Fiche TD avec le logiciel R : bgsc

Structures des génomes bactériens (correction)

J.R. Lobry

Correction des exercices proposés dans le cours « Bacterial Genome structures ».

1 Nombre de chromosomes et plasmides

1.1 Lecture des données

Les données avaient été récupérées sur le site de GOLD :

```
gt <- read.table("http://www.genomesonline.org/DBs/goldtable.txt", sep = "\t", com = "", header = TRUE, quote =
comment(gt) <- date()
save(gt, file = "gt.RData")
```

```
load(url("https://pbil.univ-lyon1.fr/R/donnees/bgsc/gt.RData"))
```

Le jeu de données comporte 3658 lignes et 57 colonnes. La dernière lecture du fichier a été efectuée le : Wed Apr 9 10 :46 :29 2008. Le dossier de travail utilisé ici est : `/Users/lobry/pedadoc/cours/bgsc`.

1.2 Sélection des bactéries

Nous voulons conserver les individus ayant comme valeur "Bacteria" ou "Archaea" pour la variable SUPERKINGDOM :

```
proc <- gt[gt$SUPER == "Bacteria" | gt$SUPER == "Archaea", ]
dim(proc)
```

```
[1] 2514   57
```

Nous avons donc 2514 bactéries.

1.3 Extraction des données

Nous récupérons les colonnes contenant le nom des espèces et le nombre de chromosomes et de plasmides, puis nous conservons uniquement les individus complètement documentés :

```
chromo <- proc[ , c("SPECIES", "CHROMOSOMES", "PLASMIDS")]
chromo <- chromo[complete.cases(chromo), ]
```

Nous avons donc ici 678 données disponibles.

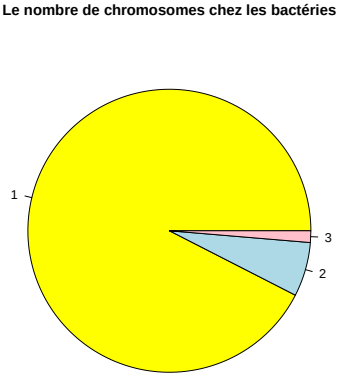


FIGURE 1 – La majorité des bactéries ne possède qu'un seul chromosome.

1.4 Les chromosomes

	Nombre d'espèces
1	627
2	42
3	9

TABLE 1 – Le nombre de chromosome chez les bactéries

Le nombre de chromosomes présents dans les espèces de notre magnifique jeu de données est indiqué dans la très belle table 1 et représenté dans la superbe figure 1.

Les espèces ayant trois chromosomes sont les suivantes : *Burkholderia ambifaria*, *Burkholderia cenocepacia*, *Burkholderia multivorans*, *Burkholderia sp. 383*, *Burkholderia vietnamiensis*, *Burkholderia xenovorans*, Les espèces ayant deux chromosomes sont les suivantes :

1.5 Les plasmides

Le nombre de plasmides présents dans les espèces de notre magnifique jeu de données est indiqué dans la très belle table 2 et représenté dans la superbe figure 2.

2 Le taux de G+C

2.1 Distribution du taux de G+C

```
tgc <- proc[, c("SPECIES", "GC...")]
tgc <- tgc[complete.cases(tgc), ]
nrow(tgc)
```

```
[1] 1362
```

