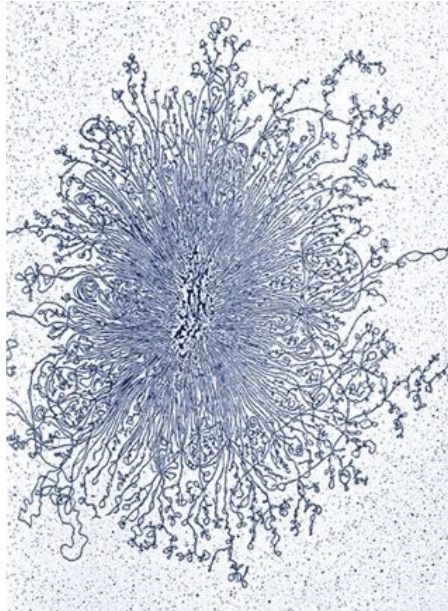


Introduction to gene prediction and functional annotation

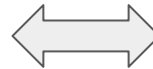
Guillaume GAUTREAU, 05/06/2024

Prokaryotic genomes...

DNA/chromosome : molecular support



E. coli nucleoid (Wang *et al.*, 2013)



Genome : the information, the sequence from the chromosome

```
GTGAAGGGCAATCAGCTGTTGCCCGTCTCACTGGTGAAAAGAAAAAC
CACCTGGCGCCCAATACGCAAACCGCCTCTCCCCGCGCGTTGGCC
GATTCATTAATGCAGCTGGCACGACAGGTTTCCCGACTGGAAAGCGG
GCAGTGAGCGCAACGCAATTAATGTGAGTTAGCTCACTCATTAGGCA
CCCCAGGCTTTACACTTTTATGCTTCCGGCTCGTATGTTGTGTGGAAT
GTGAGCGGATAACAATTTACACAGGAAACAGCTATGACCATGATTA
CGGATTCACTGGCCGTCGTTTTACAACGTCGTGACTGGGAAAACCT
GGCGTTACCCAACTTAATCGCCTTGCGACATCCCCCTTTCGCCAG
CTGGCGTAATAGCGAAGAGGCCCGCACCGATCGCCCTTCCAACAG
TTGCGCAGCCTGAATGGCGAATGGCGCTTTGCTGGTTTCCGGCAC
CAGAAGCGGTGCCGAAAGCTGGCTGGAGTGCGATCTTCTGAGGC
CGATACTGTCGTCGTCGCCCTCAAACCTGGCAGATGCACGGTTACGATG
CGCCCATCTACACCAACGTGACCTATCCATTACGGTCAATCCGCCG
TTTGTTCCCACGGAGAATCCGACGGGTTGTTACTCGCTCACATTTAAT
GTTGATGAAAGCTGGCTACAGGAAGGCCAGACGCGAATTATTTTTGA
TGGCGTTAACTCGGCGTTTCATCTGTGGTGCAACGGGCGCTGGGTC
GGTTACGGCCAGGACAGTCGTTTGCCGTCTGAATTTGACCTGAGCG
CATTTTTACGCGCCGGAGAAAACCGCCTCGCGGTGATGGTGCTGCG
CTGGAGTGACGGCAGTTATCTGGAAGATCAGGATATGTGGCGGATG
AGCGGCATTTTCCGTGACGTCTG.....
```



...annotation

According to Cambridge Dictionary



annotate

verb [T] • formal

UK  /'æn.ə.teɪt/ US  /'æn.ə.teɪt/



[often passive]

to add a short explanation or opinion to a text or image:

- *Annotated editions of Shakespeare's plays help readers to understand old words.*
- *an annotated bibliography/manuscript/edition*
- *His great-granddaughter has painstakingly transcribed and annotated his wartime diaries.*
- *The students annotate their photos, saying why they are relevant.*



COMPUTING, LANGUAGE • specialized

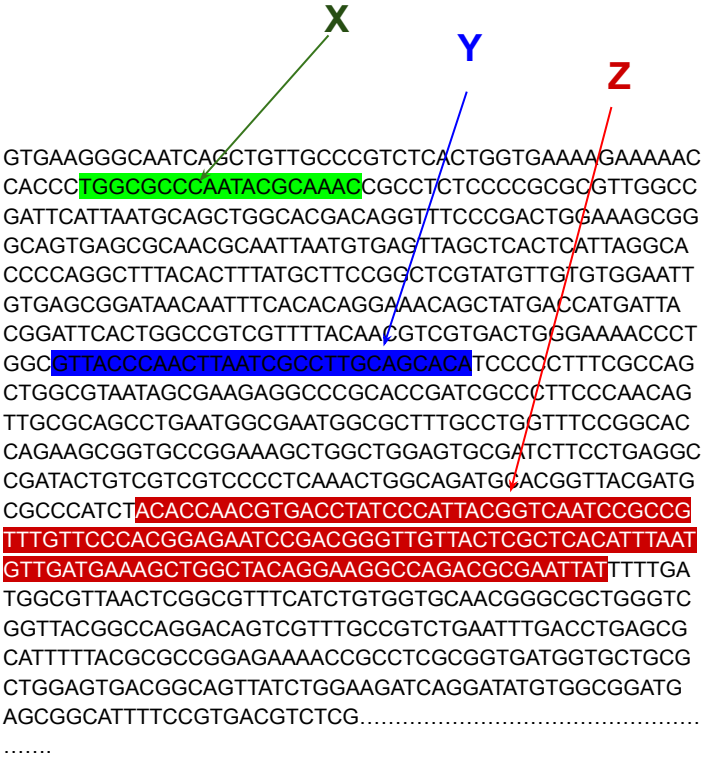
to add a description or piece of information to data, for example a label saying whether a word is a noun, a verb, etc.:

- *After the corpus was collected we annotated it.*
- *Textual or numerical data can be copied into databases, annotated, and linked to other data.*

Prokaryotic genome annotation

It means the combination of:

1. a genomic sequence location (described using coordinates)
2. a biological meaning for this sequence:
 - Is it a gene ?
 - what the function of that gene ?
 - Is it a coding gene or a non coding gene ?
 - Is it a set of co-transcribed genes ?
 - Is it regulated by a common factor ?
 - Is it a restriction site ?
 - Is it a integration hotspot ?
 - Is it a replication origin ?
 -



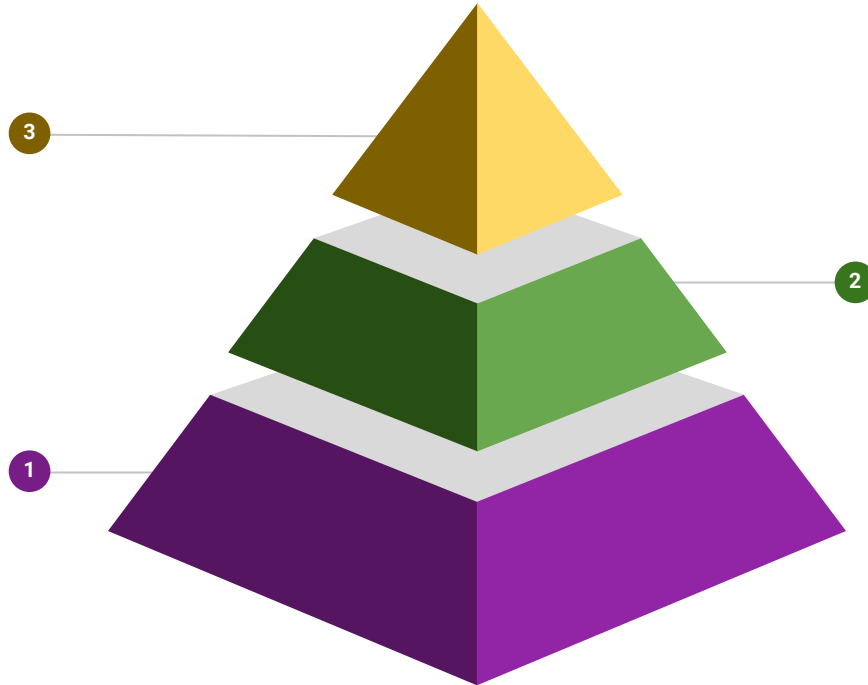
Cumulative levels of genomic annotation

Relational annotation

Which genes are associated in a common biological process ?
(how?)

Syntactic annotation

Detect the location of the genes (where ?)



Functional annotation

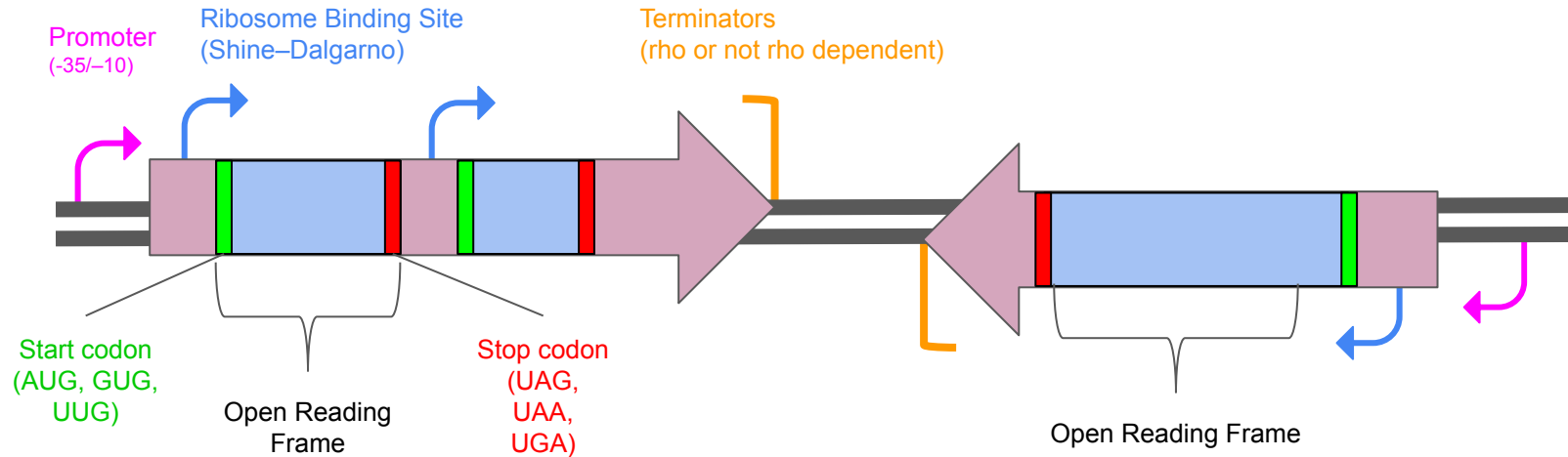
Find the biological functions of the genes (what?)

How to detect gene positions in a genome ?

Starting from an assembly, syntactic annotation is based on the identification of specific patterns:

- 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

Reminder on the genomic structure of coding genes

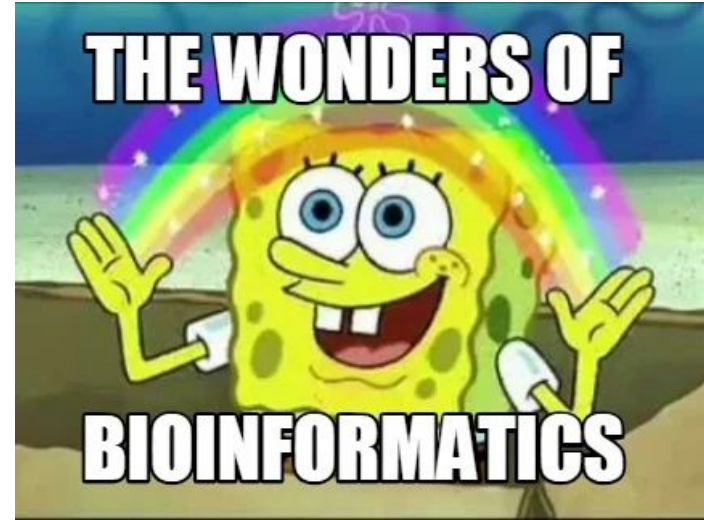


But the actual data are:

GTGAAGGGCAATCAGCTGTTGCCCGTCTCACTGGTGAAGAAAAAACCCCTGGCGCCAAATACGAAACCGCCTCTCCCCGCGGTTGGCCGATTCAATATGC
AGCTGGCACACAGGTTTCCCAGCTGGAAGCGGGCAGTGAAGCGCAACCGCAATTAATGTGAGTTAGCTCACTATTAGGCACCCAGGCTTTACACTTTATGCTTCC
GGCTCGTATGTTGTGGAAATTTGAGCCGGATAACAATTTACACAGGAAACAGCTTACCATTGATACGCGTCACTGGCCGTCGTTTACAAACGTCGTGACTGGT
AAAAACCCCTGGCGTTACCCAACTTAATCGCCTTGACAGCACCCTTTCCGGCAGCTGGCGTAATAGCGAAAGAGGCCCGCACCGGATCGCCCTTCCCAACAGTTGGC
CAGCTGAAATGGCGCAATGGCCGTTTGGCTGTTTTCCGGCAGTCCAGCGAGCGGTAAGCGGTAAGCGGTAAGCGGTAAGCGGTAAGCGGTAAGCGGTAAGCGG
CTAAACTGGCAGATGCACGGTTACGATGCGCCCATACACCAACGTGACCTATCCATTACGGTCAATCCGCGCTTTGTTCCACGGAGAATCCGACGGGTTGT
ACTCGCTACACTTAATGTTGATGAAAGCTGGCTACAGGAAGGCCAGACGCGCAATATTTTTGATGGCGTTAACTCGCGCTTTCATCTGTGGTCAACCGGGCGCTGG
GTCGGTTACGGCCAGGACAGTCGTTTTGCCGCTGTAATTTGACCTGAGCGCATTTCACGCGCCGGAGAAAAACCGCCTCGCCGTTGATGGTGTGCGCTGGGATGAC
GCAGTTATCTGGAAGATCAGGATATGTGGCGGATGAGCGCCATTTCCGTGACGTCTGTTGCTGCATAAACCGACTACACAAATCAGCGATTTCCATGTTGCCACT
GCTTTAATGATGATTTTACGCGCGCTGTACTGGAAGCTGAAGTTAGATGTGCGCGGAGTTGCGTGACTACCTACGGGTAACAGTTTCTTTATGGCAGGGTGAACG
CAGGTCGCCAGCGGCACCGCGCCTTTCCGGCGTGAATTTATCGATGAGCGTGGTGGTTATCCCGATCGCGTACACACTACGTCTGAAACGTCGAAAACCCGAAACTGT
GAGGCGCCGAAATCCGAACTCTATCGTGGCGGTTGAACTGCACACCGCCGACGGCAGCGTGATTGAAGCAGAAGCCTGCGATGTGCGTTTCCGCGAGGTTGC
GGATGAAAATGGTCTGCTGCTGCTGAAGCGCAAGCCGTTGCTGATTCAGGCGTTAACCGTACAGGATCATCTCTGATGGTCAGGTCATGGATGAGCAGAC
GATGGTGCAGGATATCTGCTGATGAAGCAGAACAACTTTAACGCCGCTGGCCTGTTCCGATTATCCGAACCATCCGCTGTGGTACACGCTGTGGCAGCCGTACGGC
CTGATGTGGTGGATGAAGCCAAATTTGAAACCCAGCGCATGGTCCAAATGCTGACCGTATCCGCGTGGTACCGGGGATGAGCGAACCGTAAACCG
GAATGGTGCAGCGCATCTAATCACCAGTGTGATCATCTGGTCCGTGGGAATGAATCAGGCCACCGCCGCTAATCAGCAGCGCTGTATCGCTGGATCAATC
TGTGATCTTTCCCGCCGGTGCAGTATGAAGCGCGGCGACACCCAGCCAGTAAATTTGCCCAGTACGCGCGCTGGATGAAGACCGAGCCCTT
CCCGCCTGTCGCAAAATGGTCCATCAAAAATGGCTTTCGCTACCTGGAGAGACGCGCCCGCTGATCTTTGCAAAATACGCCACCGCATGGTAAACAGCTTGGC
GGTTTTGCTAAATACTGCAGCGCTTTCGTCAGTATCCCGCTTTACAGCGCGCTTCGCTGGGACTGGTGGATCAGTCGCTGATTAATATGATGAAAAACGGCAA
CCCGTGGTGGCTTACGGCGGTGATTTTGGCGATACCGCGAACGATCGCCAGTCTGTATGAACGGTCTGGCTTTTGGCGACCGCACCGCATCCAGCGCTGACG
GAAGCAAAACACCCAGCAGCAGTTTTTCCAGTTCCGTTTATCCGGCAAAACCATCGAATGACCCAGCGAATACCTGTTCCGCTATAGCGATAACGAGCTCCTGCACCT
GATGGTGGCGCTGGATGGTAAGCCGCTGGCAAGCGGTGAAGTGCCTCTGGATCTGCTCCACAAGGTAACAGTTGATTGAACCTGCCTGAACTACCCGACGCGGA
GAGGCCCGGGCAACTCTGGCTACAGTACCGTAGTGAACCGCAACCGCAGCCGATGGTCAGAAGCCGGGACATCAGCGCCTGGCAGCAGTGGCGTCTGGCGG
AAAACTCAGTGTGACGCTTCCCGCGCGCTTCCACCCGATCCCGCATCGCCATCGACCCAGCGAAATGGATTTTGCATCGAGCTGAGCGTAATAAGCGTTGGCAATTTAAC
CGCCAGTCAGGCTTTCTTTACAGATGTGGATTTGGCATAAAAAACAATGCTGACGCGCGCTGGCGCATCAGTTACCCGCTGCACCGCTGGATAACGACATTTGGCT
AAGTGAAGCGAACCCGCAATTGACCCTAACCGCTGGGTGCAAGCGTGGAGAGCGCGCGCCATTACCAGGCCGAAGCAGCGTTTGTGACGCTGCAGCGATACACT
TGCTGATGCGGCTGCTGATTACGACCGCTCACGCGTGGCAGCATCAGGGGAAAACTTATTTATCAGCGGAAAAACCTACCGGATGATGGTATGGTCAAATGGCG
ATTACCGTTGATGTTGAAGTGGCGGAGCGGATACACCGCATCCGCGCGGATGGCTGCAACTGCCAGCTGGCGCAGGTAGCAGAGCGGGTAAACTGGCTCGGATTA
GGCGCGCAAGAAAACTATCCGACCGCCTTACTGCGCGCTGTTTTGACCGCTGGGATCTGCCATTGTGAGACATGATACCCCGTACGCTTTCCCGAGCGAAAAACG
GTCTGCGCTGCGGGACCGCGGAATTAATTTGCGCCACACCCAGTGGCGGGGCGACTTCCAGTTCAACATCAGCCGCTACAGTCAACAGCACTGATGAAAACCG
CCATCGCCTCTGCTGCACGGGAAAGGACAGATGGCTGAATCGACGGTTTCCATATGGGGATTGGTGGCGGACGACTCCTGGAGCCCGTCAATGATCGCGGAA
TTCAGCTGAGCGCCGGTGCCTACCATACCAATGGTCTGGTGTCAAAAATAAATAACCCGGCAGGCCATGCTGCCCCGCTTTTCCGTAAGGAAATCCATATG
TACTATTAAAAAACAACAATTTGGATGTTGCGTTTATCTTTTCTTACTTTTATCAGTGGGAGCCTACTTCCCGTTTTTCCCGATTGGCTACATGACATCAAC
CATATCAGCAAAAGTATACCGGTATTAATTTTGGCCGATTTCTGTGTTGCTGCTATTTCCAAACCGCTGTTTGGCTGCTTTCTGACAAAACCTGGGCTGCGCAAT
ACCTGCTGGATTTACCGCATGTAGTGTGATGTTTGGCCGCTTCTTATTTTATCTTCCGGCCACTGTTACAATAACAATTTAGTAGGATCGATGTTGGTGGT
ATTTATTCAGCTTTTGGTTTAAACCGCGTGGCCAGCAGTATGAGGCAATTAATGAGAAAGTACCGCGTGCAGTAATTTGCAATTTGGTCCGCGCGGATTTGGC
TGTGTTGGCTGGCGCTGTGTGCTCGATGTGCGCATCATGTTCCACATCAATAATCAGTTGTTGTTTCTGGCTGGGCTGCGCTGTCACATCTCGCCGTTTTA
CTCTTTTTGCAAAAACCGGATGCGCCCTTCTGCCACGGTGGCAATGCGGTAGGTGCCAACCATTCGCGATTAGCCCTAAGCTGGCAGCTGGAACGTTTGCAGCA
GCCAAAACCTGTGGTTTTTGTCACTGTATGTTATGGCGTTTCTGTGACCATACGATGTTTTGACCAACAGTTTGTCTAATTTCTTACTTCGTTCTTTGCTACCGGTAAAC
AGGGTACGCGGGTATTTGGCTACGTAAACGCAATGGCGCAATTACTTAAACCGCTGATTTATGTTCTTTGCGCCACTGATCATAATCGCATCGGTGGGAAAAACCGCC
CTGCTGCTGGCTGGCAGTATGTCGCTGACGTAATTTGCGCTCATCTGCGCCACTCAGCGCTGGAAGTGGTTATCTGAAAACCGTGCATATGTTGAAGTACCG
TTCCTGCTGGTGGGCTGCTTTAAATATAATACCAGCCAGTTGAAAGCGGTTTTTCAAGCGGATTTATCTGGTCTGTTTCTGCTTTTAAAGCAACTGGCGGATGATTT
TATGCTGTACTGGCGGGCAATATGATGAAAACGATCGTTTTCCAGGCGCCTTATGTTGGTGGGCTGCGGCTGCGGCTTACCTTAATTTCCGTGTTTCCGCT
TTAGCGGCCCGCGCCGCTTTCCCTGCTGCGTGTGAGTGAATGAAGTGCCTTAAAGCAATCAATGTGCGATGCGCGCGGAGCGCCTTATCCGACCAACATATCAT
AACGGAGTGTGCGATTTGAAATGCGCAATGACCGAAAAGATAAGAGCAGGCAAGCTATTACCGATGTCGCGAAGGCTACCGGAAAAAAGACTCTGCGGAAAAAC
GTTAATGATGAGTTTAACTACTCGCATCCATCAGA

But the actual data are:

