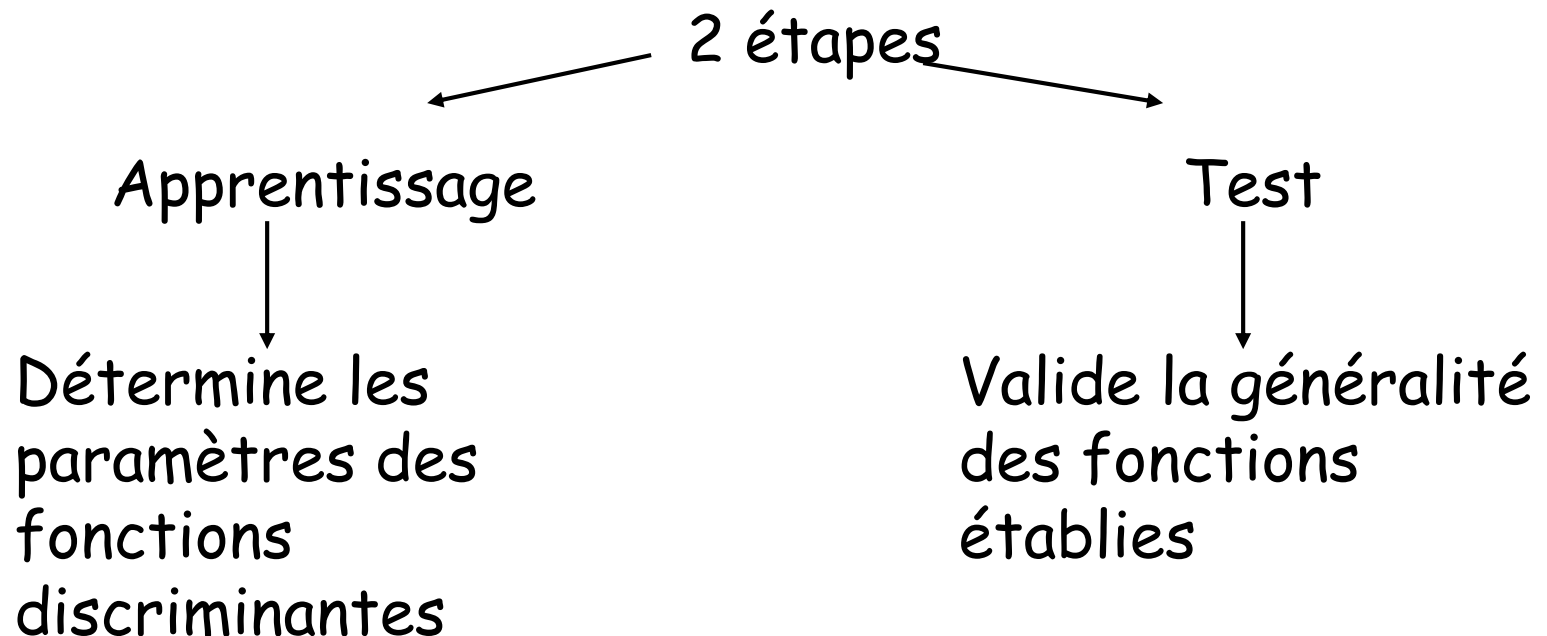


Support de cours
Annotation des génomes

Méthodes de prédiction: démarche générale

- Définir clairement l'objectif.
- Choisir les critères.
- Choisir le type d'approche :
 - sans système de référence,
 - avec système de référence.



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 an anglais

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

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

Annotation d'un génome

Identification des gènes codant pour :

- les ARNr
- les ARNt
- les protéines

Identification des unités de transcription (promoteur et terminateur)

Identification des unités de traduction

Pour les gènes codant pour les protéines, prédiction fonctionnelle par recherche de similarité de séquences (Blast) et classification en grandes classes fonctionnelles (ex: biosynthèse des acides aminés, métabolisme énergétique....)

Exemple d'annotation d'un génome *Mycoplasma genitalium*



Recherche des régions codant pour des protéines chez les procaryotes

- recherche des ORFs (Open reading frame)
- recherche des unités de traduction. 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)

La transcription

3 étapes :

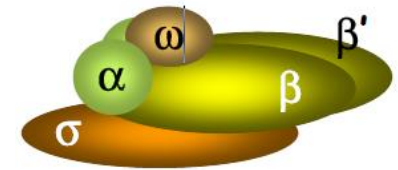
- l'initiation : reconnaissance de séquences spécifiques sur l'ADN : le promoteur
- l'élongation
- la terminaison

Transcription des gènes codant pour des protéines

Chez les bactéries :

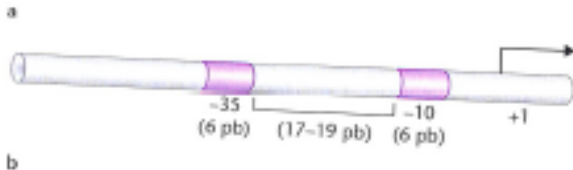
L'initiation :

Le facteur σ de la RNA polymérase reconnaît 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 gènes).

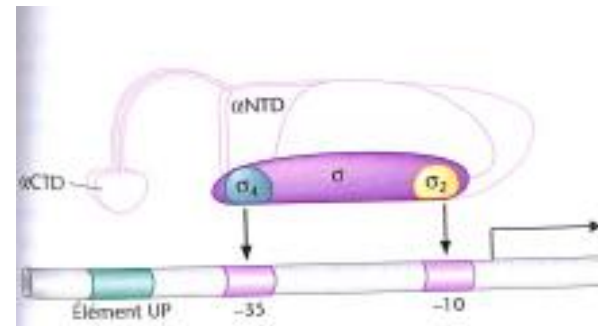


ARN polymérase procaryote

Promoteurs bactériens



Recrutement du cœur de l'ARN polymérase au promoteur

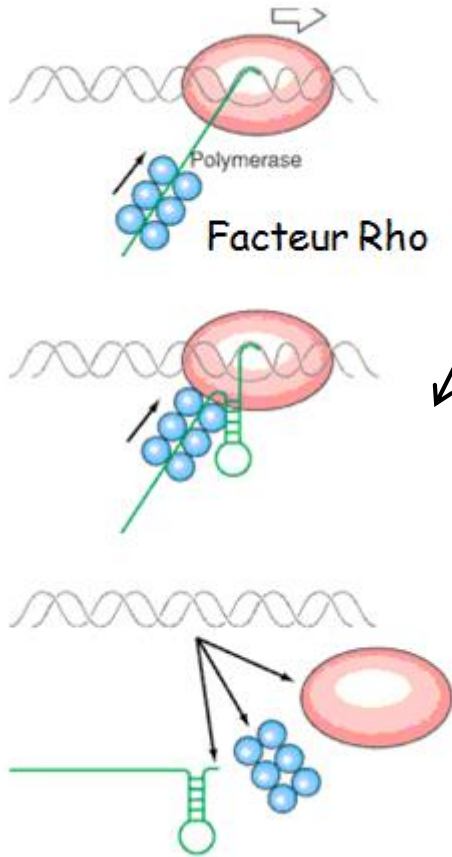


La transcription

Transcription des gènes codant pour des protéines

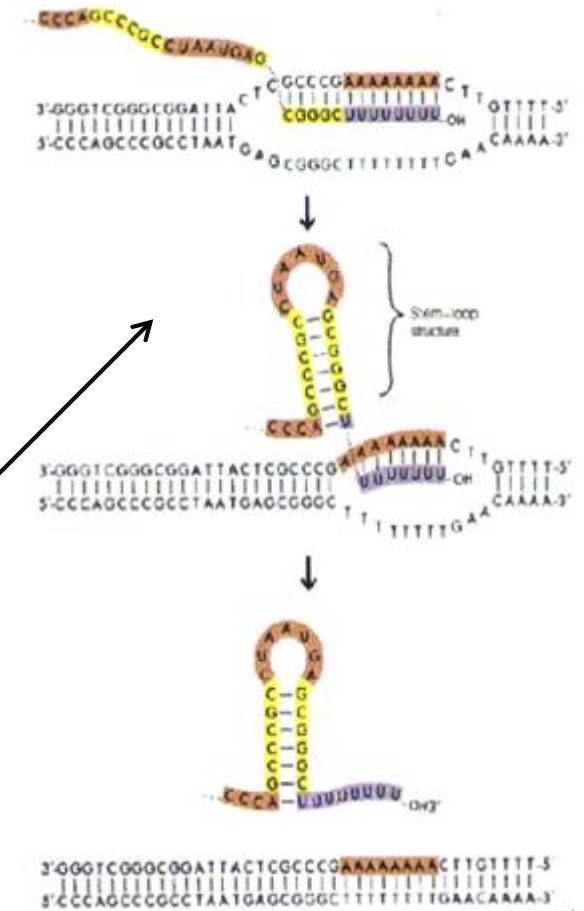
Chez les bactéries :

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.



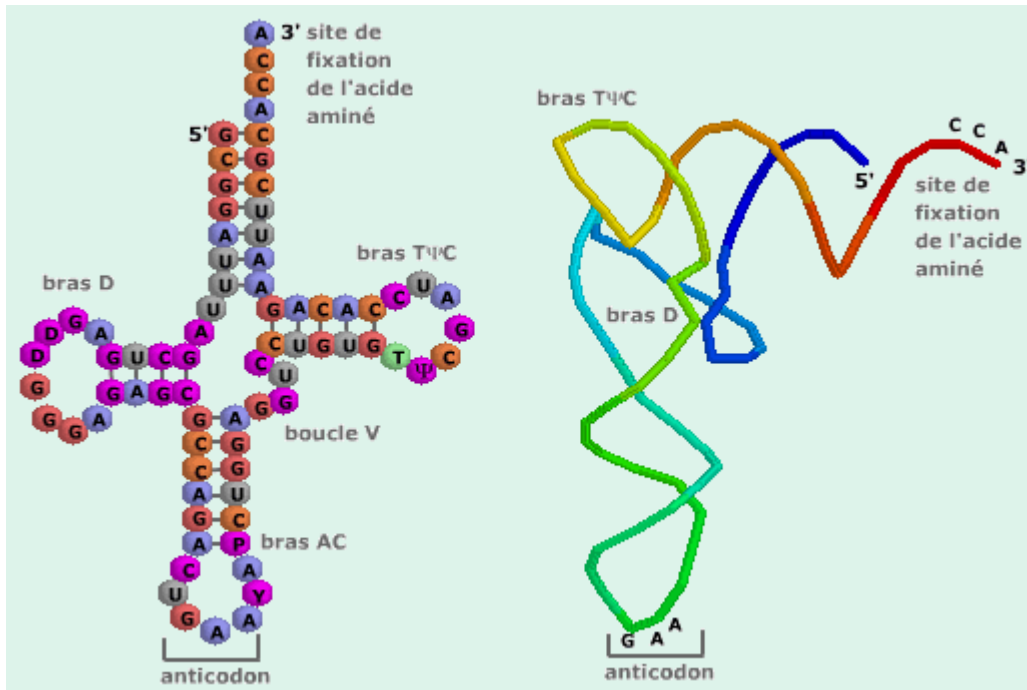
La traduction

4 composants principaux font partie de la machinerie de traduction :

- l'ARN messenger (ARNm) contenant la séquence codant la protéine ainsi que des régions de reconnaissance pour l'initiation et la terminaison de la traduction.
- les ARN de transfert (ARNt) qui une fois chargé de leur acide aminé (processus réalisé par les aminocyl-ARNt- synthétases) vont mettre en relation un codon et un acide aminé.
- les aminocyl-ARNt- synthétases permettant donc de charger l'acide aminé sur l'ARNt
- le ribosome : gros complexe formé de deux sous-unités (une grande et une petite) chacune constituées par des protéines ribosomiques et un ou plusieurs ARN ribosomique (ARNr).

La traduction : les ARNt

Ce sont de petites molécules d'ARN, d'une taille variant entre 70 et 100 nucléotides. Ils se replient en une structure secondaire caractéristique appelée feuille de trèfle et présente une structure tridimensionnelle en forme de L.



L'anticodon s'apparie avec le codon de l'ARNm. L'ARNt chargé portant à son extrémité 3' l'acide aminé correspondant permet donc de faire la correspondance codon-acide aminé. Après reconnaissance du codon, l'acide aminé est transféré à la chaîne peptidique en croissance. Ce processus est réalisé à l'intérieur du ribosome.

Structures secondaire et tertiaire d'un ARNt

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

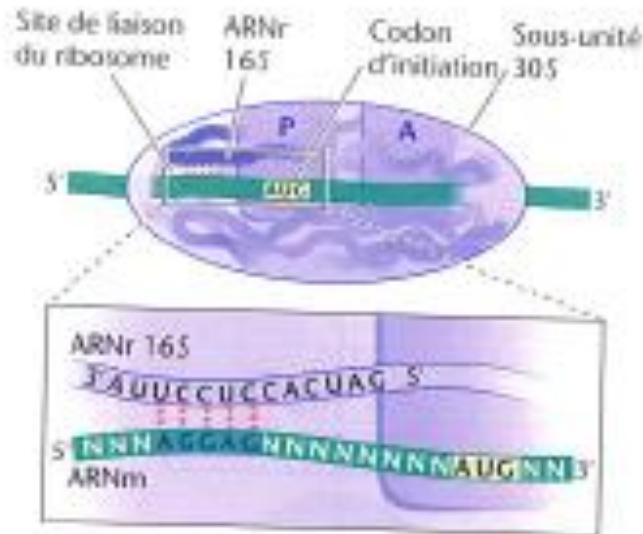
La traduction

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'apparie 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.

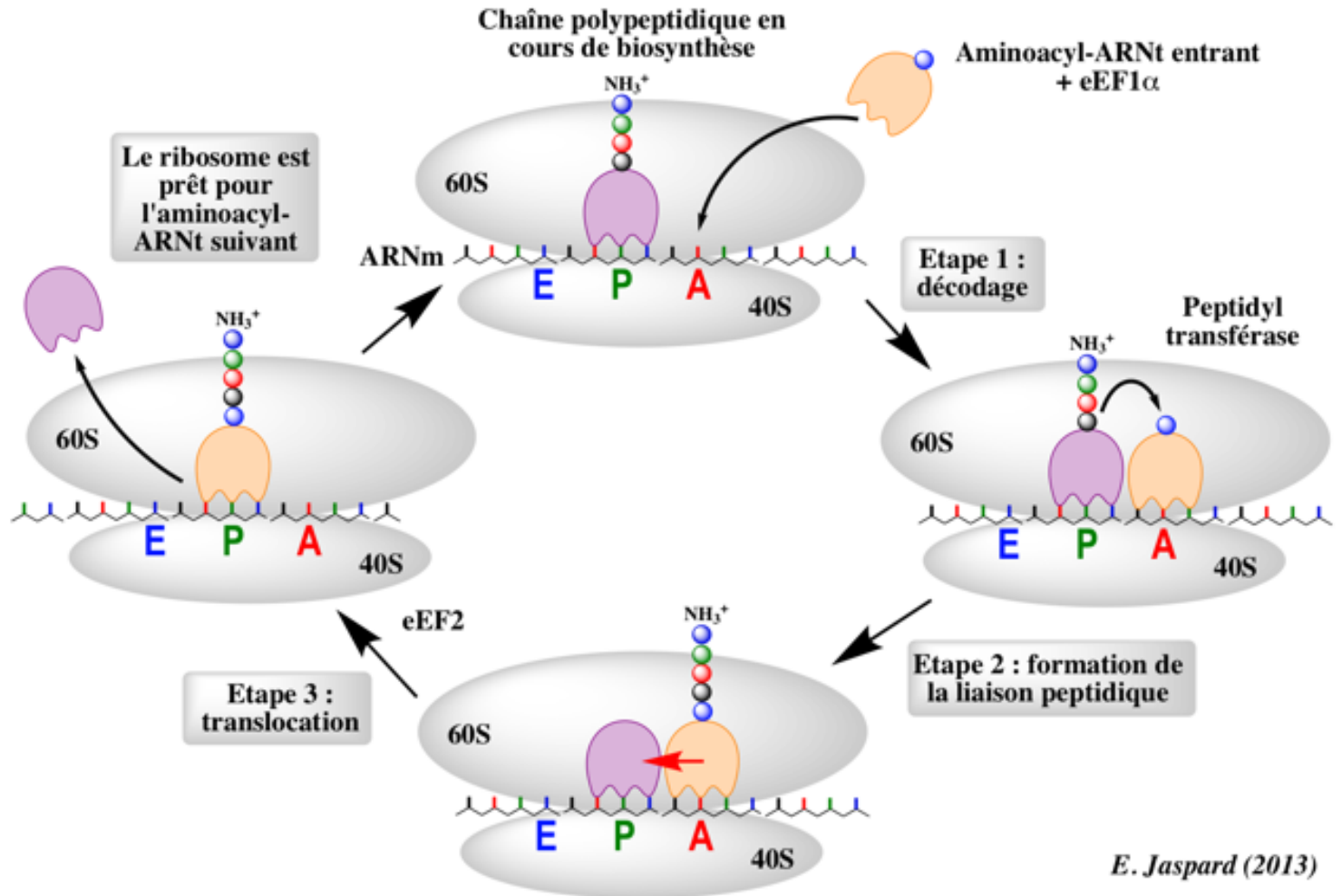


Interaction RBS et ARNr 16S pour positionnement du codon AUG au site P

RBS = Ribosome Binding Site

La traduction

L'elongation :



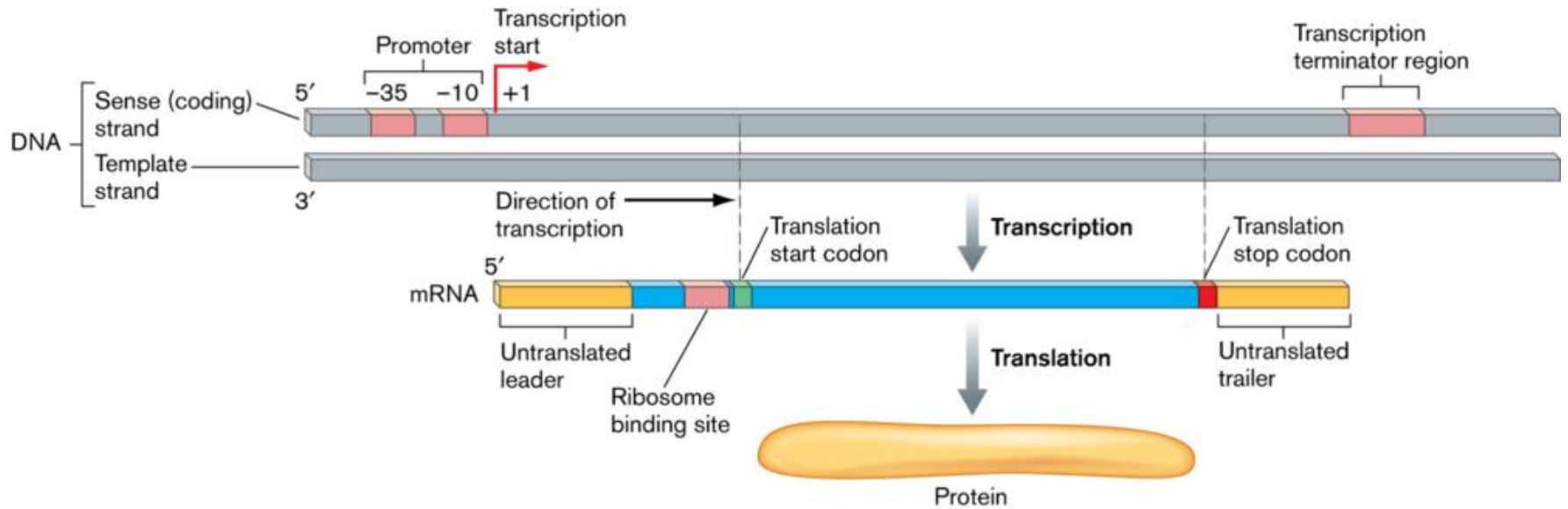
La traduction

La terminaison :

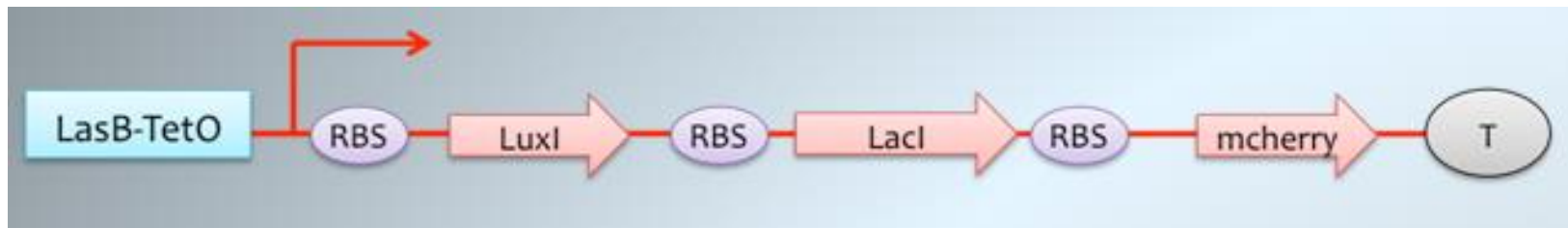
La traduction s'arrête quand le ribosome rencontre un codon stop au site A.

Ce codon stop est reconnu par un des deux facteurs de terminaison de classe 1 chez les procaryotes. Ce facteur stimule l'hydrolyse du polypeptide et du peptidyl-ARNt, libérant ainsi le peptide complet.

Structure d'un gène bactérien

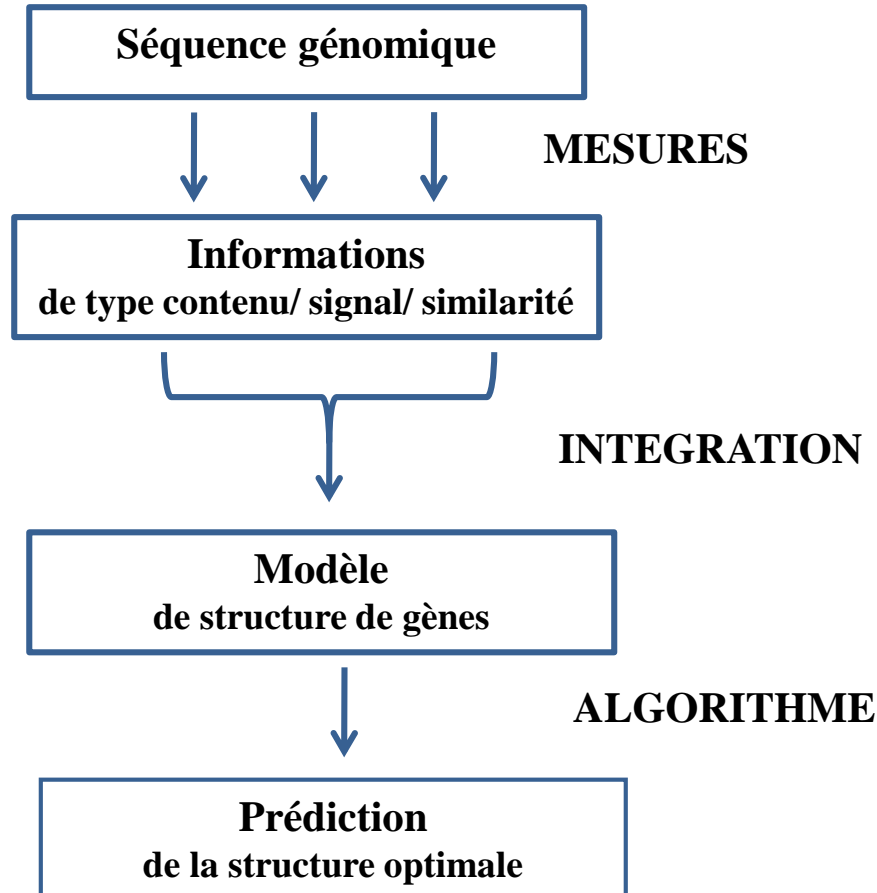


Les gènes peuvent être co-transcrits. Ils sont organisés en unité de transcription, appelée opéron entre un promoteur et un terminateur de transcription.



Recherche des régions codant pour des protéines

Fonctionnement schématique d'un logiciel de prédiction de gènes




Une méthode simple: ORFfinder (NCBI)

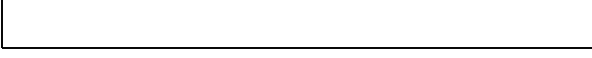
Recherche les phases ouvertes de lecture, les ORFs, dans les 6 cadres de lecture (les 3 cadres du brin direct et les 3 cadres du brin complémentaire).

Attention problème de sémantique :

Alors qu'une ORF est normalement définie entre deux codons stop

stop **XXXXXXXXXXXXXXXXXXXXX** stop

n codons

Dans ORF Finder, elle est définie entre un codon start et un codon stop

ATG **XXXXXXXXXXXXXXXXXXXXX** 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, donc lors de l'annotation d'un génome, dans un premier temps on ne retient que les ORFs de taille supérieure ou égale à 100 codons).

