Projet 2

08/11/19

À rendre pour le 22/11/19

Projet 2

- Récupérer les séquences des domaines WW des protéines humaines + créer un dataset avec maximum 60% de similarité entre ells
  - Vous devez simplement décrire la procedure dans le notebook
  - Vous pouvez indiquer le code
- Aligner ces séquences grâce à un des outils proposés (Clustal Omega, MUSCLE, TCoffee)
- Construire une matrice de profil pour le domaine WW
- Chercher les domaines dans les séquences en alignant les séquences au profil

To-be-aligned reduce

Commencer avec une sequence dans un premier groupe, puis :

- Aligner (en global) la sequence suivante avec les sequences du groupe 1
  - Si toutes les sequences ont <60% de similarité, ajouter la sequence au groupe
  - Sinon, passez au groupe suivant
- Si aucun groupe n’a pu accepter la sequence, créer un nouveau groupe

Quand vous avez fini d’aligner toutes les sequences, prenez le groupe le plus grand

Construire une matrice de profil

Plusieurs formules possibles !

```
P06241|149-246   WYFGKLGR---KDAERQLLSFGNPRGTFLIRESETTK-GAYSLSIRDWDDMKGDHV--KH
Q06124|6-102     WFHPNITG---VEAENLLLTRG-VDGSFLARPSKSNP-GDFTLSVRR-----NGAV--TH
P62993|60-152    WFFGKIPR---AKAEEMLSKQ-RHDGAFLIRESESAP-GDFSLSVKF-----GNDV--QH
P12931|151-248   WYFGKITR---RESERLLLNAENPRGTFLVRESETTK-GAYCLSVSDFDNAKGLNV--KH
P41240|82-171    WFHGKITR---EQAERLL-YPP-ETGLFLVRESTNYP-GDYTLCVS-C----DGKV--EH
P00519|127-217   WYHGPVSR---NAAEYLL-SSG-INGSFLVRESESSP-GQRSISLRY-----EGRV--YH
P20936|181-272   WYHGKLDR---TIAEERLRQAG-KSGSYLIRESDRRP-GSFVLSFLSQ----MNVV--NH
P42224|573-670   WNDGCIMGFISKERERALLKDQ-QPGTFLLRFSESSREGAITFTWVERSQNG-GE--P--
O60674|401-482   --HGPISM---DFAISKLKKAGNQTGLYVLRCSPKDF-NKYFLTFAVER---ENVIEYKH
```

$$m_{u,a}=\log \frac{q_{u,a}}{p_a} \qquad q_{u,a}=\frac{n_{u,a}+\beta p_a}{N_{seq}+\beta} \qquad q_{u,a}=\frac{\alpha f_{u,a}+\beta p_a}{\alpha+\beta} \qquad q_{u,a}=\frac{\alpha f_{u,a}+\beta g_{u,a}}{\alpha+\beta}$$

$$m_{u,a}=\sum_{b\in\{AA\}} f_{u,b}\, s_{a,b}$$

Ala (A) 8.28	Gln (Q) 3.94	Leu (L) 9.67	Ser (S) 6.50
Arg (R) 5.53	Glu (E) 6.76	Lys (K) 5.85	Thr (T) 5.32
Asn (N) 4.05	Gly (G) 7.09	Met (M) 2.43	Trp (W) 1.07
Asp (D) 5.45	His (H) 2.27	Phe (F) 3.86	Tyr (Y) 2.91
Cys (C) 1.36	Ile (I) 5.99	Pro (P) 4.68	Val (V) 6.87

$$f_{u,b}=\frac{n_{u,b}}{N_{seq}} \qquad g_{u,a}=\sum_b f_{u,b}\frac{q_{a,b}}{p_b}$$

Aligner une séquence à un profil

$$S(i,j) = max \begin{cases} S(i-1,j-1)+PSSM(seq(i),j) \\ S(i-1,j)+PSSM(\text{“-“},j) \\ S(i,j-1)+PSSM(\text{“-”},j-1) \\ 0 \end{cases}$$

