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

$n_{tot}$ : 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

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

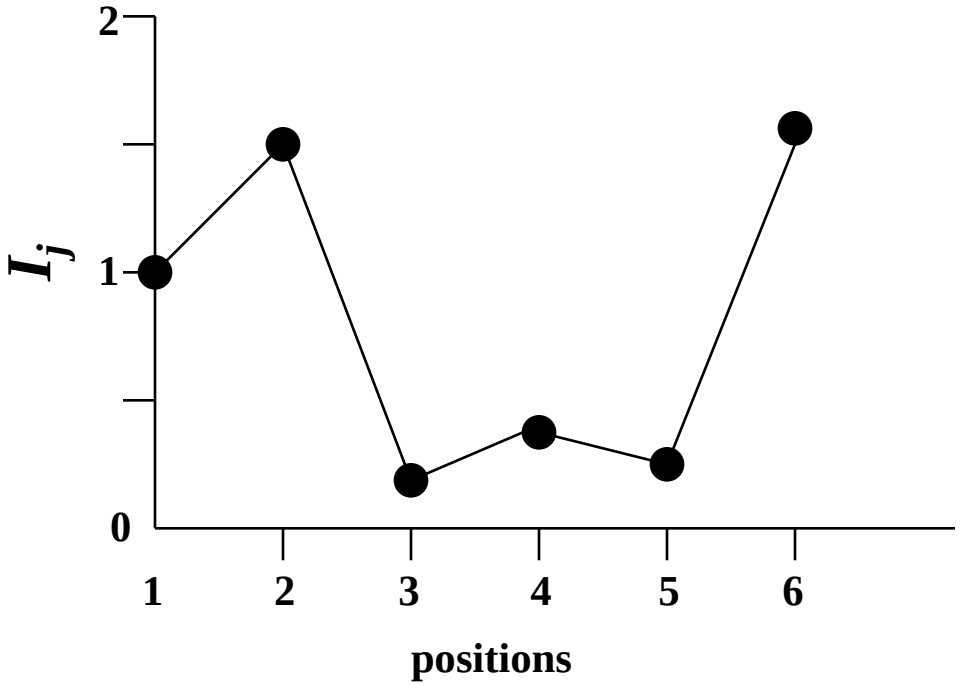
Les positions du signal qui contiendront de l'information seront celles qui auront une composition très biaisées par rapport à ce qui est attendu.

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

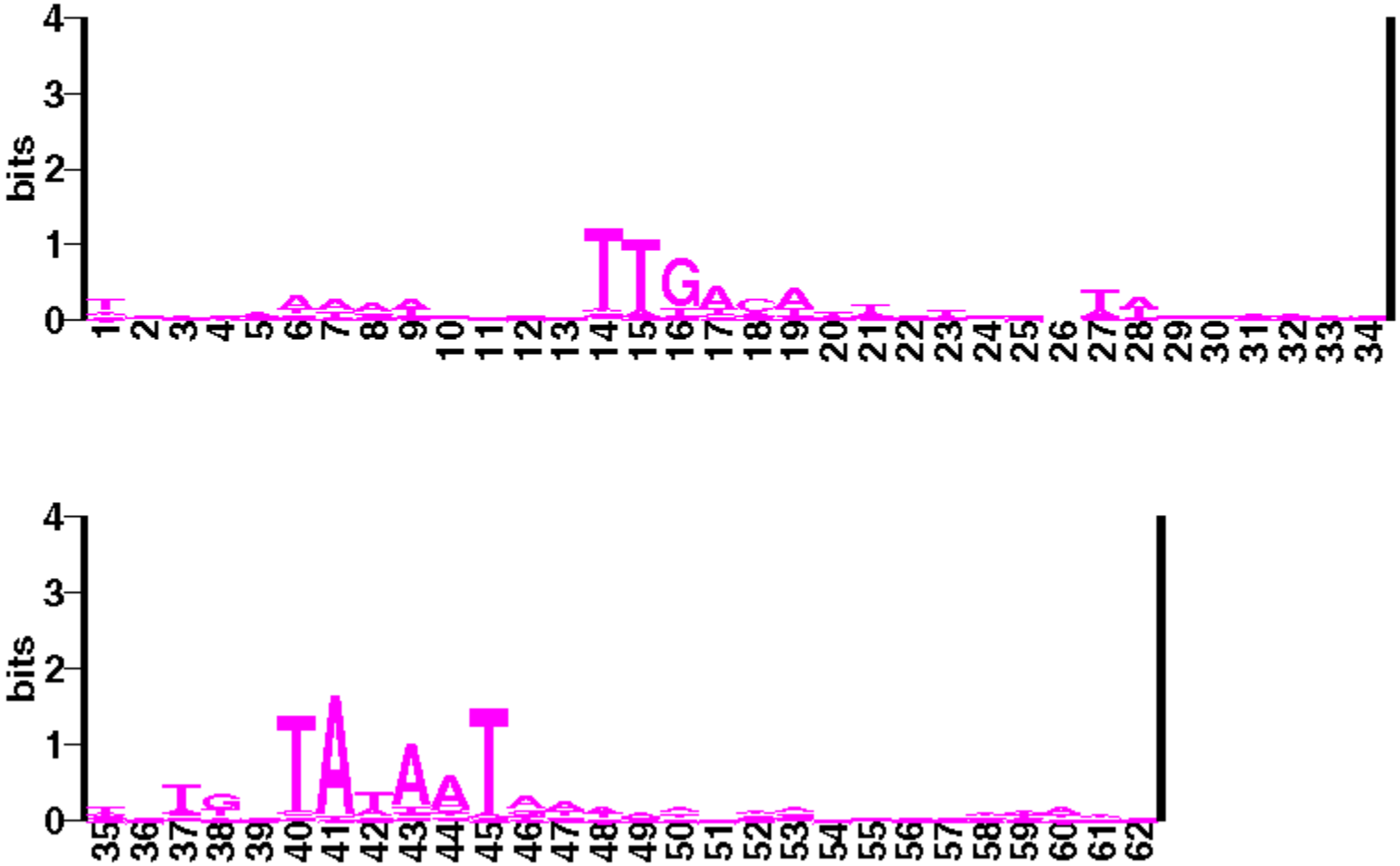
$f_{ij} \log_2 f_{ij} = 0$ et les fréquences observées des autres bases sont nulles. On aura

