

3. Prediction of coding genes

Guillaume GAUTREAU, 28/09/2022

What analyses can be done based on prokaryotic genomes ?

- Compare the similarities and differences between genomes
 - Phylogenetic:
 - Trees
 - Comparative genomics:
 - Conserved synteny plots
 - Pangenome analysis (core genome + accessory genome)
- Identify the biological features owned by an organism:
 - Syntactic annotation (synonym : structural annotation)
 - Functional annotation
 - Relational annotation

What analyses can be done based on prokaryotic genomes ?

- Compare the similarities and differences between genomes
 - Phylogenetic:
 - Trees
 - Comparative genomics:
 - Conserved synteny plots
 - Pangenome analysis (core genome + accessory genome)
- Identify the biological features owned by an organism:
 - **Syntactic annotation**
 - Functional annotation
 - Relational annotation

How to detect gene positions in a genome ?

Starting from a genome assembly (complete assembly or draft assembly):

Syntactic annotation based on the specific patterns of the genes :

- Identify the coding genes :
 - Start using a start codon (mostly “ATG”) and end with a STOP codon
 - Has a specific frequency of nucleotides (because of the genetic code constraints)
- Identify the non coding genes based on their specific patterns :
 - tRNA
 - tmRNA
 - rRNA
 - lncRNA

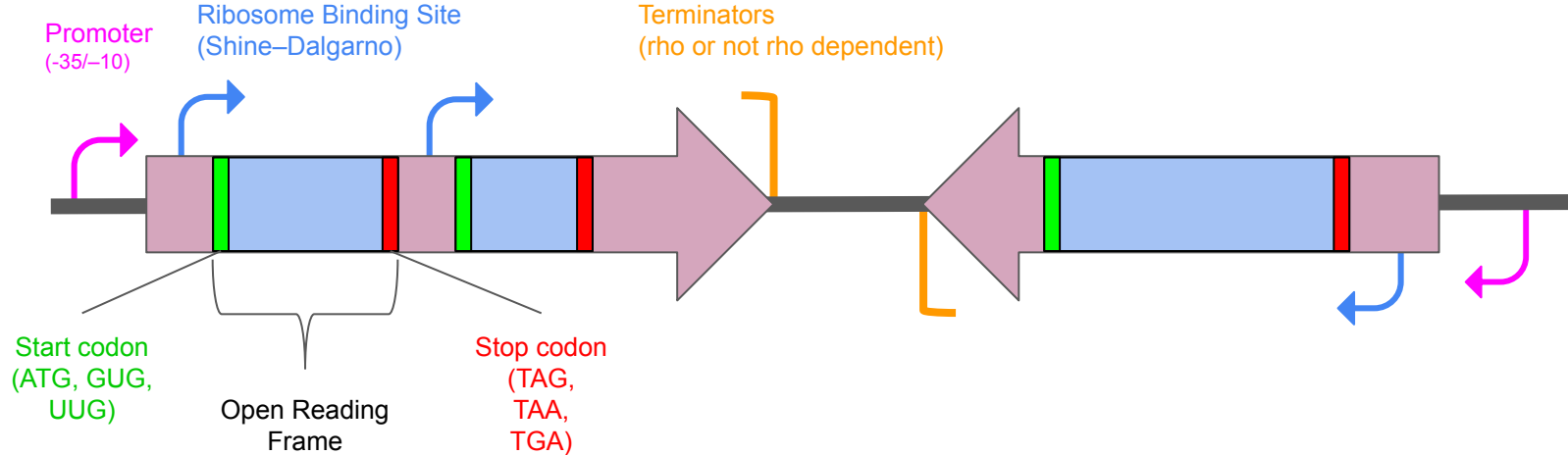
How to detect genes position in a genome ?

Starting from a genome assembly (complete assembly or draft assembly):

Syntactic annotation based on the specific patterns of the genes :

- **Identify the coding genes (main part of the genomes) :**
 - Start using a start codon (mostly “ATG”) and end with a STOP codon
 - Has a specific frequency of nucleotides (because of the genetic code constraints)
- Identify the non coding genes based on their specific patterns :
 - tRNA
 - tmRNA
 - rRNA
 - lncRNA

Reminder on the genomic structure of coding genes

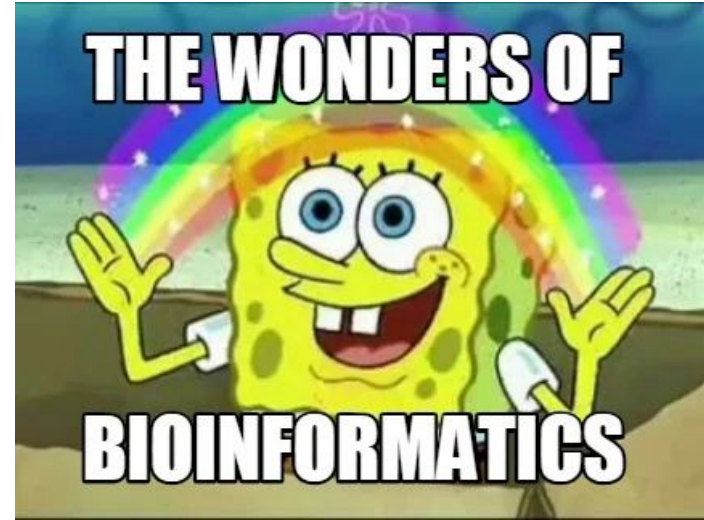


But the actual data are:

GTGAAGGGCAATCAGCTGTTGCCCGTCTCACTGGTAAAAGAAAAACCACCTGGCGCCCAATACGAAACCGCCTCTCCCCGCGGTTGGCCGATTCAATATGC
AGCTGGCACACAGGTTTCCCAGCTGAAAAGCGGGCAGTGAAGCGCAACCGCAATTAATGTGAGTTAGCTCACTTACAGCCACCCAGGCTTTACACTTTATGCTTCC
GCTCGTATGTTGTGGAAATTTGAGCCGGATAACAATTTACACAGGAAACAGCTTACCCATGATACGGATTCACTGGCCGTCGTTTACAAACGTCGTGACTGGT
AAAACCCCTGGCGTTACCCAACTTAATCGCTTGCAGCACCACCCCTTTGCCAGCGTGGCGTAAATAGCGAAAGAGGCCCGCACCGGATCGCCCTTCCCAACAGTTGGC
CAGCTGAAATGGCGCAATGGCCGTTTGGCTGTTTTCCGGCAGTCCAGGCGGAGCGGTAAGCGTGGCGTGGAGTGGCATTTCTGAGGCCGATACTCGTCGTGCC
CTAAACTGGCAGATGCACGGTTACGATGCGCCCATACACCAACGTGACCTATCCATTACGGTCAATCCGCCGTTTGTCCACGGAGAATCCGACGGGTTGT
ACTCGCTACACTTAATGTTGATGAAAAGCTGGCTACAGGAAGGCCAGACGCGAATATTTTTGATGGCGTTAACTCGCCGTTTCACTGTGGTGAACCGGGCGCTGG
GTCGGTTACGGCCAGGACAGTCGTTTCCGCTCTGAAATTTGACCTGAGCGCATTTCACGCGCCGGAGAAAAACCGCCTCGCCGTTGATGGTGTGCGCTGGAGTGAC
GCAGTTATCTGGAAGATCAGGATATGTGGCGGATGAGCGCCATTTCCGTGACGTCTGTTGCTGCATAAACCGACTACACAAATCAGCGATTCCATGTTGCCACT
GCTTTAATGATGATTTTACGCCGCTGTACTGGAAGCTGAAGTTAGATGTGCGCGGAGTTGCGTGACTACCTACGGGTAACAGTTCTTTATGGCAGGGTGAACG
CAGGTCGCCAGCGGCACCCGCGCTTTCCGGCGTGAATATTCGATGAGCGTGGTGGTTATCCGATCGCGTACACACTACGCTGAAACGTCGAAAACCCGAAACTGT
GAGGCGCCGAAATCCGAACTCTATCGTGGCGTGGTGAAGTGCACACCGCCGACGGCAGCGTGATTGAAGCAGAAGCCTGCGATGTGCGTTTCCGCGAGGTTGC
GGATTGAAAATGGTCTGCTGCTGCTGAAGCGCAAGCCGTTGCTGATTCAGGCGTTAACCCGTACAGGAGCATCCTCTGATGGTCAGGTATGGATGAGCAGAC
GATGGTGCAGGATATCCTGCTGATGAAGCAGAACAACTTTAACGCCGTCGCTGTTCCGATTATCCGAACCATCCGCTGTGGTACACGCTGTGGCAGCCGTACGGC
CTGATGTGGTGGATGAAGCCAAATTTGAAACCCAGCGCATGGTCCAAATGCTGACCCGATGTCGCGTGGTACCGGGGATGAGCGAACCGTAAACG
GAATGGTCAGCGCATGTAATCACCAGTGTGATCATCTGGTCGCTGGGAATGAATCAGGCCACGGCGCTAATCAGACGCGCTGTATCGCTGGATCAATC
TGTGATCTTTCCCGCCGGTGCAGTATGAAGCGCGCGCACACCCAGCCATGATTTTCCCGGATGACGCGCGCTGATGAAGACCGAGCCCTT
CCCGCGCTGCGGAAATGGTCCATCAAAAATGGCTTTCGCTACCTGGAGAGACGGCCCGCTGATCTTTGCGAATACGCCACCGCATGGTAAACAGCTTGGC
GGTTTCCGTAATACTGCGAGCGGTTTCCGTCAGTATCCCGGTTTACAGCCGCGCTCGCTGGGACTGGGTGGATCAGTCGCTGATTAATATGATGAAAACGGCAA
CCCGTGGTGGCTTACGGCGGTGATTTTGGCGATACCGCGAACGATCGCCAGTCTGTATGAACGGTCTGGCTTTTGGCGACCGCACCGCATCCAGCGCTGACG
GAAGCAAAACACCCAGCAGCAGTTTTTCCAGTTCCGTTTATCCGGCAAAACCATCGAAGTGACCCAGCGAATACCTGTTCCGCTATAGCGATAACGAGCTCCTGCACCT
GATGGTGGCGCTGGATGGTAAGCCGCTGGCAAGCGGTGAAGTGCCTCTGGATCTGCTCCACAAGGTAACAGTTGATTGAACCTGCCTGAACTACCCGACGCGGA
GAGGCCCGGGCAACTCTGGCTACAGTACCGTAGTGAACCGAACCGACCGGATGGTCAGAAGCCGGGACATCAGCGCCTGGCAGCAGTGGCGCTCGGGG
AAAACCTCAGTGTGACGCTTCCCGCGCGCTCCACCCGATCCCGCATCGCCATCAGCCACAGCGAAATGGATTTTGCATCGAGTAAAGCGTTGGCAATTTAAC
CGCCAGTCAGGCTTTCTTTACAGATGTGGATTGGCATAAAAAACAATGCTGACGCGCGCTGCGCGATCAGTTACCCGCTGCACCGCTGGATAACGACATTTGGCT
AAGTGAAGCGACCCCGATTGACCCTAACCGCTGGGTGCAAGCTGGTGAAGCGCGCGCCATTACCAGGCCGAAGCAGCGTTTGTGACGCTGCACGAGATACACT
TGCTGATGCGGCTGCTGATTACGACCGCTCACGCGTGGCAGCATCAGGGGAAAACTTATTATCAGCCGGAAAAACCTACCGGATTGATGGTAGTGGTCAAATGGCG
ATTACCGTTGATGTTGAAGTGGCGGAGCGATACACCCGATCCCGCGCGGATGGCTGCAACTGCCAGCTGGCGCAGGTAGCAGAGCGGGTAAACTGGCTCGGATTA
GGCGCGCAAGAAAACTATCCGACCGCCCTACTGCGCGCTGTTTTGACCGCTGGGATCTGCCATTGTGAGACATGATACCCCGTACGCTTTCCCGAGCGAAAAACG
GTCTGCGCTGCGGGACGCGCGAATTAATATGGCCACACCCAGTGGCGGGGCGACTTCCAGTTCAACATCAGCCGCTACAGTCAACAGCACTGATGAAAACCG
CCATCGCCATCTGCTGCACGGGAAAGGACAGATGGCTGAATATCGACGGTTTCCATATGGGGATTGGTGGCGCAGACTCCTGGAGCCCGTCAATGATCGCGGAA
TTCAGCTGAGCGCCGGTGCCTACCATTAACGATTTGGTCTGGTGTCAAAAATAATAAACCGGGCAGGCCATGCTGCGCCGATTTTCCGTAAGGAAATCCATATG
TACTATTAAAAAACAACAATTTGGATGTTGCGTTTATCTTTTCTTACTTTTATCAGTGGGAGCCTACTTCCCGTTTTCGCGATTGGCTAGCATGACATCAAC
CATATCAGCAAAAGTATACCGGTATATTTTTGCGGCTATTTCTGTGCTGCTATTTCCAAACCGCTGTTTGGCTGCTTTCTGACAAAACCTGGGCTGCGCAAT
ACCTGCTGGGATTATACCGCATGTAGTGTGATGTTTGGCCGCTTCTTATTTTATCTTCCGGCCACTGTTACAATAACAATTTAGTAGGATCGATGTTGGTGGT
ATTTATTCAGCTTTTGGTTTAAACCGCGGTGCCACAGCATAGAGCAATTAATGAGAAAGTACGCGCTGCAGTAATTTGCAATTTGGTTCGCGCGCGGATTTGGC
TGTGTTGGCTGGCGCTGTGTGCTCGATGTGCGCATCATGTTACCATCAATAATCAGTTGTTTCTGGCTGGGCTCGGCTGTGCACTCATCTCGCCGTTTTA
CTCTTTTTGCGAAAACGGATGCGCCCTTCTGCCACGGTGGCAATGCGGTAGGTGCCAACCATTCGCGATTAGCCTTAAGCTGGCAGCTGGAACGTTTGCAGCA
GCCAAAACCTGTGGTTTTGTCACTGTATGTTATGGCGTTTCTGTGACCATACGATGTTTTGACCAACAGTTTGTCTAATTTCTTACTTCGTTCTTTGCTACCGGTAAAC
AGGGTACGCGGGTATTTGGCTACGTAAACGACAAATGGCGGAATTACTTAAACCGCTGATTTATGTTCTTTGCGCCACTGATCATAATCGCATCGGTGGGAAAAACCGCC
CTGCTGCTGGCTGGCAGTATGTCGCTGACGTAATTTGGCTCATCTGCGCCACTCAGCCCTGGAAGTGGTTATCTGAAAACCGTGCATATGTTGAAGTACCG
TTCCTGCTGGTGGGCTGCTTTAAATATATACCAGCCAGTTGAAAGCGGTTTTTCAAGCGAGGATTATCTGGTCTGTTTGTGCTTTTAAAGCAACTGGCGGATGATTT
TATGCTGTACTGGCGGGCAATATGATGAAAACGATCGTTTCCAGGGCCCTTATGTTGGTGGGCTGCGGCTGCGGCTTACCTTAATTTCCGTGTTTCCGCT
TTAGCGGCCCGCGCCGCTTTCCCTGCTGCGTGCAGTGAATGAAGTGCCTTAAAGCAATCAATGTGCGATGCGCGCGGAGCGCCTTATCCGACCAACATATCAT
AACGGAGTGTGCGATTGAAACATGCCAATGACCGAAAAGATAAGAGCAGGCAAGCTATTACCGATGTGCGAAGGCTACCGGAAAAAAGACTCTGTGGGAAAAAC
GTTAATGATGAGTTTAACTACTCGCATCCATCAGA

But the actual data are:

GTGAAGGGCAATCAGCTGTTGCCGCTCTACTGGTGAAGAAAAAACCCCTGGCGCCAAATACGAAACCGCCTCTCCCCGCGGTTGGCCGATTCAATATGC
AGCTGGCACACAGGTTTCCCAGCTGGAAGAGCGGGCAGTGAAGCCAAACCGCAATTAATGTAGTGTAGCTCACTTAGGCACCCAGGCTTTACACTTTATGCTTCC
GGCTCGTATGTTGTGGAAATTTGAGCGGGATAACACTTTCACAGGAAACAGCTTACGACCATGATACGGATTCACTGGCCGTCGTTTACAAACGTCGTGACTGGG
AAAAACCCCTGGCGTACCCCAACTTAATCGCTTCCAGCAGTACCCCTTTCCGCGAGCTGGCCATTAAGCGAAAGAGGCCCGCACCGGATCGCCCTTCCCAACAGTTGGC
CAGCCTGAATGGCGGAATGGCGCTTGGCTGTTTTCCGATCCAGCAAGCGGTCGCGGAAAGCTGGCTGGAGTGGCATTCTCTGAGGGCCGATACTGCTGCTGCC
CTAAACCTGGCAGATGCACGGTTACGATGCGCCCATACACCAACGTGACCTATCCATACGGTCAATCCGCGCTTTGTTCCACGGGAATCCGACGGGTTGT
ACTCGCTCACATTTAATGTTGATGAAAGCTGGCTACAGGAAGGCCAGACGCGAATATTTTTGATGGCGTTAACTCGCGGTTTCATCTGTGGTCAACGGCGCTGG
GTCGGTTACGGCCAGGACAGTCGTTTTGCCGCTGTAATTTGACCTGAGCGCATTTCACGCGCCGAGAAAAACCGCCTCGCGGTTGATGGTGTGCGGCTGGAGTGACG
GCAGTTATCTGGAAGATCAGGATATGTGGCGGATGAGCGCGCATTTCCTGTGACGTCTGCTTTGCTGCATAAACCGACTACACAAATCAGCGATTCCATGTTGCCACTC
GCTTTAATGATGATTTTACGCGCGCTGTACTGAGGCTGAAGTTAGATGTGCGCGGAGTTCGCTGACTACCTACGGGTAACAGTTCTTTATGGCAGGGTGAAACG
CAGGTCGCCAGCGGCACCGCGCCTTTCCGCGGTGAATATTCGATGAGCGTGGTGGTTATCCGATCGCGTACACTACGCTGAAACGTCGAAACCCGAAACTGT
GAGCGCCGAAATCCGAACTCTATCGTGGCGGTTGAACTGCACACCGCCGACCGCCAGCTGATTGAAGCAGAAAGCCTGCGATGTGCGTTTCCGCGAGGTCG
GGATGAAATGGTCTGCTGCTGAAGCGCAAGCCGTTGCTGATTCCGAGGCTAACCGCTCACGAGCATCATCTGATGGTCAGGTGATGGATGAGCAGAC
GATGGTGCAGGATATCTGCTGATGAAGCAGAACTTAACCGCGTGGCTGTTCCGATTATCCGAACTACCGCTGTGGCAGCCGTAGCGCCGCTAGCGC
CTGATGTGGTGGATGAAGCCAAATTTGAAACCCAGCGCATGTCGCAATGAATCGCTGACCCGCTGCGGCTGACTACCGGGGATGAGCGAAACCGTAACCG
GAATGGTCAGCGGATCGTATACCCGAGTGTGATCATCTGCTGCTGGGAATGAATCAGGCCAGCGGCTATACGACGCGCTGATCGCTGGATCAAACT
TGCTGATCTTCCGCGCGTACGATGAAGCGCGGAGCGCACCCGACCCAGCCAGCATATTTGCCCAGTACGCGCGCGTGATGAAGACCCAGCCCTT
CCGCGCTGTCCGAAATGGTCCATCAAAAATGGCTTTCGCTACCTGGAGAGACGCGCCCGCTGATCTTTGCGAATACGCCACCGCATGGGTAACAGTCTTGGC
GGTTTTCCGTAATACTGGCAGCGCTTTCGTCAGTATCCCGCTTTACAGCGCGCTTCGCTGGGACTGGTGGATCAGTCGCTGATTAATATGATGAAACCGGCA
CCCGTGTGCTGGCTACGGCGGATTTTTGGCGATCGCCGCAAGCAGTCCGCGATTCTGTATGAACGGTCTGGCTTTTCCGCGACCGCAACCGCATCCAGCGCTGACG
GAAGCAAAACACCCAGCAGCATTTTTCCAGTTCCGTTATTCGCGGCAAAACCTGGAATGACCCAGCGAATACCTGTTCCGTCATGCGATAACGAGCTCCTGCACTG
GATGGTGGCGCTGGATGGTAAGCCGCTGGCAAGCGGTTGAAGTGCCTGTGATGTCGCTCCACAAGGTAACACAGTTGATGAAGTCGCTGAACTACCGCAGCCGA
GAGCGCCGGGCAACTCTGGCTCACAGTACCGCTAGTGAACCGCAAGCGCCGATGGTTCAGAAAGCCGGGACATCAGCGCCTGGCAGCAGTGGCGCTCGCGG
AAAACTCAGTGTGACGCTCCCCGCGCTCCACCGCATCCCGCATCGACCCAGCGAAATGGATTTTTGCATCGAGCTGGGTAATAAGCGTTGGCAATTAAC
CGCCAGTCAGGCTTTTCCAGATGTGGATTGGCATAAAAAAACAATGCTGACCGCGCTGCGCGATCAGTTACCCGCTGCACCCTGGATAACGACATTTGGCCT
AAGTGAAGCGAACCGCATTGACCCCTAACCGCTGGGTGCAACCGCTGGAAGCGCGGCTGCGCCATTACCAGCCGCAAGCAGCGTTGTTGCACTGCACGGCAGATACCT
TGCTGATGCGGTGCTGATTCAGCCGCTACCGGCTGGCAGCATCAGGGGAAAACTTATTTATCAGCCGAAAAACCTACCGGATTGATGGTAGTGGTCAAATGGCG
ATTACCGTTGATGTTGAAGTGGGAGCGGATACCCGATCCGCGCGGATTTGGCTGAACTGCCAGTGGCGCAGGTAGCAGAGCGGGGTAACCTGGCTCGGATTA
GGCGCGAAGAAAACTATCCGACCGCCTACTGCGCGCTGTTTTGACCGCTGGGATTCGCCATTGTCAGACATGATACCCCGTACGCTTTCCGAGCGAAAAACG
GTCTGCGCTGCGGAGCGCGGAATGAATATGGCCACACCAAGTGGCGGGGCACTCCAGTTCACATCAGCCGCTACAGTCAACAGCAACTGATGAAACCG
CCATCGCATCTGCTGCACGGGAAGAGGCACATGGCTGAATCGACGGTTTCCATATGGGATTTGGTGGCGGACACTCCTGGAGCCGCTCAGTATCGCGGGA
TTCAGCTGAGCGCGGCTGCTACCATACCAATGGCTGCTGGTGTCAAAAATAATAAACCCGCGAGCCATGCTCGCCGATTTCCGCTAAGGAAATCCATTAG
TACTATTAAAAAACAACAATTTGGATGTGCGGTTATCTTTTCTTTACTTTTATCATGGGAGCTACTTCCGCTTTTCCCGATTGGCTACATGACATCAAC
CATATCAGCAAAAGTGATACCGGATTAATTTTTGCGGCTATTTCTGCTGCTATTTACCAACCGCTGTTTGGCTGCTTTCTGACAACTCGGGCTCGCAAT
ACCTGCTGGGATATACCGCATGTAGTGATGTTGCGCGCTTCTTATTTTACTTTCGCGGCACTGTTACAATAACAATTTAGTAGGATGATGTTGGTGGT
ATTTATTAGCTTTTTGTTTAAACCGCGGTGCGCCAGCAAGGCAATTAATGCAAAAGTACGCGCTGCGCATATTTGCAATTTGCTGCGCGCGGATTTGGC
TGTGTTGGCTGGCGCTGTGCTCCTGATGTGCGCATCATGTTACCATCAATAATCAGTTGTTTCTGGCTGGGCTGCTGGCTGCACTCATCTCGCGTTTTA
CTCTTTTTGCGAAAAAGCGGATCGCCCTTCTGCCACGGTGGCAATGCGGTAGTGGCCACCATTCGCGATTAGCCCTTAAAGCTGGCAGCTGGAACGTTGAGACA
GCTTTACTGTGGTTTTGTCACCTGTATGTTTGGCTTTCTGCGACCATACGATGTTTGTGACCAACAGTTTCTGCTAATTTCTTTACTTCGTTCTTTGCTACCGGTGAAC
AGGGTACGCGGGATTTGGCTACGTAAACGACATGGCGGAATTAACCGCCCTGATTTGATGTTCTTTCGCGCACTGATCATTAATCGCATCGGTGGGAAAAACCGC
CTGCTGCTGGCTGGCAGTATGTCGCTGACGATTTAGGCTCATCTGCGCACCTCAGCGCTGGAAGTGGTTATCTGAAACCGCTGCATATGTTGAAAGTACCG
TTCTGCTGGTGGGCTGCTTTAAATATATACCAGCCAGTTGAAAGCGGTTTTGACGGAGCATTATCTGGTCTGTTTCTGCTTTTAAAGCACTGGCGATTGATTT
TATGCTGACTGGCGGGCAATATGATGAAAGCATCGTTTTCCAGGGCGCTTATCTGGTGGGCTGCGGCTGCTGGCGCTGGGCTTACCTTAATTTCCGCTGTTCCG
TTAGCGCGCCCGCGCCCTTTCCCTGCTGCGTGTGAGTGAATGAAGTGGCTTAAAGCAATCAATGTCGGATGCGCGCGGAGCGCCTATCCGACCAACATATCAT
AACGGAGTGTGCTGATGAACTGAACTGAACTGAAAGAAATAAGAGCAGGCAAGCTAATTTACCGATGTCGCAAGGCTTACCGGAAAAAGACTCTGCGGAAAAAC
GTTAATGATGAGTTTAACTACTCGCATCCATCAGA



