Dynamics of Epigenetic Reader Proteins and their Interplay with Expression in Development

Dissertation
zur Erlangung des Grades

des Doktors der Naturwissenschaften (Dr. rer. nat.)

der Naturwissenschaftlich-Technischen, Fakultät III
Chemie, Pharmazie, Bio- und Werkstoffwissenschaften
der Universität des Saarlandes

von
M.Sc. Siba Shanak

Saarbrücken

01.12.2014

The work of this PhD thesis has been supported by the DAAD Scholarship

Tag des Kolloquiums: 15. Januar 2015

Dekan: Prof. Dr.-Ing. Dirk Bähre

Berichterstatter: Prof. Dr. Volkhard Helms

Prof. Dr. Albrecht Ott

Vorsitz: Prof. Dr. Uli Müller

Akademischer Mitarbeiter: Dr. Sonja Kessler

Table of Contents

Acknowledgements

First, I would like to express my gratitude to Prof Dr. Helms who was of great help for me in getting the chance to accomplish my masters and PhD studies in Germany and the Helms group. He was one of the few to mention making this dream come true. He was so cooperative, nice, humane, and understanding. Beyond this, he enriched my studies with fruitful scientific discussions.

I am indebt to my colleagues in work who have made a relaxing environment and introduced well the teamwork. I owe my sincere thanks in this respect to Ahmad Barghash who was greatly helpful and nice in scientific and social lives. I also thank Özlem Ulucan who was quite salutary in terms of my research and provided a great assistance in my work. I also want to thank deeply the 'Mensa club', namely Maryam and Ruzi for the continuous moral and spiritual support. I would also like to thank Thorsten Will and Kerstin Gronow-Pudelek for their helpful suggestions to the thesis, the German translations and the great cooperation. Not to forget all my friends and colleagues in the workgroup for the great support as well as the scientific collaborators in the workgroup of Prof Dr. Ott-Biophysics.

Special favor belongs to mom, dad, and siblings for sharing me the determination to make the impossible possible. I also owe a significant debt to my best beloved innocent kids, Sharif and Liane, who were supporting me like adults, and by whom I gained the inspiration and determination to go through this.

I owe my deepest gratitude to my husband. Hussein's emotional guidance and steadfast support have been invaluable over the years and I feel incredibly privileged to have him in my life. Hussein, I grant you my success.

Finally yet importantly, thank you God for making this stage go steadily and smoothly, and for giving me the nice people to surround and support me.

Abstract

DNA methylation is a crucial epigenetic signal that modulates gene expression in time and space.

Two pivotal forms of DNA methylation are C5-cytosine methylation and N6-adenine methylation. The emergence of distinct forms of DNA methylation poses the question why nature utilized several forms of DNA methylation to modulate gene expression. Thus, we present here conventional MD simulations and free energy calculations for the two forms of DNA methylation.

First, we showed through free energy calculations that not all forms of methylated nucleic acid bases are more hydrophobic than the respective non-methylated bases. We used this insight to understand the consequent stability of naturally methylated DNA sequences and conducted further free energy calculations, which showed that nature favors specific sequence contents as targets for DNA methylation.

We extended our scope to the specific binding of proteins to methylated DNA and showed that some proteins induce structural rearrangements of the DNA that are beneficial for the modulation of gene expression. We also shed light on the diverse enthalpic and entropic contributions to the binding process.

Finally, we investigated the epigenetic codes that can modulate splicing and correlated our knowledge to exon expression. We studied this in human tissues across developmental stages. We found that alternative splicing correlates well with alternative read numbers of several epigenetic marks in genes crucial for development.

Zusammenfassung

DNA-Methylierung ist ein wesentliches epigenetisches Signal um Genexpression anzupassen.

Die beiden grundsätzlichen Methylierungsarten von DNA sind die C5-Cytosin-Methylierung und N6-Adenin-Methylierung. Die Entstehung verschiedenartiger Ausprägungen dieser Modifikation wirft die Frage auf warum sich die Natur mehrere Arten der Modifikation zur Expressionsregulierung zunutze macht. Dazu zogen wir Moleküldynamik-Simulationen als auch Freie Energie-Berechnungen für beide Methylierungsarten zu Rate.

Zuerst zeigten wir anhand Berechnungen der Freien Energie dass nicht alle Formen methylierter Nukleinsäurebasen hydrophober als ihre nicht-methylierten Basen sind. Wir nutzten diese Erkenntnis um die daraus folgende Stabilität natürlich methylierter DNA-Sequenzen zu verstehen.

Wir erweiterten unsere Untersuchungen auf das gezielte Binden von Proteinen an methylierter DNA und erkannten, dass einige Proteine strukturelle Umwandlungen der DNA herbeiführten, die die Regulierung der Genexpression begünstigen. Zusätzlich konnten wir Aufschluss über die unterschiedlichen enthalpischen und entropischen Beiträge des Bindeprozesses geben.

Zuletzt überprüften wir epigenetische Marker auf der Ebene der DNA die das Spleißen steuern und korrelierten diese mit der Expression von Exons in menschlichem Gewebe über mehrere Entwicklungsstadien. Dabei stellten wir fest, dass das alternative Spleißen von Genen, die wichtig für die Entwicklung sind, eng mit einigen epigenetischen Signalen zusammenhängt.

Chapter 1

Introduction and background

As a starting point of our lives, the zygote condenses our genetic heritage that we get from our parents. Still nature has a lot to do in terms of changes that allocate to our genes in all orders of magnitude. This is the 'epigenetic fashion'. It starts right with the first mitotic cell divisions, goes on through the different stages of development and makes man. It also affects human characteristics and behaviors based on input from the life style and the environmental conditions. It is moreover the gigantic fellow of all living organisms. The first chapter introduces the term of epigenetics that is the main target of this thesis and explains some of the different concepts associated with this 'cell fashion'. It closes with the objective of this work.

1.1 Epigenetics

Epigenetics can be explained as the study of the heritable changes that allocate to the organismal genomes (namely seen through differential phenotypic expressions) but cannot be explained in terms of changes in DNA sequence (1). Such changes are very well associated with the modifications in DNA sequence through evolution. Thus, it would be expected that organisms do 'learn' by experience through acceptance or refusal of such changes to be allocated to their genomes.

The mechanisms of epigenetic changes allocated to the genome of an organism are explained in figure 1.1 (2). These variations include DNA methylation, histone modifications and other chromatin remodeling events, plus the effect induced by non-coding RNAs. In this thesis, DNA methylation was the main tackled epigenetic modification. For this aim, we will go briefly over the other forms of epigenetic modifications and end with DNA methylation.

1.1.1 Non-coding RNA

Whereas the eukaryotic genomes transcribe up to 90% of the genomic DNA (3), only 1-2% represent genes that code for proteins. The vast majority of the genomes are transcribed as non-coding RNAs (nc-RNAs), crucial effectors across developmental stages (4). Their regulation can be either infrastructural, such as the several forms bound to transcription, and regulatory, such as for instance, micro-RNA (miRNA),

small-interfering RNA (siRNA), and long non-coding RNA, all known for their role in DNA silencing (5,6).

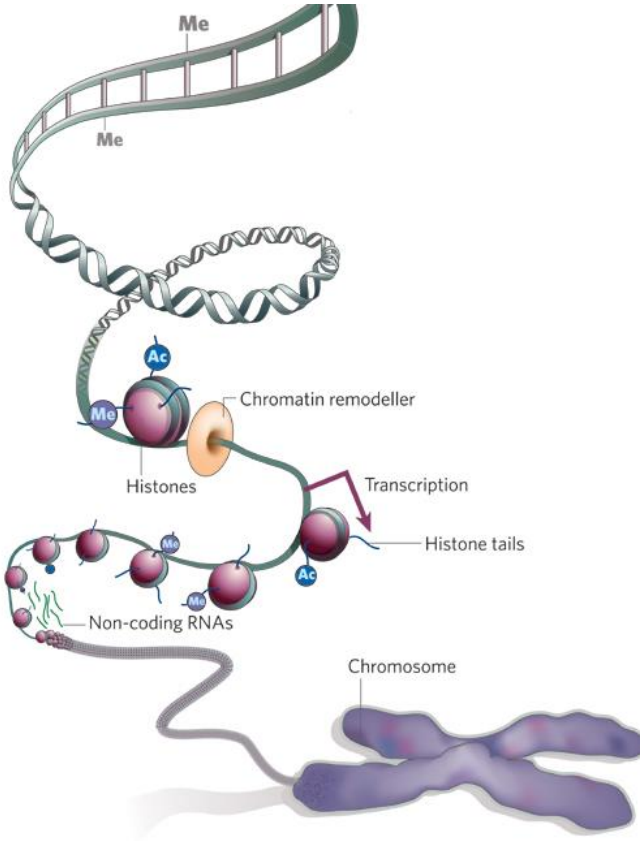


Figure 1.1 Mechanism of epigenetic imprinting. The genome is prone to DNA direct methylation, histone modifications; which include histone acetylation and methylation. Other chromatin remodelers also come into play. Additionally, noncoding RNAs play a major role in DNA targeting by silencing or different mechanisms. The figure was taken from (2).

1.1.2 Chromatin remodeling

As the name implies, the term chromatin remodeling includes all changes in chromatin architecture and nucleosome positioning imposed by chromatin remodeler proteins (see figure 1.2). These proteins include the covalent histone modifiers, i.e.; histone methyl-transferases (e.g., H3K27me3, H3K36me3, H3K4me1, H3K4me3, H3K9me3), histone-deacetylases (e.g., H3K27ac, H3K9ac), ATP-dependent activities of proteins that target the chromatin, and ubiquitination (7). Such changes in chromatin design make it more/less plausible for the different players in gene transcription, e.g.; DNA-dependent RNA polymerase plus the several transcription factors to reach the vicinity of the gene promoter and induce DNA transcription. Chromatin remodelers also play an indispensable role in reforming the genome for the several cell cycle events, such that the genome is either highly packed for the cell division stages, or is semi (heterochromatin) to full open (euchromatin) for possible DNA replication, transcription and translation events through the G1, G2

and S phases of cell cycle. A study made by Schwartz and colleagues (8) showed that chromatin organization also plays a crucial role in marking the exon-intron structure. This finding also unearthed the higher enrichment in nucleosome protein assembly, especially the histone marks in the exon context than on the level of the introns.

1.1.3 DNA Methylation

A 'transcription factor only' model for the effect on gene expression can hardly explain how the cell would switch or reverse the usage of a plethora of the available transcription factors once the cell changes its developmental state. Thus, there has to be a key memory player that introduces the stages to the different genetic and epigenetic players across the stages of development in what is called cell differentiation. This player is DNA methylation. There are several DNA methylation forms. Two of these modifications are crucial methylation forms that have been reported across species, namely N6-adenine and C5-cytosine methylation, which have been termed the fifth and sixth bases of DNA due to their great importance in regulation of cellular processes.

N6-methyladenine

This form of DNA methylation is more common in bacterial genomes where it protects bacteria against invading microorganisms by marking the methylated genome as 'self' (9). An alternative view was put forward taking into account evolution and the consequent battle of species. According to this view, an invading genome, which contains 6mA, is marked as foreign and is targeted for degradation (10). Accumulating evidence suggests that N6-adenosine methylation also plays a role in eukaryotic genome (11-13), but this putative role has sofar remained largely unknown. A global investigation on the occupancy of N6mA in the DNA of several prokaryotic and eukaryotic genomes was performed by Ratel et al (9). The comparison gave roughly similar ratios for this type of methylation in the so-far investigated species of both domains. A conclusion was that cytosine methylation might hinder the adenine methylation in the human genome. It is, however, currently known that this kind of methylation exists at least on the level of RNA in human (14) and is suggested to control gene expression on the level of the so-called epitranscriptome (15).

C5-methylcytosine

Unlike adenine DNA methylation, this form of DNA methylation shows large variations in its methylation ratio in the genomic DNA across species (16). C5-

cytosine methylation has an indispensable role in development, well seen through the reprogramming effect of the methyl groups by the erasure and re-establishment of methylation marks during the early stages of development (17), namely in the zygote and in primordial germ cells, progenitors of sperm or oocyte. This puts forward the question of the stability of several DNA sequences in the context of their propensity to integrate or remove the C5-methyl marks. Whether DNA in all sequence contexts accepts C5-cytosine methylation or reverses the binding by demethylating their sequences within differential levels is still of question. Additionally, the active/passive interplay for the effect of methylation/demethylation for specific DNA sequences is rather ambiguous. It is worth to mention that the removal of methyl groups during the erasure stage of reprogramming is proposed to be not complete, but is rather replaced by the oxidation event of the methyl mark (18). On the other hand, several *de novo* (19) and maintenance (20) methyltransferases serve to actively methylate DNA. However, for the exact sequence context this is a major question. DNA methylation at the cytosine level may serve to either suppress (21) or activate (22) gene expression.

1.2 Cross-talks between DNA methylation and chromatin remodelling: meDNA:Protein binding

DNA methylation and the various histone marks may work in a concerted manner in some scenarios. For example, site-specific *de novo* methyltransferases (DNMT3 proteins) are often involved with an initial binding of trans-acting elements that are able to recruit histone modifications (23). Figure 2 shows an example of an interplay between the several epigenetic marks and the mode of expression (on/off). Whereas DNA methylation may serve to recruit histone deacetylases (HDACs) in the on/off switch of gene expression for genes that were already acetylated by histone acetylases (HATs), histone methyltransferases and/or deacetylases are predicted to target the maintenance DNA methyltransferase DNMT1 for gene silencing (24).

Because of such interplay, transcription factors can be also recruited for further modularity of the transcriptional regulatory machinery.

Here, we will concentrate on two different forms of DNA methylation-binding proteins, namely those binding to the fifth/sixth letters of the DNA alphabet, N6-methyladenine and C5-methylcytosine binding proteins.

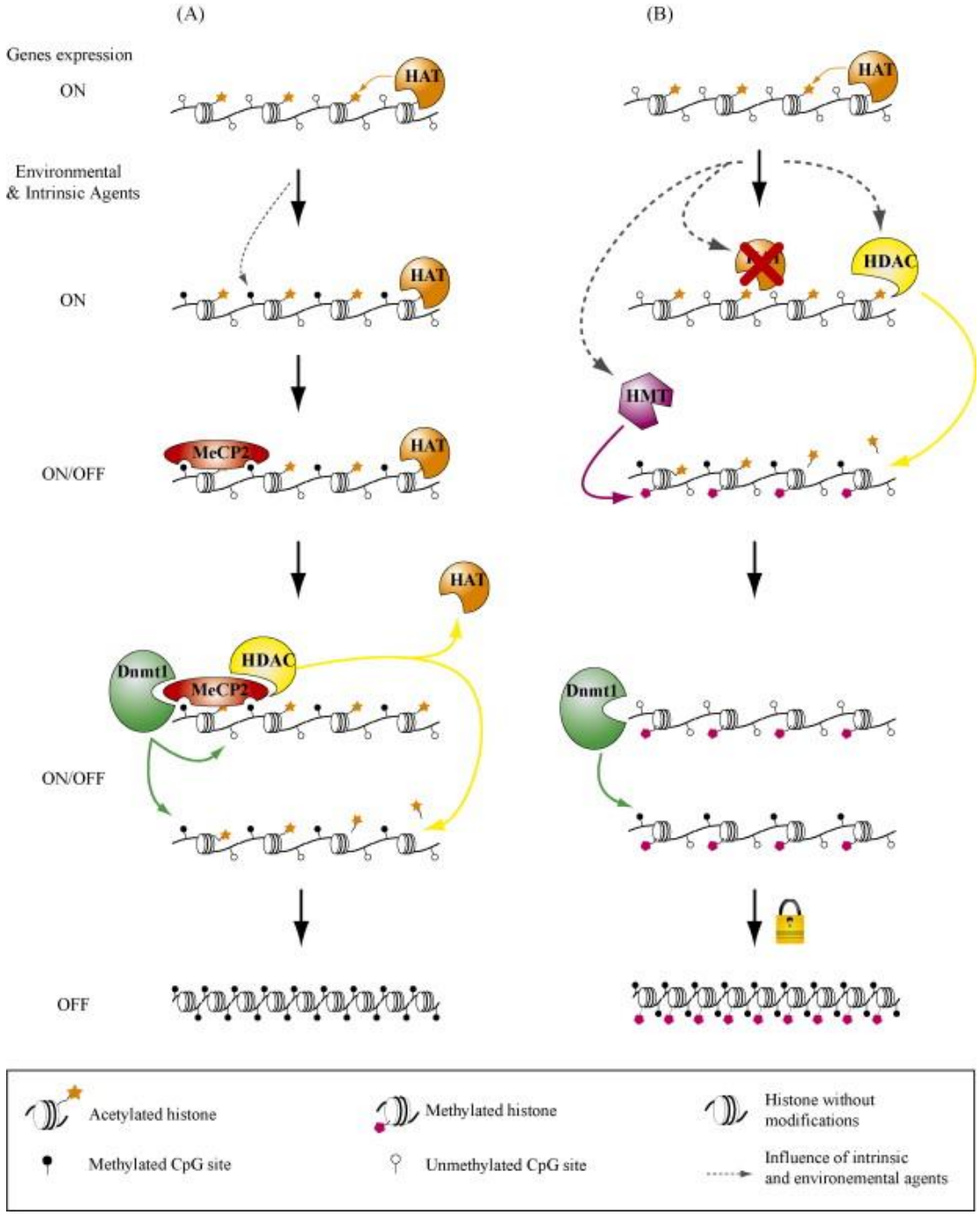


Figure 1.2 taken from (24). The epigenetic interplay between DNA methylation and histone modifications in gene silencing. (A) DNA methylation targets histone modifications for the on/off switch of expression, (B) histone modifications target methyl-specific proteins for gene silencing.

In a first sense, N6-methyladenine (m6A) and C5-methylcytosine are usually targeted by two classes of DNA methyltransferases: those associated with restriction-modification (R-M) systems (25), and solitary methyltransferases that do not have a restriction-enzyme counterpart, such as the bacterial N6-adenine methyltransferases

Dam and the C5-methylcytosine transferases Dcm (26). In alternative bacterial forms, where no Dam/Dcm is found, restriction methyltransferases do occur to mark self as non-methylated and target non-self, being methylated, to be cleaved by cellular enzymes, as for example the bacterial R.DpnI enyzme, that can be isolated from *Streptococcus pneumoniae* (10). This bacterium is gram positive and inhabits the upper respiratory system of human. Diseases in children have a very high association with this bacterium (27). One interesting fact about the *R.dpnI* system is that it is *per se* an R-M system. Thus, the study of the mechanism of action for this protein can serve well to understand the mechanism of binding to N6-methyladenine containing DNA. This can also help to understand the so far discovered proteins that bind the N6-methylated RNA in human and put forward suggestions for possible mechanisms in terms of DNA.

In another sense, nature has reasons to include more than one methylation form. This also indicates the possibility that 5mC-containing DNA and the consequent binding of methyl-recognition proteins perform differently than the 6mA-containing DNA, even though the sole difference with methylation in both contexts is the shared addition of the hydrophobic methyl group. For example, even though the methyl is in nature 'water hating' by itself, recognition of the C5-methylated DNA by the MeCP2 totally depends upon hydration in a at methyl-CpG (28).

1.3 DNA structure

DNA is one of the major macromolecules in the cell and is synthesized through the polymerization and the consequent condensation (dehydration) reaction of the monomer units, the nucleotides, composed of bases, a pentose sugar and the phosphate group. Phosphate and sugars make up the backbone of the DNA double helix. Monomers are connected through their phosphate-pentose backbone via a phosphodiester bond.

DNA exists typically in a double helical form. The tertiary structure of DNA can be in most cases classified into one of three structures, A-, B- and Z-DNA forms (see figure 1.3). A- and B- forms of DNA are right-handed in terms of twisting. Z-DNA shows left-handedness. The earliest detection of DNA structure was after the discovery of X-ray diffraction. Then, a race started to characterize the possible structure of DNA. Franklin and Gosling were thus the first ones to introduce the A-DNA form (29), which is the DNA structure to show the highest abundance in a 'dry environment'. This was followed by several other proposals for the DNA structure until Watson and Crick (30) discovered the iconic right-handed, anti-parallel, double-stranded B-DNA form,

which is the most predominant form of DNA in solution. They combined chemistry knowledge with insight from X-ray crystallography to make a prediction of the structure.

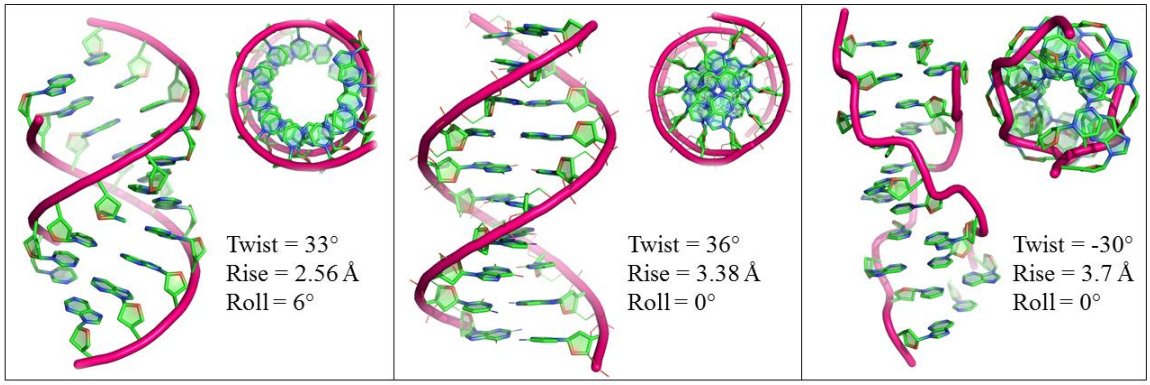


Figure 1.3 A- , B- and Z-DNA forms shown from left to right. Longitudinal and cross views of the DNA. In the right-bottom corner, some basepair step parameters are indicated. See main text for explanation.

The Z-DNA form becomes predominant in solution at high salt concentrations that can overcome the high energy barrier for the B-Z DNA transition (31). The salt concentration required for this is a function of salt type, DNA sequence content and methylation ratio. The CpG-rich DNA sequence content is more prone to this B-Z transition than AT runs. Additionally, C5-cytosine methylation induces lower barriers for the transition (31). Z-DNA form is often found in the enhancer regions of several genes where negative supercoiling is predicted to take place (32).

