Markov Models part 1: nothing to hide so far. Solutions

Computational biology

Jacques van Helden

2019-12-18

Contents


```
if (!require("gplots")) {
   install.packages("gplots", dependencies = TRUE)
   library(gplots)
   }
if (!require("RColorBrewer")) {
   install.packages("RColorBrewer", dependencies = TRUE)
   library(RColorBrewer)
   }
```

CpG island length distribution

```
## Load CpG island coordinates
CpG.coord <- read.delim(file = "results/hg38/hg38_CpG_islands.bed.gz", header = FALSE)

names(CpG.coord) <- c("chr", "start", "end", "id")
## Note: bed convention : the end marks the first position
## **after** the feature, so the length is simply the difference
CpG.coord$length <- CpG.coord$end - CpG.coord$start
```

```
genome.size <-  3e+9

## Compute statistics about CpG island lengths
CpG.length.stat <- data.frame(
  nb = nrow(CpG.coord),
  min = min(CpG.coord$length),
  Q1 = quantile(CpG.coord$length, probs = 0.25),
  median = median(CpG.coord$length),
  mean = mean(CpG.coord$length),
  Q3 = quantile(CpG.coord$length, probs = 0.75),
  pc95 = quantile(CpG.coord$length, probs = 0.95),
  max = max(CpG.coord$length),
  MB =  sum(CpG.coord$length) / 1e6,
  cov.percent = sum(CpG.coord$length) / genome.size * 100
)


## Print the stats
kable(CpG.length.stat, caption = "Statistics about CpG island lengths", digits = 3, row.names = FALSE)
```

Table 1: Statistics about CpG island lengths

nb	min	Q1	median	mean	Q3	pc95	max	MB	cov.percent
31144	201	325	569	777.05	959	1947	45712	24.2	0.807

The currently annotated CpG islands cover 24.2 Mb, which represent 0.8% of the genome.

```
par(mfrow = c(2,1))


par(mar = c(4.1, 5.1, 1, 1))

hist(CpG.coord$length,
     breaks = seq(from = 0,
                  to = CpG.length.stat$max + 50,
                  by = 50),
     main = NA, las = 1,
     xlab = "CpG length", ylab = "Number of CpGs")

## Draw the histogram with a limited X axis to see the relevant part of the distribution
hist(CpG.coord$length,
     breaks = seq(from = 0,
                  to = CpG.length.stat$max + 50,
                  by = 50),
     xlim = c(0,3000),
     main = NA, las = 1,
     xlab = "CpG length (truncated X scale)", ylab = "Number of CpGs", col = "gray")
```

```
par(mar = c(4.1, 5.1, 4.1, 2.1))
par(mfrow = c(1,1))
```

Loading the Markov models

We start by loading the two Markov models previously computed from the RSAT tool create background model.

```
## Load a transition matrix from the RSAT result
readMarkovModel <- function(file) {
  model <- read.delim(
    file = file,
    row.names = 1)
  names(model) <- toupper(names(model))
  rownames(model) <- toupper(rownames(model))
  prior <- model[nrow(model),1:4] ## Last row contains priors as comments
  names(prior) <- names(model)[1:4]
  transitions <- rbind(model[1:4, 1:4], "B" = prior)
  return(transitions)
}

## Load the CpG transition table from the RSAT result
transitions.CpG <- readMarkovModel(file = "results/hg38/hg38_CpG_transitions_m1.tsv")

## Load the genomic background transition table from the RSAT result
transitions.Bg <- readMarkovModel(file = "results/hg38/hg38_genomic-bg_transitions_m1.tsv")
```

