M1 - BIOINFORMATIQUE DE BASE - ANALYSE DE SÉQUENCES
Eric CORTIAL - Génie Biologique / EIDD

I) Introduction

a) présentation des analyses et leur but

Notre objectif est d'identifier les oligomères TATA et CGGC dans les séquences suivantes :

- YAP_up800 : correspondant aux régions promotrices des gènes dont l'expression est induite par la protéine Yap1P.
- MET_up800 : correspondant aux régions promotrices des gènes réprimés par la méthionine.

Nous allons ensuite, pour confronter nos résultats, observer s'ils sont sous ou sur-représentés dans ces séquences.

Ensuite, nous allons analyser les séquences avec les motifs YAP1 et SPT15 issue de la base de données JASPAR et vérifier si ils sont aussi sur ou sous-représentés dans ces séquences

b) description des méthodes et données

Pour la suite, voici notre protocole :

- Dans un premier temps, après avoir téléchargé nos motifs/séquences, nous y compterons le nombre d'occurrences d'oligomères sur Python. Puis, nous générons des séquences aléatoires en utilisant le modèle shuffling, pour savoir comment nos oligomères sont représentés dans des séquences aléatoires, et grâce à une analyse sur R déterminer si les oligomères sont sur ou sous représentées dans nos séquences en calculant le z-score et la p-values
- Dans un second temps, après avoir téléchargé les motifs à partir de la base de donnée Jaspar, nous allons générer et afficher un logo représentant les motifs. Puis, nous allons générer une PSSM et une PWM pour chaque motif (python). Enfin, on va scanner les séquences avec des matrices et déterminer quels sont les hits significatifs et voir si ils sont sur ou sous représentés dans les séquences.

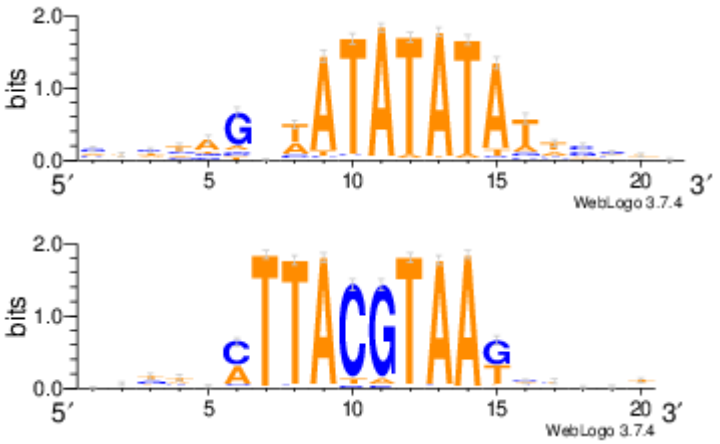
II) Resultats et Analyses

a) description des résultats

Après réalisation sur python, on observe que :

- TATA est présent 78 fois et CGGC 13 fois dans la séquence MET_up800
- TATA est présent 86 fois et CGGC 23 fois dans la séquence YAP_up800
- Grâce au shuffling, on observe sur 1000 séquences aléatoires que:
    - TATA est représentée à peu près également dans des séquences aléatoires, cad entre 70 et 80 fois : ce qui signifie que TATA dans MET est ni sur, ni sous représenté.
    - CGGC est lui aussi ni sur ni sous représentée dans MET puisqu'il y apparait entre 10 et 20 fois à peu près aussi.
    - TATA est représenté au moins entre 100 fois minimum et 120-130 fois maximum dans les séquences aléatoires : ce qui signifie que TATA est sous représentée dans YAP
    - CGGC est quand à lui légèrement sur représenté dans YAP puisque la moyenne d'occurrences dans les seq aléatoires est d'environ 14-15 fois.
- Concernant les motifs YAP1 et SPT15, voici ce que nous obtenons :

