Support de cours
Annotation des génomes

Annotation d'un génome

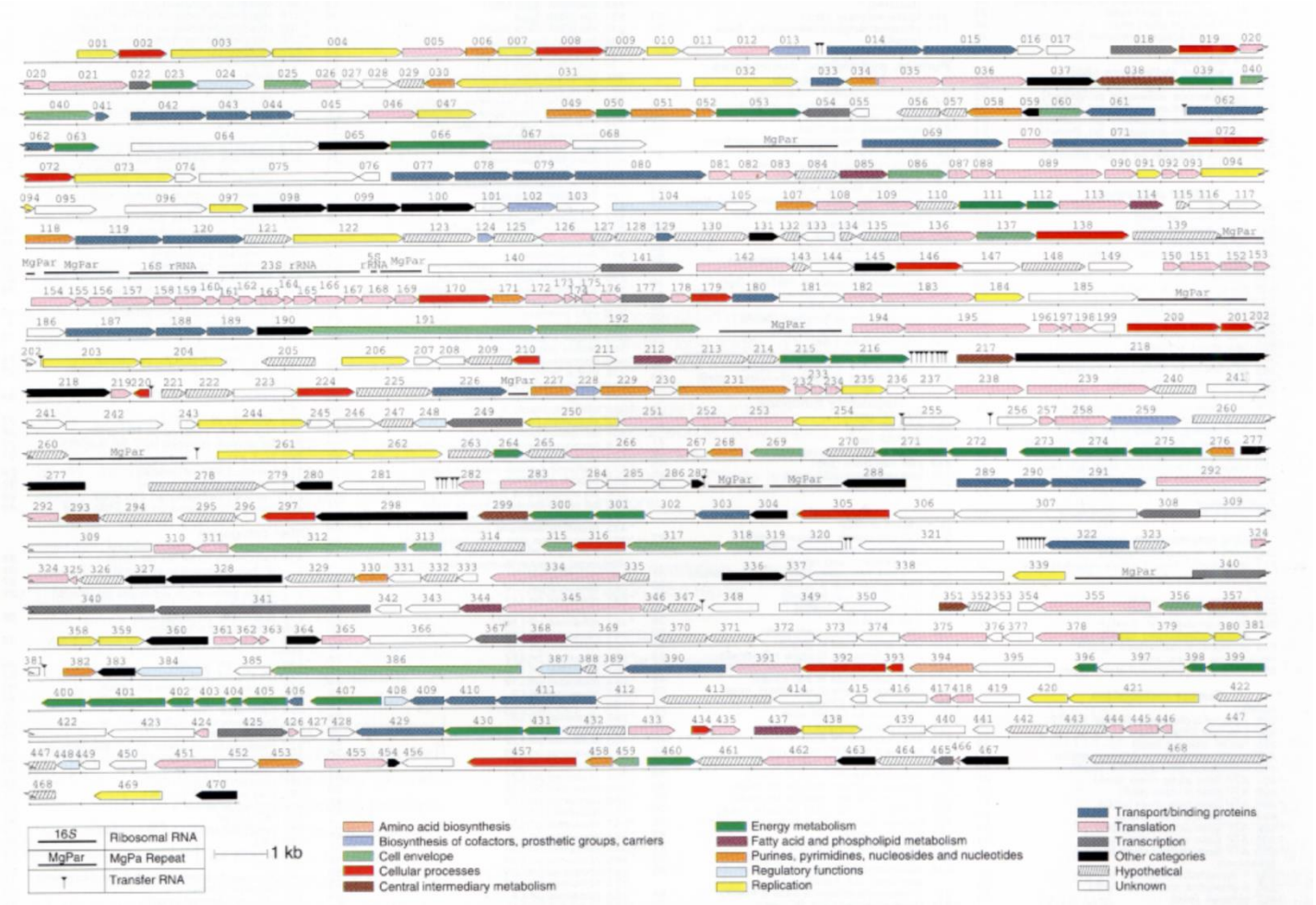
Identification des gènes codant pour :

- les ARNr
- les ARNt
- les petits ARN non codants
- les protéines

Identification des unités de transcription (promoteur et terminateur)

Pour les gènes codant pour les protéines, prédiction fonctionnelle par recherche de similarité de séquences (Blast) et de domaines fonctionnels (Pfam, InterPro etc.)

Exemple d'annotation d'un génome
Mycoplasma genitalium



UNIVERSITÉ TOULOUSE III PAUL SABATIER

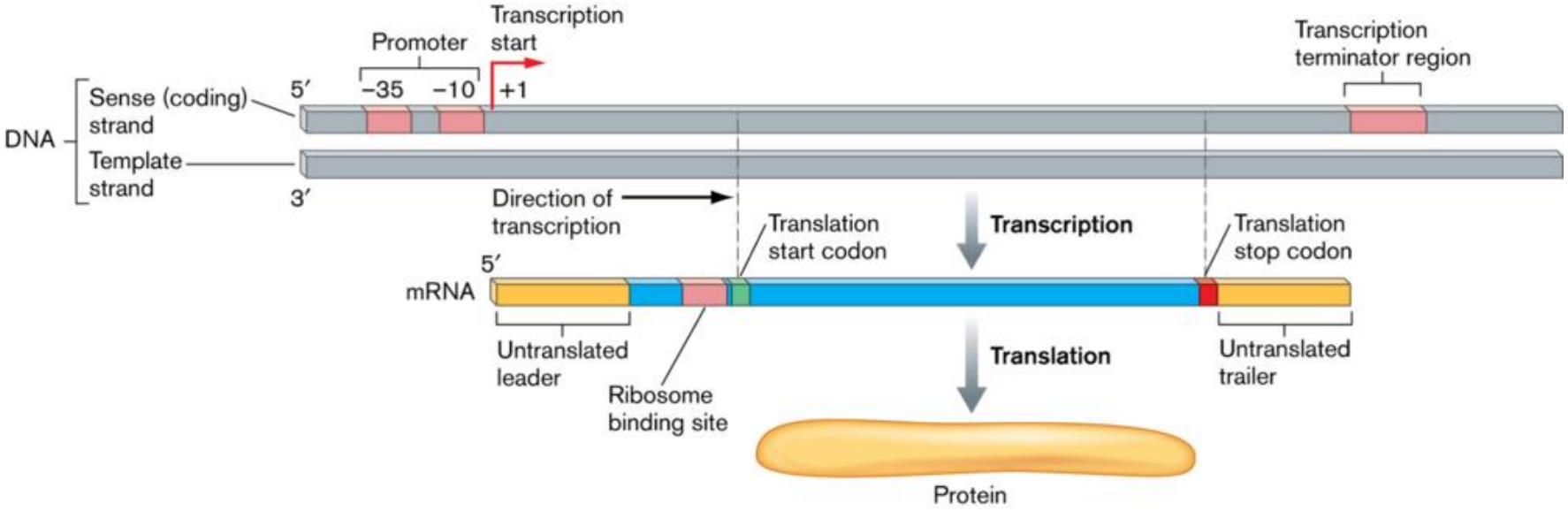
Annotation GenBank

```
LOCUS       L43967                580076 bp    DNA     circular BCT 31-JAN-2014
DEFINITION  Mycoplasma genitalium G37, complete genome.
ACCESSION   L43967 U39679-U39729
VERSION     L43967.2
SOURCE      Mycoplasmoides genitalium G37
  ORGANISM  Mycoplasmoides genitalium G37
            Bacteria; Mycoplasmatota; Mycoplasmoidales; Metamycoplasmataceae;
            Mycoplasmoides.
REFERENCE   1  (bases 1 to 580076)
  AUTHORS   Fraser,C.M., & al.
  TITLE     The minimal gene complement of Mycoplasma genitalium
  JOURNAL   Science 270 (5235), 397-403 (1995)
   PUBMED   7569993
FEATURES             Location/Qualifiers
     source          1..580076
                     /organism="Mycoplasmoides genitalium G37"
                     /mol_type="genomic DNA"
                     /strain="G-37"
                     /type_material="type strain of Mycoplasma genitalium"
                     /db_xref="ATCC:33530"
                     /db_xref="taxon:243273"
     gene            174690..174793
                     /gene="rrfA"
                     /locus_tag="MG_rrnA5S"
     rRNA            174690..174793
                     /gene="rrfA"
                     /locus_tag="MG_rrnA5S"
                     /product="5S ribosomal RNA"
     gene            175805..179146
                     /locus_tag="MG_140"
     CDS             175805..179146
                     /locus_tag="MG_140"
                     /note="identified by similarity to EGAD:57280"
                     /codon_start=1
                     /transl_table=4
                     /product="conserved hypothetical protein"
                     /protein_id="AAC71358.1"
                     /translation="MNDWQWLKNRLVNSKTKSVSFWLPQTSSNIIDIAELIKCCSELK
                     ...........KLNPIGVISKIRSSLAVHQNQHEEI"
     gene            179151..180746
                     /gene="nusA"
                     /locus_tag="MG_141"
     CDS             179151..180746
                     /gene="nusA"
```

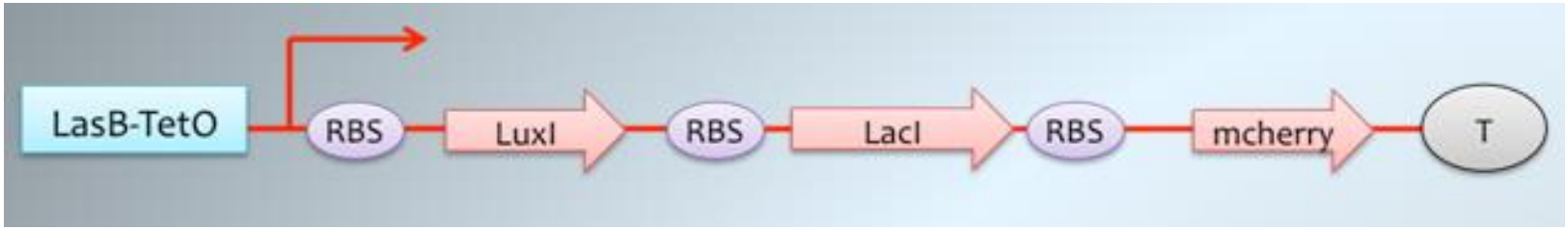
Recherche des régions codant pour des protéines chez les procaryotes

- recherche des ORFs (Open reading frame)
- recherche des CDS . Même si les gènes sont co-transcrits, ils sont en général traduits de façon indépendante (recherche des Shine Dalgarno en 5' du codon initiateur). Permet d'identifier le « bon » codon initiateur.
- recherche des unités de transcription. Chez les procaryotes, certains gènes sont co-transcrits donc recherche de la structure en opérons (promoteurs et terminateurs de transcription)

Structure d'un gène bactérien



Les gènes peuvent être co-transcrits. Ils sont organisés en unité de transcription, appelée opéron entre un promoteur et un terminateur de transcription.



La traduction : les ribosomes

Le ribosome : une usine de synthèse des protéines.

Il contient 3 sites de fixation pour les ARNt :

- le site A lie les ARNt chargés de leur acide aminé (ARNt aminoacylés)
- le site P lie les ARNt liés à la chaîne peptidique en cours de synthèse (peptidyl ARNt)
- le site E lie les ARNt libres (déchargé et décroché la chaîne peptidique) avant leur sortie du ribosome (E pour « exit »).

La traduction comprend également 3 étapes :

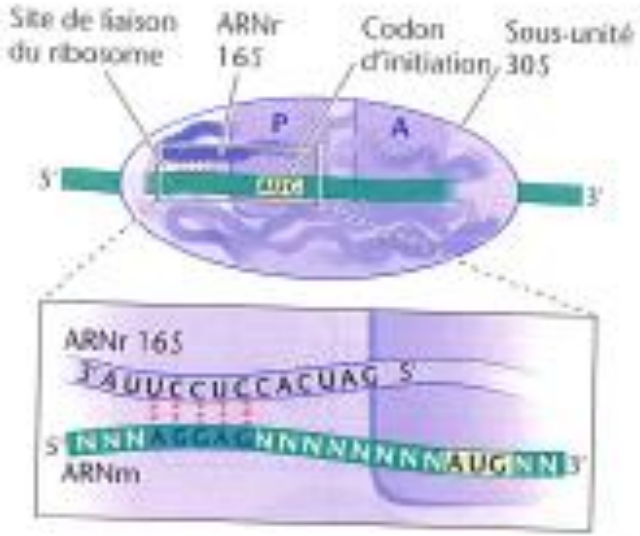
- l'initiation qui charge le ribosome sur l'ARNm
- l'élongation qui conduit à la synthèse de la protéine
- la terminaison conduisant à la dissociation du ribosome de l'ARNm

La traduction

L'initiation :

➢ chez les bactéries :

La petite sous-unité est chargée en premier sur l'ARNm, ceci grâce à un appariement des bases d'une région de l'ARNr 16S avec le RBS. Pour un RBS idéalement situé, le codon initiateur (AUG, GUG ou UUG) se trouve situé au site P du ribosome (et non au A comme pour l'élongation). Ceci requiert un ARNt particulier, appelé ARNt initiateur. Cet ARNt ne porte ni la méthionine, ni la valine, ni la leucine comme acide aminé, mais une méthionine modifiée (N-formylméthionine) d'où son nom ARNt fMet.
Quand l'ARNt fMet s'apparient au site P, il y a un changement de conformation de la petite sous-unité qui fait que la grande sous-unité peut se lier à elle pour former le ribosome 70S.

