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**

Génomes eucaryotes

- **Très longues régions intergéniques**
  - Gènes : moins de 5% d'un génome humain

- Les gènes sont interrompus par des **régions non-codantes**, appelées **introns**


exon
intron
exon
intron
exon
intron
exon
ARN

Génomes eucaryotes

- **Très longues régions intergéniques**
  - Gènes : moins de 5% d'un génome humain

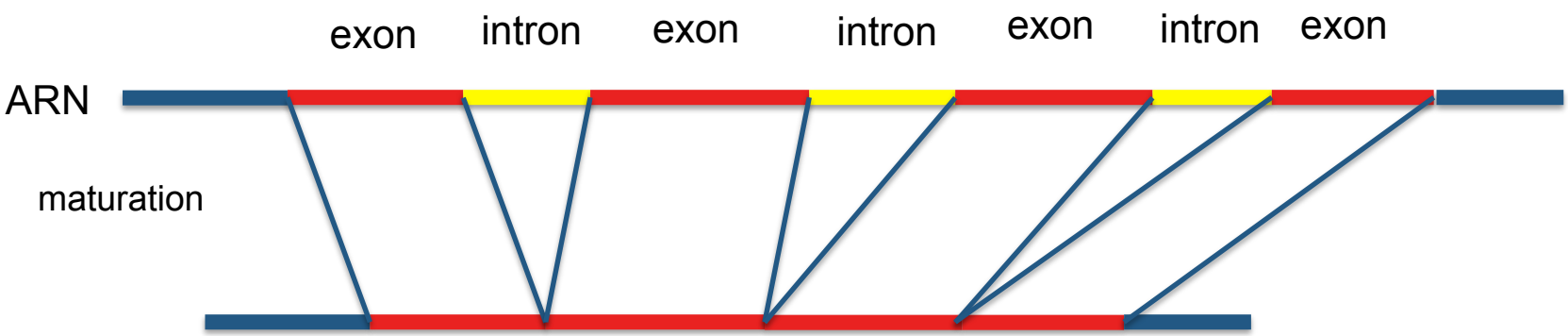
- Les gènes sont interrompus par des **régions non-codantes**, appelées **introns**
  - Moins de 2,5% d'un génome humain est codant

Génomes eucaryotes

- **Très longues régions intergéniques**
  - Gènes : moins de 5% d'un génome humain

- Les gènes sont interrompus par des **régions non-codantes**, appelées **introns**
  - Moins de 2,5% d'un génome humain est codant

- Les introns sont **excisés** lors de la **phase de maturation de l'ARNm**

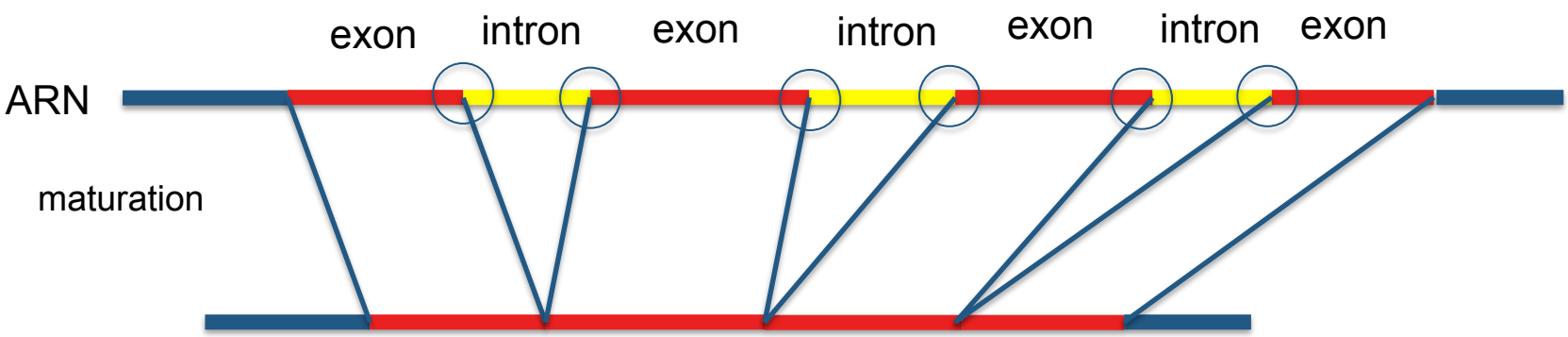

exon
intron
exon
intron
exon
intron
exon
ARN
maturation

Génomes eucaryotes

- **Très longues régions intergéniques**
  - Gènes : moins de 5% d'un génome humain

- Les gènes sont interrompus par des **régions non-codantes**, appelées **introns**
  - Moins de 2,5% d'un génome humain est codant