The geometry of a DNA double helix is defined by several parameters. We concentrate here on the parameters that we used in later chapters of this thesis. Torsion angles of the DNA backbone define one set of those terms. Figure 4 shows the list of the terms. Starting from the 5'-end (O5') of the base, naming of the torsion angles begins with α around the phosphorus atom of the base and its O5'. Further torsion angles are named consecutively by walking over the bonds of the base (see figure 1.4). ε and ζ are the two torsion angles that define the DNA:protein binding process. *X* torsion is used to define the anti and syn conformations of the nucleotide (see figure 1.5). These two conformations are associated with the B-DNA and the Z-DNA forms, respectively. As can be seen in figure 5, this torsion angle controls the relative orientation of the sugar and base rings. Base flipping, also a characteristic of the B-Z DNA transition, involves transitions in the *X* torsion angle. The anti conformation has a torsion in the 90° - 180° range, while the syn conformation has a *X* torsion in the $^{-}$90° - $^{+}$90° range.

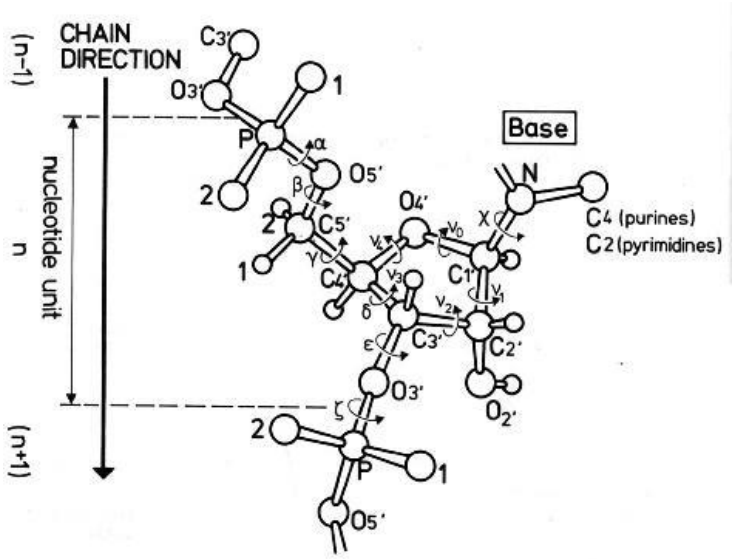


Figure 1.4 Torsion angles for the phosphate sugar backbone, taken from (33).

BI and BII states are two conformational states of DNA with altered positions of the phosphates (34). The shift of the BI/BII equilibrium also plays a key role for the specificity of protein-DNA binding (35). The BI/BII transition is a consequence of the change of the difference between the two torsion angles ε (C4'-C3'-03'-P) and ζ (C3'-03'-P-05') from ε-ζ of about -90° in the BI conformation to about +90° in the BII form. Whereas the BI conformation is rather symmetric in the phosphate position with respect to the minor and the major grooves, the BII conformation results in a shift of the phosphates to the minor groove. This causes a kink therein, and a consequent shift of the bases towards the major groove. A simple definition for the major/minor groove can be taken from the relative orientation of the bases towards the 5' end of the sequence (around the larger major groove), and towards the 3' end of the sequence (around the smaller minor groove). It is worthwhile to mention that the relative major/minor groove width is more defined in terms of the B-DNA than the Z-DNA.

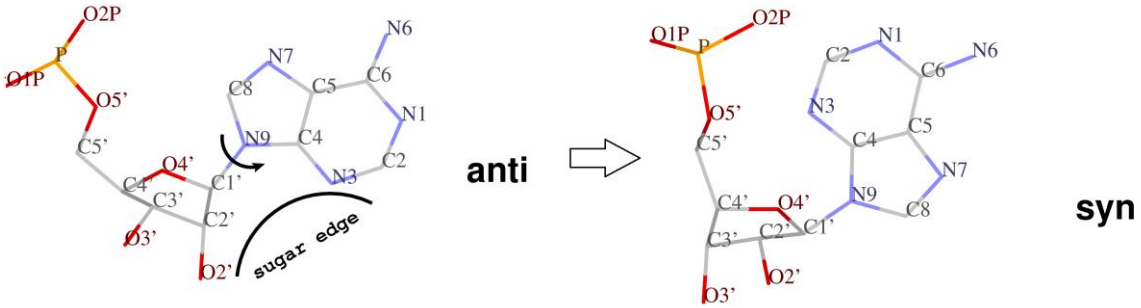


Figure 1.5 anti and syn conformations of the DNA, taken from (33)

One further set of geometric terms that can define the DNA structure is the base-pair parameters. These parameters include those sets that define the relative orientation of the bases in a single base pair, or the orientation of a pair of basepairs to each other (basepair step). Figure 1.6 illustrates these parameters. These parameters are very dependent on the sequence content of DNA, however, they can introduce a general idea of DNA properties related to the study under question.

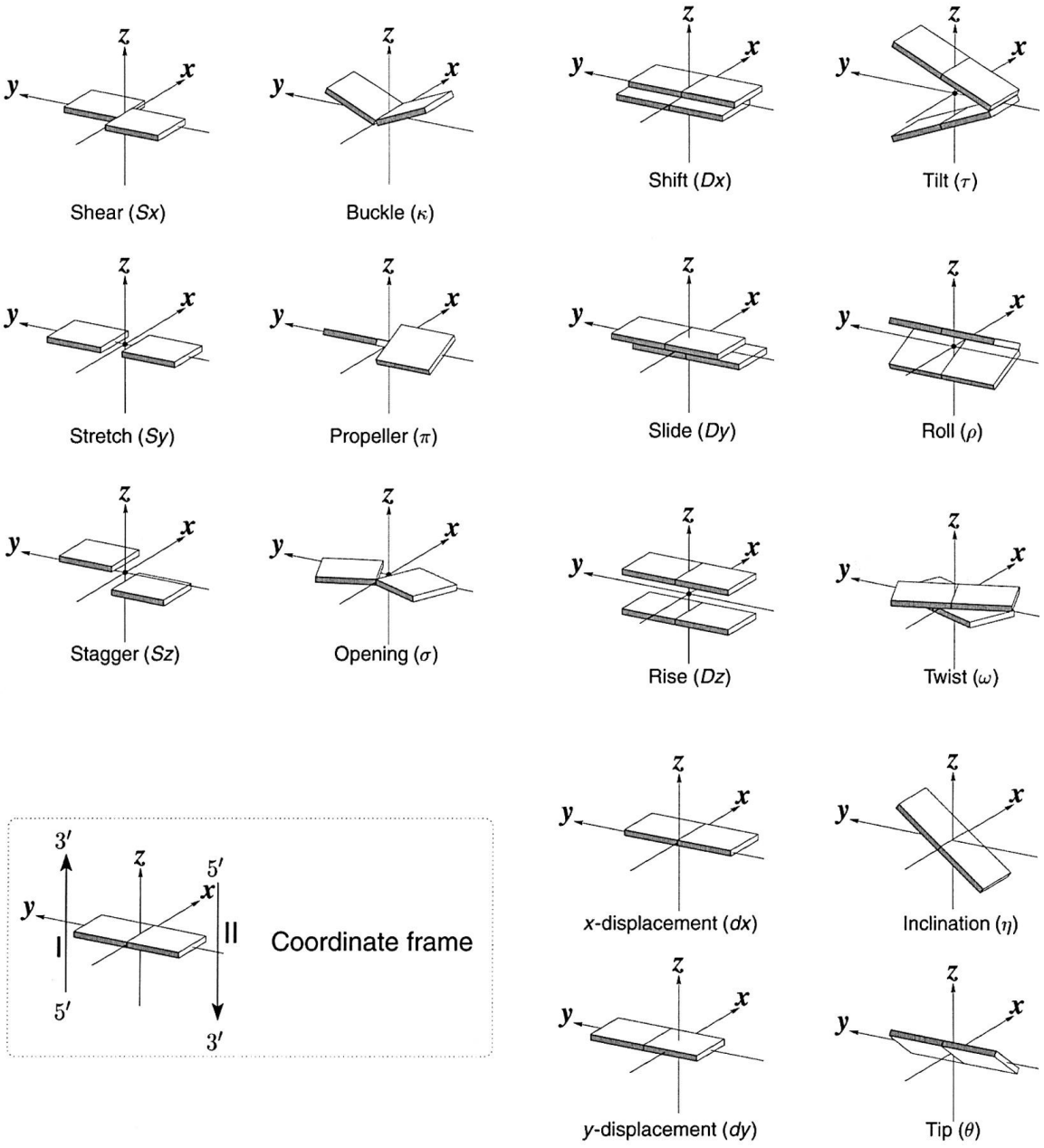


Figure 1.6 Base-pair parameters, taken from (36)

1.4 MD Simulations

The main aim of molecular dynamics simulations is to understand structural, dynamic, and energetic properties of molecules and of their assemblies. Simulations can thus serve to complement the experiment where information is missing and cannot be retrieved elsewhere. This provides the bridging of the micoscopic lengths and time scales with the macroscopic world of experiment. Computer simulations provide the possibility understanding of the studied properties of the system in atomic detail and on very short time scales, because the number of variables is minimized and the study is confined in the sense of the target variables. There are two main families of simulation techniques: molecular dynamics (MD) and Monte Carlo (MC) simulations. Throughout the simulations performed in this thesis, we utilized MD simulations. MD simulations provide a full view for the 3N-6 degrees of freedom for the total number of atoms being simulated as a function of time. Thus, the dynamics of the system can be understood. Furthermore, structural properties can be analyzed and predictive sampling of the configurational landscape can be met easily for the target degrees of freedom under the miscroscope.

The molecular dynamics simulations method was first introduced by Alder and Wainwright in 1959 (37), who studied the interactions of hard spheres. Next, Rahman carried out the first simulation using a realistic potential of liquid argon (38). The first realistic system to be simulated was liquid water by Rahman and Stillinger in 1974 (39). The first protein simulation appeared shortly after in 1977 by McCammon and colleagues, who presented a 6-ps long simulation of the bovine pancreatic trypsin inhibitor protein in vacuum (40). Nowadays, MD simulations are performed in many fields of science, including material science and biological sciences. In terms of the biological systems, one can see simulations for solvated proteins, protein-protein complexes, DNA, RNA, protein:DNA complexes, and several others. The thermodynamics of binding and the folding processes of several proteins has been also addressed. When extended to QM/MM approaches, simulations can also involve the study of reactions, e.g., enzymatic reactions.

1.5 Next generation sequencing and the human epigenome roadmap

With the exponential growth of digital data, fast and compact investigation methods are needed for a useful use of this data. To this aim, data integration and analysis emerged. Data integration combines data from different resources and unites them in a sense that the user can homogeneously use them under a unified view. Data

mining, on the other hand, aims at knowledge discovery in databases (KDD) (41). At an abstract level, voluminous data of low-level processing are further treated and mapped into more compact, abstract and useful forms.

Next generation sequencing (NGS) is a new generation of non-Sanger sequencing technologies. This form of DNA sequencing outperforms the Sanger sequencing method in read length and the number of reads in a single machine (42). In essence, next generation sequencing (NGS) provides an overplus of DNA reads that are sequenced at an ultrafast speed and on a low level of annotation for further analysis and use. To this aim, several databases have emerged for the robust analysis and visualization of the NGS data for an easier understanding by the end user. An example is Galaxy; a tool suite that manipulates FASTQ formats variants and filters data for use by the end user (43).

Applications of NGS include RNA-Seq (44), a technique for characterizing transcriptome data. This method of RNA sequencing outdates the genomic microarrays and the serial analysis of gene expression (SAGE) (45) by the fast sequence retrieval, the very low background noise, ability to distinguish different isoforms and allelic expression, the low amount of RNA needed and the relatively low cost equipment. In principle, a population of RNA is converted to a library of cDNA fragments, with adaptors attached to one (single-end) or both ends (pair-end). After that, all molecules are then sequenced. Read lengths can range typically from 30-400 bp, depending on the purpose.

Chromatin immunoprecipitation coupled to sequencing (ChIP-Seq) is one further application of NGS (46). This technique enables the possibility of genome-wide mapping of protein-DNA interactions. Examples for mapped proteins include transcription factors, core transcriptional machinery, histone modifications and chromatin remodellers, DNAse I hypersensitivity sites (47), and other proteins. DNAse I hypersensitivity gives the chance to check for an open chromatin structure. The ChIP-seq technique selects a set of proteins that binds DNA *in vivo*, and then targets them by specific antibodies.

Robust detection of C5-methylation in the genomes is one further NGS technique that has emerged in recent years. C5-methylation is known to be actively involved in human development (48). Several approaches can detect C5-methylation in genomic sequences. These include bisulfite sequencing (49), reduced representation bisulfite sequencing (RRBS) (50), and methylated DNA immunoprecipitation-sequencing (MeDIP-Seq) (51). Bisulfite sequencing and the RRBS techniques aim at finding the

methylated cytosines in the target sequences at a base-pair resolution. Bisulfite treatment of DNA converts unmethylated cytosine to uracil, leaving methylated cytosines intact (52). After that, the cytosine/thymine ratio is detected per base via several methods, including for example pyrosequencing (49). A slight difference betweenRRBS and bisulfite sequencing is the additional enzymatic digestion of the target DNA by a methylation-insensitive restriction enzyme (50). MeDIP-Seq is a ChIP-Seq method that uses specific antibodies to target the C5-methylated DNA sequences (51).

NGS data for expression analysis and the human epigenome atlas is nowadays integrated in the NIH Roadmap Epigenomics Mapping Consortium and consists of several early and late developmental stages (e.g.; stem cells and primary *ex vivo* tissues) (53). Data for this Consortium are integrated for further analysis by end users in the Human Epigenome Atlas (54).

1.6 Goal of the work

We aim in this work at obtaining a deeper mechanistic understanding of DNA methylation as an epigenetic mark. This variable is studied on different levels. On one end, and in its simplest picture, the relative thermodynamic stability of naturally occurring methylated/nonmethylated DNA sequences is studied. Additionally, structural investigation of MD results helped in characterizing water properties around methylated and non-methylated DNA, such that the combined effect of methylation/sequence specificity is grasped (Chapter 3). We next build on this understanding by further studying an example of specific protein:meDNA interactions. The target protein for our study is the C5-methyl-binding domain (MBD) of MeCP2 (methyl-CpG binding protein 2) (55). Deep structural investigation is held in a collaboration project with the experimental biophysics group of Prof. Albrecht Ott/UdS (Chapter 4). Next, we studied the binding thermodynamics for two methylated/nonmethylated protein:DNA complexes (Chapter 5). Finally, genome-wide analysis of NGS data for correlated expression/methylation and several epigenetic marks is studied across developmental stages. We aim to understand the impact of single epigenetic modifications in development (Chapter 6).

Chapter 2

Theory and Methods

In this chapter, we introduce the basics of statistical mechanics and algorithmic aspects related to MD simulations plus the statistical approaches that we have applied during our genome-wide analysis of NGS data.

2.1 MD Simulations

Molecular dynamics simulations in essence solve numerically the classical equation of motion formulated by Isaac Newton.

$$m_i \ddot{r}_i = f_i \qquad f_i = -\frac{d}{dr_i}\boldsymbol{u} \tag{2.1}$$

where the forces f_i acting on the atoms are derived from a potential energy function $\boldsymbol{u}(r^N)$, where $r^N = (r_1, r_2, r_3, \ldots, r_N)$ represents the complete set of 3N atomic coordinates.

2.1.1 Evaluation of the potentials

The Born-Oppenheimer approximation states that the movement of the quantum-mechanical particles that make up an atom, nuclei and electrons, can be treated separately due to their large mass imbalance. Thus, the movement of the much heavier atom nuclei can be well represented as classical point particles that follow the classical mechanics of Newtonian laws of motion (56). As a result, the electronic wave function depends only on the positions of the nuclei and not on their momenta (56). Force-field based molecular dynamics simulations follow this classic view. In MD simulations, and since we deal with the system at the microstate, the potential energy can be calculated by summing up the bonded and non-bonded interactions for the participating atoms. Force fields are thus presented to provide an approximation for the set of interactions. The most widely used molecular mechanical force fields are AMBER (57), CHARMM (58), GROMOS (59) and OPLSAA (60).

2.1.1.1 Non-bonded Interactions

The part of the potential energy $\boldsymbol{u}_{\text{non-bonded}}$ represents non-bonded interactions between atoms. This can be split into 1-body, 2-body, 3-body, etc terms:

$$\boldsymbol{u}_{\text{non-bonded}}(\mathrm{r^N}) = \sum_i u(r_i) + \sum_i \sum_{j>i} (r_i, r_j) + \ldots \tag{2.2}$$

Where $\boldsymbol{u}(\mathrm{r})$ is an externally applied potential field or the effect of the container walls. This term is usually dropped for periodic simulations and is not relevant for the work described in this thesis.

Two potentials constitute the components of the non-bonded form of the potential energy. The first one is the Lennard Jones potential. The most-commonly used form of this potential is:

$$v^{LJ}(r) = 4\varepsilon \left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} \right]. \tag{2.3}$$

Where $\left(\frac{\sigma}{r}\right)^{12}$ is the long-ranged repulsive term and $\left(\frac{\sigma}{r}\right)^{6}$ is the short-ranged attractive term. The two parameters σ and ε represent the diameter and the well depth, respectively.

The second non-bonded potential is the electrostatic Coulomb potential that accounts for interactions between charged particles:

$$v^{Coulomb}(r) = \frac{Q_1 Q_2}{4\pi\epsilon_0 r}, \tag{2.4}$$

where Q_1 and Q_2 are the charges, r the distance between them, and ϵ_0 is the dielectric permittivity of the free space.

2.1.1.2 Bonded Potentials

An electronic structure (also termed quantum-chemistry) calculation of a single molecule can be used to estimate the electron density throughout the molecule. However, such calculations require a high numerical effort that is prohibitive for dynamic simulations of solvated biomolecules. In the molecular mechanics picture of molecules, that is the basis for molecular force-fields, bonded interactions in a molecule are accounted for via the inclusion of several terms, namely the bonds, bend

angles and the torsion angles. Harmonic potentials are estimated per each term via the following equation:

$$\mathcal{U}_{intramolecular} = \frac{1}{2}\sum_{bonds} k_{ij}^{r}\left(r_{ij} - r_{eq}\right)^2 \qquad (2.5a)$$

$$+ \frac{1}{2}\sum_{\substack{bend\\angles}} k_{ijk}^{\theta}\left(\theta_{ijk} - \theta_{eq}\right)^2 \qquad (2.5b)$$

$$+ \frac{1}{2}\sum_{\substack{torsion\\angles}}\sum_{m} k_{ijkl}^{\emptyset,m}(1 + \cos(m\emptyset_{ijkl} - \gamma_m)) \qquad (2.5c)$$

The bonds are defined in terms of the distance between the adjacent pair of atoms, or the vector length $\vec{r}_{ij}$, with the equilibrium displacement (r_{eq}) taken under a harmonic assumption. The bend angle θ_{ijk} is defined via the angle between successive bond vectors $\overrightarrow{r_{ij}}$, $\overrightarrow{r_{jk}}$, and therefore includes three atomic coordinates. The torsion angle $\emptyset_{ijkl}$ is defined by three connected bonds and the normal to the plain defined by each pair of bonds. According to the force field under use, equilibrium bonds, angles and the various force constants are defined.

2.1.2 The MD Algorithms

Together with the potential energy introduced in the previous section, the Hamiltonian of the system consists additionally of the kinetic energy. Thus, for a system with coordinates $r^N = (r_1, r_2, \dots, r_N)$, and a potential energy $\mathcal{U}(r^N)$, the atomic momenta are defined as $p^N = (p_1, p_2, \dots, p_N)$. The kinetic energy can thus be expressed as a function of the momenta

$$\mathrm{K}(p^N) = \sum_{i=1}^{N} |p_i|^2/2m_i . \qquad (2.6)$$

Then the Hamiltonian is defined as the sum of the kinetic and the potential terms $H = \mathrm{K} + \boldsymbol{\mathcal{U}}$. Also, the classical equation of motion can be defined in terms of coupled ordinary differential equations as:

$$p = mv \dashrightarrow \dot{r}_i = p_i/m_i \ and \ \dot{p}_i = f_i \qquad (2.7)$$

For a rapid sampling of the phase space, simulation algorithms tend to be of low order. This allows a large enough time step without violating the conservation of the total energy.

2.1.2.1 The Verlet Algorithm

The Verlet algorithm is a numerical method used to integrate Newton's equation of motion (61). Several versions of the Verlet algorithm have been set. In principle, the basic Störmer-Verlet can be integrated with or without direct velocity calculations in the iteration over the timesteps. The more commonly used variants are the Velocity Verlet algorithm (62) and the related leapfrog algorithm (63). Whereas positions and velocities are updated simultaneously in the Velocity Verlet algorithm (61), they are updated at interleaved time points in the leapfrog algorithm (63).

The standard implementation for the Velocity Verlet algorithm is:

$$p\left(t+\frac{1}{2}\Delta t\right)=\vec{p}_i(t)+\frac{1}{2}\Delta t f_i(t) \qquad (2.8a)$$

$$\vec{r}_i(t+\Delta t)=\vec{r}_i(t)+\dot{r}_i(t)\Delta t+\frac{1}{2}\ddot{r}_i(t)\Delta t^2 \qquad (2.8b)$$

$$\vec{p}_i(t+\Delta t)=\vec{p}_i\left(t+\frac{1}{2}\Delta t\right)+\frac{1}{2}\Delta t f_i(t+\Delta t) \qquad (2.8c)$$

Where step 3 is dependent on step 2 in the sense that a force evaluation is carried out to calculate the new momentum. The algorithm iterates until the given number of steps is executed.

2.1.2.2 Constraints

Representing intramolecular bond lengths by simple harmonic terms in the potential function can have a large, undesirable effect on the MD calculations, especially when changes in bond lengths may include vibrational frequencies that should arguably be correctly treated in a quantum mechnical way rather than by classical mechanics. For sake of simplicity, and to allow for a longer time step of 1 or 2 fs rather than 0.5 fs, constraints are applied to the bond lengths in addition to the above Verlet algorithm. It has been shown that these constraints hardly affect the dynamics of the system in the other degrees of freedom since the bond vibrations have much higher frequencies than the other covalent terms. . Taking the function of the constraint and applying

Lagrangian multipliers implements the applied constraint $\mathcal{X}$. Written in a simple form, the equation for the constraint is applied with a free term, such that the constraint is constant:

$$\mathcal{X}(r_1, r_2) = (r_1 - r_2).(r_1 - r_2) - b^2 = 0 \qquad (2.9a)$$

The partial differential equation in terms of the two replacement variables is applied such that the margin is minimized:

$$\dot{\mathcal{X}}(r_1, r_2) = 2(v_1 - v_2).(r_1 - r_2) = 0 \qquad (2.9b)$$

The undetermined multiplier can thus be added to the equation of force calculations such that:

$$m_i \ddot{r}_i = f_i + \Lambda g_i \qquad (2.10a)$$

Solving the Lagrangian multipliers ends up in:

$$g_1 = -\frac{d\mathcal{X}}{dr_1} = 2(r_1 - r_2) \qquad g_2 = -\frac{d\mathcal{X}}{dr_2} = -2(r_1 - r_2) \qquad (2.10b)$$

However, what is needed is not an exact solution, but rather a solution that is guaranteed to be satisfied by the end of each timestep. Two different constraints are used in the GROMACS software, namely LINCS/P-LINCS that allows to apply holonomic constraints such that a larger timestep can be integrated (64), and SHAKE that applies geometric constraints to bonds (65). Throughout our simulations, we used the LINCS/P-LINCS algorithms.

