Université Pierre et Marie Curie

Analyse de séquences génomiques

Rapport de projet

Pablo Donato
Alexandre Doussot
3I005 – Probabilités, Statistiques et Informatique
Paris, le 12 mars 2017

Professeurs : Hugues Richard
Thomas Gerald

Préliminaires : données et lecture de fichiers

Question 4

On cherche à calculer la fréquence d'apparition de chaque lettre dans le génome de *S. cerevisae*.

Pour cela nous avons écrit la fonction `freq_letters` dans le fichier `genoseq.py` qui renvoie la liste des fréquences d'apparition des lettre ACGT dans le génome (liste d'entiers) passé en paramètre.

On récupère les séquences dans les fichiers `regulatory_seq_PHO.fasta`, `regulatory_seqs_MET.fasta` et `regulatory_seqs_GAL.fasta`, puis on les concatène en une unique séquence que l'on passe à la fonction `freq_letters`. On obtient alors les fréquences suivantes :

A	C	G	T
0.3165	0.1956	0.1860	0.3017

Annotation des régions promoteurs

Description empirique, préliminaires

Question 2

On cherche à connaître le nombre théorique attendu d'occurrences d'un mot w dans une séquence de longueur l, sachant qu'on connaît les fréquences d'apparition des lettres dans le génome.

Soit $a \in \{0, 1, 2, 3\}$ la variable correspondant à la notation en nombre entier d'une lettre parmi $\{A, C, G, T\}$.
On note p_a la probabilité pour a d'être tirée dans un modèle aléatoire de la séquence, qui correspond ici à la fréquence d'apparition de a dans le génome.
On note n_a^w le nombre d'occurrences de a dans w.

On remarque que l'ordre des lettres dans w n'a aucune importance ici, seul le nombre d'occurrences influe sur la probabilité d'apparition de w dans la séquence : on s'attend en effet à avoir moins d'occurrences du mot si celui-ci contient un plus grand nombre de lettres.

On peut alors calculer la probabilité de tirer w à une position donnée dans la séquence en multipliant les probabilités de tirer chaque lettre a n_a^w fois :

$$\prod_a (p_a)^{n_a^w}$$

Pour déterminer le nombre théorique d'occurrences N_w^l de w dans la séquence, on cumule la probabilité que l'on vient de calculer pour chaque position possible de w.

Sachant que w est de longueur $\sum_a n_a^w$ et qu'on a donc $l - \sum_a n_a^w + 1$ positions possibles, on obtient :

$$N_w^l = \prod_a (p_a)^{n_a^w} \times \left(l - \sum_a n_a^w + 1 \right)$$

La formule est implémentée dans la fonction `expected_counts` du module `fixation.py` qui l'applique à tous les mots de longueur k.

Question 3

On cherche à évaluer à quel point les nombres d'occurrences des mots de longueur k observés sur les séquences PHO, MET et GAL s'écartent des nombres d'occurrences théoriques attendus.

Pour cela nous avons implémenté la fonction `plot_counts` dans le module `fixation.py` qui nous permet d'afficher un graphe de dispersion du nombre observé en fonction du nombre attendu pour différentes séquences et différentes valeurs de k. On obtient les résultats suivants :

