

Épreuve aMIG - Contrôle terminal - 28 avril 2009

J.R. Lobry, L. Duret

18 février 2017

Les deux problèmes sont complètement indépendants.

(Durée : 2 heures)

Documents autorisés

1 Problème 1 (J.R. Lobry)

On utilise principalement ici le génome complet de la bactérie *Chlamydia trachomatis* qui est distribué avec le paquet `seqinr`, il n'y a donc pas besoin de connexion internet pour récupérer ces données si le paquet `seqinr` a déjà été installé.

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

1.1 Question 1

Quelle est la taille du génome, exprimée en paires de bases (bp), de *Chlamydia trachomatis*? Donner le code  permettant d'obtenir ce résultat. Comment se situe la taille de ce génome par rapport aux autres génomes bactériens?

1.2 Question 2

Quel est le taux de G+C, exprimé en %, de ce génome? Donner le code  permettant d'obtenir ce résultat (sans utiliser la fonction pré-définie `GC()` du paquet `seqinr`). Comment se situe le taux de G+C ce génome par rapport aux autres génomes bactériens?

1.3 Question 3

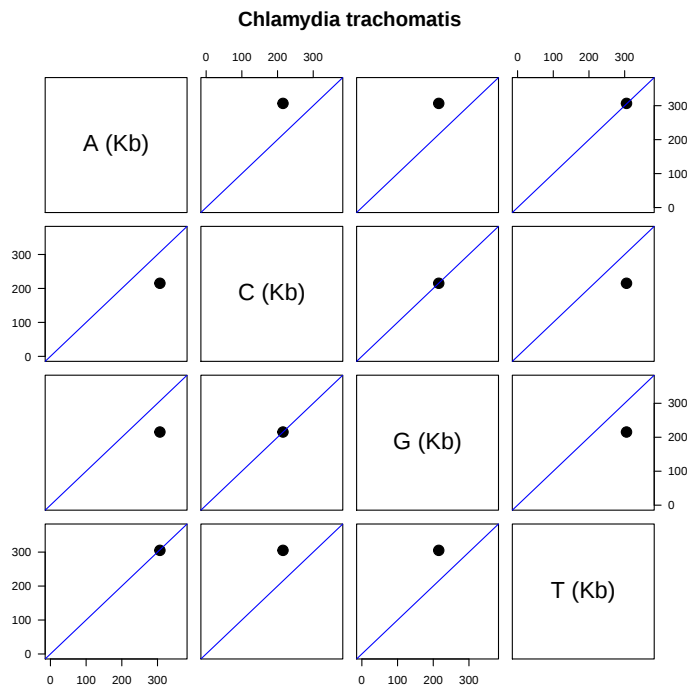
On s'intéresse au nombre total des 4 bases, exprimé en Kb. Que montre le graphique suivant?

```
na <- sum(myseq == "a")/10^3
nc <- sum(myseq == "c")/10^3
ng <- sum(myseq == "g")/10^3
nt <- sum(myseq == "t")/10^3
mydf <- data.frame(list(na=na,nc=nc,ng=ng,nt=nt))
lims <- c(0,1.2*max(mydf))
```

```

panel.lm <- function(x, y, ...)
{
  points(x, y, ...)
  abline(coef = c(0,1), col = "blue")
}
pairs(mydf,xlim=lims,ylim=lims, pch = 19, cex = 2,
      las = 1, lower.panel = panel.lm, upper.panel = panel.lm,
      labels = c("A (Kb)", "C (Kb)", "G (Kb)", "T (Kb)"),
      main = "Chlamydia trachomatis")

```



1.4 Question 4

Pour savoir si la propriété mise en évidence dans la question précédente présente un caractère de généralité on récupère les données suivantes :