- Alignement local à pénalité linéaire
- Utiliser la PSSM au lieu des matrices de substitution
- Plusieurs sous-alignements possibles

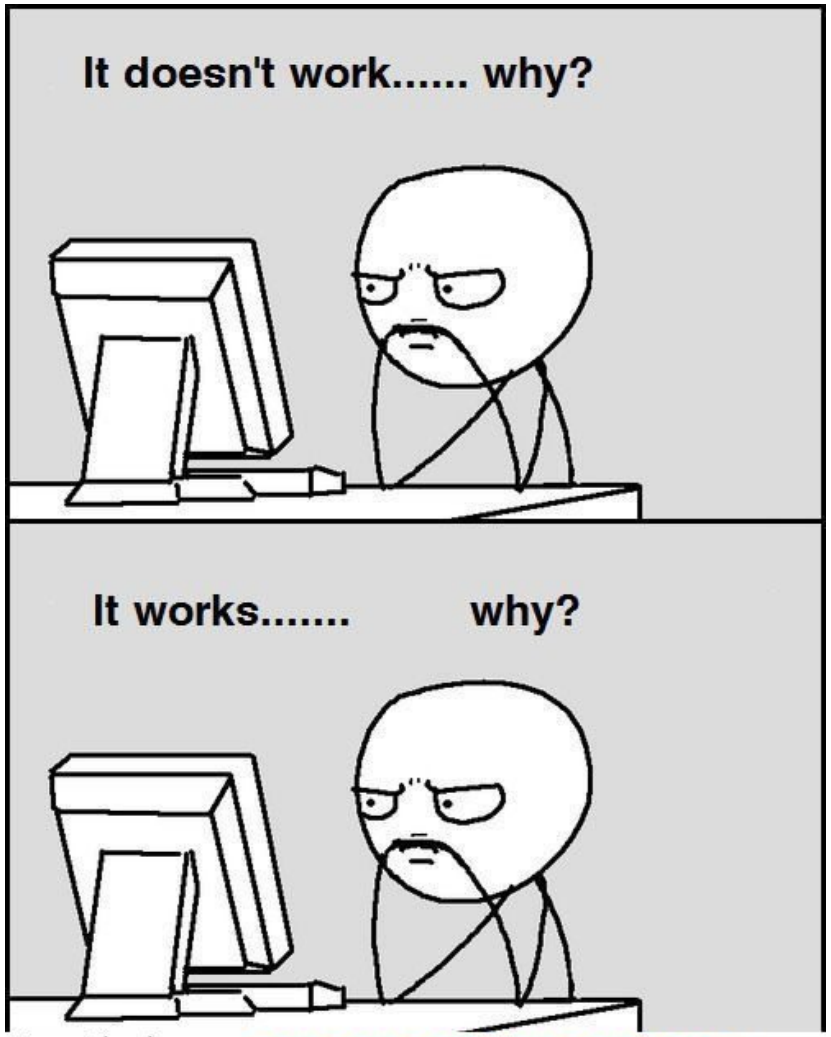
Instructions supplémentaires

- Fournir les fichiers nécessaires avec le jupyter notebook dans un fichier zip
  - 2 fichiers de sequences (complet et réduit)
  - Au minimum 2 fichiers d'alignements
  - Weblogo (2, 1 pour chaque alignement multiple)
  - HMM logo

- Justifiez vos choix de formules, implementations, etc.

- Affichez bien les PSSM, weblogos et HMM ! N'oubliez pas non plus de bien montrer vos résultats des alignements de sequences avec les PSSMs

- Ne pas hésitez à mettre des images pour illustrer vos explications ! (exemple de calcul, etc.)



Happy programming !

Séances de TP :

- 15/11/19
- 22/11/19 (Dernière chance !)