Support de cours
Motifs et Profils
Domaines fonctionnels

Motifs et profils

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

• correspondent en général à des zones fonctionnelles

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

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

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

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

Comment représenter cette variabilité ?

- séquence consensus
- matrice de poids

Représentation : Séquence consensus

Exemples des boîtes CRE:

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

Motif dans les séquences de Maltose Binding Proteins

YvfK_Bs	PTPNIPEMNEIW
YvfK_Bs	PTPNIPEMAEVW
MalX_Sp	PLPNISQMSAVW
MalE_Sc	PRPALPEYSSLW
MalE_Tm	PMPNVPEMAPVW
MalE_Dr	PMPNIPEMGAVW
CymE_Ko	AMPSIPEMGYLW
MalE_Ea	IMPNIPQMSAFW
MalE_Sy	IMPNIPQMSAFW
MalE_Ec	IMPNIPQMSAFW

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 du nombres d'occurrences de chaque base *b* à chaque position *i* ($n_{b,i}$) du motif -10 (6 positions) :

Pos.	1	2	3	4	5	6
A	9	214	63	142	118	8
C	22	7	26	31	52	13
G	18	2	29	38	29	5
T	193	19	124	31	43	216

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 Weight Matrix, PWM)

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

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

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

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

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

Utilisation d'une matrice de poids sur une séquence

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

A CTATAATCG

AC TATAATCG

ACT ATAATCG

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

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

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

Exemples de fonction pour le calcul du score

Soit *l* le nombre de positions dans le motif, $f_{b,i}$ la fréquence de la base b observée à la position *i* dans la séquence analysée et $f_{max,i}$ la fréquence de la base la plus fréquente à la position *i* dans la matrice de poids :

$$S = \frac{\sum_{i=1}^{l} f_{b,i}}{\sum_{i=1}^{l} f_{\max,i}}$$

La valeur du score *S* va varier entre 0 et 1, quelque soit la longueur du motif étudié et la matrice de poids établie.
On retient la séquence comme motif putatif si *S* ≥ seuil.

$$D = \sum_{i=1}^{l} \ln\left(\frac{f_{\max,i} + 0.5}{f_{b,i} + 0.5}\right)$$

***D* est un indice de disimilarité établi par Berg and Von Hippel. Plus la valeur de D sera élevée, plus la séquence analysée est éloignée de la séquence consensus. On ajoute 0.5 pour éviter la division par 0 quand $f_{b,i}$ est nulle.**
On retient la séquence comme motif putatif si *D* ≤ seuil.

Théorie de l 'information

Shannon et Weaver (1949).

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

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

où :

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

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

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

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

donc

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

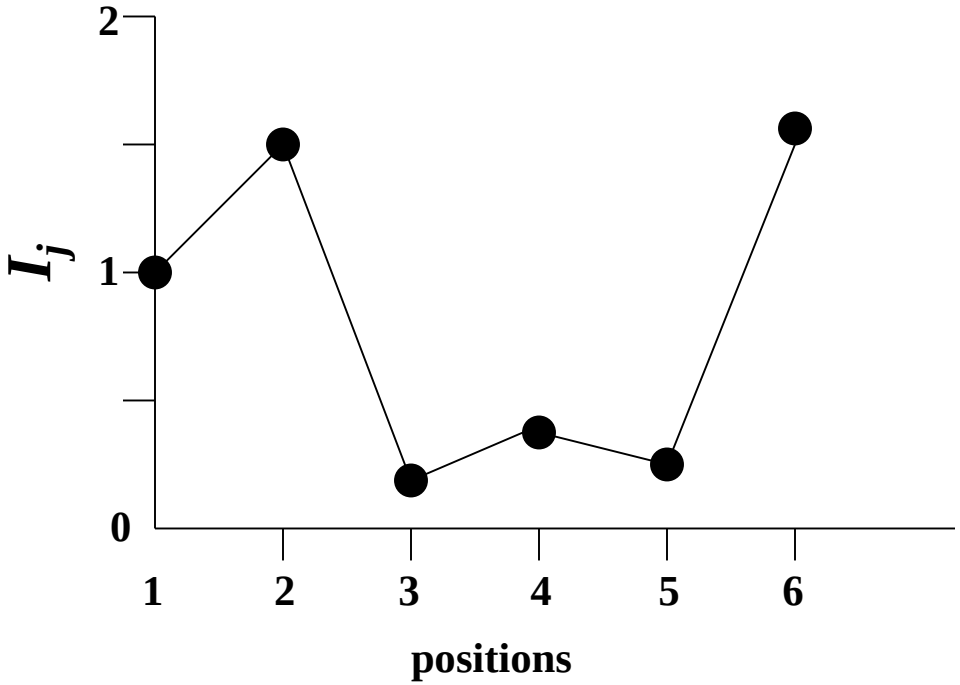
Les positions du signal qui contiendront de l'information seront celles qui auront une composition très biaisées par rapport à ce qui est attendu.

Si à une position j du signal, présence d'une seule base invariante i alors $f_{ij} = 1$ et $\log_2 f_{ij} = 0$ donc

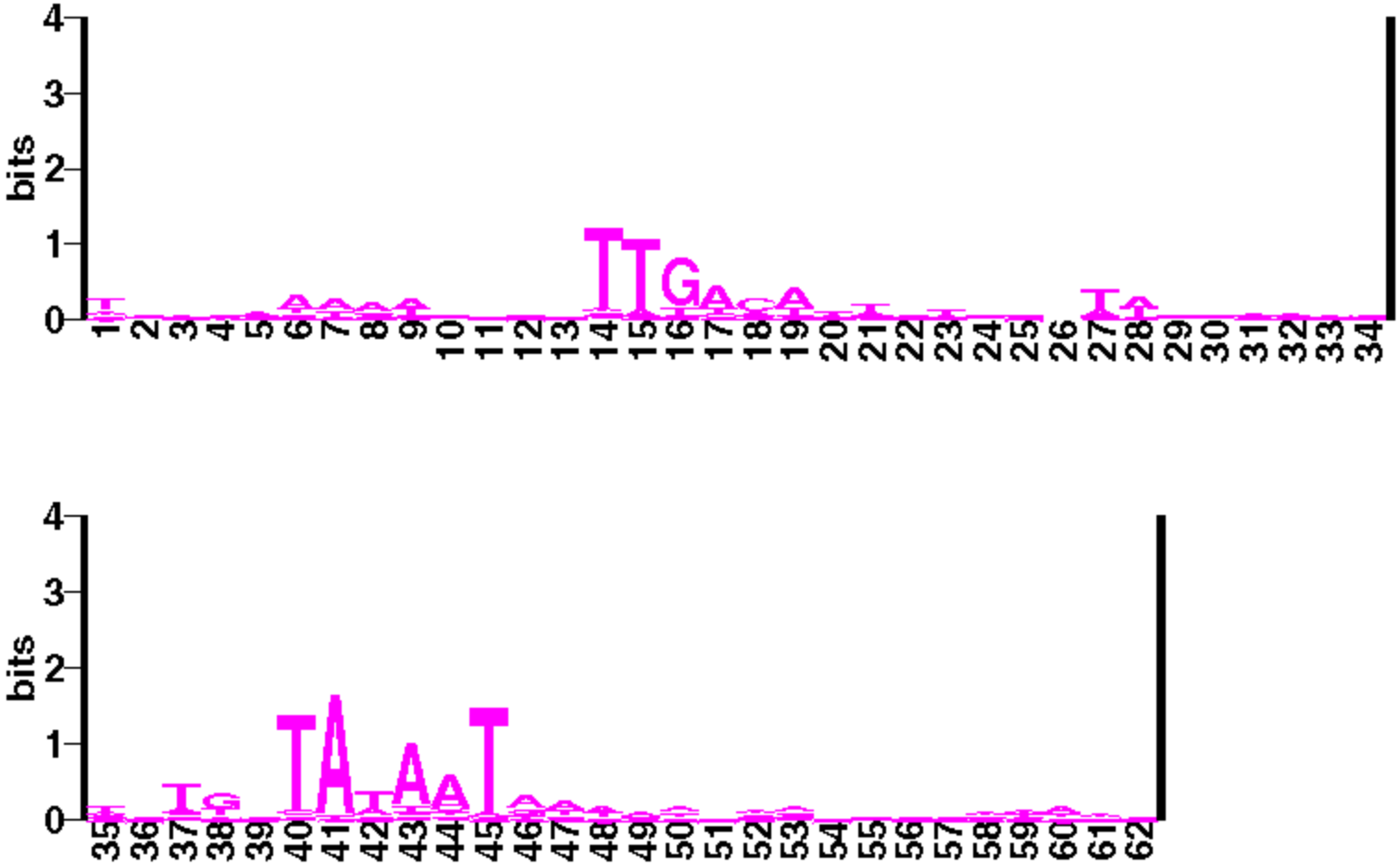
$f_{ij} \log_2 f_{ij} = 0$ et les fréquences observées des autres bases sont nulles. On aura

