

Estimation of epimutation rates using Markov Chains and the EM-algorithm

Topics in computational biology

Olle Holmberg
Cristina Cipriani

Supervisor: Dr. Maria Colome-Tatche

Content

- Biological introduction
- Data
- Discrete-time homogeneous Markov model
- Estimation: Missing Data and the EM-algorithm
- Statistical significance: Bootstrap
- Results
- Discussion

Biological introduction

Epigenetics is the study of stable changes in gene expression that are NOT caused by changes in the DNA sequence. One of the main mechanism is **DNA methylation** which consists of chemical modifications of a cytosine base in the genome.

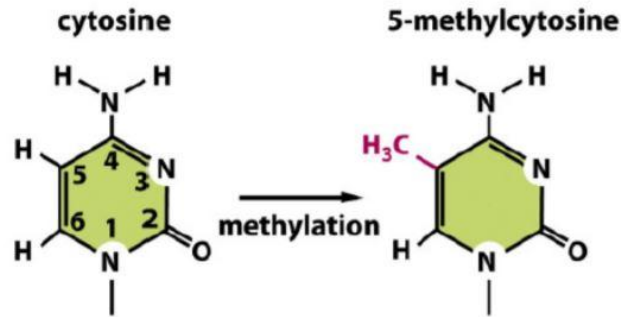
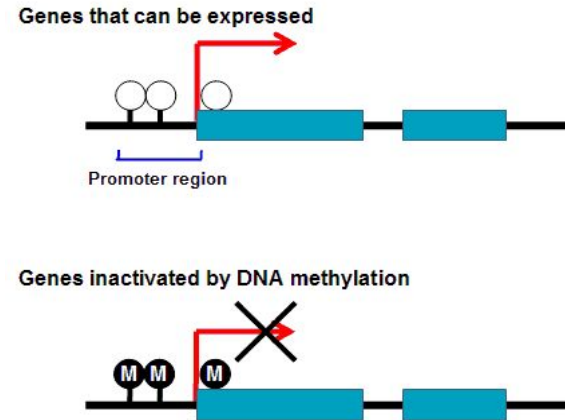


Figure 7-79 *Molecular Biology of the Cell* (© Garland Science 2008)

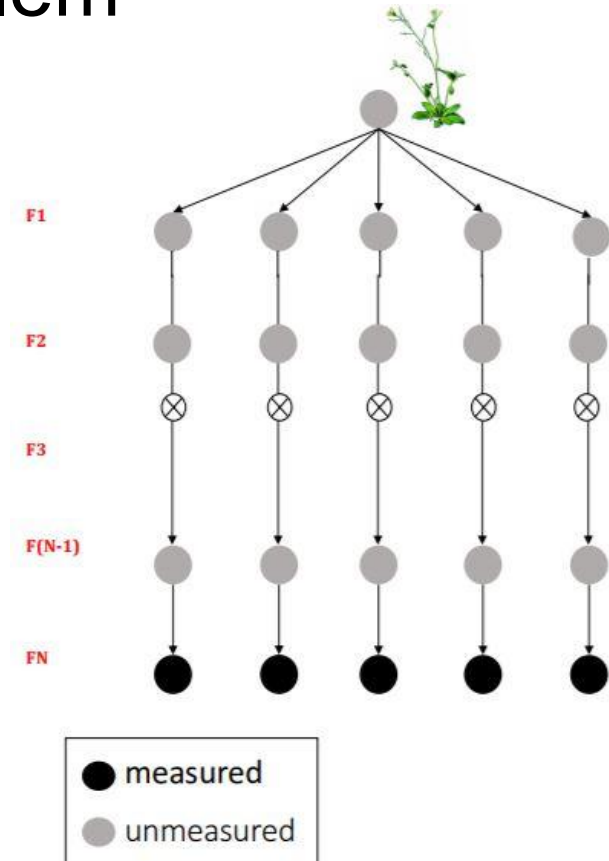


Methylation losses or gains can emerge spontaneously in an event termed **epimutation**. Cytosine methylation is maintained by different biological pathways in three different **contexts** (CG, CHG, CHH) where context tell us if the cytosine lie next to a guanine base or not.

