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

Réponses à la série de questions 1

Dans la mesure où les comptages à effectuer ne mettent en jeu qu'une seule table, un simple `select count(...)` permet de réaliser les comptages demandés, soient les instructions :

```
# comptage global

  select count(pr_id) from dbdbprot ;

# sans pont

  select count(pr_id) from dbdbprot where pr_nbr_intra=0 and pr_nbr_inter=0 ;

# avec pont

  select count(pr_id) from dbdbprot where pr_nbr_intra>0 OR  pr_nbr_inter>0 ;

# ponts intra seuls

  select count(pr_id) from dbdbprot where pr_nbr_intra>0 and pr_nbr_inter=0 ;

# ponts inter seuls

  select count(pr_id) from dbdbprot where pr_nbr_intra=0 and pr_nbr_inter>0 ;

# avec les deux types de ponts

  select count(pr_id) from dbdbprot where pr_nbr_intra>0 and pr_nbr_inter>0 ;
```

Pour trouver les pourcentages, c'est un peu plus compliqué suivant la version de *sql*. En *mysql*, par exemple, on ne peut pas effectuer de `"select"` imbriqués et donc il faut rentrer les valeurs "à la main" dans une section interactive comme :

```
# le comptage global donne 453 donc pour les pourcentages  de "sans pont"

  select  concat(round(100.0*(count(pr_id)/453))," % ")
          from dbdbprot where pr_nbr_intra=0 and pr_nbr_inter=0 ;
```

ou même seulement `select 100.0*(52/453) ;` si on connait toutes les valeurs pour calculer le pourcentage.

Par contre si on dispose de *postgresql* ou d'*oracle*, il est possible de diviser un `"select count"` par un autre `"select count"` soit une instruction comme :

```
select round( 100.0 * (count(pr_id) / (select count(pr_id) from dbdbprot)))
       from dbdbprot where pr_nbr_intra=0 and pr_nbr_inter=0 ;
```

En *php* les choses sont plus simples car les résultats des requêtes peuvent être mis dans des variables donc le pourcentage correspond à un simple calcul entre variables :

```
<?
   # le comptage global renvoie 453 dans $nbpr
   # donc pour les pourcentages  de "sans pont"

     $rq    = "select count(pr_id) from dbdbprot " ;
     $rq   .= "       where pr_nbr_intra=0 and pr_nbr_inter=0  " ;
     $er    = mysql_query(" $rq ") ;
     $ligr  = mysql_fetch_array($er) ;
     $sansp = $ligr["count(pr_id)"] ;
     $pct   = sprintf("%3d",100.0*$sansp/$nbpr)
?>
```

Tous calculs effectués, on peut donc présenter le tri à plat ordonné de la variable qualitative "type de protéines" de la façon suivante :

Modalité	Pourcentage
Avec pont	89 %
dont pont intra seul	81 %
dont pont inter seul	5 %
dont intra et inter	2 %
Sans pont	11 %

Remarque : les valeurs sont arrondies.

La base de données contient donc (et c'est heureux) presque exclusivement des protéines avec pont. Les protéines à ponts sont très trés majoritairement (366/401 soit 91 %) des protéines avec des ponts intra seulement et il y a très très peu de protéines (un peu moins de 3 %) avec les deux types de ponts.

Pour exporter vers un fichier-texte, au lieu de taper *mysql* puis de taper les requêtes en interactif, on peut utiliser la redirection des entrées sorties, c'est à dire mettre la requête dans un fichier (par exemple `td2p1.sql`) et taper

```
mysql < td2p1.sql > td2p1.txt
```

L'affichage ne se fait plus à l'écran mais dans le fichier indiqué soit `td2p1.txt`.

Quelle requête exécuter ?

Si on écrit directement

```
select pr_id, pr_nbr_intra,pr_nbr_inter from dbdbprot ;
```

le résultat sera difficilement exploitable car on aura

```
+-------+--------------+--------------+
| pr_id | pr_nbr_intra | pr_nbr_inter |
+-------+--------------+--------------+
|     1 |            0 |            1 |
|     2 |            5 |            1 |
|     3 |            1 |            0 |
|     4 |            3 |            0 |
|     5 |            4 |            0 |
...
```

et il faudra ensuite regrouper les valeurs strictement positives ensemble.