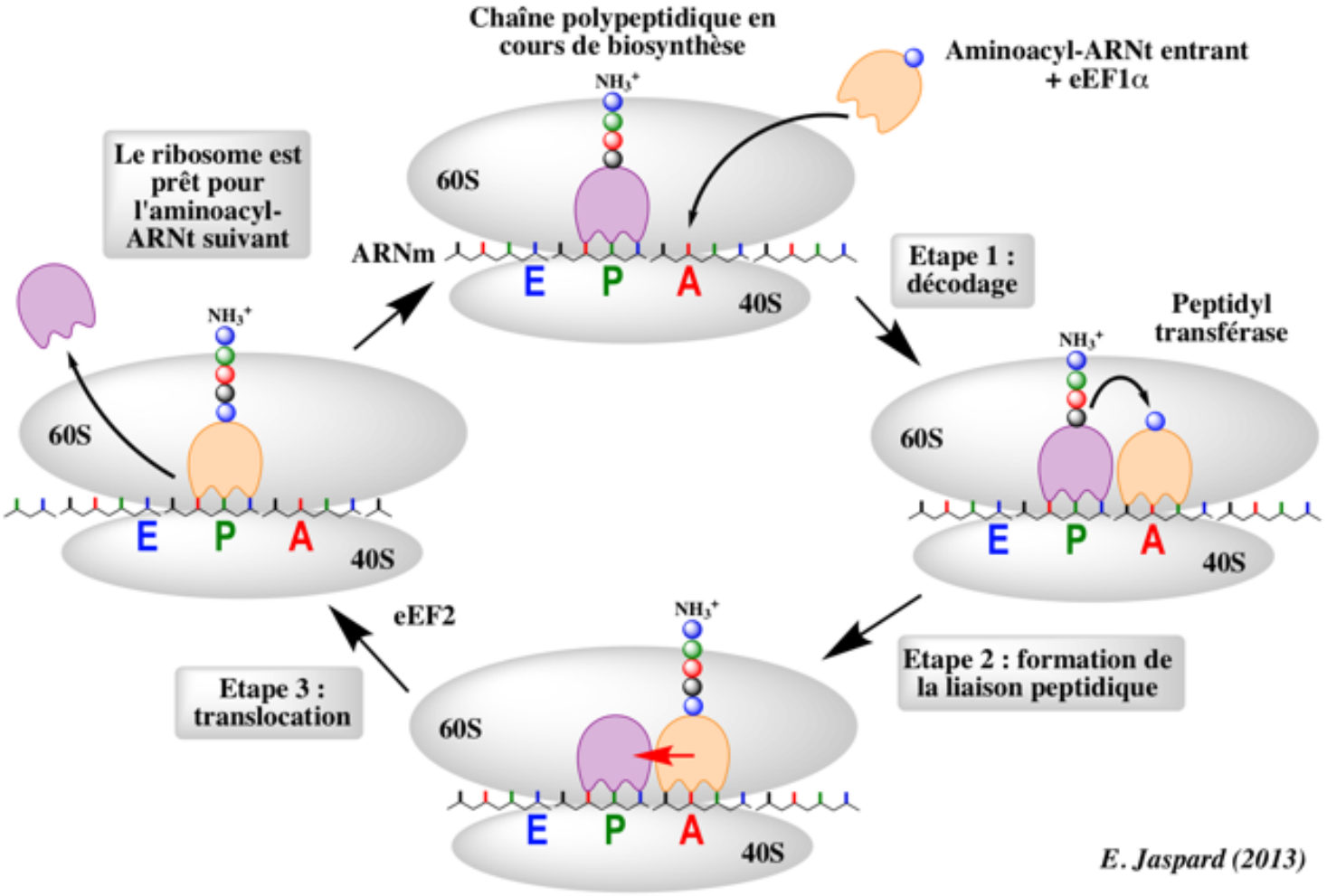


Interaction RBS et ARNr 16S pour positionnement du codon AUG au site P

RBS = Ribosome Binding Site

La traduction

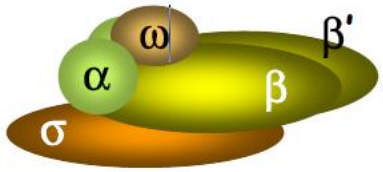
L'élongation :



La transcription

3 étapes :

- l'initiation : reconnaissance de séquences spécifiques sur l'ADN : le promoteur
- l'élongation
- la terminaison



ARN polymérase procaryote

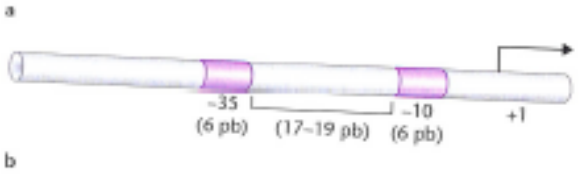
Transcription des gènes codant pour des protéines

Chez les bactéries :

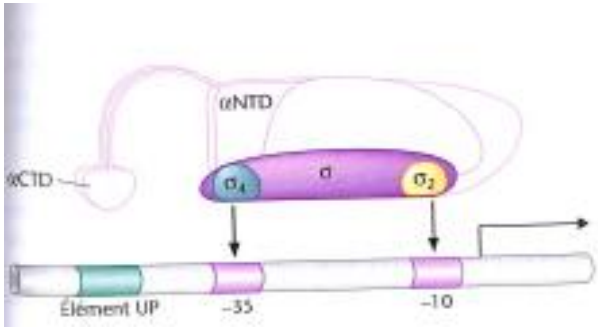
L'initiation :

Le facteur σ de la RNA polymérase reconnait deux régions constituant le promoteur : la région -35 et la région -10 (TATAAT box) séparées par une région de taille variable (17 à 19 pb). Le promoteur se trouve en amont du début du (des) gène(s).

Promoteurs bactériens



Recrutement du cœur de l'ARN polymérase au promoteur

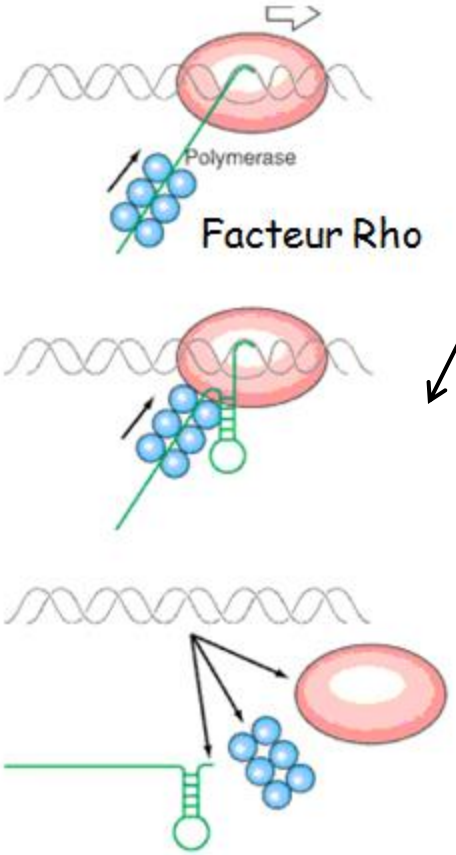


La transcription

Transcription des gènes codant pour des protéines

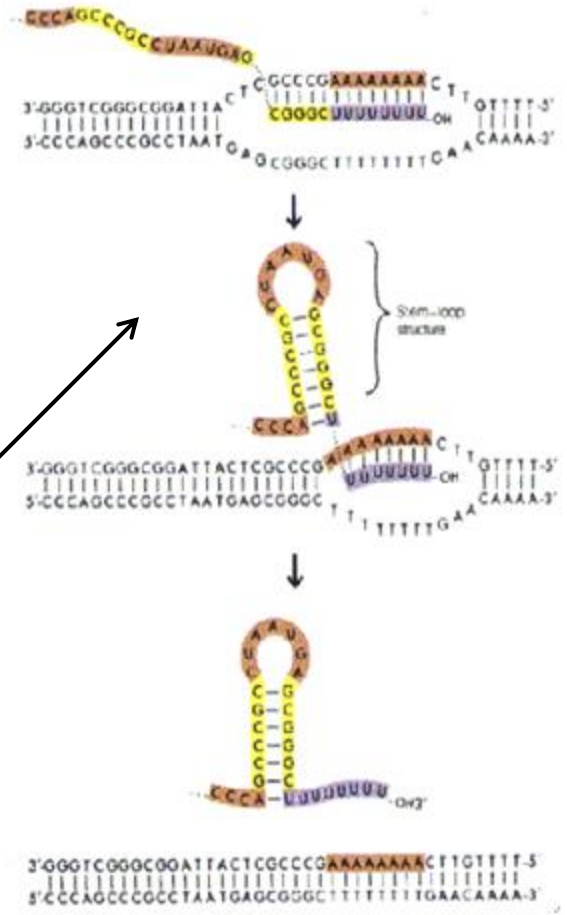
Chez les bactéries :

La terminaison : deux types de terminateurs Rho-dépendants et Rho-indépendants

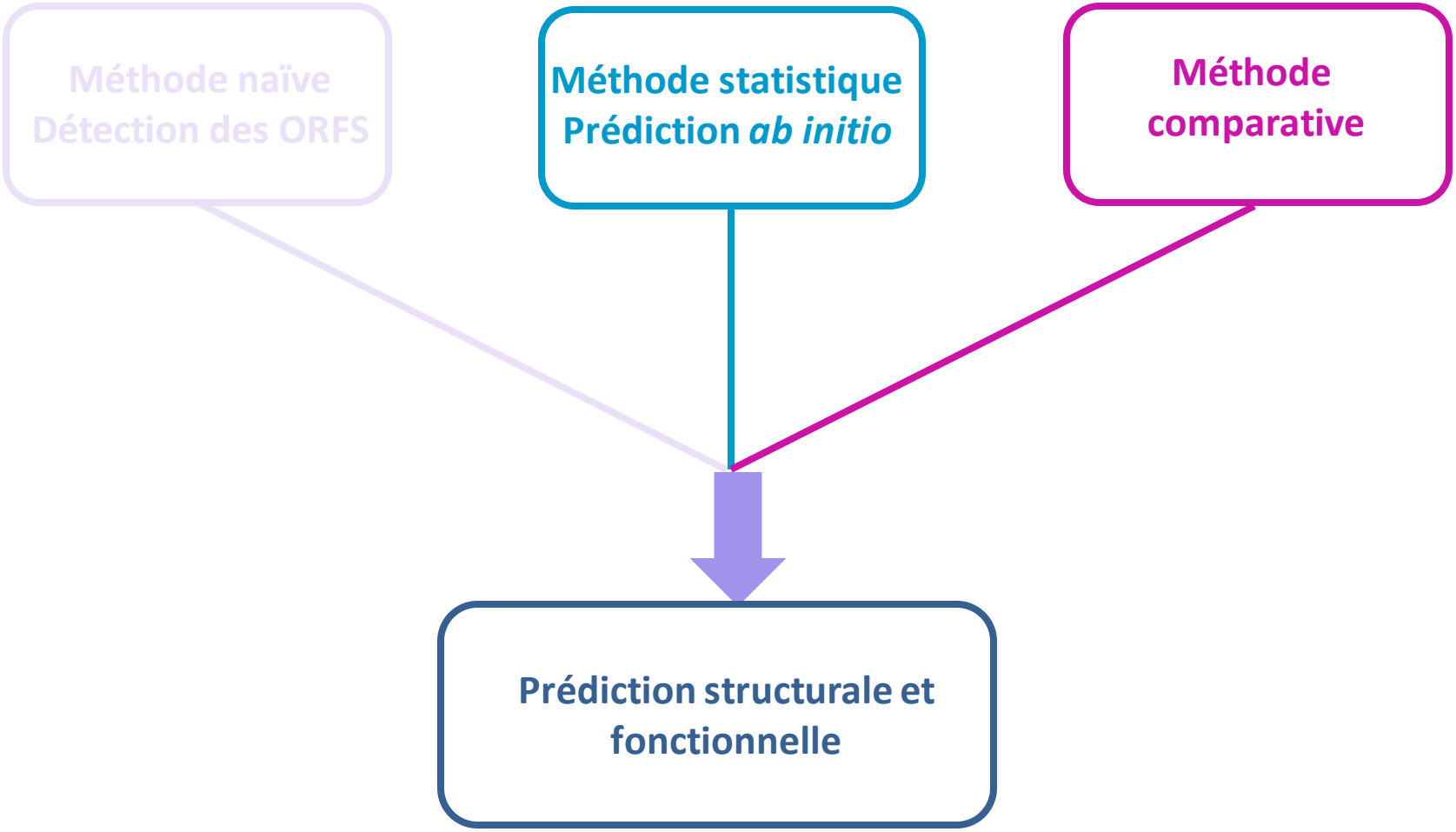


Rho-dépendant : L'ATPase Rho, une protéine qui se déplace le long du transcrit naissant jusqu'à rattraper la polymérase, stimule la terminaison de la transcription

Rho-indépendant : deux éléments une courte séquence répétée inversée suivie d'une série d'environ 8 paires de bases A-T. Le transcrit est capable de former une structure secondaire tige-boucle qui provoque la désorganisation de l'ARN polymérase. La suite du transcrit est faiblement apparié au brin ADN ce qui conduit à sa libération.



Recherche des régions codant pour des protéines chez les procaryotes



Une méthode naïve : ORFfinder (NCBI)

Recherche les phases ouvertes de lecture, les ORFs, dans les 6 cadres de lecture (les 3 cadres du brin direct et les 3 cadres du brin complémentaire).

Attention problème de sémantique :

Une ORF peu être définie entre deux codons stop

stop XXXXXXXXXXXXXXXXXXXXXXXXXXXXX stop

n codons

Ou entre un codon initiateur (start) et un codon stop

ATG XXXXXXXXXXXXXXXXXXXXXXXX stop

n codons

On considère en général que les ORFs supérieures à 100 codons (300 pb) comme étant potentiellement codantes (analyse statistique a montré que bien que des gènes de taille inférieure à 100 codons existent, la majorité des petites ORFs étaient des faux positifs).

