Support de cours
Motifs et Profils

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 $\boxed{f_{b,i} = n_{b,i} / n_{tot}}$ **n_{tot} : nombre total de séquences analysées**

Représentation : Matrice de poids

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 $S \geq$ 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 $D \leq$ 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

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

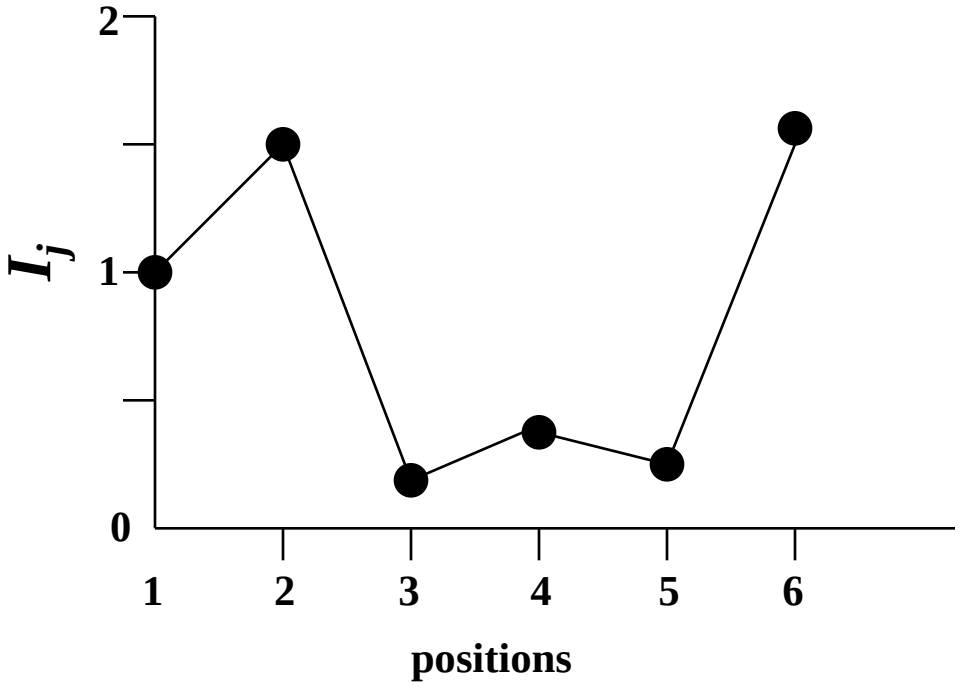
Les positions du signal qui contiendront de l'information seront celles qui auront une composition très biaisées par rapport à ce qui est attendu.

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

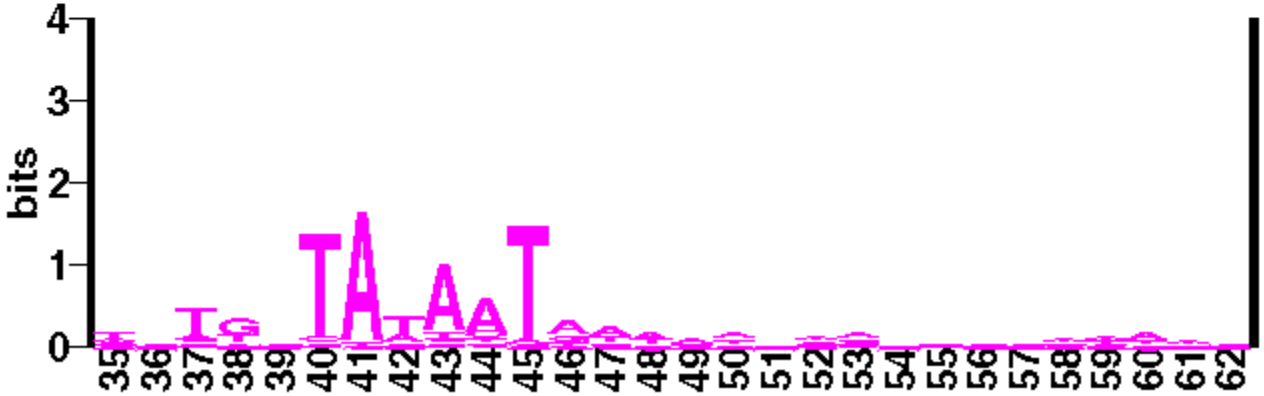
$f_{ij} \log_2 f_{ij} = 0$ et les fréquences observées des autres bases sont nulles. On aura