2.1.2.3 Periodic Boundary Conditions

To account for possible surface effects in the system, periodic boundaries are applied. Periodic images of each atom along the Cartesian axes are virtually coupled to the atom in the simulation box, and a minimum image convection is to be applied. As the longest allowed cutoff for the potentials must be shorter than half of the length of each box dimension, no atom can interact with any of its images. Additionally, periodic boudaries are assumed such that if an atom leaves the box from one end, its periodic image replaces it and enters from the other end. Also, for the calculations of next step, a neighbour list is established for each atom. At each iteration, the set of neighbour atoms that have the minimum periodic distance is calculated. However,

since the number of atoms can be exceedingly large, an r_{cutoff} is established such that the closest atoms within a cutoff distance are retrieved (66).

2.1.2.4 Ensembles for the MD simulations

MD simulations are advantageous over the alternative technique of Monte Carlo simulations in the sense that both static and dynamic quantities can be studied. However, early MD simulations suffered the problem that the conditions did not mimic those in experiment. In 1981, Nosé proposed a method of MD simulations that ensures that the generated configurations belong either to the canonical (NVT; constant temperature, constant volume) ensemble or to the constant temperature constant pressure (NPT) ensemble. The physical system of interest consists of N particles, but is connected to an external heat reservoir. This allows the total energy of the physical system to fluctuate (67). Before that, several attempts had been made to recover the conditions in experiments, as for example, the method proposed by Andersen in 1977 to account for make the pressure constant by allowing the volume to fluctuate (68). Parrinello and Rahman have later extended the method to allow changes in the shape of the MD cell (69). To keep the kinetic energy of the system constant, Hoover and Ladd proposed a method in which an external velocity term is added to the kinetic energy (70).

The method proposed by Nose introduces an additional degree of freedom to the system, such that the total energy of the system is allowed to fluctuate. The added degree of freedom to the potential energy realizes the condition that the system follows the canonical ensemble by scaling the system's velocities (67).

2.1.3 Free Energy Perturbation

Finding free energy differences between two different states is of major importance for a variety of biologically relevant question, namely for drug design, and several others. When the intersection in Hamiltonians of two states has a very low probability, very poor sampling and insufficient results for free energy arise. Several approaches have been proposed to calculate free energy differences for alchemical transitions between two different states, where the Hamiltonians can be divided into artificial substates with a high probability for the Hamiltonian to cover the closest defined ones.

2.1.3.1 Approaches for Free Energy Perturbation

The statistical mechanics perturbation theory was first developed by Zwanzig and has been applied by several workers to dense fluids (71). The Hamiltonian of a system is separated into two parts:

$$H = H_0 + H_1 \qquad (2.11a)$$

where H_0 is the reference state and H_1 is a perturbation from the reference state. When the configurational partition functions of the reference and the total states are used, the perturbation free energy is:

$$G - G_0 = G_1 = -RT \ln\langle \exp(-H_1/RT)\rangle_0 \qquad (2.11b)$$

where $\langle \quad \rangle_0$ is the ensemble or time average over the reference system.

Free energy differences of solvation, relative changes in binding, relative stabilities of several protein-DNA complexes, etc, can be calculated via the free energy perturbation method. The crucial step is to define a Hamiltonian for the solutes in states A and B. These Hamiltonians are linked by a coupling parameter λ (72):

$$H_\lambda = \lambda H_A + (1-\lambda)H_B \qquad 0 \le \lambda \le 1 \qquad (2.12)$$

H_A is the Hamiltonian for state A, and H_B is that for B. At λ=1, $H_\lambda = H_A$. However, at λ=0, $H_\lambda = H_B$. At intermediate values of λ, the solute is hypothetically a mixed environment of A and B. In practice, this coupling ensures a smooth conversion between the states A and B . The single perturbed Hamiltonian can be further subdivided into several Hamiltonians. This can be accomplished by further considering each perturbed Hamiltonian as its own reference state, and the target state to be the next perturbed Hamiltonian.

Several algorithms have been proposed to connect the results from the intermediate stages for different λ values of the perturbed Hamiltonian. In 'multicanonical' thermodynamic integration (73), or simply thermodynamic integration (TI), the equilibrium ensemble average of the derivative of the Hamiltonian H with respect to λ is computed at a number of points along the path from state A to state B. The ensemble is then integrated numerically to obtain the free energy difference. Whereas one typically requires that the system is sampled in statistical equilibrium at every intermediate λ value, the Jarzynski equality can also be applied to derive the

free energy difference from perturbations applied to systems that are sampled away from equilibrium at intermediate points. This equality assumes that the nonequilibrium work W takes a number of systems in thermal equilibrium from an initial Hamiltonian to a different final Hamiltonian, which can be averaged over the entire Boltzmann-weighted initial ensemble. Then, the equilibrium work, here the free energy, between the two states is given by $\Delta F = -\beta^{-1}\ln\langle\exp(-\beta W)\rangle$ (74).

An alternative method for free energy perturbation is the Bennett's acceptance ratio method (75). Taking the phase space as a function of its internal variables, e.g., momenta and position in space, the Hamiltonian $H_\lambda(z)$ gives the total energy of the system at a particular point of the variable z. Given the partition function Z and the free energy function F as a function of z, and thus of β, and by performing a Metropolis move, the probability of transition between two neighbouring states can thus be used to calculate the free energy difference of the two states. Briefly, Bennett has shown that the free energy for the forward and reverse transitions satisfy:

$$\sum_{i=1}^{n_P} \frac{1}{1+\exp(-\beta(M+W_i-\Delta F))} - \sum_{j=1}^{n_R} \frac{1}{1+\exp\left(-\beta\left(M+W_j-\Delta F\right)\right)} = 0 \qquad (2.13a)$$

where $M = kTln\, n_f/n_r.\, n_f$ and n_r are the number of values from the forward and reverse distributions of work, respectively. The free energy among free energy estimates generally essentially satisfies the detailed balance condition:

$$\exp(-\beta\Delta F) = \frac{\langle f(W)\rangle_F}{\langle f(-W)\exp(-\beta W)\rangle_R}, \qquad (2.13b)$$

where $f(W)$ is an arbitrary function and the averages are over the two end states. Thus the Bennett acceptance ratio can be considered as the maximum likelihood estimator for the free energy, thus:

$$ln\left(\frac{P_F(W)}{P_R(-W)}\right) = \beta(W-\Delta F), \qquad (2.13c)$$

where $P_F(W)$ and $P_R(-W)$ are probability distributions for the work of nonequilibrium processes in opposite direction, arbitrarily named F and R for the forward and reverse transitions, respectively.

2.1.3.2 Soft-core potentials

Alchemical free energy perturbation is a method for free energy calculations where the system is alchemically perturbed such that the electrostatic potentials and/or the Lennard Jones potentials are introduced or annihilated. This can introduce singularities when possible artificial undesirable interactions of atoms are introduced. In practice, one observes the most severe problems when the interactions of an annihilated group of atoms become so small that neighboring molecules, e.g. water molecules, can approach them very closely. At such short atom-atom distances, small displacements during an MD time step can suddenly generate an exceedingly large repulsion due to the remaining repulsive term of the LJ interaction. Using a soft-core potential yields a convenient solution to this problem. An example is the introduction of the soft-core terms α, σ and p at states A and B, such that $r_A = (r_{ij}^6 + \alpha\sigma_A^6\lambda^p)^{1/6}$ and $r_B = (r_{ij}^6 + \alpha\sigma_B^6(1-\lambda)^p)^{1/6}$, respectively. Then, at state A, the new potential then takes the form (76):

$$V_{ij}(r_{ij}) = \frac{q_i q_j}{4\pi\varepsilon_0\varepsilon_r r_A} + \left(\frac{C_{ij}^{(12)}}{r_A^{12}} - \frac{C_{ij}^{(6)}}{r_A^6}\right) \qquad (2.14)$$

Figure 2.1 represents an example (77). In the figure, the –NH2 group is alchemically grown. At some point during the growth, the radii for distances between a water molecule and the growing –NH2 surface decrease until the water molecule collides with the target molecule and singularities arise. A further singularity may arise when adding/removing the electrostatic potential and the LJ term simultaneously. As a result, at very distances, the repulsive term $\frac{C_{ij}^{(12)}}{r_A^{12}}$of the equation 14 exceedingly grows over the attractive term$\frac{C_{ij}^{(6)}}{r_A^6}$, and a sudden uncontrollable jump in the potential takes place as a cummulative effect of the additional electrostatic term $\frac{q_i q_j}{4\pi\varepsilon_0\varepsilon_r r_A}$. At very close distances, however, the repulsive term can be no longer strong to hold atoms against each other, and a possible 'trap' for energy minimum may arise. Consequently, the separation of the electrostatic and LJ terms is the solution for such an effect and it can be compulsory. As a result, soft-core potentials are only applied to the LJ term, where wells of minima can be reached.There appears no need to apply them to the continuously growing electrostatic potential from the initial to the final states.

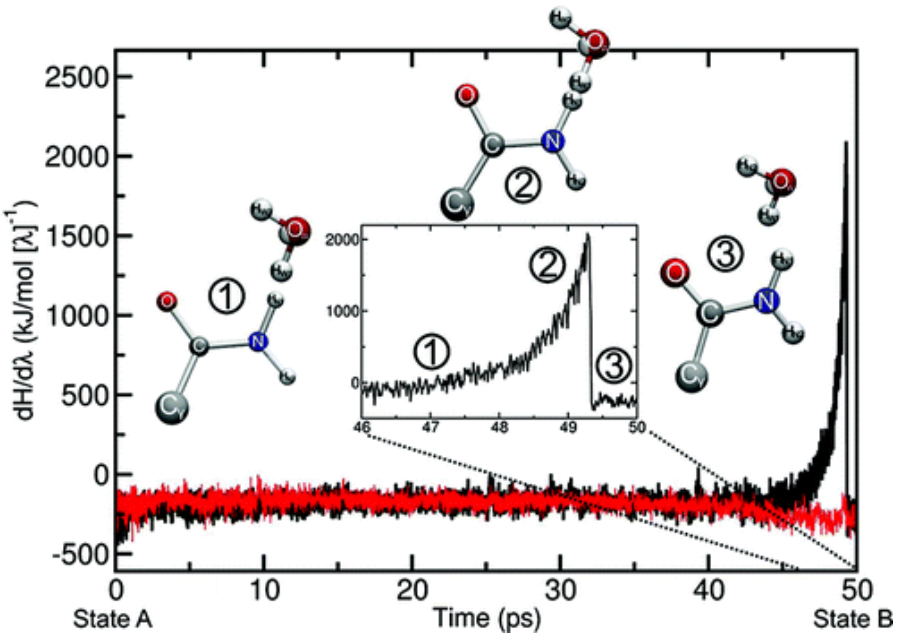


Figure 2.1 Free energy perturbation with the growth of the –NH2 from state A to state B, taken from (77). Whereas the red line introduces no singularites, the black line does introduce, due to the different initial structures. Water molecule approaches the molecule at the given peak (2). See text for further explanation

2.2 Statistical Thermodynamics of binding and standard states

After the introduction of the Particle-Mesh Ewald (PME) summation for computing long-range interactions in periodic systems, calculation of the free energy term got more possible in MD simulations. PME is an N.log(N) method for evaluating the long-range interactions via performing a Fourier transform (78).

When two biological species bind, saturation occurs and the equilibrium of binding is reached such that:

$$A + B \leftrightharpoons AB \tag{2.15}$$

Since we have biological systems, such transitions take place in an aqueous electrolyte. At equilibrium (79):

$$\mu_{sol,A} + \mu_{sol,B} = \mu_{sol,AB} \tag{2.16}$$

where each $\mu_{sol,i}$ is the chemical potential of species i = A,B or AB in solution. For species i in solution, the chemical potential $\mu_{sol,i}$ is described by:

$$\mu_{sol,i} = \mu^{o}_{sol,i} + RTln\frac{\gamma_i C_i}{C^o} \quad (2.17)$$

Where $\mu^{o}_{sol,i}$ C_i and γ_i are the standard chemical potential, the concentration of species *i*, and the activity coefficient of species *i*, respectively.

$$\Delta G^{o}_{AB} = \mu^{o}_{sol,AB} - \mu^{o}_{sol,A} - \mu^{o}_{sol,B} \quad (2.18a)$$

$$= -RTln\left(\frac{\gamma_{AB}}{\gamma_A\gamma_B}.\frac{C^o C_{AB}}{C_A C_B}\right)_{eq} = -RTlnK_{AB} \quad (2.18b)$$

The Gibbs free energy constitutes of the contribution of the enthalpic and the entropic components of the system, H and S, respectively in an isothermic isobaric system; given an absolute temperature T:

$$G = H - T.S \quad (2.19a)$$

$$\Delta G = \Delta H - T.\Delta S \quad (2.19b)$$

2.2.1 Win some, lose some: enthalpy-entropy compensation

Favorable free energy changes can be understood in terms of favorable enthalpic contributions, an increased entropy of the system, or both. When one of the two terms contributes unfavorably to the change in free energy, the other one may compensate with a favorable effect of comparative absolute magnitude. This is the enthalpic-entropic compensation. However, if the two terms do contribute to the free energy gain in a favorable fashion, their individual contributions will seem to be mild (80). Where large amount of strain happens to the structure of the target molecule, enthalpic contribution is highly costly (not favorable) and the entropic contribution is highly favorable. In such a case, this entropic enthalpic compensation can be viewed not as a cause for binding, but rather as the ultimate purpose (80).

Enthalpic entropic compensation can be also understood in terms of environmental effects to the system under study, as for example the temperature effect. One example is the relationship of entropy and enthalpy for a melting solid. Whereas the free energy of melting should equal zero, the final entropic contribution to fusion should be the enthalpic contribution to the fusion divided by the melting temperature $\Delta S_f = \Delta H_f / T_f$, hence the compensation follows (81).

2.2.2 Calculations of Configurational Entropy

Experimental evaluation of entropy by calorimetric studies makes it conceivable to achieve only the total change in entropy of the whole system. However, individual contribution of the several components cannot be accessed. This is one of the strengths of molecular dynamics simulations, where the individual contribution of one entity to the whole free energy change can be extracted.

One approach for entropy calculations is the Schlitter approach (82), where the contribution of the individual entity can be extracted from the covariance matrix of atomic positional coordinates throughout the simulation time. Given 3N different x-, y-, and z- coordinates of all atoms in a system, and introducing a time-series of atomic positions along the simulation time, the covariance matrix is the $\sum_{3N\times 3N}$. matrix of the covariance of all the 3N variables along the simulation trajectory. A formulation of the resulting matrix is:

$$\sum = \begin{bmatrix} E[(X_1-\mu_1)(X_1-\mu_1)] & \cdots & E[(X_1-\mu_1)(X_{3n}-\mu_{3n})] \\ \vdots & \ddots & \vdots \\ E[(X_{3n}-\mu_{3n})(X_1-\mu_1)] & \cdots & E[(X_{3n}-\mu_{3n})(X_{3n}-\mu_{3n})] \end{bmatrix} \quad (2.20)$$

The entropy of the system can be most easily expressed in terms of probabilities, where:

$$S = -k\sum p_n ln p_n \quad (2.21)$$

The average and variance for the coordinates are then taken and the maximum probability of entropy is estimated. Lagrange parameters are used to get the maximum entropy and a condition is added such that the system satisfies the harmonic oscillator with the Hamiltonian. Finally the entropy of the system is derived:

$$S < S' = 0.5k\ ln\ det[1 + (kTe^2/\hbar^2)M\sigma] \quad (2.22)$$

where σ is the covariance matrix of concerted atomic motions, $\boldsymbol{M}$ is the mass matrix that contains masses at the diagonal and is zero elsewhere (82).

Another approach for entropy calculations is the normal mode or quasiharmonic approximation, which is based on the assumption that the harmonic approximation

holds for energy calculations. The entropy is thus calculated in terms of the vibrational degrees of freedom. This form thus adds the wave vector and and frequency calculations associated with the wave vector (83). Thus PCA analysis fits in effective harmonic potentials on the observed coordinate covariance, smoothing out any anharmonicity. Thus eigenvalues represent a series of uncorrelated harmonic oscillators.

2.3 Bioinformatics: next-generation sequencing

2.3.1 Data processing

Heraclitus said, 'you cannot step twice into the same river'. This very well applies to NGS data. Even though performing several technical replicates is usually a 'must' to reproduce the results, there are still variations that can still be seen across replicates. For example, identical twin mice living in the same conditions can never be exact copies in terms of gene expression, nor is the one mouse the same when studied at different time intervals. While we are not interested to suppress the effect of the biological variations, an assessment of the noise due to technical variations should be considered. In this respect, several algorithms have been introduced to account for technical variation (dispersion) and background noise.

Variance in data is usually a sum of two components, one is the sample-to-sample variation due to biological differences plus the uncertainty in concentrations that can contaminate the real count reads. This variation accounts usually for 20% of differences across replicates. The other form of variation, known as shot noise or Poisson noise, is the one that dominates in differential expression inference (84).

Regression analysis can be used to model the relationships between a dependent and an independent variable. Several approaches are proposed for regression of two different data sets, according to the distributions that the data points follow. Regression can altogether be used to account for noise in the single data set and introduce possible dissimilarities between two data sets. For processing of next generation sequencing reads, we now discuss logistic regression, MA regression, Poisson regression, and negative binomial regression.

In Poisson regression, data points in two different data sets are modeled via the Poisson distribution, where the probability of observing some count, y of an outcome Y is given by:

$$\Pr(Y = y) = \frac{\lambda^y e^{-y}}{y!} \tag{2.23}$$

During the modelling process, when the expected value is equal to the mean of the data points, this is a good indication that data points do indeed follow the Poisson distribution. We wish then to check whether there exists a difference in the means of the two samples. Data points are thus fitted to reduce the noise to the corresponding Poisson distribution, and a simple log-linear model can be used to parametrize the relationship between samples. Given two different samples X and Y, the relationship can be modelled such that:

$$\log\big(\lambda(X)\big) = a + bX \tag{2.24}$$

This means that sample X is modelled via the Poisson process (e^a), and sample Y is modelled via the Poisson (e^{a+b}). Thus the incident rate ratio is (e^b).

Even though the Poisson distribution provides strong assumptions, but the standard error using the Poisson model can be exceedingly small and biased. Additionally, subjects are assumed to have the same rate of outcome. So an alternative regression method is the use of the negative binomial regression. The negative binomial probability distribution of Y is thus:

$$P(Y = y) = \left(\frac{\alpha}{\alpha + \lambda}\right)^y \frac{\Gamma(\alpha + y)}{\Gamma(y + 1)\Gamma(\alpha)} \left(\frac{\lambda}{\alpha + \lambda}\right)^y, \tag{2.25}$$

where Γ is the gamma function. Thus the mean for the negative binomial distribution is λ, and the variance is $\lambda + \lambda^2/\alpha$, α is the dispersion parameter. In expression data, changes can also be seen between samples due to erroneous concentration effect. This, however, can also be accounted for during data analysis by adding one more factor, the size factor s to the modelling of the binomial distribution. According to Anders and colleagues, the variance of count data can thus be modelled via (84):

$$v = s\lambda + s^2\lambda^2/\alpha, \tag{2.26}$$

Logistic regression is based on the probability outcome of binary classifiers. This classifier can be used to model methylation ratio in Bisulfite-Seq and RRBS data, where information is based on the number of methylated vs. unmethylated Cs at a given region. The methylation proportion P_i can be modelled for samples i=1,…,n

(where n is the number of biological samples). The logistic regression is then defined as (85):

$$\log(P_i/(1-P_i)) = B_0 + B_1 * T_i \quad (2.27)$$

T_i can be modelled as the treatment indicator for sample i, with T_i=0 for control and T_i=1 for the control group. B_0 is the log odds of the control group and B_1 is the log odds ratio of the treatment group. According to the null hypothesis H_0, $B_1 = 0$. Thus rejecting the null hypothesis would imply a differential methylation ratio between treatment and control groups.

One further method for data regression is MA-normalization proposed by Shao and colleagues (86). This method especially targets ChIP-Seq data and is based on the assumption that if a chromatin-associated protein has a large number of peaks shared in two conditions; similar global intensities of binding should show up across samples. Based on this assumption, the $\log_2$ read density between two samples (M) is calculated and plotted against the average $\log_2$ read density (A) for all peaks. After that, robust linear regression is applied to fit the global dependence between the M-A values of common peaks (86).

2.3.2 Data analysis

Cluster analysis is used to group data according to the full representation of pair-wise gene similarity. Datasets can thus be partitioned into different clusters in a recursive manner. Two general forms of hierarchical clustering can be used. In the first "agglomerative" type, the clustering algorithm starts with the single data points as the starting clusters and combines them iteratively. This can be achieved by finding the most similar clusters in each step and merging them, until all data points are treated. The other "decisive" style considers initially the whole dataset as one cluster, and then iteratively splits the whole cluster into two smaller sub-clusters until we end in a singleton per cluster (87).

Several clustering methods build on the agglomerative method. The resulting cluster shape follows the distance criteria. Normally, the simple Eucledian distance is the method of choice, and it can be used to extract the spherical clusters, where the Eucledian distance for two points a and b in a multidimensional space can be defined as:

$$\|a-b\| = \sqrt{\sum_i (a_i - b_i)^2} \quad (2.28)$$

Mahalanobis distance can be used in the detection of ellipsoidal clusters, where:

$$\|a - b\| = \sqrt{(a - b)S^{-1}(a - b)} \quad (2.29)$$

where a and b are the observations and S^{-1} is the covariance matrix.

Several versions of agglomerative methods exist that affect the results in the final cluster. These include single-linkage, complete-linkage and average-linkage clustering. In *single-linkage* clustering (the *connectedness* or *minimum* method), the distance between any two clusters is defined as the shortest distance from any member of one cluster to any member of the other cluster. In c*omplete-linkage* clustering (also called the *diameter* or *maximum* method), in contrast to single-linkage, the distance between one cluster and another is equal to the largest distance from any member of one cluster to any member of the other cluster. In *average-linkage* clustering, we consider the distance between one cluster and another to be equal to the average distance from any member of one cluster to any member of the other cluster (88).