Heureusement *mysql* dispose d'une fonction `IF` ce qui permet d'écrire :

```
select pr_id,
       if(pr_nbr_intra>0,"APA","SPA"),
       if(pr_nbr_inter>0,"APR","SPR")
       from dbdbprot ;
```

où `APA` signifie "Avec Pont intrA" et `APR` signifie "Avec pont Inter".

L'affichage correspondant est

```
+-------+--------------------------------+--------------------------------+
| pr_id | if(pr_nbr_intra>0,"APA","SPA") | if(pr_nbr_inter>0,"APR","SPR") |
+-------+--------------------------------+--------------------------------+
|     1 | SPA                            | APR                            |
|     2 | APA                            | APR                            |
|     3 | APA                            | SPR                            |
|     4 | APA                            | SPR                            |
|     5 | APA                            | SPR                            |
...
```

Il suffit alors d'importer le fichier sous *Excel* ou de le lire avec `read.table` en *Rstat*.

Pour trouver le nombre de protéines avec à la fois un ou des ponts intra et un ou des ponts inter, on peut utiliser la même technique en *mysql* soit la requête

```
select pr_id,
if(pr_nbr_intra>0 and pr_nbr_inter>0,"AVEC","SANS")
from dbdbprot ;
```

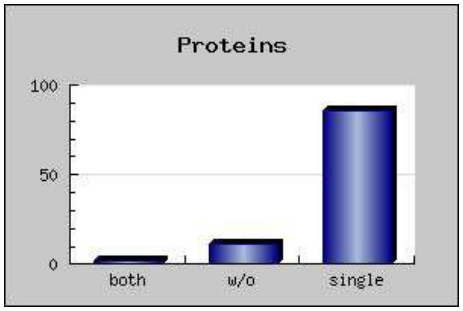
Si on n'y a pas pensé, on peut effectuer un test similaire sous *Excel* à l'aide de la fonction `SI`. Par exemple si la colonne B correspond à `APA/SPA` et si la colonne C correspond à `APR/SPR` on peut mettre en colonne D à la ligne i

```
= SI(Bi="APA";SI(Ci="APR";"AVEC";"SANS");"SANS")
```

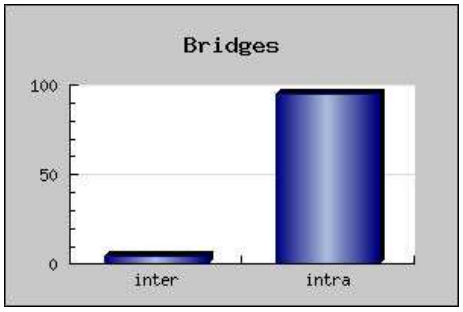
et on peut ensuite utiliser le calcul de tri à plat et tri croisé d'Excel via le menu Données / Rapport de tabelau croisé dynamique.

Si on compte le nombre de ponts en tout (car une protéine peut avoir plusieurs ponts) on trouve les mêmes résultats (heureusement) avec presque exclusivement de ponts intra. Nous reproduisons ici les résultats et histogrammes de fréquences fournis (en dynamique) par le script *php* de la page : `http://www.info.univ-angers.fr/pub/gh/Idas/adbdb.php`

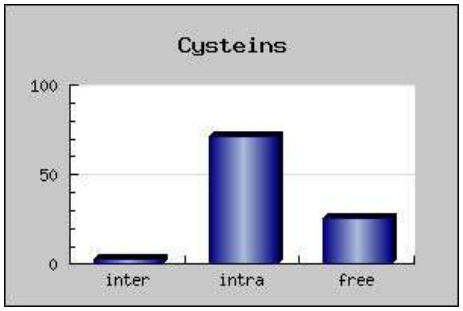
Protéines		
total	453	100 %
inter et intra	11	2 %
sans pont	52	11 %



