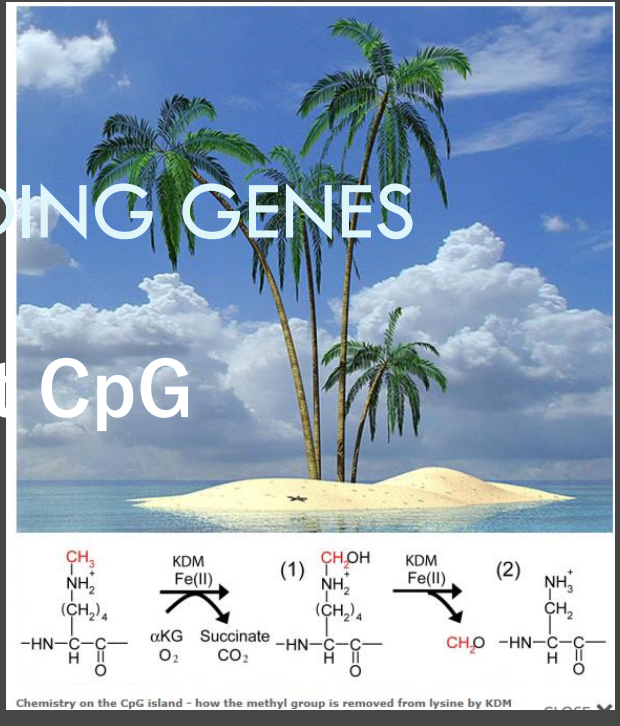
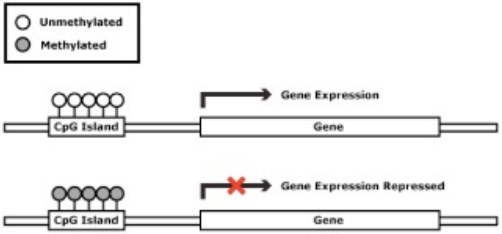
SESSION 12. FINDING GENES

Going ashore at CpG islands



CpG islands on genome ocean

- Human genome projects were completed back in 2003.
- What are the genetic information represented in the three billion bases ?
- Regions that give rise to RNAs that in turn encode proteins Protein-coding genes (1.5~2.0% for human) and non-coding genes. E.coli contains as much as 83% of coding sequence.
- Analogy to finding a needle in a haystack → **require a highly specific method**
- CpG island well defines the transcription start site (TSS)



	T.	F
+	TP	FP
-	FN	TN

SN = TP / (TP+FN)

SP = TN/ (FP+TN)

Eukaryotic transcription regulation

- Transcription of genes takes place with the help of the RNA polymerases
  - RNA pol I, II, and III
  - RNA pol II is responsible for the transcription of PCGs, snRNA, microRNA, and lncRNA genes.
  - RNA pol I is responsible for the transcription of rRNAs
  - RNA pol III is responsible for the transcription of tRNA, 5s rRNA,... other ncRNAs

Table 21-1 Properties of Eukaryotic RNA Polymerases

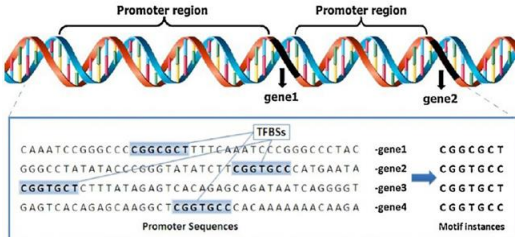
RNA Polymerase	Location	Main Products	α-Amanitin Sensitivity
I	Nucleolus	Precursor for 28S rRNA, 18S rRNA, and 5.8S rRNA	Resistant
II	Nucleoplasm	Pre-mRNA and most snRNA	Very sensitive
III	Nucleoplasm	Pre-tRNA, 5S rRNA, and other small RNAs	Moderately sensitive*
Mitochondrial	Mitochondrion	Mitochondrial RNA	Resistant
Chloroplast	Chloroplast	Chloroplast RNA	Resistant

*In mammals.

