Bioinformatique, partie Statistiques (L3) TD1 : Etude des longueurs de chaines polypeptidiques

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. On s'intéresse ici à l'ensemble de ces protéines sous l'angle de "chaines polypeptidiques". Le fichier `chp.fasta` contient ces chaines au format *Fasta*, le fichier `chp.lng` contient les longueurs de ces chaines et le fichier `chp.cnt` dénombre de plus les acides aminés par chaine.

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.

Série de Questions 1

Combien y a-t-il de chaines polypeptidiques dans le fichier `chp.fasta` ?

Quelle est la nature de la variable statistique LNG ?

Quels calculs statistiques descriptifs faut-il effectuer globalement sur l'ensemble de la population pour cette variable LNG ? et par sous-population ?

Les effectuer avec *Rstat* puis commenter les résultats sans oublier de réaliser les graphiques correspondant. On discutera de l'automatisation des calculs. Pourquoi ne faut-il pas utiliser *Excel* ou *OpenOffice/Calc* ?

Série de Questions 2

Quelle est la nature des 20 variables statistique CntA, CntC... ?

Quels calculs statistiques descriptifs faut-il effectuer globalement sur l'ensemble de la population pour ces variables ? et par sous-population ?

Les effectuer avec *Excel* et *Rstat* puis commenter les résultats sans oublier de réaliser les graphiques correspondant. On discutera de l'automatisation des calculs.

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

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

Pour le logiciel R, on charge l'archive des programmes par

```
source("http://www.info.univ-angers.fr/pub/gh/wstat/statgh.r")
```

el la lecture des données (fichier *.chp) peut se faire par

```
chp <- lit.dar("http://www.info.univ-angers.fr/pub/gh/Bis/chp.lng")
names(chp)
attach(chp)
lng <- LNG
```

Réponses à la série de questions 1

Si on est sous *Windows*, il faudrait ouvrir le fichier `chp.fasta` à l'aide d'un éditeur de texte évolué comme PsPad (gratuit, disponible sur internet) pour voir le nombre de lignes. A défaut on peut utiliser *Excel* mais surtout pas *Word* ou *Notepad* : *Notepad* n'affiche pas les numéros de ligne en standard et il semblerait qu'il "coupe" les lignes au-delà de 1024 caractères alors que *Word* redéploie les lignes bizarrement. Si on fait "Fichier/Ouvrir" avec *Excel* et qu'on lui indique le chemin d'accès et le nom du fichier à l'aide du panneau d'ouverture de fichier, *Excel* affiche un panneau nommé "Assistant d'importation de texte". Il suffit alors de cliquer sur "Terminer" dès l'étape 1 pour voir toutes les lignes avec *Excel*, sauf peut-être la dernière ligne du fichier, qui est une ligne vide et qui se confond avec les lignes vides d'*Excel*.

Sous *Unix*, c'est beaucoup plus facile car il suffit d'écrire `wc -l chp.fasta` pour obtenir la réponse. Attention toutefois à wc (*word count*) sans option qui afficherait

```
1362    908     98251
```

c'est à dire respectivement le nombre de lignes, de mots (!) et de caractères.

On trouve donc qu'il y a 1362 lignes dans le fichier `chp.fasta`. Comme chaque séquence prend 3 lignes (1 pour l'identifiant, 1 pour la séquence, 1 ligne vide), il y a 1362/3 soit 454 séquences. Toutefois, certaines séquences sont identiques, comme `1UR5:A` et `1UR5:C`. Le fichier `chp.lng` contient 415 longeurs correspondant à 415 séquences distinctes.

La variable statistique `LNG` correspond à la longueur de la chaine polypeptidique c'est à dire au nombre d'acides aminés (en abrégé `aa`) ou plus exactement au nombre de résidus d'acides aminés car lors de la formation de la chaine polypeptidique un acide aminé perd une molécule d'eau. On peut additionner des nombres d'acides aminés donc `LNG` est une variable quantitative (QT) pour laquelle on doit calculer taille, moyenne, écart-type etc.

On trouvera sur la page suivante les principaux résultats. Le fichier `chplng.xls` permet de voir comment ces calculs ont réalisés sous *Excel* et le fichier `chplng.r` fournit pour le détail des calculs pour le logiciel *R*.

Nom	*Notation*	*Valeur*	*Unité*
Taille	n	415	chaines
Moyenne	m	176	aa
Ecart-type	σ	160	aa
Coef. de variation	σ/m	90	%
Minimum		11	aa
Maximum		1245	aa
1er quartile	q_1	61	aa
2ème quartile (médiane)	q_2	129	aa
3ème quartile	q_3	220	aa
Distance interquartile	$q_3 - q_1$	160	aa

On peut donc dire que pour ces données la longueur moyenne est donc d'environ 176 aa. Cette longueur ne varie pas beaucoup puisque son coefficient de variation est de moins de 100 %. A l'aide des quartiles, on peut dire que

- 25 % des protéines ont une longueur inférieure à 61 aa,
- 50 % des protéines ont une longueur comprise entre 61 et 220 aa,
- 25 % des protéines ont une longueur supérieure à 220 aa.

Si *Excel* permet d'effectuer ces calculs, il ne permet pas de tous les calculer d'un coup (sauf à programmer une macro). C'est pourquoi on doit souvent se tourner vers un logiciel statistique (la plupart du temps payant) comme par exemple *Statbox* qui s'intégre sous forme de menu à *Excel*. Les autres logiciels statistiques payants importants sont *Sas*, *Statistica* et *Spss*. Par contre l'excellent logiciel statistique R est gratuit.

Pour continuer la description statistique, il faudrait tracer des graphiques comme la courbe des valeurs, son diagramme en "tige et feuilles", sa "boite à moustaches". Là encore *Excel* ne dispose d'aucune fonction de base, contrairement aux logiciels cités.

Valeurs triées des longueurs (extrait)

11	13	16	17	17	17	17	21
21	21	25	27	29	29	30	30
30	30	30	31	31	31	32	32
32	32	32	32	33	33	34	34
34	35	35	35	35	35	36	36
...							
460	464	467	475	475	477	478	478
478	478	490	499	501	503	504	513
538	543	548	549	558	583	605	656
659	660	677	720	821	1045	1245	

Diagramme en "tige et feuilles" des longueurs

```
   0 | 1122222222333333333333333333333334444444444444444444444444444444444444444
     | 5555555555555555555566666666666666666666677777777777777788888888899999999
 100 | 0000000000000111111111111111111111111111111222222222222222333333333333333
     | 3334444444445555555555566666666666666667777778888888888888999999999999999
 200 | 00000000001111111111122222222333333334444455566667777888889
 300 | 00011111233333344455555677788
 400 | 01123333445567888888889
 500 | 00001445568
 600 | 16668
 700 | 2
 800 | 2
 900 |
1000 | 5
1100 |
1200 | 5
```

Tracé des longueurs

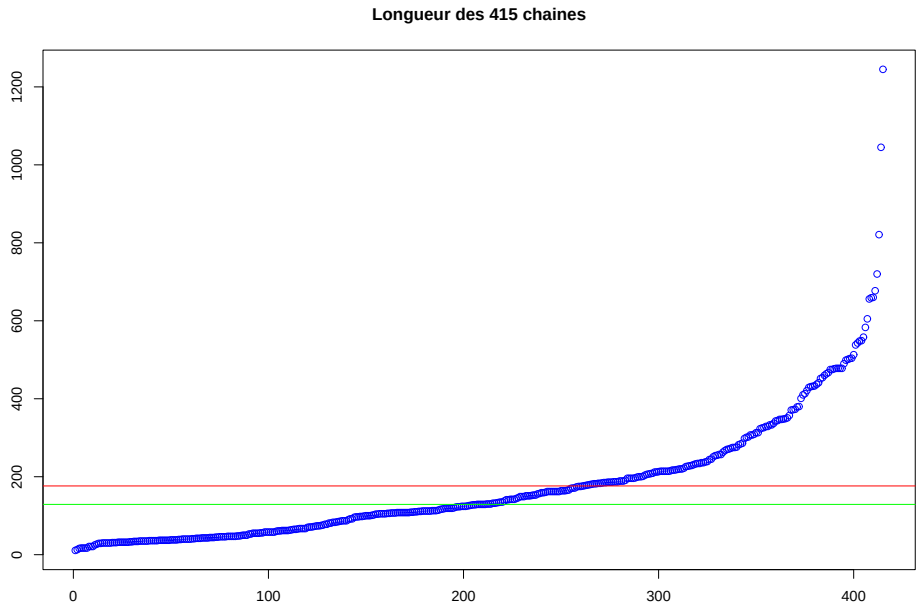
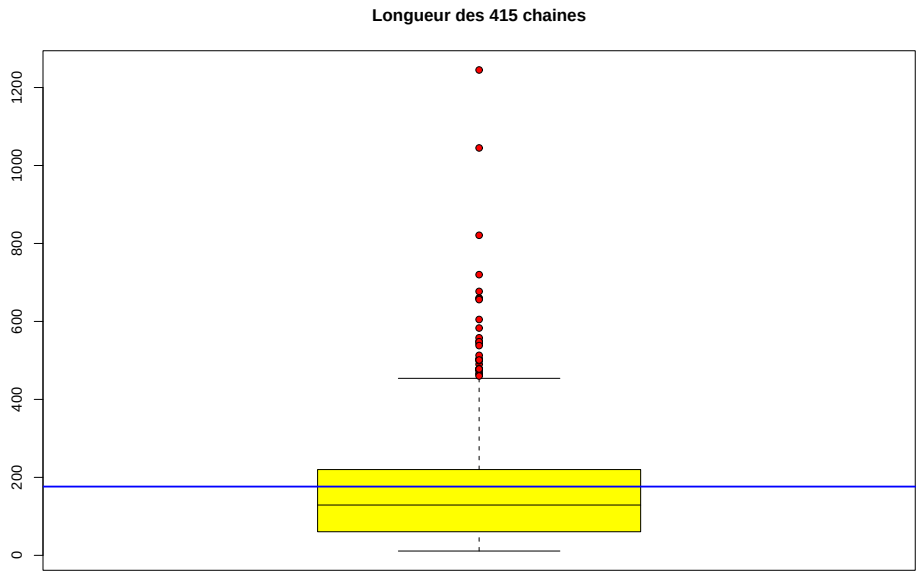


