Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

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

Olle Holmberg [1], Cristina Cipriani [2]

Abstract

Motivation: Stochastic changes in cytosine methylation are a source of heritable epigenetic and phenotypic diversity in plants. Using the model plant Arabidopsis thaliana, we derived estimates of the rate at which methylation is spontaneously gained (forward epimutation) or lost (backward epimutation) transgenerationally at context CG, CHH and CHG. Our results indicate significant differences in epimutation rates between the three studied contexts, reflecting the different molecular pathways that are responsible for DNA methylation maintenance in the different contexts.
Contact: ga85gac@mytum.de, cristina.cipriani@tum.de
Supplementary information: R code available at https://github.com/Gottfrid91/epimutation_analysis_R

1 Introduction

Epigenetics is the study of stable changes in gene expression not caused by changes in the DNA sequence. One of the main mechanisms of changes in gene expressions is DNA methylation, which consists of chemical modifications of a cytosine base in the genome. Such modifications can in addition change the expression of the genes in the neighbouring regions.
For this reason, plant genomes make extensive use of cytosine methylation to control the expression of transposable elements (TEs) and genes. Despite its tight regulation, methylation losses or gains at individual cytosines or clusters of cytosines can emerge spontaneously, in an event termed "epimutation" (1). Many examples of segregating epimutations have been documented in experimental and wild populations of plants and in some cases contribute to heritable variation in phenotypes independently of DNA sequence variation.
The aim of this project is to estimate these losses and gains in methylation for the plant Arabidopsis thaliana using plants that have been propagated for many generations in a perfect environment, so called mutation accumulation or MA lines. Obtaining precise estimates of these rates is necessary to to be able to quantify the long-term dynamics of epigenetic variation under laboratory or natural conditions, and to understand the molecular mechanisms that drive methylome evolution.

2 Methods

2.1 Data

DNA methylation can be measured experimentally using a technique called whole genome bisulfite sequencing. This technique produces a measure that allows us to determine how likely is every cytosine in the genome to be methylated or unmethylated. The data used for this project is the methylation calls, noted as 1 (for methylated) or 0 (for unmethylated) for each cytosine in the genome.
In order to study the transgenerational stability of DNA methylation we analyze data from an MA line that has been propagated from a founder for 8 generations (Figure 1). From these generations, methylation was measured in generations G0 (the founder), G1, G2, G4, G5 and G8. For all analysis we split the cytosines in the genome depending on what "context" they belong to. We consider three different contexts: CG, CHG and CHH (where H = anything but a G). We partition the genome this way because cytosine methylation is maintained by different biological pathways in these three contexts (3), which could lead to different stability rates between them.
In order to study methylation dynamics in this population, we construct what is called longitudinal data sets (Table 1). These datasets contain, for every cytosine in the genome, the sequence of methylated "1" and unmethyated "0" state at every generation. For the generations where no data was obtained, methylation status is labelled NA.
The goal is to estimate the probabilities to transition from unmethylated to methylated and vice versa using the Discrete-time homogeneous Markov chain framework.

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

2.2 Discrete-time homogeneus Markov model

A Markov chain is a discrete-time stochastic process (X_n, n ≥ 0) such that each random variable X_n takes values in a discrete set S (S = $\mathbb{N}$, typically) and

$$P(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, X_{n-2} = i_{n-2}, ..., X_0 = i_0) = P(X_{n+1} = j \mid X_n = i) \quad \forall n \geq 0, j, i, i_n, ..., i_0 \in S$$

That is, as time goes by, the process loses the memory of the past.
Moreover, if P(X_{n+1} = j | X_n = i) = p_{ij} is independent of n, then X_n is said to be a time homogeneous Markov chain. These probabilities constitute the transition matrix of the chain defined as P = (p_{ij}) i,j $\in$ S.
A Markov chain is a simple yet powerful model to describe progression through states, easy to construct and study through matrix analysis. The typical discrete-time Markov chain limits the description of each subject′s history to equally spaced time points instead of modeling the possibility of progression at every instant in time. The interval between these time points is known as the cycle length. This is often set to an interval associated with the model and inference of the transition matrix is drawn from observational cohort data where each subject is observed at common intervals.
For the epimutation rate data, DNA methylation is measured at defined discrete time points, which are the different generations, the cycle lenght is defined as one generation. Then, the subjects are the individual cytosines which are observed at different discrete generation times.

