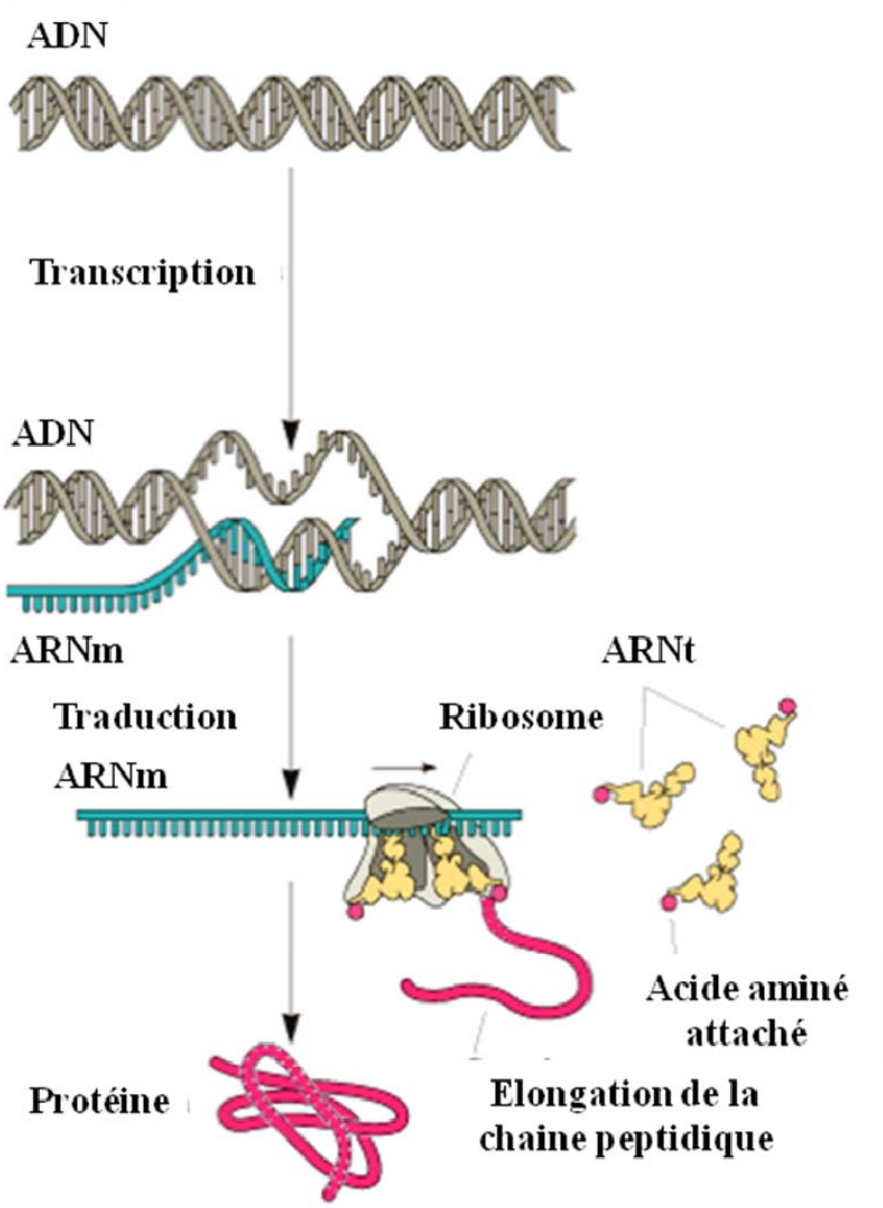
Annotation d'une séquence bactérienne !

cf. Cours d'Hélène DEBAT

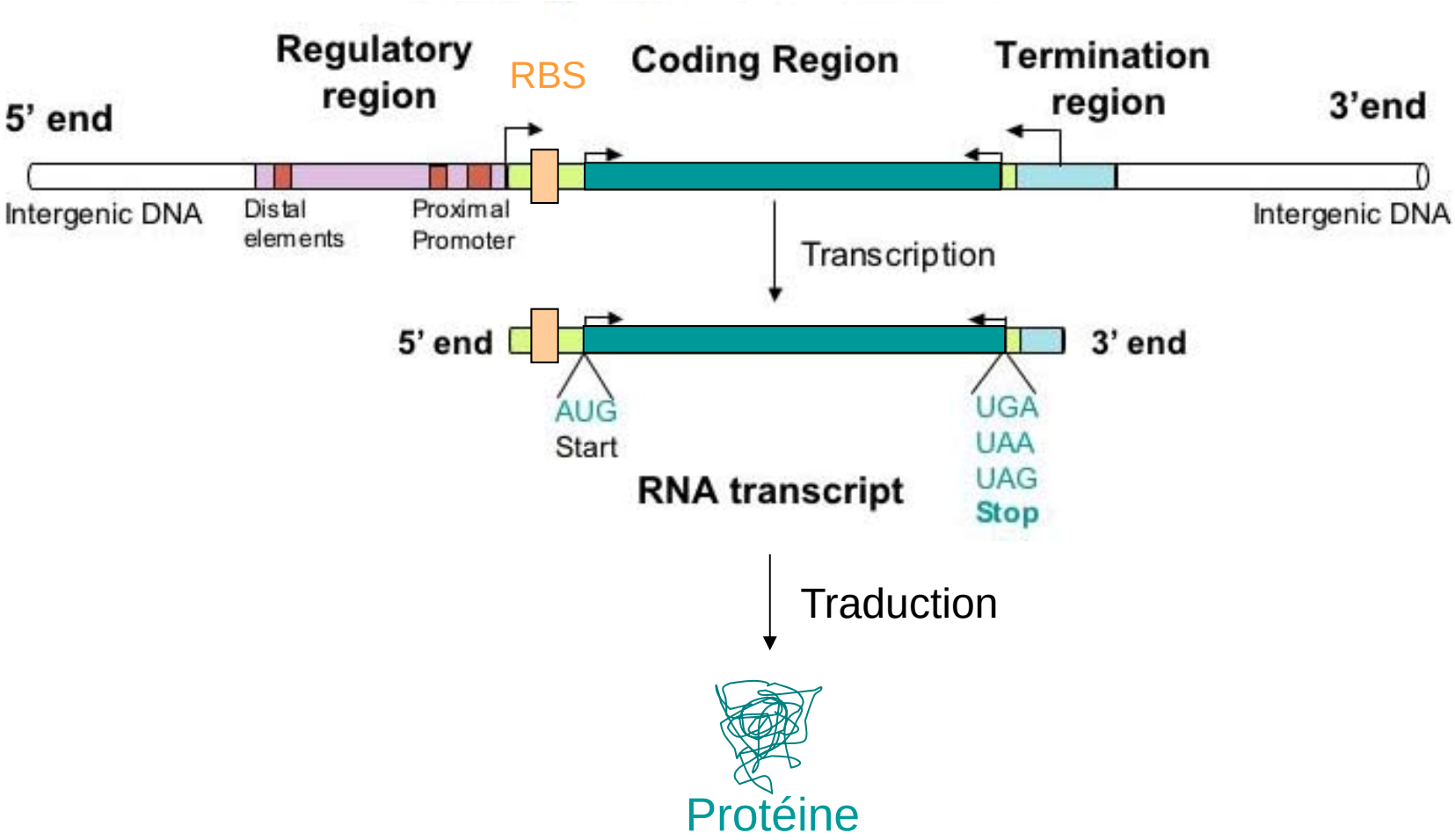
(Annotation des génomes)