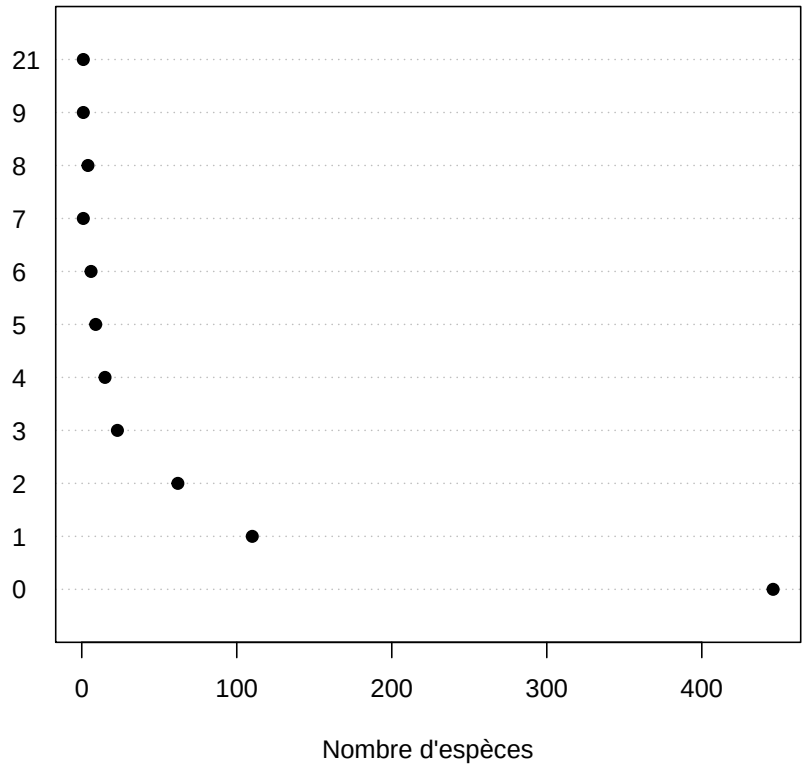


FIGURE 2 – La majorité des bactéries n'a pas de plasmide

	Nombre d'espèces
0	446
1	110
2	62
3	23
4	15
5	9
6	6
7	1
8	4
9	1
21	1

TABLE 2 – Le nombre de plasmides chez les bactéries

Nous avons donc 1362 données disponibles.
Le top-10 des espèces ayant le plus fort taux de G+C sont :

```
tail(tgc[order(tgc$GC), ], n = 10)
```

```
                          SPECIES GC...
1738                Opitutus terrae  74.0
1870  Pseudonocardia dioxanivorans  74.0
1895           Actinosynnema mirum  74.0
1898         Cellulomonas flavigena  74.0
2197 Anaeromyxobacter dehalogenans  74.0
2198         Anaeromyxobacter sp. K  74.0
160       Kineococcus radiotolerans  74.2
436  Anaeromyxobacter dehalogenans  74.9
138                 Vibrio harveyi  85.0
137         Enterobacter sakazakii  87.0
```

Le top-10 des espèces ayant le plus fort taux de A+T sont :

```
head(tgc[order(tgc$GC), ], n = 10)
```

```
                          SPECIES GC...
341  Candidatus Carsonella ruddii  16.0
335           Buchnera aphidicola  20.1
669    Wigglesworthia glossinidia  22.0
1338      Clostridium ljungdahlii  22.0
1924   Candidatus Sulcia muelleri  22.0
447         Mycoplasma capricolum  23.8
1649          Mycoplasma mycoides  23.8
606           Mycoplasma mycoides  24.0
1311   Candidatus Sulcia muelleri  24.0
588             Mycoplasma mobile  24.9
```

La distribution du taux de G+C est donnée dans la figure 3.

```
hist(tgc$GC, col = grey(0.7),
main = paste("Distribution du taux de G+C pour",
  nrow(tgc), "bactéries"), xlim = c(0, 100),
  xlab = "Taux de G+C [%]", ylab = "Nombre d'espèces",
  las = 1, labels = TRUE,ylim = c(0, 250),
  breaks = seq(0, 100, by = 5))
```

2.2 Taux de G+C et température

```
selectT <- c("Psychrophile", "Mesophile", "Thermophile", "Hyperthermophile")
gcet <- proc[, c("SPECIES", "GC...", "TEMPERATURE.RANGE")]
gcet <- gcet[complete.cases(gcet), ]
gcet <- gcet[gcet$TEMPERATURE.RANGE %in% selectT, ]
gcet$TEMPERATURE.RANGE <- factor(gcet$TEMPERATURE.RANGE, levels =
  selectT, ordered = TRUE)
table(gcet$TEMPERATURE.RANGE)
```