>BS 1-8301

tttcgaggaaaatgtgcaataaccaactcatttcccgggcaattccgccc
gttccgaatgatacgaacaactgagactgagccgcaaattgggttcagtctt
tttacatggcagccagagggctttgtgcaacttgacatttgtgaaaaagaa
agtaaaatattttactaaaacaatgcgagctgaataatggaggcagatac
aatggcgacaattaaagatatcgcgaggaagcgggattttcaatctcaa
ccgtttcccgcggttttaaaataacgatgaaagcctttctgttcctgatgag
acacgggagaaaatctatgaagcggcggaaaagctcaattaccgcaaaaa
aacagtaaggccgctgggtgaaacatattgcggtttttatattggctgacag
ataaagaagaattagaagatgtctattttaaaacgatgagattagaagta
gagaaactggcgaaagcattcaatgtcgatatgaccactataaaatagc
ggatggaatcgagagcattcctgaacatacgggaagggtttattgccgtcg
gcacattttcagatgaagagctggctttcctcagaaatctcactgaaaac
ggcgtgttcacgattcaactcctgatcccgatcattttgactcggtaag
gcccgatttggcacaatgacaaggaagacggtaaacatcctgactgaga
aggggcataagagcatcgggttttatcggcggcacatacaaaaatccgaat
accaatcaggatgaaatggacatccgtgaacaaaccttcagatcctatat
gagggaaaaagccatgctggacgagcgtatatattttctgtcatcgcggat
tctctgtagaaaacggctaccgcctgatgtcagcagcgcgatcgacacatta
ggcgatcagcttccgactgcttttatgattgcagcggaccgattgcagt
gggctgtctgcaagccctgaacgaaaaaggaattgccatacacaacaggg
taagcattgtgagtatcaacaacatcagcttcgcgaagtatgtctgcct
cctctgacgacgtttcatattgatatacatgaattatgtaaaaacgctgt
tcaattactgcttgaacaagtgcaggacaagagaagaacggtaaaaacat
tatatgtgggcgagaattaatcgtcaggaagagtatgaattaaggatga
cttaggacactaagtcatTTTTTTTtaggtaaaaaaatttactctatga
agtaaatagtttgtttacacattttctcaggcatgctatattatctttaa
agcgctttcattcctaccgaaagggtgacaatcaatgaaaatggcaaaaa
agtgttccgtattcatgctctgcgcagctgtcagtttatccttggcggct
tgcggcccaaaggaaagcagcagcgccaaatcgagttcaaaaagggtcaga
gcttgttgtatgggaggataaagaaaagagcaacggcattaagacgctg
tggctgcatttgaaaaagagcatgatgtgaaggtcaaagtcgttgaaaa
ccgtatgccaaagcagattgaagatttgcgaaatggatggaccggccggcac
aggccctgacgtgttaacaatgccaggggaccaaatcggaaccgctgtca
cggaaggattactcaaggaattacatgtcaaaaaagacgttcaatcactt
tatactgacgcttccattcagctctcaaatggtagatcaaaaagctttatgg
actgccaaaagcggctcgaaacgactgtgcttttttacaacaaagatctca
tcacagaaaaggaattgcccaaaaacgctggaagagtgggtacgactattcc

**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
tttcgaggaaaatgtgcaataaccaactcatttccgggcaattccgcccgttccgaatg
atcgaacaactgagactgagccgcaaatggttcagtccttttacatggcagccagaggg
ctttgtgcacttgacatttggaaaaagaaagtaaaatcttactaaaacaatgcgagc
tgaataatggaggcagatacaatggcgacaattaaagatatcgcgaggaagcgggattt
tcaatctcaaccgtttcccgcttttaataacgatgaaagcctttctgttctgatgag
acacgggagaaaatctatgaagcggcgaaaagctcaattaccgcaaaaaaacagtaagg
ccgtggtgaaacatatgctgttttatattggctgacagataaagaattagaagat
gtctattttaaaccgatgagattagaagttagagaaactggcgaagcattcaatgtcgat
atgaccacttataaaatagcggatggaatcgagagcattcctgaacatacggagggtt
attgccgtcggcacatttcagatgaagagctggcttctcagaaatctcactgaaaac
```

From: To:

Choose Search Parameters

Minimal ORF length (nt):

Genetic code:

ORF start codon to use:

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

Ignore nested ORFs:

Start Search / Clear



Résultat de ORFfinder : ORFs de plus de 300 pb

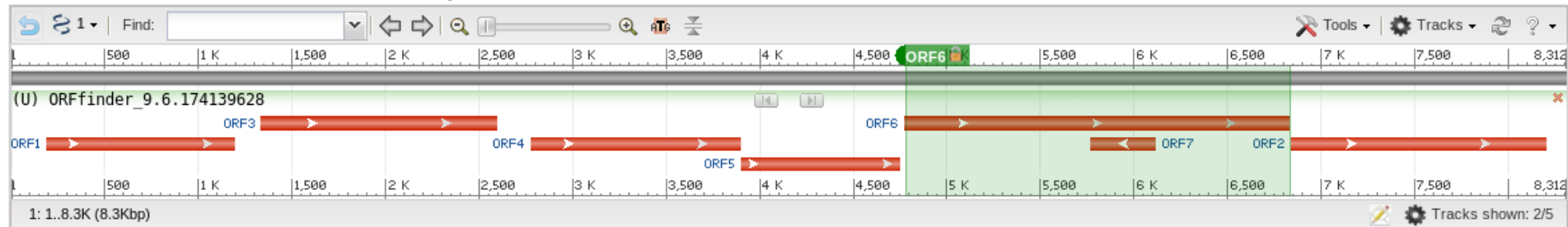
- Options : - ATG only
 - Ignore nested ORF pas coché

Open Reading Frame Viewer

Help

Sequence

ORFs found: 7 Genetic code: 1 Start codon: 'ATG' only



Six-frame translation...

ORF6 (686 aa)

Display ORF as...

Mark

Mark subset...

Marked: 0

Download marked set

as

Protein FASTA

```
>lcl|ORF6
MSKLEKTHVTKAKFMLHGGDYNPDQWLDLDRPDILADDIKMLKLSHTNTFSV
GIFAWSALEPEEGVYQFEWLDDIFERIHSIGGRVILATPSGARPAWLSQT
YPEVLRVNASRVKQLHGGRHNCCLTSKVYREKTRHINRLLAERYGHPAL
LMWHISNEYGGDCHCDLQAHAFREWLKSKYDNSLKTLMHAWWTPFWSHTF
NDWSQIESPSPIGENGLHGLNLDWRRFVTDQTSIFYENEIIPLKELTPDI
PITTFMADTPDLIPYQGLDYSKFAKHVDAISWDAYPVVHNDWESTADLA
MKVGFINDLYRSLKQPFLLMECTPSAVNWHNVNKAARPGMNLSSMQMI
AHGSDSVLYFQYRKSRSSEKLGAVVDHNSPKNRVFEVAKVGETLER
LSEVVGTKRPAQTAILYDWHENHWALEDAQGFATKRYPTLQQHRTFW
EHDIPVDVITKEQDFSPYKLLIVPMLYLISEDVSRKKAFTADGGTLVMT
YISGVVNEHDLTYTGGWHPDLQAFVGEPLETDTLYPKDRNAVSYRSQIY
EMKDYATVIDVKTASVEAVYQEDFYARTPAVTSHEYQQGKAYFIGARLED
QFQDFYEGRLITDLSLSPVFPVRRHGKGVSVQARQDQDNDYIFVMNFTEEK
QLVTFDQSVKDIINTGDILSGDLTMEKYEVRIVVNTH
```

Label	Strand	Frame	Start	Stop	Length (nt aa)
ORF6	+	3	4773	6833	2061 686
ORF2	+	1	6838	8202	1365 454
ORF3	+	3	1335	2600	1266 421
ORF4	+	3	2778	3896	1119 372
ORF1	+	1	187	1194	1008 335
ORF5	+	3	3900	4751	852 283
ORF7	-	3	6117	5770	348 115

ORF6

Marked set (0)

SmartBLAST

SmartBLAST best hit titles...

BLAST

BLAST

Résultat de ORFfinder : ORFs de plus de 300 pb

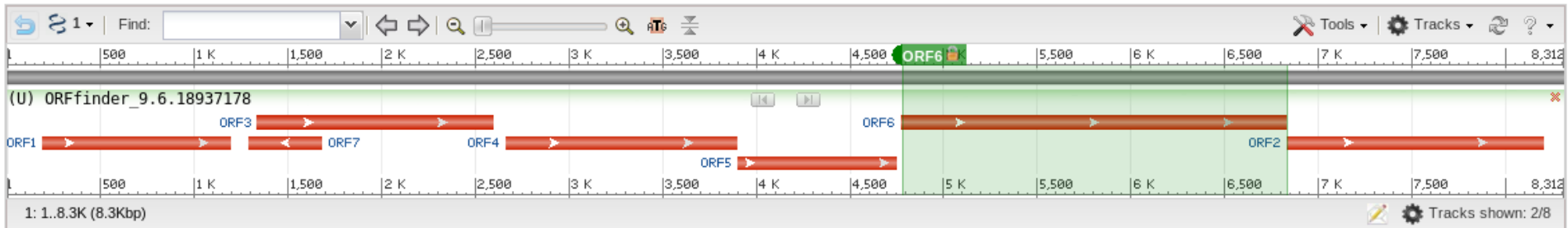
Options : - ATG and alternative initiation codons
 - Ignore nested ORF coché

Open Reading Frame Viewer

Help

Sequence

ORFs found: 7 Genetic code: 1 Start codon: 'ATG' and alternative codons Nested ORFs removed



Six-frame translation...

ORF6 (686 aa)

Display ORF as...

Mark

Mark subset...

Marked: 0

Download marked set

as Protein FASTA

```
>lcl|ORF6
MSKLEKTHVTKAKFMLHGGDYNPDQWLDLRDPDILADDIKMLKLSHTNTFSV
GIFAWSALEPEEGVYQFEWLDIFERIHSIGGRVILATPSGARPALWSQT
YPEVLRVNASRVKQLHGGRHNHCLTSKVYREKTRHINRLLAERYGHHPAL
LMWHISNEYGGDCHCDLQAHAFREWLKSKYDNSLKTLNHAWWTPFWSHTF
NDWSQIESPSPIGENGLHGLNLDWRRFVTDQTIISFYENEIIPLKELTPDI
PITTFMADTPDLIPYQGLDYSKFAKHVDAISWDAYPVWHNDWESTADLA
MKVGFINDLYRSLKQPFLLMECTPSAVNWHNVNKAKRPGMNLSSMQMI
AHGSDSVLYFYQRKSRGSEKLGAVVDHNSPKNRVQEVAKVGETLER
LSEVVGTRPAQTALYDWHENWHALEDAOGFAKATKRYPTLQOHYRTFW
EHDIPVDVITKEQDFSPYKLLIVPMLYLISETVSRKKAFTADGGTLVMT
YISGVVNEHDLTYTGGWHPDLQAI FGVEPLETDTLYPKDRNAVSYRSQIY
EMKDYATVIDVKTASVEAVYQEDFYARTPAVTSHEYQQGKAYF IGARLED
QQRDFYEGLIITDLSLSPVFPVRHGGKGVSVQARQDQDNDYIFVMNFTEEK
QLVTFDQSVKDIMTGDILSGDLTMEKYEVRIVVNTH
```

Label	Strand	Frame	Start	Stop	Length (nt aa)
ORF6	+	3	4773	6833	2061 686
ORF2	+	1	6835	8202	1368 455
ORF3	+	3	1335	2600	1266 421
ORF4	+	3	2661	3896	1236 411
ORF1	+	1	187	1194	1008 335
ORF5	+	3	3900	4751	852 283
ORF7	-	2	1681	1286	396 131

ORF6

Marked set (0)

SmartBLAST

SmartBLAST best hit titles...

BLAST

BLAST

Résultat de ORFfinder : ORFs de plus de 150 pb

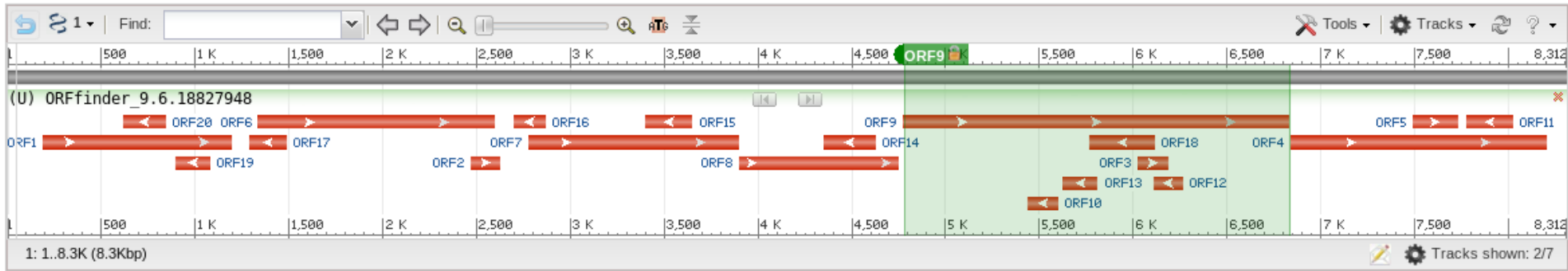
- Options : - ATG and alternative initiation codons
 - Ignore nested ORF coché

Open Reading Frame Viewer

Help

Sequence

ORFs found: 20 Genetic code: 1 Start codon: 'ATG' only



ORF9 (686 aa)

Display ORF as...

Mark

Mark subset...

Marked: 0

Download marked set

as Protein FASTA