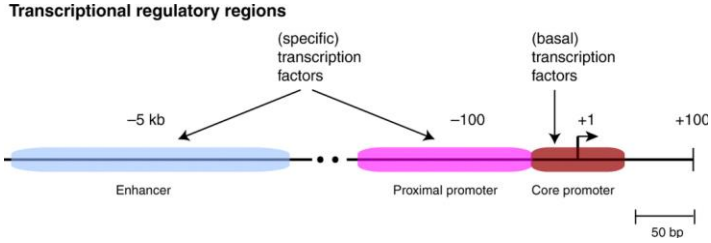
- Transcription factors are also required for initiation of RNA pol II.
  - General TF – for initiation of transcription of all protein genes
  - Specific TF – for regulation of subset of genes through enhancer or repressor

Cis-regulatory element of transcription

- Specific TF binding site (enhancer or repressor)
  - Ex, Oestrogen receptor recognizes a sequence "AGGTCANNNTGACCT"
  - Motif (k-mer) enrichment analysis in a specific set of genes
  - ChIP-seq (chromatin IP followed by sequencing)

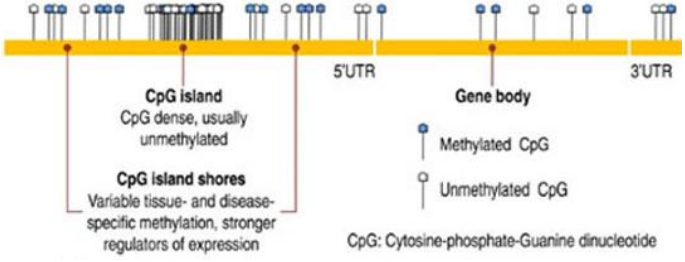


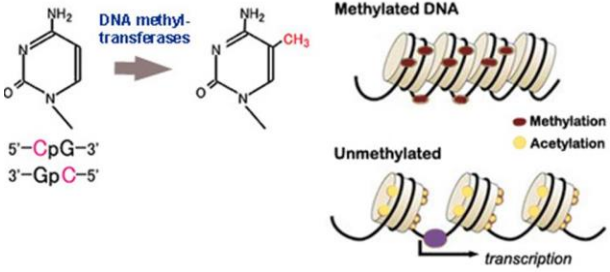
- General TF binding site (core + proximal promoter) *for identifying TSS*
  - TATA box: 10~25% of all human genes have this TATA box, recognized by TFIID



CpG islands

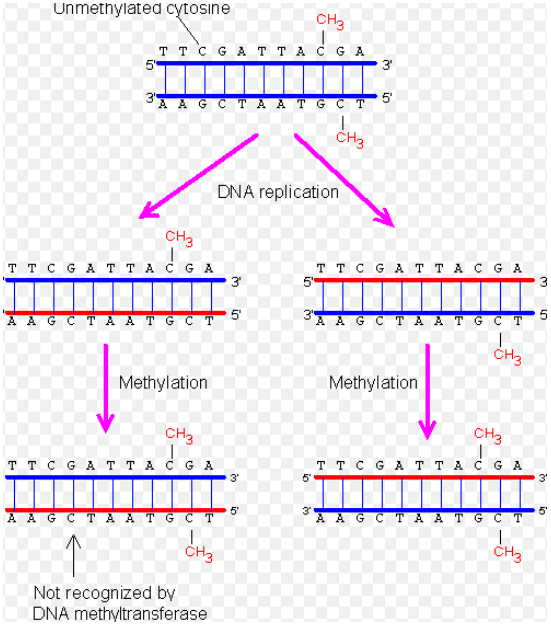
- A more common feature of transcription initiation regions are CpG islands
- The regions close to the TSS of ~60-70% of all human genes.
- CpG → C + phosphodiester bond + G (dinucleotide)
- CpG islands where the frequency of CpG sites is much higher than the background frequency of this dinucleotide
- C: 0.19 G:0.19 → CpG: 0.026 (observed) / 0.19X0.19=0.036 (expected)





CpG islands

- CpG → methylation on C → TpG mutation over evolution → depletion of CpG
- CpG near TSS → low methylation on C → high freq. of CpG
- The majority (70%) of CpG dinucleotides in a mammalian genome are methylated but the CpG near TSSs are typically not methylated.
- **Epigenetic inheritance**: the methylation pattern in a specific cell may be transmitted to the progeny cells.
- DNA methylation is also an important basis for genomic imprinting of either maternal or paternal origin.