GTGAAGGGCAATCAGCTGTTGCCGCTCTACTGGTGAAGAAAAAACCCCTGGCGCCCAATACGAAACCGCCTCTCCCCGCGGTTGGCCGATTCAATATGC
AGCTGGCACACAGGTTTCCCAGCTGGAAGAGCGGGCAGTGAAGCCAAACCGCAATTAATGTGAGTTAGCTCACTTAGGCACCCAGGCTTTACACTTTATGCTTCC
GGCTCGTATGTTGTGGAAATTTGAGCGGATAACAAITTCACAGGAAACAGCTTACGACATGATACGGATTCACTGGCCGTCGTTTACAAACGTCGTGACTGGG
AAAAACCTGGCGTACCCAACTTAATCGCTTCCAGCAGTCCCCCTTTCGGCAGCGTGGCCTAATAGCGAAAGAGGCCCGCACCGGATCGCCCTTCCCAACAGTTGGC
CAGCCTGAATGGCGAATGGCGCTTTGGCTGTTTTCCGATCCAGGAGCGGAGCGGTCGCGGAAAGCTGGCTGGAGTGGCATCTTCTGAGGGCGATACTGCTGCTGCC
CTAAACCTGGCAGATGCACGGTTACGATGCGCCCATACACCAACGTGACCTATCCATACGGTCAATCCGCGCTTTGTTCCACGGGAATCCGACGGGTTGT
ACTCGCTCACATTTAATGTTGATGAAAGCTGGCTACAGGAAGGCCAGACGCGAATATTTTTGATGGCGTTAACTCGCGGTTTCATCTGTGGTCAACGGCGGCTGG
GTCGGTTACGGCCAGGACAGTCGTTTCCGCTGTAATTTGACCTGAGCGCATTTCACGCGCCGAGAAAAACCGCCTCGCGGTGATGGTGTGCGGCTGGAGTGACG
GCAGTTATCTGGAAGATCAGGATATGAGCGGATGAGCGCGATTTCCTGTGACGTCTGCTTGTGCATAAACCGACTACACAAATCAGCGATTCCATGTTGCCACTC
GCTTAATGATGATTTTACGCGCGCTGTACTGAGGCTGAAGTTAGATGTCGGCGAGTTCGCTGACTACCTACGGGTAACAGTTTCTTATGGCAGGGTGAAACG
CAGGTCGCCAGCGGCACCGCGCTTTCGGCGGTGAATATCGATGAGCGTGGTGGTTATCCGATCGCGTACACTACGCTGAAACGTCGAAACCCGAAACTGT
GAGCGCCGAAATCCGAACTCTATCGTGGCGGTTGAACTGCACACCGCCGACCGCAGCGTATTGAAGCAGAAAGCCTGCGATGTCGTTTCCGCGAGGTCG
GGATGAAATGGTCTGCTGCTGAAGCGCAAGCGGTTGCTGATTGAGCGGTTAACCGCTAACCGCTACGAGCATCATCTGATGGTCAGGTGATGGATGAGCAGAC
GATGGTGCAGGATATCTGCTGATGAAGCAGAACTTAACCGCGTGGCTGTTCCGATTATCGAACCATCCGCTGGTACACGCTGCGGACCGCTAGCGG
CTGATGTGGTGGATGAAGCAATTTGAAACCCAGCGCATGTCGCAATGAATCGCTGACCCGCTGCGGCTGACTACCGGGATGAGCGAAACCGTAAACG
GAATGGTCAGCGGATCGTATACCCGAGTGTGATCATCTGCTGCTGGGAATGAATCAGCCACCGCGCTATACGACGCGCTGATCGCTGGATCAAACT
TGTCATCTTCCCGCCGGTACGATGAAGCGCGGAGCGACACCGCAACCGCATATATTTGCCGATGACGCGCGCTGGATGAAGACACGCCCCT
CCCGCTGTCCGCAAAATGGTCCATCAAAAATGGCTTTCGCTACCTGGAGAGACGCGCCCGCTGATCTTTGCCAATACGCCACCGCATGGTAAACGCTTGGC
GGTTTCGCTAAACTCGCAGCGCTTTCGTCAGTATCCCGCTTTACAGCGCGCTTCGCTGGGACTGGTGGATCAGTCGCTGATTAATATGATGAAACGGCAA
CCCGTGTGCTGGCTTACGGCGGTGATTTGGCGATACCGCGCAAGCAGTCCGCGATTCTGTATGAACGGTCTGGCTTTGCCGACCGCAACCGCATCCAGCGCTGACG
GAAGCAAAACACCCAGCAGCATTTTCCAGTTCCGTTATTCGGGCAAAACCGAATGACCAAGCGAATACCTGTTCCGTCATGCGATAACGAGCTCCTGCACCT
GATGGTGGCGCTGGATGGTAAGCCGCTGGCAAGCGGTTGAAGTCCCTGCTGCTCCACAAGGTAACACAGTTGATGAAGTCGCTGAACTACCGCAGCCGGA
GAGCGCCGGGCAACTCTGGCTCACAGTACCGGTAGTGAACCGCAAGCGCGGATGGTCAGAAGCCGGGACATCAGCGCCTGGCAGCAGTGGCGCTCGCGG
AAAACTCAGTGTGACGCTCCCGCGCGCTCCACCGCATCCCGCATCGACACCGCAAAATGGATTTTGCATCGAGCTGGGTAATAAGCGTTGGCAATTAAC
CGCCAGTCAGGCTTTTCCAGATGGTGGTGGCATAAAAAACAACTGCTGACCGCGCTGCGCGATCAGTTACCCGCTGCACCGCTGGATAACGACATTTGGCCT
AAGTGAAGCGACCGCATTGACCCCTAACCGCTGGGTGCAACCGCTGGAAGCGCGGCTTACCAGCCGCAAGCAGCGTTGTTGCAGTGCACGGCAGATACCT
TGCTGATGCGGTGCTGATTCAGCAGCGCTACCGGTTGGCAGCATCAGGGGAAAACTTATTTATCAGCGGAAAAACCTACCGGATTGATGGTAGTGGTCAAATGGCG
ATTACCGTTGATGTTGAAGTGGGAGCGGATACCCGATCCCGCGCGGATGGCTGCACTGCCAGTGGCGCAGGTAGCAGAGCGGGGTAACCTGGCTCGGATTA
GGCGCGAAAGAAAATATCCGACCGCTTACTGCGCGCTGTTTACCGCTGGGATTCGCCATGTCAGACATGATACCCCGTACGCTTCCCGAGCGAAAAACG
GTCTGCGCTGCGGACCGCGGAATGAATATGGCCACACAGTGGCGGGGCACTCCAGTTCACATCAGCCGCTACAGTCAACAGCAACTGATGAAACCG
CCATCGCATCTGCTGCACGGGAAGGACACATGGCTGAATCGACGGTTCCATATGGGATGGTGGCGGACGCTCCTGGAGCCGCTCAGTATCGCGGAA
TTCAGCTGAGCGCGGCTGCTACCATACCAATGGTGTGGTGTGCAAAAATAAATAAACCCGGCAGGCCATGCTGCGCGTATTCGCGTAAAGAAATCCATTAG
TACTATTAAAAAACAACAACTTTGGATGTCGGTTTATCTTTTCTTACTTTTATCATGGGAGCTACTTCCGCTTTTCCCGATTGGCTACATGACATCAAC
CATATCAGCAAAAGTGATACCGGATTAATTTTGGCCGATTTCTGCTGCTGCTATTTACCAACCGCTGTTTGGCTGCTTTGACAAACTCGGGCTGCGCAAT
ACCTGCTGGGATATACCGCATGTAGTGATGTTGGCGGCTTCTTATTTTACTTTCGGCCGCTGTTACAATAACAACTTTAGTAGGATGATGTTGGTGGT
ATTTATTAGCTTTTGGTTTAAACCGCGGTGCCCAAGCAATGACCGAAAGATAAGAGCAGGCAAGCTATTACCGATGTCGCAAGGCTTACCGGAAAAAGACTCTGGGAAAAA
TGTTGGCTGGCGCTGTGCTCGTATGTCGGCATCATGTTACCATCAATAATCAGTTGTTTCTGGCTGGGCTGCTGGCTGTCACATCATCTCGCGTTTTA
CTCTTTTTGCCAAAACGGATCGCGCTTCTGCCAGGTTGCCAATGCGGTAGTGGCCACCATTCGCGATTAGCCCTTAAAGCTGGCAGCTGGAACGTTGCAGACA
GCTTTACTGTGGTTTTGTCACTGTATGTTTGGCTTCTCTGCCACTACGATGTTTGTGACCAACAGTTTGTCTAATTTCTTACTCTGTTCTTGTACTCCGGTGAAC
AGGGTACGCGGGTATTTGGCTACGTAAACGACAAATGGCGGAATTAACCGCCCTGATTTGTTCTTGGCCACTGATCATTAATCGCATCGGTGGGAAAAACCGCC
CTGCTGCTGGCTGGCACTATATGCTGCTGACTATTTTGGCTCATCTGCGCCACTCAGCGCTGGAAGTGGTTATCTGAAAACCGTGCATATGTTGAAGTACCG
TTCTGCTGGTGGGCTGCTTTAAATATATACCAAGCAATGTTGAAGTGGGTTTTACGGAGCATTATCTGGTCTGTTTGTCTTTTAAAGCAACTGGCGATTGATTT
TATGCTGACTGGCGGGAATATGATGAAAGCATCGTTTCCAGGGCGCTTACTGGTGGTCTGGTCTGGTGGCGCTGGGCTTACCTTAATTTCCGTTCTCACCG
TTAGCGCGCCCGCGCCCTTCCCTGCTGCGTGCAGGTGAATGAAGTGCCTTAAAGCAATCAATGTCGGATGCGCGCGGACCGCTTATCCGACCAACATATCAT
AACGGAGTGCATGCAATTAACCATGCAATGACCGAAAGATAAGAGCAGGCAAGCTATTACCGATGTCGCAAGGCTTACCGGAAAAAGACTCTGGGAAAAA
GTTAATGATGAGTTTAACTACTCGCATCCATCAGA



Bioinformatics solutions

```
GTGAAGGGCAATCAGCTGTTGCCGCTCTACTGGTGAAGAAAAAACCCCTGGCGCCAAATACGCAACCGCCTCTCCCCGCGGTTGGCCGATTCAATATGC
AGCTGGCACGACAGGTTTCCCAGCTGGAAGACGGGGCAGTGAAGCCAAACCGCAATTAATGTGAGTTAGCTCACTTATTAGGCACCCAGGCTTTACACTTATGCTTCC
GGCTCGTATGTTGTGGAAATTTGAGCCGGATAACAATTTACACAGGAAGAACAGCTTACCATTGATACGGATTCACTGGCCGTCGTTTACAAACGTCGTGACTGGG
AAAACCCCTGGCGTTACCCAACTTAATCGCCTTGACAGCATAACCCCTTTCCGCGAGCTGGCGTAATAGCGAAAGAGGCCCGCACCGGATCGCCCTTCCCAACAGTTGGC
CAGCCTGAATGGCGCAATGGCCGTTTGGCTGTTTCCGTCACCCAGCAAGGCGGTGGCCGAAAGCTGGCGTGGAGTGGCATTCTTCCGAGGCCGATACTGCTCGTCC
CTAAACTGGCAGATGCACGGTTACGATGCGCCCATACACCAACGTGACCTATCCATTACGGTCAATCCCGCTTGTGCCACGGGAATCCGACGGGTTGT
ACTCGCTCACATTAATGTTGATGAAAGCTGGCTACAGGAAGGCCAGACGCGAATATTTTTGATGGCGTTAACTCGCCGTTTCACTGTGGTGAACGGCCGCTGG
GTCGGTTACGGCCAGGACAGTCGTTTCCGCTCTGAAATTTGACCTGAGCGCATTTCACGCGCCGAGAAAAACCGCCTCGCCGTTGATGGTGCCTGGCTGGAGTGAC
GCAGTTATCTGGAAGATCAGGATATGTTGGCGGATGAGCGCCATTTCCGTGACGCTCGTTGCTGCATAAACCGACTACACAAATCAGCGATTTCCATGTTGCCACTC
GCTTAATGATGATTTTACGCGCGCTGACTGGAAGCTGAAGTTAGATGCGCGGAGTTGCGTGACTACCTACGGGTAACAGTTTCTTTATGGCAGGGTGAACG
CAGGTCGCCAGCGGCACCCGCGCTTTCCGCGGTGAATATCGATGAGCGTGGTGGTTATGCCGATCGCGTACACTACGCTGAAACGTCGAAAACCCGAAACTGT
GGAGCGCCGAAATCCCGAATCTATCGTGGCGTGGTGAACGTGCACACCGCCGACGGCAGCGCTGATTGAAGCAGAAGCCTGCCGATGTCGTTTCCGCGAGGTCG
GGATTGAAAATGGTCTGCTGCTGAACGGCAAGCCGTTGCTGATTGAGGCGTTAACCGTACAGGATCATCTCTGATGGTCAGGTATGGATGAGCAGAC
GATGGTGCAGGATATCTGCTGATGAAGCAGAACAACTTTAACCGCTGGCTGTTCCGATTATCGAACCATCCGCTGTTGGTACACGCTGCGGACCCGCTAGGCG
CTGATGTTGGTGAAGCAATATTTGAAACCCAGCGCATGGTCCAACTGCTGACCCGATGTCGCGGTGGCTACCGGGGATGAGCGCAACGTAACCG
GAATGGTCAGCGCATGTAATCCCGAGTGTGATCATCTGGTCCGTTGGGAATGAATCAGGCCAGCGGCTAATCAGACGCGCTGATCGCTGGATCAAACT
TGTCATCTTCCGCGCGTAGTGAAGCCGGCGAGCGACACCGCCACCGAATATTTGCCGATGACGCGCGCTGGATGAAGACCCAGCCCTT
CCCGGCTGCGCAAAATGGTCCATCAAAAATGGCTTTCGCTACCTGGAGAGACGCGCCGCTGATCCTTTGCAATACGCCACGCGATGGTAAACAGTCTTGGC
GGTTTCGCTAAACTGCGACGGCTTTCGTCAGTATCCCGCTTTACAGCCGCGCTTCGCTGGGACTGGTGGATCAGTCGCTGATTAATATGATGAAAACGGCAA
CCCGTGGTGGCTTACGGCGGTGATTTGGCGATACCGCGAACGATCGCCGATCTGTATGAACGGTCTGGCTTTGCCGACCGCACCGCATCCAGCGCTGACG
GAAGCAAAACACCAGCAGCAGTTTCCAGTTCCGTTTATCCGGCAAAACCGAATCGAATGACCCAGCGAATACCTGTTCCGTCATAGCGATAACGAGCTCCTGCAC
GATGGTGGCGCTGGATGGTAAGCCGCTGGCAAGCGGTGAAGTCCCTGTCGCTCCACAAGGTAAACAGTTGATTGAACTCGCTGAACTACCGCAGCCGGA
GAGCGCCGGGCAACTCTGGCTACAGTACGCGTAGTGAACCGCAACCGCAGCCGATGGTCAGAAGCCGGGACATCAGCGCCTGGCAGCAGTGGCGTCTGGCGG
AAAACCTCAGTGTGACGCTCCCCGCGCTTCCGCGCCGATCCCGCATCCCGCATCCACCCAGCGAAATGGATTTTGCATCGAGCTGGGTAATAAGCGTTGGCAATTAAC
CGCCAGTCAGGCTTTCTTTCACAGATGGTGGTGGGCAAAAAAACAATGCTGACGCGCGCTGGCGCATCAGTTACCCGCTGCACCGCTGGATAACGACATTTGGCT
AAGTGAAGCGAACCCGCAATTGACCCCTAACCGCTGGCTGCAACGCTGGAAAGCGCGGCGCCATTACCAGCGCCGAAGCAGCGTTTGGCAGTGCACGACATACACT
TGCTGATGGCGTCTGATTACGACCGCTCAGCGGTGGCAGCATCAGGGGAAAACTTATTTATCAGCCGAAAAACCTACCGGATTGATGGTAGTGGTCAAATGGCG
ATTACCGTTGATGTTGAAGTGGCGAGCGATACACCGCATCCGCGCGGATGGCTGAACTGCCAGCTGGCCGAGGTAGCAGAGCGGGTAAACTGGCTCGGATTA
GGCGCGCAAGAAAACTATCCGACCGCCTTACTGCGCGCTGTTTGAACCGCTGGGATGTCGCATGTCAGACATGATACCCCGTACGCTTCCCGAGCGAAAAACG
GTCTGGCTGCGGGACCGCGGAATTAATATGGCCACACCGCAGTGGCGGGGCACTCCAGTTCAACATCAGCCGCTACAGTCAACAGCACTGATGGAACCAAC
CCATCGCCTCTGCTGCACGGGAAGGACACATGGCTGAATATCGACGGTTTCCATATGGGGATGGTGGCGGAGCTCCTGGAGCCGCTCAGTATCCGGCGGAA
TTCAGCTGAGCCCGGTGCTACCATACCAATGGCTGTTGTTGCTGCAAAAATAATAAACCCGGGACGGCCATGCTGCCCGTATTTCCGTAAGGAAATCCATATG
TACTATTAAAAAACCAAACTTTGGATGTTGCGTTTATCTTCTTACTTTTATCATGGGAGGCTACTTCCGCTTTCGCAATTTGGCTAGCATGACATCAAC
CATATCAGCAAAAGTACAGCGGATTAATTTTGGCCGATTTCTGCTGCTGCTATTTACCCGCTGTTTGGCTGCTTTTCGAAACTCAGGCTGGGCTGGCAAACT
ACCTGCTGGATATACCGCATGTTAGTGATGTTTGGCCGCTTCTTATTTTATCTTCCGGCCACTGTTACAATAACAATTTAGTAGGATGATGTTGGTGGT
ATTTATCAGCTTTTGGTAAACCGCGTGGCCAGCAAGCAATTAATGAAAGATCAGCCGCTGCAGTAAITTCGAAATTTGGCTGGCGCGGATGTTGGC
TGTGTTGGCTGGCGCTGTGTCCTGATGTCGGCATCATGTTACCATCAATAATCAGTTGTTTCTGGCTGGGCTCGGCTGTCACATCATCTCGCCGTTTTA
CTCTTTTTGCAAAACGGATGCGCCCTCTTCCGACGCTGTCACATGCGGTAGTGGCCACCACTCCGCATTTAGCCCTAAAGCTGGCAGCTGGAACGTTGCAGCA
GCCCCAATCTGGTTTTGTCACCTGTATGTTATGGCTTCTCTGACCATACGATGTTTTGACCAACAGTTTGCATATTTCTTACTCTGTTCTTGTCTACCGGTAAC
AGGGTACGCGGGTATTTGGCTACGTAAACGCAATGGGCGAATTACTTAAACCGCTCGATTATGTTCTTCCGCACTGATCATTAATCGCATCGTGGGAAAAACCGCC
CTGCTGCTGGCTGGCACTATATGCTGCTGACTGATTTGGCTCATCTGCGCCACTCAGCGCTGGAAGTGGTATTCTGAAAAACCGTGCATATGTTGAAGTACCG
TTCTGCTGGTGGGCTGCTTTAAATATATACCAACCGCAATTTGAAAGCGGTTTTTCAAGCGAGGATTTATCGGCTGTTTCTGCTTCTTTAAGCAACTGGCGATTGATTT
TATGCTGACTGGCGGGCAATATGATGAAAGCATCGTTTCCAGGCGCCTTATCGTGGTGGGCTGCTGGTGGCGCTGGGCTCACCTTAATTTCCGCTGTTCCAGC
TTAGCGGCCCGCGCCGCTTTCCCTGCTGCGTGCAGGTGAATGAAGTCGCTTAAAGCAATCAATGTCGGATGCGCGCGGACCGCCTATCCGACCAACATATCAT
AACGGAGTGCATGCAATTAACATGCAATGACCGAAAGATAAGAGCAGGCAAGCTAATTTACCGATGTCGCAAGGCTTACCGGAAAAAGACTCTGCGGGAAAAAC
GTTAATGATGAGTTAATCACTCGCATCCATCAGA
```

Two kinds of approaches:

- Alignments methods:

Aligning our sequence over 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

```
GTGAAGGGCAATCAGCTGTTGCCGCTCTACTGGTGAAGAAAAAACCCCTGGCGCCCAATACGCAAAACCGCTCTCCCCGCGGTTGGCCGATTCAATATGC
AGCTGGCACGACAGGTTTCCCAGCTGGAAGACGGGGCAGTGAAGCCAAACGCAATTAATGTGAGTTAGCTCACTTATTAGGCACCCAGGCTTTACACTTATGCTTCC
GGCTCGTATGTTGTGGAAATTTGAGCGGATAACAATTTACACAGGAAACAGCTTACCATTGATACGCGTACTGGCCGTCGTTTACAAACGTCGTGACTGGG
AAAAACCCCTGGCGTTACCCAACTTAATCGCCTTGACAGCATAACCCCTTTGCCAGCGTGGCGTAAATAGCGAAAGAGGCCCGCAGCGATCGCCCTTCCCAACAGTTGGC
CAGCCTGAATGGCCGAATGGCCCTTTGGCTGAGTTTCCGTCAGCCACGCAAGCGGAGCGGAAAGCTGGCGTGGAGTGGCATCTTCTGAGGGCGATACTGCTGCTGCC
CTAAACCTGGCAGATGCACGGTTACGATGCGCCCATACACCAACGTGACCTATCCATTACGGTCAATCCCGCTTTGTTCCACGGGAATCCGACGGGTTGT
ACTCGCTCACATTTAATGTTGATGAAAGCTGGCTACAGGAAGGCCAGACGCGAATATTTTTGATGGCGTTAACTCGCGCTTTCATCTGTGGTCAACGGCGCTGG
GTCGGTTACGGCCAGGACAGTCGTTTCCGCTGTAATTTGACCTGAGCGCATTTCACGCGCCGAGAAAAACCGCTCGCCGTTGATGGTGCCTGGCTGGAGTGACG
GCAGTTATCTGGAAGATCAGGATATGTTGGCGGATGAGCGGCATTTCCGTGACGTCTGCTTTGCTGCATAAACCGACTACACAAATCAGCGATTTCCATGTTGCCACTC
GCTTAAATGATGATTTTACGCGCGCTGACTGGAAGCTGAAGTTAGATGTCGGCGAGTTGCGTGACTACCTACGGGTAACAGTTTCTTATGGCAGGGTGAACG
CAGGTCGCCAGCGGCACCCGCGCTTTCCGCGGTGAATATTCGATGAGCGTGGTGGTTATGCCGATCGCGTACACTACGCTGAAACGTCGAAAACCCGAAACTGT
GGAGCGCCGAAATCCGAACTCTATCGTGGCGTGGTTGAACGTGCACACCGCCGACGGCAGCGTGATTGAAGCAGAAGCCTGCCGATGTCGTTTCCGCGAGGTGC
GGATGAAAATGGTCTGCTGCTGAACGGCAAGCCGTTGCTGATTGAGGGCTTAAACCGTACAGGATCATCTCTGATGGTCAGGTATGGATGAGCAGAC
GATGGTGCAGGATATCTGCTGATGAAGCAGAACAACTTTAACCGCTGGCTGTTCCGATATCGAACCATCCGCTGGTGAACCGCTGCGGACCCGACAGGC
CTGATGTTGGTGAAGCAATATTTGAAACCCAGCGCATGGTCCAAATGCTGACCCGATGTCGGCTGGCTACCGGGATGAGCGAACCGTAAACG
GAATGGTCAGCGCATGTAATCCCGAGTGTGATCATCTGGTCCGTTGGGAATGAATCAGCCACGGCCGCTAATCAGACGGCTGATCGCTGGATCAATC
TGTCATCTTCCGCGCAGTATGAAGCGCGGAGCGGCACACCCGATATATTTGCCGATGACGCGCGCTGGATGAAGACCCAGCCCTT
CCCGCTGTCGCAAAATGGTCCATCAAAAATGGCTTTCGCTACCTGGAGAGACGGCCCGCTGATCCTTTGCAATACGCCACGCGATGGTAAACGCTTGGC
GGTTTCGCTAAACTGCGAGCGGTTTCGTCAGTATCCCGCTTTACAGCCGCGCTTCGCTGGGACTGGTGGATCAGTCGCTGATTAATATGATGAAAACGGCAA
CCCGTGGTGGCTTACGGCGGTGATTTTGGCGATACCGCGAACCGACTGCGCAGTCTGTATGAACGGTCTGGCTTTTCCGCAACCGCACCCGCTCCAGCGCTGACG
GAAGCAAAACCCAGCAGCAGTTTCCAGTTCGCTTATCCGCGCAAAACCGAATCGAATGACCAAGCAATACCTGTTCCGCTACATCGATAAAGGAGCTCTGCACCTG
GATGGTGGCGCTGGATGGTAAGCCGCTGGCAAGCGGTGAAGTGCCTCTGCTCCACAAGGTAACAGTTGATTGAACCTGCCTGAACTACCCGACGGCAG
GAGCGCCGGGCAACTCTGGCTACAGTACCGTAGTGCAACCGAACCGGACCGGATGGTCAGAAGCCGGGACATCAGCGCCTGGCAGCAGTGGCGCTGGCGG
AAAAACCTCAGTGTGACGCTTCCCGCGCGCTCCACCGCATCCCGCATGACACCCAGCGAAATGGATTTTGCATCGAGCTGGGTAATAGCGTTGGCAATTTAAC
CGCCAGTCAGGCTTTTCTTACAGATGGTGGTGGGCAAAAAAACAATGCTGACGCGCGCTGGCGATCAGTTACCCGCTGCACCCGCTGGATAACGACATTTGGCT
AAGTGAAGCGACCCGCAATTGACCCCTAACCGCTGGCTGCAAGCGTGGAAAGCGCGGCGCCATTACCAGGCCGAAGCAGCGTTTGGCATGCGCATACATCACT
TGCTGATGGCGTCTGATTACGACCGCTCAGCGGTGGCAGCATCAGGGGAAAACTTATTATCAGCCGAAAAACCTACCGGATGATGGTAGTGGTCAAATGGCG
ATTACCGTTGATGTTGAAGTGGCGAGCGATACCCGATCCCGCGCGGATGGCTGAACTGCCAGCTGCCAGCTAGCAGAGCGGGTAAACTGGCTCGGATTA
GGCGCGCAAGAAAAATATCCGACCGCCTTACTGCGCGCTGTTTGAACCGTGGGATGTCGCATGATACCCGCTGACGCTTCCCGAGCGAAAAACG
GTCTGGCTGCGGGACCGCGAATTGAATATGGCCACACCGAGTGGCGGGGACTTCCAGTTCAACATCAGCCGCTACAGTCAACAGCAGCATGATGGAACCCAG
CCATCGCCTCTGCTGCACGGGAAGGACAGATGGCTGAATATCGACGGTTTCAATATGGGGATTGGTGGCGAGCCTCCTGGAGCCGCTCAGTATCGCGGGAA
TTCAGACTGAGCGCCGGTGCCTACCATACCAATGGCTGGTGTCAAAAATATAAACCCGGCAGGCCATGCTGCGCCGATTTCCGCTAAGGAAATCCATATG
TACTATTAAAAAACATAACTTTGGATGTTGGTATTTCTTCTTACTTTTATCAGGGAGCTACTTCCGCTTTTCCCGATTGGCTACATGACATCAAC
CATATCAGCAAAAGTACAACGGATATTTTTGCGCGTATTTCTGCTGCTGCTATTTCAACCCGCTTTTGGTCTGCTTTCTGCAAACTCGGGCTGGCAAA
ACCTGCTGGATATACCGCATGTAGTGTGTTTGGCGGCTTCTTATTTTCTTCCGGCCACTGTTACAATAACAATTTAGTAGGATGATGTTGGTGGT
ATTTATCAGCTTTTGGTTAAGCGCGGTGCCCGAGCAGTGAAGCAATTTAGGAAAGTACGCGCTGCAGTAATTTGCAATTTGGTGGCGCGGATGTTTGGC
TGTGTTGGCTGGCGCTGTGCTCGTATGTCGGCATCATGTTACCATCAATAATCAGTTGTTTCTGGCTGGGCTCGGCTGTCACATCATCTCGCGTTTTA
CTCTTTTTCCGCAAAACGGTGGCCCTCTTCCGCAAGTGGCCAGTGGTGGCCACCATTCGCGATTTAGCCCTTAAGCTGGCAGTGGCAACTGTTGAGCA
GCCAAAACCTGGTTTTGTCAGTGTATTTGGCTGTTCTGCGCACTTCGCTGTTTGGCAACAGTTTGGCTAATTTCTTACTTCGTTCTTGTACTCGGTGTAAC
AGGGTACCGCGGATTTGGCTACGTAAACGACAAATGGGCGAATTAACCGCCCTCGATTGTTCTTTGGCCGCACTGATCATAATCGCATCGTGGGAAAAACCGCC
CTGCTGCTGGCTGGCAATATGTCGCTGACGATATTTGGCTCATCTGCGCCACTCAGCGCTGGAAGTGGTATTTCTGAAAACCGTGCATATGTTGAAGTACCG
TTCCTGCTGGTGGGCTGCTTTAAATATATACCAGCCAGTTGAAAGTGGGTTTTTCAAGCGAGTTATCTGGCTGTTTCTGCTTCTTTAAGCAACTGGGATGATTTT
TATGCTGTACTGGCGGCAATGTATGAAAGCATCGTTTTCCAGGCGCCTTATGTTGGTGGTCTGGTGGCGCTGGGCTTCACTTAAITTCGTTTACGCTTACG
TTAGCGCGCCCGCCGCTTTCCTGCTGCGTGCAGTGAATGAAGTGCCTTAAAGCAATCAATGTCGGATGCGCGCGGACCGCCTTATCCGACCAATATCAT
AACGGAGTGTGCTGATTAAGCATGCCAATGACCGAAAGATAAGAGCAGGCAAGCTATTACCAGATGTGCGAAGGCTTACCGGAAAAAGACTTCTGGGAAAAA
GTTAATGATGAGTTAATCACTCGCATCCATCAGA
```