>BS 1-8301
tttcgaggaaaatgtgcaataaccaactcatttcccgggcaattccgccg
gttccgaatgatacgaacaactgagactgagccgcaaatggttcagtctt
tttacatggcagccagagggctttgtgcacttgacatttgtgaaaaagaa
agtaaaatattttactaaaacaatgcgagctgaataatggaggcagatac
aatggcgacaattaaagatatcgcgcaggaagcgggattttcaatctcaa
ccgtttcccgcgttttaaataacgatgaaagcctttctgttcctgatgag
acacgggagaaaatctatgaagcggcggaaaagctcaattaccgcaaaaa
aacagtaaggccgctggtgaaacatattgcgtttttatattggctgacag
ataaagaagaattagaagatgtctattttaaaacgatgagattagaagta
gagaaactggcgaaagcattcaatgtcgatatgaccacttataaaatagc
ggatggaatcgagagcattcctgaacatacggaagggtttattgccgtcg
gcacattttcagatgaagagctggctttcctcagaaatctcactgaaaac
ggcgtgttcatcgattcaactcctgatcccgatcattttgactcggtaag
gcccgatttggcacaaatgacaaggaagacggtaaacatcctgactgaga
aggggcataagagcatcggttttatcggcggcacatacaaaaatccgaat
accaatcaggatgaaatggacatccgtgaacaaaccttcagatcctatat
gagggaaaaagccatgctggacgagcgctatattttctgtcatcgcggat
tctctgtagaaaacggctaccgcctgatgtcagcagcgatcgacacatta
ggcgatcagcttccgactgcttttatgattgcagcggacccgattgcagt
gggctgtctgcaagccctgaacgaaaaaggaattgccataccaaacaggg
taagcattgtgagtatcaacaacatcagcttcgcgaagtatgtctcgcct
cctctgacgacgtttcatattgatatacatgaattatgtaaaaacgctgt
tcaattactgcttgaacaagtgcaggacaagagaagaacggtaaaaacat
tatatgtgggcgcagaattaatcgtcaggaagagtatgaattaaggatga
cttaggacactaagtcatttttatttaggtaaaaaaatttactctatga
agtaaatagtttgtttacacattttctcaggcatgctatattatctttaa
agcgctttcattcctaccgaaagggtgacaatcaatgaaaatggcaaaaa
agtgttccgtattcatgctctgcgcagctgtcagtttatccttggcggct
tgcggcccaaaggaaagcagcagcgccaaatcgagttcaaaagggtcaga
gcttgttgtatgggaggataaagaaaagagcaacggcattaaagacgctg
tggctgcatttgaaaaagagcatgatgtgaaggtcaaagtcgttgaaaaa
ccgtatgccaagcagattgaagatttgcgaatggatggaccggccggcac
aggccctgacgtgttaacaatgccaggggaccaaatcggaaccgctgtca
cggaaggattactcaaggaattacatgtcaaaaaagacgttcaatcactt
tatactgacgcttccattcagtctcaaatggtagatcaaaagctttatgg
actgccaaaagcggtcgaaacgactgtgctttttacaacaaagatctca
tcacagaaaaggaattgcccaaaacgctggaagagtggtacgactattcc

Exemple traité : fragment de 8300 pb du génome de *Bacillus subtilis*

Open Reading Frame Finder

ORF finder searches for open reading frames (ORFs) in the DNA sequence you enter. The program returns the range 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.

This 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 nucleotide sequence in FASTA format:

```
>BS 1-8301
tttcgaggaaaatgtgcaataaccaactcatttcccgggcaattccgccggttccgaatg
atacgaacaactgagactgagccgcaaatggttcagtctttttacatggcagccagaggg
ctttgtgcacttgacatttgtgaaaaagaaagtaaaatattttactaaaacaatgcgagc
tgaataatggaggcagatacaatggcgacaattaaagatatcgcgcaggaagcgggattt
tcaatctcaaccgtttcccgcgttttaaataacgatgaaagcctttctgttcctgatgag
acacgggagaaaatctatgaagcggcggaaaagctcaattaccgcaaaaaaacagtaagg
ccgctggtgaaacatattgcgtttttatattggctgacagataaagaagaattagaagat
gtctattttaaaacgatgagattagaagtagagaaactggcgaaagcattcaatgtcgat
atgaccacttataaaatagcggatggaatcgagagcattcctgaacatacggaagggttt
attgccgtcggcacattttcagatgaagagctggctttcctcagaaatctcactgaaaac
```

From: To:

Choose Search Parameters

Minimal ORF length (nt): 300

Genetic code: 1. Standard

ORF start codon to use:

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

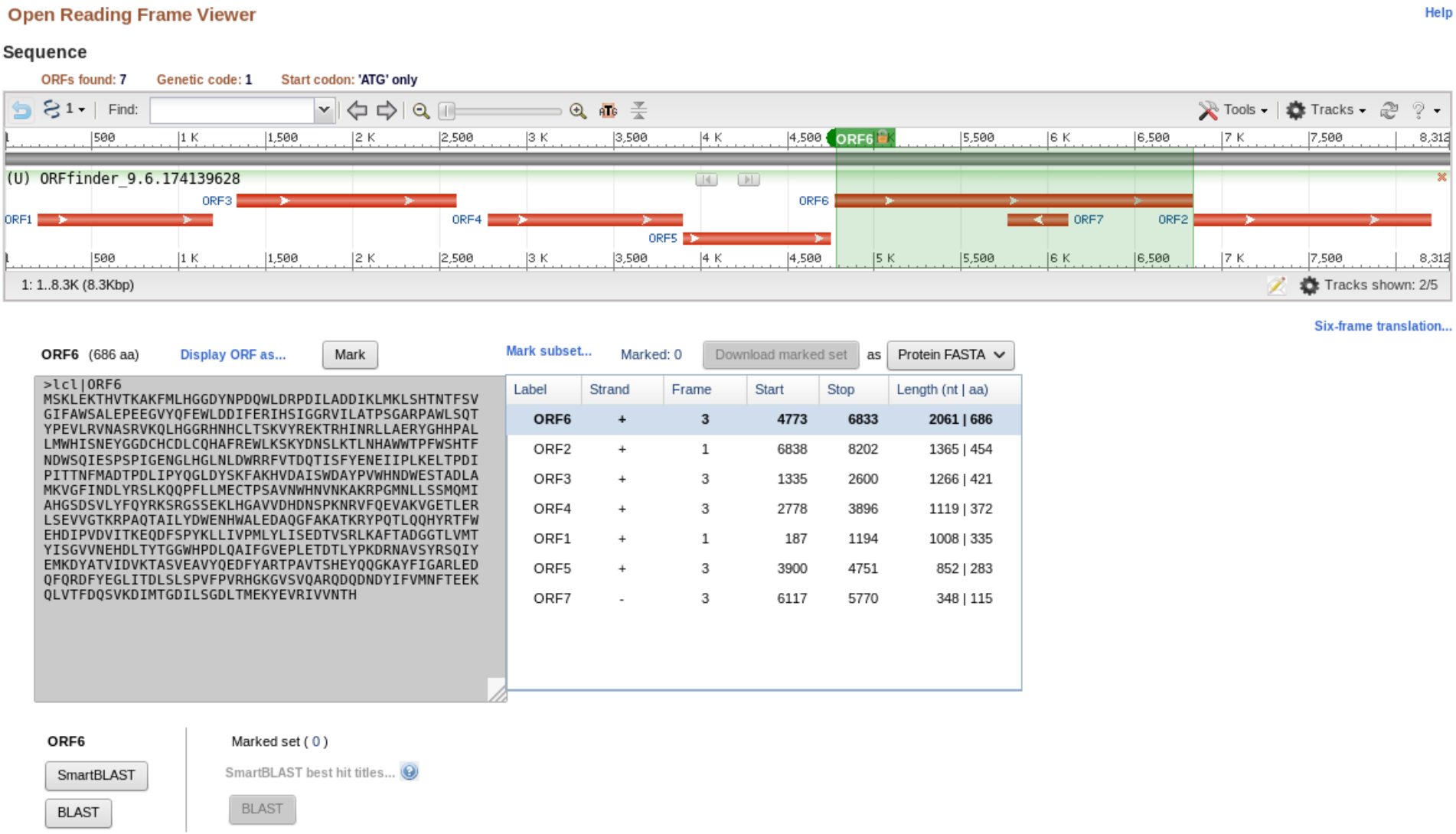
Ignore nested ORFs:

Start Search / Clear

Submit Clear

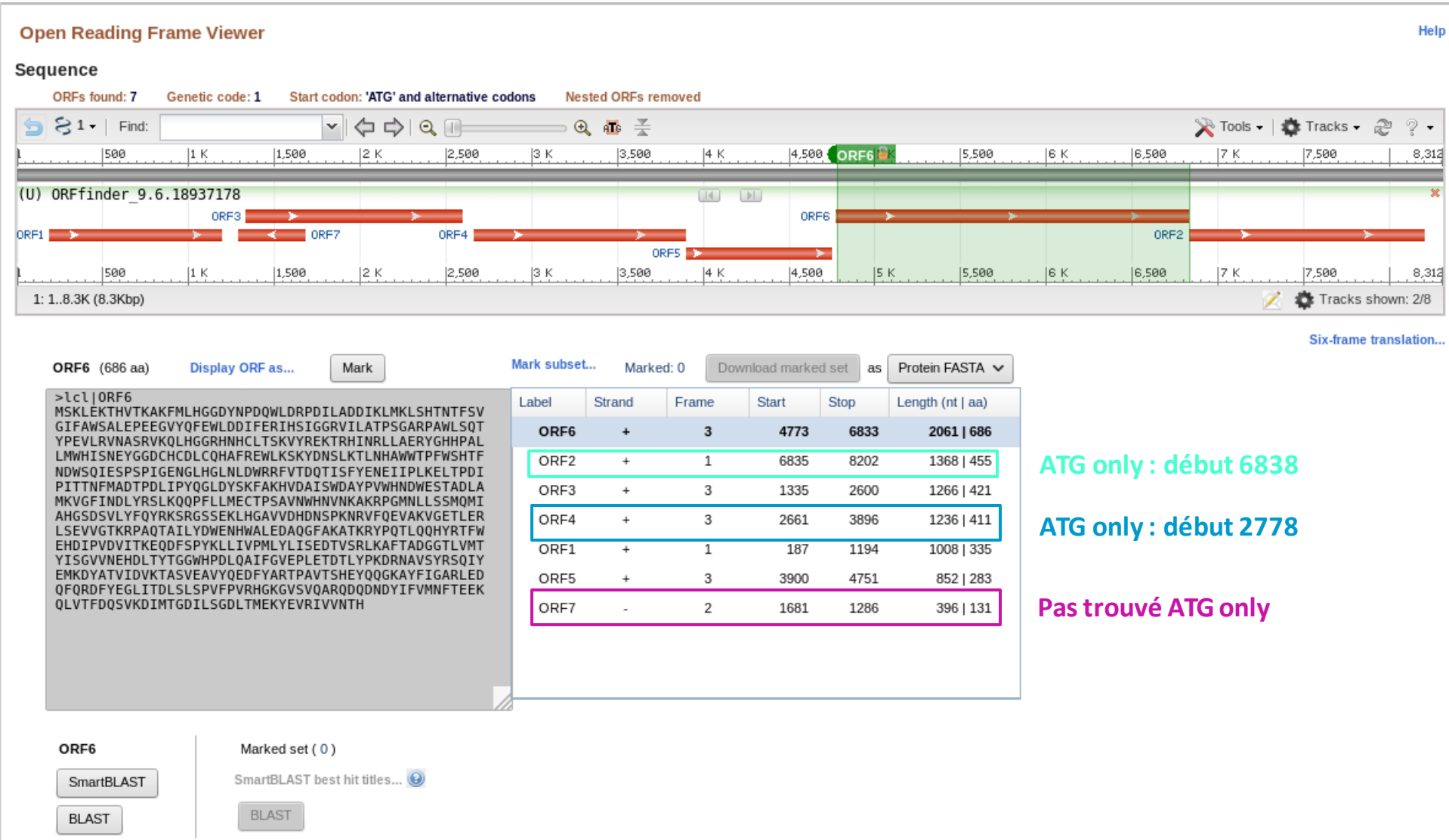
Résultat de ORFfinder : ORFs de plus de 300 pb

Options : - ATG only
- Ignore nested ORF pas coché



Résultat de ORFfinder : ORFs de plus de 300 pb

Options : - ATG and alternative initiation codons
- Ignore nested ORF coché



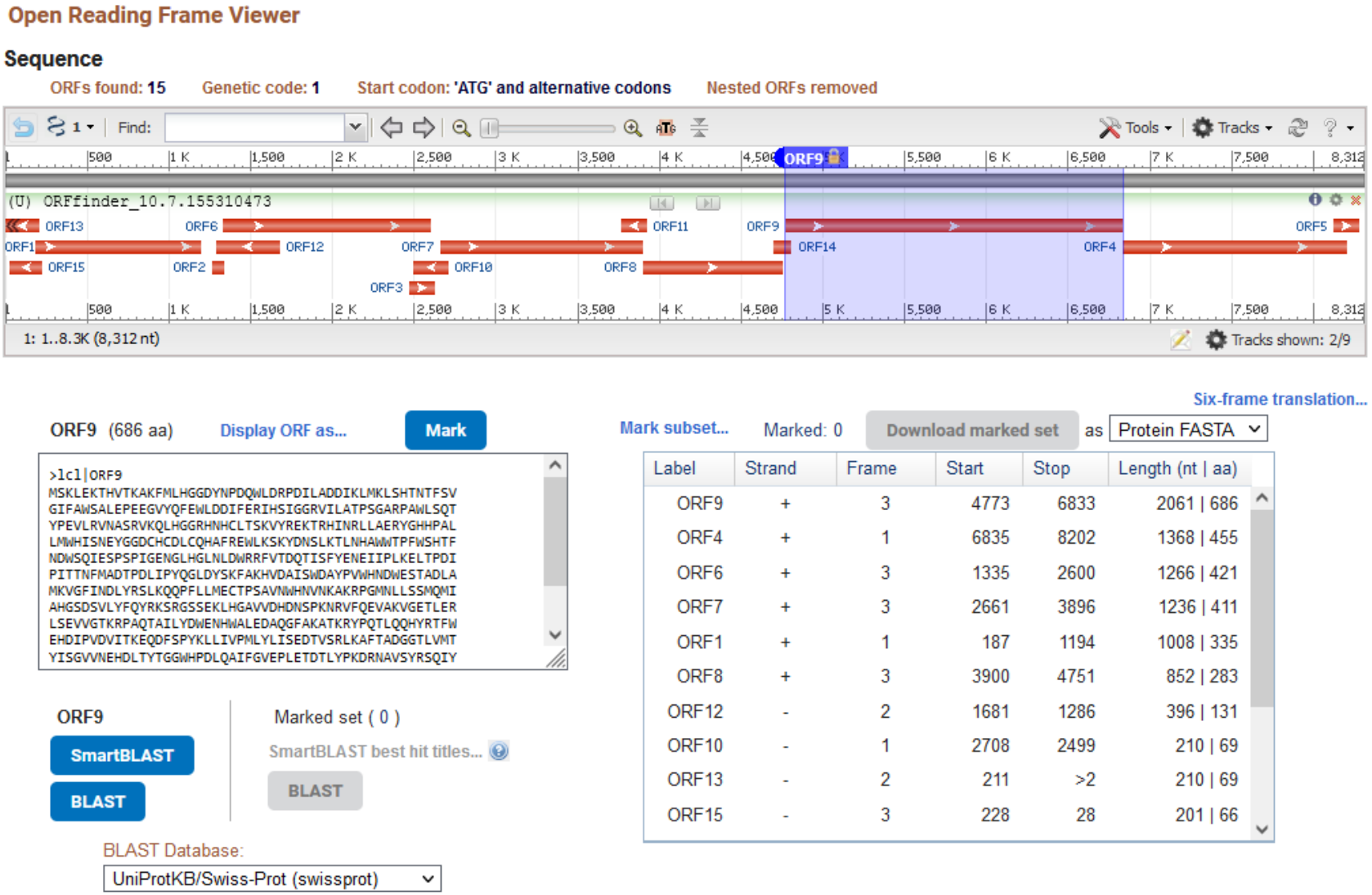
ATG only : début 6838

ATG only : début 2778

Pas trouvé ATG only

Résultat de ORFfinder : ORFs de plus de 75 pb

Options : - ATG and alternative initiation codons
- Ignore nested ORF coché



Limites d'ORFfinder :

- Seul critère la taille, mais tous les ORF ne sont pas des régions codantes (Coding Sequence : CDS) quand ils sont de petites tailles.