$I(j) = 2$ information maximale

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

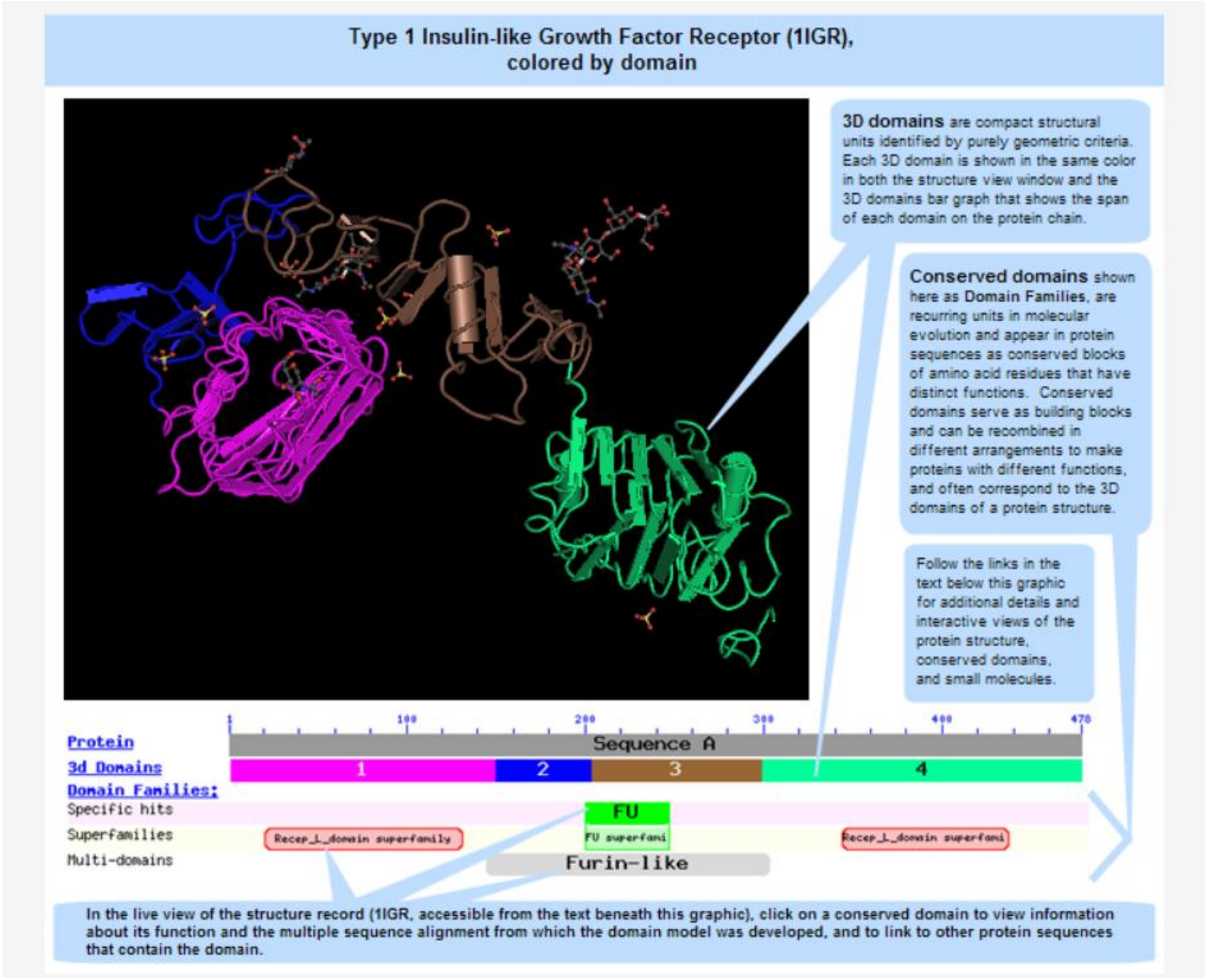


Domaines fonctionnels

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 recombinée 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



(Extrait de la documentation sur le site du NCBI https://www.ncbi.nlm.nih.gov/Structure/cdd/cdd_help.shtml#CDWhat)

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)

ScanProsite tool

http://prosite.expasy.org/scanprosite/

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).

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]