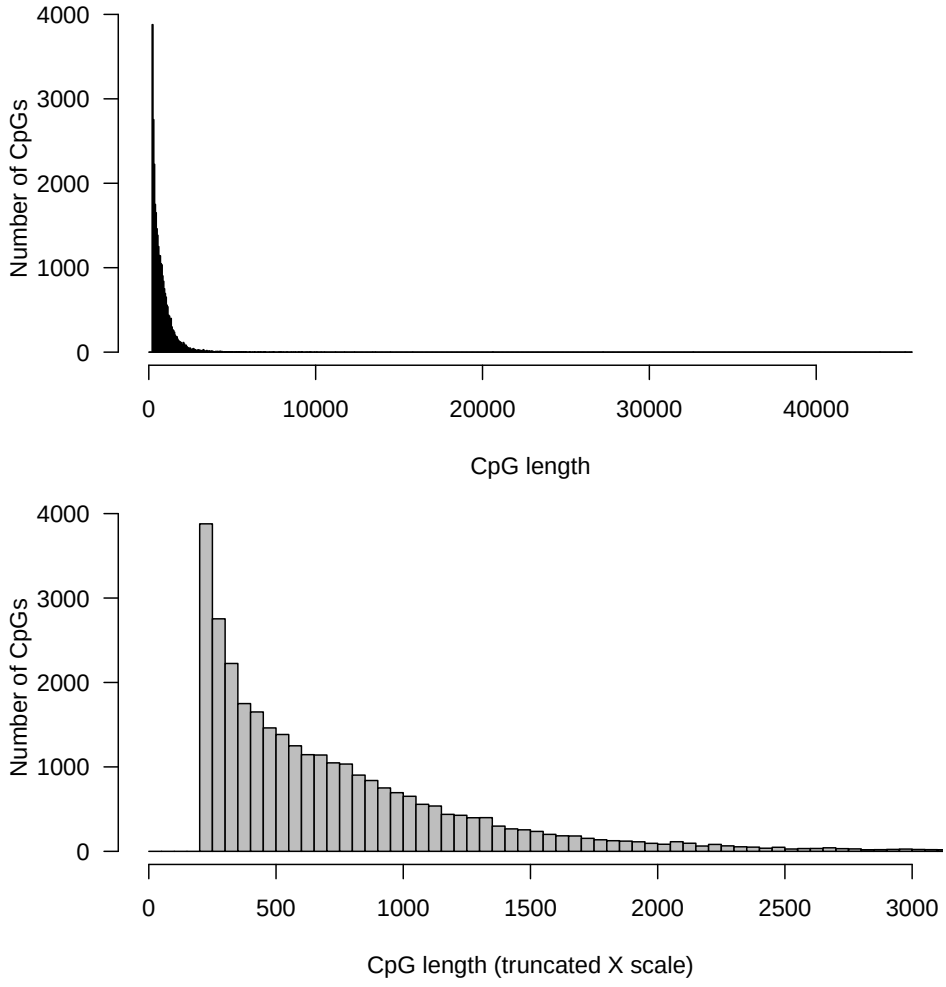

Figure 1: Distribution of lengths for all the CpG islands annotated in the Human genome.

```
## Check that the rows of the transition matrix sum to 1
## (given some rounding errors)
# apply(transitions.CpG[2:5], 1, sum)
# apply(transitions.Bg[2:5], 1, sum)
```

```
## Define a function that draws a heatmap from a transition matrix
transition.heatmap <- function(
  transitions,
  main = "Transitions",
  col.palette = gray.colors(n = 100, start = 1, end = 0),
  breaks = (0:length(col.palette))/(n.colors*2)) {


  n.colors <- length(col.palette)
  h <- heatmap.2(
    as.matrix(transitions),
    main = main,
    cellnote = round(digits = 3, as.matrix(transitions)),
    trace = "none",
    margins = c(4, 4),
    #    offsetRow = -10,
    breaks = breaks,
    key = FALSE,
    srtCol = 0,
    notecol = "black",
    notecex = 1.2,
    col = col.palette,
    scale = "none",
    las = 1,
    Rowv = FALSE, Colv = FALSE, dendrogram = "none")
  return(h)
  #  grab_grob()
}
```

```
## Draw transition heatmap
heatmap.CpG <- transition.heatmap(transitions.CpG, "CpG islands")
```

```
## Draw transition heatmap
heatmap.Bg <- transition.heatmap(transitions.Bg, "Genomic background")
```

Discrimination between two models

Problem: for a given sequence of events (e.g. a nucleotidic sequence), identify the most likely Markov model.

Approach: compute the log-likelihood ratios (log-odd ratios) of the sequence probabilities computed using respectively the CpG island and genomic background models.

$$P_{\text{CpG}}(S) = P_{\text{CpG}}(S_1) \cdot \prod_{i=1}^{n-1} P_{\text{CpG}}(S_{i+1}|S_i)$$

$$P_{\text{Bg}}(S) = P_{\text{Bg}}(S_1) \cdot \prod_{i=1}^{n-1} P_{\text{Bg}}(S_{i+1}|S_i)$$