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

Bioinformatics solutions

```
GTGAAGGGCAATCAGCTGTTGCCGCTCTACTGGTGAAGAAAAAACCCCTGGCGCCAAATACGCAAAACCGCTCTCCCCGCGGTTGGCCGATTCAATATGC
AGCTGGCACGACAGGTTTCCCAGCTGGAAGACGGGGCAGTGAAGCCAAACGCAATTAATGTGAGTTAGCTCACTTATTAGGCACCCAGGCTTTACACTTATGCTTCC
GGCTCGTATGTTGTGGAAATTTGAGCGGATAACAATTTACACAGGAAGAACAGCTTACCATTGATACGGATTCACTGGCCGTCGTTTACAAACGTCGTGACTGGG
AAAAACCCCTGGCGTTACCCAACTTAATCGCCTTGACAGCATCCCCCTTTCCGCGAGCTGGCGTAATAGCGAAAGAGGCCCGCACCGATCGCCCTTCCCAACAGTTGGC
CAGCTCGAATGGCCGAATGGCCGTTTGGCTGTTTTCCGTCACCCAGCAAGCGGAGTGGCCGAAGAGCTGGCGTGGAGTGGCATTCTTCCGAGGCCGATACTGTCGTGCC
CTCAAACCTGGCAGATGCACGGTTACGATGCGCCCATACACCAACGTGACCTATCCATTACGGTCAATCCCGCTTTGTTCCACGGGAATCCGACGGGTTGTT
ACTCGCTCACATTTAATGTTGATGAAAGCTGGCTACAGGAAGGCCAGACGCGAATATTTTTGATGGCGTTAACTCGCCGTTTCACTGTGGTGAACCGGGCGCTGG
GTCGGTTACGGCCAGGACAGTCGTTTCCGCTCTGAATTTGACCTGAGCGCATTTCACGCGCCGAGAAAAACCGCTCGCCGTTGATGGTGCCTGGCTGGAGTGACG
GCAGTTATCTGGAAGATCAGGATATGTTGGCGGATGAGCGCCATTTCCGTGACGTCTGCTTGGCTGCATAAACCCGACTACACAAATCAGCCGATTTCCATGTTGCCACTC
GCTTAAATGATGATTTTACGCGCGCTGTACTGGAAGCTGAAGTTAGATGTCGGCGAGTTGCGTGACTACCTACGGGTAACAGTTTCTTATGGCAGGGTGAACG
CAGGTCGCCAGCGGCACCCGCGCTTTCCGCGGTGAATATCGATGAGCGTGGTGGTTATGCCGATCGCGTACACTACGCTGAAACGTCGAAAACCCGAAACTGT
GGAGCGCCGAAATCCCGAATCTCTATCGTCCGGTGGTTGAACGTGCACACCGCCGACGGCAGCGTGATTGAAGCAGAAGCCTGCCGATGTCGTTTCCGCGAGGTCG
GGATTGAAATGGTCTGCTGCTGCTGAACGGCAAGCCGTTGCTGATTCCAGGCGTTAACCGCTCACGAGCATCATCTCGCATGGCTCAGGTCATGGATGAGCAGAC
GATGGTGCAGGATATCTGCTGATGAAGCAGAACAACTTTAACCGCTGGCTGTTCCGATTATCGAACCATCCGCTGGTGAACCGCTGCGGACCCGTCAGGCC
CTGATGTTGGTGAAGCAACATTTGAAACCCAGCGCATGGTCCAAATGCTGACCGATATCCCGGCTGGCTACCGGGATGAGCGAACCGTAAACCG
GAATGGTGCAGCGCATGTAATCCCGAGTGTGATCATCTGGTCCGTTGGGAATGAATCAGGCCACGGCCGCTAATCAGACCGCTGATCGCTGGATCAAACT
TGTCATCTTCCCGCCGTCAGTATGAAGCGCGGAGCGGCACACCCGATCATGACCCGATATTTGGCCGATGACGCGCGCTGGATGAAGACCCAGCCCTT
CCCGCTGTGCCAAATGGTCCATCAAAAATGSCCTTTCGCTACCTGGAGAGACGGCCCGCTGATCCTTTGCCAATACGCCACCGATGGTAAACGATCTTGGC
GGTTTTCCGTAATACTGCGAGCGGTTTCCGTCAGTATCCCGCTTTACAGCCGCGCTTCGCTGGGACTGGTGGATCAGTCGCTGATTAATATGATGAAAAACGGCAA
CCCGTGGTCCGCTTACGGCGGTGATTTTGGCGATCCCGGCAACGATCGCCAGTCTGTATGAACGGTCTGGCTTTTCCGCAACCGCACCCGATCCAGCGCTGACG
GAAGCAAAACCCAGCAGCAGTTTTTCCAGTTCGCTTATCCGCGCAAAACCATCGAAGTGACCAAGCAATACCTGTTCCGCTCATAGCGATAACGAGCTCTGCACCTG
GATGGTGGCGCTGGATGGTAAAGCCGCTGGCAAGCGGTGAAGTGCCTCTGATGTCGCTCCACAAGGTAACAGTTGATTGAACCTGCCTGAACTACCCGACGGCGGA
GAGGCGCCGGGCAACTCTGGCTACAGTACCGTAGTGCAACCGAACCGGACCGGATGGTCAGAAGCCGGGACACATCAGCGCCTGGCAGCAGTGGCGCTCGGGCG
AAAACTCAGTGTGACGCTTCCCGCCGCTCCACCCGATCCCGCATCCGACCCAGCGAAATGGATTTTGCATCGAGCTGGGTAATAGCGCTTGGCAATTTAAC
CGCCAGTCAGGCTTTCTTTCACAGATGGTGGTGGGCAAAAAAACAATGCTGACGCGCGCTGGCCGATCAGTTACCCGCTGCACCCGCTGGATAACGACATTTGGCT
AAGTGAAGCGACCCCGCATTGACCCCTAACCGCTGGCTGCAACCGCTGGAAGCGCGGCTTACCAGGCCGAAGCAGCGTTTGGCAGTCAACAGACATACACT
TGCTGATGCGGCTGCTGATTACGACCGCTCACGCGTGGCAGCATCAGGGGAAAACTTATTATCAGCCGGAAAAACCTACCGGATGATGGTAGTGGTCAAATGGCG
ATTACCGTTGATGTTGAAGTGGCGGAGCGATACCCGATCCGCGCGGATGGCTGAACTGCCAGCTGCCAGCTGGCGAGGTAGCAGAGCGGGTAAACTGGCTCGGATTA
GGGCGCAAGAAAACTATCCGACCGCCTTACTGCGCGCTGTTTTGACCGCTGGGATCTGCCATTGTGACAGCATGATACCCCGTACGCTTCCCGAGCGAAAAACG
GTCTGCGCTGCGGGACCGCGAATTGAATATGGCCACACAGTGGCGGGGCACTCCAGTTCAACACTCAGCGCTACAGTCAACAGCAACTGATGGAACGCAAG
CCATCGCCATCTGCTGCACGGGAAGGACAGATGGCTGAATATCGACGGTTTCCATATGGGGATTGGTGGCGAGCACTCCTGGAGCCGCTCAGTATCGGCCGAA
TTCAGCTGAGCGCCGGTGCCTACCATTAACAGTTGGTCTGGTGCATAAAATAAATACCCGGGAGGCCATGCTGCGCCGATTTCCGCTAAGGAAATCCATTA
TACTATTAAAAAACACAACCTTTGGATGTTGCGGTTATGCTTTTCTTACTTTTATCATGGGAGCTACTTCCGCTTTTCCCGATTGGCTACATGACATCAAC
CATATCAGCAAAAGTACACGGATATTTTTGCGGCTATTTCTGCTGCTGCTATTTCCAAACCGCTTTTGGTCTGCTTTCTGCAAACTCGGGCTGGCAAACT
ACCTGCTGGGATATACCGCATGTAGTGTGTTTGGCGGCTTCTTATTTTATCTTCCGGCCACTGTTACAATAACAATTTAGTAGGATGATGTTGGTGGT
ATTTATCTAGCTTTTGTTTAACCGCGGTGCCCCAGCAGTAGAGCAATTTAAGCAAAATCAGCCGCTGCAGTAAITTCGAATTTGGCTGCGCGGATGTTTGGC
TGTGTTGGCTGGGCGCTGTGCTCCGATGTCGGCATCATGTTACCATCAATAATCAGTTGTTTCTGGCTGGGCTCGGCTGTGCACTCATCTCGCCGTTTTA
CTCTTTTTCCGCAAAACGGATGGCCCTTCTCCGACGGTGGCCAAATCGCGTAGTGGCCACCACTTCGCGATTAGCCCTTAAGCTGGCAGCTGTTGACAG
GCCAAAACCTGGTTTTTGTCACTGTATGTTTGGCTTCTCGCACCTACGATGTTTGGACCAACAGTTTGGCTAATTTCTTACTTCGTTCTTGTCTACCGGTGAAC
AGGGTACCGCGGGTATTTGGCTACGATGACACAAATGGGCGAATCTAACCGCCCTGATGTTGTTCTTGGCCCACTGATCATTAATCGCATCGGTGGGAAAAACCGC
CTGCTGCTGGCTGGCAATATGCTGTGACGATATTTGGCTCATCTGTCGCCACTCAGCCGCTGGAAGTGGTATTCTGAAAACCGCTGCATATGTTGAAGTACCG
TTCCTCGTGGTGGGCTGCTTAAATATATCCAGCCGATTTGAAAGCGGTTTTTCCAGCGAGGATTTATCTGGCTGTTTCTGCTTCTTAAAGCAACTGGGCGATGATTT
TATGCTGTGATGGCGGCAATGATGAAAGCATCGGTTTCCAGGCGCCTTATGTTGGTCTGGTCTGGTGGGCTGTTGGGCGCTGGGCTCAATTTTCCGTTTACCG
TTAGCGGCCCGCGCCGCTTTCCCTGCTGCGTGCAGTGAATGAAGTGCCTTAAAGCAATCAATGTCGGATGCGCGCGGAGCCGCTTATCCGACCAACATATCAT
AACGGAGTGTGCTGATTAAGCATGCCAATGACCGAAAGATAAGAGCAGGCAAGCTAATTTACCGATGTGCGGAAGGCTTACCGGAAAAAGACTTCGTTGGGAAAA
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 abnormal genes¹²

Bioinformatics solutions

```
GTGAAGGGCAATCAGCTGTTGCCGCTCTACTGGTGAAGAAAAAACCCCTGGCGCCAAATACGCAAAACCGCCTCTCCCCGCGGTTGGCCGATTCAATATGC
AGCTGGCACGACAGGTTTCCCAGCTGGAAGACGGGGCAGTGAAGCCGCAACCGCAATTAATGTGAGTTAGCTCACTCATTAGGCACCCAGGCTTTACACTTTATGCTTCC
GGCTCGTATGTTGTGGAAATTTGAGCCGGATAACAATTTACACAGGAAACAGCTTACGCCATGATACGGATTCACTGGCCGTCGTTTACAAACGTCGTGACTGGG
AAAACCCCTGGCGTTACCCAACTTAATCGCCTTGACAGCATAACCCCTTTGCCGAGCTGGCGTAATAGCGAAAGAGGCCCGCACCGGATCGCCCTTCCCAACAGTTGGC
CAGCCTGAAATGGCGCAATGGCCGTTTGGCTGTTTCCGTCACCCAGGAGCGGTCGGGAAAGCTGGCGTGGAGTGGCATTCTTCTGAGGGCCGATACTGTCTGCTGCC
CTCAAACCTGGCAGATGACAGGTTACGATGCGCCCATACACCAACGTGACCTATCCATTACGGTCAATCCGCGCTTTGTTCCACGGGAAATCCGACGGGTTGT
ACTCGCTCACATTTAATGTTGATGAAAGCTGGCTACAGGAAGGCCAGACGCGAATATTTTTGATGGCGTTAACTCGCGCTTTCATCTGTGTGCAACGGGCGCTGG
GTCGGTTACGGCCAGGACAGTCGTTTCCGCTCTGAAATTTGACCTGAGCGCATTTCACGCGCCGAGAAAAACCGCCTCGCCGTTGATGGTGTGCGCTGGGATGACG
GCAGTTATCTGGAAGATCAGGATATGTTGGCGGATGAGCGGCATTTCCGTGACGTCTGCTTTGCTGCATAAACCCGACTACACAAATCAGCGATTTCCATGTTGCCACTC
GCTTAATGATGATTTTACGCGCGCTGTACTGGAAGCTGAAGTTAGATGTCGGCGGAGTTGCGTGACTACCTACGGGTAACAGTTTCTTTATGGCAGGGTGAACG
CAGGTCGCCAGCGGCACCCGCGCCTTTCCGCGGTGAAATTCATGATGAGCGTGGTGGTTATGCCGATCGCGTCACACTACGCTCGAACCTCGAAAACCCGAAACTGT
GGAGCGCCGAAATCCCGAATCTCTATCGTCCGGTGGTTGAACTGCACACCGCCGACGGCAGCGTGATTGAAGCAGAAGCCTGCCGATGTCGTTTCCGCGAGGTTGC
GGATGAAAATGGTCTGCTGCTGCTGAACGGCAAGCCGTTGCTGATTCCGAGGCGTTAACCGTACAGGATCATCTCTGATGCTCAGGTGATGGATGAGCAGAC
GATGGTGGCAGGATACCTGCTGATGAAGCAGAAACAATTTAACCGCGTGGCCTGTTCCGATTATCCGAACCATCCGCTGTGGTACACGCTGTGGCAGCCGTAGCGG
CTGATGTGGTGGATGAAGCCAAATTTGAAACCCAGCGCATGGTCCAAATGCTGACCCGATGATCCGCGTGGCTGATCCGCGGATGAGCGGATGAGCGGTAACGC
GAATGGTGCAGCGCATGTAATCCCGAGTGTGATCATCTGGTCCGTCGGGAATGAATCAGGCCAGCGGCTAATCAGACGCGCTGTATCGCTGGATCAAACT
TGTCATCTTCCGCGCAGTATGAAGCCGGCAGCCGACACCCGATCCGATTAATTTGCCGATGACGCGCGCTGATGAAAGACAGCCCTT
CCGCGCTGAGCAGAAATGGTCCATCAAAAATGSCCTTTCGCTACCTGGAGAGACGCGCCGCTGATCCTTTGCAATACGCCACGCGATGGGTAACAGTCTTGGC
GGTTTCGCTAAACTGCGAGCGGTTTCCGTCAGTATCCCGCTTTACAGGCGGCTTCGCTGGGACTGGTGGATCAGTCGCTGATTAATATGATGAAAACGGCAA
CCCGTGGTCCGCTTACGGCGGTGATTTTGGCGATCCGCGCAACGATCGCCGCTGTGATGAAACGGTCTGGCTTTTCCGAGCCGCAACCGCCGATCCAGCGCTGACG
GAAGCAAAACCCAGCAGCAGTTTCCAGATCCGTTATCCGCGCAAAACCATCGAAGTGACCAAGCGAATACCTGTTCCGCTCATAGCGATAACGAGCTCTGCACCTG
GATGGTGGCGCTGGATGGTAAGCCGCTGGCAAGCGGTGAAGTGCCTGTGCTCCACAAGGTAAACAGTTGATTGAACCTGCCTGAACTACCCGACGCGGA
GAGGCGCCGGCAACTCTGGCTACAGTACGCGTAGTGCAACCGAACGCGACCGGATGGTCAGAAGCCGGGACATCAGCGCCTGGCAGCAGTGGCGCTCGGGG
AAAACCTCAGTGTGACGCTTCCCGCGCGCTCCACCCGATCCCGCATCCCGCATGACCCACGCGAAATGGATTTTGCATCGAGCTGGGTAATAGCGCTTGGCAATTTAAC
CGCCAGTCAGGCTTTTCTTACAGATGTGGATTGGGCATAAAAAACAATGCTGACGCGCGCTGCCGATCAGTTACCCGCTGCACCGCTGGATACAGCATTGGCCT
AAGTGAAGCGACCCCGCATTGACCCCTAACCGCTGGGTGCAAGCTGGGAGCGCTGGAAGGCGCGCTCCAGTACCAGCGCGAAGCAGCGTTGTCGACAGATAC
TGCTGATGCGGCTGATTACGACCGCTCAGCGGTGGCAGCATCAGGGGAAAACTTATTTATCAGCGGAAAAACCTACCGGATTGATGGTATGGTCAAATGGGG
ATTACCGTTGATGTTGAAGTGGCGGAGCGATACCCGATCCCGCGCGGATGGCTGAACTGCCAGCTGCCAGCTAGCAGAGCGGGTAAACTGGCTCGGATTA
GGGCGCAAGAAAACTATCCGACCGCCTTACTGCCGCTGTTTTGACCGCTGGGATGTCGCCATTGTGAGACATGATACCCCGTACGCTTCCCGAGCGAAAAACG
GTCTGCGCTGCGGGACCGCGGAATTTGAAATATGGCCACACAGTGGCGGGGACTTCCAGTTCAACACTCAGCGCTACAGTCAACAGCAACTGATGGAACACG
CCATCGCCATCTGCTGCACGGGAAAGGACAGATGGCTGAATATCGACGGTTTCCATATGGGGATTGGTGGCGAGCCTCCTGGAGCCGCTCAGTATCGGCGGAA
TTCAGCTGAGCGCCGGTGCCTACCATTAACCAATGGCTGGTGTGCTGTCAAAATAATAAACCCGGGAGGCCATGCTCGCCGCTATTTCCGTAAGGAAATCCATTA
TACTATTAATAAACACAACTTTGGATGTTGCGGTTATGCTTTTCTTACTTTTATCATGGGAGCCTACTTCCGCTTTTTCCCGATTGGCTAGCATGACATCAAC
CATATCAGCAAAAGTACAACGGATATTTTTGCGCGTATTTCTGCTGCTGCTATTAACACCGCTTTTGGTCTGCTTTTTCTGCAAACTCGGCTGGCGGCAAA
ACCTGCTGTGGATTATACCGCATGTAGTGTGTTTGGCGGCTTCTTATTTTATCTTCCGCGCACTGTTACAATAACAATTTAGTAGGATGATGTTGGTGGT
ATTTATCAGCTTTTGGTTTAAACCGCGGTGCCCCAGCAGTAAAGCAATTTAAGCAAAAGTACGCGCTGCAGTAAITTCGAATTTGGCTGCGCGGATGTTTGGC
TGTGTTGGCTGGGCGCTGTGCTCCGATGTCGGCATCATGTTCCACATCAATAATCAGTTGTTTCTGGCTGGGCTCGGCTGTGCACTCATCTCGCGTTTTA
CTCTTTTTCCGCAAAACGGATGGCGCCCTTCTCCGACGCTTGGCAATGCGGTAGTGGCCACCATTCGCGATTAGCCCTTAAAGCTGGCAGCTGTTGAGCA
GCCAAAACCTGTGGTTTTGTCACTGTATGTTTGGCTTTCTCGACCATCAGATGTTTGTGACCAACAGTTTGCATATTTCTTACTTCGTTCTTTGCTACCGGTGAAC
AGGGTACGCGGGATTTGGCTACGTAAACGACAAATGGGCGAATCTTAAACCGCTCGATTATGTTCTTTGCGCCACTGATCATTAAATCGCATCGGTGGGAAAAACCGC
CTGCTGCTGGCTGGCACTATGATGCTGTGACTGATTTGGCTCATCTGCGCCACTCAGCCGCTGGAAGTGGTATTCTGAAAACCGCTGCATATGTTGAAGTACCG
TTCCTGCTGGTGGGCTGCTTTAAATATATCCAGCCGATTTGAAATGGGTTTTTCAAGCGAGGATTATCTGGTCTGTTTCTGCTTCTTTAAGCAACTGGGCGATTGTTT
TATGCTGTGATGGCGGCAATGATGAAAGCATCGGTTTCCAGGCGCCTTATGTTGGTGGTCTGTTGGGCTGTTGGTGGGCTGGGCTGAACTTAAITTCGTTTACCG
TTAGCGGCGCCCGCCGCTTTCCCTGCTGCGTGCAGGTGAATGAAGTGCCTTAAAGCAATCAATGTCGGATGCGCGCGGACCGCCTATCCGACCAACATATCAT
AACGGAGTGTGCTGATGAACTGAACTGAACTGAAAGATAAGAGCAGGCAAGCTATTACCGATGTCGCAAGGCTTACCGGAAAAAGACTTCGTGGGAAAAAC
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 abnormal genes¹³