```
>lcl |ORF9
MSKLEKTHVTKAKFMLHGGDYNPDQWLDRPDILADDIKLMLSHTNTFSV
GIFAWSALEPEEGVYQFEWLDDIFERIHSIGGRVILATPSGARPAWLSQT
YPEVLRVNASRVKQLHGGRHNCSTSKVYREKTRHINRLLAERYGHHPAL
LMWHISNEYGGDCHCDLCOHAFREWLKSKYDNSLKTLNHAWWTPFWSHTF
NDWSQIESPPIGENGLHGLNLDWRRFVTDQTIISFYENEIIPKELTPDI
PITTNFMADTPDLIPYQGLDYSKFAKHVDAISWDAYPVVHNDWESTADLA
MKVGFINDLYRSLKQPFLLMECTPSAVNWHNVNKA R PGMNLLSSMQMI
AHGSDSVLYFQYRKS RGSSEKHLHGAVVDHNSPKNRV FQEVAKVGETLER
LSEVVGTKRPAQTAILYDWHENHWALEDAQGFAKATKRYPQTLQQHYRTFW
EHDIPVDVITKEQDFSPYKLLIVPMLYLISEDVSRKAFADGGTLVMT
YISGVVNEHDLTYTGGWHPDLQAI FGVPELETDTLYPKDRNAVSYRSQIY
EMKYATVIDVKTASVEAVYQEDFYARTPAVTSHEYQQGKAYFIGARLED
QFQRDFYEGLITDLSLSPVFPVRHKGVS VQARQDQNDYIFVMNFTEEK
QLVTFDQSVKDIMTGDILSGDLTMEKYEVRIVVNTH
```

Label	Strand	Frame	Start	Stop	Length (nt aa)
ORF9	+	3	4773	6833	2061 686
ORF4	+	1	6838	8202	1365 454
ORF6	+	3	1335	2600	1266 421
ORF7	+	3	2778	3896	1119 372
ORF1	+	1	187	1194	1008 335
ORF8	+	3	3900	4751	852 283
ORF18	-	3	6117	5770	348 115
ORF14	-	2	4627	4349	279 92
ORF15	-	2	3652	3401	252 83
ORF11	-	2	8026	7775	252 83

ORF9

Marked set (0)

SmartBLAST

SmartBLAST best hit titles... ?

BLAST

BLAST

Six-frame translation...

Les codons initiateurs alternatifs chez les procaryotes sont GTG et TTG (chez *B. subtilis* **GTG** 13%, TTG 9%)

Limites d'ORFfinder :

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

Traitement de l'information de type contenu

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 ou de la région génomique (isochores chez les vertébrés) (riche ou pauvre en C+G)
 - 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).

Exemples d'usage des codons chez les procaryotes

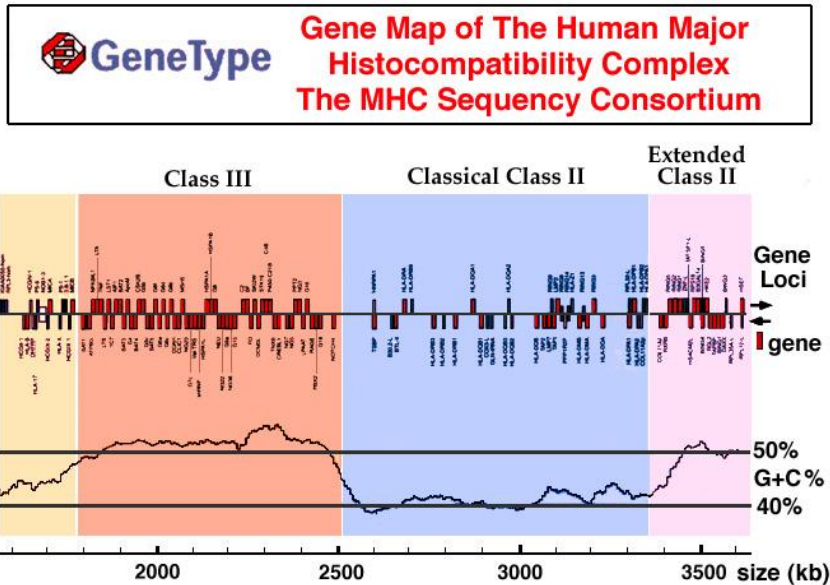
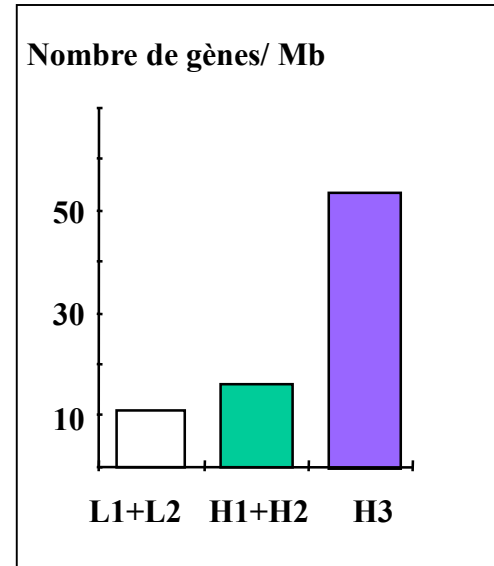
Espèce	GC% codant	GC% 1 ^{ère} pos. codon	GC% 2 ^{ème} pos. codon	GC% 3 ^{ème} pos. codon
<i>Synechocystis sp.</i>	48.25	55.82	39.74	49.19
<i>Streptomyces coelicolor</i>	72.30	72.67	51.39	92.83
<i>Escherichia coli</i> 0157:H7	51.54	58.44	41	55.17
<i>Bacillus subtilis</i>	44.36	52.10	36.08	44.91

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

Modèle de la structure en isochores chez les vertébrés

Isochores : régions > 300 kb homogène dans sa composition en bases
 5 types d'isochores en fonction de leur pourcentage en G+C (2 légers et 3 lourds)

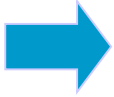
Isochore	%C+G	% total genomic DNA
L1+L2	33%-44%	62 %
H1+H2	44%-51%	31%
H3	51%-60%	3-5%



MHC locus (3.6 Mb) (The MHC sequencing consortium 99)

- Class I, class II (H1-H2 isochores): 20 gènes/Mb, beaucoup de pseudogènes
- Class III (H3 isochore): 84 gènes/Mb, pas de pseudogène

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 suivantes seront abordées :

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

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:

- des bases précédant cette position
- de sa localisation dans le codon

Modélisé par

modèle de Markov homogène pour les régions non-codantes.

modèle de Markov non-homogène pour les séquences codantes.

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).

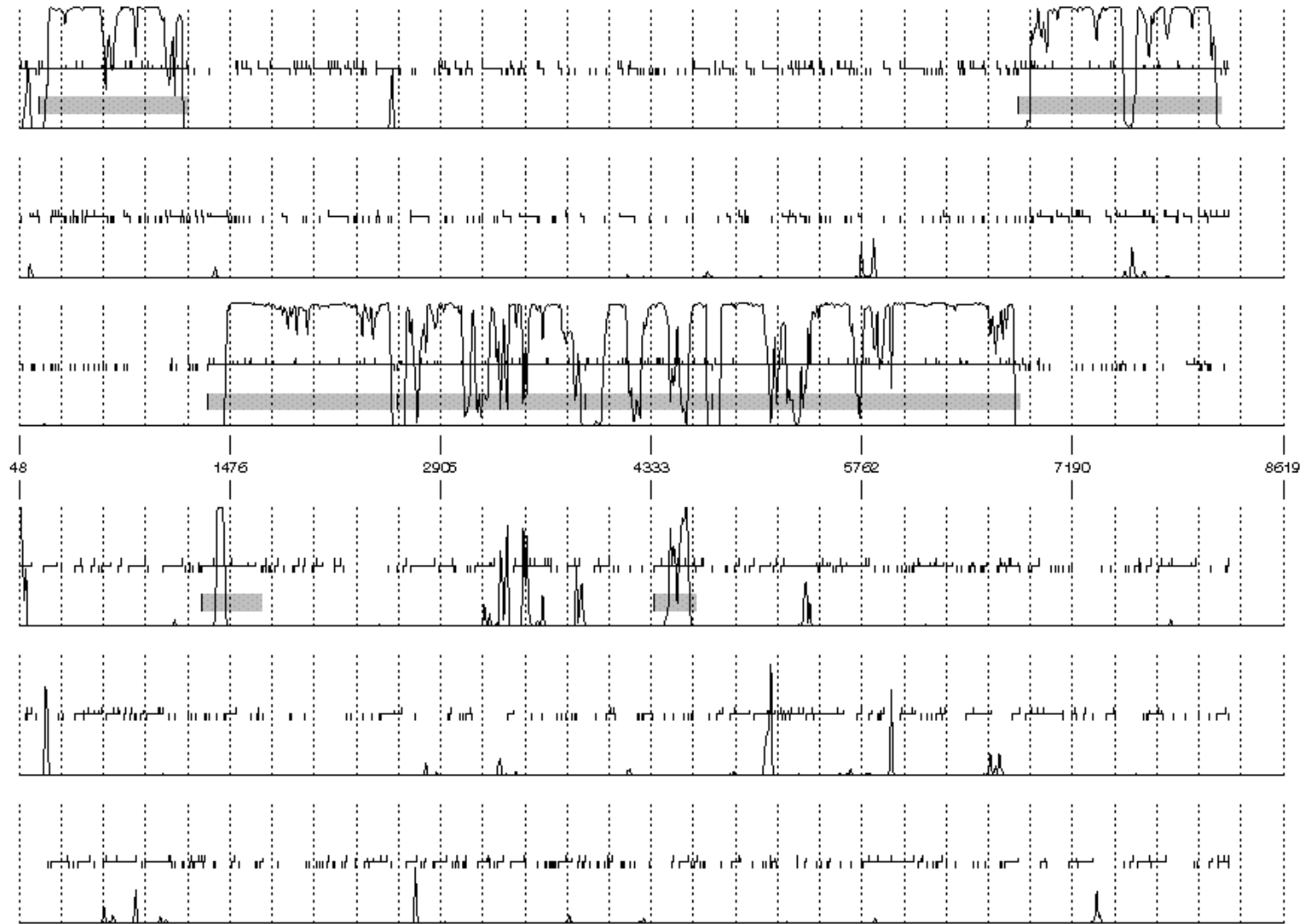
Modèle de Markov

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

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

Modèle probabiliste qui représente une séquence comme un processus qui peut être décrit comme une séquence de variable aléatoire X_1, X_2, \dots où X_i correspond à la position i de la séquence. Chaque variable aléatoire X_i prend une valeur dans l'ensemble des bases (A,C,G,T). La probabilité que va prendre la variable X_i dépend du contexte c'est à dire des bases immédiatement adjacentes à la base à la position i

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



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

List of Open reading frames predicted as CDSs, shown with alternate starts

(regions from start to stop codon w/ coding function >0.50)

Left end	Right end	DNA Strand	Coding Frame	Avg Prob	Start Prob	
-----	-----	-----	-----	-----	-----	-----
187	1194	direct	fr 1	0.80	0.99	-> ORF Finder
202	1194	direct	fr 1	0.81	0.89	
367	1194	direct	fr 1	0.82	0.29	
436	1194	direct	fr 1	0.81	0.03	
481	1194	direct	fr 1	0.80	0.02	
1335	2600	direct	fr 3	0.85	0.01	-> ORF Finder
1341	2600	direct	fr 3	0.85	0.00	
1365	2600	direct	fr 3	0.87	0.08	
1500	2600	direct	fr 3	0.93	0.07	
1527	2600	direct	fr 3	0.93	0.00	
1581	2600	direct	fr 3	0.92	0.03	
2631	3896	direct	fr 3	0.73	0.67	
2640	3896	direct	fr 3	0.73	0.77	
2778	3896	direct	fr 3	0.76	0.53	-> ORF Finder
2814	3896	direct	fr 3	0.75	0.02	
2868	3896	direct	fr 3	0.74	0.40	
3900	4751	direct	fr 3	0.65	0.17	-> ORF Finder
3912	4751	direct	fr 3	0.66	0.02	
3966	4751	direct	fr 3	0.71	0.34	
4116	4751	direct	fr 3	0.71	0.11	
4137	4751	direct	fr 3	0.70	0.02	
4158	4751	direct	fr 3	0.69	0.06	
4770	6833	direct	fr 3	0.85	0.76	
4773	6833	direct	fr 3	0.85	0.82	-> ORF Finder
4815	6833	direct	fr 3	0.86	0.12	
4890	6833	direct	fr 3	0.85	0.05	
5226	6833	direct	fr 3	0.85	0.01	
6838	8202	direct	fr 1	0.79	0.03	-> ORF Finder
6877	8202	direct	fr 1	0.82	0.86	
6913	8202	direct	fr 1	0.83	0.67	
6925	8202	direct	fr 1	0.83	0.01	
6931	8202	direct	fr 1	0.83	0.01	
6952	8202	direct	fr 1	0.84	0.00	
7009	8202	direct	fr 1	0.85	0.63	
7057	8202	direct	fr 1	0.86	0.28	

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

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 probabilité

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

Pour certains kmers rares même avec un grand jeu d'apprentissage comme un génome entier, il peut être difficile d'obtenir des estimations précises et inversement certains kmers fréquents même avec un modèle de markov d'ordre élevé des estimations précises peuvent être obtenues.

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 un exon, un intron ou à une région intergénique.
- 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.

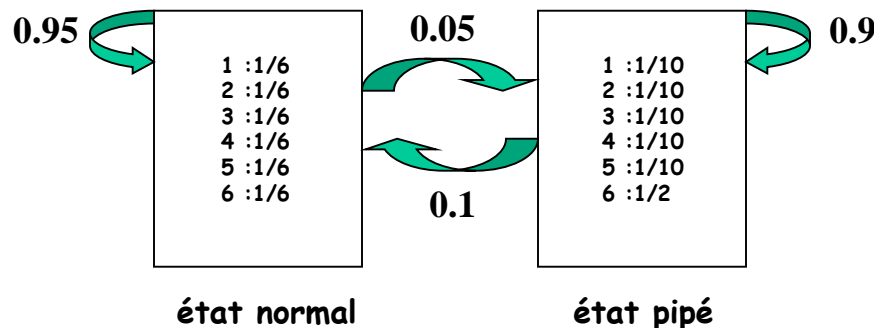
La probabilité de changer du dé normal au dé pipé avant chaque jet est de 0.05, et celle de passer du dé pipé au dé normal est de 0.1.

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.

Les GHMM : Generalized Hidden Markov Model

Les HMM vu précédemment ne permettent pas de prendre en compte la longueur des régions à identifier (CDS, exons, introns etc.).

L'idée est que la probabilité d'observation d'une suite de nucléotides d'un même état n'est pas uniquement le produit des probabilités associées à chaque nucléotides. Elle dépend aussi de la longueur de cette suite, c'est-à-dire que la probabilité qu'un segment de séquence soit dans un état donné est aussi fonction de sa longueur.

Pour pouvoir prendre cela en compte, il faut recourir à des GHMM.

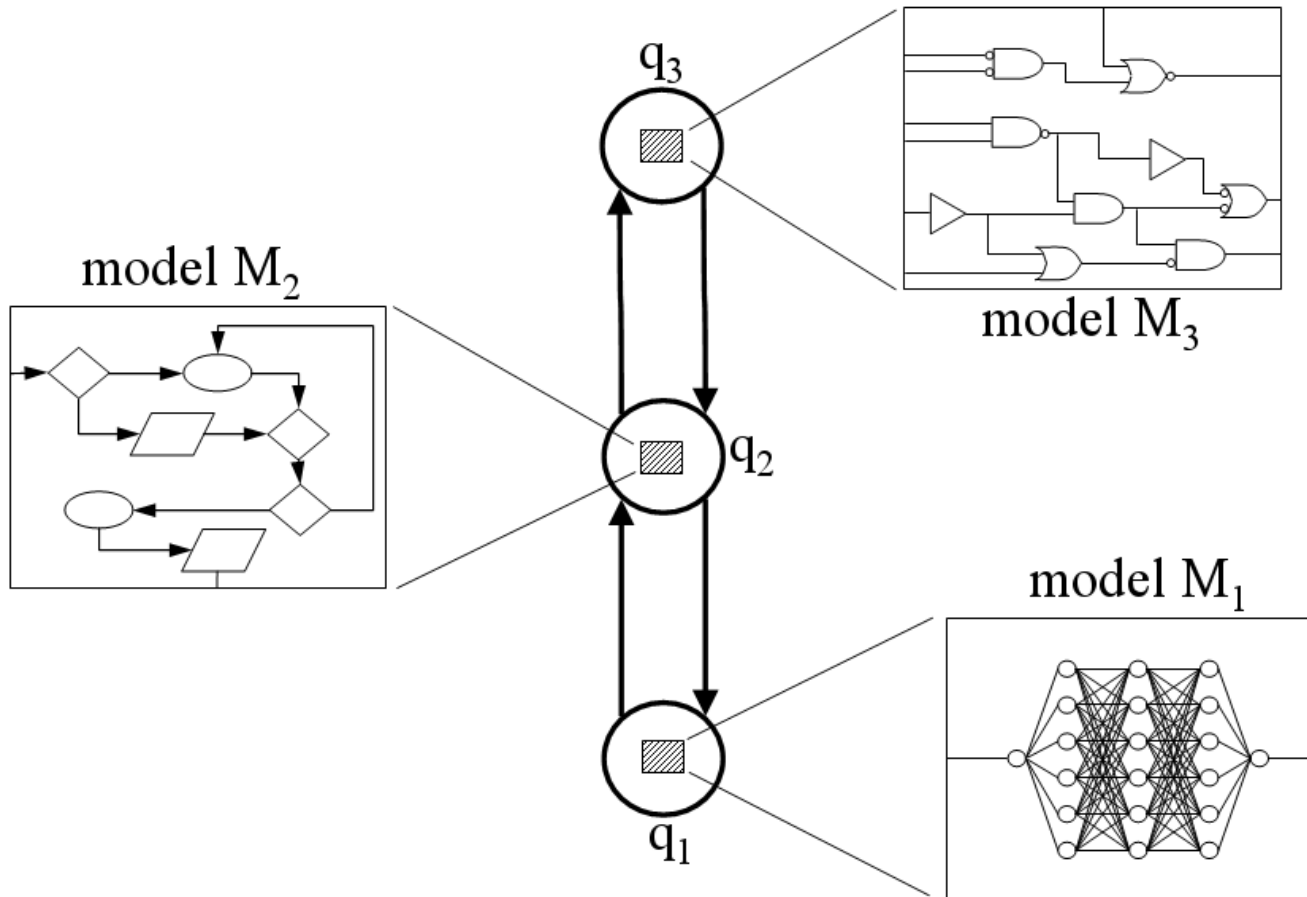
Un GHMM est défini par :

- un ensemble fini d'états $Q = \{q_0, q_1, \dots, q_m\}$
- Un vecteur de probabilités initiales $\Pi = (\pi_i)$
- un vecteur de probabilités de transition $A = (a_{ij})$
- une matrice de probabilités d'émission
- une distribution des longueurs (duration)

GHMM et HMM : principales différences

- Chaque état émet maintenant une sous séquence et non plus juste un symbole
- Les longueurs sont maintenant explicitement modélisées
- Les probabilités d'émission peuvent maintenant être obtenues par n'importe quel modèle probabiliste (des modèles différents peuvent être utilisés suivant l'état à identifier)
- Cela tend à réduire le nombre d'état, donc plus simple et plus facile à modifier

GHMM



Avantages :

- * abstraction des sous-modèles
- * architecture plus simple
- * modélisation de la durée des états

Désavantage :

- * complexité du décodage

modèle de genMark.hmm : « GHMM »

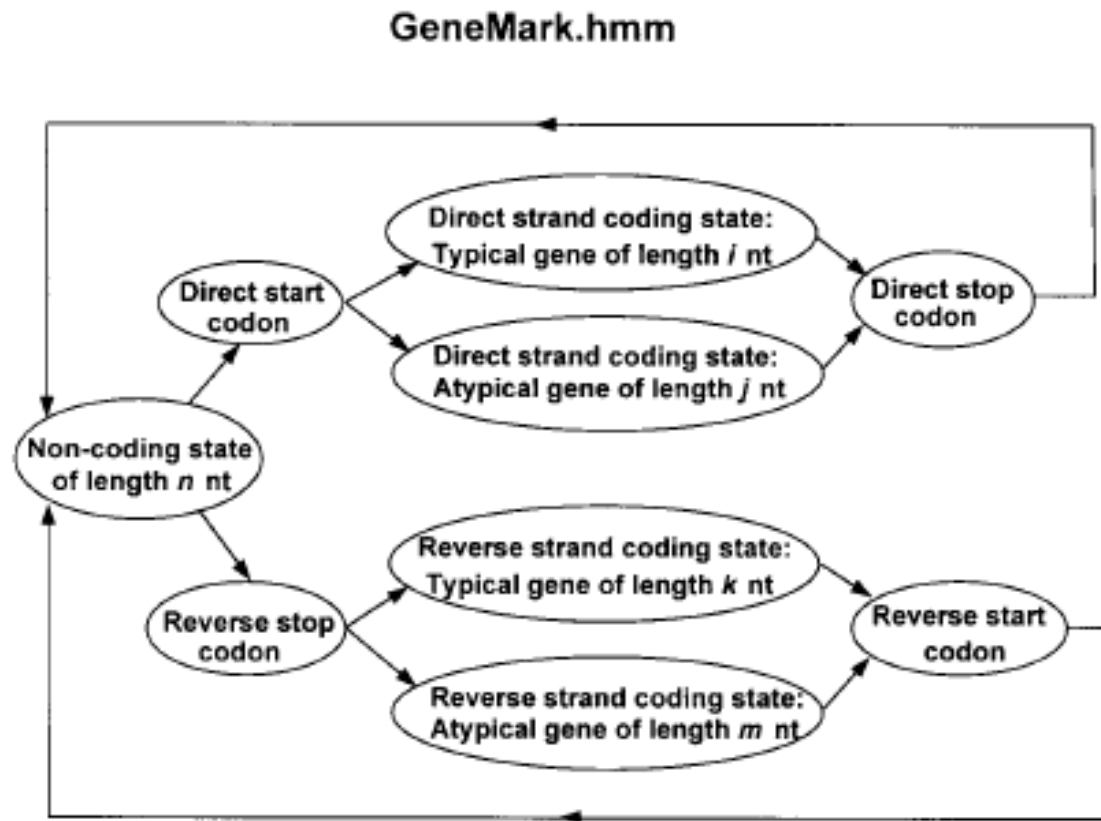


Figure 1. Hidden Markov model of a prokaryotic nucleotide sequence used in the GeneMark.hmm algorithm. The hidden states of the model are represented as ovals in the figure, and arrows correspond to allowed transitions between the states.

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

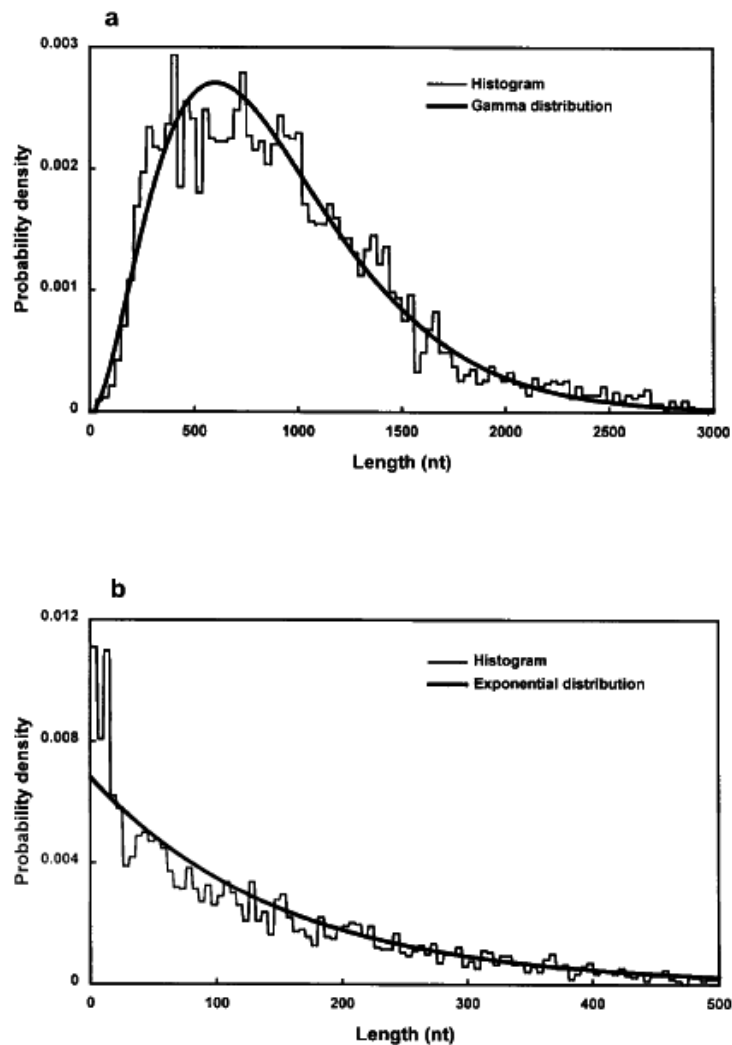
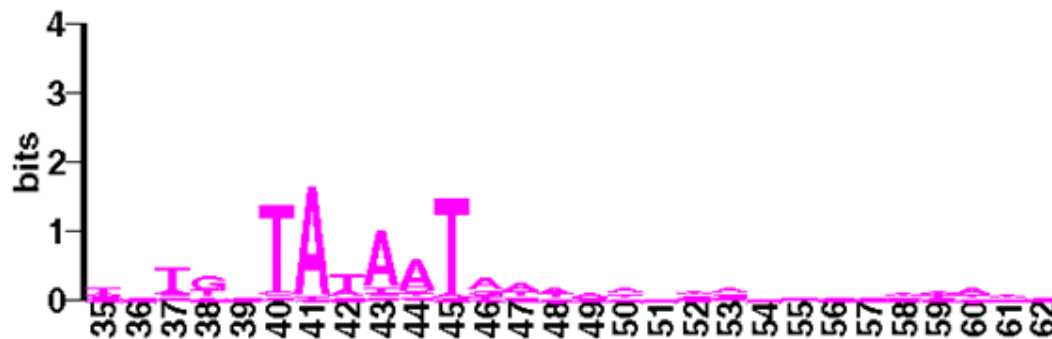
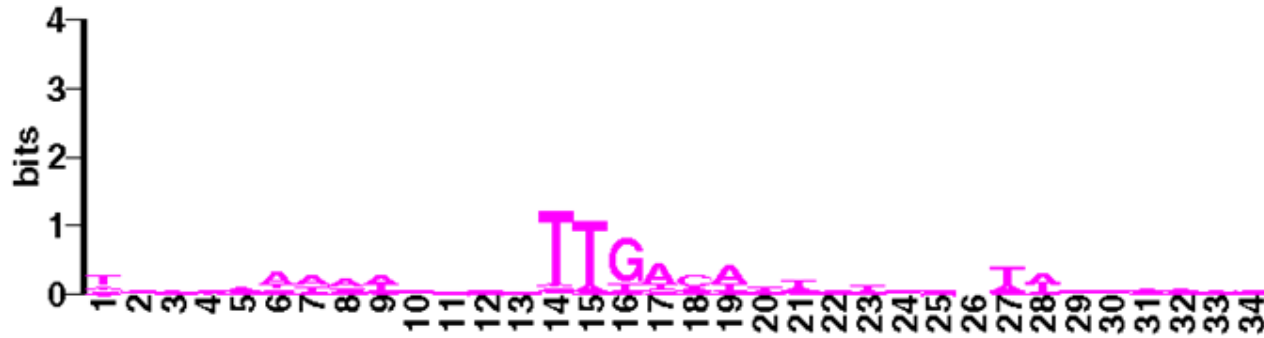


Figure 2. Length distribution probability densities of protein-coding and non-coding regions derived from the annotated *E.coli* genomic DNA (histograms). (a) Coding regions; the solid curve is the approximation by γ distribution $g(d) = N_c(d/D_c)^2 \exp(-d/D_c)$, where d is the length in nt, $D_c = 300$ nt, N_c is the coefficient chosen to normalize the distribution function on the interval from 30 nt (the minimal length of coding region) to 7155 nt (the maximal length). (b) Non-coding regions; the solid curve is the approximation by exponential distribution $f(d) = N_n \exp(-d/D_n)$, where $D_n = 150$ nt. The coefficient N_n normalizes the distribution function on the interval from 1 to 1000 nt.

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.

Compilation of *Bacillus subtilis* sigma A-dependent promoter elements



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,
 - boîte 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 boîtes 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:

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

Motif dans les séquences de Maltose Binding Proteins

YvfK_Bs	PT P NIPEMNEIW
YvfK_Bs	PT P NIPEMAEVW
MalX_Sp	PL P NISQMSAVW
MalE_Sc	PR P ALPEYSSLW
MalE_Tm	PM P NVPEMAPVW
MalE_Dr	PM P NIPEMGAVW
CymE_Ko	AM P SIPEMGYLW
MalE_Ea	IM P NI PQMSAFW
MalE_Sy	IM P NI PQMSAFW
MalE_Ec	IM P NI PQMSAFW

Signature PROSITE :

[PAI]-[TLRM]-P-[NAS]-[ILV]-[PS]-[EQ]-[MY]-[NASG]-[EASPY]-[ILVF]-W

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

$$\text{Score1} = -15 - 17 + 1 - 10 + 10 - 29 = -60$$

AC TATAATCG

$$\text{Score2} = 17 + 18 + 10 + 12 + 10 + 18 = 85$$

ACT ATAATCG

$$\text{Score3} = -28 - 17 + 1 + 12 - 5 - 22 = -59$$

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

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

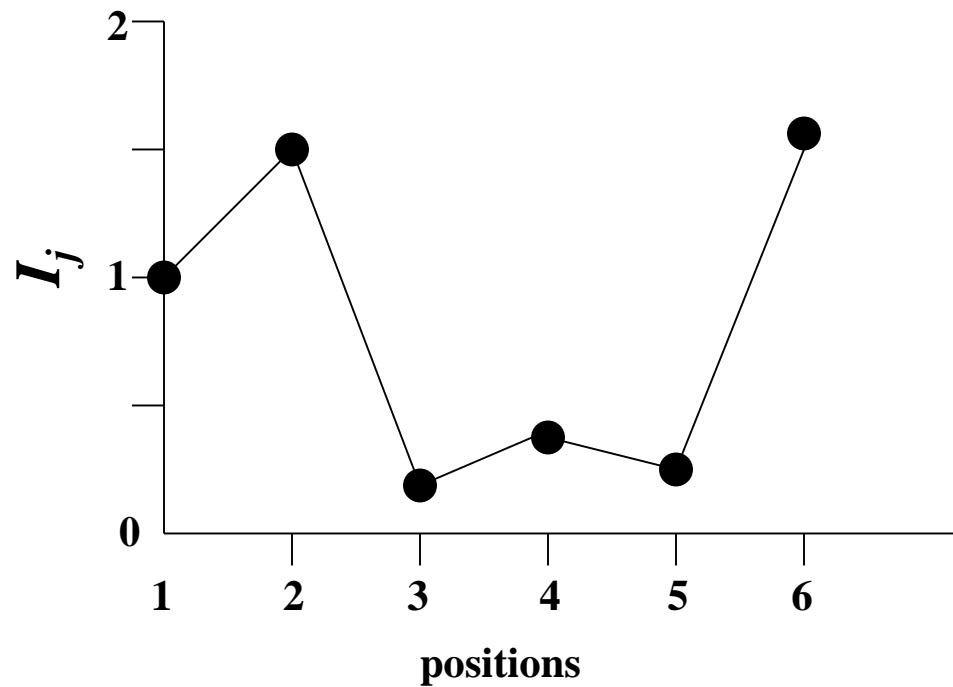
Les positions du signal qui contiendront de l'information seront celles qui auront une composition très biaisées par rapport à ce qui est attendu.

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

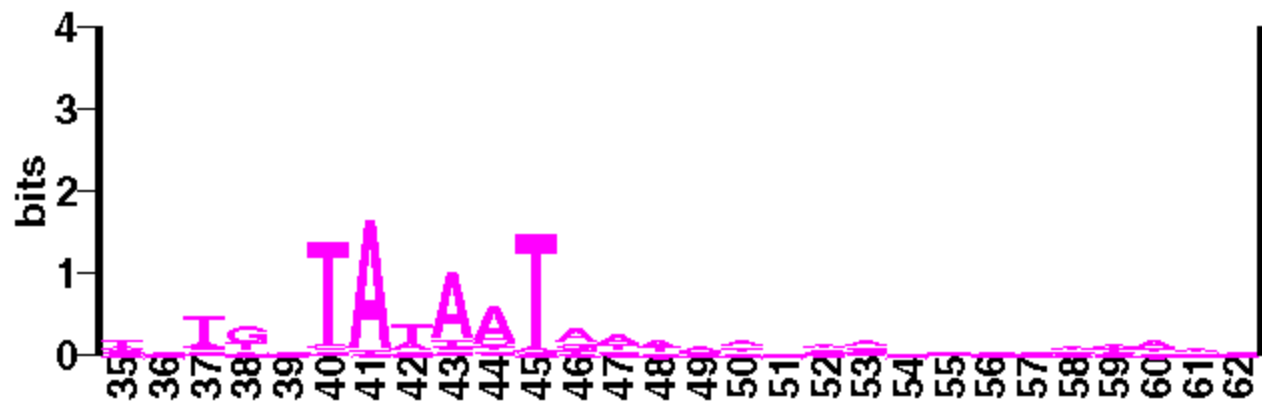
$f_{ij} \log_2 f_{ij} = 0$ et les fréquences observées des autres bases sont nulles. On aura

$$I(j) = 2 \text{ information maximale}$$

Valeurs de l'information I_j à chaque position j du motif -10 des promoteurs d'*E. coli*.



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 autre représentation (matrice de poids) car ils sont modifiés.

```
AAGGAGGTG      consensus
GAAAGGGTG      7  ATG pour B
AGA GAGGTG      6  GTG pour C
GGGGGGATG      5  ATG pour F
```

Unités de traduction prédites

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

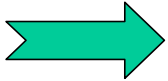
Grand nombre de promoteurs de type sigma A identifiés expérimentalement chez *B. subtilis*:



matrices de poids position (PWM)

Résultats de la recherche des promoteurs

Utilisation du programme Scan_For_Matches et de la matrice de poids



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

Recherche des terminateurs de transcription

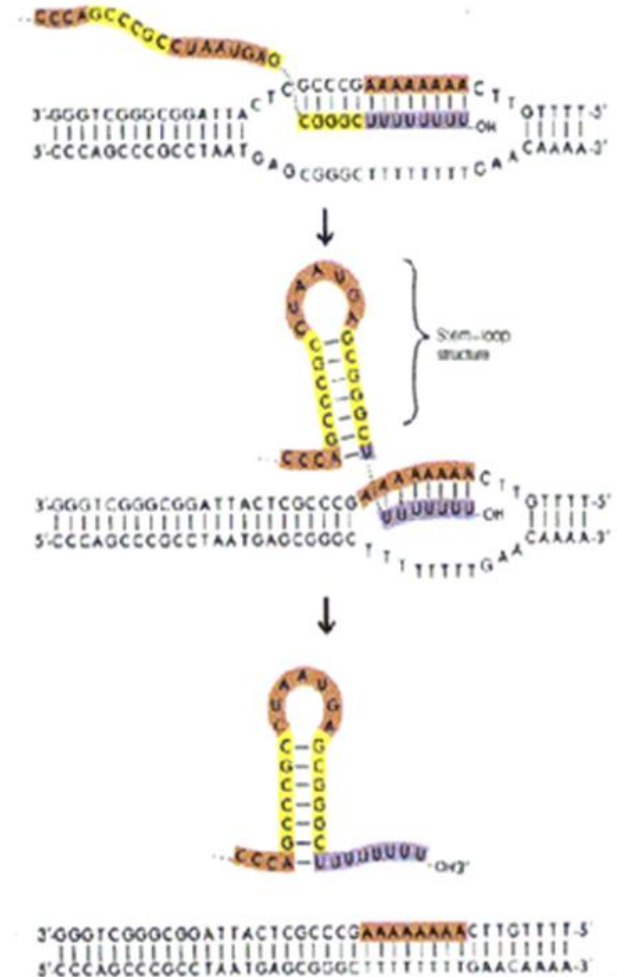
Au niveau séquence, on ne sait modéliser que les terminateurs Rho indépendants.

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).



Termineurs rho indépendant

Deux classes de termineurs:

- 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.

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

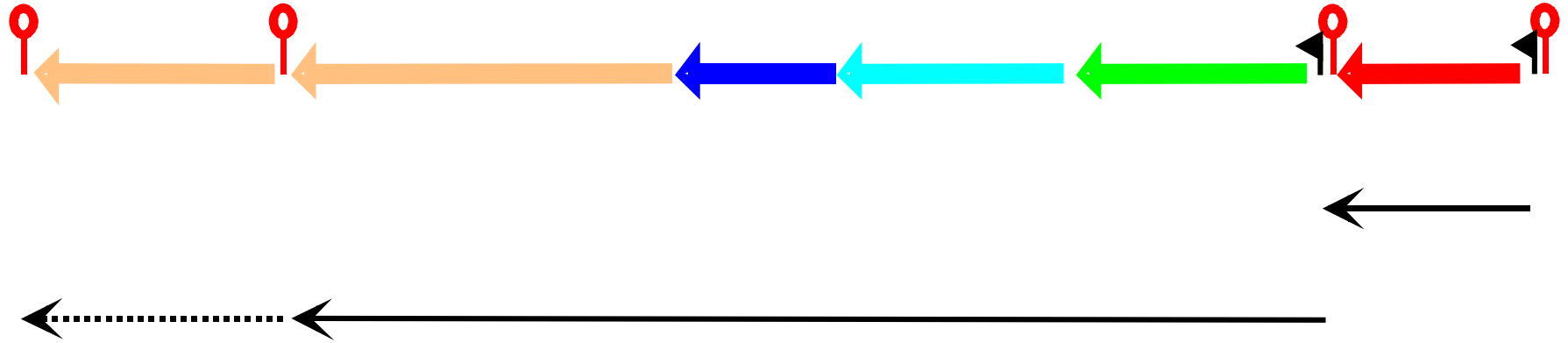
```