$I(j) = 2$ 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

Description Technical section References Copyright Miscellaneous

## Description

Bacterial high affinity transport systems are involved in active transport of solutes across the cytoplasmic membrane. The protein components of these traffic systems include one or two transmembrane protein components, one or two membrane-associated ATP-binding proteins (ABC transporters; see <PDOC00185>) and a high affinity periplasmic solute-binding protein. The later are thought to bind the substrate in the vicinity of the inner membrane, and to transfer it to a complex of inner membrane proteins for concentration into the cytoplasm.

In gram-positive bacteria which are surrounded by a single membrane and have therefore no periplasmic region the equivalent proteins are bound to the membrane via an N-terminal lipid anchor. These homolog proteins do not play an integral role in the transport process per se, but probably serve as receptors to trigger or initiate translocation of the solute through the membrane by binding to external sites of the integral membrane proteins of the efflux system.

In addition at least some solute-binding proteins function in the initiation of sensory transduction pathways.

On the basis of sequence similarities, the vast majority of these solute-binding proteins can be grouped [1] into eight families of clusters, which generally correlate with the nature of the solute bound.

Family 1 currently includes the following proteins:

- Periplasmic maltose/maltodextrin-binding proteins of Enterobacteriaceae (gene malE) and homologous lipoprotein in Streptococcus pneumoniae (gene malX).
- Periplasmic multiple oligo-saccharide binding protein of Streptococcus mutans (gene msmE), which is involved in the uptake of melibiose, raffinose and isomaltotriose.
- Periplasmic glycerol-3-phosphate-binding protein of Escherichia coli, a component of the phosphate-limitation inducible uptake system for sn-glycerol-3-phosphate and glycerophosphoryl diesters.
- Periplasmic iron-binding protein from Serratia marcescens (gene sfuA) and the homologous proteins (gene fbp) from Haemophilus influenzae and Neisseria, which are part of the iron-acquisition system.
- Periplasmic thiamine-binding protein (gene tbpA) from Escherichia coli and Haemophilus influenzae.

The signature pattern for this family is located in the central section of the mature proteins.

Last update:

December 2004 / Pattern and text revised.