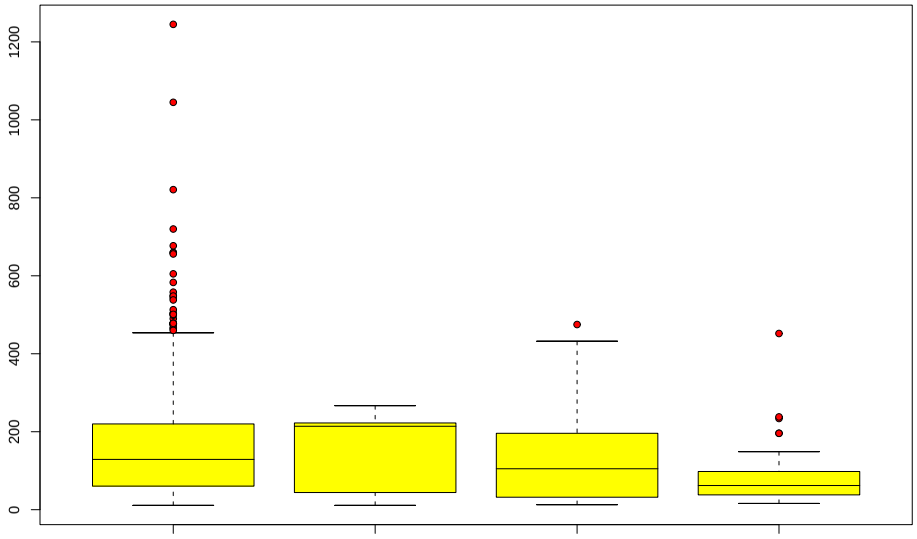
Diagramme en "boxplot" des longueurs



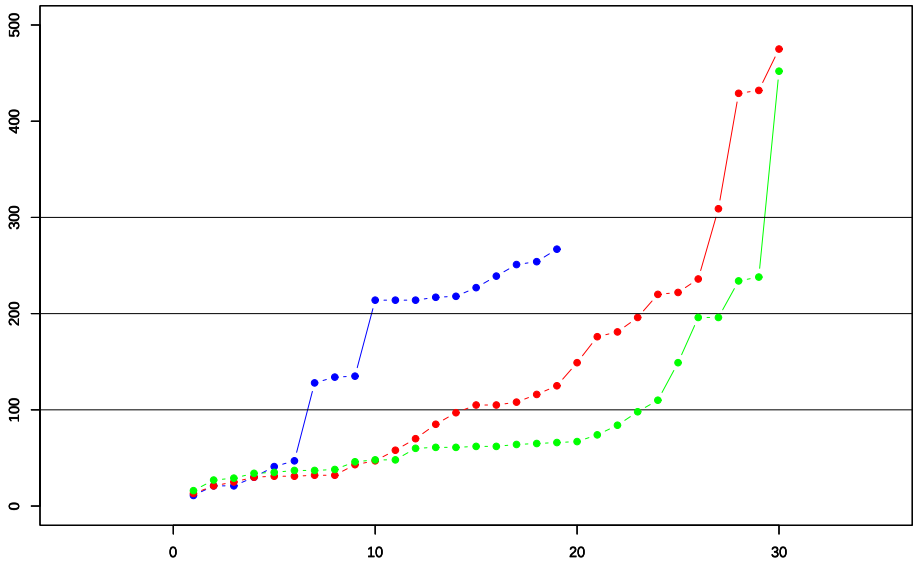
Les résultats pour les trois sous-populations sont assez différents puisqu'on trouve des chaines en général plus courtes, notamment pour la population J4, ce qui est "normal" car ce sont des toxines.

```
Sous-population CHP1

       Taille                    19    chaines        Tige et feuilles

       Moyenne                  152    aa               0 | 122345
       Ecart-type                92    aa             100 | 334
       Coef. de variation        61    %              200 | 1112234557
       1er Quartile              44    aa             300 |
       Médiane                  214    aa             400 |
       3eme Quartile            223    aa
       Minimum                   11    aa
       Maximum                  267    aa

Sous-population CHP2

       Taille                    30    chaines        Tige et feuilles

       Moyenne                  139    aa               0 | 1233333345679
       Ecart-type               126    aa             100 | 011123588
       Coef. de variation        90    %              200 | 0224
       1er Quartile              35    aa             300 | 1
       Médiane                  105    aa             400 | 338
       3eme Quartile            192    aa
       Minimum                   13    aa
       Maximum                  475    aa

Sous-population CHP4

       Taille                    30    chaines        Tige et feuilles

       Moyenne                   93    aa               0 | 2333444455566666677778
       Ecart-type                89    aa             100 | 015
       Coef. de variation        96    %              200 | 0034
       1er Quartile              40    aa             300 |
       Médiane                   62    aa             400 | 5
       3eme Quartile             95    aa
       Minimum                   16    aa
       Maximum                  452    aa
```

Boites à moustaches pour CHP et ses trois sous-séries



Tracés des longueurs pour les 3 séries



On peut se demander si nos longueurs moyennes reflètent les "vraies" longueurs de protéines. Voici quelques extraits de réponses trouvées sur le net :

http ://en.wikipedia.org/wiki/Protein_structure

A certain number of residues is necessary to perform a particular biochemical function, and around 40-50 residues appears to be the lower limit for a functional domain size. Protein sizes range from this lower limit to several thousand residues in multi-functional proteins. However, the current estimate for the average protein length is around 300 residues.

http ://www.pubmedcentral.nih.gov/articlerender.fcgi ?tool=pubmed&pubmedid=12520037

Analysis of full-length proteins (fragments excluded) : (a) Homo sapiens average protein length : 469 amino acid residues ; size range : 3-34350 amino acid residues. (b) Mus musculus average protein length : 416 amino acid residues ; size range : 10-7389 amino acid residues.

http ://www.biocomp.unibo.it/biowulf/W35report.htm

The released used contains 705,002 sequences ; corresponding to 222,117,092 total residues for an entry average length of 315 residues.

http ://cubic.bioc.columbia.edu/papers/2002_target/paper.html

The average protein length in PDB is clearly lower than the average length (300) of the proteins found in entirely sequenced proteomes.

http ://140.115.156.84/statistic/pld.htm

The average protein length of whole data is 366 aa for 125744 proteins (Swiss-Prot).

http ://www.web.virginia.edu/Heidi/home.htm

EBook : Biochemistry Garret & Grisham, chap. 5.

Polypeptide chains of proteins range in length from about 100 amino acids to 1800 (the number found in each of the two polypeptide chains of myosin, the contractile protein of muscle). However, titin, another muscle protein, has nearly 27000 amino acid residues and a molecular weight of 2.8 $\times 10^6$. The average molecular weight of polypeptide chains in eukaryotic cells is about 31700, corresponding to about 270 amino acid residues.

Comparaisons de moyennes

Nous venons de dire que les résultats pour les sous-populations indiquent que nous avons affaire à des chaines plus courtes. Mais comment le prouver ? Toute simplement en utilisant les tests de comparaison de moyennes fournis par les statistiques théoriques. *Excel* ne dispose pas de fonction pour le faire mais *R* dispose de la fonction *t.test*. Elle est incluse dans notre fonction *compMoyData* qui a l'avantage de fournir des résultats en français...

Voici ce que donne la comparaison des longueurs de `chp.lng` et de `j4.lng` auquel nous avons adjoint un test de comparaison de variances pour être complet :

```
> compMoyData(" CHP et J4",lng,lng4)

    Variable    nbVal      Moyenne    Variance     Ecart-type     Cdv
      lng        415      176.342    25519.081      159.747        91  %
      lng4        30       93.133     8267.844       90.928        98  %
    Différence réduite :  4.5321 ; au seuil de 5 % soit 1.96,
    on peut rejeter l'hypothèse d'égalité des moyennes.

> t.test(lng,lng4)

   Welch Two Sample t-test
      t = 4.5321, df = 43.234, p-value = 4.570e-05
      alternative hypothesis: true difference in means is not equal to 0
      95 percent confidence interval:
        46.18833 120.22934
      sample estimates:
      mean of x mean of y
      176.34217  93.13333

> var.test(lng,lng4)

   F test to compare two variances
      F = 3.0865, num df = 414, denom df = 29, p-value = 0.0005269
      alternative hypothesis: true ratio of variances is not equal to 1
      95 percent confidence interval:
      1.685930 4.976394
      sample estimates:
      ratio of variances
             3.086546
```

Pour un(e) statisticien(ne) ces résultats montrent clairement que les protéines de `j4` ont en moyenne une longueur trés différente de l'ensemble des protéines. Par contre ce n'est pas le cas pour `j1` et `j2` au vu de l'écart-réduit ou des *p-values* :

```
Comparaison CHP et j1

  différence réduite     1.0684
  p-value                0.2694

Comparaison CHP et j2

  différence réduite     1.4694
  p-value                0.1505

Comparaison CHP et j4

  différence réduite     4.5321
  p-value                4.570e-05
```

Automatisation

Excel ne disposant pas des fonctions de base pour effectuer nos calculs, il est difficile de parler d'automatisation, sauf à programmer en VBA tout ce qui manque ! On trouvera par exemple sur nos pages Web de quoi tracer des "stemleafs" et des "boxplots".