Ab initio methods

```
GTGAAGGGCAATCAGCTGTTGCCCGTCTCACTGGTGAAGAAAAAACCCCTGGCGCCCAATACGCAAAACCGCTCTCCCCGCGGTTGGCCGATTCAATATGC
AGCTGGCACACAGGTTTCCCAGCTGGAAGCGGGGAGTGAAGCGCAACGCAATTAATGTGAGTTAGCTCACTTATTAGGCACCCAGGCTTTACACTTTATGCTTCC
GGCTCGTATGTTGTGGAAATTTGAGCGGATAACAATTTACACAGGAAACAGCTTACCATTGATACGCGTACTGGCCGTCGTTTACAAACGTCGTGACTCGGG
AAAAACCCCTGGCGTTACCCAACTTAATCGCCTTGACAGCATCCCCCTTTGCGGACCGCTGGCCGTAATAGCAGAAAGGCGCCGACCGATTGCGCCTTCCCAACAGTTGGC
CAGCCTGAATGGCGGAATGGCCGCTTTGGCTCGCTTGGCCAGTACCCAGGAGCGGTCGGGAAAGCTGGCTGGAAGTGGCATTCTTCCGAGCCGATACTGCTCGTCCC
CTAAACTGGCAGATGCACGGTTACGATGCGCCCATACACCAACGTGACCTATCCATTACGGTCAATCCGCGCTTTGTTCCACGGGAATCCGACGGGTTGTT
ACTCGCTCACATTTAATGTTGATGAAAGCTGGCTACAGGAAGGCCAGACGCGAATATTTTTGATGGCGTTAACTCGCCGTTTCACTGTGGTCAACGGCGGCTGG
GTCGGTTACGGCCAGGACAGTCGTTTCCGCTGTAATTTGACCTGAGCGCATTTTACGCGCCGGAGAAAAACCGCTCGCCGTTGATGGTGCCTGGCTGGAGTGACC
GCAGTTATCTGGAAGATCAGGATATGTTGGCGGATGAGCGCCATTTCCGTGACGTCTGCTTGGTGCATAAAACCGACTACACAAATCAGCGATTCCATGTTGCCACTC
GCTTTAATGATGATTTTACGCGCGCTGACTGGAAGCTGAAGTTAGATGCGCGGAGTTGCGTGACTACCTACGGGTAACAGTTTCTTTATGGCAGGGTGAACCG
CAGGTCGCCAGCGGCAACCGCCGCTTTCCGCGGTGAATTTATCGATGAGCGTGGTGGTTATCCGATCGCGTCAACACTACGCTGAAACGTCGAAAACCCGAAACTGT
GGAGCGCCGAAATCCGAATCTCTATCGTCCGGTGGTTGAACGTGCACACCGCCGACGCGCAGCTGATTGAAGCAGAAGCCTGCCGATGTCGTTTCCGCGAGGTTGC
GGATGAAAAATGGTCTGCTGCTGCTGAAGCGCAAGCCGTTGCTGATTCAGGCGCTTAAACCGTCAAGGAGTCACTCTGATGGTCAGGTGATGGATGAGCAGAC
GATGGTGCAGGATATCTGCTGATGAAGCAGAAACAATTTAACGCCGTCGCTGATTCGCAATATCCGAACCATCCGCTGTGGTACACGCTGTGGCAGCCGTAGGGC
CTGATGTTGGTGGATGAAGCCAAATTTGAAACCCAGCGCATGGTCCAAATGCTGACCCGATGATCCGCGTGGCTACCGGGGATGAGCGAACCGGTAACCG
GAATGGTCAGCGCGATGTAATCACCAGAGTGTGATCATCTGGTCCGTCGGGAATGAATCAGGCCAGCGGCTAATCAGACGCGCTGTATCGCTGGATCAATC
TGTGATCTTCCCGCCGGTGCAGTATGAAGCGCGGAGCGACACCCAGATTAATTTGCCGATGACGCGCGGTGATGAAGACCGCCCTT
CCCGCTGTGCCGAAATGGTCCATCAAAAATGGCTTTCGCTACCTGGAGAGACGCGCCGCTGATCTTTCGCAATACGCCACGCGATGGGTAACAGTCTTGGC
GGTTTCGTAATACTGGCAGCGGTTTCCGTCAGTATCCCGCTTACAGCCGCGCTCGCTGGGACTGGTGGATCAGTCGCTGATTAATATGATGAAAAACGGCAA
CCCGTGGTCCGCTTACGCGCGTATTTGGCGATACCGCCGAAACGATCGCCAGTCTGTATGAACGGTCTGGCTTTTGGCGACCGCAACCGCATCCAGCGCTGACG
GAAGCAAAAACACCAGCAGCAGTTTTTCCAGTCCGTTTATCCGCGCAAAACCATGGAATGACCCAGCGAATACCTGTTCCGTCATAGCGATAACGAGCTCTGCACCTG
GATGGTGGCGCTGGATGGTAAGCCGCTGGCAAGCGGTGAAGTGCCTCTGATGTCGCTCCACAAGGTAACACAGTTGATTGAACCTGCCGTAACCTACCCGACGCGGA
GAGGCCCGGGCAACTCTGGCTCACAGTACGCGTAGTGAACCGCAACCGCCGATGGTGCAGAAAGCGGGCACATCAGCGCCTGGCAGCAGTGGCGCTCGGGCG
AAAAACCTCAGTGTGACGCTTCCCGCCGCGTCCACCCGATCCGCGTACGACCCAGCAGAAATGGATTTTGCATCGAGTGGGTAATAGCGTTGGCAATTTAAC
CGCCAGTCAGGCTTTTCCAGATGTTGGATGGGCAATAAAAAAAGTGTGACGCGCGTGGCGATCAGTTACCCGCTGCACCGCTGGATAACGACATTTGGCGT
AAGTGAAGCGAACCCGATTGACCCCTAACCGCTGGGTGCAAGCGTGGGTCAGCCGCTGGAAGCGCGGCGCCATTACCAGCGCGAAGCAGCGTTGTTGCACTGCACGGCAGATACACT
TGCTGATGCGGCTGATTACGACCGCTACCGGTTGGCAGCATCAGGGGAAAACTTATTATCAGCCGGAAAAACCTACCGGATTGATGGTATGGTCAAATGGCG
ATTACCGTTGATGTTGAAGTGGCGAGCGATACCCGATCCGCGCGGATGGCTGAACTGCCAGCTGGCGAGGTAGCAGAGCGGGTAAACTGGCTCGGATTA
GGCGCGCAAGAAAAATATCCGACCGCCCTACTGCGCGCTGTTTTGACCGCTGGGATGTCGCCATTGTCAGACATGATACCCCGTACGCTTCCCGAGCGAAAAACG
GTCTGCGCTGCGGGACGCGGGAATTAATATGGCCACACAGTGGCGGGGCGGACTCCAGTTCAACATCAGCCGCTACAGTCAACAGCAACTGATGAAACCGA
CCATCGCCATCTGCTGCACGGGAAAGGCAAGTCAAGTGGTGAATCGACGGTTTCCATATGGGGATGGTGGCGGACGACTCCTGGAGCCGCTCAGTATCGGGCGAA
TTCAGCTGAGCGCCGCTGCTACCATTAACAGTTGGTCTGGTGTCAAAAATAATAAACCCGGCAGGCCATGCTGCGCGTATTTCCGCTAAGCAATCCATTAAG
TACTATTAATAAAACACAACCTTTGGATGTTCCGTTTATCTTTTCTTACTTTTATATCGGGGACCTACTTCCCGTTTTCGCGATTTGGCTACGATGACATCAAC
CATATCAGCAAAAGTATACCGGATTAATTTTGGCCGATCTTCTGCTGCTGCTATTCACCCGCTGTTGGTCTGCTTCTGACAAACTCGGGCTGGCAAAAT
ACCTGCTGGGATATACCGCATGTTAGTGTGTTTGGCCGCTTCTTATTTTATCTTCCGGCCACTGTTACAATAACAATTTAGTAGGATCGATTGGTGGTGT
ATTTATCAGCTTTTATTTAACCGCGTGGCCAGCAGTAGAGCAATTAATGAAGAAATCAGCCGCTGCGCAATTTGCAATTTGGCTGGCGCGGATGTTGGC
TGTGTTGGCTGGCGCTGTGCTCGTATGTCGGCATCATGTTCCACATCAATATCAGTTGTTTCTGGCTGGGCTCGGCTGTCACATCATCTCGCGTTTTA
CTCTTTTTCCGCAAAACGGATGCGCCCTTCTGCGACGGTGGCAATGCGGTAGTGGCAACCATCCGCTTATAGCCTTTAGCCTGGCAGCTGGAACGTTGTGCAGCA
GCCAAAACCTGGTTTTTGTCACTGTATGTTATGGCGTTTCTGTGCGACTAGATGTTTTGACCAACAGTTTGTCTAATTTCTTACTCTGTTTCTTGTACCGGTGAAC
AGGGTACGCGGGTATTTGGCTACGTAAACGACAAATGGCGGAATTACTTAAACCGCTGATGTTGTTTTCGCGCACTGATCATTAATCGCATCGTGGGAAAAACCGCC
CTGCTGCTGGCTGGCAGTATGTCGCTGACGATTTAATGCTCATCTGCGCCACTCAGCGCTGGAAGTGGTTAATCTGAAAACCGCTGCATATGTTGAAGTACCG
TTCTGCTGGTGGGCTGCTTTAAATATATACCAAGCAGATTTGAAGTGGGTTTTTCAAGCGGACTTATCTGGTCTGTTTCTGCTTCTTAAAGCAACTGGCGATTGATTT
TATGCTGACTGGCGGGCAATATGATGAAAGCATCGGTTTCCAGGCGGCTTATGTTGGTGGTGGTCTGTTGGCGGCTGGGCTTACCTTAATTTCCGTTGTTCCAGC
TTAGCGCGCCCGCCGCTTTCCCTGCTGCGTGTGAGTGAATGAAGTGGTAAAGCAATCAATGTCGGATGCGCGCGGACCGCTTATCCGACCAACATATCAT
AACGGAGTGTGCTGATTAAGCATGCAATGACCGAAAGATAAGAGCAGGCAAGCTATTACCGATGTGCGAAGGCTTACCGGAAAAAGACCTTCTGGGAAAAAC
GTTAATGATGAGTTAATCACTCGCATCCATCAGA
```

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
AGCTGGCACACAGGTTTCCCAGCTGGAAGCGGGCAGTGAAGCGCAACGCAATTAATGTAGTGTAGCTCACTATTAGGCACCCAGGCTTTACACTTATGCTTCC
GGCTCGTATGTTGTGGAATTTGTGAGCGGATAACAATTTACACAGGAAACAGCTTACCATTGATACGGATTACTGGCCGTCGTTTACAAACGTCGTGACTGGG
AAAAACCTGGCGTACCCAACTTAATCGCTTGCAGCAGATCCCCTTTGCCGAGCTGGCCGTAATAGCGAAAGAGCCCGCAGCGATCGCCCTTCCCAACAGTTGGC
CAGCCTGAATGGCGAATGGCCGCTTTGGCTGTTTGGCCGAGCAAGCGGTCGGCGAAAGCTGGCTGGAGTGGCATCTTCCGAGCCGATACTGCTGCTGCC
CTAAACTGGCAGATGCACGGTTACGATGCGCCCATACACCAACGTGACCTATCCATTACGGTCAATCCGCGCTTTGTTCCACGGGAATCCGACGGGTTGT
ACTCGCTCACATTTAATGTTGATGAAAGCTGGCTACAGGAAGGCCAGACGCGCAATTTTTGATGGCGTTAACTCGCCGTTTCACTGTGGTCAACGGCGCTGG
GTCGGTACGGCCAGGACAGTCGTTTCCGCTCTGAATTTGACCTGAGCGCATTTCACGCGCCGAGAAAAACCGCTCGCCGTTGATGGTGCCTGGCTGGAGTGAC
GCAGTTATCTGGAAGATCAGGATATGTTGGCGGATGAGCGCCATTTCCGTGACGTCTCGTTGCTGCATAAACCGACTACACAAATCAGCGATTCCATGTTGCCACT
GCTTTAATGATGATTTTCCAGCCGCTGACTGGAGGCTGAAGTTAGATGCGCGGAGTTGCGTGACTACCTACGGGTAACAGTTTCTTTATGGCAGGGTGAACG
CAGGTCGCCAGCGGCAACCGCCGCTTTCCGGCGTGAATTTATCGATGAGCGTGGTGGTTATCCGATCGCGTCAACACTACGCTGAAACGTCGAAAACCCGAAACTG
GGAGCGCCGAAATCCGAATCTCTATCGTGGCGTGGTGAACGTGCACACCGCCGACGGCAGCGTGATTGAAGCAGAAGCCTGCGATGTCGTTTCCGCGAGGTCG
GGATTGAAAATGGTCTGCTGCTGCTGAAGCGCAAGCCGTTGCTGATTCGAGCGCTTAAACCGTACAGGATCATCTCTGATGGTCAGGTATGGATGAGCAGAC
GATGGTGCAGGATATCTGCTGATGAAGCAGAAACAATTTAACGCCGTCGCTGTTCCGATTATCCGAACCATCCGCTGTTGGTACACGCTGCGGACCCGACGGC
CTGTATGGTGGATGAAGCCCAATTTGAAACCCAGCGCATGGTCCAAATGCTGACCGATGATCCGCGTGGCTACCGGCGATGAGCGAACCGTAAACGC
GAATGGTCAGCGGATGTAATCACCAGTGTGATCATCTGGTCGCTGGGAATGAATCAGGCCAGCGGCTAATCAGACGCGCTGTATCGCTGGATCAATC
TGTGATCTTTCCCGCCGGTGCAGTATGAAGCGCGGAGCGCACACCCAGATTAATTTTCCCGGATGACGCGCGCTGATGAAGACAGCCCTT
CCCGCTGTCCGAAATGGTCCATCAAAAATGCTTTCCGCTACTGGAGAGACGCGCCGCTGATCTTTCCGAAATACGCCACGCGATGGTAAACAGTCTTGGC
GGTTTCCGTAATACTGGCAGCGGTTTCCGCTAGTATCCCGCTTACAGGCGGCTTCGCTGGGACTGGTGATCAGTCGCTGATTAATATGATGAAAAAGCCGCA
CCCGTGGTCCGCTTACGGCGGTGATTTTGGCGATACCGCGAACGATCGCCAGTCTGTATGAACGGCTGGTCTTTGCCGACCCGCAACCGATCCAGCGCTGAG
GAAGCAAAACACCAGCAGCAGTTTTTCCAGTCCGTTTATCCGCGCAAAACATCGAATGACCAAGCAATACCTGTTCCGCTATAGCGATAACGAGCTCTGCATCT
GATGGTGGCGCTGGATGGTAAAGCCGCTGGCAAGCGGTGAAGTGCCTCTGGATGCTGCTCCACAAGGTAACAGTTGATTGAACCTGCCGTAACACTACCGCAGCCGA
GAGGCCCGGCAACTCTGGCTACAGTACGCGTAGTGCAACCGCAAGCCGCGGATGGTCAGAAGCCGGGACATCAGCGCCTGGCAGCAGTGGCGCTCGCCGG
AAAACTCAGTGTGACGCTTCCCGCCGCTCCACCCGATCCCGCATCGACACCCAGCGAAATGGATTTTTGCATCGAGTGGGTAATAGCGTTGGCAATTTAAC
CGCCAGTCAGGCTTTCTTCCAGATGGTGGTGGCATAAAAAAACAATGCTGACGCGCGCTGCGCGATCAGTTACCCGCTGCACCCGCTGGATAACGCATTTGGCGT
AAGTGAAGCCAGCCGATTAACCGCTGGGTGCAAGCGTGGGAGCGGCGGCGCCATTACCAGCCGGAAGCAGCGTTGTTGCAGTGCACGGCAGATACACT
TGCTGATGCGGCTGCTGATTACGACCGCTACCGGTTGGCAGCATCAGGGGAAAACTTATTATACGCGGAAAAACCTACCGGATTGATGGTAGTGGTCAAATGGCG
ATTACCGTTGATGTTGAAGTGGCGGAGCGGATACACCGCATCCGCGCGGATGGCCGTAACCTGCCAGCTGGCAGGATAGCAGAGCGGGTAAACTGGCTCGGATTA
GGCGCGCAAGAAAACTATCCGACCGCCTTACTGCGCGCTGTTTTGACCGCTGGGATGTCGCCATTGTGACAGCATGTATACCCCGTACGCTTCCCGAGCGAAAAAGC
GTCTGCGCTGCGGGACGCGGAAATTGAATATGGCCACACAGCTGGCGGCGGACTCCAGTTCAACATCAGCCGCTACAGTCAACAGCAACTGATGAAAACCG
CCATCGCATCTGCTGCACGGGAAAGGCAAGTGGCTGAATATCGACGTTTCCATATGGGGATTGGTGGCGAGGACTCCCTGGAGCCCGCTCAGTATCGCGGAA
TTCCAGCTGAGCGCCGCTGCTACCATTAACAGTTGCTGTGGTGTCAAAAATTAATAAACCGGCAAGCCATGCTGCGCGTATTTCCGTAAGGAAATCCATTATG
TACTATTAAAAAACACAACCTTTGGATGTTCCGTTTATCTTTTCTTACTTTTTTATCATGGGAGCGCTACTTCCCGTTTTTCCCGGATTGGCTACGATGACATCAAC
CATATCAGCAAAAGTACAGGTTATTTTTGCGGCTATTTCTGTCTGCTATTTCCAAACCGCTGTTTGGTCTGCTTTCTGACAAAACCTGGGCTGGCAGAAAT
ACCTGCTGAGGATATACCGCATGTAGTGTGTTGCGCGCTTTCTTATTTTATCTTCCGGCCACTGTTACAATAACAATTTAGTAGGATCGATTGTTGGTGT
ATTTATTAGCCTTTTGTTTAACGCGCGTGCGCCAGTAGAGCAATTTAGGAAAGTACGCGCTGCAGTAAITTCGAATTTGGTGCAGCGCGGATGTTGGC
TGTGTTGGCTGGCGCTGTGCTCGTATGTCGGCATCATGTTACCATCAATAATCAGTTGTTTCTGGCTGGGCTCGCTGTGCATCATCTCGCGTTTTTA
CTCTTTTTTCCGCAAAAGGATGGCCCTCTTCCGCAAGTGTGCCAATGCGGTAGTGGCAACCATCCGCTTATAGCCTTAAAGCTGGCAGCTGGAACGTTTGCAGACA
GCCAAAACCTGGTTTTTGTCACTGTATGTTATGGCGTTTCTGTGCACTACGATGTTTGGCAACAGTTTGTCTAATTTCTTTACTCTGTTCTTGTCAACCGGTGAAC
AGGGTACGCGGGTATTTGGCTACGTAAACGCAATGGCGGAATTACTTAAACCGCTGATTTGTTCTTTGCCCACTGATCATTAATCGCATCGTGGGAAAAACCGCC
TCTGCTGCTGGCTGGCATATATGCTGCTGACTGATTTTGGCTCATCTGCTCCACCTCAGCGCTGGAAGTGGTTATCTGAAAACCGCTGCATATGTTGAAGTACCG
TTCTGCTGGTGGGCTGCTTTAAATATATACCAGCCAGTTTGAAGTGGGTTTTTCAAGCGAGGATTATCTGGTCTGTTTGTCTTTTAAAGCAACTGGCGGATGATTT
TATGTCTGACTGGCGGGAATATGATGAAAAGCATCGTTTCCAGGCGGCTTATCTGGTGGTCTGGTCTGGTGGCGCTGGGCTTACCTTAAITTCGTTCCGTTCCAGC
TTAGCGCGCCCGCCGCTTTCCCTGCTGCGTGCAGGTGAATGAAGTGCCTTAAAGCAATCAATGTCGGATGCGCGCGGACGGCCTTATCCGACCAACATATCAT
AACGGAGTGTGCGATTTGAACATGCCAATGACCGAAAAGATAAGAGCAGGCAAGCTATTACCGATATGTCGGAAGGCTTACCGGAAAAAGACTCTGGGAAAAAC
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
AGCTGGCACAGAGGTTTCCCAGCTGGAAGACGGGGCAGTGAGCGCCAAACCGAATTAATGTGAGTTAGCTCACTATTAGGCACCCAGGCTTTACACTTTATGCTTCC
GGCTCGTATGTTGTGGAAATTTGAGCGGATAACAATTTACACAGGAAACAGCTATGACCATGATACGGATTCACTGGCCGTCGTTTTACAAACGTCGTGACTGGG
AAAACCTTGGCGTTACCCAACCTAATCGCCTTGCCAGCACATCCCTCTTCCGGCAGCTGGCGTAATAGCGAAAGAGCGCCGACGATCGCCCTTCCCAACAGATTGGC
CAGCCTGAATGGCAATGGCGCTTTGCCCTGGTTTTCCGCACACGAAGCGGTGGCGGAAAGCTGGCTGGAGTGGCATCTTCTGAGGGCCGATAGCTGTCGTGGCC
CTAAACCTGGCAGATGCACGGTTACGATGCGCCCATACACCAACGTCCTATCCCATACGGTCAATCCGCCGTTTGTCCACGGGAAGTCCGACGGGTTGTT
ACTCGCTCACATTTATGTTGATGAAAGCTGGCTACAGGAAGGCCAGCGAAGTAATTTTTGATGGCGTTAACTCGCGCTTTCATCTGGTGCACACGGGGCGCTGC
GTCGGTTACGGCCAGGACAGTCGTTGCCCTCTGAATTTGACCTGCGCATTTCACGGCCGGAGAAAAACCGCCTCGCGGTGATGGTGGCTGGGCTGGAGTGACC
GCAGTTATCTGGAAGATCAGGATATGTGGCGGATGAGCGGCATTTCCTGTGACGTCTCGTTGCTGCATAAACCCGACTACACAAATCAGCGATTTCCTGTTGCCACTC
GCTTTAATGATGATTTACGCCCGCTGTACTGGAGCGTGAAGTTCAGATGTGGCGGAGTTCGCTGACTACCTACCGGTAACAGTTTTCTTATGGCAGGGTGAACG
CAGGTGCGCCAGCGCACCGCGCCTTTCCGGCGTGAATATTCGATGAGCGTGGTGGTTATCCGATCGCGTACACACTACGTCTGAACGTCGAAAAACCGAAACTGT
GAGGCGCCGAAATCCCAGTCTCTATCGTGGCGTGGTTGAACCTGCACACCGCCGACCGCACCGTATTGAAGCAGAAACGCTGGCATTGTCGGTTCCCGCAGGTTGC
GGATTGAAAAATGGTCTGCTGCTGCTGAACGGCAAGCGGTTGCTGATTCGAGGGCTTAAACCGTACGAGCATCATCTCTGATGGTCAAGTCAATGGATGAGCAGAC
GATGGTGCAGGATATCTGCTGATGAAGCAGAAACAACTTTAACCGCTTGGCTGTTCCGATTCGGAACCAATCCGCTGGTGAACCGCTGGGGACCGCTAGCGG
CTGATGTGGTGGATGAAGCCAAATTTGAAACCCAGCGCAATGGCCAAATGCTGACCCGATGATCCGGCTGGTACCGGGGATGAGCGAACCGTAAACCG
GAATGGTGCAGCGCATCTAATCACCAGTGTGATCATCTGGTCCGTGGGGAATGATCAGGCCACGGCCATATCAGCAGCGCTGTATCGCTGGATCAAAAT
TGTCGATCTTTCCCGCCGGTCCAGTATGAAGCGCGCGAGCGCACACACCGACCGAATATTTGCCGATGTACGGCGCGGTGGATGAAGACGAGCCCTT
CCCGCGCTGACGAAATGGTCCATGAAAAATGGCTTTCGCTACCTGGAAGAGACGGCGCCGCTCCCTTTCCGAAATACGCCACGCGATGGTGAACAGCTTTGCC
GGTTTCGTAATAACTGCGAGCGGTTTCGTCAGTATCCCGCTTTACAGGCGCGCTTCGCTGGGACTGGTGGATCAGTCGCTGATTAATATGAAAAACGGCAA
CCCGTGGTGGCTTACGGCGGTGATTTGGCGAATACGGCAAGCTCCGACTGCTGTATGTAACGGTCTGGTCTTTGCCGACCGCACCGCCGATCCAGCGCTGACG
GAAGCAAAACACAGCAGCAGTTTCCAGTTCGCTTATCCGGCAAAACCTGGAATGACCCAGCGAATACCTGTTCCGCTCATAGCGATAACGAGCTCCCTGCACCTG
GATGGTGGCGCTGGATGGTAAGCCGCTGGCAAGCGGTGAAGTGCCTCTGGATCTGCTCCACAAAGGTAACAGTTGATTGAACCTGCCTGAACTCAGCCCGCGGA
GAGGCCCGGGCAACTCTGGCTCACAGTACGGTAGTGCAACCGCAACGACCGGATGGTTCAGAAAGCCGGGCACATCAGCGCCTGGCAGCAGTGGCGCTCGGGCG
AAAACCTCAGTGTGACGCTCCCCGGCGCTCCACCCGATCCCGCATCGACCCAGCAGGAAATGGATTTTGCATCGAGTAAAGCGTTGGAATTTAAC
CGCCAGTCAGGCTTTCTTTCAGATGTTGGATGGGCATAAAAAAACAATGCTGACGGCGCTGGCCGATCAGTTACCCGCTGCACCCTGGATAACGACATTTGGCT
AAGTGAGCGAACCCGCAATTGACCCTTAAGCCGTTGGTGCAGCCGCTGGAAGGGCCGCTTACCAGCCGAAAGCAGCGTTTGTGAGTGCACAGACATACACT
TCTGATGGCGTCTGATTACGACCGCTCACCGGTGGCAGCATCAGGGGAAAAACCTTATTATCAGCCGAAAAACCTACCGGATTCATGCTGATGGTCAAATGGCC
ATTCAGCTTATGATTTGAAGTGGCGAGCGGATACCCCGCATCCCGCGGATTTGGCCCTGACTGCCAGCTGGCGCAGGTAGCAGAGCGGGTAAACTGGCTCGGATTA
GGGCGCAAGAAAACTATCCGACCGCCTTACTGCCGCTGTTTTGACCGCTGGGATCTGCCATTGTGACAGATGATACCCCGTACGCTTTCCCGAGCGAAAAACG
GTCTGCGCTGCGGGACGCGGGAATTAATGATGGCCACACAGTGGCGGGGGACTTCCAGTTCACACATCAGCCGCTACAGTCAACAGCACTGATGAAACCCAG
CCATCGCCTCTGCTGACCGCGAAGAAAGCCAGATGGCTGAATATCGACGGTTTCAATATGGGGATGGTGGCGCAGACTCCTGGAGCCCGCTCAGTATCGCGGAA
TTCCAGCTGAGCGCCGGTGGCTGCTGACTTACCAATGGCTGGTGTGTCAAAAATAAATAACCCGGCAGGCCATGCTGCCCCGATTTCCGCTAAGGAAATCCATTAT
TACTATTAAAAAACAACAACCTTTGGATGTTCCGTTTATCTTTCTTTACTTTTTATCATGGGAGGCTACTTCCCGTTTTTCCCGATTGGCTACATGACATCAAC
CATATCAGCAAAAGTACAGCGGATTAATTTTCCCGCATCTCTGTCGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTG
ACCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTG
ATTTTCTAGCTTTTGTGTTTAAAGCCGCTGGCCAGCAGTAAAGCAATTAATGACAAAGTACAGCCGCTGCGCAATTTGCAATTTGGCTGCGCGCGGATTTGGC
TGTGTTGGCTGGGCGCTGTGCTCCGATGTGCGGCATCATGTTACCATCAATATCAGTTGTTTCTGGCTGGGCTCTGGCTGTGCACTCATCTCGCGTTTTA
CTCTTTTTGCGCAAAACGGATGCGCCCTTCTGCCACGTTGCCAATGGCGATGGTGGCAACCATTCGGCTTAGCCCTAAGCTGGACCTGGAACCTGTTGAGACA
GTTCAAAACTGTGGTTTTTGTCACTGTGGTGGCGTTTCCGTCACCTAGGATTTTTGACCAACAGTTTGTCTTCTTCTGCTGCTGCTGCTGCTGCTGCTGCTGCTG
AGGGTACCGCGGATTTGGCTACCTGACACATGGGCGAATCTTAAACCGCTCGATTATGTTCTTTGCCCGACTGATCATTAACTCGCATCGGTGGGAAAAACCGC
CTGCTGCTGGCTGGCATAITGTCGTGACGATATTTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTG
TTCTGCTGGTGGGCTGCTTTAAATATATACCAAGCAATTTGAAAGCGGTTTTTCAAGCGAGGATTTATCTGGTCTGTTTCTGCTTCTTAAAGCAACTGGCGATTGATTT
TATGCTGACTGGCGGGCAATATGATGAAAGCATCGGTTTTCAAGGCGCTTATCTGGTGGGCTGCTGGTGGGCTGCTGGTGGGCTGCTGGTGGGCTGCTGGTGGG
TTAGCGGCCCGCGCCCTTTCCCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTG
AACGGAGTATCGCATTTGAAACATGCCAATGACCGAAAGAAATAAGAGCAGGCAAGCTATTTACCGATGTGCGAAGGCTTACCGGAAAAAGACTCTGTTGGGAAAA
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 -1