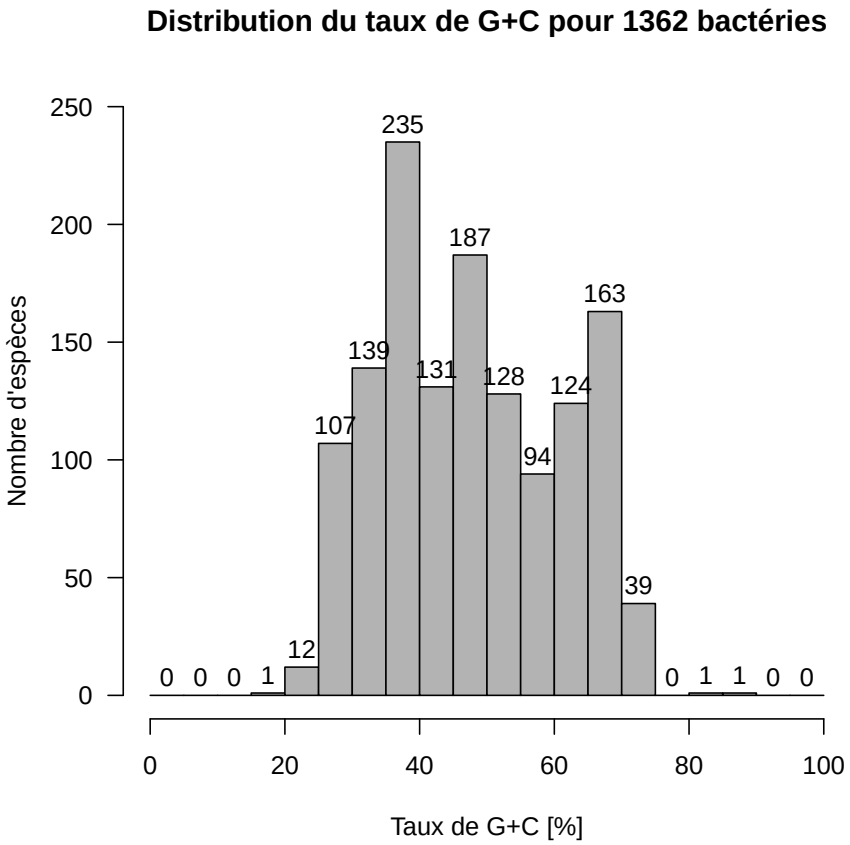

Distribution du taux de G+C pour 1362 bactéries
Nombre d'espèces
Taux de G+C [%]
0
50
100
150
200
250
0
20
40
60
80
100
0 0 0 1 12 107 139 235 131 187 128 94 124 163 39 0 1 1 0 0

FIGURE 3 –

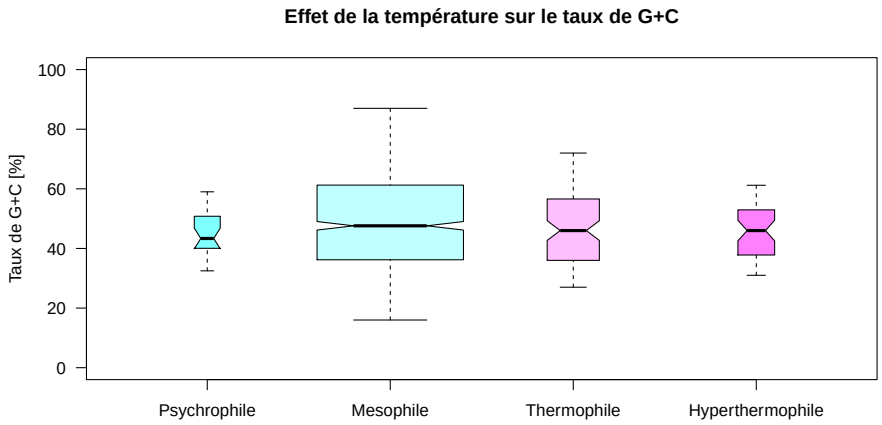


FIGURE 4 – Il n'y a pas ou peu d'effet de la température sur le taux de G+C chez les bactéries (n = 933).

```
Psychrophile         Mesophile       Thermophile Hyperthermophile
          24               764                97                48
```

L'absence d'effet de la température sur le taux de G+C est illustré par la figure 4.

```
suppressWarnings(boxplot(gcet$GC~gcet$TEMP, col = cm.colors(4),
notch = TRUE, varwidth = T, ylim = c(0,100),
las = 1,
main = "Effet de la température sur le taux de G+C",
ylab = "Taux de G+C [%]"))
```

