Saarland University
Center for Bioinformatics
Master's Program in Bioinformatics

Master's Thesis in Bioinformatics

Design and calibration of stochastic models for DNA methylation patterns

submitted by

Andrea Kupitz

on March 2019

Supervisor
Prof. Dr. Verena Wolf

Advisor
M.Sc. Alexander Lück

Reviewers
Prof. Dr. Verena Wolf
Prof. Dr. Volkhard Helms

Zentrum für Bioinformatik

Kupitz, Andrea
Design and calibration of stochastic models for DNA methylation patterns
Master's Thesis in Bioinformatics
Universität des Saarlandes
Saarbrücken, Germany
March 2019

Declaration

I hereby confirm that this thesis is my own work and that I have documented all sources used.

I hereby declare that the submitted digital and hardcopy versions of this thesis correspond to each other. I give permission to the Universität des Saarlandes to duplicate and publish this work.

Saarbrücken, on March 2019

Andrea Kupitz

Abstract

The expression of genes in the human genome is not only based on the DNA sequence; it relies on epigenetic modifications like DNA-methylations. Hereby, gene expression is inactivated by binding of methyl-groups to a cytosine-phosphate-guanine (CpG) dinucleotide at the promoter region of the concerned gene. The binding is performed by specific enzymes - the DNA Methyltransferases (DNMTs). The specific function of DNMTs is not fully determined.

In the following, the methylation of DNA is modelled using an Markov Chain Monte Carlo (MCMC) algorithm and parameters are estimated by Maximum Likelihood Estimation (MLE). Alternatively, parameter estimation is performed with an implementation of the Approximative Bayesian Computation (ABC) method. Moreover, a method to compare distributions of methylation patterns is provided.

We find differences in the function of the different DNMTs that are consistent with current biological findings. Thus, using this model, some parameters are difficult to identify because they seem to be conditionally dependant.

Acknowledgments

I would like to thank deeply Prof. Dr. Verena Wolf and Alexander Lück, who introduced me to my master theme and always offered me their advice and support which I appreciated very much.

I owe thanks to both of them for the opportunity to develop this master thesis at the *Center for Bioinformatics*.

Furthermore I would like to thank all other people who put in touch with me during the development of this master thesis and who were very helpful in providing information.

Contents

List of Tables

List of Figures

List of Abbreviations

ABC	Approximative Bayesian Computation	ii
CGI	CpG island	1
CpG	cytosine-phosphate-guanine	ii
DNMT	DNA Methyltransferase	ii
HMM	Hidden Markov Model	2
MCMC	Markov Chain Monte Carlo	ii
MLE	Maximum Likelihood Estimation	ii
RV	Random Variable	2

Chapter 1

Introduction

1.1 DNA methylation

Epigenetics is the name of the science that studies the heritable information not relying on changes in the DNA sequence and influencing the organisms phenotype. There exist two kinds of epigenetic modifications: Chromatin and DNA modifications. Chromatin is the three-dimensional arrangement of DNA and a histone protein. The modification of Chromatin is performed by binding of an amino group or RNA.[1]

The most common DNA modification is the addition of one methyl group to the fifth position in the cytosine ring of DNA. Methylation occurs mainly at CpGs, where a cytosine nucleotide (C) is followed by a guanine nucleotide (G) in the DNA sequence.[2] Whereas the majority of CpGs in mammals are methylated, so called CpG island (CGI) are rather unmethylated. The CGIs are associated to gene regulation as they are often located at the promoter region of genes. Hereby, methylation inhibits the gene expression; ensuring that changes in the methylation pattern of the DNA are effecting diseases like cancer.[3] Hypomethylation of repeat elements for example results in an unstable DNA and may increase the risk of cancer; as also the hypermethylation of cancer suppressor genes.[2] Finally, DNA methylation is responsible for genomic imprinting and X-chromosome inactivation, whereby one of the two alleles respectively is transcribed and the other inactivated by methylation. Dysregulation may contribute to diseases like Prader-Willi syndrome, Angelman syndrome and cancer. [4]

The transfer of the methyl group to the DNA is performed by DNMTs. Five different DNMTs are distinguished in mammals: DNMT1, DNMT2, DNMT3A, DNMT3B and DNMT3L.[2]
DNMT1 is known as maintanance methyltransferase as its activity is associated with the DNA replication process. Thereby, not only the DNA of the cell is transmitted from one cell generation to another, but also the methylation patterns. DNMT1 has a preference for hemimythelated DNA, which means that one of the opposite CpGs is methylated and the other unmethylated. Subsequently, DNMT1 transfers the methylation by methylating the positions on

the daughter strand that are methylated at the same position on the parental strand.[2]
DNMT2 is negligible if human DNA methylation is considered, because it methylates small RNA at the anticodon loop.[**DNMT2**]
DNMT3A and DNMT3B are de novo methyltransferases, which work during the early embryonic development to synthesize new methylations. Hereby, the enzymes do not show any preference neither for hemi- or unmethylated DNA nor for a DNA-strand. In other words, DNMT3A/B may add a methyl-group to any non-methylated CpG at any DNA-strand. DNMT3L does not catabolize methylation, but increases the binding ability of DNMT3A/B and thus is required for establishing genomic imprinting.[2]
The loss of one of the methyltransferases leads to embryonic lethality.[2]

