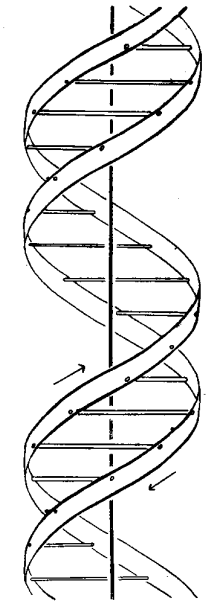
Fiche TD avec le logiciel : tdr312

Taux de G+C des chromosomes bactériens

J.R. Lobry & A.-B. Dufour

Examen et solution, BMS 2002

1 Introduction



J.D. Watson and F.H.C. Crick. A structure for deoxyribose nucleic acid. *Nature*, 171 :737–738, 1953

Dans la structure en double hélice de l'ADN il y a deux types de paires de bases : les paires A•T et les paires G•C. Il y a deux liaisons hydrogène dans les paires A•T et trois liaisons hydrogène dans les paires G•C. Plus l'ADN est riche en paires G•C plus l'ADN résiste à la dénaturation par l'augmentation de la température. Le taux de G+C d'une molécule d'ADN est la fréquence relative, exprimée généralement en pour cent, dans cet ADN. Par exemple, dans la molécule suivante :

```
5'-ACGT-3'
3'-TGCA-5'
```

le taux de G+C est de $100 \cdot \frac{2}{4} = 50$ %. Dans la molécule suivante :

```
5'-CCGG-3'
3'-GGCC-5'
```

le taux de G+C est de $100 \cdot \frac{4}{4} = 50$ %, elle sera donc plus résistante à la chaleur que la molécule précédente.

2 Premier fichier

Le premier fichier est extrait de : N. Galtier and J.R. Lobry. Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. *J Mol Evol*, 44 :632–635, 1997

```
data1 <- read.table("ftp://pbil.univ-lyon1.fr/pub/datasets/JME97/species")
```

C'est un tableau à 772 lignes (les chromosomes bactériens) et 4 variables :

1. nom du genre.
2. nom de l'espèce.
3. le taux de G+C du chromosome de la bactérie.
4. T_{opt}, la température optimale de croissance de la bactérie exprimée en degrés Celsius. C'est la température pour laquelle le taux de croissance en phase exponentielle est maximal. On distingue trois grands groupes de bactéries en fonction de leur température optimale de croissance :
   - $T_{opt} < 20\,°C$ Bactérie psychrophile
   - $T_{opt} > 45\,°C$ Bactérie thermophile
   - $20\,°C \leq T_{opt} \leq 45\,°C$ Bactérie thermophile

3 Deuxième fichier

Le deuxième fichier à analyser est extrait de : H. Naya, H. Romero, A. Zavala, B. Alvarez, and H. Musto. Aerobiosis increases genomic GC% in prokaryotes. *Journal of Molecular Evolution*, 55 :260–264, 2002

```
data2 <- read.table("http://pbil.univ-lyon1.fr/R/donnees/gc02.txt")
```

C'est un tableau à 551 lignes (les chromosomes bactériens) et 4 variables :

1. nom du genre.
2. nom de l'espèce.
3. le taux de G+C du chromosome de la bactérie.
4. une variable qualitative ayant deux modalités selon que la bactérie est strictement aérobie (`Aerobic`)) ou strictement anaérobie (`Anaerobic`). Une bactérie strictement aérobie ne pousse qu'en présence d'oxygène, une bactérie strictement anaérobie qu'en son absence.

4 Consignes

1. Tous documents, papiers, informatiques, calculatrices, neurones, etc. autorisés.
2. Explicitez clairement la question biologique que vous désirez traiter pour chaque fichier, les résultats et la conclusion.
3. 4 pages maximum (times 12 points interligne simple).
4. Envoyez le fichier à la fin de la séance en mettant [BMS] dans le sujet du message.

5 Proposition de solution

Remarque importante. Le sujet étant ouvert, il n'y a pas de correction canonique de cet exercice. Il ne s'agit que d'une illustration de ce qu'il était possible de faire ici.