- Les introns sont **excisés** lors de la **phase de maturation de l'ARNm**
  - Les jonctions introns-exons doivent être prédites


exon
intron
exon
intron
exon
intron
exon
ARN
maturation

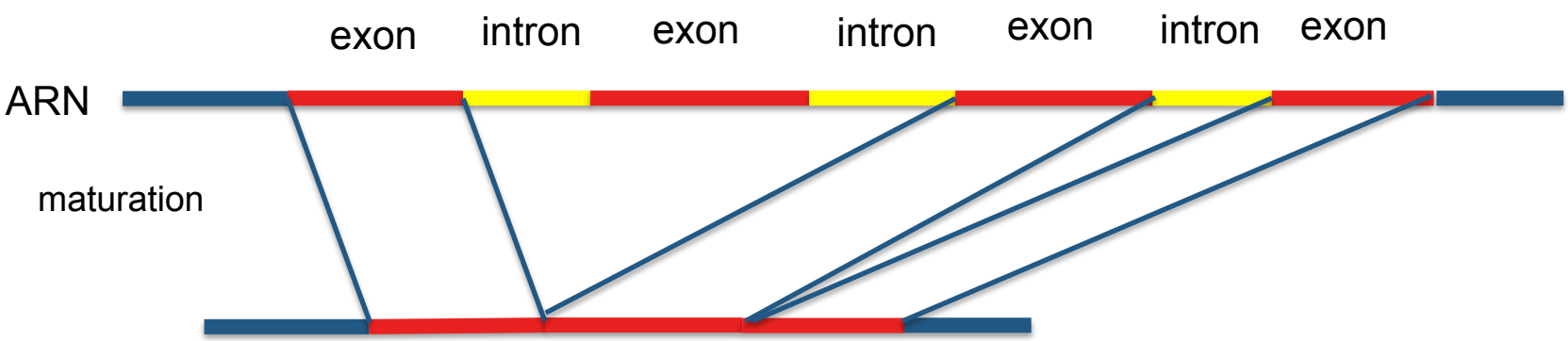
Génomes eucaryotes

- **Très longues régions intergéniques**
  - Gènes : moins de 5% d'un génome humain

- Les gènes sont interrompus par des **régions non-codantes**, appelées **introns**
  - Moins de 2,5% d'un génome humain est codant

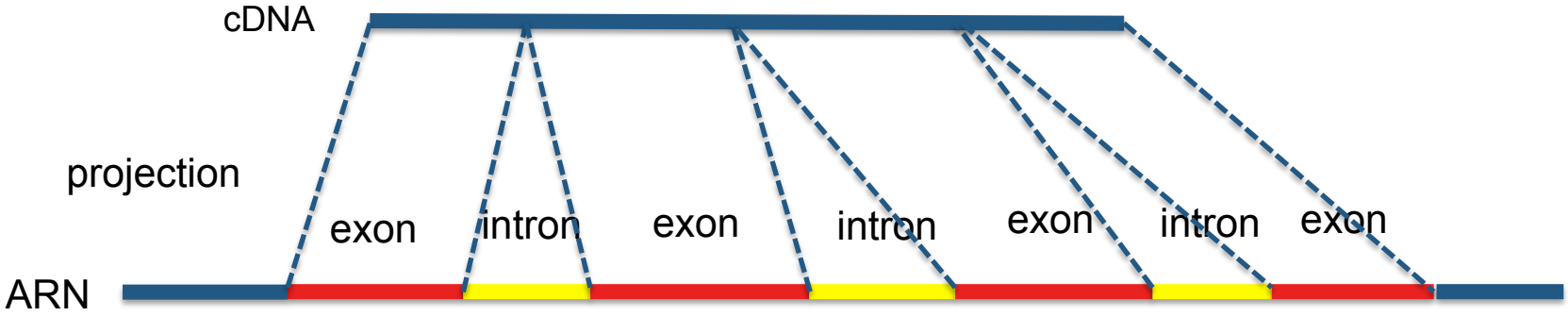
- Les introns sont **excisés** lors de la **phase de maturation de l'ARNm**
  - Les jonctions introns-exons doivent être prédites

- **Epissage alternatif**
  - Un gène → plusieurs protéines


exon
intron
exon
intron
exon
intron
exon
ARN
maturation

Utiliser toutes les connaissances disponibles

- **Modèles de Markov cachés** (*Hidden Markov Models* : HMM)
  - Autant de modèles reliés entre eux que de types de régions :
    régions codantes, non-codantes, intergéniques, etc.
- Motifs caractérisant les **jonctions exon/intron**
- **Données expérimentales**
  - ADN complémentaire (cDNA)


cDNA
projection
exon
intron
exon
intron
exon
intron
exon
ARN

Un problème encore mal résolu

- La **prédiction des gènes eucaryotes** n'a pas encore reçu de solution générale

- Les prédicteurs de gènes **ne simulent pas les mécanismes cellulaires**
  - « Ils font feu de tout bois »

- Les prédictions bioinformatiques sont des **prédictions**