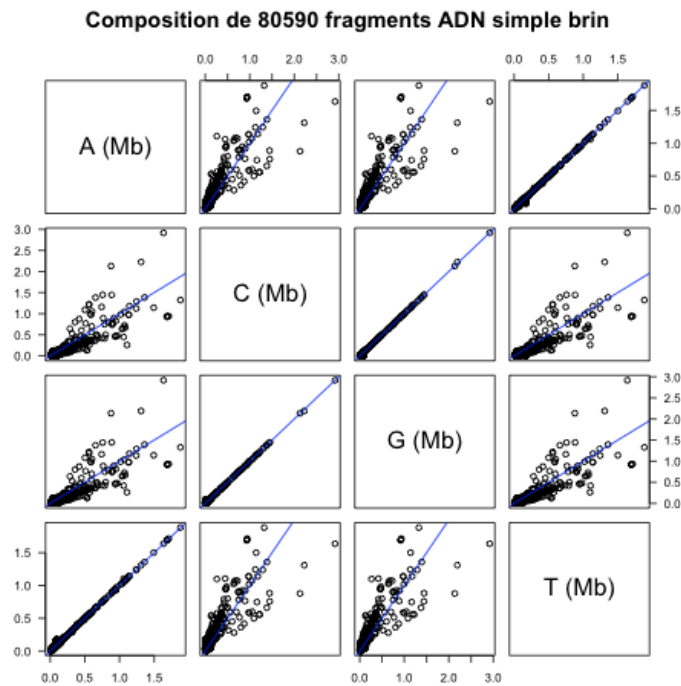
```

load(url("http://pbil.univ-lyon1.fr/R/donnees/big50.RData"))
dim(big50)
[1] 80590      7
head(big50)

```

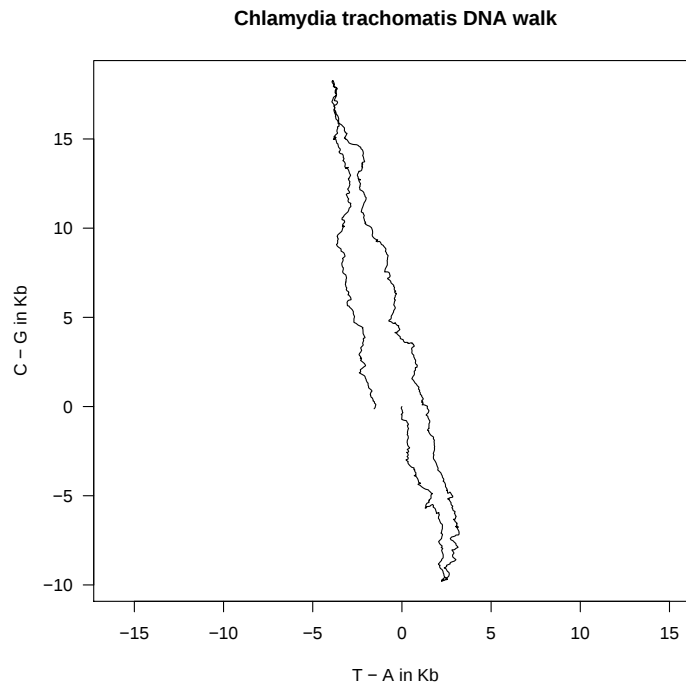
	mnemo	bp	A	C	G	T	0
1	A48542	133894	39195	27151	27347	40201	0
2	A69720	53789	6707	19183	20504	7395	0
3	A79350	320040	66253	93201	94558	66028	0
4	A79351	236165	48909	68589	69313	49354	0
5	A93002	320040	66253	93201	94558	66028	0
6	A93003	236165	48909	68589	69313	49354	0

Il s'agit du nombre de bases nucléotidiques dans 80590 fragments d'ADN simple brin de plus de 50 Kb (extraits de GenBank le 24 novembre 2004). Donner le code  permettant de produire le graphique ci-dessous et votre interprétation du résultat obtenu.



