

## Mini projet 3 : Les profils PSSM

Professeur Tom Lenaerts  
Assistante : Catharina Olsen

Information additionnelle sur :  
[http://www.ulb.ac.be/di/map/tlenaert/Home\\_Tom\\_Lenaerts/INFO-F-208.html](http://www.ulb.ac.be/di/map/tlenaert/Home_Tom_Lenaerts/INFO-F-208.html)


Le but de ce projet est de construire un profil pour la famille de séquences WW et de comparer votre résultat avec le profil qu'on peut trouver sur le site web PFAM pour la même famille. Comme avant, n'oubliez pas de bien expliquer comment vous avez trouvé votre solution dans votre document Jupyter.

### Les données

Un ensemble de séquences qui représente la famille est disponible dans la base de données SMART (<http://smart.embl.de>) en utilisant la mode « normal ». Après vous avez choisis le mode normal, vous arrivez à un autre page et, dans la boîte avec la titre « *Domains detected by SMART* », il faut insérer le mot « WW » et cliquer « *Search* ». Vous obtenez maintenant le page suivant :

The screenshot shows the SMART database interface for the WW domain. At the top, there's a navigation bar with links like HOME, SETUP, FAQ, ABOUT, GLOSSARY, WHAT'S NEW, and FEEDBACK. The main header displays 'SMART' in large letters. Below this, the domain 'WW' is highlighted, described as 'Domain with 2 conserved Trp (W) residues'. The page includes a 'SMART MODE:' section with options for Simple, Modular, Architecture, Research, and Tool. A search bar with 'keywords...' and a 'Search SMART' button is also present. The main content area provides detailed information about the WW domain, including its accession number (SM00456), description, and a list of associated proteins. A section titled 'Interpro abstract (IPRO01202)' contains a detailed description of the domain's structure and function. At the bottom, there's a section for 'GO function' and 'Family alignment' options. The page concludes with a statement that there are 18519 WW domains in SMART's nrdb database and a list of links for further information, such as Evolution, Cellular role, Literature, Metabolism, Structure, and Links.

Sur ce page vous voyez toutes les informations pertinentes pour la domaine WW. Vous voyez qu'il y a 18519 domaines du type WW. Si vous cliquez ce 18519, le système cherche pour les protéines liées aux domaines WW. Vous obtenez le page suivant :



**SMART MODE:**  
 NORMAL  
 GENOMIC

Simple  
 Modular  
 Architecture  
 Research  
 Tool

keywords...  
 Search SMART

[HOME](#)
[SETUP](#)
[FAQ](#)
[ABOUT](#)
[GLOSSARY](#)
[WHAT'S NEW](#)
[FEEDBACK](#)

### All proteins containing WW domain

There are 10656 proteins matching your query. Select an option below and press the action button. To include only a subset of proteins, mark the checkboxes next to protein names. Selecting a checkbox next to a taxonomic node will select all proteins in all its subnodes.


**Action** display the domain architecture display SMART bubbligrams of all or selected proteins  
**Selection** 10656 proteins

Display proteins

Expand all nodes search for a taxonomic node...
 

- ☐ Eukaryota (10656)
  - ☐ Fungi (1553)
    - ☐ Ascomycota (1210)
    - ☐ Basidiomycota (299)
    - ☐ Chytridiomycota (6)
    - ☐ Microsporidia (7)
    - ☐ undefined phylum (31)
  - ☐ Metazoa (7677)
    - ☐ Annelida (41)
    - ☐ Arthropoda (925)
    - ☐ Chordata (6220)
    - ☐ Cnidaria (58)
    - ☐ Ctenophora (1)
    - ☐ Echinodermata (41)
    - ☐ Hemichordata (28)
    - ☐ Mollusca (53)
    - ☐ Nematoda (213)
    - ☐ Placozoa (19)
    - ☐ Platyhelminthes (43)
    - ☐ Porifera (35)
  - ☐ Viridiplantae (494)
    - ☐ Chlorophyta (131)
    - ☐ Streptophyta (363)
    - ☐ undefined kingdom (932)
  - ☐ Apicomplexa (100)
    - ☐ Bacillariophyta (58)
    - ☐ Eustigmatophyceae (3)
    - ☐ Phaeophyceae (51)
    - ☐ undefined phylum (720)

On utilisera ce page pour chercher les 136 séquences WW qui sont liées aux domaines WW dans des protéines humaines. Pour qu'on puisse obtenir cette information il faut d'abord suivre dans la hiérarchie des espèces le branchement indiqué dans l'écran en bas. Après il faut choisir dans « Action » l'option « download protein sequences as fasta files ». En plus vous insérez dans « Options -- specific domain only : » le nom de la domaine, c.-à-d. WW.



**SMART MODE:**  
 NORMAL  
 GENOMIC

Simple  
 Modular  
 Architecture  
 Research  
 Tool

keywords...  
 Search SMART

[HOME](#)
[SETUP](#)
[FAQ](#)
[ABOUT](#)
[GLOSSARY](#)
[WHAT'S NEW](#)
[FEEDBACK](#)

### All proteins containing WW domain

There are 10656 proteins matching your query. Select an option below and press the action button. To include only a subset of proteins, mark the checkboxes next to protein names. Selecting a checkbox next to a taxonomic node will select all proteins in all its subnodes.