GTGAAGGGCAATCAGCTGTTGCCGCTCTACTGGTAAAAGAAAAACCACCTGGCGCCAAATACGAAACCGCCTCTCCCGCGCGTTGGCCGATTCAATATGC
AGCTGGCACGACAGGTTTCCCAGCTGAAAAGCGGGCAGTGAAGCGCAACCGCAATTAATGTGAGTTAGCTCACTTATAGGCACCCAGGCTTTACACTTATGCTTCC
GGCTCGTATGTTGTGGAAATTTGAGCGGATAACAATTTACACAGGAAACAGCTTACGCCATGATACGGATTCACTGGCCGTCGTTTACAAACGTCGTGACTGGG
AAAACCTTGGCGTTACCCAACTTAATGCCCTTGACGACATCCCTCTTGGCCAGCTGGCGTAATAGCGAAAGAGGCCCGCACCGATGCCCTTCCCAACAGTTGGC
CAGCCTGAATGGCGAAATGGCGCTTTGGCTGAGTTTCCGACACAGGAAAGCGCGGAAAGCTGGCGTGGAGTGGCATCTTCCAGAGGCCGCTCTGCTGCTCC
CTCAAACCTGGCAGATGCACGGTTACGATGCGCCCATACACCAACGTGACCTATCCATTACGGTCAATCCCGCGTTTGTCCACGGAGAATCCGACGGGTTGT
ACTCGCTCACATTTAATGTTGATGAAAAGCTGGCTACAGGAAAGCCAGACGCAATATTTTTGATGGCGTTACTCGGGCTTCACTGTGGTGAACCGGGCGCTGG
TCCGTTACGGCCAGGACAGTCGTTGCCCTCTGAATTTGACCTGAGCCGATTTTTACGCGCCGGAGAAAACCGCCTCGCGGTGATGGTGTGCGCTGGGATGAC
GCAGTTATCTGGAAGATCAGGATATGTTGGCGGATGAGCGGCATTTCCGTGACGTCTGTTGCTGCATAAACCGACTACACAAATCAGCGATTTCCATGTTGCCACT
GCTTTAATGATGATTTTACGCGCGCTGACTGGAAGCTGAAGTTAGATGCGGGGAGTTGCGTACTACCTACGGGTAACAGTTTCTTATGGCAGGGTAAAAC
CAGGTCCGACCGGACCCGCGCTTTCCGCGGTGAAATATCGATGAGCGTGGTGGTTATCCGATCGCGTACACACTACGTCTGAACGTCGAAAACCCGAAACTGT
GAGGCGCCGAAATCCGAATCTCTATCGTGGCGTGGTGAACGTGCACCCGCGACCGCAGCGTATTGAAGCAGAAACGCTGCGATGTCGTTTCCGCGAGGTGC
GGATGAAAATGGTCTGCTGCTGCTGAACGGCAAGCCGTTGATTGCGAGGCTTAAACCGTACGAGCATCATCTCTGATGGTCAAGGTATGGATGAGCAGAC
GATGGTGCAGGATATCTGCTGATGAAGCAGAAACAATTTAACGCCGTGGCTGTTCCGATTTCCGAAACCATCCGCTGGTGAACCGCTGCGGACCCGTAGCGC
CTGATGTGGTGGATGAAGCCAAATTTGAAACCCAGCCGATGGTCCAAATGCTGACCCGATGTCGCGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGT
GAATGGTGCAGCGCATGTAATCACCAGTGTGATCATCTGGT
TGTCGATCTTCCCGCCGGTGCAGTATGAAGCGCGGACGACACCCGACCCAGATATTTTCCCGCATGTACGCGCGCTGGATGAAGACACGCCCCTT
CCCGCTGTCCGAAATGGTCCATCAAAAATGGCTTTCGCTACTGGAGAGACGCGCCCGCTGATCCTTTCCGAAATACGCCACCGCATGGTGAACGCTTGGC
GGTTTCGCTAAACTGCAGCGGCTTTCGTCAGTATCCCGCTTTACAGGGCGGCTTCGCTGGGACTGGTGGATCAGTCCGCTGATTAATATGATGAAAACGGCAA
CCCGTGGTCCGCTTACGGCGGTGATTTTGGCGATCCGCGCAACGATCCGCACTGCTGATGAACGGTCTGGCTTTTCCCGACCCGACCCGCGCATCCAGCGCTGACG
GAAGCAAAACACCAGCAGCAGTTTCCAGTTCCGTTTATCCGCGCAAAACCATCGAAGTGACCAAGCAATACCTGTTCCGTCATAGCGATAACGAGCTCCTGCACGT
GATGGTGGCGCTGGATGGTAAGCCGCTGGCAAGCGGTGAAGTGCCTCTGGATGTCGCTCCACAAGGTAAACAGTTGATTGAACCTGCCTGAACTGCCTGGCA
GAGGCGCCGGCAACTCTGGCTCACAGTACGCGTAGTGAACCGCAAGCCGACCGGATGGTGCAGAAAGCCGGGACATCAGCGCCTGGCAGCAGTGGCGTCTGGCG
AAAACCTCAGTGTGACGCTTCCCGCGCGCTTCCACCCGATCCCGCATCTGACCCACGCAAGCAATGGATTTTGCATCGAGTAAAGCGTTGGCAATTTAAC
CGCCAGTCAGGCTTTCTTTCAGATGTTGGATGGGCAAAAAACAATGCTGACGCGCGTGGCGATCAGTTACCCGCTGCACCCTGGATAACGACATTTGGCT
AAGTGAAGCCACCCGCAATTGACCCTTAAGCCCTGGGTGCAAGCGTGGAAAGCGCGGCGCCATTACCAGCCGAAAGCAGCGTTTGTGAGTGCACGACATACACT
TGCTGATGCGGCTGCTGATTACGACCGCTCACGCGTGGCAGCATCAGGGGAAAACCTTATTATCAGCCGGAAAACCTACCGGATGATGGTATGGTCAAATGGCG
ATTACCGTTGATGTTGAAGTGGCGGACGACATACCCGATCCCGCGGATGGCTGACCTGCCAGCTGGCGCAGGTAGCAGAGCGGGTAAACTCGGATTA
GGCGGCAAGAAAATATCCGACCGCCTTACTGCGCGCTGTTTACGCGCTGGGATGTCGCCATGTGACAGCATGATACCCCGTACGCTTCCCGAGCGAAAACG
GTCTGCGCTGCGGGACGCGCAATTAATATGGCCACACCACTGGCGGGGCACTCCAGTTCAACAACA

CGCGCTACAGTCAACAGCAACTGATGAAAACCG
CAATCGCCACTCTGCTGCACGGGAAAGGACCAATGGCTGAATATCGACGGTTCATATGGGGATGGTGGGACGACTCCTGGAGCCCGTCAAGTCCGGCGAA
TTCCAGCTGACCGCCGGTCCGCTACCATTACCAATTTGGTCTGGGTGCAAAAATAAATAAACCGGGCAGGC

- In the other direction :

ATG → CAT
TAA → TTA
TGA → TCA
TAG → CTA

- here : 2 ORFs found

Example of ORF finding : phase +2

- 3 ORFs found

GTGAAGGGCAATCAGCTGTTGCCGTCTCACTGGTAAAAGAAAAACCACCTGGCGCCAATACGAAACCGCTCTCCCCGCGGTTGGCCGATTCAATATGC
AGCTGGCACACAGGTTTCCCAGCTGAAAAGCGGGCAGTGAGCGCAACCGCAATTAATGTAGTGTAGCTCACTTATAGGCACCCAGCGCTTACACCTTATGCTTCC
GGCTCGTATGTTGTGGAAATGTAGCCGGATAACAATTTACACAGGAAACAGTACGACCATGATACGGATTCACTGGCCGTCTGTTTACAAACGTCGTGACTGGG
AAAACCCCTGGCGTACCCAACTTAATCGCTTGCAGCAGTCCCCCTTTCGGCAGCTGGCGTAATAGCGAAAGAGGCCCGCACCGGATCGCCCTTCCCAACAGTTGGC
CAGCTGAAATGGCGCAATGGCCGTTTGGCTGTTTCCGCTGAAATTCATCGATGAGCGTGGTGGTTATGCCGATCGCGTACACACTACGCTGAAACGTCGAAAACCCGAAACTG
CTAAACTGGCAGATGCACGGTTACGATGCGCCCATACACCAACGTGACCTATCCATTACGGTCAATCCGCCGTTTGTCCACGGAGAATCCGACGGGTTGT
ACTCGCTCACATTAATGTTGATGAAAAGCTGGCTACAGGAAGGCCAGACGCGAATATTTTTGATGGCGTTAACTCGCGCTTTCATCTGTGGTGAACCGGGCGCTGG
GTCGGTACGGCCAGGACAGTCGTTTCCGCTGTAATTTGACCTGAGCGCATTTCACGCGCCGAGAAAAACCGCTCGCCGTTGATGGTGTGCGCTGGGATGACG
GCAGTTATCTGGAAGATCAGGATATGTGGCGGATGAGCGGCATTTCCTGTGACGTCTGCTTGTGCATAAACCGACTACACAAATCAGCGATTCCATGTTGCCACT
GCTTAAATGATGATTTTACGCCGCTGACTGGAAGCTGAAGTTAGATGTGGCGGAGTTCGCTGACTACCTACGGGTAACAGTTCTTTATGGCAGGGTGAACG
CAGGTCGCCAGCGGCACCGCGCCTTTCCGGCGTGAATTCATCGATGAGCGTGGTGGTTATGCCGATCGCGTACACACTACGCTGAAACGTCGAAAACCCGAAACTG
GAGGCGCCGAAATCCGAACTCTATCGTGGCGTGGTGAACGTGCACACCAGCGCAGCGCTGATTGAAGCAGAAAGCCTGCGATGTGCGTTTCCGCGAGGTGC
GGATTGAAAATGGTCTGCTGCTGCTGAACGGCAAGCCGTTGACTTTCAGGCGTTAACCGTCAAGGATCATCTCTGATGTCAGGTGATGGATGAGCAGAC
GATGGTGCAGGATATCTGCTGATGAAGCAGAAACAATTTAACGCCGCTGGCCTGTTCCGATTCGGAACCATCCGCTGTTGATACACGCTGTGGACCCGCTAGCGC
CTGATGTGGTGGATGAAGCCAAATTTGAAAACCCAGCGCTGGTCCAAATGCTGACCGATGTCGCCGCTGGTGGTGAACCGGGATGAGCGAAACGCTAACGC
GAATGGTGCAGCGCATGTAATCACCAGTGTGATCATCTGGTCCGTCGGGAATGAATCAGGCCAGCGGCTAATCACACGCGCTGATCGCTGGATCAAACT
TGTGATCTTCCCGCCGGTGCAGTATGAAGCGCGGAGCGACCCGACCCAGATATTTTCCCGCATGACGCGCGCTGGATGAAGACCGAACCTT
CCCGCTGTGCCAAAGCTGATCAAAAATGGCTTTCGCTAGCTGGAGAGCGGGCCGCTGATCGTTTGGCAATACGCCACGGCATGGGACAGCTTTGGC
GGTTTCGTAATACTGGCAGCGCTTTCGTCAGTATCCCGCTTACAGGGCGGCTTCGCTGGGACTGGGTGGATCAGTCGCTGATTAATATGATGAAAACGGCAA
CCCGTGGTGGCTTACGGCGGTGATTTTGGCGATACCGCCGAACGATCGGCACTGTATGATGAACGGCTGGTCTTTGCCGACCCGACCCGATCCAGCGCTGACG
GAAGCAAAACACCCAGCAGCAGTTTTCAGTTCGCTTATCCGGCAAAACCATGGAATGACCCAGCGAATACCTGTTCCGCTATAGCGATAACGAGCTCCTGCACGT
GATGGTGGCGCTGGATGGTAAGCCGCTGGCAAGCGGTGAAGTGCCTCTGGATCTGCTCCCAAGGTAACAGTTGATTGAACCTGCCTGAACTACCCGACGCGGA
GAGGCCGGGGCAACTCTGGCTCACAGTACGGCTAGTGCACAAACGACCGGACCGCTGGTGCAGAAGCCGGGCACATCAGCGCTGGCAGCAGTGGCGTCTGGCG
AAAACCTCAGTGTGACGCTTCCCGCGCGCTTCCGACCATCCCGCATCTGACCCAGCGCAAAATGGATTTTGCATCGAGCTGGGAAATAGCGGTTGGCAATTTAC
CGCCAGTCAGGCTTTCTTTCACAGATGGTGGTGGGCAAAAAACAATGCTGACGCGCGCTGGCGCATCAGTTACCCGTCGACCGCTGGATAACGACATTTGGCT
AAGTGAAGCGAACCCGATTAACCCCTAACCGCTGGGTGCAAGCTGGGAGGCGCTGGAAAGCGCGGGCCATTACCAGGCCGAAGCAGCGTTGTCAGTACGGCAGATAC
TGCTGATGGCGTCTGATTACGACCGCTCACCGCTGGCAGCATCAGGGGAAAACTTATTTATCAGCCGAAAAACCTACCGGATGATGGTATGGTCAAATGGCG
ATTACCGTTGATGTTGAAGTGGCGGAGCGGATACCCGATCCCGCGCGGATGGCCGAACTGCCAGCTGGCGCAGGTAGCAGAGCGGGTAAACTCGGATTA
GGCGGCAAGAAAATATCCGACCGCTTACTGCGCGCTGTTTTCAGCGCTGGGATCTGCCATTGTGACAGATGATACCCCGTACGCTTCCCGAGCGAAAAACG
GTCTGCGCTGCGGGACCGCGGAATGAATATGGCCACACCACTGGCGGGCGACTCCAGTCAACATCAGCCGCTACAGTCAACAGCACTGATGAAAACCG
CCATCGCCATCTGCTGCACGGGAAAGAACGACCTGGCTGAATCGACGGTTTCCATATGGGGATGGTGGCGGAGCTCTGGAGCCCGTCAGTATCGCGGAA
TTCCAGCTGACCGCGGCTCGCTACCATACCAATTTGGCTGGGTGCAAAAATGAATAACCGGGCAGGCCATGCTGCCGCTATTTCGCGATGAAGGAAATCCATTATG
TACTATTAAAAAACCAAACTTTGGATGTTCCGTTTTCCTTTTACTTTTATCATGTTATGATGGAGGCTACTTCCCGTTTTCGCGATTGGCTACATGCACTAAC
CATATCAGCAAAAGTACGCGGATATTTTTCGCGGCTATTTCTGCTGCTGATTTCCAACCGGCTGTTGGTCTGCTTCTGACAAAACCTGGCGTGGCGGCTGC
ACCTGCTGGGATTATACCGCATGTAGTGATGTTTGCGCCGTTCTTATTTATCTTCCGGCCACTGTTACAATAACAATTTAGTAGGATGATGTTGGTGGT
ATTTATCAGCTTTTGGTTTAAAGCCGCTGGCCACAGATAGAGCAATTAATGAGAAAGTCAAGCCGCTGCAGTAATTTGCAATTTGGTTCGCGCGCGGATTTGGC
TGTGTTGGCTGGCGCTGTGCTCCTGATGTGCGCATCATGTTACCATCAATATCAGTTGTTTCTGGCTGGGCTCGGCTGTCACATCATCTCGCGTTTTA
CTCTTTTTTCGCAAAACGGATGCGCCCTTCTGCCAGGTTGCCAATGCGGTAGGTGCCAACCATTCGCGATTAGCCCTTAAAGTGGCAGCTGGAACGTTGCAGAA
GCCAAACTGTGGTTTTTGTCACTGTATGTTATGGCGTTTCTGTGACCACTACGATGTTTTGACCAACAGTTTCTGTAATTTCTTACTTCGTTCTTGTCTACCGGTAAAC
AGGGTACGCGGGTATTTGGCTACGTAAACGACAAATGGCGGAATTACTTAAACCGCTGATGTTGTTCTTTCGCCACTGATCATAATCGCATCGGTGGGAAAAACCGCC
TGCTGCTGGCTGGCAGTATATGCTGCTGACGATTTTGGCTCATCTGTCGCCACTCAGCCCTGGAAGTGGTTATCTGAAAACCGTGCATATGTTGAAGTACCG
TTCTGCTGGTGGGCTGCTTTAAATATATACCAAGCCAGTTTGAAGTGGGTTTTTCAGCGGACTTATCTGGTCTGTTTCTGCTTCTTAAAGCACTGGCGGATGATTT
TATGCTGACTGGCGGGCAATATGATGAAAACGATCGGTTTCCAGGGCCCTTATGTTGGTGGGCTGGGCTGCTGGGCTTCACTTAAATTTCCGTTCTCAGC
TTAGCGGCCCGCGCCGCTTTCCCTGCTGCGTGCAGTGAATGAAGTGCCTTAAAGCAATCAATGTCGGATGCGCGCGGACCGCTTATCCGACCAACATATCAT
AACGGAGTGTGATGATTAAGCATGCGCAATGACCGAAAAGATAAGAGCAGGCAAGCTATTTCCGATATGTGCGAAGGCTTACCGGAAAAAAGACTCTGTGGGAAAA
GTTAATGATGAGTTAATCACTCGCATCCATCAGA

Example of ORF finding : phase -2

- 7 ORFs found

GTGAAGGGCAATCAGCTGTTGCCGCTCTACTGGTAAAAGAAAAACCACCTGGCGCCAAATCGCAAAACCGCTCTCCCCGCGGTTGGCCGATTCAATATGC
AGCTGGCCACGACAGGTTTCCCAGCTGAAAAGCGGGCAGTGGAGCCAAACCGCAATTAATGTAGTTAGCTCACTTATAGGCACCCAGGCTTTACACTTTATGCTTCC
GGCTCGTATGTTGTGGAAATGTGAGCCGGATAACAATTCACACAGGAAGAACAGCTTACACCATGATACGGATTCACTGGCCGTCGTTTACAAACGTCGTGACTGGG
AAAACCCCTGGCGTACCCAACTTAATCGCTTGGACGACATCCCCCTTTCGGCAGCTGGCCGTAAATAGCAGAAAGGCGCCGACCCGATCGCCCTTCCCAACAGTTGGC
CAGCCGTAAGTGGCGCAATGGCCGCTTTGGCTGTTTGGCCGATCCAGCCAGCAAGGCGGTCGGGAAAGCTGGCGTGGAGTGGCATCTTCCGAGGCGGATACTGCTCGTCC
CTAAACTGGCAGATGCACGGTTACGATGCGCCCATACACCAACGTGACCTATCCATACGGTCAATCCGCGCTTGTCCACGGAGAATCCGACGGGTTGT
ACTCGCTACACTTAATGTTGATGAAAAGCTGGCTACAGGAAGGCCAGACGCGAATATTTTTGATGGCGTTAACTCGCGCTTTCATCTGGTGGCAACGGGCGCTGG
GTCGGTACGGCCAGGACAGTCGTTTCCGCTGTAATTTGACCTGAGGCGCATTTTACGCGCCGGAGAAAACCGCCTCGCCGTTGATGGTGCCTGGCTGGAGTGACG
GCAGTTATCTGGAAGATCAGGATATGTTGGCCGATGAGCGGCATTTCCTGTCAGCTGCTGTTGCTGCATAAACCGACTACACAAATCAGCGATTTCAGTTGTTGCCACTC
GCTTAAATGATGATTTTACGCGCGCTGACTGGAAGCTGAAGTTCAGATGTCGGCGGAGTTGCGTGACTACCTACGGGTAACAGTTTCTTATGGCAGGGTGAACG
CAGGTCGCCAGCGGCACCCGCGCTTTCGGCGGTGAAATTCATCGATGAGCGTGGTGGTTATGCCGATCGCGTACAATACGCTGTAACCTCGAAAACCCGAAACTG
GGAGCGCCGAAATCCCAGTCTCTATCGTCCGCTGGTGAACCTGCACACCCCGCAGCCGACCGCTGATTGAAGCAGAAAGCCTGCGGATGTCGTTTCCGCGAGGTTGC
GGATTGAAAATGGTCTGCTGCTGCTGAACGGCAAGCCGTTGCTGATTGAGGCGTTAACCGTCAAGGATCATCTCTGATGGTCAAGTGGATGAGCAGAG
GATGGTGCAGGATATCTGCTGATGAAGCAGAACAACTTAAACCGCTGGCCGCTGTCGCAATATCCGAACCAACGCTGTGGTACACCGCTGTGGACCCGCTACGGC
CTGATGTTGGTGGATGAAGCCAAATTTGAAACCCAGCCGATGGCCAAATGATCGTCACTGACCGATATCCGCGCTGGTGGTGAACCGGATGAGCGAAACCGTAAACG
GAATGGTCAGCGCGATGTAATCACCAGAGTGTGATCATCTGGTCCGCGGAATGAATCAGGCCACCGCCGCTAATCAGACCGCTGTATCGCTGGATCAAACT
TGTCTATCTTCCCGCCGGTGCAGTATGAAGCCGGCGGAGCCACACCCAGCCAGATATATTTTCCCGCATGACGCGCGCTGATGAAGACCGACTT
CCGCGCTGTCCGAAATGTTCCATCAAAAATGGCTTTCGCTACTGGAGAGACGGCCCGCTGATCTTTCGCAATACGCCACCGCATGGGTAACAGTCTTGGC
GGTTTCGCTAAACTGCGAGCGGTTTGGTTCAGTATCCCGCTTACGAGGCGCTGCTGGGACTGGTGGATCAGTCCCTGATTAATATGATGAAAACGGCA
CCCGTGGCTCGCTACGCGCGTGAATTTGGCCATACGCCAACGATCGCCAGTCTGCTGATGAACGCTGCTGGCTTTGCCGACCGCAACCGCCGCTCAGCGCTGACG
GAAGCAAAACACGAGCAGCAGTTTTCAGTTCCTGTTATCCGGCAAAACCATCGAAGTACCCAGCGCAATACCTGTTCCGCTCATAGCGATAACGAGCTCCTGCACCTG
GATGGTGGCGCTGGATGGTAAGCCGCTGGCAAGCGGTGAAGTGCCTCTGGATGCTGCTCCACCAAGGTAACAGTTGATTGAACCTGCCTGAACTACCGACCGCCGGA
GAGGCCCGCCGCAACTCTGGCTCAGATACCGCTAGTGGCAACCGCAACCGGACCGGATGGTCCGAAAGCCGCGTGAAGCCGGGACATCAGCGCCTGGCAGCAGTGGCGTCTGGCG
AAAACCTCAGTGTGACGCTGCCGCGCGCTCCACCGCCATCCCGCACTGCAACCCAGCGAAATGGATTTTGCATCGAGCTGGGTAATAAGCGTTGGCAATTTAAC
CGCCAGTCAGGCTTTTCTTACAGATGTGGATGGGCAATAAAAAAACAATGCTGACGCGCGCTGGCCGATCAGTTACCCCGTGCACCGCTGGATAACGACATTTGGCT
AAGTGAAGCGACCCCGCATTGACCCCTAACCGCTGGGTGCAAGCCTGGAAAGCGCGGCCCATACCAGCCGAAGCAGCGTGTGGCATGCAAGCAGATACACT
TCTGATGGCGCTGCTGATTACGACCGCTCAGCGTGGCAGCAAGGGGAAAAACCTTATTTATCAGCCGGAACCACTACCGGATGATGGTATGGTCAAAATGGCC
ATACCGTTGATGTTGAAGTGGCGGAGCGGATACCCGCACTCCGCGCGGATGGCCGTAAGTGCACCTGCGCAGGTAGCAGAGCGGGTAACTGGCTCGGATTA
GGCCGCAAGAAAACTATCCGACCGCCTTACTGCGCGCTGTTTGAACCGCTGGGATGCTGCCATTGTCAGACATGATACCCCGTACGCTTCCCGAGCGAAAAACG
STCTGCGCTGCGGGACGCGCGAATGAAATATGGCCACACCGCACTGGCGGGCGACTTCCAGTCAACATCAGCGCTACAGTCAACAGCAACTGATGAAAACCG
CCATCGCCTCTGCTGACGCGGGAAGGCAACATGGCTGAATATCGACGGTTTCCAGATGCGGATTTAGGGGATTTGGTGGCGACGACTCCTGGAGCCGCTCAGTATCGCCGAA
TTCAGCTGAGCGCCGCTGCTACCATACCAATTTGGCTGTTGCTGTCAAAATAAATAAACCCGGGCAAGCCATGCTGCGCCGATTTTCGCTAAGGAAATCCGATATG
TACTATTAAAAAACAACAATTTGGATGTTCCGTTTATCTTTTCTTACTTTTATCAGTGGGAGCCTACTTCCGCTTTTCCCGATTGGCTAGCATGACATCAAC
CATATCAGCAAAAGTATACCGGATATATTTTGGCCGATTTCTGTTGCTGATATCCAAACCGCTGTTGGTCTGCTTCTGACAAAACCTGGCTGCGCAAA
ACCTGCTGGATATACCGCACTGTAGTGTGTTTGGCCGCTTCTTATTTATCTTCCGGCCACTGTTACAATAACAATTTAGTAGGATGATGTTGGTGGT
ATTTATCAGCTTTTGGTTTAAACCGCGTGGCCAGCAGTACGCAATTAATGAGGAAAGTACGCGCTGCAGTAATTTGCAATTTGGTCCGCGCGGATTTGGC
TGTGTTGGCTGGCGCTGTGCTCGATGTCGGCATCATGTTACCATCAATAATCAGTTGTTTCTGGCTGGGCTGCTGGCTGTCACATCACTCGCCGTTTTA
CTCTTTTTCCGCAAAACGGATGCGCCCTTCTGCCAGCTTGGCAATGCGGTAGTGGCAACCATTCGGCATTTAGCCCTTAAAGCTGGCAGCTGGAACCTGTGAGCA
GCCAAAACCTGGGTTTTTGTCACTGTATGTTATGGCTTTCTGCGACCATACGATGTTTGTGACCAACAGTTTGTCTAATTTCTTACTCTGTTCTTGTCTACCGGTAAAC
AGGGTACGCGGGATTTGGCTACGTAAACGACAAATGGGCGAATTACTTAAACCGCTGATTTATGTTCTTTCGCGCACTGATCATTAATCGCATCGGTGGGAAAAACCGCC
TCTGCTGGCTGGCAGCATATGTCGCTGACGATTTTGGCTCATCTGTCGCCACCTCAGCCGCTGGAAGTGGTTATCTGAAAACCGTGCATATGTTGAAAGTACCG
TTCTGCTGGTGGGCTGCTTTAAATATATACCAGCCAGTTGAAAGTGGGTTTTTCAAGCAGGATTTATCGGCTGTTTCTGCTTCTTAAAGCACTGGCGGATGATTT
TATGCTGACTGGCGGGCAATATGATGAAAGCATCGGTTTCCAGGCGCCTTATGTTGGTGGGCTGGGCTGCTGAGGCGGCTCACCTTAAATTTCCGTTGTTCCAGC
TTAGCGGCCCGCGCCGCTTTCCTGCTGCGTGCAGTGAATGAAGTGCCTTAAAGCAATCAATGTCGGATGCGCGCGGAGCGCCTATCCGACCAACATATCAT
AACGGAGTGCATGCAATTTGAACTGCGCAATGACCGGAAAGATAAGAGCAGGCAAGCTAATTTACCGATGTCGCGAAGGCTTACCGGAAAAAAGACTCTGCGGAAAAAC
GTTAATGATGAGTTAATCACTCGCATCCATCAGA