We also analyzed the Gene Ontology for gene sets to account for biological processes that can be significant in terms of our analysis. The Gene Ontology is a structured vocabulary where functional associations are annotated to individual genes. To this aim, we used the GOSim R-package (89). Genes of interest are analyzed, and the GO terms are retrieved. This can be done by following the strategy of 'disjunctive common ancestors', which was followed by Couto et al (90). In this algorithm, taking into account that the GO total set is represented as a DAG, consider the ontologies a_1 and a_2. These represent the disjunctive ancestors of c if there is a path from a_1 to c not passing through a_2 and a path from a_2 to c not passing through a_1:

$$Dis\ j\ Anc(c) = \{(a_1, a_2)|$$

$$(\exists p: (p \in Paths(a_1, c)) \wedge (a_2 \notin p)) \wedge$$

$$(\exists p: (p \in Paths(a_2, c)) \wedge (a_1 \notin p))\}. \quad (2.30)$$

Now given the two GO terms c_1 and c_2, their common disjunctive ancestors are the most informative common ancestor of disjunctive ancestor of c_1 and c_2. This means that a_1 is a common disjunctive ancestor of c_1 and c_2 if for each ancestor a_2 more informative than a_1, a_1 and a_2 are disjunctive ancestors of c_1 or c_2 *(90)*.

$$CommonDisjAnc(c_1, c_2) = \{a_1|$$

$$a_1 \in CommonAnc(c_1, c_2) \wedge$$

$$\forall\, a_2: \left[\left(a_2 \in CommonAnc(c_1, c_2)\right) \wedge \left(IC(a_1) \leq IC(a_2)\right)\right] \Rightarrow$$

$$[(a_1, a_2) \in (DisjAnc(c_1) \cup DisjAnc(c_2))]\}. \qquad (2.31)$$

Chapter 3

Hydration properties of natural and synthetic DNA sequences with methylated adenine or cytosine bases in the R.DpnI target and BDNF promoter studied by molecular dynamics simulations (91)

Abstract

Adenine and cytosine methylation are two important epigenetic modifications of DNA sequences at the levels of the genome and transcriptome. To characterize the differential roles of methylating adenine or cytosine with respect to their hydration properties, we performed conventional MD simulations and free energy perturbation calculations for two particular DNA sequences, namely the BDNF promoter and the R.DpnI-bound DNA that are known to undergo methylation of C5-methyl cytosine and N6-methyl adenine, respectively. We found that a single methylated cytosine has a clearly favorable hydration free energy over cytosine since the attached methyl group has a slightly polar character. In contrast, capping the strongly polar N6 of adenine with a methyl group gives a slightly unfavorable contribution to its free energy of solvation. Performing the same demethylation in the context of a DNA double-strand gave quite similar results for the more solvent-accessible cytosine, but much more unfavorable results for the rather buried adenine. Interestingly, the same demethylation reactions are far more unfavorable when performed in the context of the opposite (BDNF or R.DpnI target) sequence. This suggests a natural preference for methylation in a specific sequence context. In addition, free energy calculations for demethylating adenine or cytosine in the context of B-DNA vs. Z-DNA suggest that the conformational B-Z transition of DNA transition is rather a property of cytosine methylated sequences but is not preferable for the adenine-methylated sequences investigated here.

3.1 Introduction

DNA methylation plays a major role in a wide variety of biological processes, including, for example, the regulation of gene expression and self-recognition. C5-cytosine methylation of DNA, on one hand, is one of the most important modifications of eukaryotic genes and plays an essential role in mammalian gene expression (92). N6-adenine methylation, on the other hand, is the foremost methylation type in bacteria and protects bacteria against the attack by foreign nucleic acid sequences. This defense mechanism can either be realized by marking the methylated DNA as their own, thus preventing their own degradation(93), or by labeling the attacking phages as 'foreign' and targeting them for degradation by cellular enzymes(94). DNA N6-adenine methylation has also recently been discovered in eukaryotes, such as plants and arthropods (9). Besides, mammalian RNA is apparently adenine methylated by the methyltransferase-like METTL3–METTL14 heterodimer complex (95) and it regulates gene expression via the 'epitranscriptome' model (15).

Mechanistically, methylation of DNA at cytosine or adenine bases can specifically affect the helical nature of DNA. C5-cytosine-methylated sequences have been observed to undergo a structural transition from the right-handed B-DNA fiber to the left-handed Z-DNA counterpart already at lower salt concentration than the respective unmethylated DNA (96). Liu et al explored in experiments the active role that Z-DNA fiber plays in modulating the inhibitory chromatin structure (97). N6-adenine methylation, on the other hand, facilitates the structural transition of B-DNA to the X-DNA form (96). In bacteria, N6-adenine methylation of the DNA sequence motif 5'-GATC-3' by the enzyme deoxyadenosine methylase (Dam) plays an important role in the timing of initiation of DNA replication, as well as in the coordination of cellular events, DNA mismatch repair, and gene regulation (98,99). An example of the specific proteins targeting this sequence is the R.DpnI enzyme (10) that targets N6-adenine-methylated sequences as foreign. Brain-derived neurotrophic factor (BDNF), a member of the nerve growth factor family of neurotrophins, has vital roles in the development, pathology and physiology of the nervous system (100). Gene regulation mediated by binding of the methyl-binding domain (MBD) of MeCP2 to the BDNF promoter is a eukaryotic transcriptional repression system, involving C5-cytosine methylation of the central CpG motif (78,99).

Furmanchuk and colleagues have carried out *ab initio* molecular dynamics simulations to investigate the structural properties of the four nucleic acid bases (NAB) in the gas state and upon hydration (101). Whereas hydration of adenosine was found to increase the flexibility of the rings due to the highly exposed water

structure, the three other nucleic acid bases showed restricted mobility upon hydration, with an increased base planarity. Here, we aim at extending the focus of such theoretical studies to the methylated forms of the cytosine and adenine nucleic acid bases.

Hydration of the methylated CpG step and the structural consequences of methylation have already been discussed before (102). Yet, a systematic comparison between the naturally occurring C5-cytosine methylated DNA and the nonmethylated form is lacking so far. Additionally, and to the best of our knowledge, the hydration properties of naturally occurring N6-adenine-methylated sequences have not been studied so far. We contrast here the hydration properties of the naturally occurring sequences with that of synthetic sequences where we swapped the central dinucleotide steps of the two methylated sequences plus their methylation state between the two studied natural sequences. Additionally, we compared the hydration properties and free energy difference of methylating N6-adenine- or C5-cytosine related with the conformational B-Z transition of R.DpnI target sequence and BDNF promoter sequence. Based on the structural and energetic parameters derived from our molecular dynamics studies, we derive a global picture how the hydration properties of different DNA sequences affect the preferences for either the methylated or non-methylated forms.

3.2 Methods

MD simulations were performed with the GROMACS 4.6 package (103) using the CHARMM27 force field (104) and the TIP3P water model (105). At first, parameterization of the nonstandard N6-methylated adenosine residue was done as follows: The H62 atom of adenine carrying a charge of +0.38e in CHARMM27 was replaced by a methyl group with the same formal charge. For this, every methyl-hydrogen was assigned the standard CHARMM27 charge of +0.07e, and the remaining charge of +0.17e was placed on the methyl-carbon atom. We followed in this charge assignment the original CHARMM philosophy of a modular nature of the distribution of charges. Charges on the replaced hydrogen atoms are summed into their original parent heavy atoms, thereby preserving the integer charge of the molecule. Typically, adding a functional group to an existing molecule would require no additional parameters (58). The equilibrium bond lengths, bond angles, dihedrals and force constants between methyl group and adenosine were determined with the VMD plugin Paratool (http://www.ks.uiuc.edu/Research/vmd/plugins/paratool/) which produces input values for an energy minimization and frequency calculation performed using Gaussian 3 (106). The Lennard Jones parameters of the methyl group were set to the standard values of a methyl group in CHARMM27. The

parameters of methylated cytosine already exist in the CHARMM27 force field distributed by the MacKerell group.

Several DNA systems were prepared to perform free energy perturbation and plain MD simulations (see below). First, we computed the relative free energy differences upon demethylating single N6-methyl-adenine and C5-methyl-cytosine. The molecular structures of the methylated forms of these compounds are illustrated in figure 3.1. Then, we simulated the B-DNA conformation of 8 DNA sequences, namely the methylated and nonmethylated DNA versions of (1) the bacterial R.DpnI target sequence d(CTGG(N6-meA)TCCAG) (2) the BDNF promoter d(TCTGGAA(C5-meC)GGAATTCTTCGA) (3) the bacterial R.DpnI target sequence, with the central AT dinucleotide pair replaced by the CG run, d(CTGG**(C5-meC)G**CCAG) and (4) the BDNF promoter with the central CG dinucleotide pair replaced by an AT dinucleotide pair d(TCTGGAA**(N6-meA)T**GAATTCTTCGA). Additionally, we simulated the Z-DNA conformation of the 4 natural DNA sequences mentioned before, namely the methylated and the non-methylated forms of the bacterial R.DpnI target sequence and the BDNF promoter sequence. Starting conformations were generated in ideal B- or Z-DNA geometries using the 3DNA package (107).

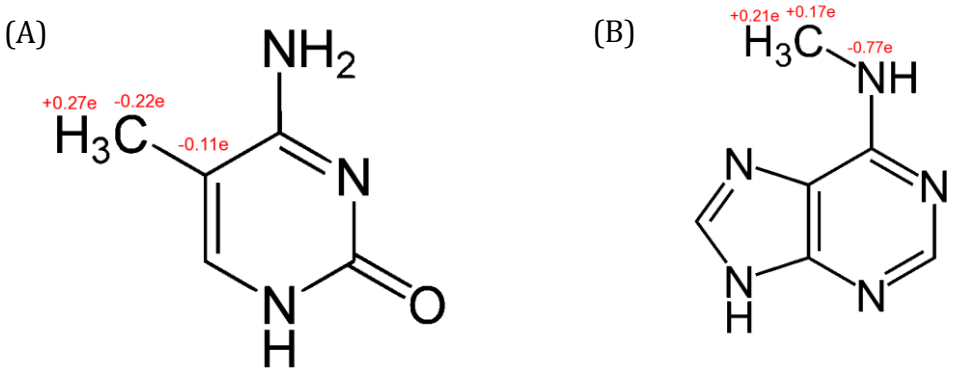


Figure 3.1 A schematic representation of (A) C5-methylated cytosine and (B) N6-methylated adenine. Partial atomic charges of the methyl carbon, methyl hydrogens, and the atoms these are bonded to are marked in red.

We simulated adenine and the modified N6-methylated-adenine bases in a cubic water box with 3.02 nm dimension and 0.10 mol/l of NaCl added, and a total number of 2720 atoms. Cytosine and C5-methylated-cytosine were both simulated in a cubic water box with 2585 atoms, a final size of about 2.90 nm dimension, and with 0.10 mol/l of NaCl added. R.DpnI target sequences, natural (in the B- or Z-DNA forms), or synthetic, methylated or non-methylated, were placed in a cubic water box of around

5.90 nm box dimensions with 0.10 mol/l NaCl added. These systems had a total size of 20320 atoms. The longer sequences of the BDNF promoter, natural (in the B- or Z-DNA forms), synthetic, methylated or non-methylated, were placed in a cubic water box of around 9.30 nm box dimensions and with the same salt concentration added. The total size of the simulated BDNF systems was 56200 atoms for DNA solvated in water. Periodic boundary conditions were used. Coulombic interactions were computed using a short-range cut-off of 10 Å and long-range interactions were treated by the particle-mesh Ewald (PME) summation method (78). The nonbonded Lennard-Jones interactions were calculated using a smooth cutoff of 10 Å. The integration time step was set to 1 fs. The temperature was maintained at 310 K by employing leap-frog stochastic dynamics forces (108) with a damping coefficient of 0.1 ps^{-1}.

As preparatory stages for plain MD simulations and free energy calculations, all starting conformations were first energy-minimized for 50000 steps using the steepest descent algorithm followed by a second energy minimization for 10000 steps using a quasi-Newtonian algorithm with the low-memory Broyden-Fletcher-Goldfarb-Shanno approach. The tolerance was set to 1.0 kJ mol^{-1} nm^{-1}. Next, the systems were heated to 310 K during 4 ps. After that, all systems were equilibrated during a 1-ns run in the NVT ensemble with harmonic restraints with a force constant of 1000 kJ mol^{-1} nm^{-2} used for all DNA atoms. With restraints kept, each system was further equilibrated for 500 ps in the NPT ensemble, and then for another 500 ps without restraints. For the resulting structures, two types of simulations were performed, alchemical free energy perturbations and plain MD simulations.

Alchemical FEP calculations were applied employing the Bennett acceptance ratio with error bars (BAR) (75,109) to compute the difference in free energy of solvation between the methylated and non-methylated forms of the single bases cytosine and adenine and when they were part of the BDNF promoter sequence, the R.DpnI-binding sequence, or the synthetic sequences mentioned above. In one type of calculations, the methyl group 'alone' was first perturbed into a dummy group and the interactions of the respective hydrogen atom were turned on subsequently. In such calculations, the target methyl group attached to adenine or cytosine was annihilated in two stages (110). In the first stage, the electrostatic interactions of each methyl group were switched off in a step-wise manner and its respective charge was assigned to the N6 atom or C5 atom of adenine or cytosine, respectively. For this, the system Hamiltonian was coupled to a coupling parameter λ where $\lambda = 0$ corresponds to the reference state and $\lambda = 1$ to the perturbed state. No soft-core potential was used in this step. In the second stage, LJ potential of the methyl group was turned off. In a

final step of this stage, we mutated the methyl group into the respective hydrogen as follows: the hydrogen atoms of the methyl group were turned into non-interacting dummy atoms (by switching their epsilon and sigma Lennard-Jones parameters to zero) and the dummy methyl carbon was mutated into the respective dummy hydrogen. In this stage, a soft core potential (111) was used where soft-core alpha was set to 0.5, the soft-core power to 1.0, and soft-core sigma to 0.3. To complete the free energy cycle, Lennard Jones interactions, and then the electrostatic interactions of the 'dummy' hydrogen atoms of the non-methylated adenine or cytosine were turned on. For the simulations with the single nucleic acid bases (NAB), each of the four stages was decomposed into 21 intermediates states ($\Delta\lambda = 0.04$). Simulations of all intermediate states were started from the equilibrated conformation at the reference state with $\lambda = 0$ with the same two-step energy minimization, followed by equilibration over 1 ns in the NVT ensemble and 500 ps equilibration in the NPT ensemble with harmonic restraints, and 500 ps without any restraints. Data were collected during another 1.5 ns for each window. This yields a total simulation time of 21 x 3.5 ns = 73.5 ns for each unidirectional simulation.

For the perturbation of the full-length double strand DNA sequences, we used the same simulation protocol, with 26 windows instead of 21 for each of the main four stages. For the R.DpnI target sequence (natural and synthetic), both methyl groups on the two strands were perturbed simultaneously because the upper and lower DNA strands of the R.DpnI target sequence have the identical sequence (detailed above). For the BDNF promoter, the same perturbation process had to be split into two stages with the involvement of a hemimethylated intermediate because the upper and lower strands have different sequences.

As a consistency check of the free energy calculations, we performed a second set of simulations where we computed the solvation free energies of the single NABs (see figure 3.2). Again, these validation calculations were split into four stages. First, the methylated NABs were annihilated in water into dummies in two stages, a discharging stage and a LJ elimination stage. Next, the interactions of the nonmethylated NABs were switched off, also in two stages, a discharging stage and a LJ stage. For these systems, a total of 21 windows per stage were performed. The difference of solvation of the methylated and the nonmethylated forms of DNA was then taken.

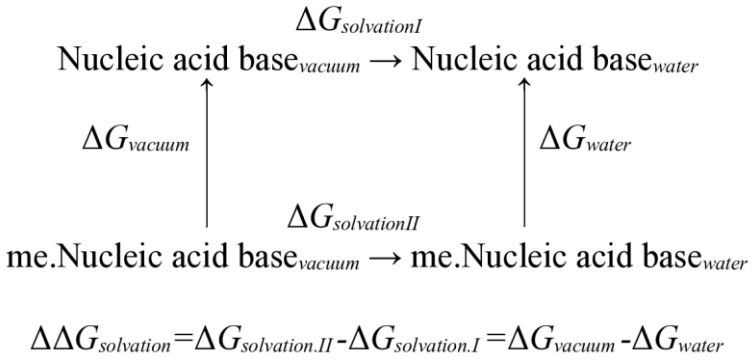


Figure 3.2 A scheme illustrating free energy calculations for the differential stability of the methylated and the nonmethylated form of the nucleic acid base (NAC) upon solvation.

As mentioned before, the set-up systems were also used to perform plain MD simulations for each studied entity. Simulations for the NAB were all performed for 10 ns each, and for the larger DNA sequence systems for 30 ns.

3.3 Results

In this study, we performed conventional MD simulations and free energy calculations to study how the aqueous microenvironment affects the stability of naturally occurring methylated DNA double-stranded sequences and the respective non-methylated variants.

3.3.1 Free Energy Perturbation

First, free energy perturbation was performed to determine the differential water preferences of the single methylated and non-methylated cytosine/adenine bases. Table 3.1 summarizes the results from free energy perturbation calculations for converting the methyl group of solvated N6-methylated-adenine into a hydrogen atom, the same process for the solvated C5-methylated cytosine, and consistency checks for the difference in the free energy of solvation for the two NABs, where solvation free energies of the NABs in the methylated and the non-methylated forms were calculated explicitly. Whereas the methylated form of cytosine is strongly favored over the unmethylated form in water, the unmethylated adenine is about as favorable in water as the methylated one. The free energy differences for methylation (-26.73 and 0.40 kcal/mol) were within 1 kcal/mol from the computed differences in solvation free energies (-26.16 and -0.52 kcal/mol).

Table 3.1 Results from free energy calculations (kcal/mol) (A) for perturbing the methylated adenine into non-methylated adenine in water and vacuum and for the solvation free energies of 6-N-methyl-adenine and adenine, (B) for perturbing the methylated cytosine into non-methylated cytosine in water and vacuum and for the solvation free energy of 5-methyl-cytosine and cytosine. In (A+B), the free energy of mutating the hydrogen atom to the methyl (-1.67 kcal/mol) residue cancels out in vacuum and water. Values in brackets are statistical errors reported by GROMACS.

(A) adenine	$\Delta G_{discharging}$	$\Delta G_{turning\ LJ\ off}$	ΔG_{total}	$\Delta\Delta G$
met. adenine$_{vacuum}$ **→*adenine***$_{vacuum}$	4.21 (±0.02)	4.26 (±0.01)	8.48 (±0.02)	
met.adenine$_{water}$ **→ *adenine***$_{water}$	4.08 (±0.02)	4.00 (±0.02)	8.08 (±0.03)	0.40 (±0.02)
met. adenine$_{water}$ **→ *met. adenine***$_{vacuum}$	-22.82 (±0.02)	-0.80 (±0.02)	-23.62 (±0.02)	
adenine$_{water}$ **→ *adenine***$_{vacuum}$	-22.15 (±0.02)	-0.95 (±0.02)	-23.10 (±0.02)	-0.52 (±0.02)

(B) cytosine	$\Delta G_{discharging}$	$\Delta G_{turning\ LJ\ off}$	ΔG_{total}	$\Delta\Delta G_{total}$
met. cytosine$_{vacuum}$ **→ *cytosine***$_{vacuum}$	0.00 (±0.00)	1.62 (±0.00)	1.62 (±0.0)	
met.cytosine$_{water}$ **→ *cytosine***$_{water}$	26.82 (±0.02)	1.53 (±0.03)	28.35 (±0.03)	-26.73 (±0.03)
met.cytosine$_{water}$ **→ *met. cytosine***$_{vacuum}$	12.80 (±0.04)	0.00 (±0.08)	12.80 (±0.06)	
cytosine$_{water}$ **→ *cytosine***$_{vacuum}$	38.93 (±0.16)	0.03 (±0.12)	38.96 (±0.14)	-26.16 (±0.11)

Next, we calculated the free energy changes (ΔG) for the same process of demethylation in the context of the DNA double strand helix (in the B- or Z-DNA fiber forms; see table 3.2). Precisely, we mutated the methylated B-DNA fiber in water to the non-methylated B-DNA, either in the sequence context of the R.DpnI target sequence, the BDNF promoter, C5-methylated cytosine introduced into the R.DpnI target sequence instead of N6-meA, or N6-methylated adenine introduced into the BDNF promoter instead of the C5-methylcytosine. The result for demethylating two cytosine bases in the context of the DNA double strand helix in B-DNA conformation (66.6 kcal/mol) is comparable to the computed difference in solvation free energies and the computed free energy difference for de-methylation of two cytosine bases (13.14 kcal/mol larger). This reflects our expectation that the methyl-group attached to the C5 atom of cytosine is well exposed to solvent in the DNA context. In contrast, the computed free energy for demethylating N6-methylated adenine in the R.DpnI target sequence (31.6 kcal/mol) is far more unfavorable than when performing the same perturbation of the single base in solution. This likely affects the buried nature of an N6-attached methyl group (see below). Interestingly, the results from the free energy perturbation calculations performed in the context of the synthetic sequences show that DNA has a strong preference for methylation in the naturally occurring

sequences when compared to the synthetic ones. This is especially pronounced for C5-methylated cytosine sequences.

Table 3.2 Free energy differences for demethylating (kcal/mol) adenine (A) cytosine (B). Results are given for perturbations performed in the native sequence content (R.DpnI target sequence for adenine and BDNF promoter for cytosine) both modeled into B-DNA and into Z-DNA conformation. For comparison, we replaced adenine (or cytosine) by cytosine (or adenine) in the respective DNA sequences and computed the free energy for demethylation in the sequence context of the BDNF (or R.dnpI) promoter. The perturbation of the methyl group was conducted in one step for the R.DpnI sequence context (natural, B- and Z-DNA forms, and synthetic), and in two stages involving a hemi-methylated intermediate for the BDNF promoter (see text). Values in brackets are the statistical errors reported by GROMACS.