Bioinformatics solutions

```
GTGAAGGGCAATCAGCTGTTGCCGCTCTACTGGTGAAGAAAAAACCCCTGGCGCCCAATACGCAAAACCGCCTCTCCCCGCGGTTGGCCGATTCAATATGC
AGCTGGCACGACAGGTTTCCCAGCTGGAAGACGGGGCAGTGAAGCCGCAACCGCAATTAATGTGAGTTAGCTCACTTATTAGGCACCCAGGCTTTACACTTTATGCTTCC
GGCTCGTATGTTGTGGAAATTTGAGCCGGATAACAATTTACACAGGAAGAACAGCTTACGACATGATACGGATTCACTGGCCGTCGTTTACAAACGTCGTGACTGGG
AAAACCCCTGGCGTTACCCAACTTAATCGCCTTGACAGCATAACCCCTTTGCGGACGCTGGCCGTAAATAGCGAAAGAGGCCCGCAGCCGATCGCCCTTCCCAACAGTTGGC
CAGCCTGAATGGCCGAATGGCCCTTTGGCTGTTTTGGCTGAGTTCCGCTACCCAGCAAGGCGGCGGAAAGCTGGCGTGGAGTGGCATTCTTCTGAGGGCGATACTGCTGCTGCC
CTAAACTGGCAGATGCACGGTTACGATGCGCCCATACACCAACGTGACCTATCCATTACGGTCAATCCGCGCTTTGTTCCACGGGAATCCGACGGGTTGT
ACTCGCTCACATTTAATGTTGATGAAAGCTGGCTACAGGAAGGCCAGACGCGAATATTTTTGATGGCGTTAACTCGCCGTTTCACTGTGGTGAACGGCCGCTGG
GTCGGTTACGGCCAGGACAGTCGTTTCCGCTCTGAAATTTGACCTGAGCGCATTTTACGCGCCGGAGAAAAACCGCCTCGCCGTTGATGGTGTGCGCTGGGATGACG
GCAGTTATCTGGAAGATCAGGATATGTGGCGGATGAGCCGCTTTCCGTGACGCTGCTGTTGCTGCATAAACCGACTACACAAATCAGCGATTCCATGTTGCCACTC
GCTTAAATGATGATTTTACGCGCGCTGTACTGGAAGCTGAAGTTAGATGTCGCGGAGTTGCGTGACTACCTACGGGTAACAGTTTCTTATGGCAGGGTGAACG
CAGGTCGCCAGCGGCACCCGCGCTTTCCGCGGTGAAATATCGATGAGCGTGGTGGTTATGCCGATCGCGTACACTACGCTGAAACGTCGAAAACCCGAAACTGT
GGAGCGCCGAAATCCCGAATCTATCGTGGCGTGGTTGAACTGCACACCGCCGACGGCAGCGCTGATTGAAGCAGAAGCCTGCCGATGTCGTTTCCGCGAGGTCG
GGATTGAAAATGGTCTGCTGCTGCTGAACGGCAAGCCGTTGCTGATTGAGGGCTTAAACCGTACAGGATCATCTCTGATGGTCAGGTCATGGATGAGCAGAC
GATGGTGCAGGATATCTGCTGATGAAGCAGAAACAATTTAACCGCTGGCTGTTGCGATTATCGAACCATCCGCTGGTGAACAGCTGCGGACCCGCTAGCGG
CTGATGTTGGTGAAGCAATATTTGAAACCCAGCGCATGGTCCAACTGCTGACCGATATCCGCGTGGCTAGCCGGGATGAGCGCAACCGTAAACG
GAATGGTCAGCGCATGTAATCCCGAGTGTGATCATCTGGTCTGGGAATGAATCAGCCAGCGCCGCTAATCAGCAGCGCTGATCGCTGGATCAAACT
TGTCATCTTCCGCGCGTAGTGAAGCCGGCAGCGACACCGCCAGCCGATGATTTGCCGATGACGCGCGCTGGATGAAGACCCAGCCCTT
CCGCGCTGAGCAAAATGGTCCATCAAAAATGGCTTTCGCTACCTGGAGAGACGCGCCGCTGATCCTTTGCCAATACGCCACGCGATGGTAAACAGTCTTGGC
GGTTTCGCTAAACTGCGAGCGCTTTCGTCAGTATCCCGCTTTACAGCCGCGCTTCGCTGGGACTGGTGGATCAGTCGCTGATTAATATGATGAAAACGGCAA
CCCGTGGTCCGCTTACGGCGGTGATTTGGCGATACCGCGAACGATCGCCGATCTGTATGAACGGTCTGGCTTTGCCGACCGCACCCGATCCAGCGCTGACG
GAAGCAAAACCCAGCAGCAGTTTCCAGTTCCGTTTATCCGGCAAAACCGAATCGAATGACCAAGCAATACCTGTTCCGCTCATAGCGATAACGAGCTCCTGCACCTG
GATGGTGGCGCTGGATGGTAAGCCGCTGGCAAGCGGTGAAGTCCCTGTCGCTCCACAAGGTAAACAGTTGATTGAACCTGCCTGAACTACCGCAGCCGA
GAGCGCCGGGCAACTCTGGCTCACAGTACGCGTAGTGAACCGAACCGCAGCCGATGGTCAGAAGCCGGGACATCAGCGCCTGGCAGCAGTGGCGCTGGCGG
AAAACCTCAGTGTGACGCTCCCCGCGCTTCCGCGCGGATCCCGCATCCGCACTGACCCACCGCAAAATGGATTTTGCATGAGCTGGGTAATAGCGTTGGCAATTTAA
CCGCGAGTCAGGCTTTCTTTCAGATGTGGATTGGCATAAAAAACAATGCTGACGCGCGCTGGCGATCAGTTACCCGCTGCACCGCTGGATAACGACATTTGGCT
AAGTGAAGCGAACCCGCAATTGACCCCTAACCGCTGGCTGCAACGCTGGAAGCGCGTCCGCGCCATTACCAGCGCCGAAGCAGCGTTTGGCAGTCCGACATACACT
TGCTGATGGCGTCTGATTACGACCGCTCACGCGTGGCAGCATCAGGGGAAAACTTATTTATCAGCCGAAAAACCTACCGGATTGATGGTAGTGGTCAAATGGCG
ATTACCGTTGATGTTGAAGTGGCGAGCGATACACCGCATCCGCGCGGATGGCTGAACTGCCAGCTGGCCGAGGTAGCAGAGCGGGTAAACTGGCTCGGATTA
GGCGCGCAAGAAAACTATCCGACCGCCCTACTGCGCGCTGTTTTGACCGCTGGGATGTCGCATTGTGAGACATGATACCCCGTACGCTTCCCGAGCGAAAAACG
GTCTGCGCTGCGGGACCGCGGAATTTGAAATATGGCCACACCGACTGGCGGGCGACTTCCAGTTCAACATCAGCCGCTACAGTCAACAGCACTGATGGAACCA
CCATCGCCATCTGCTGACCGGGAAAGGACACATGGCTGAATATCGACGGTTTTCCATATGGGGATTGGTGGCGGAGCTCCTGGAGCCGCTCAGTAGTCCGGCGAA
TTCAGCTGAGCCCGGTGCTACCATACCAATGGCTGTTGGTGTCAAAAATAATAAACCCGGCAGCCATGCTGCCCCGTTATCGCTAAGGAAATCCATATG
TACTATTAAAAAACAACAATTTGGATGTGCGTTTATCTTTTCTTACTTTTATCATGGGAGGCTACTTCCGCTTTTCCCGATTGGCTAGATGACATCAAC
CATATCAGCAAAAGTACAGCGGATTAATTTTGGCCGATTTCTGCTGCTGCTATTTACCCGCTATTTACCAACCGCTGTTGGTCTGCTTTTCGAAAACTCGGGCTGGCAAA
ACCTGCTGGATTATACCGCATGTAGTGATGTTTGGCCGCTTCTTATTTTATCTTCCGGCCACTGTTACAATAACAATTTAGTAGGATCGATTGGTGGT
ATTTATCAGCTTTTGGTAAACCGCGTGGCCAGCAAGCAATTAATGAAAGATCAGCCGCTGCAGTAAITTCGAAATTTGGCTGGCGCGGATGTTGGC
TGTGTTGGCTGGCGCTGTGTCCTCGATTGTCGGCATCATGTTACCATCAATAATCAGTTGTTTCTGGCTGGGCTCGGCTGTCACATCATCTCGCCGTTTTA
CTCTTTTTGCCAAAACGGATGCGCCCTCTTCCGACGCTGTCACATGCGGTAGTGGCCACCACTCCGCTATTAGCCCTTAAAGCTGGCAGCTGGAACGTTGACAGA
CCTTAACTGTGGTTTTGTCACCTGTATGTTATGGCTGTTCTGTGACCAACAGTGTTTGACCAACAGTTTGCATATTTCTTACTCTGTTCTTGTCTACCGGTAAC
AGGGTACGCGGGTATTTGGCTACGTAAACGCAATGGCGCAATTAACCGCCCTGATTTATGTTCTTCCGCACTGATCATTAATCGCATCGGTGGGAAAAACCGCC
CTGCTGCTGGCTGGCACTATATGCTGCTGACTGATTTGGCTCATCTGCGCCACTCAGCCCTCAGCGCTGGAAGTGGTATTCTGAAAAACCGTGCATATGTTGAAGTACCG
TTCTGCTGGTGGGCTGCTTTAAATATATACCAAGCCAGTTGAAAGCGGTTTTGACGGAGCATTATCTGGTCTGTTTCTGCTTCTTAAAGCAACTGGCGATTGATTT
TATGCTGACTGGCGGGCAATATGATGAAAGCATCGTTTCCAGGGCCCTTATGTTGGTGGGCTGCTGGTGGCGCTGGGCTCACCTTAATTTCCGTTGTTCCAGC
TTAGCGCGCCCGCCGCTTTCCCTGCTGCGTGTGAGTGAATGAAGTGGCTTAAAGCAATCAATGTCGGATGCGCGCGGACCGCCTATCCGACCAACATATCAT
AACGGAGTGTGCTGATGAACTGAACTGACCGAAAGATAAGAGCAGGCAAGCTATTACCGATATGTCGGAAGGCTTACCGGAAAAAGACTCTGTTGGGAAAA
GTTAATGATGAGTTTAACTCGCATCCATCAGA
```

Two kinds of approaches:

- **Alignments methods:**

Aligning our sequence over gene data banks of known genes

- **Ab initio methods:**

Detection of signals in our sequence corresponding to a set of rules describing a gene

Bioinformatics solutions

```
GTGAAGGGCAATCAGCTGTTGCCGCTCTACTGGTGAAGAAAAAACCCCTGGCGCCAAATACGCAAAACCGCTCTCCCCGCGCTTGGCCGATTCAATATGC
AGCTGGCACGACAGGTTTCCCAGCTGGAAGACGGGGCAGTGAAGCCAAACGCAATTAATGTGAGTTAGCTCACTTATAGGCACCCAGGCTTTACACTTATGCTTCC
GGCTCGTATGTTGTGGAAATTTGAGCCGATAACAATTTACACAGGAAGAACAGCTTACGACATGATACGGATTCACTGGCCGTCGTTTACAAACGTCGTGACTGGG
AAAACCCCTGGCGTTACCCAACTTAATCGCCTTGACAGCATAACCCCTTTGCCGAGCTGGCGTAATAGCGAAAGAGGCCCGCACCGATCGCCCTTCCCAACAGTTGGC
CAGCCTGAATGGCGCAATGGCCGTTTGGCTGTTTCCGTCACCCAGCAAGCGAGTGGCCGAAAGCTGGCGTGGAGTGGCATCTTCTGAGGGCCGATACTGCTGCTGCC
CTCAAACCTGGCAGATGCACGGTTACGATGCGCCCATACACCAACGTGACCTATCCATTACGGTCAATCCGCGCTTGTCCACGGGAATCCGACGGGTTGT
ACTCGCTCACATTAATGTTGATGAAAGCTGGCTACAGGAAGGCCAGACGCGAATATTTTTGATGGCGTTAACTCGCCGTTTCACTGTGGTGAACGGGGCGCTGG
GTCGGTTACGGCCAGGACAGTCGTTTCCGCTCTGAATTTGACCTGAGCGCATTTCACGCGCCGAGAAAAACCGCCTCGCCGTTGATGGTGCCTGGCTGGAGTGACG
GCAGTTATCTGGAAGATCAGGATATGTTGGCGGATGAGCGCCATTTCCGTGACGCTCGCTTGTGCATAAACCCGACTACACAAATCAGCGATTCCATGTTGCCACTC
GCTTAATGATGATTTTACGCGCGCTGACTGGAAGCTGAAGTTAGATGCGCGGAGTTGCGTGACTACCTACGGGTAACAGTTTCTTATGGCAGGGTGAACG
CAGGTCGCCAGCGGCACCCGCGCTTTCCGCGGTGAATATCGATGAGCGTGGTGGTTATGCCGATCGCGTACACTACGCTGAAACGTCGAAAACCCGAAACTGT
GGAGCGCCGAAATCCCGAATCTATCGTGGCGGTTGAACTGCACACCGCCGACGGCAGCGTGATTGAAGCAGAAGCCTGCCGATGTCGTTTCCGCGAGGTCG
GGATTGAAAATGGTCTGCTGCTGAACGGCAAGCCGTTGCTGATTGAGGGCTAAACCGTACGAGCATCATCTCGATGGTCAGGTATGGATGAGCAGAC
GATGGTGCAGGATATCTGCTGATGAAGCAGAACAACTTTAACCGCGTGGCCTGTTCCGATTATCGAACCATCCGCTGGTGAACCGCTGGCGACCCGTCAGGCC
CTGATGTTGGTGAAGCAAGCAATATGAAACCCAGCGCATGGTCCAACTGCTGACCCGATGTCGCGCTGGCTACCGGGATGAGCGAACCGTAACCG
GAATGGTCAGCGCATGTAATCCCGAGTGTGATCATCTGGTCCGTTGGGAATGAATCAGCCAGCGCCGCTAATCAGACGCGCTGATCGCTGGATCAAACT
TGTCATCTTCCGCGCGTAGTGAAGCCGCGGAGCGACACCGCCACCGCATATATTTGCCGATGACGCGCGCTGGATGAAGACCCAGCCCTT
CCGCGCTGAGCAAAATGGTCCATCAAAAATGGCTTTCGCTACCTGGAGAGACGCGCCGCTGATCCTTTGCAATACGCCACGCGATGGTGAACGCTTGGC
GGTTTCGCTAAACTGCGAGCGCTTTCGTCAGTATCCCGCTTTACAGCCGCGCTTCGCTGGGACTGGTGGATCAGTCGCTGATTAATATGATGAAAACGGCAA
CCCGTGGTCCGCTTACGGCGGTGATTTGGCGATACCGCGAACGATCGCCGATCTGTATGAACGGCTGGTCTTTGCCGACCGCACCCGATCCAGCGCTGACG
GAAGCAAAACCCAGCAGCAGTTTCCAGTTCCGTTTATCCGGCAAAACCGAATCGAATGACCCAGCGAATACCTGTTCCGCTCATAGCGATAACGAGCTCCTGCACCTG
GATGGTGGCGCTGGATGGTAAGCCGCTGGCAAGCGGTGAAGTCCCTGTCGCTCCACAAGGTAAACAGTTGATTGAACCTGCCTGAACTACCGCAGCCGGA
GAGCGCCGGGCAACTCTGGCTCACAGTACGCGTAGTGAACCGCAACCGCAGCCGATGGTGAAGCCGGGGCACATCAGCGCCTGGCAGCAGTGGCGCTCGCCG
AAAACCTCAGTGTGACGCTCCCGCGCGCTCCACCCGCGATCCCGCATCCGACCCAGCGAAATGGATTTTGCATCGAGCTGGGTAATAAGCGTTGGCAATTTAAC
CGCCAGTCAGGCTTTCTTTCAGATGTTGGATGGGCAATAAAAACAATGCTGACGCGCGCTGGCGATCAGTTACCCGCTGCACCCGCTGGATAACGACATTTGGCT
AAGTGAAGCGACCCCGCATTGACCCCTAACCGCTGGCTGCAACGCTGGAAAGCGCGGCGCCATTACCAGCGCCGAAGCAGCGTTGTTGACGATGCGCAGATACACT
TGCTGATGCGGCTGATTACGACCGCTCAGCGCTGGCAGCATCAGGGGAAAACTTATTTATCAGCCGAAAAACCTACCGGATTGATGGTAGTGGTCAAATGGCG
ATTACCGTTGATGTTGAAGTGGCGAGCGATACACCGCATCCGCGCGGATGGCTGAACTGCCAGCTGGCCGAGGTAGCAGAGCGGGTAAACTGGCTCGGATTA
GGGCGCAAGAAAACTATCCGACCGCCCTACTGCGCGCTGTTTGAACCGCTGGGATGTCGCATTGTGAGACATGATACCCCGTACGCTTCCCGAGCGAAAAACG
GTCTGCGCTGCGGGACCGCGGAATTAATATGGCCACACCGCATGGCGGGGACTTCCAGTTCAACATCAGCCGCTACAGTCAACAGCACTGATGGAACCAAC
CCATCGCCATCTGCTGCACGGGAAGGACACATGGCTGAATATCGACGGTTTCCATATGGGGATGGTGGCGGAGCTCCTGGAGCCGCTCAGTATCGGGCGAA
TTCAGCTGAGCGCGGCTGCTACCATACCAATGGCTGGTGGTGTCAAAAATAATAAACCCGGGACGGCCATGCTGCCCGTATTTCCGTAAGGAAATCCATATG
TACTATTAAAAAACAACAATTTGGATGTTGGCTTATCTTCTTACTTTTATCATGGGAGGCTACTTCCGCTTTCGCAATTTCCGATTTGGCTAGATGACATCAAC
CATATCAGCAAAAGTACAGCGGATTAATTTTGGCCGATTTCTGCTGCTGCTATTTACCCGCTATTTACCAACCGCTGTTGGCTGCTCTTTCGAAACCTCGGGCTGGCAAA
ACCTGCTGGATTTACCGCACTGTAGTGATGTTTGGCCGCTTCTTATTTTATCTTCCGGCCACTGTTACAATAACAATTTAGTAGGATGATGTTGGTGGT
ATTTATCAGCTTTTGGTAAACCGCGTGGCCAGCAAGCAATTAATGAAAGATCAGCCGCTGCAGTAATTTGCAATTTGGCTGGCGCGGATGTTGGC
TGTGTTGGCTGGCGCTGTGCTCGTATGTCGGCATCATGTTACCATCAATAATCAGTTGTTTCTGGCTGGGCTCGGCTGTCACATCATCTCGCCGTTTTA
CTCTTTTTGCAAAACGGATGCGCCCTCTTCCGACGCTGTCACATGCGGTAGTGGCCACCATTCGCGATTAGCCCTAAAGCTGGCAGCTGGAACGTTGACAGA
CCTTTACTGTGGTTTTGTCACCTGTATGTTGGCTTATGGCTGTTCTCTGACCAACAGTTTGGTGAACCAAGTTGCTAATTTCTTACTTCGTTCTTGTCTACCGGTAAC
AGGGTACGCGGGTATTTGGCTACGTAAACGCAATGGCGGAATTAACCGCCCTGATTTATGTTCTTCCGCACTGATCATTAATCGCATCGGTGGGAAAAACCGCC
CTGCTGCTGGCTGGCACTATATGCTGCTGACTGATTTGGCTCATCTGCGCCACTCAGCGCTGGAAGTGGTATTCTGAAAAACCGTGCATATGTTGAAGTACCG
TTCTGCTGGTGGGCTGCTTTAAATATATACCAAGCAATGTTGAAGTGGGTTTTCAAGCGAGGATTATCTGGTCTGTTTCTGCTTCTTAAAGCAACTGGCGATTGATTT
TATGCTGACTGGCGGGCAATATGATGAAAGCATCGTTTCCAGGGCCCTTATGTTGGTGGGCTGCTGGTGGCGCTGGGCTCACCTTAATTTCCGTTGTTCCAGC
TTAGCGCGCCCGCCGCTTCCCTGCTGCGTGCAGGTGAATGAAGTGCCTTAAAGCAATCAATGTCGGATGCGCGCGGACCGCCCTATCCGACCAACATATCAT
AACGGAGTATCGCATTTGAACTGCGCAATGACCGAAAGATAAGAGCAGGCAAGCTAATTTACCGATATGTCGGAAGGCTTACCGGAAAAAGACTCTGTTGGGAAAA
GTTAATGATGAGTTAATCACTCGCATCCATCAGA
```