[x] 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

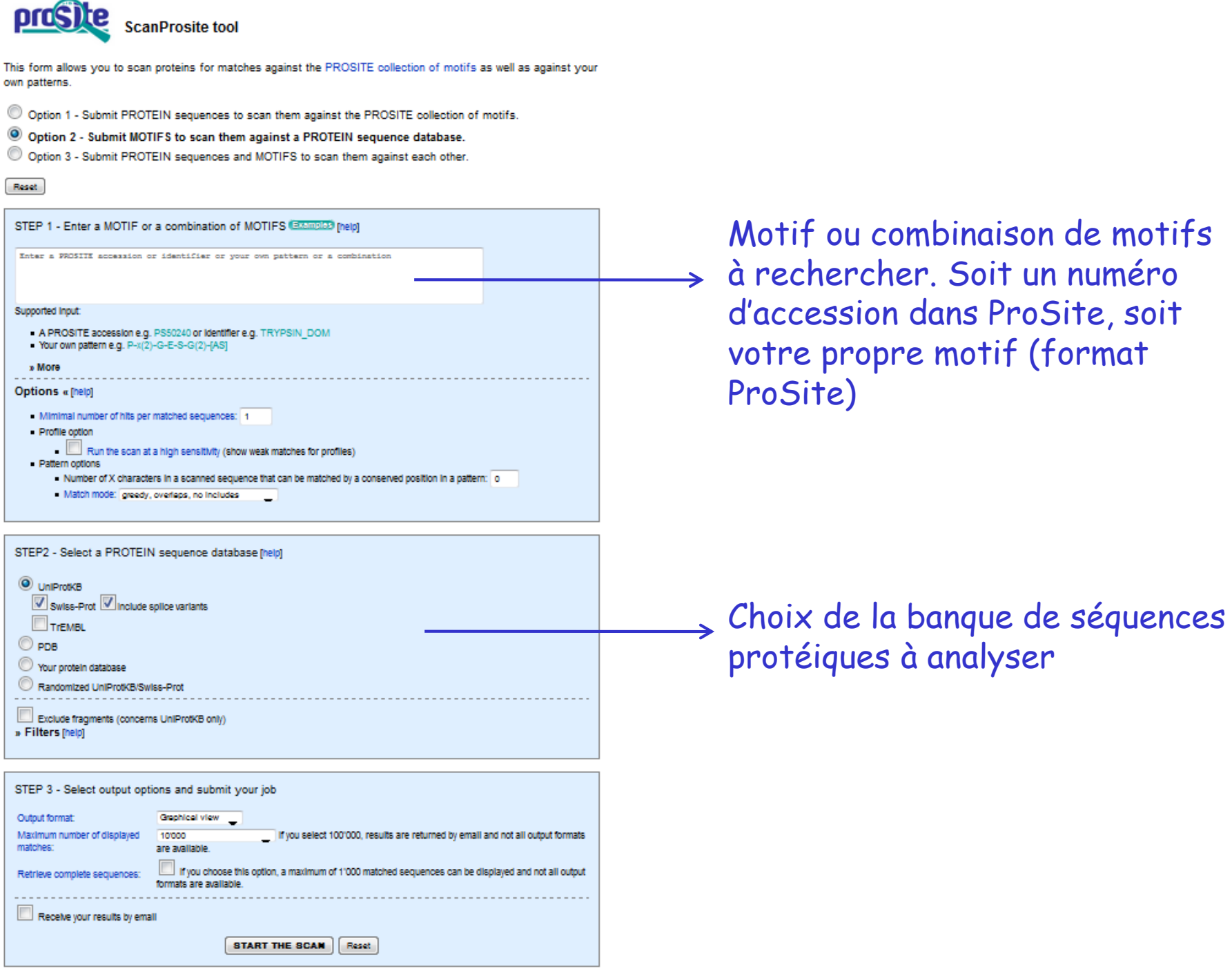
Output format: Graphical view

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

[ ] Receive your results by email

START THE SCAN Reset

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



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

Choix de la banque de séquences protéiques à analyser

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
MKFGKRHYRPQVDQMDCGVASLAMVFGYYGSYYFLAHLRELAKTTMDGTTALGLVKVAEE
IGFETRAIKADMTLFDLPDLTFPFVAHVLKEGKLLHYYVVTGQDKDSIHIADPDPGVKLT
KLPRERFEEEWTGVTLFMAPSPDYKPHKEQKNGLLSFIPILVKQRGLIANIVLATLLVTV
INIVGSYYLQSIIDTYVPDQMRSTLGIISIGLVIVYILQQILSYAQEYLLLVLGQRLSID
VILSYIKHVFHLPMSFFATRRTGEIVSRFTDANSIIDALASTILSIFLDVSTVVIISLVL
FSQNTNLFFMTLLALPIYTVIIFAFMKPFEKMNRDTMEANAVLSSSIIEDINGIETIKSL
TSESQRYQKIDKEFVDYLKKSFTYSRAESQQKALKKVAHLLLNVGILWMGAVLVMDGKMS
LGQLITYNTLLVYFTNPLENIINLQTKLQTAQVANNRLNEVYLVASEFEEKKTVEDLSLM
KGDMTFKQVHYKYGYGRDVLSDINLTVPQGSKVAFVGISGSGKTTLAKMMVNFYDPSQGE
ISLGGVNLNQIDKKALRQYINYLPQQPYVFNGTILENLLLGAKEGITQEDILRAVELAEI
```

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]

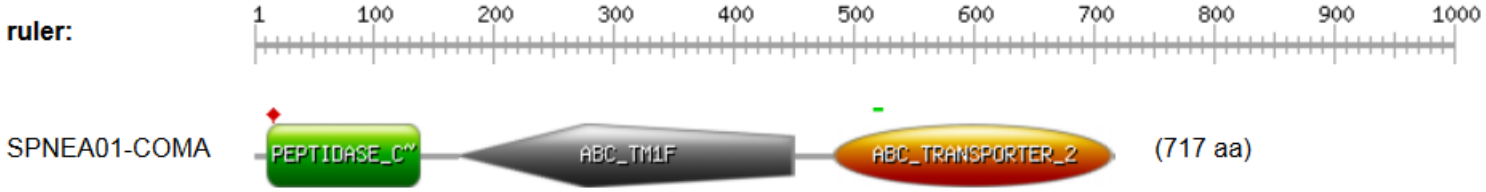
- [x] Exclude motifs with a high probability of occurrence from the scan
- [ ] Exclude profiles from the scan
- [ ] Run the scan at high sensitivity (show weak matches for profiles)

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]

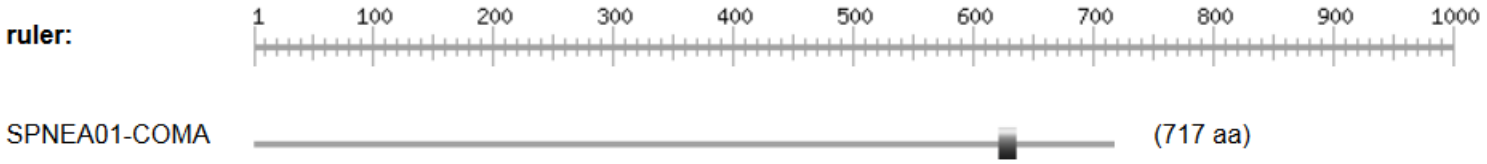
Upper case represents match positions, lower case insert positions, and the '-' symbol represents deletions relative to the matching profile.



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* :

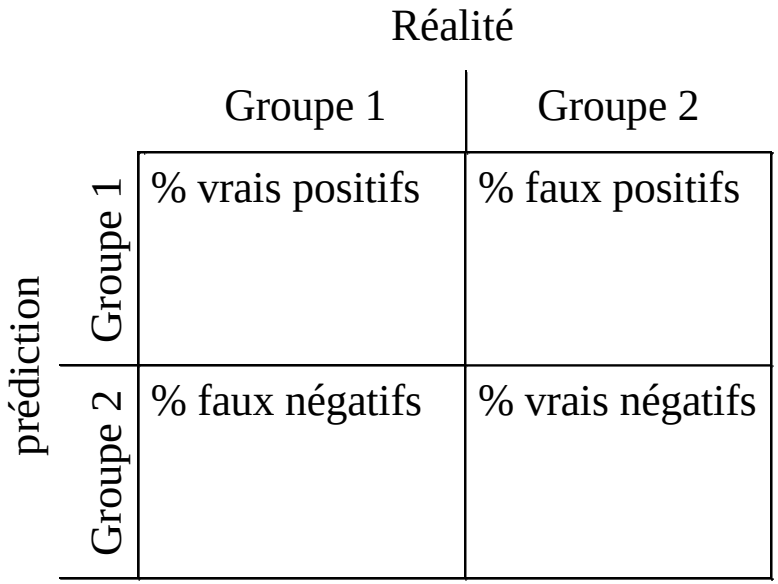


622 - 636: [confidence level: (0)] ISGGQRQRIALARAL

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)**



Groupe 1 : exemples
Groupe 2 : contre-exemples

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]

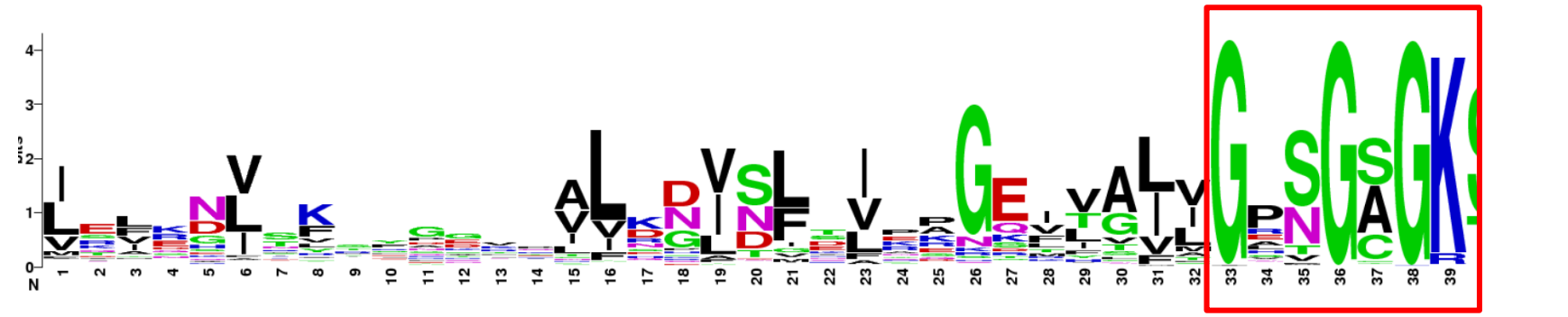
```
/GENERAL_SPEC: ALPHABET='ABCDEFGHIKLMNPQRSTVWYZ'; LENGTH=240;
/DISJOINT: DEFINITION=PROTECT; N1=6; N2=235;
/NORMALIZATION: MODE=1; FUNCTION=LINEAR; R1=5.7291522; R2=0.0066693; TEXT='NScore';
/NORMALIZATION: MODE=-1; FUNCTION=LINEAR; R1=27511.1894531; R2=19.6396694; TEXT='Heuristic 5.0%';
/CUT_OFF: 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