$I(j) = 2$ 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



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) sensitivity an anglais

Spécificité = VP/(VP+FP) specificity en anglais

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

http://www.expasy.ch/tools/scanprosite/

Home ScanProsite ProRule Documents Downloads Links Funding

ScanProsite

Sequence(s) to be scanned:

Enter:

- UniProtKB(Swiss-Prot and TrEMBL) AC and/or ID *(e.g. P00747, ENTK_HUMAN)*
- PDB identifier(s)
- your own protein sequence(s)

Clear

☑ Exclude motifs with a high probability of occurrence

☐ Do not scan profiles

Motif(s) to scan for:

Enter:

- PROSITE AC and/or ID *(e.g. PS50808, CHEB)*
- your own pattern(s)

Clear

Protein database(s):

☑ UniProtKB/Swiss-Prot ☑ including splice variants

☐ UniProtKB/TrEMBL ☐ PDB

randomize databases no

☐ excluding fragments

Filter(s):

- On taxonomy: *(e.g.Eukaryota; Escherichia coli;)*
- On description: *(e.g.protease)*

Pattern option(s):

- Allow at most 0 X sequence characters to match a conserved position in the pattern
- Match mode greedy, overlaps, no includes

Output:

- Format Graphical rich view
- Show only sequences with at least hit(s)

Maximum of matched sequences 1000

☐ Show low level score

☐ Retrieve complete sequences

Your e-mail:

START THE SCAN reset

ScanProsite

Scan SWISS-PROT with a PROSITE pattern
SWISS-PROT Release 40.10 of 11-Feb-2002: 105224 entries

User-entered pattern
Pattern: [PAI]-[TLRM]-P-[NAS]-[ILV]-[PS]-[EQ]-[MY]-[NASG]-[EASPY]-[ILVF]-W

>P02928 (MALE_ECOLI) Maltose-binding periplasmic protein precursor (Maltodextrin-binding protein) (MMBP) [Escherichia coli, Escherichia coli O157:H7]. [PDB: 1a7l 1mdp 1iud 1anf 1dmb]

355 - 366 IMPNIPQMSAFW

>P18815 (MALE_ENTAE) Maltose-binding periplasmic protein precursor (Maltodextrin-binding protein) (MMBP) [Enterobacter aerogenes (Aerobacter aerogenes)].

355 - 366 IMPNIPQMSAFW

>P19576 (MALE_SALTY) Maltose-binding periplasmic protein precursor (Maltodextrin-binding protein) (MMBP) [Salmonella typhimurium].

355 - 366 IMPNIPQMSAFW

>P29850 (MALX_STRPN) Maltose/maltodextrin-binding protein precursor [Streptococcus pneumoniae].

374 - 385 PLPNISQMSAVW

4 hits in 4 sequences from 4 entries

Exemple d'entrée PROSITE

PROSITE: PS00594

```
ID   AROMATIC_AA_PERMEASE_1; PATTERN.
AC   PS00594;
DT   DEC-1991 (CREATED); NOV-1997 (DATA UPDATE); JUL-1998 (INFO UPDATE).
DE   Aromatic amino acids permeases signature.
PA   I-G-[GA]-G-M-[LF]-[SA]-x-P-x(3)-[SA]-G-x(2)-F.
NR   /RELEASE=40.7,103373;
NR   /TOTAL=8(8); /POSITIVE=8(8); /UNKNOWN=0(0); /FALSE_POS=0(0);
NR   /FALSE_NEG=0; /PARTIAL=0;
CC   /TAXO-RANGE=???P?; /MAX-REPEAT=1;
DR   P22306, MTR_ECOLI , T; P44614, MTR_HAEIN , T; P23173, TNAB_ECOLI, T;
DR   P28785, TNAB_PROVU, T; Q47825, TUTB_ERWHE, T; P18199, TYRP_ECOLI, T;
DR   P44727, TYRP_HAEIN, T; P44747, TYRQ_HAEIN, T;
DO   PDOC00513;
//
```

La banque de données PROSITE est disponible sur le site Expasy: http://www.expasy.ch/

