Mini projet 2 : L'alignement et les PSSM

Professeur : Tom Lenaerts (Tom.Lenaerts@ulb.ac.be)
Assistant : Charlotte Nachtegael (Charlotte.Nachtegael@ulb.ac.be)

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

Date limite : le 22 nov. 2019 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 aussi le domaine WW comme exemple (Fig. 1).

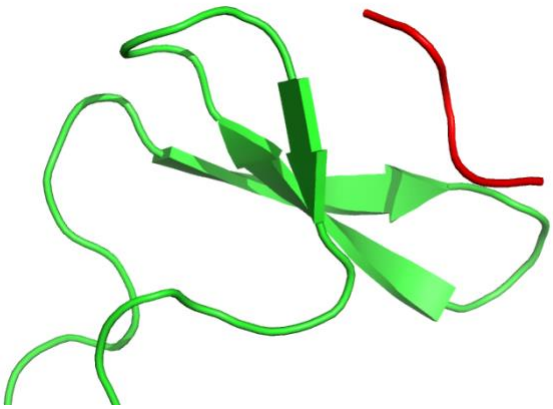

Figure 1 : Un domaine WW typique interagissant avec un peptide (PDB ID : 1K9R). Pour plus de détails sur cette structure regardez http://www.rcsb.org/structure/1K9R.

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 ne sont pas acceptables et vous ne pouvez pas faire du *copy-paste* des diapos du cours.
5. **Les explications en 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

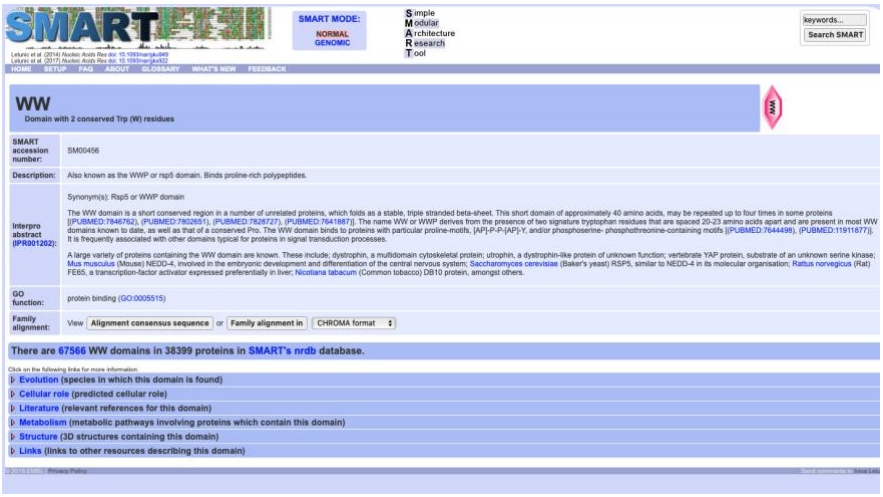

Figure 2 : Information liée aux domaines WW sur le site SMART.

Un ensemble de séquences qui représentent la famille WW est disponible dans la base de données SMART[1] qui doit être utilisée en mode « *normal* » (voir la page d'accueil du site web).

[1] http://smart.embl.de

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 pour le domaine WW, visualisée dans la Figure 2. Sur cette page, vous pouvez voir toutes les informations pertinentes pour le domaine WW. Vous pouvez constater qu'il y a 67566 domaines du type WW. Si vous cliquez ce 67566, 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 177 séquences 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 177 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.

