EPiC Series in Computing

Volume 104, 2024, Pages 51–83

Proceedings of 3rd International Workshop on
Mathematical Modeling and Scientific Computing

Computational and Analytical Approaches for DNA Methylation Pattern Modeling

Yifan Chen[1], Veronika Hofmann[2], Anneliese Riess[3,4], Tarini Singh[5], Suman Majumder[6], and Sona John[7]

[1] Technical University of Munich, Munich, Germany
yifan.chen@tum.de
[2] Technical University of Munich, Munich, Germany
veronika.hofmann@ma.tum.de
[3] Helmholtz Munich, Neuherberg, Germany
anneliese.riess@helmholtz-munich.de
[4] Technical University of Munich, Munich, Germany
[5] Helmholtz Munich, Neuherberg, Germany
tarini.singh@helmholtz-munich.de
[6] Amity University Uttar Pradesh, Noida, India
smajumder@amity.edu
[7] Technical University of Munich, Munich, Germany
sona.john@tum.de

Abstract

DNA methylation is a modification of the biochemical environment of a nucleotide that can occur at so-called CpG sites in the DNA strand. Just as a genetic mutation, it can benefit or harm the organism, depending on where exactly it happens and to what extent. This work focuses on two questions regarding the pattern evolution of methylation in certain DNA sequences, since the impact of methylation has been observed to depend on these patterns: does the size of (de-)methylated CpG clusters depend on reactions with other CpG sites? And can these reactions alter epigenetic variation, i.e. population-wide methylation patterns? To describe the methylome evolution within one individual (on a single cell basis), but also inter-generational developments, we formulate two mathematical models and corresponding master equations: one considering the influence of a single neighboring CpG site and one regarding both nearest neighbors. As the master equations can only be solved for certain parameter values, we use numerical simulations for further analysis. The simulation is compared to the analytical solution for validation, and then it is used for the investigation of the aforementioned questions. We find that for the chosen parameters, the cluster size increases if neighboring interactions are involved, independently of methylation status. Our results suggest that the epigenetic variation is larger in the case of the models which include neighboring interactions.

1 Introduction

Methylation of nucleotides is a molecular way to reversibly mark genomic deoxyribonucleic acid (DNA). This phenomenon occurs on the cytosines of CpG dinucloetides in vertebrates and plants. A CpG site is a DNA region where a cytosine nucleotide is followed by a guanine nucleotide in the linear sequence of bases along its 5' → 3' direction. In the case of plants, cytosine can also be methylated in other settings, but the methylation of cytosine at a CpG site is the most frequent scenario [13]. A methylated cytosine molecule has one hydrogen atom (H) replaced by a methyl group (H_3C) [18].

As a form of epigenetic mutation, methylation is one of the main influencing factors of gene expression in both, plants and vertebrates. It can benefit the individual who is the recipient of such a mutation, e.g. by improving resistance against certain diseases [19], but – and this is generally the more likely case – it can harm the affected subject and lead to illnesses such as cancer [17]. Methylations of plant and animal DNA are either inherited or occur newly during one individual's lifetime, for example as a consequence of certain lifestyle decisions or environmental influences, as well as during the process of ageing [4, 16, 17]. There are observations particularly in tumors suggesting that accumulations of methylated CpG sites attract other sites to methylate as well, whereas in regions where methylations are rather rare they do not occur as much [27]. A similar effect is reported for plants [20]. We call this effect neighboring interaction, as it is a matter of mutual influence by CpG sites in the neighborhood of each other. See Fig. 1 for a schematic depiction of methylation and the neighboring interactions. Methylations cause the largest impact when they occur cumulatively, in so-called clusters. Clusters are agglomerations of CpG sites with the same methylation status (either methylated or "demethylated", meaning that the site of interest is not methylated) that are not interrupted by sites of a different status.

Due to its pathologic consequences, but also possibilities, this clustering effect obviously provides a motivation for biochemists to investigate methylation mechanisms more thoroughly, but the evolution of methylation patterns is an interesting task in the field of mathematical modeling and scientific computing as well. The probabilistic nature of the neighboring interactions and the resulting reactions serve as a suitable area of application for master equations and stochastic simulations. We formulate our first hypothesis.

Neighboring interactions create larger clusters than exclusively spontaneous processes, i.e. methylations and demethylations without any influence from nearby sites.

As mentioned before, methylation can be – and most of the time *is* – inherited, at least over multiple cell divisions [12]. In the field of genetic inheritance, there is the phenomenon of genetic linkage, meaning that genes which lie close to each other are more likely to be inherited together [21]. It is also possible that genetic inheritance happens without this linkage, which means that the chance for a gene to be passed onto the next generation depends solely on its frequency in the parents, not on the genes close to it and their chances to be inherited. Both phenomena have been observed in epigenetic inheritance as well [8, 9, 10].

"Methylome" is the term which describes the state of methylation for all CpG sites in one genome. When comparing individuals of a population, their methylomes can be very similar (a sign for low epigenetic variation), or very different (high epigenetic variation). Just like genetic variation, a high epigenetic variation is an important evolutionary property as it can help a population to develop and then spread useful (epi-)mutations, and it makes it easier to avoid inbreeding. We speculate that the maintenance of this high variation is the reason for the inheritance of methylations, and formulate our second hypothesis.

Neighboring interactions combined with unlinked inheritance lead to more epigenetic vari-

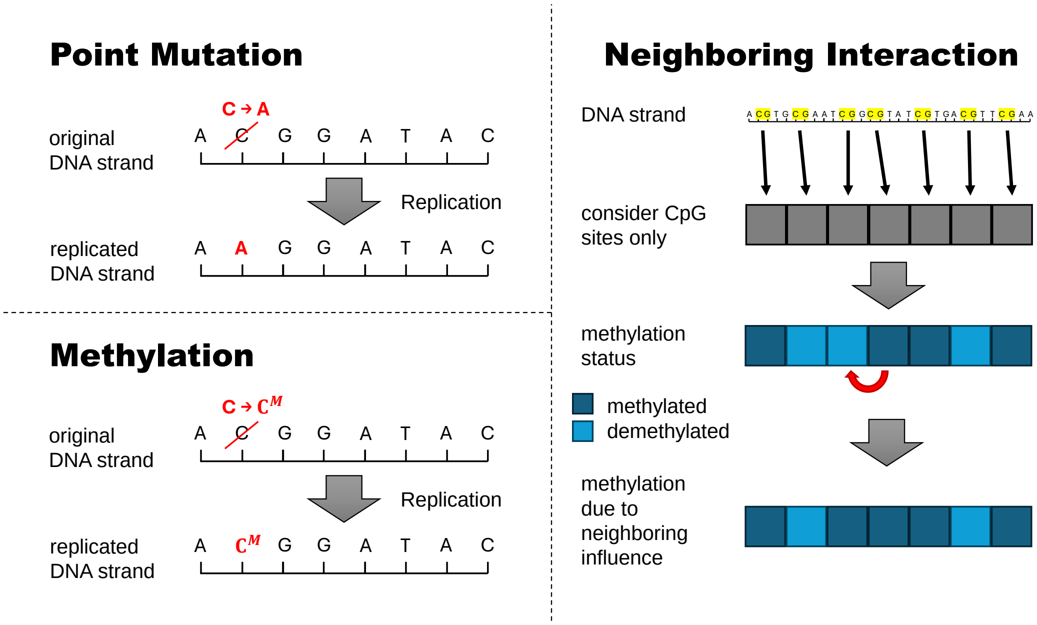

Figure 1: On the left, the difference between a genetic mutation (such as the point mutation) and the epigenetic mutation "methylation" that we consider here is shown. On the right, it is depicted how one obtains the methylation pattern from a DNA strand and what is meant by neighboring influence.

ation than exclusively spontaneous processes.

We approach this investigation by starting with the definition of a mathematical model for the neighboring interactions in Section 2, beginning with a single CpG site as the nearest neighbor and also giving an impression into what a model including the second nearest neighbor could look like. Furthermore, the scenario of inheritance is considered. To practically assess the validity of our hypotheses, a corresponding numerical simulation is introduced in Section 3. Lastly, in Section 4, a comparison between the theoretical results and the simulation is performed, along with a statistical analysis of the cluster size under various model conditions and the examination of the epigenetic variation.

2 Analytical model

In the following section, we consider a model approach on a single cell level to better understand the effect of neighboring interaction between CpG sites. For simplification, we assume a single DNA strand and that a CpG site is only influenced by the CpG sites to its right and left that are closest to it. We neglect the distance between two CpG sites and assume the nearest-neighbor interactions to be the same for all neighboring CpG sites. We now consider a sequence of L CpG sites which represents a segment of the DNA strand. It is denoted by a random vector $X = (X_1, ..., X_L)$ where each of the X_l, $l \in \{1, ..., L\}$, is a random variable assuming values in $\{0, 1\}$. We say that a site X_l is methylated if $X_l = 1$ and demethylated if $X_l = 0$. To incorporate the neighboring influence, we begin with a simple model which only accounts for the left nearest neighbor and later extend it to a model accounting for both left and right nearest neighbors. Note that without loss of generality the choice of accounting only for the *left* nearest neighbor is arbitrary.

Model formulation. Consider a single CpG site X_l, $l \in \{1, ..., L\}$. Spontaneous, non-collaborative changes of methylation state can occur naturally at any site X_l. However, X_l can additionally change its methylation state due to the influence of its left neighbor, denoted by X_{l-1}. We assume that the methylation state of any CpG site X_l is solely influenced by its left neighbor. Moreover, we assume these two processes to be independent. In addition, we introduce a periodic boundary condition such that the first and the last sites of the sequence can interact with one another, i.e. for X_1, its left neighbor is $X_{1-1} = X_L$. The reaction system looks as follows

$$\begin{aligned} 1 &\xrightarrow{k_1} 0, \\ 0 &\xrightarrow{k_2} 1, \\ 0,1 &\xrightarrow{k_3} 0,0, \\ 1,0 &\xrightarrow{k_4} 1,1, \end{aligned}$$

where the two processes at the top describe the spontaneous switches of single CpG sites, and the two bottom processes describe how a CpG site is influenced by its left neighbor and hence changes its status. The reaction rates are given by $k_1 := s$, $k_2 := sy$, $k_3 := as$, and $k_4 := asy$. Motivated by [11], that defined a similar, more complex DNA methylation model, $s > 0$ denotes the spontaneous, non-collaborative demethylation rate per site while $y > 0$ denotes the strength of methylation versus demethylation. Particularly, if $y < 1$, we assume that demethylation is the stronger reaction, and $y = 1$ means that both reactions are equally likely. Further, $a \geq 0$ measures the strength of neighboring influence. As we can see, if $a = 0$, the strength of neighboring influence would be zero leading to a model where only spontaneous changes of methylation are possible.

Master equation. One can derive a master equation for these reactions (for details consult Appendix A.1):

$$\begin{aligned} \frac{\mathrm{d}P(X_l(t) = 1)}{\mathrm{d}t} &= aysP(X_{l-1}(t) = 1, X_l(t) = 0) \\ &+ ys \\ &- asP(X_{l-1}(t) = 0, X_l(t) = 1) \\ &- (s + ys)P(X_l(t) = 1). \end{aligned} \tag{1}$$

This analytical result agrees with our understanding of how the probability measure $P(X_l(t) = 1)$ should change with time. We can illustrate this by considering the following two extreme cases:

1. Let us assume that at time $t = 0$ it is $X(t) = (0, ..., 0)$, i.e. $P(X_l(0) = 1) = 0 \; \forall \; l \in \{1, ..., L\}$. Then, as a result of the non-collaborative spontaneous changes, the CpG site X_l, as well as other sites in the sequence, can change their methylation status with a strictly positive probability, at least for a short period. Thus, $P(X_l(t) = 1)$ strictly increases for $t \in (0, \delta)$ for $\delta > 0$ small enough. This agrees with the master equation, which yields $\frac{\mathrm{d}}{\mathrm{d}t}P(X_l(0) = 1) = ys > 0$.

2. On the other hand, let us assume that at time $t = 0$ it is $X(t) = (1, ..., 1)$, i.e. $P(X_l(0) = 1) = 1$. Then as a result of the spontaneous demethylation process, the CpG site X_l, as well as other sites in the sequence, can change their methylation status with a strictly positive probability, but in this case from methylated to demethylated. Thus, $P(X_l(t) = 1)$ decreases for $t \in (0, \delta)$ for $\delta > 0$ small enough, agreeing with the master equation $\frac{\mathrm{d}}{\mathrm{d}t} P(X_l(0) = 1) = ys - (s + ys) = -s < 0$.

The joint probabilities which occur in (1) can be approximated assuming pairwise independence of X_{l-1} and X_l. Then, (1) simplifies to

$$\begin{aligned} \frac{\mathrm{d}P(X_l(t) = 1)}{\mathrm{d}t} \approx\, & as(1-y)P(X_l(t) = 1)^2 \\ & - \left(as(1-y) + s(1+y)\right) P(X_l(t) = 1) \\ & + ys. \end{aligned} \tag{2}$$

It should be noted that the independence-assumption only holds for very small values of a. The parameter a is a measure for the strength of the neighbor's influence, and true independence of X_{l-1} and X_l is only given if there is no such influence.

Stationary states and solution. The stationary states of (2) are given by

$$\begin{aligned} x_0 :=& \frac{1}{2} \text{ if } y = 1, \\ x_{1/2} :=& \frac{1}{2\alpha_1} \left((\alpha_1 + \alpha_2) \pm \sqrt{(\alpha_1 + \alpha_2)^2 - 4\alpha_1\alpha_3} \right) \text{ if } y \in (0, 1), \end{aligned}$$

where $x := P(X_l(t) = 1)$, $\alpha_1 := as(1 - y)$, $\alpha_2 := s(1 + y)$, and $\alpha_3 := ys$. It is sufficient to consider only $y \in (0, 1]$ as for $y > 1$, methylation would be the stronger process and we can consider $P(X_l(t) = 0)$ for a similar behavior due to $1 = P(X_l(t) = 0) + P(X_l(t) = 1)$. A stability analysis reveals that x_0 and x_2 are stable equilibria, while x_1 is unstable.