2.3 Taux de G+C et a(n)érobiose

```
selectO <- c("Obligate anaerobe", "Anaerobe", "Facultative",
             "Aerobe", "Obligate aerobe")
gcetO <- proc[, c("SPECIES", "GC...", "OXYGEN.REQUIREMENTS")]
gcetO <- gcetO[complete.cases(gcetO), ]
gcetO <- gcetO[gcetO$OXYGEN.REQUIREMENTS %in% selectO, ]
gcetO$OXYGEN.REQUIREMENTS <- factor(gcetO$OXYGEN.REQUIREMENTS,
  levels = selectO, ordered = TRUE)
levels(gcetO$OX) <- c("Anaerobe", "Anaerobe", "Facultative",
                      "Aerobe", "Aerobe")
table(gcetO$OX)
```

```
  Anaerobe Facultative       Aerobe
       289         449         432
```

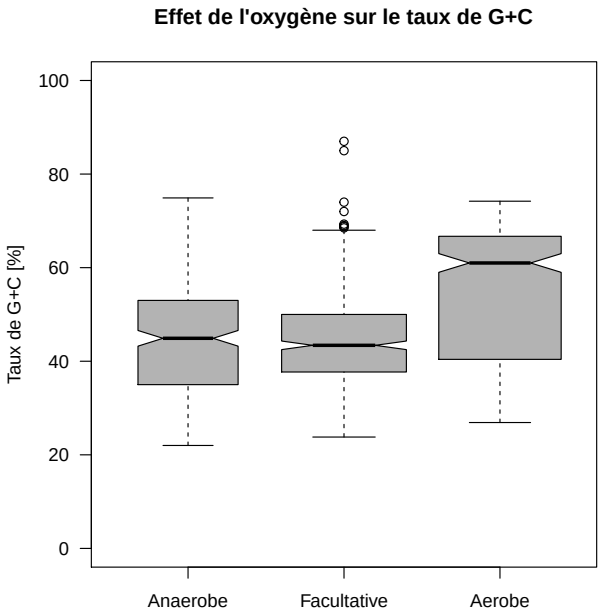


FIGURE 5 – Il y a un effet très net de l'(an)aérobiose sur le taux de G+C (n = 1170 bactéries).

Nous disposons ici en tout de 1170 données, l'effet de l'aérobiose sur le taux de G+C est représenté sur la figure 5.

```
boxplot(gcet0$GC~gcet0$OX, col = grey(0.7),
notch = TRUE, varwidth = T, ylim = c(0,100),
las = 1,
main = "Effet de l'oxygène sur le taux de G+C",
ylab = "Taux de G+C [%]")
```

2.4 Taux de G+C et fréquences en acides-aminés

Lecture des données :

```
load(url("https://pbil.univ-lyon1.fr/R/donnees/bgsc/uco739.RData"))
```

```
dim(uco739)
```

```
[1] 739  61
```

```
comment(uco739)
```

```
[1] "Wed Apr  9 10:46:30 2008"
```

Calcul du taux de G+C :

```
library(seqinr)
sapply(colnames(uco739), function(x) sum(s2c(x) %in% c("c", "g"))) -> tmpgc
tmpgc
```

```
aaa aac aag aat aca acc acg act aga agc agg agt ata atc atg att caa cac cag cat cca
  0   1   1   0   1   2   2   1   1   2   2   1   0   1   1   0   1   2   2   1   2
ccc ccg cct cga cgc cgg cgt cta ctc ctg ctt gaa gac gag gat gca gcc gcg gct gga ggc
  3   3   2   2   3   3   2   1   2   2   1   1   2   2   1   2   3   3   2   2   3
ggg ggt gta gtc gtg gtt tac tat tca tcc tcg tct tgc tgg tgt tta ttc ttg ttt
  3   2   1   2   2   1   1   0   1   2   2   1   2   2   1   0   1   1   0
```

```
gcpc <- 100*as.matrix(uco739) %*% tmpgc/(3*rowSums(uco739))
head(gcpc)
```

```
                                [,1]
ACHROMOBACTER DENITRIFICANS    61.88166
ACHROMOBACTER XYLOSOXIDANS     63.37462
ACIDIANUS AMBIVALENS           37.59371
ACIDITHIOBACILLUS FERROOXIDANS 58.72497
ACINETOBACTER BAUMANNII        43.92444
ACINETOBACTER CALCOACETICUS    42.96045
```