The data for this project contains missing values for generation 3, 6, and 7, see table 1. This means that the data contains unequal observation intervals. One could use only generations 0,1,2 and 4,5 to estimate the transition probabilities but this is not ideal as it would mean to throw away useful data. Instead, it is appropriate to handle missing values by imputation. This can be achieved in an EM-algorithm, discussed below.

2.3 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.
The EM iteration alternates between performing an expectation (E) step, which creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters, and a maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E step. This is repeated until the result is stable.

For this project, this involves imputing the states in the unobserved cycles, in the E-step, tallying the expected number of transitions, and then apply the maximum likelihood method, M-step, to obtain a new estimate of the transition matrix. We retrieve our initial matrix values for the EM-algorithm M_1, M_2 and M_3 by calculating the transition matrices of 1 cycle transitions (calculated between generations 0 and 1, generations 1 and 2 and generations 4 and 5), of 2 cycle transitions (calculated between generations 2 and 4), and of 3 cycle transitions (calculated between generations 5 and 8), then taking $M_2^{1/2}$ and $M_3^{1/3}$ to approximate the corresponding one cycle transition probabilities. This enables verification that one initialization does not get stuck at a local maximum and that the algorithm converges.

2.4 Confidence intervals using a bootstrap approach

In addition to the matrix estimate itself, one is often interested in summaries that are functions of this matrix. To assess the uncertainty and construct

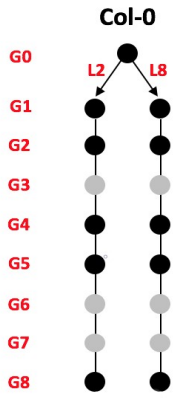


Fig. 1. Schematic of two mutation accumulation lines, where every dot represents a plant in a given generation (from G0 to G8) and the black lines represent a parent-to-offspring relationship. In our analysis we consider line L2 and measured generations are marked in black, while unmeasured generations are marked in gray.

confidence intervals of these summaries, the bootstrap is suggested (2). While sensitivity analysis, which involves varying single or multiple transition parameters, is frequently used to investigate the behavior of the transition matrix, it should not be used to form confidence intervals because it does not adequately account for the complex relationship among the transition probabilities(2).

For this purpose, Efrons bootstrap is recommended. With this method, other possible data sets, the same size as the original, are formed by sampling with replacement from the original data set. This is done by addressing each row of the transition count matrix separately. Letting n_r denote the total number of transitions for row r, bootstrapping row r simply involves sampling n_r transitions with replacement from the observed n_r transitions. In other words n_r draws are taken from a Multinomial distribution with probabilities derived from the transition matrix to generate a new set of transition counts for row r.(2) Combining the results of each row forms a new transition count matrix and thus another possible transition probability matrix M^* .
The collection of bootstrapped transition matrices approximates the sampling distribution. From this distribution, one can assess the uncertainty of each probability using the standard deviation between the collection of bootstrap estimates.

3 Results

In order to estimate epimutation rates we implemented an EM algorithm as the one described in (2). For the implementation we used the programming language R.
Briefly, we calculated all the observed one-cycle transitions between methylation states, as well as all the two-cycle and three-cycle transitions. We then estimated the most likely one-cycle transitions as the number of observed one-cycle transition plus the contribution to every one-cycle transition from the two- and three-cycle transitions with the current update of the transition matrix. With these summarized one-cycle transition values, we applied maximum likelihood to update the transition matrix. After running multiple initialization of the EM-algorithm, ensuring for convergence, we can conclude that the results clearly show a difference in loss and gain of methylation between the different contexts.
The results of the estimations are the following:

Table 2. Estimation results

context	Loss	C.I	Gain	C.I
CG	0.0282	0.0046	0.0076	0.0012
CHH	0.1219	0.0025	0.0169	0.0008
CHG	0.0861	0.0102	0.0118	0.0014

the C.I's are the 95 percent CI's derived using the t-distribution

The differences in probability for losses (from m to u) and gains (from u to m) in methylation are displayed per context below.