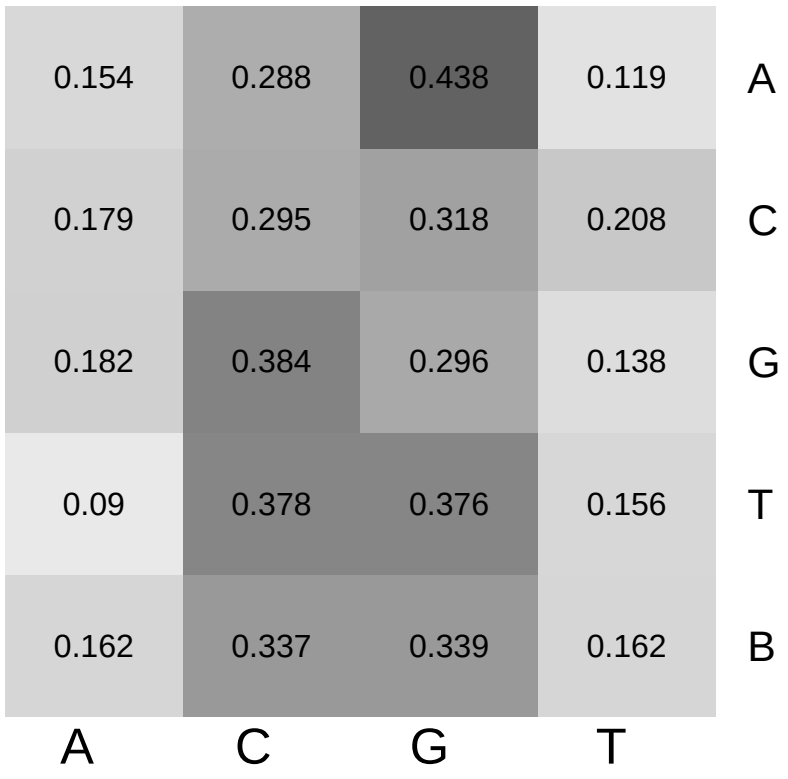

Figure 2: Heatmap of the transition matrices from 1st order Markov models trained on CpG islands.

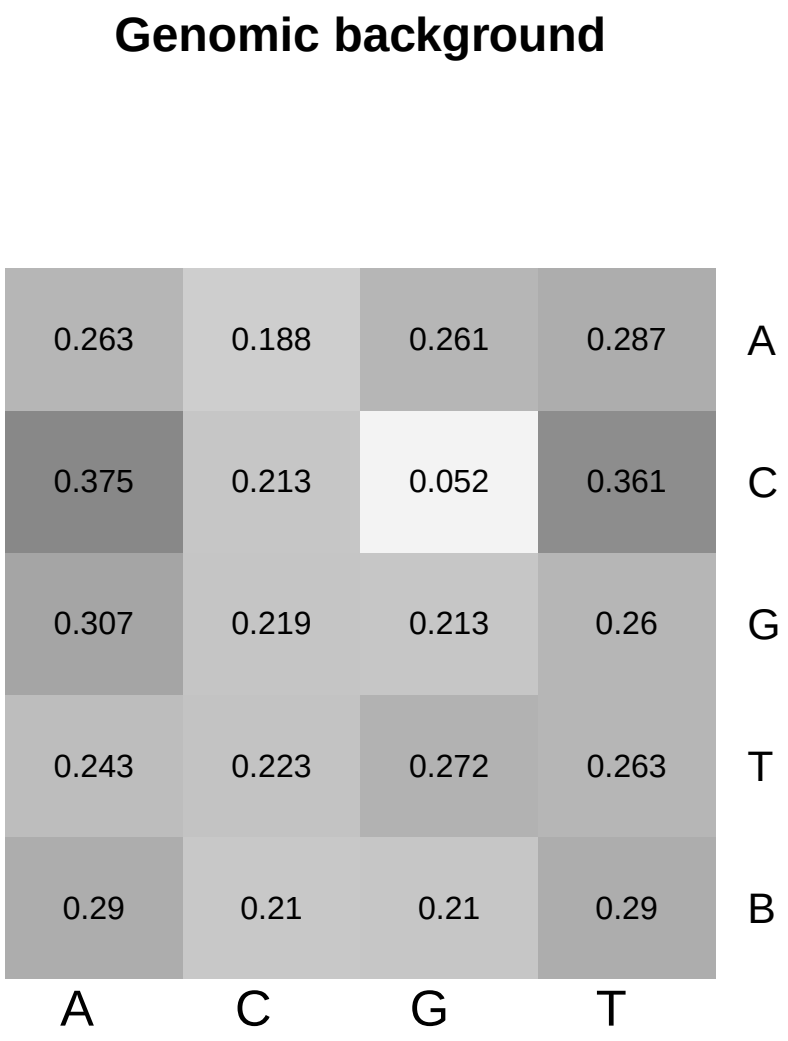

Figure 3: Heatmap of the transition matrices from 1st order Markov models trained on genomic background.

$$L(S) = log\left(\frac{P_{\text{CpG}}(S)}{P_{\text{Bg}}(S)}\right)$$

where

- $L(S)$ is the log-likelihood of the sequence S,
- $P_{\text{CpG}}(S)$ the probability for this sequence to be generated by the CpG island model, and
- $P_{\text{Bg}}(S)$ its probability to be generated by the background model.