PROSITE: PDOC00513 (documentation)
****
*** Aromatic amino acids permeases signature ***
****

It has been shown [1] that some proteins involved in the transport of aromatic amino acids in Escherichia coli and related bacteria are evolutionary related; these permeases are:

- **Tryptophan-specific transport protein (tryptophan permease) (gene mtr).**
- **Low affinity tryptophan permease (gene tnaB or trpP).**
- **Tyrosine-specific transport protein (tyrosine permease) (gene tyrP).**
- **Tyrosine permease from Erwinia herbicola (gene: tutB).**

These permeases are proteins of about 400 to 420 amino acids which probably contain 11 or 12 transmembrane regions. The best conserved domain is a stretch of 20 residues which seems to be located in a cytoplasmic loop between the first and second transmembrane region.

-Consensus pattern: I-G-[GA]-G-M-[LF]-[SA]-x-P-x(3)-[SA]-G-x(2)-F
-Sequences known to belong to this class detected by the pattern: ALL.
-Other sequence(s) detected in SWISS-PROT: NONE.
-Last update: November 1997 / Pattern and text revised.

[1] Sarsero J.P., Wookey P.J., Gollnick P., Yanofsky C., Pittard A.J.
J. Bacteriol. 173:3231-3234(1991).

Protein: *TF3B_YEAST* (P29056)

Summary

TF3B_YEAST

This is the summary of UniProt entry TF3B_YEAST (P29056).

Description:	Transcription factor IIIB 70 kDa subunit
Source organism:	Saccharomyces cerevisiae (Baker's yeast) (NCBI taxonomy ID 4932) View Pfam proteome data.
Length:	596 amino acids

Please note: when we start each new Pfam data release, we take a copy of the UniProt sequence database. This snapshot of UniProt forms the basis of the overview that you see here. It is important to note that, although some UniProt entries may be removed *after* a Pfam release, these entries will not be removed from Pfam until the *next* Pfam data release.

Pfam domains

This image shows the arrangement of the Pfam domains that we found on this sequence. Clicking on a domain will take you to the page describing that Pfam entry. The table below gives the domain boundaries for each of the domains.

Source	Domain	Start	End
Pfam A	TF_Zn_Ribbon	2	46
Pfam A	TFIIB	87	157
Pfam A	TFIIB	182	255
low_complexity		188	206
low_complexity		390	421
low_complexity		457	479
coiled_coil		465	485
Pfam A	BRF1	466	595
low_complexity		517	528

Family: *TFIIB* (PF00382)

Domain organisation

Below is a listing of the unique domain organisations or architectures in which this domain is found. More...

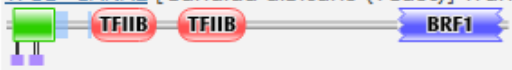
There are 334 sequences with the following architecture: TF_Zn_Ribbon, TFIIB x 2

TF2B_ARCFU [Archaeoglobus fulgidus] Transcription initiation factor IIB (326 residues)

Show all sequences with this architecture.

There are 73 sequences with the following architecture: TF_Zn_Ribbon, TFIIB x 2, BRF1

TF3B_CANAL [Candida albicans (Yeast)] Transcription factor IIIB 70 kDa subunit (553 residues)



Show all sequences with this architecture.

There are 53 sequences with the following architecture: TF_Zn_Ribbon, TFIIB

Q8ZVU4_PYRAE [Pyrobaculum aerophilum] Transcription initiation factor IIB, conjectural (159 residues)

Show all sequences with this architecture.

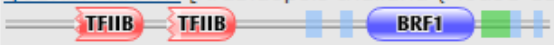
There are 51 sequences with the following architecture: TFIIB

Q9HQ02_HALSA [Halobacterium salinarium (Halobacterium halobium)] Putative uncharacterized protein (103 residues)

Show all sequences with this architecture.

There are 48 sequences with the following architecture: TFIIB x 2, BRF1

Q9SR27_ARATH [Arabidopsis thaliana (Mouse-ear cress)] Putative transcription factor (600 residues)



Show all sequences with this architecture.

There are 42 sequences with the following architecture: TFIIB x 2

TF2B_AERPE [Aeropyrum pernix] Transcription initiation factor IIB (322 residues)

Show all sequences with this architecture.