(A) adenine-N6-methylation	**DNA conformation**	**DNA sequence**	$\Delta G_{discharging}$	$\Delta G_{turning}$ LJ off	ΔG_{total}	$\Delta\Delta G$
met. DNA →DNA	B-DNA	R.DpnI	24.40 (±0.65)	7.15 (±0.33)	31.55 (±0.73)	
met. DNA→ DNA	Z-DNA	R.DpnI	16.22 (±0.69)	7.24 (±0.34)	23.46 (±0.54)	
met. DNA → hemi-methylated DNA	B-DNA	BDNF	-8.80 (±1.49)	4.29 (±0.85)	-4.51 (±1.21)	8.77 (±1.16)
hemi-methylated DNA→ nmDNA	B-DNA	BDNF	11.32 (±1.28)	1.96 (±0.91)	13.28 (±1.11)	

(B) cytosine-C5-methylation	**DNA conformation**	**DNA sequence**	$\Delta G_{discharging}$	$\Delta G_{turning}$ LJ off	ΔG_{total}	$\Delta\Delta G$
met. DNA→ hemi-methylated DNA	B-DNA	BDNF	31.04 (±1.01)	7.26 (±0.68)	38.30 (±0.86)	66.60 (±0.97)
hemi-methylated DNA→ nmDNA	B-DNA	BDNF	19.19 (±1.29)	9.11 (±0.79)	28.30 (±1.07)	
met. DNA→ hemi-methylated DNA	Z-DNA	BDNF	25.18 (±0.55)	-11 .54 (±0.58)	13.64 (±0.56)	67.56 (±0.63)
hemi-methylated DNA→ nmDNA	Z-DNA	BDNF	46.45 (±0.74)	7.47 (±0.65)	53.92 (±0.70)	
met. DNA→ DNA	B-DNA	R.DpnI	-8.87 (±0.80)	8.82 (±0.20)	-0.05 (±0.59)	

We also checked for the differential propensities of the methylated and the non-methylated DNA sequences to undergo a B-Z DNA transition. For this, we focused on the native sequences, namely the C5-cytosine methylated BDNF promoter, and the N6-adenine methylated R.DpnI target sequence. As is discussed before for B-DNA, we

now mutated the methyl groups into hydrogen atoms in the Z-DNA conformation. The contribution of methylation to the free energy difference between B-DNA and Z-DNA can thus be calculated by closing the free energy cycle shown in figure 3.3. Our calculations showed that C5-cytosine methylation makes a slightly favorable contribution to the B/Z DNA transition (ΔΔG = (66.60 - 67.56) kcal/mol = -0.96 kcal/mol). In contrast, N6-adenine methylation does not contribute favorably to the B-Z DNA transition (ΔΔG = (31.55 - 23.46) kcal/mol = 8.09 kcal/mol). We noticed that upon demethylation in the Z-DNA form, the DNA fiber rearranges the conformation such that a complete opening of the strands takes place at the end of the demethylation process.

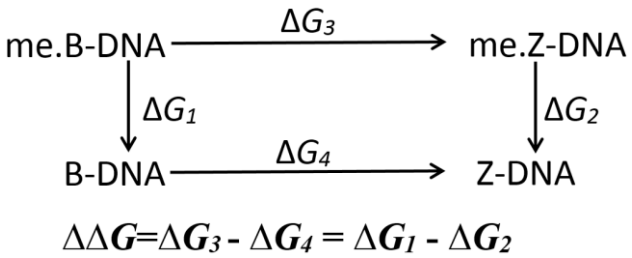


Figure 3.3 A scheme illustrating free energy calculations for the contribution of DNA methylation to the B/Z-transition of DNA.

3.3.2 Differential hydration properties of the methylated and non-methylated DNA

To obtain a molecular picture of the determinants characterizing the energetics of these demethylation processes, we analyzed the hydration properties of DNA by analysis of plain MD simulations. First, we computed the solvent accessible surface area (SASA) of the methyl groups. We found that the methyl groups attached to cytosine and adenine bases have SASA values of 0.55 nm^2 and 0.65 nm^2, respectively. Table 3.3 lists the SASA for the methyl groups (Watson and Crick strands) in the 8 DNA systems we studied. The two C5-cytosine methyl groups in the B-DNA form of the natural BDNF and in the synthetic R.DpnI target sequences in the B-DNA form showed a SASA of 70% and 67% compared to that of a single methylated cytosine base in solution, respectively. In the Z-DNA fiber form of the BDNF promoter, the two C5-cytosine methyl groups showed a comparable SASA of 66%. On the other hand, the two N6-adenine methyl groups of the natural R.DpnI target and the synthetic BDNF sequence in the B-DNA fiber form had only a SASA of 44% and 40% compared to that of a methylated adenine base in solution, respectively. Surprisingly, the two N6-adenine methyl groups showed a SASA of 71.5% of the methylated adenine in solution.

Table 3.3 Solvent accessible surface areas (nm^2) for the methyl groups of N6-methylated adenine and C5-methylated cytosine bases in a DNA double-strand.

	R. DpnI target sequence		BDNF promoter sequence	
Methylated Sequence	Upper Strand	Lower Strand	Upper Strand	Lower Strand
Natural	0.29 (±0.05)	0.28 (±0.05)	0.40 (±0.06)	0.36 (±0.0
Synthetic	0.37 (±0.05)	0.36 (±0.05)	0.26 (±0.06)	0.27 (±0.0
Z-DNA fiber	0.47 (±0.09)	0.46 (±0.10)	0.39 (±0.06)	0.34 (±0.0

Next, we calculated the number of coordinating waters to characterize the structuring of water around the DNA sequence. As is commonly done, the first hydration shell was defined to range up to the first minimum of the water density distribution that was found to be 3.56 Å away from the DNA surface. Table 3.4 shows the coordination number of water in the first hydration shell and the SASA around the 10 basepairs long DNA sequence of the R.DpnI system and around the BDNF promoter. For consistency, we also selected the central 10 basepairs of the BDNF promoter sequence for this analysis. For all systems, the methylated forms of DNA showed small increases of the water coordination number. As expected, all methylated forms of DNA also had slightly larger SASA than the non-methylated forms (see table 3.4). Thus, the small increases in the number of coordination waters can be traced back to the increased molecular surface due to the attached methyl groups. Introducing the synthetic methylation variant had only a small effect on the water coordination around the R.DpnI target sequence. In contrast, for the BDNF promoter target sequence, mutating the central dinucleotide basepair of C5-me.CG into N6-me.AT led to a more sizeable increase in the number of coordinating waters, although this is still not statistically significant. On average, 3.50 (±1.64) ions coordinated to DNA in the first hydration shell (what amounts to a local concentration of 0.43 M NaCl in the first hydration shell). The ion contribution can be mostly attributed to Na^+ ions, because Cl^- only made up a small fraction with an average of 0.006 atom. Interestingly, whereas the Z-DNA conformation introduced a much higher coordination of waters around the BDNF promoter sequence, this was not the case for the R.DpnI target sequence. Nevertheless, enlarged SASA values were observed for both sequences in the Z-DNA conformation. Additionally, the local ion concentration in the first hydration shell around Z-DNA is 8 times higher (3.3 M) compared to its local concentration around the B-DNA fiber form (for this, see also the ion radial distribution below).

Next, we computed the water density in the first hydration shell. As a reference, the density of water in a box with bulk water was computed as about 0.90 g/mL using the double cubic lattice method (112) implemented in GROMACS. For the DNA sequences

studied here, all sequences showed a roughly 25% lower density of pure water in the first hydration shell of (0.62-0.63) ± (0.01-0.02) g/mL than in bulk water. However, no significant changes in water density were found when changing the methylation status, neither in the natural sequences nor in the synthetic ones. For the joint density of water and ions, we obtained slightly higher values of (0.64-0.65) ± (0.01-0.02) g/mL.

Table 3.4 Coordination water number and solvent accessible surface areas (nm^2) in the first hydration shell around the methylated and nonmethylated DNA in the 12 studied DNA systems

(A) Total DNA Sequence		R. DpnI target sequence		BDNF promoter sequence	
Methylation status		✓	✖	✓	✖
Natural	Coordination No.	289.81 (±7.29)	288.02 (±7.33)	258.32 (±6.78)	257.4 (±6.7
	SASA (nm^2)	38.31 (±0.45)	38.16 (±0.44)	35.44 (±0.44)	35.20 (±0.4
Synthetic	Coordination No.	289.91 (±7.27)	289.28 (±7.41)	263.70 (±6.62)	262.41 (±6.6
	SASA (nm^2)	38.18 (±0.45)	38.03 (±0.47)	35.56 (±0.44)	35.46 (±0.4
Z-DNA fiber	Coordination No.	281.18 (±7.69)	276.28 (±7.57)	269.76 (±7.98)	269.82 (±7.4
	SASA (nm^2)	42.81 (±1.19)	41.32 (±1.00)	39.78 (±0.79)	41.27 (±0.9

Next, we characterized water ordering in the first hydration shell by computing the tetrahedral order parameter. As a reference, we first computed the tetrahedral order parameter for a periodic box with around 30000 TIP3P water molecules and obtained 0.512. In the DNA simulations, water in the first hydration shell had tetrahedral order parameters of about 0.40 (R.DpnI) and 0.34 (BDNF), respectively, what is lower compared to that for bulk water (figure 3.4). In the context of the natural DNA sequences, the values found for the methylated and the non-methylated forms of the bases were highly similar. In contrast, figure 3.4 shows that for the synthetic R.DpnI target sequence, noticeable deviations were observed at the end of the simulation of the nonmethylated form but not the methylated form. Interestingly, in the simulations of the synthetic BDNF promoter sequence both methylated and non-methylated forms gave about 0.04 lower values than observed for the natural sequence. Furthermore, waters around the Z-DNA fiber of the R.DpnI target sequence showed 0.06 lower values for the tetrahedral order of water in comparison to the B-DNA fiber form, with a clear deviation for the methylated sequence at the end of the simulation. For the BDNF promoter, the tetrahedral order of water around the Z-DNA fiber was very similar to that for the synthetic sequences in the B-DNA form.

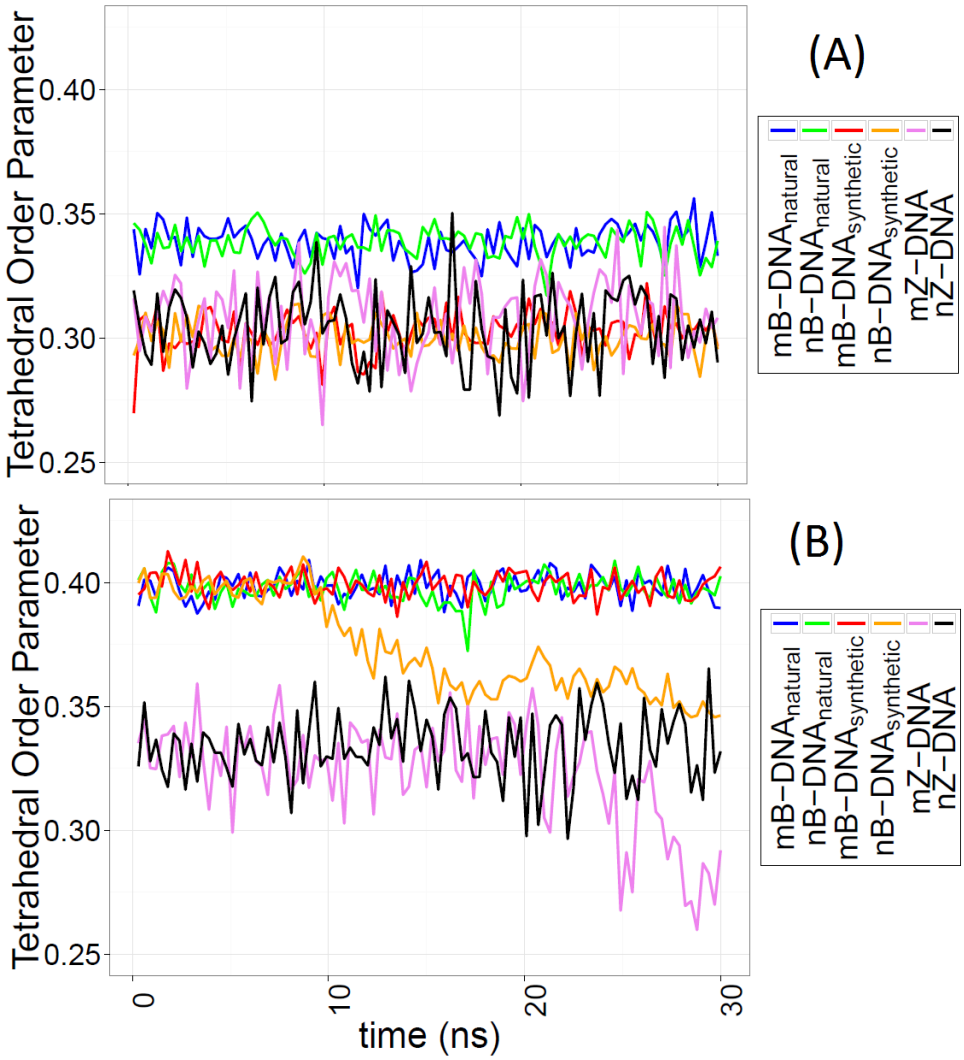


Figure 3.4 The tetrahedral order parameter of water molecules in the first hydration shell of the central dinucleotides (A) in the BDNF promoter sequence, and (B) in the R.DpnI target sequence.

Finally, figure 3.5 illustrates the radial distributions of water oxygens and the Na^+ ions around the negative phosphate of the total DNA sequence. Na^+ ions were observed to strongly compete with water molecules for locations in the immediate surrounding of the phosphate group of DNA. This tendency was particular pronounced for the Z-DNA conformation that exposes the charged groups of the DNA backbone to the solution more strongly than other DNA conformations. Even at far distances (1.5 nm), the ion concentration around Z-DNA is ca. 37% larger than around B-DNA.

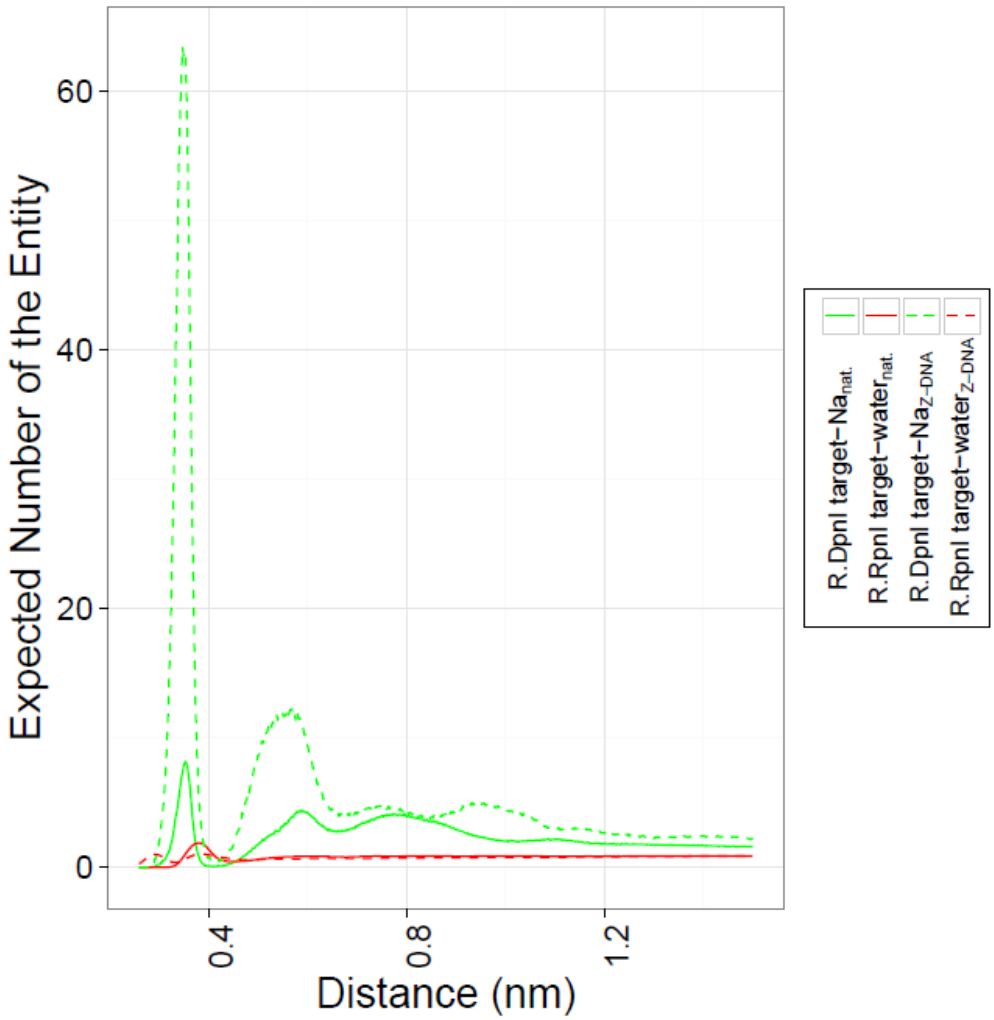


Figure 3.5 Radial distribution function of water oxygens and Na^+ around DNA phosphate groups in simulations of naturally occurring sequences studied in the methylated form. See legend for further illustration.

3.4 Discussion

In this study, we performed plain MD simulations and free energy perturbation calculations to investigate the relationship between DNA sequence, its methylation state, and its hydration properties.

First, free energy perturbation calculations showed that methylating the hydrophobic C5 atom of cytosine leads to more favorable solvation in water. The opposite result was obtained when methylating the polar N6 atom of adenine. As a consistency check, we performed two sorts of FEP calculations. In the first set, the methyl group was mutated to the respective hydrogen bound to the NAB. In the second set, the whole NAB was mutated into non-interacting dummy atoms either in the methylated or in

the nonmethylated form, and then the difference of solvation free energy was computed. Resultant calculated free energy differences obtained from both calculation types were nicely consistent. We observed no drastic change in free energy perturbation results between the base and nucleotide forms of the nucleic acids (data not shown). We acknowledge that the calculated free energy difference for mutating cytosine into C5-methylcytosine appears a bit high. Thus, re-parameterization of the force field according to the strategy described by Vanommeslaeghe et al (58) may be warranted. For calibration, it would be desirable to determine experimental solvation free energies of methylated nucleic acids.

Next, a similar set of free energy perturbation calculations was carried out in the context of the native or non-native DNA sequence context. Interestingly, we found that the native sequences had a strong energetic preference for the methylated forms compared to the synthetic sequences what suggests that nature apparently has a strong selection for specific DNA sequences to code for methylation. One can expect that this preference likely results from cumulative effects of variations in internal structural parameters of DNA and according rearrangements of ionic and water preferences around the DNA. Furthermore, the Z-DNA conformation showed a slightly higher preference for C5-cytosine methylated DNA than for non-methylated DNA, but is clearly disfavored for N6-adenine methylated DNA.

To get further insight into this behavior, we analyzed various water properties around DNA using snapshots from plain MD simulations. For example, we analyzed the SASA of the methyl groups in the free C5-methylated cytosine, N6-methylated adenine, and the 12 studied DNA systems. The SASA of the methyl groups of C5-methylcytosine in the natural DNA strand was 70% of that when attached to the free cytosine in solution. When assuming a proportionality between the degree of solvent-exposure and the demethylation free energy, this would make up for ~41kcal/mol of the computed 66.6 kcal/mol for the free energy of mutating the methylated DNA sequence to the non-methylated form. On the other hand, the C5-methyl cytosine containing synthetic R.DpnI target sequence showed only a slight decrease in the exposed surface area (3%) of the two methyl groups, but the same mutation accounted only for 0.05 kcal/mol here. In contrast to cytosine, the methyl groups in the N6-methylated adenine of the natural sequence had a clearly smaller degree of solvent exposure (only 43% of the SASA computed for a single adenine base in water) than cytosine, what may be related to the overall small change of free energy found in the FEP calculation. We also noticed a slight decrease in the two methyl surface area (~4%) for the synthetic sequences when compared to the natural ones. Moreover, a strongly increased exposure of the N6-adenine methyl groups was

noticed in the Z-DNA conformation of the R.DpnI target sequence. This likely contributes to the lower preference of the N6-adenine-methylated R.DpnI target sequence to be in the Z-DNA form than in the B-DNA form. For comparsion, the SASA for the C5-cytosine methylated BDNF sequence in Z-DNA conformation was similar to the SASA for its B-DNA conformation.

We suggest that the process of demethylation should be considered as a combined process of (a) creating a cavity inside a DNA double strand and (b) of performing the demethylation reaction in water. The fraction between both processes can be expected to vary depending on the degree of solvent exposure of each base. One complication, though, is the fact that non-methylated DNA can of course adopt its conformation with respect to the methylated form. Obtaining deeper insights into such processes will require investigating more DNA systems in a similar fashion.

The coordination number of water around the central 10 basepairs in DNA was slightly lower in the nonmethylated form than in the methylated form. This decrease was nicely consistent with the decrease of the SASA of the 10 basepairs, so that one can invoke a good degree of proportionality between the exposed surface of the 10 base-pairs and the coordination water number. Also, we computed the water density in the first hydration shell. We noticed that coordination waters that surround DNA have a roughly 25% lower density compared to bulk water. Not much change in density was found when also including the presence of Na^+ ions. However, Na^+ ions *per se* had an about four times increased molarity in the first hydration shell when compared to its bulk characterization that was set to 0.1 M. This reflects the competition between the two polar species water and sodium ions for energetically favorable positions along the highly negatively charged DNA (figure 3.5).

Finally, the tetrahedral order parameter of water decreased markedly around DNA compared to bulk water (given the same temperature for both simulations). This is a typical behavior of hydration water when it is in contact with mostly hydrophobic solutes (113). A study performed by Kumar et al showed that the tetrahedral order parameter is inversely proportional to temperature (114). Upon increasing the average distance between water molecules up to some threshold, the tetrahedral order of water first dropped, then increased again to reach a constant plateau at very sparse distances (beyond approximately 0.45 nm). Here, all simulations were performed at a constant temperature of 310 K. We ascribe the observed lowering of the tetrahedral order parameter for water around DNA to its structural adaptation and the resulting lower density around the hydrophobic portions of the DNA bases and to the high competition with salt ions for energetically favorable positions at the

phosphate backbone. Interestingly, the tetrahedral order parameter of water was further decreased with respect to the native DNA sequences around the synthetic sequences of the BDNF promoter, around the nonmethylated form of the synthetic R.DpnI target sequence, and around the Z-DNA form for both biological sequences. Since the confinement properties of water molecules around DNA sequences have been related to protein-DNA binding processes before (115), such effects may contribute to the binding specificity to proteins of particular methylated or non-methylated DNA sequence motifs, and may favor specific DNA conformations.

3.5 Conclusion

MD simulations and free energy perturbation calculations were performed to check for differential hydration properties of methylated and nonmethylated forms of DNA, for naturally occurring sequences in the B- and Z-DNA forms and for synthetic ones. We found that the specific sequence of DNA has a larger effect on the structure of nearby than methylation/demethylation of a central base. Hence, the free energy of demethylation depends strongly on the sequence content that leads to small, but distinct structural variations of DNA and the conformation of coordinating ionic water. Methylation and DNA sequence content altogether seem to have substantial effects on the properties of water surrounding DNA, so that the specific sequence code appears to be tightly coordinated with its respective methylation status.