The master equation (2) can be solved analytically and yields for some initial value $P(X_l(0) = 1) = x_0$

$$P(X_l(t) = 1) = \frac{1}{2\alpha_1} \left(\sqrt{A} \tanh \left(-\frac{\sqrt{A}}{2} t + \operatorname{artanh} \left(\frac{2\alpha_1 x_0 - (\alpha_1 + \alpha_2)}{\sqrt{A}} \right) \right) + (\alpha_1 + \alpha_2) \right) \tag{3}$$

for $y \in (0, 1)$ with $A := (\alpha_1 + \alpha_2)^2 - 4\alpha_1\alpha_3$. This solution can be simplified by reducing the master equation (2) further, i.e. by neglecting the inhomogeneous term ys. The solution of the resulting equation is given by

$$P(X_l(t) = 1) \approx \frac{1}{\left(\frac{1}{x_0} - \frac{\alpha_1}{\alpha_1 + \alpha_2} \right) e^{(\alpha_1 + \alpha_2)t} + \frac{\alpha_1}{\alpha_1 + \alpha_2}}. \tag{4}$$

For the case $y = 1$, no simplification of (2) is necessary to obtain a relatively simple expression:

$$P(X_l(t) = 1) = \frac{1}{2} - \frac{1}{2} e^{-2st} + x_0 e^{-2st}. \tag{5}$$

See Appendix A.2 for more details on the solution.

Model extension to both nearest neighbors. Keeping the previously defined boundary condition, we once again consider a single CpG site X_l, $l \in \{1, ..., L\}$. The left neighbor of X_l is X_{l-1}, and we analogously define the nearest right neighbor as X_{l+1}. To extend our model to also account for the right neighbor, we make the following assumptions:

1. For methylation status $X_{l-1} \neq X_{l+1}$ or $X_{l-1} = X_l = X_{l+1}$, we assume that a change of the status at site X_l can only occur due to a spontaneous, non-collaborative reaction.

2. For $X_{l-1} = X_{l+1}$ and $X_{l-1} \neq X_l$, we assume that a change of the methylation status at site X_l can occur due to a spontaneous, non-collaborative reaction and additionally, due to the influence of the left and right neighbors. We assume these two processes to be independent.

The resulting reaction system is described by

$$\begin{aligned} 1 &\xrightarrow{k_1} 0 \\ 0 &\xrightarrow{k_2} 1 \\ 0,1,0 &\xrightarrow{\tilde{k}_3} 0,0,0 \\ 1,0,1 &\xrightarrow{\tilde{k}_4} 1,1,1, \end{aligned}$$

where the two bottom reactions describe the influence of left and right neighbor on the status of the CpG site in the middle. It is $k_1 = s$, $k_2 = sy$, $\tilde{k}_3 = as$, and $\tilde{k}_4 = asy$. The parameters $s > 0$ and $y > 0$ are defined the same way as in the previous section. Similar to the previous model, $a \geq 0$ measures the strength of neighboring influence. Note that the parameter a that is used in this model is not necessarily the same as the parameter a defined in the previous model because now, we need to consider the influence of two neighbors at the same time. Thus, $\tilde{k}_3$ and $\tilde{k}_4$ do not necessarily have the same quantitative influence as k_3 and k_4 from the previous model.

As earlier, we can derive a master equation for the system:

$$\begin{aligned} \frac{\mathrm{d}P\left(X_l(t)=1\right)}{\mathrm{d}t} = & \, aysP\left(X_{l-1}(t)=1, X_l(t)=0, X_{l+1}(t)=1\right) \\ & + ys \\ & - asP\left(X_{l-1}(t)=0, X_l(t)=1, X_{l+1}(t)=0\right) \\ & - (s+ys)P\left(X_l(t)=1\right), \end{aligned}$$

which can be simplified by using again the critical independence assumption of X_{l-1}, X_l, X_{l+1}:

$$\begin{aligned} \frac{\mathrm{d}P\left(X_l(t)=1\right)}{\mathrm{d}t} \approx & \, \left(-as-(s+ys)\right)P(X_l(t)=1) \\ & + \left(ays+2as\right)P(X_l(t)=1)^2 \\ & + \left(-ays+as\right)P(X_l(t)=1)^3 \\ & + ys. \end{aligned}$$

For details of the derivation, consult Appendix A.3.

Inheritance. In the following, we incorporate multiple generations into our models. For this, we use a discrete Wright-Fisher model [15], in which we consider a finite size population of N individuals (on a single cell level), where each of them carries a genome sequence of L CpG sites. In particular, we assume that all N individuals are born simultaneously at time $t = 0$ and that after one generation, denoted by t_{gen}, all N individuals die and are replaced by their offspring. Note that the number of individuals throughout generations remains constant. The methods by which methylations are passed on to the next generation are called between-sequence reactions. During each generation, each individual experiences changes in its methylome, e.g. as described by the models above. These reactions are called within-sequence reactions. In the following, the within-sequence reactions occur according to the model which only considers the left nearest neighbor. Let $X_{n,t}$ denote the sequence of individual n at time t and $X_{n;t;l}$ denote the methylation status of the l-th site of individual n at time t.

We begin with the simple case assuming no linkage between sites. Consider the end of the m-th generation and the beginning of the $m+1$-th generation. For a newborn n in the $m+1$-th generation, $n \in \{1, ..., N\}$, the probability for the CpG site l, $l \in \{1, ..., L\}$, to be methylated is

$$\frac{\sum_{n=1}^{N} X_{n;m\cdot t_{\text{gen}};l}}{N}.$$

After the replacement takes place, the offspring undergo the same within-sequence reactions as before until the end of the generation.

Now we continue with the case of linked inheritance. The main difference of this type of inheritance to unlinked type is that now the methylation status of a newborn's CpG site directly depends on the status of the corresponding sites of two specific individuals from the previous generation. This means that the methylation pattern of an individual n in generation $m+1$ is a mixture of the methylation patterns of the parents of individual n. We choose the parents i and j randomly from generation m (we assume there to be two parents necessary to generate a new individual; also, we do not differ between male or female) and we add break points in their sequences $X_{i,m\cdot t_{\text{gen}}}$, $X_{j,m\cdot t_{\text{gen}}}$ after equidistant numbers of CpG sites, say after each p-th site (p needs to be a divisor of L). This procedure results in $\frac{L}{p}$ sequence-"snippets" of length p for each parent. Now there are two possibilities for the sequence $X_{n,m\cdot t_{\text{gen}}}$ of their offspring: either $X_{n,m\cdot t_{\text{gen}}}$ is an alternating sequence of the parents' snippets starting with i:

$$X_{n,m\cdot t_{\text{gen}}} = (X_{i;m\cdot t_{\text{gen}};1}, ..., X_{i;m\cdot t_{\text{gen}};p}, X_{j;m\cdot t_{\text{gen}};p+1}, ..., X_{j;m\cdot t_{\text{gen}};2p}, X_{i;m\cdot t_{\text{gen}};2p+1}, ...),$$

or starting with j:

$$X_{n,m\cdot t_{\text{gen}}} = (X_{j;m\cdot t_{\text{gen}};1}, ..., X_{j;m\cdot t_{\text{gen}};p}, X_{i;m\cdot t_{\text{gen}};p+1}, ..., X_{i;m\cdot t_{\text{gen}};2p}, X_{j;m\cdot t_{\text{gen}};2p+1}, ...).$$

The whole population of generation m can be replaced with this procedure, and continue as generation $m+1$, undergoing the same within-sequence reactions as their parents.

3 Simulation

To better analyze our model, we have implemented Gillespie's algorithm [6] to conduct stochastic simulations. The numerical implementation of the model also bears the advantage that we do not need to rely on analytical results, which sometimes are only available after strong simplifications, as it was the case with the solution for the master equation.

We start with the algorithm for one individual within one generation, and a DNA strand with L CpG sites. In each step of Gillespie's algorithm, all potential reactions of each individual CpG site are identified. The time, after which a reaction occurs, is exponentially distributed with expected value corresponding to the inverse of the sum of all possible transition rates which are identified previously. After choosing a time point, a reaction is then randomly selected. The corresponding CpG site changes its methylation status while all other CpG sites preserve theirs. This procedure is repeated until a predefined stopping time has been reached. A detailed implementation using Python [23] of both models described in Section 2 can be found in our GitHub repository[1].

Now consider a cohort of N individuals, each with a sequence of L CpG sites. We distinguish between within- and between-sequence reactions. Until the end of one generation, the N sequences undergo independently of each other within-sequence reactions corresponding to the process described in the previous paragraph. Further, let $i_{m,n}$, $m \in \{1, ... \lfloor t_{\text{end}}/t_{\text{gen}} \rfloor\}$ and $n \in \{1, ..., N\}$, denote the last time point at which a reaction takes place in sequence n, before the m-th generation ends. In our simulation, we define a Boolean variable, which keeps track of whether the time point at which a reaction should happen lies still inside the time of one generation. The first time a time point exceeds the current generation, $i_{m,n}$ will be set to the last time point, at which a reaction occurred.

At the end of generation m, the sequences undergo one between-sequence reaction, either governed by linked or unlinked inheritance.

4 Results from the simulation

Comparison with the stationary analytical solution. In the following, we consider the model for only one generation (and one influencing neighbor). We want to compare the long-term behavior of the simulation to the stable stationary state x_2 from the analytical model. This will provide some information regarding the parameter ranges for which the simplified master equation is adequate, and if we use small a-values, it should also allow to check whether the simulation behaves as expected. Hence, this comparison should justify the further use of the numerical model for the investigation of our hypotheses.

The expected value of $P(X_l(t) = 1)$ from the analytical model – and therefore the value of x_2 – can be approximated using the mean methylation level. It is equal to the average number of methylated CpG sites per sequence at a certain point in time. From data observations, the mean methylation level averaged over multiple sequences seems to stabilize after 2000 time units, where the initial sequences are randomly generated with equal probability for 0 and 1. Furthermore, we fix the sequence length to $L = 200$ CpG sites and the spontaneous demethylation rate $s = 1.47 \cdot 10^{-3}$ (taken from [22] as an example for methylation in plants). For $y \in \{0.1, 0.3, 0.6, 0.9\}$, and $a \in \{0.1, 0.5, 1, 2, 10\}$, Fig. 2 shows the mean methylation level over 30 runs and the corresponding analytically approximated stable stationary state x_2. We conclude that the approximated stable stationary state is more or less able to describe our expectation even for $a = 1$. In particular, we observe that x_2 and the simulated long-term behavior coincide well for small a, which can be interpreted as a sign of the adequateness of the simulation.

As expected, for $y = 0.6$ and $y = 0.9$, x_2 seems to deviate significantly from the observed value of $P(X_l(t) = 1)$ in the case of $a = 10$. However, we have to keep in mind that x_2 only describes the case $a \approx 0$.

Next, for the same fixed parameter values and the same ranges for y and a, Fig. 3 shows the mean methylation level over 30 runs and the corresponding solution (3) of the approximated

[1] https://github.com/yifanchn/methylation_pattern_modelling_pub.git

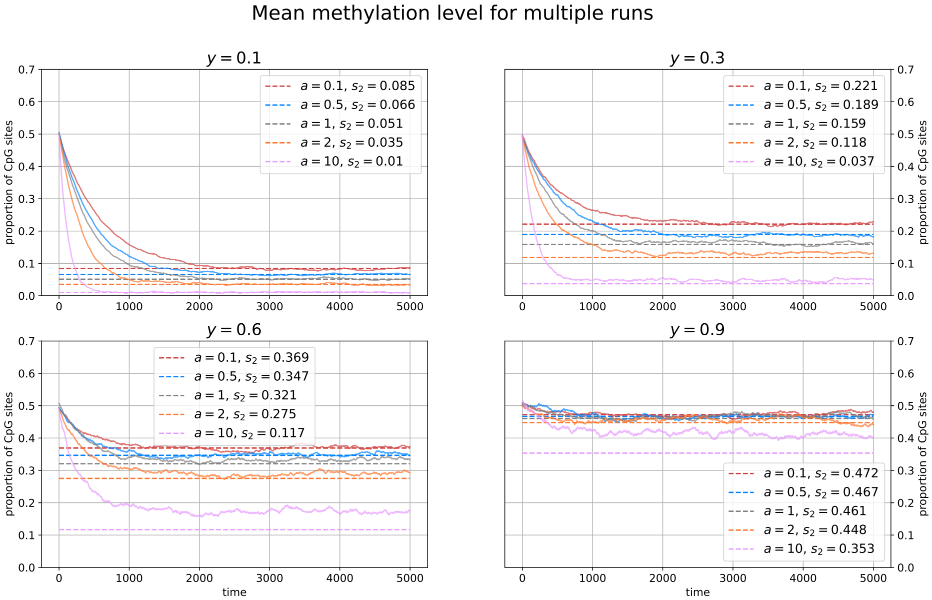

Figure 2: Mean methylation level over 30 runs (solid lines) including the approximated analytical stable stationary state x_2 (dashed lines). Each panel shows the result for a different y-value, $y < 1$. Same color means same a-value.

master equation (2). We observe that similarly to what we have seen with the stationary point, our analytical estimation of $P(X_l(t) = 1)$ is able to capture the observed behavior of the mean methylation level also for bigger values of a.

Comparisons between the simulated mean methylation level and Equations (4) and (5) were performed as well and can be found in Appendix A.4, together with a proposal for a general fitting function. In conclusion, the simulation and the analytical results agree as expected, allowing us to continue our investigations with the simulation.

Cluster Size Evolution. As mentioned before, clusters are agglomerations of CpG sites with the same methylation status that are not interrupted by sites of a different status. By our definition of the term, even single CpG sites can be called clusters: they form clusters of length 1. For instance, consider the sequence of $L = 5$ CpG sites given by $(0, 1, 1, 1, 0)$. This sequence features two clusters: one demethylated cluster of length 2 (coming from the periodic boundary condition that links the last site of the sequence to the first) and one methylated cluster of length 3. We refer to the length of a cluster also as “cluster size”, to avoid confusion with the length of the sequence L.

