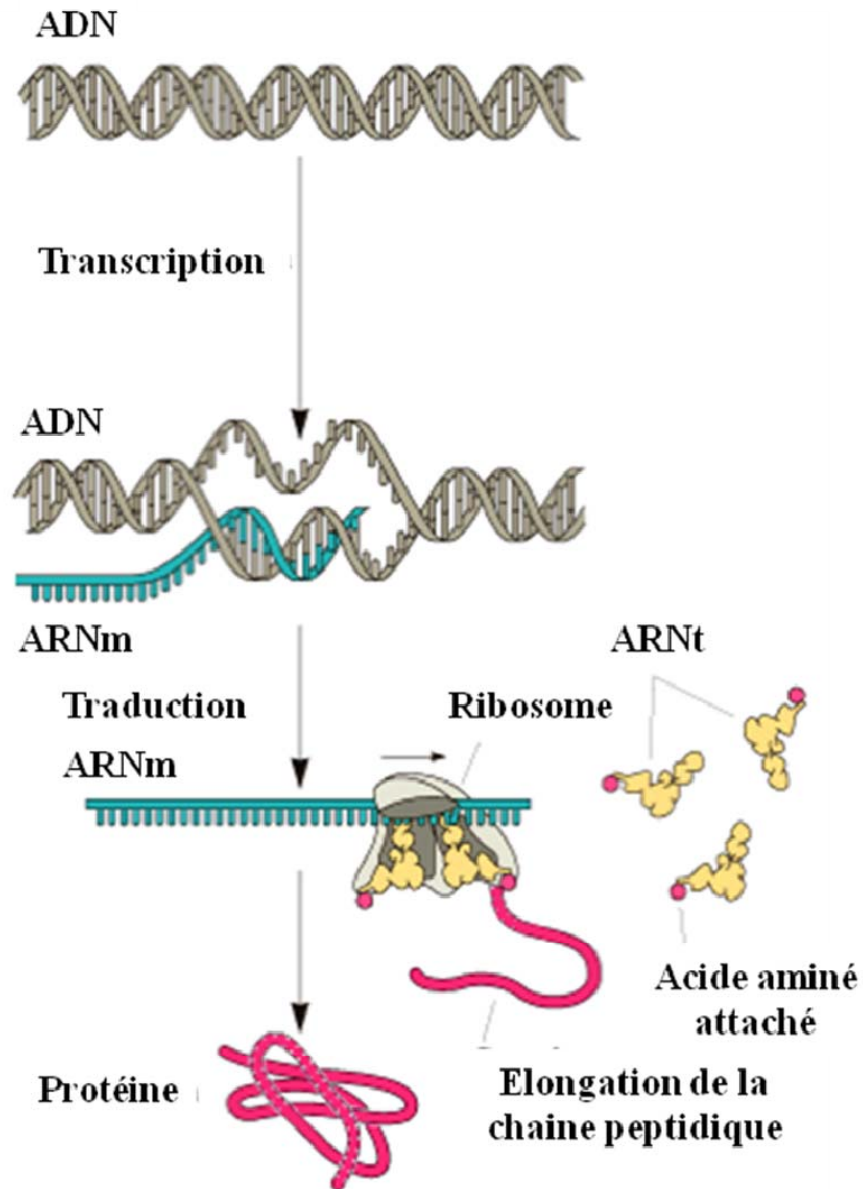


