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

La nécessité de disposer d'une référence

- **Comparer les résultats d'un prédicteur avec un génome bien annoté**
  - pour lequel les prédictions ont été confirmées expérimentalement
  - par exemple, un génome d'*E. coli*

La nécessité de disposer d'une référence

- **Comparer les résultats d'un prédicteur avec un génome bien annoté**
  - pour lequel les prédictions ont été confirmées expérimentalement
  - par exemple, un génome d'*E. coli*

En pratique, très peu de génomes ont été annotés et confirmés
par des démarches expérimentales

Comparaison de deux listes de gènes

- Appliquer l'algorithme de prédiction de gènes sur la séquence de référence
- Comparer les positions des codons start et stop des deux annotations, celle de référence et celle produite par l'algorithme

- **Vrais positifs** (VP) : les gènes prédits par l'algorithme et confirmés sur l'annotation de référence
- **Faux positifs** (FP) : les gènes prédits, qui ne se retrouvent pas sur la référence
- **Faux négatifs** (FN) : les gènes qui ne sont pas prédits par l'algorithme, alors qu'ils existent sur la référence

Sensibilité et précision

Sensibilité = VP / (VP + FN)

Précision = VP / (VP + FP)

Exemple

- Notre algorithme de prédiction de gène se comporte raisonnablement bien sur le génome de *B. subtilis*

- Il prédit correctement 3 500 gènes sur les 4 100 attendus
- Mais il prédit aussi 1 200 faux positifs