Dans ce TD nous allons tenter de retrouver

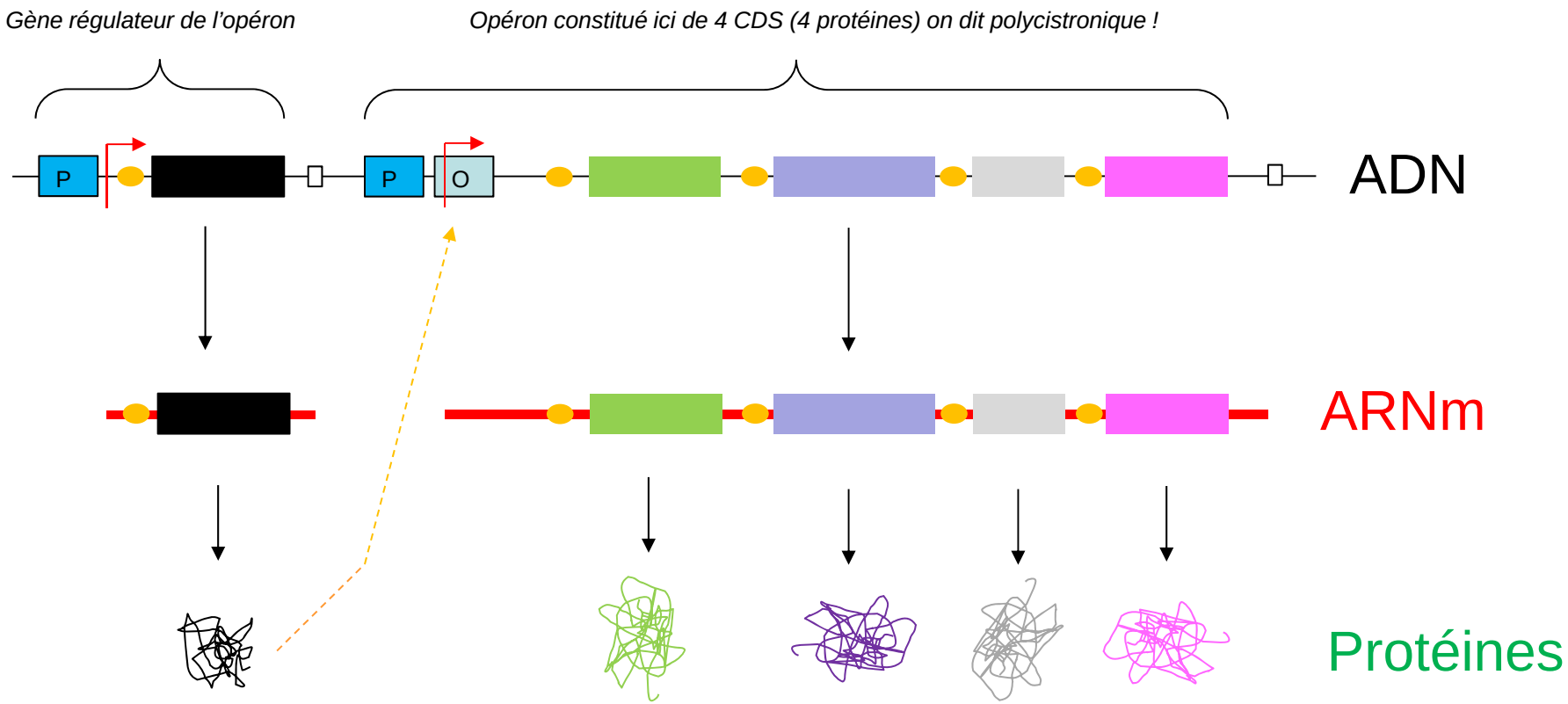
la structure d'un gène bactérien !