Calcul des fréquences en acides-aminés :

```
sapply(colnames(uco739), function(x) aaa(translate(s2c(x)))) -> tmpaa
tmpaa
```

```
  aaa   aac   aag   aat   aca   acc   acg   act   aga   agc   agg   agt   ata   atc
"Lys" "Asn" "Lys" "Asn" "Thr" "Thr" "Thr" "Thr" "Arg" "Ser" "Arg" "Ser" "Ile" "Ile"
  atg   att   caa   cac   cag   cat   cca   ccc   ccg   cct   cga   cgc   cgg   cgt
"Met" "Ile" "Gln" "His" "Gln" "His" "Pro" "Pro" "Pro" "Pro" "Arg" "Arg" "Arg" "Arg"
  cta   ctc   ctg   ctt   gaa   gac   gag   gat   gca   gcc   gcg   gct   gga   ggc
"Leu" "Leu" "Leu" "Leu" "Glu" "Asp" "Glu" "Asp" "Ala" "Ala" "Ala" "Ala" "Gly" "Gly"
  ggg   ggt   gta   gtc   gtg   gtt   tac   tat   tca   tcc   tcg   tct   tgc   tgg
"Gly" "Gly" "Val" "Val" "Val" "Val" "Tyr" "Tyr" "Ser" "Ser" "Ser" "Ser" "Cys" "Trp"
  tgt   tta   ttc   ttg   ttt
"Cys" "Leu" "Phe" "Leu" "Phe"
```

```
t(apply(uco739, 1, function(x) tapply(x,tmpaa,sum) )) -> freqaa
head(freqaa)
```

```
                               Ala  Arg  Asn  Asp Cys  Gln  Glu  Gly  His  Ile  Leu
ACHROMOBACTER DENITRIFICANS   2783 1601  566 1204 206  903 1339 1828  543 1045 2196
ACHROMOBACTER XYLOSOXIDANS    5031 2828 1090 2062 330 1637 2322 3192  941 1714 3794
ACIDIANUS AMBIVALENS          1297  700  849  853 218  482 1184 1319  283 1592 1919
ACIDITHIOBACILLUS FERROOXIDANS 5876 3794 1559 2705 623 2174 3138 4492 1391 2822 5179
ACINETOBACTER BAUMANNII       7428 4257 3823 4476 745 4252 5442 5608 2050 5660 9074
ACINETOBACTER CALCOACETICUS   2973 1599 1474 1903 413 1779 2004 2331  927 2145 3440
                               Lys  Met  Phe  Pro  Ser  Thr  Trp  Tyr  Val
ACHROMOBACTER DENITRIFICANS    907  498  733 1094 1266 1083  300  541 1621
ACHROMOBACTER XYLOSOXIDANS    1574  966 1309 2053 2138 2050  590  999 2859
ACIDIANUS AMBIVALENS          1381  499  925  923 1159  954  328  931 1412
ACIDITHIOBACILLUS FERROOXIDANS 1984 1460 1915 2734 2763 2784  783 1480 3922
ACINETOBACTER BAUMANNII       4871 1831 3518 3850 5756 4665 1466 2955 6092
ACINETOBACTER CALCOACETICUS   1884  774 1442 1433 1970 1845  430 1180 2211
```

```
aapc <- 100*freqaa/rowSums(freqaa)
head(aapc)
```

```
                                    Ala      Arg      Asn      Asp       Cys
ACHROMOBACTER DENITRIFICANS    12.503931 7.193243 2.543020 5.409534 0.9255515
ACHROMOBACTER XYLOSOXIDANS     12.743484 7.163302 2.760962 5.223030 0.8358874
ACIDIANUS AMBIVALENS            6.752395 3.644315 4.420033 4.440858 1.1349438
ACIDITHIOBACILLUS FERROOXIDANS 10.967188 7.081265 2.909776 5.048714 1.1627907
ACINETOBACTER BAUMANNII         8.458306 4.847470 4.353272 5.096847 0.8483358
ACINETOBACTER CALCOACETICUS     8.703926 4.681324 4.315367 5.571332 1.2091226
                                    Gln      Glu      Gly      His       Ile
ACHROMOBACTER DENITRIFICANS    4.057151 6.016085 8.213146 2.439682 4.695152
ACHROMOBACTER XYLOSOXIDANS     4.146508 5.881608 8.085311 2.383546 4.341549
ACIDIANUS AMBIVALENS           2.509371 6.164098 6.866930 1.473344 8.288213
ACIDITHIOBACILLUS FERROOXIDANS 4.057636 5.856882 8.384038 2.596215 5.267087
ACINETOBACTER BAUMANNII        4.841777 6.196837 6.385862 2.334347 6.445075
ACINETOBACTER CALCOACETICUS    5.208303 5.867026 6.824370 2.713939 6.279826
                                    Leu      Lys      Met      Phe      Pro
ACHROMOBACTER DENITRIFICANS    9.866559 4.075122 2.237498 3.293346 4.915308
ACHROMOBACTER XYLOSOXIDANS     9.610172 3.986930 2.446870 3.315687 5.200233
ACIDIANUS AMBIVALENS           9.990629 7.189713 2.597876 4.815702 4.805289
ACIDITHIOBACILLUS FERROOXIDANS 9.666281 3.703012 2.724999 3.574228 5.102841
ACINETOBACTER BAUMANNII       10.332616 5.546636 2.084970 4.005967 4.384017
ACINETOBACTER CALCOACETICUS   10.071142 5.515707 2.266007 4.221682 4.195333
                                    Ser      Thr      Trp      Tyr      Val
ACHROMOBACTER DENITRIFICANS    5.688098 4.865885 1.347891 2.430696 7.283102
ACHROMOBACTER XYLOSOXIDANS     5.415537 5.192634 1.494465 2.530459 7.241825
ACIDIANUS AMBIVALENS           6.033944 4.966681 1.707622 4.846939 7.351104
ACIDITHIOBACILLUS FERROOXIDANS 5.156967 5.196163 1.461421 2.762328 7.320169
ACINETOBACTER BAUMANNII        6.554390 5.312062 1.669343 3.364875 6.936995
ACINETOBACTER CALCOACETICUS    5.767485 5.401528 1.258893 3.454636 6.473051
```

