TD L3 ASG 2019

Séances 3, 4, 5, 6, 7

Récupération et alignement de séquences

1. A l'aide des numéros d'accession de la Table 1, de l'article Suzuki 1997, récupérer séquences de GP Ebola sur le site de Genbank, au format fasta. Ne prenez pas les séquences de Marburg (trop éloignées). Extrayez spécifiquement les séquences ADN de GP à partir du format gb : « send to » + « coding sequence », puis choisissez le gène de la protéine GP. Créez un fichier multi fasta.

2. A l'aide du site Phylogeny.fr, réalisez un alignement multiple de ces séquences et un arbre phylogénétique, avec les arguments par défaut.

3. Inspectez l'alignement multiple et l'arbre. Retrouvez-vous la topologie de Suzuki 1997 ? Sauvegardez l'alignement au format Fasta.

4. Rechercher sur le site du NCBI une séquence Ebola « Makona variant ». Il s'agit du variant de l'épidémie du Sierra Leone de 2014-15. Extrayez la séquence GP, ajoutez là au fichier fasta et refaites un alignement multiple et arbre. La structure de l'arbre est-elle toujours cohérente avec l'article Suzuki 1997 ? Sauvegardez l'alignement multiple en format fasta.

Lecture de séquences dans R

5. Dans R, créez un nouveau script. Insérez la fonction R fournie `readfasta.r`.

6. A l'aide de cette fonction, lisez dans R les fichiers :
   - Fasta brut
   - Fasta aligné

7. Affichez et comptez les séquences pour vérifier le bon fonctionnement du programme.

Traduction ADN>protéine

8. Chargez dans votre programme le code R fourni dans `genet.code.r`.
   Voyez comment fonctionne le tableau aa. Testez-le dans la console pour traduire un codon quelconque.

9. Créez une fonction pour traduire un vecteur de nucléotide en vecteur d'aminoacides, en utilisant le tableau aa. Testez cette fonction avec une séquence aléatoire. Fonctions utiles:

`aa[codon], aa['ATG']` >dans le tableau aa, la valeur de la case d'index codon ou d'index 'ATG'
`codon=substr(seq, i, i+2)` >extrait la sous-chaîne de I à i+2 de la chaîne seq et l'affecte à la variable 'codon'
`seq <- paste0 (seq, 'xyz')` >ajoute 'xyz ' à la chaine 'seq'

10. Ecrire les séquences protéiques dans un fichier au format Fasta. Fonctions utiles :

`write (x, file="test.aa",append=T)` >écrit dans un fichier en mode « ajout »

`sink(file="test.aa")` …. `sink()` >alternative à print. Ouvre un fichier pour écriture, puis le referme

`cat (x, « \n »)` >écrit la variable x suivi de retour

11. Réalisez l'alignement multiple par Muscle sur le site Phylogeny.fr. Visualisez l'alignement au format Clustalw. Retrouvez-vous les sites de fixation de Glycanes (Figure 1, article Lennemann 2014) dans l'alignement? Sauvez l'alignement protéique au format fasta.

Analyse d'alignement, entropie

12. Lisez les deux alignements (ADN et protéique) sous R. Réalisez un programme pour convertir ces alignements en un tableau de caractères. On utilisera la fonction suivante pour convertir une chaine de caractères « l » en vecteur de caractères « v »:

```
v <- unlist (strsplit(l,""))
```

13. Affichez le nombre de lignes et de colonnes du tableau contenant l'alignement. Affichez une seule colonne.

14. Nous commencerons par analyser les séquences ADN : créez une fonction calculant et affichant le nombre de nucléotides "A", "T", "G", "C" dans un vecteur de nucleotides, dont les indices portent la lettre du nucléotide (par exemple vec["A"]=3). Testez la fonction sur un vecteur de votre choix.

15. Sur cette base, créez une fonction calculant les fréquences des bases du vecteur. *Attention* : vous avez des séquences alignées contenant des indels. Que se passe-t-il si une colonne contient des indels? Doivent-ils être comptés dans le calcul de fréquence ?

16. Sur cette base, créez une fonction calculant l'entropie du vecteur. Testez-là sur un vecteur de votre choix et sur une colonne de l'alignement. (sur une colonne conservée à 100%, l'entropie devrait être être nulle).

17. A l'aide de la fonction ci-dessus, calculez les entropies de toutes les colonnes de l'alignement et stockez les résultats dans un vecteur H. Tracez H.

18. Lissez la courbe H : à l'aide de la fonction smooth/spline qui lisse une courbe définie par deux vecteurs x (absisses) et y (ordonnées) calculez des points lissées.

    ```
    lisse = smooth.spline(x, y, spar=0.35)
    ```

    Il faudra ensuite tracer la courbe H sans les points :

    ```
    plot (H,type="n")
    ```

    Puis utiliser la fonction « line » pour tracer la courbe « lisse »

19. Identifiez les positions d'entropie la plus élevée/basse.

20. Adaptez vos fonctions aux séquences protéiques et refaites l'analyse d'entropie pour les protéines.

21. Sur l'alignement de séquences protéiques : affichez les positions d'entropie la plus élevée/basse, les 10% des positions d'entropie la plus élevée/basse. Liez cette information à la structure des GP1 et GP2.

22. Pour la suite, nous aurons besoin d'un fichier à 3 colonnes comprenant : position dans l'alignement, aminoacide, entropie, selon le modèle suivant. Produisez ce fichier.

```
1  M     1.52
2  L     0.75
3  C     0.70
…
```

Analyse structurale

23. Récupérez sur le site de la pdb, la structure 5FHC (complexe GP1, GP2, 2 anticorps). A quelle molécule correspond chaque chaîne A, B, C…?

24. Repérez dans le fichier les chaînes peptidiques, les coordonnées atomiques, les B-factors (l'atténuation liée a la diffraction des rayons X provoquée par l'agitation thermique).

25. Ouvrez la structure pdb avec Pymol. Fonctions Pymol à connaître :
    a. Les objets « all » et les chaînes. Les sélectionner. Afficher une seule chaine. Zoomer.
    b. Les boutons ASHLC
    c. La commande split_chains
    d. Les couleurs. Colorer avec option spectrum + Bfactor
    e. Le bouton « S » (séquence) + sélection de résidu
    f. Affichage cartoon et surface
    g. Colorez les chaines de GP et des anticorps de couleur différentes.
    h. A quel résidu commencent les structures GP1/2 de 5FHC ?
    i. Sauvegardez la session

26. Récupérez les séquences peptidiques correspondant GP1 et GP2 de 5FHC sur le site de la pdb, et ajoutez la concaténation de ces séquences à votre fichier de protéines Fasta. Alignez le tout avec Muscle pour obtenir l'alignement de vos protéines avec la protéine GP de 5FHC. Ceci nous permettra d'établir la correspondance entre la numérotation de la structure PDB et la numérotation de notre alignement. (attention : vérifier que l'alignement n'a pas changé par rapport au précédent).