1.5 Question 5

Le graphique ci-dessous a été obtenu à partir des données sur le génome de *Chlamydia trachomatis*. Donner le code  permettant de produire ce graphique. Que montre ce graphique ?



2 Problème 2 (L. Duret)

2.1 Protéine de souris

La protéine MaProt (de souris) a été caractérisée en 1998. Cette protéine a été comparée à la banque de données SwissProt à l'aide du logiciel BLASTP. La liste des séquences similaires détectées à l'époque par BLASTP est indiquée ci-dessous :

```
#####
BLASTP 2.0.5 [May-5-1998]

Query= MaProt

Database: SwissProt
        600,231 sequences; 186,808,058 total letters

Sequences producing significant alignments:

Seq1  264 Human Seq1 protein.          1851  0
Seq2  193 Chicken Seq2 protein         138  1e-55
Seq3  351 Zebrafish Seq3 protein.       92  1e-03
Seq4  558 Drosophila Seq4 protein.      31  0.7
Seq5  531 Caenorhabditis elegans Seq5 protein. 42  0.9
#####
```

En 2009, on a refait la recherche de similarité avec BLASTP sur la dernière version de la banque de données SwissProt. On a obtenu le résultat suivant :

```
#####
BLASTP 2.1.9 [April-15-2008]
```

Query= MaProt

Database: SwissProt

6,8102,836 sequences; 191,518,738,896 total letters

Sequences producing significant alignments:			Score (bits)	E Value
Seq1	264 Human Seq1 protein.		1851	0
Seq2	193 Chicken Seq2 protein		138	1e-54
Seq3	351 Zebrafish Seq3 protein.		92	0.01
Seq4	558 Drosophila Seq4 protein.		31	7
Seq5	531 Caenorhabditis elegans Seq5 protein.		42	9

#####

1. Donnez la liste des homologues identifiés par BLAST en 2009 et ceux identifiés en 1998 (vos choix doivent être justifiés).
2. Donnez la définition de la "E-value".
3. Pourquoi la "E-value" est-elle différente en 2009 par rapport à 1998 ?
4. Quelle méthode de recherche de similarité pourrait-on utiliser pour gagner en sensibilité ?