Ponts		
total	951	100 %
intra	907	95 %
inter	44	4 %



Cystéines		
total	2567	100 %
intra	1814	70 %
inter	88	3 %
libres	665	25 %



L'analyse du fichier `chp.aa1` peut se faire en R à l'aide du programme

```
# chargement des fonctions (gH)
  source("statgh.r")

# lecture des données
  aa   <-   read.table("chp1.aa1",header=TRUE) ;
  dims   <-   dim(aa) ;
  nbl    <-   dims[1] ;
  nbc    <-   dims[2] ;
  maa    <- aa[1:nbl,2:nbc]  ;

# écriture directe en R du tri à plat

  pre         <- as.factor(aa[,2])  ;
  table(pre)

# tri à plat plus évolué :

  numNomAA    <- c("A","C","D","E","F","G","H","I","K","L","M",
                   "N","P","Q","R","S","T","V","W","Y","X")
  levels(pre) <-  numNomAA
  round(100.0*sort(table(pre),decreasing=TRUE)/length(pre))

# tracé de l'histogramme des fréquences

  histEffectifs(" CHP1",table(pre),1:21)

# à l'aide de la fonction triAplatAvecOrdre (gH)

  pre      <- aa[,2]  ;
  triAplatAvecOrdre("CHP1",pre,numNomAA)

  print(numNomAA)
```

On obtient comme résultat :

```
>table(pre)
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21
73 13 34 27  6 33  6 12 17 22 60  4 10 15 12 25 23 13  1  4  5

>round(100.0*sort(table(pre),decreasing=TRUE)/length(pre))
 A  M  D  G  E  S  T  L  K  Q  C  V  I  R  P  F  H  X  N  Y  W
18 14  8  8  7  6  6  5  4  4  3  3  3  3  2  1  1  1  1  1  0

>triAplatAvecOrdre("CHP",pre,numNomAA)

 QUESTION :  CHP

                   A  M  D  G  E  S  T  L  K  Q  C  V  I  R  P F H X N Y W  Total
  Effectif        73 60 34 33 27 25 23 22 17 15 13 13 12 12 10 6 6 5 4 4 1    415
  Fréquence (en %) 18 14  8  8  7  6  6  5  4  4  3  3  3  3  2 1 1 1 1 1 0     99
```

ce qui permet de voir la prépondérance de l'acide aminé A suivi de M. Une étude similaire pour `chp1`, `chp2` et `chp4` fournit comme résultats :

```
 QUESTION :  CHP1
                    D  C  F  G  H  I  K A E L  Total
  Effectif          4  3  2  2  2  2  2 1 1 1     20
  Fréquence (en %) 20 15 10 10 10 10 10 5 5 5    100

 QUESTION :  CHP2
                    H  F  C  D E A G I K L M  Total
  Effectif         12  4  3  3 2 1 1 1 1 1 1     30
  Fréquence (en %) 40 13 10 10 7 3 3 3 3 3 3     98

 QUESTION :  CHP4
                    H  A  E  K  M C F I L D G  Total
  Effectif          5  4  4  3  3 2 2 2 2 1 1     29
  Fréquence (en %) 17 14 14 10 10 7 7 7 7 3 3     99
```

ce qui semble indiquer des profils différents.

Pour prouver que les profils sont différents, on pourrait comparer acide aminé par acide aminé à l'aide d'une comparaison de pourcentages. Pour comparer globalement, il faudrait restructurer les données pour effectuer un calcul de chi^2.

Si on s'intéresse maintenant aux fichies de type `aa3`, comme on dispose de trois colonnes A1, A2 et A3 correspondant aux trois premiers acides aminés, il faut effectuer le tri à plat de chacune des colonnes et tous les tris croisés possibles. On peut effectuer ce traitement en *Rstat* avec le programme qui suit avant de récapituler "à la main" :

```
# chargement des fonctions (gH)

  source("statgh.r")

# lecture des données

  aa   <-   read.table("chp.aa3",header=TRUE) ;
  dims   <-   dim(aa) ;
  nbl    <-   dims[1] ;
  nbc    <-   dims[2] ;
  maa    <- aa[1:nbl,2:nbc]  ;

  numNomAA    <- c("A","C","D","E","F","G","H","I","K","L","M",
                   "N","P","Q","R","S","T","V","W","Y","X")

  # tris à plat "à la main"

  a1 <- aa[,2]  ;
  triAplatAvecOrdre(" A1 : ",a1,numNomAA)

  a2 <- aa[,3]  ;
  triAplatAvecOrdre(" A2 : ",a2,numNomAA)

  a3 <- aa[,4]  ;
  triAplatAvecOrdre(" A2 : ",a3,numNomAA)

  # tris croisés "à la main avec chi2"

  triCroiseAvecMarges("A1",a1,numNomAA,"A2",a2,numNomAA)
  chi2IndepTable(table(a1,a2))
```

```
triCroiseAvecMarges("A1",a1,numNomAA,"A3",a3,numNomAA)
chi2IndepTable(table(a1,a3))

triCroiseAvecMarges("A2",a2,numNomAA,"A2",a2,numNomAA)
chi2IndepTable(table(a2,a3))
```