- Problème d'identification du « vrai » codon initiateur car plusieurs possibilités en fonction du choix « ATG » ou « codon initiateur alternatif », GTG et TTG chez les bactéries avec comme fréquence 13% et 9% chez *Bacillus subtilis* par exemple.

- ne prend pas en compte le biais de l'utilisation des triplets existant dans les phases codantes car structurées en codons.

Utilisation de méthode statistique

Biais des codons chez les bactéries

Prise en compte du biais de l'utilisation des triplets existant dans les phases codantes par rapport aux régions non codantes car structurées en codons.

Biais dans l'utilisation des codons dus à :

- la différence de fréquence des acides aminés (Leu plus fréquent que Trp par exemple)
- la dégénérescence du code génétique (61 codons -> 20 aa)
- pour un acide aminé donné, certains codons peuvent être plus fréquemment utilisés que d'autres. Ces préférences varient en fonction :
  - la composition en bases de l'organisme (riche ou pauvre en C+G) : **usage du code** différent
  - du taux d'expression du gène : il a été montré chez *E. coli* que les gènes fortement exprimés utilisaient préférentiellement certains codons correspondant aux ARNt les plus abondants dans la cellule (efficacité de la traduction, coadaptation codons/ARNt. (Calcul du **Codon Adaptation Index (CAI**) pour prédire gène fortemnt exprimé ou pas)

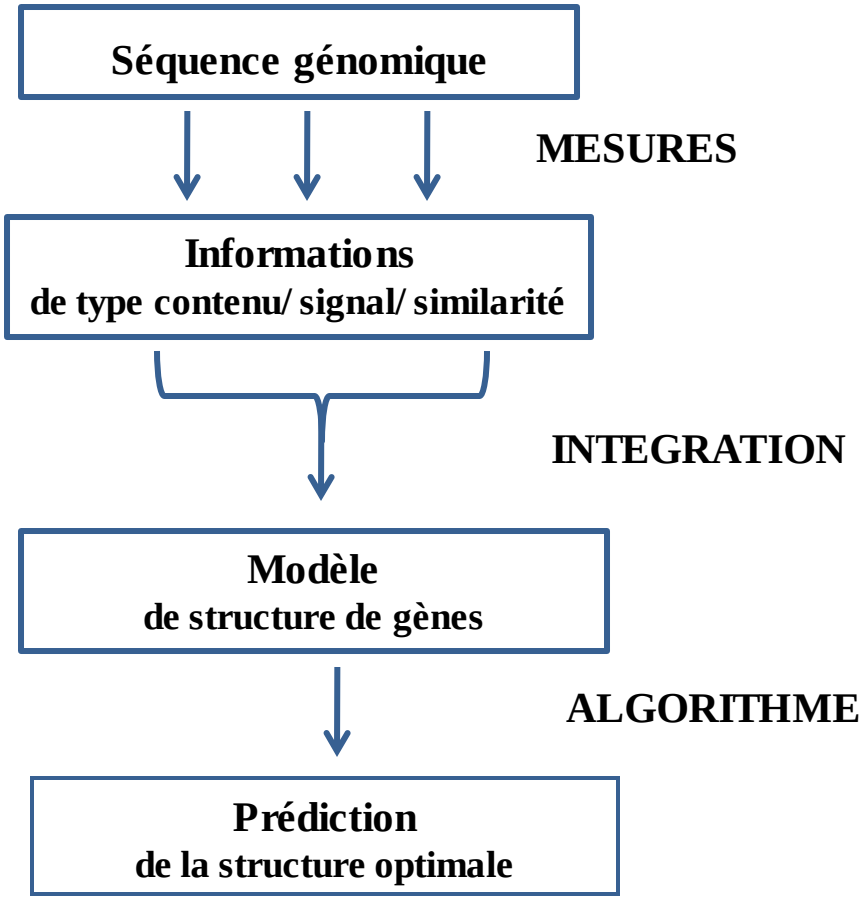
Exemples d'usage des codons chez les bactéries

Espèce	GC% codant	GC% 1ère pos. codon	GC% 2ème pos. codon	GC% 3ème pos. codon
Synechocystis sp.	48.25	55.82	39.74	49.19
Streptomyces coelicolor	72.30	72.67	51.39	92.83
Escherichia coli 0157:H7	51.54	58.44	41	55.17
Bacillus subtilis	44.36	52.10	36.08	44.91

A.A.	codon	% S. sp. (cyano)	% S. coelicolor	% E. coli	% B. subtilis
Gly	GGG	0.24	0.19	0.164	0.16
Gly	GGA	0.18	0.075	0.123	0.315
Gly	GGT	0.27	0.096	0.331	0.187
Gly	GGC	0.31	0.64	0.382	0.337
Glu	GAG	0.264	0.846	0.325	0.32
Glu	GAA	0.736	0.154	0.675	0.68
Asp	GAT	0.646	0.05	0.631	0.636
Asp	GAC	0.354	0.95	0.369	0.364

Recherche des régions codant pour des protéines

Fonctionnement schématique d'un logiciel de prédiction de gènes



Méthode statistique
Traitement de l'information de type contenu

Utilisation de méthodes statistiques prenant en compte ces biais d'utilisation des codons. Plus récemment avec l'augmentation des données pour établir les systèmes de référence, prise en compte de la composition en hexanucléotides (mots de longueur 6).

Les méthodes statistiques couramment utilisées :

- Modèles de Markov
- Modèles de Markov interpolés (IMM)
- Modèles de Markov caché (HMM)

Modèle de Markov

Un modèle de Markov d'ordre *k* appliqué aux séquences ADN est entièrement défini par les deux probabilités suivantes :

$$\begin{cases} P_0(w_1^k) \longrightarrow \text{Probabilité initiale du mot } w^k \\ P(x / w^k) \longrightarrow \text{Probabilité d'observer } x \text{ sachant que le mot } w^k \text{ le précède} \end{cases}$$

Le modèle est donc caractérisé par la probabilité initiale de chaque mot (exemple si *k* vaut 2 par la fréquence du dinucléotide observé) et par la probabilité d'observer une base *x* en fonction du mot précédent.
Exemple : probabilité d'observer la base A sachant le mot précédent GT = probabilité d'observer le triplet GTA = fréquence du triplet GTA

Modèle de Markov : Présentation de GeneMark

(Borodovsky et al., Nucleic Acids Res.,22,4756-67)

La méthode repose sur le modèle probabiliste suivant appelé modèle de Markov:

Hypothèse 1: La probabilité d'observer une base à une position donnée dépend :

- Régions non codantes
  - ✓ des bases précédant cette position
  - ✓ modélisé par un modèle de Markov homogène

- Régions codantes
  - ✓ des bases précédant cette position
  - ✓ de sa localisation dans le codon
  - ✓ Modélisé par un modèle de Markov non-homogène

Hypothèse 2: Une région particulière ne peut être que dans un des 7 états suivants:

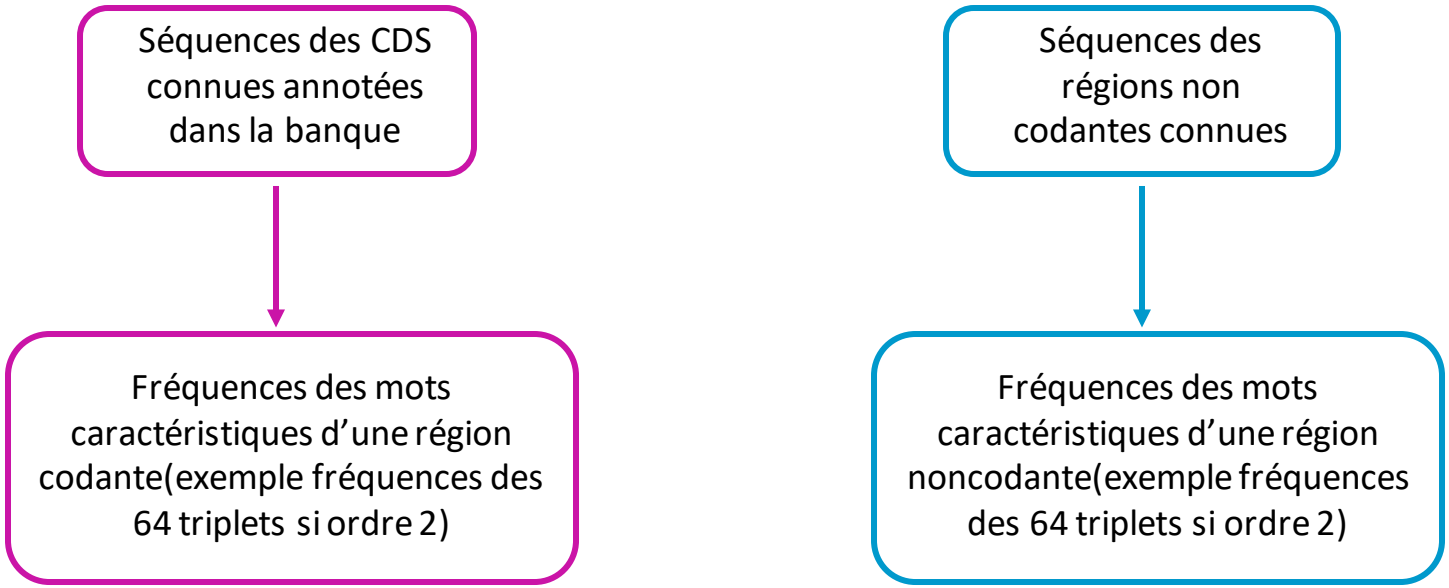
- 1. codant en phase 1 sur le brin direct
- 2. codant en phase 2 sur le brin direct
- 3. codant en phase 3 sur le brin direct
- 4. codant en phase 4 sur le brin indirect
- 5. codant en phase 5 sur le brin indirect
- 6. codant en phase 6 sur le brin indirect
- 7. non-codant

Prédiction : calculer les probabilités d'observer la région dans un état *i* sachant que l'un des 7 états est réalisé (formule de Bayes).

Mise en oeuvre

Nécessite d'avoir des tables de référence pour calculer la fréquence de chaque mots.
Annotation d'un fragment génomique issu d'une espèce bactérienne *B* :

- Des données sont disponibles dans les bases de données comme GenBank ou EMBL pour l'espèce *B* ou une espèce proche dans l'évolution -> utilisation d'une référence externe
  - ✓ création de deux ensemble d'apprentissage et calcul des tables de référence



  - ✓ Prédiction sur notre séquence : comparaison des mots rencontrés dans notre fragment avec chaque table : calcul de la probabilité que la portion de la séquence soit dans chacun des 7 états. Choix de l'état le plus probable

- Pas de données disponibles pour l'espèce *B :*
  - ✓ Tables obtenues en utilisant un modèle heuristique

Résultat de GeneMark sur le fragment de *B. subtilis*

Entête du fichier :

```
Sequence: EMBOSS_001 Reversed:
Sequence file: seq.fna
Sequence length: 8312
GC Content:  45.19%
Window length: 96
Window step: 12
Threshold value: 0.500
---
Matrix: Bacillus_subtilis_168
Matrix author: -
Matrix order: 4
```

Fin du fichier :

List of Regions of interest
(regions from stop to stop codon w/ a signal in between)

