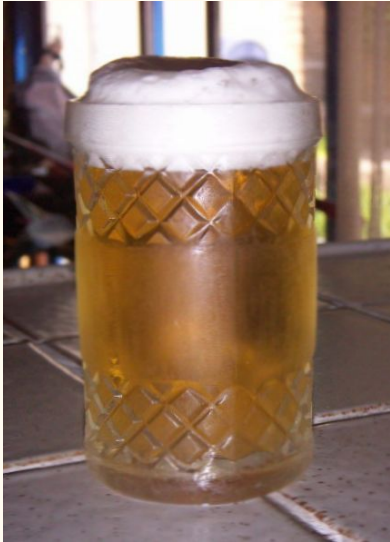


# Analyse de séquences génomiques

—

## Projet 3

*S. cerevisiae*



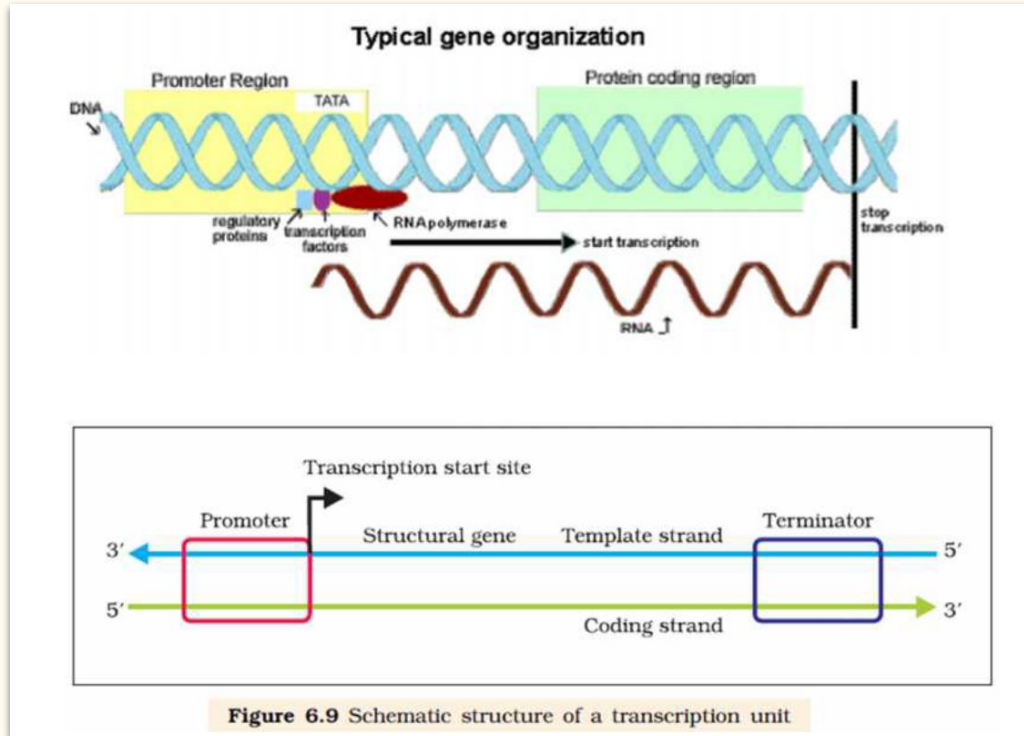
# *Saccharomyces cerevisiae*

- champignon unicellulaire (= levure) de 5–10  $\mu\text{m}$  de diamètre
  - **organisme modèle** en biologie cellulaire et en génétique
    - premier eucaryote à être séquencé, en 1997
    - Le séquençage de l'être humain c'est 10 ans plus tard.
  - On estime que l'être humain partage 23% de ses gènes avec cette levure.
- 
- En tant qu'Eucaryote, *S. cerevisiae* partage la structure cellulaire interne complexe des plantes et des animaux sans le pourcentage élevé d'ADN non codant qui peut perturber la recherche chez les eucaryotes supérieurs.