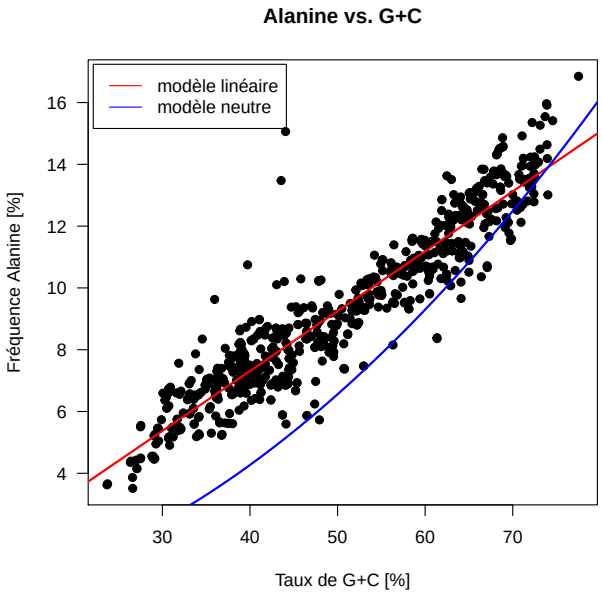


FIGURE 6 – La fréquence en Alanine augmente fortement avec le taux de G+C des génomes bactériens.

```
x <- gcpc ; y <- aapc[,"Ala"]
plot(x, y, pch = 19, las = 1,
  xlab = "Taux de G+C [%]",
  ylab = "Fréquence Alanine [%]",
  main = "Alanine vs. G+C")
abline(lm(y~x), col = "red", lwd = 2)
legend("topleft", inset = 0.01, lty = 1,
  legend = c("modèle linéaire", "modèle neutre"), col = c("red","blue"))
neutre <- function(xpc){
  x <- xpc/100
  y <- 2*x^2/(8 - (1-x)^2*(1+x))
  return(100*y)
}
xseq <- seq(from = 0, to = 100, lenght = 255)
mn <- sapply(xseq, neutre)
points(xseq, mn, col = "blue", type = "l", lwd = 2)
```

L'influence du taux de G+C sur la fréquence en Alanine est donné dans la figure 6.