There are 8 sequences with the following architecture: TFIIB, BRF1

Q9S9Q7_ARATH [Arabidopsis thaliana (Mouse-ear cress)] F26G16.4 protein (294 residues)

Family: *TFIIB* (PF00382)

15 architectures | 1120 sequences | 2 interactions | 186 species | 20 structures

Summary
Domain organisation
Alignments
HMM logo
Trees
Curation & models
Species
Interactions
Structures

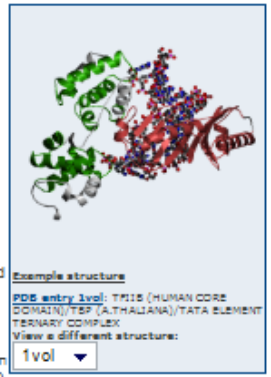
Jump to...
enter ID/acc Go

Summary

Transcription factor TFIIB repeat

Add annotation

No Pfam abstract.



InterPro entry IPR013150

Cyclins are eukaryotic proteins that play an active role in controlling nuclear cell division cycles PUBMED:12910258, and regulate cyclin dependent kinases (CDKs). Cyclins, together with the p34 (cdc2) or cdk2 kinases, form the Maturation Promoting Factor (MPF). There are two main groups of cyclins, G1/S cyclins, which are essential for the control of the cell cycle at the G1/S (start) transition, and G2/M cyclins, which are essential for the control of the cell cycle at the G2/M (mitosis) transition. G2/M cyclins accumulate steadily during G2 and are abruptly destroyed as cells exit from mitosis (at the end of the M-phase). In most species, there are multiple forms of G1 and G2 cyclins. For example, in vertebrates, there are two G2 cyclins, A and B, and at least three G1 cyclins, C, D, and E.

Cyclin homologues have been found in various viruses, including Saimiriine herpesvirus 2 (Herpesvirus saimiri) and Human herpesvirus 8 (HHV-8) (Kaposi's sarcoma-associated herpesvirus). These viral homologues differ from their cellular counterparts in that the viral proteins have gained new functions and eliminated others to harness the cell and benefit the virus PUBMED:11056549.

In eukaryotes, transcription initiation of all protein encoding genes involves the polymerase II system. This sytem is modulated by both general and specific transcription factors. The general factors (which include TFIIA, TFIIB, TFIID, TFIIE, TFIIF, TFIIG and TFIIH) operate through common promoter elements, such as the TATA box. Transcription factor IIB (TFIIB) is of central importance in transcription of class II genes. It associates with TFIID-TFIIA bound to DNA (the DA complex) to form a ternary TFIID-IIA-IBB (DAB) complex, which is recognised by RNA polymerase II PUBMED:1876184, PUBMED:1949150. TFIIB comprises ~315-340 residues and contains an imperfect C-terminal repeat of a 75-residue domain that may contribute to the symmetry of the folded protein. The basal archaeal transcription machinery resembles that of the eukaryotic polymerase II system and includes a homologue of TFIIB PUBMED:7597027.

This entry represents a cyclin-like domain which is found repeated in the C-terminal region of a variety of eukaryotic TFIIB's and their archaeal counterparts. These domains individually form the typical cyclin fold, and in the transcription complex they straddle the C-terminal region of the TATA-binding protein - an interaction essential for the formation of the transcription initiation complex PUBMED:9177165, PUBMED:10619841.

Clan

This family is a member of clan Cyclin (CL0065), which contains the following 7 members:

CDK5_activator | Cyclin | Cyclin_C | Cyclin_N | RB_A | RB_B | TFIIB

Gene Ontology

Cellular component	nucleus (GO:0005634)
Molecular function	translation initiation factor activity (GO:0003743)
	protein binding (GO:0005515)
Biological process	translational initiation (GO:0006413)

