

Mini projet 2 : L'alignement et les PSSM

Professeur : Tom Lenaerts (Tom.Lenaerts@ulb.ac.be)

Assistant : Charlotte Nachtegaele (Charlotte.Nachtegaele@ulb.ac.be)

Information liée au cours : <http://www.ulb.ac.be/di/map/tlenaert/>

Date limite : le 10 nov. 2017 à 12h

Dans le premier projet de votre portfolio, vous avez créé un outil bio-informatique qui construit des alignements entre des paires de séquences. Nous avons vu dans la partie théorique du cours que les alignements, construits par cet outil, ne sont pas toujours les meilleurs. Les alignements peuvent être améliorés en utilisant plusieurs séquences, qui peuvent être représentées par des profils, encodés par des *position-specific scoring matrices* (PSSM).

Dans ce nouveau projet, nous allons étendre l'outil d'alignement vers un système qui peut aligner des séquences à des profils. Cette approche est expliquée dans le cours mais vous pouvez trouver des informations additionnelles dans l'article « *RM profiles and alignments.pdf* ». Le nouvel outil permettra à l'utilisateur d'identifier si un domaine particulier, représenté par la PSSM, est présent dans une séquence protéique donnée. Pour cette partie, nous utiliserons le domaine WW comme exemple (Fig. 1).

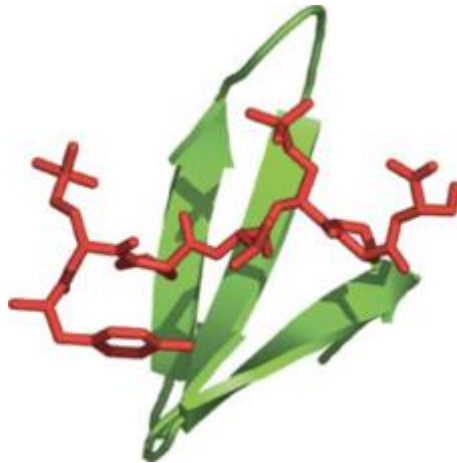


Figure 1 : Un domaine WW typique interagissant avec un peptide (en rouge). Pour plus de détails voir l'article « *WW and SH3 domains: two different scaffolds to recognize proline-rich ligands* » (2002) par Macias, Wiener et Sudol.

Exigences

1. Le Jupyter notebook que vous construisez est un rapport, ce qui signifie que vous devriez le structurer comme un rapport, même si le code est directement disponible.

2. Un rapport se compose d'une introduction du problème, d'une explication des méthodes (et leurs implémentations), d'une discussion sur les résultats et enfin d'une conclusion sur les résultats que vous avez obtenus.
3. Toutes les questions posées dans ce document doivent être clairement répondues et les résultats doivent être présentés afin qu'ils puissent être reproduits dans le Jupyter notebook (pas d'exécution dans un terminal)
4. Des captures d'écran de la sortie du terminal sont pas acceptable et vous ne pouvez pas faire du *copy-paste* des diapos du cours.
5. **Les explications dehors du code ne sont pas une documentation du code mais une description explicative d'algorithme : qu'est-ce que la fonction ou l'ensemble de fonctions fait ? Telles explications contiennent des exemples qui illustrent vos propos.**
6. **Un rapport est un document formel. On utilise donc la première personne du pluriel, pas la première personne du singulier.**

Évaluation

L'évaluation sera basée sur les critères suivants :

1. La compréhension générale des instructions et exigences,
2. L'utilisation correcte du langage de programmation,
3. La structure du rapport et l'organisation des blocs de code dans le *Jupyter notebook*,
4. L'efficacité et l'exactitude de l'algorithme mis en œuvre,
5. La clarté et la pertinence des commentaires par bloc de code et en général,
6. La clarté des exemples utilisés pour l'illustration du fonctionnement de votre code,
7. La clarté de la comparaison faite avec d'autres outils,
8. Les illustrations graphiques.

Partie 1, Collecte des données



SMART

SMART MODE: **NORMAL** **GENOMIC**

Simple Multiple Alignment Search Tool

WW

Domain with 2 conserved Trp (W) residues

SMART accession number: **SM00336**

Description: Also known as the WWP or wip domain. Binds proline-rich polypeptides.