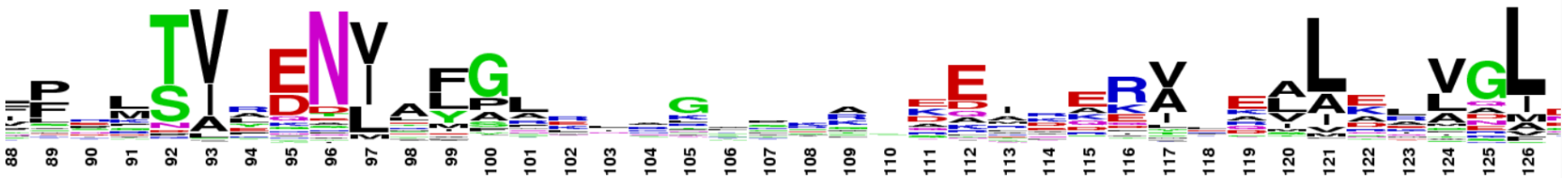
ABC_TRANSPORTER_1, PS00211; ABC transporters family signature (PATTERN)

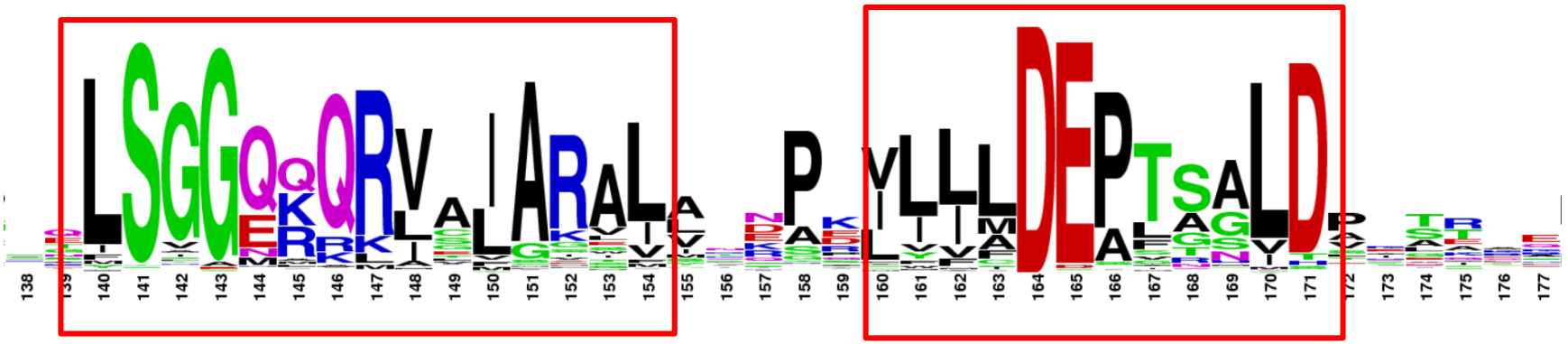
- Consensus pattern:
  [LIVMFYC]-[SA]-[SAPGLVFYKQH]-G-[DENQMW]-[KRQASPCLIMFW]-[KRNQSTAVM]-[KRACLVM]-[LIVMFYPAN]-{PHY}-[LIVMFW]-[SAGCLIVP]-{FYWHP}-{KRHP}-[LIVMFYWSTA]
- Sequences in UniProtKB/Swiss-Prot known to belong to this class: 4003
  - detected by PS00211: 3668 (true positives)
  - undetected by 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.