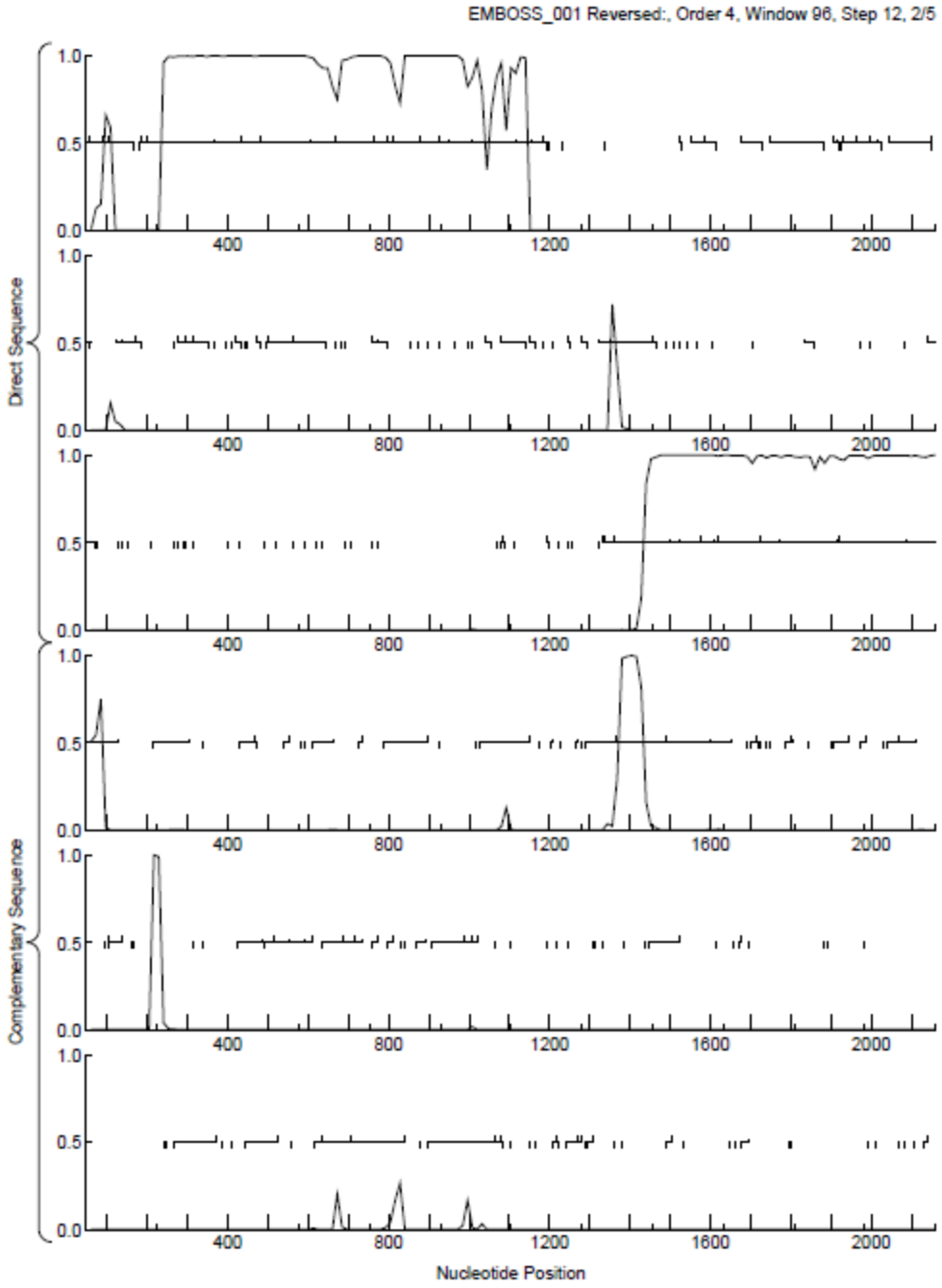
LEnd	REnd	Strand	Frame
181	1194	direct	fr 1
1286	1693	complement	fr 1
1326	2600	direct	fr 3
2610	3896	direct	fr 3
3894	4751	direct	fr 3
4749	6833	direct	fr 3
6820	8202	direct	fr 1

List of Open reading frames predicted as CDSs, shown with alternate starts
(regions from start to stop codon w/ coding function >0.50)

UNIVERSITÉ TOULOUSE III PAUL SABATIER

Left end	Right end	DNA Strand	Coding Frame	Avg Prob	Start Prob	
187	1194	direct	fr 1	0.85	0.99	→ ORF Finder
202	1194	direct	fr 1	0.86	0.90	
367	1194	direct	fr 1	0.89	0.10	
436	1194	direct	fr 1	0.88	0.02	
481	1194	direct	fr 1	0.87	0.01	
1335	2600	direct	fr 3	0.87	0.04	→ ORF Finder
1341	2600	direct	fr 3	0.87	0.02	
1365	2600	direct	fr 3	0.89	0.08	
1500	2600	direct	fr 3	0.95	0.04	
1527	2600	direct	fr 3	0.95	0.02	
1581	2600	direct	fr 3	0.95	0.01	
2631	3896	direct	fr 3	0.80	0.75	
2640	3896	direct	fr 3	0.80	0.87	
2778	3896	direct	fr 3	0.82	0.28	→ ORF Finder (ATG only)
2814	3896	direct	fr 3	0.81	0.03	2661 start alternatif
2868	3896	direct	fr 3	0.80	0.49	
3900	4751	direct	fr 3	0.74	0.26	→ ORF Finder
3912	4751	direct	fr 3	0.75	0.01	
3966	4751	direct	fr 3	0.80	0.65	
4116	4751	direct	fr 3	0.80	0.27	
4137	4751	direct	fr 3	0.80	0.05	
4158	4751	direct	fr 3	0.79	0.14	
4770	6833	direct	fr 3	0.90	0.82	
4773	6833	direct	fr 3	0.90	0.79	→ ORF Finder
4815	6833	direct	fr 3	0.92	0.13	
4890	6833	direct	fr 3	0.92	0.01	
5226	6833	direct	fr 3	0.93	0.02	
6838	8202	direct	fr 1	0.85	0.06	→ ORF Finder (ATG only)
6877	8202	direct	fr 1	0.88	0.71	6835 start alternatif
6913	8202	direct	fr 1	0.89	0.74	
6925	8202	direct	fr 1	0.89	0.02	
6931	8202	direct	fr 1	0.89	0.01	

Résultat graphique de GeneMark sur le fragment de *B. subtilis*



Interpolated Markov Model (IMM)

Glimmer (Salzberg et *al.*, Nucleic Acids Res.,26,544-48)

Modèle de Markov d'ordre *k* :apprendre 4^{k+1} probabilités
Dans le cadre de la prédiction des CDS prise en compte des 6 cadres de lecture, donc nécessité d'apprendre 6 * 4^{k+1} probabilités
Si modèle de Markov d'ordre 5 : 4096 probabilités à définir (hexamères)
Si on considère les 6 cadres de lecture : 24 576 probabilités

Plus l'ordre du modèle est élevé, moins l'estimation des paramètres du modèle va être fiable.

Pour certains *k*mers rares même avec un grand jeu d'apprentissage comme un génome entier, il peut être difficile d'obtenir des estimations précises et inversement pour certains *k*mers fréquents un modèle de markov d'ordre élevé donnera des estimations plus précises.

Souhait : un modèle de Markov qui utilise les ordres les plus élevés quand il y a assez de données disponibles et des ordres moins élevés dans les cas où les données sont insuffisantes.

Interpolation des modèles de Markov

Modèle de Markov Caché (HMM Hidden Markov Model)

En biologie on recherche souvent à mettre le bon label sur chaque résidu d'une séquence.
Par exemple :

- définir si un résidu appartient à une CDS, à une région intergénique etc..
- déterminer si une nouvelle séquence protéique appartient à une famille de protéines donnée
- etc…

Les modèles de Markov cachés (HMM) permettent de réaliser des modèles probabilistes d'une suite de problèmes linéaires labellisés.

Ils sont utilisés pour :

- déterminer la structure en gènes d'un fragment génomique
- réaliser des alignements multiples
- déterminer des profils
- identifier des sites de régulations
- etc…

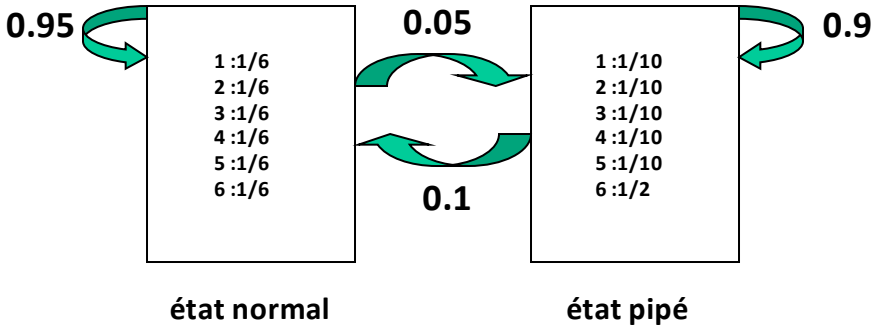
Modèle de Markov Caché (HMM hidden Markov Model) Un exemple simple non biologique

Première étape : modéliser le problème en terme d'états

Exemple simple : dans un casino, ils utilisent la plupart du temps un dé normal, mais occasionnellement aussi un dé pipé. Le dé pipé a une probabilité de 0.5 pour le 6 et de 0.1 pour les autres chiffres.
La probabilité de changer du dé normal au dé pipé avant chaque jet est de 0.05, et celle de passer du dé pipé au dé normal est de 0.1.
Le changement de dé suit donc un processus de Markov.
Dans chaque état du processus, le résultat d'un jet de dé est associé à des probabilités différentes.

L'ensemble du processus décrit peut être modélisé par un HMM :

On a deux états : dé normal, dé pipé



Qu'est ce qui est caché :

Observation : résultat du jet de dé
On ne sait pas quel dé est utilisé

L'état de la séquence d'observation est caché

Modèle de Markov Caché (HMM hidden Markov Model)

Le modèle est décrit par deux ensembles de probabilités :

- probabilités de passer d'un état à l'autre : probabilités de transition
- probabilités d'observer un symbole pour un état donné : probabilités d'émission

A ceci s'ajoute le choix de l'état initial.

Un HMM est donc défini par :

- Un vecteur de probabilités initiales $\Pi = (\pi_i)$
- un vecteur de probabilités de transition $A = (a_{ij})$
  (probabilité de passer de l'état *i* à l'état *j*)
- une matrice de probabilités d'émission $E = (e_i(b))$
  (probabilité que le symbole *b* soit observé dans l'état *i*)

La probabilité d'une séquence d'observation *x* et d'une séquence d'état (chemin) π est donnée par :

$$P(x,\pi) = a_{0\pi_1} \prod_{i=1}^{L} e_{\pi_i}(x_i) a_{\pi_i \pi_{i+1}}$$

Problème : en général on ne connaît pas π. On cherche à l'estimer.

Modèle de GenMark.hmm

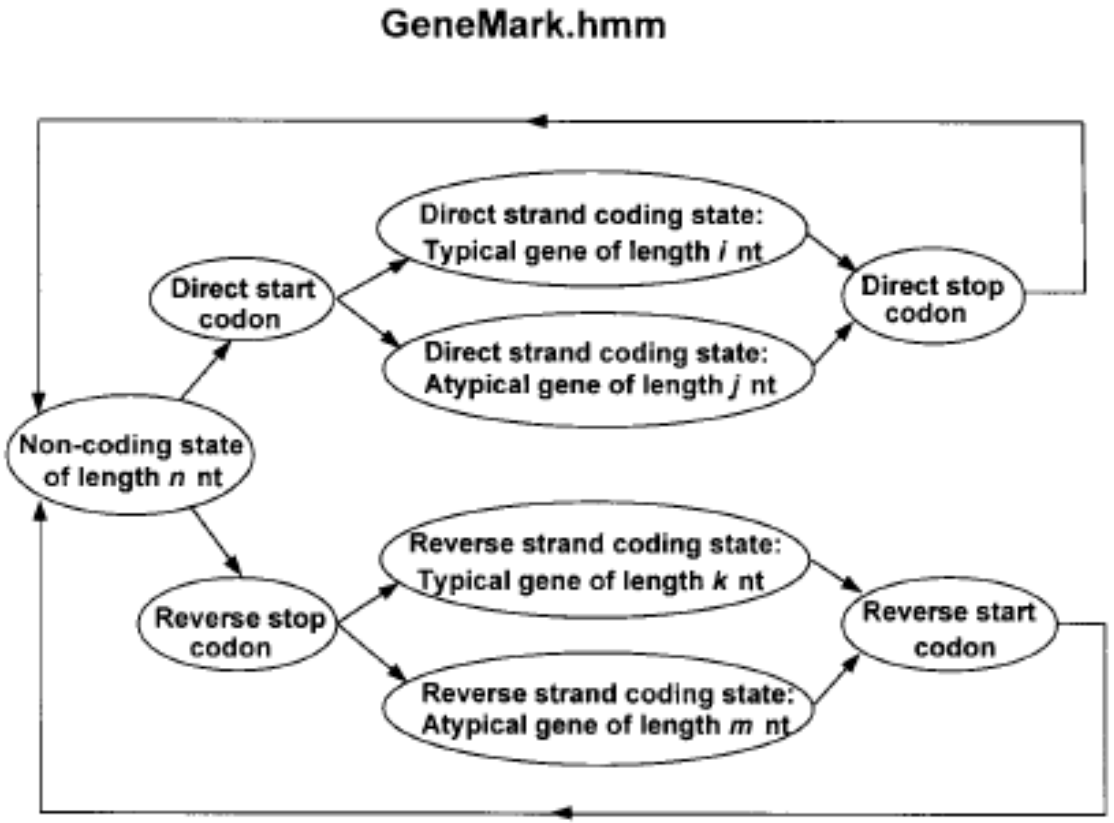
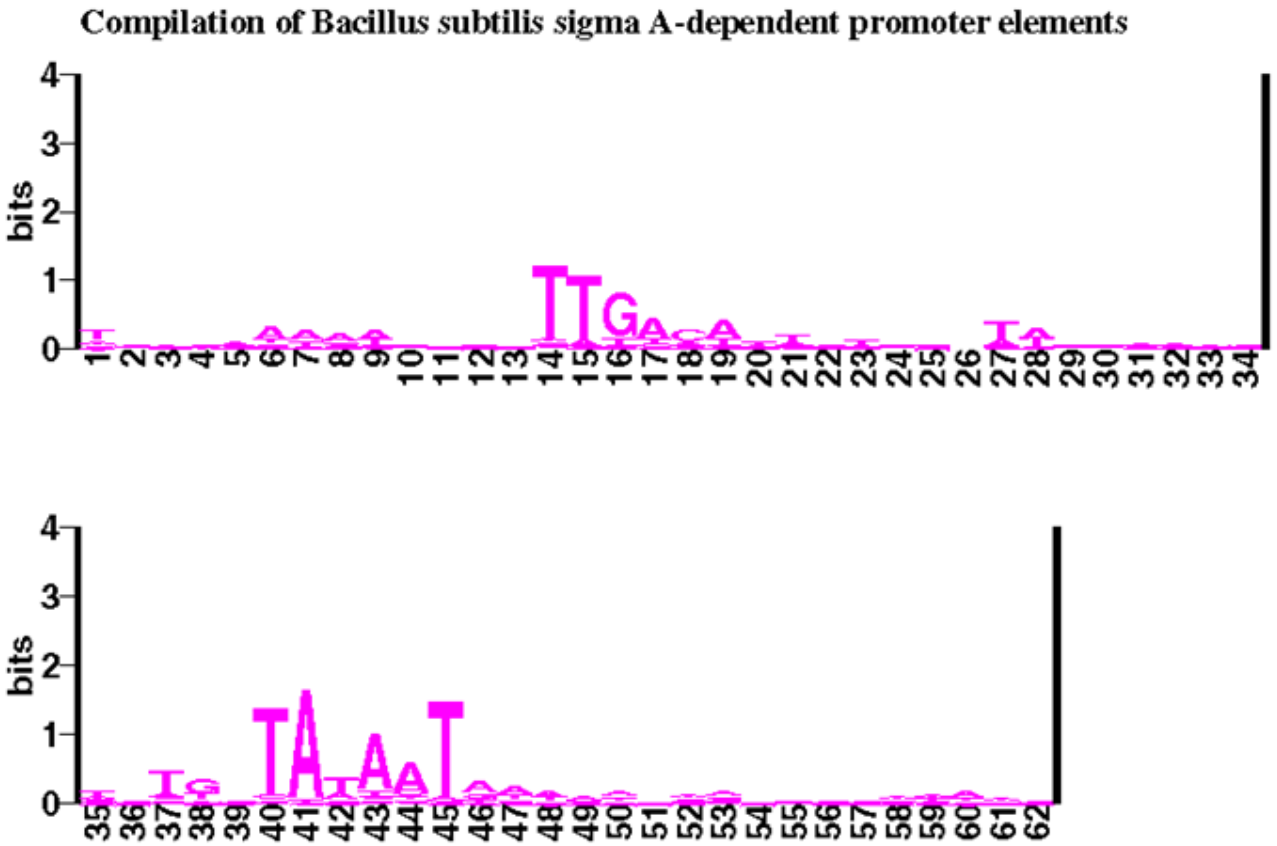


Figure 1. Hidden Markov model of a prokaryotic nucleotide sequence used in the GeneMark.hmm algorithm. The hidden states of the model are represented as ovals in the figure, and arrows correspond to allowed transitions between the states.