Example of ORF finding : phase +3 and -3

GTGAAGGGCAATCAGCTGTTGCCGCTCTACTGGTAAAAGAAAAACCACCTGGCGCCAAATACGCAAAACCGCTCTCCCCGCGGTTGGCCGATTCAATATGC
AGCTGGCACACAGGTTTCCCAGCTGAAAAGCGGGCAGTGAAGCGCAACCGCAATTAATGTAGTGTAGCTCACTATTAGGCACCCAGGCTTTACACTTTATGCTTCC
GGCTCGTATGTTGTGGAAATTTGAGCCGGATAACAATTTACACAGGAAACAGCTTACGCCATGATACCGGATCACTGGCCGTCGTTTACAAACGTCGTGACTGGG
AAAACCCCTGGCGTTACCCAACTTAATCGCTTGCAGCAGCATCCCTCTTGGCCAGCTGGCGTAATAGCAGAAAGCGCCGACAGCATCGCCCTTCCCAACAGTTGGC
CAGGCTGAATGGCGCAATGGCGCTTTGGCTGCTTCCGACCCAGCAAGCGCGCGGAAAAGCTGGCCGAGTGGCATCTTCCGAGGCCGATACTGCTCGTGGCC
CTAAACTGGCAGATGCACGGTTACGATGCGCCCATACACCAACGTGACCTATCCATACGGTCAATCCGCGCTTGTCCACGGAGAATCCGACGGGTTGT
ACTCGCTCACATTAATGTTGATGAAAAGCTGGCTACAGGAAGGCCAGACGCGAATATTTTTGATGGCGTTAACTCGCGCTTTCATCTGTGGTCAACGGCGCTGG
GTCGGTTACGGCCAGGACAGTCGTTTCCGCTGTAATTTGACCTGAGCGCATTTCACGCGCCGAGAAAAACCGCTCGCCGTTGATGGTGTGCGCTGGGATGACG
GCAGTTATCTGGAAGATCAGGATATGTGGCGGATGAGCGGCATTTCCGTGACGTCGTTGCTGCATAAACCGACTACACAAATCAGCGATTCCATGTTGCCACT
GCTTTAATGATGATTTACGCGCGCTGTACTGAGGCTGAAGTTAGATGTCGGCGGAGTTCGCTGACTACCTACGGGTAACAGTTCTTTATGGCAGGGTAAAACG
CAGGTCGCCAGCGCACCGCGCTTTCCGCGGTGAAATATCGATGAGCGTGGTGGTATGCCGATCGCGTACACTACGCTGAAACGTCGAAAACCCGAAACTGT
GAGGCGCCGAAATCCGAATCTCTATCGTGGCGTGGTGAACGTGCACACCGCCGACCGCAGCGTATTGAAGCAGAAAGCTGCGATGTCGTTCCGCGAGGTGC
GGATGAAAATGGTCTGCTGCTGCTGAAACGGCAAGCGCTTACTGATTCCAGGCGTTAAACCGTACAGGATCATCTCTGATGGTCAAGGTGATGATGAGCAGAC
GATGGTGCAGGATATCTGCTGATGAAGCAGAAACACTTTAACCGCTGCGCTGTTCCGATTATCCGAACCATCCGCTGGTACACGCTGGGACCGCTAGCGG
CTGATGTTGGTGGATGAAGCCAAATTTAAAACCCAGCGCATGGTCCAAATGCTGACCCGATGTCGCGCTGGCTACCGGGGATGAGCGAAACCGGTAACCG
GAATGGTGCAGCGCATGTAATACCCAGTGTGATCATCTGGTCCGTCGGGAATGAATCAGGCCACCGCGCTAATCAGACGCGCTGATCGCTGGATCAAACT
TGTCATCTTCCCGCCGTCAGTGAAGCGCGCGGACGACACCGCCAGCCAGCATATTTTCCCGGATGACGCGCGCTGATGAGAACCGACCCCTT
CCCGCTGTGCGCAAAATGGTCCATCAAAAATGCTTTCGCTACCTGGAGAGACGCGCCGCTGATCCTTTCCGAAATACCGCCACGCGATGGTAAACAGTCTTGGC
GGTTTCGTAATACTGCAGCGCTTTCGTCAGTATCCCGCTTACAGCGCGGCTTCGCTGGGACTGGTGGATCAGTCGCTGATTAATATGATGAAAACGGCAA
CCCGTGGTGGCTTACGGCGGTGATTTTGGCGATACCGCGCAACGATCGCCAGTTCGTGATGAACGCTGTCGTTTGGCGACCGCACCGCATCCAGCGCTGACG
GAAGCAAAACACCCAGCAGCATTTTCCAGTTCCGTTTATCCGCGCAAAACCATGGAATGACCAAGCAATACCTGTTCCGTCATAGCGATAACGAGCTCCTGCACCT
GATGGTGGCGCTGGATGGTAAGCCGCTGGCAAGCGGTGAAGTGCCTCTGGATGTCGCTCCACAAGGTAAACAGTGTGATGAACTGCCTGAACTACCCGACGCGGA
GAGGCGCGGGCAACTCTGGCTACAGTACCGTAGTGAACCGCAACCGCAGCGGATGGTCAAGAACCGGGCACATCAGCGCTGGCAGCAGTGGCGTCTGGCGG
AAAACCTCAGTGTGACGCTTCCCGCGCGCTTCCCGCGCGTCCAGCATCCCGCATCGACCCAGCGAAATGGATTTTGCATCGAGTAAGCGTAAAGCGTTGGCAATTTAAC
CGCCAGTCAGGCTTTCTTTACAGATGGTGGTGGGCAAAAAAACAATGCTGACGCGCGCTGGCGCATCAGTTACCCGCTGCACCGCTGGATAACGACATTTGGCT
AAGTGAAGCGAACCCGCAATTGACCCCTAACCGCTGGGTCGAAAGCGTGGCGGGCCGACCTACCAGCGCGAAGCAGCGTGTGTCAGTGCACGCGGACATAC
TGCTGATGGCGTCTGATTACGACCGCTCAGCGTGGCAGCATCAGGGGAAAACTTATTTATCAGCGGAAAAACCTACCGGATGATGGTAGTGGTCAAATGGCG
ATTACCGTTGATGTTGAAGTGGCGGAGCGGATACACCGCATCCGCGCGGATGGCTCAGACTGCCAGCTGGCGCAGGTAGCAGAGCGGGTAACTGGCTCGGATTA
GGCGCGCAAGAAAACTATCCGACCGCTTACTGCGCGCTGTTTACCGCTGGGATGTCGCCATGTCAGACATGATACCCCGTACGCTTCCCGAGCGAAAAACG
GTCTGCGCTGCGGACCGCGGAATTAATGATGGCCACACCGAGTGGCGGGGACTTCCAGTTCAACATCAGCGCTACAGTCAACAGCAACTGATGAAAACCG
CCATCGCCTCTGCTGCACGGGAAAGGCAAGTGCATGGCTGAATATCGACGGTTTCAATATGGGGATGGTGGCGCAGACTCCTGGAGCCCGTCAAGTATCGCGGAA
TTCAGCTGAGCGCGGTGCTGCTACCATTAACATTTGGTCTGGTGTCAAAAATAAATAAACCCGGCAGCCATGCTGCGCGTATTTCCGTAAGGAAATCCATATG
TACTATTAAAAAACAACAATTTGGATGTTCCGTTTATCTTTTCTTACTTTTATCAGTGGGAGCTACTTCCCGTTTTCGCGATTGGCTACATGACATCAAC
CATATCAGCAAAAGTATACCGGATATTTTTGGCCGCTATCTCTGCTGCTGCTATTTCAACCGGCTGTTGGTCTGCTTCTGACAAAACCTGGGCTGCGCAAT
ACCTGCTGGATTTACCGCGATGTAGTGATGTTGGCGGCTTCTTATTTATCTTCCGGCCACTGTTACAATAACAATTTAGTAGGATCGATGTTGGTGGT
ATTTATCTAGCTTTTGGTTTAAACCGCGTGGCCAGCAGTGAAGCAATTAATGCAAAAGTACGCGCTGCAAGTAAITTCGAATTTGGTCCGCGCGGATTTGGC
TGTGTTGGCTGGCGCTGTGCTCGTATGTCGGCATCATGTTCCACATCAATATCAGTTGTTTCTGGCTGGGCTGCTGGCTGTCACATCTCGCGCTTTTA
CTCTTTTTCCGCAAAACGGATGCGCCCTTCTGCCAGGTGTCCAATGCGGTAGGTGCCAACCATTCGCGATTAGCCCTAAGCTGGCAGCTGGAACGTTGCAGACA
GCCAAACTGTGGTTTTGTCAGCTGATGTTATGGCGTTTCTGCGACCATAGCATTTTTGACCAACAGTTTGTCTAATTTCTTACTCTGTTCTTGTCTACCGGTGAAC
AGGCTACGCGGGTATTTGGCTACGTAAACGACATTTGGCGGAATTACTTAAACCGCTGATGTTGTTCTTGGCCACTGATCATAATCGCATCGTGGGAAAAACCGCC
TGCTGCTGGCTGGCAGCATATGTCGCTGACGATTTAGGCTCATCTGTCGCCACTCAGCGCTGGAAGTGGTATTTCTGAAAACCGTGCATATGTTGAAGTACCG
TTCTGCTGGTGGGCTGCTTTAAATATATACCAGCCAGTTGAAAGTGGGTTTTACGCGAGGATTTATCGGCTGTTTCTGCTTCTTAAAGCAACTGGCGGATGATTT
TATGCTGACTGGCGGGAATATGATGAAAACGATCGTTTCCAGGCGGCTTATGTTGGTGGGCTGGGCTGCTGGGCTGCGCTTCAATTTAATTTCCGCTGTTCCG
TTAGCGCGCCCGCGCCCTTTCCCTGCTGCGTGCAGGTGAATGAAGTGCCTTAAAGCAATCAATGTCGGATGCGCGCGGACCGCTTATCCGACCAACATATCAT
AACGGAGTGCATGCAATTTAAACGCTGCAATGACCGAAAAGATAAGAGCAGGCAAGCTAATTTACCGATGTCGCAAGGCTTACCGGAAAAAGACTCTGCGGAAAAAC
GTTAATGATGAGTTAATCACTCGCATCCATCAGA



Example of ORF finding : phase +3 and -3

```
GTGAAGGGCAATCAGCTGTTGCCGCTCTACTGGTAAAAGAAAAACCACCTGGCGCCAAATACGAAACCGCCTCTCCCGCGCGTGGCCGATTCAATATGC
AGCTGGCACACAGGTTTCCCAGCTGAAAAGCGGGCAGTGAAGCCAAACGCAATTAATGTAGTGTAGCTCACTTATAGGCACCCAGGCTTTACACTTATGCTTCC
GGCTCGTATCTTGTTGGAAATTTGAGCGGATAACAATTTACACAGAGAAACAGCTTACGACATGATACGGATTCACTGGCCGTCGTTTACAAACGTCGTGACTGGG
AAAACCGCTGGCGTTACCCAACTTAATCGCCTTGACAGCAATCCCCCTTTCGGCAGCGTGGCGTAATAGCAGAAAGGCGCCGACCGATCGCCCTTCCCAACAGTTGGC
CAGGCTGAATGGCGCAATGGCGCTTTGGCTGCTTTCGGCAGCCAGCAAGCGCGCGCGAAAGCTGGCGTGAAGTGGCATCTTCTGAGGCGGATACTGTCTGCTGCC
CTCAAACCTGGCAGATGCACGGTTACGATGCGCCCATACACCAACGTGACCTATCCATACGGTCAATCCGCGCTTGTCCACGGAGAATCCGACGGGTTGT
ACTCGCTCACATTAATGTTGATGAAAAGCTGGCTACAGGAAGGCCAGACGCGAATATTTTTGATGGCGTTAACTCGCGCTTTCATCTGTGGTCAACCGGGCGCTGG
GTCGGTTACGGCCAGGACAGTCGTTTCCGCTGTAATTTGACCTGAGCGCATTTCACGCGCCGAGAAAAACCGCCTCGCGGTGATGGTGTGCGCTGGGATGACG
GCAGTTATCTGGAAGATCAGGATATGTGGCGGATGAGCGGCATTTCCGTGACGTCTCGTTGCTGCATAAACCGACTACACAAATCAGCGATTCCATGTTGCCACT
GCTTTAATGATGATTTTACGCGCGCTGTACTGAGGCTGAAGTTAGATGCGGGGAGTTCGCTGACTACCTACGGGTAAACAGTTCTTTATGGCAGGGTGAACAG
CAGGTCGCCAGCGGCACCGCGCCTTTCCGCGGTGAATATTCGATGAGCGTGGTGGTTATCCGATCGCGTACACACTACGCTGAAACGTCGAAAACCCGAAACTGT
GAGGCGCCGAAATCCGAACTCTATCGTGGCGGTGTTGACTGCACACCGCCGACGGCAGCGTATTGAAGCAGAAAGCCTGCGATGTGCGTTCCGCGAGGTGC
GGATGAAAATGGTCTGCTGCTGCTGAAAGCGCAAGCGGTTGCTGTTCCAGGCGTTAACCGTACAGGATCATCTCTGATGGTCAAGGTGATGGATGAGCAGAG
GATGGTGCAGGATATCTGCTGATGAAAGCAGAAACACTTTAACCGCGTGGCTGATTCGCAATATCGAACCATCCGCTGGTACACCGCTGGGACCGCTAGCGG
CTGATTTGGTGGATGAAGCCAAATTTAAAACCCAGCGCAATTTGACCAACCGCTGACCGTGTACCGGATGATCCGCGGATGAGCGAAACCGGTAACCG
GAATGGTGCAGCGCATGTAATCCCGAGTGATCATCTGCTGCTGGGAATGAATCAGCCACCGCGCTAATCAGACCGCTGTATCGCTGGATCAATC
TGCTGATCTTCCCGCCGCTGAGTGAAGCGCGGAGCGGACGACCCAGCCAGCCATATTTTCCCGATGACGCGCGGTGATGAAGACAGCCGCTT
CCCGCTGAACTGGCAAAATGGTCCATCAAAAATGGCTTTCGCTACCTGGAGAGACGGCCCGCTGATCCTTTCCGAAATACCGCCACCGATGGTAAACAGTCTTGGC
GGTTTCGCTAAACTGGCAGCGCTTTCGCTGATCCCGCTTTACAGGCGCGCTTCGCTGGGACTGGTGGATCAGTCCGCTGATTAATATGATGAAAACGGCA
CCCGTGTGCGCTTACGGCGGTGATTTTGGCGATACCGCCGAAACGCGAGTCTGTATGAACGGTCTGGCTTTTCCGAGCCGCAACCGCCGATCCAGCGCTGAGC
GAAGCAAAACACCCAGCAGCATTTTCCAGTTCCGTTTATCCGGCAAAACATCGAATGACCAAGCGAATACCTGTTCCGCTATAGCGATAACGAGCTCCGCACT
GATGGTGGCGCTGGATGGTAAAGCCGCTGGCAAGCGTGAAGTGCCTGTCCTCCACAAGGTAACAGTTGATTGAATCGCTGAACTACCCGACGCGGA
GAGGCGCCGGGAACTCTGGCTCACAGTACGGCTAGTGAACCGCAAGCGCCGATGGTGCAGAAAGCGGGCACATCAGCGCTGGCAGCAGTGGCGCTGGCGG
AAAACCTCAGTGTGACGCTCCCGCGCGCTCCACCGCATCCCGCATCTGACACCAAGCGAAATGGATTTTGCATCGAGTGGGTAATAGCGTTGGCAATTTAAC
CGCCAGTCAGGCTTTCTTCCAGATGTGGATTGGGCAATAAAAAACAATGCTGACGCGCGCTGGCCGATCAGTTCAACCGTGCACCCGCTGGATAACGACATTTGGCT
AAGTGAAGCGCCGCAATTGACCCCTAACCGCTGGGTGCAAGCTGGGCGAGCGCGGCAATACAGGCGCGAAGCAGCGTTTGGATGTCACGGGAGATACACT
TGCTGATGGCGTCTGATTACGACCGCTACGCGTGGCAGCATCAGGGGAAAACTTATTATACGCGGAAAAACCTACCGGATTGATGGTATGGTCAAATGGCG
ATTACCGTTGATGTTGAAGTGGCGGAGCGGATACACCGCATCCGCGCGGATTTGGCTGAACCTGCCAGCTGGCGCAGGTAGCAGAGCGGGTAAACTGGCTGGATTA
GGGCGCAAGAAAACTATCCGACCGCCTTACTGCGCGCTGTTTGAACCGCTGGGATCTGCCATGTGACAGACATGATACCCCGTACGCTTCCCGAGCGAAAAACG
GTCTGCGCTGCGGACGCGGCAATTTGAATATGGCCACACAGTGGCGGGGACTTCCAGTTCAACATCAGCGCTACAAGTCAACAGCAACTGATGGAAACG
CCATCGCCATCTGCTGACCGGGAAGGACACATGGCTGAATCGACGGTTTCCATATGGGATGGTGGCGGACGATCTCCGGAGCCGCTCAGTATCGCGGAA
TTCAGCTGAGCGCGGCTGCTACCATACCAATGGCTGGTGTGATGTCAAAATAATAAACCGGGGAGCCCATGCTGCCCGTATTTCCGTAAGGAAATCCATTAT
TACTATTAATAAAACAACAATTTGGATGTTCCGTTTACTTCTTTTACTTCTTTATCATGGGAGCCTACTTCCCGTTTTCCCGATTGGCTACATGACATCAAC
CATATCAGCAAAATGATCGGGTATTAATTTGCGGCTATTTCTGCTGCTGCTATTTCAACCGCTGTTGGTCTGCTTTGACAAAACCTGGCTGGCCAAAT
ACCTGCTGGATATACCGCATGTAGTGTGATGTTGCGCGCTTCTTATTTTACTTCTTTCGCGGCACTGTTACAATAACAATTTAGTAGGATCGATTTGGTGGT
ATTTATCAGCTTTTGGTTTAAAGCGCGTGGCCAGCAGTAGAGCAATTTAGGAAAGTACCGCTGCGCAATTTTGAATTTGGCTGCGCGCGGATGTTGGC
TGTGTTGGCTGGCGCTGTGCTCCTGATGTGGCATCATGTTCCACATCAATATCAGTTGTTTCTGGCTGGGCTGGCTGTGCACTCATCTCGCGTTTTTA
CTCTTTTCCGCAAAACGGATGGCGCCTTCTGCGCAGTGTGCCAATGGCGGATGGTGCACCACTTCGCGCTTAGCCCTTAAAGCTGGCAGCTGGAACGTTGCAGCA
GCCAAAACCTGGTTTTTGTCACTGTATGTTTATGGCCTTCTCTGCGCACTGAGTGTGTTGACCAACAGTTTGTCTAATTTCTTACTTCGTTCTTGTCTACCGGTGAAC
AGGATACCGCGGATTTGGCTACGTAAACGACATGGGCGCAATTTCAACCGCTCGATGATGTTCTTTCGCGCACTGATCATTAATCGCATCGTGGGAAAAACCGC
CTGCTGCTGGCTGGCACTATATGCTGTGACGTAATTTGGCTCATCTGCGCACCTCAGCGCTGGAAGTGGTTATCTGAAAACCGTGCATATGTTGAAAATCCG
TTCTGCTGGTGGGCTGCTTAAATATATACCAAGCAATTTGAAAGCGGTTTTGACGGAGCTTTATCGGAGCTTTATCGGCTGTTTCTGCTCTTTAAGCAACTGGGATTT
TATGCTCTACCTGGCGGCAATATGATGAAAACGATCGTTTTCCAGGCGGCTTATCTGGTGGTGGTCTGTTGGCGCTGGGCTTCAATTTAATTTCCGTTGTTACGG
TTAGCGCGCCCGCGCCCTTTCCCTGCTGCGTGTGAGTGAATGAAGTGGCTTAAAGCAATCAATGTCGGATGCGCGCGGAGCGCCTTATCCGACCAAGATATCAT
AACGGAGTGTATCCGCTTGAACATGCCAATGACCGAAAAGATAAGAGCAGGCAAGCTATTACCGATATGTCGGAAGGCTTACCGGAAAAAGACCTTGGGAAAAAC
GTTAATGATGAGTTTAACTACTCGCATCCATCAGA
```



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

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

The screenshot shows the ORFfinder web interface. At the top, there is a search bar with the text "Enter Query Sequence". Below it, there is a section for "Enter accession number, gi, or nucleotide sequence in FASTA format:" with a large text input area. Below the input area, there are fields for "From:" and "To:". Underneath, there is a section for "Choose Search Parameters" with several options: "Minimal ORF length (nt):" set to 75, "Genetic code:" set to "Standard", "ORF start codon to use:" with radio buttons for "ATG" only (selected), "ATG" and alternative initiation codons, and "Any sense codon"; and "Ignore nested ORFs:" with an unchecked checkbox. At the bottom, there are "Start Search / Clear" buttons and a "Submit" button. The page number "21" is visible in the bottom right corner.