Chapter 4

Epigenetic Switching: Transcriptional gene regulation with the methyl-CpG binding domain of MeCP2 studied in an E. coli-based in vitro expression system **

** This project described in this chapter was carried out in collaboration with Marc Schenkelberger in the group of Prof. Albrecht Ott/chair of experimental biophysics, Saarland University. All the experimental part was performed by MS in the Ott group (116). Simulations were performed by the author.

Abstract

Recent advances in systems biotechnology have led to cell free expression systems that not only produce vital proteins but also enable the in vitro study of functional molecular scaffolds and their dynamic interaction. Cytosine methylation is a hot spot in the epigenetic regulation of eukaryotic gene expression. The human methyl-CpG binding domain (MBD) is a master regulator that binds to CpG islands. However, the methylation dependent recognition by MBD is still poorly understood. Here we report the first MBD CpG 'epigenetic switch' in a totally endogenous cell-free expression system using the *E. coli* molecular expression machinery. This is achieved by integrating the CpG methylation-binding domain of the human repressor MeCP2 as a transcriptional regulator for gene expression with the specific mammalian BDNF promoter into the bacterial cell-free extract. We combine our study with molecular dynamics simulations for a deeper understanding of the specificity of the epigenetic regulation of gene expression by MBD. Plausible conformational changes, including, for example, an opening in the B-DNA structure due to the binding of MBD match well the mutation effects seen in our experiments. Simulations at varying ionic strength confirm that the selected conditions of the extract are well-suited to study the physiologically relevant methylation-dependent binding of MeCP2. Our results demonstrate that the molecular cooperative functions of mammalian epigenetic transcription regulation can be reproduced in vitro and further investigated on the molecular and dynamic levels for a more detailed understanding of MBD:DNA binding specificity when compared what can be achieved in an organism.

4.1 Introduction

Escherichia coli cell-free expression systems are widely used in biomedical research and *in vitro* synthetic biology, e.g., for the synthesis of human proteins. The molecular machinery present in the cytoplasmic extract performs transcription and translation and there is no need to add further enzymes. Compared to *in vivo* methods, the cell free system has the advantage that the experimental conditions, including molecular composition and concentration, DNA-sequence, and temperature can be controlled very easily. The *in vitro* system is devoid of the unknowns of an organism, what introduces a less complex environment with easily controllable variables. Additionally, there is precise knowledge of all transcriptional elements that come into play. Shin and colleagues produced a new system that utilizes the endogenous *E. coli* RNA polymerase and sigma factor 70 (117). It presents an important advantage over earlier systems using the bacteriophages, including unlocking of transcription modularity. The extract produces recombined proteins in the micro molar range in a few hours and GFP enables quantitative measurement of the expression levels without the need for further protein purification.

Methylation of DNA is one of the most important modifications of eukaryotic genes. It plays an essential role in mammalian gene expression. The enzymatic addition of a methyl group to the DNA base cytosine usually takes place at the CpG islands (CGIs), a genomic region with a high frequency of CpG dinucleotides. About 40 % of mammalian genes contain CGIs in their promoters and exonic regions, whereby promoter CGIs are normally unmethylated (92). While cytosine methylation is predominant only in eukaryotic genomes, prokaryotes are known to predominantly exhibit adenosine methylation. However, a recent study (118) has shown that cytosine methylation is associated with stationary phase prokaryotic gene expression plus a perceived marginal effect in regulating the exponential growth of bacterial cells.

The human protein MeCP2 belongs to a family of DNA binding proteins that can mediate gene silencing by binding specifically to methylated CpG sites and cause transcriptional repression. MeCP2 possesses a methyl-binding domain (MBD) that directly interacts with the DNA methyl group through strong electrostatic interactions (119). For example, MBD shows high affinity for binding to the promoter III of the mouse-brain-derived neurodrophic factor (BDNF), which contains a single central CpG pair (99),(28),(120),(121). Recent crystallographic results of Ho and colleagues for the MeCP2 protein bound to the methylated BDNF promoter (28) showed that binding of MBD to DNA causes a perturbation of the perfect B-DNA form

at the bound interface, involving a narrowing of the minor groove at the binding interface and a drift from the ideal B-DNA form (28). Khrapunov et al showed that solution conditions at physiological or higher salt concentrations are necessary for the discrimination by the MBD protein between the mCpG and CpG with high specificity (119). Moreover, it is known (122) that the MeCP2 does not strongly recognize specific sequences *per se,* but rather, it detects the methylation status. Thus, hypermethylated promoters are enriched with this protein, whereas the nonmethylated promoters are devoid of it.

Besides the crystallographic evidence, molecular dynamics simulations are becoming an accepted technique to reveal the conformational and dynamic characteristics of DNA-containing systems. For example, conventional MD simulations and alchemical free energy perturbation calculations were conducted by Zou and colleagues (123) for the MBD:DNA complex. Thereby they elucidated the beneficial effect of DNA methylation on the binding specificity of the MBD:DNA complex by means of minimizing the methyl surface area being exposed to the solvent upon binding as well as by strengthening the interaction between mDNA and MBD proteins.

In this work, we investigated the role of the methyl-CpG binding domain of MeCP2 as a transcriptional repressor for eukaryotic gene expression with the BDNF promoter. To this aim, we used the *Escherichia coli* cell free expression system with the endogenous *E. coli* RNA polymerase and sigma factor 70 (117). We combined experiments with molecular dynamic simulations to study how the binding of the methyl-binding domain (MBD) of the MeCP2 protein affects the structure of the wild-type DNA and several mutant DNA forms that we test. The combined evidence from experiment and simulations contribute to a better understanding of the binding specificity of MBD:mDNA and the effect of the environmental conditions therein.

4.2 Methods

4.2.1 Preparation of the cell free extract

The BDNF promoter (see table 4.1) was cloned into a plasmid. The methylated promoter controled the transcription of a gene that holds the information for the enhanced green fluorescent protein (eGFP) and mediated transcriptional repression. eGFP is used as a fluorescent reporter for the expression (116). The cell-free expression system used in this study was developed by Shin, et al (117). It is a modification of the protocol presented by Kigawa et al (124).

Table 4.1 DNA sequences of the studied BDNF promoters. All promoters contain a central CpG motif (blue), which can either be methylated or unmethylated. In the case of the four mutated versions of the BDNF promoter (M1-M4) the red bases indicate the sequence that is mutated compared to the wt. Mutants were selected by us.

Promoter	**Sequence (5'-3')**
BDNF-wt	CTG-GAA-CGG-AAT-TCT-TTC
BDNF-M1	CTG-GAA-CGC-AAT-TCT-TTC
BDNF-M2	CTG-GAA-CGG-AGC-CCT-TTC
BDNF-M3	CTG-GGG-CGG-AAT-TCT-TTC
BDNF-M4	CTG-GGG-CGG-AGC-CCT-TTC

4.2.2 MD Simulations

As structural reference for the MeCP2:DNA complex, we used the X-ray structure of the BDNF promoter bound to the methyl-binding domain (RCSB:3C2I; (28)).

The MD simulations were performed with the GROMACS 4.5.5 package (103) using the CHARMM27 force field (104) and the TIP3P water model (105). The parameters for 5-methyl-cytosine were used as defined in the CHARMM force field. Systems with unbound DNA duplex strands or protein:dsDNA complexes were placed in a dodecahedral water box of 16 nm box dimensions with 0.10 mol/l (or 0.20 mol/l) of KCl or NaCl added, so that the system had an overall zero electrostatic charge. The total size of the simulated systems was approximately 56190 atoms for DNA solvated in a water box and 56370 atoms for the solvated protein-DNA complex. Periodic boundary conditions were employed. Long-ranged Coulombic interactions evaluated beyond a cut-off of 13 Å were computed by the particle-mesh Ewald (PME) summation method (78). The nonbonded Lennard-Jones interactions were computed using a smooth cutoff of 13 Å. The integration time step was set to 1 fs.

At first, each simulated system was energy-minimized for 50000 steps using the steepest descent algorithm followed by a second energy minimization for 10000 steps using a quasi-Newtonian algorithm with the low-memory Broyden-Fletcher-Goldfarb-Shanno approach. The tolerance was set to 1.0 kJ mol^{-1} nm^{-1}. After that, the system was heated to 310 K during 4 ps. Then, each system was subjected to 2.0 ns-equilibration in the NVT ensemble with harmonic restraints applied to all protein and DNA heavy atoms. The temperature was kept at 310 K by applying leap-frog stochastic dynamics forces (78) with a damping coefficient of 0.1 ps^{-1}. With restraints kept, each system was further equilibrated for 0.5 ns in the NPT ensemble, and then for another 1.0 ns without restraints.

Conventional MD simulations were performed for different systems. These included the fully methylated DNA (the wild type), the non-methylated DNA, the mutants showing the most significant inhibition (or activation) of the GFP expression in the *in vitro* experiments (namely M2, both in the CpG+ and the CpG- forms), and the Z-DNA form of the wild type. Mutations to the wild type and the Z-DNA form of the BDNF promoter sequence were generated using the 3DNA package (36,107,125) . All of the above simulations were performed in 0.10 M KCl. For the wild type unbound DNA and when bound to the MeCP2 protein, additional simulations were conducted in 0.20M KCl, 0.10M NaCl, 0.20M NaCl salt conditions; or in a mixed environment of 0.26M KCl and 0.15M NaCl (nuclear salt conditions). Simulations were conducted for 100 ns in two replicates each. The results of the two replicates were almost indistinguishable. Thus the results for the second replicate are only shown in one table, but not in the figures. For the computation of RMSF fluctuations, the trajectories of the two replicates (100 ns each) were concatenated and fitted to the initial structure.

The collective variable of handedness was used to describe the dynamics of DNA upon MeCP2 binding. It is a good choice for the detection of the B-Z DNA transition, as it can be used to introduce the helical twisting of the right-handed B-DNA to the left-handed Z-DNA. According to the definition by Moradi et al (126), given a set of basepairs, starting at basepair *n* and ending at basepair *m*; the following sequence of atoms is used to describe the collective variable: P^1_n , P^2_n, P^1_{n+1}, P^2_{n+1},..., P^1_m, P^2_m; where P^1_n is the atom starting from the 5' position in the n^{th} basepair. As such, the sum can also be started starting at the 5' nucleotide triphosphate from the other end. The total collective variable of handedness for the DNA strand is thus the polynomial sum of the handednesss terms, starting at one phosphorus atom each, and ending three bases thereafter (e.g; P^1_n , P^2_n, P^1_{n+1}, P^2_{n+1} + P^2_n, P^1_{n+1}, P^2_{n+1},P^1_{n+2} +...+ P^1_{m-1}, P^2_{m-1},P^1_m, P^2_m). Thus, the position of these atoms defines the handedness via:

$$H(p_1p_2p_3...p_n)=\sum_{k=1}^{n-3} H(p_ip_{i+1}p_{i+2}p_{i+3}) \qquad (4.1)$$

Given a sequence of points *A*, *B*, *C*, and *D*; the single handedness term is defined as (figure 4.1):

$$H(ABCD)=\frac{\overrightarrow{AB}}{|\overrightarrow{AB}|}\times\frac{\overrightarrow{CD}}{|\overrightarrow{CD}|}\cdot\frac{\overrightarrow{EF}}{|\overrightarrow{EF}|} \qquad (4.2)$$

Where the points define the vectors $\overrightarrow{AB}$ and $\overrightarrow{CD}$. The vector $\overrightarrow{EF}$ defines the vector matching the midpoints of $\overrightarrow{AB}$ and $\overrightarrow{CD}$.

To address the global effect on DNA handedness, we included a further DNA sequence, the $(CpG)_6$ dinucleotide repeat run, for which one can expect an amplified effect of the change in handedness during 100 ns of simulations. This sequence was studied in the non-methylated form and in the hyper-methylated form where each base pair consists of a 5-methyl-cytosine. Hyper-methylated DNA was shown to have a high propensity to interconvert between the B-DNA and the Z-DNA forms (96). Both sequence types were simulated in the bound and unbound forms.

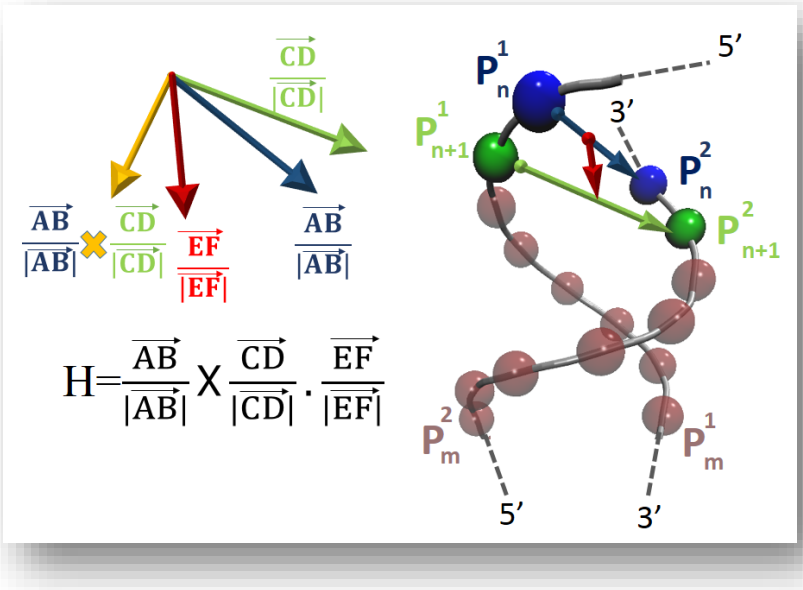


Figure 4.1 A schematic representation for the collective variable of handedness (126). On the right-hand side, the figure shows the vectors involved in this definition. P atoms are named in context of the residue number and the strand, with P^1_n,P^2_n atoms representing the n^{th} basepair P atoms; with 1 and 2 representing the Watson and Crick strands, running in the 5′→3′ direction, and the 3′→5′ direction, respectively. For the handedness term including the four atoms $P^1_n.P^2_n.P^1_{n+1}P^2_{n+1}$, represented in the (ABCD) order here as well, the vectors connecting the atoms and contributing to handedness are defined on the right-hand side, with the blue vector connecting the first two atoms and the green vector connecting the second two atoms. The red vector defines the connection between the two midpoints connecting either vector, in the given direction. On the left-hand side, the definition of handedness is given, in terms of the units vectors of the vectors defined on the right-hand scheme. Vector multiplication, followed by a dot product is assumed, and the final handedness term retrieved. For more details about the definition of the global handedness term in terms of a long run of bases, please refer to the text.

4.3 Results

4.3.1 MBD mediated repression of eGFP with the wt version of the BDNF promoter

We studied MBD mediated repression of eGFP as a function of CpG methylation. Figure 4.2 shows the repression efficiency of MBD upon binding to the wt version of the BDNF promoter. Adding recombined MBD proteins to the reaction led to an almost complete repression of eGFP in the case of CpG methylated reporter plasmids (no MBD: 23 μM of recombined eGFP, 6 μM of MBD: 0.75 μM of recombined eGFP, 97 % repression). MBD bound with high specificity to the methylated BDNF promoter. On the other hand, in the case of the unmethylated BDNF promoter, the eGFP expression was barely modified as a function of MBD concentration (no MBD: 24 μM of recombined eGFP, 6 μM of MBD: 20 μM of recombined eGFP, 20 % repression)(116).

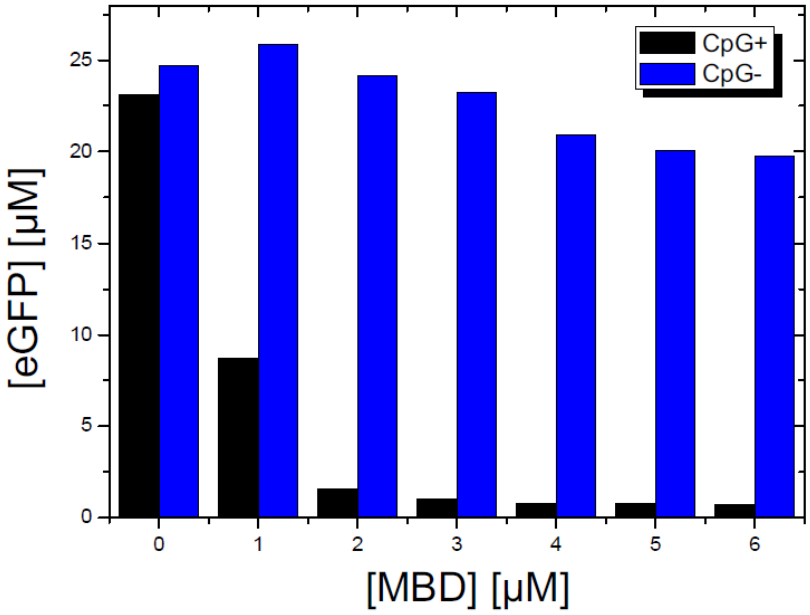


Figure 4.2 Methylation response of MBD mediated repression of eGFP using the wt of the BDNF promoter. The bars refer to the concentration of recombined eGFP as a function of the concentration of MBD protein. CpG+ (black) refers to the fully methylated promoter, CpG- to the unmethylated one. The concentration of the reporter plasmid pBEST-BDNF-UTR1-eGFP-T500 is 5 nM. The concentrations of amino acids, magnesium glutamate and potassium glutamate added to the reaction are 0.5 mM, 3 mM, and 50 mM, respectively. For a MBD concentration of about 3-4 μM, eGFP is almost completely repressed in the CpG+ case, while the expression level is only remotely modified in the CpG- case (116).

MD simulations of the MBD:DNA complex and the unbound DNA revealed stable conformations for methylated and non-methylated DNA (RMSD ~ 2.0 Å from the

average structure). The conformational dynamics of DNA bound to MBD was analyzed in terms of root mean square fluctuations (RMSF) of the heavy atoms in DNA backbone (CpG+ and CpG-) and compared to that of free DNA. Figure 4.3 shows that methylated DNA bound to MeCP2 showed the smallest fluctuations, followed by the two unbound forms and the nonmethylated DNA bound to MeCP2. In the protein:DNA complexes, we observe a reduced BI/BII ratio of the DNA double helix as is often found for protein:DNA complexes (figure 4.4). both in the methylated and non-methylated forms. The density distributions of DNA in the bound and the unbound forms are significantly different (KS-test; D of 0.188, p-value < 2.2e-16 in the methylated DNA; and D of 0.1651 p-value < 2.2e-16 in the unmethylated DNA).

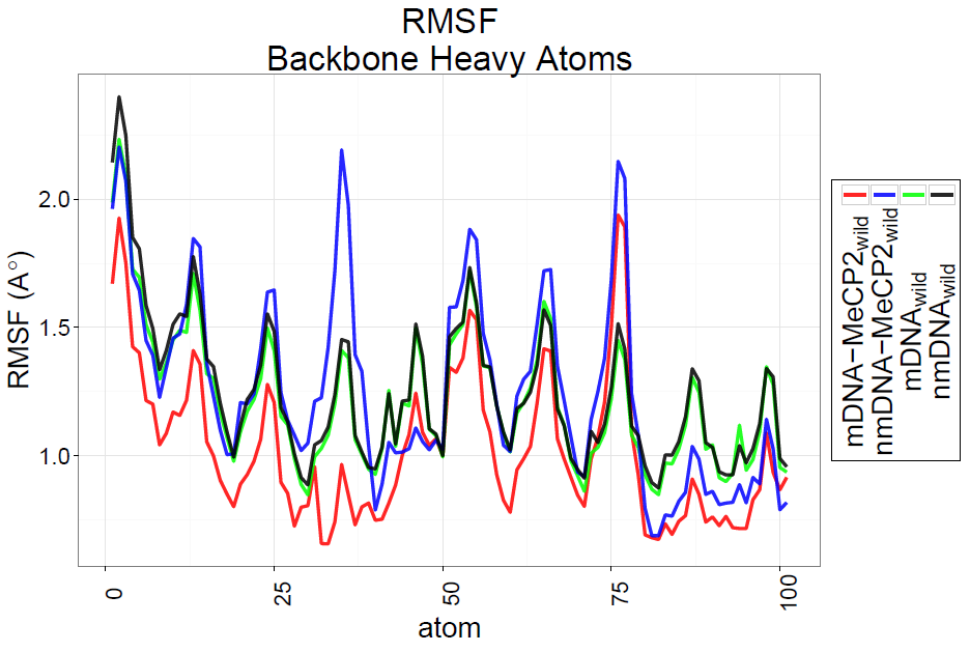
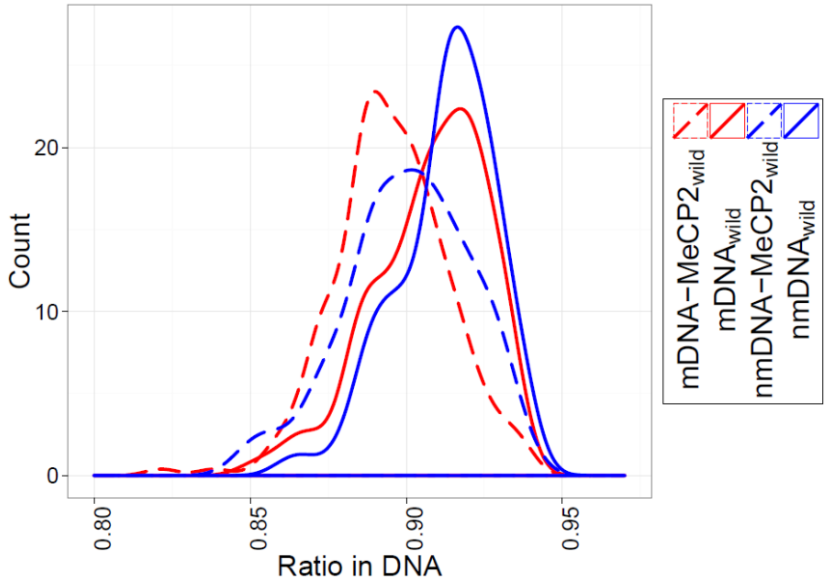


Figure 4.3 Root mean square fluctuations (RMSF) of the heavy atoms in DNA backbone (CpG+ and CpG-), in the free form (green and black, respectively) and after binding to MBD (red and blue, respectively). Trajectories of the two replicates (100 ns each) were concatenated and fitted to the initial structure. RMSF results in the figure show that the DNA atoms are stabilized the most in terms of the methylated complex when compared to the results of the free DNA and the nonmethylated complex. This indicates the significance of the methyl group in stabilizing the bound counterparts in the protein:DNA complex.



	KS-Test D	p-value
0.1M KCl wild type CpG+	0.188	< 2.2e-16
0.1M KCl wild type CpG-	0.165	< 2.2e-16

Figure 4.4 a) Density distribution of the BI ratio in the DNA entities in the wild type (CpG+ and CpG-; bound and unbound) during 100 ns simulation time. b) the table shows the KS test for the difference in the density distribution for the bound vs. the unbound entities.

