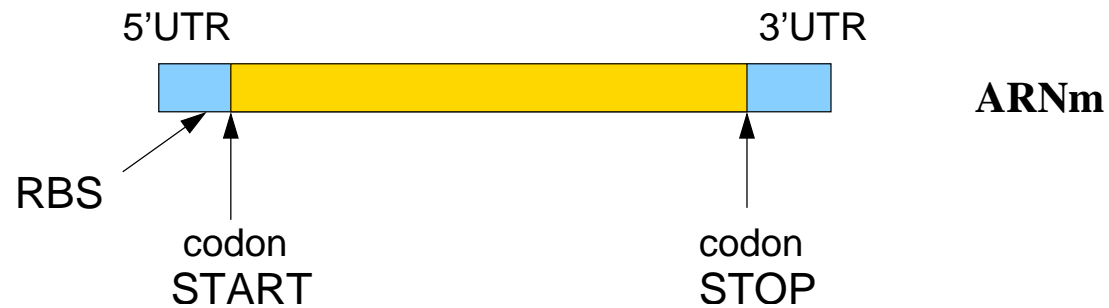
1. prédiction par similarité

- EST
- BLASTN
- BLASTX

traduction suivant les 6 cadres de lecture

2. prédiction *de novo*, ou *ab initio*

Structure d'un gène procaryote



UTR : *UnTranslated Region*

RBS : *Ribosome Binding Site*

Comment localiser les gènes ?

En regardant le génome

- Signaux ADN

Promoteur, RBS, codons START et STOP

- Composition en codons de la région codante

Table d'usage des codons

Les promoteurs

Site d'initiation de la transcription

- ▷ Reconnus 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

- ▷ Difficulté algorithmique (Distance variable entre les deux boites)
- ▷ Le signal peut être très dégradé

	position					
	1	2	3	4	5	6
<i>A</i>	0.04	0.88	0.26	0.59	0.49	0.03
<i>C</i>	0.09	0.03	0.11	0.13	0.22	0.05
<i>G</i>	0.07	0.01	0.12	0.16	0.12	0.02
<i>T</i>	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 ?

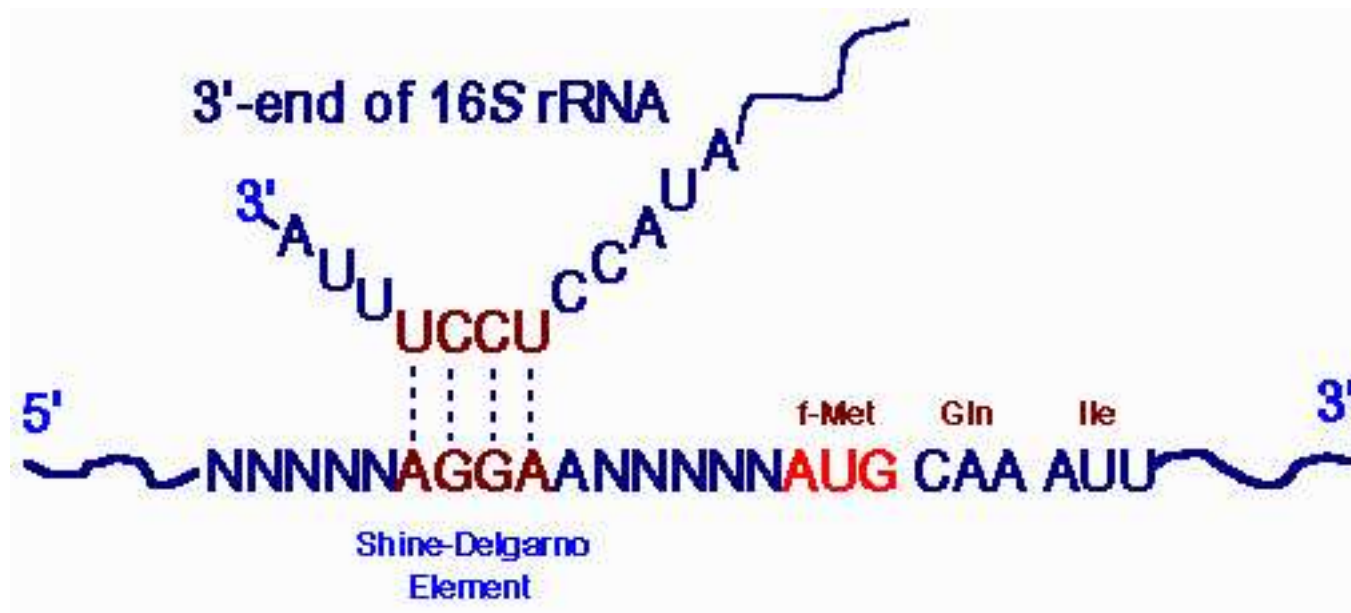
- ▷ Existence d'opérons
- ▷ Pas de prise en compte de la structure de l'ADN : accessibilité du site

