GENOMES AND ALGORITHMS

Computer analysis of genetic information

François
Rechenmann

GENOMES AND ALGORITHMS

1. Genomic texts
2. Genes and proteins
3. **Gene prediction**
4. Sequence comparison
5. Phylogenic tree construction

3. Gene prediction

- **All genes end on a stop codon**
- A simple algorithm for gene prediction
- Searching for start and stop codons
- Predicting all the genes in a sequence
- Making the predictions more reliable
- Boyer-Moore algorithm
- Index and suffix trees
- Probabilistic methods
- Benchmarking the prediction methods
- Gene prediction in eukaryotic genomes

All genes end on a stop codon

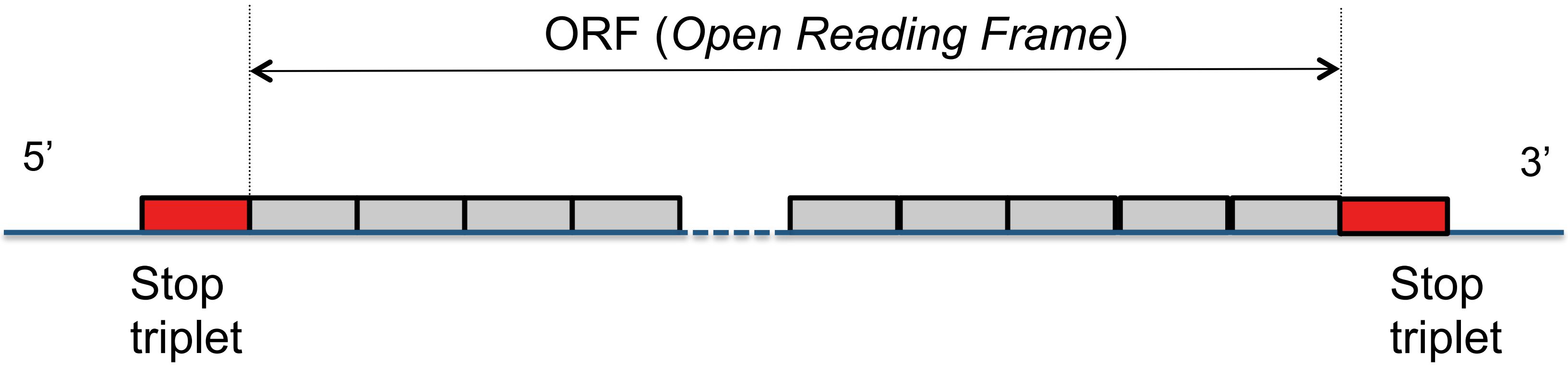
Necessary conditions

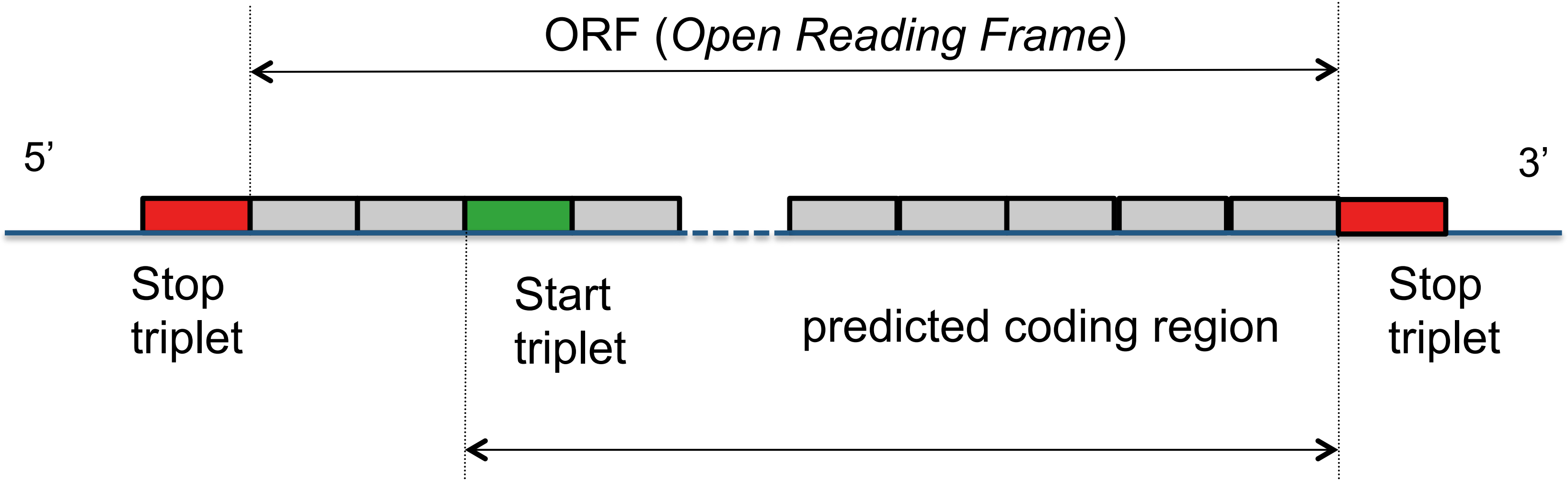
- A gene can only be found between two consecutive stop triplets in the same phase
- A minimal distance is required in order to be translated into an protein
  - Typically 300 nucleotides (100 AA)

Open Reading Frame (ORF)

- On every phase, on the two strands (i.e. 6 different sequences)

1. Look for stop triplets
2. If distance between 2 consecutive stop triplets > 300, then record the ORF
3. In the ORF, look for the “left”-most start codon, so that the coding region length is maximal


ORF (*Open Reading Frame*)
5’
3’
Stop triplet
Stop triplet


ORF (*Open Reading Frame*)
5’
3’
Stop triplet
Start triplet
predicted coding region
Stop triplet