Synonym(s): Rpf1 or WWP domain

The WW domain is a short conserved region in a number of unrelated proteins, which folds as a stable, type II beta-sheet. This short domain of approximately 40 amino acids, may be repeated up to four times in some proteins (PUBMED:784762, PUBMED:782551, PUBMED:782871, PUBMED:7841887). The name Wip or WWP derives from the presence of two signature tryptophan residues that are spaced 20-23 amino acids apart and are present in most WW domains known to date, as well as that of a conserved Pro. The WW domain binds to proteins with particular proline motifs, (X)(Y)(X)(X)(Y)(X), and/or phosphoserine- phosphothreonine-containing motifs (PUBMED:784468, PUBMED:781817). It is frequently associated with other domains typical for proteins in signal transduction processes.

A large variety of proteins containing the WW domain are known. These include: dystrophin, a multidomain cytoskeletal protein; atrophin, a dystrophin-like protein of unknown function; vesicle WIP protein, subunit of an unknown protein complex; Mus musculus (mouse) WIP-4, involved in the embryonic development and differentiation of the central nervous system; *Search for WW domain* (Rafin's yeast) RFPs, similar to WIP-4 in its molecular organization; *Rafin* homologous (Raf F88), a transcription factor activator expressed preferentially in liver; *Neutrophin* (Neutrophin) (GPI-3) protein, amongst others.

GO function: protein binding (GO:0005616)

Family alignment: View **Alignment consensus sequence** or **Family alignment in** CHROMA format

There are **18519** WW domains in 10556 proteins in SMART's nrdb database.

Find in the following tree for more information:

- Evolution (species in which this domain is found)
- Cellular role (predicted cellular role)
- Literature (relevant references for this domain)
- Metabolism (metabolic pathways involving proteins which contain this domain)
- Structure (3D structures containing this domain)
- Links (links to other resources describing this domain)

Figure 2 : Information liée aux domaines WW sur le site SMART.

Un ensemble de séquences qui représentent la famille WW est disponible dans la base de données SMART² qui doit être utilisée en mode « *normal* » (voir la page d'accueil du site web). Après avoir choisi le mode normal, vous arrivez à une autre page qui est composée de 4 parties. Dans la boîte avec le titre « Domains detected by SMART », il faut insérer le mot « WW » et cliquer sur « Search ».

Vous obtenez maintenant la page de la Figure 2. Sur cette page, vous pouvez voir toutes les informations pertinentes pour le domaine WW. Vous pouvez constater qu'il y a 18519 domaines du type WW. Si vous cliquez ce 18519, le système cherche pour les protéines possédant des domaines WW. Vous obtenez la page de la Figure 3.



Figure 3 : SMART page de sélection de protéines

On utilisera cette page pour chercher les 136 séquences WW qui sont liées aux domaines WW des protéines humaines. Pour obtenir cette information, il faut d'abord suivre dans la hiérarchie des espèces le branchement indiqué dans la Figure 3. La Figure 4 indique où trouver l'espèce humaine exactement dans cette hiérarchie. En cliquant sur les symboles « + », vous pouvez descendre dans l'arbre au niveau correct. Vous verrez le numéro 136 à côté de l'espèce « homo sapiens », indiquant le nombre de séquences WW trouvées dans cette espèce.

Une fois que vous avez coché la case avant « homo sapiens », vous devez retourner au début de la page et sélectionner dans la boîte avec le titre « *Action* » l'option « *download protein sequences as fasta files* ». En plus, vous devez ajouter dans « *Options -- specific domain only :* » le nom du domaine, c.-à-d. WW.

² <http://smart.embl.de>



Figure 4 : Où trouver l'espèce humaine dans l'arbre des espèces.

Après avoir cliqué sur « *Download FASTA* », vous obtenez une page avec tous les 136 domaines WW qu'on peut trouver dans des protéines humaines en format FASTA. Copiez et collez l'information que vous trouvez sur ce page dans un fichier avec le nom « *t-o-b-e-a-l-i-g-n-e-d-f-a-s-t-a* ». Dans l'étape suivant de ce projet, il faut aligner ces séquences.