4.3.2 Sequence mutation of the BDNF promoter

We tested four mutated versions of the BDNF promoter for the transcriptional regulation involving CpG methylation. The sequences of the mutants are listed in table 4.1. M1 and M3 were designed to study the influence of mutating the flanking bases of the central CpG motif. M2 breaks the run of AT bases adjacent to the CpG motif (120). M4 combined the effects for M1, M2 and M3 within one single promoter. The central CpG motif of the BDNF promoter was not changed in all four mutants. For studying the repression efficiency of MBD in the case of mutants M1-M4, we expressed MBD from a plasmid preparation (see table 4.1). This enabled the careful evaluation of the interaction between MBD and the mutated BDNF promoters at low protein concentration. Mutant 2 showed a lowered expression profile compared to the wt version of the BDNF promoter even before binding of MBD (see figure 4.5 and figure 4.2 for the cases w/o the addition of MBD protein/plasmid: CpG+/wt: 23 µM, CpG+/M2: 8 µM, CpG-/wt: 24 µM, CpG-/M2: 17 µM). Additionally, we observed an activator-like behavior of MBD for plasmid concentrations up to 0.5 nM in the case of the CpG methylated M2 promoter. In this concentration range, the eGFP expression level increased as a function of MBD concentration. However, for concentrations between 1 µM and 6 µM of the recombined MBD protein, eGFP was repressed as observed for the unmutated BDNF promoter (see figure 4.6). For all other mutants

M1, M3, and M4 we found almost identical repression results as observed for the wt version of the BDNF promoter (116).

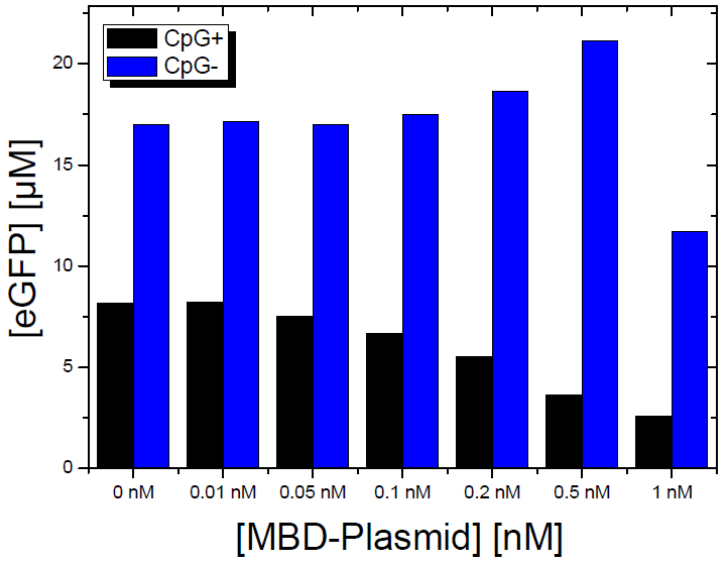


Figure 4.5 Methylation response of MBD mediated repression of eGFP using the mutant 2 of the BDNF promoter. The bars refer to the concentration of recombined eGFP as a function of the concentration of MBD expressed from the plasmid p15a-OR2-OR1-Pr-UTR1-MBD-T500. CpG+ (black) refers to the fully methylated promoter, CpG- to the unmethylated one. The concentration of the reporter plasmid pBEST-BDNF-M2-UTR1-eGFP-T500 is 5 nM. The concentrations of amino acids, magnesium glutamate and potassium glutamate added to the reaction are 0.5 mM, 3 mM, and 50 mM, respectively. MBD activates eGFP expression for plasmid concentrations up to 0.5 nM in the case of the CpG methylated M2 promoter (116)

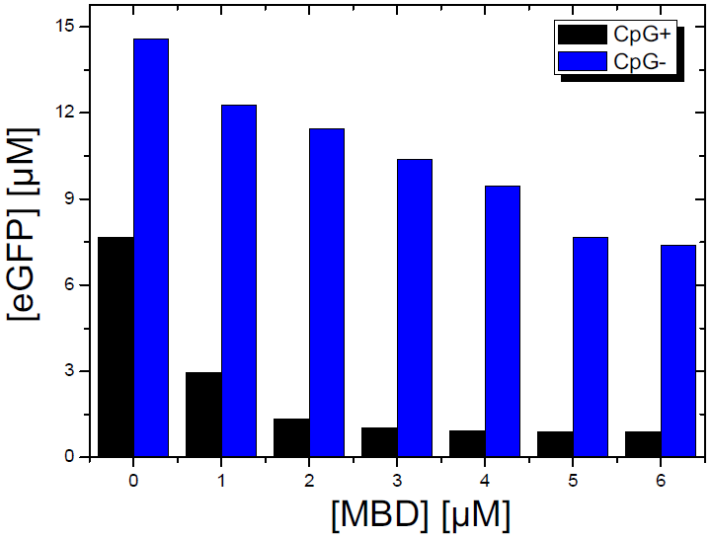


Figure 4.6 done by MS Same experiment as shown in figure 4.5 but for the recombined version of MBD protein. In this case, eGFP is repressed as observed for the unmutated version of the BDNF promoter (figure 4.2). (116).

For wild type and mutant M2, as well as for the wild type DNA under different salt conditions and ionic strengths, MD simulations showed that protein binding induced a widening of the major groove at the binding interface (from 15 Å to about 20 Å; figure 4.7-A). This well-characterized effect of protein binding (127) increased the accessibility of the functional groups of DNA and favors specific protein-DNA contacts. The 0.1M NaCl environments showed the smallest change. On the other hand, protein binding had no large effect on the width of the minor groove of DNA. Most simulations of wild-type CpG+ and CpG- DNA bound to MBD in 0.1M KCl, 0.1M NaCl, 0.2M KCl, 0.2M NaCl showed a decrease in the B-fiber ratio during the first half of the simulation to a stable conformation with 10-20% lower B-fiber ratio (figure 4.8-A; KS test of bound vs. unbound forms for the several complexes:0.15 ≤D≤ 0.35, p-value < 2.2e-16). For the mutant M2 CpG-, the negative shift in the B-form between the unbound and bound states was much smaller than for all other systems (KS test of bound vs. unbound forms for mutant M2 CpG-: D = 0.05, p-value = 7.453e-06).

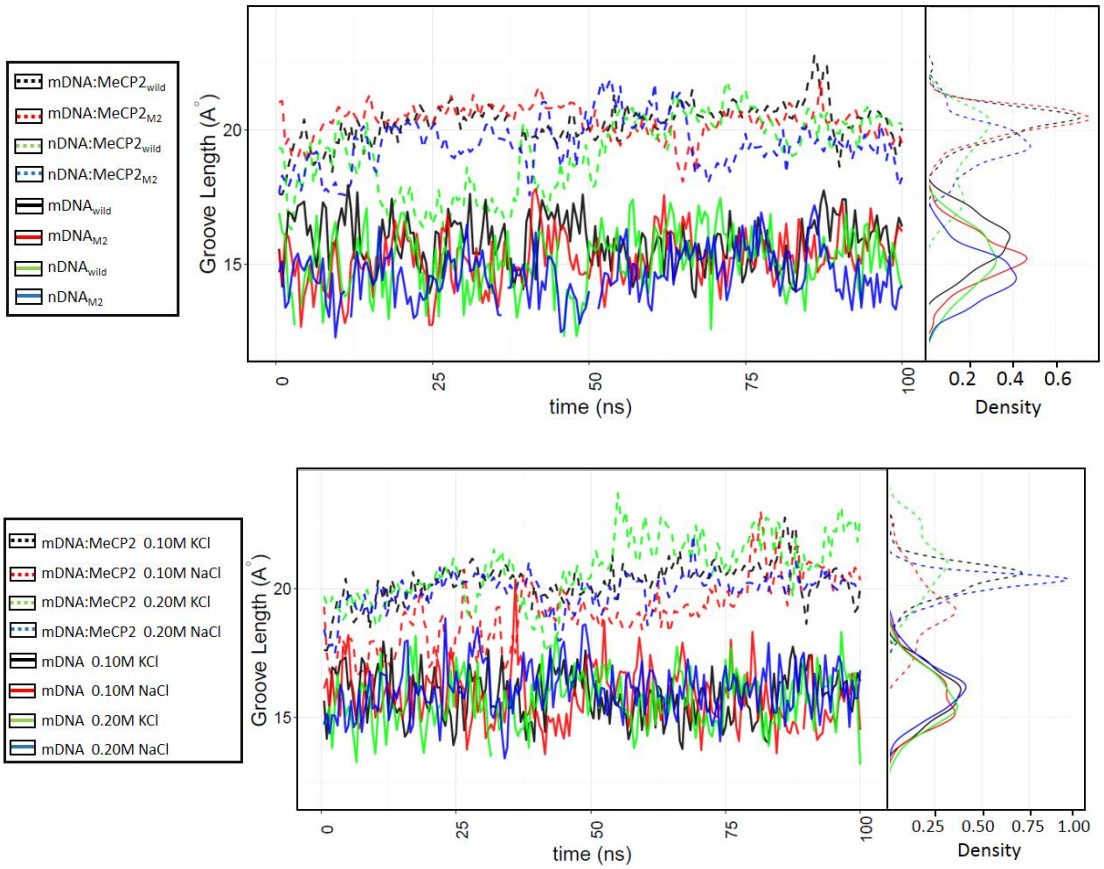


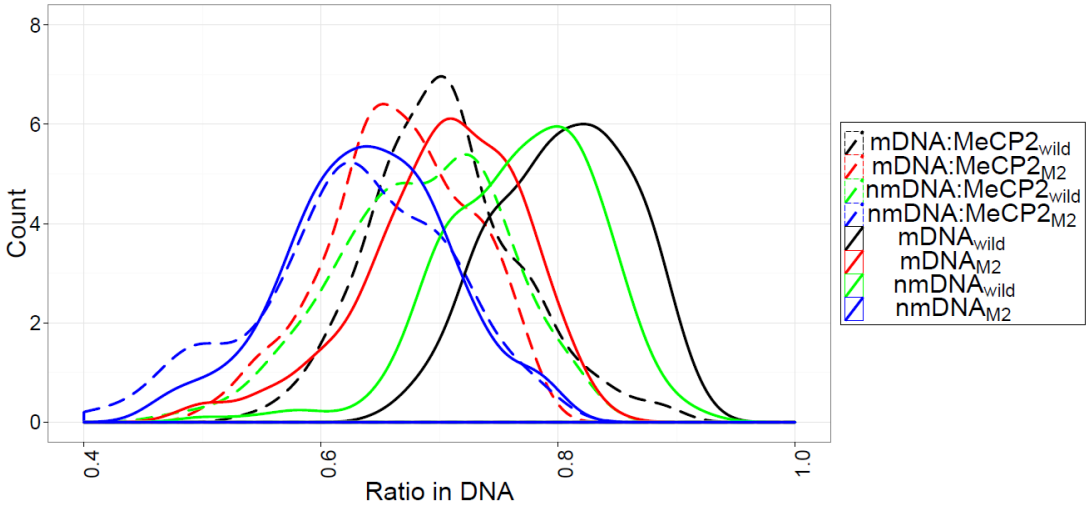
Figure 4.7 The width of the major groove at the binding interface. The width of the major groove at the binding interface was detected for the bound (dashed lines) and the unbound DNA (solid lines). To the left is the timestamp change in groove width during the 100 ns simulation. To the right is the density distribution for the range of the widths during the 100 ns simulation time. Upper figure is for the effect by the several mutants. The lower one is for the salt effect.

The central basepair, namely the 5-methyl-cytosine in the upper strand and the adjacent guanine in the lower strand, showed an untwisting of the helix away from the perfect B-DNA fiber upon binding of the MeCP2 protein (table 2; figure 4.9). This effect was weaker in the surrounding base pairs. The methylated form of the complex showed this shift, to smaller handedness values (see figure 4.9), much more strongly than in the case of the non-methylated wild-type DNA (KS test: D = 0.78, p-value < 2.2e-16; for methylated DNA; and D = 0.23, p-value < 2.2e-16, for the non-methylated DNA). The negative shift in handedness in the case of the mutant M2 (CpG+ and CpG-), was quite similar to the effect found for the wild-type DNA CpG+ (the CpG+ with KS test: D = 0.81, p-value < 2.2e-16; and the CpG- mutants; D = 0.70, p-value < 2.2e-16; respectively; figure 4.9-A). For comparison, changing the sequence content and the methylation status in the unbound form apparently had no effect on the handedness term in the studied salt ranges (figure 4.9-B).

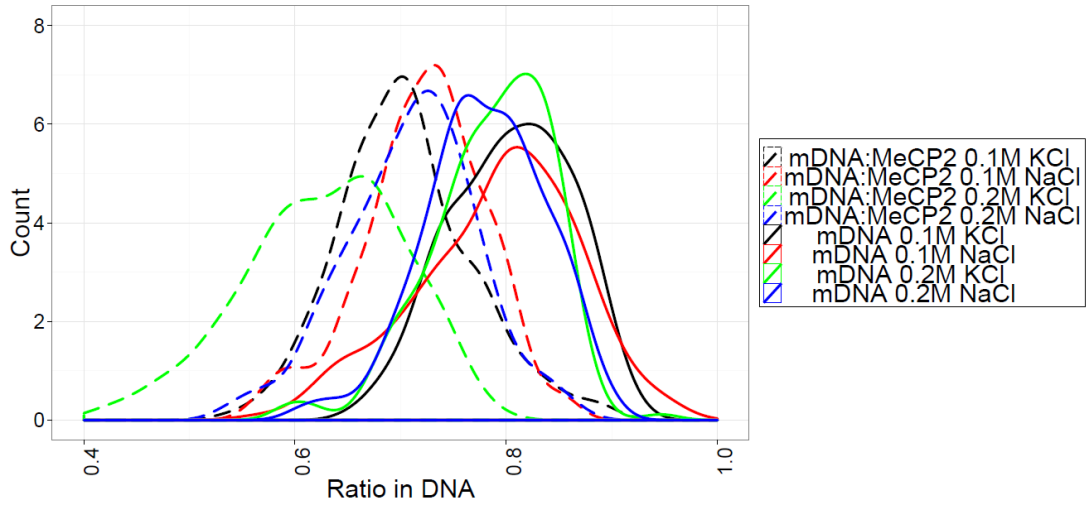
Table 4.2 Mean (standard error) calculations for the handedness terms in the central dinucleotide steps, namely the (36T-7A-35T-85CM) and the (7DA-35DT-85CM-34DG), bound and unbound, under different conditions and in replicates during the 100 ns simulations times.

Handedness in the Central Dinucleotide Steps (100 ns long simulations: Average (SD))				
	Complexes		*DNAs*	
	Central Step 1 (36T-7A-35T-85CM)	***Central Step 2 (7DA-35DT-85CM-34DG)***	***Central Step 1 (36T-7A-35T-85CM)***	***Central Step 2 (7DA-35DT-85CM-34DG)***
Wild Type	0.408 (±0.057)	0.459 (±0.045)	0.524 (±0.043)	0.582 (±0.038)
Methylated	0.441 (±0.067)	0.476 (±0.063)	0.521 (±0.042)	0.580 (±0.037)
Mutant 2	0.402 (±0.052)	0.462 (±0.038)	0.523 (±0.042)	0.579 (±0.041)
Methylated	0.403 (±0.052)	0.462 (±0.038)	0.522 (±0.044)	0.576 (±0.042)
0.1M NaCl	0.454 (±0.069)	0.502 (±0.061)	0.522 (±0.042)	0.584 (±0.038)
Methylated	0.394 (±0.069)	0.435 (±0.048)	0.522 (±0.042)	0.582 (±0.038)
0.2M KCl	0.389 (±0.071)	0.446 (±0.057)	0.5233 (±0.042)	0.580 (±0.039)
Methylated	0.403 (±0.053)	0.443 (±0.047)	0.520 (±0.043)	0.579 (±0.041)
0.2M NaCl	0.395 (±0.058)	0.439 (±0.047)	0.525 (±0.042)	0.584 (±0.038)
Methylated	0.405 (±0.064)	0.458 (±0.046)	0.516 (±0.042)	0.581 (±0.038)
Original (0.1M KCL)	0.492 (±0.055)	0.540 (±0.048)	0.518 (±0.044)	0.575 (±0.042)
non-methylated	0.411 (±0.084)	0.477 (±0.058)	0.523 (±0.047)	0.576 (±0.043)
Mutant 2	0.405 (±0.071)	0.459 (±0.055)	0.526 (±0.046)	0.577 (±0.039)
non-methylated	0.408 (±0.048)	0.451 (±0.038)	0.452 (±0.200)	0.500 (±0.214)
Z-DNA	-0.423 (±0.068)	-0.352 (±0.064)	-0.599 (±0.059)	-0.078 (±0.100)
non-methylated	-0.257 (±0.080)	-0.329 (±0.084)	-0.585 (±0.064)	-0.097 (±0.099)
0.26M KCl	0.401 (±0.064)	0.460 (±0.051)	0.525 (±0.043)	0.585 (±0.036)
0.15M NaCl	0.439 (± 0.064)	0.492 (±0.053)	0.523 (±0.043)	0.580 (±0.040)
$(5mCpG)_6$	0.401 (±0.050)	0.465 (±0.048)	0.519 (±0.046)	0.587 (±0.036)
	0.484 (0.080)	0.540 (0.064)	0.526 (±0.043)	0.592 (±0.030)
$(CpG)_6$	0.476 (±0.090)	0.525 (±0.062)	0.531 (±0.050)	0.585 (±0.036)
	0.375 (±0.063)	0.469 (±0.045)	0.530 (±0.052)	0.585 (±0.038)

(a)



(b)



(c)

	D	p-value
0.1M KCl wild type CpG+	0.3446	< 2.2e-16
0.1M KCl wild type CpG-	0.2776	< 2.2e-16
0.1M KCl M2 CpG+	0.1408	< 2.2e-16
0.1M KCLl M2 CpG-	0.05	$7.45*10^{-6}$
0.2M KCl wild type CpG+	0.3515	< 2.2e-16
0.1M NaCl wild type CpG+	0.2484	< 2.2e-16
0.2M NaCl wild type CpG+	0.2462	< 2.2e-16

Figure 4.8 Density distribution of the B-DNA strand ratio in the DNA entities during the second half of the 100 ns simulations (a) bound and unbound DNA, wild type and mutant M2, CpG+ and CpG-; all in 0.10M KCl. (b) bound and unbound DNA, wild type CpG+; under several ionic strengths (0.1M KCl, 0.1M NaCl, 0.2MKCl, 0.2M NaCl). Table (c) shows the KS test for the difference in the density distribution for the bound vs the unbound entities.

As mentioned before, the binding specificity of MBD to methylated vs. unmethylated DNA was experimentally shown to depend strongly on the ionic strength of the solution (119). In MD simulations of the protein-DNA complexes as well as of unbound DNA, we found that K^+ ion bound preferentially to the electronegative atoms of the bases, whereas Na^+ showed a preferential binding to the phosphorus atoms (see figure 4.10). An increase in salt concentration and/or changing ion type exerted almost no effect on the B-fiber as well as on the handedness (figure 4.9-B) in the unbound form. Importantly, the DNA in the bound form showed varied effects on the B-fiber ranging from mild (in the presence of NaCl, both 0.1M, KS test: D = 0.248, p-value < 2.2e-16, and 0.2M concentrations; D = 0.2462, p-value < 2.2e-16), to moderate perturbations (0.1M KCl; KS test: D = 0.3446, p-value < 2.2e-16; and 0.2M KCl; KS test: D = 0.3515, p-value < 2.2e-16; figure 4.8b).

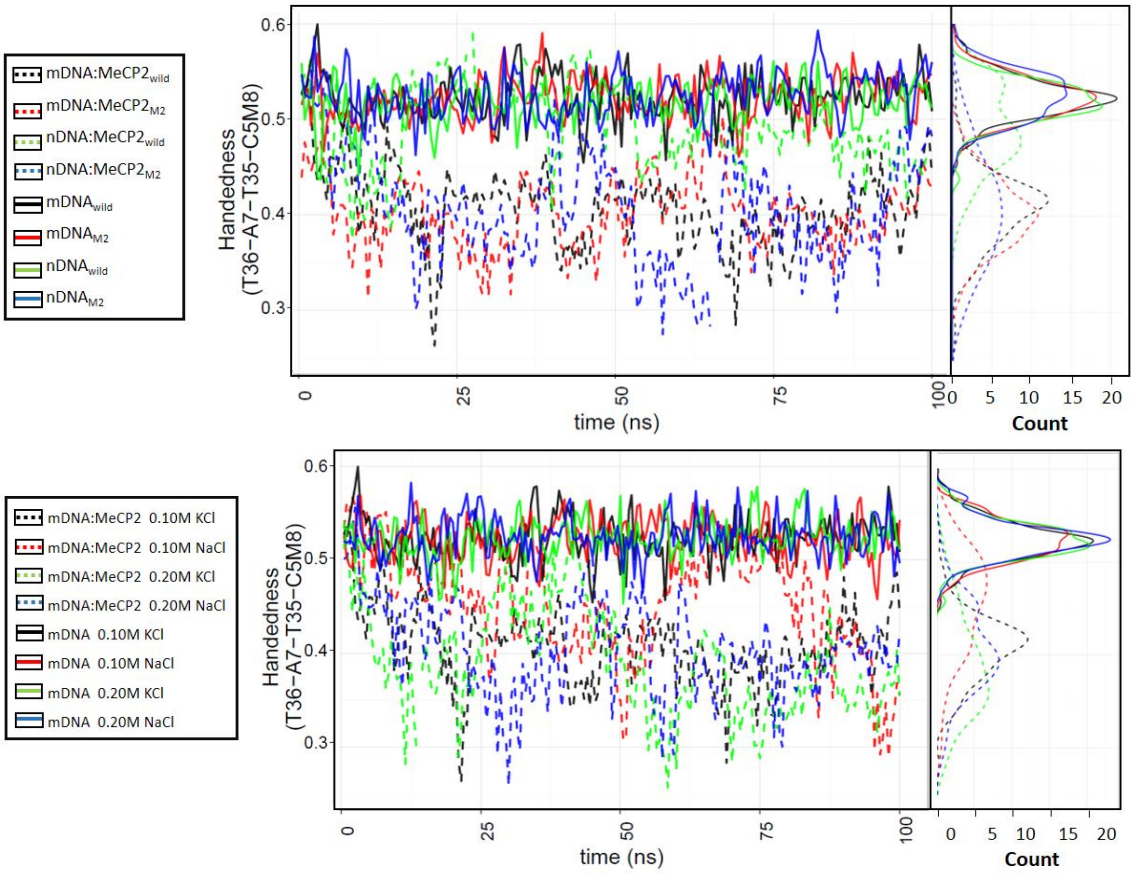


Figure 4.9 The handedness term for (36DT-7DA-35DT-85CM) phosphorus atoms during 100 ns simulations. for the (a) bound and unbound DNA, wild type and mutant M2, CpG+ and CpG-; all in 0.10M KCl; and the (b) bound and unbound DNA, wild type CpG+; under several ionic strengths (0.1M KCl, 0.1M NaCl, 0.2MKCl, 0.2M NaCl). For (a-b), the left panel corresponds timestamp change, while the right panel corresponds to the density distribution during the 100 ns simulations.