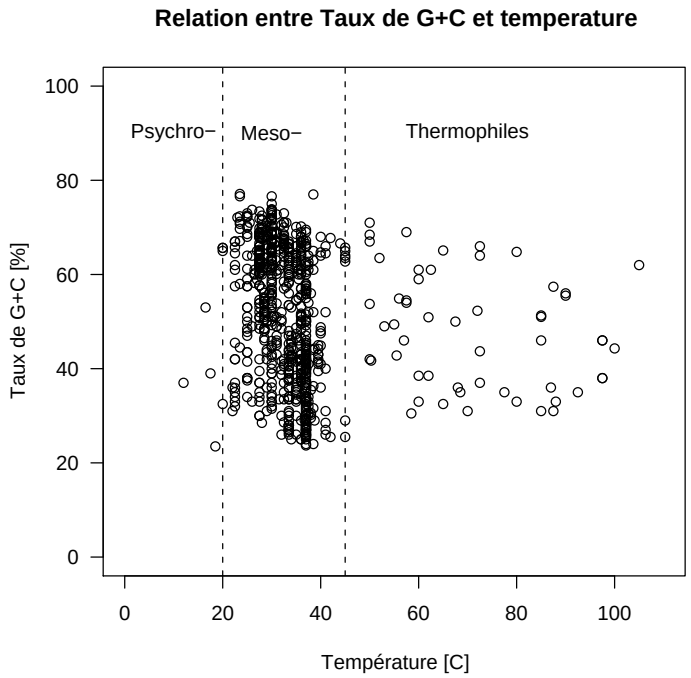
5.1 Objectif biologique

La mise en évidence d'une relation entre une caractéristique génomique et une variable environnementale est un des moyens utilisés pour détecter les effets de la sélection naturelle. La caractéristique génomique qui nous intéresse ici est le taux global en bases G+C des chromosomes bactériens. La première variable environnementale est la température, les bactéries ne régulant pas leur température, on supposera que leur température optimale de croissance en est un bon reflet. La deuxième variable environnementale est la présence ou non d'oxygène.

5.2 Existe-t-il une augmentation du taux de G+C avec la température ?

La simple représentation des données montre que si augmentation il y a, elle est loin d'être évidente :

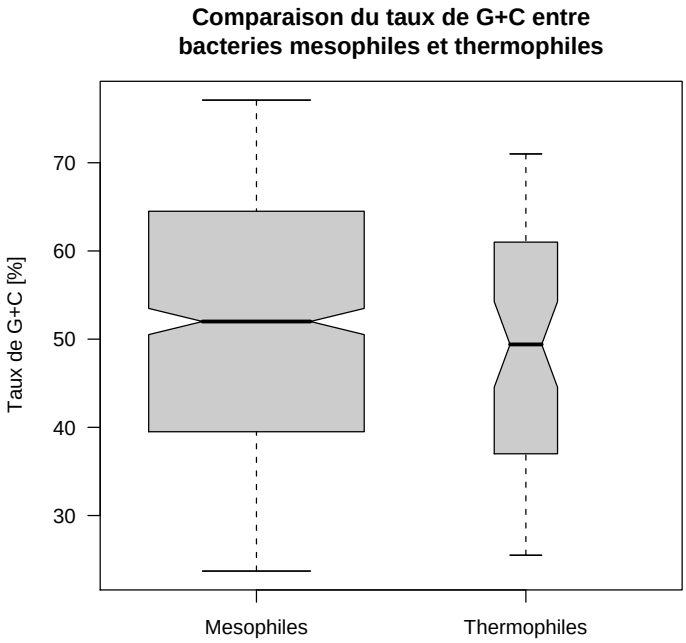
```
plot(data1$V4, data1$V3, ylab = "Taux de G+C [%]", xlab = "Température [C]",
    ylim = c(0, 100), xlim = c(0, 110), main = "Relation entre Taux de G+C et temperature",
    las = 1)
abline(v = 20, lty = 2)
abline(v = 45, lty = 2)
text(x = 10, y = 90, labels = "Psychro-")
text(x = 30, y = 90, labels = "Meso-")
text(x = 70, y = 90, labels = "Thermophiles")
```



On remarque que la gamme des températures couvre assez bien l'intervalle biologiquement admissible, l'eau étant en phase liquide entre 0 et 100 °C aux pressions ordinaires. Les psychrophiles, avec seulement 4 individus, sont mal représentés. Il y a une très forte sur-représentation des bactéries mésophiles, sans doute liée à un biais d'échantillonnage en faveur des bactéries pathogènes

pour l'homme. Pour pallier ce biais, on décide de recoder la température en une variable qualitative à deux modalités : mésophile et thermophile.

```
thermo <- data1[data1$V4 >= 45, ]
meso <- data1[(data1$V4 >= 20) & (data1$V4 < 45), ]
boxplot(meso$V3, thermo$V3, names = c("Mesophiles", "Thermophiles"),
    ylab = "Taux de G+C [%]", las = 1, varwidth = TRUE, notch = TRUE,
    col = grey(0.8), main = "Comparaison du taux de G+C entre\nbacteries mesophiles et thermophiles")
```

