

a/ La signature occupe les positions 107 à 113 (centrée sur 110, position de la sérine active). Le nombre d'heptapeptides dans la séquence, hors la signature, entre les positions 1 à 106 et 114 à 320, est : $N=(106-7+1)+((320-114+1)-7+1) = 301$
En toute rigueur, il faut enlever à cette valeur 2 fois 7 possibilités dues à la présence de l'HIS et de l'ASP de la triade (en positions 270 et 295) qu'il ne faut pas inclure dans cette recherche.

D'où $N=301-2*7=287$ (la valeur d'environ 300 heptapeptides est acceptée!)

La probabilité d'obtenir par le hasard une séquence ZXYsXXB dans un heptapeptide est : $\Pi = 18/20 \times 1/20 \times 1/20 \times 1/20 \times 1 \times 1 \times 3/20 = 54/160000 = 0,0003375$

Soit X : "Nombre d'heptapeptides ZXYsXXB de part et d'autre de la signature, centrée sur la sérine active des zylinase"

X suit une loi binomiale de paramètre $N=287$ et $\Pi = 54/160000$; $X \sim \mathcal{B}(287; 54/160000)$

Espérance : $\mu = N\Pi = 287 * 54/160000 = 0,09686$

Variance : $\sigma^2 = N\Pi(1-\Pi) = 287 * 54/160000 * 159946/160000 = 0,09683 \sim N\Pi$

$N\Pi \ll 5$; mais $\sigma^2 \sim \mu \Rightarrow$ on peut donc ici approcher les calculs en utilisant une loi de Poisson de paramètre $\lambda = 0,1$; $\mathcal{P}(0,1)$.

La probabilité recherchée est $P(X > 0) = 1 - P(X=0)$

En consultant la table de la loi de Poisson à la colonne $\lambda=0,1$ et à ligne 0, on obtient :

$P(X > 0) = 1 - 0,9048 = 0,0952$

Cette probabilité est loin d'être négligeable.

b/ La proportion observée $P_o = 5/253 = 0,0198$ dans un échantillon de 253 séquences est à comparer à la proportion exacte $P(X \geq 1)$

avec X suivant une loi Poisson $\mathcal{P}(0,1)$ de paramètre $\lambda = 0,1$

(si le motif ZXYsXXB est rencontré plus d'1 fois dans la séquence, il est rencontré au moins 1 fois ailleurs que centré sur la sérine active)

X suivant une loi Poisson $\mathcal{P}(0,1)$, $P(X \geq 1) = 1 - P(X=0) = 1 - 0,9048 = 0,0952$

On peut au choix traiter la question en faisant un test d'hypothèse ou en construisant un intervalle de confiance sur la proportion observée $P_o = 0,0198$ ou encore en construisant un intervalle de pari sur la proportion exacte attendue $\Pi = 0,0952$

Test d'hypothèse

H_o : la proportion d'heptapeptide ZXYsXXB dans la population des séquences est $\Pi = 0,0952$ (les différences observées sont dues aux fluctuations d'échantillonnage de P_o)

Sous H_o , P_o suit une loi binomiale $\mathcal{B}(253; 0,0952)$.

$N\Pi = 253 \times 0,0952 = 24,09 > 5$

$N(1-\Pi) = 253 \times 0,9048 = 228,91 > 5$

Espérance : 0,0952 ; Ecart-type : $((0,0952 \times 0,9048)/253)^{0,5} = 0,0185$

Les calculs peuvent être approchés par une loi Normale $\mathcal{N}(0,0952 ; 0,0185)$ de même espérance et écart-type que la loi binomiale.

$\alpha_o = P(P_o < 0,0198)$

$= P(\varepsilon_o < (0,0198 - 0,0952)/0,0185)$

$= P(\varepsilon_o < -4,0757) = 0,0000225$

$\alpha_0 \sim 0$; α_0 est très inférieur à 5%, risque seuil standard

On rejette H_0 avec un risque de première espèce pratiquement nul! Les différences sont très significatives. La proportion observée n'est pas conforme à ce que l'on pourrait attendre selon une répartition aléatoire des heptapeptides dans le domaine catalytique des zylinases.

Le résultat n'est pour autant pas si surprenant que cela : les séquences de protéines ne sont pas construites au hasard!...

(Remarque : zylinase est un nom inventé dans cet exercice)