      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

```

ATCATT

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

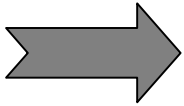
- secondaire
- tertiaire

Recherche de liens fonctionnelles

- réseaux de régulation
- voies métaboliques
- interactions moléculaires

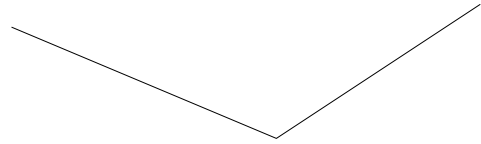
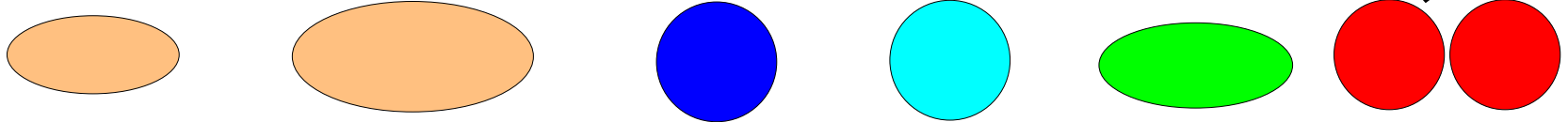
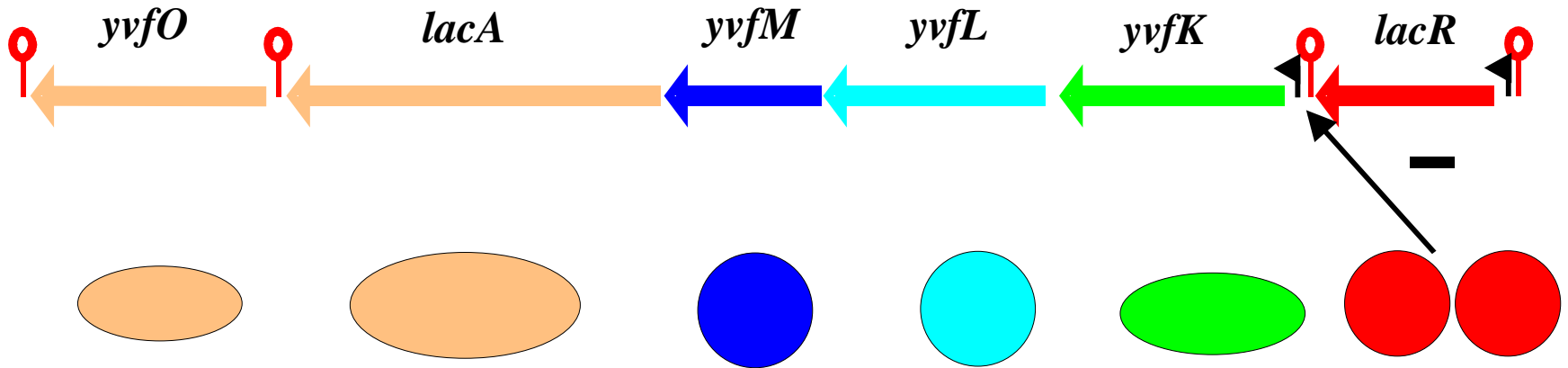
Prédiction fonctionnelle

Recherche par similitude dans les bases de données: programme BLAST

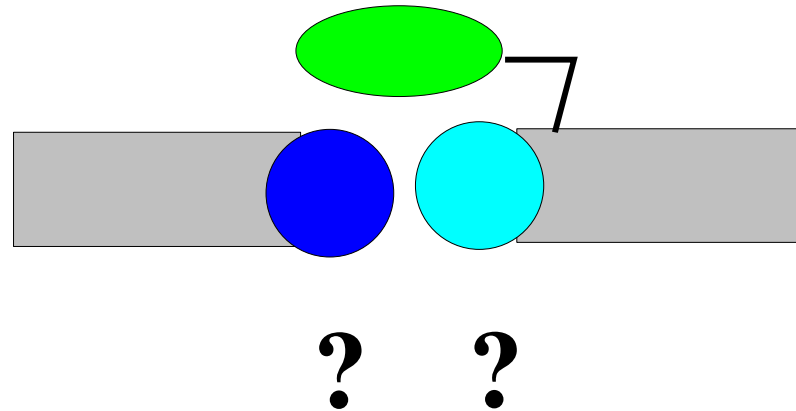


- **A LACR** protéine régulatrice de type LacI/GalR
- **B YVFK** protéine affine d'un ABC transporteur
- **C YVFL** perméase d'un ABC transporteur
- **D YVFM** perméase d'un ABC transporteur
- **E LACA** galactosidase
- **F YVFO** arabino-galactosidase

Synthèse des résultats



Systeme 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

- ✓ **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.

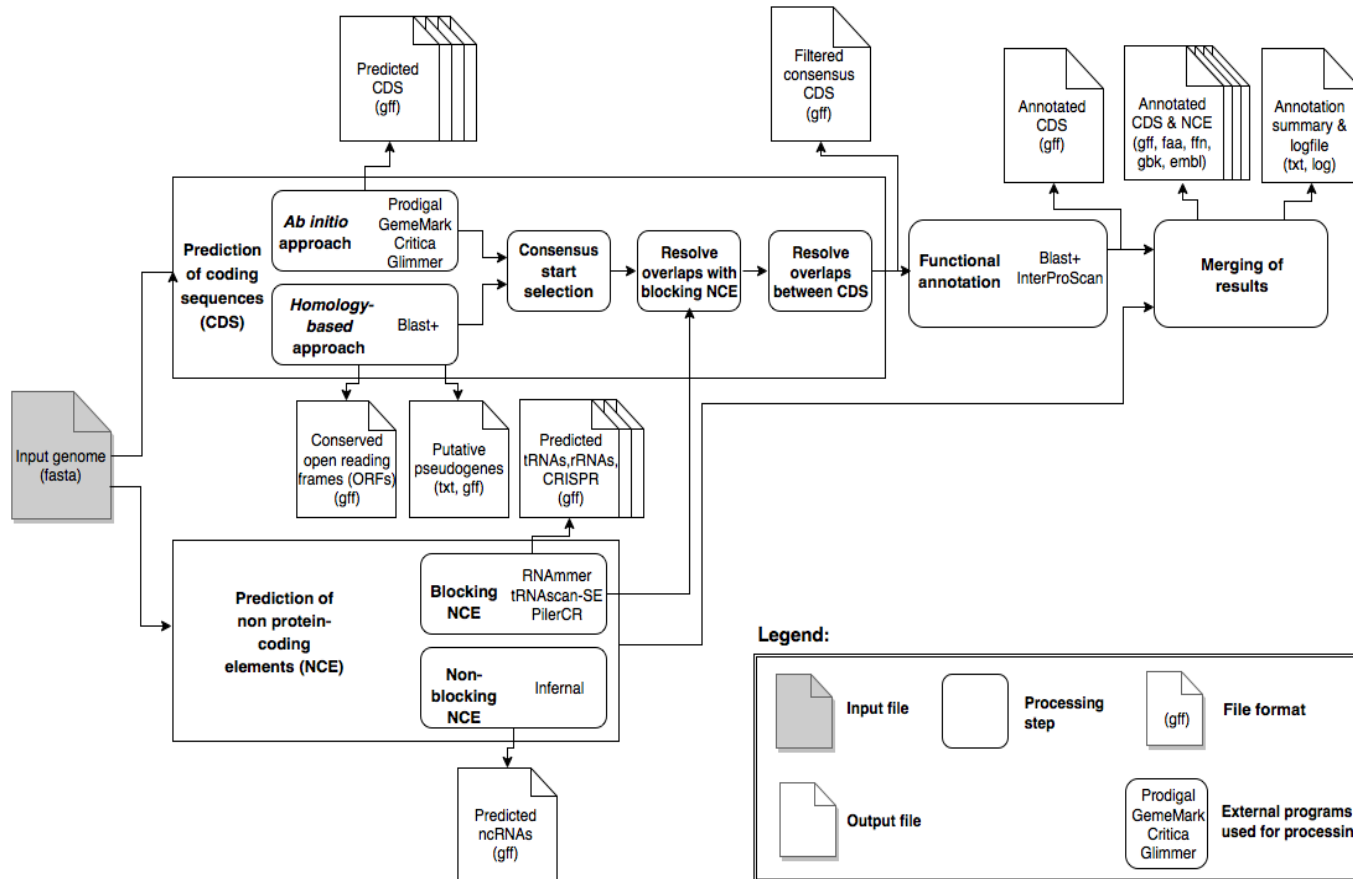
- ✓ **ConsPred** (Bioinformatics. 2016, 32:3327-29)

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.

Rappel sur la prédiction fonctionnelle

- Recherche par similarité dans les banques de données de séquences
- Recherche de Domaines fonctionnels

Recherche par similarité dans les banques/bases de données

La suite Blast (Basic Local Alignment Search Tool)

- Approche développée pour faire des recherches rapides des séquences homologues à notre séquences d'intérêt.
- L'alignement fourni n'est pas un alignement exact car utilisation d'heuristiques

Pourquoi comparer des séquences (nucléiques ou protéiques) ?

Hypothèse 1 : si deux ou plusieurs séquences possèdent des résidus conservés (bases ou acides aminés), cela signifie qu'elles ont une histoire évolutive commune. Elles ont évolué à partir d'une séquence ancêtre commune.

On dit qu'elles sont **homologues**.

Hypothèse 2 : si deux séquences sont homologues, alors elles doivent avoir des fonctions similaires.

Le pourcentage de similitude entre deux séquences est considéré comme reflétant la distance évolutive existant entre ces deux séquences. Les différences observées sont dues à l'accumulation de mutations au cours du temps. Les mutations prises en compte sont les substitutions et les insertions/délétions (indels).

Homologie - orthologie- paralogie

- Deux gènes sont **homologues** s'ils ont divergé à partir d'une séquence ancêtre commune.
- Deux gènes sont **orthologues** si leur divergence est due à la spéciation (le gène ancêtre commun se trouvait dans l'organisme ancêtre).
- Deux gènes sont **paralogues** si leur divergence est due à la duplication du gène ancêtre.

Donc deux séquences sont ou ne sont pas homologues.

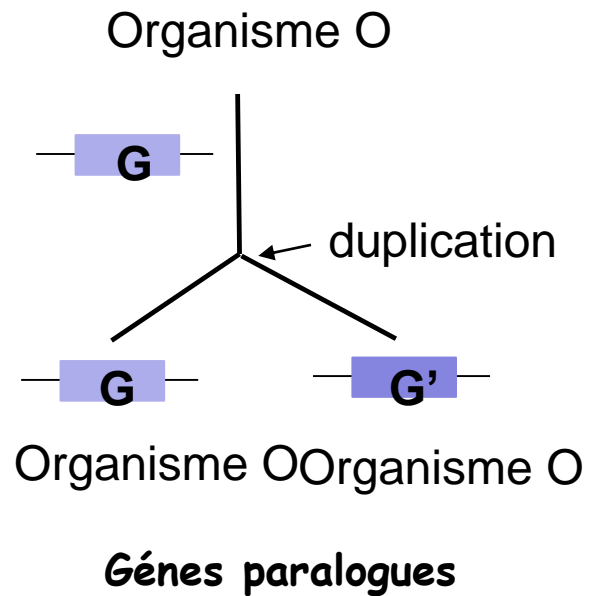
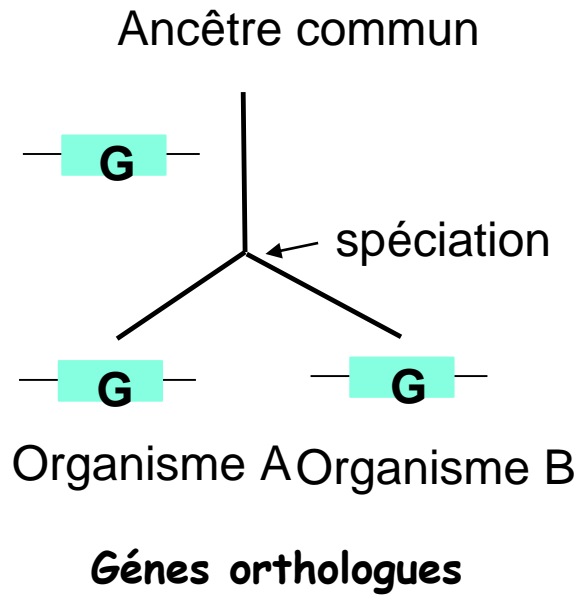
Dire que la protéine X a 80% d'homologie avec la protéine Y est donc

incorrect:

soit:

- les deux protéines présentent 80% d'identité (résidus identiques)
- les deux protéines présentent 80% de similarité (résidus similaires)

Homologie - orthologie- paralogie



Alignement de deux séquences

L'alignement de deux séquences permet de détecter les régions conservées entre deux séquences et de quantifier les similarités observées

→ Calcul d'un score

Petit exemple d'alignement :

```
AACT--GGTAACCG  
AGCTACGGT--CCG
```

Le score de l'alignement doit prendre en compte toutes les positions alignées : identités, substitutions et indels. Chacun de ces événements va recevoir un poids, appelé score élémentaire s_e . Le score de l'alignement correspondra à la somme des scores élémentaires correspondant aux positions alignées.

$$S = \sum_{i=1}^l s_e(i)$$

exemple: $l = 14$

s_e identité = +2

s_e substitution = -1

s_e indels = -2



$S = 9$

Où l est le nombre de positions alignées

Alignement de deux séquences : algorithme de programmation dynamique

Etant donné un système de score, garantit l'obtention de l'alignement optimal

Hypothèse : l'évolution est parcimonieuse

Signification: pour trouver l'alignement optimal, l'algorithme va rechercher le chemin permettant de passer d'une séquence à l'autre avec le minimum de changements

Deux types de score en fonction des algorithmes :

- score d'homologie: la valeur du score diminue avec le nombre de différences observées entre les deux séquences
- score de distance: la valeur du score augmente avec le nombre de différences observées entre les deux séquences

Systeme de score (acides nucléiques) : matrices de substitution

	Score d'homologie	Score de distance
identité	+1	0
mismatch	-1	+1
indel	-2	+2



	A	T	C	G	-
A	+1	-1	-1	-1	-2
T	-1	+1	-1	-1	-2
C	-1	-1	+1	-1	-2
G	-1	-1	-1	+1	-2
-	-2	-2	-2	-2	

	A	T	C	G	-
A	0	+1	+1	+1	+2
T	+1	0	+1	+1	+2
C	+1	+1	0	+1	+2
G	+1	+1	+1	0	+2
-	+2	+2	+2	+2	

Les matrices de substitution permettent de spécifier le coût/score de chaque substitution possible (A avec C, A avec T, ...) de manière indépendante

Alphabet étendu pour les nucléotides

- Problème de séquençage
- Polymorphisme

L'alphabet étendu appelé code IUPAC (International Union of Pure and Applied Chemistry) permet de modéliser l'incertitude sur une séquence : le nucléotide à une position n'est pas clairement identifié ou peut varier.

Symbol	Meaning	Nucleic Acid
A	A	Adenine
C	C	Cytosine
G	G	Guanine
T	T	Thymine
U	U	Uracil
M	A or C	
R	A or G	purine
W	A or T	
S	C or G	
Y	C or T	pyrimidine
K	G or T	
V	A or C or G	not T
H	A or C or T	not G
D	A or G or T	not C
B	C or G or T	not A
X	G or A or T or C	any
N	G or A or T or C	any
.	G or A or T or C	any

Problème :

un mismatch entre A et C
n'a pas le même coût qu'un
mismatch entre A et M !

Matrice pour les nucléotides (alphabet étendu)

NUC4.4 pour BLAST ou EDNAFULL pour EMBOSS

This matrix was created by Todd Lowe 12/10/92

#

Uses ambiguous nucleotide codes, probabilities rounded to
nearest integer

#

Lowest score = -4, Highest score = 5

#

	A	T	G	C	S	W	R	Y	K	M	B	V	H	D	N
A	5	-4	-4	-4	-4	1	1	-4	-4	1	-4	-1	-1	-1	-2
T	-4	5	-4	-4	-4	1	-4	1	1	-4	-1	-4	-1	-1	-2
G	-4	-4	5	-4	1	-4	1	-4	1	-4	-	-	-	-	-
C	-4	-4	-4	5	1	-4	-4	1	-4	1	-	-	-	-	-
S	-4	-4	1	1	-1	-4	-2	-2	-2	-2	-	-	-	-	-
W	1	1	-4	-4	-4	-1	-2	-2	-2	-2	-	-	-	-	-
R	1	-4	1	-4	-2	-2	-1	-4	-2	-2	-	-	-	-	-
Y	-4	1	-4	1	-2	-2	-4	-1	-2	-2	-	-	-	-	-
K	-4	1	1	-4	-2	-2	-2	-2	-1	-4	-	-	-	-	-
M	1	-4	-4	1	-2	-2	-2	-2	-4	-1	-	-	-	-	-
B	-4	-1	-1	-1	-1	-3	-3	-1	-1	-3	-	-	-	-	-
V	-1	-4	-1	-1	-1	-3	-1	-3	-3	-1	-	-	-	-	-
H	-1	-1	-4	-1	-3	-1	-3	-1	-3	-1	-	-	-	-	-
D	-1	-1	-1	-4	-3	-1	-1	-3	-1	-3	-	-	-	-	-
N	-2	-2	-2	-2	-1	-1	-1	-1	-1	-1	-	-	-	-	-

	A	T	G	C	S	W	R	Y	K	M	B	V	H	D	N
A	5	-4	-4	-4	-4	1	1	-4	-4	1	-4	-1	-1	-1	-2
T	-4	5	-4	-4	-4	1	-4	1	1	-4	-1	-4	-1	-1	-2
G	-4	-4	5	-4	1	-4	1	-4	1	-4	-	-	-	-	-
C	-4	-4	-4	5	1	-4	-4	1	-4	1	-	-	-	-	-
S	-4	-4	1	1	-1	-4	-2	-2	-2	-2	-	-	-	-	-
W	1	1	-4	-4	-4	-1	-2	-2	-2	-2	-	-	-	-	-
R	1	-4	1	-4	-2	-2	-1	-4	-2	-2	-	-	-	-	-
Y	-4	1	-4	1	-2	-2	-4	-1	-2	-2	-	-	-	-	-
K	-4	1	1	-4	-2	-2	-2	-2	-1	-4	-	-	-	-	-
M	1	-4	-4	1	-2	-2	-2	-2	-4	-1	-	-	-	-	-
B	-4	-1	-1	-1	-1	-3	-3	-1	-1	-3	-	-	-	-	-
V	-1	-4	-1	-1	-1	-3	-1	-3	-3	-1	-	-	-	-	-
H	-1	-1	-4	-1	-3	-1	-3	-1	-3	-1	-	-	-	-	-
D	-1	-1	-1	-4	-3	-1	-1	-3	-1	-3	-	-	-	-	-
N	-2	-2	-2	-2	-1	-1	-1	-1	-1	-1	-	-	-	-	-

	A	T	G	C	S	W	R	Y	K	M	B	V	H	D	N
A	5	-4	-4	-4	-4	1	1	-4	-4	1	-4	-1	-1	-1	-2
T	-4	5	-4	-4	-4	1	-4	1	1	-4	-1	-4	-1	-1	-2
G	-4	-4	5	-4	1	-4	1	-4	1	-4	-	-	-	-	-
C	-4	-4	-4	5	1	-4	-4	1	-4	1	-	-	-	-	-
S	-4	-4	1	1	-1	-4	-2	-2	-2	-2	-	-	-	-	-
W	1	1	-4	-4	-4	-1	-2	-2	-2	-2	-	-	-	-	-
R	1	-4	1	-4	-2	-2	-1	-4	-2	-2	-	-	-	-	-
Y	-4	1	-4	1	-2	-2	-4	-1	-2	-2	-	-	-	-	-
K	-4	1	1	-4	-2	-2	-2	-2	-1	-4	-	-	-	-	-
M	1	-4	-4	1	-2	-2	-2	-2	-4	-1	-	-	-	-	-
B	-4	-1	-1	-1	-1	-3	-3	-1	-1	-3	-	-	-	-	-
V	-1	-4	-1	-1	-1	-3	-1	-3	-3	-1	-	-	-	-	-
H	-1	-1	-4	-1	-3	-1	-3	-1	-3	-1	-	-	-	-	-
D	-1	-1	-1	-4	-3	-1	-1	-3	-1	-3	-	-	-	-	-
N	-2	-2	-2	-2	-1	-1	-1	-1	-1	-1	-	-	-	-	-

	A	T	G	C	S	W	R	Y	K	M	B	V	H	D	N
A	5	-4	-4	-4	-4	1	1	-4	-4	1	-4	-1	-1	-1	-2
T	-4	5	-4	-4	-4	1	-4	1	1	-4	-1	-4	-1	-1	-2
G	-4	-4	5	-4	1	-4	1	-4	1	-4	-	-	-	-	-
C	-4	-4	-4	5	1	-4	-4	1	-4	1	-	-	-	-	-
S	-4	-4	1	1	-1	-4	-2	-2	-2	-2	-	-	-	-	-
W	1	1	-4	-4	-4	-1	-2	-2	-2	-2	-	-	-	-	-
R	1	-4	1	-4	-2	-2	-1	-4	-2	-2	-	-	-	-	-
Y	-4	1	-4	1	-2	-2	-4	-1	-2	-2	-	-	-	-	-
K	-4	1	1	-4	-2	-2	-2	-2	-1	-4	-	-	-	-	-
M	1	-4	-4	1	-2	-2	-2	-2	-4	-1	-	-	-	-	-
B	-4	-1	-1	-1	-1	-3	-3	-1	-1	-3	-	-	-	-	-
V	-1	-4	-1	-1	-1	-3	-1	-3	-3	-1	-	-	-	-	-
H	-1	-1	-4	-1	-3	-1	-3	-1	-3	-1	-	-	-	-	-
D	-1	-1	-1	-4	-3	-1	-1	-3	-1	-3	-	-	-	-	-
N	-2	-2	-2	-2	-1	-1	-1	-1	-1	-1	-	-	-	-	-

	A	T	G	C	S	W	R	Y	K	M	B	V	H	D	N
A	5	-4	-4	-4	-4	1	1	-4	-4	1	-4	-1	-1	-1	-2
T	-4	5	-4	-4	-4	1	-4	1	1	-4	-1	-4	-1	-1	-2
G	-4	-4	5	-4	1	-4	1	-4	1	-4	-	-	-	-	-
C	-4	-4	-4	5	1	-4	-4	1	-4	1	-	-	-	-	-
S	-4	-4	1	1	-1	-4	-2	-2	-2	-2	-	-	-	-	-
W	1	1	-4	-4	-4	-1	-2	-2	-2	-2	-	-	-	-	-
R	1	-4	1	-4	-2	-2	-1	-4	-2	-2	-	-	-	-	-
Y	-4	1	-4	1	-2	-2	-4	-1	-2	-2	-	-	-	-	-
K	-4	1	1	-4	-2	-2	-2	-2	-1	-4	-	-	-	-	-
M	1	-4	-4	1	-2	-2	-2	-2	-4	-1	-	-	-	-	-
B	-4	-1	-1	-1	-1	-3	-3	-1	-1	-3	-	-	-	-	-
V	-1	-4	-1	-1	-1	-3	-1	-3	-3	-1	-	-	-	-	-
H	-1	-1	-4	-1	-3	-1	-3	-1	-3	-1	-	-	-	-	-
D	-1	-1	-1	-4	-3	-1	-1	-3	-1	-3	-	-	-	-	-
N	-2	-2	-2	-2	-1	-1	-1	-1	-1	-1	-	-	-	-	-

	A	T	G	C	S	W	R	Y	K	M	B	V	H	D	N
A	5	-4	-4	-4	-4	1	1	-4	-4	1	-4	-1	-1	-1	-2
T	-4	5	-4	-4	-4	1	-4	1	1	-4	-1	-4	-1	-1	-2
G	-4	-4	5	-4	1	-4	1	-4	1	-4	-	-	-	-	-
C	-4	-4	-4	5	1	-4	-4	1	-4	1	-	-	-	-	-
S	-4	-4	1	1	-1	-4	-2	-2	-2	-2	-	-	-	-	-
W	1	1	-4	-4	-4	-1	-2	-2	-2	-2	-	-	-	-	-
R	1	-4	1	-4	-2	-2	-1	-4	-2	-2	-	-	-	-	-
Y	-4	1	-4	1	-2	-2	-4	-1	-2	-2	-	-	-	-	-
K	-4	1	1	-4	-2	-2	-2	-2	-1	-4	-	-	-	-	-
M	1	-4	-4	1	-2	-2	-2	-2	-4	-1	-	-	-	-	-
B	-4	-1	-1	-1	-1	-3	-3	-1	-1	-3	-	-	-	-	-
V	-1	-4	-1	-1	-1	-3	-1	-3	-3	-1	-	-	-	-	-
H	-1	-1	-4	-1	-3	-1	-3	-1	-3	-1	-	-	-	-	-
D	-1	-1	-1	-4	-3	-1	-1	-3	-1	-3	-	-	-	-	-
N	-2	-2	-2	-2	-1	-1	-1	-1	-1	-1	-	-	-	-	-

	A	T	G	C	S	W	R	Y	K	M	B	V	H	D	N
A	5	-4	-4	-4	-4	1	1	-4	-4	1	-4	-1	-1	-1	-2
T	-4	5	-4	-4	-4	1	-4	1	1	-4	-1	-4	-1	-1	-2
G	-4	-4	5	-4	1	-4	1	-4	1	-4	-	-	-	-	-
C	-4	-4	-4	5	1	-4	-4	1	-4	1	-	-	-	-	-
S	-4	-4	1	1	-1	-4	-2	-2	-2	-2	-	-	-	-	-
W	1	1	-4	-4	-4	-1	-2	-2	-2	-2	-	-	-	-	-

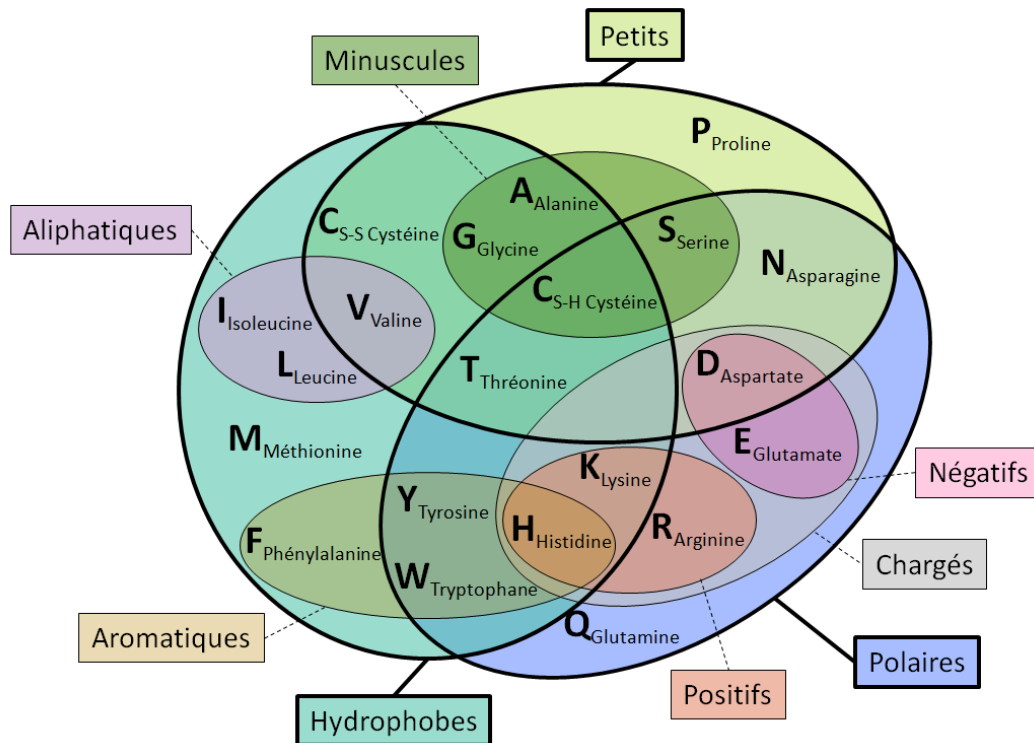
Alignement de deux séquences protéiques

Les acides aminés composant une protéine peuvent avoir des propriétés physico-chimiques similaires.



La structure 3D dépend de ces caractéristiques

Une similitude au niveau de ces propriétés sera suffisante pour permettre la substitution d'un acide aminé en un autre sans perturber la fonction de la protéine (par exemple, échange de l'acide aminé hydrophobe valine en leucine).



Acide aminé polaire ou hydrophile, localisé le plus souvent à la surface des protéines (certains sont chargés)

Acide aminé hydrophobe ou apolaire ou aliphatique tendent à occuper le cœur de la protéine.

Les acides aminés aromatiques possèdent un cycle dans leur chaîne latérale.

La cystéine permet d'établir des ponts disulfures

Diagramme de Venn des propriétés des acides aminés

Alignement de deux séquences protéiques

Lors de la comparaison de deux séquences protéiques, il faut prendre en compte en plus de l'identité et de la différence, la similarité qui peut exister entre deux acides aminés.

Comment quantifier la similarité entre deux acides aminés ?

- calculer une distance entre acides aminés basée sur leurs caractéristiques
- estimer la fréquence de substitution de l'acide aminé X en Y au cours de l'évolution

Les deux approches donnent une matrice (20,20) symétrique par rapport à la diagonale. Cependant, les matrices les plus utilisées ont été obtenues par la seconde approche et sont appelées « matrices de substitution »

Approches basée sur les fréquences de substitutions des a.a. au cours de l'évolution

Principe :

- les séquences homologues ont conservées des fonctions similaires
- deux a.a. se ressembleront d'autant plus que la fréquence de substitution observée est grande puisque ces substitutions n'auront pas modifié la fonction de la protéine
- il est possible d'estimer la fréquence avec laquelle un a.a. est remplacé par un autre au cours de l'évolution à partir de séquences homologues alignées

Principales approches :

- Comparaison directe des séquences (alignement global) : matrices PAM (Dayhoff, 1978)
- Comparaison des domaines protéiques (régions les plus conservées) : matrices **BLOSUM** (Henikoff et Henikoff, 1992)

Dans les deux cas, on obtient une famille de matrices car les probabilités de mutation d'un acide aminé en un autre dépend de la distance évolutive entre les séquences comparées.

Matrices PAM

La matrice « de base » construite par Dayhoff correspond à la 1-PAM dans laquelle ont a 1 mutation acceptée pour 100 acides aminés. Comment estimer les probabilités de mutation des paires d'acides aminés pour des distances évolutives plus grande ?

On a fait l'hypothèse que la probabilité de mutation d'un acide aminé est indépendante de ce qui s'est produit à cette position dans le passé, donc pour obtenir les probabilités de mutation pour des intervalles d'évolution plus grands on va multiplier la PAM1 avec elle-même. Une PAMk sera obtenue en multipliant la PAM1 k fois par elle-même (k mutations acceptées pour 100 sites)

$$\text{PAM2} = \text{PAM1} \times \text{PAM1} = \text{PAM1}^2$$

intervalle d'évolution : 2 mutations acceptées pour chaque 100 résidus

$$\text{PAM40} = \text{PAM1}^{40}$$

intervalle d'évolution : 40 mutations acceptées pour chaque 100 résidus

$$\text{PAM120} = \text{PAM1}^{120}$$

intervalle d'évolution : 120 mutations acceptées pour chaque 100 résidus

$$\text{PAM250} = \text{PAM1}^{250}$$

intervalle d'évolution : 250 mutations acceptées pour chaque 100 résidus

divergence



Matrices BLOSUM (Henikoff et Henikoff, 1992)

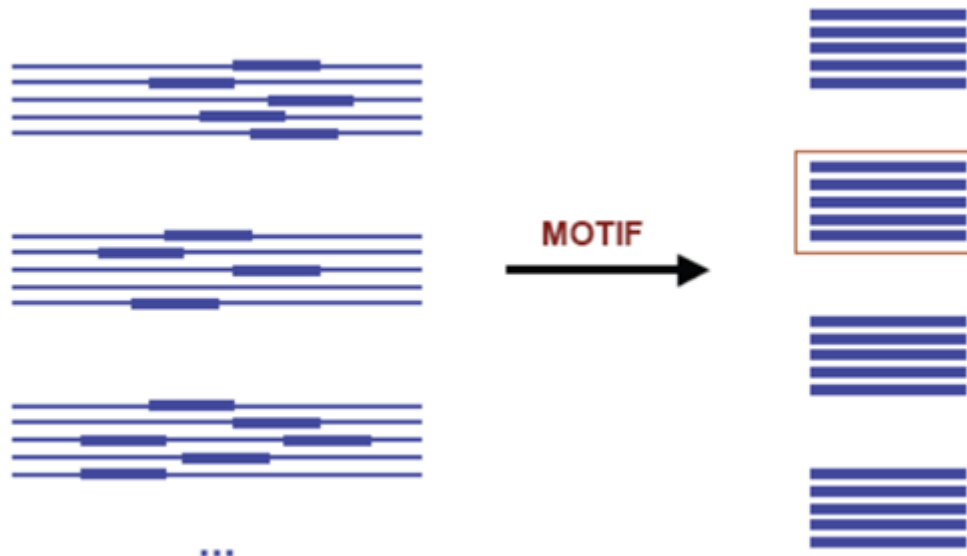
BLOSUM : BLOcks SUBstituion Matrix

Principe :

- Obtention à partir de blocs de séquences alignées (alignement multiple sans brèche)
- Pour une paire d'a.a. : $\log(\text{fréquence observée} / \text{fréquence attendue})$

Avantages par rapport aux matrices PAM :

- contrairement aux matrices PAM, les matrices BLOSUM pour différentes distances évolutives sont obtenues directement avec des séquences plus ou moins divergentes
- l'utilisation de blocs plutôt que de séquences complètes : modélise les contraintes uniquement sur les régions conservées
- obtenues à partir d'un plus grand jeu de données (>2000 blocks, > 500 familles)



504 groupes de protéines apparentées

2205 blocs (domaines alignement sans gap)

Choix des Matrices de substitution

Famille de matrices correspondant à différentes distances évolutives entre les séquences :

PAM120 et BLOSUM80 : estimation des fréquences de substitution entre acides aminés pour des séquences proches dans l'évolution (courtes distances)

PAM250 et BLOSUM45 : estimation des fréquences de substitution entre acides aminés pour des séquences distantes dans l'évolution (longues distances)

PAM160 et BLOSUM62 : estimation des fréquences de substitution entre acides aminés pour des séquences ayant des distances évolutives intermédiaires.

longueur séquence	matrice	ouverture de gap	extension de gap
≥300	BLOSUM50	-10	-2
85-300	BLOSUM62	-7	-1
50-85	BLOSUM80	-16	-4
≥300	PAM250	-10	-2
85-300	PAM120	-16	-4

distance %	PAM
1	1
25	30
50	80
80	246

Recommandations (à adapter)

Utilisation des Matrices de substitution

Ces matrices sont utilisées comme paramètres dans :

- les programmes d'alignement de deux séquences
- les recherches par similitude dans les bases de données
- les programmes d'alignement multiple

Quelle matrice doit-on utiliser ?

Les matrices BLOSUM sont le plus souvent proposées comme matrices par défaut car les fréquences de substitution sont directement calculées à partir de l'alignement.

La BLOSUM62 est utilisée comme matrice par défaut car elle offre un bon compromis quand les distances évolutives entre les séquences ne sont pas connues.

La BLOSUM80 donnera de meilleurs résultats pour des séquences proches dans l'évolution. Elle tend à trouver des alignements courts fortement similaires.

La BLOSUM45 donnera de meilleurs résultats pour des séquences éloignées dans l'évolution. Elle trouvera de plus longs alignements locaux de faible conservation.

Recherche de similarité avec la suite Blast

En fonction de la nature des séquences sonde et banque, utilisations de différents programmes de la suite Blast :

programme	séquence requête	Banque
BlastN	nucléique	nucléique
BlastP	protéique	protéique
BlastX	nucléique (séquence traduite dans les 6 cadres de lecture)	protéique
tblastN	protéique	nucléique (séquences de la banque traduites dans les 6 cadres de lecture)
tblastX	nucléique (séquence traduite dans les 6 cadres de lecture)	nucléique (séquences de la banque traduites dans les 6 cadres de lecture)

Recherche de similarité avec la suite Blast

Quelque soit le programme une même approche est utilisée qui permet de passer très peu de temps à analyser les séquences de la banque qui ne présentent pas de conservation avec la séquence sonde (requête).

- Choix d'une matrice de substitution pour réaliser les comparaisons
- Utilisation d'une heuristique pour fournir l'alignement final entre les deux séquences (alignement local qui va renvoyer les deux sous-régions les plus conservées entre la séquence requête et séquence de la banque)

Page d'accueil de la suite Blast (site NCBI)

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. [Learn more](#)

**N
E
W
S**

A new version IgBLAST (1.15.0) is here.

We've released a new version of IgBLAST with four new improvements.

Wed, 22 Jan 2020 16:00:00 EST

[More BLAST news...](#)

Web BLAST



Nucleotide BLAST
nucleotide ▶ nucleotide



blastx
translated nucleotide ▶ protein



tblastn
protein ▶ translated nucleotide



Protein BLAST
protein ▶ protein

BLAST Genomes

Human Mouse Rat Microbes

Exemple d'interface du programme BlastP (site NCBI)

The image shows a screenshot of the NCBI BLASTP web interface in a Mozilla Firefox browser. The browser's address bar shows the URL: `http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&BLAST_PROGRAMS=blastp&PAGE_TYPE=BlastSearch&SHOW_DEFAULTS=on&LINK_LOC=blastp`. The page title is "Protein BLAST: search protein databases using a protein query - Mozilla Firefox".

The interface is titled "BLAST Basic Local Alignment Search Tool" and includes navigation links: Home, Recent Results, Saved Strategies, and Help. There are also links for "My NCBI" with "Sign In" and "Register" options.

The main content area is titled "BLASTP programs search protein databases using a protein query. more...". It features several sections:

- Enter Query Sequence:** A large text input field for "Enter accession number, gi, or FASTA sequence". To its right are "Clear" and "Query subrange" options, with "From" and "To" input fields. A blue arrow points from the text "Votre séquence sonde" to the main input field.
- Or, upload file:** A file upload button labeled "Parcourir...".
- Job Title:** A text input field for "Enter a descriptive title for your BLAST search".
- Align two or more sequences:** A checkbox option.
- Choose Search Set:** A section with three sub-sections:
 - Database:** A dropdown menu currently set to "Non-redundant protein sequences (nr)". A blue arrow points from the text "Choix de la banque" to this dropdown.
 - Organism:** A text input field for "Enter organism name or id--completions will be suggested".
 - Entrez Query:** A text input field for "Enter an Entrez query to limit search".
- Program Selection:** A section with radio buttons for "blastp (protein-protein BLAST)", "PSI-BLAST (Position-Specific Iterated BLAST)", and "PHI-BLAST (Pattern Hit Initiated BLAST)".

At the bottom, there is a "BLAST" button and a search bar containing the word "complexity". Below the search bar are navigation buttons: "Précédent", "Suivant", "Surligner tout", and "Respecter la casse". The status bar at the very bottom indicates "Terminé".

Les paramètres « cachés »

Algorithm parameters

General Parameters

Max target sequences

100

Nombre de séquences cibles

Select the maximum number of aligned sequences to display

Short queries

Automatically adjust parameters for short input sequences

Expect threshold

10

Seuil E-value

Word size

3

Taille des mots pour construire la liste des mots voisins

Max matches in a query range

0

Scoring Parameters

Matrix

BLOSUM62

Choix de la matrice de substitution

Gap Costs

Existence: 11 Extension: 1

Pondération des gaps : ouverture et extension

Compositional adjustments

Conditional compositional score matrix adjustment

Filters and Masking

Filter

Low complexity regions

Mask

Mask for lookup table only

Mask lower case letters

La suite Blast

Possibilité d'appliquer des filtres et masques (paramètres de l'algorithme) :

- masquer les séquences de faible complexité (proposé pour l'ensemble des programmes de la suite Blast)
- dans le cas d'une recherche avec une séquence d'acides nucléiques contre une banque de séquences nucléiques (BlastN), masquer les séquences répétées (ex: les séquences Alu chez les primates).

Une première analyse compare la séquence sonde à une banque de séquences d'éléments répétés. Les zones de la séquence sonde s'alignant avec les séquences d'éléments répétés sont masquées pour la recherche de similarité dans la banque.

Exemple de recherche BlastP (site NCBI)

BLAST » blastp suite

Standard Protein BLAST

blastn blastp blastx tblastn tblastx

BLASTP programs search protein databases using a protein query [more...](#)

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Query subrange