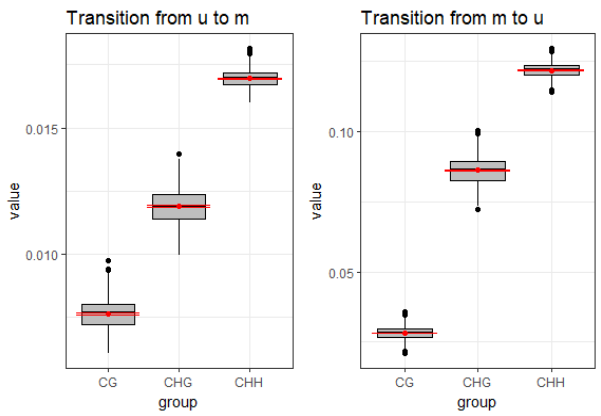
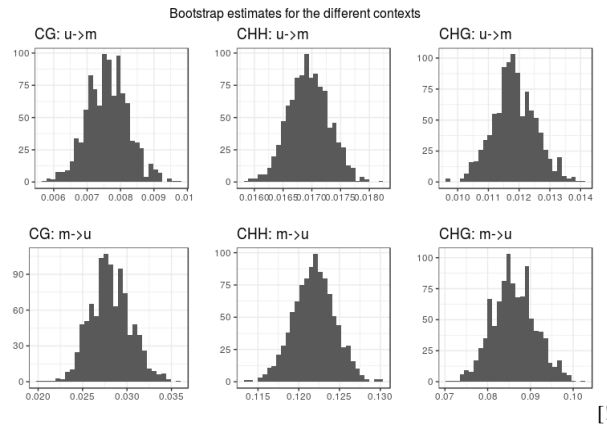


Fig. 2. Box plots of the estimated gains and losses of methylation for each context

The estimates indicate significant differences in loss and gain of methylation between contexts, and per context also significant differences between methylation gains and losses (see standard error in red). Viewing the three estimates and their bootstrap distributions, see plot below, it can be observed that the bootstrap estimates follow an approximate Normal distribution which motivates constructing C.I's using the t-statistic.



[!h]

Fig. 3. Bootstrap results for the estimated gains and losses of methylation for each context

4 Discussion

We may notice from the table 2 and the figure 2 that the estimated transition probabilities of loosing methylation are smaller than the ones of gaining methylation in all of the three contexts. This is consistent with the previous studies of the topic (1) and with the fact that the genome is mainly unmethylated. Previous studies on the same plant showed that in Arabidopsis thaliana, genome-wide DNA methylation levels of approximately 24%, 6.7% and 1.7% were observed for CG, CHG and CHH contexts, respectively, and the biological reason is that in plants methylation in the different contexts is maintained by three different pathways. Methylation at symmetrical CG and CHG nucleotide groups is maintained during subsequent rounds of DNA replication through the action of maintenance methyltransferases (enzimes that catalyze the transfer of a methyl group) which recognize methylation on one DNA strand but not the other. Hence our result are biologically reasonable since they also show that the methylation probabilities are statistically different for all of the three contexts and that in the context CG it's more likely to not have transitions since the methylation of the cytosines is simply duplicated without any stochasticity. According to the fact that CHG is maintained by a different set of enzymes than CG but that is more faithful at producing non-noisy methylation states compared to the enzymes that target CHH, our results show that in this context the probabilities of transition are higher than in the context CG, but still lower than CHH. Finally,in the case of CHH which does not rely on copying DNA methylation status on the other strand and leads to more noise in the methylation levels, the measured transition probabilities are higher.

The results achieved using the EM-algorithm and Markov chain formulation seem biologically reasonable, thus motivating the suitability for using this statistical framework when estimating epimutation rates.

Acknowledgements

We thank our supervisor Dr. Maria Colome-Tatche for introducing us to all Biological context of this problem and helping to provide interpretation of results through a biologial perspective.

References

1. Rate, spectrum, and evolutionary dynamics of spontaneous epimutations, Adriaan van der Graaf, RenÃ© Wardenaar, Drexel A. Neumann, Aaron Taudt, Ruth G. Shaw, Ritsert C. Jansen, Robert J. Schmitz, Maria ColomÃ©-TatchÃ©, and Frank Johannes, a Groningen Bioinformatics Centre, University of Groningen, 9747 AG Groningen, The Netherlands; b Department of Genetics, University of Georgia, Athens,GA 30602; c European Institute for the Biology of Aging, University of Groningen, University Medical Centre Groningen, 9713 AV Groningen, The Netherlands; and Department of Ecology, Evolution and Behavior, University of Minnesota, Minneapolis, MN 55455

2.Estimation of the transition matrix of a discrete-time Markov chain, Bruce A. Craig a, * and Peter P. Sendi,Department of Statistics, Purdue University, West Lafayette, USA, Internal Medicine Outpatient Department, University of Basel, Switzerland

3. Law, J.A., Jacobsen, S.E.: Establising, maintaining and modifying DNA methylation patterns in plants and animals. Nat Rev Genet. 11(3), 204âŁ“220 (2010). doi:10.1038/nrg2719.Establishing

York, 1993.