L'automatisation sous *R* se réalise sans problème : chacune des instructions écrites en interactif peut être mise dans un fichier et il suffit d'utliser la redirection d'entrée pour exécuter l'ensemble des lignes qu'on nomme "script". Si on est déja sous *R*, on utilise la fonction `source` pour charger et exécuter ces lignes.

On peut aller plus loin : il est possible d'écrire une fonction qui analyse le fichier passé en paramètre qu'on applique à nos 4 fichiers.

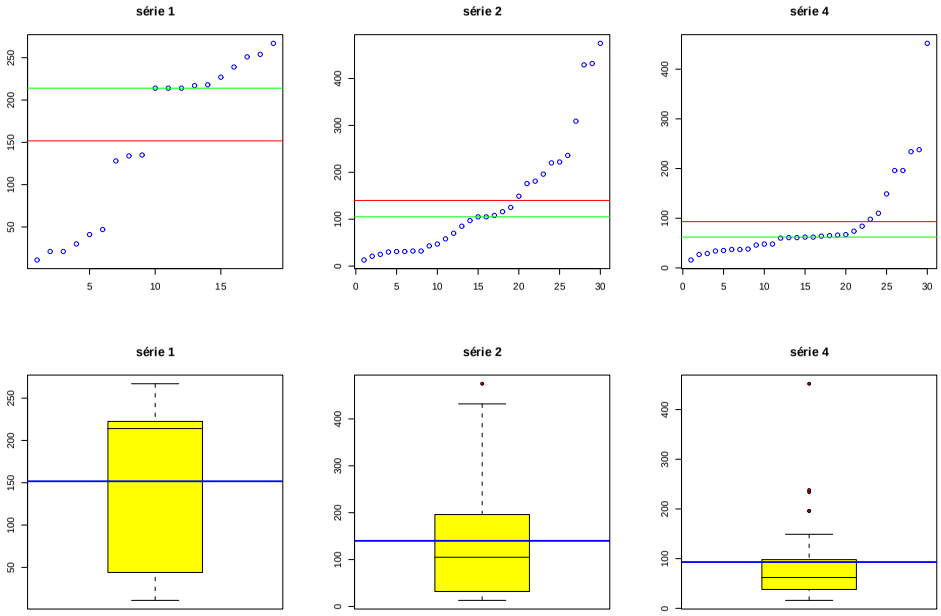
Voici une version simpliste de la fonction mise dans le fichier `flng.r`

```
analyseLongueurs <- function( nomfic ) {

    # lecture du fichier, les longueurs sont en colonne 2

    lngdata <- read.table(nomfic,header=TRUE) ;
    lng     <- lngdata[,2]

    # calculs minimaux : nombre de valeurs, moyenne et écart-type

    nbrval <- length(lng)
    moyval <- mean(lng)
    stdval <- sd(lng)

    # Affichage

    cat(" Fichier ",nomfic," soit ",nbrval," valeurs\n",
        " moyenne ",moyval," écart-type ",stdval,"\n\n")

} ; # fin de la fonction analyseLongueurs

cat("vous pouvez utiliser analyseLongueurs( nomfichier ) \n" )
```

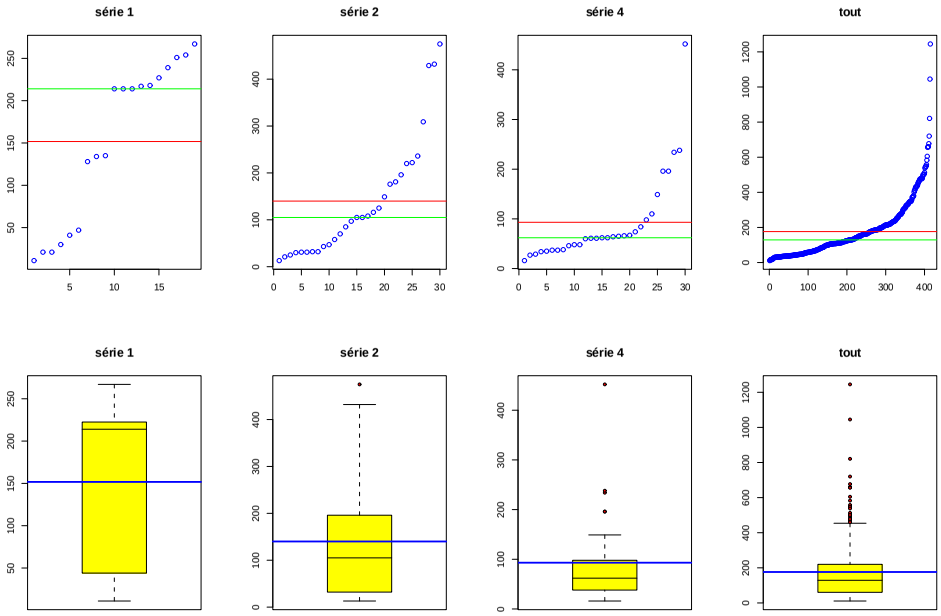
et de son utilisation :

```
> source("flng.r")

    vous pouvez utiliser analyseLongueurs( nomfichier )

> analyseLongueurs("chp.lng")

    Fichier  chp.lng  soit  415  valeurs
    moyenne  176.3422  écart-type  159.7469

> analyseLongueurs("j4.lng")

    Fichier  j4.lng  soit  29  valeurs
    moyenne  95.03448  écart-type  91.92834
```

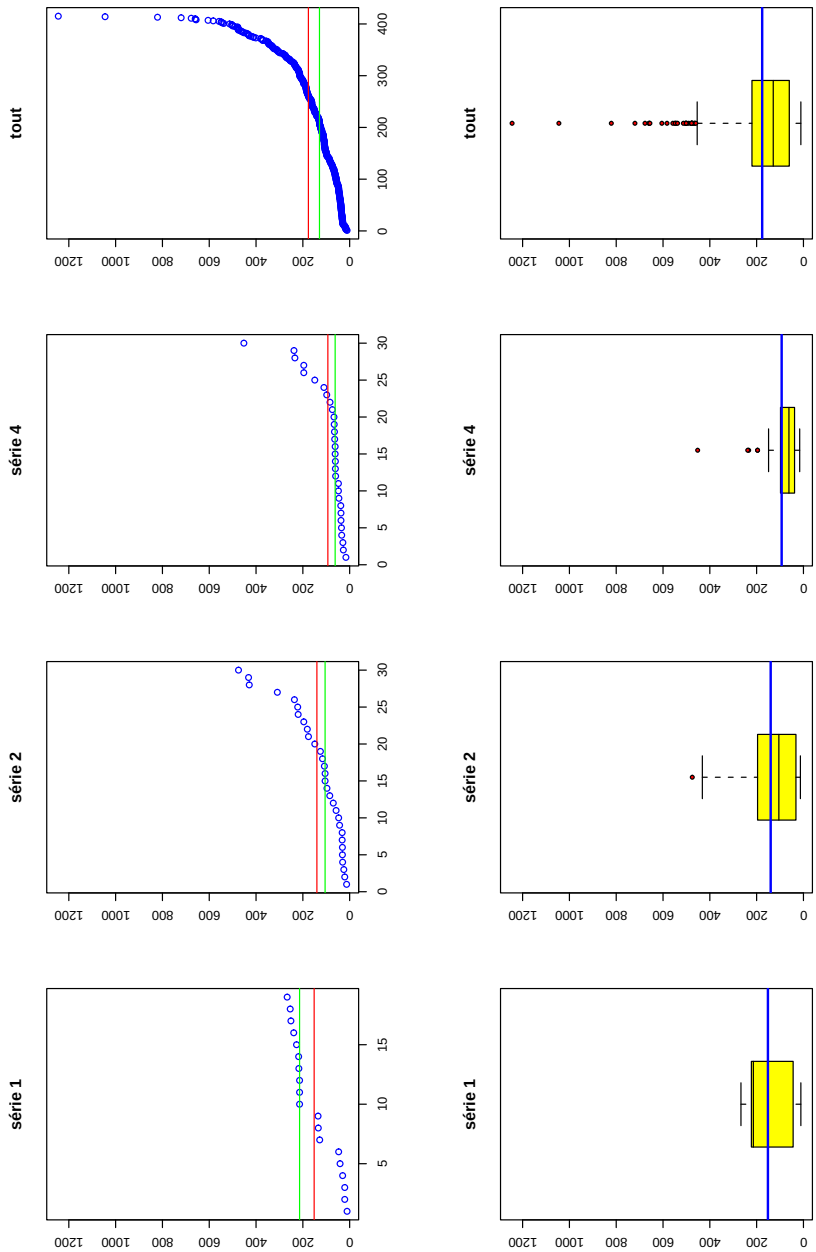
Un tracé incomplet



Un tracé complet qui induit en erreur



Le bon tracé complet



Remarques sur les calculs

En toutes rigueur, le calcul de comparaison de moyennes est faux car on ne peut comparer `CHP` et `CHP1` puisque `CHP` contient `CHP1`. On trouvera ici les vrais chiffres de comparaison (qui heureusement, différent peu des précédents) :

```
Comparaison J1 et  le reste de CHP
  différence réduite     1.1145
  p-value                0.2764

Comparaison J2 et  le reste de CHP
  différence réduite     1.5757
  p-value                0.1237

Comparaison J4 et  le reste de CHP
  différence réduite     4.8365
  p-value                1.586e-05
```