Structure d'un gène procaryote



Oui mais au final c'est très souvent sous forme d'opéron …

Structure d'un opéron



RBS

P = région promotrice (constituée des boites -35 et -10)

O = opérateur (région de régulation de l'opéron)

Début de transcription (point +1 de transcription)

Terminateur de la transcription

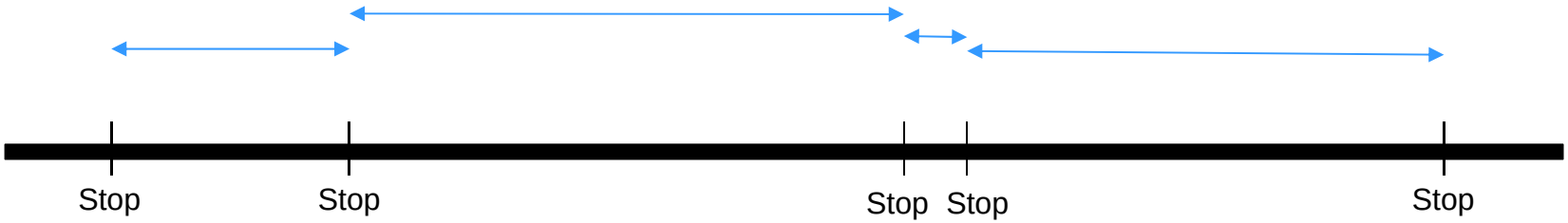
But : annoter votre séquence

- Recherche d'ORF et donc de séquence codante CDS (ORF Finder et GeneMark)
- Recherche de signaux de transcription (motifs de promoteurs grâce à DNA Pattern)
- Recherche de signaux de traduction (recherche de RBS avec DNA Pattern)
- Recherche de signaux de régulation (avec DNA Pattern aussi)

1. Recherche d'ORF (Open Reading Frame)

<u>Définition</u> : une ORF commence après un codon stop, se termine avec un codon stop et contient une séquence codante (CDS) potentielle …

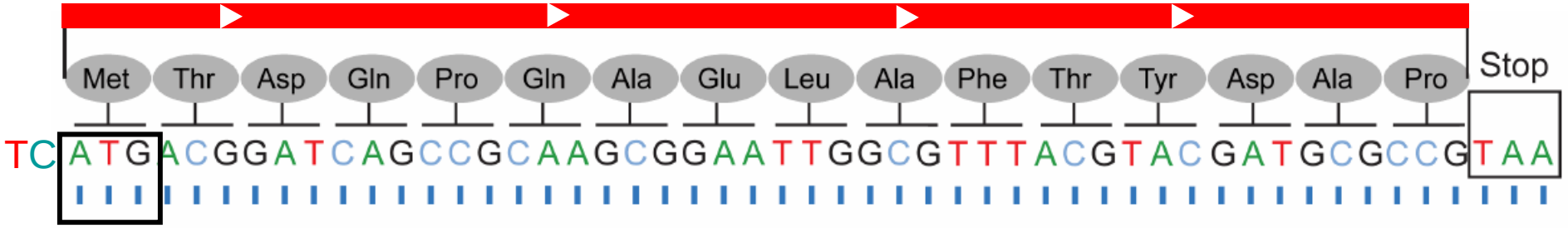
ORF contenant une CDS potentielle ??



« ORF finder » cherche les ORF (entre 2 STOP) et donne la séquence codante trouvée à l'intérieur et la signale en trait plein rouge.

1. Recherche d'ORF (Open Reading Frame)

Dès que « ORF finder » trouve une CDS dans l'ORF
il l'encadre en rouge et donne son sens de lecture
par des flèches !!