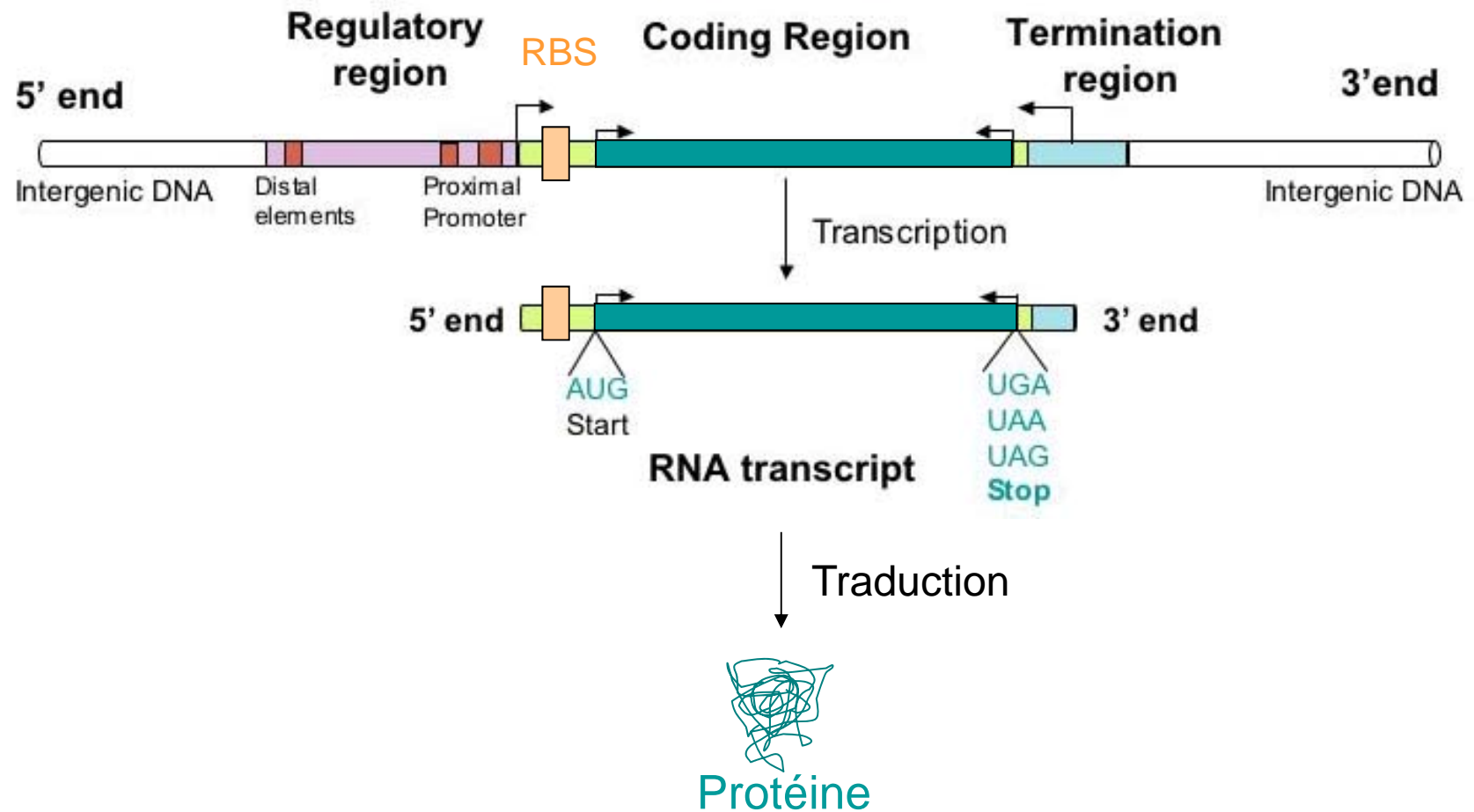
Annotation d'une séquence bactérienne !