Internal database links

SCOOP:	YscJ_FliF

External database links

HOMSTRAD:	transcript_fac2
PANDIT:	PF00382
PRINTS:	PR00685
PROSITE:	PDOC00624
Pseudofam:	PF00382
SCOP:	1vol
SYSTERS:	TFIIB

wellcome trust sanger institute

Seed sequence alignment for PF00382

```
TF2B_PYRWO/116-186          LDRITAQLKLPR...HVEEEAARLYREAVRKGLIRGRSIESVMAACVYAACRLLKVPRTLDEIADIARVDKKEI
TF2B_PYRWO/116-186 (SS)     HHHHHHHHT--H...HHHHHHHHHHHHHHC--------HHHHHHHHHHHHHHHHT----HHHHHHHTTS-TTT-
TF3B_KLULA/193-266          IQHFAEKLELGDKKIKVIRDAVKLAQTMSRDWMYEGRRPAGIAGACLLLACRMNNLRRTHSEIVAISHVAEETL
TF3B_CANAL/190-263          IQHFVEKLDFKDKATKVAKDAVKLAHRMAADWIHEGRRPAGIAGACVLLAARMNNFRRSHAEIVAVSHVGEETL
TF2B_HALVA/30-100           VPRFASELELSE...EVQSKANEIIDTTAEQGLLSGKSPTGYAAAAIYAASLLCNEKKTQREVADVAQVTEVTI
TF2B_PYRWO/212-282          VNKFADELGLSE...KVRRRAIEILDEAYKRGLTSGKSPAGLVAAALYIASLLEGEKRTQREVAEVARVTEVTV
TF2B_PYRWO/212-282 (SS)     HHHHHHH----H...HHHHHHHHHHHHHHH------------HHHHHHHHHH------HHHHHHHTS-HHHH
TF2B_DROME/214-284          MCRFCANLDLPN...MVQRAATHIAKKAVEMDIVPGRSPISVAAAAIYMASQASEHKRSQKEIGDIAGVADVTI
TF2B_KLULA/245-315          IPRFCSHLGLSV...QVANAAEYIAKHSKDVNVLAGRSPITIAAAAIYMATLLFKLNISPTRISQTLQVTEGTV
TF3B_CANAL/95-165           IKRIAAALKIPD...YIAEAAGEWFRLALTLNFVQGRRSNNVLATCLYVACRKERTHHMLIDFSSRLQISVYSL
TF3B_KLULA/98-168           LKAVSYALNIPE...YVTDAAFQWYRLALSNNFVQGRKSQNVIAACLYIACRKERTHHMLIDFSSRLQVSVYSI
TF2B_DROME/120-190          ISSMADRINLPK...TIVDRANNLFKQVHDGKNLKGRSNDAKASACLYIACRQEGVPRTFKEICAVSKISKKEI
TF2B_KLULA/137-207          ITMMCDAAELPK...IVKDCAKEAYKLCFEERVLKGKSQESIMASVILVGCRRAEVGRSFKEILSLTNVRKKEI
TF2B_YEAST/133-203          ITMLCDAAELPK...IVKDCAKEAYKLCHDEKTLKGKSMESIMAASILIGCRRAEVARTFKEIQSLIHVKTKEF
```

This alignment is coloured according to the ClustalX colouring scheme:

- Glycine (G)
- Proline (P)
- Small or hydrophobic (A,V,L,I,M,F,W)
- Hydroxyl or amine amino acids (S,T,N,Q)
- Charged amino-acids (D,E,R,K)
- Histidine or tyrosine (H,Y)

For UniProt-based alignments, we also add some additional mark-up to the alignments where appropriate. Active site information is shown as follows:

- Active site (residue annotated in SwissProt as an active site)
- Predicted active site (residue aligns in a Pfam alignment with a SwissProt active site)
- Predicted active site (residue annotated in SwissProt as a potential active site)

Some UniProt sequences can be mapped to protein structures, in which case we also show the secondary structure definition. These lines are shown below the sequence to which they apply and are marked (SS). The meaning of each of the symbols is as follows:

Family: *TFIIB* (PF00382)