Two kinds of approaches :

- Alignments methods:

Aligning our sequence over a set of banks of known genes

- *Ab initio* methods:

Detection of signals in our sequence corresponding to a set of rules describing a gene

Bioinformatics solutions

```
GTGAAGGGCAATCAGCTGTTGCCGCTCTACTGGTGAAGAAAAAACCCCTGGCGCCAAATACGCAAAACCGCTCTCCCCGCGCTTGGCCGATTCAATATGC
AGCTGGCACGACAGGTTTCCCAGCTGGAAGACGGGGCAGTGAAGCCAAACCGCAATTAATGTGAGTTAGCTCACTTATAGGCACCCAGGCTTTACACTTATGCTTCC
GGCTCGTATGTTGTGGAAATTTGAGCGGATAACAATTTACACAGGAAGAACAGCTTACCATTGATACGGATTCACTGGCCGTCGTTTACAAACGTCGTGACTGGG
AAAAACCCCTGGCGTACCCCAACTTAATCGCCTTGACAGCATCCCCCTTTGCCGAGCTGGCCGTAATAGCGAAAGAGGCCCGCACCGATCGCCCTTCCCAACAGTTGGC
CAGCTCGAATGGCCGAATGGCCGTTTGGCTGTTTTCCGTCACCCAGCAAGCGAGTGGCCGAAAGCTGGCGTGGAGTGGCATTCTTCTGAGGGCCGATACTGCTCGTCCC
CTAAACTGGCAGATGCACGGTTACGATGCGCCCATACACCAACGTGACCTATCCATTACGGTCAATCCCGCTTTGTTCCACGGGAATCCGACGGGTTGTT
ACTCGCTCACATTTAATGTTGATGAAAGCTGGCTACAGGAAGGCCAGACGCGAATATTTTTGATGGCGTTAACTCGCCGTTTCACTGTGGTGAACCGGGCGCTGG
GTCGGTTACGGCCAGGACAGTCGTTTCCGCTCTGAAATTTGACCTGAGCGCATTTCACGCGCCGAGAAAAACCGCTCGCCGTTGATGGTGCCTGGCTGGAGTGACG
GCAGTTATCTGGAAGATCAGGATATGTTGGCGGATGAGCGGCATTTCCGTGACGTCTGCTTTGCTGCATAAACCCGACTACACAAATCAGCGATTTCCATGTTGCCACTC
GCTTAAATGATGATTTTACGCGCGCTGACTGGAAGCTGAAGTTAGATGTCGGCGAGTTCGCTGACTACCTACGGGTAACAGTTTCTTATGGCAGGGTGAACG
CAGGTCGCCAGCGGCACCCGCGCTTTCCGCGGTGAATATCGATGAGCGTGGTGGTTATGCCGATCGCGTACACTACGCTGAAACGTCGAAAACCCGAAACTGT
GGAGCGCCGAAATCCCGAATCTCTATCGTCCGGTGGTTGAACGTGCACACCGCCGACGGCAGCGTATTGAAGCAGAAGCCTGCCGATGTCGTTTCCGCGAGGTGC
GGATGAAAATGGTCTGCTGCTGCTGAACGGCAAGCCGTTGCTGATTCCAGGGCTTAAACGCTCACGAGCATCATCTCTGCATGGTCAGGTATGGATGAGCAGAC
GATGGTGCAGGATATCTGCTGATGAAGCAGAACAACTTTAACCGCTGGCTGTTCCGATTATCGAACCATCCGCTGGTGAACCGCTGCGGACCCGCTAGCGG
CTGATGTTGGTGAAGCAGCAATATTTGAAACCCAGCGCATGGTCCAAATGCTGACCCGATGTCGCGGCTGACCGGGATGAGCGCAACCGTAAACGC
GAATGGTGCAGCGCATGTAATCCCGAGTGTGATCATCTGGTCCGTTGGGAATGAATCAGGCCACGGCCGCTAATCAGACCGGCTGATCGCTGGATCAAACT
TGTCATCTTCCCGCCGTCAGTATGAAGCCGCGGAGCGCACCCAGCCAGCCATGATATTTGCCGATGACGCGCGCTGATGAAAGACAGCCCTT
CCCGGCTGAGCAGAAATGGTCCATCAAAAATGSCCTTTCGCTACCTGGAGAGACGGCCCGCTGATCCTTTGCCAATACGCCACCGATGGTAAACGATCTTGGC
GGTTTTCCGTAATACTGCGACGGCTTTCGTCAGTATCCCGCTTTACAGCCGGCTTCGCTGGGACTGGTGGATCAGTCGCTGATTAATATGATGAAAAACGGCAA
CCCGTGGTCCGCTTACGGCGGTGATTTTGGCGATACCGCGAACCGATCGCCAGTCTGTATGAACGGTCTGGCTTTTCCGAGCCGCAACCGCCGATCCAGCGCTGACG
GAAGCAAAACCCAGCAGCAGTTTTCCAGTTCGCTTTATCCGGCAAAACCGAATCGAATGACCAAGCAATACCTGTTCCGCTATAGCGATAACGAGCTCTGCACCTG
GATGGTGGCGCTGGATGGTAAAGCCGCTGGCAAGCGGTGAAGTCCCTGATGTCGCTCCACAAGGTAAACAGTTGATTGAACCTGCCTGAACTACCCGACGGCGGA
GAGGCGCCGGGCAACTCTGGCTACAGTACCGTATGTCAAACGACGCGACCGGATGGTTCAGAAAGCCGGGACATCAGCGCCTGGCAGCAGTGGCGCTCGGGG
AAAAACCTCAGTGTGACGCTTCCCGCCGCTCCACCCGATCCCGCATCCGACCCAGCGAAATGGATTTTGCATCGAGCTGGGTAATAGCGCTTGGCAATTTAAC
CGCCAGTCAGGCTTTCTTTCACAGATGGTGGTGGGCAAAAAAACAATGCTGACGCGCGCTGGCCGATCAGTTACCCGCTGCACCCGCTGGATAACGACATTTGGCT
AAGTGAAGCGACCCGCAATTGACCCCTAACCGCTGGCTGCAACCGCTGGAAGCGCGGCGCCATTACCAGGCCGAAGCAGCGTTTGGCAGTGCACGACATACACT
TGCTGATGCGGCTGATTACGACCGCTCAGCGGTGGCAGCATCAGGGGAAAACTTATTTATCAGCCGAAAAACCTACCGGATGATGGTATGGTCAAATGGCG
ATTACCGTTGATGTTGAAGTGGCGAGCGATACACCGCATCCGCGCGGATGGCTGAACTGCCAGTCCGAGGTAGCAGAGCGGGTAACTGGCTCGGATTA
GGGCGCAAGAAAACTATCCGACCGCCTTACTGCGCGCTGTTTTGACCGCTGGGATGTCGCCATTGTGAGACATGATACCCCGTACGCTTCCCGAGCGAAAAACG
GTCTGCGCTGCGGGACCGCGGAATTTGAATATGGCCACACCAAGTGGCGGGGACTTCCAGTTCAACACTCAGCCGCTACAGTCAACAGCAACTGATGGAACCCAG
CCATCGCCATCTGCTGCACGGGAAAGGACAGATGGCTGAATATCGACGTTTTCATATGGGGATTGGTGGCGAGCACTCCTGGAGCCGCTCAGTATCGGCGGAA
TTCAGCTGAGCGCCGGTGCCTAGCATTAACCAATGGTCTGGTGTCAAAAATTAATAAACCCGGGAGGCCATGCTCGCCGCTATTTCCGTAAGGAAATCCATTA
TACTATTAAAAAACATAAATTTGGATGTTGCGTTATGTTCTTTACTTTTATCAGTGGGAGCTACTTCCGCTTTTCCCGATTGGCTACATGACATCAAC
CATATCAGCAAAAGTACACGGATATTTTTGCGCGTATTTCTGCTGCTGCTATTTCCAAACCGCTTTTGGTCTGCTTTCTGCAAACTCGGGCTGGCAAACT
ACCTGCTGGGATATACCGCATGTAGTGTGTTTGGCGGCTTCTTATTTATCTTCCGGCCACTGTTACAATAACAATTTAGTAGGATGATGTTGGTGGT
ATTTACTAGCTTTTGTAAAGCCGCTGGCCAGCAGTAGAGCAATTTAAGCAAAAGTACGCGCTGCAGTAATTTGCAATTTGGCTGCGCGGATGTTTGGC
TGTGTTGGCTGGGCGCTGTGCTCGCTGATGTCGGCATCATGTTACCATCAATAATCAGTTGTTTCTGGCTGGGCTCGGCTGTCACATCATCTCGCGTTTTA
CTCTTTTTCCGCAAAACGGATGGCCCTTCTCAGCAGGTTGCCAATCGCGTAGGTGCCAACCATTCGGCATTTAGCCATTAAGCTGGCAGTGTTCAGACA
GCCAAAACCTGTGGTTTTGTCACTGTATGTTTGGCGTTTCTGCGACCATCAGATGTTTTCGACCAACAGTTTGGCTAATTTCTTACTTCGTTCTTGTCTACCGGTGAAC
AGGGTACCGCGGGATTTGGCTACGTAAACGACAAATGGGCGAATTAACCGCCCTGATTTATGTTCTTTGCCACACTGATCATTAACTCGCATCGGTGGGAAAAACCGC
CTGCTGCTGGCTGGCAATATGCTGCTGACGATATTTGGCTCATCTGTCGCCACCTCAGCCGCTGGAAGTGGTATTCTGAAAACCGCTGCATATGTTGAAGTACCG
TTCCTGCTGGTGGGCTGTTTTAAATATATCCAGCCGATTTGAAAGTGGGTTTTTCCAGCGAGGATTTATCTGGCTGTTTCTGCTCTTTTAAAGCAACTGGGCGATTTAT
TATGCTGTGATGGCGGCAATGATGAAAGCATCGGTTTTCCAGGCGGCTTATGTTGGTCTGGTCTGGTGGGCTGGGCTGAGCTTAAITTCGTTTTCAGCG
TTAGCGGCGCCCGCCGCTTTCCCTGCTGCGTGCAGTGAATGAAGTGCCTTAAAGCAATCAATGTCGGATGCGCGCGGAGCCGCTTATCCGACCAACATATCAT
AACGGAGTGTGCTGATTAAGCATGCAATGACCGAAAGATAAGAGCAGGCAAGCTAATTTACCGATGTGCGGAAGGCTTACCGGAAAAAGACTTCGTGGGAAAAAC
GTTAATGATGAGTTTAACTACTCGCATCCATCAGA
```

Two kinds of approaches :

- Alignments methods:

Aligning our sequence over data banks of known genes

⇒ unable to find new genes

- *Ab initio* methods:

Detection of signals in our sequence corresponding to a set of rules describing a gene

⇒ can miss exotic genes

Ab initio methods

```
GTGAAGGGCAATCAGCTGTTGCCGCTCTACTGGTGAAGAAAAAACCCCTGGCGCCAAATACGCAAAACCGCCTCTCCCCGCGGTTGGCCGATTCAATATGC
AGCTGGCACACAGGTTTCCCAGCTGGAAGACGGGGCAGTGAAGCCAAACGCAATTAATGTGAGTTAGCTCACTTATTAGGCACCCAGGCTTTACACTTTATGCTTCC
GGCTCGTATGTTGTGGAAATTTGAGCGGATAACAATTTACACAGGAAACAGCTTACCATTGATACGCGTACTGGCCGTCGTTTTACAAACGTCGTGACTCGGG
AAAACCCCTGGCGTTACCCAACTTAATCGCCTTGACAGCATCCCCCTTTCCGCGAGCTGGCGTAATAGCAGAAAGGCCCGCACCGATTCCGCCCTTCCCAACAGTTGGC
CAGCTGAAATGGCGAAATGGCCGCTTTGGCTCGCTTGGCCAGTACCCAGGAGCGGTCGCGGAAAGCTGGCTGGAAGTGGCATTCTTCCGAGCCGATACTGCTGCTGCC
CTAAACTGGCAGATGCACGGTTACGATGCGCCCATACACCAACGTGACCTATCCATTACGGTCAATCCCGCCTTTGTTCCACGGGAATCCGACGGGTTGT
ACTCGCTCACATTTAATGTTGATGAAAGCTGGCTACAGGAAGGCCAGACGCGAATATTTTTGATGGCGTTAACTCGCCGTTTCACTGTGGTCAACGGCGGCTGG
GTCGGTTACGGCCAGGACAGTCGTTTCCGCTGTAATTTGACCTGAGCGCATTTCACGCGCCGAGAAAAACCGCCTCGCCGTTGATGGTGCCTGGCTGGAGTGACG
GCAGTTATCTGGAAGATCAGGATATGTTGGCGGATGAGCGCCATTTCCGTGACGTCTGCTTTGCTGCATAAACCGACTACACAAATCAGCGATTCCATGTTGCCACTC
GCTTAATGATGATTTTACGCGCGCTGACTGAGGCTGAAGTTAGATGCGGGGAGTTCGCTGACTACCTACGGGTAACAGTTTCTTTATGGCAGGGTGAACCG
CAGGTCGCCAGCGGCACCGCCGCTTTCCGCGGTGAATTTATCGATGAGCGTGGTGGTTATCCGATCGCGTCAACACTACGCTGAAACGTCGAAAACCCGAAACTGT
GGAGCGCCGAAATCCGAATCTCTATCGTCCGGTGGTTGAACGTGCACACCGCCGACCGCACGCTGATTGAAGCAGAAAGCTGCGATGTCGTTTCCGCGAGGTCG
GGATTGAAAATGGTCTGCTGCTGCTGAAGCGCAAGCCGTTGCTGATTCAGGCGCTTAAACCGTCAAGGATCATCTCTGATGGTCAGGTATGGATGAGCAGAC
GATGGTGCAGGATATCTGCTGATGAAGCAGAAACAATTTAACGCCGTCGCTGATTCGCAATATCCGAACCATCCGCTGTGGTACACGCTGTGGCAGCCGTAGGGC
CTGATGTGGTGGATGAAGCCAAATTTGAAACCCAGCGCATGGTCCAAATGCTGACCGATGATCCGCGTGGCTACCGGGATGAGCGAACCGGTAACCG
GAATGGTCAGCGCGATCTAATCACCAGTGTGATCATCTGGTCCGTGGGAATGAATCAGGCCACGGCCGCTAATCAGACGCGCTGTATCGCTGGATCAATC
TGTGATCTTCCCGCCGGTGCAGTATGAAGCGCGGAGCGACACCCAGCATTAATTTGCCGATGACGCGCGGTGATGAAAGACCGCCCTT
CCCGCCTGTCGCAAAATGGTCCATCAAAAATGGCTTTCGCTACTGGAGAGACGGCCCGCTGATCTTTCGCAATACGCCACCGCATGGGTAACAGTCTTGGC
GGTTTCGTAATACTGGCAGCGGTTTCGTCAGTATCCCGCTTACAGCCGGCTTCGCTGGGACTGGTGGATCAGTCGCTGATTAATATGATGAAAACGGCAA
CCCGTGGTCCGCTTACGGCGGTGATTTTGGCGATACCGCCGAAACGATCGCCAGTCTGTATGAACGGTCTGGCTTTTGGCGACCGCAACCGCATCCAGCGCTGACG
GAAGCAAAACACCAGCAGCAGTTTTTCCAGTCCGTTTATCCGGCAAAACCATGGAATGACCCAGCGAATACCTGTTCCGTCATAGCGATAACGAGCTCTGCACCTG
GATGGTGGCGCTGGATGGAAGCCGCTGGCAAGCGGTGAAGTGCCTCTGATGTCGCTCCACAAGGTAACACGTTGATTGAACCTGCTGAACTACCCGACGCGGA
GAGGCCCGGGCAACTCTGGCTACAGTACCGTAGTGCAACCGAACCGCCGATGGTGCAGAAAGCCGGGCACATCAGCGCCTGGCAGCAGTGGCGCTCGGGCG
AAAACCTCAGTGTGACGCTTCCCGCCGCTCCACCCGATCCCGCATCGACACCCAGCAGAAATGGATTTTGCATCGAGTGGGTAATAGCGTTGGCAATTTAAC
CGCCAGTCAGGCTTTTTCACAGATGGTGGATGGGCATAAAAAAACAATGCTGACCGCCGCTGGCCGATCAGTTACCCGCTGCACCGCTGGATAACGCATTTGGCGT
AAGTGAAGCCAGCCGATTAACCGCTGGGTGCAAGCGTGGTGGAGCGGCTGGCCAGCCGATACCAGCCGGAAGCAGCGTTGTTGCAGTGCACGGCAGATACACT
TGCTGATGCGGCTGATTACGACCGCTACCGGTTGGCAGCATCAGGGGAAAACTTATTATCAGCCGGAAAACTACCGGATTGATGGTAGTGGTCAAATGGCG
ATTACCGTTGATGTTGAAGTGGCGAGCGATACCCGATCCGGCGGATGGCTGAACTGCCAGCTGGCGAGGTAGCAGAGCGGGTAAACTGGCTCGGATTA
GGCGCGCAAGAAAACTATCCGACCGCCTTACTGCGCCGCTTTTACCGCTGGGATGTCGCCATTGTCAGACATGATACCCCGTACGCTTCCCGAGCGAAAAACG
GTCTGCGCTGCGGGACGCGGGAATTAATATGGCCACACCAAGTGGCGGGGCACTCCAGTTCAACATCAGCCGCTACAGTCAACAGCAACTGATGAAAACCG
CCATCGCCATCTGCTGCACGGGAAGGACAGATGGCTGAATCGACGGTTTCCATATGGGGATGGTGGCGGACGACTCCTGGAGCCGTCAGTATCGGGCGAA
TTCAGCTGAGCGCCGGTGCCTACCATTAACCATTTGGTGTGGTGTCAAAAATAATAAACCCGGCAGGCCATGCTGCCCCGTTTCCGTAAGCAATCCATTGAG
TACTATTAATAAAACACAACCTTTGGATGTTCCGTTTATCTTTTCTTACTTTTATCAGTGGGAGCCTACTTCCCGTTTTCGCGATTGGCTACGATGACATAAC
CATATCAGCAAAAGTATACCGGTATTAATTTTGGCCGATTTCTGCTGCTGCTATTCACCCGCTGTTGGCTGCTTTCTGACAAACTCGGGCTGGCAAAAT
ACCTGCTGGGATATACCGCATGTAGTGTGTTTGGCCGCTTCTTATTTTATCTCCGGCCACTGTTACAATAACAATTTAGTAGGATCGATTGGTGGTGT
ATTTATCAGCTTTTATTTAACCGCGTGGCCAGCAGTAGAGCAATTAATGAAAGTCAAGCCGTCGCACTAATTTGCAATTTGGTCCGCGCGGATGTTGGC
TGTGTTGGCTGGCGCTGTGTCCTCGATGTCGGCATCATGTTACCATCAATAATCAGTTGTTTTCGCTGGCTGGGCTCGGCTGTCACATCATCTCGCCGTTTTA
CTCTTTTTCCGCAAAAGCGATGCGCCCTTCTGCCACGGTGGCAATCGGATGGTGGCAACCATCCGCTTAGCCTTTAGCCTGGCAGCTGGAACGTTGTGCAGCA
GCCAAACTGTGGTTTTTGTCACTGTATGTTATGGCGTTTCTGTGCACTACGATGTTTGGACCAACAGTTTGTCTAATTTCTTACTCGTTTCTTGTCAACCGGTGAAC
AGGGTACGCGGGTATTTGGCTACGTAAACGACAAATGGCGCAATTAACCGCCCTGATTTATGTTCTTTCGCGCACTGATCATTAATCGCATCGGTGGGAAAAACCGCC
CTGCTGCTGGCTGGCAGTATGTCGCTGACGATTTTGGCTCATCTGTCGCCACTCAGCCGCTGGAAGTGGTTATCTGAAAACCGCTGGCATATGTTGAAAGTACCG
TTCTGCTGGTGGGCTGCTTTAAATATAATACCAGCCAGTTGAAAGCGGTTTTACGGAGCATTATCTGGTCTGTTTCTGCTTCTTAAAGCAACTGGCGATTGATTT
TATGCTGACTGGCGGGCAATATGATGAAAGCATCGTTTCCAGGGCCCTTATGTTGGTGGTGGTCTGTTGGCGGCTGGGCTCACCTTAATTTCCGTGTTCCAGC
TTAGCGGCCCGCCGCTTCCCTGCTGCGTCGTCAGGTGAATGAAGTGGTAAAGCAATCAATGTCGGATGCGCGCGGACCGCCTATCCGACCAACATATCAT
AACGGAGTGTGCTGATTAACCATGCAATGACCGAAAAGATAAGAGCAGGCAAGCTATTACCGATGTGCGGAAAGGCTACCGGAAAAAGACCTCTGGGAAAAAC
GTTAATGATGAGTTTAACTACTCGCATCCATCAGA
```

