Bioinformatique, partie Statistiques (L3) TD2 : Répartition des protéines, des ponts et des premiers acides aminés

1. Présentation des données

La base de données DBDB disponible à l'adresse

http ://www.info.univ-angers.fr/pub/richer/rec/bio/dbdb/

contient de nombreuses protéines avec des ponts disulfure (comme au TD1).

2. Série de Questions 1

Sachant qu'une protéine est répérée dans la table dbdbprot de la base ddbdb par un champ identifiant nommé `pr_id`, quelle instruction SQL faut-il écrire pour trouver le nombre de protéines en tout ?

Les champs `pr_nbr_intra` et `pr_nbr_inter` donnent respectivement le nombe de ponts intra et de pont inter pour une protéine donnée. Quelle(s) instruction(s) SQL faut-il écrire pour compter les protéines avec pont et celles sans pont ? Comment en déduire les pourcentages correspondant ? Serait-ce plus simple en PHP ? Comment exporter vers un fichier texte (pour $Excel$, $Rstat$...) ?

Sachant que les résultats sont consignés dans le tableau suivant, commenter ces résultats avant de tracer le graphique correspondant.

Protéines avec pont intra ou inter	401
Protéines avec pont intra et pont inter	11
Protéines avec pont intra seulement	366
Protéines avec pont inter seulement	24
Protéines sans pont	52

Effectuer le même genre d'analyse pour les divers types de ponts toutes protéines confondues et pour les cystéines présentes dans les chaines polypeptidiques. Au passage combien y a-t-il de chaines par protéine en moyenne ? et combien de ponts par protéine ?

Quelles(s) variable(s) statistique(s) QL pourrait-on définir ? et quelle(s)s QT ?

3. Série de Questions 2

On s'intéresse maintenant à l'ensemble des protéines de la DBDB sous l'angle de séquences d'acides aminés (ou plutôt de "résidus"). Le fichier `chp.aa1` contient à raison d'une protéine par ligne l'identifiant PDB de la protéine et le premier acide aminé de la séquence *Fasta* associée. Le fichier `chp.aa3` contient l'identifiant et les 3 premiers acides aminés de la séquence *Fasta* associée.

Les fichiers chp$i$.* avec i=1, 2 ou 4 contiennent les mêmes informations pour des sous-populations choisies de la DBDB :

- $i = 1$: protéines avec ponts inter et intra ;
- $i = 2$: protéines avec ponts inter sans pont intra ;
- $i = 4$: toxines avec ponts intra sans pont inter.

Quelle sont les variables statistiques qualitatives dans les fichier `*.aa1` ?

Quels calculs statistiques descriptifs à une dimension faut-il effectuer globalement sur l'ensemble de la population pour ces variables ? Et pour les sous-populations ?

Les effectuer avec *Excel* et *Rstat* puis commenter les résultats sans oublier de réaliser les graphiques correspondant.

Quels calculs statistiques descriptifs à deux dimensions faut-il effectuer globalement sur l'ensemble de la population pour les variables des fichiers `*.aa3` ?

Les effectuer avec *Excel* et *Rstat* puis commenter les résultats. sans oublier de réaliser les graphiques correspondant.

Pourquoi est-il plus exact de parler de résidu plutôt que d'acide aminé ?

4. Série de Questions 3

Quels calculs statistiques peut-on effectuer avec deux variables, l'une qualitative, l'autre quantitative ? On pensera par exemple aux fichiers utilisés dans le TD1 et le TD2.

Effectuer ces calculs avec *Excel* et *Rstat* puis commenter les résultats. sans oublier de réaliser les graphiques correspondant.

Rappel : Les fichiers de données sont disponibles à l'adresse

http ://www.info.univ-angers.fr/pub/gh/Bis/bis.htm