So far a lot is known about the function of DNMTs, but some properties still remain unexplained:

- Why and where do the DNMTs bind?
- On which conditions does the methylation event depend?
- Which enzymes are able to methylate multiple CpGs in a row without deassociating from the DNA?
- How much de novo and maintenance methylation do DNMT1 and DNMT3A/B perform?

To study the function of DNMTs, a computational model is designed and fitted using real biological data. These models are based on Markov models.

1.2 Markov models

A discrete-time Markov model is a stochastic model that fulfils the Markov property. If the next state can only be determined by the current state and not by the previous one, a chain holds the Markov property.

A **Markov chain** is the simplest Markov model. This chain is a sequence of variables $X_1, X_2, ...$ whose outcomes are random, so called **Random Variables (RVs)**. Each RV has a number of possible outcomes, the state space. Given a set of states $S = \{s_1, ..., s_n\}$ with transients between the states, p_{ij} is the transition probability of s_i to s_j. And P of size $|S| \cdot |S|$ is the transition matrix containing the transition probabilities between all states.
Let X_0 be the initial distribution of the chain X s.t. $\sum_{x \in X_0} x = 1$, then $X_1 = X_0 * P$ is the distribution of X at time t=1. We call π the **equilibrium distribution**; the distribution of X that does not change from one time step to another or formally: $\pi * P = \pi$.
Hidden Markov Models (HMMs) are an extension of Markov chains. There are two kinds of states in this model; hidden states S and observable output states O. Similarly, there is a transition probability matrix between the states in S and a matrix of output probabilities B between the hidden and the output states.
This model is used, when there is information about the output of a process,

but no knowledge about the states of the system.

1.3 Bayesian statistics

1.4 Problem

In the context of DNMTs, Markov models are used to simulate the function of the enzymes, making use of results from biological experiments. Here, few is known about the properties of methyltransferases, but the methylation patterns before and after the catabolism of the DNMTs. Each methylation state (unmethylated, hemimethylated, methylated) of one CpG is an output state, whereas the binding state of the DNMT at each CpG is a hidden state. Thus the traversing probabilities between the hidden states P are equal to the probabilities of the DNMTs to bind/fall of and the output probabilities represent the different methylation probabilities:

$$O_i := \text{methylation state of CpG i}$$

$$\tag{1.1}$$

$$O_i = \begin{cases} 0 & \text{unmethylated} \\ 1 & \text{hemimethylated} \\ 2 & \text{hemimethylated} \\ 3 & \text{unmethylated} \end{cases} \tag{1.2}$$

$$S_i := \text{binding state of DNMT at CpG i} \tag{1.3}$$

$$S_i = \begin{cases} 0 & \text{unbound} \\ 1 & \text{bound} \end{cases} \quad P := \text{probability of DNMT to change binding state}$$

$$\tag{1.4}$$

$$B := \text{methylation probabilities} \tag{1.5}$$

$$\tag{1.6}$$

Figure 1.1 shows the four kind of methylation patterns.

Given the output sequence, we are able to determine the most likely sequence of S. Moreover, P and B can be reconstructed by computer simulation of the Discrete-time Markov Chain (DTMC). $Onekindofsimulationis\mathbf{MCMC}$.

Recently, different models and computational approaches were studied in order to predict the methylation process. Here we present some of them.

(a) *(b)* *(c)* *(d)*

Figure 1.1: *Kind of methylation patterns; each figure shows two opposite CpGs on the parental and the daughter strand; each circle represents one CpG; plane red circles are methylated CpGs; (a) unmehtylated, (b) hemimethylated (upper strand unmethylated, lower strand methylated), (c) hemimethylated (vice versa), (d) fully methylated CpGs*

1.5 Related work

Chapter 2

Methods

Chapter 3

Evaluation

Chapter 4

Discussion

todo

Appendix A

Regulation

Bibliography

[1] B. Weinhold. “Epigenetics: The Science of Change”. In: *Environment Health Perspective* 114.3 (2006). DOI: 10.1289/ehp.114-a160.

[2] B. Jin et al. “DNA Methylation”. In: *Genes Cancer* 2.6 (2011). DOI: 10.1177/1947601910393957.

[3] K. Janitz and M. Janitz. “Handbook of Epigenetics”. In: Academic Press, 2011. Chap. 12. DOI: 10.1016/B978-0-12-375709-8.00012-5.

[4] W. Reik and J. Walter. “Genomic imprinting: parental influence on the genome”. In: *Nature Reviews Genetics* 2 (2001).