2 main steps :

1) List Open Reading Frames (OFRs) over the 6 reading frames :

- starting with a start codon (mostly ATG)
- ending with a stop codon (TAA, TAG, TGA)

2) Select valid ORFs

Ab initio methods

```
GTGAAGGGCAATCAGCTGTTGCCGCTCTACTGGTGAAGAAAAAACCCCTGGCGCCAAATACGCAAAACCGCTCTCCCCGCGGTTGGCCGATTCAATATGC
AGCTGGCACACAGGTTTCCCAGCTGGAAGACGGGGCAGTGAAGCCAAACCGCAATTAATGTGAGTTAGCTCACTATTAGGCACCCAGGCTTTACACTTTATGCTTCC
GGCTCGTATGTTGTGGAAATTTGAGCGGATAACAATTTACACAGGAAACAGCTTACCATTGATACGCGTACTGGCCGTCGTTTACAAACGTCGTGACTGGG
AAAAACCTGGCGTACCCAACTTAATCGCTTGCAGACATCCCCCTTTCGGCAGCTGGCCGTAATAGCGAAAGAGCCCGCAGCGATCGCCCTTCCCAACAGTTGGC
CAGCCTGAATGGCGAAATGGCCGCTTGGCTGTTTGGCTGAGTACACCCAGGAGCGGTCGGCGAAAGCTGGCTGGAGTGGCATCTTCCGAGCCGATACTGCTGCTGCC
CTAAACTGGCAGATGCACGGTTACGATGCGCCCATACACCAACGTGACCTATCCATTACGGTCAATCCCGCTTGTGCCACGGGAATCCGACGGGTTGT
ACTCGCTCACATTTAATGTTGATGAAAGCTGGCTACAGGAAGGCCAGACGCGAATATTTTTGATGGCGTTAACTCGCCGTTTCACTGTGGTCAACGGCGCTGG
GTCGGTTACGGCCAGGACAGTCGTTTCCGCTGTAATTTGACCTGAGCGCATTTCACGCGCCGAGAAAAACCGCTCGCCGTTGATGGTGCCTGGCTGGAGTGACC
GCAGTTATCTGGAAGATCAGGATATGTTGGCGGATGAGCGCCATTTCCGTGACGTCTCGTTGCTGCATAAACCGACTACACAAATCAGCGATTCCATGTTGCCACTC
GCTTAATGATGATTTTACGCGCGCTGACTGAGGCTGAAGTTAGATGCGCGGAGTTGCGTGACTACCTACGGGTAACAGTTTCTTTATGGCAGGGTGAACG
CAGGTCGCCAGCGGACCCGCGCTTTCGGCGGTGAATTTATCGATGAGCGTGGTGGTTATCCGATCGCGTCAACACTACGCTGAAACGTCGAAAACCCGAAACTGT
GGAGCGCCGAAATCCGAATCTCTATCGTGGCGGTTGAACTGCACACCCGCGCAGCCAGCGTGATTGAAGCAGAAGCCTGCCGATGTCGTTTCCGCGAGGTTGC
GGATGAAAAATGGTCTGCTGCTGCTGAAGCGCAAGCCGTTGCTGATTCAGGCGCTTAAACCGTACAGGACATCATCTGATGGTCAGGTCATGGATGAGCAGAC
GATGGTGCAGGATATCTGCTGATGAAGCAGAAACAATTTAACGCCGTCGCTGTTCCGATTTACCGAACCATCCGCTGTGGTACACGCTGCGGACCCGTCAGGCC
CTGATGTTGGTGGATGAAGCCAAATTTGAAACCCAGGCACTGGCCAAATGCTGACCGATGATCCGCGTGGCTACCGGGGATGAGCGAACCGGTAACCG
GAATGGTCAGCGCATGTAATCACCAGTGTGATCATCTGGTCGCTGGGAATGAATCAGGCCACGGCCGCTAATCAGACGCGCTGATCGCTGGATCAATC
TGTGATCTTCCCGCCGGTGCAGTATGAAGCGCGGAGCGACACCCAGCCGATATTTTCCCGGATGACGCGCGCTGATGAGAACCCAGCCCTT
CCCGCTGTCCGAAATGGTCCATCAAAAATGCTTTCGCTACCTGGAGAGACGGCCCGCTGATCTTTCGCAATACGCCACCGCATGGGTAACAGCTTGGC
GGTTTCGTAATACTGGCAGCGGTTTCGTCAGTATCCCGCTTACAGGCGGCTTCGCTGGGACTGGTGGATCAGTCGCTGATTAATATGATGAAAAACGGCAA
CCCGTGGTCCGCTTACGGCGGTGATTTTGGCGATACCGCCGAACGATCGCCAGTCTGTATGAACGGTCTGGCTTTTCCGACCCGCAACCGCATCCAGCGCTGAG
GAAGCAAAACACCCAGCAGCATTTTTCCAGTTCGCTTATCCGGCAAAACCTGAAAGTACAGCGCAATACCTGTTCCGCTATAGCGATAACGAGCTCTGCATCT
GATGGTGGCGCTGGATGGTAAGCCGCTGGCAAGCGGTGAAGTGCCTCTGATGTCGCTCCACAAGGTAACAGTTGATTGAACCTGCCGTAACACTACCCGACGCGGA
GAGGCCCGGGCAACTCTGGCTACAGTACGCGTAGTGCAACCGAACGCGCCGATGGTGCAGAACGGGGCACATCAGCGCTGGCAGCAGTGGCGCTCGGGCG
AAAACTCAGTGTGACGCTTCCCGCCGCTCCACCCGCTCCGCGATCGACACCCAGCGAAATGGATTTTTGCATCGAGTGGGTAATAGCGTTGGCAATTTAAC
CGCCAGTCAGGCTTTCTTTCACAGATGGTGGTGGCATAAAAAAACAATGCTGACGCGCGCTGCGCGATCAGTTACCCGCTGCACCCGCTGGATAACGACATTTGGCGT
AAGTGAAGCCAGCCGATTAACCGCTGGGTGCAAGCGTGGGAGCGGCGGACCTACCAGCCGGAAGCAGCGTTGTTGCAGTGCACGGGAGATACAT
TGCTGATGCGGCTGATTACGACCGCTACCGGTTGGCAGCATCAGGGGAAAACTTATTATCAGCCGGAAAAACCTACCGGATTGATGGTAGTGGTCAAATGGCG
ATTACCGTTGATGTTGAAGTGGCGGAGCGGATACACCGCATCCGCGCGGATGGCTGAACTGCCAGCTGGCGAGGTAGCAGAGCGGGTAAACTGGCTCGGATTA
GGCGCGCAAGAAAAATATCCGACCGCCCTACTGCGCGCTGTTTTCGACCGTGGGATGTCGCCATGTCAGACATGATACCCCGTACGCTTCCCGAGCGAAAAACG
GTCTGCGCTGCGGACCGCGGAATTAATGTTGGCCACACAGTGGCGGGGACTTCCAGTTCAACATCAGCCGCTACAGTCAACAGCAACTGATGAAACCCAG
CCATCGCATCTGCTGCACGGGAAAGGACAGTGGCTGAATCGACGTTTCCATATGGGGATTGGTGGCGGACGACTCCTGGAGCCCGCTCAGTATCGGGCGGAA
TTCAGCTGAGCGCCGGTGCCTACCATTAACATTTGGTCTGGTGTCAAAAATTAATAAACCCGGGACGGCCATGTCGCGCGTATTTCCGTAAGGAAATCCATTATG
TACTATTAATAAAACATAAATTTGGATGTTCCGTTTATCTTTTCTTACTTTTATCAGTGGGAGGCTACTTCCCGTTTTCCCGGATTGGCTACGATGACATAAC
CATATCAGCAAAAGTACACGGTATTAATTTGCGGCTATTTCTGCTGCTGCTATTTCCAAACCGCTGTTGGTCTGCTTTCTGCAAACTCGGGCTGGCAGAAAT
ACCTGCTGGGATATACCCGATGTTAGTGTGTTGCGCGCTTCTTATTTTATCTTCGGGCCACTGTTACAATAACAATTTAGTAGGATGATGTTGGTGGT
ATTTATCAGCTTTTGTTTAACCGCGGTGCCGACGATGAGCAATTAATGAAAGTACAGCCGTCGCACTAATTTGCAATTTGGCTGGCGCGGATGTTGGC
TGTGTTGGCTGGCGCTGTGTCCTCGATGTCGGCATCATGTTACCATCAATAATCAGTTGTTTTCGCTGGCTGGGCTCGGCTGTCACATCATCTCGCGTTTTA
CTCTTTTTTCGCAAAACGGATGGCCCTCTTCCGACCGTGTGCCAATGCGGTAGTGGCAACCATTCGGCTATTAGCCTTAAAGCTGGCAGCTGGAACGTTGTGCAGCA
GCCAAAACCTGGTTTTTGTCACTGTATGTTATGGCGTTCCTGTGCGACTAGATGTTTGGACCAACAGTTTGTCTAATTTCTTACTCTGTTCTTGTACCGGTGAAC
AGGGTACGCGGGTATTTGGCTACGTAAACGACATGGCGGAATTACTTAAACCGCTGATGTTGTTTTCGCGCACTGATCATTAACTCGCATCGGTGGGAAAAACCGCC
TCTGCTGCTGGCTGGCATATATGCTGCTGACGATTTAGGCTCATCTGCGCACCTCAGCGCTGGAAGTGGTATTCTGAAAACCGCTGCATATGTTGAAAGTACCG
TTCTGCTGGTGGGCTGCTTTAAATATATACCAAGCCAGTTGAAAGCGGTTTTTCAAGCGGACTTATCTGGTCTGTTTCTGCTTCTTAAAGCAACTGGCGATTGATTT
TATGCTGACTGGCGGGCAATATGATGAAAGCATCGTTTCCAGGGCCGCTTATGTTGGTCTGGGCTGCTGGTGGCGGCTGGGCTTACCTTAAATTCGTTCCGTTCCAGC
TTAGCGCGCCCGCCGCTTTCCTGCTGCGTGCAGTGAATGAAGTGCCTTAAAGCAATCAATGTCGGATGCGCGCGGACGGCCCTATCCGACCAACATATCAT
AACGGAGTGTGCTGATTTGAAATGCGCAATGACCGAAAGATAAGAGCAGGCAAGCTATTACCGATGTGCGGAAAGGCTTACCGGAAAAAGACTCTGGGAAAAAC
GTTAATGATGAGTTTAACTACTCGCATCCATCAGA
```

2 mains steps :

1) List Open Reading Frames (OFRs) over the 6 reading frames :

- starting with a start codon (mostly ATG)
- ending with a stop codon (TAA, TAG, TGA)

2) Select valid ORFs

Example of ORF finding* : phase +1

GTGAAGGGCAATCAGCTGTTGCCCGTCTCACTGGTGAAGAAAAAACCCCTGGCGCCCAATACGAAACCGCCTCTCCCCGGCGTTGGCCGATTCAATATGC
 AGCTGGCACAGAGGTTTCCCAGCTGGAAGACGGGGCAGTGTAGCGCCAAACCGAATTAATGTGAGTTAGCTCACTATTAGGCACCCAGGCTTTACACTTTATGCTTCC
 GGCTCGTATGTTGTGGAAATTTGAGCGGATAACAATTTACACAGGAAACAGCTATGACCATGATACGGATTCACTGGCCGTCGTTTACAAACGTCGTGACTCGGG
 AAAACCCCTGGCGTTACCCAACCTAATCGCCTTGCCAGCACAATCCCTCTGGCCAGCTGGCGTAATAGCGAAAGAGCGCCGACGATCGCCCTTCCCAACAGATTGGC
 CAGCCTGAATGGCAATGGCGCTTTGCCCTGGTTTCCCGCACCAAGACGGCTGGCGGAAAGCTGGCTGGAGTGGCATCTTCTGAGGGCGATAGCTGCTGCTGGC
 CTAAACCTGGCAGATGCACGGTTACGATGCGCCCATACACCAACGTCCTATCCCAATACGGTCAATCCGCCGTTTGTCCACGGGAATCCGACGGGGTTGT
 ACTCGCTCAACATTAATGTTGATGAAAGCTGGCTACAGGAAGGCCAGCGAATTAATTTTATGAGCGTTAACTCGCGCTTTCATCTGGTGCACACGGGGCGCTGC
 GTCGGTTACGGCCAGGACAGTCGTTGCCCTCTGAATTTGACCTGCGCATTTCACGGCCGAGAAAAACCGCCTCGCGGTGATGGTGGCTGGGCTGGAGTGACG
 GCAGTTATCTGGAAGATCAGGATATGTGGCGGATGAGCGGCATTTCCGTGACGTCTCGTTGCTGCATAAACCCGACTACACAAATCAGCGATTTCCTGTTGCCACTC
 SCTTTAATGATGATTTACGCCCGCTGTACTGGAGGCTGAAGTTCAGATGTGGCGGAGTTCGCTGACTACCTACCGGTAACAGTTTCTTATGGCAGGGTGAACG
 CAGGTCGCCAGCGCACCGCGCCTTTCCGGCGTGAATATTCGATGAGCGTGGTGGTTATCCGATCGCGTACACACTACGTCGAAACGTCGAAAACCCGAAACTGT
 GAGGCGCCGAAATCCGAATCTCTATCGTGGCGTGGTTGAACGTACACCGCCGACGGCCAGCGTATTGAAGCAGAAACGCTGCGATGTCGGTTCCCGCAGGTTGC
 GGATGAAAATGGTCTGCTGCTGCTGAACGGCAAGCCGTTGCTGATTCGAGGCGTTAACCGTACGAGCATCATCTCTGATGGTCAAGTCAATGGATGAGCAGAC
 GATGGTCAGGATATCTGCTGATGAAGCAGAAACACTTTAACGCCGTGGCCTGTTCCGATATCCGAACCATCCGCTGGTGAACAGCTGTGGACCCGCTAGCGG
 CTGATGTGGTGGATGAAGCCAAATTTGAAACCCAGCCGATGGCCAAATGCTGACCCGATGATCCGGCTGGCTAACCAGGATGAGCGAACCGTAAACCG
 GAATGGTCAGCGCATCTAATCACCAGTGTGATCATCTGGTCCGTGGGGAATGATCAGCCAGCCGCTAATCAGCAGCCGCTGTATCGCTGGATCAATAT
 TGTCCGCTCTCCCGCCGTCAGATGAAGCGCGCGAGCCGACACCGACCCGATATTTTGGCCGATGACGCGCGCTGGATGAAGACAGCCGCTT
 CCGCGCTGTCCGAAATGGTCCATGAAATAAGCTTTGGCTACCTGGAGAGACGGCCCGCTCCCTTCGGAATACGCCACGCGATGGTGAACAGCTTTGGC
 GGTTCGTAATAACTGCGAGCGGTTTCGTCAGTATCCCGCTTTACAGGCGCGCTTCGCTGGGACTGGTGGATCAGTCGCTGATTAATATGATAAAACGGCA
 CCGCTGGCTCGCTTACGGCGGTGATTTGGCGAATACCGCAACCTGCGCAGTCTGTGAACGCTGTCGTTTCCCGACCGCACCGCCGATCCAGCGCTGACG
 GAAGCAAAACACAGCAGCAGTTTCCAGTTCGCTTATCCGGCAACCACTCGAATGACCCAGCGAATACCTGTTCCGCTCATAGCGATAACGAGCTCCCTGCACCT
 GATGGTGGCGCTGGATGGTAAGCCGCTGGCAAGCGGTGAAGTGCCTCTGGATCTGCTCTCCACAAAGGTAACAGTTGATTGAACCTGCCTGAACTCCGACCGGGA
 GAGGCCCGGGCAACTCTGGCTCACAGTACGGTAGTGAACCGCAACCGACCGGATGGTTCAGAAAGCCGGGACATCAGCGCCTGGCAGCAGTGGCGCTCGGGG
 AAAACCTCAGTGTGACGCTCCCGCGCGCTCCACCCGATCCCGCATCGACACCAAGCGAAATGGATTTTGCATCGGTAATAGCGTTTGGCAATTTAAC
 CGCCAGTCAGGCTTTCTTTCAGATGTTGGATGGGCAATAAAAAACAATGCTGACGCGCGCTGGCCGATCAGTTACCCGCTGCACCCGCTGGATAACGACATTTGGCT
 AAGTGAAGCGACCCGCAATTGACCCCTAACCGCTGGGTCGAAAGCTGGGAGGCGCCATTACCAGCCGAAAGCAGCGTTTGTGAGTGCACAGACATACACT
 TCGTATGCGGCTGCTGATTACGACCGCTCACCGGTGGCAGCATCAGGGGAAAAAATCTTATTATCAGCCGAAAAACCTACCGGATTCATGCTGATGGTCAAATGGCC
 ATTCAGCTTGTGATTTGAAGTGGCGAGCGGATACCCGCTACCGCGCGGATTTGGCCCTGACTGCCAGCTGGCGCAGGTAGCAGAGCGGGTAACTGGCTCGGATTA
 GGGCGCAAGAAAAATATCCGACCGCCTACTGCGCGCTGTTTGAACCGCTGGGATCTGCCATTGTGAGACATGATACCCCGTACGCTTCCCGAGCGAAAAACG
 GTCTGCGCTGCGGACCGCGGAATTAATATGGCCACACAGTGGCGGGGGACTTCCAGTTCAACATCAGCCGCTACAGTCAACAGCAACTGATGAAACCCAG
 CCATCGCCATCTGCTGACCGCGAAGAAAGCCAGATGGCTGAATATCGACGGTTTCAATATGGGGATGGTGGCGCAGACTCCTGGAGCCGCTCAGTATCGCGGAA
 TTCAGCTGAGCGCCGGTGCCTGACTTACCAATGGCTGTGGTGTCAAAAATAATAAACCCGGCAGGCCATGCTGCCCCATTTCCGCTAAGGAAATCCATTA
 TACTATTAATAAAACCAAACTTTGGATGTTCCGTTTATCTTTTCTTACTTTTATATCGGGAGGCTACTTCCCGTTTTCGCGATTGGCTACATGACATCAAC
 CATATCAGCAAAAGTATACCGGATTAATTTTCCCGCATCTCTGTTTCTGTTTCCGCTGATTTCCAAACCGCTGTTTGGCTGCTGTTCTGACAAAACCTGGGCTGCGCA
 ACCGCTGCTGGATATACCGGATGATGATGATGTTTGGCCGTTCTTATTTATTTCCGGCCACTGTTACAATAACAATTTAGTAGGATCGATTGTTGGTGGT
 ATTTTCTAGCTTTTGTTTTAAACCGCGTGCGCCAGAGTGAAGCAATTAATGAAAGTCAAGCCGTCGCACTAATTCGAATTTGGCTCGCGCGGATTTGGC
 TGTGTTGGCTGGCGCTGTGCTCCGATGTCGGCATCATGTTACCATCAATAATCAGTTGTTTCTGGCTGGGCTCTGGCTGTGCACTCATCTCGCGTTTTA
 CTCTTTTTCCGCAAAACGGATGCGCCCTTCTGCCACGGTGTCCAAATGGCGATGGTGCACCACTTCGGCTTAGCCCTAAGCTGGACATGGAACGTTGCAGACA
 GCAAAAACCTGGTTTTTGTCACTGTGGTGGGCTTTCCTGACACCTAGGATTTTGGACCAACAGTTTGTCTTCTTACTTCGTTCTTCTGACCGGTTGAA
 ACGGCTACCGGGTATTTGGCTACCTGACACATGGGCGAATCTCTAACCGCTCGATTGTTCTTTTGGCCACTGATCATAATCGCATCGGTGGGAAAAACCGCC
 CTGCTGCTGGCTGGCATATATGCTGTCAGTATTTGGCTTACTTTCGCCACCTCAGCCCTGGAAGTGGTTATCTGAAAACCGTGCATATGTTGAAGTACCG
 TTCTGCTGGTGGGCTGCTTTAAATATATACCAAGCCAGTTGAAAGCGGTTTTCAGCGGACTTATCTGGTCTGTTTCTGCTTCTTAAAGCACTGGCGGATGATTT
 TATGCTGACTGGCGGGCAATATGATGAAAGCATCGTTTCCAGGCGGCTTATCTGGTGGGCTGCTGGTGGCGCTGGGCTTACCTTAACTTCCGTTGTTCCAGC
 TTAGCGGCCCGCGCCGCTTTCCCTGCTGCGTGCAGTGAATGAAGTGCCTTAAAGCAATCAATGTGCGATGCGCGCGGACCGCTTATCCGACCAACATATCAT
 AACGGAGTGCATGCAATTAAGCATGCCAATGACCGAAAGATAAGAGCAGGCAAGCTATTACCGATGTGCGAAGGCTTACCGGAAAAAGACTCTCGTGGGAAAA
 GTTAATGATGAGTTAATCACTCGCATCCATCAGA