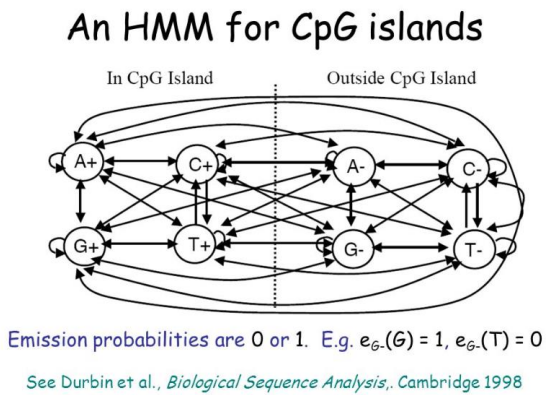
Finding CpG islands

- Original definition of CpG islands (Gardiner-Garden and Frommer, 1987)
  - A region at least 200 bp in length with a GC content of at least 50% and an observed CpG/expected CpG ratio greater than 0.6.
  - Drawback with this definition → many Alu repeats of the human genome will be incorrectly predicted as CpG (~a million copy of Alu elements)

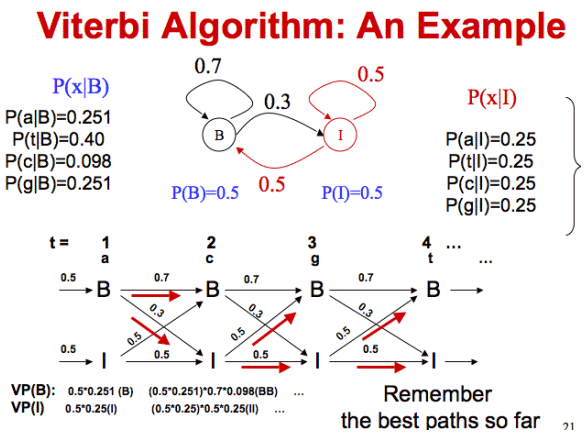
- Alternative definition of CpG islands (Takai and Jones 2002)
  - Important criterion for optimizing specificity in methods of CpG island
  - The total length of CpG island should be at least 500 nt
  - The G+C content should be at least 55%
  - The ratio b/w the frequency of observed CpG sites and the frequency of expected CpG sites should be at least 0.65.
  - $\frac{f_{CpG}}{f_C f_G} = \frac{SN_{CpG}}{N_C N_G}$
- Alternative prediction using HMM (Durbin 2007)

Finding CpG islands

- Alternative prediction using HMM (Durbin 2007)



Optimal path that maximizes the emission * the transition prob.



cpy.py

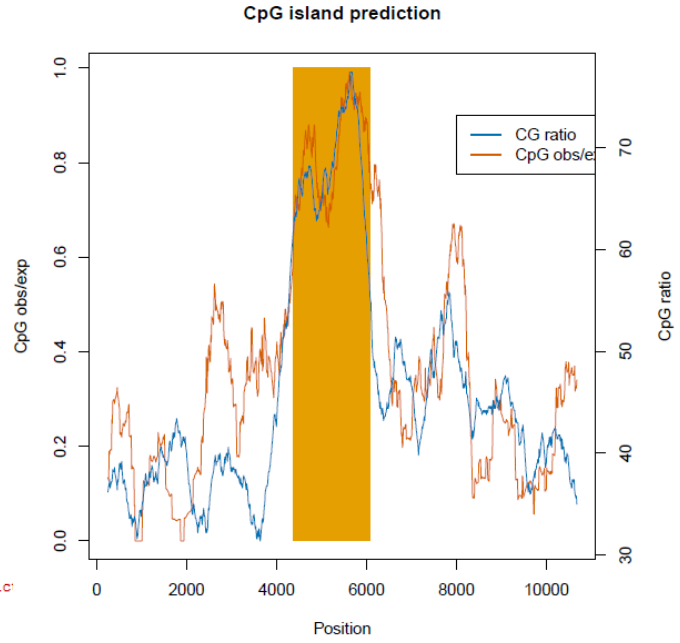
```
#!/usr/bin/python

import re, sys
filename = sys.argv[1]
win = 500
step = 10
seq = ''
for line in open(filename):  # a shorter sequence
    if not re.search('>', line):
        line = line.rstrip()
        seq = seq + line
print 'pos\tcpg\tcg_ratio\tcg_obs_exp'
for i in range(0, len(seq) - win, step):
    testseq = seq[i:i + win]
    c = float(testseq.count('C'))
    g = float(testseq.count('G'))
    cg = float(testseq.count('CG'))
    cg_ratio = (c + g) * 100 / len(testseq)
    cg_obs_exp = cg * len(testseq) / (c * g)
    pos = i + win / 2
    if cg_ratio >= 55 and cg_obs_exp >= 0.65:
        print pos, '\t', 1, '\t', cg_ratio, '\t', cg_obs_exp
    else:
        print pos, '\t', 0, '\t', cg_ratio, '\t', cg_obs_exp
```

```
[jwnam@biglab-master Session11]$ python cpg.py short.fa
pos     cpg     cg_ratio        cg_obs_exp
250     0       36.2    0.133547008547
260     0       36.4    0.131492439185
270     0       36.8    0.127567291746
280     0       37.0    0.189729319504
290     0       37.4    0.18315018315
300     0       36.6    0.191815856777
310     0       36.8    0.192604006163
320     0       37.0    0.192307692308
330     0       37.2    0.192110655738
340     0       37.2    0.256147540984
350     0       37.6    0.248385494287
360     0       38.4    0.294568163073
370     0       38.2    0.296982656213
```