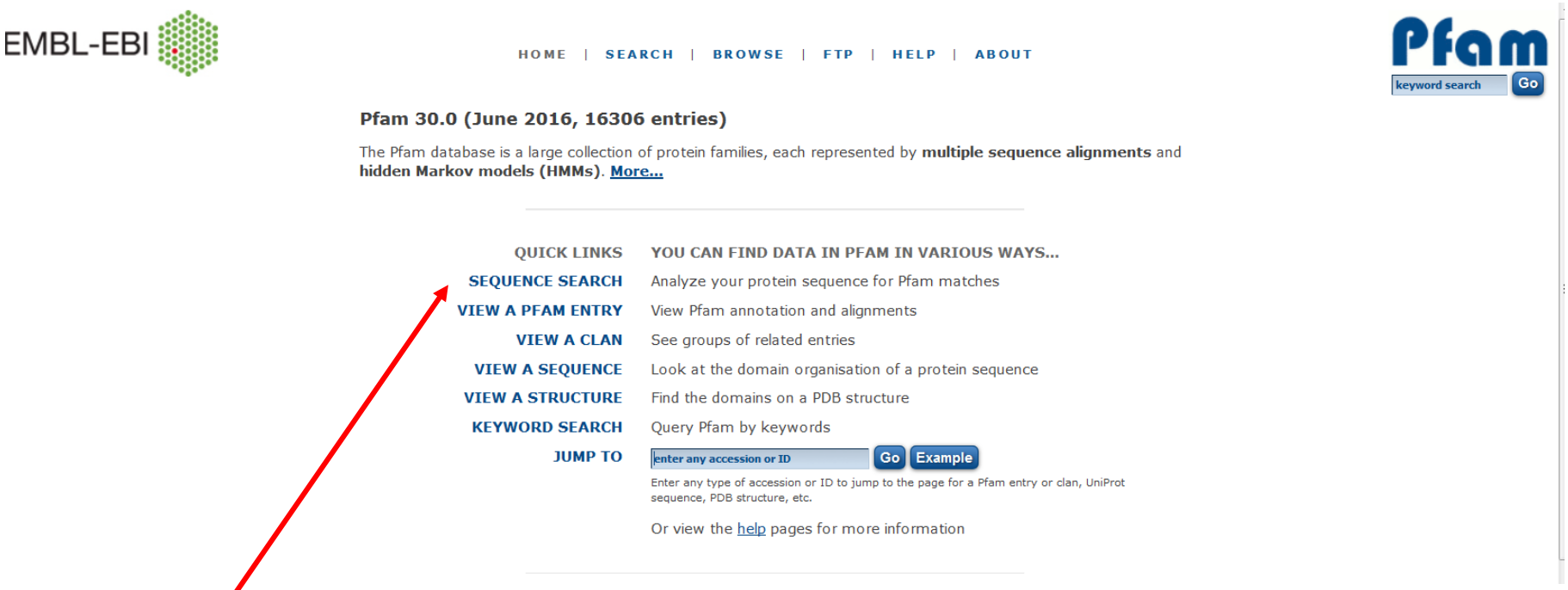
2 sections dans Pfam:

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/



Analyser le contenu en domaines d'une séquence protéique

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

EMBL-EBI

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Pfam

keyword search Go

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	Show
ABC_membrane	ABC transporter transmembrane region	Family	CL0241	168	438	170	438	3	274	274	212.1	1.1e-62	n/a	Show
ABC_tran	ABC transporter	Domain	CL0023	500	650	500	650	1	137	137	110.3	9.3e-32	n/a	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

EMBL-EBI

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Pfam keyword search Go

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 Go

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...

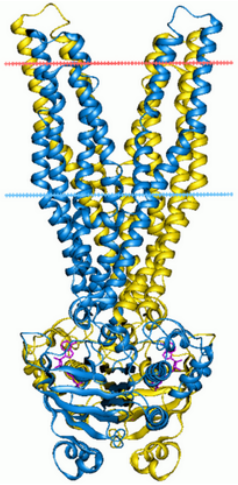
ATP-binding domain of ABC transporters

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]



Multidrug ABC transporter SAV1866, closed state

Identifiers

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

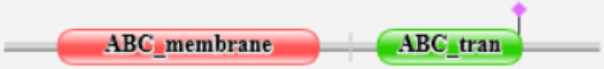
The ABC module (approximately two hundred amino acid residues) is known to bind and hydrolyze ATP, thereby coupling transport to ATP hydrolysis in a large number of biological processes. The cassette is duplicated in several subfamilies. Its primary sequence is highly conserved, displaying a typical

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.

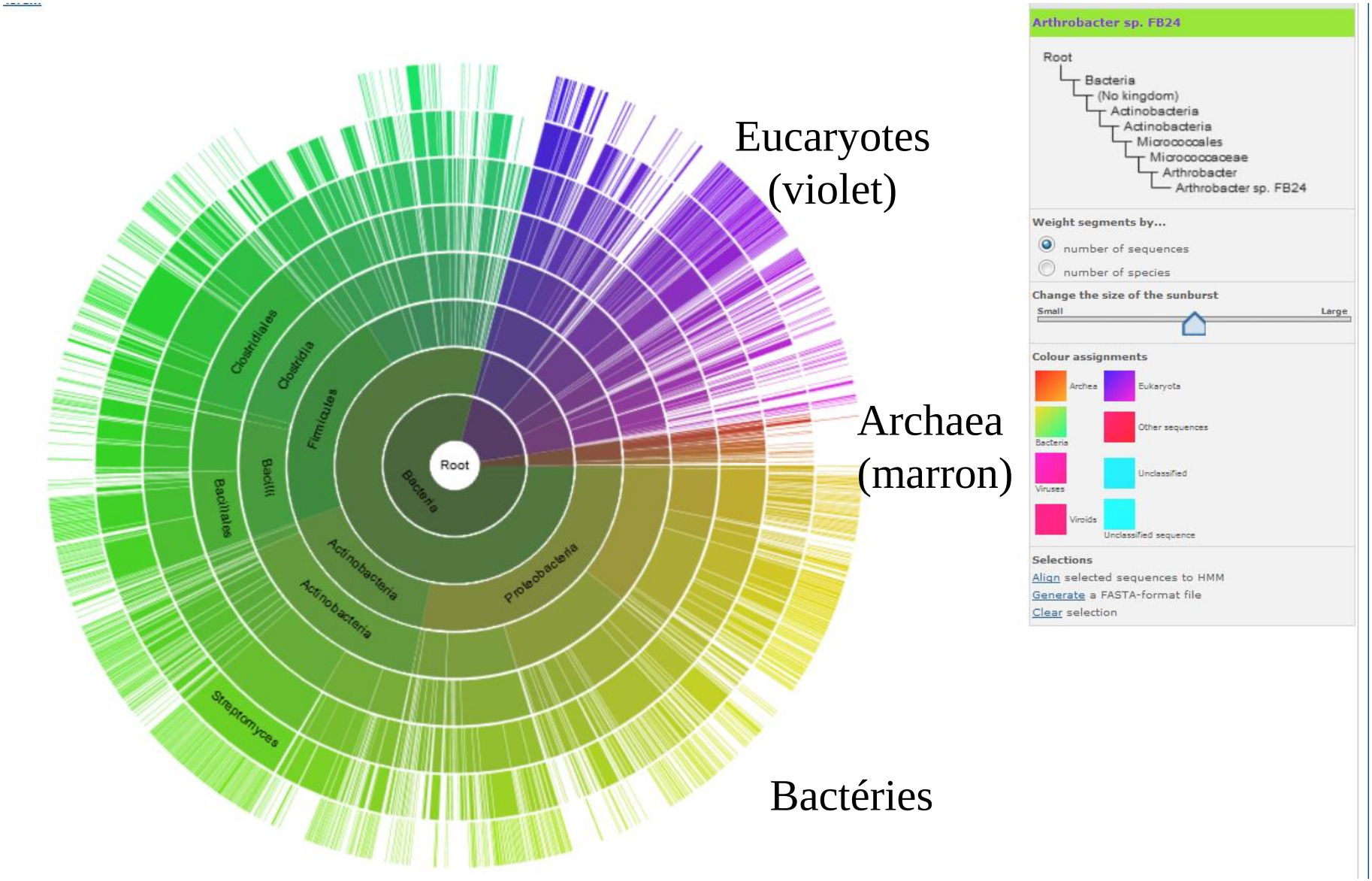
There are 8145 sequences with the following architecture: ABC_tran, oligo_HPY
W8X1U8_CASDE [Castellaniella defragrans 65Phen] Dipeptide transport ATP-binding protein DppF {ECO:0000313|EMBL:CDM26063.1} (380 residues)

Show all sequences with this architecture.