The stochastic nature of Gillespie’s algorithm leads to considerable fluctuations not only in the number of (de-)methylated CpG sites from time step to time step, but obviously also to oscillations in average cluster size. Therefore, we will observe the mean evolution of several sequences’ clusters, similar as we handled the comparisons above.

In the following, we investigate how the different models (only spontaneous reactions, one neighbor, two neighbors) influence the mean cluster size and especially whether larger clusters

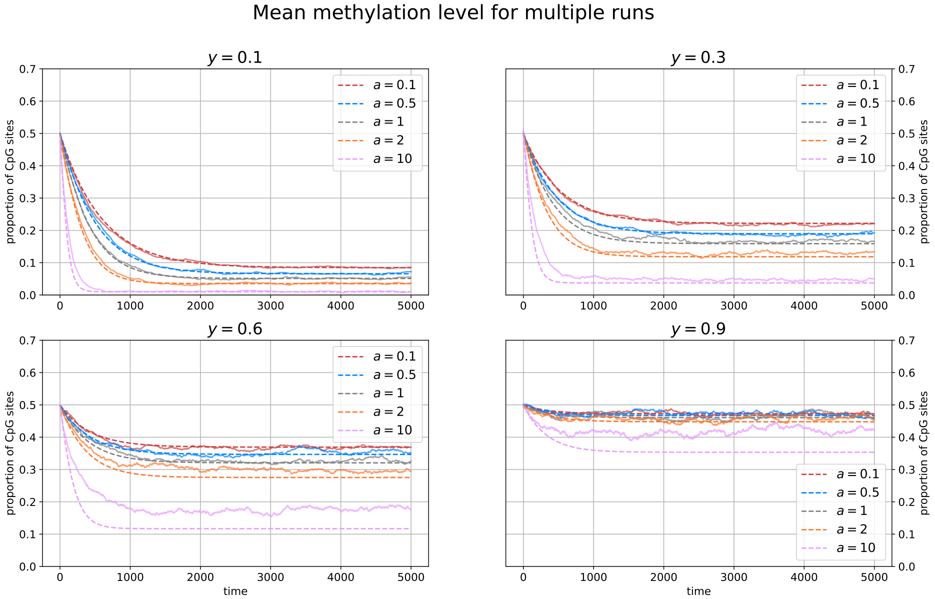

Figure 3: Mean methylation level over 30 runs (solid lines) including the exact solution (dashed lines) for the approximation of the methylation level obtained in Equation (3) for different $y < 1$.

appear in case of the models that include nearest-neighbor interactions. A model that only features spontaneous changes in methylation status is obtained by setting $a = 0$ in the one-neighbor model.

To enhance statistical significance in the later tests, we consider more and longer simulations than earlier. In particular: $N = 100$ sequences of length $L = 200$, each over a running time until $t_{\text{end}} = 5000$. The spontaneous demethylation rate is again chosen as $s = 1.47 \cdot 10^{-3}$, and we investigate three different pairings of parameters (note that the values for a are only valid for the two nearest-neighbor models since $a = 0$ in case of the purely spontaneous model):

- $a = 1$, $y = 1$: the neighboring interactions are relatively low and no status is preferred
- $a = 1$, $y = 0.5$: the neighboring interactions are relatively low and demethylation is preferred
- $a = 1$, $y = 2$: the neighboring interactions are relatively low and methylation is preferred

$a = 1$ is chosen for the two models that include nearest neighbor interactions since it is the lowest possible value of a where the influence of neighboring interactions is noticeable in the sense that we understand them (i.e. the effect of the neighbors' influence is at least as large as the effect of spontaneous reactions). We assume that if we find significant differences in mean cluster size for a value of a as small as 1, this implies significant differences for $a > 1$. For comparability, we start every run of the simulation with the same initial sequence. This initial sequence is a random combination of zeros and ones drawn from a uniform distribution.

The Figures 4, 5 and 6 show how over time the mean cluster sizes asymptotically approach equilibrium values that differ for each model and parameter combination (please note that the

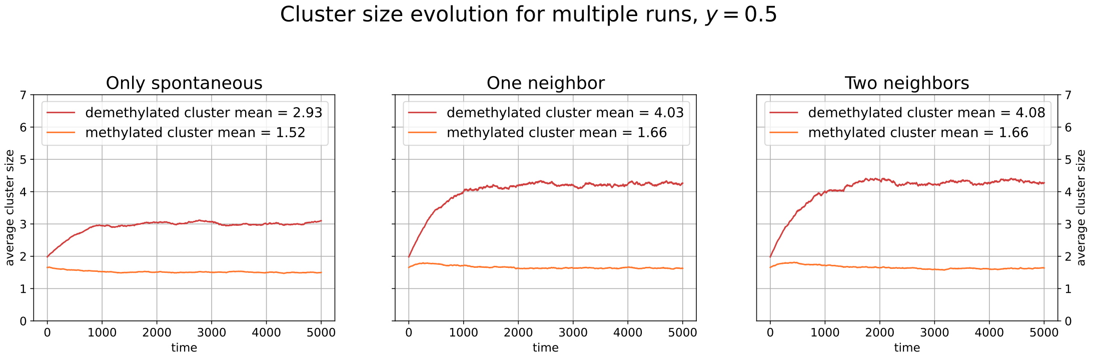

Figure 4: Cluster size evolution for $y = 0.5$ over 100 runs. From left to right: only spontaneous reactions, model including the influence of one neighbor, model including the influence of both neighbors. (red: methylated clusters, orange: demethylated clusters)

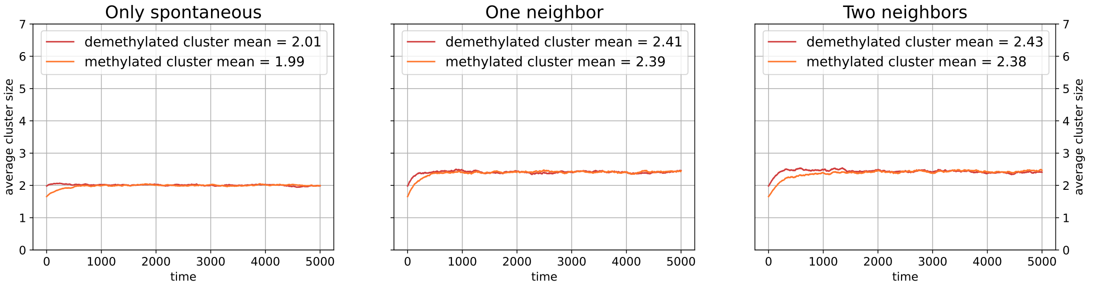

Figure 5: Cluster size evolution for $y = 1$.

vertical axes have different scaling; the mean values are indicated in the legends for orientation). To validate these differences statistically, we test the final distributions of the cluster sizes first for normality, and then conduct ANOVAs and two-sample t-tests or – if the lack of normal distribution requires it – Wilcoxon rank-sum tests. For details on the statistical tests consult Appendix A.5.

The tests lead to two results: firstly, the differences in mean cluster size between the purely

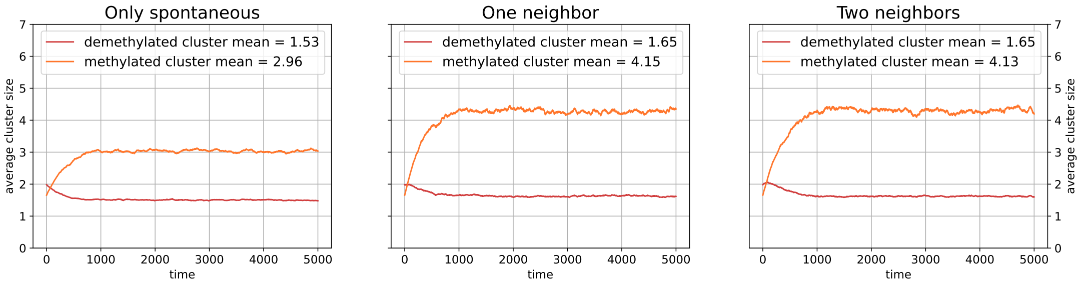

Figure 6: Cluster size evolution for $y = 2$.

spontaneous model and each of the nearest-neighbor models are significant for all parameter values. Secondly, the differences in mean cluster size between the model with one neighbor as influence and the one with both neighbors are generally not significant, with one exception for the the mean demethylated cluster size in the case of $y = 1$: here, the p-value 0.0441 lies slightly below the chosen significance level of $\alpha = 0.05$.

Although the differences between the two models that include neighboring interactions overall cannot be considered significant, we were able to show that both, the model considering only the left nearest neighbor and the model considering both nearest neighbors, have a statistical impact on the cluster sizes. A glance at Fig. 7 – which shows the mean cluster sizes after 100 runs per parameter combination and is a summary/extension of the Figs. 4, 5 and 6 – also tells us that this impact leads in mean to larger clusters for the models that consider the influence of one or two neighbors, regardless of the choice of y. A linear regression analysis reveals that mean demethylated cluster size decays with y following the power law, while the mean methylated cluster size grows exponentially with y.

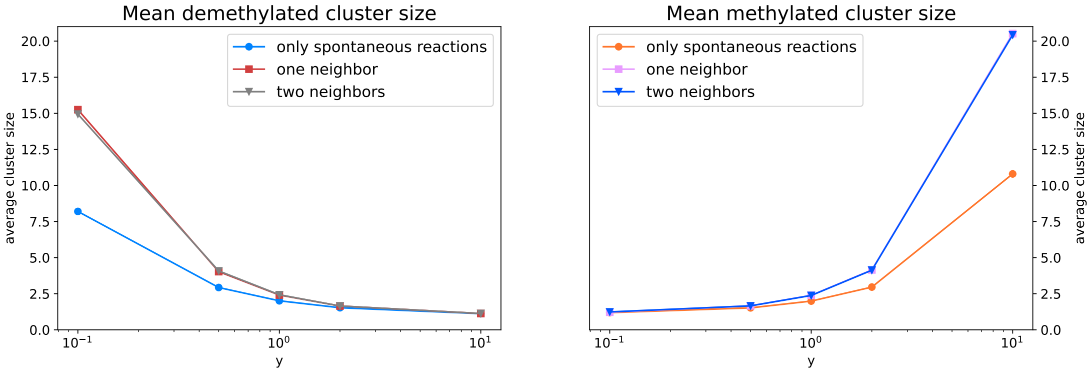

Figure 7: Mean cluster sizes for various values of the parameter y, namely $y \in \{0.1, 0.5, 1, 2, 10\}$. Note that the horizontal axis is log-scaled and the clusters in case of the neighborhood-influenced models have such similar sizes that their curves overlap.

Methylation Site Frequency Spectrum. Similar in calculation to the site frequency spectrum (or allele frequency spectrum), the methylation site frequency spectrum (mSFS) appears to be a relatively popular tool to see how common (de-)methylations are on certain CpG sites and how they are distributed on the genome when one has a specific sequence from each subject [24, 26, 28]. Compared to the cluster size, it is a rather generation-wide summary statistic and does not provide information about the clusters, but about phenomena such as epigenetic variation and drift among a population. The mSFS answers the question: how many CpG sites at the same site X_l are demethylated in exactly 1, exactly 2, exactly 80 of the subjects? On the horizontal axis of the spectrum, one can find the number of demethylated sites, and the vertical axis denotes the number of CpG sites at which this number of demethylated sites is observed among the population.

What we are interested in concerning the mSFS is whether the neighboring interactions have an influence on its shape. We hypothesize that a model with purely spontaneous processes ($a = 0$) leads to a more U-shaped mSFS than the models that consider nearest-neighbor interactions. The U-shape means that in most individuals, the same CpG sites are (de-)methylated, which is a sign for low epigenetic variation. In contrast, an A-shaped mSFS is the result of an intermediate level of methylation in most sites (meaning that for most CpG sites X_l, it is $X_l = 0$ in about 50% of individuals, and $X_l = 1$ in the other 50%), and signifies high epigenetic variation.

To see the effects of our choice of parameters in the mSFS within a reasonable number of generations, the parameters need to be chosen more extremely than in the previous section (e.g. s very small, a very large to witness nearest-neighbor interactions). The inheritance of methylation happens without linkage. We consider the following choice of parameters (they were determined experimentally): each generation comprises $N = 100$ individuals, the observed sequences have length $L = 2000$, running time per individual ends at $t_{\text{end}} = 100$, the spontaneous demethylation rate is $s = 10^{-8}$ and a is set to 0 for the model that considers only spontaneous reactions, whereas we choose $a = 10^3$ for the two models that consider nearest-neighbor influences. As we want to study the effect of different models on a selection-free methylome evolution, we choose $y = 1$. We start from an initial population with randomly chosen sequences (zeros and ones drawn from a uniform distribution), so that the mSFS in generation 0 is A-shaped.

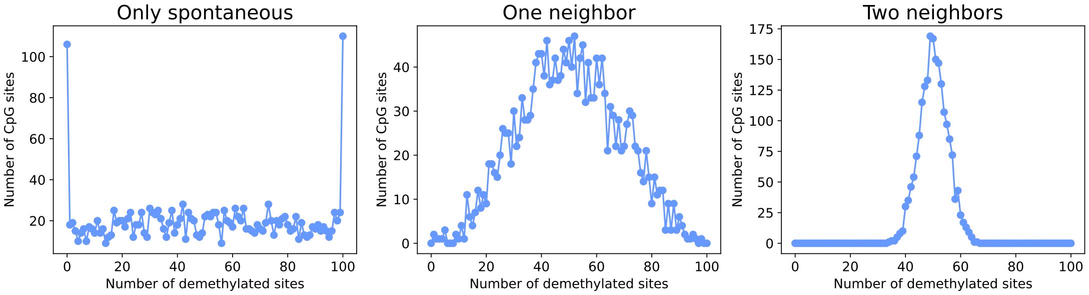

Figure 8: mSFS for various models after 50 generations. From left to right: only spontaneous reactions, model including the influence of one neighbor, model including the influence of both neighbors. Note that the vertical axis ranges differ, as this is a qualitative comparison.

