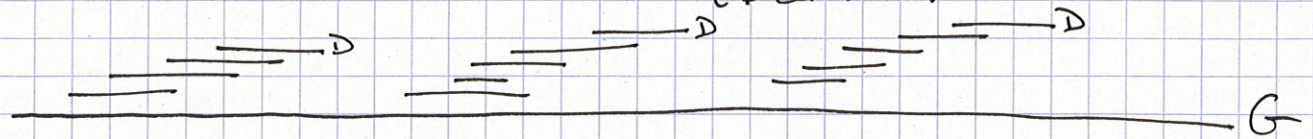
On génère N lectures à partir d'une distribution de longueurs f_L aléatoirement

La taille du génome est G.

On cherche à calculer la valeur moyenne du nombre de contigs

Pour cela on définit une v.a. Z_i $\begin{cases} 1 \text{ si lecture } i \text{ est fragment le plus à droite du contig} \\ 0 \text{ sinon.} \end{cases}$



$Z_i = 1$ pour toutes les lectures labellisées D ci-dessus.

La lecture i est le fragment le plus à droite du contig s'il n'y a aucune lecture parmi les N-1 autres qui commence dans la lecture i

Si on appelle X_j la position de départ de la lecture j on a :

$\mathbb{P}(X_j = x) = \frac{1}{G}$ $\forall x$ (les lectures sont générées de façon uniforme sur génome)

$\mathbb{P}(X_j$ se trouve dans lecture $i) = \frac{L_i}{G}$

Le nombre de lectures dont le point de départ se trouve dans la lecture i est donné par la binomiale

$\mathcal{B}(N-1, \frac{L_i}{G})$ donc

$$\mathbb{P}(Z_i = 1) = \binom{N-1}{0} \left(\frac{L_i}{G}\right)^0 \left(1 - \frac{L_i}{G}\right)^{N-1} \sim e^{-\frac{NL_i}{G}}$$

Nbre total de contigs = nbre total de lectures les plus à droite d'un contig $= \sum_{i=1}^{N} Z_i$

La moyenne du nombre de contigs : $\mathbb{E}(\sum_{i=1}^{N} Z_i) = \sum_{i=1}^{N} \mathbb{E}(Z_i)$

Si $\mathbb{E}(Z_i)$ est indépendant de i, on a $N\mathbb{E}(Z_i)$

<u>mais</u> $\mathbb{E}(Z_i) = 1\,\mathbb{P}(Z_i=1) + 0\,\mathbb{P}(Z_i=0) = \mathbb{P}(Z_i=1) = e^{-\frac{NL_i}{G}}$

Je ne comprends pas pourquoi la valeur moyenne de Z_i dépend de i

Je m'attends à trouver $Ne^{-\frac{N\mathbb{E}(L)}{G}}$ qui est une généralisation du cas où toutes les lectures ont une longueur L identique.

($\mathbb{E}(L)$ est remplacé par L)

$$\mathbb{E}(Z_i) = \sum_{k=1}^{\infty} e^{-\frac{Nk}{G}} f_L(k)$$

En décomposant $e^{-\frac{Nk}{G}}$ au voisinage du point $-\frac{N}{G}\mathbb{E}(L)$ en série de Taylor on a

$$e^{-\frac{Nk}{G}} = e^{-\frac{N\mathbb{E}(L)}{G}} + e^{-\frac{N\mathbb{E}(L)}{G}}\left(-\frac{Nk}{G} + \frac{N}{G}\mathbb{E}(L)\right) + \frac{e^{-\frac{N\mathbb{E}(L)}{G}}}{2}\left(-\frac{Nk}{G} + \frac{N}{G}\mathbb{E}(L)\right)^2 + \dots$$

$$= e^{-\frac{N\mathbb{E}(L)}{G}}\left[1 + \frac{N}{G}\left(\mathbb{E}(L) - k\right) + \frac{1}{2}\frac{N^2}{G^2}\left(k - \mathbb{E}(L)\right)^2 + \dots\right.$$

Donc quand on intègre sur la distribution de longueur

$$\mathbb{E}(Z_i) \simeq e^{-\frac{N\mathbb{E}(L)}{G}}\left[1 + \frac{1}{2}\frac{N^2}{G^2}\mathbb{V}(L)\right]$$

Au final le nombre moyen de contig est donné par

$$N\mathbb{E}(Z_i) \simeq N e^{-\frac{N\mathbb{E}(L)}{G}}\left[1 + \frac{1}{2}\frac{N^2}{G^2}\mathbb{V}(L)\right]$$

donc à l'ordre 2 on a un terme correctif qui dépend de la variance si on remplace L par $\mathbb{E}(L)$ dans la formule ~~à droite~~ qu'on obtient quand toutes les longueurs des lectures sont identiques

Bon, je n'ai pas vérifié quel était l'ordre de grandeur du reste quand on bronque à l'ordre 2 la série de Taylor