## Technical section

PROSITE method (with tools and information) covered by this documentation:

SBP_BACTERIAL_1, PS01037; Bacterial extracellular solute-binding proteins, family 1 signature (PATTERN)

- Consensus pattern:
  [GAP]-[LIVMFA]-[STAVDN]-x-{H}-x(2)-[GSAV]-[LIVMFY](2)-Y-[ND]-x(3)-[LIVMF]-x-[KNDE]
- Sequences in UniProtKB/Swiss-Prot known to belong to this class: 82
  - detected by PS01037: 38 (true positives)
  - undetected by PS01037: 44 (42 false negatives and 2 "partials")
- Other sequence(s) in UniProtKB/Swiss-Prot detected by PS01037:
  11 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 tree view of all UniProtKB (Swiss-Prot + TrEMBL) entries matching PS01037
- Retrieve a list of all UniProtKB (Swiss-Prot + TrEMBL) entries matching PS01037
- Scan UniProtKB (Swiss-Prot and/or TrEMBL) entries against PS01037
- View ligand binding statistics of PS01037
- Matching PDB structures: 1A7L 1ANF 1D9V 1DMB ... [ALL]

## Reference

| | | |
|---|---|---|
| 1 | Authors | Tam R., Saier M.H. Jr. |
| | Title | Structural, functional, and evolutionary relationships among extracellular solute-binding receptors of bacteria. |
| | Source | Microbiol. Rev. 57:320-346(1993). |
| | PubMed ID | 8336670 |

## Copyright



## Miscellaneous

View entry in original PROSITE document format
View entry in raw text format (no links)