After 50 generations, we obtain the spectra that are shown in Fig. 8. All observed mSFS appear more or less symmetrical, which is a consequence of $y = 1$. Choosing y in a way that makes demethylation or methylation the more favoured state leads to an mSFS that is skewed to the right or left, respectively, which has been observed for all models.

While the mSFS for the model that considers purely spontaneous reactions exposes a clear U-shape, the situation looks completely different for the two models featuring neighborhood interactions: the one-neighbor model leads to a wide A-shaped model, the one considering two neighbors yields a narrower A-shape. The latter is an example for extreme epigenetic variation, but also the former is a sign for eclectic methylation patterns among the population. This confirms our hypothesis, or at least shows that with the same rate for spontaneous demethylations we can provoke large differences in the shape of the mSFS depending on whether none, one, or two neighboring CpG sites have an influence on the methylome evolution.

5 Conclusion

It is well-known that mathematical models can help to describe, quantify and even predict reality. Our current study is a simplified model which is sufficient to capture essential mechanisms of DNA methylation patterns in reality. However, there is room for improvements, which of course requires a lot more attention to the details which we discuss below:

- The methylation pattern constructed in our model represents only one strand of DNA, neglecting cases where at a specific site one strand is methylated and the other one not. Considering such cases could result in a difference in the impact of neighboring interactions. On this note, we also ignored that the DNA is not just a double-stranded line, but due to its double-helix structure it is also curved and twisted. This case also suggests that even quantifying the number of influencing neighbors is a challenge.
- Assuming that there are only nearest-neighbor interactions excludes the possibility of interaction with other sites in the same region or in other domains of the DNA. Second-, third-, or n-closest CpG sites could also have an influence [14].
- We do not account for the distance between CpG sites and assume that the nearest-neighbor interactions are the same for all neighboring sites. Consequently, the change rates employed in our model represent a very simplified version of the changes that occur in nature which have been observed to be distance-dependent [1, 14].
- In our model extension to two neighbors, we assume that neighboring interactions only occur when both nearest-neighbors have the same methylation status. Hence, we ignore possible, potentially weaker, one-sided neighboring interactions when $X_{l-1} \neq X_{l+1}$.
- In the derivation of the master equations, we made the assumption that the CpG sites change their methylation status independently of each other, which of course contradicts the very basic assumptions of this work. For further work with the analytical formulation of the model, a simplification without this requirement should be used. However, we were still able to make statements about our hypotheses without relying too much on the analytical results as we used the stochastic simulations for these investigations.
- Regarding the analysis of the stochastic model, the parameters we used are mainly selected as results of experiments with the simulations, with the exception of the demethylation rate which was taken from a publication on methylation in thale cress. We speculate there is a potentially less biased way to determine them.
- In our methylation site frequency spectrum analysis, we used inheritance without linkage to model the inheritance process of methylation. Inheritance with at least a certain amount of linkage would be a more realistic approach.

The simplifications made during the analytical and numerical estimations were necessary at first to understand our model and its challenges. Our investigations serve as proof of principle, and the application of our results to specific plants or vertebrates lies beyond the scope of this work. Possible further steps should focus on refining the analytical model as well as the simulation, and/or on validating the current model and possible alternatives with real-world data to find out which simplifications affect the models' performance the most. Comparisons like this require single-cell data sets that contain DNA methylome information at different time points, in the best case over the cell's entire lifetime.

Furthermore, we aim to conduct an analytical study concerning the birth-death process in the future and compute a master equation that describes birth-death processes with linkage. Then, we require an analysis of cluster size evolution depending on the kind of inheritance employed. In Appendix A.6, we discuss whether inheritance with linkage is a more appropriate way to model inheritance of epigenetic information, particularly given the importance of cluster size in our research. We propose to study this question further. Based on the findings concerning this topic (see appendix), the hypothesis: "There is a difference in mean cluster size evolution between inheritance with and without linkage" can be rejected, but further statistical tests would be necessary to ensure this.

As described in Section 4 the differences in cluster size between our two proposed models cannot be considered significant for the values chosen for the parameter a (the measure for the strength of the neighboring interactions). On the one hand, this could indicate that the chosen a is not sufficiently large and on the other hand, it could be necessary to further extend this model to include the cases where $X_{l-1}(t) \neq X_{l+1}(t)$, but a single site, for example X_{l-1}, still influences the methylation status of its right neighbor. Under these assumptions, we would additionally introduce rates that account for these changes since these interactions would not be as strong as in the case where $X_{l-1}(t) = X_{l+1}(t)$. Lastly, further analyses are needed in order to determine whether this new model could be significantly different from the others. For a formal model description of this alternative we refer to Appendix A.7.

This work is an example of how scientific computing in the form of stochastic simulations and statistics can support model analysis in cases where analytical models cannot be handled explicitly. The comparison of the long-term behavior and the curve fitting ensured that the simulation and the analytical model agree for the parameter regime on which an approximate solution of the analytical model could be obtained (i.e. $a \ll 1$), which served as a basis for the assumption that the simulation would be adequate for other parameter values as well. The subsequent statistical analysis provided reliability to the simulation results.

With the techniques presented in this work, we obtained results in favor of our initial hypotheses. The findings regarding mean cluster size evolution illustrate a significant difference between a model that accounts for spontaneous changes and neighboring interactions versus one that exclusively considers spontaneous, non-collaborative modifications. Furthermore, we found solid hints supporting that neighboring interaction combined with unlinked inheritance leads to more epigenetic variation than exclusively spontaneous processes.

Data availability

To allow further research on this topic using the code that was created during this project, it is made available in a public GitHub repository: https://github.com/yifanchn/methylation_pattern_modelling_pub.git.

Author contributions

YC: Model formulation, analysis and solution of the master equations, numerical simulations. Writing: original draft, review & editing. VH: Numerical simulations, statistical analysis. Writing: original and final draft, review & editing. AR: Model formulation, derivation and analysis of the master equations. Writing: original draft, review & editing. TS: Model formulation, derivation of the master equations. Writing: original draft, review & editing. SM: Assistant project supervision. Writing: review & editing. SJ: Project supervision.

References

[1] O. Affinito, D. Palumbo, A. Fierro, M. Cuomo, G. de Riso, A. Monticelli, G. Miele, L. Chiariotti, and S. Cocozza. Nucleotide distance influences co-methylation between nearby CpG sites. *Genomics*, 112(1):144–150, 2020.

[2] I. N. Bronstein, K. A. Semendjajew, G. Musiol, and H. Mühlig. *Taschenbuch der Mathematik. With CD-ROM. 10th revised edition.* Europa-Lehrmittel, Haan-Gruiten, 2016.

[3] K. R. Das and A H. M. Rahmatullah Imon. A brief review of tests for normality. *American Journal of Theoretical and Applied Statistics*, 5(1):5–12, 2016.

[4] E. J. Finnegan, W. J. Peacock, and E. S. Dennis. DNA methylation, a key regulator of plant development and other processes. *Current Opinion in Genetics & Development*, 10(2):217–223, 2000.

[5] C. Ford. The wilcoxon rank sum test. https://data.library.virginia.edu/the-wilcoxon-rank-sum-test/, 2017. Accessed: 2024-08-05.

[6] D. T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22(4):403–434, 1976.

[7] M. R. Harwell, E. N. Rubinstein, W. S. Hayes, and C. C. Olds. Summarizing monte carlo results in methodological research: The one- and two-factor fixed effects ANOVA cases. *Journal of Educational Statistics*, 17(4):315–339, 1992.

[8] B. T. Hofmeister, K. Lee, N. A. Rohr, D. W. Hall, and R. J. Schmitz. Stable inheritance of DNA methylation allows creation of epigenotype maps and the study of epiallele inheritance patterns in the absence of genetic variation. *Genome biology*, 18(1):155, 2017.

[9] F. Johannes, E. Porcher, F. K. Teixeira, V. Saliba-Colombani, M. Simon, N. Agier, A. Bulski, J. Albuisson, F. Heredia, P. Audigier, D. Bouchez, C. Dillmann, P. Guerche, F. Hospital, and V. Colot. Assessing the impact of transgenerational epigenetic variation on complex traits. *PLOS Genetics*, 5(6):e1000530, 2009.

[10] S. John and W. Stephan. Important role of genetic drift in rapid polygenic adaptation. *Ecology and Evolution*, 10(3), 2020.

[11] L. Kerr, D. Sproul, and R. Grima. Cluster mean-field theory accurately predicts statistical properties of large-scale DNA methylation patterns. *Journal of the Royal Society Interface*, 19(186), 2022.

[12] M. Kim and J. Costello. DNA methylation: an epigenetic mark of cellular memory. *Experimental & Molecular Medicine*, 49(4):e322, 2017.

[13] J. A. Law and S. E. Jacobsen. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nature Reviews Genetics*, 11(3):204–220, 2010.

[14] C. Lövkvist, I. B. Dodd, K. Sneppen, and J. O. Haerter. DNA methylation in human epigenomes depends on local topology of CpG sites. *Nucleic acids research*, 44(11):5123–5132, 2016.

[15] P. W. Messer. Neutral models of genetic drift and mutation. In Richard M. Kliman, editor, *Encyclopedia of Evolutionary Biology*, pages 119–123. Academic Press, Oxford, 2016.

[16] L. D. Moore, T. Le, and G. Fan. DNA methylation and its basic function. *Neuropsychopharmacology*, 38:23–38, 2013.

[17] B. C. Richardson. Role of DNA methylation in the regulation of cell function: autoimmunity, aging and cancer. *The Journal of Nutrition*, 132(8 Suppl):2401S–2405S, 2002.

[18] D. Schübeler. Function and information content of DNA methylation. *Nature*, 517(7534):321–326, 2015.

[19] S. Tirnaz and J. Batley. DNA methylation: Toward crop disease resistance improvement. *Trends in Plant Science*, 24(12):1137–1150, 2019.

[20] R. K. Tran, J. G. Henikoff, D. Zilberman, R. F. Ditt, S. E. Jacobsen, and S. Henikoff. DNA methylation profiling identifies CG methylation clusters in arabidopsis genes. *Current Biology*,

15(2):154–159, 2005.

[21] University of Utah. Learn. Genetics. Genetic Linkage. [online]. https://learn.genetics.utah.edu/content/pigeons/geneticlinkage/. Accessed on August 08, 2024.

[22] A. van der Graaf, R. Wardenaar, D. A. Neumann, A. Taudt, R. G. Shaw, R. C. Jansen, R. J. Schmitz, M. Colomé-Tatché, and F. Johannes. Rate, spectrum, and evolutionary dynamics of spontaneous epimutations. *Proceedings of the National Academy of Sciences of the United States of America*, 112(21):6676–6681, 2015.

[23] G. Van Rossum and F. L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009.

[24] A. Vidalis, D. Živković, R. Wardenaar, D. Roquis, A. Tellier, and F. Johannes. Methylome evolution in plants. *Genome Biology*, 17(1):264–278, 2016.

[25] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, and P. van Mulbregt. SciPy 1.0: Fundamental algorithms for scientific computing in python. *Nature Methods*, 17(3):261–272, 2020.

[26] J. Wang and C. Fan. A neutrality test for detecting selection on DNA methylation using single methylation polymorphism frequency spectrum. *Genome Biology and Evolution*, 7(1):154–171, 2014.

[27] G. Xu, J. Lyu, Q. Li, H. Liu, D. Wang, M. Zhang, N. M. Springer, J. Ross-Ibarra, and J. Yang. Unique DNA methylome profiles in CpG island methylator phenotype colon cancers. *Genome Research*, 22(2):283–291, 2012.

[28] G. Xu, J. Lyu, Q. Li, H. Liu, D. Wang, M. Zhang, N. M. Springer, J. Ross-Ibarra, and J. Yang. Evolutionary and functional genomics of DNA methylation in maize domestication and improvement. *Nature Communications*, 11(1):5539–5550, 2020.

A Appendix

A.1 Derivation of the master equations (single-neighbor model)

For $l \in \{1, ..., L\}$, let $X_l(t)$, $t \geq 0$, denote the state of the l-th site at time t. The main focus of this section is to approximate $P\left(X_l(t) = 1\right)$ by observing the dynamics of the process at X_l. Since this behavior is identical for all X_l, and we assume a periodic boundary condition, it follows that $P\left(X_l(t) = 1\right)$ is equal for all $l \in \{1, ..., L\}$.