# Rappels de biologie

- Un **génome** peut être vu comme une chaîne de caractères écrite dans un alphabet à 4 lettres (A, C, G ou T).
- On va s'intéresser à 2 types d'éléments le long des génomes, les **gènes** et les **séquences promoteurs**.

# Rappels de biologie




# Partie préliminaire

Données et lecture des fichiers

---

# Fichier fasta

>NC\_001133.9 *Saccharomyces cerevisiae* S288c chromosome I, complete sequence 

ccaacaccacacccacacacccacacaccacaccacacaccacacccacacacacacatCCTAACACTACCCTAAC  
ACAGCCCTAATCTAACCCCTGGCCAACCTGTCTCTCAACTTACCCTCCATTACCCTGCCTCCACTCGTTACCCTGTCCCAT  
TCAACCATAACCACTCCGAACCACCATCCATCCCTCTACTTACTACCACCTACCCACCGTTACCCTCCAATTACCCATATC  
CAACCCACTGCCACTTACCCTACCATTACCCTACCATCCACCATGACCTACTCACCATACTGTTCTTTCTACCCACCATAT  
TGAAACGCTAACAAATGATCGTAAATAACACACACGTGCTTACCCTACCACCTTTATACCACCACCACATGCCATACTCAC  
CCTCACTTGTATACTGATTTTACGTACGCACACGGATGCTACAGTATATACCATCTCAAACCTACCCTACTCTCAGATTTC  
CACTTCACTCCATGGCCCATCTCTCACTGAATCAGTACCAAATGCACTCACATCATTATGCACGGCACTTGCCTCAGCGG  
TCTATACCCCTGTGCCATTTACCCATAACGCCCATCATTATCCACATTTTGATATCTATATCTCATTGGCGGTcccaaat  
attgtataaCTGCCCTTAATACATACGTTATACCACCTTTTGCACCATACTTACCACCTCCATTTATATACACTTATGTC

...

Nom de la séquence

[illegible]

1 ligne = 60, 70 ou 80 caractères

# Manipulation des séquences

- Recoder la séquence avec :  $\{ 'A' : 0, 'C' : 1, 'G' : 2, 'T' : 3 \}$
  - Compter le **nombre d'occurrences** des 4 lettres dans le texte recodé
  - Calculer les **fréquences** d'apparition de chaque lettres
- 
- Calculer la **log-probabilité** d'une séquence sachant les fréquences d'occurrence des lettres

