BIOINFORMATIQUE, Unité libre 2007
Partie "Statistiques" - Enoncés
(à rédiger sur une copie séparée)

Question 1

Dans le fichier **http ://www.info.univ-angers.fr/ gh/Bis/bistd2.pdf**, on trouve écrit tout en bas de la page 5 qu'il y a un peu moins de 3 % de protéines avec les deux types de ponts pour les protéines à ponts de la DBDB. Donner la fraction qui permet d'obtenir cette information et sa valeur en % avec deux décimales.

Question 2

Deux ensembles distincts de protéines nommés A et B, soigneusement choisis et rapatriés contiennent respectivement nb(A) =80 et nb(B) = 110 protéines. Les moyennes respectives de la longueur des chaines sont moy(A) = 350 aa et moy(B) = 380 aa ; les médianes sont respectivement med(A) = 300 aa et med(B) = 290 aa. Quel(s) test(s) statistiques peut-on effectuer pour savoir si les deux ensembles A et B sont significativement différents pour la variable "longueur des chaines" si on ne dispose que ces informations ? Si vous savez calculer ce ou ces tests, qu'en conclut-on au risque α de première espèce égal à 5 % ?

Question 3

Pourquoi trouve-t-on souvent le mot "rang" (rank en anglais) dans les noms de tests non-paramétriques ?

BIOINFORMATIQUE 2007
Partie "Statistiques" - Solution

Réponse à la question 1

En TD, nous avions vu qu'il y avait 11 protéines avec les deux types de ponts et qu'au total on disposait de 401 "protéines à pont". La proportion demandée est donc `11 / 401` soit 2,74 % si on l'exprime en pourcent avec deux décimales (avec plus de décimales ce serait 2,743142).

Réponse à la question 2

Malheureusement avec aussi peu de renseignements on ne peut faire aucun test. Une comparaison de moyennes (vue en TD) suppose que l'on utilise les variances qui ne sont pas fournies ici. Pour une comparaison de médianes (non vue, mais facile à trouver via *Google*) on se rend compte qu'il faut trier les valeurs pour compter celles avant et après la médiane. Si on ne dispose que de la médiane, ce test n'est pas applicable non plus.

Réponse à la question 3

Dans de nombreux tests non paramétriques, on utilise les rangs des valeurs plutôt que les valeurs elles-mêmes. Par exemple au lieu de 12, 5 et 28 on utilise les valeurs 2, 1 et 3 (car 12 est la deuxième valeur, 5 la première et 28 la troisième par ordre croissant). C'est pourquoi le mot *rank* apparait souvent.