A more efficient approach : rather than computing the two sequence probabilities (as the product of transition probabilities), we can compute once and forever a matrix with tbe log-odds of residue transitions.

$$L(r_j|r_i) = log\left(\frac{P_{\text{CpG}}(r_j|r_i)}{P_{\text{Bg}}(r_j|r_i)}\right)$$

```
my.palette <- colorRampPalette(c("blue", "white", "red"))(n = 100)

## Compute log-odds of transition frequencies
transition.log.odds <- log2(transitions.CpG / transitions.Bg)

max.lor <- ceiling(max(abs(range(transition.log.odds))))

heatmap.logodds <- transition.heatmap(
  transition.log.odds, "CpG / Bg log-odds",
  col.palette = my.palette,
#  breaks = seq(from = -max.lor, to = +max.lor, length.out = length(my.palette) + 1)
  breaks = length(my.palette) + 1
)
```

```
# library(gridGraphics)
# library(grid)
# library(gplots)
# library(gridExtra)
#
# grab_grob <- function(){
#   grid.echo()
#   grid.grab()
# }
# heatmaps <- list(
#   CpG = one.heatmap(transitions.CpG, "CpG islands"),
#   Bg = one.heatmap(transitions.Bg, "Genomic background")
# )
# grid.newpage()
# grid.arrange(grobs = heatmaps, ncol = 2, clip = TRUE)
```

We can then use this log-odds matrix to compute the log-odds of a sequence as the sum of the transition log-odds.

$$L(S) = L_{\text{B}}(S_1) \cdot \sum_{i=1}^{n-1} L(S_{i+1}|S_i)$$

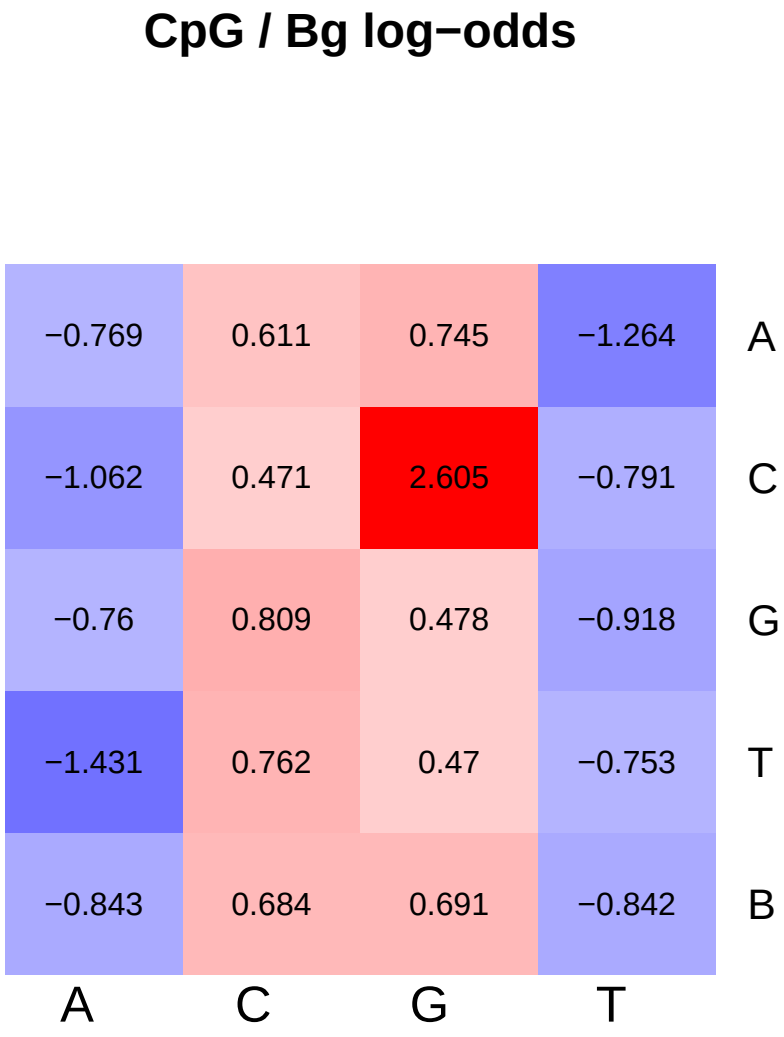

Figure 4: Heatmap of the log-odds between CpG and genomic background.