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

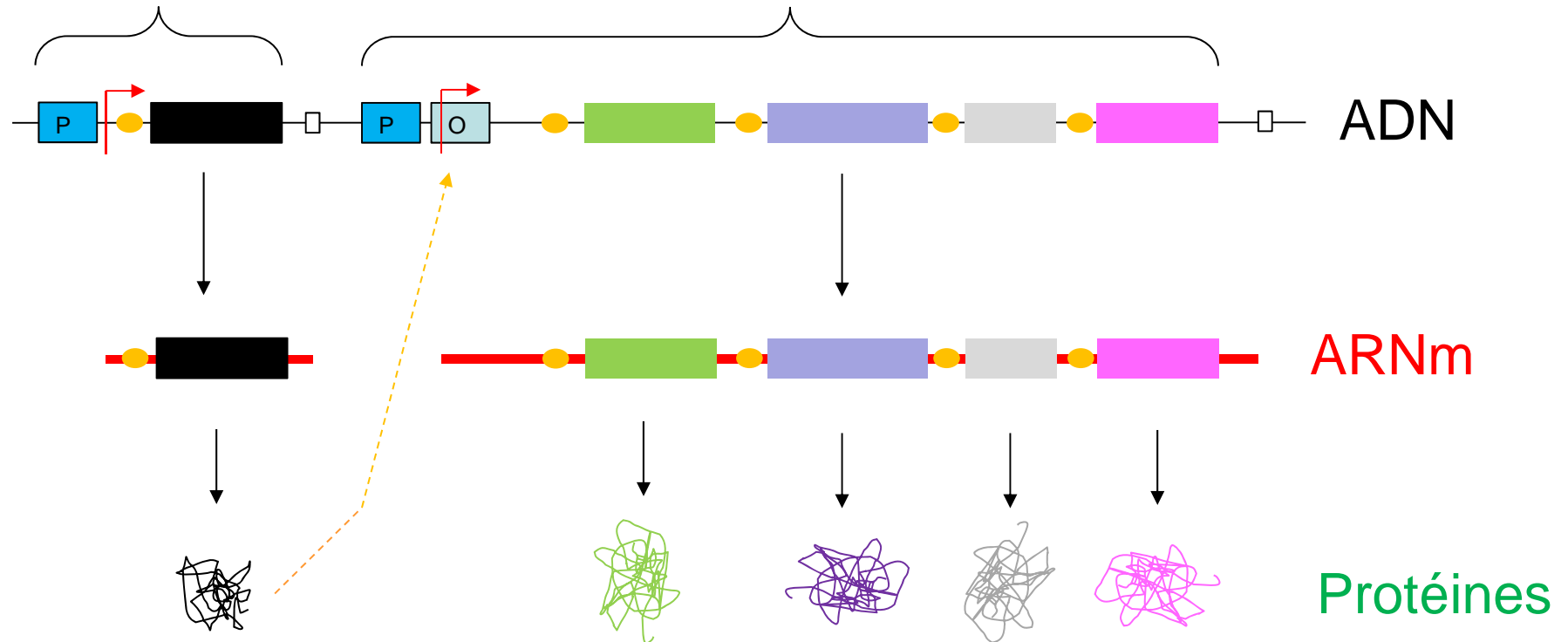


Oui mais au final c'est très souvent sous forme d'opéron ...

Structure d'un opéron

Gène régulateur de l'opéron

Opéron constitué ici de 4 CDS (4 protéines) on dit polycistronique !



● RBS

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

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



Début de transcription (point +1 de transcription)



Termineur de la transcription

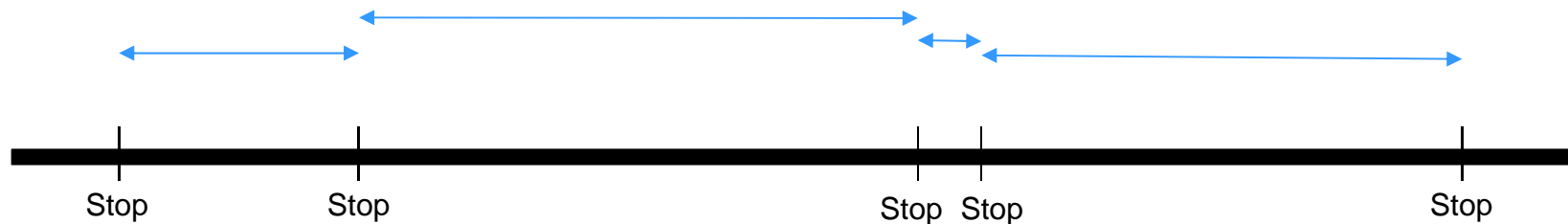
But : annoter votre séquence

- 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)

Définition : une ORF commence après un codon stop, se termine avec un codon stop et contient une séquence codante (CDS) potentielle ...