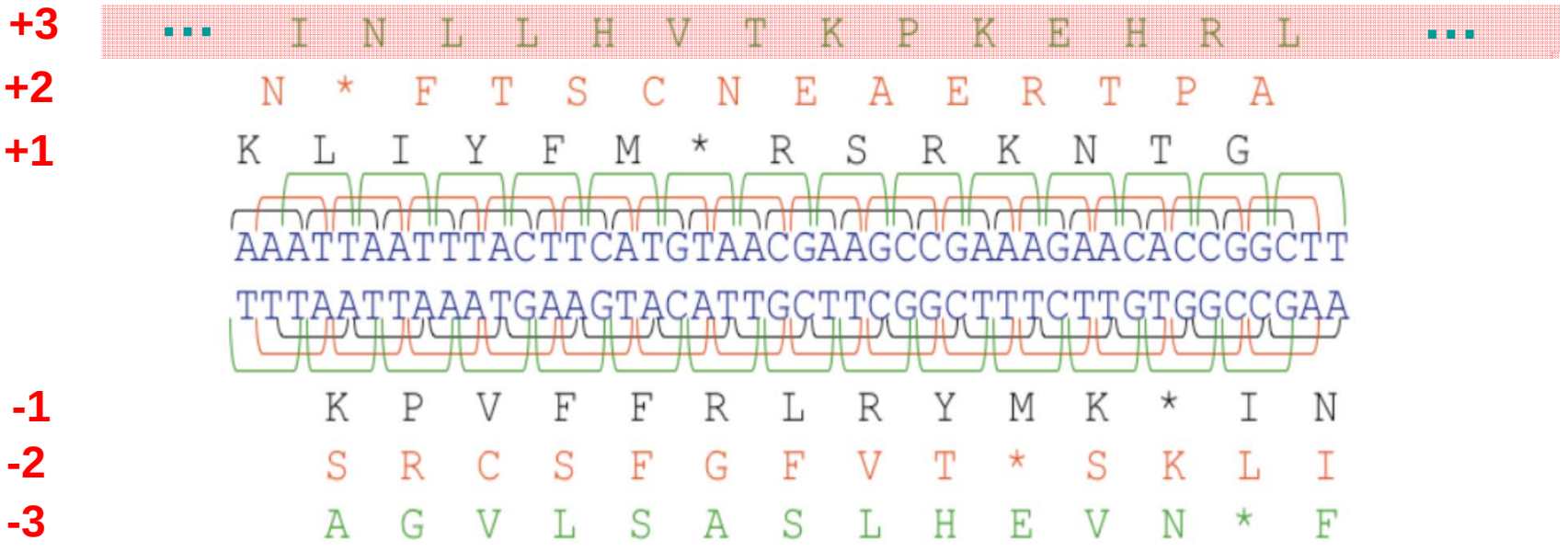


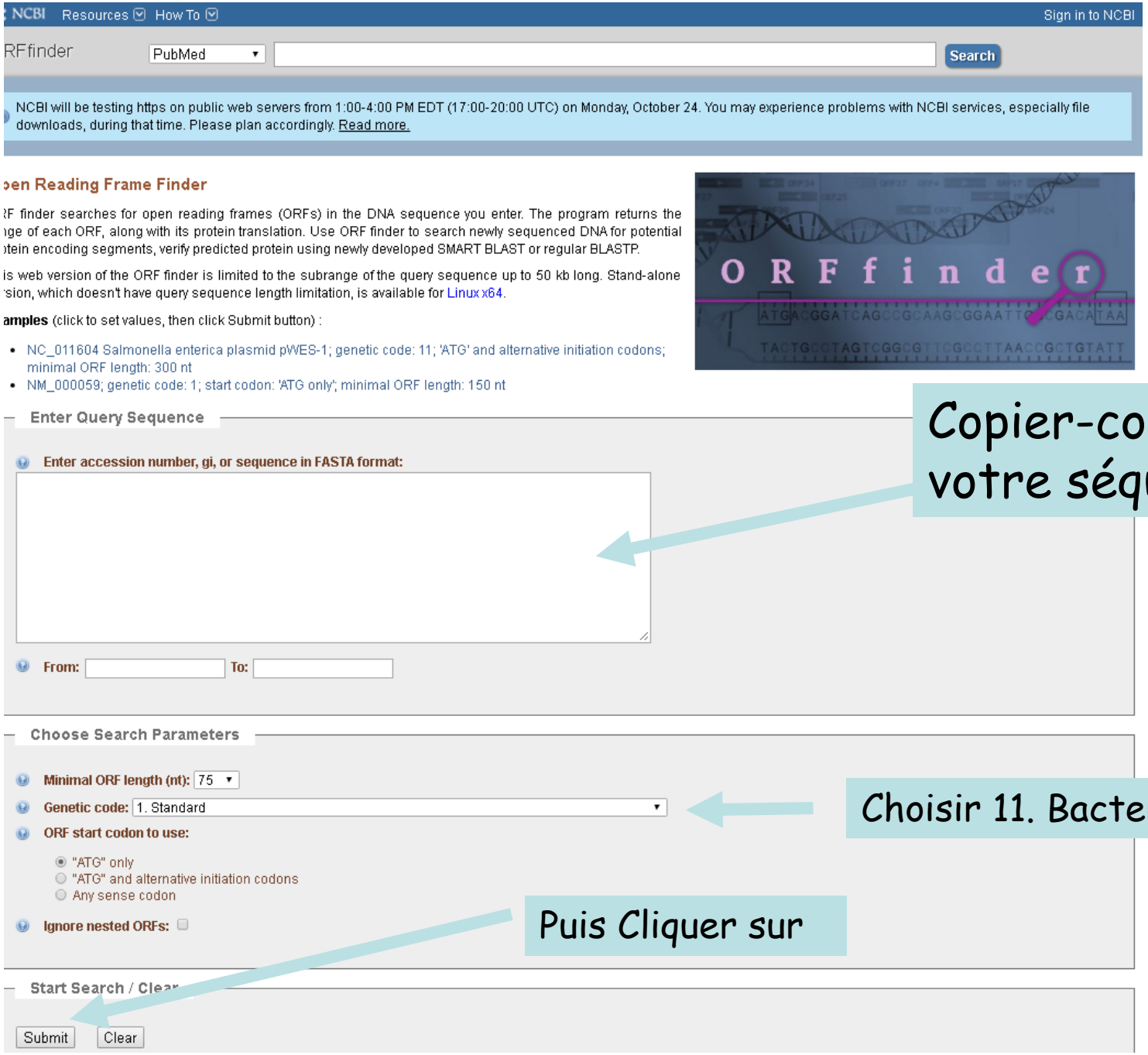
Codon d'initiation

Ce que fait ORF finder

- Traduction à l'aveugle

  6 phases de lecture = 6 séquences protéiques possibles




NCBI Resources How To
Sign in to NCBI
RFfinder
PubMed
Search
NCBI will be testing https on public web servers from 1:00-4:00 PM EDT (17:00-20:00 UTC) on Monday, October 24. You may experience problems with NCBI services, especially file downloads, during that time. Please plan accordingly. Read more.
pen Reading Frame Finder
F finder searches for open reading frames (ORFs) in the DNA sequence you enter. The program returns the ge of each ORF, along with its protein translation. Use ORF finder to search newly sequenced DNA for potential tein encoding segments, verify predicted protein using newly developed SMART BLAST or regular BLASTP.
is web version of the ORF finder is limited to the subrange of the query sequence up to 50 kb long. Stand-alone sion, which doesn't have query sequence length limitation, is available for Linux x64.
amples (click to set values, then click Submit button) :
NC_011604 Salmonella enterica plasmid pWES-1; genetic code: 11; 'ATG' and alternative initiation codons; minimal ORF length: 300 nt
NM_000059; genetic code: 1; start codon: 'ATG only'; minimal ORF length: 150 nt
ORFfinder
Enter Query Sequence
Enter accession number, gi, or sequence in FASTA format:
From:
To:
Choose Search Parameters
Minimal ORF length (nt): 75
Genetic code: 1. Standard
ORF start codon to use:
"ATG" only
"ATG" and alternative initiation codons
Any sense codon
Ignore nested ORFs:
Start Search / Clear
Submit
Clear
Copier-coller votre séquence
Choisir 11. Bacterial Code
Puis Cliquer sur