```
>Eco1A01.YRAP
MKALSPFIAVLISALLLGGCVAAAIVGTAAVGTAKAATDFRSVGTQVDDGTELVFVNSALSK
DEQIKKEARINVTAYQGRVLLVQGSFNAEL SARAKIANGVDGANEVYNEIRQQPIGLG
EASNDTWITTKVRSQQLLTSDLVKSNNVKVTEENGEVFLMGLVTEREAKAAADIASRVGV
KRVTAFTTIIK
```

From To

Or, upload file Aucun fichier sélectionné.

Job Title Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set

Database

Organism Exclude

Optional Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude Models (XM/XP) Non-redundant RefSeq proteins (WP) Uncultured/environmental sample sequences

Optional

Program Selection

Algorithm

Quick BLASTP (Accelerated protein-protein BLAST)

blastp (protein-protein BLAST)

PSI-BLAST (Position-Specific Iterated BLAST)

PHI-BLAST (Pattern Hit Initiated BLAST)

DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

BLAST Search database nr using Blastp (protein-protein BLAST)

Show results in a new window

Algorithm parameters **Note: Parameter values that differ from the default are highlighted in yellow and marked with * sign**

General Parameters

Max target sequences Select the maximum number of aligned sequences to display

Short queries Automatically adjust parameters for short input sequences

Expect threshold

Word size

Max matches in a query range

Scoring Parameters

Matrix

Banque nr : non redondante. Toutes les protéines non redondantes correspondant à la traduction des CDS de Genbank + PDB + SwissProt + PIR-NBRF

Recherche restreinte aux alphaproteobacteria

Nombre de séquences cibles : 500 (nombre de hits)

Page de résultat BlastP (site NCBI)

BLAST® » blastp suite » results for RID-38G2CGA901R

[Home](#) [Recent Results](#) [Saved Strategies](#) [Help](#)

[< Edit Search](#) [Save Search](#) [Search Summary](#) ▾

[How to read this report?](#) [BLAST Help Videos](#) [Back to Traditional Results Page](#)

i Your search is limited to records that include: Alphabacteria (taxid:28211)

Job Title EcoIA01.YRAP

RID [38G2CGA901R](#) *Search expires on 02-01 21:57 pm* [Download All](#) ▾

Program BLASTP [?](#) [Citation](#) ▾

Database nr [See details](#) ▾

Query ID Icl|Query_4796

Description EcoIA01.YRAP

Molecule type amino acid

Query Length 191

Other reports [Distance tree of results](#) [Multiple alignment](#) [MSA viewer](#) [?](#)

Filter Results

Organism *only top 20 will appear* exclude

Type common name, binomial, taxid or group name

[+ Add organism](#)

Percent Identity to

E value to

Query Coverage to

[Filter](#) [Reset](#)

Descriptions [Graphic Summary](#) [Alignments](#) [Taxonomy](#)

Sequences producing significant alignments

[Download](#) ▾ [Manage Columns](#) ▾ Show [?](#)

select all 371 sequences selected

[GenPept](#) [Graphics](#) [Distance tree of results](#) [Multiple alignment](#)

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
<input checked="" type="checkbox"/>	osmotically-inducible protein OsmY [Thalassospira xiamenensis]	159	159	93%	5e-47	48.80%	PTC01376.1
<input checked="" type="checkbox"/>	osmotically-inducible protein OsmY [Rickettsiales bacterium]	137	137	96%	1e-38	39.67%	MBB20570.1
<input checked="" type="checkbox"/>	BON domain-containing protein [Rickettsiales bacterium]	123	123	99%	7e-33	39.69%	MBB86425.1
<input checked="" type="checkbox"/>	MULTISPECIES: BON domain-containing protein [Proteobacteria]	107	107	99%	1e-26	36.32%	WP_124735979.1
<input checked="" type="checkbox"/>	phospholipid-binding protein [Rhodospirillaceae bacterium]	105	105	94%	9e-26	34.62%	MBD21055.1
<input checked="" type="checkbox"/>	BON domain-containing protein [Alphaproteobacteria bacterium]	103	103	99%	2e-25	33.88%	NBR40331.1
<input checked="" type="checkbox"/>	transport-associated protein [Rickettsiales bacterium]	101	101	87%	2e-24	37.13%	MBV35553.1
<input checked="" type="checkbox"/>	Uncharacterized protein in gshII_5' region [Candidatus Jidaibacter acanthamoeba]	100	100	98%	1e-23	33.51%	KIE04947.1
<input checked="" type="checkbox"/>	hypothetical protein BTY25_04340 [Alphaproteobacteria bacterium 16-39-46]	100	100	90%	3e-23	33.91%	OYZ37433.1
<input checked="" type="checkbox"/>	BON domain-containing protein [Candidatus Jidaibacter acanthamoeba]	98.8	98.8	89%	4e-23	35.47%	WP_039456942.1
<input checked="" type="checkbox"/>	BON domain-containing protein [endosymbiont of Acanthamoeba sp. UWCS]	98.8	98.8	89%	4e-23	34.88%	WP_038537893.1
<input checked="" type="checkbox"/>	hypothetical protein COV36_02350 [Alphaproteobacteria bacterium CG11_big_fil_rev_8_21_14_0_20_44_7]	96.7	96.7	92%	2e-22	34.46%	PIR33644.1

Page de résultat : résumé graphique

Descriptions

Graphic Summary

Alignments

Taxonomy

[hover to see the title](#) [click to show alignments](#) Show Conserved Domains

Alignment Scores

■ < 40

■ 40 - 50

■ 50 - 80

■ 80 - 200

■ >= 200

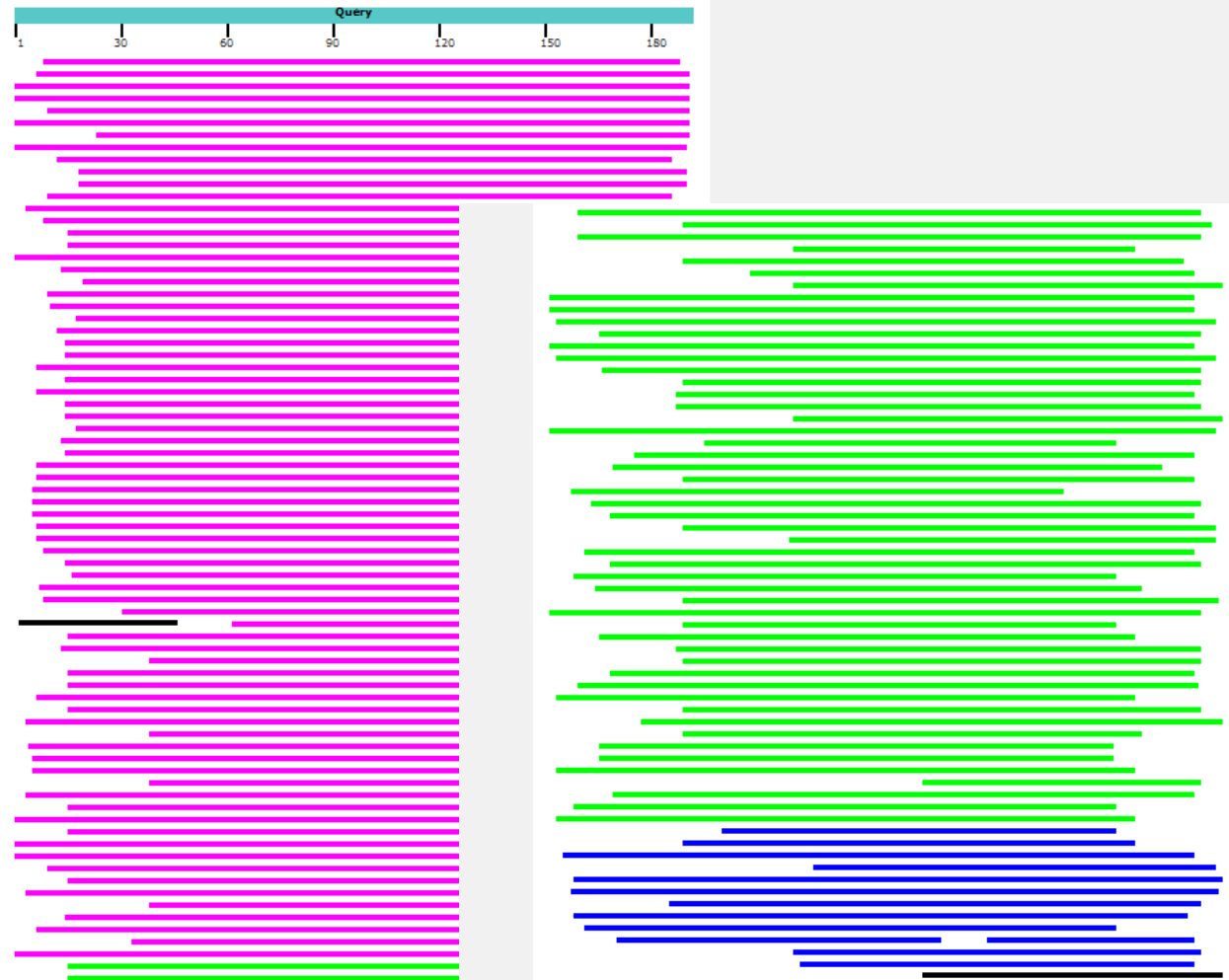


371 sequences selected [?](#)

Putative conserved domains have been detected, click on the image below for detailed results.

[CD search result summary](#)

Distribution of the top 371 Blast Hits on 371 subject sequences



Page de résultat : alignement deux à deux

Descriptions Graphic Summary **Alignments** Taxonomy

Alignment view Pairwise ?

Download ▼

371 sequences selected ?

[Download](#) ▼ [GenPept](#) [Graphics](#)

[Next](#) [Previous](#) [Descriptions](#)

osmotically-inducible protein OsmY [Thalassospira xiamenensis]

Sequence ID: [PTC01376.1](#) Length: 196 Number of Matches: 1

Range 1: 17 to 193 [GenPept](#) [Graphics](#) [Next Match](#) [Previous Match](#)

Score	Expect	Method	Identities	Positives	Gaps		
159 bits(403)	5e-47	Compositional matrix adjust.	87/179(49%)	116/179(64%)	2/179(1%)		
Query 9	VLISALLLQGCVA	AAAVVGTAAVGT	KAATDPRSVGTQ	VDDGTL	LEVRVNSALS	SKDEQIKKEA	68
	+LIS L+LQGC A	A+VG AVG	KAATD RS+GTQ+DD	T+E R A+S+	A A		
Sbjct 17	LLISVLMQLQGC	-AVALVGAGAV	GKAAATDRRSLGT	QLDDQTIEYRAQ	RAISEGSAFDN	-A	74
Query 69	RINVTAYQGGK	VLLVGGSPNAELS	SARAKQIAMGVD	GANEVYNEIRQ	GGPIGLGEAS	NDTWI	128
	R+ AY +VLLVGG	+P+ E+ +A+	G +V+NE+R +G	+ DTWI			
Sbjct 75	RVLAVAYNDRV	LVLVGGQAPSDE	IRRQAE LAVREAV	GERQVHNEVR	IRSRVGF	TTKTEDTWI	134
Query 129	TTKVRSQLLTS	DLVKSSNVKVT	TENGEVFLMGL	VTEREAKAAAD	IASRVSGVKR	VTTAF	187
	T+KV++ LL+ +	V S +KV	TENGEVFLMG+V	E + A DIA	+GV RV TAF		
Sbjct 135	TSKVKTALLS	DERVDGSGQIK	VITENGEVFLMG	IIVDAEEGR	IATDIARNTAG	VVRVVITAF	193

[Download](#) ▼ [GenPept](#) [Graphics](#)

[Next](#) [Previous](#) [Descriptions](#)

osmotically-inducible protein OsmY [Rickettsiales bacterium]

Sequence ID: [MBB20570.1](#) Length: 192 Number of Matches: 1

Range 1: 10 to 192 [GenPept](#) [Graphics](#) [Next Match](#) [Previous Match](#)

Score	Expect	Method	Identities	Positives	Gaps			
137 bits(346)	1e-38	Compositional matrix adjust.	73/184(40%)	113/184(61%)	1/184(0%)			
Query 7	I AVLISALLLQ	GCVA	AAAVVGTAAVGT	KAATDPRSVGTQ	VDDGTL	LEVRVNSALS	SKDEQIKK	66
	I L++ LLQGC A	V AA D	RS+GTQ+DD T+	R++ ALS ++ I +				
Sbjct 10	IGALLALSLLQ	GCAALVVGAG	VGA-ASAADR	RRSLGTQLDDK	TVSSRISIALS	DNKAIDE	68	
Query 67	EARINVTAYQ	GKVVLLVGGSP	NAELSARAKQ	IAMGVDGANE	VYNEIRQGGP	IGLGEASNDT	126	
	+A I+V +G	VLLVGG+PN	L +A+ A V	+++N+IR	G P+ ++D			
Sbjct 69	QANIHVHVFN	GSVLLVGGAP	NEALINQAQ	VAAATSVK	NIEKLNQIR	LGSFVSA	STTTHDV	128
Query 127	WITTKVRSQL	LTS	DLVKSSNVKVT	TENGEVFLMGL	VTEREAKAAAD	IASRVSGVKR	VTTA	186
	W+T+K+++ L+	+ ++ V	EN EVFLMGLV+	+EA+ AA +A	+ GVK+V A			
Sbjct 129	WLTSKIKANL	VADKRIDGF	HIHVAVENSE	VFLMGLVST	KEA	VAATVARN	IDGVKQVIKA	188
Query 187	FTFI	190						
	F ++							
Sbjct 189	FEYL	192						


Related Information

[Identical Proteins](#) - Identical proteins to MAD17062.1

Page de résultat : distribution taxonomique des hits

Descriptions Graphic Summary Alignments **Taxonomy**

Reports Lineage Organism **Taxonomy**

371 sequences selected 

Taxonomy	Number of hits	Number of Organisms	Description
<input type="checkbox"/> Bacteria	508	246	
<input type="checkbox"/> Proteobacteria	1	246	Proteobacteria hits
<input type="checkbox"/> Alphaproteobacteria	507	245	
<input type="checkbox"/> Rhodospirillales	236	94	
<input type="checkbox"/> Rhodospirillaceae	222	88	
<input type="checkbox"/> Thalassospira	12	7	
Thalassospira xiamenensis	1	1	Thalassospira xiamenensis hits
Thalassospira marina	3	1	Thalassospira marina hits
<input type="checkbox"/> unclassified Thalassospira	4	2	
Thalassospira sp. MCCC 1A01428	2	1	Thalassospira sp. MCCC 1A01428 hits
Thalassospira sp. TSL5-1	2	1	Thalassospira sp. TSL5-1 hits
Thalassospira australica	1	1	Thalassospira australica hits
Thalassospira profundimaris	1	1	Thalassospira profundimaris hits
Thalassospira mesophila	2	1	Thalassospira mesophila hits
<input type="checkbox"/> unclassified Rhodospirillaceae	52	5	
Rhodospirillaceae bacterium	48	1	Rhodospirillaceae bacterium hits
Rhodospirillaceae bacterium LM-1	1	1	Rhodospirillaceae bacterium LM-1 hits
Rhodospirillaceae bacterium TMED23	1	1	Rhodospirillaceae bacterium TMED23 hits
Rhodospirillaceae bacterium TMED256	1	1	Rhodospirillaceae bacterium TMED256 hits
Rhodospirillaceae bacterium TMED63	1	1	Rhodospirillaceae bacterium TMED63 hits
<input type="checkbox"/> Azospirillum	1	28	Azospirillum hits
<input type="checkbox"/> unclassified Azospirillum	1	16	unclassified Azospirillum hits
Azospirillum sp. TSH58	3	1	Azospirillum sp. TSH58 hits
Azospirillum sp. TSO5	2	1	Azospirillum sp. TSO5 hits
Azospirillum sp. TSH7	1	1	Azospirillum sp. TSH7 hits
Azospirillum sp. TSH20	1	1	Azospirillum sp. TSH20 hits
Azospirillum sp. TSA2s	1	1	Azospirillum sp. TSA2s hits
Azospirillum sp. Sh1	2	1	Azospirillum sp. Sh1 hits
Azospirillum sp. TSA6c	2	1	Azospirillum sp. TSA6c hits
Azospirillum sp. M2T2B2	2	1	Azospirillum sp. M2T2B2 hits
Azospirillum sp. TSH100	3	1	Azospirillum sp. TSH100 hits
Azospirillum sp. TSH64	2	1	Azospirillum sp. TSH64 hits
Azospirillum sp. B506	1	1	Azospirillum sp. B506 hits
Azospirillum sp. B21	2	1	Azospirillum sp. B21 hits
Azospirillum sp. TSO35-2	2	1	Azospirillum sp. TSO35-2 hits
Azospirillum sp. K2W22B-5	2	1	Azospirillum sp. K2W22B-5 hits

Page de résultat : application d'un filtre

i Your search is limited to records that include: Alphabacteria (taxid:28211)
Your results are filtered to match records that include: Rickettsiales (taxid:766)

Job Title	EcolA01.YRAP		
RID	38G2CGA901R	Search expires on 02-01 21:57 pm	Download All ▾
Program	BLASTP ?	Citation ▾	
Database	nr	See details ▾	
Query ID	lcl Query_4796		
Description	EcolA01.YRAP		
Molecule type	amino acid		
Query Length	191		
Other reports	Distance tree of results	Multiple alignment	MSA viewer ?

Filter Results

Organism *only top 20 will appear* exclude

Rickettsiales (taxid:766)

[+ Add organism](#)

Percent Identity

 to

E value

 to

Query Coverage

 to

Filter

Reset

Descriptions | [Graphic Summary](#) | [Alignments](#) | [Taxonomy](#)

Sequences producing significant alignments

[Download](#) ▾ [Manage Columns](#) ▾ Show [?](#)

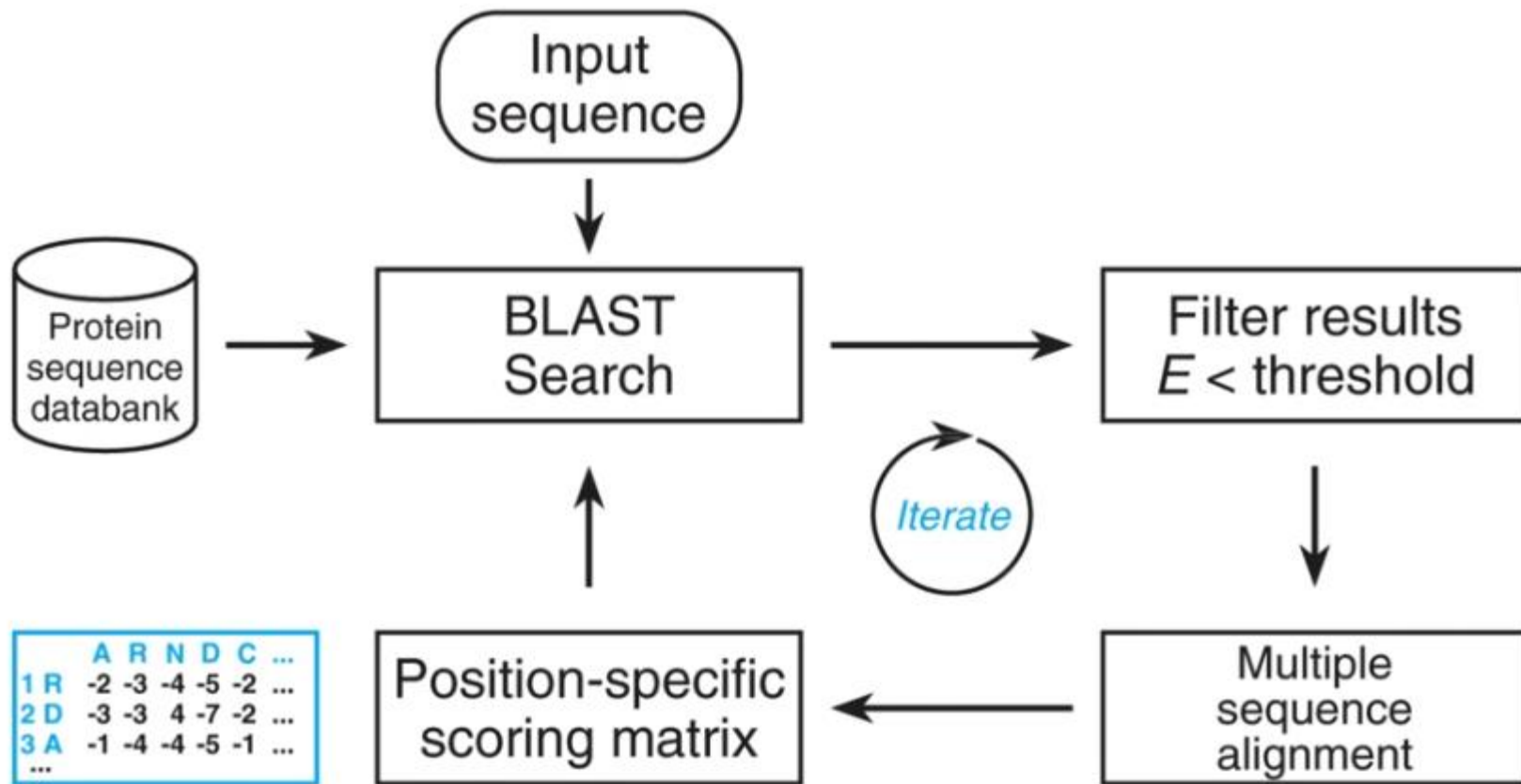
select all 20 sequences selected

[GenPept](#) | [Graphics](#) | [Distance tree of results](#) | [Multiple alignment](#)

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
<input checked="" type="checkbox"/>	osmotically-inducible protein OsmY [Rickettsiales bacterium]	137	137	96%	1e-38	39.67%	MBB20570.1
<input checked="" type="checkbox"/>	BON domain-containing protein [Rickettsiales bacterium]	123	123	99%	7e-33	39.69%	MBB66425.1
<input checked="" type="checkbox"/>	transport-associated protein [Rickettsiales bacterium]	101	101	87%	2e-24	37.13%	MBV35553.1
<input checked="" type="checkbox"/>	Uncharacterized protein in gshII 5' region [Candidatus Jidaibacter acanthamoeba]	100	100	98%	1e-23	33.51%	KIE04947.1
<input checked="" type="checkbox"/>	BON domain-containing protein [Candidatus Jidaibacter acanthamoeba]	98.6	98.6	89%	4e-23	35.47%	WP_039456942.1
<input checked="" type="checkbox"/>	BON domain-containing protein [Candidatus Aquarickettsia rohweri]	82.4	82.4	91%	5e-17	32.22%	WP_128044820.1
<input checked="" type="checkbox"/>	Osmotically-inducible protein Y precursor [Candidatus Arcanobacter lacustris]	82.0	82.0	94%	8e-17	30.27%	KKB96100.1
<input checked="" type="checkbox"/>	BON domain-containing protein [Candidatus Phycoarickettsia trachydisci]	80.5	80.5	96%	2e-16	30.16%	WP_106874613.1
<input checked="" type="checkbox"/>	BON domain-containing protein [Candidatus Hepatobacter penaei]	74.3	74.3	87%	7e-14	32.94%	WP_031934417.1
<input checked="" type="checkbox"/>	hypothetical protein [Rickettsiales bacterium]	74.3	74.3	96%	9e-14	31.75%	MBO90079.1
<input checked="" type="checkbox"/>	BON domain protein [Neorickettsia helminthoeca str. Oregon]	70.1	70.1	96%	2e-12	28.57%	AHX11031.1
<input checked="" type="checkbox"/>	BON domain-containing protein [Neorickettsia helminthoeca]	70.1	70.1	96%	3e-12	28.57%	WP_051579482.1
<input checked="" type="checkbox"/>	phospholipid-binding protein [Rickettsiales bacterium TMED174]	68.6	68.6	94%	8e-12	25.41%	OUW24413.1
<input checked="" type="checkbox"/>	BON domain-containing protein [Rickettsia typhi]	68.9	68.9	92%	8e-12	29.12%	WP_011190840.1
<input checked="" type="checkbox"/>	BON domain-containing protein [Candidatus Midichloria mitochondrii]	67.0	67.0	89%	4e-11	28.07%	WP_013951652.1
<input checked="" type="checkbox"/>	hypothetical protein [Rickettsiales bacterium]	61.6	61.6	92%	4e-09	20.79%	MBH43598.1
<input checked="" type="checkbox"/>	hypothetical protein [Rickettsiales bacterium]	60.8	60.8	93%	6e-09	23.76%	MAY90257.1
<input checked="" type="checkbox"/>	hypothetical protein [Rickettsiales bacterium]	56.2	56.2	73%	3e-07	30.28%	MAZ46716.1
<input checked="" type="checkbox"/>	hypothetical protein [Rickettsiales bacterium]	55.1	55.1	96%	8e-07	25.27%	MBS91946.1
<input checked="" type="checkbox"/>	hypothetical protein [Rickettsiales bacterium]	53.1	53.1	86%	4e-06	28.74%	MBC11294.1

PSI-BLAST (Position-Specific Iterative Basic Local Alignment Search Tool)

Psiblast



Extrait de la page résultat du premier run

<input checked="" type="checkbox"/>	BON domain-containing protein [Magnetospira sp. QH-2]	45.8	45.8	80%	0.001	29.84%	WP_048020069.1	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	BON domain-containing protein [Rhodospirillaceae bacterium RKSG073]	45.8	45.8	94%	0.002	23.63%	WP_155154364.1	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	BON domain-containing protein [Caenispirillum salinarum]	45.8	45.8	51%	0.002	28.00%	WP_009540523.1	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	BON domain-containing protein [SAR116 cluster bacterium]	45.8	45.8	78%	0.002	26.00%	RCL78867.1	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	BON domain-containing protein [Magnetospirillum sp. ME-1]	45.8	45.8	76%	0.002	28.05%	WP_085372472.1	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	BON domain-containing protein [Alphaproteobacteria bacterium]	45.4	45.4	59%	0.002	30.43%	NCF45983.1	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	BON domain-containing protein [Alphaproteobacteria bacterium]	45.4	45.4	82%	0.002	25.95%	NBR38031.1	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	BON domain-containing protein [Pelagibacteraceae bacterium]	45.4	45.4	83%	0.002	21.88%	MBH71125.1	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	BON domain-containing protein [Alphaproteobacteria bacterium]	45.4	45.4	49%	0.002	29.79%	MQA65521.1	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	phospholipid-binding domain-containing protein [Rhodospirillaceae bacterium]	45.1	45.1	83%	0.003	26.06%	MAI12765.1	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	BON domain-containing protein [Roseospirillum parvum]	45.1	45.1	79%	0.003	27.33%	WP_092617644.1	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	hypothetical protein CFH02_01694 [Alphaproteobacteria bacterium MarineAlpha3_Bin1]	44.7	44.7	80%	0.004	27.04%	PPR66808.1	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	BON domain-containing protein [Magnetospirillum moscoviense]	43.1	78.2	87%	0.004	31.07%	WP_068499403.1	<input checked="" type="checkbox"/>

Run PSI-BLAST Iteration 2 with max number of sequences

Sequences with E-value WORSE than threshold

select all 72 sequences selected

PSI-BLAST iteration 1

	Description	Max score	Total score	Query cover	E value	Per. Ident	Accession	Select for PSI blast	Used to build PSSM	Newly added
<input checked="" type="checkbox"/>	BON domain-containing protein [Ensifer sp. WSM1721]	44.3	44.3	39%	0.005	40.00%	WP_051440915.1	<input type="checkbox"/>		
<input checked="" type="checkbox"/>	Osmotically-inducible protein OsmY, contains BON domain [Roseospirillum parvum]	44.7	44.7	79%	0.005	27.33%	SDH05704.1	<input type="checkbox"/>		
<input checked="" type="checkbox"/>	BON domain-containing protein [Enhydrobacter aerosaccus]	42.7	42.7	38%	0.008	36.00%	SKA25257.1	<input type="checkbox"/>		
<input checked="" type="checkbox"/>	hypothetical protein [Rhodospirillaceae bacterium]	43.9	43.9	50%	0.008	27.45%	MAF49290.1	<input type="checkbox"/>		
<input checked="" type="checkbox"/>	BON domain-containing protein [Alphaproteobacteria bacterium]	42.7	42.7	49%	0.008	29.47%	MPY71064.1	<input type="checkbox"/>		

Extrait de la page résultat du premier run

<input checked="" type="checkbox"/>	BON domain protein [Rickettsia amblyommatis str. Darkwater]	151	151	78%	3e-44	26.67%	KJV94938.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	BON domain-containing protein [Rhodospira trueperi]	153	153	84%	3e-44	25.15%	WP_092780429.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	BON domain-containing protein [Alphaproteobacteria bacterium]	152	203	95%	3e-44	24.46%	NBQ83556.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	BON domain-containing protein [Candidatus Micropelagos thuwalensis]	153	198	98%	4e-44	25.26%	WVP_021777646.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	putative phospholipid-binding domain protein [Wolbachia pipientis wUni]	150	150	73%	4e-44	32.87%	ONI56871.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	21 kDa hemolysin protein, putative [Rhodospirillum centenum SW]	152	152	86%	4e-44	28.12%	ACJ00819.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	BON domain-containing protein [alpha proteobacterium IMCC14465]	152	202	96%	6e-44	27.27%	WVP_008518969.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	Transport-associated domain [Ehrlichia minasensis]	151	151	90%	8e-44	28.00%	CEI85043.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	BON domain protein [Rickettsia amblyommatis str. Ao/Pa]	150	150	76%	8e-44	26.67%	KJV82027.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	BON domain-containing protein [Marinicauda pacifica]	151	197	96%	8e-44	29.57%	WVP_135945713.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	Transport-associated domain [Ehrlichia canis str. Jake]	151	151	90%	9e-44	26.86%	AAZ68703.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	BON domain-containing protein [PS1 clade bacterium]	153	199	98%	1e-43	23.68%	RCL83276.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	hypothetical protein [Rhizobiales bacterium]	153	199	98%	1e-43	23.68%	MAV18645.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	hypothetical protein [Rickettsiales bacterium]	151	151	100%	1e-43	23.96%	MAI83460.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	BON domain-containing protein [Amphiplicatus metriotherophilus]	150	150	79%	1e-43	24.34%	WVP_089410959.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	Putative phospholipid-binding domain protein (modular protein) [uncultured Alphaproteobacteria bacterium]	152	192	97%	1e-43	22.87%	SBW08860.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	RlpA-like protein [Rickettsia endosymbiont of Proechinophirus fluctus]	149	184	95%	1e-43	23.17%	KYP96821.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	TPA: hypothetical protein [Rhodobiaceae bacterium]	152	199	98%	2e-43	23.68%	HCN31600.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	BON domain-containing protein [Oleomonas sp. K1W22B-S]	151	210	80%	2e-43	32.90%	RJF89784.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	hypothetical protein [Rhodospirillaceae bacterium]	149	149	90%	2e-43	24.57%	MAJ92173.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	hypothetical protein COC03_03675 [Robiginitomaculum sp.]	153	198	98%	2e-43	25.79%	PHQ59960.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	hypothetical protein COB92_02500 [Robiginitomaculum sp.]	153	198	98%	2e-43	26.32%	PHQ67920.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	BON domain-containing protein [Henriciella marina]	151	151	97%	2e-43	32.96%	WVP_156807951.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	phospholipid-binding protein [Pelagibacteraceae bacterium]	150	150	91%	2e-43	25.84%	MAV49112.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	hypothetical protein [Rickettsiales bacterium]	150	188	99%	3e-43	22.51%	MBS91946.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	BON domain-containing protein [alpha proteobacterium HIMB114]	150	150	91%	3e-43	26.97%	WVP_009373058.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	phospholipid-binding protein [Candidatus Pelagibacter sp.]	150	150	89%	3e-43	28.65%	MBL60633.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	BON domain-containing protein [Oleisolibacter albus]	150	150	84%	3e-43	27.78%	WVP_114394999.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	hypothetical protein COA91_06815 [Robiginitomaculum sp.]	153	200	98%	3e-43	24.06%	PHS39033.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	hypothetical protein [Rhodospirillaceae bacterium]	150	150	92%	4e-43	18.08%	MBB42584.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	hypothetical protein [Rickettsiales bacterium]	149	149	97%	4e-43	24.06%	MAT32797.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	BON domain-containing protein [Oleomonas sp. K1W22B-S]	149	207	80%	7e-43	32.90%	WVP_158596205.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	BON domain-containing protein [Novispirillum itersonii]	149	149	89%	1e-42	17.92%	WVP_019645911.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	BON domain-containing protein [Marinicauda algicola]	148	148	76%	2e-42	28.19%	WVP_135995419.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	hypothetical protein COB56_07460 [Robiginitomaculum sp.]	150	194	96%	3e-42	26.32%	PHS75645.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	BON domain-containing protein [Rhizobiales bacterium]	147	147	100%	3e-42	37.19%	RYE75761.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	MULTISPECIES: BON domain-containing protein [unclassified Azospirillum]	147	147	57%	3e-42	33.33%	WVP_160105555.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	BON domain-containing protein [Haematospirillum jordaniae]	149	149	87%	4e-42	23.86%	WVP_082611849.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	TPA: phospholipid-binding protein [Hyphomonas atlantica]	146	146	68%	4e-42	35.38%	HBH44370.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>

En jaune les nouvelles séquences détectées avec la PSSM

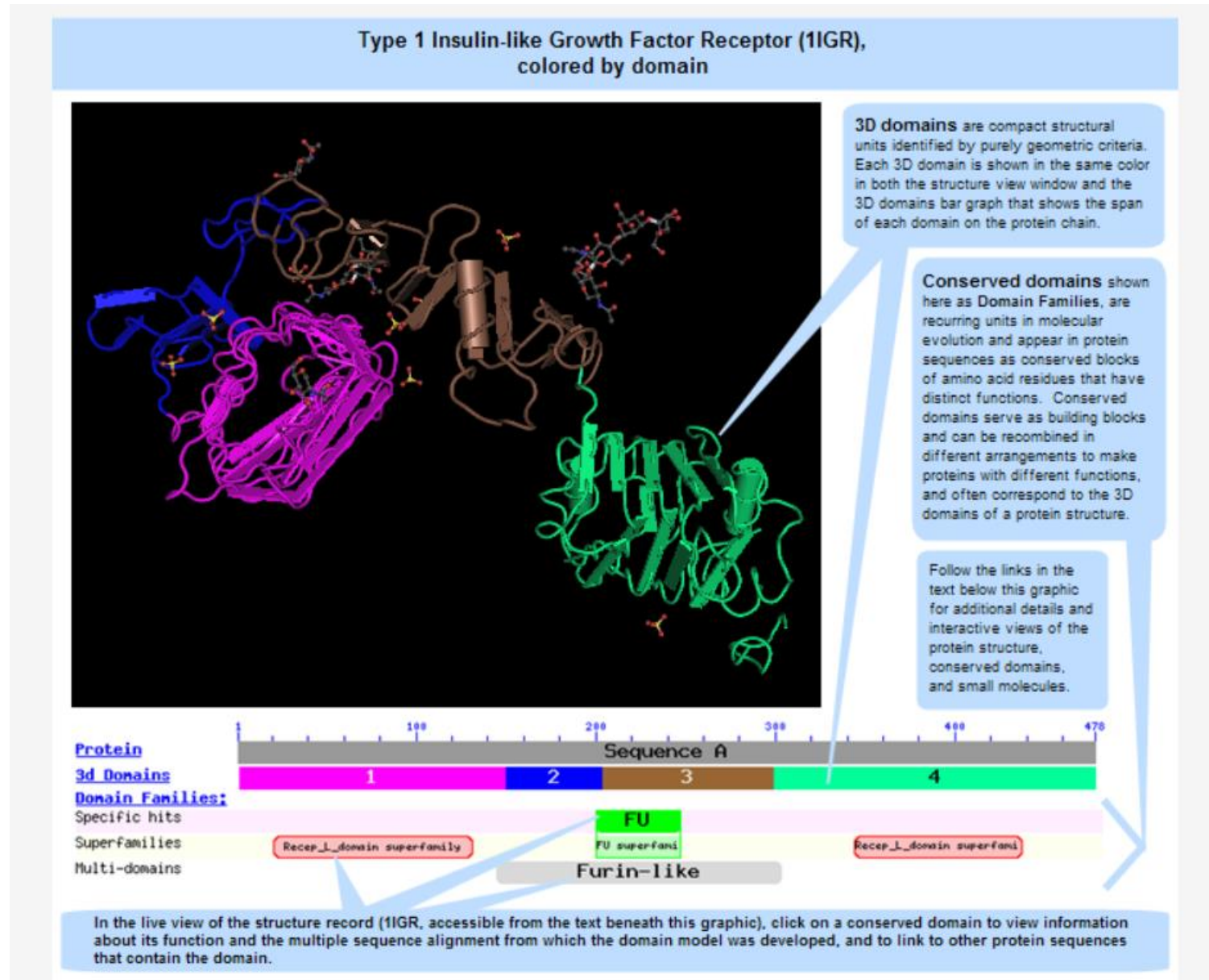
Domaines fonctionnels

Deux définitions des domaines : domaines structuraux et domaines fonctionnels. Ils constituent des parties de la protéines. Un domaine structural est une partie de la chaîne polypeptidique qui se replie indépendamment. Un domaine fonctionnel est une région de la protéine qui présente une conservation de séquence mise en évidence par alignement multiple auquel on peut associer une fonction. Il forme une « brique » qui a pu être recombinaisonnée dans différents arrangements pour moduler l'expression des protéines au cours de l'évolution.

Les deux classifications des domaines fonctionnels et structuraux coïncident assez souvent.

Les domaines fonctionnels peuvent contenir des motifs (ou pattern) fonctionnels. La différence est souvent liée à leur taille, plus petite pour les motifs.

Exemple de coïncidence entre domaines fonctionnels et structuraux



Banque de données ProSite

ProSite consiste en un ensemble d'entrées décrivant les domaines protéiques et les motifs caractéristiques de fonctions ou de familles protéiques.

Une entrée Prosite est constituée de deux parties :

- une fiche qui fournit une description des domaines et des motifs fonctionnels et renseigne sur la fonction associée au domaine ou motif. Cette fiche a un préfixe PDOC (exemple : PDOC00185)
- Une fiche décrivant le motif ou le domaine qui a un préfixe PS (exemple PS50893)

This form allows you to scan proteins for matches against the PROSITE collection of motifs as well as against your own patterns.

- Option 1 - Submit PROTEIN sequences to scan them against the PROSITE collection of motifs.
- Option 2 - Submit MOTIFS to scan them against a PROTEIN sequence database.
- Option 3 - Submit PROTEIN sequences and MOTIFS to scan them against each other.

STEP 1 - Submit PROTEIN sequences [\[help\]](#)

- Submit PROTEIN sequences (max. 10) [Examples](#)
- Submit a PROTEIN database (max. 16MB) for repeated scans (The data will be stored on our server for 1 month).

Enter UniProtKB accessions or identifiers or PDB identifiers or sequences in FASTA format

Supported input:

- UniProtKB accessions e.g. P98073 or Identifiers e.g. ENTK_HUMAN
- PDB Identifiers e.g. 4DGJ
- Sequences in FASTA format

Séquence(s) à analyser
(max 10)

STEP 2 - Select options [\[help\]](#)

- Exclude motifs with a high probability of occurrence from the scan
- Exclude profiles from the scan
- Run the scan at high sensitivity (show weak matches for profiles)

Exclure de l'analyse les motifs
avec une forte probabilité
d'occurence

STEP 3 - Select output options and submit your job

Output format:

Retrieve complete sequences: if you choose this option, not all output formats are available.

Receive your results by email

