BIOINFORMATIQUE, septembre 2011
Master TPV, Partie "Statistiques"
(à rédiger sur une feuille simple séparée)

La base de données DISPROT recense des protéines *intrinsèquement désordonnées*, c'est-à-dire sans structure 3D stable. Son site Internet est à l'adresse :

`http://www.disprot.org/index.php`

On voudrait comparer les longueurs des protéines de cette base avec celles de la LEAPDB. Le fichier displea.dar contient un extrait de chacune de ces deux bases. Vous le trouverez sur Internet à l'adresse :

`http://forge.info.univ-angers.fr/~gh/Bism/displea.dar`

et aussi localement, à l'endroit habituel (sur le disque K :, dans le répertoire stat_ad). Chaque ligne après la première, contient le numéro d'accession de la protéine, son origine (1 pour LEADB, 2 pour DISPROT) et sa longueur.

1. Combien y a-t-il de protéines pour chacun des deux échantillons ? Quelle est la moyenne des longueurs pour chacun des deux échantillons ?

2. Réaliser une comparaison statistique entre les longueurs pour les deux échantillons. On fournira une conclusion "propre" sur la présence d'une différence significative ou non pour ces longueurs.

On trouvera au dos de cette page un extrait des données.

Extrait des données

id	base	lng
1906384B	1	110
1YYCA	1	174
A2XG55	1	333
A2ZDX4	1	151
A2ZDX6	1	164
A2ZDX8	1	164
A2ZDX9	1	172
...		
P46527	2	198
P46532	1	93
P49799	2	205
P49913	2	37
P49918	2	316
P50224	2	295
P50440	2	423
P51123	2	2068
...		
Q9Y9L0	2	250
XP_00227	1	93
XP_00228	1	191
XP_00231	1	93
YP_00279	1	133