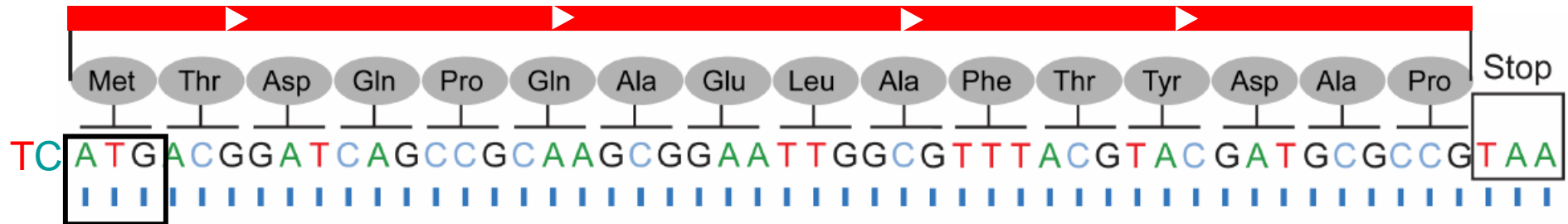
ORF contenant une CDS potentielle ??



« 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 !!

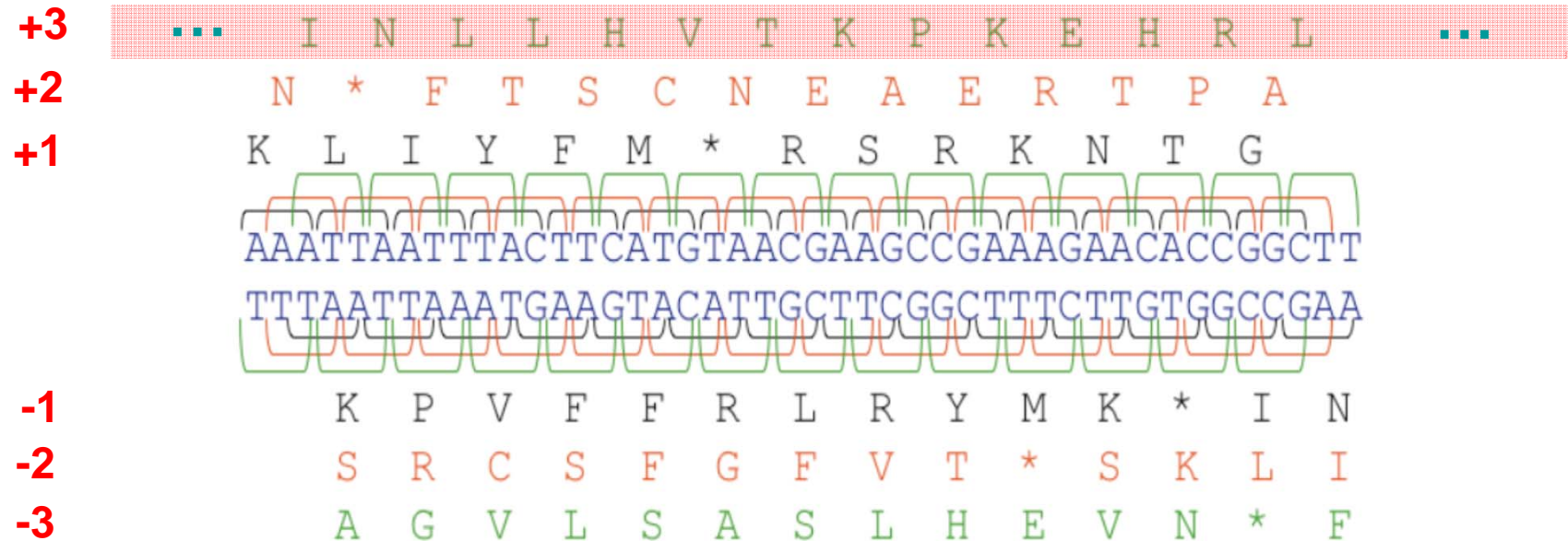


Codon d'initiation

Ce que fait ORF finder

- Traduction à l'aveugle

6 phases de lecture = 6 séquences protéiques possibles



Open Reading Frame Finder

ORF finder searches for open reading frames (ORFs) in the DNA sequence you enter. The program returns the location of each ORF, along with its protein translation. Use ORF finder to search newly sequenced DNA for potential protein encoding segments, verify predicted protein using newly developed SMART BLAST or regular BLASTP.