```
ID ASN_GLYCOSYLATION; PATTERN.  
AC PS00001;  
DE N-glycosylation site.  
PA N-{P}-[ST]-{P}
```

This form allows you to scan proteins for matches against the [PROSITE collection of motifs](#) as well as against your own patterns.

- Option 1 - Submit PROTEIN sequences to scan them against the PROSITE collection of motifs.
- Option 2 - Submit MOTIFS to scan them against a PROTEIN sequence database.**
- Option 3 - Submit PROTEIN sequences and MOTIFS to scan them against each other.

[Reset](#)

STEP 1 - Enter a MOTIF or a combination of MOTIFS [Examples](#) [\[help\]](#)

Enter a PROSITE accession or identifier or your own pattern or a combination

Supported input:

- A PROSITE accession e.g. [PSS0240](#) or Identifier e.g. [TRYPSIN_DOM](#)
- Your own pattern e.g. [P-x\(2\)-G-E-S-G\(2\)-\[AS\]](#)

» More

Options « [\[help\]](#)

- Minimal number of hits per matched sequences:
- Profile option
 - Run the scan at a high sensitivity (show weak matches for profiles)
- Pattern options
 - Number of X characters in a scanned sequence that can be matched by a conserved position in a pattern:
 - Match mode:

Motif ou combinaison de motifs à rechercher. Soit un numéro d'accèsion dans ProSite, soit votre propre motif (format ProSite)

STEP2 - Select a PROTEIN sequence database [\[help\]](#)

- UniProtKB
 - Swiss-Prot Include splice variants
 - TrEMBL
- PDB
- Your protein database
- Randomized UniProtKB/Swiss-Prot

Exclude fragments (concerns UniProtKB only)

» [Filters](#) [\[help\]](#)

Choix de la banque de séquences protéiques à analyser

STEP 3 - Select output options and submit your job

Output format:

Maximum number of displayed matches: If you select 100'000, results are returned by email and not all output formats are available.

Retrieve complete sequences: if you choose this option, a maximum of 1'000 matched sequences can be displayed and not all output formats are available.

Receive your results by email

[START THE SCAN](#) [Reset](#)

Domaines et motifs fonctionnels par l'exemple



ScanProsite tool

Analyse de la séquence protéique ComA de *Streptococcus pneumoniae* souche R6

This form allows you to scan proteins for matches against the PROSITE collection of motifs as well as against your own patterns.

- Option 1 - Submit PROTEIN sequences to scan them against the PROSITE collection of motifs.
- Option 2 - Submit MOTIFS to scan them against a PROTEIN sequence database.
- Option 3 - Submit PROTEIN sequences and MOTIFS to scan them against each other.

Reset

STEP 1 - Submit PROTEIN sequences [help]

- Submit PROTEIN sequences (max. 10) [Examples](#)
- Submit a PROTEIN database (max. 16MB) for repeated scans (The data will be stored on our server for 1 month).