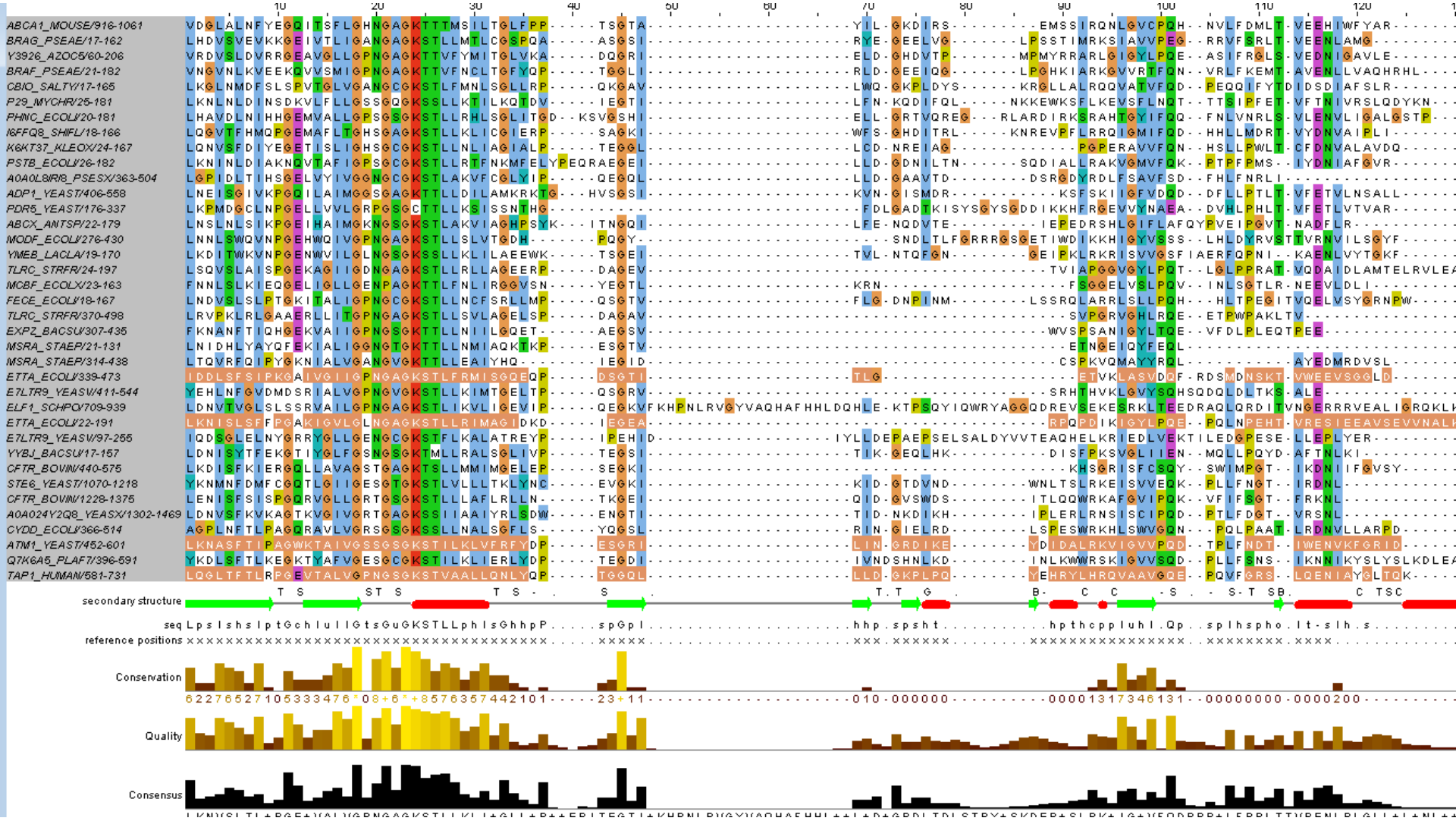
There are 7095 sequences with the following architecture: ABC_tran, TOBE_2
D5RQ25_9PROT [Roseomonas cervicalis ATCC 49957] ABC transporter, ATP-binding protein {ECO:0000313|EMBL:EFH10598.1} (347 residues)

Show all sequences with this architecture.

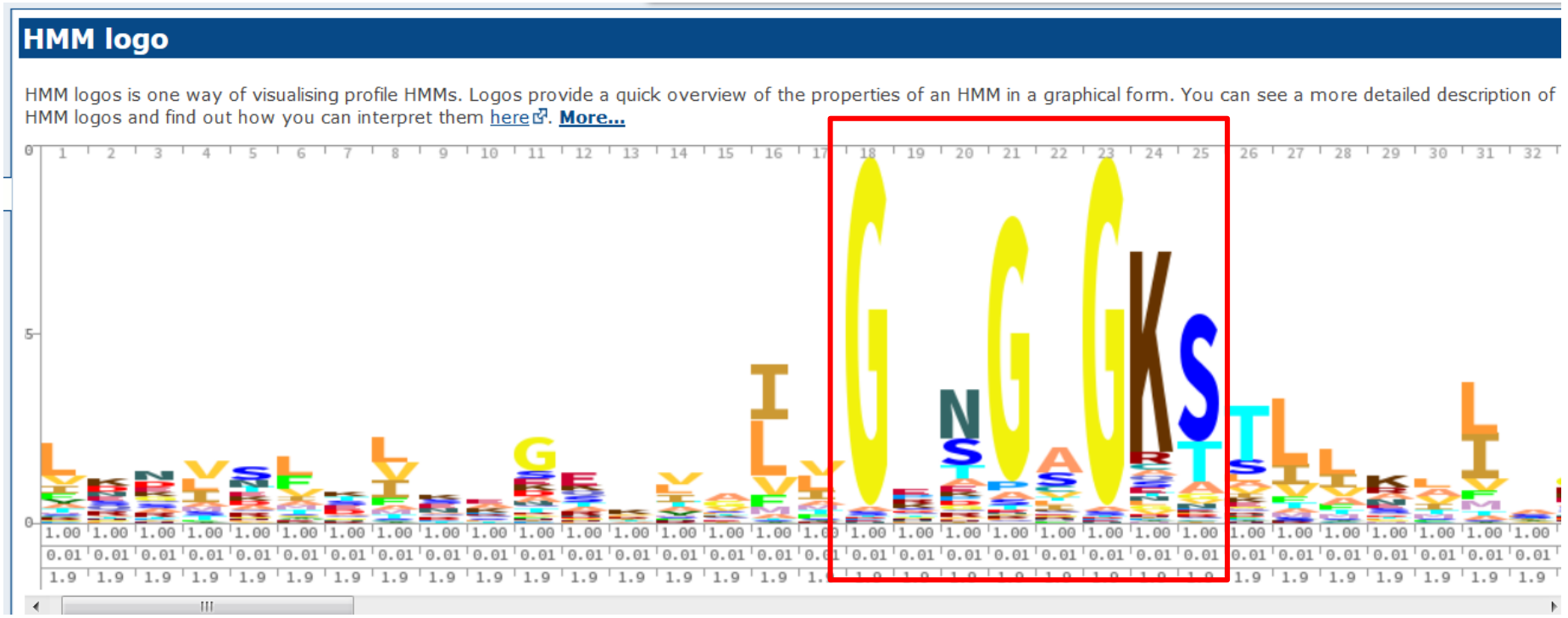
Visualisation graphique simple de cette famille de protéines au sein des espèces