The web version of the ORF finder is limited to the subrange of the query sequence up to 50 kb long. Stand-alone version, which doesn't have query sequence length limitation, is available for [Linux x64](#).

Examples (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



Enter Query Sequence

Enter accession number, gi, or sequence in FASTA format:

From: To:

Copier-coller
votre séquence

Choose Search Parameters

Minimal ORF length (nt):

Genetic code:

ORF start codon to use:

- ☒ "ATG" only
☐ "ATG" and alternative initiation codons
☐ Any sense codon

Ignore nested ORFs: ☐

Choisir 11. Bacterial Code

Puis Cliquer sur

Start Search / Clear

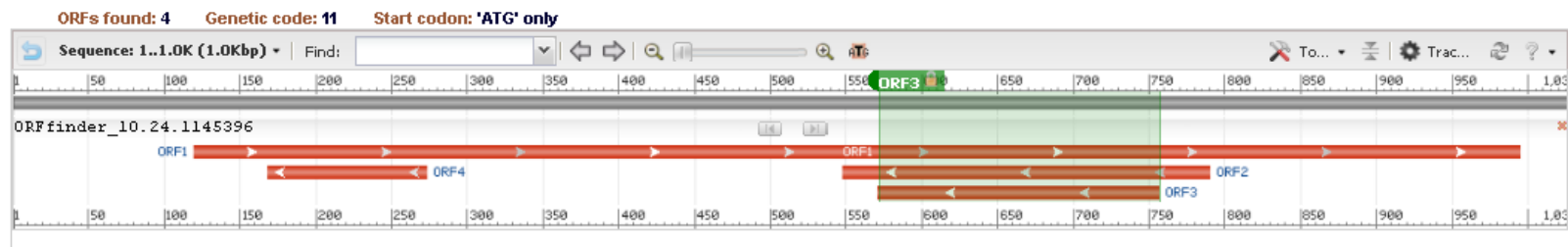
Submit

Clear

Open Reading Frame Viewer

Sequence inconnue au format Fasta

Représentation schématique des CDS sur la séquence inconnue



Traduction de la CDS sélectionnée

ORF3 (61 aa)
Mark

```
>1c1|ORF3
MSPAAPVGC R PMLYPASDSA EFA PLPEVKW
YHAD R PEWRCLTSARYGRDGT FRLPAVGRT
R
```

SmartBLAST ORF3
BLAST ORF3
BLAST marked set

BLAST Database:
UniProtKB/Swiss-Prot (swissprot)

Mark subset
Marked: 0
Download marked set as FASTA
Add six-frame translation track

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 !

GeneMarkS

John Besemer, Alexandre Lomsadze and Mark Borodovsky

[GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions.](#)

Nucleic Acids Research (2001) 29, pp 2607-2618

This webpage provides accesses to version 4.28 of gene prediction program GeneMarkS. This version combines the original 2001 prokaryotic GeneMarkS with later development, which extended the unsupervised gene prediction to intron-less eukaryotes, eukaryotic viruses, phages and EST/cDNA sequences.

[Browse GeneMarkS manual](#)

Input sequence

Enter sequence (FASTA or multi FASTA format)



or, upload file:

Choisissez un fichier

Aucun fichier choisi

Action

Start GeneMarkS

Réinitialiser

Options

Sequence type	Output format for gene prediction	Output options	Optional: results by E-mail
<input checked="" type="radio"/> Prokaryotic <input type="radio"/> Intronless eukaryotic <input type="radio"/> Virus <input type="radio"/> Phage <input type="radio"/> EST/cDNA	<input checked="" type="radio"/> LST <input type="radio"/> GFF	<input type="checkbox"/> Protein sequence <input type="checkbox"/> Gene nucleotide sequence Coding potential graph (not for multi FASTA) <input type="checkbox"/> PDF <input type="checkbox"/> PostScript	E-mail <input type="text"/> Subject GeneMarkS <input type="checkbox"/> Compress files

