# Mini projet 3: Les profils PSSM

Professeur Tom Lenaerts Assistante : Catharina Olsen

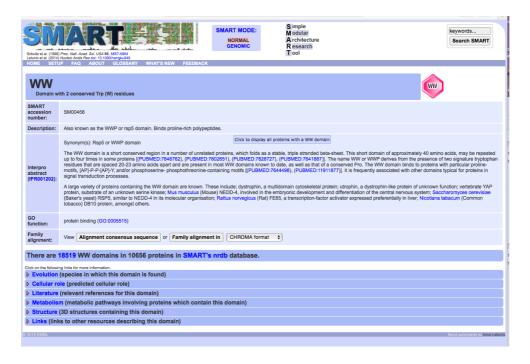
Information additionnelle sur:

http://www.ulb.ac.be/di/map/tlenaert/Home\_Tom\_Lenaerts/INFO-F-208.html

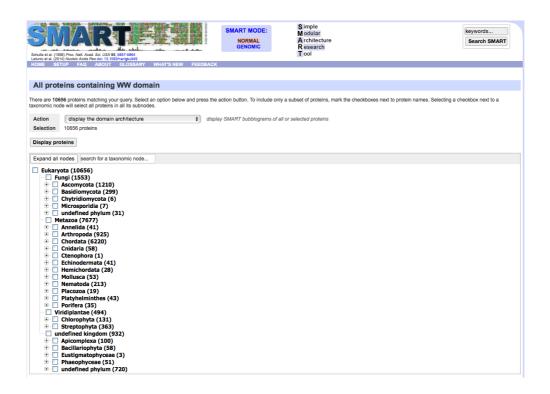
Le but de ce projet est de construire un profil pour la famille de séquences WW et de comparer votre résultat avec le profil qu'on peut trouver sur le site web PFAM pour la même famille. Comme avant, n'oubliez pas de bien expliquez comment vous avez trouvé votre solution dans votre document Jupyter.

#### Les données

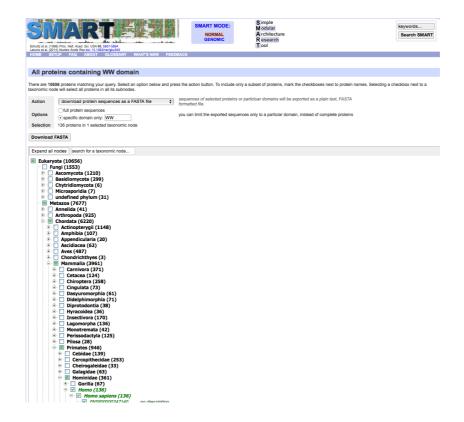
Un ensemble de séquences qui représente la famille est disponible dans la base de données SMART (<a href="http://smart.embl.de">http://smart.embl.de</a>) en utilisant la mode « normal ». Après vous avez choisis le mode normal, vous arrivez à un autre page et, dans la boite avec la titre « <a href="mailto:Domains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



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 :



On utilisera ce page pour cherchez 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.



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 fichierto-be-aligned.fasta en utilisant deux des outils suivants :

- 1. CLUSTAL Omega: <a href="http://www.ebi.ac.uk/Tools/msa/clustalo/">http://www.ebi.ac.uk/Tools/msa/clustalo/</a>
- 2. TCoffee: http://www.ebi.ac.uk/Tools/msa/tcoffee/
- 3. MUSCLE: <a href="http://www.ebi.ac.uk/Tools/msa/muscle/">http://www.ebi.ac.uk/Tools/msa/muscle/</a>

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 utilisez 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 (<a href="http://weblogo.threeplusone.com">http://weblogo.threeplusone.com</a>) 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 (<a href="http://pfam.xfam.org">http://pfam.xfam.org</a>) pour les deux familles (choissisez « 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 expliquez 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 ou on peut trouver dans ces deux séquences les domaines WW
- 3) Vérifier sur UNIPROT (<u>www.uniprot.org</u>) si vos solutions pour les deux protéines sont correctes. Notez qu'il y a plusieurs WW dans les séquences du fichier test.fasta.