Pour comparer `CHP1`, `CHP2` et `CHP4` ensemble sans faire référence aux autres protéines de CHP, il faut en principe effectuer une ANOVA qui est une comparaison de moyennes utilisant les variances. Voici les résultats fournis par SAS (qui montrent qu'ensemble, CHP1, CHP2 et CHP3 ne peuvent pas être considérés comme différents en moyenne) :

```
THE SAS SYSTEM -- The ANOVA Procedure

Dependent Variable: LNG

   Source             DF   Somme des carrés   Carré moyen   Valeur F   Pr > F
   Model               2      50766.8698      25383.4349       2.19    0.1186
   Error              76     879766.1175      11575.8700
   Corrected Total    78     930532.9873


   R-carré     Coeff Var    Racine MSE    LNG Moyenne
   0.054557      86.06426      107.5912       125.0127
```

Il n'est pas surprenant que CHP soit différent de `CHP1`, `CHP2` et `CHP4` car ces trois sous-ensembles de données ne représentent qu'une faible partie de l'ensemble des données :

Sous-ensemble	Protéines	% dans CHP
CHP1	19	4.6
CHP2	30	7.2
CHP3	30	7.2
CHP	415	100

Il resterait bien d'autre questions à se poser sur le fichier `CHP.LNG` comme par exemple :

- *combien y a-t-il de chaines en moyennes pour une protéine ?*
- *le nombre de chaines par protéine est-il différent pour les différentes groupes (en moyenne) ?*

Les réponses à ces questions sont moins faciles à calculer que précédemment car il s'agit de calculs pondérés qu'on doit effectuer à partir des valeurs

Chaines	Protéines
1	367
2	18
3	1
4	1
5	1

qui fournissent un total de 415 chaines pour 388 protéines ce qui donne donc une moyenne de 1.07 chaines par protéines avec un écart-type de 0.34 soit le très faible coefficient de variation de 32 % (la médiane étant bien sur 1.0).

Réponses à la série de questions 2

Les variables statistiques `LNG` et `CNT*` sont des variables quantitatives (QT) pour laquelle on doit calculer taille, moyenne, écart-type etc. en dimension 1 (on parle d'*analyse séparée*). En dimension 2 (*"analyse conjointe"*), il faut calculer la matrice des coefficients de corrélation linéaire, les formules linéaires pour les meilleurs corrélations...

Voici une partie des résultats (on pourra consulter `chpaa.xls` et `chpaa.r` pour le détail des calculs), `chpaa.txt` pour l'ensemble des résultats.

`Résultats par cdv décroissant`

Nom	Num	Taille	Moyenne	Ecart-type	Cdv	Minimum	Maximum
22	X	415	0.106	0.693	653.68	0.000	10.000
10	H	415	3.882	4.883	125.78	0.000	44.000
19	W	415	2.918	3.663	125.54	0.000	25.000
14	M	415	3.101	3.822	123.26	0.000	29.000
15	F	415	6.882	7.819	113.62	0.000	50.000
3	R	415	7.701	8.647	112.28	0.000	68.000
20	Y	415	7.265	8.005	110.18	0.000	47.000
16	P	415	8.863	9.646	108.84	0.000	89.000
11	I	415	8.246	8.712	105.66	0.000	57.000
2	A	415	13.352	13.911	104.19	0.000	74.000
9	G	415	14.181	14.730	103.88	0.000	117.000
5	D	415	10.178	10.560	103.75	0.000	74.000
8	E	415	9.723	10.035	103.21	0.000	79.000
12	L	415	13.800	14.046	101.78	0.000	103.000
21	V	415	11.672	11.806	101.15	0.000	83.000
4	N	415	8.641	8.645	100.05	0.000	49.000
7	Q	415	6.923	6.834	98.71	0.000	57.000
18	T	415	11.388	11.233	98.64	0.000	86.000
17	S	415	12.795	12.316	96.26	0.000	92.000
13	K	415	9.595	8.960	93.38	0.000	71.000
1	LNG	415	176.342	159.554	90.48	11.000	1245.000
6	C	415	5.130	4.253	82.89	1.000	50.000

Résultats par moyenne décroissante

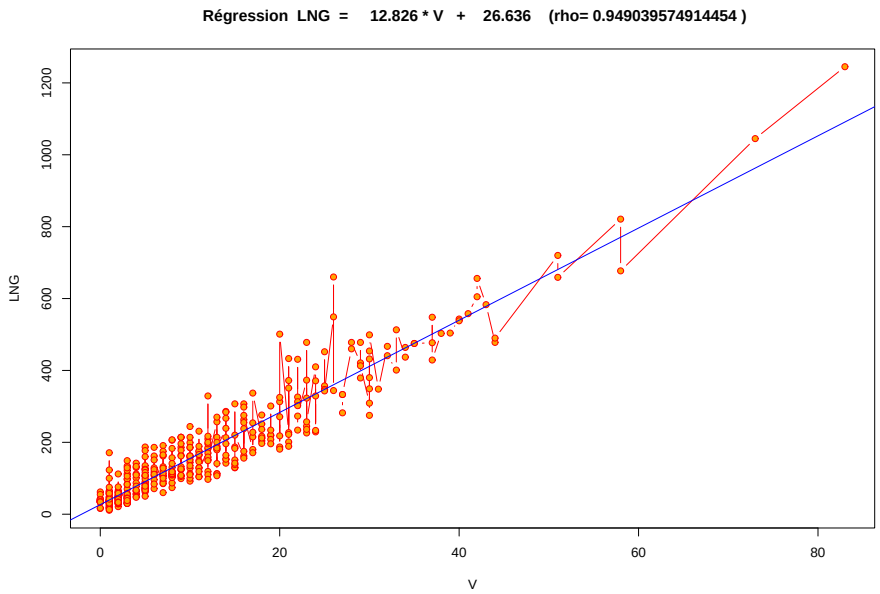
Nom	Num	Taille	Moyenne	Ecart-type	Cdv	Minimum	Maximum
1	LNG	415	176.342	159.554	90.48	11.000	1245.000
9	G	415	14.181	14.730	103.88	0.000	117.000
12	L	415	13.800	14.046	101.78	0.000	103.000
2	A	415	13.352	13.911	104.19	0.000	74.000
17	S	415	12.795	12.316	96.26	0.000	92.000
21	V	415	11.672	11.806	101.15	0.000	83.000
18	T	415	11.388	11.233	98.64	0.000	86.000
5	D	415	10.178	10.560	103.75	0.000	74.000
8	E	415	9.723	10.035	103.21	0.000	79.000
13	K	415	9.595	8.960	93.38	0.000	71.000
16	P	415	8.863	9.646	108.84	0.000	89.000
4	N	415	8.641	8.645	100.05	0.000	49.000
11	I	415	8.246	8.712	105.66	0.000	57.000
3	R	415	7.701	8.647	112.28	0.000	68.000
20	Y	415	7.265	8.005	110.18	0.000	47.000
7	Q	415	6.923	6.834	98.71	0.000	57.000
15	F	415	6.882	7.819	113.62	0.000	50.000
6	C	415	5.130	4.253	82.89	1.000	50.000
10	H	415	3.882	4.883	125.78	0.000	44.000
14	M	415	3.101	3.822	123.26	0.000	29.000
19	W	415	2.918	3.663	125.54	0.000	25.000
22	X	415	0.106	0.693	653.68	0.000	10.000

Matrice des corrélations (extrait)

	LNG	A	R	N	D	C	Q	E	G	...	X
LNG	1.000										
A	0.881	1.000									
R	0.867	0.763	1.000								
N	0.852	0.707	0.653	1.000							
D	0.943	0.829	0.803	0.825	1.000						
C	0.325	0.169	0.317	0.261	0.288	1.000					
Q	0.860	0.750	0.732	0.739	0.791	0.276	1.000				
E	0.855	0.712	0.842	0.624	0.801	0.301	0.676	1.000			
G	0.926	0.859	0.746	0.797	0.875	0.357	0.778	0.705	1.000		
...											
X	0.155	0.155	0.168	0.098	0.163	0.099	0.085	0.132	0.181	...	1.000

```
Meilleure corrélation  0.9490396  pour  V  et  LNG

Formules  LNG  =     12.826 * V   +    26.636
      et  V  =      0.070 * LNG       -0.711

Coefficients de corrélation par ordre décroissant (extrait)

    0.949  pour         21  et      1  soit  V        LNG
    0.943  pour          5  et      1  soit  D        LNG
    0.926  pour          9  et      1  soit  G        LNG
...
    0.883  pour         21  et      9  soit  V        G
    0.881  pour          2  et      1  soit  A        LNG
...
    0.209  pour         19  et      6  soit  W        C
    0.204  pour         14  et      6  soit  M        C
    0.196  pour         13  et      6  soit  K        C
...
    0.069  pour         22  et     14  soit  X        M
    0.066  pour         22  et     17  soit  X        S
```

Tracé de la meilleure corrélation



Globalement, on peut dire la distribution de chaque acide aminé varie relativement peu puisque tous les coefficients de variation se situent entre 84 et 116 %. La colonne X qui correspond à un acide aminé indéterminé est par contre très différente des autres colonnes puisqu'elle contient très peu de valeurs non nulles :

Valeurs de X	0	1	2	3	5	10	14
Occurences	393	11	2	3	4	1	1

Les acides aminés ne sont pas tous présents en même quantité : on trouve très souvent la leucine (L), l'alanine (A) et la glycine (G) alors qu'il y a assez peu (4 à 5 fois moins) de cystéine (C), histidine (H) méthionine, (M) tryptophane (W).

On note aussi un "effet taille" qui se traduit par la dépendance linéaire entre la longueur de la chaine et un certain nombre d'acides aminés, à savoir par ordre décroissant de liaison :

0.949	V	LNG
0.943	D	LNG
0.926	G	LNG
0.920	P	LNG
0.919	L	LNG
0.907	F	LNG
0.903	T	LNG
0.896	I	LNG
0.895	Y	LNG
0.883	V	G
0.881	A	LNG

A l'opposé, la cystéine (C) n'est pratiquement pas liée de façon linéaire à la longueur ($\rho = 0.366$).