Peu convaincant

RBS = Ribosome Binding Site

Séquence de Shine-Dalgarno

Site d'initiation de la traduction



- ▷ Signal bref et dégradé (Trop de faux positifs)
- ▷ Distance entre le RBS et le codon START variable (≈ -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**, toujours dans la même phase

longueur moyenne d'un ORF ?

longueur moyenne d'une protéine ?

Approche statistique : biais de composition

▷ *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
AAG	1.1	1.6
AAC	2.4	1.4
AAT	1.4	1.3
AGA	0.1	1.6
AGG	0.1	1.8
AGC	1.6	1.7
AGT	0.7	1.5
ACA	0.5	1.4
ACG	1.4	1.7
ACC	2.5	1.5
ACT	0.9	1.4
ATA	0.3	1.3
ATG	2.5	1.5
ATC	2.7	1.4
ATT	2.8	1.3

CAA	1.3	1.4
CAG	3.0	1.7
CAC	1.1	1.5
CAT	1.2	1.4
CGA	0.3	1.7
CGG	0.4	2.0
CGC	2.4	1.8
CGT	2.5	1.6
CCA	0.8	1.5
CCG	2.6	1.8
CCC	0.4	1.6
CCT	0.6	1.5
CTA	0.3	1.4
CTG	5.7	1.6
CTC	1.0	1.5
CTT	0.9	1.4

GAA	4.3	1.6
GAG	1.8	1.8
GAC	2.2	1.7
GAT	3.2	1.5
GGA	0.6	1.8
GGG	1.0	2.2
GGC	3.2	2.0
GGT	2.8	1.8
GCA	2.0	1.7
GCG	3.6	2.0
GCC	2.5	1.8
GCT	1.6	1.6
GTA	1.1	1.5
GTG	2.7	1.8
GTC	1.5	1.6
GTT	1.9	1.5

TAA	*	*
TAG	*	*
TAC	1.4	1.4
TAT	1.5	1.3
TGA	*	*
TGG	1.4	1.8
TGC	0.7	1.6
TGT	0.5	1.5
TCA	0.6	1.4
TCG	0.8	1.6
TCC	0.9	1.5
TCT	0.9	1.4
TTA	1.1	1.3
TTG	1.2	1.5
TTC	1.8	1.4
TTT	1.9	1.2

Exemple: Table d'usage des codons (*E. coli*)

1^{ère} colonne: fréquence observée (gènes connus)

2^{ème} colonne: fréquence théorique (modèle de base)

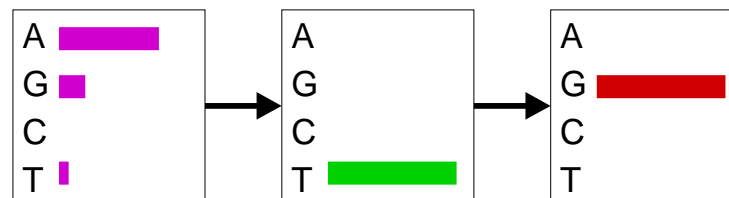
Régions codantes : Modèle de Markov basé sur la table d'usage des codons

Régions intergéniques : *En première approche*: Modèle de Markov avec indépendance des bases

$$\begin{aligned}P(A) &= 0,237 \\P(C) &= 0,253 \\P(G) &= 0,279 \\P(T) &= 0,231\end{aligned}$$