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

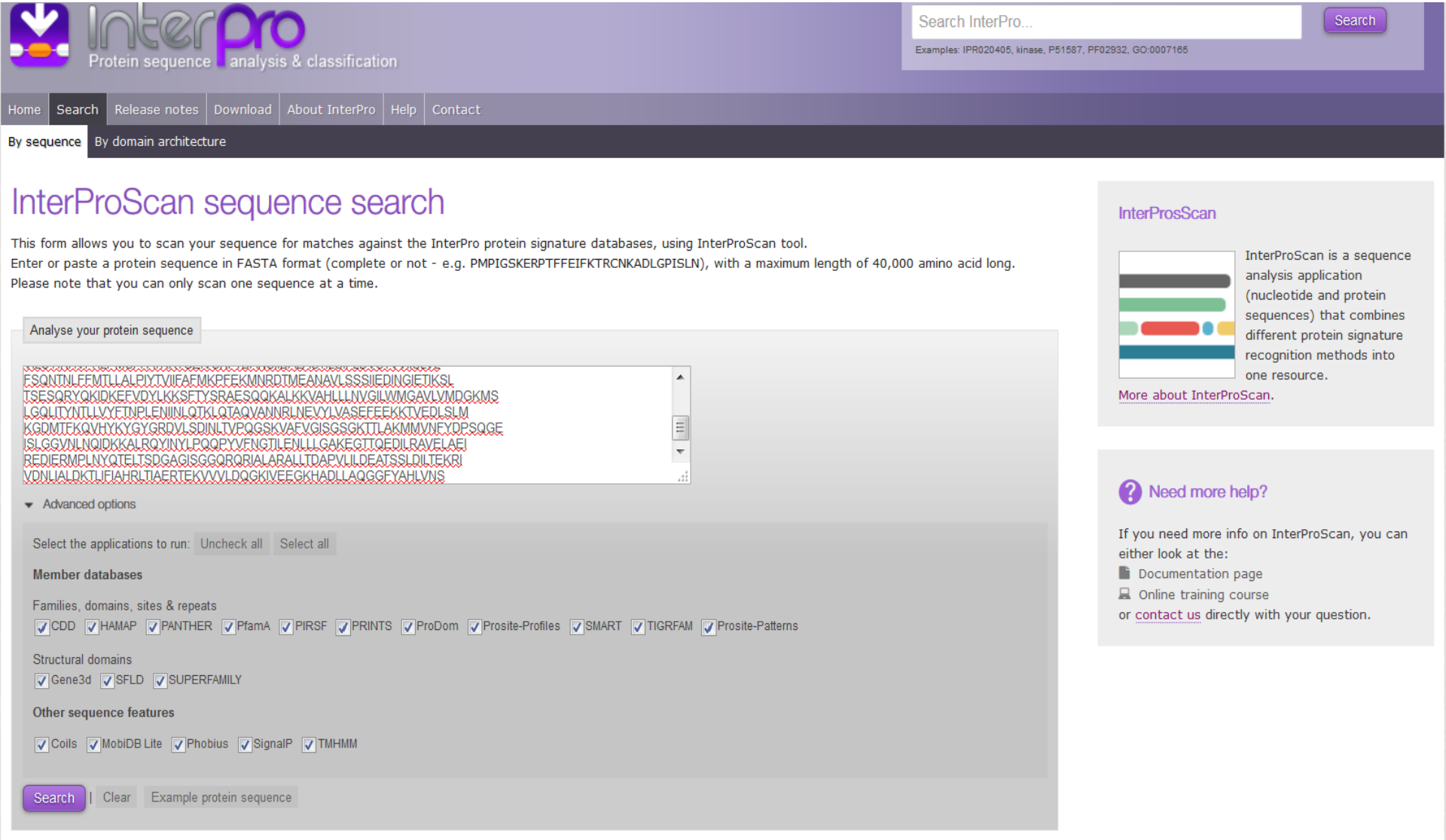


Correspond à la zone fortement conservée de l'alignement précédent et représente le motif Walker A de liaison de l'ATP

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*



https://www.ebi.ac.uk/interpro/search/sequence-search

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

Protein family membership

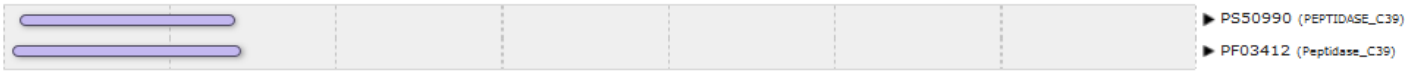
F Peptidase C39, ABC-type bacteriocin transporter (IPR005897)