For $\Delta > 0$ but small, let us recall the law of total probability

$$P(X_l(t) = i) = \sum_{j=0}^{1} P(X_l(t) = i | X_l(t - \Delta) = j) P(X_l(t - \Delta) = j). \tag{6}$$

Let A be the event where X_l changes its state due to a neighboring interaction, and let B denote the event where a spontaneous change occurs and thus, X_l changes its state. Keeping in mind that the events A and B are independent of each other, it follows that

$$\begin{aligned}
&P(X_l(t) = 1|X_l(t-\Delta) = 0)P(X_l(t-\Delta) = 0)\\
=&P(\{X_l(t) = 1\} \cap \{X_l(t-\Delta) = 0\} \cap A)\\
&+ P(\{X_l(t) = 1\} \cap \{X_l(t-\Delta) = 0\} \cap B)\\
=&P(\{X_l(t) = 1\} \cap A|X_{l-1}(t-\Delta) = 1, X_l(t-\Delta) = 0)P(X_{l-1}(t-\Delta) = 1, X_l(t-\Delta) = 0)\\
&+ P(\{X_l(t) = 1\} \cap B|X_l(t-\Delta) = 0)P(X_l(t-\Delta) = 0)\\
=&[ays\Delta + O(\Delta^2)]P(X_{l-1}(t-\Delta) = 1, X_l(t-\Delta) = 0)\\
&+ [ys\Delta + O(\Delta^2)]P(X_l(t-\Delta) = 0),
\end{aligned}$$

as well as

$$\begin{aligned}
&P(X_l(t) = 0|X_l(t-\Delta) = 1)P(X_l(t-\Delta) = 1)\\
=&P(\{X_l(t) = 0\} \cap \{X_l(t-\Delta) = 1\} \cap A)\\
&+ P(\{X_l(t) = 0\} \cap \{X_l(t-\Delta) = 1\} \cap B)\\
=&P(\{X_l(t) = 0\} \cap A|X_{l-1}(t-\Delta) = 0, X_l(t-\Delta) = 1)P(X_{l-1}(t-\Delta) = 0, X_l(t-\Delta) = 1)\\
&+ P(\{X_l(t) = 0\} \cap B|X_l(t-\Delta) = 1)P(X_l(t-\Delta) = 1)\\
=&[as\Delta + O(\Delta^2)]P(X_{l-1}(t-\Delta) = 0, X_l(t-\Delta) = 1)\\
&+ [s\Delta + O(\Delta^2)]P(X_l(t-\Delta) = 1).
\end{aligned}$$

Further, from

$$P\left(X_l(t) = 1|X_l(t-\Delta) = 1\right) + P\left(X_l(t) = 0|X_l(t-\Delta) = 1\right) = 1$$

we obtain

$$\begin{aligned}
&P\left(X_l(t) = 1|X_l(t-\Delta) = 1\right)P\left(X_l(t-\Delta) = 1\right)\\
=&(1 - P\left(X_l(t) = 0|X_l(t-\Delta) = 1\right))P\left(X_l(t-\Delta) = 1\right)\\
=&P\left(X_l(t-\Delta) = 1\right) - P\left(X_l(t) = 0|X_l(t-\Delta) = 1\right)P\left(X_l(t-\Delta) = 1\right).
\end{aligned} \tag{7}$$

Therefore, using (6), it follows that

$$\begin{aligned}
P(X_l(t) = 1) =&[ays\Delta + O(\Delta^2)]P(X_{l-1}(t-\Delta) = 1, X_l(t-\Delta) = 0)\\
&+ [ys\Delta + O(\Delta^2)]P(X_l(t-\Delta) = 0)\\
&+ P(X_l(t-\Delta) = 1)\\
&- [as\Delta + O(\Delta^2)]P(X_{l-1}(t-\Delta) = 0, X_l(t-\Delta) = 1)\\
&- [s\Delta + O(\Delta^2)]P(X_l(t-\Delta) = 1).
\end{aligned} \tag{8}$$

Since $P(X_l(t) = 0) + P(X_l(t) = 1) = 1$, by subtracting $P(X_l(t-\Delta) = 1)$ from both sides of Equation (8), dividing both sides of (8) by Δ, and lastly letting $\Delta \to 0$, it follows

$$\begin{aligned}
\frac{dP(X_l(t) = 1)}{dt} =&aysP(X_{l-1}(t) = 1, X_l(t) = 0)\\
&+ ys\\
&- asP(X_{l-1}(t) = 0, X_l(t) = 1)\\
&- (s + ys)P(X_l(t) = 1).
\end{aligned} \tag{9}$$

Let us further analyze (9) and we begin by finding an approximation for the probabilities $P(X_{l-1}(t) = 1, X_l(t) = 0)$ and $P(X_{l-1}(t) = 0, X_l(t) = 1)$. We set

$$\begin{aligned} P\left(X_{l-1}(t) = 1, X_l(t) = 0\right) &\approx P(X_{l-1}(t) = 1)P(X_l(t) = 0) \\ &= P(X_l(t) = 1) - P(X_l(t) = 1)^2, \end{aligned}$$

analogously

$$\begin{aligned} P\left(X_{l-1}(t) = 0, X_l(t) = 1\right) &\approx P(X_{l-1}(t) = 0)P(X_l(t) = 1) \\ &= P(X_l(t) = 1) - P(X_l(t) = 1)^2, \end{aligned}$$

since $P(X_l(t) = 1) + P(X_l(t) = 0) = 1$, and $P(X_{l-1}(t) = 1) = P(X_l(t) = 1)$ by the construction of our model. Note that these simplifications are better suited for scenarios where no neighboring interactions occur or are very weak, i.e. $a = 0$ or $a \approx 0$, since we are assuming independence of X_{l-1} and X_l. Inserting these approximations in Equation (9), we obtain

$$\begin{aligned} \frac{dP(X_l(t) = 1)}{dt} \approx &+ as(1-y)P(X_l(t) = 1)^2 \\ &- \left(as(1-y) + s(1+y)\right)P(X_l(t) = 1) \\ &+ ys =: f(x), \end{aligned} \tag{2}$$

where $x := P(X_l(t) = 1)$.

To compute the stationary states of (2), we first note that it suffices to consider only $y \in (0, 1]$ as for $y > 1$, methylation would be the stronger process and we can simply consider $P(X_l(t) = 0)$ for a similar behavior due to $1 = P(X_l(t) = 0) + P(X_l(t) = 1)$.
For $y = 1$,

$$\begin{aligned} \frac{dP(X_l(t) = 1)}{dt} \approx &+ as(1-y)P(X_l(t) = 1)^2 \\ &- \left(as(1-y) + s(1+y)\right)P(X_l(t) = 1) \\ &+ ys \\ = &- 2sP(X_l(t) = 1) + s =: g(x). \end{aligned}$$

Setting that last result to zero and solving for $P(X_l(t) = 1)$ yields $P(X_l(t) = 1) = \frac{1}{2}$. Further,

$$\frac{dg}{dx} = -2s < 0$$

since $s > 0$. Thus, we conclude that $P(X_l(t) = 1) = \frac{1}{2}$ is a stable stationary state. This result reinforces how we expect the probability measure $P(X_l(t) = 1)$ to behave for $t \to \infty$ for the case where no selection is present in the model, i.e. $y = 1$.
Now, let $y < 1$. Setting (2) to zero and solving it for x yields

$$x_{1/2} = \frac{1}{2\alpha_1}\left((\alpha_1 + \alpha_2) \pm \sqrt{(\alpha_1 + \alpha_2)^2 - 4\alpha_1\alpha_3}\right),$$

where $\alpha_1 = as(1-y)$, $\alpha_2 = s(1+y)$, and $\alpha_3 = ys$. Further,

$$\begin{aligned} (\alpha_1 + \alpha_2)^2 - 4\alpha_1\alpha_3 =& \alpha_1^2 + +2\alpha_1\alpha_2 + \alpha_2^2 - 4\alpha_1\alpha_3 \\ =& \alpha_1(\alpha_1 + 2\alpha_2 - 4\alpha_3) + \alpha_2^2 \\ >& 0, \end{aligned}$$

since

$$\begin{aligned}\alpha_1 + 2\alpha_2 - 4\alpha_3 =& 2s(1+y) - 4sy + as(1-y)\\ =& 2s - 2sy + as(1-y)\\ >& 0\end{aligned}$$

as $y < 1$. Consequently, $x_{1/2} \geq 0$ holds, as all parameter values are positive.

To check the stability of the stationary states, we use the linearisation of the function $f(x)$. With

$$f'(x) = \frac{df}{dx} = 2\alpha_1 x - (\alpha_1 + \alpha_2),$$

it holds that

$$f'(x_1) = +\sqrt{(\alpha_1+\alpha_2)^2 - 4\alpha_1\alpha_3} > 0$$

and thus, x_1 is unstable. Whereas

$$f'(x_2) = -\sqrt{(\alpha_1+\alpha_2)^2 - 4\alpha_1\alpha_3} < 0$$

holds and thus, x_2 is stable.

A.2 Solution of the master equations (single-neighbor model)

Solution of (2). For a sufficiently small, we can assume that the sites X_l and X_{l-1} are independent. Thus, for a sufficiently small,

$$\begin{aligned}\frac{dP(X_l(t)=1)}{dt} \approx& as(1-y)P(X_l(t)=1)^2\\ &- (as(1-y) + s(1-y))\, P(X_l(t)=1)\\ &+ ys.\end{aligned} \tag{2}$$

Since we assume that $P(X_l(0)=1) = \frac{1}{2}$, we have a differential equation of the following form:

$$\frac{dx}{dt} = \alpha_1 x^2 - (\alpha_1+\alpha_2)x + \alpha_3, \quad f(0) = \frac{1}{2},$$

for $x = P(X_l(t) = 1)$, $\alpha_1 = as(1-y)$, $\alpha_2 = s(1+y)$, and $\alpha_3 = ys$. Additionally, we set $A := (\alpha_1+\alpha_2)^2 - 4\alpha_1\alpha_2$.

Let us consider the case where $y < 1$. As $a \geq 0$ and $s > 0$, $A > 0$ holds due to

$$\begin{aligned}(\alpha_1+\alpha_2)^2 - 4\alpha_1\alpha_3 =& \alpha_1^2 + +2\alpha_1\alpha_2 + \alpha_2^2 - 4\alpha_1\alpha_3\\ =& \alpha_1(\alpha_1 + 2\alpha_2 - 4\alpha_3) + \alpha_2^2\\ >& 0,\end{aligned}$$

since

$$\begin{aligned}\alpha_1 + 2\alpha_2 - 4\alpha_3 =& 2s(1+y) - 4sy + as(1-y)\\ =& 2s - 2sy + as(1-y)\\ >& 0\end{aligned}$$

as $y < 1$. Separation of variables yields

$$\begin{aligned} &\int\limits_{x_0}^{x} \frac{1}{\alpha_1 s^2 - (\alpha_1 + \alpha_2)s + \alpha_3} ds = \int_0^t ds \\ \iff &- \frac{2}{\sqrt{A}} \operatorname{artanh}\left(\frac{2\alpha_1 x - (\alpha_1 + \alpha_2)}{\sqrt{A}}\right) + \frac{2}{\sqrt{A}} \operatorname{artanh}\left(\frac{2\alpha_1 x_0 - (\alpha_1 + \alpha_2)}{\sqrt{A}}\right) = t. \end{aligned} \tag{10}$$

We computed this result using $A > 0$ and (21.7.1.2) from [2].

Furthermore, (10) is equivalent to

$$\begin{aligned} \operatorname{artanh}\left(\frac{2\alpha_1 x - (\alpha_1 + \alpha_2)}{\sqrt{A}}\right) &= -\frac{\sqrt{A}}{2} t + \operatorname{artanh}\left(\frac{2\alpha_1 x_0 - (\alpha_1 + \alpha_2)}{\sqrt{A}}\right) \\ \iff \frac{2\alpha_1 x - (\alpha_1 + \alpha_2)}{\sqrt{A}} &= \tanh\left(-\frac{\sqrt{A}}{2} t + \operatorname{artanh}\left(\frac{2\alpha_1 x_0 - (\alpha_1 + \alpha_2)}{\sqrt{A}}\right)\right), \end{aligned}$$

which is equivalent to

$$x(t) = \frac{1}{2\alpha_1}\left(\sqrt{A} \tanh\left(-\frac{\sqrt{A}}{2} t + \operatorname{artanh}\left(\frac{2\alpha_1 x_0 - (\alpha_1 + \alpha_2)}{\sqrt{A}}\right)\right) + (\alpha_1 + \alpha_2)\right),$$

which yields a solution of (2), for a sufficiently small and $y < 1$.

Now we consider the case where $y = 1$, and consequently $\alpha_1 = 0$, $\alpha_2 = 2s$ and $\alpha_3 = s$. In this case,

$$\frac{dx}{dt} = -2sx + s.$$

Multiplying both side of the previous equation by e^{2st}, we obtain

$$\begin{aligned} & e^{2st}\frac{dx}{dt} + 2sxe^{2st} = se^{2st} \\ \iff & \frac{d}{dt}\left(xe^{2st}\right) = se^{2st} \\ \iff & \int_0^t \frac{d}{dz}\left(xe^{2sz}\right) dz = \int_0^t se^{2sz} dz \\ \iff & x(t)e^{2st} - x_0 = \frac{1}{2}e^{2st} - \frac{1}{2}, \end{aligned}$$

which implies

$$x(t) = \frac{1}{2} - \frac{1}{2}e^{-2st} + x_0 e^{-2st}.$$

Solution of (2) without inhomogeneous term. If we consider the case where $y < 1$, (3) (the exact solution of (2)) is hard to analyze at first glance. Therefore, we further simplify Equation (2) by leaving out the inhomogeneous part (ys) of the equation. Note that this is applicable for y and s sufficiently small. Additionally, for a sufficiently small, it follows that

$$\begin{aligned} \frac{dP(X_l(t) = 1)}{dt} \approx &+ as(1 - y)P(X_l(t) = 1)^2 \\ &- (as(1 - y) + s(1 + y))\, P(X_l(t) = 1). \end{aligned}$$

Using $f(t) := P(X_l(t) = 1)$, we have a differential equation of the following form:

$$\frac{df(t)}{dt} = -(\alpha_1 + \alpha_2)f(t) + \alpha_1 f(t)^2, \quad f(0) = x_0 \neq 0.$$

To solve this differential equation, we first set

$$g(t) = \frac{1}{f(t)},$$

and obtain

$$\begin{aligned} \frac{dg(t)}{dt} &= -\frac{1}{f(t)^2}\frac{d}{dt}f(t) \\ &= (\alpha_1 + \alpha_2)\frac{1}{f(t)} - \alpha_1 \\ &= (\alpha_1 + \alpha_2)g(t) - \alpha_1, \end{aligned} \tag{11}$$

with $g(0) = \frac{1}{f(0)}$.