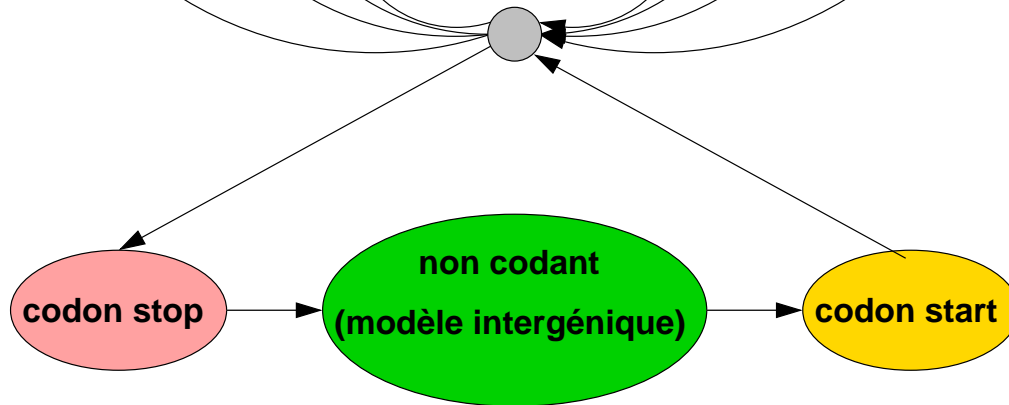
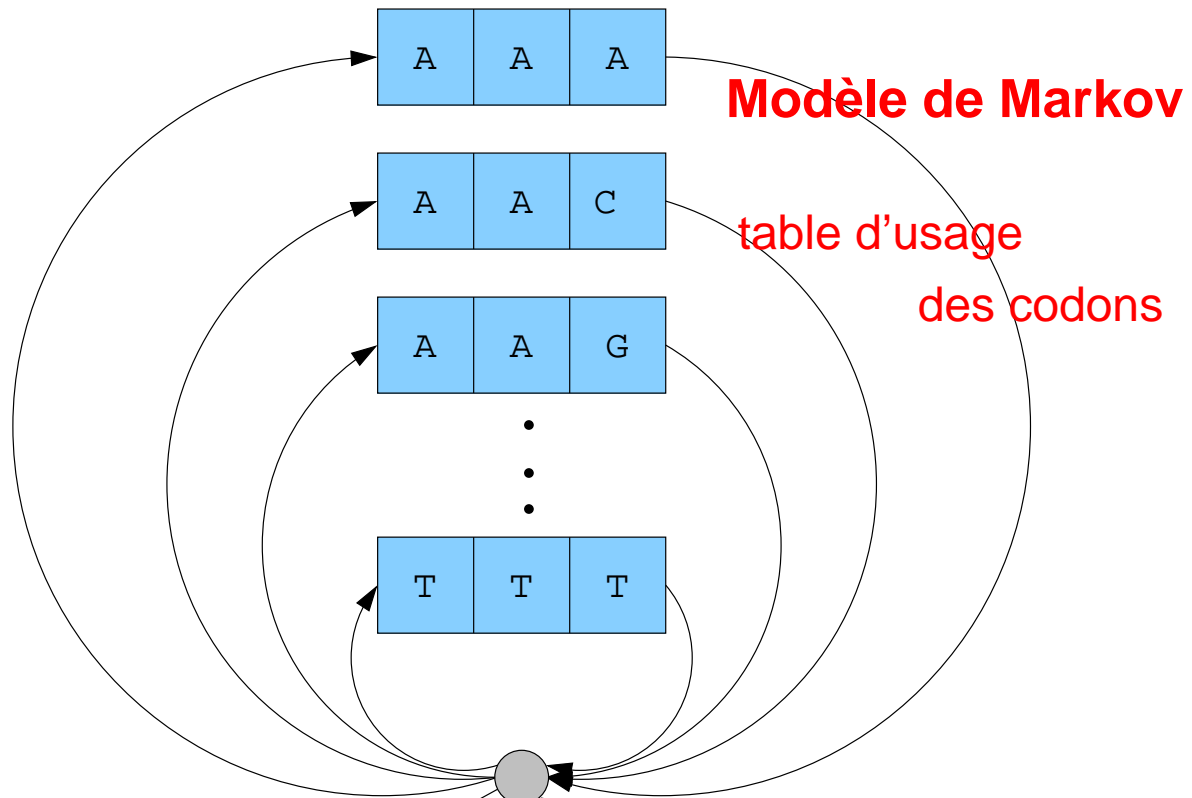
Codon start

$$\begin{aligned}P(ATG) &= 0.905 \\P(GTG) &= 0.090 \\P(TTG) &= 0.005\end{aligned}$$