2.2 Fragment du génome du poulet

Un fragment du génome du poulet (*Gallus gallus*) a été séquencé :

```
cat(readLines("http://pbil.univ-lyon1.fr/R/donnees/poulet4.fasta"), sep = "\n")
```

```
>poulet4
GATAAAAGTTTAAAGTGAATTTTTCAGTGATATAATTTCCATACAGGAAAAGTGTATGATA
GATTTGAAAAAGGAATTAATGTAATCTGATTCATTTTAAAGGAGCAGATGATTTTAG
GTAAGCTTCTAAGAGAGCTATTGAAATAATCTATACGTTATAGACATACAGAGATTGGTT
TTAGATGTTTATTTTCTGGAAAGATATACTAATGTGTAATAGAATACTTACTATGGCAAC
AGCATGTGAAATGTTTTTAAAAACAAAGTAATTTTTTAATAGGTCTTACAAGAACCTTA
AGTATAAATTCAGCTGAATCAGCATTGAAAAAGCTCAAGGAGAGAGATTACGTTGCCA
TGTACCTTTGAACCTCTCAGAAGAGGATGTAGGCACACTAGATATTGAGTGGGTTTGATA
CCAGCAGATATTCAGAAGAAGGAAGAAACAGTAAGTAAAACTTCTTAATTTTGGCAGC
TGTAGTCAAGTCATTAGGGAAGTCAATGTGTGAACCTTGGTAGTGACTCTTTGACTTT
CACCTTTTCAAGATGGTATGCGTGTGATCTTAAATTAATTTGAAGCACCATGTGGAAGT
TGTAAATAGTGAAGTGAATGTTACAATATTTTGTGCTTAATCTCTATTCACTGTGAACAAT
TCTAAATTAATAAATGATTGAGAGAACAGACAACCTGCTTTTGCAGACCTACACTATG
TTTCTGATTGAAAGAAGTCGTTTATAAGAATTTGGTCTTAAATGTAGTATGCTTTTGTTA
GTTATTAGCTCTTTCACTTATTCTGAAGTCTGAGAACCTCTTTAGCAGAGATTTACAGA
ATTAAGATTAGATAAATAGTTTTCTTTAAGACAGAAGACAGTGTGACATCTTTCACTT
GAACAGGTATTGCAATTTGGAGTATTTGTTAGAGCGTTTGTGTGGTGTATTTTGAAGAAA
GGTACGTCATGCCCTGGGATTACTTTAAACAGTAAAGGGAGGCTGCTAAATCTGAAGA
TGCTTTTATTTTCTTCAATGTTTCAATATGTCAAGTAGATTCTGAACCGTATTTCTT
CTTTTACAGATAATCTATATTCTGGAGATAGGATTATAATCATTATCATCTGCTCTG
GCTGGGCGCTGCAATTTACTAGTTCTGATCCCAATCTGGTATGGTTCACTGGGATATC
CTGAATTTAAAGTCAGCAGATACTGGCAGATACCAAGTGAAGTGAAGAAGGCTCTGGA
GTTGAAGGCTTAAATACAGTTGAACGTAAGTGAAGTGAAGTGAAGTGAAGTGAAGTGAAG
TGAATTTCAAGTGGCATGTTTGTGTTTATGCTATTTTTCATTGATATATAAATGAAGAA
GAATGTAGGCGAAAGTCATACTGAATCTACAATAATGCATCCATTATTAAGTCTTCAGGA
CAGGGGCGAGGCAAGTAAATAATAGAGTAATGAGCACTTTTAAATTTGTGCATCTGAT
TCGATTTTCATAGTATTTTTCATGCTTACAAAGGCTTTGTTGTACAAATTTGCTATAG
TAAATGTGCTGATGAATTAATGTTACGATTTTGAATGTGACTTATTAATTCATAAT
ACTTTGTTTATCAGTAAGCCAGCAAGCACTAAATGCTCCATTGAAGGATCACAGGAGAT
TGGAAGGACATTGATTGAAGTGTGCATCACAAGAAAGCAACCCACTTTTGTATTATGA
CTGGAGAAGAGTAGTAACCTGGCACACAAGGACTTCCTGCCACTTCCTGACTAAGTATTGT
CAAACTTTAAGTATTTTCAAGATAAATAAATAATACAGATTGTATTCATATTTTGGACATAT
TTTGACTTGTAGCAGCTTATGCGTGCTTTAATATAGAGCAGTCTATAGAAGTGTCACT
TCTATCCTTCCAGTGGATAGTAAACTGACATTTAAACAAGCTACTCCGAGTTGAGAGAG
ATTCTTACCATTAGGTTTGCAGAAATCATGCTAATCTCTGCATACAAATGTTTGAAGTCT
CTTCTGTTGATTCTTGGCTGCTTTTCTGTTCCAGATGGATGGGTTGAAGATCAGAATA
GAGTATAGACCTGGTTTGAAGTGTATATTTCAATTGGAATGAAATTTAAATTTTAGTA
TAGATTGAAGTGAAGCTTCAGTTTCTGGTCAAGTGAAGTGAAGTGAAGTGAAGTGAAGTGAAG
GAGAGAATACATGCTCTATGGAAGAAAAATTTGTCATGAGTAAAGGAAATGAGAAGGTTGTC
TCTCTTCTTATGTGACACATCCAGGAAAGAAATTTCTTAAAAAACTGGCTTATAAAT
GAGAATTGTTTATTTGTTGTTGTTGTTGTTGTTGTTGTTGTTGTTGTTGTTGTTGTTGTTGTT
TTGCTGAAGTTTATTAGAGAAGTTGATTTTACTCATCAGTTTACTTCATTGCTAGTTTGT
GAGTTCTGACTTTGAAAGTTATCCTGGAATATTTTATGTAATATGCTTTGCAGTACCAA
TGCTTTTCTTACTTTTGTAAATTAATAAAAAAAGCTTTAAAGCCTTGAATGGTAGCC
CTTAAGTGCAGGTATAGCTCTTAAGTATTTGCTTGAAGTGTTTTAACTCAGATTGCA
```