- When several start codons are available, the longest ORF is reported
- Here : 8 ORFs found

* to simplify, only ORFs starting with an ATG start codon are shown

Example of ORF finding : phase +2

- 3 ORFs found

GTGAAGGGCAATCAGCTGTTGCCCGTCTCACTGGTAAAAGAAAAACCACCTGGCGCCCAATACGAAACCGCCTCTCCCGCGCGTTGGCCGATTCAATATGC
AGCTGGCACACAGGTTTCCCAGCTGAAAAGCGGGCAGTGAAGCCAAACGCAATTAATGTAGTTAGCTCACTTATAGGCACCCAGGCTTTACACCTTATGCTTCC
GGCTCGTATGTTGTGGAAATTTGAGCCGGATAACAATTTACACAGGAAACAGTATGACCATGATACGGATTCACTGGCCGTCTGTTTACAAACGTCGTGACTGGG
AAAACCCCTGGCGTACCCAACTTAATCGCCTGGACGACATCCCCCTTTCGGCAGCTGGCGTAAATAGCGAAAGAGGCCCGCACCGGATCGCCCTTCCCAACAGTTGGC
CAGCCTGAATGGCGCAATGGCCCTTTGGCTGTTTCCGCTCCAGCAAGGCGGTCGGGAAAGCTGGCGTGGAGTGGCATCTTCCGAGGCCGATACTCGTCGTCC
CTAAACTGGCAGATGCACGGTTACGATGCGCCCATACACCAACGTGACCTATCCATTACGGTCAATCCGCGCTTGTCCACGGAGAATCCGACGGGTTGT
ACTCGCTACATTTAATGTTGATGAAAAGCTGGCTACAGGAAAGCCAGACGCGAATATTTTTGATGGCGTTAACTCGCGCTTTCATCTGTGGTGAACCGGGCGCTGG
GTCGGTTACGGCCAGGACAGTCGTTTCCGCTCTGAAATTTGACCTGAGCGCATTTCACGCGCCGAGAAAAACCGCCTCGCCGTTGATGGTGTGGCTGGGATGACG
GCAGTTATCTGGAAGATCAGGATATGTTGGCGGATGAGCGGCATTTCCGTGACGTCTGCTTGTGCATAAACCGACTACACAAATCAGCGATTTCCATGTTGCCACT
GCTTTAATGATGATTTTACGCGCGCTGACTGGAAGCTGAAGTTAGATGCGCGGAGTTGCGTGACTACCTACGGGTAACAGTTTCTTTATGGCAGGGTGAACG
CAGGTCGCCAGCGGCACCCGCGCTTTCCGCGGTGAAATATCGATGAGCTGGTGGTTATGCCGATCGCGTACACACTACGCTGAAACGTCGAAAACCCGAAACTG
GAGGCGCCGAAATCCGAACTCTATCGTGGCGGTTGACTGCACACCCGACGCGCACGCTGATTGAAGCAGAAGCCTGCGATGTGCGTTTCCGCGAGGTTGC
GGATTGAAAATGGTCTGCTGCTGCTGAACGGCAAGCCGTTGACTTCCGAGGCTTAAACCGTACGAGCATCATCTCTGATGTCAGGTCAGGTCATGGATGAGCAGAC
GATGGTGCAGGATATCTGCTGATGAAGCAGAAACAATTTAACCGCTGGCCTGTTCCGATTATCCGAACCATCCGCTGTTGGTACACGCTGTCGGACCCGTAGCGC
CTGATGTGGTGGATGAAGCCAAATTTGAAACCCAGCGGTTGGTCCAAATGCTGACCCGATGTCGGCTGGTGGTACCGGGGATGAGCGAAACCGTAAACG
GAATGGTGCAGCGCATGTAATCACCAGTGTGATCATCTGGTCCGTCGGGAATGAATCAGGCCACGGCGCTAATCAGACGCGCTGATCGCTGGATCAAACT
TGTGCATCTTCCCGCCGGTGCAGTATGAAGCGCGGAGCGACCCGACCCAGATATTTTCCCGCATGACGCGCGCGTGGATGAAGACCGAACTT
CCCGCTGTGCCGAAAGTGCATCAAAAATGGCTTTGGCTAGCTGGAGAGCGGGCCGCTGATCGTTTGGCAATACGCCACGGCATGGGACAGTCTTGGC
GGTTTCGTAATACTGGCAGCGCTTTCGTCAGATCCCGCTTACAGGGCGGCTTCGCTGGGACTGGGTGGATCAGTCGCTGATTAATATGATGAAAACGGCAA
CCCGTGGTGGCTTACGGCGGTGATTTTGGCGATACCGCCGAAACGATCTGCTGATGAACGGTCTGGCTTTTGGCGACCCGACCCGATCCAGCGCTGACG
GAAGCAAAACACCCAGCAGCAGTTTCCAGTTCGCTTATCCGGCAAAACCATGGAATGACCCAGCGAATACCTGTTCCGCTATAGCGATAACGAGCTCCTGCACCT
GATGGTGGCGCTGGATGGTAAGCCGCTGGCAAGCGGTGAAGTGCCTCTGGATGTCCTGCCACCAAGGTAACAGTTGATTGAACCTGCCTGAACTACCCGACGGCA
GAGGCCCGGGCAACTCTGGCTCACAGTACGGCTAGTGCACCAACCGACCGGACCGCTGGTGCAGAAGCCGGGACATCAGCGCTGGCAGCAGTGGCGTCTGGCG
AAAACCTCAGTGTGACGCTTCCCGCGCGCTCCGACCATCCCGCATCTGACCCAGCCGCAAAATGGATTTTGCATCGAGCTGGGATAAGCGTTGGCAATTTAAC
CGCCAGTCAGGCTTTCTTTCACAGATGGATTTGGCATAAAAAACAATGCTGACGCGCGCTGGCGATCAGTTACCCGTCGACCCGCTGGATAACGACATTTGGCT
AAGTGAAGCGACCCCGATTGACCCTAAAGCCCTGGGTGCAAGCTGGGAGGCGCTGGAAGGCGCGGCGCCATTACCAGGCCGAAAGCAGCGTTGTCAGTACGGACATAC
TGCTGATGGCGTCTGATTACGACCGCTCACCGCTGGCAGCATCAGGGGAAAACTTATTTATCAGCCGAAAAACCTACCGGATGATGGTATGGTCAAATGGCG
ATTACCGTTGATGTTGAAGTGGCGGAGCGGATACCCGATCCGCGCGGATGGCTGAACTGCCAGCTGGCGCAGGTAGCAGAGCGGGTAAACTCGGATTA
GGCGGCAAGAAAATATCCGACCGCTTACTGCCGCTGTTTTGACCGCTGGGATCTGCCATTGTGACAGATGATACCCCGTACGCTTCCCGAGCGAAAAACG
GTCTGGCTGCGGGACCGCGGAATGAATATGGCCACACCACTGGCGGGGACTTCCAGTCAACATCAGCCGCTACAGTCAACAGCACTGATGAAAACCG
CCATCGCCATCTGCTGCACGGGAAAGGAGCACATGGCTGAATATCGACGGTTTCCATATGGGGATGGTGGCGGACGACTCTGGAGCCGCTCAGTATCGCGGAA
TTCCAGCTGACCGCGGCTCGCTACCATACCAATTTGGCTGGTGAACCAATATAACCGGGCAGGCCATGCTGCCGCTATTTCGCGGATTAAGGAAATCCATATG
TACTATTAAAAAACCAAACTTTGGATGTTCCGTTATCTTTTCTTACTTTTATGATGGGAGGCTACTTCCCGTTTCCCGATTGGCTACAATGCATCAAC
CATATCAGCAAAAGTACGCGGATATTTTTGCCGCTATTTCTGCTGCTGATTTCCAAACCGCTGTTGGTCTGCTTCTGACAAAACCTGGCGTGGCGGAT
ACCTGCTGGATTTACCGCATGTTAGTGTGATGTTTGGCCGCTTCTTATTTTACTTCCGGCCACTGTTACAATACAACATTTAGTAGGATGATGTTGGTGGT
ATTTATCAGCTTTTGGTTTAAAGCCGCTGGCCACGATAGAGCAATTTAGTGAAGTCAAGCCGCTGCGAGTAATTTGCAATTTGGTTCGCGCGCGGATTTGGC
TGTGTTGGCTGGCGCTGTGCTCGATGTCGGCATCATGTTCCACATCAATATCAGTTGTTTCTGGCTGGGCTCGGCTGTGCATCATCTCGCGTTTTA
CTCTTTTTCCGAAAACGGATGCGCCCTTCTGCCAGGTTGCCAATGCGGTAGGTGCCAACCATTCGGCTTAGCCCTTAAAGTGGCAGCTGGAACCTGTTGAGAA
GCCAAAACCTGTGGTTTTGTCAGCTGTATGTTATGGCTGTTCTGCGACCATACGATGTTTTGACCAACAGTTTGTCTAATTTCTTACTCGTTCTTTGCTACCGGTAAAC
AGGGTACGCGGGTATTTGGCTACGTAAACGACAAATGGCGGAATTACTTAAACCGCTGATGTTGTTCTTGGCCACTGATCATTAATCGCATCGGTGGGAAAAACCGCC
CTGCTGCTGGCTGGCATAITATGCTGTGACTGATTTTGGCTCATCTGTCGCCACTCAGCCCTGGAAGTGGTTATCTGAAAACCGTGCATATGTTGAAAGTACCG
TTCTGCTGGTGGGCTGCTTTAAATATATACCAGCCAGTTTGAAGTGGGTTTTTCAAGCAGGATTTATCTGGTCTGTTTCTGCTTCTTAAAGCACTGGCGGATGATTT
TATGCTGACTGGCGGGCAATATGATGAAAACGATCGGTTTCCAGGCGGCTTATGTTGGTGGGCTGGGCTGCTGGGCTTACCTTAATTTCCGTTGTTCCAGC
TTAGCGGCCCGCGCCGCTTTCCCTGCTGCGTGTGAGTGAATGAAGTGGCTTAAAGCAATCAATGTCGGATGCGCGCGGACCGGCTTATCCGACCAACATATCAT
AACGGAGTGTGATGATTTGAACTGCGCAATGACCGAAAAGATAAGAGCAGGCAAGCTATTTCCGATATGTCGGAAGGCTTACCGGAAAAAAGACTCTGCGGAAAA
GTTAATGATGAGTTAATCACTCGCATCCATCAGA

Example of ORF finding : phase -2

- 7 ORFs found

GTGAAGGGCAATCAGCTGTTGCCCGTCTACTGGTAAAAGAAAAACCACCTGGCGCCCAATACGAAACCGCCTCTCCCGCGCGTTGGCCGATTCAATATGC
AGCTGGCACACAGGTTTCCCAGCTGAAAGCGGGCAGTGGAGCCAAACCGCAATTAATGTAGTTAGCTCACTTATAGGCACCCAGGCTTTACACTTTATGCTTCC
GGCTCGTATGTTGTGGAAATGTGAGCCGGATAACAATTCACACAGGAAGAACAGCTTACACCATGATACGGATTCACTGGCCGTCGTTTACAAACGTCGTGACTGGG
AAAACCCCTGGCGTACCCAACTTAATCGCCTGGACGACATCCCCCTTTCGGCAGCTGGCCGTAAATAGCGAAAGAGGCCCGCCAGCCGATCGCCCTTCCCAACAGTTGGC
CAGCCTGAAATGGCGCAATGGCCGTTTGGCTGTTTGGCCGATCCAGCCAGCAAGCGGAGCGCGGAAAGCTGGCGTGGAGTGGCATCTTCCGAGGCCGATACTGCTCGTCC
CTAAACTGGCAGATGCACGGTTACGATGCGCCCATACACCAACGTGACCTATCCATACGGTCAATCCCGCGTTTGTCCACGGAGAATCCGACGGGTTGT
ACTCGCTACACTTAATGTTGATGAAAGCTGGCTACAGGAAGGCCAGACGCGAATATTTTTGATGGCGTTAACTCGCGCTTTCATCTGGTGGCAACGGCGCTGG
GTCGGTACGGCCAGGACAGTCGTTTCCGCTGTAATTTGACCTGAGCGCATTTTACGCGCCGGAGAAAACCGCCTCGCCGTTGATGGTGTGCGCTGGAGTGACG
GCAGTTATCTGGAAGATCAGGATATGTTGGCGGATGAGCGGCATTTCCTGTCAGCTGCTGTTGCTGCATAAACCGACTACACAAATCAGCGATTTCAGTTGTTGCCACT
GCTTTAATGATGATTTTACGCGCGCTGTACTGAGGCTGAAGTTAGATGTCGGCGGAGTTGCGTGACTACCTACGGGTAACAGTTTCTTTATGGCAGGGTGAACG
CAGGTCGCCAGCGGCACCCGCGCTTTCGGCGGTGAAATTCATCGATGAGCGTGGTGGTTATGCCGATCGCGTACAACGCTCTGAACCTCGAAAACCCGAAACTG
GGAGCGCCGAAATCCCAGTCTCTATCGTCCGCTGGTTGAACTGCACACCCCGCAGCCAGCGCTGATTGAAGCAGAAAGCCTGCGGATGTCGTTTCCGCGAGGTTGC
GGATTGAAAATGGTCTGCTGCTGCTGAACGGCAAGCGCTGCTGATTGAGGCGTTAACCGTACAGGATCATCTCTGATGGTCAGGTGATGGATGAGCAGAG
GATGGTGCAGGATATCTGCTGATGAAGCAGAACAACTTAAACCGCTGGCCGCTGTCGCAATATCCGAACCAACGCTGTGGTACACCGCTGTGGACCCGCTACGGC
CTGATGTTGGTGGATGAAGCCAAATTTGAAACCCAGCCGATGGCCAAATGATCGTCTGACCGATATCCCGGCTGGTGGTGAACCGGATGAGCGAAACCGTAAACG
GAATGGTCAGCGCGATGTAATCACCAGAGTGTGATCATCTGGTCCGCGGAATGAATCAGGCCAGCCGCTAATCAGACCGCTGTATCGCTGGATCAAACT
TGTCTATCTTCCCGCCGGTGCAGTATGAAGCGCGGAGCGACACCCAGCCAGCCAGCCAGCCAGCCAGCCAGCCAGCCAGCCAGCCAGCCAGCCAGCCAGCC
CCCGCGTGTCCGGAATGTTCCATCAAAAATGGCTTTCGCTACTGGAGAGACGGCCCGCTGATCTTTCGGAATACGCCACCGCATGGGTAACAGTCTTGGC
GGTTTCGCTAAACTGCGAGCGTTCGTCAGTATCCCGCTTACGAGCGCTCGCTGGGACTGGTGGATCAGTCGCTGATTAATATGATGAAAACGGCAAC
CCCGTGGCTCGCTACGCGCGTGAATTTGGCCGATACGCCAACGATCGCCAGTCTGCTGATGAACGCTGCTGGCTTTTCCGAGCCGCAAGCCGCGCTCCAGCGCTGACG
GAAGCAAAACACAGCAGCAGTTTTCCAGTTCCTGTTATCCGGCAACCACTCGAAGTGAACAGCGAAATACCTGTTCCGCTCATAGCGATAACGAGCTCCTGCACGT
GATGGTGGCGCTGGATGGTAAGCCGCTGGCAAGCGGTGAAGTGCCTCTGGATGTCCTGCTGCTCCACAAGGTAACAGTTGATTGAACCTGCCTGAACTACCGACCGCGGA
GAGGCCCGCCGCAACTCTGGCTCACAGTACCGCTAGTGGCAACCGCAACCGGACCGGATGGTCCGAAAGCCGCGGACATCAGCGCCTGGCAGCAGTGGCGTCTGGCGG
AAAACCTCAGTGTGACGCTCCCGCCGCGCTCCACCGCCATCCCGCACTGCAACCCAGCGAAATGGATTTTTGCATCGAGCTGGGTAATAAGCGTTGGCAATTTAAC
CGCCAGTCAGGCTTTTCTTACAGATGTGGATGGGCAATAAAAAACAATGCTGACGCGCGCTGGCCGATCAGTTACCCCGTGCACCGCTGGATAACGACATTTGGCGT
AAGTGAAGCGACCCCGCATTGACCCCTAACCGCTGGGTGCAAGCCTGGAAAGCGCGGCGCCATACCAGCCGGAAGCAGCGTGTGTCAGTGCACCGGACATACACT
TCTGATGGCGGTGCTGATTACGACCGCTCAGCGTGGCAGCAAGGGGAAAAACCTTATTTATCAGCCGGAACCACTACCGGATGATGGTGGTGGTCAAAATGGCC
ATACCGTTGATGTTGAAGTGGCGGAGCGGATACCCGCAACCGCGCGCGGATGGGCTGAACGCGGATGCGGATGCGGATGCGGATGCGGATGCGGATGCGGATG
GGCCGCAAGAAAATATCCGACCGCTTACTGCGCGCTGTTTGAACCGCTGGATGCTGCCATTGTCAGACATGATACCCCGTACGCTTCCCGAGCGAAAAACG
GTCTGCGCTGCGGGACGCGCGAATGAAATATGGCCACACCGCACTGGCGGGGCACTTCCAGTTCAACATCAGCCGCTACAGTCAACGCAACTGATGAAAACCG
CCATCGCCTCTGCTGACGCGGGAAGGCGCAATGGCTGAATATCGACGGTTTCCAGTATGGGGATTGGTGGCGACGACTCCTGGAGCCGCTCAGTATCGCCGGAAT
TTCAGCTGAGCCCGCGTGCCTACCATACCAATTTGGCTGTTGCTGTTCAAAAATAAATAAACCCGGGCAAGCCATGCTGCGCCGATTTTCGCGTAAGGAAATCCGATG
TACTATTAAAAAACAACAATTTGGATGTTCCGTTTATCTTTTCTTACTTTTATCAGTGGGAGCCTACTTCCCGTTTTCGCGATTGGCTAGCATGACATCAAC
CATATCAGCAAAAGTATACCGGATATTTTTGCGCGTATTTCTGTTGCTGCTTATCCAAACCGCTGTTGGTCTGCTTCTGACAAAACCTGGGCTGCGCAAA
ACCTGCTGGATATACCGCGATGTAGTGTGTTGGCCGCTTCTTATTTATCTTCCGGCCACTGTTACAATAACAATTTAGTAGGATGATGTTGGTGGT
ATTTATCAGCTTTTGGTTTAAACCGCGTGGCCAGCAGTACAGGCAATTTAAGGAAAGTACAGCCGCTGCAGTAATTTGCAATTTGGTGGCGCGCGGATTTGGC
TGTGTTGGCTGGCGCTGTGCTCCTGATGTCGGCATCATGTTACCATCAATAATCAGTTGTTTCTGGCTGGGCTGCTGGCTGTCACATCACTCGCCGTTTTA
CTCTTTTTTCGCAAAACGGATGCGCCCTTCTGCCAGCTTGGCAATGCGGTAGTGGCCACCATTCCGCTTTAGCCCTTAAAGCTGGCAGCTGGAACCTGTGAGCA
GCCAAAACCTGGGTTTTTGTCACTGTATGTTATGGCGTTTCTGCGACCATACGATGTTTGTGACCAACAGTTTGTCTAATTTCTTACTCTGTTCTTGTCTACCGGTAAAC
AGGGTACGCGGGATTTGGCTACGTAAACGACAAATGGCGCAATTAACCGCCCTGATTTATGTTCTTTCGCGCACTGATCATTAATCGCATCGGTGGGAAAAACCGCC
TCTGCTGGCTGGCAGCATATGTCGCTGACGATTTTGGCTCATCTGTCGCCACTCAGCCCTGGAAGTGGTTATCTGAAAACCGTGCATATGTTGAAAGTACCG
TTCTGCTGGTGGGCTGCTTTAAATATATACCAGCCAGTTGAAAGTGGGTTTTTCAAGCGAGTATCTGGTCTGTTTCTGCTTCTTAAAGCACTGGCGATGATTT
TATGCTGACTGGCGGGCAATATGATGAAAGCATCGGTTTCCAGGCGCTTATCTGGTGGTGGTCTGGTGGCGCTGCGCTTACCTTAAATTTCCGTTCTCACCG
TTAGCGGCCCGCGCCGCTTTCCTGCTGCGTGCAGTGAATGAAGTGCCTTAAAGCAATCAATGTCGGATGCGCGCGGAGCGCTTATCCGACCAACATATCAT
AACGGAGTGCATGCAATTTGAACTGCGCAATGACCGGAAAGATAAGAGCAGGCAAGCTAATTTACCGATGTCGCGAAGGCTTACCGGAAAAAAGACTCTGGGAAAAAC
GTTAATGATGAGTTAATCACTCGCATCCATCAGA