Codon stop : idem

Observations pour *E.coli*



Modèle minimal

Ecoparse

(1994)

▷ Courtes régions intergéniques (< 10)

▷ Longues régions intergéniques (> 10)

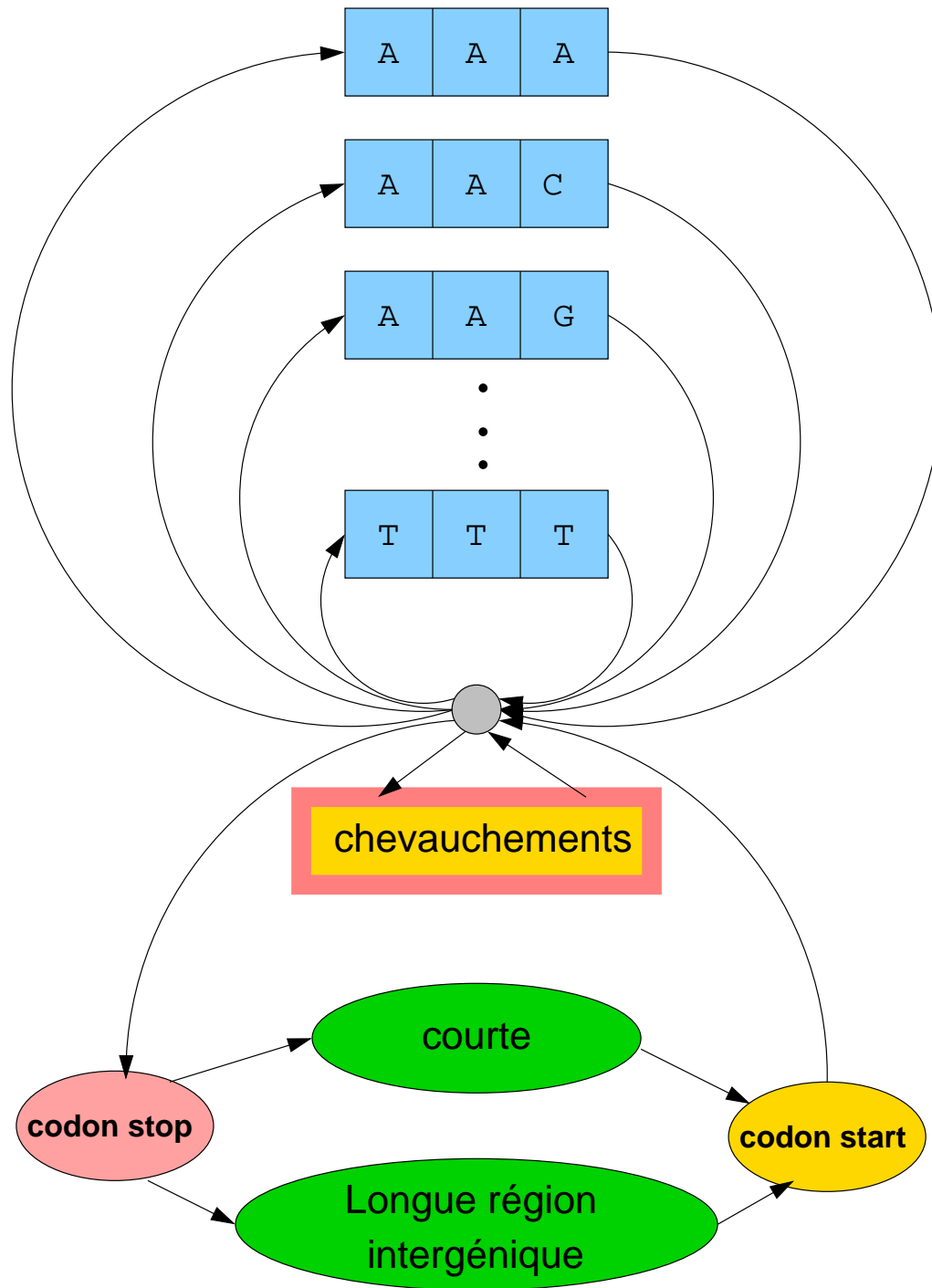
Apparition de motifs connus

- après le codon STOP (*Repetitive Extragenic Palindromic sequences*)
- avant le codon START (RBS)

