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. Chaque lettre a possède ainsi une probabilité $(p_a)^{n_a^w}$ d'apparaître n_a^w fois dans w.

On peut alors calculer la probabilité d'apparition de w à une position donnée dans la séquence en multipliant les probabilités d'apparition de chaque lettre dans w :

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

Pour déterminer finalement l'espérance théorique du nombre 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

Simulation de séquences aléatoires

Question 2

Question 3

Question 4

Probabilités de mots

Question 1

Question 2

Question 3

Question 5