Il n'y a pas non plus de liaison linéaire négative fortement marquée entre la longueur et un acide aminé B qui se serait interprétée *"plus la chaine est grande, moins il y a d'acide aminé B"*.

Etude de la sous-population 1 (inter et intra)

L'analyse des coefficients de variation montre que la distribution par acide aminé varie peu. On note ici que `X` est constant et vaut 0. Les `aa` les plus fréquents sont ici S, G, T, L, V, A et les moins fréquents H, W, M. De nombreux acides aminés croissent linéairement en même temps que la longueur, dont notamment (et dans cet ordre) G, L, D, V.

Etude de la sous-population 2 (inter sans intra)

Un coefficient de variation est nettement plus important que les autres pour cette population, à savoir celui de W. A un degré un peu moindre, les `aa` Y, P, H varient aussi un peu plus que les autres `aa`. Il y a encore de nombreuses relations linéaires entre des `aa` et la longueur dont principalement V, R, D, G.

Etude de la sous-population 4 (intra sans inter)

C'est maintenant au tour de I, L, M, N de présenter des forts coefficients de variation, les `aa` les plus présents étant G, K, S, C alors que H, M, W sont très peu présents. les meilleures corrélations avec la longueur ont lieu ici pour S, F, Y, L.

On notera que pour cette sous-population la cystéine semble être négativement liée à tous les autres `a` ce qui n'était absolument pas le cas pour les deux autres sous-populations :

		LNG	A	R	N	D	C	...
J1	C	0.359	0.227	0.252	0.334	0.404	1.000	...
J2	C	0.585	0.531	0.630	0.429	0.689	1.000	...
J4	C	-0.478	-0.464	-0.436	-0.384	-0.525	1.000	...

L'automatisation sous *Excel* requiert de savoir programmer en VBA ce qui ne se fait pas en 10 minutes. Par contre pour *Rstat* il n'y a qu'à changer le nom des fichiers dans l'instruction `read.table` pour passer d'une population à une autre.

En effet, pour *Excel* il y a de nombreux calculs à effectuer, trier, présenter. Nous fournissons donc un fichier-modèle nommé `ASGQT.XLT` qui contient des macros. Il suffit de recopier les données (y compris les noms des lignes et des colonnes) dans l'onglet "Données" puis de cliquer sur le bouton "Départ" dans l'onglet "Statistiques" pour voir *Excel* faire tout le travail de calcul.

Il ne reste plus alors qu'à lire, comprendre, interpréter, rédiger...

De la même façon, pour *R* nous fournissons le fichier `statgh.r` qui contient de nombreuses fonction. Ainsi pour traiter toutes ces variables qualitatives, on se contentera d'écrire :

```
allQT(maa,colnames(aa)[-1],rep("aa",22)) ;
```

si les données ont été lues et mises dans la matrice maa. On consultera les textes fournis en annexe pour voir le détail de l'exécution de ce programme.

Pour aller plus loin

Il est clair qu'au vu de ces nombreux résultats on peut être un peu "embrouillé(e)" par tous les tableaux de chiffres qui ne viennent explorer que les relations à 1 ou 2 dimensions. Pour avoir une vue plus synthétique sur nos données, il faudrait avoir recours à des statistiques multidimensionnelles qu'on nomme en France Analyse des données.

Ainsi une AFC (*Analyse Factorielle des Correspondances*) couplée à une CAH (*Classification Ascendante Hiérarchique*) devrait être profitable.

Hélas ! Qui dit "statistiques multidimensionnelles" dit "calculs vectoriels", "décomposition de l'inertie", "axes factoriels"... et il est difficile d'appréhender en quelques minutes le cadre théorique, les formules et l'utilisation des logiciels.

C'est pourquoi nous nous contenterons ici de montrer les résultats de l'AFC pour notre tableau issu du fichier `CHP4.CNT` en vous laissant découvrir par vous-mêmes ce qu'on peut voir sans calcul, en vous laissant imaginer ce qu'on ne peut pas voir directement à cause des projections planes pour des points dans des espaces de dimensions élévees...

```
****************************************
*                                      *
*  B I B L I O T H E Q U E   A D D A D *
*                                      *
* menhir 16b pour MsWindows 3.1x, 95, NT *
*                                      *
* programme:  15 Novembre 1998   93L8p  *
* exécution le: 22/ 2/2006  à: 17:27: 6  *
****************************************

A D D A D  - 89 -

ANALYSE DES CORRESPONDANCES (ANCORR)
D'APRES : YAGOLNITZER ET TABET


INS.  1 - TITRE :
          TITRE AFC CHP4AA.DBF ;

INS.  2 - PARAM (PARAMETRES GENERAUX) : NI,NJ,NF,NI2,NJ2,LECIJ,STFI,STFJ
          PARAM NI=30 NJ=21 NF=5 LECIJ=1 STFI=1 STFJ=1 ;

INS.  3 - OPTIONS : IOUT,IMPVP,IMPFI,IMPFJ,NGR
          OPTIONS IOUT=1 IMPFI=1 IMPFJ=1 NGR=5;

INS.  5 - GRAPHE (NGR DEMANDES DE GRAPHIQUES) : X,Y,GI,GJ,NCHAR,OPT,NPAGE,CADRE
          GRAPHE X=1 Y=2 GI=0 GJ=3 OPT=3 CADRE=1 ;
          GRAPHE X=1 Y=2 GI=3 GJ=0 OPT=3 CADRE=1 ;
          GRAPHE X=1 Y=2 GI=3 GJ=3 OPT=3 CADRE=1 ;
          GRAPHE X=1 Y=3 GI=0 GJ=3 OPT=3 CADRE=1 ;
          GRAPHE X=2 Y=3 GI=0 GJ=3 OPT=3 CADRE=1 ;

INS.  6 - LISTE (LECTURE DU TABLEAU DES DONNEES - A,F) :
          LISTE IDEN(1,4)  LNG (7,8) CNTA(15,5) CNTR(20,5) CNTN(25,5) CNTD(30,5)
           CNTC(35,5) CNTQ(40,5) CNTE(45,5) CNTG(50,5) CNTH(55,5) CNTI(60,5) CNTL(65,
           CNTK(70,5) CNTM(75,5) CNTF(80,5) CNTP(85,5) CNTS(90,5) CNTT(95,5) CNTW(100
           CNTY(105,5) CNTV(110,5) ;
```

LES POIDS DES LIGNES ET DES COLONNES SONT MULTIPLIES PAR 10 ** -1

NOMJ(J)	LNG	CNTA	CNTR	CNTN	CNTD	CNTC	CNTQ	CNTE	CNTG	CNTH	CNTI	
PJ(J)	279	16	14	18	15	19	9	13	23	6	14	559
1N8M	38	2	4	2	1	8	1	1	4	0	3	8
1DL0	37	5	2	1	2	8	0	2	3	0	1	7
1QUZ	34	2	4	2	2	8	0	1	2	0	0	7
5EBX	62	0	3	4	1	8	4	4	5	1	4	12
2SN3	65	3	0	3	2	8	1	6	9	0	0	13
1AOM	16	0	1	3	2	4	0	0	1	0	0	3
1A8D	452	18	16	49	34	4	7	17	26	7	41	90
1ABT	74	5	3	2	2	10	1	4	4	2	2	15
1ACW	29	2	0	1	3	6	2	3	0	1	1	6
2SH1	48	6	2	1	5	6	0	2	4	0	3	10
1NBT	66	2	4	4	3	10	5	1	3	1	4	13
1FFJ	60	3	1	3	2	8	0	0	2	1	1	12
1G6M	62	0	5	7	2	8	3	4	8	1	2	12
1JE9	61	1	4	6	2	8	3	3	5	2	2	12
1BMR	67	3	1	3	3	8	1	4	11	4	3	13
1NOR	61	0	4	6	2	8	3	3	5	2	2	12
1SIS	35	1	2	3	2	8	1	0	4	0	0	7
1ANS	27	0	1	1	0	6	1	1	5	0	0	5
1AHO	64	3	3	4	4	8	1	3	7	2	1	13
1AQZ	149	7	6	10	11	4	6	4	14	8	5	30
1ATX	46	3	2	4	1	6	1	1	8	0	3	9
1AXH	37	0	1	5	1	6	2	3	2	0	1	7
1PRT	234	27	22	13	7	2	7	18	17	5	8	47
1PRT	196	12	12	5	8	6	7	8	21	4	11	39
1PRT	196	21	13	4	8	6	9	7	20	2	13	39
1PRT	110	9	6	1	4	4	4	5	7	1	2	22
1PRT	98	8	3	3	5	4	3	6	7	3	3	20
1F53	84	2	6	7	9	2	2	2	9	3	8	17
3SEB	238	5	5	20	24	2	8	12	9	5	9	48
1EHS	48	6	2	1	2	4	3	2	6	1	2	10