NCBI Resources How To Sign in to NCBI

ORFfinder PubMed Search

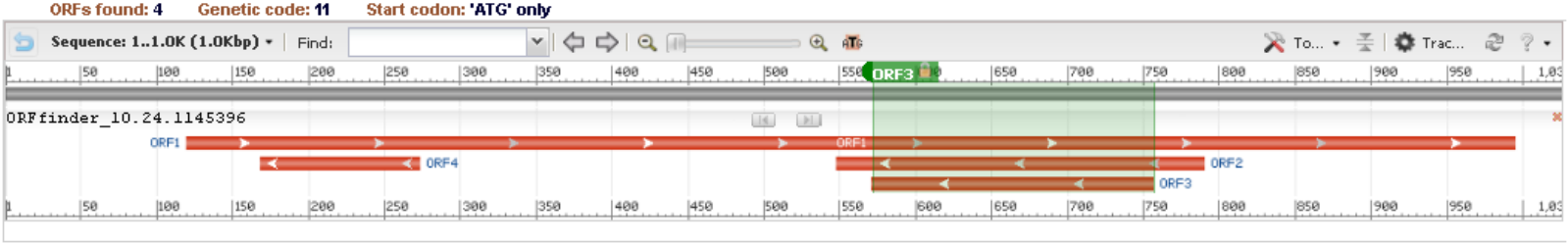
NCBI will be testing https on public web servers from 1:00-4:00 PM EDT (17:00-20:00 UTC) on Monday, October 24. You may experience problems with NCBI services, especially file downloads, during that time. Please plan accordingly. Read more.

Open Reading Frame Viewer

Sequence inconnue au format Fasta

Représentation schématique des CDS sur la séquence inconnue

ORFs found: 4 Genetic code: 11 Start codon: 'ATG' only



Traduction de la CDS sélectionnée

ORF3 (61 aa) Mark

```
>lcl|ORF3
MSPAAPVGCRPMLYPASDSAEFAPLPEVKW
YHADRPEWRCLTSARYGRDGTFRLPAVGRT
R
```

SmartBLAST ORF3

BLAST ORF3 BLAST marked set

BLAST Database: UniProtKB/Swiss-Prot (swissprot)

Add six-frame translation track

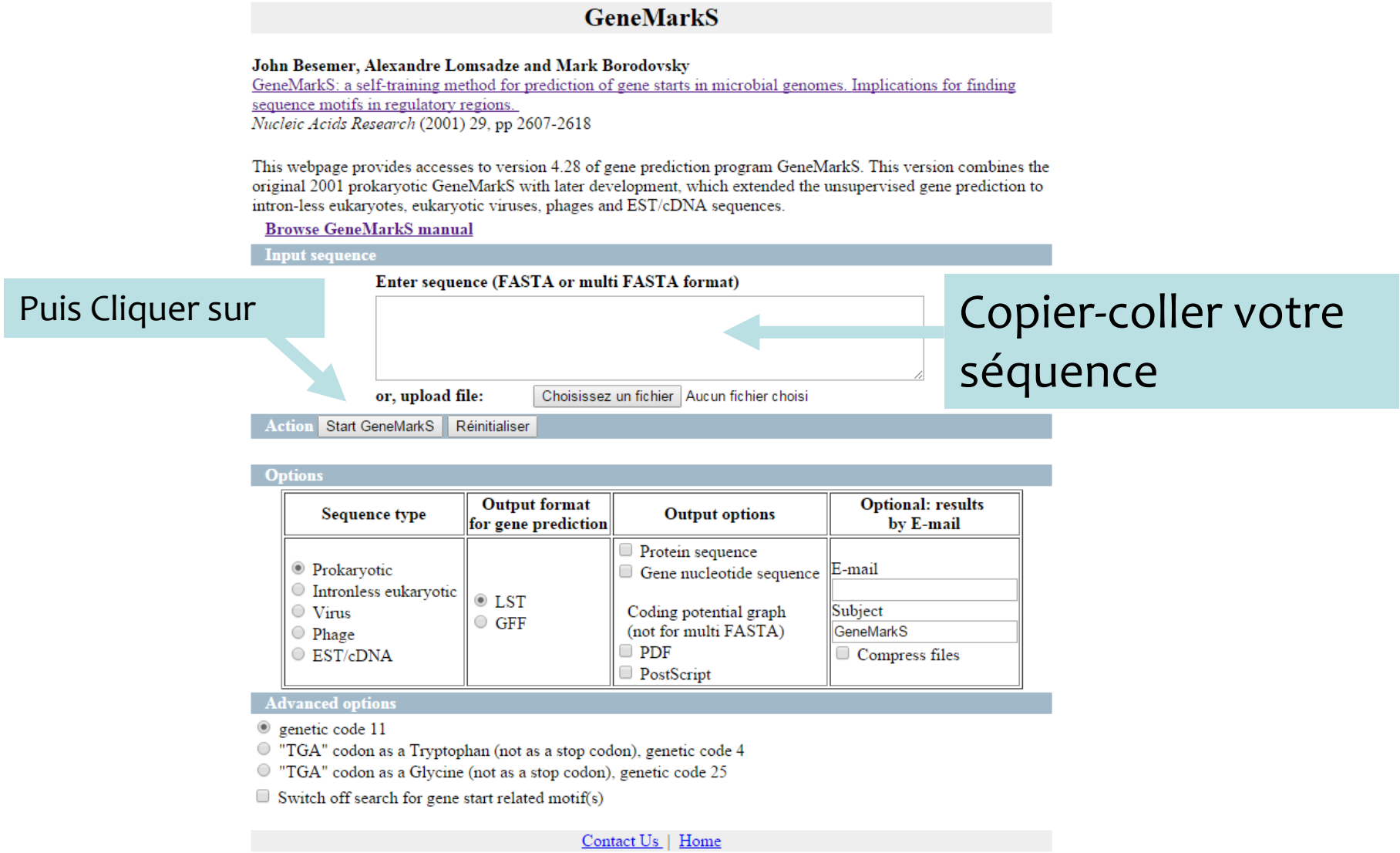
Mark subset Marked: 0 Download marked set as FASTA

