BIOINFORMATIQUE, avril 2015

Master BPV, Partie "Statistiques"

(à rédiger sur une copie séparée)

Pour répondre aux questions posées, vous utiliserez les données à l'adresse :

`http://forge.info.univ-angers.fr/~gh/Bism/shsp4609.dar`

Le fichier correspondant contient un peu plus de quatre mille six cent protéines considérées comme des "petites" protéines HSP.

1. Si on devait fournir un indicateur de tendance centrale pour la variable pI (point isoélectrique), que choisiriez-vous ? La moyenne ou la médiane ? Et surtout, pourquoi ?

   Essayez ensuite de décrire complétement cette variable pI avec des indicateurs statistiques numériques. Vous compléterez cette liste d'indicateurs par la rédaction détaillée et soignée d'un unique paragraphe. On ne demande pas de reproduire de graphique.

2. Comparer la variable pI de ces 4609 protéines aux pI des 1493 protéines de la LEAPdb dont les données sont à l'adresse :

   `http://forge.info.univ-angers.fr/~gh/Bism/leadb1493.dar`

   Indiquez clairement quel test vous utilisez et comment vous l'avez choisi. On rédigera une conclusion statistique et une conclusion "métier".

(suite au dos)

3. Peut-on envisager une relation linéaire entre la variable pI des protéines sHSP et la variable pI des protéines LEA ? Pourquoi ?

4. On voudrait maintenant savoir s'il y a une corrélation linéaire significative entre la variable pI des protéines sHSP et l'une des variables 3 à 14 des données (de MW à Transmembr). Quelle(s) instructions R faut-il écrire ?

   Quel est alors, parmi ces variables, la variable la plus corrélée à pI ? Et quelle est la relation linéaire associée ? Là encore, on ne fournira pas de graphique.

   Par contre, on pourra réfléchir à une relation de causalité possible et – bien que ce ne soit pas statistique – essayer de formuler une explication quant à l'existence de cette relation linéaire.

– FIN DU SUJET –