(extrait de Nucleic Acids Res. (1998), 26, 1107-1115)

Traitement de l'information de type signal

Différentes façon de représenter la conservation des séquences impliquées dans un processus donné (promoteur lors de la transcription, ribosome binding site lors de la traduction, jonction d'épissage etc…) et ensuite de rechercher ces « signaux » dans une nouvelle séquence.



Petit rappel : Motifs

Définition : zone d'une séquence nucléique ou protéique présentant une conservation quand on compare plusieurs séquences.

• correspondent en général à des zones fonctionnelles

• ADN et ARN : aussi appelé **signal**, ces zones interviennent souvent dans des systèmes de régulation, ex :

- ➢ -10 et -35 des promoteurs chez les procaryotes, jonction d'épissage,
- ➢ boite CRE (catabolite repression element) : après mise en évidence de certains gènes soumis à la répression catabolique chez *B. subtilis*, l'identification du signal permet de rechercher dans le génome complet les boites CRE et donc les gènes qui pourraient être soumis à la répression catabolique.

• différents des signaux reconnus par les enzymes de restrictions qui reconnaissent des séquences exactes, ex: GAATTC pour ECOR1.

• Les motifs et profils présentent une certaine **variabilité** (souvent impliquée dans la variabilité de la régulation par une reconnaissance plus ou moins forte des partenaires)

Comment représenter cette variabilité ?

- ➢ séquence consensus
- ➢ matrice de poids

Représentation : Séquence consensus

Exemples des boîtes CRE:

acsA	TGAAAGCGTTACCA
acuA	TGAAAACGCTTTAT
amyE	TGTAAGCGTTAACA
gntR	TGAAAGCGGTACCA
hutP	TGAAACCGCTTCCA
licS	AGAAAACGCTTTCA
xylA	TGGAAGCGTAAACA
xylA	TGAAAGCGCAAACA
xylA	AGTAAGCGTTTACA
ackA	TGTAAGCGTTATCA
consensus	TGAAAGCGNTAACA
	T TC

Représentation : Matrice de poids

Exemples de 242 séquences de promoteurs (-10) chez *E. coli* :

Matrices des fréquences de chaque base *b* à chaque position *i* ($f_{b,i}$) du motif -10 (6 positions) :

Pos.	1	2	3	4	5	6
A	0.04	0.88	0.26	0.59	0.49	0.03
C	0.09	0.03	0.11	0.13	0.22	0.05
G	0.07	0.01	0.12	0.16	0.12	0.02
T	0.80	0.08	0.51	0.13	0.18	0.89

Avec $f_{b,i} = n_{b,i} / n_{tot}$

n_{tot} : nombre total de séquences analysées

Représentation : Matrice de poids position (Position Weight Matrix, PWM)

Exemples de 242 séquences de promoteurs (-10) chez *E. coli* :

Normalisation de la matrice : log matrice $\log_2(f_{b,i}/P_b)$

$f_{b,i}$ = fréquence observée de la base *b* à la position *i* dans toutes les séquences
P_b = fréquence de cette base dans l'ensemble du génome

Pos.	1	2	3	4	5	6
A	-2.76	1.88	0.06	1.23	0.96	-2.92
C	-1.46	-3.11	-1.22	-1.00	-0.22	-2.21
G	-1.76	-5.00	-1.06	-0.67	-1.06	-3.58
T	1.67	-1.66	1.04	-1.00	-0.49	1.84

Le rapport $f_{b,i}/P_b$ est une mesure de l'écart entre fréquence observée et attendue.

Utilisation d'une matrice de poids sur une séquence

Pos.	1	2	3	4	5	6
A	-28	18	1	12	10	-29
C	-15	-31	-12	-10	-2	-22
G	-18	-50	-11	-7	-11	-36
T	17	-17	10	-10	-5	18

A CTATAATCG

AC TATAATCG

ACT ATAATCG

Score1 = -15-17 +1-10+10-29 = -60

Score2 = 17+18+10+12+10+18 = 85

Score3 = -28-17+1+12-5-22 = -59

Théorie de l ’information : obtention de WebLogo

Shannon et Weaver (1949).

La valeur de l'information *I* à la position *j* d'un signal est donnée par :

$$I(j) = \sum_i f_{ij} \log_2 f_{ij} - \sum_i P_i \log_2 P_i$$

où :

P_i (*i* = 1 à 4) est la fréquence de la base *i* dans l'ensemble du génome (probabilité théorique)

f_{ij} est la fréquence observée de la base *i* à la position *j* d'un signal sur un ensemble d'exemples.

Les P_i étant estimées à 0.25 pour chacune des 4 bases on a :

$$\sum_i P_i \log_2 P_i = -2$$

donc

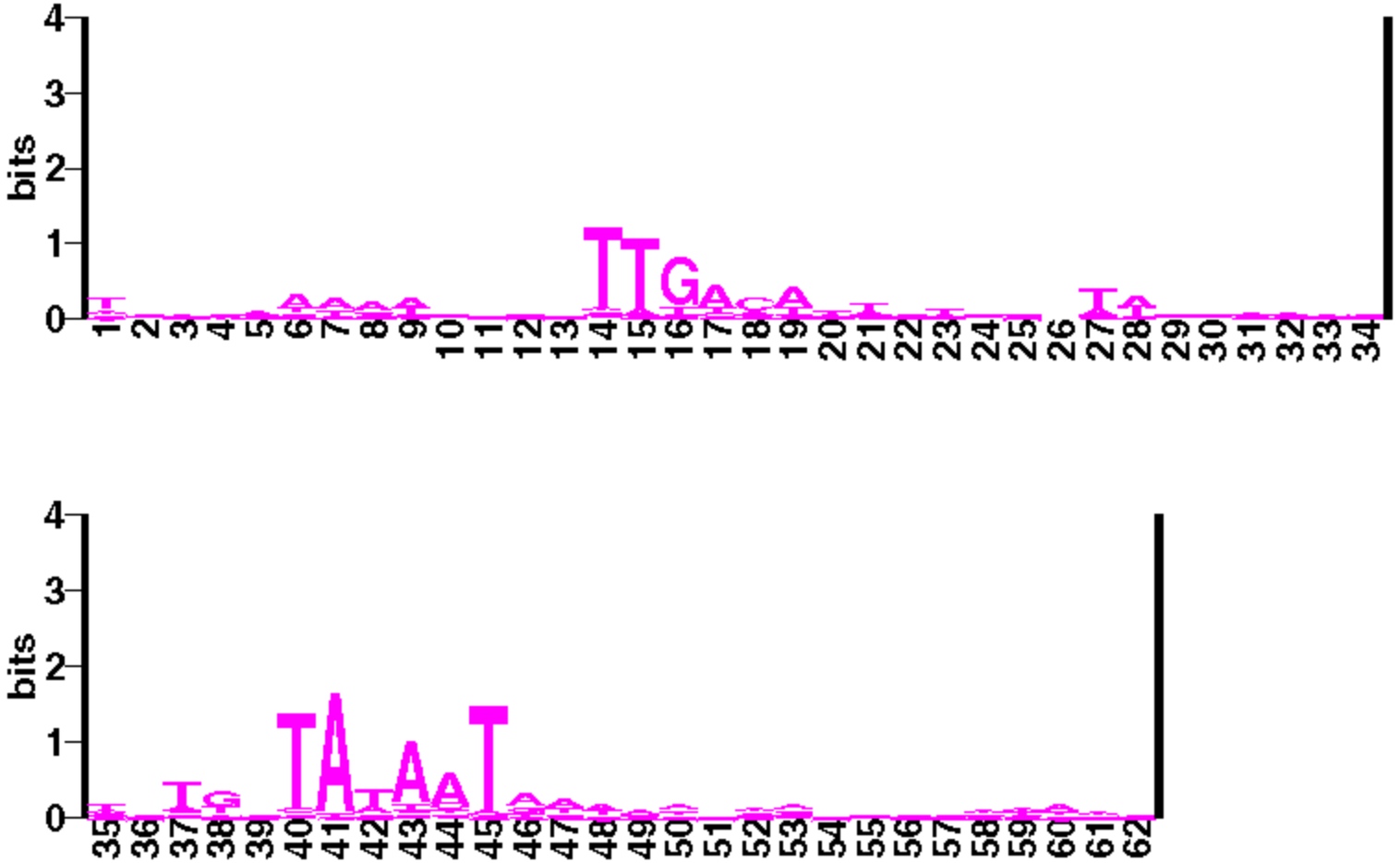
$$I(j) = \sum_i f_{ij} \log_2 f_{ij} + 2$$

Les positions du signal qui contiendront de l'information seront celles qui auront une composition très biaisées par rapport à ce qui est attendu.

Si à une position *j* du signal, présence d'une seule base invariante *i* alors $f_{ij} = 1$ et $\log_2 f_{ij} = 0$ donc $f_{ij} \log_2 f_{ij} = 0$ et les fréquences observées des autres bases sont nulles. On aura

$I(j) = 2$ information maximale

Compilation of Bacillus subtilis sigma A-dependent promoter elements



Recherche des signaux d'initiation de la traduction

Programme utilisé: Scan_For_Matches

Motif du Shine-Dalgarno recherché : **GGAGG 6...11 DTG** correspond à la présence de la séquence GGAGG à 6 ou 11 pb en amont d'un codon AUG, GUG ou UUG.

Résultats:

```
BS:[189,204]   :  ggagg cagataca   atg    ->    A
BS:[3175,3192]:  ggagg tcgacttttt  ttg    ->    dans le gène C
BS:[3887,3902]:  ggagg cataaggt    atg    ->    D
BS:[4760,4775]:  ggagg agaatgtg    atg    ->    E
BS:[7501,7516]:  ggagg atttgccg    gtg    ->    dans le gène F
```

Donc:

Gène A : début en 202
Gène D : début en 3900
Gène E : début en 4773

Les autres SD des gènes B, C et F trouvés avec une matrice de poids car ils sont modifiés

```
202        1194  direct      fr 1    -> A
1335       2600  direct      fr 3    -> B
2640       3896  direct      fr 3    -> C
3900       4751  direct      fr 3    -> D
4773       6833  direct      fr 3    -> E
6877       8202  direct      fr 1    -> F
```

Recherche des unités de transcription

Chez *B. subtilis*, l'initiation de la transcription fait intervenir le facteur sigma A qui reconnaît une séquence spécifique localisée environ en -10 et -35 pb **du +1 de transcription.**

Séquence consensus: TTGACA 16…35 TATAAT

Grand nombre de promoteurs de type sigma A identifiés expérimentalement chez *B. subtilis :*

matrices de poids position (PWM)

Utilisation du programme Scan_For_Matches et de la PWM :

	-35	-10
BS:[1264,1292]:	**tttaca**	**cattttctcaggcatgc tatatt**
BS:[131,158] :	**ttgaca**	**tttgtgaaaaagaaag taaaat**