▷ Traitement 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

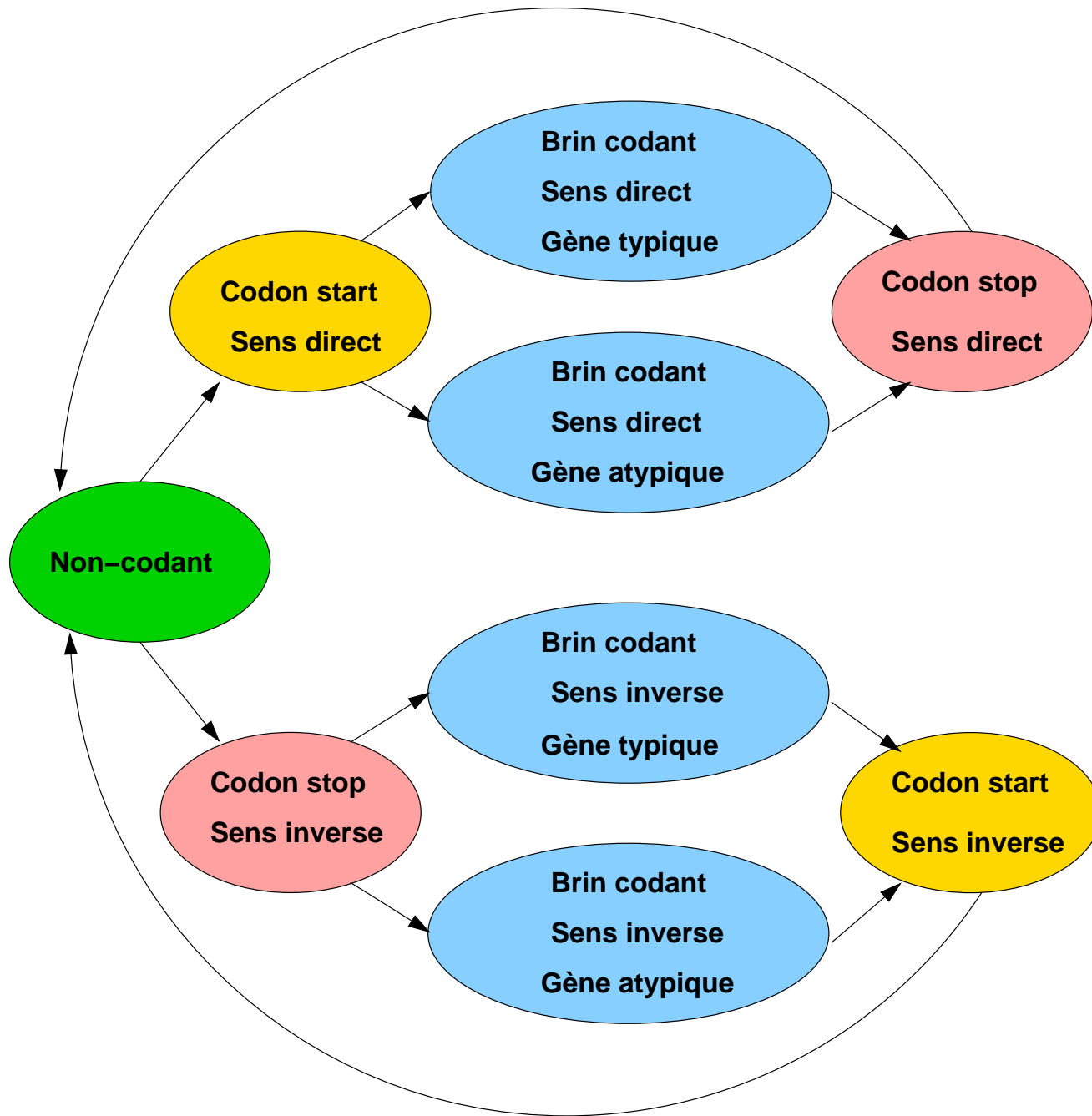
- ▷ Gènes **typiques** et **atypiques**

typique : 90% des gènes connus

atypique : transfert horizontal

- ▷ 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

- ▷ **Apprentissage** à partir des gènes détectés pour la prédiction de nouveaux gènes
 - Signaux (transcription, traduction)
 - Composition en codons

- ▷ Comment constituer l'ensemble d'apprentissage?
 - Homologie
 - Expériences

- ▷ ? Problème de biais ? (On prédit ce qu'on connaît)