Project start: previous problem

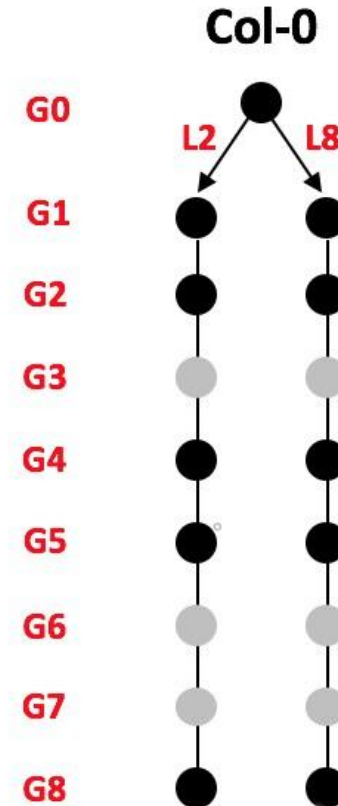
Data with unmeasured founder plant.

Estimated epimutation rates through measuring differences between measured plants.



Project start: our problem

Data with measured founder plant.
Only two lines and eight generations.



In order to study methylation dynamics in this population, we construct what is called **longitudinal data sets**. These datasets contain, for every cytosine in the genome, the sequence of methylated "1" and unmethylated "0" state at every generation. For the generations where no data was obtained, methylation status is labelled NA.

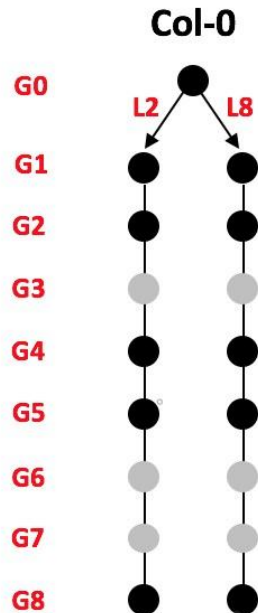


Table 1. Data format of one contex (E.g. CG)

cytosine	G0	G1	G2	G3	G4	G5	G6	G7	G8
1	1	1	0	NA	0	1	NA	NA	0
2	1	1	0	NA	0	1	NA	NA	0
.	-	-	-	-	-	-	-	-	-
.	-	-	-	-	-	-	-	-	-
N	1	1	0	NA	0	1	NA	NA	0

Discrete-time homogeneous Markov model

A **Markov chain** is a discrete-time stochastic process $(X_n, n \geq 0)$ such that each random variable X_n takes values in a discrete set S and

$$\mathbb{P}(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0) = \mathbb{P}(X_{n+1} = j \mid X_n = i) \\ \forall n \geq 0, j, i, i_{n-1}, \dots, i_0 \in S$$

Moreover, if $P(X_{n+1} = j \mid X_n = i) = p_{ij}$ is independent of n , then X_n is said to be a time **homogeneous Markov chain**.

The interval between these time points is known as the **cycle length**.

In our case the cycle length we are interested in is one generation.

Table 1. Data format of one contex (E.g. CG)

cytosine	G0	G1	G2	G3	G4	G5	G6	G7	G8
1	1	1	0	NA	0	1	NA	NA	0
2	1	1	0	NA	0	1	NA	NA	0
.	-	-	-	-	-	-	-	-	-
.	-	-	-	-	-	-	-	-	-
N	1	1	0	NA	0	1	NA	NA	0

Estimation: the EM-algorithm

The EM algorithm is an iterative method to find maximum likelihood estimates of parameters in statistical models, where the model depends on unobserved variables.