Liste des logiciels pour la prédiction des promoteurs chez les bactéries

TABLE 1 General information on the tools used here

Tool	Method	Training sequence data set[a]	No. of *E. coli* sigma factors	Availability	Yr	Reference	No. of citations (Google Scholar)[b]
BPROM	Weight matrices of different motifs combined with linear discriminant analysis	Positive: Experimentally validated promoters from *E. coli* (14). Negative: Inner regionsof protein-coding ORFs.	70	Web server http://www.softberry.com/berry.phtml?topic=bprom&group=programs&subgroup=gfindb	2011	33	427
bTSSfinder	Position weight matricesfor promoter elements, oligomer frequencies, physicochemical properties as features, and Mahalanobis distance for feature selection and with neural network for classification	Positive: Experimentally validated TSSs from Regulon DB. [−200, +51]. Negative: Genomic regions withno experimental evidence for the presence of TSSs.	24, 28, 32, 38, 70	Stand-alone and Web server http://www.cbrc.kaust.edu.sa/btssfinder/	2016	23	26
BacPP	Weighted rules extracted from neural network	Positive: Regulon DB available promoters. [−60, +20]. Negative: randomly generated sequences (with established nucleotide frequencies) and intergenic sequences.	24, 28, 32 38, 54, 70	Web server http://www.bacpp.bioinfoucs.com/home	2011	17	22
Virtual Footprint	PWMs from different available databases			Web server http://www.prodoric.de/vfp/vfp_promoter.php	2005	36	370
IBBP	Image-based and evolutionary approach which generates "images" (template-image strings that keep features of spatial sequence relationships)	Positive: sigma 70 promoters from Regulon DB. [−60, +20]. Negative: randomly generated from protein-coding sequences.	70 (expand-able approach)	Source code https://github.com/hahatcdg/IBPP	2018	35	1
iPro70-FMWin	22,595 features extracted from sequence and AdaBoost to select the most representatives among then; logistic regression classifier	Positive: Regulon DB annotated promoters. [−60, 20]. Negative: randomly generated from protein- coding and intergenic region sequences.	70	Webserver http://ipro70.pythonanywhere.com/	2019	38	4
70ProPred	Support vector machine using position-specific tendencies of trinucleotide and electron-ion interaction pseudopotentials as features	Positive: promoters from Regulon DB. [−60, 20]. Negative: randomly generated from coding and noncoding sequences.	70	Webserver http://server.malab.cn/70ProPred/	2017	39	33
CNNProm	Convolutional neural networks	Positive: promoters from Regulon DB. [−60, 20]. Negative: the opposite chain of randomly selected protein-coding genes.	70	http://www.softberry.com/berry.phtml?topic=index&group=programs&subgroup=deeplearn	2017	34	60
MULTiPly	Support vector machine using biprofile Bayes, KNN features, k-tuple nucleotide compositions, and dinucleotide-based auto-covariance as features	Positive: promoters from Regulon DB. [−60, 20].	24, 28, 32, 38, 54, 70	Web server and stand-alone http://flagshipnt.erc.monash.edu/MULTiPly/	2019	41	30
iPromoter-2L	Multiwindow-based pseudo k-tuple nucleotide composition with physicochemical properties as features and Random Forest as a predictor	Positive: promoters from Regulon DB. [−60, 20]. Negative: randomly extracted from the middle regions of long coding sequences and convergent intergenic region (none of the promoters in each set has more than 0.8 pairwise sequence identity)	24, 28, 32. 38, 54, 70	Web server http://bioinformatics.hitsz.edu.cn/iPromoter-2L/	2018	40	180

[a]Positive, positive sequences, sequences expected to be promoters. Negative, negative sequences, sequences expected to not include promoters. The interval of the sequence with the boundary numbers related to a TSS is indicated within brackets ([−60, +20], [−60, +19], or [−200, +51]).
[b]Citations checked on 3 May 2020.

Extrait de : Cassiano MHA, Silva-Rocha R. Benchmarking Bacterial Promoter Prediction Tools: Potentialities and Limitations. mSystems. 2020 Aug 25;5(4):e00439-20. doi: 10.1128/mSystems.00439-20.

Liste des logiciels pour la prédiction des promoteurs chez les bactéries

UNIVERSITÉ TOULOUSE III PAUL SABATIER

TABLE 2 Usage characteristics of the tools analyzed here

Tool	Multifasta	Big files	Shows promoter core	Score or probability	Uppercase only	Output format	Execution time	Follow up	Interface	Comment(s)
BPROM	No	Yes	Yes	Yes	No	Text on screen	Fast	Progress on screen	Webform, simple, intuitive	Multifasta not supported and sequence boxes are not shown; difficult to process the results
bTSSfinder	Yes	Yes	Yes	Yes	No	Text file, GFF file, BED file	Fast	Progress on screen	Login needed, webform, simple, intuitive	Flexible configurations of cutoff values; results saved for 1 week; Linux tool available for download; it needs a large promoter sequence (−200, +50, related to the putative TSS)
BacPP	Yes	No	No	Yes	No	Text on screen or text file	Fast	N	Login needed, webform, simple, intuitive	Short tests per time
Virtual Footprint	No	Yes	Yes	Yes	No	Text on screen	Medium fast	Progress on screen	Webform, many fields, and option in the screen	Integrated with a large PWM database of TFBS; applicable to many species; limited to the position weight matrix available
IBBP	No	Yes	It shows the putative TSS	Yes	No	Text file	Fast	Progress on screen	Command line	Windows SO only execution; requires the manual input files; training and test procedures are separated; fast for big files; it can be used as an approach to the initial prediction of any type of promoters
iPro70-FMWin	Yes	Yes	No	Yes	No	Text on screen	Fast	No	Webform, simple, intuitive	High accuracy
70ProPred	Yes	Yes	No	No	Yes	Text on screen	Fast	No	Webform, simple, intuitive	High accuracy; it does not accept a file as input, just text on a form
CNNProm	Yes	Yes	No	Yes	Yes	Text on screen	Long time (for genomes), fast for multifasta	No	Webform, simple, intuitive	Useful for large sequences (genomes); without a follow-up display; multifasta not supported and sequence boxes are not shown; difficult to process the results
MULTiPly	Yes	Yes	No	No	No	Text on screen or text file	Medium time	Progress on screen, job ID to find the result later	Webform, simple, intuitive	Good accuracy; it saves the result; time-consuming for large sequences
iPromoter-2L	Yes	Yes	No	No	No	Text on screen	Fast	Progress on screen	Webform, simple, intuitive	Good accuracy

Extrait de : Cassiano MHA, Silva-Rocha R. Benchmarking Bacterial Promoter Prediction Tools: Potentialities and Limitations. mSystems. 2020 Aug 25;5(4):e00439-20. doi: 10.1128/mSystems.00439-20.

Recherche des terminateurs de transcription

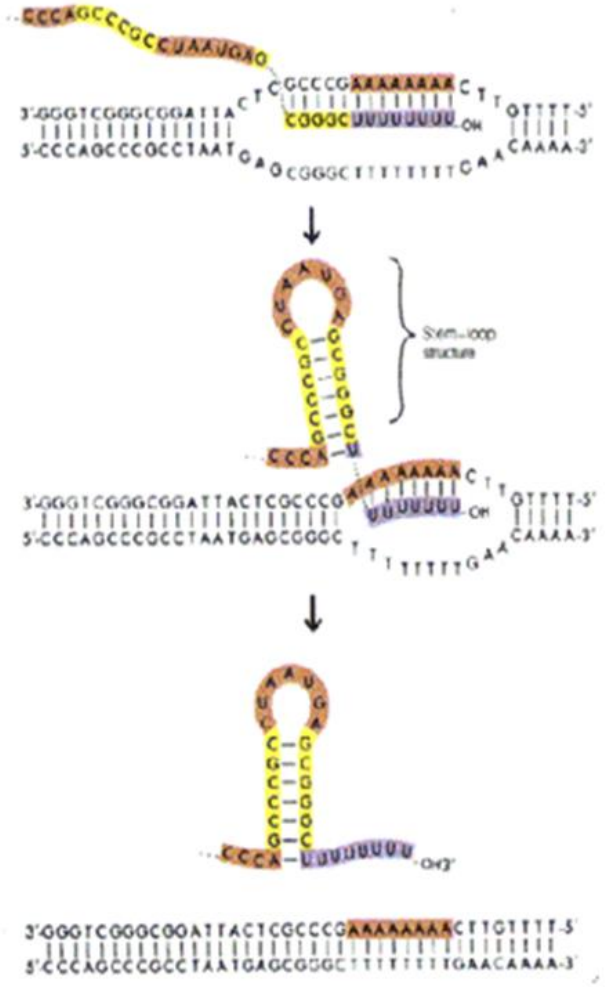
Au niveau séquence, on ne sait modéliser que les terminateurs Rho indépendants.

Mécanisme proposé pour les terminateurs Rho indépendants.

Quand l'ARN est en cours de transcription, on a une hybridation ARN/ADN sur environ 12 pb.

Le site de terminaison de la transcription est précédé par une séquence capable de former une structure secondaire stable. Il y a compétition entre la formation de cette structure et l'appariement avec l'ADN. La présence d'un poly(U) en cours de synthèse déplace l'équilibre en faveur de la tige-boucle et il y a alors décrochage de l'ARN et arrêt de la transcription.

Dans les séquences, on va donc rechercher des séquences répétées inversées suivies d'un poly(U).



Terminateurs rho indépendant

Deux classes de terminateurs:

- petite tige de 5 à 7 pb très stable et d'une boucle de 4 pb suivie d'une région riche en U.

- une longue tige qui peut se décomposer en deux tiges imbriquées l'une dans l'autre.
    - La première plus stable doit faire au moins 3 pb de long avec un appariement GC à son pied.
    - La seconde est incluse dans la première et comporte au moins 3 appariements. Elle est généralement moins stable que la première. La boucle est de 3 à 7 pb de long.

Utilisation de scan_for-matches
Un site éventuel pour prédire les terminateurs de transcription ou au moins vérifier la probabilité de ceux sélectionnés avec scan_for_matches
http://lin-group.cn/server/iTerm-PseKNC/predictor.php

Résultat de la recherche des terminateurs sur le fragment de *B. subtilis*

1199-1223

```
  A C
G   A
  G-C
  T-A
  T-A
  C-G
  A-T
  G-C
  T-A
    T
    T
    T
    T
    T
    T
```

6843-6866

```
    T
A     A
  G.T
  C-G
  T.G
  C-G
  G-C
  C-G
  C-G
    A
    T
    T
    C
    T
    T
    T
```

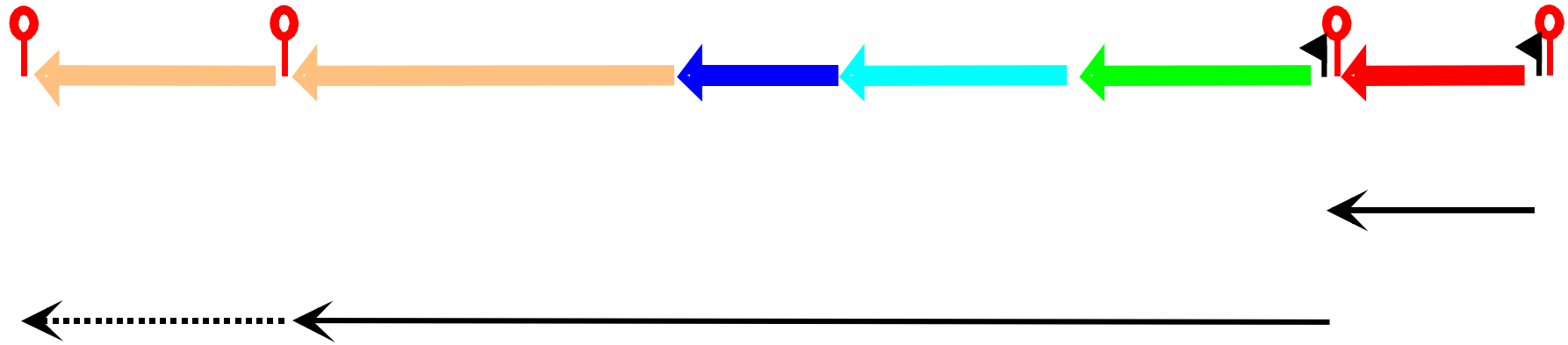
75-103

```
  A A
C     A
  G.T
  C-G
  C-G
  G.T
  A-T
  G-C
  T-A
  C-G
  A-T
  G-C
    T
    T
    T
    T
    T
```

8215-8256

```
    T C
A       A
  A-T
  A-T
  C-G
  A-T
  A-T
  T-A
  G.T
  T.G
  A-T
  G-C
  A-T
  G-C
  T-A
  A-T
  C-G
  C-G
    ATCATTT
```

Prédiction des unités de traduction et de transcription



terminateur rho-indépendant

promoteurs de transcription de type sigma

transcrit putatif

Prédictions fonctionnelles

Identification
- homologues
- motifs
- domaines

Localisation cellulaire
- fragments trans-membranaires
- peptide signal

Structure
- secondaire
- tertiaire

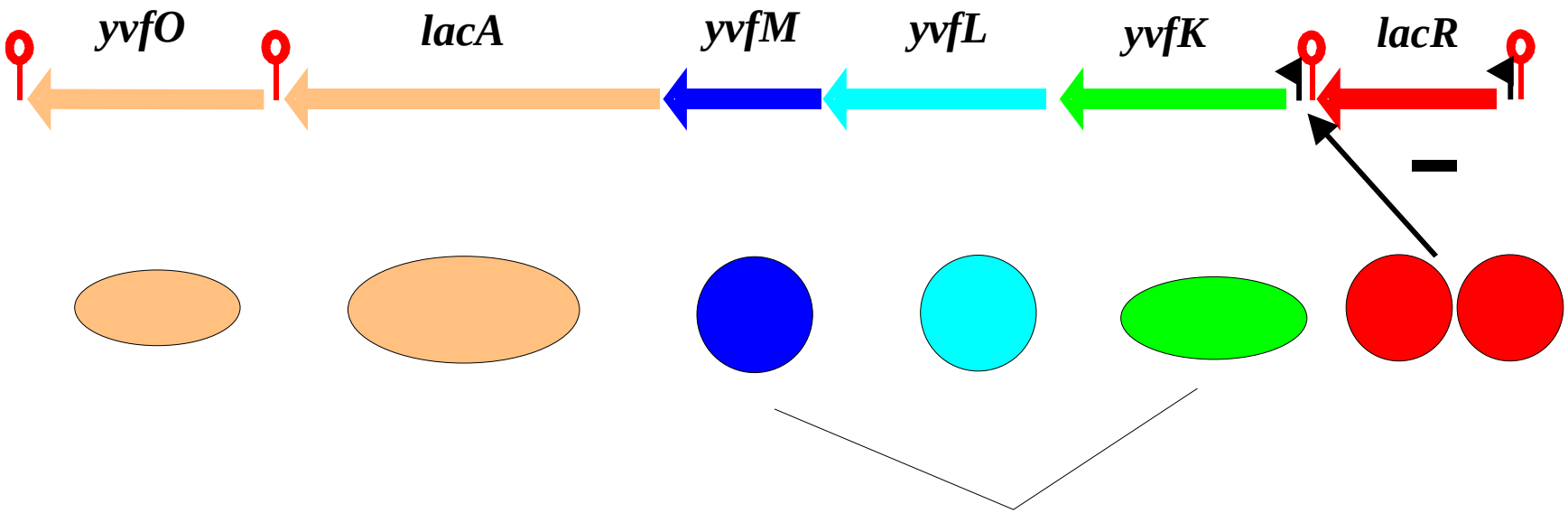
Recherche de liens fonctionnelles
- réseaux de régulation
- voies métaboliques
- interactions moléculaires

Prédictions fonctionnelles

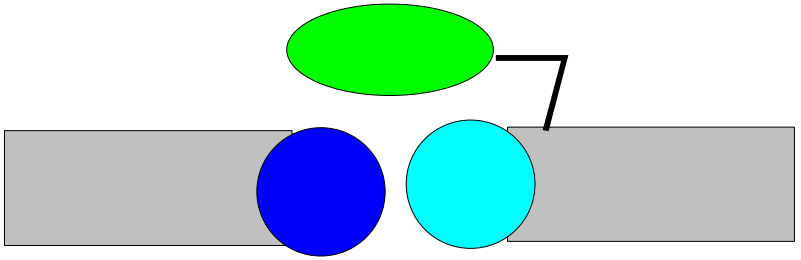
Recherche par similitude dans les bases de données: programme BLAST

- A LACR protéine régulatrice de type LacI/GalR
- B YVFK protéine affine d'un ABC transporteur
- C YVFL perméase d'un ABC transporteur
- D YVFM perméase d'un ABC transporteur
- E LACA galactosidase
- F YVFO arabino-galactosidase

Synthèse des résultats



Système à la membrane



? ?