Visualization with R

```
# plot the results of cpg island prediction

# define some colours
rgb <- c("#000000", "#E69F00", "#56B4E9", "#009E73",
    "#F0E442", "#0072B2", "#D55E00", "#CC79A7")

# read file being output from Perl script
data <- read.table("cpg_short.out", sep = "\t", header = TRUE)

# Specify the number of lines of margin to the four sides
# of the plot. In this case we want to make room for text
# at the right axis.

pdf("cpg_plot.pdf")
par(mar = c(5, 4, 4, 5) + 0.1)

# make first plot, which is empty
plot(data$pos, data$cg_obs_exp, type = "n", xaxt = "n",
    yaxt = "n", xlab = "", ylab = "")

# plot lines to indicate where CpG islands are predicted
for (i in 1:length(data$pos)) {
    if (data$cpg[i] == 1) {
        lines(c(data$pos[i], data$pos[i]), c(0, 1), col = rgb[2])
    }
}

# before another plot, prevent R from clearing
# the graphics device
par(new = TRUE)

# make second plot with the cg_obs_exp data
plot(data$pos, data$cg_obs_exp, type = "l", main = "CpG island predic
    xlab = "Position", ylab = "CpG obs/exp", col = rgb[7])

# print legend
legend(8000, 0.9, c("CG ratio", "CpG obs/exp"), col = c(rgb[6],
    rgb[7]), lwd = 2)
```

```
par(new = TRUE)
```



Scale up for the search

```r
# plot the results of CpG island prediction

#define some rgb colours
rgb <- c("#000000", "#E69F00", "#56B4E9", "#009E73",
    "#F0E442", "#0072B2", "#D55E00", "#CC79A7")

# read data which is output of Perl script
data <- read.table("cpg_chr4.out", sep = "\t", header = TRUE)
pdf("cpg_chr4.pdf")
# make an emtpy plot
plot(0, type = "n", xlim = c(72009075, 72009075 +
    1520980), ylim = c(1.6, 2.3), xlab = "Position", ylab = "",
    yaxt = "n", main = "CpG island prediction")

# plot the predicted CpG islands
for (i in 1:length(data$pos)) {
    if (data$cpg[i] == 1) {
        # convert the position numbers to chromosomal positions
        data$pos[i] <- data$pos[i] + 72009075
        #print (data$pos[i])
        lines(c(data$pos[i], data$pos[i]), c(1.7, 1.8), col = rgb[7])
    }
}

# Read file with chr4:72,009,075-73,530,562 region
# annotation. This information was obtained with
# the Table Browser of the UCSC browser

annot <- read.table("chr4_annotation.txt", sep = "\t",
    header = TRUE)

color <- 1
prevname <- ""
lines <- length(annot$chr)  # number of lines in the annotation file
```

```r
for (i in (1:lines)) {

    if (annot$category[i] == "exon") {
        # if we consider an exon
        # if it a different gene as compared to the previous line,
        # change the colour
        if (annot$gene_id[i] != prevname) {
            color <- color + 1
        }
        prevname <- annot$gene_id[i]

        # draw rectangles for the exons
        rect(annot$beg[i], 1.9, annot$end[i], 2.1, border = rgb[color],
```