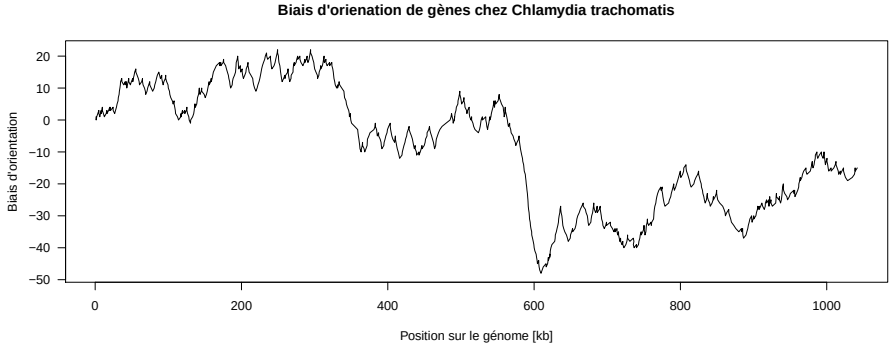


FIGURE 7 – Le génome de *Chlamydia trachomatis* est un exemple montrant qu'il n'y a pas toujours de biais d'orientation de gènes chez les bactéries.

3 Biais d'orientation des gènes

Nous allons illustrer cette partie avec le génome complet de *Chlamydia trachomatis* :

```
oriloc.res <- oriloc()
names(oriloc.res)
```

```
[1] "g2num"      "start.kb"   "end.kb"     "CDS.excess" "skew"       "x"
[7] "y"
```

Les résultats sont donnés dans la figure 7.

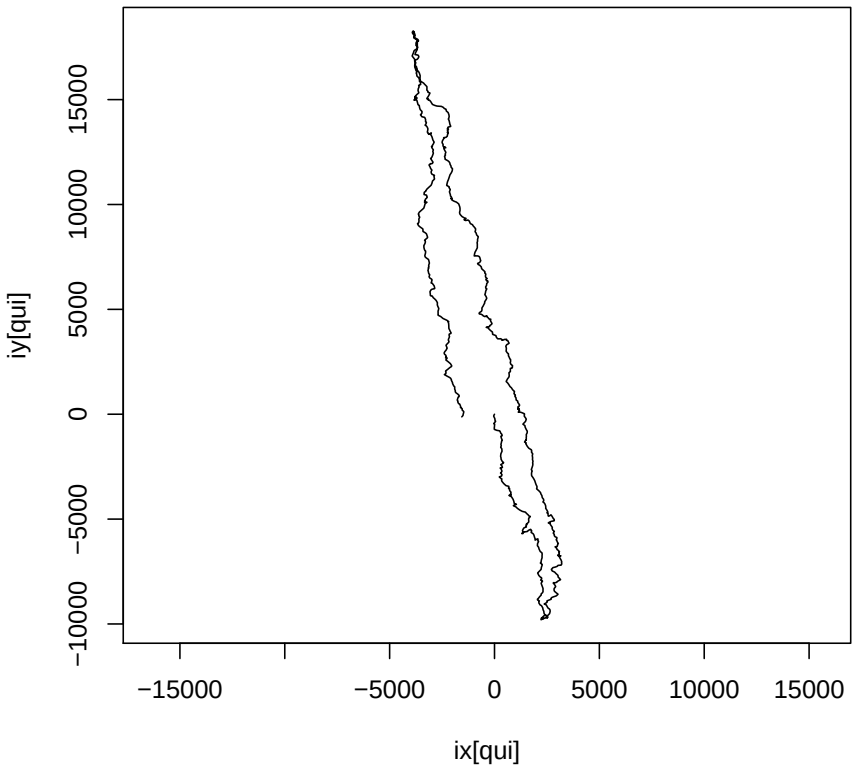
```
plot(x = oriloc.res$end.kb, y = oriloc.res$CDS.excess,
type = "l",
main = "Biais d'orienation de gènes chez Chlamydia trachomatis",
xlab = "Position sur le génome [kb]",
ylab = "Biais d'orientation",
las = 1)
```

4 Les chirochores

```
library(seqinr)
ctf <- system.file("sequences/ct.fasta.gz", package = "seqinr")
myseq <- read.fasta(ctf)[[1]]
length(myseq)
```

```
[1] 1042519
```

```
ifelse(myseq == "t", +1, 0) -> it
ifelse(myseq == "a", -1, 0) -> ia
ifelse(myseq == "c", +1, 0) -> ic
ifelse(myseq == "g", -1, 0) -> ig
ix <- cumsum(ia + it)
iy <- cumsum(ic + ig)
qui <- seq(from = 1, to = length(myseq), by = 1000)
plot(ix[qui], iy[qui], type = "l", asp = 1)
```


iy[qui]
15000
10000
5000
0
−5000
−10000
−15000
−5000
0
5000
10000
15000
ix[qui]