Domains and repeats

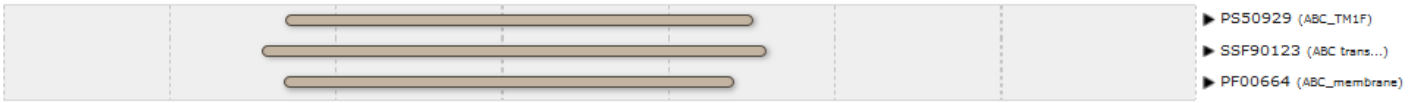
Detailed signature matches

F IPR005897 Peptidase C39, ABC-type bacteriocin transporter

D IPR005074 Peptidase C39, bacteriocin processing

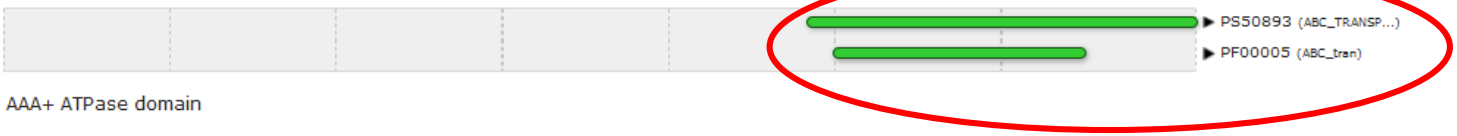


D IPR011527 ABC transporter type 1, transmembrane domain

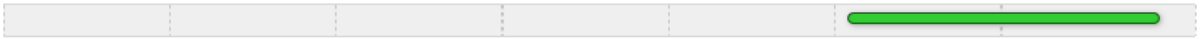


D IPR027417 P-loop containing nucleoside triphosphate hydrolase

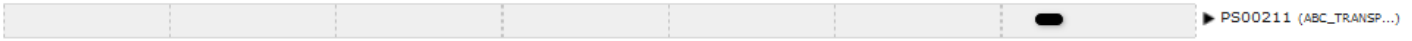
D IPR003439 ABC transporter-like



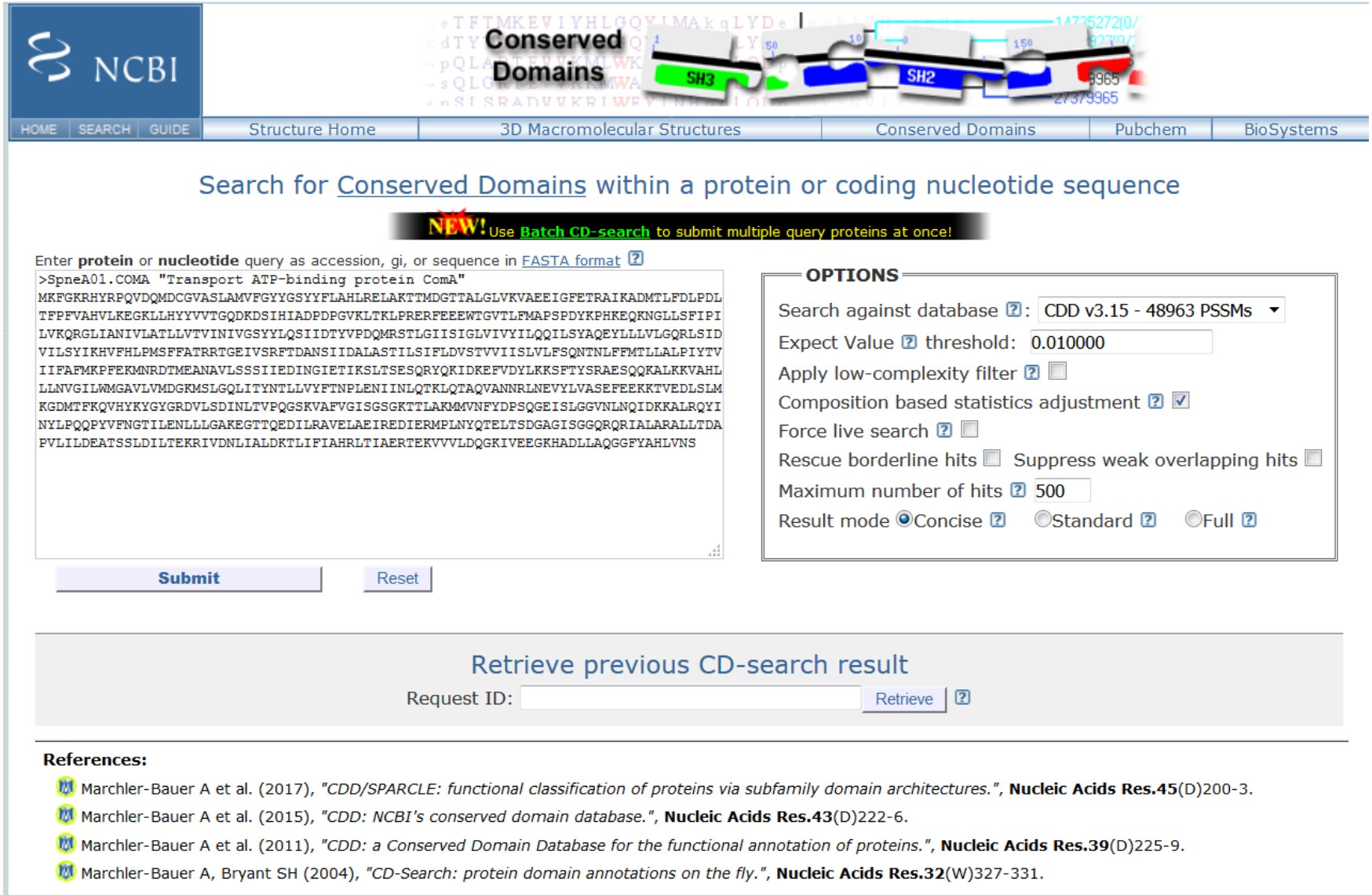
D IPR003593 AAA+ ATPase domain



S IPR017871 ABC transporter, conserved site



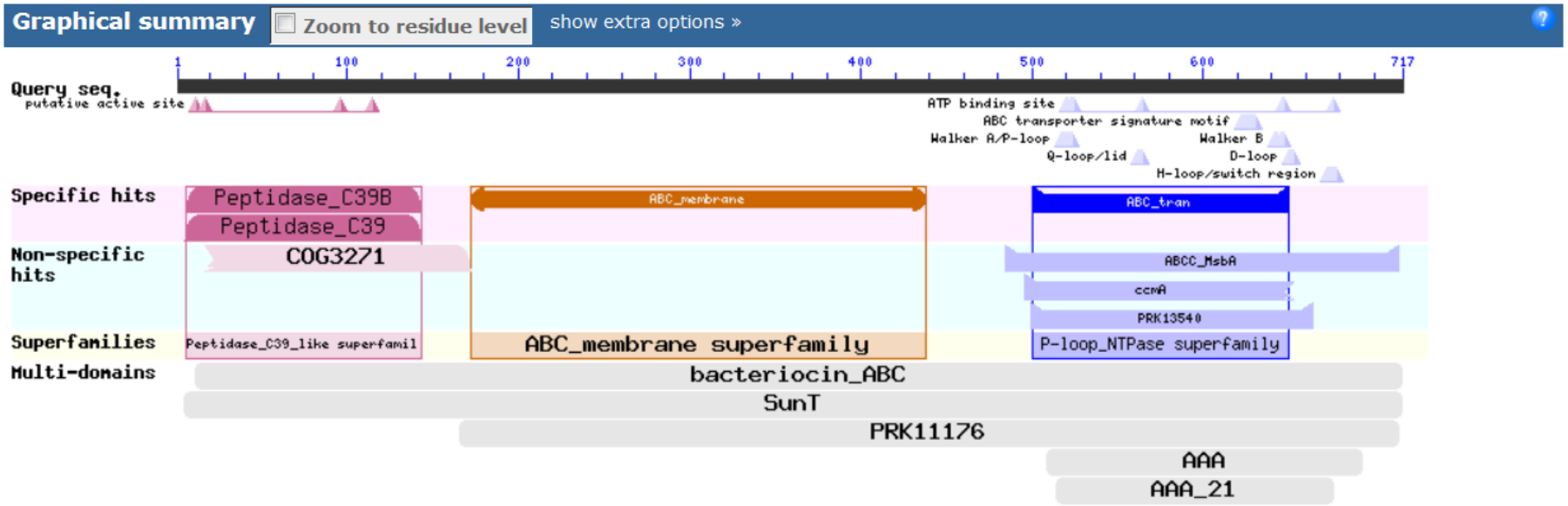
Recherche des domaines fonctionnels dans la séquence ComA de *S. pneumoniae* dans la banque « Conserved Domain Database » (CDD) maintenue au NCBI (CD_search)



https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi

Résultat de la recherche dans la banque CDD

protein containing domains Peptidase_C39B, ABC_membrane, and P-loop_NTPase



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 LSDINLTVPQGSKVAFVGISGSGKTTLAKMMVNFYDPSQGEISLGGVNLNQIDKKALRQYINYLPQQPYVFNG-TILENL 578
Cdd:pfam00005                                   1 LKNVSLTLNPGEILALVGPNGAGKSTLLKLIAGLLSPTEGTILLDGQDLTDDERKSLRKEIGYVFQDPNLFPRlTVRENL  80

                                          90       100       110       120       130       140       150
                                  ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....
seqsig_MKFGK_e2ea59cba8ae2f892bff9c30e69b60cc 579 LLGAKEgttqEDILRAVELAEIREDIERMPLNYQ--TELTSDGAGISGGQRQRIALARALLTDAPVLILDEATS 650
Cdd:pfam00005                                  81 RLGLRL----KGLSKREKDARAEEALEKLGLGDLldRPVGENPGTLSGGQKQRVAIARALLTKPKLLLLDEPTA 150
```