# 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.

Exemple d'une entrée ProSite

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

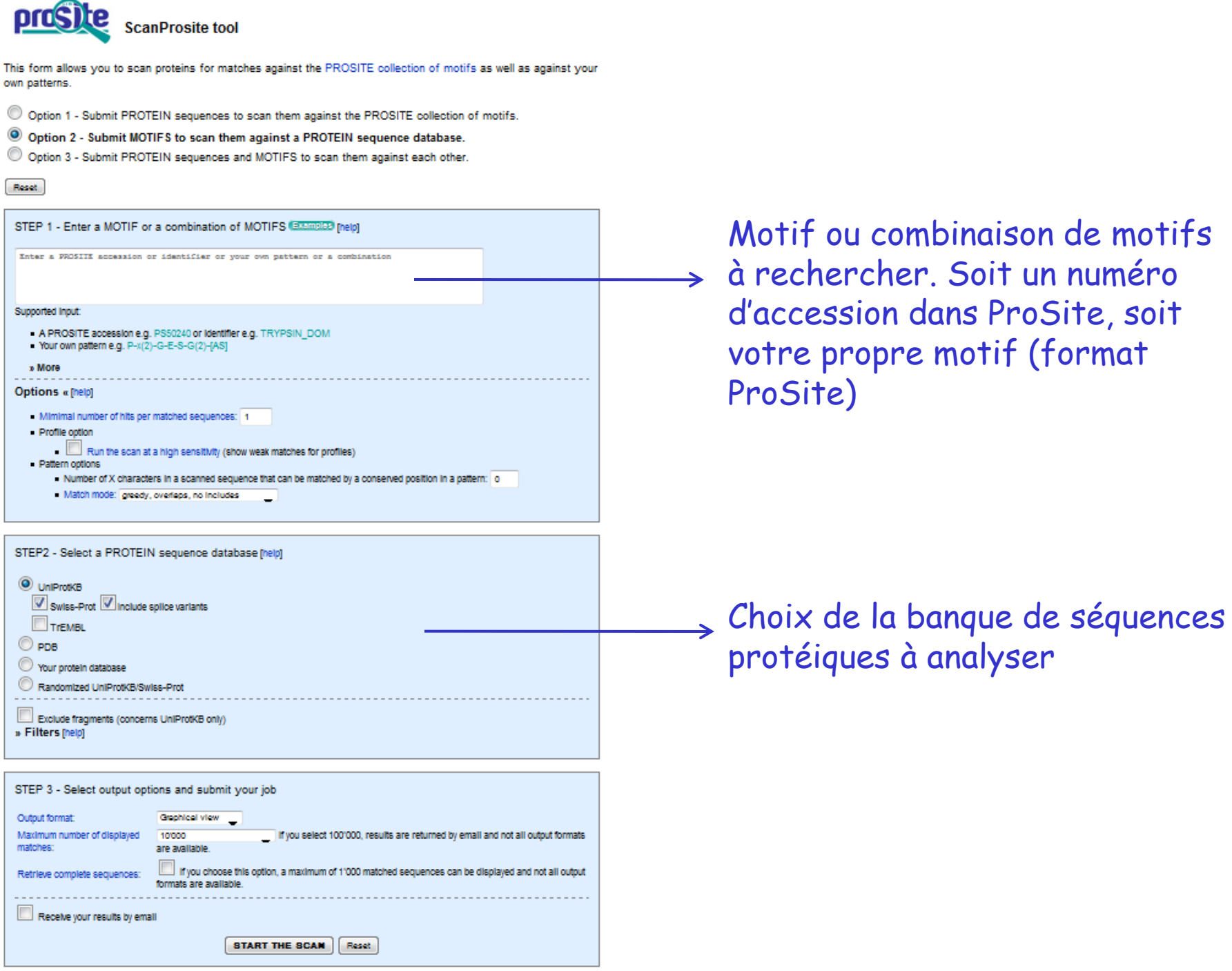
Output format: Graphical view

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}
```



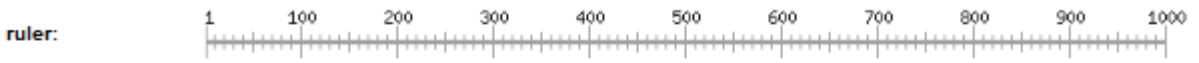

Motif ou combinaison de motifs à rechercher. Soit un numéro d'accession dans ProSite, soit votre propre motif (format ProSite)

Choix de la banque de séquences protéiques à analyser

# Résultat d'une recherche de motifs avec ScanProsite

**hits by patterns:** [8 hits (by 1 pattern) on 8 sequences]

Hits by **USERPAT1** :
Pattern: **[PAI]-[TLRM]-P-[NAS]-[ILV]-[PS]-[EQ]-[MY]-[NASG]-[EASPY]-[ILVF]-W**
Approximate number of expected random matches [Ref: PMID 11535175] in ~ 100'000 sequences (50'000'000 residues): 6.659160e-04

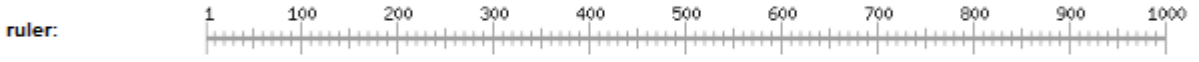



O07009
(CYCB_BACSU) (421 aa) View all PROSITE motifs hits on sequence

**Cyclodextrin-binding protein.** *Bacillus subtilis (strain 168)*

**371 - 382:** PTPNIPEMNEIW

---




P0AEY0
(MALE_ECO57) (396 aa) View all PROSITE motifs hits on sequence

**Maltose-binding periplasmic protein.** *Escherichia coli O157:H7*
Hits on PDB 3D structures: [3VD8-A, 4MY2-A]

**355 - 366:** IMPNIPQMSAFW

---


P0AEX9
(MALE_ECOLI) (396 aa) View all PROSITE motifs hits on sequence

**Maltose-binding periplasmic protein.** *Escherichia coli (strain K12)*
Hits on PDB 3D structures: [3C4M-B, 3SEW-A, 4KI0-E]

**355 - 366:** IMPNIPQMSAFW

---

# Banque de données Pfam

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.

# Recherche des domaines fonctionnels Pfam d'une protéine

## Protéine :TF3B_YEAST Transcription factor IIIB (AC: P29056 )

wellcome trust sanger institute