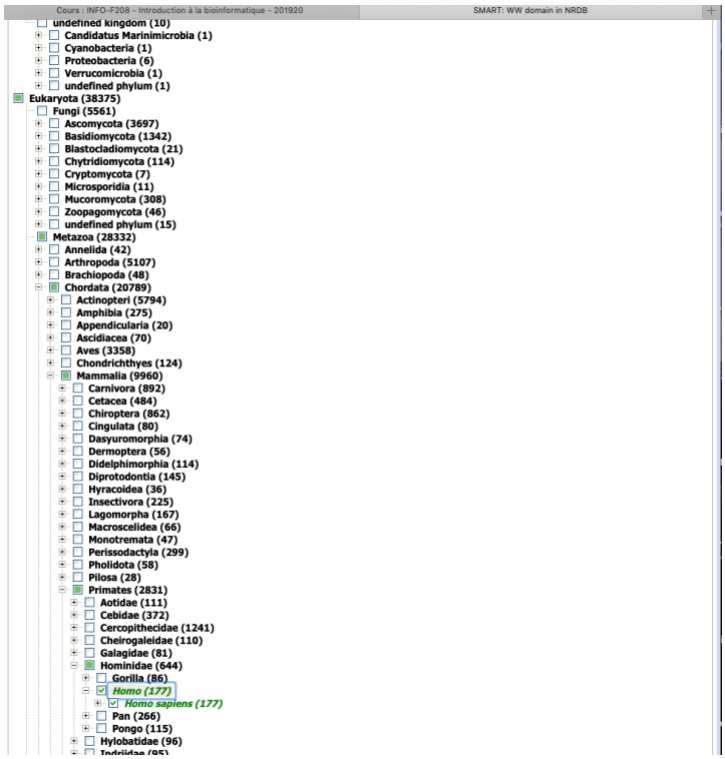

Figure 4 : Où trouver l'espèce humaine dans l'arbre des espèces.

Après avoir cliqué sur « *Download FASTA* », vous obtenez une page avec tous les domaines WW qu'on peut trouver dans des protéines humaines en format FASTA. Copiez et collez l'information que vous trouvez sur cette page dans un fichier avec le nom « `to-be-aligned.fasta` ».

Avant de faire l'alignement vous devez d'abord créer un deuxième fichier. Il est possible qu'il y ait des séquences trop similaires entre elles dans ces 177 domaines. Créez un ficher « `to-be-aligned-reduced.fasta` » dans lequel vous gardez tous les séquences qui ont un maximum de 60% de similarité entre eux. Expliquez bien dans votre notebook comment vous avez résolu ce problème.

IMPORTANT : Quand vous déposez votre mini projet 2, il est nécessaire que vous déposiez aussi ces deux fichiers.

Partie 2, L'alignement de plusieurs séquences

Alignez maintenant les séquences au sein du fichier `to-be-aligned.fasta` et `to-be-aligned-reduced.fasta` en utilisant un des outils suivants. Mentionnez clairement dans votre Jupyter notebook quel outil vous avez utilisé.

1. CLUSTAL Omega[2]
2. TCoffee[3]
3. MUSCLE[4]

Enregistrez le premier alignement en format FASTA dans un fichier nommé `msaresults-<nom d'outil MSA>.fasta`. Le deuxième, dans le fichier `msaresults-reduced-<nom d'outil MSA>.fasta`

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

Partie 3, Construction du profil

Implémentez un logiciel qui construit deux profils en utilisant les deux alignements que vous avez construits. 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 des PSSM dans le document Jupyter.

Quand vous avez construit les 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 Weblogo[5] pour les deux MSA 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 des deux Weblogos?
2) Comparez vos résultats avec le HMM-logo que vous trouvez sur le site PFAM[6] pour le domaine WW. Quand vous écrivez « WW » dans la boite « *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 vos Weblogos et vos PSSM ?

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] http://www.ebi.ac.uk/Tools/msa/clustalo/
[3] http://www.ebi.ac.uk/Tools/msa/tcoffee/
[4] http://www.ebi.ac.uk/Tools/msa/muscle/
[5] http://weblogo.threeplusone.com
[6] http://pfam.xfam.org

2) Dans le document « *RM profiles and alignments.pdf* » il est aussi expliqué comment 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` aux deux PSSM. Montrez où on peut trouver dans ces deux séquences les domaines WW. Est-ce qu'il y a des différences entre les résultats pour les deux PSSM ?
4) Vérifiez sur UNIPROT[7] si vos solutions pour les trois protéines sont correctes. Trouvez-vous par exemple les mêmes positions de départ et de fin pour les domaines ? 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'autre[8], ou tout type de travail, sans devis ou référence explicite.

[7] www.uniprot.org
[8] http://www.bib.ulb.ac.be/fr/aide/eviter-le-plagiat/ et http://www.plagiarism.org/