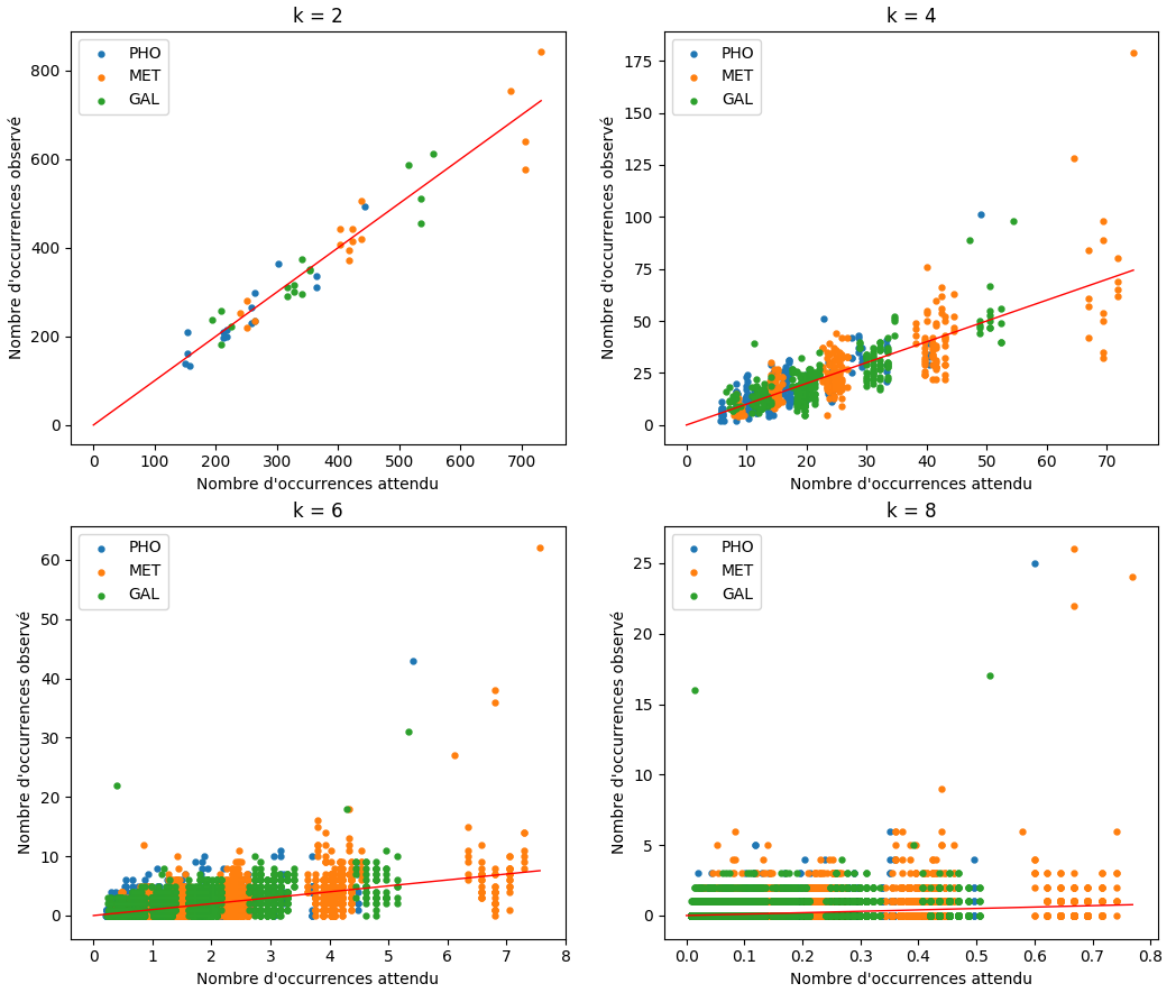

FIGURE 1 – Nombres d'occurrences des mots de taille k dans le génome de *S. cerevisae*

La droite rouge correspond à l'enrichissement relatif : les points situés au-dessus (resp. en-dessous) de la droite sont les mots dont le nombre d'occurrences observé est supérieur (resp. inférieur) au nombre d'occurrences attendu.

On voit que le nombre d'occurrences observés suit globalement le nombre d'occurrences attendu. Aussi, comme nous l'avions prévu, plus la longueur du mot k est grande, et plus sa fréquence d'apparition est proche de 0.

Néanmoins on remarque un fait intéressant : plus k est grand, et plus on observe un petit nombre de mots bien plus fréquents que les autres.
On peut conjecturer que ces mots très fréquents sont caractéristiques du génome de *S. cerevisae*.