We first solve the differential equation for the function $g(t)$. Let us multiply both sides of (11) by $e^{-(\alpha_1+\alpha_2)t}$:

$$\begin{aligned} & e^{-(\alpha_1+\alpha_2)t}\frac{dg(t)}{dt} = e^{-(\alpha_1+\alpha_2)t}(\alpha_1 + \alpha_2)g(t) - e^{-(\alpha_1+\alpha_2)t}\alpha_1 \\ \Longleftrightarrow\ & e^{-(\alpha_1+\alpha_2)t}\frac{dg(t)}{dt} - e^{-(\alpha_1+\alpha_2)t}(\alpha_1 + \alpha_2)g(t) = -e^{-(\alpha_1+\alpha_2)t}\alpha_1 \\ \Longrightarrow\ & \frac{d}{dt}\left(e^{-(\alpha_1+\alpha_2)t}g(t)\right) = -e^{-(\alpha_1+\alpha_2)t}\alpha_1. \end{aligned}$$

Thus,

$$\begin{aligned} \int_0^t \frac{d}{ds}\left(e^{-(\alpha_1+\alpha_2)s}g(s)\right)\,\mathrm{ds} &= \int_0^t -e^{-(\alpha_1+\alpha_2)s}\alpha_1\,\mathrm{ds} \\ \Longrightarrow e^{-(\alpha_1+\alpha_2)t}g(t) - g(0) &= \frac{\alpha_1}{\alpha_1+\alpha_2}e^{-(\alpha_1+\alpha_2)t} - \frac{\alpha_1}{\alpha_1+\alpha_2} \\ \Longrightarrow g(t) = e^{(\alpha_1+\alpha_2)t}g(0) &+ \frac{\alpha_1}{\alpha_1+\alpha_2} - \frac{\alpha_1}{\alpha_1+\alpha_2}e^{(\alpha_1+\alpha_2)t}. \end{aligned}$$

Inserting $g(0) = \frac{1}{f(0)}$ and solving for $f(t)$ we obtain the solution:

$$f(t) = \frac{1}{\left(\frac{1}{x_0} - \frac{\alpha_1}{\alpha_1+\alpha_2}\right)e^{(\alpha_1+\alpha_2)t} + \frac{\alpha_1}{\alpha_1+\alpha_2}},$$

for $x_0 = f(0) = P(X_l(0) = 1)$.

A.3 Derivation of the master equations (both-neighbors model)

In this section, we once again consider a single site X_l, for $l \in L$, where $X_l(t)$ denotes the state of this site at time $t \geq 0$, and concentrate on approximating $P(X_l(t) = 1)$ by observing the dynamics of the process at $X_l(t)$. As before, we let A be the event where X_l changes its value due to a neighboring interaction, and let B denote the event where a spontaneous change occurs and thus X_l changes its value, and assume A and B to be independent of each other.

We first compute $P\left(X_l(t)=1|X_l(t-\Delta)=0\right)P\left(X_l(t-\Delta)=0\right)$:

$$\begin{aligned}
&P\left(X_l(t)=1|X_l(t-\Delta)=0\right)P\left(X_l(t-\Delta)=0\right)\\
=&P\left(\{X_l(t)=1\}\cap\{X_l(t-\Delta)=0\}\cap A\right)+P\left(\{X_l(t)=1\}\cap\{X_l(t-\Delta)=0\}\cap B\right)\\
=&P\left(\{X_l(t)=1\}\cap\{X_l(t-\Delta)=0\}\cap\{X_{l-1}(t-\Delta)=1,X_{l+1}(t-\Delta)=1\}\cap A\right)\\
&+P\left(\{X_l(t)=1\}\cap\{X_l(t-\Delta)=0\}\cap B\right)\\
=&P\left(\{X_l(t)=1\}\cap A|X_{l-1}(t-\Delta)=1,X_l(t-\Delta)=0,X_{l+1}(t-\Delta)=1\right)\\
&\quad\times P\left(X_{l-1}(t-\Delta)=1,X_l(t-\Delta)=0,X_{l+1}(t-\Delta)=1\right)\\
&+P\left(\{X_l(t)=1\}\cap B|X_l(t-\Delta)=0\right)P\left(X_l(t-\Delta)=0\right)\\
=&\left(ays\Delta+O(\Delta^2)\right)P\left(X_{l-1}(t-\Delta)=1,X_l(t-\Delta)=0,X_{l+1}(t-\Delta)=1\right)\\
&+\left(ys\Delta+O(\Delta^2)\right)P\left(X_l(t-\Delta)=0\right).
\end{aligned}$$

Now, we compute $P\left(X_l(t)=0|X_l(t-\Delta)=1\right)P\left(X_l(t-\Delta)=1\right)$:

$$\begin{aligned}
&P\left(X_l(t)=0|X_l(t-\Delta)=1\right)P\left(X_l(t-\Delta)=1\right)\\
=&P\left(\{X_l(t)=0\}\cap\{X_l(t-\Delta)=1\}\cap A\right)+P\left(\{X_l(t)=0\}\cap\{X_l(t-\Delta)=1\}\cap B\right)\\
=&P\left(\{X_l(t)=0\}\cap\{X_l(t-\Delta)=1\}\cap\{X_{l-1}(t-\Delta)=0\}\cap\{X_{l+1}(t-\Delta)=0\}\cap A\right)\\
&+P\left(\{X_l(t)=0\}\cap\{X_l(t-\Delta)=1\}\cap B\right)\\
=&P\left(\{X_l(t)=0\}\cap A|X_{l-1}(t-\Delta)=0,X_l(t-\Delta)=1,X_{l+1}(t-\Delta)=0\right)\\
&\quad\times P\left(X_{l-1}(t-\Delta)=0,X_l(t-\Delta)=1,X_{l+1}(t-\Delta)=0\right)\\
&+P\left(\{X_l(t)=0\}\cap B|X_l(t-\Delta)=1\right)P\left(X_l(t-\Delta)=1\right)\\
=&\left(as\Delta+O(\Delta^2)\right)P\left(X_{l-1}(t-\Delta)=0,X_l(t-\Delta)=1,X_{l+1}(t-\Delta)=0\right)\\
&+\left(s\Delta+O(\Delta^2)\right)P\left(X_l(t-\Delta)=1\right).
\end{aligned}$$

Once again, using the law of total probability (6), and Equation (7) as well as the two results above, we obtain

$$\begin{aligned}
P\left(X_l(t)=1\right)=&\left(ays\Delta+O(\Delta^2)\right)P\left(X_{l-1}(t-\Delta)=1,X_l(t-\Delta)=0,X_{l+1}(t-\Delta)=1\right)\\
&+\left(ys\Delta+O(\Delta^2)\right)P\left(X_l(t-\Delta)=0\right)\\
&+P\left(X_l(t-\Delta)=1\right)\\
&-\left(as\Delta+O(\Delta^2)\right)P\left(X_{l-1}(t-\Delta)=0,X_l(t-\Delta)=1,X_{l+1}(t-\Delta)=0\right)\\
&-\left(s\Delta+O(\Delta^2)\right)P\left(X_l(t-\Delta)=1\right).
\end{aligned}\tag{12}$$

Since $P(X_l(t)=0)+P(X_l(t)=1)=1$, subtracting $P\left(X_l(t-\Delta)=1\right)$ from both sides of (12), dividing both sides by Δ and letting it go to zero, we obtain the following differential equation

$$\begin{aligned}
\frac{dP\left(X_l(t)=1\right)}{dt}=&aysP\left(X_{l-1}(t)=1,X_l(t)=0,X_{l+1}(t)=1\right)\\
&+ys\\
&-asP\left(X_{l-1}(t)=0,X_l(t)=1,X_{l+1}(t)=0\right)\\
&-(s+ys)P\left(X_l(t)=1\right).
\end{aligned}\tag{13}$$

As it was the case with the single-neighbor-model, this analytical result also corresponds to our understanding of how $P(X_l(t)=1)$ should change in time, but now considering both the left and right nearest-neighbors. As a consequence of how we extended our model, there are now stronger assumptions for neighboring interactions to occur, namely both neighbors must have the same methylation status.

Before solving (13) it is important to find an approximation for the probabilities $P(X_l(t) = 0, X_{l-1}(t) = 1, X_{l+1}(t) = 1)$ and $P\left(X_l(t) = 1, X_{l-1}(t) = 0, X_{l+1}(t) = 0\right)$. Let us assume

$$\begin{aligned} &P\left(X_{l-1}(t) = 1, X_l(t) = 0, X_{l+1}(t) = 1\right) \\ \approx &P(X_{l-1}(t) = 1)P(X_l(t) = 0)P(X_{l+1}(t) = 1) \\ = &P(X_l(t) = 1)^2 - P(X_l(t) = 1)^3, \end{aligned}$$

and

$$\begin{aligned} &P\left(X_{l-1}(t) = 0, X_l(t) = 1, X_{l+1}(t) = 0\right) \\ \approx &P(X_{l-1}(t) = 0)P(X_l(t) = 1)P(X_{l+1}(t) = 0) \\ = &P(X_l(t) = 1)^3 - 2P(X_l(t) = 1)^2 + P(X_l(t) = 1), \end{aligned}$$

since $P(X_l(t) = 1) + P(X_l(t) = 0) = 1$, and $P(X_{l-1}(t) = 1) = P(X_l(t) = 1) = P(X_{l+1}(t) = 1)$ by construction of the model. Note, once again, that these simplifications are better suited for the cases where there is no neighboring interaction or almost no neighboring interaction, i.e. $a = 0$ or $a \approx 0$, since we assume independence of $X_l - 1, X_l, X_l + 1$. Inserting these approximations in (13), we obtain

$$\begin{aligned} \frac{dP\left(X_l(t) = 1\right)}{dt} \approx &\left(-as - (s + ys)\right) P(X_l(t) = 1) \\ &+ \left(ays + 2as\right) P(X_l(t) = 1)^2 \\ &+ \left(-ays + as\right) P(X_l(t) = 1)^3 \\ &+ ys \\ &=: f(x), \end{aligned} \tag{14}$$

for $x = P(X_l(t) = 1)$. Thus, we have

$$f(x) = \beta_3 x^3 + \beta_2 x^2 + \beta_1 x + \beta_4,$$

for $\beta_1 = -as - (s + ys)$, $\beta_2 = ays + 2as$, $\beta_3 = -ays + as$, and $\beta_4 = ys$.

As done in the single-neighbor-model, we can compute the roots of the function $f(x)$ and potentially analyze the stability of the stationary points, given they are in $[0, 1]$ and, thus meaningful to us. Additionally, if we further simplify (14) we can also potentially find an approximation of $P(X_l(t) = 1)$ that we can compare with the results of our simulation.

A.4 Further comparisons between the simulation and the analytical results

Comparison with the Equations (4) and (5). Note that we once again only consider the left-neighbor model. For $y \in \{0.1, 0.3, 0.6, 0.9\}$, $a \in \{0.1, 0.5, 1, 2, 10\}$, and the usual $s = 1.47 \cdot 10^{-3}$, Fig. 9 shows the observed mean methylation level over 30 simulated runs as well as the approximated solution obtained in Equation (4). Note that for Equation (4), we assumed y and s to be small and thus, omitted the term ys. In fact, we can indeed observe that the approximation works better for small values of the parameter y. In particular, as for $t \to \infty$,

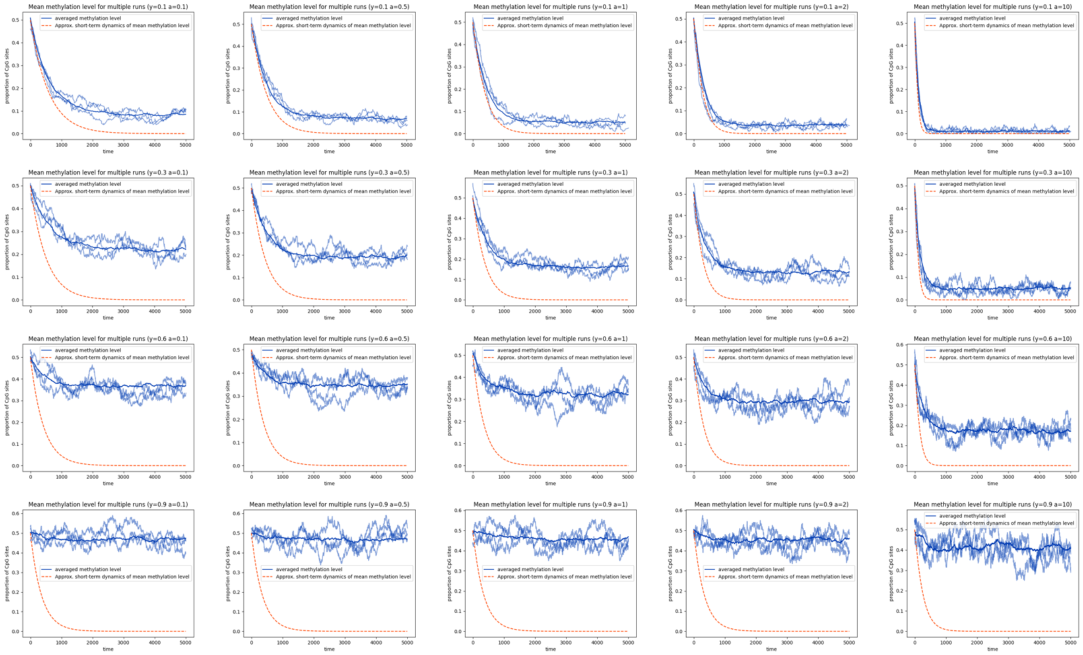

Figure 9: Mean methylation level over 30 runs including the approximated solution for the methylation level obtained in Equation (4) (red).

Approximation (4) tends to 0, thus (4) is more suitable in describing the short-term behavior of the underlying process. For small values of y, this can also be observed in Fig. 9.

Further, for the case where $y = 1$ (i.e. methylation and demethlation are equally likely) and the same s, Fig. 10 shows the mean methylation level over 30 runs and the corresponding solution (5) of the approximated master equation (2) for $a \in \{0.1, 0.5, 1, 2, 10\}$. We conclude that our analytical solution is able to more or less describe the mean methylation level, even for $a = 10$.

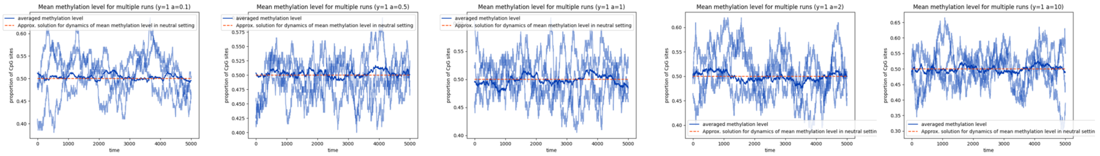

Figure 10: Mean methylation level over 30 runs including the exact solution for the approximation of the methylation level obtained in Equation (5) for $y = 1$ (red).