-



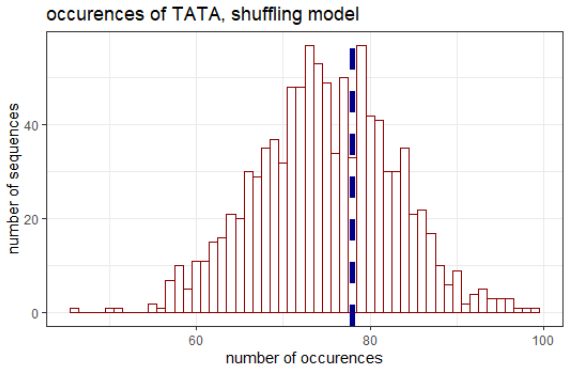
Pour les motifs YAP1 et SPT15 les sites du milieu semblent être très importants, et les sites en 3' et 5' sont peu informatifs.

b) interprétation des résultats

On s'intéresse donc à valider nos résultats précédents : grâce au logiciel R et à nos données précédentes, nous représentons dans un diagramme en bâton le nombre d'occurrences de TATA / CGGC en fonction du nombre aléatoire de séquences.
Pour la suite nous indiquons que :
Le **z-score** correspond au nombre d'**écart type** par rapport à la moyenne d'une valeur.
La **p-value** correspond à la probabilité que le modèle observé a été créé par un processus aléatoire. Plus elle est basse, plus cela signifie qu'il est très improbable que le modèle observé soit le résultat d'un processus aléatoire.

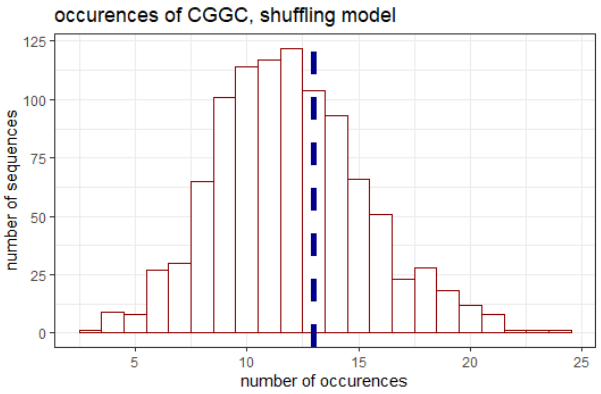
Voici ce que nous obtenons :

- TATA pour MET :



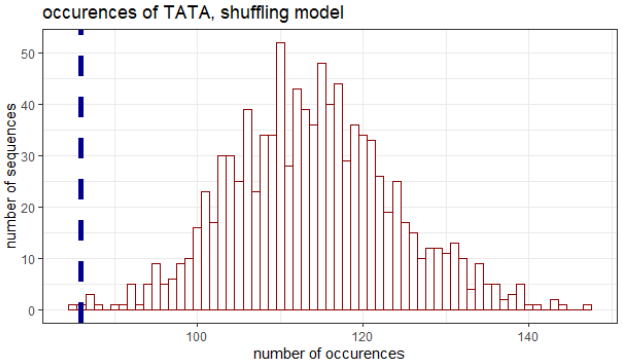
interprétation : TATA est légèrement surreprésenté dans la séquence MET
Z-score : 0.3942421
p-value : 0.6532989

- CGGC pour MET



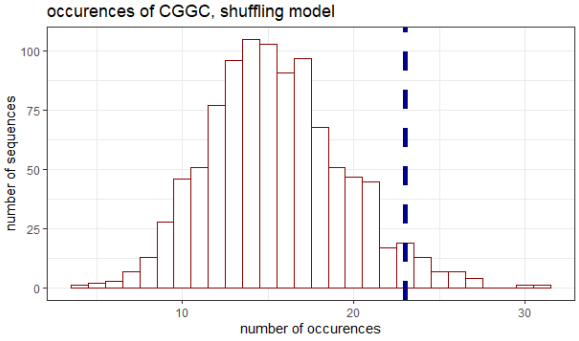
interprétation : CGGC est ni sur, ni sous représenté dans la séquence
Z-score : 0.3120072
p-value : 0.6224825