Par la suite nous avons décidé d'expérimenter un peu plus avec ces valeurs en dessinant la distribution des nombres d'occurrences observés selon leur indice lexicographique :

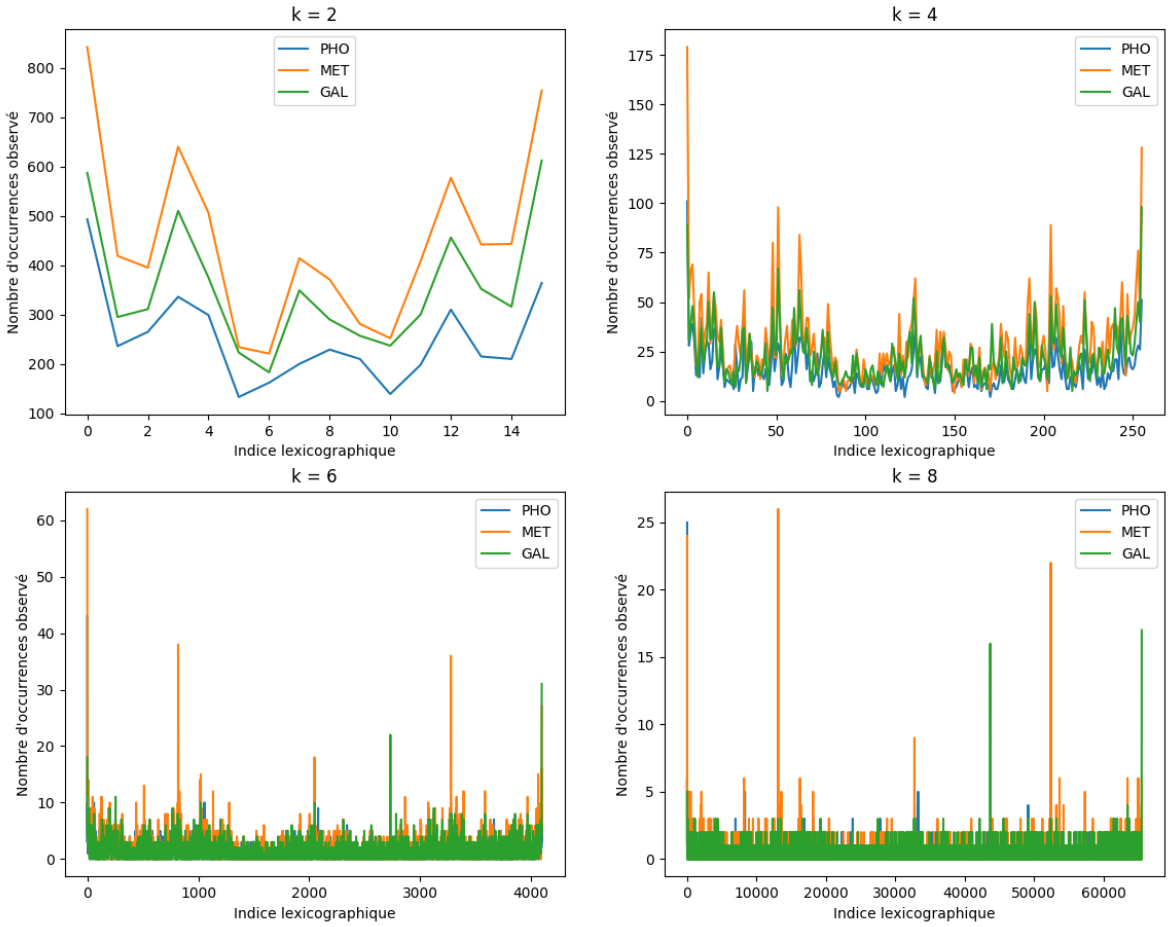

FIGURE 2 – Distribution des nombres d'occurrences observés des mots de taille k dans le génome de *S. cerevisae*

On remarque alors que quelque soit la valeur de k, la distribution est très similaire, ce qui ce voit tout particulièrement sur les positions relatives des pics, qui correspondent aux mots caractéristiques trouvés précédemment.

Simulation de séquences aléatoires

Question 2

Question 3

Question 4

Probabilités de mots

Question 1

Question 2

Question 3

Question 5