15 architectures 1120 sequences 2 interact

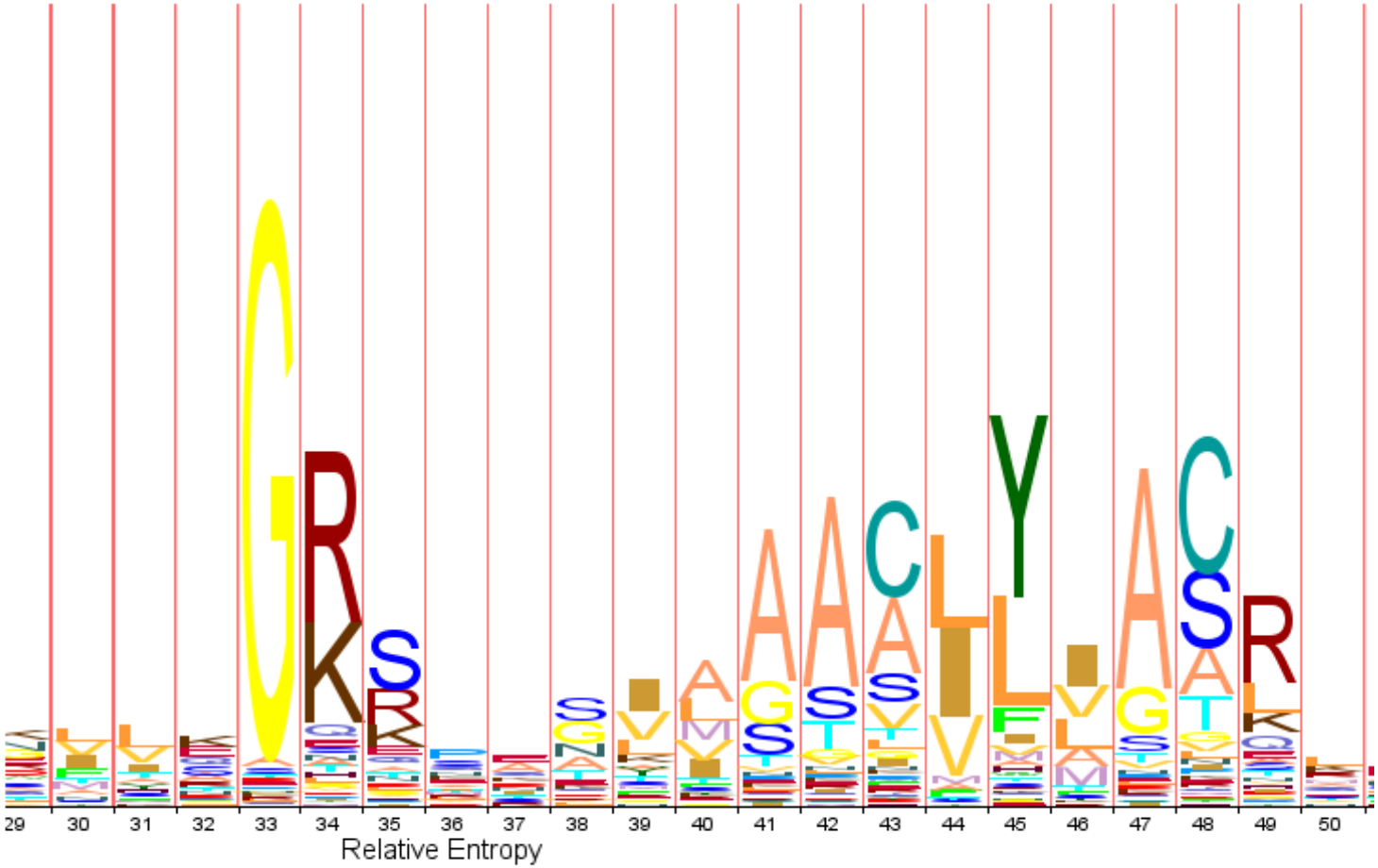
- Summary
- Domain organisation
- Alignments
- HMM logo
- Trees
- Curation & models
- Species
- Interactions
- Structures

Jump to...

enter ID/acc Go

HMM logo

HMM logos is one way of visualising profile HMMs. Logos provide a quick overview of the properties of an HMM in a detailed description of HMM logos and find out how you can interpret them here. **More...**



Species distribution

The tree shows the occurrence of this domain across different species. **More...**