**Action** download protein sequences as a FASTA file sequences of selected proteins or particular domains will be exported as a plain text, FASTA formatted file.  
**Options** ☐ full protein sequences ☒ specific domain only: WW you can limit the exported sequences only to a particular domain, instead of complete proteins  
**Selection** 136 proteins in 1 selected taxonomic node

Download FASTA

Expand all nodes search for a taxonomic node...
 

- ☒ Eukaryota (10656)
  - ☐ Fungi (1553)
    - ☐ Ascomycota (1210)
    - ☐ Basidiomycota (299)
    - ☐ Chytridiomycota (6)
    - ☐ Microsporidia (7)
    - ☐ undefined phylum (31)
  - ☒ Metazoa (7677)
    - ☐ Annelida (41)
    - ☐ Arthropoda (925)
    - ☒ Chordata (6220)
      - ☐ Actinopterygii (1148)
      - ☐ Amphibia (107)
      - ☐ Appendicularia (20)
      - ☐ Ascidiacea (62)
      - ☐ Aves (487)
      - ☐ Chondrichthyes (3)
      - ☒ Mammalia (3961)
        - ☐ Carnivora (371)
        - ☐ Cetacea (124)
        - ☐ Chiroptera (258)
        - ☐ Cingulata (73)
        - ☐ Dasyuromorphia (61)
        - ☐ Didelphimorphia (71)
        - ☐ Diprotodontia (38)
        - ☐ Hyracoidea (36)
        - ☐ Insectivora (170)
        - ☐ Lagomorpha (136)
        - ☐ Monotremata (42)
        - ☐ Perissodactyla (125)
        - ☐ Pilosa (28)
        - ☒ Primates (946)
          - ☐ Cebidae (139)
          - ☐ Cercopithecidae (253)
          - ☐ Cheirogaleidae (33)
          - ☐ Galagidae (63)
          - ☒ Hominidae (361)
            - ☐ Gorilla (67)
            - ☒ Homo (136)
              - ☒ Homo sapiens (136)
 ENCCGNNNNNN7871.61 run:blastsearch

Après, quand vous avez cliqué « *Download FASTA* », vous obtenez un page avec tous les domaines WW qu'on peut trouver dans des protéines humaines dans le format FASTA. Copiez et collez l'information que vous trouvez sur ce page dans un fichier avec le nom `to-be-aligned.fasta`. Dans l'étape suivant de ce projet, il faut aligner ces séquences.

**IMPORTANT:** Quand vous déposez votre mini projet 3, il est nécessaire que vous déposiez aussi ce fichier.

### L'alignement

Alignez maintenant les séquences au sein du fichier `to-be-aligned.fasta` en utilisant deux des outils suivants :

1. CLUSTAL Omega : <http://www.ebi.ac.uk/Tools/msa/clustalo/>
2. TCOFFEE : <http://www.ebi.ac.uk/Tools/msa/tcoffee/>
3. MUSCLE : <http://www.ebi.ac.uk/Tools/msa/muscle/>

Enregistrez chaque alignement en format FASTA dans un fichier nommé `msaresults-<nom d'outil MSA>.fasta`. Ce fichier contient maintenant l'alignement entre toutes les séquences qui a été produit par un des deux outils que vous avez choisis (CLUSTAL, TCOFFEE ou MUSCLE).

**IMPORTANT:** Quand vous déposez votre mini projet 3, il est nécessaire de déposer aussi les deux fichiers avec les alignements.

### Implémentation Jupyter

Implémentez un logiciel qui construit un profil pour chaque fichier (donc 1 PSSM par outil d'alignement). Regardez les slides 23-34 dans « L7 Alignement de plusieurs séquences ». N'oubliez pas d'utiliser les *pseudo-counts* (expliquez la méthode que vous avez utilisée dans le document Jupyter)

### Validation et présentation

Quand vous avez construit les deux PSSM, il faut les analyser et comparer (ajoutez vous réponses aux questions suivantes dans le document Jupyter. N'hésitez pas d'insérer des images ou illustrations dans votre document).

- 1) Construisez un Weblogo (<http://weblogo.threeplusone.com>) pour la famille WW et l'autre famille que vous avez choisies. Comparez ce Weblogo avec les informations dans vos PSSMs. Quelles sont les positions conservées ?
- 2) Comparez vos résultats avec le HMM-logo que vous trouvez sur le site PFAM (<http://pfam.xfam.org>) pour les deux familles (choisissez « view a PFAM entry »). Quand vous écrivez WW et tapez « go » vous obtenez le

page PF00397 et sur ce page vous pourriez voir le HMM logo (regardez le menu). Quelles sont les différences et similarités avec votre weblogo ?

### **Aligner une séquence au PSSM**

Comme expliqué dans le cours vous pouvez maintenant adapter votre code du premier mini-projet dans un tel façon qu'il pourrait aligner une séquence au PSSM.

- 1) Faites cette adaptation pour votre code qui fait l'alignement local.
- 2) Alignez les deux séquences dans le fichier `test.fasta` à vos deux PSSM et montrez où on peut trouver dans ces deux séquences les domaines WW.
- 3) Vérifier sur UNIPROT ([www.uniprot.org](http://www.uniprot.org)) si vos solutions pour les deux protéines sont correctes. Notez qu'il y a plusieurs WW dans les séquences du fichier `test.fasta`.