In our case the algorithm is the following:

- Start with an approximation of the transition matrix;
- **E-step**: impute the states in the unobserved cycles, considering all the possible transitions and calculate the expected number of transitions;
- **M-step**: apply the maximum likelihood method to obtain a new estimate of the transition matrix.
- Repeat until convergence.

Since convergence is not guaranteed, we use the algorithm with different initial transition matrices to verify that one initialization does not get stuck at a local maximum.

The E-step:

Count the number of transitions for each different interval of length 1,2,3:

Table 1. Data format of one context (E.g. CG)

cytosine	G0	G1	G2	G3	G4	G5	G6	G7	G8
1	1	1	0	NA	0	1	NA	NA	0
2	1	1	0	NA	0	1	NA	NA	0
.	-	-	-	-	-	-	-	-	-
.	-	-	-	-	-	-	-	-	-
N	1	1	0	NA	0	1	NA	NA	0




N₁:	0	1
0	11774	148
1	140	4850



N₃:	0	1
0	3900	41
1	38	1625



N₂:	0	1
0	3874	63
1	38	1629


$$n_{00} = N_1(0,0) + 2 N_2(0,0) \left(\frac{\theta_{00}\theta_{00}}{\theta_{00}\theta_{00} + \theta_{01}\theta_{10}} \right) + \dots$$


where θ_{ij} is the probability of going from i to j , where $i, j \in \{0,1\}$

All possible paths from 0 to 0 in the cycle length 2:

$0 \rightarrow 0 \rightarrow 0$


$0 \rightarrow 1 \rightarrow 0$

$$n_{00} = N_1(0,0) + 2 N_2(0,0) \left(\frac{\theta_{00}\theta_{00}}{\theta_{00}\theta_{00} + \theta_{01}\theta_{10}} \right) + N_2(0,1) \left(\frac{\theta_{00}\theta_{01}}{\theta_{00}\theta_{01} + \theta_{01}\theta_{11}} \right) + \dots$$


All possible paths from 0 to 1 in the cycle length 2:

$0 \rightarrow 0 \rightarrow 1$

$0 \rightarrow 1 \rightarrow 1$

$$n_{00} = N_1(0,0) + 2 N_2(0,0) \left(\frac{\theta_{00}\theta_{00}}{\theta_{00}\theta_{00} + \theta_{01}\theta_{10}} \right) + N_2(0,1) \left(\frac{\theta_{00}\theta_{01}}{\theta_{00}\theta_{01} + \theta_{01}\theta_{11}} \right) + N_2(1,0) \left(\frac{\theta_{10}\theta_{00}}{\theta_{10}\theta_{00} + \theta_{11}\theta_{10}} \right) + \dots$$


All possible paths from 1 to 0 in the cycle length 2:

$1 \rightarrow 0 \rightarrow 0$

$1 \rightarrow 1 \rightarrow 0$

$$n_{00} = N_1(0,0) + 2 N_2(0,0)\left(\frac{\theta_{00}\theta_{00}}{\theta_{00}\theta_{00}+\theta_{01}\theta_{10}}\right) + N_2(0,1)\left(\frac{\theta_{00}\theta_{01}}{\theta_{00}\theta_{01}+\theta_{01}\theta_{11}}\right) + N_2(1,0)\left(\frac{\theta_{10}\theta_{00}}{\theta_{10}\theta_{00}+\theta_{11}\theta_{10}}\right) + N_2(1,1)(0) + \dots$$



All possible paths from 1 to 1 in the cycle length 2:

$1 \rightarrow 0 \rightarrow 1$

$1 \rightarrow 1 \rightarrow 1$

$$n_{00} = N_1(0,0) + 2 N_2(0,0)\left(\frac{\theta_{00}\theta_{00}}{\theta_{00}\theta_{00}+\theta_{01}\theta_{10}}\right) + N_2(0,1)\left(\frac{\theta_{00}\theta_{01}}{\theta_{00}\theta_{01}+\theta_{01}\theta_{11}}\right) + N_2(1,0)\left(\frac{\theta_{10}\theta_{00}}{\theta_{10}\theta_{00}+\theta_{11}\theta_{10}}\right) + N_3(0,0)\left(\frac{\theta_{01}\theta_{10}\theta_{00} + \theta_{00}\theta_{01}\theta_{10}}{\theta_{00}\theta_{00}\theta_{00} + \theta_{00}\theta_{01}\theta_{10} + \theta_{01}\theta_{10}\theta_{00} + \theta_{01}\theta_{10}\theta_{00} + \theta_{01}\theta_{11}\theta_{10}}\right) + \dots$$



All possible paths from 0 to 0 in the cycle length 3:

$0 \rightarrow 0 \rightarrow 0 \rightarrow 0$

$0 \rightarrow 1 \rightarrow 0 \rightarrow 0$

$0 \rightarrow 0 \rightarrow 1 \rightarrow 0$

$0 \rightarrow 1 \rightarrow 1 \rightarrow 0$

Statistical significance: Bootstrap

If we measure differences in different parts of the genome: How do we know that the differences are statistically significant?

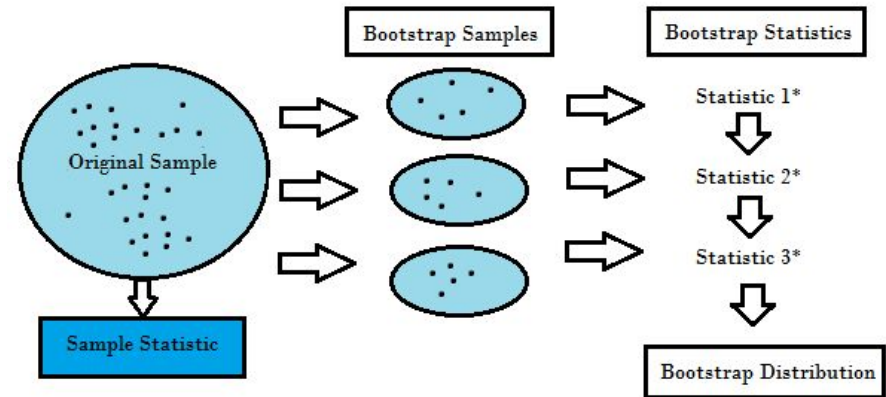
Problem: no easy statistical distribution for which we can quantify our uncertainty!

Solution: Bootstrap

Statistical significance: Bootstrap

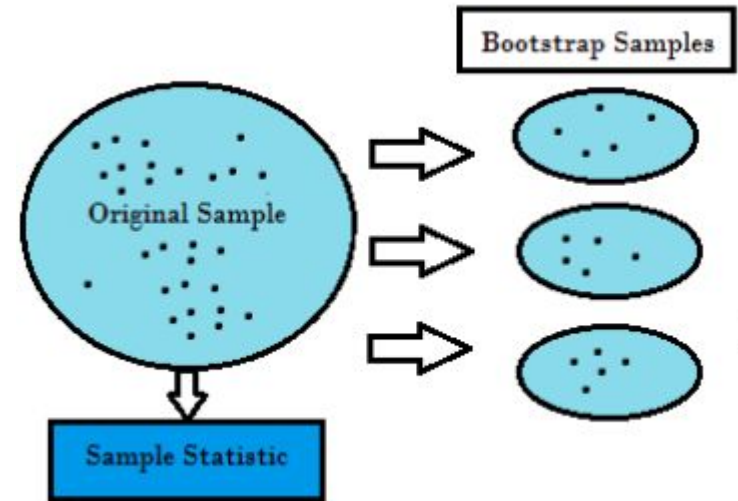
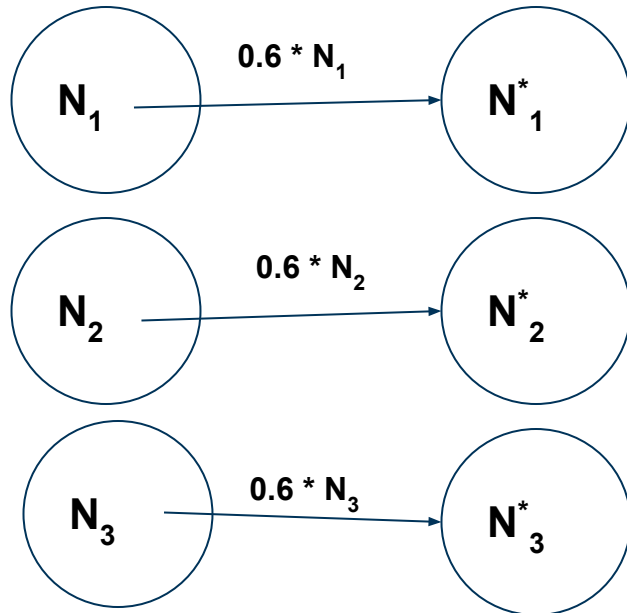
- How Bootstrap works:
 - create several different data sets by random sub-sampling our original data
 - Evaluate our statistics on each new data set