Advanced options

- ☒ genetic code 11
- ☐ "TGA" codon as a Tryptophan (not as a stop codon), genetic code 4
- ☐ "TGA" codon as a Glycine (not as a stop codon), genetic code 25
- ☐ Switch off search for gene start related motif(s)

[Contact Us](#) | [Home](#)

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

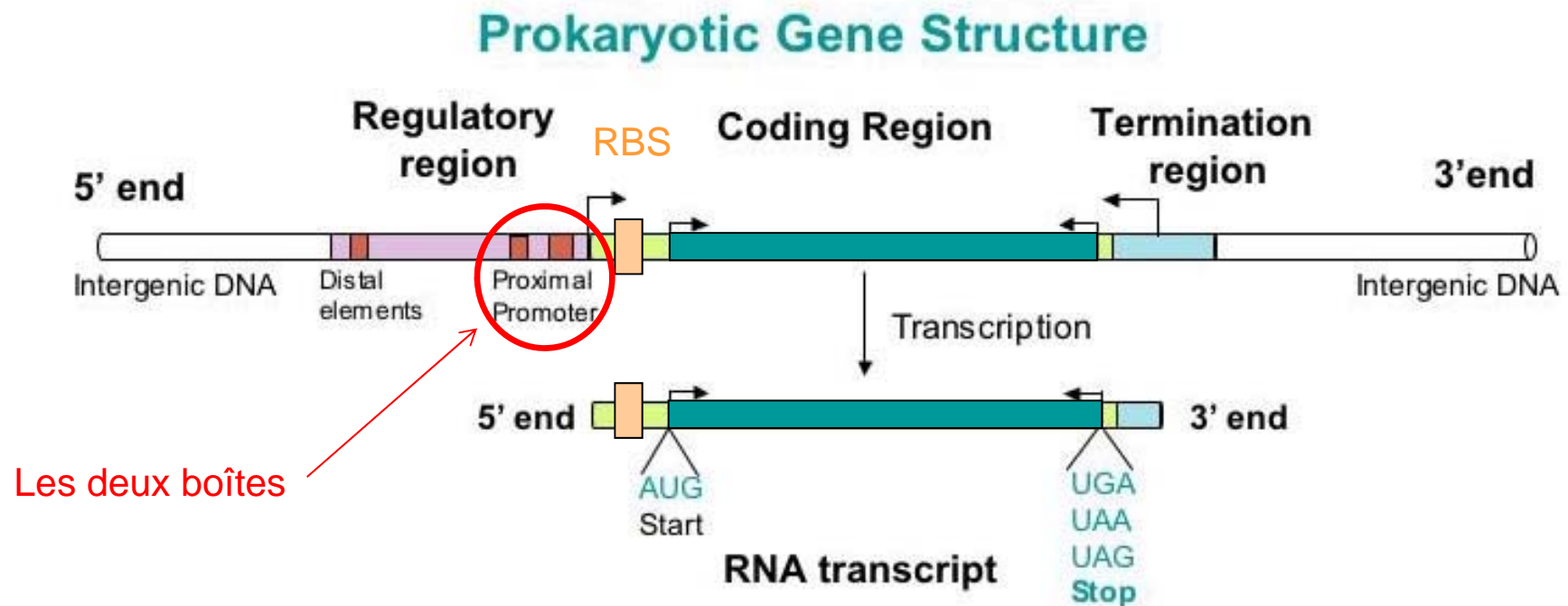
FASTA definition line: text

Predicted genes

Gene #	Strand	LeftEnd	RightEnd	Gene Length	Class
1	+	<1	555	555	1
2	+	25301	25422	1092	1
3	+	63545	63619	3075	1