```
>SpneA01.COMA|
MKFGKRHYRPOVDQMDCGVASLAMVFGYYGSYYFLAHLRELAKTMDGTTALGLVKVAEE
IGFETRAIKADMILFDLPDLTFPFVAHVLEKGGKLLHYVVVTGQDKDSIHIADFPQVKLI
KLPRERFEEEWIGVTLFMAPSDDYKPKHEQRNGLLSFTIPILVKQRGLIANIVLAILLVTV
INIVGSSVYVLSIIDIYVDPQMRSTLGIISIGLVIVVYILQQILSYAQEYLLLVLGQRLSID
VILSYIKHVHFLPMSFFATRRTGEIVSRETDANSIIDALASTILSIFLDVSTVVIISIVL
FSQNTNLFMTLLALPFIYTVIIFAFMKPFERKQNRDTMEANAVLSSSIIEDINGIETIKSL
TSESQRVYQKIDKEFVDYLLKSFYTSRAESQQKALKVVAHLLLNVCILWMCVAVLMDGKMS
LQQLITVNTLLVYFNPLENIINLQTKLQTAQVANNRLNEVYLVASEFEEKKTVEDLSIM
KGDMTFKQVHYKYGYGRDVLSDINLIVPQSGKVAEFGISGSGKITLAKMMVNEYDPSQGE
ISLGGVNLNQIDKKALRQYINVLQQQPVVFNGLILENLLGAKEGTQRDILRAVELASL
RSDTFRMPLVYQTEFTEGACTECCGQRIALABATLFRAPHTTDEATSEETDITTEPRT
```

Supported input:

- UniProtKB accessions e.g. P98073 or identifiers e.g. ENTK_HUMAN
- PDB identifiers e.g. 4DGJ
- Sequences in FASTA format

STEP 2 - Select options [help]

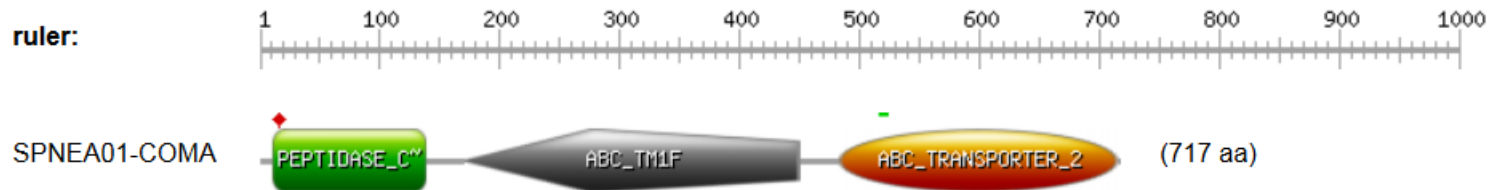
- Exclude motifs with a high probability of occurrence from the scan
- Exclude profiles from the scan
- Run the scan at high sensitivity (show weak matches for profiles)

Domaines et motifs fonctionnels par l'exemple

Domaines fonctionnels identifiés dans la séquence ComA de *S. pneumoniae*

hits by profiles: [3 hits (by 3 distinct profiles) on 1 sequence]

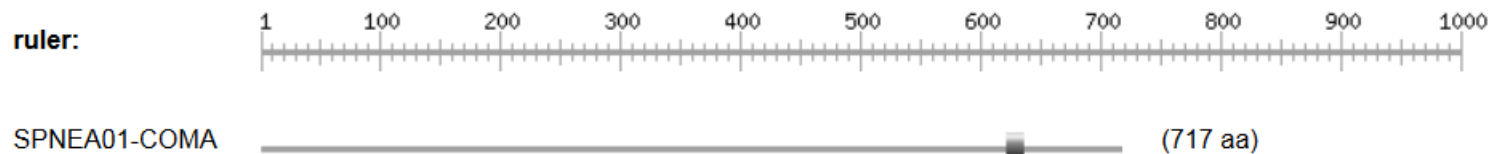
Upper case represents match positions, lower case insert positions, and the '-' symbol represents deletions relative to the matching profile.



Motif fonctionnel identifié dans la séquence ComA de *S. pneumoniae*

hits by patterns: [1 hit (by 1 pattern) on 1 sequence]

Hits by [PS00211](#) ABC_TRANSPORTER_1 ABC transporters family signature :



622 - 636: [confidence level: (0)] ISGGQRQRIALARAL

Entrée ProSite associées au domaine fonctionnel ABC_TRANSPORTER_2

ABC_TRANSPORTER_2, [PS50893](#); ATP-binding cassette, ABC transporter-type domain profile (MATRIX)

- Sequences in UniProtKB/Swiss-Prot known to belong to this class: 3998
 - detected by PS50893: 3983 (true positives)
 - undetected by PS50893: 15 (5 false negatives and 10 'partials')
- Other sequence(s) in UniProtKB/Swiss-Prot detected by PS50893:
NONE.
- [Domain architecture view of Swiss-Prot proteins matching PS50893](#)



- Retrieve an alignment of UniProtKB/Swiss-Prot true positive hits:
[Clustal format, color, condensed view](#) / [Clustal format, color](#) / [Clustal format, plain text](#) / [Fasta format](#)
- [Retrieve the sequence logo from the alignment](#)
- [Taxonomic distribution of all UniProtKB \(Swiss-Prot + TrEMBL\) entries matching PS50893](#)
- [Retrieve a list of all UniProtKB \(Swiss-Prot + TrEMBL\) entries matching PS50893](#)
- [Scan UniProtKB \(Swiss-Prot and/or TrEMBL\) entries against PS50893](#)
- [View ligand binding statistics of PS50893](#)
- [Matching PDB structures: 1B0U 1F3O 1G29 1G6H ... \[ALL\]](#)

Extrait de la matrice (profil) ProSite associées au domaine fonctionnel ABC_TRANSPORTER_2

Matrix / Profile
[info]

```
/GENERAL_SPEC: ALPHABET='ABCDEFGHIJKLMNPQRSTVWYZ'; LENGTH=240;
/DISJOINT: DEFINITION=PROTECT; N1=6; N2=235;
/NORMALIZATION: MODE=1; FUNCTION=LINEAR; R1=5.7291522; R2=0.0066693; TEXT='NScore';
/NORMALIZATION: MODE=-1; FUNCTION=LINEAR; R1=27511.1894531; R2=19.6396694; TEXT='Heuristic 5.0%';
/CUT_OFF: LEVEL=0; SCORE=431; H_SCORE=35976; N_SCORE=8.6; MODE=1; TEXT='!';
/CUT_OFF: LEVEL=-1; SCORE=116; H_SCORE=29789; N_SCORE=6.5; MODE=1; TEXT='?';
/DEFAULT: M0=-8; D=-20; I=-20; B0=*; B1=*; E0=*; E1=*; MI=-105; MD=-105; IM=-105; DM=-105;

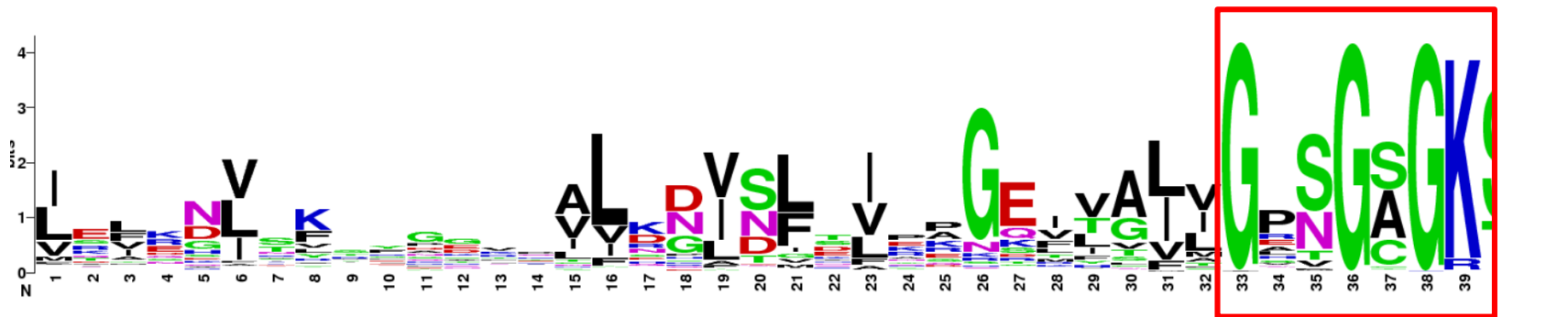
          A  B  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y  Z
/I:      B0=0; B1=0; BI=-105; BD=-105;
/M: SY='I'; M= -6,-30,-18,-32,-26, 4,-32,-26, 29,-26, 29, 16,-28,-28,-24,-22,-20, -6, 29,-24, -4,-26;
/M: SY='R'; M=-16, -4,-30, -4, 6,-26,-20, 19,-28, 22,-22, -6, 2,-16, 21, 38, -8,-12,-24,-22, -5, 10;
/M: SY='L'; M=-10,-10,-18,-11,-14, 8,-25,-18, 2,-19, 11, 0,-13,-22,-20,-16, -9, 4, 5,-22, -2,-16;
/M: SY='E'; M=-10, 1,-27, 5, 12,-24,-23, -9, -8, 3,-12, -6, -4,-12, 9, -3, -6, -8, -7,-27,-12, 10;
/M: SY='D'; M=-12, 38,-23, 41, 10,-30, -5, 2,-30, -2,-30,-25, 32,-13, 0, -7, 10, -2,-27,-40,-20, 5;
/M: SY='V'; M= -5,-30,-17,-32,-27, 3,-32,-27, 31,-25, 25, 15,-28,-28,-25,-22,-19, -5, 33,-25, -5,-27;
/M: SY='F'; M= -8,-18,-21,-21,-17, 7,-20, -4, -6,-17, -9, -6,-13,-22,-16,-14, -3, 1, 0, 3, 6,-16;
/M: SY='K'; M=-13,-10,-27,-14, -4, 8,-23,-13,-20, 22,-16, -7, -7,-17, -7, 13,-13,-10,-13,-10, 4, -4;
/M: SY='T'; M= 3,-12,-17,-15,-12, -5,-17, -8, -6, -5,-10, -6, -9,-19, -8, -1, 4, 6, 1,-15, 6,-12;
/I:      I=-5; MI=0; MD=-27; IM=0; DM=-27;
/M: SY='F'; M=-17,-20,-25,-22, -9, 28,-27, -8, -6,-11, 6, -1,-15,-24,-15, 0,-17,-10, -7, -3, 19,-12;
/M: SY='G'; M= -5,-12,-30, -9, -6,-23, 25,-17,-25,-16,-15,-13, -9, 1,-13,-17, -7,-15,-23,-23,-24,-10;
/M: SY='E'; M= -9, 5,-30, 11, 15,-28, 8,-10,-24, -7,-20,-16, 1,-11, -2,-12, -3,-13,-22,-27,-20, 6;
/M: SY='V'; M= -9,-20,-18,-26,-22, 12,-28, -6, 10,-20, 3, 4,-15,-24,-21,-17, -8, 2, 16,-20, 4,-22;
/M: SY='E'; M=-10,-10,-27, -8, 3,-12,-23, -9,-12, 2, -7, -6,-11, 2, -3, -3, -9, -1,-13,-18, -1, -1;
/M: SY='A'; M= 25,-20,-10,-25,-20,-10,-15,-25, 10,-15, 0, 0,-20,-20,-20,-20, 0, 0, 25,-25,-15,-20;
/M: SY='L'; M= -8,-30,-18,-30,-22, 8,-30,-22, 22,-28, 44, 18,-30,-30,-22,-20,-27, -8, 16,-22, -2,-22;
/M: SY='K'; M=-14, 9,-30, 13, 12,-32,-18, -3,-30, 29,-27,-13, 4,-12, 17, 24, -6,-10,-24,-24,-12, 14;
/M: SY='G'; M= -6, 10,-29, 12, 2,-30, 33,-10,-36, -9,-28,-22, 11,-15, -8,-13, 1,-14,-30,-28,-25, -3;
/M: SY='V'; M= -5,-28,-17,-32,-27, 2,-30,-24, 30,-22, 20, 21,-27,-27,-22,-20,-17, -5, 33,-25, -5,-25;
/M: SY='S'; M= 2, 12,-13, 4, -2,-18, -4, -6,-18, -7,-26,-18, 23,-13, -2, -7, 27, 19,-14,-38,-18, -2;
/M: SY='F'; M=-13,-30,-22,-36,-26, 32,-32,-22, 20,-30, 30, 13,-24,-28,-27,-22,-24,-10, 11,-10, 10,-26;
/M: SY='E'; M= -4, 7,-21, 4, 12,-24,-13, -5,-21, 7,-23,-14, 10,-10, 12, 3, 12, 9,-19,-30,-15, 12;
/M: SY='V'; M= 3,-27,-18,-31,-25, -1,-28,-27, 28,-23, 17, 12,-24,-24,-22,-23,-14, -5, 29,-24, -7,-25;
/M: SY='R'; M=-13, 7,-28, 4, 13,-25,-16, -1,-28, 26,-25,-14, 13,-14, 10, 32, -4, -8,-24,-26,-14, 10;
/M: SY='K'; M= -4, 2,-29, 7, 12,-28,-15, -9,-27, 14,-24,-16, -2, 8, 3, 10, -4, -8,-22,-26,-18, 6;
/M: SY='G'; M= 0,-10,-30,-10,-20,-30, 70,-20,-40,-20,-30,-20, 0,-20,-20,-20, 0,-20,-30,-20,-30,-20;
/M: SY='E'; M=-14, 23,-30, 35, 39,-35,-16, 2,-32, 6,-24,-20, 7, -5, 19, -2, 0,-10,-30,-32,-18, 29;
/M: SY='V'; M= -7,-30,-15,-33,-28, 25,-30,-25, 20,-25, 17, 9,-27,-30,-31,-20,-16, -5, 29,-17, 3,-28;
/M: SY='T'; M= -5,-17,-16,-24,-19, -2,-26,-20, 14,-18, 12, 13,-15,-19,-14,-16, -5, 15, 14,-25, -5,-17;
/M: SY='A'; M= 16,-15,-15,-20,-12,-10,-13,-19, 3,-17, 2, -2,-10,-16,-11,-19, 5, 2, 4,-26,-13,-12;
/M: SY='I'; M= -8,-30,-23,-35,-27, 3,-35,-27, 37,-28, 28, 18,-25,-25,-22,-25,-22, -8, 27,-22, -2,-27;
/I:      I=-6; MD=-32;
```

Entrée ProSite associées au motif fonctionnel ABC_TRANSPORTER_1

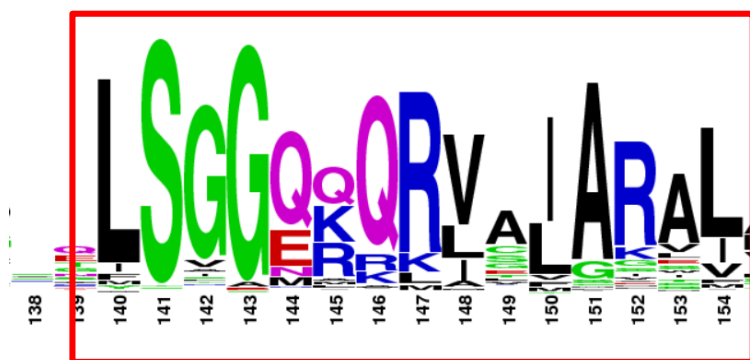
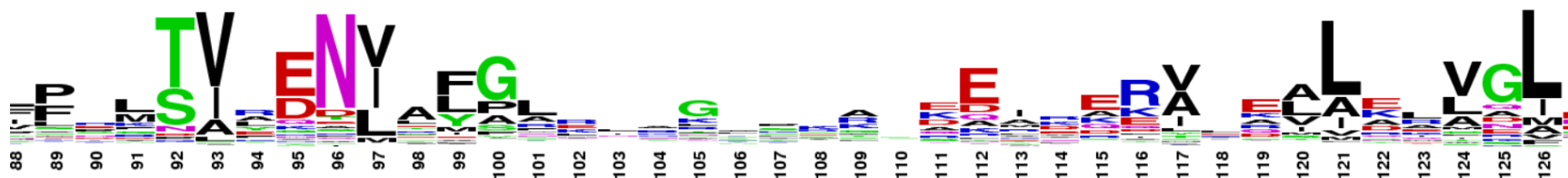
ABC_TRANSPORTER_1, [PS00211](#); ABC transporters family signature (PATTERN)

- Consensus pattern:
[LIVMFYC]-[SA]-[SAPGLVFKQH]-G-[DENQMW]-[KRQASPCLIMFW]-[KRNQSTAVM]-[KRACLVM]-[LIVMFYPAN]-{PHY}-
[LIVMFW]-[SAGCLIVP]-{FYWHP}-{KRHP}-[LIVMFYWSTA]
- Sequences in UniProtKB/Swiss-Prot known to belong to this class: 4003
 - detected by PS00211: [3668](#) (true positives)
 - undetected by PS00211: 335 ([327](#) false negatives and 8 'partials')
- Other sequence(s) in UniProtKB/Swiss-Prot detected by PS00211:
[201](#) false positives.
- Retrieve an alignment of UniProtKB/Swiss-Prot true positive hits:
[Clustal format, color, condensed view](#) / [Clustal format, color](#) / [Clustal format, plain text](#) / [Fasta format](#)
- [Retrieve the sequence logo from the alignment](#)
- [Taxonomic distribution of all UniProtKB \(Swiss-Prot + TrEMBL\) entries matching PS00211](#)
- [Retrieve a list of all UniProtKB \(Swiss-Prot + TrEMBL\) entries matching PS00211](#)
- [Scan UniProtKB \(Swiss-Prot and/or TrEMBL\) entries against PS00211](#)
- [View ligand binding statistics of PS00211](#)
- [Matching PDB structures: 1B0U 1F3O 1G29 1G6H ... \[ALL\]](#)

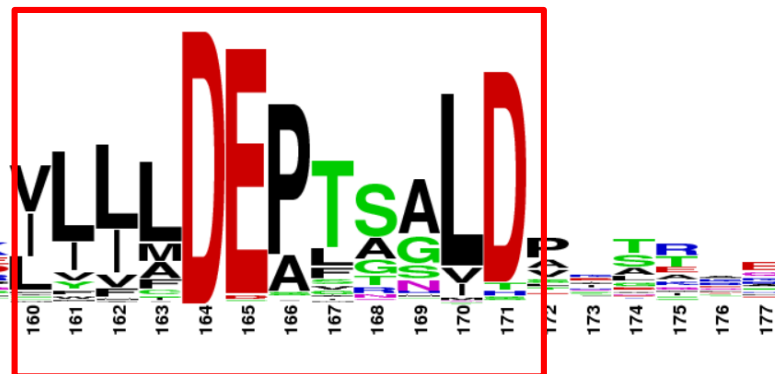
Extrait du logo du domaine fonctionnel ABC_TRANSPORTER_2 de ProSite



Motif Walker A



Motif ABC transporters family signature



Motif Walker B

Banque de données Pfam : banque de domaines fonctionnels

La banque de données Pfam est une large collection de familles de protéines représentées par des alignements multiples et des modèles de Markov cachés.

Les protéines sont généralement composée d'une ou plusieurs régions fonctionnelles, appelées domaines. Différentes combinaisons de domaines donnent naissance aux différentes protéines trouvées dans la nature. L'identification des domaines présents dans une protéine permet donc d'avoir des idées sur sa fonction.

2 sections dans Pfam:

Pfam-A : entrées de très grande qualité produite par des experts

Pfam-B : entrées produites par une procédure automatisée.

Page d'entrée de Pfam

<http://pfam.xfam.org/>



[HOME](#) | [SEARCH](#) | [BROWSE](#) | [FTP](#) | [HELP](#) | [ABOUT](#)



Pfam 30.0 (June 2016, 16306 entries)

The Pfam database is a large collection of protein families, each represented by **multiple sequence alignments** and **hidden Markov models (HMMs)**. [More...](#)

QUICK LINKS

[SEQUENCE SEARCH](#)

[VIEW A PFAM ENTRY](#)

[VIEW A CLAN](#)

[VIEW A SEQUENCE](#)

[VIEW A STRUCTURE](#)

[KEYWORD SEARCH](#)

[JUMP TO](#)

YOU CAN FIND DATA IN PFAM IN VARIOUS WAYS...

Analyze your protein sequence for Pfam matches

View Pfam annotation and alignments

See groups of related entries

Look at the domain organisation of a protein sequence

Find the domains on a PDB structure

Query Pfam by keywords

Enter any type of accession or ID to jump to the page for a Pfam entry or clan, UniProt sequence, PDB structure, etc.

Or view the [help](#) pages for more information

Analyser le contenu en domaines d'une séquence protéique

Analyse de la séquence protéique ComA de *Streptococcus pneumoniae* souche R6

Sequence search results

[Show](#) the detailed description of this results page.

We found **3** Pfam-A matches to your search sequence (**all** significant)



[Show](#) the search options and sequence that you submitted.

[Return](#) to the search form to look for Pfam domains on a new sequence.

Significant Pfam-A Matches

[Show](#) or [hide](#) all alignments.


Family	Description	Entry type	Clan	Envelope		Alignment		HMM		HMM length	Bit score	E-value	Predicted active sites	Show/hide alignment
				Start	End	Start	End	From	To					
Peptidase_C39	Peptidase C39 family	Family	CL0125	5	143	7	142	3	132	133	140.0	3.4e-41	n/a	<input type="button" value="Show"/>
ABC_membrane	ABC transporter transmembrane region	Family	CL0241	168	438	170	438	3	274	274	212.1	1.1e-62	n/a	<input type="button" value="Show"/>
ABC_tran	ABC transporter	Domain	CL0023	500	650	500	650	1	137	137	110.3	9.3e-32	n/a	<input type="button" value="Show"/>

Comments or questions on the site? Send a mail to pfam-help@ebi.ac.uk.
European Molecular Biology Laboratory


Extrait de la description du domaine ABC_tran


Family: *ABC_tran* (PF00005)

Loading page components (2 remaining)...

 1119 architectures

 228719 sequences

 14 interactions

 3422 species

 507 structures

Summary

Domain organisation

Clan

Alignments

HMM logo

Trees

Curation & model

Species

Interactions

Structures

Jump to...

enter ID/acc

Summary: ABC transporter

Pfam includes annotations and additional family information from a range of different sources. These sources can be accessed via the tabs below.

[Wikipedia: ATP-binding domain of ABC transporters](#)

[Pfam](#)

[InterPro](#)

This is the Wikipedia entry entitled "[ATP-binding domain of ABC transporters](#)". [More...](#)

ATP-binding domain of ABC transporters [Edit Wikipedia article](#)

In molecular biology, **ATP-binding domain of ABC transporters** is a water-soluble [domain](#) of transmembrane [ABC transporters](#).

ABC transporters belong to the [ATP-Binding Cassette superfamily](#), which uses the hydrolysis of [ATP](#) to translocate a variety of compounds across [biological membranes](#). ABC transporters are minimally constituted of two conserved regions: a highly conserved ATP binding cassette (ABC) and a less conserved transmembrane domain (TMD). These regions can be found on the same protein or on two different ones. Most ABC transporters function as a dimer and therefore are constituted of four domains, two ABC modules and two TMDs.

Contents [\[hide\]](#)

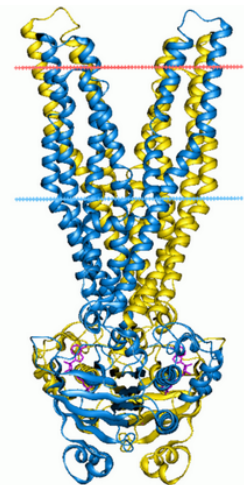
- [1 Biological function](#)
- [2 Amino acid sequence](#)
- [3 3D structure](#)
- [4 Human proteins containing this domain](#)
- [5 References](#)

Biological function

ABC transporters are involved in the export or import of a wide variety of [substrates](#) ranging from small ions to macromolecules. The major function of ABC import systems is to provide essential nutrients to bacteria. They are found only in prokaryotes and their four constitutive domains are usually encoded by independent polypeptides (two ABC proteins and two TMD proteins). Prokaryotic importers require additional extracytoplasmic binding proteins (one or more per systems) for function. In contrast, export systems are involved in the extrusion of noxious substances, the export of extracellular toxins and the targeting of membrane components. They are found in all living organisms and in general the TMD is fused to the ABC module in a variety of combinations. Some eukaryotic exporters encode the four domains on the same polypeptide chain.

Amino acid sequence

The ABC module (approximately two hundred amino acid residues) is known to bind and hydrolyze ATP, thereby coupling transport to ATP hydrolysis in a large number of biological processes. The cassette is duplicated in several subfamilies. Its primary sequence is highly conserved, displaying a typical



Multidrug ABC transporter SAV1866, closed state

Identifiers

Différentes architectures protéiques possédant le domaine ABC_tran (extrait)

There are 90080 sequences with the following architecture: ABC_tran

[X6PF59_RETFI](#) [Reticulomyxa filosa] ATP-binding cassette protein {ECO:0000313|EMBL:ETO36694.1} (332 residues)



[Show](#) all sequences with this architecture.

There are 21754 sequences with the following architecture: ABC_membrane, ABC_tran

[X4ZNP7_9BACL](#) [Paenibacillus sabinae T27] ABC transporter ATP-binding protein {ECO:0000313|EMBL:AHV98210.1} (617 residues)



[Show](#) all sequences with this architecture.

There are 10217 sequences with the following architecture: ABC_tran x 2

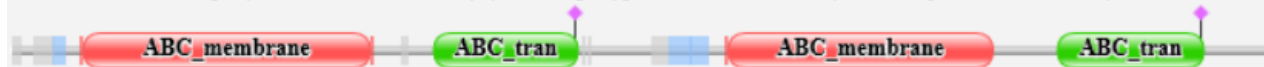
[W9SF44_9ROSA](#) [Morus notabilis] ABC transporter F family member 4 {ECO:0000313|EMBL:EXC49943.1} (726 residues)



[Show](#) all sequences with this architecture.

There are 8474 sequences with the following architecture: ABC_membrane, ABC_tran, ABC_membrane, ABC_tran

[W5N0E5_LEPOC](#) [Lepisosteus oculatus (Spotted gar)] Uncharacterized protein {ECO:0000313|Ensembl:ENSLOCP00000014104} (1310 residues)



[Show](#) all sequences with this architecture.

There are 8145 sequences with the following architecture: ABC_tran, oligo_HPY

[W8X1U8_CASDE](#) [Castellaniella defragrans 65Phen] Dipeptide transport ATP-binding protein DppF {ECO:0000313|EMBL:CDM26063.1} (380 residues)



[Show](#) all sequences with this architecture.

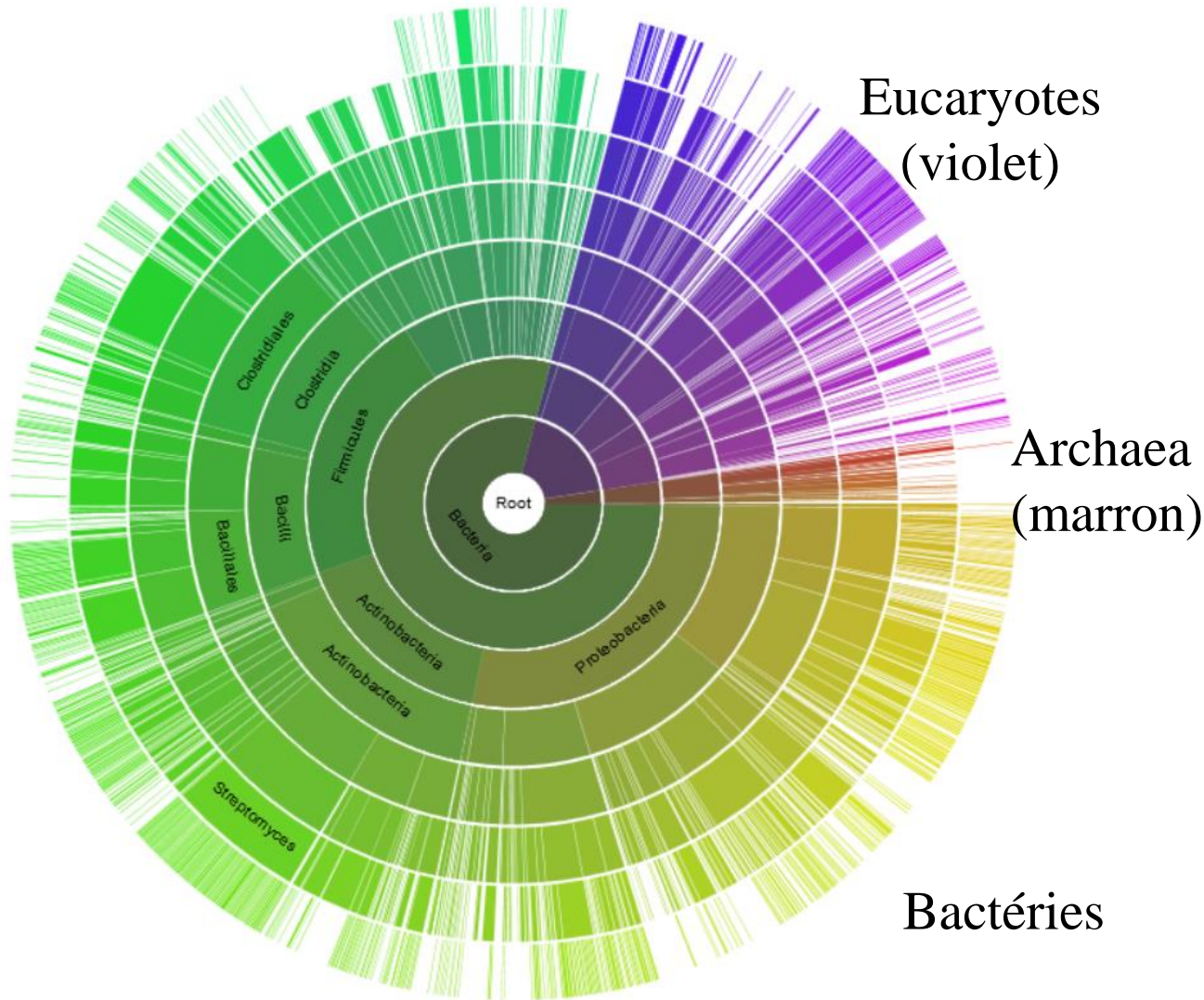
There are 7095 sequences with the following architecture: ABC_tran, TOBE_2

[D5RQ25_9PROT](#) [Roseomonas cervicalis ATCC 49957] ABC transporter, ATP-binding protein {ECO:0000313|EMBL:EFH10598.1} (347 residues)



[Show](#) all sequences with this architecture.

Visualisation graphique simple de cette famille de protéines au sein des espèces



Arthrobacter sp. FB24