Ab initio methods

```
GTGAAGGGCAATCAGCTGTTGCCGCTCTACTGGTGAAGAAAAAACCCCTGGCGCCCAATACGCAAAACCGCTCTCCCCGCGGTTGGCCGATTCAATATGC
AGCTGGCACACAGGTTTCCCAGCTGGAAGACGGGGCAGTGAAGCCGCAACGCAATTAATGTGAGTTAGCTCACTATTAGGCACCCAGGCTTTACACTTATGCTTCC
GGCTCGTATGTTGTGGAAATTTGTGAGCGGATAACAATTTACACAGGAAACAGCTTACCATTGATACGCGTACTGGCCGCTGTTTACAAACGTCGTGACTGGG
AAAAACCTGGCGTACCCAACTTAATCGCTTGCAGACATCCCCCTTTCGGCAGCTGGCCGTAATAGCGAAAGAGCCCGCAGCGATCGCCCTTCCCAACAGTTGGC
CAGCCTGAATGGCGAAATGGCGCTTTGGCTGTTTGGCCGATCCCAAGCGGAGCGGCGGAAAGCTGGCTGGAGTGGCATCTTCTGAGGCGGATACTGCTGCTGCC
CTAAACTGGCAGATGCACGGTTACGATGCGCCCATACACCAACGTGACCTATCCATTACGGTCAATCCCGCTTGTGCCACGGGAATCCGACGGGTTGT
ACTCGCTCACATTTAATGTTGATGAAAGCTGGCTACAGGAAGGCCAGACGCGAATATTTTTGATGGCGTTAACTCGCGGTTTCATCTGTGGTCAACGGCGGCTGG
GTCGGTTACGGCCAGGACAGTCGTTTCCGCTGTAATTTGACCTGAGCGCATTTCACGCGCCGAGAAAAACCGCTCGCCGTTGATGGTGCCTGGCTGGAGTGACC
GCAGTTATCTGGAAGATCAGGATATGTTGGCGGATGAGCGCCATTTCCGTGACGTCTCGTTGCTGCATAAACCGACTACACAAATCAGCGATTCCATGTTGCCACTC
GCTTAATGATGATTTTACGCGCGCTGACTGAGGCTGAAGTTAGATGCGCGGAGTTGCGTGACTACCTACGGGTAACAGTTTCTTATGGCAGGGTGAACG
CAGGTCGCCAGCGGCACCGCGCTTTCGGCGGTGAATTTATCGATGAGCGTGGTGGTTATCCGATCGCGTACACTACGCTGAAACGTCGAAAACCCGAAACTGT
GGAGCGCCGAAATCCGAATCTCTATCGTGGCGGTTGAACTGCACACCGCCAGCGCAGCGTGATTGAAGCAGAAGCCTGCCGATGTCGTTTCCGCGAGGTCG
GGATGAAAAATGGTCTGCTGCTGCTGAAGCGCAAGCCGTTGCTGATTCGAGGCGTTAACCGTACAGGACATCATCTGCAATGGTCAGGTCATGGATGAGCAGAC
GATGGTGCAGGATATCTGCTGATGAAGCAGAACTTAACGCGCTGGCTGTTCCGATTTCCGAACCATCCGCTGTGGTACACGCTGCGGACCGCTAGCGG
CTGATGTTGGTGGATGAAGCCAAATTTGAAACCCAGCGCATGGCCAAATGCTGACCGATGATCCGCGTGGCTACCGGCGATGAGCGAACCGTAAACGC
GAATGGTCAGCGCGATCTAATCACCAGTGTGATCATCTGGTCCGTCGGGAATGAATCAGGCCACGGCGCTAATCAGACGCGCTGTATCGCTGGATCAATC
TGTGATCTTCCCGCCGGTGCAGTATGAAGCGCGGAGCGACACCCAGCATATTTTCCCGGATGACGCGCGCTGGATGAAGACCGCCCTT
CCCGCTGTGCCAAATGTTCCATCAAAAATGCTTTCGCTACCTGGAGAGACGGCCCGCTGATCTTTCGCAATACGCCACCGCATGGTAAACAGCTTTCGCG
GGTTTCGTAATACTGGCAGCGGTTTCGTCAGTATCCCGCTTACAGGGCGGCTTCGCTGGGACTGGTGGATCAGTCGCTGATTAATATGATGAAAAACGGCAA
CCCGTGGTCCGCTTACGGCGGTGATTTTGGCGATACCGCGAACCGATCCGCGACTGCTGTATGAACGGTCTGGCTTTTGGCGACCGCAACCGCATCCAGCGCTGAG
GAAGCAAAACACCAGCAGCAGTTTTTCCAGTTCCGTTTATCCGGCAAAACCATGGAATGACCCAGCGAATACCTGTTCCGCTATAGCGATAACGAGCTCCTGCACCT
GATGGTGGCGCTGGATGGTAAAGCCGCTGGCAAGCGGTGAAGTGCCTCTGGATGTCGCTCCACAAGGTAACAGTTGATTGAACCTGCCGTAACACTACCGCAGCCGA
GAGGCCCGGGCAACTCTGGCTCACAGTACGCGTAGTGAACCGAACCGCCGATGGTGCAGAAAGCCGGGACATCAGCGCTGGCAGCAGTGGCGCTCGGGCG
AAAACTCAGTGTGACGCTCCCGCGCGCTCCACCGCCATCCGCTCAGCCACCGCAAGCAATGGATTTTGCATCGAGTGGGTAATAGCGTTGGCAATTTAAC
CGCCAGTCAGGCTTTCTTACAGATGGTGGTGGGCAATAAAAAAAGTGTGACGCGCGCTGGCGATCAGTTACCCGCTGCACCGCTGGATAACGACATTTGGCGT
AAGTGAAGCGAACCCGATTAACCGCTGGGTGCAAGCGTGGGAGCGGCGGCGCCATTACCAGCCGGAAGCAGCGTTGTTGCAGTGCACGGGAGATACACT
TGCTGATGCGGCTGATTACGACCGCTACCGGTTGGCAGCATCAGGGGAAAACTTATTATACGCGGAAAAACCTACCGGATTGATGGTATGGTCAAATGGCG
ATTACCGTTGATGTTGAAGTGGCGGAGCGGATACACCGCATCCGCGCGGATGGCTGAACTGCCAGCTGGCGAGGTAGCAGAGCGGGTAAACTGGCTCGGATTA
GGCGCGCAAGAAAACTATCCGACCGCCCTACTGCGCGCTGTTTACCGCTGGGATGTCGCCATTGTCAGACATGATACCCCGTACGCTTCCCGAGCGAAAAACG
GTCTGCGCTGCGGACCGCGGAATTAATATGGCCACACAGTGGCGGGGCGGACTTCCAGTTCAACATCAGCCGCTACAGTCAACAGCAACTGATGAAACCGA
CCATCGCCATCTGCTGCACGGGAAAGGACAGTGGCTGAATCGACGTTTCCATATGGGGATTGGTGGCGGAGCTCCTGGAGCCCGCTCAGTATCGGGCGGAA
TTCAGCTGAGCGCGGCTGCTACCATTAACAGTTGGTCTGGTGTCAAAAATTAATAAACCGGCAAGCCATGCTCGCCGATTTTCGCTAAGGAAATCCATTATG
TACTATTAAAAAACATAAATTTGGATGTTCCGTTTATCTTTTACTTTTATCTTTATCAGGGAGCTACTTCCCGTTTTCGCGATTGGCTACGATGACATAAC
CATATCAGCAAAAGTACACGGTATTAATTTGCGGCTATTTCTGCTGCTGCTATTTCCAAACCGCTGTTGGCTGCTTTCTGACAACTCGGGCTGGCAAAAT
ACCTGCTGGGATATACCGCATGTTAGTGTGTTGCGCGCTTCTTATTTATCTTCGGGCCACTGTTACAATAACAATTTAGTAGGATGATGTTGGTGGT
ATTTATCAGCTTTTGTTTTAAACCGCGTGGCCAGCAGTAGAGCAATTAATGAAGAAATCAGCGCTGCGCAATTTGCAATTTGGCTGGCGCGGATGTTGGC
TGTGTTGGCTGGCGCTGTGTCCTCGATGTCGGCATCATGTTACCATCAATAATCAGTTGTTTTCGCTGGCTGGGCTCGGCTGTCACATCATCTCGCGTTTTA
CTCTTTTTTCGCAAAAGGATGCGCCCTTCTCGCCAGGTTGCCAATGCGGTAGTGGCCAACTCCGCTATTAGCCTTAAAGCTGGCAGCTGGAACGTTGTGCAGCA
GCCAAACTGTGGTTTTTGTCACTGTATGTTATGGCGTTTCTGTGCACTACGATGTTTGAACCAAGTTTGTCTAATTTCTTACTCTGTTCTTGTACCGGTGAAC
AGGGTACGCGGGTATTTGGCTACGTAAACGACATGGCGGAATTACTTAAACCGCTGATGTTGTTTTCGCGCACTGATCATTAATCGCATCGTGGGAAAAACCGCC
TCTGCTGCTGGCTGGCATAITATGCTGCTGACTGATTTTGGCTCATCTGCGCCACTCAGCGCTGGAAGTGGTATTCTGAAAACCGCTGCATATGTTGAAGTACCG
TTCTGCTGGTGGGCTGCTTTAAATATAATACCAGCCAGTTGAAAGCGGTTTTACGCGGAGCTTATCTGGTCTGTTTGTGCTTTTAAAGCAACTGGCGGATGATTT
TATGCTGACTGGCGGGCAATATGATGAAAGCATCGTTTCCAGGCGGCTTATGTTGGTGGGCTGTTGGTGGCGGCTGGGCTTACCTTAATTTCCGTGTTTACCG
TTAGCGCGCCCGGCGCCCTTTCCTGCTGCTGCTGCTGAGTGAATGAAGTGGCTTAAAGCAATCAATGTCGGATGCGCGCGGAGCGCCCTATCCGACCAACATATCAT
AACGGAGTGTGCTGATTAAGCATGCAATGACCGAAAGATAAGAGCAGGCAAGCTATTACCGATGTGCGAAGGCTTACCGGAAAAAGACTCTGGGAAAAAC
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

```
GTGAAGGGCAATCAGCTGTTGCCGCTCTACTGGTGAAGAAAAAACCCCTGGCGCCCAATACGCAAAACCGCTCTCCCCGCGCTTGGCCGATTCAATATGC
AGCTGGCACACAGGTTTCCCAGCTGGAAGACGGGGCAGTGAAGCCCAACGCAATTAATGTGAGTTAGCTCACTATTAGGCACCCAGGCTTTACACTTTATGCTTCC
GGCTCGTATGTTGTGGAAATTTGAGCGGATAACAATTTACACAGGAAACAGCTTACCATTGATACGCGTACTGGCCGCTGTTTACAAACGTCGTGACTCGGG
AAAACCCCTGGCGTTACCCAACTTAATCGCTTGGACGACATCCCCCTTTCGGCAGCTGGCCGTAATAGCAGAAAGGCCCGCACCGATCGCCCTTCCCAACAGTTGGC
CAGCTTGAATGGCGAAATGGCCGTTTGGCTGAGTGAAGAGCTGGCTACACAGGAGCGGAGCGGTAAGCTGGCTGGAAGTGGCATTCTTCCGAGCCGATACTCGTCCGCC
CTCAAACCTGGCAGATGCACGGTTACGATGCGCCCATACACCAACGTGACCTATCCATTACGGTCAATCCCGCTTGTCCACGGGAATCCGACGGGTTGT
ACTCGCTCACATTTAATGTTGATGAAAGCTGGCTACAGGAAGGCCAGACGCGAATATTTTTGATGGCGTTAACTCGCCGTTTCACTGTGGTCAACGGCGCTGG
GTCGGTTACGGCCAGGACAGTCGTTTCCGCTGTAATTTGACCTGAGCGCATTTCACGCGCCGAGAAAAACCGCTCGCCGTTGATGGTGCCTGGCTGGAGTGACC
GCAGTTATCTGGAAGATCAGGATATGTTGGCGGATGAGCGCCATTTCCGTGACGTCTGCTTGGCTGCATAAACCGACTACACAAATCAGCGATTCCATGTTGCCACTC
GCTTAATGATGATTTTACGCGCGCTGACTGGAAGCTGAAGTTAGATGCGCGGAGTTCGCTGACTACCTACGGGTAACAGTTTCTTATGGCAGGGTGAACG
CAGGTCGCCAGCGGACCCGCGCTTTCGGCGGTGAATTTATCGATGAGCGTGGTGGTTATCCGATCGCGTCAACACTACGCTGAAACGTCGAAAACCCGAAACTGT
GGAGCGCCGAAATCCGAATCTCTATCGTGGCGTGGTGAACGTGCACACCCGCGCAGCCGACGCTGATTGAAGCAGAAAGCTGCGATGTCGTTTCCGCGAGGTGC
GGATGAAAAATGGCTGCTGCTGCTGAAGCGCAAGCCGTTGCTGATTCGAGCGCTTAAACCGTACAGGACATCATCTGATGGTCAGGTGATGGATGAGCAGAC
GATGGTGCAGGATATCTGCTGATGAAGCAGAAACAATTTAACGCCGTGGCTGATTCGCAATATCCGAACCATCCGCTGTGGTACACGCTGCGGACCCGTAGGGC
CTGTATGGTGGGATGAAGCCAAATTTGAAACCCAGCGCATGGTCCAAATGCTGACCGATGATCCGCGTGGCTACCGGGGATGAGCGAACCGGTAACGC
GAATGGTCAGCGCGATGTAATCACCAGTGTGATCATCTGGTCGCTGGGAATGAATCAGGCCACGGCGCTAATCAGACGCGCTGTATCGCTGGATCAATC
TGTGATCTTCCCGCCGGTGCAGTATGAAGCCGGGAGCGGACACCCAGCCGATGTAATTTGCCGATGACGCGCGCGTGATGAAAGACAGCCCTT
CCCGCGCTGCGCAAAATGGTCCATCAAAAATGGCTTTCGCTACCTGGAGAGACGGCCCGCTGATCTTTCGCAATACGCCACGCGATGGGTAACAGCTTGGC
GGTTTCGTAATACTGGCAGCGGTTTCGTCAGTATCCCGCTTACAGCCGGCTTCGCTGGGACTGGTGGATCAGTCGCTGATTAATATGATGAAAAACGGCAA
CCCGTGGTCCGCTTACGGCGGTGATTTTGGCGATCCCGGCAACGATCGCCAGTCTGTATGAACGGCTGGTCTTGGCGACCGCAACCGCATCCAGCGCTGAGC
GAAGCAAAACACCCAGCAGCATTTTTCCAGTCCGTTTATCCGGCAAAACATCGAATGACCCAGCGAATACCTGTTCCGTCATAGCGATAACGAGCTCTGCACCTG
GATGGTGGCGCTGGATGGTAAGCCGCTGGCAAGCGGTGAAGTGCCTCTGATGTCGCTCCACAAGGTAACACGTTGATTGAAGCTGCTGAACTACCCGACGCGGA
GAGGCCCGGGCAACTCTGGCTCACAGTACGCGTAGTGCAACCGCAACCGCCGATGGTGCAGAAAGCCGGGACATCAGCGCTGGCAGCAGTGGCGCTCGGGCG
AAAACCTCAGTGTGACGCTTCCCGCGCGCTCCACCCGATCCCGCATCGACACCCAGCGAAATGGATTTTGCATCGAGTGGGTAATAGCGTTGGCAATTTAAC
CGCCAGTCAGGCTTTCTTCCAGATGGTGGATGGGCAATAAAAAAAGTGTGACGCGCGCTGGCCGATCAGTTACCCGCTGCACCCGCTGGATAACGACATTTGGCCT
AAGTGAAGCGAACCCGATTAACCCGCTGGGTGCAAGCGCTGGGAGCGGCTGGGAGCGGCGCCATTACCAGCCGGAAGCAGCGTTTGGCAGTGCACGGGAGATACAT
TGCTGATGCGGCTGCTGATTACGACCGCTACCGGTTGGCAGCATCAGGGGAAAACTTATTATCAGCCGGAAAAACCTACCGGATTGATGGTAGTGGTCAAATGGCG
ATTACCGTTGATGTTGAAGTGGCGGAGCGATACCCGATCCGGCGCGGATGGCTGAACTGCCAGCTGGCGAGGTAGCAGAGCGGGTAAACTGGCTCGGATTA
GGCGCGCAAGAAAAATATCCGACCGCCCTACTGCGCGCTGTTTTCGACCGCTGGGATGTCGCCATTGTCAGACATGATACCCCGTACGCTTCCCGAGCGAAAAACG
GTCTGCGCTGCGGACCGCGGAATTAATGATGGCCACACAGTGGCGGGGGACTTCCAGTTCAACATCAGCCGCTACAGTCAACAGCAACTGATGAAACCCAG
CCATCGCCTGCTGTCACGGGGAAGGACAGATGGCTGAATCGACGGTTTCCATATGGGGATTGGTGGCGGAGCTCCTGGAGCCGCTCAGTATCGGGCGGAA
TTCAGCTGAGCGCCGGTGCCTACCATTAACAGTTGGCTGGTGTCAAAAATAATAAACCGGCAAGCCGATGTCGCGCGTATTCGCGTAAGCCGATCCATTAAG
TACTATTAAAAAACAACAATTTGGATGTTCCGTTATGCTTTTCTTACTTTTATCAGTGGGAGGCTACTTCCCGTTTTCGCGATTGGCTCAGTGCATGACATAAC
CATATCAGCAAAAGTATACCGGATATATTTTGGCCGATTTCTGCTGCTGCTATTCACACCGCTGTTGGCTGCTTTCTGACAAAACCTGGGCTGGCGCAAT
ACCTGCTGGGATATACCGCATGTAGTGTGATGTTGGCCGCTTCTTATTTATTCCTGGGCCACTGTTACAATAACAATTTAGTAGGATGATGTTGGTGGT
ATTTATAGCTTTTGTTTAACCGCGGTGCCGCCAGTAGAGCAATTAATGACAAAGTCAAGCCGCTGCAGTAATTTGCAATTTGGCTGGCGCGGATGTTTGGC
TGTGTTGGCTGGCGCTGTGCTCGTATGTCGGCATCATGTTACCATCAATAATCAGTTGTTTTCGCTGGCTGGGCTCGGCTGTCACATCATCTCGCGTTTTA
CTCTTTTTCCGCAAAACGGATGCGCCCTTCTCGCCAGGTTGCCAATGCGGTAGTGGCAACCATTCGCGATTAGCCCTTAAGCTGGCAGTGAAGCTTGCAGACA
GCCAAAACCTGGTTTTGTCAGCTGTATGTTATGGCGTTTCTGTGCACTACGATGTTTGGCAACAGTTTGTCTAATTTCTTACTCTGTTCTTGTACCGGTGAAC
AGGGTACGCGGGTATTTGGCTACGTAAACGACAAATGGCGGAATTACTTAAACCGCTGATGATGTTCTTTCGCGCACTGATCATTAATCGCATCGTGGGAAAAACCGCC
TCTGCTGCTGGCTGGCAGTATGTCGCTGACGATTTAGGCTCATCTGTCGCCACTCAGCCGCTGGAAGTGGTTATCTGAAAACCGCTGCATATGTTGAAGTACCG
TTCTGCTGGTGGGCTGCTTTAAATATATACCAGCCAGTTGAAAGCGGTTTTTCAGCGGACTTATCTGGTCTGTTTCTGCTTCTTAAAGCAACTGGCGGATGATTT
TATGCTGACTGGCGGGCAATATGATGAAAGCATCGGTTTCCAGGCGGCTTATCTGGTGGTGGTCTGTTGGCGGCTGGGCTCACCTTAATTTCCGTGTTTCCG
TTAGCGGCGCCCGCCGCTTTCCTGCTGCGTGCAGTGAATGAAGTGCCTTAAAGCAATCAATGTCGGATGCGCGCGGAGCCGCTATCCGACCAACATATCAT
AACGGAGTGTGCTGATTAAGCATGCGCAATGACCGAAAGATAAGAGCAGGCAAGCTATTACCGATGTGCGGAAGGCTTACCGGAAAAAGACTCTGTTGGGAAAA
GTTAATGATGAGTTTAACTACTCGCATCCATCAGA
```

2 mains steps :

1) List Open Reading Frames (ORFs) 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
GCTGGATCAAATCTGTGATCCTTCCCGCCCGGTGCAGTATGAA
GGCGGGCGGAGCCGACACCACGGCCACCGATATTATTTGCCCGA
TGTACGCGCGCGTGGATGAAGACCAGCCCTTCCCGGCTGTGCC
GAAATGGTCCATCAAAAATGGCTTTCGCTACCTGGAGAGACGC
GCCCGCTGATCCTTTGCGAATACGCCACGCGATGGGTAAACAG
```

...

But overlaps can be valid too...

Select valid ORFs because of overlaps between phases

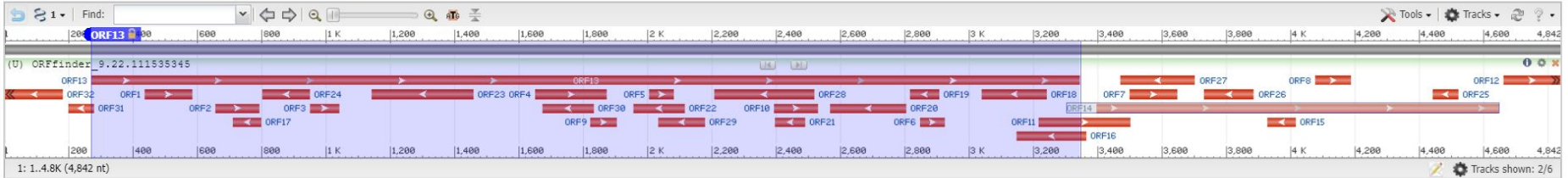
< ORFfinder submitting page

> Help

Open Reading Frame Viewer

Sequence

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



Six-frame translation...

ORF13 (1024 aa) Display ORF as... **Mark**

Mark subset... Marked: 0 Download marked set as Protein FASTA

```
>|c1|ORF13
MTMITHDLSLAVVLQRDDWENPGVTQLNRLAAHPPFASWRISSEEARDRPSQ
QLRSLNGEVRFAHFPAPFAVPSVILECDLPEADTVVPSM@PHGYDAP1
YTMVYPTIVNIPFPVTEINPTGYSYSLTFWVDESILQESQTRIFDQVISA
FHLNCGNRWVGYGQSRLPSEFDLSAFLRAGENLRVWLRMSDQSVLED
QMWNRMSGIFRQVSLHKPTTQISDFHVRFRNDFSRVLAEVQKCGE
LRDYLRVTVSLKQGETQVASGTAPFGGEIIDERGGYADRVTLRNINPK
LHSAEIPNLYRAVVELHTADGTLIEAEACDVGFREVRIEHGLLLNGKPL
LISGVRHREHPLHGGVDEQTRVQDILLKWKQWIFUAVRCSHPHPLHY
TLCDRYGLVYVDEANIEETHGSRWPMRLTDDPNLPAHSRVRTRVQRDRN
HPSVIIISLGNESGHGHDALYRNIKSDPSPRPVQEGGADTTADII
CPMYARVDEQDQFPVAPKWSIKKWLSPGETRPLLCEYAHAMGNSLGGF
```

ORF13 Marked set (0)
SmartBLAST SmartBLAST best hit titles...
BLAST BLAST

BLAST Database:
 UniProtKB/Swiss-Prot (swissprot)

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

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)

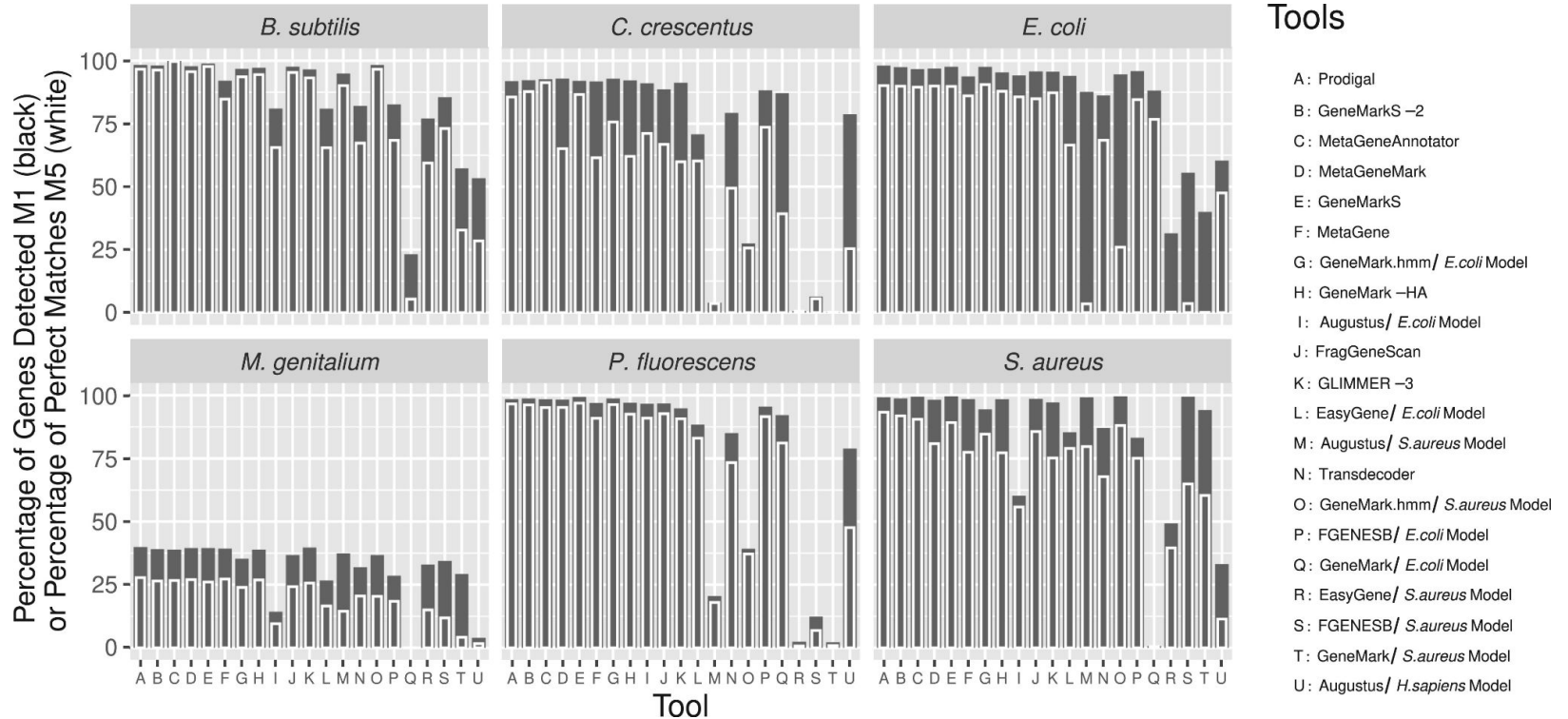
How to select valid ORFs in *ab initio* methods ?