- How does this work for our data?



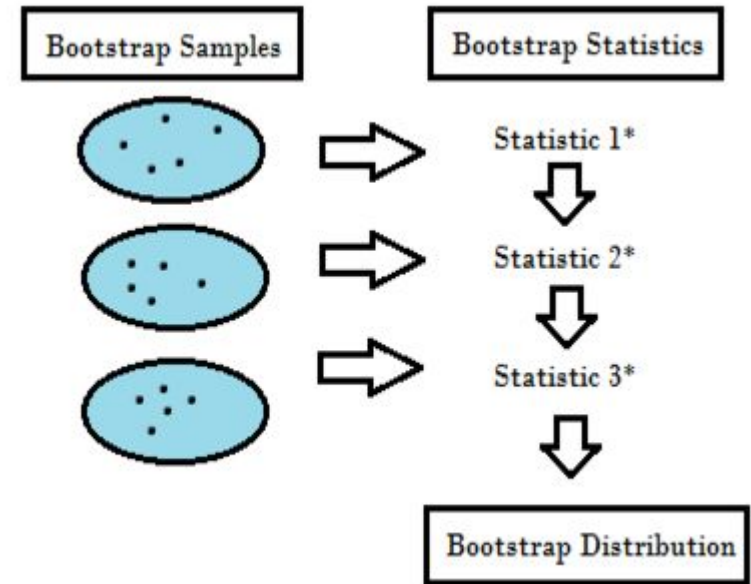
Statistical significance: Bootstrap

- Randomly subsample using a multinomial distribution with probabilities from our sample statistic creating 1000 random bootstrap samples



Statistical significance: Bootstrap

- Once one 1000 sets of \mathbf{N}_1^* , \mathbf{N}_2^* , \mathbf{N}_3^* are obtained we run the EM-algorithm and obtain 1000 sets of estimates.
- Now we can measure uncertainty and construct confidence intervals



Results

- From table:
overall probability of gaining methylation is less than losing methylation
- From boxplot:
there are clear differences between the contexts
- All differences are statistically significant