```
Root
├── Bacteria
│   ├── (No kingdom)
│   ├── Actinobacteria
│   │   ├── Actinobacteria
│   │   ├── Micrococcales
│   │   │   ├── Micrococaceae
│   │   │   └── Arthrobacter
│   │   │       └── Arthrobacter sp. FB24
```

Weight segments by...

number of sequences
 number of species

Change the size of the sunburst

Small Large

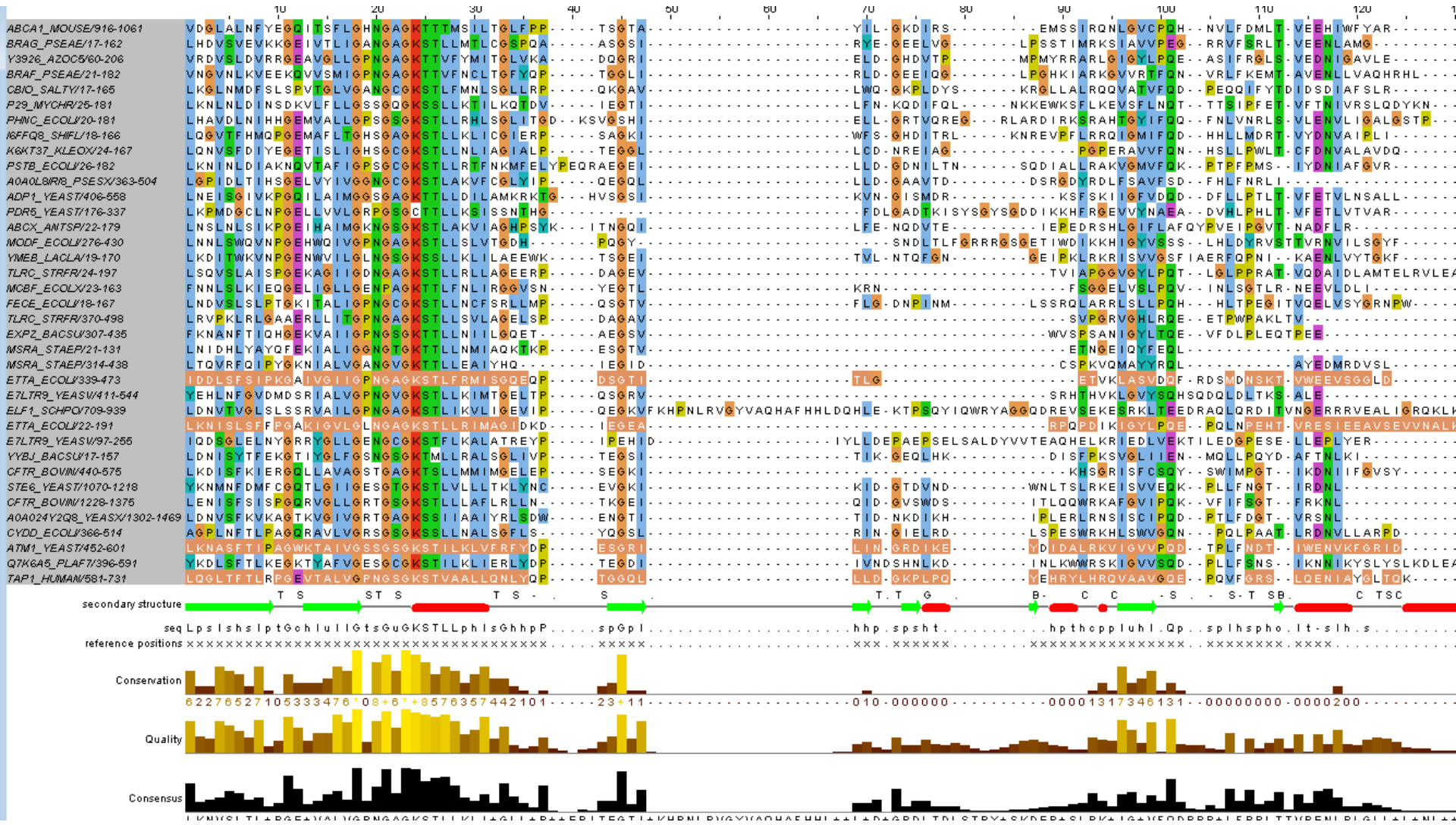
Colour assignments

<input type="checkbox"/> Arches	<input type="checkbox"/> Eukaryote
<input type="checkbox"/> Bacteria	<input type="checkbox"/> Other sequences
<input type="checkbox"/> Viruses	<input type="checkbox"/> Unclassified
<input type="checkbox"/> Virioids	<input type="checkbox"/> Unclassified sequence

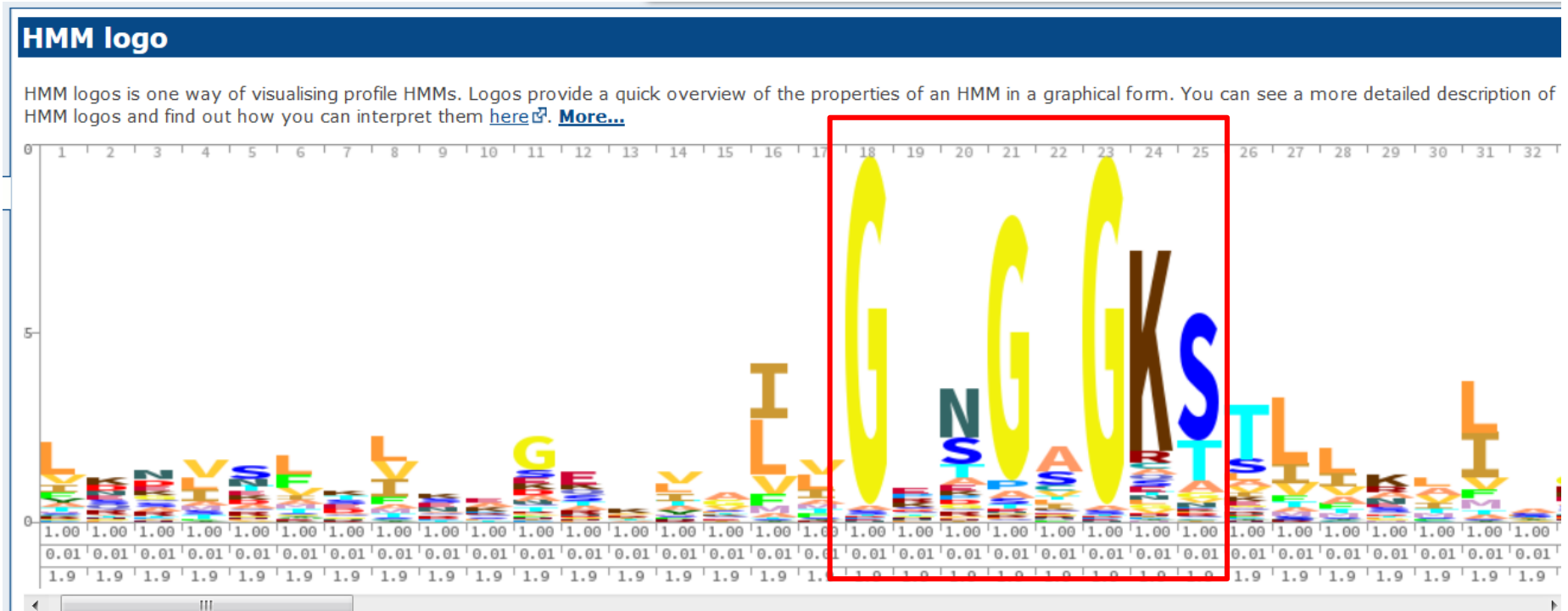
Selections

[Align](#) selected sequences to HMM
[Generate](#) a FASTA-format file
[Clear](#) selection

Extrait de l'alignement multiple correspondant au domaine fonctionnel ABC_tran (sous Jalview) sur les séquences « seed »



Extrait du logo correspondant au domaine fonctionnel ABC_tran



Correspond à la zone fortement conservée de l'alignement précédent et représente le motif Walker A de liaison de l'ATP

InterPro

Interpro permet la classification des protéines en fonction de la présence de domaines fonctionnels, répétitions, et signaux grâce à une recherche automatisée dans plusieurs bases de données (CATH-Gene3D, HAMAP, PANTHER, Pfam, PIRSF, PRINTS, ProDom, PROSITE, SMART, SUPERFAMILY, TIGRFAMs).

Page d'entrée d'InterPro : analyse de la séquence ComA de *S. pneumoniae*



Search InterPro... Search

Examples: IPR020405, kinase, P51587, PF02932, GO:0007165

Home Search Release notes Download About InterPro Help Contact

By sequence By domain architecture

InterProScan sequence search

This form allows you to scan your sequence for matches against the InterPro protein signature databases, using InterProScan tool.
Enter or paste a protein sequence in FASTA format (complete or not - e.g. PMPIGSKERPTFFEIFKTRCNKADLGPISLN), with a maximum length of 40,000 amino acid long.
Please note that you can only scan one sequence at a time.

Analyse your protein sequence

```
ESQNTLFFMTLLALPIYTVIIFAFMKPFKMNDRDIMEANAVLSSSIEDINGIETKSL  
TSESQRYQKIDKEEVDYLKKSFTYSRAESQKQKALKKVAHLLNVGILWVGAVLVMDGKMS  
LGQLTYNTLVYFTNPLENINLQTKLQTAQVANNRLNEVYLVAEFEFEKKTVEDLSLM  
KGDMTFKQVHYKYGYGRDVLSDINLTPQGSKVAFVGSISGSKTLAKMMVNFYDPSQGE  
ISLGGVNLNQIDKKALRQYINYLPOQPYYFNGTLENLLGAKEGTTQEDILRAVELAEI  
REDIERMPLNYQTELTSDGAGISGGQRORIALARALLTDAPVILDEATSSLDILTEKRI  
VDNLIALDKTLFIAHRLTIAERTEKVVVLDQGGKIVEEGKHADLLAQGGFYAHLVNS
```

Advanced options

Select the applications to run: Uncheck all Select all

Member databases

Families, domains, sites & repeats

CDD HAMAP PANTHER PfamA PIRSF PRINTS ProDom Prosite-Profiles SMART TIGRFAM Prosite-Patterns

Structural domains


Gene3d SFLD SUPERFAMILY

Other sequence features

Coils MobiDB Lite Phobius SignalP TMHMM

Search | Clear

InterProScan



InterProScan is a sequence analysis application (nucleotide and protein sequences) that combines different protein signature recognition methods into one resource.

[More about InterProScan.](#)

? Need more help?

If you need more info on InterProScan, you can either look at the:

- Documentation page
- Online training course

or [contact us](#) directly with your question.

Résultats de l'analyse de la séquence ComA de *S. pneumoniae*

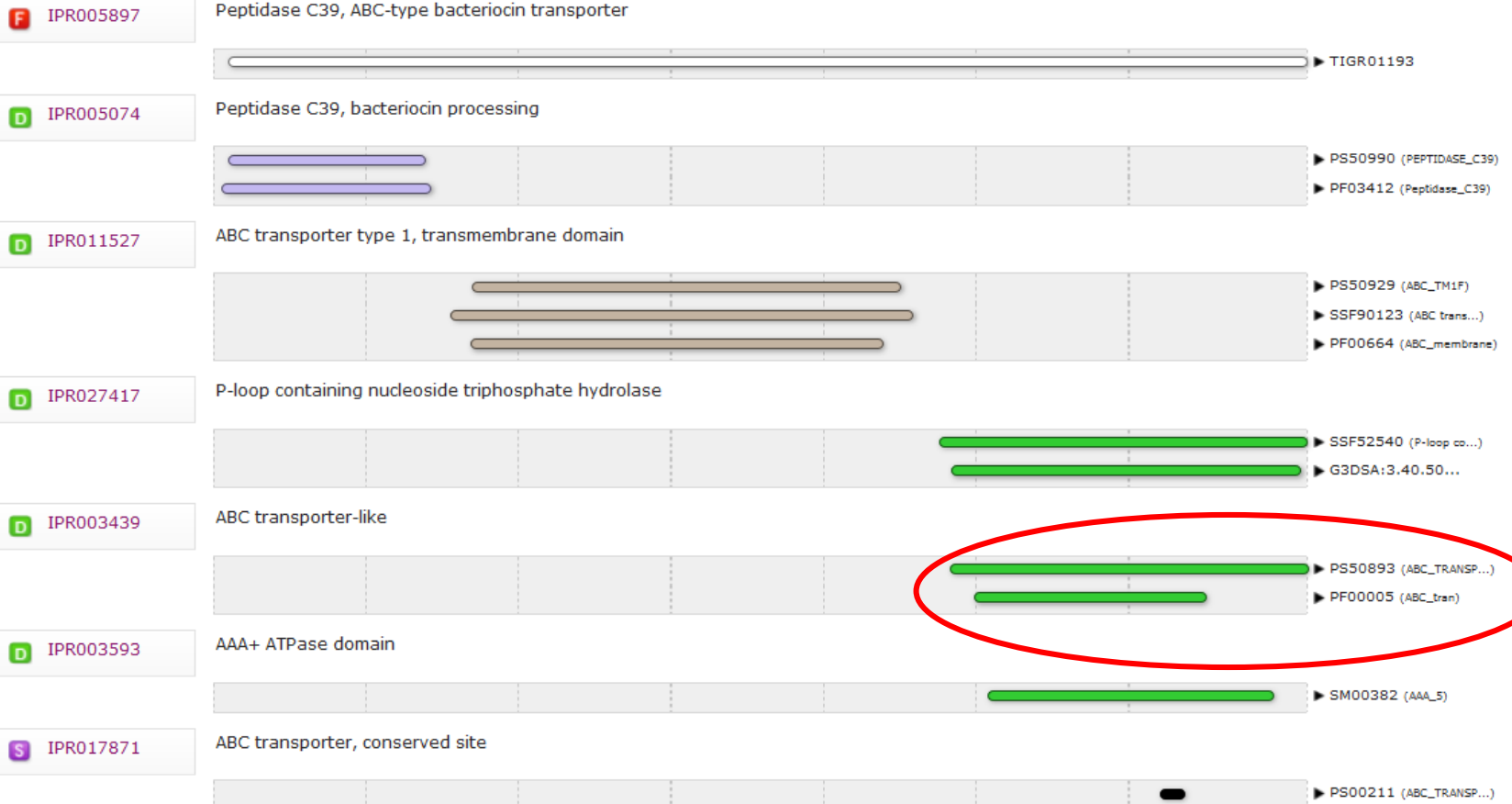
Protein family membership

F Peptidase C39, ABC-type bacteriocin transporter (IPR005897)

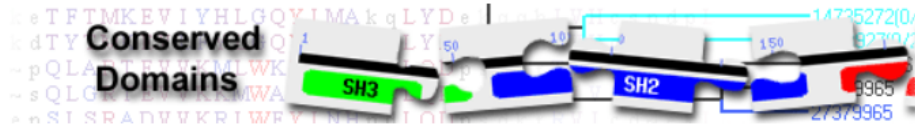
Domains and repeats



Detailed signature matches



Recherche des domaines fonctionnels dans la séquence ComA de *S. pneumoniae* dans la banque « Conserved Domain Database » (CDD) maintenue au NCBI (CD_search)



HOME SEARCH GUIDE

Structure Home

3D Macromolecular Structures

Conserved Domains

Pubchem

BioSystems

Search for Conserved Domains within a protein or coding nucleotide sequence

NEW! Use **Batch CD-search** to submit multiple query proteins at once!

Enter **protein** or **nucleotide** query as accession, gi, or sequence in **FASTA format** [?](#)

```
>SpneA01.COMA "Transport ATP-binding protein ComA"
MKFGKRHRYPQVDMDCGVASLAVFGYYSYFLAHLRELAKTTMDGTTALGLVKVAEEIGFETRAIKADMTLFDLPDL
TFPFVAHVLEKGLLHYVVTGQDKDSIHADDPDGVKLTKLPRERFEEWTGVTLFMAPSPDYKPKHEQKNGLLSFIPI
LVKQRGLIANIVLATLLVTVINIVGSYYLQSIIDTYVPDQMRSTLGIISIGLVIVYILQQIILSYAQEYLLLVLGQRLSID
VILSYIKHVFLPMSFFATRRRTGEIVSRFTDANSIIDALASTILSIFLDVSTVVIISLVLFSQNTNLFMFTLLALPIYTV
IIFAFMKPFKEMNRDTMEANAVLSSSIIEDINGIETIKSLTSESQRYQKIDKEFVDYLKKSFTYSRAESQQKALKKVAHL
LLNVGILWMGAVLVMDGKMSLGLITINTLLVYFTNPLENIINLQTKLQTAQVANNRLNEVYLVASEFEERKTVEDLSLM
KGDMTFKQVHYKYGYGRDVLSDINLTVPGSKVAFVIGSGSGLTTLAKMMVNFYDPSQGEISLGGVNLNQIDKKALRQYI
NYLPQQPYVFNGTILENLLGAKEGTTQEDILRAVELAETREDIERMPLNYQTELTSDGAGISGGQRQRIALARALLTDA
PVLILDEATSSLDILTERRIVDNLIALDKTLIFIAHRLTIAERTEKVVVLDQKIVEEGKHADLLAQGGFYAHLVNS
```

Submit

Reset

OPTIONS

Search against database [?](#): CDD v3.15 - 48963 PSSMs [v](#)

Expect Value [?](#) threshold: 0.010000

Apply low-complexity filter [?](#)

Composition based statistics adjustment [?](#)

Force live search [?](#)

Rescue borderline hits Suppress weak overlapping hits

Maximum number of hits [?](#) 500

Result mode Concise [?](#) Standard [?](#) Full [?](#)

Retrieve previous CD-search result

Request ID: [Retrieve](#) [?](#)

References:

- Marchler-Bauer A et al. (2017), "CDD/SPARCLE: functional classification of proteins via subfamily domain architectures.", **Nucleic Acids Res.**45(D)200-3.
- Marchler-Bauer A et al. (2015), "CDD: NCBI's conserved domain database.", **Nucleic Acids Res.**43(D)222-6.
- Marchler-Bauer A et al. (2011), "CDD: a Conserved Domain Database for the functional annotation of proteins.", **Nucleic Acids Res.**39(D)225-9.
- Marchler-Bauer A, Bryant SH (2004), "CD-Search: protein domain annotations on the fly.", **Nucleic Acids Res.**32(W)327-331.

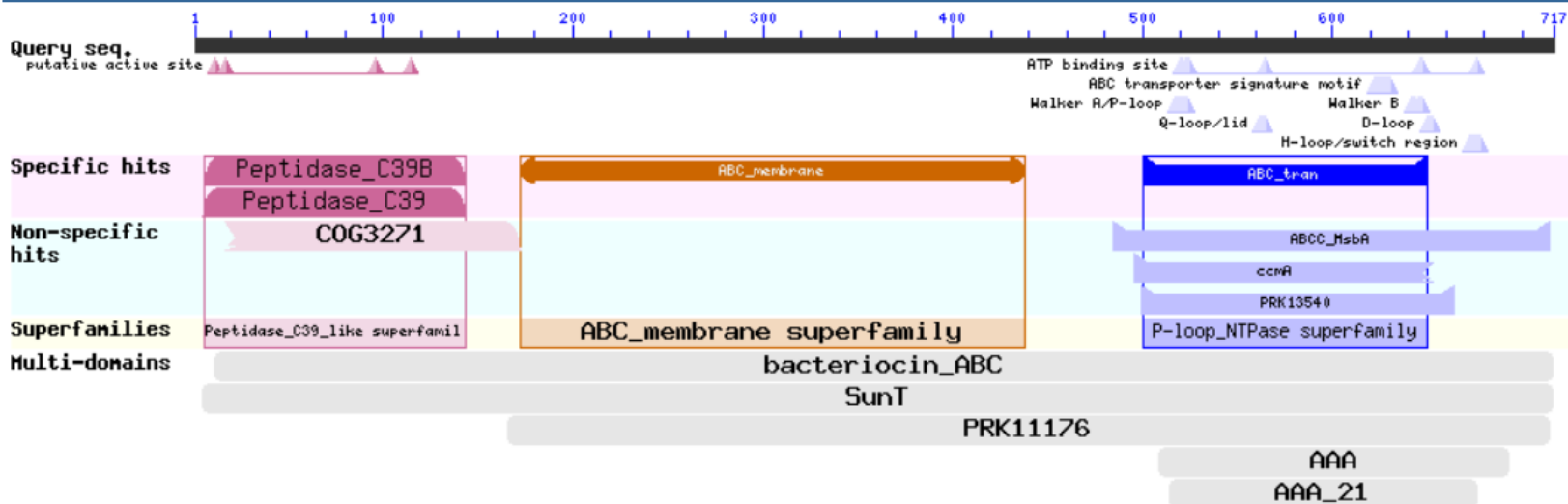
Résultat de la recherche dans la banque CDD

protein containing domains Peptidase_C39B, ABC_membrane, and P-loop_NTPase

Graphical summary

Zoom to residue level

[show extra options >](#)



[Search for similar domain architectures](#) ?

[Refine search](#) ?

List of domain hits

+	Name	Accession	Description	Interval	E-value
[+]	ABCC_MsbA	cd03251	ATP-binding cassette domain of the bacterial lipid flippase and related proteins, subfamily C; ...	485-714	9.00e-79
[+]	Peptidase_C39B	cd02418	A sub-family of peptidase family C39. Peptidase family C39 mostly contains ...	6-143	1.66e-68
[+]	ABC_membrane	pfam00664	ABC transporter transmembrane region; This family represents a unit of six transmembrane ...	172-438	2.87e-64
[+]	Peptidase_C39	pfam03412	Peptidase C39 family; Lantibiotic and non-lantibiotic bacteriocins are synthesized as ...	5-143	1.20e-50
[+]	ABC_tran	pfam00005	ABC transporter; ABC transporters for a large family of proteins responsible for translocation ...	500-650	5.05e-39
[+]	ccmA	TIGR01189	heme ABC exporter, ATP-binding protein CcmA; This model describes the cyt c biogenesis protein ...	496-653	6.96e-19
[+]	PRK13540	PRK13540	cytochrome c biogenesis protein CcmA; Provisional	499-664	4.40e-10
[+]	COG3271	COG3271	Predicted double-glycine peptidase [General function prediction only];	17-172	4.87e-05
[+]	bacteriocin_ABC	TIGR01193	ABC-type bacteriocin transporter; This model describes ABC-type bacteriocin transporter. The ...	11-716	0e+00
[+]	SunT	COG2274	ABC-type bacteriocin/lantibiotic exporters, contain an N-terminal double-glycine peptidase ...	5-716	0e+00
[+]	PRK11176	PRK11176	lipid transporter ATP-binding/permease protein; Provisional	166-714	9.99e-76
[+]	AAA	smart00382	ATPases associated with a variety of cellular activities; AAA - ATPases associated with a ...	509-693	8.93e-09
[+]	AAA_21	pfam13304	AAA domain, putative AbiEii toxin, Type IV TA system; Several members are annotated as being ...	514-676	1.42e-03

Résultat détaillée de la détection du domaine fonctionnel Pfam00664

ABC transporter; ABC transporters for a large family of proteins responsible for translocation of a variety of compounds across biological membranes. ABC transporters are the largest family of proteins in many completely sequenced bacteria. ABC transporters are composed of two copies of this domain and two copies of a transmembrane domain pfam00664. These four domains may belong to a single polypeptide or belong in different polypeptide chains.

Pssm-ID: 278435 Cd Length: 150 Bit Score: 140.09 E-value: 5.05e-39

```

          10          20          30          40          50          60          70          80
          .....*.....|.....*.....|.....*.....|.....*.....|.....*.....|.....*.....|.....*.....|.....*.....|
seqsig_MKFGK_e2ea59cba8ae2f892bff9c30e69b60cc 500 LSDINLTVPQGSKVAFVGLSGSGKTLAKMMVNFYDPSQGEISLGGVNLNQIDKKALRQYINYLPPQPYVFNG-TILENL 578
Cdd:pfam00005 1 LKNVSLTLNPGEILALVGPNGAGKSTLLKLIAGLLSPTEGTILLDGGDLTDDERKSLRKEIGYVFQDPNLFPRLTVRENL 80

          90          100          110          120          130          140          150
          .....*.....|.....*.....|.....*.....|.....*.....|.....*.....|.....*.....|.....*.....|.....
seqsig_MKFGK_e2ea59cba8ae2f892bff9c30e69b60cc 579 LLGAKEggttqEDILRAVELAEIREDIERMPINLQ--TELTSDGAGISGGQRQRIALARALLTDAPVLLIIDEATS 650
Cdd:pfam00005 81 RLGLRL---KGLSKREKDARAEALEKLGLDLdRPVGENPGTSSGGQRQVAIARALLTKPKLLLLDEPTA 150
```