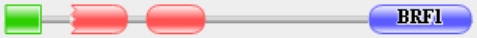
HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Pfam keyword search Go

### Sequence search results

**Show** the detailed description of this results page.

We found **4** Pfam-A matches to your search sequence (**all** significant). You did not choose to search for Pfam-B matches.




**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 | | | | | |
| TF_Zn_Ribbon | TFIIB zinc-binding | Domain | CL0167 | 2 | 46 | 3 | 45 | **2** | **42** | 43 | 40.9 | 8.8e-11 | n/a | Show |
| TFIIB | Transcription factor TFIIB repeat | Domain | CL0065 | 87 | 157 | 89 | 157 | **3** | 71 | 71 | 86.4 | 7.8e-25 | n/a | Show |
| TFIIB | Transcription factor TFIIB repeat | Domain | CL0065 | 182 | 255 | 182 | 255 | 1 | 71 | 71 | 94.6 | 2.3e-27 | n/a | Show |
| BRF1 | Brf1-like TBP-binding domain | Family | n/a | 466 | 595 | 467 | 595 | 1 | 97 | 97 | 113.7 | 3.3e-33 | n/a | Show |

# Exemple d'une entrée Pfam

wellcome trust sanger institute

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Pfam keyword search Go

## Family: *TFIIB* (PF00382)

23 architectures | 2462 sequences | 2 interactions | 536 species | 21 structures

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

Jump to... enter ID/acc Go

### Summary: Transcription factor TFIIB repeat

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

No Wikipedia article | Pfam | InterPro

This tab holds the annotation information that is stored in the Pfam database. As we move to using Wikipedia as our main source of annotation, the contents of this tab will be gradually replaced by the Wikipedia tab.