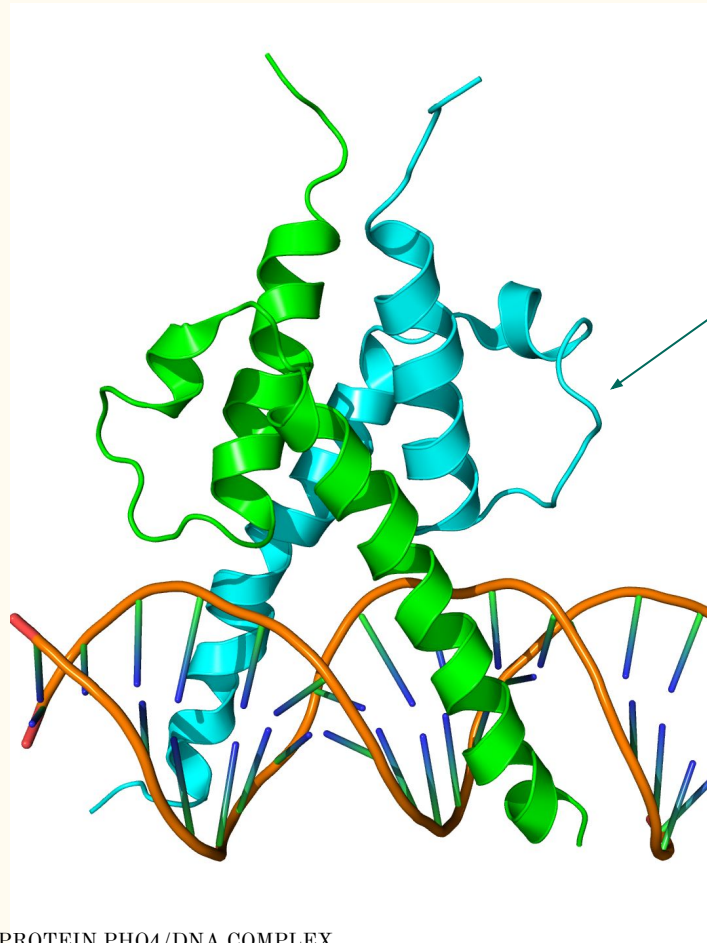


# Annotation des régions promoteurs

---

# Régions promoteurs

- Les séquences promoteurs sont situées **avant les gènes**.
- Ce sont des mots de **6 à 10 lettres**
- Ce sont des sites de fixation pour des protéines appelées **facteur de transcription**
- "**interrupteurs**" qui lancent la production d'une protéine



Protéine se liant à  
l'ADN

Séquence promotrice  
= promoteur

**Notre but est de détecter des régions promoteurs  
par des méthodes statistiques**

# “Mots”

On attend que les “mots” promoteurs vont apparaître plus souvent que les autres mots

Combien existe-il de mots différents de taille  $k$  ?

# “Mots”

On attend que les “mots” promoteurs vont apparaître plus souvent que les autres mots

Combien existe-il de mots différents de taille  $k$  ?  $\longrightarrow 4^k$  mots

Avec  $k=2$  :

AA	AT	AC	AG
TA	TT	TC	TG
CA	CT	CC	CG
GA	GT	GC	GG

# Compter les mots d'une séquence

- Ecrire la fonction qui compte le nombre d'occurrences pour tous les mots de taille  $k$  dans une séquence d'ADN.
- On comptera les **occurrences chevauchantes**
- Par exemple pour  $k = 2$  :

ATCAT

AT=1 ,

# Compter les mots d'une séquence

- Ecrire la fonction qui compte le nombre d'occurrences pour tous les mots de taille  $k$  dans une séquence d'ADN.
- On comptera les **occurrences chevauchantes**
- Par exemple pour  $k = 2$  :

A**TC**AT

AT=1, TC=1,



# Compter les mots d'une séquence

- Ecrire la fonction qui compte le nombre d'occurrences pour tous les mots de taille  $k$  dans une séquence d'ADN.
- On comptera les **occurrences chevauchantes**
- Par exemple pour  $k = 2$  :

ATCAT

AT=1 , TC=1 , CA=1

# Compter les mots d'une séquence

- Ecrire la fonction qui compte le nombre d'occurrences pour tous les mots de taille  $k$  dans une séquence d'ADN.
- On comptera les **occurrences chevauchantes**
- Par exemple pour  $k = 2$  :

ATCAT

AT=2, TC=1, CA=1

- Ensuite, comparer le nombre d'occurrences et le nombre d'occurrences théoriques

# Simulation de séquences aléatoires

- Ecrire une fonction qui génère une séquence aléatoire d'une composition donnée
- Comparer le comptage attendu et le comptage observé

# Modèle de dinucléotides

Le modèle précédent est très simple, il ne peut pas prendre en compte le fait certaines combinaisons de nucléotides ont plus de chance d'apparaître que d'autres.

## Nouveau modèle :

- $M$  est une matrice de taille  $4 \times 4$ , avec  $M(i, j) = P(j | i)$  la probabilité de la lettre  $j$  sachant qu'on est à la lettre  $i$ .
- la distribution de probabilité initiale est donnée par la table de fréquence de nucléotides

# Petits conseils

- Coder en Python
- Bien commenter vos codes
- Utiliser Jupyter
- Enregistrer régulièrement votre travail

