BIOINFORMATIQUE, Unité libre 2006 (1)
Partie "Statistiques" - Enoncés

Nous avons vu en cours que le nombre de chaines pour les protéines de la DBDB, traité comme une QT fournissait les résultats suivants

```
Nombre de valeurs : 388     protéines
Moyenne           :   1.07  chaines
Ecart-type        :   0.34  chaines
Cdv               :  32     %
```

Si on regarde maintenant la répartition du nombre de chaines, on trouve

```
 415  chaines pour  388  protéines

      367  protéines avec  1  chaine   soit 94.59 %
       18  protéines avec  2  chaines        4.64 %
        1  protéines avec  3  chaines        0.26 %
        1  protéines avec  4  chaines        0.26 %
        1  protéines avec  5  chaines        0.26 %
```

On se propose de définir une QL "classe de nombre de chaines". Au vu des résultats précédents, que proposez comme nombre de modalités et comme modalités ? Quel est alors le tri à plat de cette variable ? Est-ce cohérent avec l'analyse de la QL "Type de pont" (inter/intra) faite en cours ?

Aucun calcul n'est exigé. Par contre la rédaction devra dépasser 10 lignes et être très précise et très soignée.

BIOINFORMATIQUE, Unité libre 2006 (1)
Partie "Statistiques" - Solution

Au vu des résultats que l'on peut résumer par la phrase *"presque la totalité des protéines n'ont qu'une seule chaine"*, on peut définir la variable CNC (classe de nombre de chaines) par les formules : CNC=1 si la protéine n'a qu'une seule chaine, CNC=2 si la protéine a plus d'une chaine.

Le tri à plat associé est alors

```
CNC  1_seule_chaine 95 %  plus_d_une_chaine 5 %
```

Rappelons le résultat sur la variable "Type de pont" : on avait une majorité de ponts *intra*. Il n'y a pas d'incohérence avec cette majorité de protéines à une seule chaine. Par contre, si on avait eu une majorité de ponts *inter* on aurait pu conclure à une majorité de protéines avec plus d'un pont puisqu'il faut deux chaines pour une même protéine pour faire un pont *inter*.