Fit with a General Function. Besides the analytical approaches, we also try to compute the average long-term stationary state of our model computationally and to fit a function to our simulated data, which describes the dynamics of the mean methylation level. In particular, our goal is to formulate a general function, which can describe the mean behavior of the process, dependent on the parameters a, s, and y. This will play a major role in understanding the impact of stronger neighboring interactions (a big) on the joint distribution of two neighboring sites. Especially since our current analytical results are all based on the assumption of a being small, i.e. independence between neighboring sites. Similar as above, we perform multiple runs and compute the average methylation level over these runs. From data observation, we propose a function of the form

$$f_{b,\beta,c}(t) := b \cdot (1 - e^{-\beta t}) + ce^{-\beta t} \tag{15}$$

to fit our data using the curve fit function from the SciPy Python library [25], which uses non-linear least squares. Results for $s = 1.47 \cdot 10^{-3}$, $y \in \{0.1, 0.3, 0.6, 0.9\}$, and $a \in \{0.1, 0.5, 1, 2, 10\}$ are shown in Fig. 11. Note that c describes the initial methylation level while b describes the long term mean methylation level.

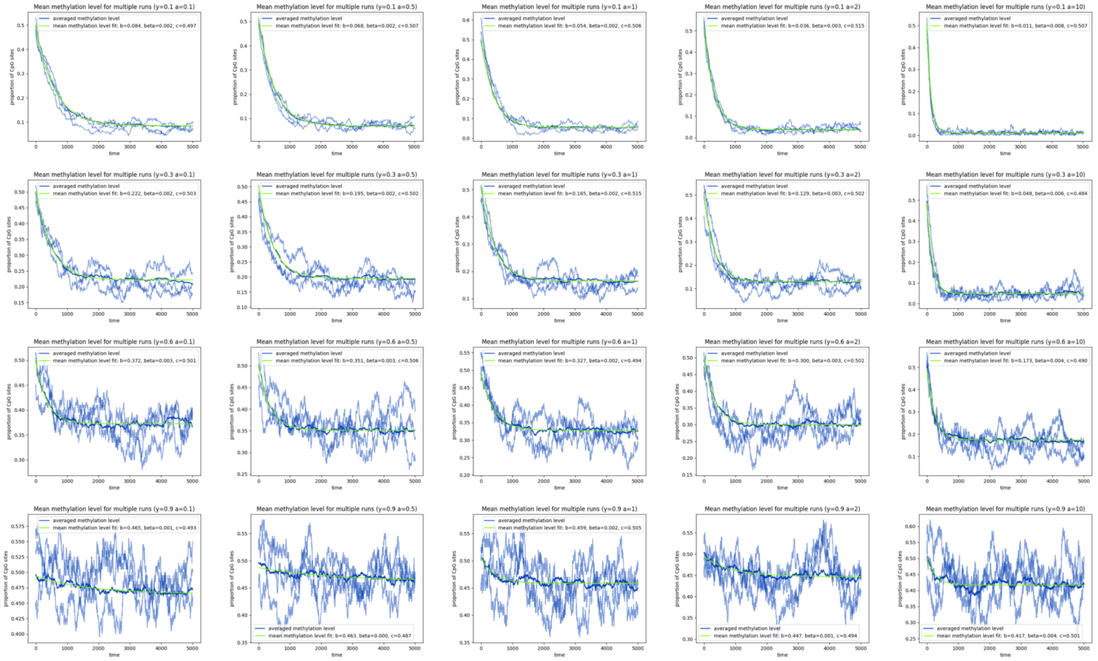

Figure 11: Averaged methylation level over 30 sequences and the fitted function (green). The parameter a varies from left to right in order $\{0.1, 0.5, 1, 2, 10\}$ and y varies from top to bottom in order $\{0.1, 0.3, 0.6, 0.9\}$. Note that the scaling of the y-axis is not the same for all plots.

Without analysis of the residuals, it is clear at first sight that the function described in (15) performs better for larger a-values than (3). This highlights again the impact of our independence-assumption.

A.5 Statistical tests for the cluster size

To validate the differences in mean cluster size for the three model formulations statistically, we need to start by answering the question: is the random variable $D_{a,y}$ that stands for "mean demethylated cluster size of a single sequence after a run of the simulation with parameter values a and y" nearly normally distributed (and analogously for the methylated clusters)? The answer to this question is interesting because it tells us whether we can perform other statistical tests, such as t-tests and ANOVAs. There are multiple possibilities to test for normal distribution. We start by using the D'Agostino-Pearson Omnibus test, which evaluates the skewness (a measure of symmetry to the mean) and kurtosis ("heaviness" of the tails) of the samples and compares them to those of a corresponding Gaussian distribution [3]. To use this test, we define a null-hypothesis of the form

$$H_0 : D_{a,y} \sim \mathcal{N}(\mu, \sigma^2),$$

where μ and σ^2 are the corresponding mean and variance, respectively. This first test will serve as a filter telling us which models and parameter combinations need to be inspected more closely: for this purpose, we impede passing the test by choosing a strict significance level. The error which we would like to avoid is the type II error (false negative), meaning that we do not want to falsely conclude that the mean (de-)methylated cluster size is normally distributed. Hence, we set the significance level to $\alpha = 0.1$ (H_0 is rejected if $p < \alpha$).

The results of the test for normal distribution yield the following p-values in case of demethylated clusters only: the model with $a = 0$ returns $p = 0.8506$ ($y = 0.5$), $p = 0.3018$ ($y = 1$), and $p = 0.3472$ ($y = 2$). For the model which only considers the influence of one neighbor we have $p = 0.7868$ ($a = 1$, $y = 1$), $p = 0.0674$ ($a = 1$, $y = 0.5$), and $p = 0.8394$ ($a = 1$, $y = 2$). The model including the influence of both neighbors yields the following p-values: $p = 0.5959$ ($a = 1$, $y = 1$), $p = 0.1242$ ($a = 1$, $y = 0.5$), and $p = 0.3879$ ($a = 1$, $y = 2$). Hence, H_0 must be rejected on the basis of the given samples in the cases of $y = 0.5$ and $y = 2$ for the one-neighbor-model. For the purely spontaneous model, the two-neighbor-model and the remaining parameter combination of the one-neighbor-model, H_0 can be accepted according to the test with the given samples.

For the mean methylated cluster sizes we obtain the following p-values: the spontaneous model yields $p = 0.6428$ ($y = 0.5$), $p = 0.8390$ ($y = 1$), and $p = 0.3783$ ($y = 2$). For the one-neighbor model we obtain $p = 0.0697$ ($a = 1$, $y = 1$), $p = 0.5877$ ($a = 1$, $y = 0.5$), and $p = 0.0808$ ($a = 1$, $y = 2$). Finally, the p-values resulting from the test for the two-neighbor model are given by $p = 0.2224$ ($a = 1$, $y = 1$), $p = 0.7865$ ($a = 1$, $y = 0.5$), and $p = 0.5746$ ($a = 1$, $y = 2$). Hence, regarding the methylated cluster sizes we have to reject H_0 in case of the one-neighbor model for the parameter choices $y = 1$, $y = 2$.

At this point, there are four groups associated with the one-neighbor model which we have to classify as non-normally distributed: the mean demethylated cluster size for the parameter sets $a = 1$, $y = 0.5$ and $a = 1$, $y = 2$, as well as the mean methylated cluster size for the parameter combinations $a = 1$, $y = 1$ and $a = 1$, $y = 2$. To find out whether it is possible to assess the influence of the type of model on the mean cluster sizes using ANOVAs we further need to examine the data from these groups. According to [7], non-normality is not necessarily problematic for the performance of an ANOVA as long as the data appears symmetric.

Figure 12 shows the histograms for these groups. It is visible that especially the three histograms on the right expose skewness, which can be quantified (from left to right) as −0.09, 0.13, 0.54 and 0.43. These values are still acceptable to fulfill the criterion of nearly-normal distribution [7], and we are also able to assume independence of the groups. The third assumption that is necessary for the performance of an ANOVA is the equality of standard deviations, but similar to the criterion of normally distributed data similarity is sufficient especially in case of equal sample sizes. With ANOVAs we can test the following null-hypotheses:

- H_0^0: $\mu^0_{y=1} = \mu^1_{a=1,y=1} = \mu^2_{a=1,y=1}$
- H_0^1: $\mu^0_{y=0.5} = \mu^1_{a=1,y=0.5} = \mu^2_{a=1,y=0.5}$

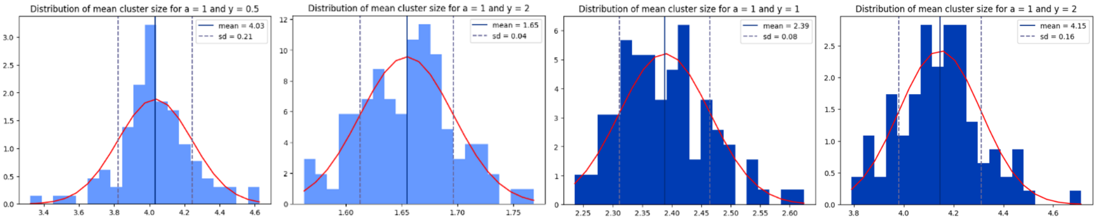

Figure 12: Distribution of the mean cluster sizes for the parameter combinations that did not pass the D'Agostino-Pearson Omnibus test. Data from the model that considers the influence of one neighbor ($a = 1$ in all cases). From left to right: $y = 0.5$, $y = 2$ (demethylated clusters), $y = 1$ and $y = 2$ (methylated clusters). The red line shows the probability density function of a normally distributed random variable with the same respective mean and standard deviation.

- H_0^2: $\mu_{y=2}^0 = \mu_{a=1,y=2}^1 = \mu_{a=1,y=2}^2$

where μ_y^0 is the true mean of demethylated clusters calculated by the purely spontaneous model, $\mu_{a,y}^1$ is the true mean of the demethylated cluster size for the model considering only one neighbor, and $\mu_{a=1,y=1}^2$ is the true mean for the two-neighbor-model. The null-hypotheses for the methylated cluster sizes were chosen analogously. We choose a significance level α that reduces the probability of a type I error, hence a lower significance level than before: $\alpha = 0.05$.

For the demethylated cluster sizes we obtain the following results: in the case of $y = 0.5$, we have a p-value of $p = 4.0010 \cdot 10^{-148}$, for $y = 1$ it is $p = 1.5689 \cdot 10^{-124}$, and for $y = 2$ p is given as $p = 3.0848 \cdot 10^{-82}$.

The methylated cluster sizes yield these p-values: for $y = 0.5$: $p = 1.7395 \cdot 10^{-63}$, for $y = 1$ we obtain $p = 3.4493 \cdot 10^{-128}$ and for $y = 2$ it is $p = 1.1959 \cdot 10^{-176}$.

These p-values tell us that we can reject each of our null-hypotheses, but they do not reveal which of the equality-assumptions should be rejected (e. g. in the case of H_0^0 it could be $\mu_{y=1}^0 \neq \mu_{a=1,y=1}^1 \neq \mu_{a=1,y=1}^2$, but also for instance $\mu_{y=1}^0 = \mu_{a=1,y=1}^1 \neq \mu_{a=1,y=1}^2$). To further assess the differences in mean cluster size we continue with pairwise tests. In the cases where both groups that are tested were accepted as being normally distributed in the D'Agostino-Pearson Omnibus test we use a two-sample t-test. In the remaining cases, we will use the Wilcoxon rank-sum test which tests for identical distribution. Identical distribution also means "same true mean", and hence the Wilcoxon rank-sum test serves as a good replacement for the two-sample t-test in the cases where we were unable to accept normal distribution under the chosen significance level [5].

For all tests the same significance level $\alpha = 0.05$ is fixed. Table 1 shows the p-values that were obtained from the pairwise testing.

A.6 Influence of inheritance on the cluster size

We have seen that there are multiple possibilities of how epigenetic information can be inherited. In the analysis of the simulation we have only considered no linkage between sites, which tears clusters apart if they are not present in enough individuals of the cohort. But as soon as there is linkage between sites, clusters can be maintained even if they are only present in single individuals. An interesting question in this context is whether this effect is large enough to create a difference in mean cluster size evolution over multiple generations.

To serve as a basis for discussion, we perform a small investigation of the model that considers the influence of one neighbor. In order to provide a substantial answer to the question,

y-value	Tested pair	Type of test	p-value
$y = 0.5$	0 vs. 1	Wilcoxon rank-sum	2.5239e-34 (D)
		2-sample t-test	8.4613e-52 (M)
	0 vs. 2	2-sample t-test	2.1112e-115 (D)
		2-sample t-test	1.6751e-46 (M)
	1 vs. 2	Wilcoxon rank-sum	0.1426 (D)
		2-sample t-test	0.5814 (M)
$y = 1$	0 vs. 1	2-sample t-test	7.9096e-89 (D)
		2-sample t-test	6.7889e-94 (M)
	0 vs. 2	2-sample t-test	2.4825e-95 (D)
		2-sample t-test	9.4430e-92 (M)
	1 vs. 2	2-sample t-test	0.0441 (D)
		Wilcoxon rank-sum	0.9357 (M)
$y = 2$	0 vs. 1	Wilcoxon rank-sum	8.1107e-34 (D)
		Wilcoxon rank-sum	2.5239e-34 (M)
	0 vs. 2	2-sample t-test	1.6152e-63 (D)
		2-sample t-test	3.3653e-129 (M)
	1 vs. 2	Wilcoxon rank-sum	0.8950 (D)
		Wilcoxon rank-sum	0.7250 (M)

Table 1: Choice and p-values of the pairwise tests. “M” stands for methylated clusters, “D” means demethylated clusters. Notation of the models: “0”: only spontaneous reactions, “1”: influence by one neighbor, “2”: influence by both neighbors.

more statistical analysis would be necessary as well as perhaps also the manipulation of more parameters than in the following.