AATTACATACAGGATGAAACTATCATGATATGTTCAAGAGCAGCCATAGTCTTGGGAGA
GAGATTAGGAACATGTGTGGATCTTCTCATAATGATCTCCTGCTCCTGTCATTTATG
CTTTTAGACCTTCCAGAGGGAGATACGGTTTCCATTCAATTGACCTAAACCAACAAAAG
TGACTATCTGAACCTGTGCTTTGTACTTCTCCAGTCTACTTCAAACCTGTAGTCCAGTAGTA
TCTGCATGTCTTAGGGGACATAAATCCTAAAAATTAATTGCATCTGAAAAACCTTGTACT
GTTCTTTGAGCTTGCCTTGTGGTAGTGTGTTTGATACATCCTATTTCTTGTTTTGACAAC
AAGCAGTTAGTTACTAATGTCTGTGTATCATTCAAGACCTTAAGCTTCTATTGTTCTC
TCTTTATCCACATTGCAAAATTTGTTTCATAAAATAAATTGATAGTTTTAGTGTGGACAAG
AATGAGGAAGTATACAGAAATGGTGTGTAGCATCATTAAAGTACTGTTGCTTTCGGCTTG
TGATTGTAGGTTTCATGGGCACTGAATGTGCTGTCAAGAACCTAGATGCTGCTTCACTTCA
GCTGAAGGAGCTAGCTAGGTCTTTGTGTGATTCTTTAGTACTGAGATGTAGTCTCTTGAC
AGCTTGGGATTTAAGTGTGTGATGGCAAACTATGAGCAATCTGAGTGTCAAGCCATTTG
AAGTTTCAGATAGTAATATCTCAAAGGTAGCTACCAGTTGCTTTGGAAGCCTTAGGAAGAA
ACTTTTTCTGTTTTCGCTCAGTCAGCCTATCTAGAGTTAAATGGTCACTATTGGATCAG
CCTTTGTGTTGAACGTGAGTTCTGCAGACATCTGAGATGGTGCATAGAATTACTTCA
ATACAGAAAGAGAAGGAGATTCCTTTGGGATGCAGCTCATCCACTGTATTTAAAACTCCA
GGTTAGGGTGAGAGAACTTGCCATCTTAAGTATCCTTTCTTCCATTGGCAGAGCAATT
TGAAAGCCTTGTACTAAAGTCAATTTCTGTGGCATAAATTTCTACTGAAGTAAACGGTT
TTTACAAATAGTAGTGTCTGTGGTGGAGTAACTTTTGGTTTGTAGTATGTTTCAAT
ATTGCAATTTTGTAGTTAAGATTTCATGAATGTGCTTGATTAATAATTCTGTGCGTGTGG
ATTGATAATTCTGTTCATTGAATTAGAAATTAACCTTTGGTAAATGGTAAACCTTTTGC
TTCTTATTTCTTGTAGACAAAAATACAGGGGAACCTCTCTTGAATAATGCCTCTAAAGAC
TATTTCTGGTACATACAGTTGTGTGCTTCAAACCGAGTTGGCAGAGATGAATGTTCTGT
GAGCTGAATGTACCCCTCGTAAGTGTCTACTGTGTAGTAGTCACATAGTGTGTTGTGTA
TTTCTGTGATCTGTTGCTGTAAGAAGCATAGAGAGAAGAAATATGAGAAAGAAGTACAT
CATGAATCAGGTAATGACATGCCTAGCTTGAGAGGTTCTGCTTGAAATTAGCTGCTAA
ATGACTTTATGAAGGAAGATGACTAGTAGTTGGGATTTCTGTTTTAATCTGGTAAATTT
ATTAACCACTAACAGTTTCATCTGAGAACCTGTTTGGATATGGTGTAGGTAGTCTATGTC