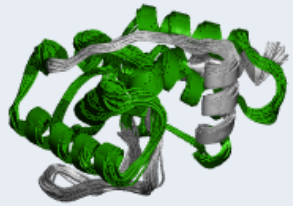
### Transcription factor TFIIB repeat

Provide feedback

No Pfam abstract.

### External database links

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



Example structure

PDB entry 1TFB: NMR STUDIES OF HUMAN GENERAL TRANSCRIPTION FACTOR TFIIB: DYNAMICS AND INTERACTION WITH VP16 ACTIVATION DOMAIN, 20 STRUCTURES

View a different structure: 1TFB

# Family: *TFIIB* (PF00382)

23 architectures | 2462 sequences | 2 interactions | 536 species | 21 structures

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

Jump to...
enter ID/acc Go

## Summary: Transcription factor TFIIB repeat

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

No Wikipedia article | Pfam | InterPro

This tab holds annotation information from the InterPro database.

### 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].

**Gene Ontology**

The mapping between Pfam and Gene Ontology is provided by InterPro. If you use this data please cite InterPro.

| | |
|---|---|
| **Molecular function** | translation initiation factor activity (GO:0003743) |
| **Biological process** | translational initiation (GO:0006413) |

# 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

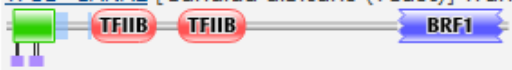
## Domain organisation

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.

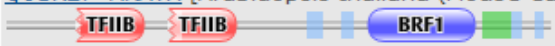
**There are 53 sequences with the following architecture: TF_Zn_Ribbon, TFIIB**
Q8ZVU4_PYRAE [Pyrobaculum aerophilum] Transcription initiation factor IIB, conjectural (159 residues)


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.

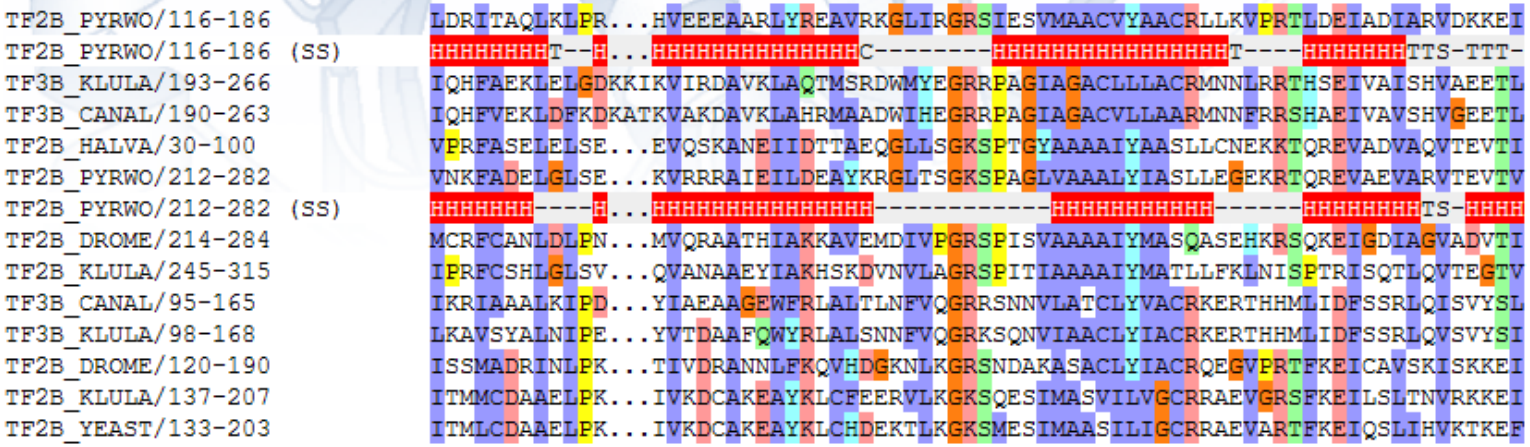