- unclassifiedsequences 1 1 2
  - uncultured marine microorganism HF4000_APKG7N23 1 1 2
- Eukaryota 152 381 686
  - Alveolata 20 39 61
  - Euglenozoa
    - Kinetoplastida
      - Trypanosomatidae 5 5 10
        - Trypanosoma brucei 1 1 2
        - Leishmania
          - Leishmania braziliensis 1 1 2
        - Leishmania infantum 1 1 2
        - Leishmania major 1 1 2
        - Trypanosoma
          - Trypanosoma cruzi 1 1 2
  - stramenopiles
  - Metazoa 36 97 175
  - Amoebozoa 3 4 8
  - Fungi 61 140 272
    - Microsporidia 4 13 25
    - Dikarya 57 127 247
      - Ascomycota 52 113 220
        - Pezizomycotina 30 61 117
        - Taphrinomycotina
        - Saccharomycotina
          - Saccharomycetes
            - Saccharomycetales 20 48 95
              - Metschnikowiaceae
              - mitosporicSaccharomycetales 3 12 24
              - Saccharomycetaceae 15 32 64
                - Debaryomyces hansenii (Yeast) (Torulaspora hansenii) 1 2 4
                - Ashbya gossypii (Yeast) (Eremothecium gossypii) 1 2 4
                - Nakaseomyces
                  - Candida glabrata (Yeast) (Torulopsis glabrata) 1 2 4
                - Saccharomyces cerevisiae (strain YJM789) (Baker's yeast) 1 2 4
                - Kluyveromyces lactis (Yeast) (Candida sphaerica) 1 2 4
                - Pichia stipitis (Yeast) 1 2 4
                - Vanderwaltozyma polyspora (strain ATCC 22028 / DSM 70294) (Kluyveromyces polysporus) 1 2 4
                - Zygosaccharomyces rouxii (Candida mogii) 1 3 6
                - Lachancea thermotolerans (Yeast) (Kluyveromyces thermotolerans) 1 2 4
                - Saccharomyces cerevisiae (strain RM11-1a) (Baker's yeast) 1 2 4
                - Saccharomyces cerevisiae (strain AWRI1631) (Baker's yeast) 1 2 4
                - Saccharomyces cerevisiae (Baker's yeast) 1 3 6

Tree controls Hide

Fully expand tree

Fully collapse tree

Expand to depth 21 Go

Annotation

Hide highlighting of species in seed

Hide summaries

Key: species, sequences, regions

Download tree

Save a text representation

Selected sequences

(Uncheck all)

View

- graphically
- as an alignment

Download

- sequence accessions
- sequences in FASTA format

InterPro

www.ebi.ac.uk/interpro/IEntry?ac=IPR000254

EMBL-EBI | EB-eye Search | All Databases | Enter Text Here | Go | Reset | Advanced Search | Give us feedback

Databases | Tools | EBI Groups | Training | Industry | About Us | Help | Site Index

EBI > Databases > InterPro

Search InterPro: >

Jump to: InterProScan Databases Documentation FTP site Help Advanced search

IPR000254 Cellulose-binding domain, fungal

Protein matches

UniProtKB Matches: 699 proteins	Overview: sorted by AC, sorted by name, of known structure, proteins with splice variants Detailed: sorted by AC, sorted by name, of known structure proteins with splice variants Table: For all matching proteins, of known structure Architectures Accession List Matches in BioMart
Accession	IPR000254 Cellulose-bd_dom_fun
Type	Domain

Signatures

Database	ID	Name	Proteins
ProDom	PD001821	CBD_fun	447
Pfam	PF00734	CBM_1	680
PROSITE pattern	PS00562	CBM1_1	552
PROSITE profile	PS51164	CBM1_2	658
SMART	SM00236	fCBD	619
SuperFamily	SSF57180	CBD_fun	663

Signatures in BioMart

GO Term annotation

Process	GO:0005975 carbohydrate metabolic process
Function	GO:0004553 hydrolase activity, hydrolyzing O-glycosyl compounds GO:0030248 cellulose binding

Quelques adresses utiles

- **PBIL (Pôle Bioinformatique Lyonnais) http://biom1.univ-lyon1.fr**
- **Pasteur http://www.pasteur.fr**
- **Génopole Toulouse : http://bioinfo.genopole-toulouse.prd.fr/**
- **Expasy (Suisse) http://www.expasy.ch (nombreux logiciels pour l'analyse des séquences protéiques, banques SwissProt et Prosite)**
- **NCBI (Etats-Unis) http://www.ncbi.nlm.nih.gov (Blast et PubMed)**
- **EMBL (laboratoire européen, Cambridge) http://www.ebi.ac.uk**