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

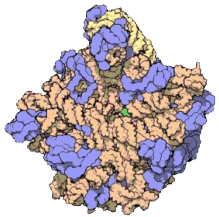
- Coding genes are detected via *ab initio* methods by finding all ORFs and selecting the most relevant one based on complex empirical rules
- This approach allows predicting **even unknown coding genes** contrary to an alignment approach, but :
 - some genes can be missed
 - some genes can be truncated 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)

Prediction of non coding genes

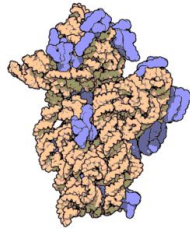
Different kinds of non coding genes : rRNA

- Several clusters composed of 3 rRNA genes can be detected in the genome

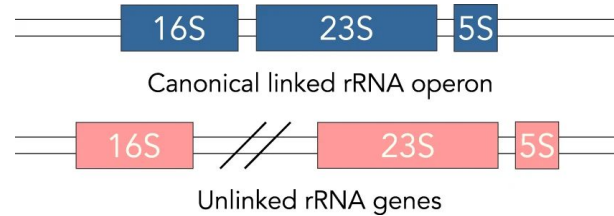
credit: Wikipedia



Large subunit 50S
(23S+5S+proteins)



small subunit 30S
(16S+proteins)



Brewer *et al.*, 2020

- One or some tRNA genes often lie between 16S and 23S genes (case of linked rRNA genes)
- 16S rRNA genes are conserved in some regions and variables in other ones (V1,V2...):
 - Largely used for phylogeny (Carl Woese, 1977)
 - And for metabarcoding to describe the biological content of a sample (“metagenomic”)

16S rRNA genes



credit: <https://www.repertoire.co.jp/en/research/technology/16srrnainfo/>

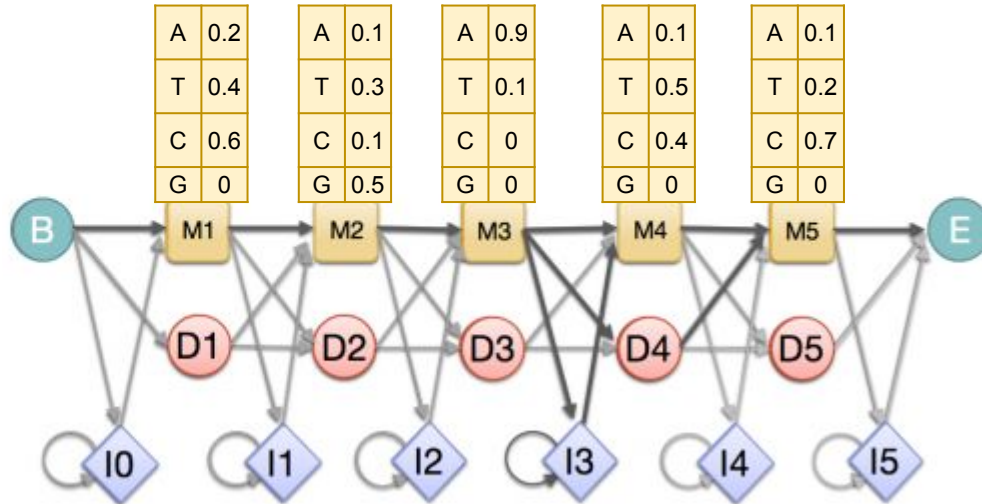
Different kinds of non coding genes : rRNA

- Detection of rRNA can be performed using RNAmmer
 1. patterns of rRNA genes are stored as Hidden Markov Model profiles (HMM) :

An alignment
of rRNA genes:

AAATT
TTATC
CTACT
ATACC
CCTTC
TGATA
CGAAC
TGACC
CGACC
CGATC

modeled
into
HMM



Different kinds of non coding genes : rRNA

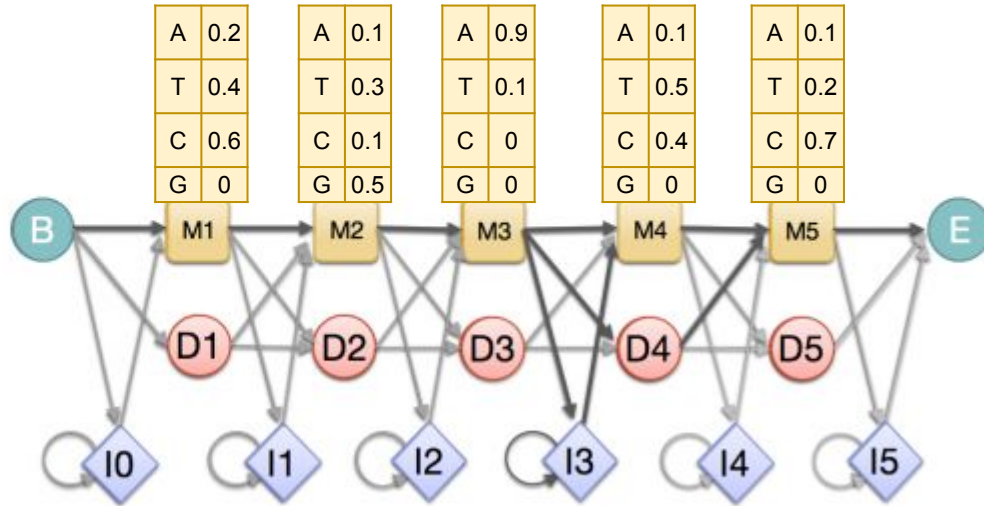
- Detection of rRNA can be performed using RNAmmer
 2. sequences of our genome are aligned on the HMM profiles :

sequences from the genome

...CTAGTC...

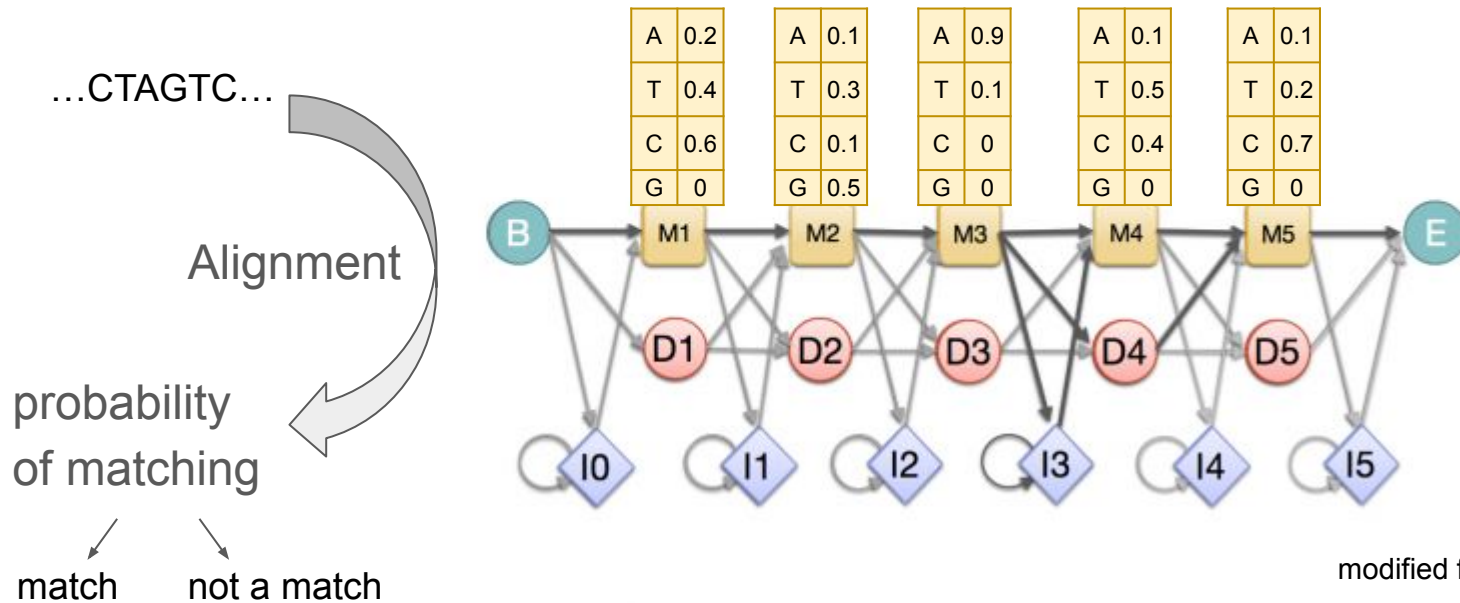
Alignment

probability
of matching



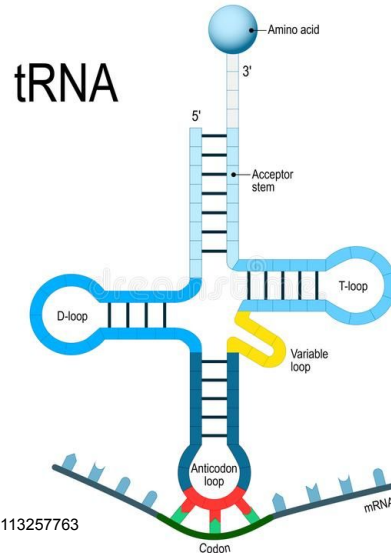
Different kinds of non coding genes : rRNA

- Detection of rRNA can be performed using RNAmmer
 3. the sequence having the best probability of matching corresponds to the genes encoded in the HMM profiles



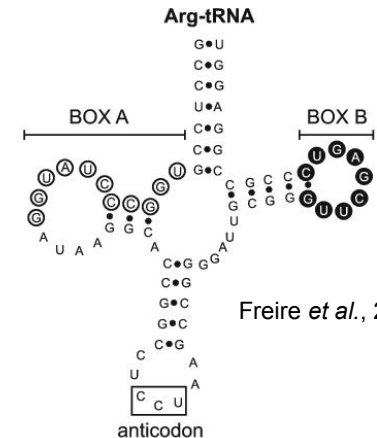
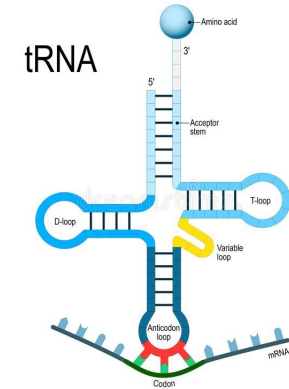
Different kinds of non coding genes : tRNA

- tRNA : several clusters of tRNA scattered all along the genome
 - One or several tRNAs for each codon
 - The pool of tRNA drives the codon usage
 - Very stable across bacterial genomes (except the variation loop and, of course, the anticodon)



Different kinds of non coding genes : tRNA

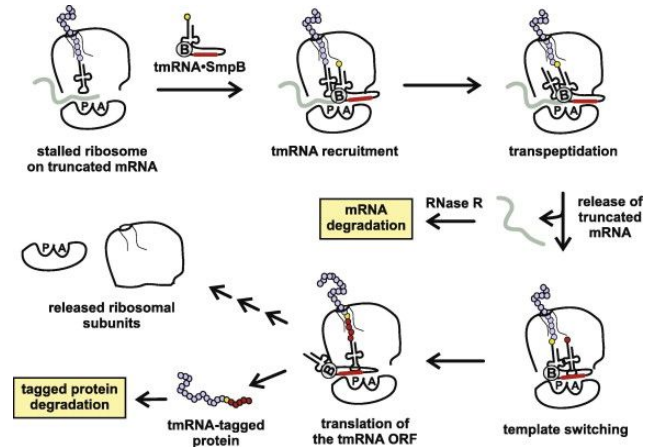
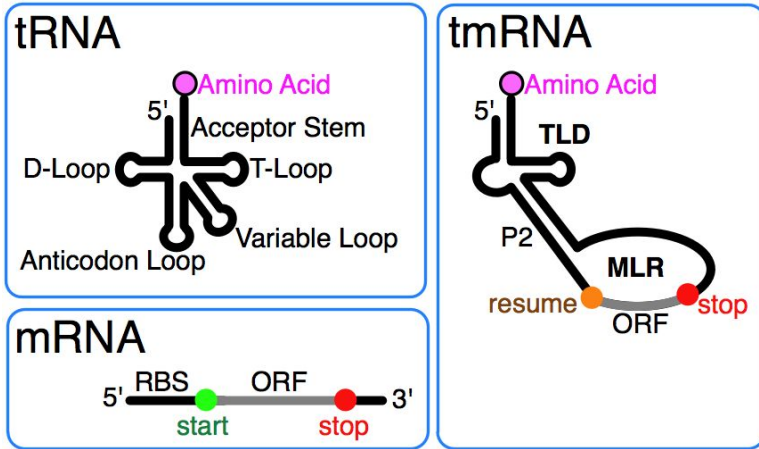
- Detection of tRNA can be performed using ARAGORN
- Here is the algorithm (an expert recipe):
 1. Searches for partially mismatched, non-gapped, occurrences of the sequence GTTC (subset of the B box consensus sequence)
 2. Around each hit, the algorithm attempts to construct a T-loop from five to nine bases long and a T-stem from 4 to 5 bp long.
 3. To detect tRNA genes, the sequence is searched from 28 to 85 bases upstream of this T-stem for the sequence motif TRGYNAA, a subset of the A box consensus sequence which allows for a D-stem from 3 to 4 bp long.
 4. Around the motif, a D-loop from five to 11 bases long and containing the sequence A- - - -GG-R is constructed.
 5. A 7–9 bp long A-stem is constructed using the sequence from two to three bases upstream of the D-stem and immediately downstream of the T-stem.
 6. The length of the V-loop is allowed to vary from three to 25 bases, upstream of the T-stem.
 7. Finally, the 3' end of the C-stem is constructed by searching between the D- and T-stems for a sequence that is complementary to the 5' end of the C-stem, immediately downstream of the D-stem and spacer base.
 8. Limited use of tertiary structures contacts between the T-loop, V-loop and D-loop is made.
 9. If position 55 in the T-loop is a non-consensus G, then a non-consensus TT at positions 18 and 19 in the D-loop is given an improved score.
- Useless to have a complete understanding (to my mind)



Different kinds of non coding genes : tmRNA

- tmRNA allows recycling stalled ribosomes during the translation of a truncated mRNA (so without any stop codon)

credit: wikipedia



Hayes et al., 2010

- Detection of tRNA can also be performed using ARAGORN using an expert recipe

Different kinds of non coding genes : ncRNA

- Non coding RNAs correspond to a set of RNA (having varieties of functions) which are not mRNA, rRNA, tRNA, and tmRNA
- Detection of tRNA can be performed using Infernal based on Rfam, which is database of HMM profiles of ncRNA



```

UUCCUCCCUCAACCGCUCCACGU5'GU GUCCC UCCCGAAGCUCCGCGCU CGG
UUCCUCCCUCAACCGCUCCACGU5'GU GUCCC UCCCGAAGCUCCGCGCU CGG
UUCCUCCCUCAACCGCUCCACGU5'GU GUCCC UCCCGAAGCUCCGCGCU CGG
UUCCUCCCUCAACCGCUCCACGU5'GU GUCCC UCCCGAAGCUCCGCGCU CGG
UUCCUCCCUCAACCGCUCCACGU5'GU GUCCC UCCCGAAGCUCCGCGCU CGG
UUCCUCCCUCAACCGC-CACGU5'GU GUCCC UCCCGAAGCUCCGCGCU CGG
>>>.....<<<<<<<<<<<<<<. <<.....>>>>>>>>>>>>>>>>
    
```

Conclusion

- Coding genes still correspond to the main part of genomes but non coding genes can't be ignored
- Detection of non coding genes can help annotation of coding genes (overlaps)
- The diversity of non coding genes is detected via 2 main kinds of approaches
 - Alignment on a database of HMM profiles
 - Expert recipes ruling a specific pattern to match

Introduction to functional annotation using Bakta

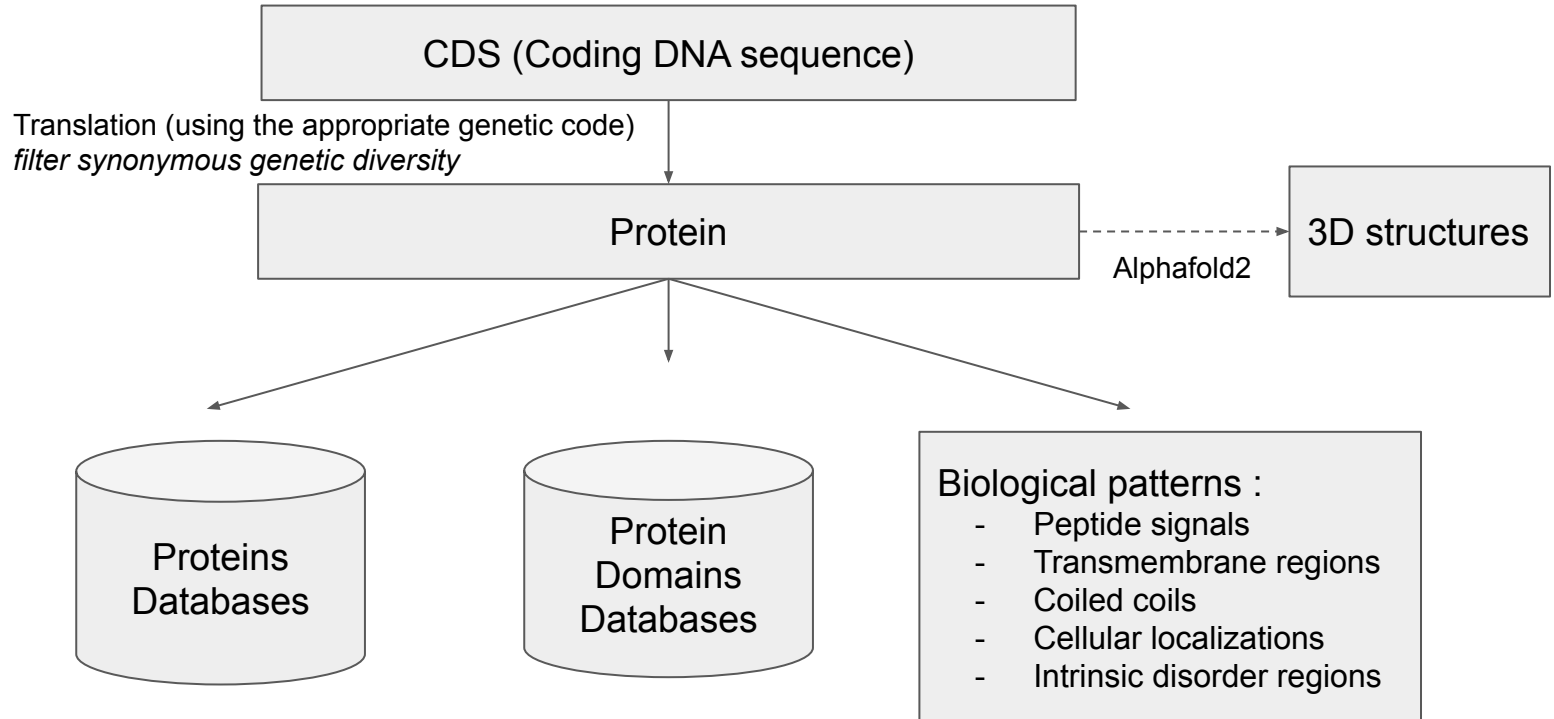
What's in an annotation?

- Location
 - which sequence? *chromosome 2*
 - where on the sequence? *100..659*
 - what strand? *-ve*
- Feature type
 - what is it? *protein coding gene*
- Attributes
 - protein product? *alcohol dehydrogenase*
 - enzyme code? *EC:1.1.1.1*
 - subcellular location? *cytoplasm*
 - note? *beer processing*

Source : <https://training.galaxyproject.org/training-material/topics/genome-annotation/tutorials/annotation-with-prokka/slides.html#9>

What is the functional role of an identified coding genes ?

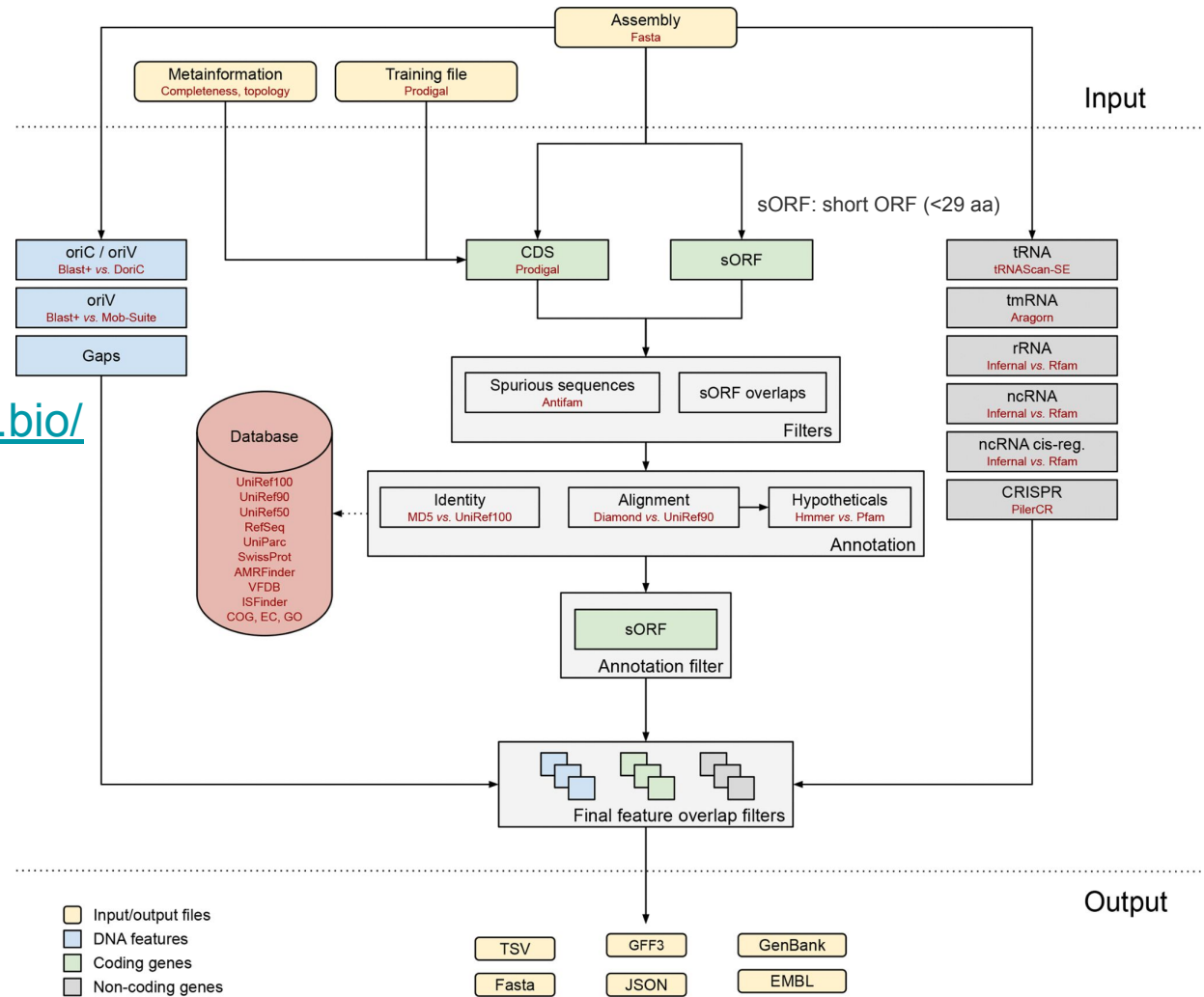
Simplify analysis pipeline :



Bakta

- successor of Prokka
- available in Galaxy

<https://bakta.computational.bio/>



Bakta

- successor of Prokka
- available in Galaxy

<https://bakta.computational.bio/>

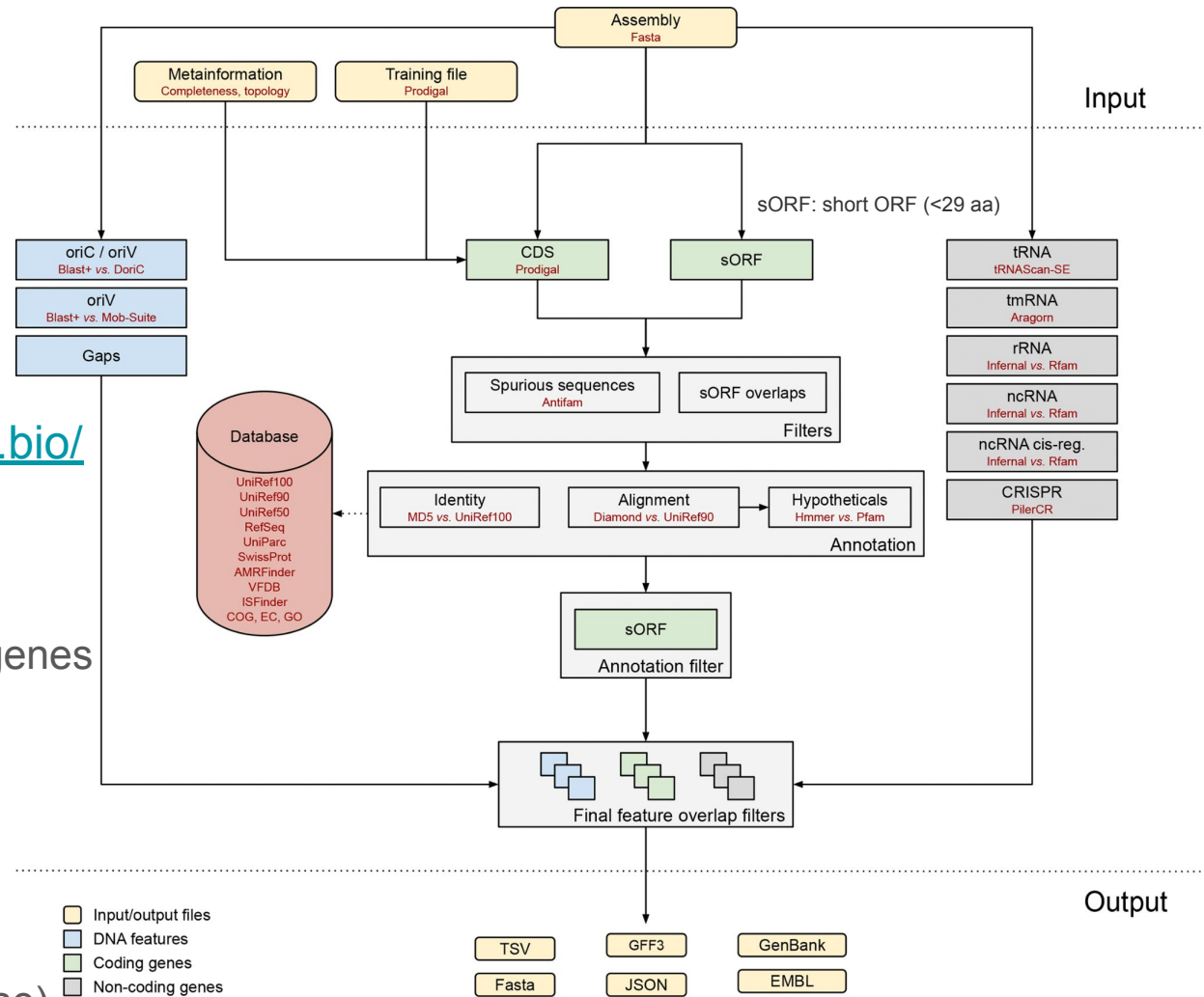
- + Peptide signal (DeepSig)
- + Antimicrobial resistance genes

(AMRFinderPlus) :

- NCBI tool
- AMR proteins: 6990
- Point mutations: 1309
- HMMs: 743

+ VFDB

(Virulence Factor Database)



Tutorials

<https://training.galaxyproject.org/training-material/topics/genome-annotation/tutorials/bacterial-genome-annotation/tutorial.html#is-inserion-sequence-elements>

Bakta,

Version de Bakta 1.8.2

The bakta database : v5.0_2023-02-20

The amrfinderplus database : v3.11_2023-04-17.1

PlasmidFinder,
IntegronFinder,
ISEScan,
JBrowse