Hidden Markov Model

We already dispose of the emission probabilities of the residues in the two respective states (CpG islands and genomic background). We now need to compute the transition probabilities between the states.

We can estimate these probabilities based on the following parameters

- genome size: $S_{Bg} = 3 \times 10^9$
- total size of the annotated CpG islands: $S_{CpG} = 24,200,434$
- number of CpG islands: $N_{CpG} = 31,144$

```
## Estimate transition matrix for CpG islands

stats <- list(L.G = 3e+9, ## Genome size
              L.CpG = 24200434, # Total size of CpG islands (label: +)
              nb.CpG = 31144 # Number of CpG islands
)
stats$L.non.CpG <- stats$L.G - stats$L.CpG # Total size of non-CpG Islands (label: -)
stats$CpG.mean.length <- stats$L.CpG / stats$nb.CpG

library(knitr)
library(pander)
pander(as.data.frame(stats),
       caption = "Parameters used to compute the CpG island transition matrix.",
       big.mark = ",")
```

Table 2: Parameters used to compute the CpG island transition matrix.

L.G	L.CpG	nb.CpG	L.non.CpG	CpG.mean.length
3e+09	24,200,434	31,144	2.976e+09	777

```
## Estimate transition matrix for CpG islands

stats <- list(L.G = 3e+9, ## Genome size
              L.CpG = 24200434, # Total size of CpG islands (label: +)
              nb.CpG = 31144 # Number of CpG islands
)
stats$L.non.CpG <- stats$L.G - stats$L.CpG # Total size of non-CpG Islands (label: -)
stats$CpG.mean.length <- stats$L.CpG / stats$nb.CpG


## Transition matrix
transitions <- data.frame(matrix(nrow = 2, ncol = 2))
names(transitions) <- c("+", "-")
rownames(transitions) <- c("+", "-")

## Transitions from non-CpG to CpG.
## This is the number of switches from non-CpG to CpG island (the number of nucleotide preceding a CpG i
## divided by the total length of non-CPG islands
transitions["-", "+"] <- stats$nb.CpG / stats$L.non.CpG
## The only other possible transition from "-" is to "-" so the sum is 1 (complementary events)
transitions["-", "-"] <- 1 - transitions["-", "+"]

## Transitions from CpG to non-CpG
```

```
## Same reasoning: the number of ending CpG positions divided by the total length of CpGs
transitions["+", "-"] <- stats$nb.CpG / stats$L.CpG
## The only other possible transition from "+" is to "+" so the sum is 1 (complementary events)
transitions["+", "+"] <- 1 - transitions["+", "-"]
```

For convenience, we will use the labels + and − to denote the CpG islands and non-CpG island regions, respectively.

We can compute the probability of switching from a non-CpG islands to a CpG island with the following reasoning.

- The total size of non-CpG islands is the difference between genome size (3e+09) and the total size of the CpG islands (24,200,434).

$$S_- = S_{bg} - S_+ = 3e + 09 - 24,200,434 = 2,975,799,566$$

- In total, there are $31,144$ CpG islands in the genome, and each of them is preceded by a non-CpG island region. There are thus $31,144$ transitions from non-CpG to CpG. The probability of transition from non-CpG to CpG is thus the number of non-CpG positions preceding a CpG divided by the total number of non-CpG positions.

$$P(+|-) = N_+/S_- = \frac{31,144}{2,975,799,566} = 1.0466e - 05$$

- Since there are only two possible states, the transition probability from non-CpG to non-CpG is the complement of the transition probability from non-CpG to CpG.

$$P(-|-) = 1 - P(+|-) = 0.99999$$

The same reasoning applies to compute the transition probabilities from CpG islands.

- Each CpG island has an exactly one ending nucleotide, which precedes a non-CpG island nucleotide. The number of nucleotides marking a transition from CpG to non-CpG is thus $31,144$. The probability of transition from CpG to non-CpG is this number divided by the total size of all CpGs.

$$P(-|+) = N_+/S_+ = \frac{31,144}{24,200,434} = 0.00129$$

- The probability of transition from CpG to CpG is the complement.

$$P(+|+) = 1 - P(-|+) = 0.99871$$

The results are summarised in the transition matrix below.

```
kable(transitions, digits  = 5, caption = "Transition matrix between CpG islands and non-CpG island reg
```

Table 3: Transition matrix between CpG islands and non-CpG island regions in the human genome.

	+	-
+	0.99871	0.00129
-	0.00001	0.99999