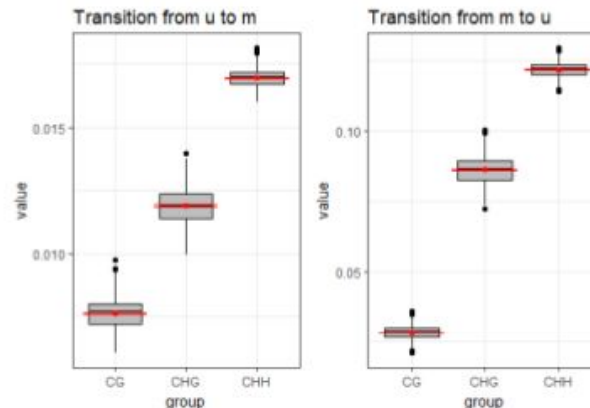


Table 2. Estimation results in percent

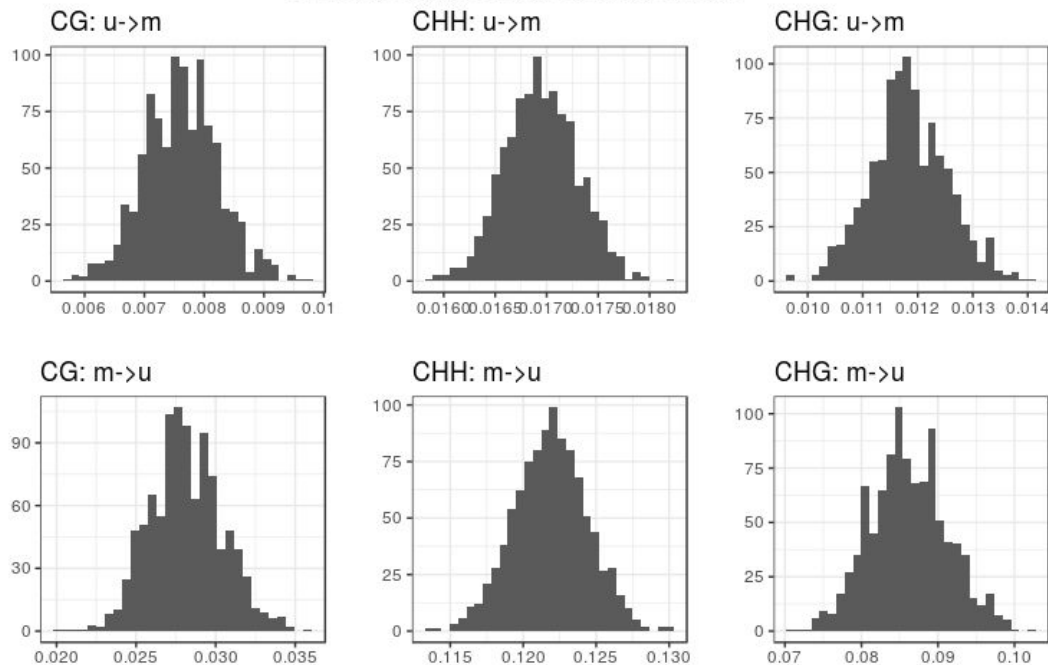
context	Loss	C.I	Gain	C.I
CG	2.83	[2.81;2.84]	0.46	[0.46;0.47]
CHH	12.33	[12.31;12.34]	1.07	[1.06;1.07]
CHG	8.70	[8.67;8.74]	0.73	[0.73;0.74]

the C.I's are the 95 percent CI's derived using the normal distribution

Results

- Bootstrap distributions:
- Seem normal
- Motivate our use of normal confidence intervals (C.I.)

Bootstrap estimates for the different contexts



Discussion

- Probability of gaining methylation is smaller than probability of losing
 - Consistent with previous studies on the topic
 - Shows that our model/methodology does not deviate from what we expect
- Differences between contexts CG,CHH,CHG
 - Biology tells us that methylation of the genome is maintained through 3 different pathways depending on the context of the cytosine
 - The order $CG < CHG < CHH$ explained by Biology
 - CG methylation is very stably maintained upon cell division; CHG and CHH are less stably maintained (although CHG more stable than CHH)
- Consistency of results speaks for applicability of our “fairly” straightforward approach

Final comments/conclusions

- Fun and challenging!
- Learned a lot about:
 - Biology and Epigenetics
 - Markov models, EM-algorithm, Bootstrap and R-programming
 - Interdisciplinary project work
 - Solving a problem from scratch (theory and programming)

Thanks to **Dr. Maria Colome-Tatche** for all the support!