Example of ORF finding : phase +3 and -3

GTGAAGGGCAATCAGCTGTTGCCGCTCTACTGGTAAAAGAAAAACCACCTGGCGCCAAATACGCAAAACCGCTCTCCCCGCGGTTGGCCGATTCAATATGC
AGCTGGCACACAGGTTTCCCAGCTGAAAAGCGGGCAGTGAAGCGCAACCGCAATTAATGTAGTGTAGCTCACTATTAGGCACCCAGGCTTTACACTTTATGCTTCC
GGCTCGTATGTTGTGGAAATTTGAGCCGGATAACAATTTACACAGGAAACAGCTTACGCCATGATACCGGATCACTGGCCGTCGTTTACAAACGTCGTGACTGGG
AAAACCCCTGGCGTTACCCAACTTAATCGCTTGCAGCAGATCCCCCTTTCGGCAGCTGGCGTAATAGCAGAAAGCGCCGACAGATCGCCCTTCCCAACAGTTGGC
CAGGCTGAATGGCGCAATGGCGCTTTGGCTGAGTTTCCGACACAGGAAAGCGCGGAAAGCTGGCCGAGAGTGGCATCTTCCGAGGCCGATACTGCTCGTGGCCC
CTAAACTGGCAGATGCACGGTTACGATGCGCCCATACACCAACGTGACCTATCCATACGGTCAATCCGCGCTTGTCCACGGAGAATCCGACGGGTTGT
ACTCGCTCACATTAATGTTGATGAAAAGCTGGCTACAGGAAAGCGGACGCGCAATTTTTGATGGCGTTAACTCGCGCTTTCATCTGTGGTCAACGGCGGCTGG
GTCGGTTACGGCCAGGACAGTCGTTTCCGCTGTAATTTGACCTGAGCGCATTTCACGCGCCGAGAAAAACCGCTCGCCGTTGATGGTGTGCGCTGGGATGACG
GCAGTTATCTGGAAGATCAGGATATGTTGGCGGATGAGCGGCATTTCCGTGACGTCTCGTTGCTGCATAAACCGACTACACAAATCAGCGATTCCATGTTGCCACT
GCTTTAATGATGATTTTACGCGCGCTGTACTGAGGCTGAAGTTAGATGTCGGCGAGTTCGCTGACTACCTACGGGTAACAGTTTCTTTATGGCAGGGTAAAACG
CAGGTCGCCAGCGGACCCGCGCTTTCGGCGGTGAATTTATCGATGAGCGTGGTGGTTATGCCGATCGCGTACACTACGCTGAAACGTCGAAAACCCGAAACTGT
GAGGCGCCGAAATCCCAGTCTCTATCGTGGCGTGGTGAACGTGCACACCGCCGACGGCAGCGCTGATTGAAGCAGAAAGCTGCGATGTCGTTTCCGCGAGGTGC
GGATGAAAATGGTCTGCTGCTGCTGAAACGGCAAGCGCTTACTGATTCCAGGCGTTAAACCGTACAGGATCATCTCTGATGGTCAAGGTGATGATGAGCAGAC
GATGGTGCAGGATATCTGCTGATGAAGCAGAAACACTTTAACCGCTGCGCTGTTCCGATTATCGAACCATCCGCTGGTGTACACGCTGGGACCGCTAGCGG
CTGATGTTGGTGGATGAAGCCAAATTTAAAACCCAGCGCATGGTCCAAATGCTGACCCGATGTCGCGCTGGCTACCGGGGATGAGCGAAACCGGTAACCG
GAATGGTGCAGCGCATGTAATACCCAGTGTGATCATCTGGTCCGTCGGGAATGAATCAGGCCACCGGCTAATCAGACGCGCTGTATCGCTGGATCAAACT
TGTGATCTTCCCGCCGTCAGTGAAGCCGCGGCGGACGACACCGCCAGCCAGATATTTTCCCGGATGACGCGCGGCTGATGAAGACCGACCTT
CCCGCTGTGCGCAAAATGGTCCATCAAAAATGCTTTCGCTACCTGGAGAGACGCGCCGCTGATCTTTCGCAATACGCCACCGCATGGTAAACAGTCTTGGC
GGTTTCGTAATACTGCAGCGCTTTCGTCAGTATCCCGCTTACAGCGCGGCTTCGCTGGGACTGGTGGATCAGTCCGCTGATTAATATGATGAAAACGGCAA
CCCGTGGTGGCTTACGGCGGTGATTTTGGCGATACCGGCAAGCAGTCCGAGTCTGTATGAACGCTGCTGGCTTTTGGCGACCGCAACCGCATCCAGCGCTGACG
GAAGCAAAACACCCAGCAGCATTTTTCCAGTTCCGTTTATCCGGCAAAACCATGGAATGACCAAGCAATACCTGTTCCGTCATAGCGATAACGAGCTCCTGCACCT
GATGGTGGCGCTGGATGGTAAGCCGCTGGCAAGCGGTGAAGTGCCTCTGGATGTCGCTCCACAAGGTAACAGTGTGATTGAATGCCTGAACTACCCGACGCGGA
GAGGCGCCGGCAACTCTGGCTACAGTACGCGTAGTGAACCGCAAGCGACCGGATGGTCAGAAGCCGGGACATCAGCGCTGGCAGCAGTGGCGTCTGGCGG
AAAACCTCAGTGTGACGCTTCCCGCGCGCTTCCCGCGGCTCAGCCAGTCCGACCCAGCGAAATGGATTTTGCATCGAGTAAGCGTAAAGCGTTGGCAATTTAAC
CGCCAGTCAGGCTTTTCTTACAGATGGTGGTGGGCAAAAAAACAATGCTGACGCGCGCTGGCGATCAGTTACCCGTCGACCGCTGGATAACGACATTTGGCT
AAGTGAAGCGAACCCGCAATTGACCCCTAACCGCTGGGTCGAAAGCGTGGCGGGCCGCAATACCAGCGCGAAGCAGCGTGTGTCAGTGCACGCGGACATAC
TGCTGATGGCGTCTGATTACGACCGCTCAGCGTGGCAGCATCAGGGGAAAACTTATTATCAGCGGAAAAACCTACCGGATTGATGGTGTGTTCAAATGGCG
ATTACCGTTGATGTTGAAGTGGCGGAGCGATAACCCGATCCCGCGGATGGCTGCAACTGCCAGTGGCGCAGGTAGCAGAGCGGGTAACTGGCTCGGATTA
GGCGCGCAAGAAAACTATCCGACCGCTTACTGCGCGCTGTTTACCGCTGGGATGTCGCCATTGTCAGACATGATACCCCGTACGCTTCCCGAGCGAAAAACG
GTCTGCGCTGCGGACGCGGCAATTAATGATGGCCACACAGGCTGGCGGGGACTTCCAGTTCAACATCAGCGCTACAGTCAACAGCAACTGATGAAAACCG
CCATCGCCTCTGCTGCACGGGAAAGGCAAGTGCATGGCTGAATATCGACGGTTTCAATATGGGGATGGTGGCGGACGACTCCTGGAGCCCGTCAAGTATCGCGGAA
TTCAGCTGAGCGCCGCTGCTACCATTAACATTTGGCTGGTGTCAAAAATAAATAAACCCGGCAGCCATGCTGCGCGTATTTCCGTAAGGAAATCCATATG
TACTATTAAAAAACAACAATTTGGATGTTCCGTTTATCTTTTCTTACTTTTATCAGTGGGAGGCTACTTCCCGTTTTCGCGATTGGCTACATGACATCAAC
CATATCAGCAAAAGTATACCGGATATTTTTGCGGCTATTTCTGTGCTGCTATTTCAACCGGCTGTTGGTCTGCTTCTGACAAAACCTGGGCTGCGCAAT
ACCTGCTGGATTTACCGCGATGTAGTGATGTTTGGCGGCTTCTTATTTATCTTCCGGCCACTGTTACAATAACAATTTAGTAGGATCGATGTTGGTGGT
ATTTATCTAGCTTTTGGTTTAAACCGCGTGGCCAGCAGTGAAGCAATTAATGCAAAAGTACGCGCTGCAAGTAAITTCGAATTTGGTCCGCGCGGATTTGGC
TGTGTTGGCTGGCGCTGTGTCCTCGATGTCGGCATCATGTTACCATCAATAATCAGTTGTTTCTGGCTGGGCTGCTGGCTGTCACATCTCTCGCGTTTTA
CTCTTTTTTCGCAAAACGGATGCGCCCTTCTGCCAGGTTGCCAATGCGGTAGTGGCCACCATTCCGCTTAGCCCTTAAAGTGGCAGCTGGAACGTTGCAGACA
GCCAAAACCTGTGGTTTTGTCAGCTGTATGTTATGGCGTTTCTGCGACCTACGATGTTTTGACCAACAGTTTGTCTAATTTCTTACTTCGTTCTTGTCTACCGGTGAAC
AGGCTACGCGGGTATTTGGCTACGTAAACGCAATGGGCGAATTACTTAAACCGCTGATGTTGTTCTTGGCCACTGATCATAATCGCATCGGTGGGAAAAACCGCC
TGCTGCTGGCTGGCAGCATATGTCGCTGACGATTTAGGCTCATCTGCGCACCTCAGCGCTGGAAGTGGTTATCTGAAAACCGTGCATATGTTGAAAGTACCG
TTCTGCTGGTGGGCTGCTTTAAATATATACCAGCCAGTTGAAAGTGGGTTTTACGCGAGGATTTATCGGCTGTTTCTGCTTCTTAAAGCAACTGGCGGATGATTT
TATGCTGACTGGCGGGAATATGATGAAAACGATCGTTTCCAGGCGGCTTATGTTGGTGGGCTGGGCTGCTGGGCGCTCACCTTAATTTCCGCTGTTCCAGC
TTAGCGGCCCCCGCCGCTTTCCTGCTGCGTGCAGTGAATGAAGTGCCTTAAAGCAATCAATGTCGGATGCGCGCGGACCGGCTTATCCGACCAACATATCAT
AACGGAGTGCATCGATTTAAACATGCCAATGACCGAAAAGATAAGAGCAGGCAAGCTAATTTACCGATGTCGCAAGGCTTACCGGAAAAAGACTCTGCGGAAAAAC
GTTAATGATGAGTTTAACTACTCGCATCCATCAGA



Example of ORF finding : phase +3 and -3

```
GTGAAGGGCAATCAGCTGTTGCCGCTCTACTGGTAAAAGAAAAACCACCTGGCGCCAAATACGCAAAACCGCTCTCCCCGCGCGTTGGCCGATTCAATATGC
AGCTGGCACACAGGTTTCCCAGCTGAAAAGCGGGCAGTGAAGCGCAACCGCAATTAATGTAGTGTAGCTCACTTATAGGCACCCAGGCTTTACACTTATGCTTCC
GGCTCGTATCTTGTGTGGAATTTGTAGCGGGATAACAATTTACACAGAGAAACAGCTTACGACATGATACGGATTCACTGGCCGTCGTTTACAAACGTCGTGACTGGG
AAAACCGCTGGCGTTACCCAACTTAATCGCCTTGACAGCACAATCCCCCTTTCGGCAGCGTGGCGTAATAGCAGAAAGGCGCCGACCGATCGCCCTTCCCAACAGTTGGC
CAGGCTGAATGGCGCAATGGCGCTTTGGCTCGTTCGCGTTCGCGACCCAGGCGGAGCGGCGGAAAAGTGGCGTAACTTCCCTGAGGCGGATACTGTCTGCTGCC
CTCAAACCTGGCAGATGACAGGTTACGATGCGCCCATACACCAACGTGACCTATCCATACGGTCAATCCGCGCTTGTCCACGGAGAATCCGACGGGTTGT
ACTCGCTCACATTAATGTTGATGAAAAGCTGGCTACAGGAAGGCCAGACGCGCAATATTTTTGATGGCGTTAACTCGCGGTTTCACTGTGGTGAACCGGGCGCTGG
GTCGGTTACGGCCAGGACAGTCGTTTCCGCTCTGAAATTTGACCTGAGCGCATTTCACGCGCCGAGAAAACCGCCTCGCCGTTGATGGTGTCTGGCTGGGATGAC
GCAGTTATCTGGAAGATCAGGATATGTGGCGGATGAGCGGCATTTCCGTGACGTCTCGTTGCTGCATAAACCGACTACACAAATCAGCGATTCCATGTTGCCACT
GCTTTAATGATGATTTTACGCGCGCTGTACTGGAAGCTGAAGTTAGATGCGGGGAGTTCGCTGACTACCTACGGGTAACAGTTTCTTTATGGCAGGGTGAACAG
CAGGTCGCCAGCGGCAACCGCGCTTTCCGCGGTGAAATTCATGATGAGCGTGGTGGTTATCCGATCGCGTACACACTACGCTGAAACGTCGAAAACCCGAAACTGT
GAGGCGCCGAAATCCGAACTCTATCGTGGCGTGGTGAAGTGCACACCGCCGACGGCAGCGTATTGAAGCAGAAAGCCTGCGATGTCGTTTCCGCGAGGTCG
GGATGAAAATGGTCTGCTGCTGCTGAAAGCGCAAGCGGTTACTGATTCCAGGCGTTAACCGTACAGGATCATCTCTGATGGTCAAGGTCATGGATGAGCAGAG
GATGGTGCAGGATATCTGCTGATGAAAGCAGAACTTAACCGCGTGGCTGATTCGCAATTCGGAACCATCCGCTGGTACACCGCTGGGACCGCTAGCGG
CTGATTTGGTGGATGAAAGCCAAATTTAAAACCCAGCGCAATTTGACCAACCGCTGACCGTGTACCGGATGATCCGCGGATGAGCGAAACCGGTAACCG
GAATGGTGCAGCGCATGTAATCCCGAGTGATCATCTGCTGCTGGGAATGAATCAGCCACCGCGCTAATCAGACCGCTGTATCGCTGGATCAATC
TGCTGATCTTCCCGCCGCTGAGTGAAGCGCGGAGCGGACGCAACCGCATATATTTCCCGGATGACGCGCGCTGATGAGAACAGCACGCTT
CCCGGCTGAACTGGCAAAATGGTCCATCAAAAATGGCTTTCGCTACCTGGAGAGACGGCCCGCTGATCTTTCGCAATACGCCACCGCATGGTAAACAGTCTTGGC
GGTTTCGCTAAACTGGCAGCGCTTTCGCTGATCCCGCTTTACAGCGCGCTTCGCTGGGACTGGTGGATCAGTCCGCTGATTAATATGATGAAAACGGCA
CCCGTGTGCGCTTACGGCGGTGATTTTGCGATACCGCGCAACGACGCGGATCTGTATGAACGCTGCTGCTTTTCCGAGCCGCAACCGCCATCCAGCGCTGAGC
GAAGCAAAACACCCAGCAGCATTTTCCAGTTCCGTTTATCCGGCAAAACATCGAATGACCAAGCGAATACCTGTTCCGCTATAGCGATAACGAGCTCCGCACT
GATGGTGGCGCTGGATGGTAAGCCGCTGGCAAGCGTGAAGTGCCTTCGCTCCACAAGGTAACAGTTGATTGAACGCTGCTGAACTACCCGAGCCGGA
GAGGCGCCGGCAACTCTGGCTCACAGTACGGCTAGTGAACCGCAACCGCAGCGGATGGTCAGAAGCCGGGACATCAGCGCTGGCAGCAGTGGCGCTGGCCG
AAAACCTCAGTGTGACGCTCCCGCGCGCTCCACCGCATCCCGCATCTGACACCAAGCGAAATGGATTTTGCATCGAGTGGGTAATAGCGTTGGCAATTTAAC
CGCCAGTCAGGCTTTCTTCCAGATGTGGATTTGGCATAAAAAACAATGCTGACGCGCGCTGGCCGATCAGTTCAACCGTGCACCCGCTGGATAACGACATTTGGCT
AAGTGAAGCCGCAATTCAGCCCTAACCGCTGGGTGCAAGCTGGGCGAGCGCGGCAATCAGCGCCGAAGCAGCGTTTGGATGCTGACGCGGAGATACACT
TGCTGATGGCGTCTGATTACGACCGCTACGCGTGGCAGCATCAGGGGAAAACCTTATTATACGCGGAAAACCTACCGGATTGATGGTATGGTCAAATGGCG
ATTACCGTTGATGTTGAAGTGGCGGAGCGGATACACCGCATCCGCGCGGATTTGGCTGAACGCTGCCAGCTGGCGCAGGTAGCAGAGCGGGTAAACTGGCTGGATTA
GGGCGCAAGAAAATATCCGACCGCCTTACTGCGCGCTGTTTGAACCGCTGGGATGTCGCCATGTCAGACATGATACCCCGTACGCTTCCCGAGCGAAAACG
GTCTGCGCTGCGGACGCGGCAATGAAATATGGCCACACCAAGTGGCGGGGACTTCCAGTTCAACATCAGCGCTACAAGTCAACAGCAACTGATGGAAGCAAC
CCATCGCCATCTGCTGACGCGGAAAGGACACATGGCTGAATCGACGGTTTCCATATGGGATGGTGGCGGACGATCTCCGGAGCCGCTCAGTATCGCGGAA
TTCAGCTGAGCGCGGCTGCTACCATACCAATGGCTGGTGTGATGTCAAAATAATAAACCGGCAAGCCATGCTGCCCCGTTTCCGTAAGGAAATCCATTA
TACTATTAAAAAACCAAACTTTGGATGTCGCTTATCTTTTCTTACTTTTATCATGGGAGCCTACTTCCCGTTTTCGCGATTGGCTACATGACATCAAC
CATATCAGCAAAAGTATGACGGGTATTAATTTGCGGCTATTTCTGCTGCTGCTATTTCAACCGCTGTTGGTCTGCTTTGACAAAACCTGGCTGGCGCAAT
ACCTGCTGGATATACCGCATGTAGTGTGATGTTGGCGGCTTCTTATTTTCTTTCGCGGCACTGTTACAATAACAATTTAGTAGGATCGATTTGGTGGT
ATTTATCAGCTTTTGTTTTAAAGCCGCTGGCCAGCAGTAGAGCAATTTAGGAAAAGTACCGCTGCGCAATTTGCAATTTGGCTGCGCGCGGATGTTGGC
TGTTGGCTGGCGCTGTGCTCCTGATGTCGGCATCATGTTACCATCAATAATCAGTTGTTTCTGGCTGGGCTGCTGCTGCACTCATCTCGCGTTTTA
CTCTTTTCCGCAAAACGGATGGCGCCTTCTGCGCAGTGTGCCAATGGCGGATGGTCCACCACTTCGCGCTTAGCCCTTAAAGCTGGCAGCTGGAACGTTGTCAGCA
GCCAAAACCTGTGTTTTGTCACTGTATGTTTATGGCCTTCTGCGCACTTCAGTATGTTTGAACCAACAGTTTGTCTAATTTCTTACTTCGTTCTTTGCTACCGCTGAA
AGGATACCGCGGATTTGGCTACGTAAACGACATGGGCGGAATTCATTAACCGCTCGATATGTTCTTTCGCGCACTGATCATTAATCGCATCGGTGGGAAAACCGCC
CTGCTGCTGGCGCACTATATGCTGTCAGTATTTGGCTCATCTGCGCCACTCAGCGCTGGAAGTGGTTATCTGAAAACCGCTGCATATGTTGAAAATGCC
TTCTGCTGGTGGGCTGCTTTAAATATATACCAAGCAGTTTGAAGTGGGTTTTTCAAGCGAGTTTATCTGGTCTGTTTCTGCTCTTTAAGCAACTGGGATGATTT
TATGCTCTACTGGCGGCAATATGATGAAAACGATCGTTTCCAGGCGGCTTACTGTTGGTGGTGGTCTGTTGGCGCTGGGCTTAAATTTCCGTTGTTACCG
TTAGCGCGCCCGCGCCCTTTCCCTGCTGCGTGTGAGTGAATGAAGTGGCTTAAAGCAATCAATGTCGGATGCGCGCGGAGCGCCTTATCCGACCAAGCATATCAT
AACGGAGTGTATCCGCTTAAACATGACCAAGAAATAAGAGCAGGCAAGCTATTTCCGATATGTCGGAAGGCTTACCGGAAAAGGACTCTGGTGGAAAAC
GTTAATGATGAGTTTAACTACTCGCATCCATCAGA
```