IMPORTANT : Quand vous déposez votre mini projet 2, il est nécessaire que vous déposiez aussi ce fichier.

Partie 2, L'alignement de plusieurs séquences

Alignez maintenant les séquences au sein du fichier *t-o-b-e-a-l-i-g-n-e-d-f-a-s-t-a* en utilisant un des outils suivants. Mentionnez clairement dans votre Jupyter notebook quel outil vous avez utilisé.

1. CLUSTAL Omega³
2. TCOFFEE⁴
3. MUSCLE⁵

Enregistrez votre alignement en format FASTA dans un fichier nommé *msaresults- <nom d'outil> MS A>.fasta*.

³ <http://www.ebi.ac.uk/Tools/msa/clustalo/>

⁴ <http://www.ebi.ac.uk/Tools/msa/tcofee/>

⁵ <http://www.ebi.ac.uk/Tools/msa/muscle/>

IMPORTANT : Quand vous déposez votre mini projet 2, il est nécessaire de déposer aussi le fichier avec le MSA.

Partie 3, Construction du profil

Implémentez un logiciel qui construit un profil en utilisant l'alignement que vous avez construit. Regardez les diapos et l'article « *RM profiles and alignments.pdf* » pour les détails. N'oubliez pas d'utiliser les *pseudo-counts*. Expliquez la méthode que vous avez utilisée pour la construction du PSSM dans le document Jupyter.

Quand vous avez construit le PSSM, vous devriez valider vos résultats avec ce qu'on sait des domaines WW. Répondez aux questions suivantes dans le document Jupyter. N'hésitez pas à insérer des images ou illustrations.

- 1) Construisez un Weblogo⁶ pour la famille WW et comparez-le avec les informations dans votre PSSM. Quelles sont les positions conservées et est-ce qu'elles correspondent à l'information au sein du Weblogo?
- 2) Comparez vos résultats avec le HMM-logo que vous trouvez sur le site PFAM⁷ pour le domaine WW. Quand vous écrivez WW dans la boîte « *view a PFAM entry* » et tapez « go », vous obtenez la page PF00397. Sur cette page, vous pourrez voir le HMM logo. Quelles sont les différences et similarités avec votre Weblogo et votre PSSM ?

Il devrait être clair maintenant d'où vient le nom « WW domain ». Veuillez l'expliquer brièvement dans votre rapport.

Partie 4, l'alignement du profil aux séquences

Comme expliqué dans le cours vous pourriez maintenant adapter votre code du premier mini-projet de telle façon que vous pourriez aligner une séquence au PSSM.

- 1) Faites cette adaptation pour votre alignement local avec la pénalité linéaire. Regardez aussi le document « *RM profiles and alignments.pdf* ».
- 2) Dans le document « *RM profiles and alignments.pdf* » il est aussi expliqué comment le faire pour la pénalité affine. Pour **des points supplémentaires**, vous pouvez également fournir cette extension. N'oubliez pas de souligner clairement comment vous avez implémenté cette extension.
- 3) Alignez les séquences dans le fichier `protein-sequences.fasta` à votre PSSM. Montrez où on peut trouver dans ces deux séquences les domaines WW.

⁶ <http://weblogo.threeplusone.com>

⁷ <http://pfam.xfam.org>

- 4) Vérifiez sur UNIPROT⁸ si vos solutions pour les deux protéines sont correctes. Trouvez-vous par exemple les mêmes positions de départ et de fin par exemple ? Trouvez-vous tous les domaines ? Expliquez et illustrez vos résultats.

Éthique

Le plagiat sera sévèrement sanctionné. Les cas de plagiat comprennent la réutilisation du matériel écrit ou tiré de quelqu'un d'autre¹⁰, ou tout type de travail, sans devis ou référence explicite.

⁸ www.uniprot.org

¹⁰ <http://www.bib.ulb.ac.be/fr/aide/eviter-le-plagiat/> et <http://www.plagiarism.org/>