Label	Strand	Frame	Start	Stop	Length (bp | aa)
ORF1	+	3	120	995	876 | 291
ORF2	-	2	790	548	243 | 80
ORF3	-	**3**	**756**	**571**	**186 | 61**
ORF4	-	3	273	169	105 | 34

+

-

2. Utilisation de GeneMark !



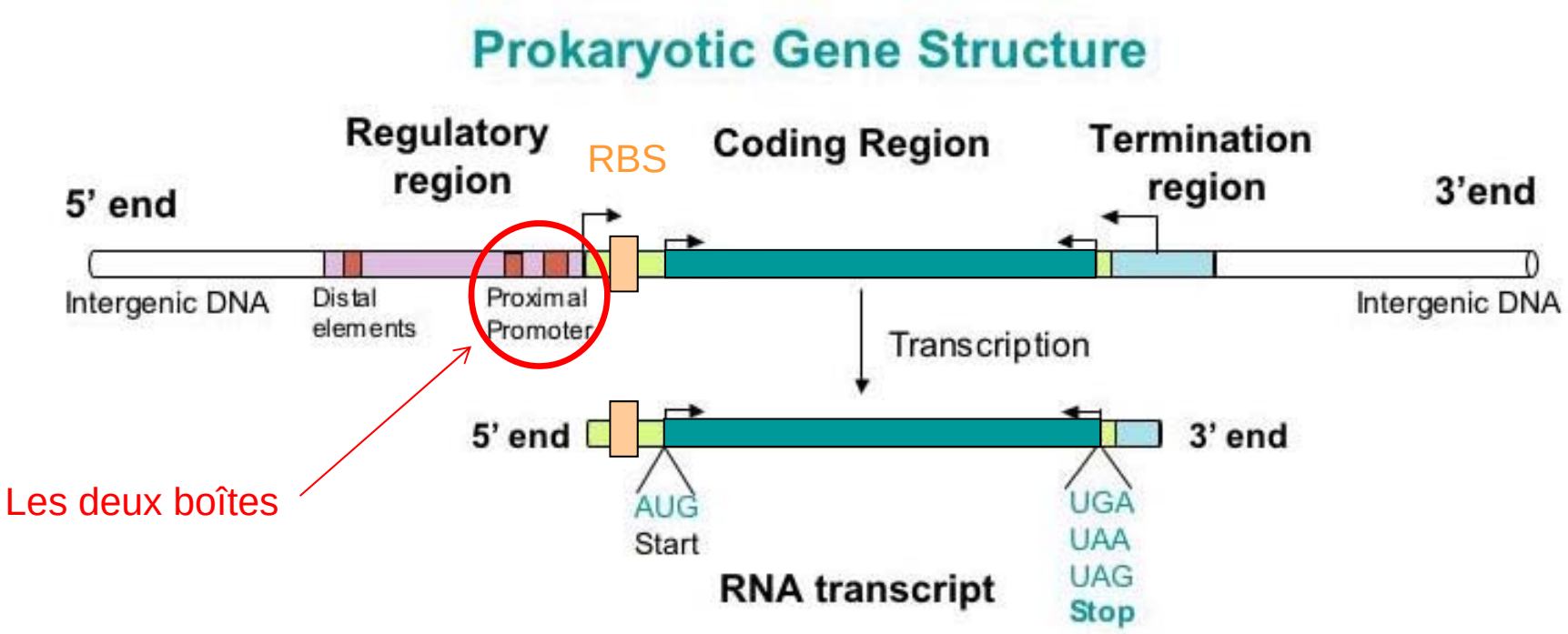
Puis Cliquer sur

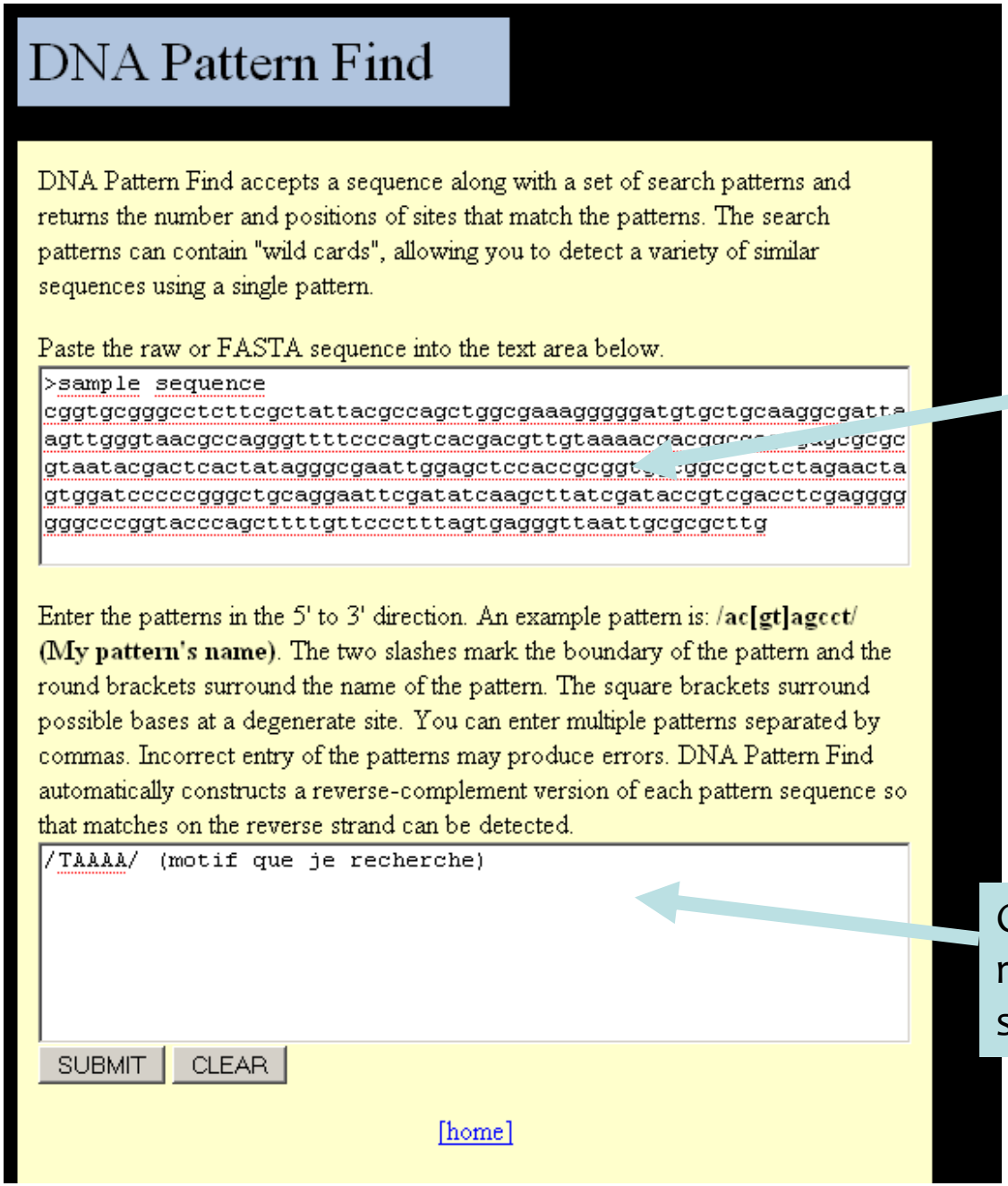
Copier-coller votre séquence

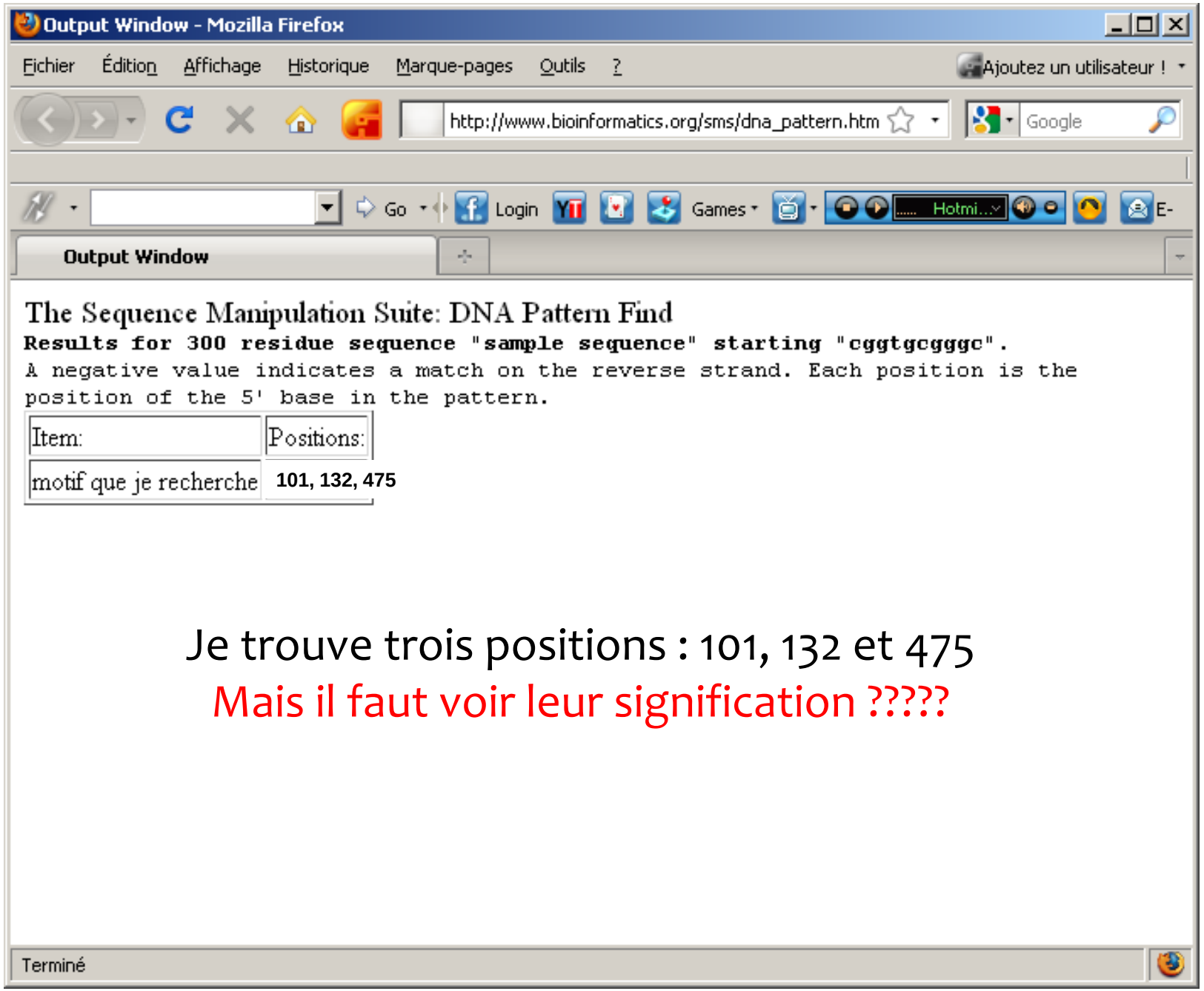
```
GeneMark.hmm PROKARYOTIC (Version 3.26)
Date: Wed Mar  4 04:10:15 2015
Sequence file name: seq.fna
Model file name: GeneMark_hmm_heuristic.mod
RBS: false
Model information: Heuristic_model_for_genetic_code_11_and_GC_52

FASTA definition line: text
Predicted genes
   Gene    Strand    LeftEnd    RightEnd       Gene     Class
    #                                        Length
    1        +          <1         555          555        1
    2        +       25301       25422         1092        1
    3        +       63545       63619         3075        1
```