<https://www.ncbi.nlm.nih.gov/orffinder/>

Actually, you can use ORFFinder, It would be easier :

Enter Query Sequence

From: To:

Choose Search Parameters

Minimal ORF length (nt):

Genetic code:

ORF start codon to use:

"ATG" only

"ATG" and alternative initiation codons

Any sense codon

Ignore nested ORFs:

Start Search / Clear

Ab initio methods

```
GTGAAGGGCAATCAGCTGTTGCCGCTCTACTGGTGAAGAAAAAACCCCTGGCGCCAAATACGCAAAACCGCTCTCCCCGCGGTTGGCCGATTCAATATGC
AGCTGGCACACAGGTTTCCCAGCTGGAAGACGGGGCAGTGAAGCCAAACCGCAATTAATGTGAGTTAGCTCACTATTAGGCACCCAGGCTTTACACTTTATGCTTCC
GGCTCGTATGTTGTGGAAATTTGTGAGCGGATAACAATTTACACAGGAAACAGCTTACGACCATGATACGGATTCACTGGCCGTCGTTTACAAACGTCGTGACTGGG
AAAAACCTGGCGTACCCAACTTAATCGCCTTGACAGACATCCCCCTTTCGGCAGCGTGGCCATATAGCGAAAGAGGCCCGCAGCGATCGCCCTTCCCAACAGTTGGC
CAGCCTGAATGGCGAAATGGCCGCTTGGCTGTTTGGCTGAGTACCCAGGAAAGCGGTCGGCGAAAGCTGGCTGGAGTGGCATCTTCTGAGGCCGATACTGCTGCTGCC
CTAAACTGGCAGATGCACGGTTACGATGCGCCCATACACCAACGTGACCTATCCATTACGGTCAATCCCGCTTGTGCCACGGGAATCCGACGGGTTGT
ACTCGCTCACATTTAATGTTGATGAAAGCTGGCTACAGGAAGGCCAGACGCGAATATTTTTGATGGCGTTAACTCGCCGTTTCACTGTGGTCAACGGCGGCTGG
GTCGGTTACGGCCAGGACAGTCGTTTCCGCTGTAATTTGACCTGAGCGCATTTCACGCGCCGAGAAAAACCGCTCGCCGTTGATGGTGCCTGGCTGGAGTGACC
GCAGTTATCTGGAAGATCAGGATATGTTGGCGGATGAGCGCCATTTCCGTGACGTCTCGTTGCTGCATAAACCGACTACACAAATCAGCGATTCCATGTTGCCACTC
GCTTAATGATGATTTTACGCGCGCTGACTGAGGCTGAAGTTAGATGTCGGCGGAGTTGCGTGACTACCTACGGGTAACAGTTTCTTTATGGCAGGGTGAACG
CAGGTCGCCAGCGGCACCGCCGCTTTCGGCGGTGAATTTATCGATGAGCGTGGTGGTTATCCGATCGCGTCAACACTACGCTGAAACGTCGAAAACCCGAAACTGT
GGAGCGCCGAAATCCGAATCTCTATCGTGGCGGTTGAACTGCACACCGCCAGCGCACGCTGATTGAAGCAGAAGCCTGCCGATGTCGTTTCCGCGAGGTTGC
GGATTGAAAAATGGTCTGCTGCTGCTGAAGCGCAAGCCGTTGCTGATTCGAGCGCTTAAACCGTACAGGACATCATCTGTCGATGGTCAGGTATGGATGAGCAGAC
GATGGTGCAGGATATCTGCTGATGAAGCAGAAACAATTTAACGCCGTCGCTGTTCCGATTATCCGAACCATCCGCTGTGGTACACGCTGCGGACCCGACGGC
CTGTATGGTGGATGAAGCCAAATTTGAAACCCAGCGCATGGTCCAAATGCTGACCGATGATCCCGCTGGCTACCGGGGATGAGCGAACCGGTAACGC
GAATGGTCAGCGCGATGTAATCACCAGTGTGATCATCTGGTCGCTGGGAATGAATCAGGCCACGGCGCTAATCAGACCGGCTGTATCGCTGGATCAATC
TGTGATCTTCCCGCCGGTGCAGTATGAAGCGCGGAGCGACACCCAGCATATTTTCCCGGATGACGCGCGCTGATGAGAACAGCCAGCCCTT
CCCGCTGTGCCGAAATGGTCCATCAAAAATGGCTTTCGCTACCTGGAGAGACGGCCCGCTGATCTTTCGCAATACGCCACCGCATGGGTAACAGCTTGGC
GGTTTCGTAATACTGGCAGCGGTTTCGTCAGTATCCCGCTTACAGGCGGCTTCGCTGGGACTGGTGGATCAGTCGCTGATTAATATGATGAAAAAGCCGCA
CCCGTGGTCCGCTTACGGCGGTGATTTTGGCGATACCGCGAACCGATCCGAGCTGCTGTATGAACGGTCTGGCTTTTGGCGACCGCACCGCCATCCAGCGCTGAGC
GAAGCAAAACACCAGCAGCAGTTTTTCCAGTTCCGTTTATCCGGCAAAACCATGGAATGACCCAGCGAATACCTGTTCCGCTACAGCGATAACGAGCTCTGCACCT
GATGGTGGCGCTGGATGGTAAGCCGCTGGCAAGCGGTTGAACTGCTGATGTCGCTCCACAAGGTAACAGTTGATTGAACCTGCCGTAACACTACCCGACGCGGA
GAGGCCCGGGCAACTCTGGCTACAGTACGCGTAGTGCAACCGAACCGCCGATGGTGCAGAAGCCGGGACATCAGCGCCTGGCAGCAGTGGCGCTCGGGCG
AAAACTCAGTGTGACGCTTCCCGCCGCTCCACCCGCTCCGACTCGACACCCAGCGAAATGGATTTTGCATCGAGTAAAGCGTTGGCAATTTAAC
CGCCAGTCAGGCTTTCTTTCACAGATGGTGGTGGCATAAAAAAACAATGCTGACGCCGCTGGCCGATCAGTTACCCGCTGCACCCGCTGGATAACGACATTTGGCCT
AAGTGAAGCGAACCCGATTGACCCCTAACGCCGTTGGTGCAGCCGCTGGAAGCGTGGCGGACCCGCAATACCAGCCGGAAGCAGCGTTGTCAGTGCACGGGATACAT
TGCTGATGGCGTCTGATTACGACCGCTACCGGTTGGCAGCATCAGGGGAAAACTTATTATACGCGGAAAAACCTACCGGATTGATGGTAGTGGTCAAATGGCG
ATTACCGTTGATGTTGAAGTGGCGGAGCGGATACACCGCATCCGCGCGGATGGCTGAACTGCCAGCTGGCGAGGTAGCAGAGCGGGTAAACTGGCTCGGATTA
GGCGCGCAAGAAAAATATCCGACCGCCTACTGCGCCGCTGTTTACCGCTGGGATGTCGCCATTGTCAGACATGATACCCCGTACGCTTCCCGAGCGAAAAACG
GTCTGGCGTCCGGGACCGCGGAATTAATATGGCCACACCAAGTGGCGGGGCACTCCAGTTCAACATCAGCCGCTACAGTCAACAGCAACTGATGAAACCGA
CCATCGCCTGCTGTCACGGGAAAGGACAGTGGCTGAATATCGACGTTTCCATATGGGGATTGGTGGCGAGGACTCCTGGAGCCCGCTCAGTATCGGGCGGAA
TTCAGCTGAGCGCCGGTGCCTACCATTAACAGTTGGTCTGGTGTCAAAAATTAATAAACCCGGGACGGCCATGCTCGCCGATTTTCGCTAAGGAAATCCATTATG
TACTATTAAAAAACATAAATTTGGATGTTCCGTTTATCTTTTACTTTTATATCGGGGAGCTACTTCCCGTTTTCGCGATTGGCTACGATGACATCAAC
CATATCAGCAAAAGTACACGGTATTAATTTGCGGCTATTTCTGCTGCTGCTATTTCAACCCGCTGTTGGCTGCTTTCTGACAAACTCGGGCTGGCAAAAT
ACCTGCTGGGATATACCGCATGTTAGTGTATTTGCGCCGTTCTTATTTTATCTTCGGGCCACTGTTACAATAACAATTTAGTAGGATGATGTTGGTGGT
ATTTTATAGCTTTTGTTTTAAACCGCGTGGCCAGCAGTAGAGCAATTAATGAAAGTACAGCCGTCGCAATTTGCAATTTGGCTGGCGCGGATGTTGGC
TGTGTTGGCTGGCGCTGTGTCCTCGATGTCGGCATCATGTTACCATCAATAATCAGTTGTTTTCGCTGGCTGGGCTCGGCTGTCACATCATCTCGCCGTTTTA
CTCTTTTTTCGCAAAAGGATGCGCCCTTCTCGCCAGGTTGCCAATGCGGTAGTGGCCAACTCCGCTATTAGCCTTAAAGCTGGCAGCTGGAACGTTGTGCAGCA
GCCAAACTGTGGTTTTTGTCACTGTATGTTATGGCTGTTCTGTGCACTACGATGTTTGGCAACAGTTTGTCTAATTTCTTACTCTGTTCTTGTACCGGTGAAC
AGGGTACGCGGGTATTTGGCTACGTAAACGACATGGCGGAATTACTTAAACCGCTGATGATGTTCTTTCGCGCACTGATCATTAATCGCATCGGTGGGAAAAACCGCC
TCTGCTGCTGGCTGGCATAITATGCTGCTGACGTAATTTGGCTCATCTGTCGCCACTCAGCCCTCAGCCGCTGGAAGTGGTATTCTGAAAACCGCTGCATATGTTGAAGTACCG
TTCTGCTGGTGGGCTGCTTTAAATATATACCAGCCAGTTGAAAGCGGTTTTTCAAGCGGATTTATCTGGTCTGTTTGTCTTCTTAAAGCAACTGGCGGATGATTT
TATGCTGACTGGCGGGCAATATGATGAAAGCATCGTTTCCAGGGCCGCTTATGTTGGTGGGCTGCTGGTGGCGGCTGGGCTTACCTTAATTTCCGTGTTTCCG
TTAGCGGCCCGCCGCTTTCCTGCTGCTGCTGCTGAGTGAATGAAGTGGCTTAAAGCAATCAATGTCGGATGCGCGCGGACGGCCCTATCCGACCAACATATCAT
AACGGAGTGTGCTGATTAAGCATGCAATGACCGAAAGATAAGAGCAGGCAAGCTATTACCGATGTGCGAAGGCTTACCGGAAAAAGACTCTGGGAAAAAC
GTTAATGATGAGTTTAACTACTCGCATCCATCAGA
```

2 mains steps :

1) List Open Reading Frames (OFRs) over the 6 reading frames :

- starting with a start codon (mostly ATG)
- ending with a stop codon (TAA, TAG, TGA)

2) Select valid ORFs

Ab initio methods

```
GTGAAGGGCAATCAGCTGTTGCCGCTCTACTGGTGAAGAAAAAACCCCTGGCGCCAAATACGCAAAACCGCTCTCCCCGCGGTTGGCCGATTCAATATGC
AGCTGGCACACAGGTTTCCCAGCTGGAAGACGGGGCAGTGAAGCCGCAACGCAATTAATGTGAGTTAGCTCACTTATTAGGCACCCAGGCTTTACACTTATGCTTCC
GGCTCGTATGTTGTGGAAATTTGAGCGGATAACAATTTACACAGGAAACAGCTTACCATTGATACGACATGATACGATGATACGATGATACGATGATACGATGAT
AAAAACCTGGCGTTACCCAACTTAATCGCTTGCAGCAGTCCCTTTGGCCAGCTGGCGTAAATAGCAGAAAGGCGCCGACAGGATCGCCCTTCCCAACAGTTGGC
CAGCTGAAATGGCGAATGGCGTTGGCTGTTTGGCTGTTTGGCTGTTTGGCTGTTTGGCTGTTTGGCTGTTTGGCTGTTTGGCTGTTTGGCTGTTTGGCTGTTTGGCT
CTAAACTGGCAGATGCACGGTTACGATGCGCCCATACACCAACGTGACCTATCCATTACGGTCAATCCGCGTGTGTTCCACGGGAATCCGACGGGTTGT
ACTCGCTCACATTTAATGTTGATGAAAGCTGGCTACAGGAAGGCCAGACGCGAATATTTTTGATGGCGTTAACTCGCGGTTTCACTGTGGTCAACGGCGCTGG
GTCGGTTACGGCCAGGACAGTCGTTTCCGCTGTAATTTGACCTGAGCGCATTTTACGCGCCGGAGAAAAACCGCTCGCGGTGATGGTGCCTGGCTGGAGTGACG
GCAGTTATCTGGAAGATCAGGATATGTTGGCGGATGAGCGCCATTTCCGTGACGTCTGCTTTGCTGCATAAACCGACTACACAAATCAGCGATTCCATGTTGCCACTC
GCTTAAATGATGATTTTACGCGCGCTGACTGAGGCTGAAGTTAGATGTCGGCGGAGTTGCGTGACTACCTACGGGTAACAGTTTCTTATGGCAGGGTGAACG
CAGGTCGCCAGCGGCACCGCGCTTTCCGCGGTGAATTTATCGATGAGCGTGGTGGTTATCCGATCGCGTCAACACTACGCTGAAACGTCGAAAACCCGAAACTGT
GGAGCGCCGAAATCCGAATCTCTATCGTGGCGGTTGAACTGCACACCGCCGACGGCAGCGTGATTGAAGCAGAAAGCTGCGATGTCGTTTCCGCGAGGTTGC
GGATGAAAATGGTCTGCTGCTGCTGAAGCGCAAGCCGTTGCTGATTCAGGCGCTTAAACCGTCAAGGATCATCTCTGATGGTCAGGTATGGATGAGCAGAC
GATGGTGCAGGATATCTGCTGATGAAGCAGAAACAATTTAACGCCGTGGCTGATTCGCAATATCCGAACCATCCGCTGTGGTACACGCTGCGGACCCGTAGGGC
CTGATGTTGGTGGATGAAGCCAAATTTGAAACCCAGCGCATGGTCCAAATGCTGACCGATGATCCGCGTGGCTACCGGGATGAGCGAACCGGTAACGC
GAATGGTGCAGCGCATGTAATCACCAGTGTGATCATCTGGTCCGTGGGAATGAATCAGGCCACGGCGCTAATCAGACGCGCTGTATCGCTGGATCAATC
TGTGATCTTCCCGCCGGTGCAGTATGAAGCGCGGAGCGACACCCAGCATTAATTTGCCGATGACGCGCGCGTGATGAAAGACAGCCCTT
CCCGCTGTCCGAAATGGTCCATCAAAAATGGCTTTCGCTACCTGGAGAGACGGCCCGCTGATCTTTCGCAATACGCCACGCGATGGGTAACAGCTTGGC
GGTTTCGTAATACTGGCAGCGGTTTCCGTCAGTATCCCGCTTACAGCGCGGCTCGCTGGGACTGGTGGATCAGTCGCTGATTAATATGATGAAAACGGCAA
CCCGTGGTCCGCTTACGGCGGTGATTTTGGCGATCCCGGCAACGATCGCCAGTCTGTATGAACGGCTGGTCTTGGCGACCGCAACCGCATCCAGCGCTGAGC
GAAGCAAAACACCCAGCAGCATTTTTCCAGTCCGTTTATCCGCGCAAAACCATGGAATGACCCAGCGAATACCTGTTCCGCTATAGCGATAACGAGCTCTGCACCTG
GATGGTGGCGCTGGATGTTAAGCCGCTGGCAAGCGGTGAACTGCTGATGTCGCTCCACAAGGTAACAGTTGATTGAACCTGCTGAACTACCCGACGCGGTA
GAGGCCCGGGCAACTCTGGCTCACAGTACGCGTAGTGCAACCGCAACCGCCGATGGTGCAGAAAGCCGGGACATCAGCGCTGGCAGCAGTGGCGCTCGGGG
AAAACTCAGTGTGACGCTTCCCGCGCGCTCCACCGCATCCGCTGACGACCCAGCAGAAATGGATTTTGCATCGAGTGGGTAATAGCGTTGGCAATTTAAC
CGCCAGTCAGGCTTTTCCAGATGTTGGATGGGCAATAAAAAAATCTGCTGACCGCGCTGGCGATCAGTTACCCGCTGCACCGCTGGATAACGACATTTGGCGT
AAGTGAAGCCAGCCGATTAACCGCTGAGCCGTTGAAAGCGTGGTGAAGCGCTGGGCAAGCGCCATTACCAGCGCGAAGCAGCGTTGTTGCAGTGCACGGAGATAC
TGTGATGCGCGTCTGATTACGACCGCTACCGGTTGGCAGCATCAGGGGAAAACTTATTATCAGCGGAAAAACCTACCGGATGATGGTATGGTCAAATGGCG
ATTACCGTTGATGTTGAAGTGGCGGAGCGATACCCGATCCGCGCGGATGGCTGAACTGCCAGCTGGCGAGGTAGCAGAGCGGGTAAACTGGCTCGGATTA
GGCGCGCAAGAAAAATATCCGACCGCTTACTGCGCGCTGTTTGAACCGTGGGATGTCGCCATTGTCAGACATGATACCCCGTACGCTTCCCGAGCGAAAAACG
GTCTGCGCTGCGGACCGCGGAATTAATGATGGCCACACAGTGGCGGGGCACTCCAGTTCAACATCAGCCGCTACAGTCAACAGCAACTGATGAAACCGA
CCATCGCTCTGCTGCACGGGAAAGGACAGTGGCTGAATCGACGGTTTCCATATGGGGATGGTGGCGGACGACTCCTGGAGCCGCTCAGTATCGGGGAA
TTCAGCTGAGCGCGGTCGCTACCATACCAATGGTGTGGTGTGCAAAAATAATAAACCGGCAAGCCATGCTGCGCGTATTCGCTGAAAGCAATCCATATG
TACTATTAATAAACACAACCTTTGGATGTTCCGTTATGCTTTTCTTACTTTTATCAGTGGGAGGCTACTTCCCGTTTTCGCGATTTGGCTAGCATGACATAAC
CATATCAGCAAAAGTATACCGGATATATTTTGGCCGATTTCTGCTGCTGCTATTTCAACCGGCTGTTGGCTGCTTTCTGCAAACTCGGGCTGGCAAAAT
ACCTGCTGGGATATACCGCATGTAGTGTGTTTGGCGGTTCTTATTTATCTTCCGGCCACTGTTACAATAACAATTTAGTAGGATGATGTTGGTGGT
ATTTATAGCTTTTGTTTAACCGCGGTGCCCGAGCAGTGAAGCAATTAATGCAAAAGTCAAGCGCTGCACTAATTTGCAATTTGGCTGGCGCGGATGTTTGGC
TGTGTTGGCTGGCGCTGTGCTCGTATGTCGGCATCATGTTACCATCAATAATCAGTTGTTTCTGGCTGGGCTCGGCTGTCACATCATCTCGCGTTTTA
CTCTTTTTCCGCAAAACGGATGCGCCCTTCTCGCCAGTGTGCCAATGCGGTAGTGGCAACCATCCGCTTATAGCCTTAAGCTGGCAGTGAAGCTTGCAGACA
GCCAAAACCTGGTTTTTGTCACTGTATGTTTGGCTGTTCTGTGCACTACGATGTTTGGCAACAGTTTGTCTAATTTCTTACTCTGTTCTTGTACCGGTGAAC
AGGGTACGCGGGTATTTGGCTACGTAAACGACAAATGGCGCAATTAACCGCCCTGATGATGTTCTTGGCCACTGATCATTAATCGCATCGTGGGAAAAACCGCC
CTGCTGCTGGCTGGCAGTATGTCGCTGACGATTTAGGCTCATCTGTCGCCACTCAGCGCTGGAAGTGGTTATCTGAAAACCGCTGCATATGTTGAAGTACCG
TTCTGCTGGTGGGCTGCTTTAAATATATACCAGCCAGTTGAAAGCGGTTTTACGCGGACTTATCTGGTCTGTTTGTGCTTTTAAAGCAACTGGCGGATGATTT
TATGCTGACTGGCGGGCAATATGATGAAAGCATCGTTTCCAGGCGGCTTATGTTGGTGGTCTGTTGGTGGTCTGTTGGCGGCTGGGCTCACCTTAATTTCCGTTTCCG
TTAGCGCGCCCGCGCCGCTTTCCCTGCTGCGTGCAGTGAATGAAGTGCCTTAAAGCAATCAATGTCGGATGCGCGCGGACGGCCCTATCCGACCAACATATCAT
AACGGAGTGTGCTGATTAAGCATGCAATGACCGAAAGATAAGAGCAGGCAAGCTATTACCGATGTGCGAAGGCTTACCGGAAAAAGACCTCTGGGAAAAAC
GTTAATGATGAGTTTAACTACTCGCATCCATCAGA
```

2 main steps :

1) List Open Reading Frames (OFRs) over the 6 reading frames :

- starting with a start codon (mostly ATG)
- ending with a stop codon (TAA, TAG, TGA)

2) Select valid ORFs

Select valid ORFs because of overlaps between phases

Example:

Phase +1 **Overlap** Phase +2

```
...GGAATGAATCAGGCCACGGCGCTAATCACGACGCGCTGTATC
GCTGGATCAAATCTGTCTGATCCTTCCCGCCCGGTGCAGTATGAA
GGCGGGCGGAGCCGACACCACGGCCACCGATATTATTTGCCCGA
TGTACGCGCGCGTGGATGAAGACCAGCCCTTCCCGGCTGTGCC
GAAATGGTCCATCAAAAATGGCTTTCGCTACCTGGAGAGACGC
GCCCGCTGATCCTTTGCGAATACGCCACGCGATGGGTAAACAG
```

...