3. Recherche de motifs : DNA Pattern

Utilisé 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
cgggtgccccctcttctgctattacgccagctggcgaaagggggatgtgctgcaaggcgatta
agttgggtaacgccagggttttccagtcacgacgttgtaaaacacacacacacacacacacac
gtaatacgactcactatagggcggaattggagctccaccggttgggcegtctctagaacta
gtggatccccccgggctgcaggaattcgatatcaagcttatcgataccgtcgacctcgagggg
gggccccgtaccagcttttgttcccttagtgagggttaattgcccgttg
```

Copier-coller la séquence

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)
```

Copier/coller la séquence à rechercher (attention à la syntaxe)

SUBMIT

CLEAR

[\[home\]](#)

Output Window - Mozilla Firefox

Fichier Édition Affichage Historique Marque-pages Outils ?

Ajoutez un utilisateur !

http://www.bioinformatics.org/sms/dna_pattern.htm

Google

Output Window

The Sequence Manipulation Suite: DNA Pattern Find

Results for 300 residue sequence "sample sequence" starting "cggcgcgggc".
A negative value indicates a match on the reverse strand. Each position is the position of the 5' base in the pattern.

Item:	Positions:
motif que je recherche	101, 132, 475

Terminé

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 boîtes -35 et -10
- Positions des RBS
- ...
- Noter les ambiguïtés !! (OFR Finder et GeneMark)

Exemple de séquence annotée ...

cgcggtgatctggccaaaagatcgtgcgccgctgattctgggtcacttacttcacccagcctcaacctaaggcaga
aagccgtcgcgatgtattagcgtcggcgggctaaaatcgtcaccgcggtgatctggccaaaagatcgtgcgccgc
tgattctgggtcacttacttcacccagcctcaacctaaggcagaaagccgtcgcgatgtattagcgtcggcgggta
aaatcgtcaccgcggtgatctggccaaaagatcgtgcgccgctgattctgggtcacttacttcacccagcctcaa
cctaaggcagaaagccgtcgcgatgtattagcgtcggcgggctaaaatcgtcaccgcggtgatctggccaaaaga
tcgtgcgccgctgattctgggtcacttacttcacccagcctcaacctaaggcagaaagccgtcgcgatgtattagc
gtcggcgggctaaaatcgtcaccgcggtgatggcgacggcaaccgtcacgctgttgtaggaagtgttgacaagcc
gctgtatgcgcaaacggcgggtataatcttgccgtgctgtacgtatcgctaaggaaattagagcggcatgtcgg
gaggcagactgggtgtggcattgattaacacagcagataaattcgcaataactttatcgtgctgatgagcgtttg
cgatgtgcagcaccagtaaaagtgtggcgcggccgcggtgctgaagaaaagtgaagcgaaaccgaatctgttaa
atcagcgagttgagatcaaaaaatctgaccttgtaactataatccgattgcggaaaagcacgtcaatgggacga
tgtcactggctgagcttagcgcggccgcgctacagtacagcgataacgtggcgatgaataagctgattgctcacg
ttggcggcccggttagcgtcaccgcgttcgcccagacagctgggagacgaaacgttccgtctcgaccgttggcgac
gtccgtatattgcctttcggaagcataaaaatcggacgcgttggtggctcgcttcaggtaaaaatattgactattcatg
ttgttgttatttcgtctcttccagaataaggaatcccatgggttaaaaaatcactgcgccagttaccgagccgacg
ttaaacaccgcatctcgggcgatccgcgtgataccacttcacctcgggcaatggcgcaaactctgcggaatctg
acgctgggttaaagcattgggcgacagccaacgggcgcagctgggtgacatggatgaaaggcaataaccacgggtgca
gcgagcattcaggctggactgcctgcttctgggttggtgggggataaaaccggcagcggtaactatggcaccacc
aacgatatcgcggtgatctggccaaaagatcgtgcgccgctgattctgggtcacttacttcacccagcctcaacct
aaggcagaaagccgtcgcgatgtattagcgtcggcgggctaaaatcgtcaccgcaggtttgtaatagcggaaacgg
aatggggaaactcattccgttttt

Position 839 = Boite -10

Position 810 = Boite -35

Position 871 = RBS

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