ou on peut profiter de notre fonction *allQL* "qui s'occupe de tout" à condition de bien préparer les données. Nous reproduisons ci-dessous le texte de la préparation de l'appel

```
# chargement des fonctions (gH)
  source("statgh.r")

# lecture des données
  aa   <-   read.table("chp.aa3",header=TRUE) ;
  dims   <-   dim(aa) ;
  nbl    <-   dims[1] ;
  nbc    <-   dims[2] ;
  maa    <- aa[1:nbl,2:nbc]  ;

  numNomAA    <- c("A","C","D","E","F","G","H","I","K","L","M",
                   "N","P","Q","R","S","T","V","W","Y","X")

  dsc  <-  matrix(nrow=length(maa),ncol=3)
  lcol <- 1:3

  dsc[1,1] <- c("A1")
  dsc[1,2] <- c("A1")
  dsc[1,3] <- lstMod(numNomAA)

  dsc[2,1] <- c("A2")
  dsc[2,2] <- c("A2")
  dsc[2,3] <- lstMod(numNomAA)

  dsc[3,1] <- c("A3")
  dsc[3,2] <- c("A3")
  dsc[3,3] <- lstMod(numNomAA)

  allQL(maa,dsc,lcol)
```

Et on trouvera aussi une partie des résultats :

```
R : Copyright 2004, The R Foundation for Statistical Computing
Version 2.0.0  (2004-10-04), ISBN 3-900051-07-0

> source("statgh.r")
(gH) version  2.44
...
> allQL(maa,dsc,lcol)

  TABLEAU RECAPITULATIF DES VARIABLES QUALITATIVES


 Affichage par mode décroissant puis par effectifs décroissants

      A1  18  % A   14  % M    8  % D
      A3  11  % C    9  % T    8  % A
      A2  10  % P    9  % V    8  % S

 ANALYSE DE TOUTES LES VARIABLES QUALITATIVES


 QUESTION :  A1

                  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y  X   Total
 Effectif        73 13 34 27  6 33  6 12 17 22 60  4 10 15 12 25 23 13  1  4  5    415
 Fréquence (en %) 18  3  8  7  1  8  1  3  4  5 14  1  2  4  3  6  6  3  0  1  1     99


 QUESTION :  A2

                  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y  X   Total
 Effectif        32 16 29 32 10 24  5 28 24 17  2 12 42 16 19 33 24 37  2 10  1    415
 Fréquence (en %)  8  4  7  8  2  6  1  7  6  4  0  3 10  4  5  8  6  9  0  2  0    100


 QUESTION :  A3

                  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y  X   Total
 Effectif        35 44 26 13  9 30  9 13 22 15  4  9 33 21 13 31 38 35  2 11  2    415
 Fréquence (en %)  8 11  6  3  2  7  2  3  5  4  1  2  8  5  3  7  9  8  0  3  0     97


 ORDRE CONSEILLE POUR LIRE LES  3  TRIS CROISES

    Variable 1        Variable 2          Chi2       Chi2Table     p-value   Signif.   Ddl
     2   A2            3   A3           587.78     447.63        0.0000000    **    400
     1   A1            3   A3           490.57     447.63        0.0013006    **    400
     1   A1            2   A2           451.29     447.63        0.0388148     *    400
```

TRI CROISE DES QUESTIONS :
A1 (en ligne)
A2 (en colonne)

Valeurs en % du total

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	X	TOTAL
A	2	1	1	1	0	1	0	0	0	1	0	0	3	1	1	1	2	1	0	1	0	18
C	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	3
D	0	0	1	1	0	0	0	1	1	0	0	0	1	0	0	0	1	1	0	0	0	8
E	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	7
F	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
G	1	0	1	0	0	1	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0	8
H	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
I	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	3
K	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
L	1	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	5
M	1	0	1	1	0	1	0	1	2	1	0	1	0	1	1	2	1	1	0	0	0	14
N	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
Q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	4
R	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	3
S	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	6
T	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	6
V	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	3
W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
X	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
TOTAL	8	4	7	8	2	6	1	7	6	4	0	3	10	4	5	8	6	9	0	2	0	100

...