**There are 42 sequences with the following architecture: TFIIB x 2**
TF2B_AERPE [Aeropyrum pernix] Transcription initiation factor IIB (322 residues)


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)

wellcome trust sanger institute

# Seed sequence alignment for PF00382




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

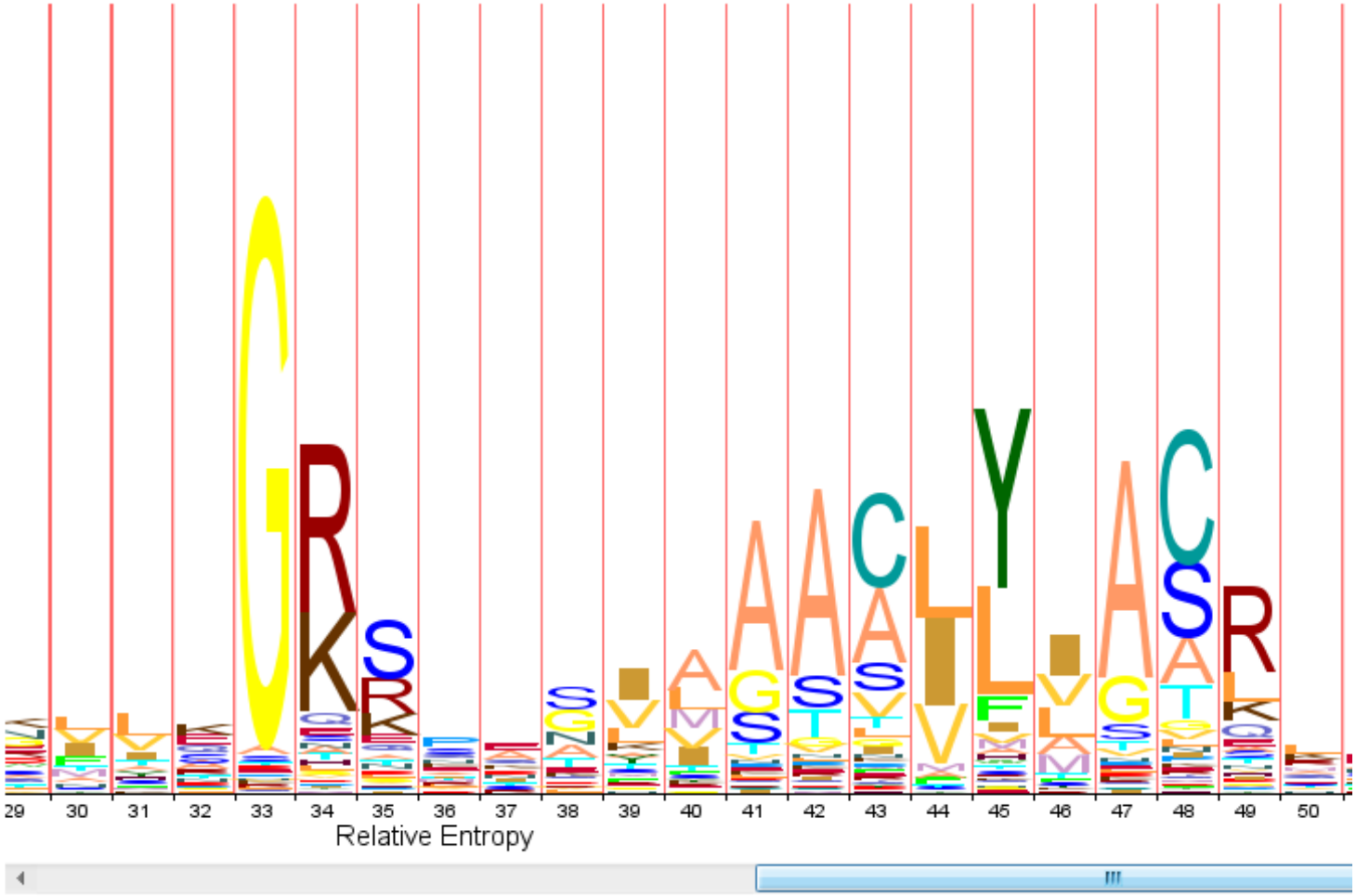


Jump to...

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...**



# Family: *TFIIB* (PF00382)



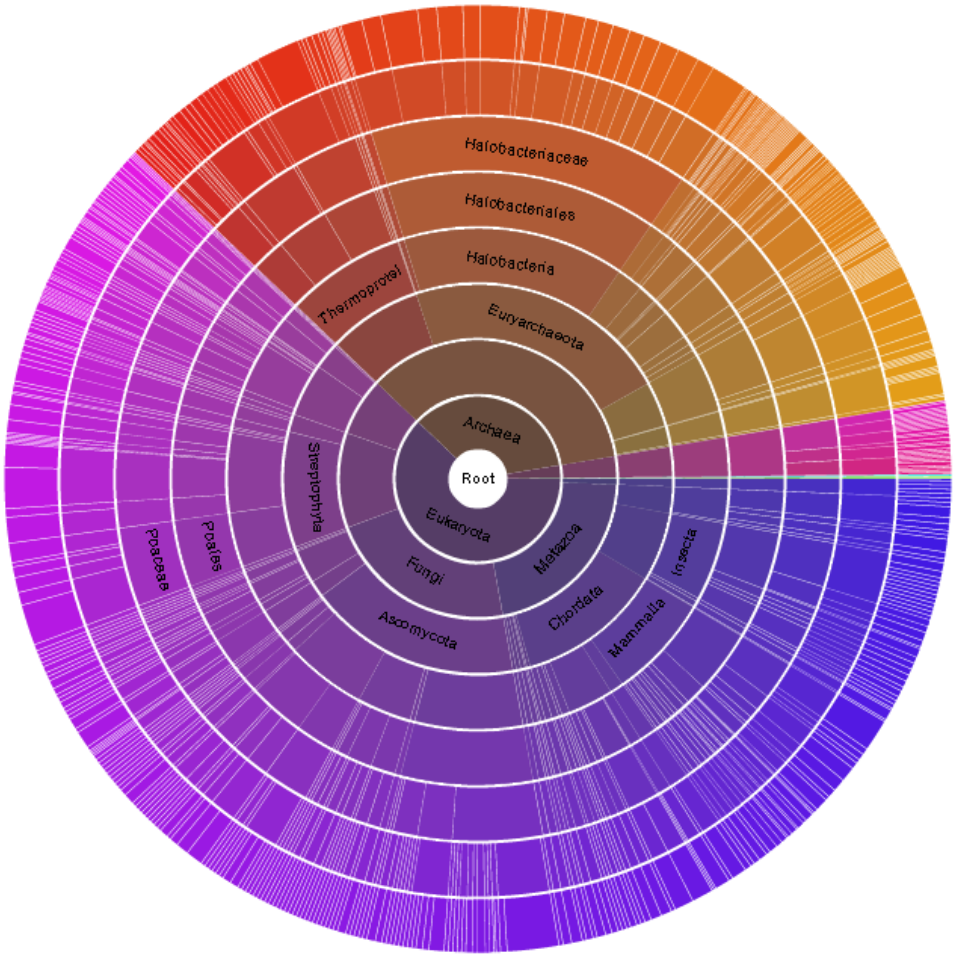


Jump to... enter ID/acc Go

## Species distribution

Sunburst | Tree

This visualisation provides a simple graphical representation of the distribution of this family across species. You can find the original interactive tree in the adjacent tab. **More...**




**Sunburst controls** Hide

**Vanderwaltozyma polyspora**

Root
Eukaryota
Fungi
Ascomycota
Saccharomycetes
Saccharomycetales
Saccharomycetaceae
Vanderwaltozyma
Vanderwaltozyma polys...

**Weight segments by...**

- (•) number of sequences
- ( ) number of species

**Change the size of the sunburst**

Small — Large

**Colour assignments**

- Archea
- Bacteria
- Viruses
- Viroids
- Eukaryota
- Other sequences
- Unclassified
- Unclassified sequence

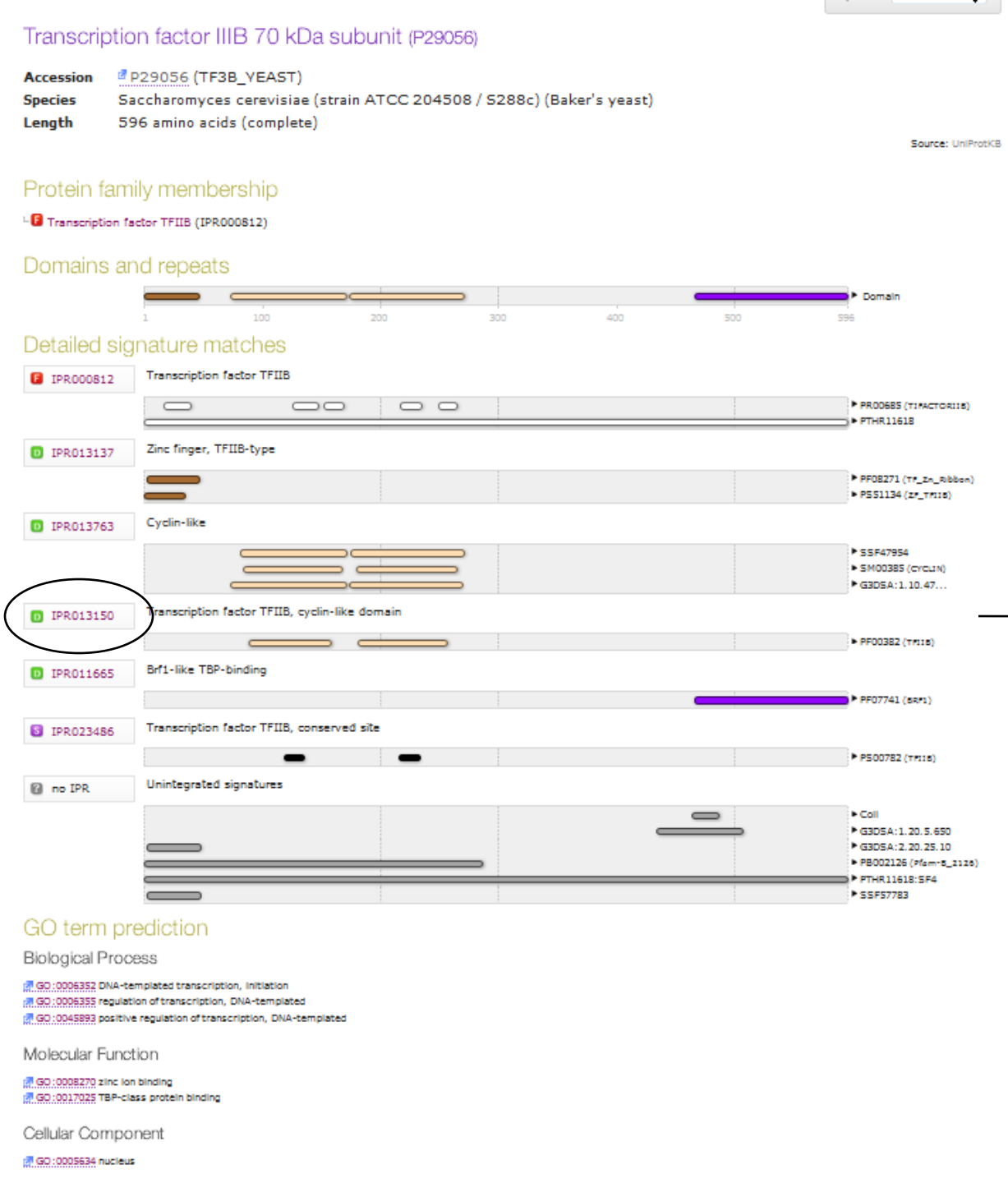
**Selections**

Align selected sequences to HMM
Generate a FASTA-format file
Clear selection

# 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).

# Exemple d'analyse d'une séquence dans InterPro




→ Détail de ce domaine Dans InterPro

# Exemple d'analyse

Transcription factor IIIB 70 kDa subunit (P29056)

Accession P29056 (TF3B_YEAST)
Species
Length

Protein

Transcri

Domain

Detailed

IPR000

IPR013

IPR013

IPR013

IPR011

IPR023

no IPR

GO terms