Environnements intégrés pour l'annotation des génomes procaryotes

✓ **RAST server** (BMC Genomics. 2008 Feb 8;9:75. doi: 10.1186/1471-2164-9-75)

DESCRIPTION:

We describe a fully automated service for annotating bacterial and archaeal genomes. The service identifies protein-encoding, rRNA and tRNA genes, assigns functions to the genes, predicts which subsystems are represented in the genome, uses this information to reconstruct the metabolic network and makes the output easily downloadable for the user (https://rast.nmpdr.org/)

✓ **Prokka** (Bioinformatics. 2014, 30:2068-69)

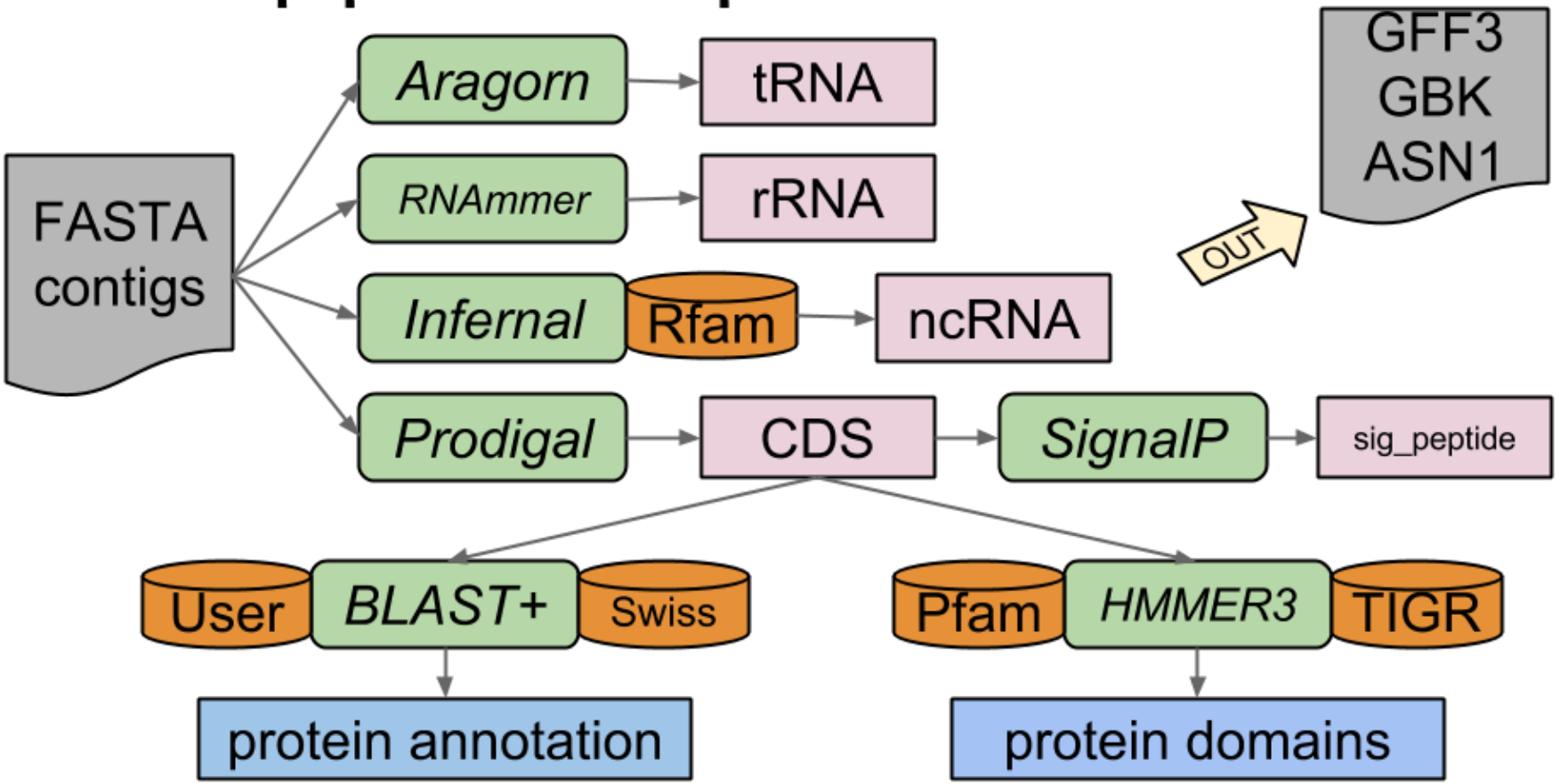
DESCRIPTION:

Prokka, a command line software tool to fully annotate a draft bacterial genome in about 10min on a typical desktop computer. It produces standards-compliant output files for further analysis or viewing in genome browsers.

Tool (reference)	Features predicted
Prodigal (Hyatt 2010)	Coding sequence (CDS)
RNAmmer (Lagesen *et al.*, 2007)	Ribosomal RNA genes (rRNA)
Aragorn (Laslett and Canback, 2004)	Transfer RNA genes
SignalP (Petersen *et al.*, 2011)	Signal leader peptides
Infernal (Kolbe and Eddy, 2011)	Non-coding RNA

Méthodes utilisées dans Prokka (issu de Bioinformatics. 2014, 30:2068-69)

Prokka pipeline (simplified)



Environnements intégrés pour l'annotation des génomes procaryotes

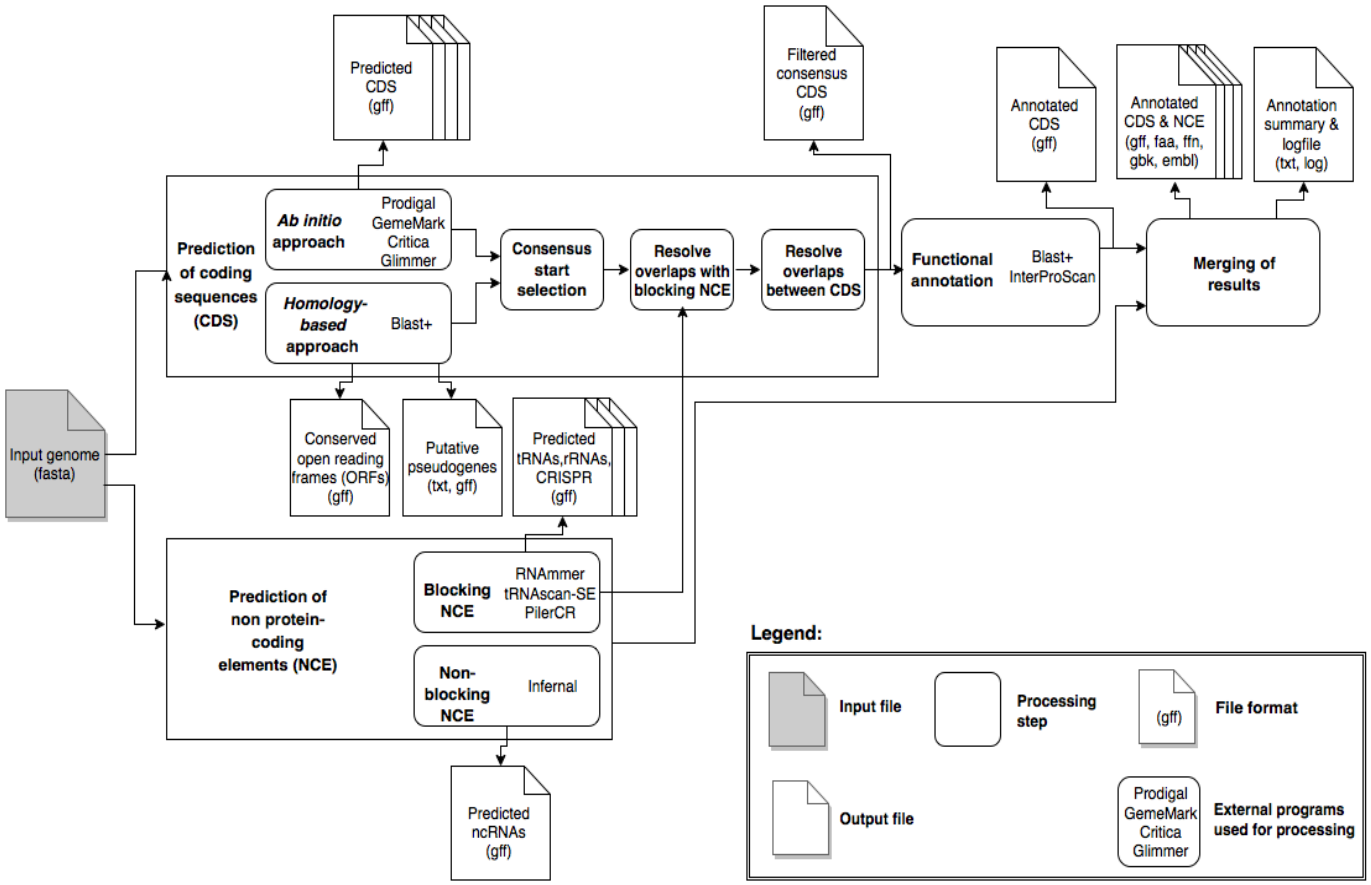
UNIVERSITÉ TOULOUSE III PAUL SABATIER

✓ **ConsPred** (Bioinformatics. 2016, 32:3327-29)

DESCRIPTION:

We present ConsPred, a prokaryotic genome annotation framework that performs intrinsic gene predictions, homology searches, predictions of non-coding genes as well as CRISPR repeats and integrates all evidence into a consensus annotation. ConsPred achieves comprehensive, high-quality annotations based on rules and priorities, similar to decision-making in manual curation and avoids conflicting predictions. Parameters controlling the annotation process are configurable by the user.

Environnements intégrés pour l'annotation des génomes procaryotes
Exemple de ConsPred (Weinmaier *et al.*, 2016, Bioinformatics, 32:3327-29)



(Extrait de Weinmaier *et al.*)

Figure S1. ConsPred workflow

Coding sequences (CDS) are predicted by combining different *ab initio* gene predictions, and conserved open reading frames (ORFs) detected by homology search against the NCBI nr database. Database entries from closely related taxa are excluded to prevent possible misannotations due to low phylogenetic distance. Putative pseudogenes are exported for user inspection. From all predicted non-protein-coding elements (NCE) those that biologically must not overlap with CDS are considered blocking NCE. CDS overlapping with blocking NCE are removed. Filtered consensus CDS are obtained from predicted CDS and conserved ORFs by using predefined weights and rules and subsequent removal of CDS that overlap with blocking NCEs. Filtered consensus CDS are functionally annotated and then merged with the NCE into the final annotation files.

Mesure du pouvoir prédictif d'une méthode

4 paramètres importants :

- pourcentage de vrais positifs (VP, True positive)
- pourcentage de faux positifs (FP, False positive)
- pourcentage de vrais négatifs (VN, True negative)
- pourcentage de faux négatifs (FN, False negative)

		Réalité	
		Groupe 1	Groupe 2
prédiction	Groupe 1	% vrais positifs	% faux positifs
	Groupe 2	% faux négatifs	% vrais négatifs

Groupe 1 : exemples
Groupe 2 : contre-exemples

Mesure du pouvoir prédictif d'une méthode

Idéal: prédire le maximum d'exemples (max VP) avec un minimum d'erreurs (min FP). Mais valeurs non indépendantes donc impossible.
Solution un compromis:

- on maximise le % de VP (donc minimise le % de FN) souvent par utilisation de critères moins stricts même si cela entraîne l'augmentation du % de FP. L'élimination des FP se fait par un autre traitement informatique ultérieur. On dit que l'on privilégie la **sensibilité** de la méthode

- inversement, on minimise le % de FP même si cela conduit à ne pas détecter certaines séquences d'intérêts (donc plus grand % de FN). On dit que l'on privilégie la **spécificité** de la méthode.

Sensibilité = VP/(VP+FN) sensibility en anglais

Spécificité = VP/(VP+FP) specificity en anglais (ou VN/VN+FP 2 définitions)

Précision = (VP+VN)/(VP+VN+FP+FN) accuracy en anglais

Mesure du pouvoir prédictif d'une méthode

Table 2. Gene prediction accuracy of the GeneMark.hmm program using the protein-coding model derived by GeneMarkS as the Typical gene model and a heuristic model as the Atypical gene model, the model for non-coding sequence is also heuristically derived

	Genes annotated	Genes detected[a]	Gene detection accuracy	
			Sn (%)	Sp (%)
A.fulgidus	2406	2583	98.5	91.8
B.subtilis	4099	4445	98.8	91.1
E.coli	4288	4397	96.9	94.5
H.influenzae	1708	1807	98.2	92.8
H.pylori	1552	1753	97.7	86.5
M.jannaschii	1714	1891	99.4	90.1
M.thermoautotrophicum	1868	1935	97.9	94.5
Synechocystis	3168	3521	98.7	88.8
Average			98.3	91.3

[a]Number of predictions that match the 3′ end of GenBank annotated genes, with possible misplacement of the 5′ end, as a percentage of the number of annotated genes.

Tableau extrait de *Nucleic Acids Research*, 2001, vol. 29, **2607-2618**