LES POIDS DES LIGNES ET DES COLONNES SONT MULTIPLIES PAR 10 ** -1

NOMJ(J)	CNTL	CNTK	CNTM	CNTF	CNTP	CNTS	CNTT	CNTW	CNTY	CNTV	
PJ(J)	17	19	5	8	15	19	18	4	16	15	559
1N8M	0	4	0	0	2	2	2	0	2	0	8
1DL0	0	2	0	0	4	3	2	0	1	1	7
1QUZ	0	5	1	0	3	1	2	0	1	0	7
5EBX	1	4	0	2	4	8	5	1	1	2	12
2SN3	5	8	0	1	4	4	3	1	6	1	13
1AOM	0	0	1	0	2	1	0	0	1	0	3
1A8D	44	33	7	17	14	35	21	9	28	25	90
1ABT	2	6	1	1	8	6	7	1	2	5	15
1ACW	0	3	0	0	2	2	1	0	0	2	6
2SH1	2	5	0	0	2	2	4	1	2	1	10
1NBT	4	2	0	3	5	6	6	0	1	2	13
1FFJ	6	10	2	2	5	3	2	0	2	7	12
1G6M	1	4	0	0	2	4	7	1	2	1	12
1JE9	2	5	0	0	3	4	6	2	1	2	12
1BMR	3	4	0	2	3	4	2	1	2	5	13
1NOR	2	5	0	0	4	4	6	2	1	2	12
1SIS	1	3	3	2	3	0	2	0	0	0	7
1ANS	0	1	0	0	4	2	0	2	2	1	5
1AHO	2	5	0	1	3	2	3	1	7	4	13
1AQZ	9	16	1	6	12	9	8	3	7	3	30
1ATX	1	2	1	1	2	4	2	2	1	1	9
1AXH	0	2	0	1	4	4	3	0	1	1	7
1PRT	8	0	5	7	9	22	17	2	19	19	47
1PRT	15	6	3	5	8	16	19	2	16	12	39
1PRT	16	5	3	5	9	10	16	1	19	9	39
1PRT	10	8	8	5	11	6	5	0	2	12	22
1PRT	15	5	2	5	5	6	7	1	4	3	20
1F53	3	3	0	2	4	6	2	4	4	6	17
3SEB	16	32	8	12	6	14	13	1	21	16	48
1EHS	2	6	1	2	0	3	2	0	1	2	10

```
 LES VALEURS  PROPRES       VAL(1)=   1.00000
-------------------------------------------------------------------------------
!NUM ! VAL PROPRE ! POURC.! CUMUL !VARIAT.!*! HISTOGRAMME DES VALEURS PROPRES
-------------------------------------------------------------------------------
!  2 !     .05162 ! 30.937! 30.937!   .000!*!***************!***************!
!  3 !     .02754 ! 16.506! 47.443! 14.431!*!***************!*
!  4 !     .02109 ! 12.640! 60.082!  3.866!*!************
!  5 !     .01031 !  6.181! 66.264!  6.459!*!******
!  6 !     .00898 !  5.380! 71.643!   .801!*!*****
!  7 !     .00854 !  5.116! 76.759!   .264!*!*****
!  8 !     .00766 !  4.593! 81.352!   .522!*!****
!  9 !     .00640 !  3.838! 85.191!   .755!*!****
! 10 !     .00555 !  3.328! 88.518!   .511!*!***
! 11 !     .00394 !  2.361! 90.879!   .967!*!**
! 12 !     .00364 !  2.182! 93.061!   .178!*!**
! 13 !     .00334 !  2.001! 95.062!   .182!*!**
! 14 !     .00260 !  1.560! 96.622!   .441!*!**
! 15 !     .00198 !  1.187! 97.809!   .373!*!*
! 16 !     .00126 !   .755! 98.564!   .431!*!*
! 17 !     .00093 !   .556! 99.121!   .199!*!*
! 18 !     .00077 !   .461! 99.582!   .095!*!
! 19 !     .00038 !   .230! 99.812!   .231!*!
! 20 !     .00031 !   .188!100.000!   .042!*!
! 21 !     .00000 !   .000!100.000!   .188!*!
-------------------------------------------------------------------
    ! I1 ! QLT POID INR!  1#F  COR CTR!  2#F  COR CTR!  3#F  COR CTR!
-------------------------------------------------------------------
   1!1N8M! 741   14  26! -448  631  53!  -35    4   1!  -69   15   3!
   2!1DL0! 800   13  29! -470  597  57! -197  105  19!  100   27   6!
   3!1QUZ! 813   12  34! -553  652  72!   54    6   1!  199   84  23!
   4!5EBX! 787   22  22! -249  366  27!   24    3   0! -155  141  25!
   5!2SN3! 587   23  28! -198  195  18!   47   11   2!   52   13   3!
   6!1A0M! 741    6  27! -512  339  29!  235   71  11!   57    4   1!
   7!1A8D! 932  162  94!  229  538 164!  154  245 140!  -85   74  55!
   8!1ABT! 653   26  21! -254  483  33!  -75   42   5!  105   83  14!
   9!1ACW! 495   10  26! -386  358  30!   93   21   3!  127   39   8!
  10!2SH1! 541   17  20! -159  133   8!  -45   11   1!   29    5   1!
  11!1NBT! 502   24  25! -227  289  24!  -15    1   0!  -46   12   2!
  12!1FFJ! 717   21  35! -122   55   6!  194  139  29!  373  513 142!
  13!1G6M! 738   22  26! -284  408  35!   24    3   0! -228  262  54!
  14!1JE9! 809   22  19! -266  487  30!   72   36   4! -165  187  28!
  15!1BMR! 417   24  23! -151  145  11!    6    0   0!  -56   20   4!
  16!1NOR! 849   22  21! -282  496  34!   96   57   7! -166  172  29!
```

17	1SIS	758	13	40	-466	403	53	131	32	8	321	191	61
18	1ANS	606	10	38	-591	529	65	-39	2	1	-164	41	12
19	1AHO	454	23	15	-155	224	11	14	2	0	-23	5	1
20	1AQZ	410	53	29	41	18	2	114	145	25	-19	4	1
21	1ATX	611	16	19	-235	283	18	-38	8	1	-165	139	21
22	1AXH	776	13	23	-369	464	35	100	34	5	-91	28	5
23	1PRT	831	84	81	159	158	41	-310	600	293	-33	7	4
24	1PRT	726	70	26	114	208	18	-169	457	73	-48	37	8
25	1PRT	799	70	45	133	164	24	-247	569	156	-3	0	0
26	1PRT	815	39	56	92	36	6	-116	57	19	391	647	285
27	1PRT	417	35	26	142	166	14	-57	27	4	118	114	23
28	1F53	595	30	32	109	67	7	93	50	10	-270	415	104
29	3SEB	820	85	77	217	313	77	238	379	176	124	102	62
30	1EHS	310	17	17	-48	14	1	-92	51	5	133	106	14
			1000			1000			1000			1000	