Biological

GO:000635
GO:000635
GO:004589

Molecular

GO:000827
GO:001702

Cellular C

GO:0005634 nucleus

EMBL-EBI

Services | Research | Training | About us

InterPro
Protein sequence analysis & classification

Search InterPro... Search

Examples: IPR020405, kinase, PS1587, PF02932, GO:0007165

Home | Release notes | Training & tutorials | FAQs | Download | About InterPro | Contact

**Overview**
- Proteins matched (2768)
- Domain organisations (47)
- Pathways & interactions
- Species
- Structures
- Literature (8)
- Cross-references

Add your annotation

## Domain

## Transcription factor TFIIB, cyclin-like domain (IPR013150)

*Short name: TFIIB_cyclin*

### Contributing signatures

Signatures from InterPro member databases are used to construct an entry.

Pfam
PF00382 (TFIIB)

### Domain relationships

Cyclin-like (IPR013763)
**Transcription factor TFIIB, cyclin-like domain (IPR013150)**

### Description

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 [PMID: 1876184, PMID: 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 [PMID: 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 [PMID: 9177165, PMID: 10619841].

### GO terms

#### Biological Process

No terms assigned in this category.

#### Molecular Function

GO:0017025 TBP-class protein binding

#### Cellular Component

No terms assigned in this category.

# Quelques adresses utiles

- **Pôle Rhône-Alpes de Bioinformatique Site Doua http://doua.prabi.fr/**
- **Pasteur http://www.pasteur.fr**
- **Génopole Toulouse : http://bioinfo.genotoul.fr/**
- **Expasy (Suisse) http://www.expasy.org (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-EBI (laboratoire européen, Cambridge) http://www.ebi.ac.uk**