BIOINFORMATIQUE : GÉNOMES ET ALGORITHMES

Analyse informatique de l'information génétique

François Rechenmann

informatiques mathématiques
Inria

GÉNOMES ET ALGORITHMES

1. ADN et séquences génomiques
2. Gènes et protéines
3. **Prédiction des gènes**
4. Comparaison de séquences
5. Arbres phylogénétiques

François Rechenmann

informatiques mathématiques
Inria

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

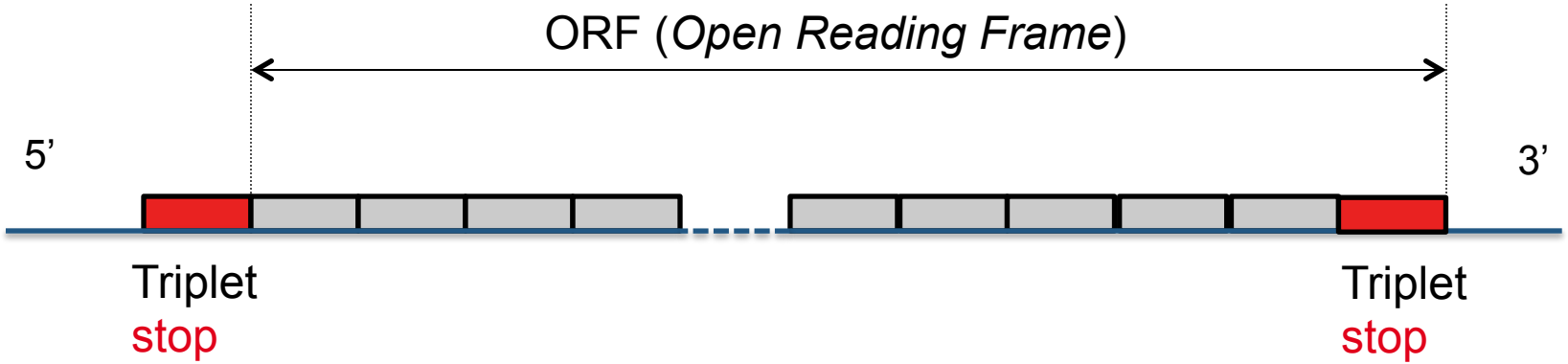
Conditions nécessaires

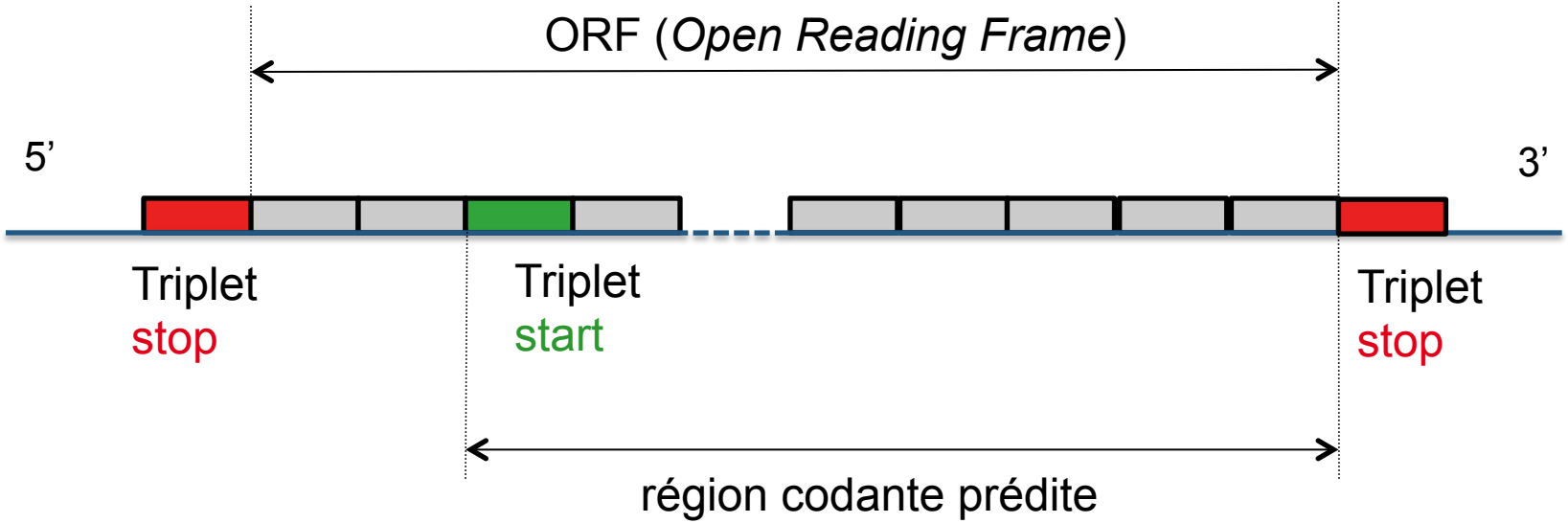
- Une **région codante** ne peut se situer qu'**entre deux stop consécutifs** dans la même phase
- Une **distance minimale entre les 2 codons stop** est requise pour coder une protéine suffisamment longue
  - Typiquement 300 nucléotides (100 AA)

Open Reading Frame (ORF)

- **Sur chaque phase**, **sur les deux brins** (c'est-à-dire sur 6 séquences différentes) :

1. Rechercher les triplets stop
2. Si la distance entre 2 triplets stop consécutifs est supérieure au seuil, enregistrer cette ORF
3. Dans cette ORF, rechercher le premier codon start, de façon à ce que la protéine codée soit de taille maximale


ORF (*Open Reading Frame*)
5’
3’
Triplet
stop
Triplet
stop


ORF (*Open Reading Frame*)
5’
3’
Triplet
stop
Triplet
start
Triplet
stop
région codante prédite