	J1	QLT	POID	INR	1#F	COR	CTR	2#F	COR	CTR	3#F	COR	CTR
1	LNG	0	500	0	0	0	0	0	0	0	0	0	0
2	CNTA	844	28	71	133	41	10	-512	615	266	198	92	52
3	CNTR	686	25	37	-57	13	2	-320	407	92	-125	62	18
4	CNTN	858	32	56	32	4	1	373	473	161	-245	204	91
5	CNTD	615	28	39	190	152	19	310	403	96	8	0	0
6	CNTC	978	34	204	-988	962	636	47	2	3	80	6	10
7	CNTQ	529	15	28	-56	10	1	-115	43	7	-79	20	5
8	CNTE	350	23	26	6	0	0	-137	98	15	-32	5	1
9	CNTG	607	41	43	-137	106	15	-145	120	31	-152	130	45
10	CNTH	443	10	29	121	30	3	99	20	4	-131	36	8
11	CNTI	690	24	46	281	250	37	80	20	6	-309	302	109
12	CNTL	617	30	61	407	496	98	84	21	8	148	66	32
13	CNTK	855	35	73	-84	20	5	436	543	240	254	184	106
14	CNTM	833	8	57	252	57	10	15	0	0	800	571	255
15	CNTF	671	15	25	349	424	35	130	58	9	203	144	29
16	CNTP	523	26	41	-316	385	51	-36	5	1	166	106	34
17	CNTS	539	35	17	37	17	1	-58	42	4	-135	224	30
18	CNTT	449	31	26	-31	7	1	-167	201	32	-69	35	7
19	CNTW	546	7	35	-63	5	1	200	46	10	-619	441	124
20	CNTY	464	28	44	288	315	45	-99	37	10	-59	13	5
21	CNTV	442	26	40	254	250	32	-77	23	6	179	124	39
			1000			1000			1000			1000	

```
AXE HORIZONTAL( 1)--AXE VERTICAL( 2)--TITRE:AFC CHP4AA.DBF

NOMBRE DE POINTS :   21
==ECHELLE :  4 CARACTERE(S) =    .077    1 LIGNE =    .032
+---------------------------------------------CNTK+-----------------------+ 0 01
!                                               !   CNTN                  ! 0 01
!                                               !                         ! 0 01
!                                               !           CNTD          ! 0 01
!                                               !                         ! 0 01
!                                               !                         ! 0 01
!                                               !                         ! 0 01
!                                           CNTW!                         ! 0 01
!                                               !                         ! 0 01
!                                               !                   CNTF! 0 01
!                                               !       CNTH        CNTL! 0 01
!                                               !               CNTI    ! 0 01
CNTC                                            !                         ! 0 01
+-----------------------------------------------LNG --------CNTM--------+ 0 01
!                               CNTP            !                         ! 0 01
!                                               !   CNTS        CNTV    ! 0 01
!                                               !               CNTY    ! 0 01
!                                               CNTQ                    ! 1 01
!                                       CNTG    CNTT                    ! 0 01
!                                               !                         ! 0 01
!                                               !                         ! 0 01
!                                               !                         ! 0 01
!                                               !                         ! 0 01
!                                               CNTR                    ! 0 01
!                                               !                         ! 0 01
!                                               !                         ! 0 01
!                                               !                         ! 0 01
!                                               !                         ! 0 01
!                                               !                         ! 0 01
!                                               !       CNTA              ! 0 01
+-----------------------------------------------+-----------------------+ 0 01

NOMBRE DE POINTS SUPERPOSES :   1

 CNTE(CNTQ)
```

```
AXE HORIZONTAL( 1)--AXE VERTICAL( 2)--TITRE:AFC CHP4AA.DBF

NOMBRE DE POINTS :   30
==ECHELLE :  4 CARACTERE(S) =    .046    1 LIGNE =    .019
+---1A0M---------------------------------------+-------------------3SEB+ 0 01
!                                              !                      ! 0 01
!                                     1FFJ     !                      ! 0 01
!                                              !                      ! 0 01
!                                              !                  1A8D! 0 01
!       1SIS                                   !                      ! 0 01
!                                              !   1AQZ               ! 0 01
!               1ACW    1NOR                   !           1F53       ! 1 01
!                           1JE9               !                      ! 0 01
1QUZ                                           !                      ! 0 01
!                               2SN3           !                      ! 0 01
!                       1G6M5EBX    1AHO       !                      ! 0 01
+----------------------------------1BMR--------+----------------------+ 0 01
!                           1NBT               !                      ! 0 01
1ANS        1N8M            1ATX    2SH1       !                      ! 0 01
!                                              !               1PRT   ! 0 01
!                           1ABT               !                      ! 0 01
!                                        1EHS!                        ! 0 01
!                                              !       1PRT           ! 0 01
!                                              !                      ! 0 01
!                                              !                      ! 0 01
!                                              !           1PRT       ! 0 01
!       1DL0                                   !                      ! 0 01
!                                              !                      ! 0 01
!                                              !                      ! 0 01
!                                              !           1PRT       ! 0 01
!                                              !                      ! 0 01
!                                              !                      ! 0 01
!                                              !               1PRT   ! 0 01
+----------------------------------------------+----------------------+ 0 01

NOMBRE DE POINTS SUPERPOSES :   1

 1AXH(1ACW)
```

```
AXE HORIZONTAL( 1)--AXE VERTICAL( 2)--TITRE:AFC CHP4AA.DBF

NOMBRE DE POINTS :   51
==ECHELLE :  4 CARACTERE(S) =    .077    1 LIGNE =    .032
+---------------------------------------------CNTK+-----------------------+ 0 01
!                                                !   CNTN                ! 0 01
!                                                !                       ! 0 01
!                                                !           CNTD        ! 0 01
!                                                !                       ! 0 01
!                                                !                       ! 0 01
!                      1AOM                      !           3SEB        ! 0 01
!                                            1FFJ!                       ! 1 01
!                                                !           1A8D        ! 0 01
!                      1SIS                      !   1AQZ            CNTF! 0 01
!                          1ACW    1NOR          !       1F53        CNTL! 2 01
!                  1QUZ            1JE9          !               CNTI    ! 0 01
CNTC                               5EBX2SN3      !                       ! 1 01
+-----------------------------------1NBT1BMR----LNG --------CNTM--------+ 1 01
!                  1ANS1N8M    CNTP1ATX2SH1       !                       ! 0 01
!                                  1ABT          !   CNTS1PRT    CNTV    ! 0 01
!                                                1EHS            CNTY    ! 0 01
!                                                CNTQ1PRT                ! 1 01
!                                     CNTG       CNTT    1PRT            ! 0 01
!                      1DLO                      !                       ! 0 01
!                                                !                       ! 0 01
!                                                !       1PRT            ! 0 01
!                                                !                       ! 0 01
!                                                CNTR    1PRT            ! 0 01
!                                                !                       ! 0 01
!                                                !                       ! 0 01
!                                                !                       ! 0 01
!                                                !                       ! 0 01
!                                                !                       ! 0 01
!                                                !       CNTA            ! 0 01
+-----------------------------------------------+-----------------------+ 0 01

NOMBRE DE POINTS SUPERPOSES :   6

 CNTW(1FFJ)  1AXH(1ACW)  CNTH(1F53)  1G6M(5EBX)  1AHO(1BMR)  CNTE(CNTQ)
 AXE HORIZONTAL( 1)--AXE VERTICAL( 3)--TITRE:AFC CHP4AA.DBF
```

```
 NOMBRE DE POINTS :   21
 ==ECHELLE :  4 CARACTERE(S) =    .077    1 LIGNE =    .032
 +-----------------------------------------------+-----------CNTM--------+ 0 01
 !                                               !                       ! 0 01
 !                                               !                       ! 0 01
 !                                               !                       ! 0 01
 !                                               !                       ! 0 01
...
 !                                               !                       ! 0 01
 !                                               !                       ! 0 01
 !                                               !                       ! 0 01
 !                                               !                       ! 0 01
 !                                               !                       ! 0 01
 !                                               !                       ! 0 01
 !                                               !                       ! 0 01
 !                                           CNTK!                       ! 0 01
 !                                               !                       ! 0 01
 !                                               !       CNTA        CNTF! 0 01
 !                              CNTP             !               CNTV    ! 0 01
 !                                               !                   CNTL! 0 01
 !                                               !                       ! 0 01
 CNTC                                            !                       ! 0 01
 !                                               !                       ! 0 01
 +-----------------------------------------------LNG --------CNTD--------+ 0 01
 !                                               CNTE                    ! 0 01
 !                                               CNTT            CNTY    ! 0 01
 !                                               CNTQ                    ! 0 01
 !                                               CNTRCNTSCNTH            ! 0 01
 !                                       CNTG    !                       ! 0 01
 !                                               !                       ! 0 01
 !                                               !                       ! 0 01
 !                                               !   CNTN                ! 0 01
 !                                               !                       ! 0 01
 !                                               !               CNTI    ! 0 01
 !                                               !                       ! 0 01
 !                                               !                       ! 0 01
 !                                               !                       ! 0 01
 !                                               !                       ! 0 01
 ...
 !                                               !                       ! 0 01
 !                                           CNTW!                       ! 0 01
 +-----------------------------------------------+-----------------------+ 0 01
```

```
AXE HORIZONTAL( 2)--AXE VERTICAL( 3)--TITRE:AFC CHP4AA.DBF

NOMBRE DE POINTS :   21
==ECHELLE :  4 CARACTERE(S) =    .053    1 LIGNE =    .022
+-----------------------------------+---CNTM----------------------------+ 0 01
!                                   !                                     ! 0 01
!                                   !                                     ! 0 01
...
!                                   !                                     ! 0 01
!                                   !                                     ! 0 01
!                                   !                                     ! 0 01
!                                   !                                     ! 0 01
!                                   !                                     ! 0 01
!                                   !                                     ! 0 01
!                                   !                                     ! 0 01
!                                   !                                     ! 0 01
!                                   !                                     ! 0 01
!                                   !                                     ! 0 01
!                                   !                                     ! 0 01
!                                   !                                     ! 0 01
!                                   !                                     ! 0 01
!                                   !                                     ! 0 01
!                                   !                                     ! 0 01
!                                   !                                     ! 0 01
!                                   !                                 CNTK! 0 01
!                                   !                                     ! 0 01
!                                   !                                     ! 0 01
CNTA                                !            CNTF                     ! 0 01
!                               CNTVCNTP                                  ! 0 01
!                                   !       CNTL                          ! 0 01
!                                   !                                     ! 0 01
!                                   !                                     ! 0 01
!                                   !   CNTC                              ! 0 01
!                                   !                                     ! 0 01
!                                   !                                     ! 0 01
!                                   !                                     ! 0 01
+-----------------------------------LNG --------------------CNTD--------+ 0 01
!                           CNTE    !                                     ! 0 01
!                                   !                                     ! 0 01
!                       CNTTCNTY    !                                     ! 0 01
!                           CNTQ    !                                     ! 0 01
!                                   !                                     ! 0 01
```

```
!            CNTR                 CNTS!       CNTH                        ! 0 01
!                        CNTG        !                                   ! 0 01
!                                    !                                   ! 0 01
!                                    !                                   ! 0 01
!                                    !                                   ! 0 01
!                                    !                           CNTN    ! 0 01
!                                    !                                   ! 0 01
!                                    !                                   ! 0 01
!                                    !       CNTI                        ! 0 01
!                                    !                                   ! 0 01
!                                    !                                   ! 0 01
!                                    !                                   ! 0 01
!                                    !                                   ! 0 01
!                                    !                                   ! 0 01
!                                    !                                   ! 0 01
!                                    !                                   ! 0 01
!                                    !                                   ! 0 01
!                                    !                                   ! 0 01
!                                    !                                   ! 0 01
!                                    !                                   ! 0 01
!                                    !                                   ! 0 01
!                                    !               CNTW                ! 0 01
+------------------------------------+-----------------------------------+ 0 01
```