But overlaps can be valid too...

Select valid ORFs because of overlaps between phases

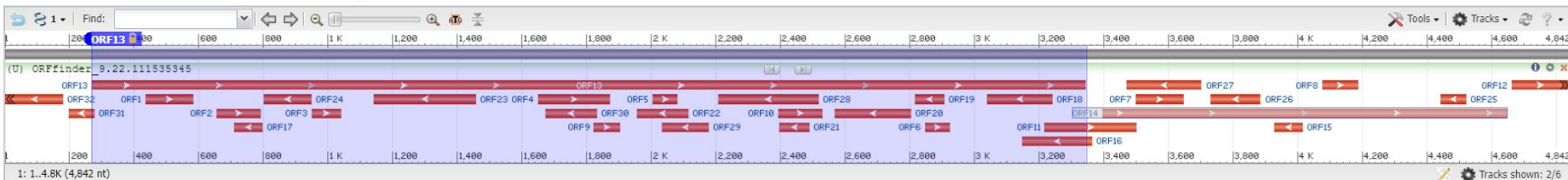
< ORFfinder submitting page

> Help

Open Reading Frame Viewer

Sequence

ORFs found: 32 Genetic code: 1 Start codon: 'ATG' only



ORF13 (1024 aa) Display ORF as...

Mark

```
>|c1|ORF13
MTMITHDLSLAVVLQRDDWENPGVTQLNRLAAHPFASWRISSEEARDRPSQ
QLRSLNGEVRFAHPFAPEAVPESVLECDLPEADTVVPSM@PHGYDAP1
YTMVYPTIVNIPFPVTEINPTGYSYSLTFWVDESILQESQTRIFDQVISA
FHLNCGNRWVGYGQDSRLPSEFDLSAFLRAGENLRVWLRMSDQSVLED
QMWNRMSGIFRQVSLLHKPTTQISDFHVATRNFDDFSAVLAEVQKCGE
LRDYLRVTVSLKQGETQVASGTAPFGGEIIDERGGYADRVTLRNINVKP
LHSAEIPNLYRAVVELHTADGTLIEAEACDVGFREVRIEHGLLLNGKPL
LIRGVRHREHPLHGGVDEQTRVQDILLKWKQWIFUAVRCSHPHPLHY
TLCDRYGLVYVDEANIEETHGSRWPMRLTDDPNLPAHSERVTRINVRDRII
HPSVIIISLGNESGHGHDALYRNIKSDPSPRVQEGGGADTTADDI
CPMYARVDEDDQPPAVPKWISIKKWLSPGETRPLLCEYAHAMGNSLGGF
```

ORF13

Marked set (0)

SmartBLAST

SmartBLAST best hit titles...

BLAST

BLAST

BLAST Database:

UniProtKB/Swiss-Prot (swissprot)

Mark subset...

Marked: 0

Download marked set

as Protein FASTA

Label	Strand	Frame	Start	Stop	Length (nt aa)
ORF27	-	3	3703	3470	234 77
ORF4	+	1	1651	1872	222 73
ORF16	-	1	3363	3148	216 71
ORF18	-	2	3242	3039	204 67
ORF32	-	3	181	>	180 59
ORF12	+	2	4664	>4840	177 58
ORF30	-	3	1831	1673	159 52
ORF26	-	2	2114	1956	159 52
ORF22	-	3	3886	3731	156 51
ORF7	+	1	3499	3648	150 49
ORF24	-	2	950	801	150 49

Six-frame translation...

How to select valid ORFs in *ab initio* methods ?

No consensual methods but here are the main tools used:

- Glimmer
- (Meta)GeneMarkS2
- Phanotate : dedicated to phage genes
- PROkaryotic DYnamic programming Gene-finding ALgorithm (PRODIGAL)

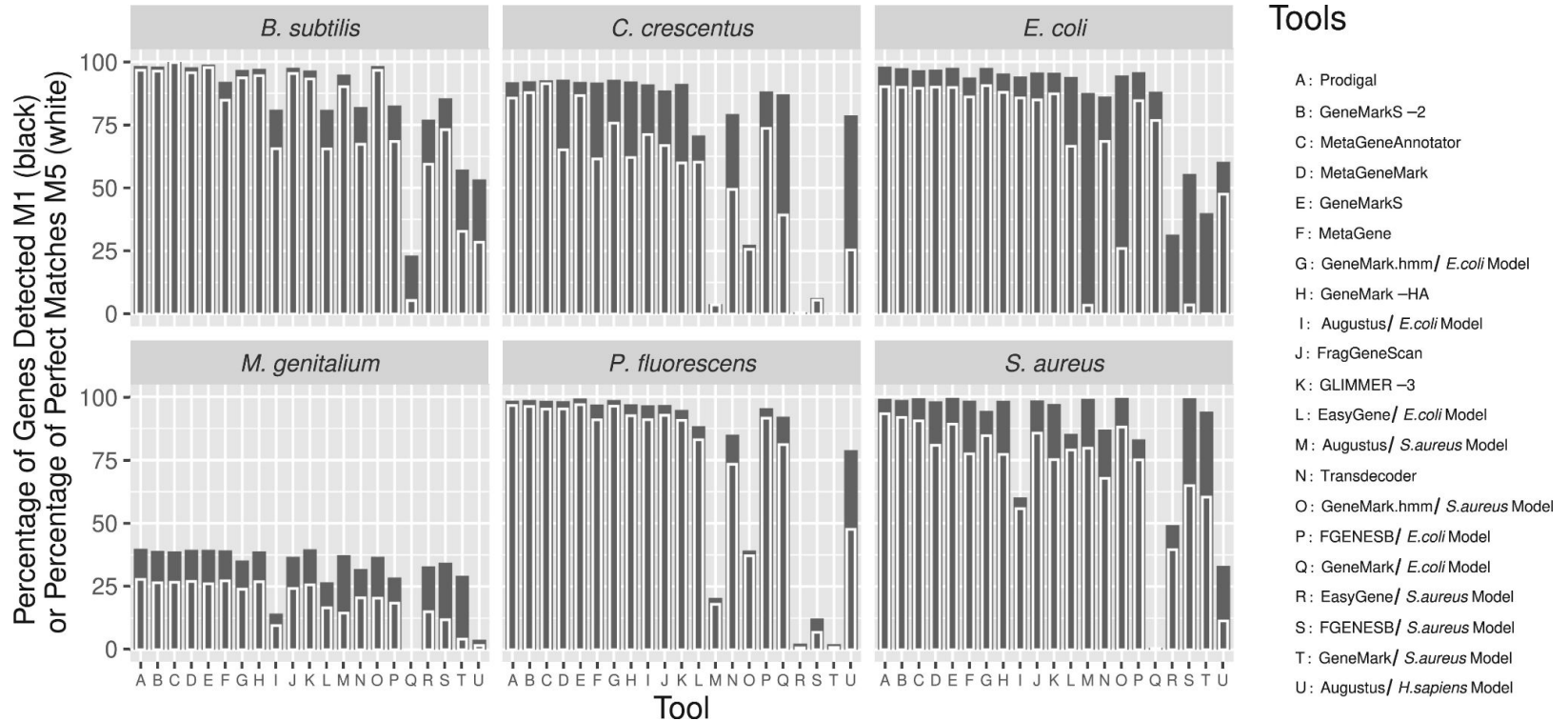
How to select valid ORFs in *ab initio* methods ?

No consensual methods but here are the main tools used :

- Glimmer
- (Meta)GeneMarkS2
- Phanotate : dedicated to phage genes
- **PROkaryotic DYnamic programming Gene-finding ALgorithm (PRODIGAL)**
 - Good results in benchmarking
 - Prodigal runs very quickly (a bacterial genome in about 10 seconds)
 - The *ab initio* prediction tool used by almost everybody (often wrapped in other tools)

Benchmark of *ab initio* gene prediction tools

Dimonaco *et al.*, 2022



The Prodigal algorithm : a nightmare to explain

1. Read in the sequence
2. Locate all starts and stops in the genome
3. Scan all open reading frames and record numbers of G's and C's in each codon position
4. Build a frame bias model based on ORF length and G/C codon position within each ORF
5. Record the highest scoring start nodes in each frame that overlap a stop codon by ≤ 60 bp
6. Do the first pass dynamic programming, connecting nodes based on frame bias scores
7. Create a hexamer background of all 6-mers in the entire sequence
8. FOR each gene model in the dynamic programming output:
 1. Gather all hexamer statistics
9. Create log table of hexamer coding scores
10. FOR each gene model in the dynamic programming output:
 1. Calculate a coding score based on hexamer statistics
 2. Penalize the score if there is a higher scoring start upstream in the same ORF
 3. IF the gene is very long but has a negative score, THEN give it a barely positive score
11. FOR 10 iterations
 1. Build a ribosomal binding site and ATG/GTG/TTG background for all nodes
 2. FOR each gene with a score of > 35.0 :
 1. Gather its Shine-Dalgarno RBS motif data and ATG/GTG/TTG data
 3. Modify RBS and ATG/GTG/TTG weights by the observations
12. IF organism is not determined to use Shine-Dalgarno THEN run the non-SD finder
13. FOR each gene model:
 1. Assign a final score of start score + coding score
 2. Penalize the final score of genes < 250 bp
14. Do the second pass dynamic programming, connecting nodes based on hexamer coding
15. FOR each gene model in the final dynamic programming:
 1. Eliminate negative scoring models
 2. Resolve very close start pairs (≤ 15 bp from each other)
16. Print final output

The Prodigal algorithm : a nightmare to explain (is it worth it ?)

1. Read in the sequence
2. Locate all starts and stops in the genome
3. Scan all open reading frames and record numbers of G's and C's in each codon position
4. Build a frame bias model based on ORF length and G/C codon position within each ORF
5. Record the highest scoring start nodes in each frame that overlap a stop codon by ≤ 60 bp
6. Do the first pass dynamic programming, connecting nodes based on frame bias scores
7. Create a hexamer background of all 6-mers in the entire sequence
8. FOR each gene model in the dynamic programming output:
 1. Gather all hexamer statistics
9. Create log table of hexamer coding scores
10. FOR each gene model in the dynamic programming output:
 1. Calculate a coding score based on hexamer statistics
 2. Penalize the score if there is a higher scoring start upstream in the same ORF
 3. IF the gene is very long but has a negative score, THEN give it a barely positive score
11. FOR 10 iterations
 1. Build a ribosomal binding site and ATG/GTG/TTG background for all nodes
 2. FOR each gene with a score of > 35.0 :
 1. Gather its Shine-Dalgarno RBS motif data and ATG/GTG/TTG data
 3. Modify RBS and ATG/GTG/TTG weights by the observations
12. IF organism is not determined to use Shine-Dalgarno THEN run the non-SD finder
13. FOR each gene model:
 1. Assign a final score of start score + coding score
 2. Penalize the final score of genes < 250 bp
14. Do the second pass dynamic programming, connecting nodes based on hexamer coding
15. FOR each gene model in the final dynamic programming:
 1. Eliminate negative scoring models
 2. Resolve very close start pairs (≤ 15 bp from each other)
16. Print final output

The Prodigal algorithm : a summary

1. Read in the sequence
2. Locate all starts and stops in the genome
3. Scan all open reading frames and record numbers of G's and C's in each codon position

Simplified

4. Look for bias in the percentage of GC% in candidates: find ORFs having a similar %GC bias
5. Look for difference in the frequency of hexamers statistics in contrast to background (region not covered by ORF)
6. Look for conserved upstream ORF patterns (Shine-Delgarno)
7. After computing all of these statistics : output predicted coding genes (Coding DNA Sequences, CDS)

Conclusion

- Coding genes are detected via *ab initio* methods by finding all ORFs and selecting the most relevant one based on complex empirical rules
- This approach allows predicting **even unknown coding genes** contrary to an alignment approach, but :
 - some genes can be missed
 - some genes can be truncated due to a partial detection
 - some genes can be longer than the reality
 - some detected genes can be biologically inactive (pseudogenes)
- Therefore, syntactic annotation must be refined using a database of known proteins (see lesson 5. Introduction to functional annotation)

2) Comment selectionner les bons ORFs avec Prodigal ?

Prodigal va faire des calculer les frequences des bases dans les condons, exemple ORF de la phase +1 :

...TGAATGGCGAATGGCGCTTTGCCTGGTTTC CGGCACCCAGAA
GCGGTGCCCGGAAAGCTGGCTGGAGTGCGATCTTCCTGAGGC
CGATACTGTCGTCCCTCAAAGTGGCAGATGCACGGTTA
CGATGCGCCCATCTACACCAACGTGACCT...

2) Comment selectionner les bons ORFs avec Prodigal ?

Prodigal va faire des calculer les frequences des bases dans les condons, exemple ORF de la phase +1 :

...TGAATGGCGAATGGCGCTTTGCCTGGTTTC CGGCACCCAGAA
GCGGTGCCCGGAAAGCTGGCTGGAGTGCGATCTTCCTGAGGC
CGATACTGTCGTCCCTCAAAGTGGCAGATGCACGGTTA
CGATGCGCCCATCTACACCAACGTGACCT...

Base	Position 1	Position 2	Position 3
A			
T			
C			
G			

2) Comment selectionner les bons ORFs avec Prodigal ?

Prodigal va faire des calculer les frequences des bases dans les condons, exemple ORF de la phase +1 :

...TGA **A**T**G****G**C**G**A**A**T**G****G**C**G**C**T**T**T**G**C**C**T****G**G**T**T**T**C**C**G**G****C**A**C****C**A**G**A**A**
AG**C**G**G**T**G**C**C**G**G**A**A**A**G**C**T****G**G**C**T**G**G**A**G**T****G**C**G**A**T**C**T**T**C****C**T**G**A**G**G
CC**G**A**T**A**C**T**G**T**C**G**T**C**G**T**C**C**C**C**T**C**A**A**A**C**T****G**G**C**A**G**A**T**G**C**A**C**G**G**T**T**
AC**G**A**T**G**C**G**C**C**C**A**T**C**T**A**C**A**C**C**A**A**C**G**G**A**C**C**T**...

Base	Position 1	Position 2	Position 3
A	15		
T	11		
C	14		
G	9		

2) Comment selectionner les bons ORFs avec Prodigal ?

Prodigal va faire des calculer les frequences des bases dans les condons, exemple ORF de la phase +1 :

...TGAATGGCGAATGGCGCTTTGCCCTGGTTTCCGGCACAGAA
AGCGGTGCCGGAAGCTGGCTGGAGTGCGATCTTCCTGAGG
CCGATACTGTCTGTCCCTCAAACTGGCAGATGCACGGTT
ACGATGCGCCCATCTACACCAACGTACCT...

Base	Position 1	Position 2	Position 3
A	15	6	
T	11	10	
C	14	18	
G	9	15	

2) Comment sélectionner les bons ORFs avec Prodigal ?

Prodigal va faire des calculer les frequences des bases dans les condons, exemple ORF de la phase +1 :

...TGAATGGCGAATGGCGCTTTGCCCTGGTTC CGGCAC CAGAA
GCGGTGCCGGAAGA GCTGGCTGGAGTGC GATCTTCCTGAGGC
CGATACTGTCGTCGTC CCTCAACTGGCAGATGCACGGTTA
CGATGCCGCCATCTACACCAACGTG CCT...

Base	Position 1	Position 2	Position 3
A	15	6	8
T	11	10	10
C	14	18	11
G	9	15	20

2) Comment sélectionner les bons ORFs avec Prodigal ?

Prodigal va faire des calculer les frequences des bases dans les condons, exemple ORF de la phase +1 :

...TGAATGGCGAATGGCGCTTTGCCCTGGTTCCTCGGCACACAGAA
GCGGTGCCGGAAAGCTGGCTGGAGTGCATCTTCCTGAGGC
CGATACTGTCGTCTCCCTCAACTGGCAGATGCACGGTTA
CGATGCCGCCCATCTACACCAACGTGCCT...

Base	Position 1	Position 2	Position 3
A	15	6	8
T	11	10	10
C	14	18	11
G	9	15	20

2) Comment sélectionner les bons ORFs avec Prodigal ?

Prodigal va calculer les fréquences des bases A/T et G/C dans les 3 positions des condons, exemple d'ORF de la phase +1 :

...TGAATGGCGAATGGCGCTTTGCCCTGGTTTC CGGCAC CAGAA
GCGGTGCCGGAAGCTGGCTGGAGTGC GATCTTCCTGAGGC
CGATACTGTCGTCCCTCAACTGGCAGATGCACGGTTA
CGATGCCGCCCATCTACACCAACGTG CCT...

Base	Position 1	Position 2	Position 3
A/T	15A+11T=26	6A+10T=16	8A+10T=18
C/G	14C+9G=23	18C+15G=28	11C+20G=31

2) Comment sélectionner les bons ORFs avec Prodigal ?

On regarde la position “gagnante” en terme de fréquence de C/G

...TGAATGGCGAATGGCGCTTTGCCCTGGTTTCCGGCACCAAGAA
GCGGTGCCGGAAGCTGGCTGGAGTGCATCTTCCCTGAGGC
CGATACTGTCGTCCCTCAAACTGGCAGATGCACGGTTA
CGATGCCCACTCTACCAACGTGCCT...

Base	Position 1	Position 2	Position 3
A/T	15A+11T=26	6A+10T=16	8A+10T=18
C/G	14C+9G=23	18C+15G=28	11C+20G=31

2) Comment sélectionner les bons ORFs avec Prodigal ?

On compte les gagnants ?

Comptage	Position 1	Position 2	Position 3
ORF où la position est gagnante	0	0	1

2) Comment sélectionner les bons ORFs avec Prodigal ?

Et avec tous les ORFs de toutes les phase ?

Comptage	Position 1	Position 2	Position 3
ORF où la position est gagnante	7	15	10

2) Comment sélectionner les bons ORFs avec Prodigal ?

Et avec tous les ORFs de toutes les phase ?

Comptage	Position 1	Position 2	Position 3
ORF où la position est gagnante	7	15	10

2) Comment sélectionner les bons ORFs avec Prodigal ?

Et avec tous les ORFs de toutes les phase ?

Comptage	Position 1	Position 2	Position 3
ORF où la position est gagnante	7	15	10