Similarly as before we choose the parameter values $L = 200$, $s = 1.47 \cdot 10^{-3}$, $a = 1$ and $t_{\text{end}} = 1000$ (because Section 4 showed that it takes about 1000 time units to reach the equilibrium cluster size). Furthermore, we will observe three different y-values: $y \in \{0.5, 1, 2\}$. As before, $N = 100$ individuals (per generation) and 50 generations are simulated, and the initial cohort is given by random combinations of ones and zeros sampled from a uniform distribution.

In case of the inheritance with linked sites, we assume that each individual has two parents that are randomly drawn from the previous generation and we add break points for the recombination after each fifth CpG site (case 1) and after each 50th site (case 2). In case 1, many snippets of the parents' sequences alternate within quite short distances, whereas in case 2 only two snippets per parent are passed on to the next generation.

The mean (de-)methylated cluster sizes over the generations are shown in Figs. 13, 14 and 15. Their evolution appears very similar and it looks as if we can answer the initial question with “no”, but to be sure we would need to perform more analysis, as is done for instance in Section 4.

A.7 Extension of the both-neighbors-model

Model formulation. We consider a single CpG site X_l, $l \in \{1, ..., L\}$. Keeping the usual periodic boundary condition, we define the nearest right neighbor as X_{l+1}. Extending our model to also account for neighboring interactions when $X_{l-1} \neq X_{l+1}$, we make the following assumptions:

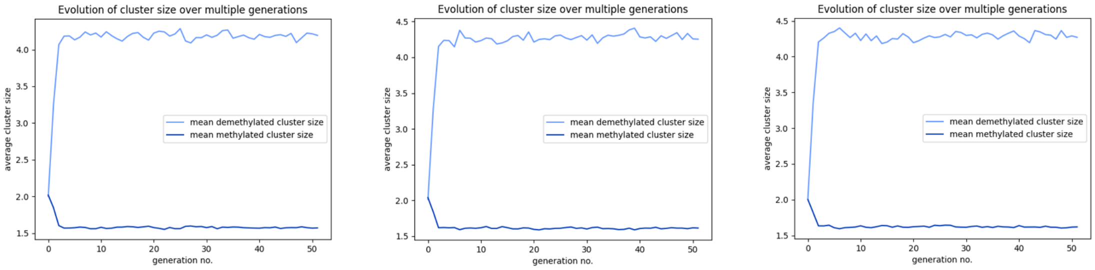

Figure 13: Evolution of the mean cluster sizes over all $N = 100$ individuals per generation for $y = 0.5$. From left to right: inheritance without linkage, inheritance with linkage and snippets of length 5, inheritance with linkage and snippets of length 50.

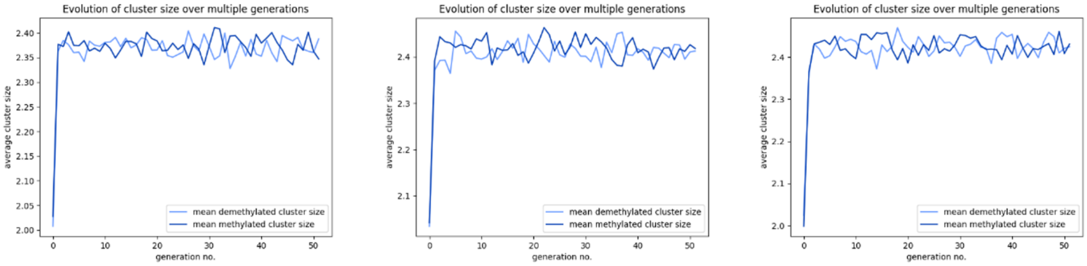

Figure 14: Cluster size evolution for $y = 1$.

1. For $X_{l-1} = X_l = X_{l+1}$, we assume that a change of the methylation status at site X_l can only occur due to a spontaneous, non-collaborative reaction.

2. For $X_{l-1} = X_{l+1}$ and $X_{l-1} \neq X_l$, we assume that a change of the methylation status at site X_l can occur due to a spontaneous, non-collaborative reaction and additionally, due to the influence of the left and right neighbors.

3. For $X_{l-1} \neq X_{l+1}$, we assume that a change of the methylation status at site X_l can occur due to a spontaneous, non-collaborative reaction. Additionally, we assume that the influences of the left and right neighbors "compete" with each other. Depending on the value of y, a change in methylation state of X_l occurs but only from a methylated to a demethylated state for $y < 1$ or from a demethylated to a methylated state for $y > 1$.

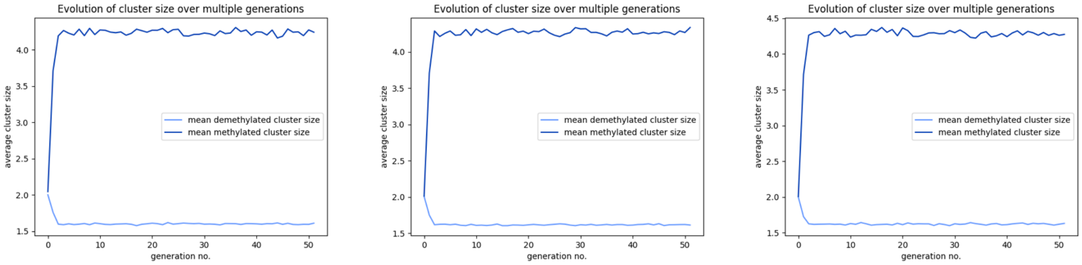

Figure 15: Cluster size evolution for $y = 2$.

For $y < 1$, the resulting reaction system is described by

$$\begin{aligned} 1 &\xrightarrow{k_1} 0 \\ 0 &\xrightarrow{k_2} 1 \\ 0,1,0 &\xrightarrow{\tilde{k}_3} 0,0,0 \\ 1,0,1 &\xrightarrow{\tilde{k}_4} 1,1,1, \\ 0,1,1 &\xrightarrow{\tilde{k}_5} 0,0,1 \\ 1,1,0 &\xrightarrow{\tilde{k}_5} 1,0,0 \end{aligned}$$

where $k_1 = s$, $k_2 = sy$, $\tilde{k}_3 = as$, $\tilde{k}_4 = asy$, and $\tilde{k}_5 = as - asy$. For $y < 1$, the resulting reaction system is described by

$$\begin{aligned} 1 &\xrightarrow{k_1} 0 \\ 0 &\xrightarrow{k_2} 1 \\ 0,1,0 &\xrightarrow{\tilde{k}_3} 0,0,0 \\ 1,0,1 &\xrightarrow{\tilde{k}_4} 1,1,1, \\ 0,0,1 &\xrightarrow{\tilde{k}_6} 0,1,1 \\ 1,0,0 &\xrightarrow{\tilde{k}_6} 1,1,0 \end{aligned}$$

where $k_1 = s$, $k_2 = sy$, $\tilde{k}_3 = as$, $\tilde{k}_4 = asy$, and $\tilde{k}_6 = asy - as$. The parameters s and y are defined the same way as Section 2. a measures the strength of neighboring influence in the same way as in the formulation of the both-neighbors-model.

Master equation. Let $y < 1$. We consider a single site X_l, for $l \in L$, where $X_l(t)$ denotes the status of this site at time $t \geq 0$. Our goal is to estimate $P(X_l(t) = 1)$ by observing the dynamics of the process at $X_l(t)$ for the case where $y < 1$. The case where $y > 1$ can be constructed analogously.

As before, we let A be the event where X_l changes its value due to a neighboring interaction, and let B denote the event where a spontaneous change occurs and thus X_l changes its value. Assume A and B to be independent of each other.

We first compute $P\left(X_l(t)=1|X_l(t-\Delta)=0\right)P\left(X_l(t-\Delta)=0\right)$:

$$\begin{aligned}
&P\left(X_l(t)=1|X_l(t-\Delta)=0\right)P\left(X_l(t-\Delta)=0\right)\\
=&P\left(X_l(t)=1,X_l(t-\Delta)=0\right)\\
=&P\left(\{X_l(t)=1\}\cap\{X_l(t-\Delta)=0\}\cap A\right)+P\left(\{X_l(t)=1\}\cap\{X_l(t-\Delta)=0\}\cap B\right)\\
=&P\left(\{X_l(t)=1\}\cap\{X_l(t-\Delta)=0\}\cap\{X_{l-1}(t-\Delta)=1,X_{l+1}(t-\Delta)=1\}\cap A\right)\\
&+P\left(\{X_l(t)=1\}\cap\{X_l(t-\Delta)=0\}\cap B\right)\\
=&P\left(\{X_l(t)=1\}\cap A|X_{l-1}(t-\Delta)=1,X_l(t-\Delta)=0,X_{l+1}(t-\Delta)=1\right)\\
&\quad\times P\left(X_{l-1}(t-\Delta)=1,X_l(t-\Delta)=0,X_{l+1}(t-\Delta)=1\right)\\
&+P\left(\{X_l(t)=1\}\cap B|X_l(t-\Delta)=0\right)P\left(X_l(t-\Delta)=0\right)\\
=&\left(asy\Delta+O(\Delta^2)\right)P\left(X_{l-1}(t-\Delta)=1,X_l(t-\Delta)=0,X_{l+1}(t-\Delta)=1\right)\\
&+\left(sy\Delta+O(\Delta^2)\right)P\left(X_l(t-\Delta)=0\right).
\end{aligned}$$

Now, we compute $P\left(X_l(t)=0|X_l(t-\Delta)=1\right)$:

$$\begin{aligned}
&P\left(X_l(t)=0|X_l(t-\Delta)=1\right)P\left(X_l(t-\Delta)=1\right)\\
=&P\left(X_l(t)=0,X_l(t-\Delta)=1\right)\\
=&P\left(\{X_l(t)=0\}\cap\{X_l(t-\Delta)=1\}\cap A\right)+P\left(\{X_l(t)=0\}\cap\{X_l(t-\Delta)=1\}\cap B\right)\\
=&P\left(\{X_l(t)=0\}\cap\{X_l(t-\Delta)=1\}\cap\{X_{l-1}(t-\Delta)=0\}\cap\{X_{l+1}(t-\Delta)=0\}\cap A\right)\\
&+P\left(\{X_l(t)=0\}\cap\{X_l(t-\Delta)=1\}\cap\{X_{l-1}(t-\Delta)\neq X_{l+1}(t-\Delta)\}\cap A\right)\\
&+P\left(\{X_l(t)=0\}\cap\{X_l(t-\Delta)=1\}\cap B\right)\\
=&P\left(\{X_l(t)=0\}\cap A|X_{l-1}(t-\Delta)=0,X_l(t-\Delta)=1,X_{l+1}(t-\Delta)=0\right)\\
&\quad\times P\left(X_{l-1}(t-\Delta)=0,X_l(t-\Delta)=1,X_{l+1}(t-\Delta)=0\right)\\
&+P\left(\{X_l(t)=0\}\cap A|X_l(t-\Delta)=1,X_{l-1}(t-\Delta)\neq X_{l+1}(t-\Delta)\right)\\
&\quad\times P\left(\{X_l(t-\Delta)=1\}\cap\{X_{l-1}(t-\Delta)\neq X_{l+1}(t-\Delta)\}\right)\\
&+P\left(\{X_l(t)=0\}\cap B|X_l(t-\Delta)=1\right)P\left(X_l(t-\Delta)=1\right)\\
=&\left(as\Delta+O(\Delta^2)\right)P\left(X_{l-1}(t-\Delta)=0,X_l(t-\Delta)=1,X_{l+1}(t-\Delta)=0\right)\\
&+\left((as-asy)\Delta+O(\Delta^2)\right)P\left(\{X_l(t-\Delta)=1\}\cap\{X_{l-1}(t-\Delta)\neq X_{l+1}(t-\Delta)\}\right)\\
&+\left(s\Delta+O(\Delta^2)\right)P\left(X_l(t-\Delta)=1\right).
\end{aligned}$$

Thus, using the law of total probability (6) and Equation (7), it follows

$$\begin{aligned}
P\left(X_l(t)=1\right)=&\left(asy\Delta+O(\Delta^2)\right)P\left(X_{l-1}(t-\Delta)=1,X_l(t-\Delta)=0,X_{l+1}(t-\Delta)=1\right)\\
&+\left(sy\Delta+O(\Delta^2)\right)P\left(X_l(t-\Delta)=0\right)\\
&-\left(as\Delta+O(\Delta^2)\right)P\left(X_{l-1}(t-\Delta)=0,X_l(t-\Delta)=1,X_{l+1}(t-\Delta)=0\right)\\
&-\left((as-asy)\Delta+O(\Delta^2)\right)P\left(\{X_l(t-\Delta)=1\}\cap\{X_{l-1}(t-\Delta)\neq X_{l+1}(t-\Delta)\}\right)\\
&-\left(s\Delta+O(\Delta^2)\right)P\left(X_l(t-\Delta)=1\right)\\
&+P\left(X_l(t-\Delta)=1\right).
\end{aligned}$$

Since $P(X_l(t)=0)+P(X_l(t)=1)=1$, subtracting $P\left(X_l(t-\Delta)=1\right)$ from both sides of the previous equation, dividing both sides by Δ and letting it go to zero, we obtain the following differential equation

$$\begin{aligned}\frac{dP\left(X_l(t)=1\right)}{dt} =& asyP\left(X_l(t)=0, X_{l-1}(t)=1, X_{l+1}(t)=1\right)\\ &+ sy\\ &- asP\left(X_l(t)=1, X_{l-1}(t)=0, X_{l+1}(t)=0\right)\\ &- (as-asy)P\left(\{X_l(t)=1\}\cap\{X_{l-1}(t)\neq X_{l+1}(t)\}\right)\\ &- (s+sy)P\left(X_l(t)=1\right).\end{aligned}$$

As done previously, we could potentially simplify this expression, compute the stationary points and check their stability behavior. Additionally, if possible, we could approximate a solution of the master equation.

We observe that the master equation derived here has an additional term that accounts for the "weaker" neighboring interactions that can occur when $X_{l-1}\neq X_{l+1}$, compared to our model extension in Section 2.

As a result of the additional interaction, we speculate that this model could create larger clusters of demethylated sites. Nonetheless, further numerical analyzes are needed in order to determine whether this new potential model is significantly different to the others.