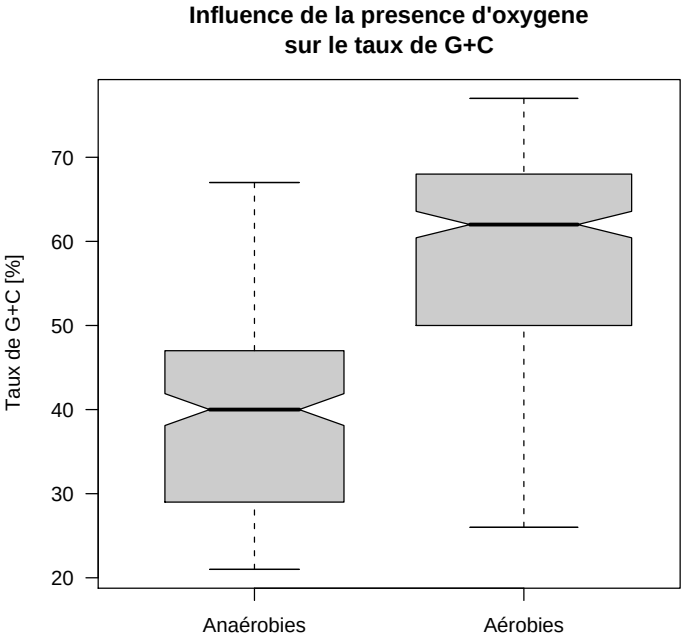


Rien n'y fait, les deux distributions sont complètement chevauchantes et la valeur médiane du taux de G+C diminue quand on passe des mésophiles aux thermophiles. Inutile donc de faire un test dans ces conditions : les données expérimentales ne nous permettent pas de mettre en évidence une augmentation globale du taux de G+C avec la température.

5.3 Le taux de G+C est il modifié par la présence d'oxygène dans le milieu ?

La représentation graphique des données nous montre qu'à l'évidence il y a un très fort effet de l'oxygène sur le taux de G+C.

```
boxplot(data2[data2$V4 == "Anaerobic", ]$V3, data2[data2$V4 == "Aerobic",
    ]$V3, names = c("Anaérobies", "Aérobies"), ylab = "Taux de G+C [%]",
    las = 1, main = "Influence de la presence d'oxygene\nsur le taux de G+C",
    varwidth = TRUE, notch = TRUE, col = grey(0.8))
```



Par acquis de conscience, on pourrait tester l'égalité des moyennes (les distributions ne sont visiblement pas normales, on utiliserait alors un test non paramétrique, comme le test sur la somme des rangs de Wilcoxon qui avec une p-value de $2.2\,10^{-16}$ nous confirmerait ici, si besoin était, que l'on peut bien rejeter l'hypothèse de l'égalité des moyennes des deux populations).

L'écart entre les taux de G+C moyens entre les deux populations est de 18.7, soit 18.7 % de l'amplitude maximale théorique, le taux de G+C pouvant varier au maximum de 0 à 100 %, soit encore 33.4 % de l'amplitude observée dans le jeu de données (de 21 % à 77 %), c'est quantitativement considérable.

5.4 Conclusion

Contrairement à ce que laissait sournoisement penser l'introduction de l'énoncé, il n'y a pas d'effet visible de la température sur le taux de G+C des génomes bactériens. Quand on met en regard les résultats obtenus pour les deux variables environnementales étudiées, l'influence de la température n'en paraît que plus illusoire. Ainsi, bien qu'il soit exact qu'un plus fort taux de G+C augmente la température de dénaturation de l'ADN, les bactéries thermophiles n'ont pas recours à cet expédient.

Les résultats montrent une très forte corrélation entre le mode de vie aérobie ou anaérobie des bactéries et le taux de G+C de leur chromosome. Corrélation n'est pas causalité, toutes les explications rationnelles sont donc acceptables à titre d'hypothèse. On peut imaginer par exemple un avantage sélectif lié à une meilleure résistance de l'ADN riche en G+C en aérobiose, ou encore une pression de mutation directionnelle vers les paires G•C induite en aérobiose.