3. Recherche de motifs : DNA Pattern

Utiisé pour rechercher la région promotrice : boîte -35 et boîte -10, rechercher le RBS (séquence consensus) ou même encore les signaux de régulation




DNA Pattern Find
DNA Pattern Find accepts a sequence along with a set of search patterns and returns the number and positions of sites that match the patterns. The search patterns can contain "wild cards", allowing you to detect a variety of similar sequences using a single pattern.
Paste the raw or FASTA sequence into the text area below.
>sample sequence
cggtgcgggcctcttcgctattacgccagctggcgaaaggggatgtgctgcaaggcgatta
agttgggtaacgccagggttttcccagtcacgacgttgtaaaacgacggccagtgccgcgc
gtaatacgactcactatagggcgaattggagctccaccgcggtggcggccgctctagaacta
gtggatcccccgggctgcaggaattcgatatcaagcttatcgataccgtcgacctcgagggg
gggcccggtacccagcttttgttccctttagtgagggttaattgcgcgcttg
Enter the patterns in the 5' to 3' direction. An example pattern is: /ac[gt]agcct/ (My pattern's name). The two slashes mark the boundary of the pattern and the round brackets surround the name of the pattern. The square brackets surround possible bases at a degenerate site. You can enter multiple patterns separated by commas. Incorrect entry of the patterns may produce errors. DNA Pattern Find automatically constructs a reverse-complement version of each pattern sequence so that matches on the reverse strand can be detected.
/TAAAA/ (motif que je recherche)
SUBMIT
CLEAR
[home]
Copier-coller la séquence
Copier/coller la séquence à rechercher (attention à la syntaxe)



Je trouve trois positions : 101, 132 et 475

Mais il faut voir leur signification ?????

Vous annoterez au fur et à mesure sur votre séquence les signaux retrouvés :

- Positions des CDS, du A de ATG, du T de TAA,TGA,TAG
- Positions des boites -35 et -10
- Positions des RBS
- …
- Noter les ambiguïtés !! (OFR Finder et GeneMark)

Exemple de séquence annotée …

Position 839 = Boite -10

```
cgcggtgatctggccaaaagatcgtgcgccgctgattctggtcacttacttcacccagcctcaacctaaggcaga
aagccgtcgcgatgtattagcgtcggcggctaaaatcgtcacccgcggtgatctggccaaaagatcgtgcgccgc
tgattctggtcacttacttcacccagcctcaacctaaggcagaaagccgtcgcgatgtattagcgtcggcggcta
aaatcgtcacccgcggtgatctggccaaaagatcgtgcgccgctgattctggtcacttacttcacccagcctcaa
cctaaggcagaaagccgtcgcgatgtattagcgtcggcggctaaaatcgtcacccgcggtgatctggccaaaaga
tcgtgcgccgctgattctggtcacttacttcacccagcctcaacctaaggcagaaagccgtcgcgatgtattagc
gtcggcggctaaaatcgtcacccacgctgatggcgacggcaaccgtcacgctgttgttaggaagtgttgacagcc
gctgtatgcgcaaacggcggtataatcttgccgtgcgtgctacgtatcgctaaaggaattagagcggcatgtcgg
gaggcagactgggtgtggcattgattaacacagcagataattcgcaaatactttatcgtgctgatgagcgctttg
cgatgtgcagcaccagtaaagtgatggccgcggccgcggtgctgaagaaaagtgaaagcgaaccgaatctgttaa
atcagcgagttgagatcaaaaaatctgaccttgttaactataatccgattgcggaaaagcacgtcaatgggacga
tgtcactggctgagcttagcgcggccgcgctacagtacagcgataacgtggcgatgaataagctgattgctcacg
ttggcggcccggctagcgtcaccgcgttcgcccgacagctgggagacgaaacgttccgtctcgaccgttggcgac
gtccgtatttgcctttcggaagcataaaatcggacgcgttgtggctcgcttcaggtaaaatattgactattcatg
ttgttgttatttcgtctcttccagaataaggaatcccatggttaaaaaatcactgcgccagttaccgagccgacg
ttaaacaccgccattccgggcgatccgcgtgataccacttcacctcgggcaatggcgcaaactctgcggaatctg
acgctgggtaaagcattgggcgacagccaacgggcgcagctggtgacatggatgaaaggcaataccaccggtgca
gcgagcattcaggctggactgcctgcttcctgggttgtgggggataaaaccggcagcggtaactatggcaccacc
aacgatatcgcggtgatctggccaaaagatcgtgcgccgctgattctggtcacttacttcacccagcctcaacct
aaggcagaaagccgtcgcgatgtattagcgtcggcggctaaaatcgtcaccgacggtttgtaatagcggaaacgg
aatggggaaactcattccgttttt
```

Position 810 = Boite -35

Position 871 = RBS

CDS : position début (887=ATG) et fin (741=TAA)