TTCCTACTTTTCAGTAAAAATCTTAAAGACAATGATAACTTCCTTTTTTTTAAAGATTAA
AAAAAGCCTTTTTTGTCTACTGAGTCAGAAATGGAAGAAACAAATGGAACGTTATTATG
CAAGCTGCTGCGTAGCAGTAACTGAACCTCTGTTGTGCAACTGTTCTTTCTTTCTTCA
GAGAAGATTTCTGCTCCAAAAAGTCGAGTTCAACAGCACGCAGCTACATAGGCAGCA
ATCGTTCTCTCTGGGTTCAATGTCTCCCTCAAATATGGAAGGATACTCCAAAACTCCAT
ATAGCCAGGTTCCAAAGTGAAGACTTTGAACGTACTTCTGGTCAAAACCAACCAATTGCAT
CTTCAAAGGTAGCTGCACCTAATTTAAGTAGAATGGGAGCTGCTCCTGTGATGATTCCAG
CACAAGCAAGATGGGTCATAGTATGAATATTAATTTAAGTCTGGGTTTTTAAAGT
GTTTGTGTAGTATTAGAGAAAACTACAGTATTCACCCCTCATTTAAACAATGGCATG
CAATTTCTTGAAGTAAATGAACATGTTAGTTTGAAGGCCCAATTCTCATTTTTTAA
TTTTAACTTATTAGTGTGTAACAGTTGAACATTTGAAAGCGTGAGAGTTCCTAAATATC
AGATACTGAAGGTGTTTGGATCTCTGGTGGCTTGTGAAGAGATGCTATTAGCTGATGTG
CAGTTCTCAGAATCTGAAAGAGCAACAACTGAGAAGTAAACAAACCAATTTCATATGTA
GACAATGAAGTTCTCAGAGTTCTGAAAGCTTTATTCTCAGTCTTGAGTAAAGGGGTAT
TTAATTTGGCCACTGTCTGATGAACCTTGCAAGTGGGCTGAGATACTGAGGAGAACTTG
AATTCGTCTCAAACCTCAGAGGACTTATGCAAGTCTTGAAGTTGTCTGGAGTGTTCCTGG
GTTGTTATCTCAGGGGCTACCAAGTCAATTTGGTTGGTTTCAAGTTAATTTAATGAATCTGTT
TGAGTTTAAATTTACAATGAGTAAAAATGCAACAGGAGATTTTGCAAGCTCCTTGAAATC
AGAATTTTACTGCATTAACTGTTCTAGTAGTGTGATAGTTAAGTTATCCTATGTAAC
ACGTCTTATCAGATTGTTACTAGTGTGATCTAGGTTCTAAATCTATAGCTCAACTGTTA
CATTTGCAAAAGTGTAGAGTGAACATGCATCTCTGAGGCTTAAAGTTAGCTGTATCTCACA
TTTTCCAAAGCCTCTTAGATTCTAAGGCAAGTGTCTTTTCTAAACCACTACCTAGAA
GTTTCACGGGGCTTGATTACAGCTGTTAAGAACTGGTATTTAATAGTTGCTCTGTGCTC

1. Recherchez le(s) gène(s) contenu(s) dans ce fragment.
2. Décrivez la structure (*i.e.* les positions des introns et des exons) du (ou des) gène(s).
3. Décrivez et justifiez chacune de vos étapes (*e.g.* pourquoi doit-on rechercher et masquer les séquences répétées?).