- TATA pour YAP



interprétation : TATA est très sous-représentée dans YAP
Z-score : -2.815906
p-value : 0.002431993

- CGGC pour YAP

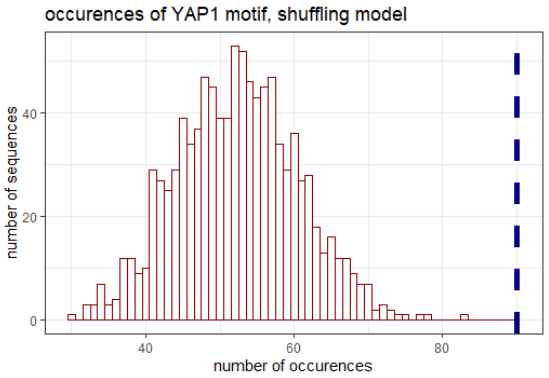


interprétation : CGGC est très surreprésenté dans YAP
Z-score : 1.880663
p-value : 0.9699911

Pour les motifs : nous cherchons à savoir si ils sont sur ou sous représentés. Pour cela, nous allons générer les séquences aléatoires sous le modèle shuffle de chaque séquence des fichier MET_up800.fasta et YAP_up800.fasta pour générer une distribution aléatoire du nombre de hits avec les motifs **YAP1** et **STP15** pour un threshold de 1.

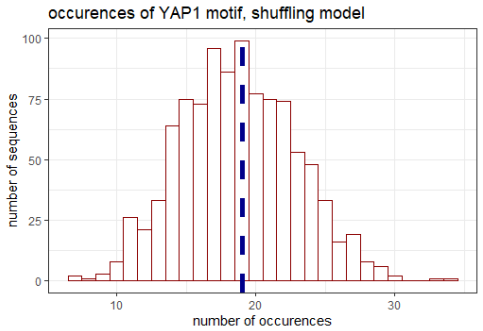
Voici ce que nous obtenons :

- YAP1 pour MET



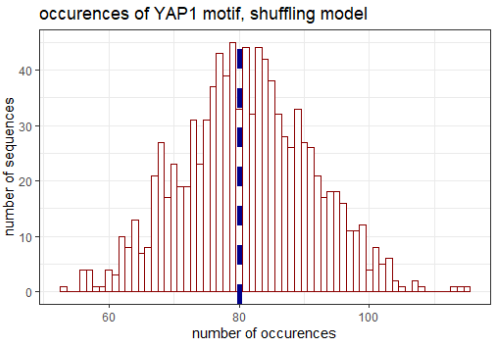
Interprétation : YAP1 est très surreprésenté dans MET
Z-score et **P-value** : 4.587861 et 2.239049e-06

- STP15 pour MET



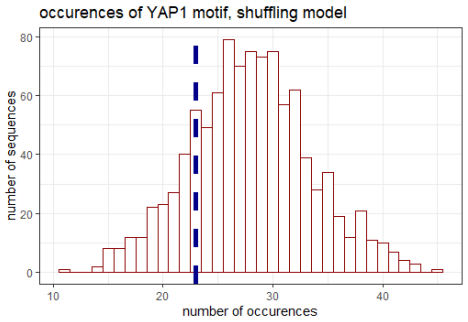
Interprétation : STP15 est ni sur ni sous-représenté dans MET
Z-score et **P-value** : 0.06274295 et 0.4749856

- YAP1 pour YAP



Interprétation : YAP1 est ni sur, ni sous-représenté dans YAP
Z-score et **P-value** : -0.1282001 et 0.5510047

- STP15 pour YAP



Interprétation : STP15 est légèrement sous-représenté dans YAP
Z-score et **P-value** : -0.9005458 et 0.8160851