# 3. Prédiction des gènes

- Tous les gènes se terminent sur un codon stop
- Un algorithme simple de prédiction de gènes
- À la recherche des codons start et stop
- Prédiction de tous les gènes d'une séquence
- Comment améliorer la qualité des prédictions ?
- L'algorithme de Boyer-Moore
- Index et arbre des suffixes
- **Des méthodes probabilistes à la rescousse**
- Comment évaluer la qualité de prédiction des méthodes ?
- La prédiction de gènes dans les génomes eucaryotes

# Utiliser les fréquences des lettres

- Dans un texte dont les lettres apparaissent de façon aléatoire, sont insérés des **passages écrits dans une langue inconnue**
- **Comment identifier ces passages de façon automatique ?**

# Fréquences des lettres dans les textes en Français et en Anglais

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Français** | 9,42 | 1,02 | 2,64 | 3,39 | 15,87 | 0,95 | 1,04 | 0,77 | 8,41 | 0,89 | 0,00 | 5,34 | 3,24 | 7,15 | 5,14 | 2,86 | 1,06 | 6,46 | 7,90 | 7,26 | 6,24 | 2,15 | 0,00 | 0,30 | 0,24 | 0,32 |
| **Anglais** | 8,08 | 1,67 | 3,18 | 3,99 | 12,56 | 2,17 | 1,80 | 5,27 | 7,24 | 0,14 | 0,63 | 4,04 | 2,60 | 7,38 | 7,47 | 1,91 | 0,09 | 6,42 | 6,59 | 9,15 | 2,79 | 1,00 | 1,89 | 0,21 | 1,65 | 0,07 |

# Utiliser les fréquences des lettres

- Dans un texte dont les lettres apparaissent de façon aléatoire, sont insérés des **passages écrits dans une langue inconnue**
- **Comment identifier ces passages de façon automatique ?**

- Calculer la fréquence des lettres dans une fenêtre glissante
- Comparer les fréquences obtenues avec les fréquences attendues
- Appliquer un test statistique ($\chi^2$) pour vérifier que les différences sont significatives

# Sur les séquences génomiques

- **Détecter les biais** dans l'usage des nucléotides
- **Chaînes de Markov**
  - Calcul des probabilités de rencontrer un nucléotide donné après avoir rencontré k nucléotides successifs juste avant (probabilités conditionnelles)
  - k est l'ordre du modèle

1. **Phase d'apprentissage** : calcul d'un modèle sous la forme d'une matrice de transition
2. **Phase de prédiction** : utiliser la matrice pour décider entre région codante et non-codante

L’association de la recherche de motifs (codons start et stop, RBS) et d’un modèle de Markov bien entraîné produit de bons résultats sur la plupart des génomes bactériens