FIN NORMALE DU PROGRAMME ANCORR

```
****************************************
*                                      *
*  B I B L I O T H E Q U E   A D D A D *
*                                      *
* menhir 16b pour MsWindows 3.1x, 95, NT *
*                                      *
* programme:  15 Novembre 1998   93L8p  *
* exécution le: 22/ 2/2006  à: 17:25:11 *
****************************************

A D D A D  - 89 -

CLASSIFICATION ASCENDANTE HIERARCHIQUE (CAH2CO)
METHODE DES VOISINS REDUCTIBLES
AUTEUR : M.JAMBU


INS.  1 - TITRE :
          TITRE CAH (COLONNES) CHP4AA.DBF ;

INS.  2 - PARAM (PARAMETRES GENERAUX) : NI,NJ,NFSTOC,IOPT,NPLACE,LECIJ,STCAH
          PARAM NI=21 NJ=3 IOPT=1 LECIJ=1 ;

INS.  3 - OPTIONS : HISTO,DESCRI,ARBRE
          OPTIONS HISTO=1 DESCRI=1 ARBRE=1 ;

SOMME DES INDICES DE NIVEAU    .10025E+00

-------------------------------------------------------------------------------
!  J  ! I(J) ! A(J)! B(J)!T(J)!T(Q)! HISTOGRAMME DES INDICES DE NIVEAU
-------------------------------------------------------------------------------
!   41!    34!   40!    6! 342! 342!****************************************
!   40!    16!   39!   37! 159! 500!******************
!   39!    14!   36!   38! 142! 642!****************
!   38!     7!   13!   33!  75! 717!*********
!   37!     7!    2!   35!  66! 783!********
!   36!     5!   29!   34!  46! 829!*****
!   35!     4!   14!   31!  38! 867!****
!   34!     3!   27!   16!  34! 900!****
!   33!     2!    5!   32!  21! 922!**
!   32!     2!   30!   11!  21! 942!**
!   31!     1!   28!   24!  15! 957!**
!   30!     1!    4!   19!  10! 967!*
```

```
!   29!     1!    1!   25!   9! 976!*
!   28!     1!   20!   21!   8! 984!*
!   27!     1!    3!   26!   6! 990!*
!   26!     1!   23!    9!   5! 996!*
!   25!     0!   10!   17!   2! 998!*
!   24!     0!   12!   15!   1! 999!*
!   23!     0!    8!   22!   1!1000!*
!   22!     0!    7!   18!   0!1000!*
```

```
-------------------------------------------------------------------------------
!  J  ! I(J) ! A(J)! B(J)! P(J)!     DESCRIPTION DES CLASSES DE LA HIERARCHIE
-------------------------------------------------------------------------------
!   41!    34!   40!    6!   21!
-------------------------------------------------------------------------------
!   40!    16!   39!   37!   20! LNG  CNTH CNTS CNTR CNTE CNTQ CNTT CNTG CNTP
!     !      !     !     !     ! CNTK CNTD CNTN CNTW CNTI CNTA CNTM CNTY CNTV
!     !      !     !     !     ! CNTL CNTF
-------------------------------------------------------------------------------
!   39!    14!   36!   38!   14! LNG  CNTH CNTS CNTR CNTE CNTQ CNTT CNTG CNTP
!     !      !     !     !     ! CNTK CNTD CNTN CNTW CNTI
-------------------------------------------------------------------------------
!   38!     7!   13!   33!    5! CNTK CNTD CNTN CNTW CNTI
-------------------------------------------------------------------------------
!   37!     7!    2!   35!    6! CNTA CNTM CNTY CNTV CNTL CNTF
-------------------------------------------------------------------------------
!   36!     5!   29!   34!    9! LNG  CNTH CNTS CNTR CNTE CNTQ CNTT CNTG CNTP
-------------------------------------------------------------------------------
!   35!     4!   14!   31!    5! CNTM CNTY CNTV CNTL CNTF
-------------------------------------------------------------------------------
!   34!     3!   27!   16!    6! CNTR CNTE CNTQ CNTT CNTG CNTP
-------------------------------------------------------------------------------
!   33!     2!    5!   32!    4! CNTD CNTN CNTW CNTI
-------------------------------------------------------------------------------
!   32!     2!   30!   11!    3! CNTN CNTW CNTI
-------------------------------------------------------------------------------
!   31!     1!   28!   24!    4! CNTY CNTV CNTL CNTF
-------------------------------------------------------------------------------
!   30!     1!    4!   19!    2! CNTN CNTW
-------------------------------------------------------------------------------
!   29!     1!    1!   25!    3! LNG  CNTH CNTS
```

28	1	20	21	2	CNTY CNTV
27	1	3	26	5	CNTR CNTE CNTQ CNTT CNTG
26	1	23	9	4	CNTE CNTQ CNTT CNTG
25	0	10	17	2	CNTH CNTS
24	0	12	15	2	CNTL CNTF
23	0	8	22	3	CNTE CNTQ CNTT
22	0	7	18	2	CNTQ CNTT

LA REPRESENTATION DE LA CLASSIFICATION HIERARCHIQUE
("DINDON" ou DENDROGRAMME) EST SUR LA PAGE SUIVANTE

```
LNG  *------*---------------*--*-------------------------------*-
     !      !               !  !                               !
CNTH !      !               !  !                               !
     !      !               !  !                               !
CNTS -      !               !  !                               !
            !               !  !                               !
CNTR *---*---               !  !                               !
     !   !                  !  !                               !
CNTE !   !                  !  !                               !
     !   !                  !  !                               !
CNTQ !   !                  !  !                               !
     !   !                  !  !                               !
CNTT -   !                  !  !                               !
         !                  !  !                               !
CNTG -   !                  !  !                               !
         !                  !  !                               !
CNTP -----                  !  !                               !
                            !  !                               !
CNTK ------------*-----------  !                               !
                 !             !                               !
CNTD --*----------             !                               !
       !                       !                               !
CNTN ! !                       !                               !
     ! !                       !                               !
CNTW - !                       !                               !
       !                       !                               !
CNTI ---                       !                               !
                               !                               !
CNTA ----------*----------------                               !
               !                                               !
CNTM -----*-----                                               !
          !                                                    !
CNTY **----                                                    !
     !!                                                        !
CNTV -!                                                        !
      !                                                        !
CNTL *-                                                        !
     !                                                         !
CNTF -                                                         !
                                                               !
CNTC -------------------------------------------------------------
FIN NORMALE DU PROGRAMME CAH2CO
```

REPRESENTATION DE LA CLASSIFICATION HIERARCHIQUE CAH (LIGNES) CHP4AA.DBF ;

```
 1F53 *------------*-----------*--------------------------------*-
      !            !           !                                !
 1A8D -            !           !                                !
                   !           !                                !
 1AQZ -*--*--*------           !                                !
       !  !  !                 !                                !
 1PRT *-  !  !                 !                                !
      !   !  !                 !                                !
 1EHS -   !  !                 !                                !
          !  !                 !                                !
 3SEB -----  !                 !                                !
             !                 !                                !
 1PRT --------                 !                                !
                               !                                !
 1PRT -*------------------------                                !
       !                                                        !
 1PRT *-                                                        !
      !                                                         !
 1PRT -                                                         !
                                                                !
 5EBX *----------*----------------------------------------------
      !          !
 1G6M !          !
      !          !
 1JE9 !          !
      !          !
 1NOR -          !
                 !
 1NBT !          !
      !          !
 1AXH -          !
                 !
 1N8M *-*-----*---
      ! !     !
 1ANS - !     !
        !     !
 1DL0 - !     !
        !     !
 1QUZ *--     !
      !       !
 1SIS -       !
```

```
             !
1AOM -       !
             !
2SN3 *---*----
     !   !
1BMR !   !
     !   !
1AHO -   !
         !
2SH1 !   !
     !   !
1ATX -   !
         !
1ABT *-*--
     ! !
1ACW - !
       !
1FFJ ---
```

FIN NORMALE DU PROGRAMME CAH2CO

ANNEXES : programmes R

Analyse individuelle des longueurs

```
source("statgh.r")

aa   <-   read.table("j1.lng",header=FALSE,skip=2) ;
dims   <-   dim(aa) ;
nbl    <-   dims[1] ;
nbc    <-   dims[2] ;
maa    <- aa[1:nbl,2:nbc]  ;
lng    <- aa[,2]  ;

print(lng) ;
print(sort(lng)) ;

stem(lng,scale=1,width=170) ;

decritQT("Longueur série 1",lng,"aa") ;

postscript("j1lng.ps")
plotQT("Longueur ",lng) ;
dev.off()

allQT(maa,colnames(aa)[-1],"aa") ;
```

Analyse commune des longueurs

```
source("statgh.r")

# lecture des données

chplngdata <- read.table("chp.lng",header=TRUE) ;
lng        <- chplngdata[,2]
```

```
chplng1    <- read.table("chp1.lng",header=TRUE) ;
lng1       <- chplng1[,2]

chplng2    <- read.table("chp2.lng",header=TRUE) ;
lng2       <- chplng2[,2]

chplng4    <- read.table("chp4.lng",header=TRUE) ;
lng4       <- chplng4[,2]

decritQT(' Longueur des chaines polypeptidiques',lng,'aa') ;
stem(lng,scale=1,width=170)

postscript("chplng1.ps")
plotQT("Longueur des 415 chaines",lng)
dev.off()

postscript("chplng2.ps")
boxplotQT("Longueur des 415 chaines",lng)
dev.off()

decritQT(' Sous-population CHP1',lng1,'aa') ;

# tracés multiples

postscript("chplngs.ps")
par(mfrow=c(3,2))

  plotQT("série 1",lng1)
  boxplotQT(" série 1",lng1)
  plotQT("série 2",lng2)
  boxplotQT("série 2",lng2)
  plotQT("série 4",lng4)
  boxplotQT("série 4",lng4)

dev.off()
par(mfrow=c(1,1))

# tracé commun

postscript("chplngs1.ps")
boxplot(lng1,lng2,lng4,main="",col="yellow",pch=21,bg="red")
```

```
dev.off()

postscript("chplngs2.ps")
plot(sort(lng1),type="b",main="",col="blue", pch=21,
     bg="blue",xlim=c(-5,35),ylim=c(0,500),xlab="",ylab="")
abline(h=100)
par(new=TRUE)
plot(sort(lng2),type="b",main="",col="red",  pch=21,
     bg="red",xlim=c(-5,35),ylim=c(0,500),xlab="",ylab="")
abline(h=200)
par(new=TRUE)
plot(sort(lng4),type="b",main="",col="green",pch=21,
     bg="green",xlim=c(-5,35),ylim=c(0,500),xlab="",ylab="")
abline(h=300)
dev.off()
```

Analyse des acides aminés

```
source("statgh.r")

aa     <-   read.table("chp.cnt",header=TRUE) ;
colnames(aa)
       <-   c("ID","LNG" ,"A" ,"R" ,"N" ,"D" ,"C" ,"Q" ,"E" ,
              "G" ,"H"   ,"I" ,"L" ,"K" ,"M" ,"F" ,"P" ,"S" ,
              "T" ,"W" ,"Y" ,"V" ,"X") ;

dims   <-   dim(aa) ;
nbl    <-   dims[1] ;
nbc    <-   dims[2] ;
maa    <-   aa[1:nbl,2:nbc]  ;

anaQT(maa,colnames(aa)[-1]) ;
```