Later, we studied the effect of ion type and concentration on handedness. The mDNA:MeCP2 complex in 0.1M KCl solution showed a clear negative drift in the handedness of the target base pair (with respect to the unbound form) with a marginal difference for the higher 0.2M salt concentration (KS test for bound vs unbound: D ~ 0.77-0.79, p-value < 2.2e-16 for either ion concentration). In contrast to this, increasing the ionic strength for NaCl had a profound effect. As for KCl, 0.1M NaCl solution resulted in a negative shift of the bound DNA when compared to the non-bound form. Yet, this effect was further induced upon increasing the salt concentration (KS test: D = 0.46, p-value < 2.2e-16; for 0.1M NaCl; and D = 0.77, p-value < 2.2e-16; for 0.2M NaCl) (figure 4.9-B). As mentioned before for the B-ratio, the salt type and the ionic strength had no effect on the handedness for free DNA in the studied salt ranges (figure 4.9-B).

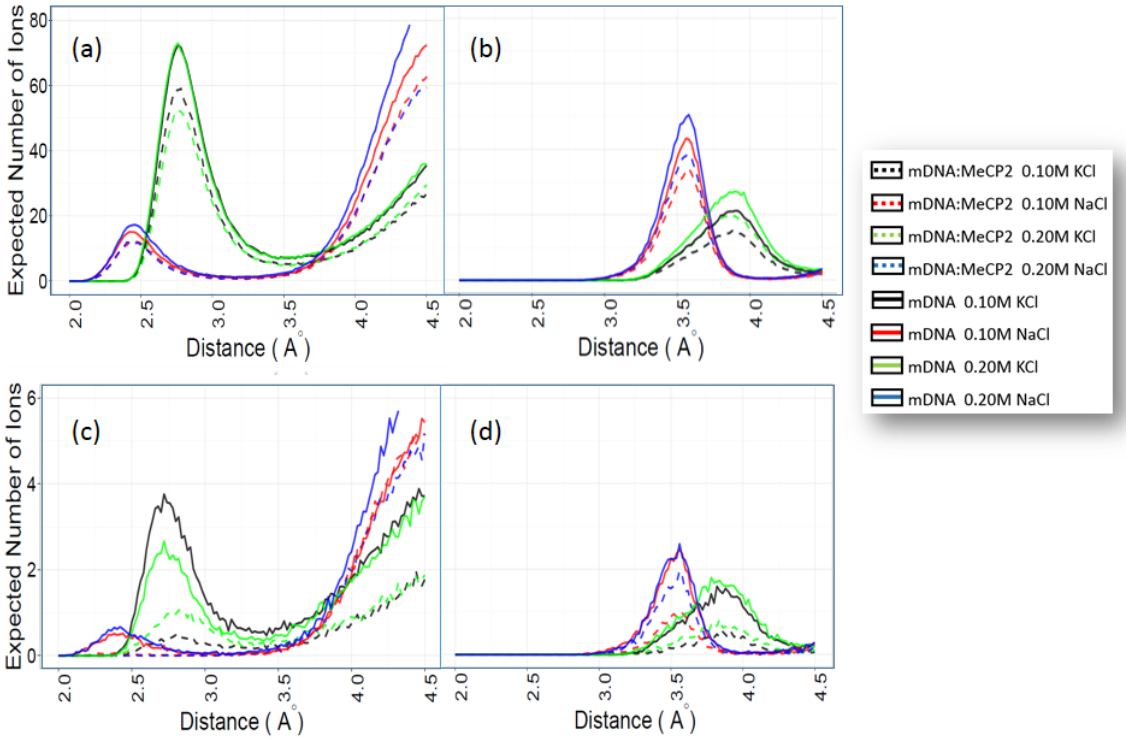


Figure 4.10 Radial distribution function for the distribution of salt cations around the DNA phosphorus atoms and the bases during 100 ns simulation, under different salt conditions (Na+ and K+) and ionic strengths. Figures (a-b) show the density distribution around the surface of the whole DNA strand in terms of (a) the surface of phosphorus atoms, and (b) the surface of the bases. Figures (c-d) show the density distribution around the surface of the central basepair (85CM-9DG), as of (a) the surface of phosphorus atoms in the backbone, and (b) the surface of the bases.

For completeness, the BDNF promoter was replaced by two DNA sequences that are known to easily shift the equilibrium in the forward direction to the Z-form; namely the (CpG)6 dinucleotide repeat and the corresponding $(5m\text{-}CpG)_6$. The handedness

was checked throughout 100 ns simulations for the bound and unbound forms (figure 4.11). No transition was seen in the unbound state for the methylated or the non-methylated DNA. Whereas only a marginal transition (similar to the BDNF promoter) was observed for the unmethylated DNA upon MBD binding (date not shown), the hyper-methylated form of DNA showed a more prominent two-state transition, which spanned several base pairs, (binding interface being step 0) rather than the central dinucleotides.

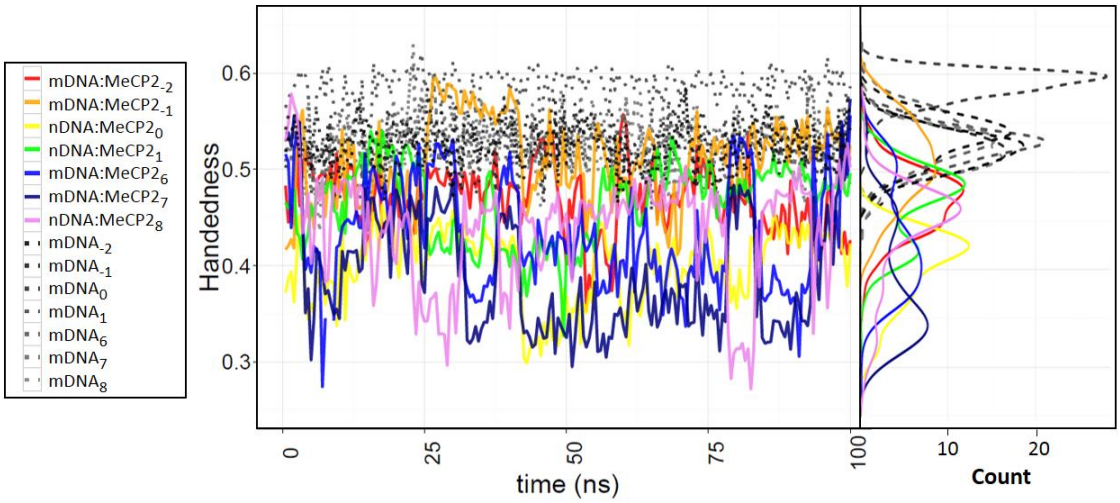


Figure 4.11 Changes in handedness for the $(5mCpG)_6$ in the bound and the unbound forms during 100 ns simulations depicted over several terms of handedness. On the left hand side, the change in handedness is depicted in the binding interface (step 0: here 19DG-6DG-185CM-75CM, step1: 6DG-185CM-75CM-17DG,…) as well as the surrounding handedness terms during 100 ns simulation. On the right-hand side, the density plot is prepared for the whole 100 ns. Colored lines refer to the bound form of DNA whereas the gray-scale lines refer to the unbound form.

4.4 Discussion

In *E. coli*, methylation of the DNA sequence motif 5'-GATC-3' by the enzyme deoxyadenosine methylase (Dam) plays an important role in the timing of initiation of DNA replication, as well as in the coordination of cellular events, DNA mismatch repair, and gene regulation (98). However, gene regulation with MBD of MeCP2 and the BDNF promoter is a purely eukaryotic transcriptional repression system, involving cytosine methylation of the central CpG motif. This sort of methylation is foreign to the *E. coli* strain BL21 used for extract preparation. Consequently, CpG methylated DNA can be attacked by the methylation sensitive restriction system mcr, which acts as a primitive immune system in bacteria (128). However, no such limitations of expression were met in our system due to methylation. This was confirmed by comparing the expression of the methylated vs. non-methylated forms of the BDNF promoter in a control system, where no MBD binding was induced.

We included four mutations into the target DNA sequence and examined them for expression efficiency. Three mutants had a broken run of AT bases adjacent to the CpG motif. This AT run has been reported essential for DNA binding selectivity of MBD (120). The central CpG motif of the BDNF promoter remained unchanged in all three mutants. We designed one further mutant to check for the influence of the local environment. Here, an mCGG stretch is replaced by a mCGC stretch. We note that the mutation effect for the double mutant (M4) shows a lower impact on repression when compared to the mutant 2.

In the MD simulations, MeCP2 binding was found to induce structural transitions of the DNA upon binding, such as the BI/BII equilibrium shift and the increased width of the major groove. Such transitions were induced more strongly by methylated DNA form than by the non-methylated form. For the mutant M2 CpG- the negative shift in the B-form from unbound to bound state was much smaller than for all other systems. This observation may serve as a likely explanation why in the non-methylated form the M2 mutant experimentally induced an activation of protein expression in experiment. This mutant also had the lowest mean ratio of the B-fiber (bound and unbound) when compared to wild type DNA (CpG+ and CpG-) and, to a lesser extent, the mutant M2 GpC+. This means that the relatively lower B-fiber ratio in M2 (CpG+ and CpG-) relative to the lower wild type DNA may also explain why the unbound form of M2 showed much lower expression than wild type DNA. With respect to the handedness of DNA, the strongest sign of untwisting was seen in the (C5-methylated-CpG) stretch that is known to induce structural transitions with a lower energy barrier. This propensity was followed by the mutant M2, the wild type CpG+, and to a lesser extent the wild type CpG-

Whereas the BDNF promoter was crystallized in conditions where the NaCl concentration is close to its physiological value of 0.1 M in the cytosol, both the ion type and the ion concentration are predicted to have an impact on the DNA structure (129). We note that the physiological condition of the nucleus is ~150 mM of NaCl and 260 mM of KCl (130,131) upon studying the salt effect of MBD binding on DNA conformation, the increase in salt concentration and/or changing ion type exerted almost no effect on the B-fiber in the unbound form. This is in perfect agreement with previous simulation studies (129). Radial distribution of the ions around the DNA bases and phosphorus corresponded nicely to previous simulations in the unbound form. On the other hand, MBD binding induced differential changes in the B-DNA fiber and in ion distribution around bases and phosphorus. To the best of our knowledge, no such effect of ionic strength and structure of the cation on the complexed DNA conformation have been reported before (129).

4.5 Conclusion

In this study, we reported the first human epigenetic switch in bacterial cell-free extract. This system provides a simple microenvironment devoid of the complexity of the crowding environment of the biological systems. This allows for an easier monitoring of biological variables crucial for the specific binding of the MBD:mDNA. To this aim, MD simulations revealed structural changes in bound and unbound DNA that could be well associated with our experimental findings on DNA mutants and methylation specificity as well as with ionic strength effect reported by others. In general, we demonstrated that cell-free extracts provide a robust technique for a better understanding of the epigenetic modifications caused by DNA-bound protein.

Chapter 5

Methylation-targeted specificity of the DNA binding proteins R.DpnI and MeCP2 studied by molecular dynamics simulations

Abstract

DNA methylation plays a major role in organismal development and the regulation of gene expression. Methylation of cytosine bases and its cellular roles in eukaryotes are well established, as well as methylation of adenine bases in bacterial genomes. Here, we present results from molecular dynamics simulations, alchemical free energy perturbation, and MM-PBSA calculations to explain the specificity of the R.DpnI enzyme for binding to adenine-methylated DNA in both its catalytic and winged-helix domains. We find that adenine-methylated DNA binds more favorably to the catalytic subunit of R.DpnI (-4 kcal/mol) and to the winged-helix domain (-1.6 kcal/mol) than unmethylated DNA. In particular, N6-adenine methylation is found to enthalpically stabilize binding to R.DpnI. In contrast, C5-cytosine methylation stabilizes binding to the MBD domain of the MeCP2 entropically with almost no difference in binding enthalpy.

5.1 Introduction

DNA methylation plays a major role in a wide variety of biological processes, including the regulation of gene expression and self-recognition. In bacteria, the dominant form is N6-adenine methylation that helps in protecting bacteria against the invasion by foreign DNA (132). The R.DpnI enzyme (see figure 5.1.A) from *Streptococcus pneumonia* is a type IIE restriction endonuclease that consists of an N-terminal catalytic domain and a C-terminal winged helix domain (residues 183-254) (10). R.DpnI protects the Dam-, R.DpnI+ bacteria against phages that have been propagated on Dam+ hosts. Both domains of R.DpnI bind highly specifically to Dam-methylated (Gm6ATC) sites (133). A recent X-ray structure determined by the Bochtler group (134) characterized how the two R.DpnI domains bind to methylated DNA. The authors noticed that the presence of the two methyl groups requires a deviation from B-DNA conformation to avoid steric conflict.

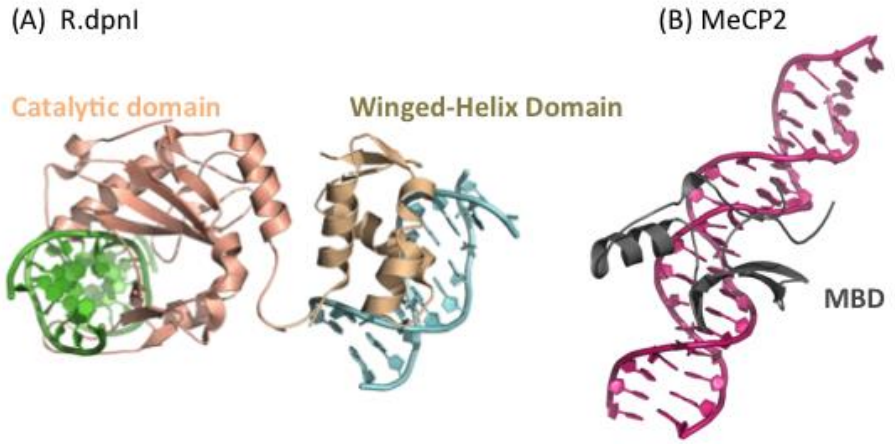


Figure 5.1 Crystal structures of methylated DNA bound to the restriction enzyme R.DpnI from *S. aureus* and human MeCP2. (A) Winged-helix domain and catalytic domain of R.DpnI, each with N6-methylated-adenine containing DNAs bound, are colored cyan and green (RCSB PDB code: 4KYW). (B) Methyl-binding domain of MeCP2 (dark grey color) bound to DNA containing C5-methylated cytosine (hot pink) (RCSB PDB code:3C2I). Figures were generated using PYMOL.

C5-cytosine methylation of DNA is one of the crucial epigenetic modifications of eukaryotic genes and plays an indispensible role in modulating mammalian gene expression. The human protein MeCP2 (see figure 5.1.B) belongs to a highly conserved family of DNA binding proteins that can mediate gene silencing by specific binding to methylated CpG sites and resulting transcriptional repression. MeCP2 possesses a hydrophobic methyl-binding domain (MBD) that directly interacts with the DNA methyl groups, as an example the promoter III of the mouse-brain-derived neurotrophic factor (BDNF; (99). A pioneering X-ray structure of the MBD domain bound to methylated DNA revealed that the methyl group of cytosine surprisingly

contacts a predominantly hydrophilic surface patch on the MBD domain that includes tightly bound water molecules (28).

As mentioned, X-ray crystallography has been instrumental in elucidating structural details of how proteins bind to methylated or non-methylated DNA sequences. When binding to DNA, proteins generally induce an increase of the width of the major DNA groove, so that the functional groups of DNA can access the protein at the binding interface more favorably and in a sequence- and methylation- specific context (127). Moreover, when proteins bind to their target DNA, they change the equilibrium between two alternating conformational states of DNA termed BI and BII that are characterized by different positions of the phosphate groups in the DNA backbone (35),(22). In the BII conformation, the bases are pushed to the major groove of DNA, making them more accessible to the bound protein, and in a specific manner (35). In addition to affecting DNA conformation, methylation of DNA may also enhance specific binding of proteins via solvent contributions or via specific interactions with protein residues (28,135). As it is typically not possible for X-ray crystallography to characterize proteins complexed with both methylated or non-methylated forms of DNA, there is an important need to apply for molecular modeling and biomolecular simulations to unravel the mechanisms behind methylation-specific binding.

In pioneering work, Zou and colleagues conducted conventional MD simulations and alchemical free energy perturbation calculations for the MBD:DNA system involving the CpG binding domain of the C5-methylcytosine binding protein MeCP2 (123). They emphasized the importance of the structural 'stair motif' consisting of mCpG dinucleotide interactions with two MBD arginine residue at the protein-binding interface for methylation-specific binding.

Here, we contrast the contribution of C5-cytosine- vs. N6-adenine- methylation to the specificity of protein binding. To this aim, we studied the R.DpnI enzyme which binds selectively to N6-adenine-methylated DNA and compared it to the MBD:BDNF promoter system that binds selectively to C5-cytosine-methylated DNA (28). We present results from conventional MD simulations, alchemical free energy perturbation and Poisson Boltzmann/ Surface Area (PB/SA) calculations to characterize the conformational changes induced in the DNA upon binding to the protein and to determine the free energy changes and the corresponding enthalpic and entropic contributions for both systems. We performed this study in the methylated or in the non-methylated DNA sequence context.

5.2 Methods

5.2.1 MD simulations

The MD simulations for all systems were performed with the GROMACS 4.5.5 package (103) using the CHARMM27 force field (104) and the TIP3P water model (105). Force field parameters for methylated adenosine were taken from (91).

Each unbound DNA or protein:DNA complex of the two R.DpnI systems (catalytic domain and winged-helix domain) was placed in a cubic water box of 6.1 or 9.9 nm box dimensions with 0.10 mol/l NaCl added. The total size of the simulated R.DpnI systems was around 23000 atoms for the unbound DNA systems and around 95400 atoms for the solvated protein-DNA complexes. The simulations of MeCP2 for the unbound DNA and the bound DNA-complex were conducted using a cubic box of 9.3 nm dimensions, also with 0.10 mol/l NaCl added. The total size of these systems was about 56200-56300 atoms. Periodic boundary conditions were employed. Coulombic interactions were evaluated using a short-range cut-off of 10 Å and long-range interactions were treated by the particle-mesh Ewald (PME) summation method (78). The non-bonded Lennard-Jones interactions were computed using a smooth cutoff of 10 Å. The integration time step was set to 1 fs. The temperature was kept at 310 K by applying leap-frog stochastic dynamics forces with a damping coefficient of 0.1 ps^{-1} (108).

At first, each simulated system was energy-minimized for 50000 steps using the steepest descent algorithm followed by a second energy minimization for 10000 steps using a quasi-Newtonian algorithm with the low-memory Broyden-Fletcher-Goldfarb-Shanno approach. The tolerance was set to 1.0 kJ mol^{-1} nm^{-1}. After that, the system was heated to 310 K during 4 ps. Then, each system was subjected to 1 ns-equilibration in the NVT ensemble with harmonic restraints applied to all protein and DNA atoms using a force constant of 1000 kJ mol^{-1} nm^{-2}. With restraints kept, each system was further equilibrated for 500 ps in the NPT ensemble, and then for another 500 ps without restraints.

The *apo* form of R.DpnI-binding DNA was simulated in the free form started from an ideal B-DNA conformation (methylated, nonmethylated and hemimethylated DNA). For methylated and nonmethylated DNA, we collected 500 ns of simulations (2 replicates with 100 ns each, plus 10 shorter replicates with 30 ns each). Simulations of the hemi-methylated forms were conducted for 40 ns each.

Simulations of the R.DpnI:DNA complexes were based on the recent crystal structure of R. DpnI with two strands of N6-adenine-methylated DNA bound to the catalytic and the winged-helix domains (134). In the simulations, we simulated the intact R.DpnI protein with DNA either bound to the catalytic or to the winged-helix domain. DNA in either domain was further mutated into the non-methylated and hemi-methylated DNA sequences (both in the sequence context of the proximal and distal methyl groups) by replacing the respective methyl groups by hydrogen atoms. For the fully methylated and nonmethylated DNA bound to either domain, we collected 500 ns of simulations (2 replicates, 100 ns each plus 10 replicates with 30ns each). Simulations of the two forms of hemimethylated DNA were conducted for 40 ns, each.

As structural reference for the simulations of the MeCP2:DNA complex, we used the X-ray structure of the BDNF promoter bound to the human methyl-binding domain (RCSB:3C2I; (28). For this system, two replicate simulations were conducted for 100 ns each for free DNA in the (a) methylated and (b) non-methylated forms, and for bound DNA in the (c) methylated and (d) non-methylated forms.

5.2.2 Free Energy Perturbation

Alchemical FEP calculations using the Bennett acceptance ratio with error bars (BAR) were employed to determine the difference in binding free energy of the R.DpnI–DNA complex upon de-methylating 5-adenosine in both DNA strands (75,109). Both in unbound DNA and in the protein:DNA complex, the two methyl groups attached to adenosine were annihilated in two stages (110), see figure 5.2. In the first stage (corresponding to the transition from panel 1→2), the electrostatic interactions of each methyl group were switched off in a step-wise manner and its respective charge was assigned to the N6 atom. For this, the system Hamiltonian was coupled to a coupling parameter λ wherey λ = 0 corresponds to the reference state and λ = 1 to the perturbed state. No soft-core potential was used in this step. In the second stage (2→3), the atoms of the methyl group were turned into dummy atoms (by switching their epsilon and sigma Lennard-Jones parameters to zero). In this stage, a soft core potential was used where soft-core alpha was set to 0.5, the soft-core power to 1.0, and soft-core sigma to 0.3 (111). To complete the free energy cycle, the 'dummy' hydrogen atom of the non-methylated adenosine was turned into an interacting hydrogen by switching on its Lennard-Jones interactions (stage 3; 4→5), and then the electrostatic interactions (stage 4; 5→6). Each of the four stages was decomposed into 26 intermediates states (Δλ = 0.04). As for the reference state at λ = 0, the simulation of each intermediate state started with a double energy minimization, followed by equilibration over 1 ns in the NVT ensemble and 500 ps equilibration in the NPT ensemble with harmonic restraints, and 500 ps without any restraints. Data was

collected during another 1.5 ns for each window. This yields a total simulation time of 26 x 3.5 ns = 91 ns for each unidirectional simulation.

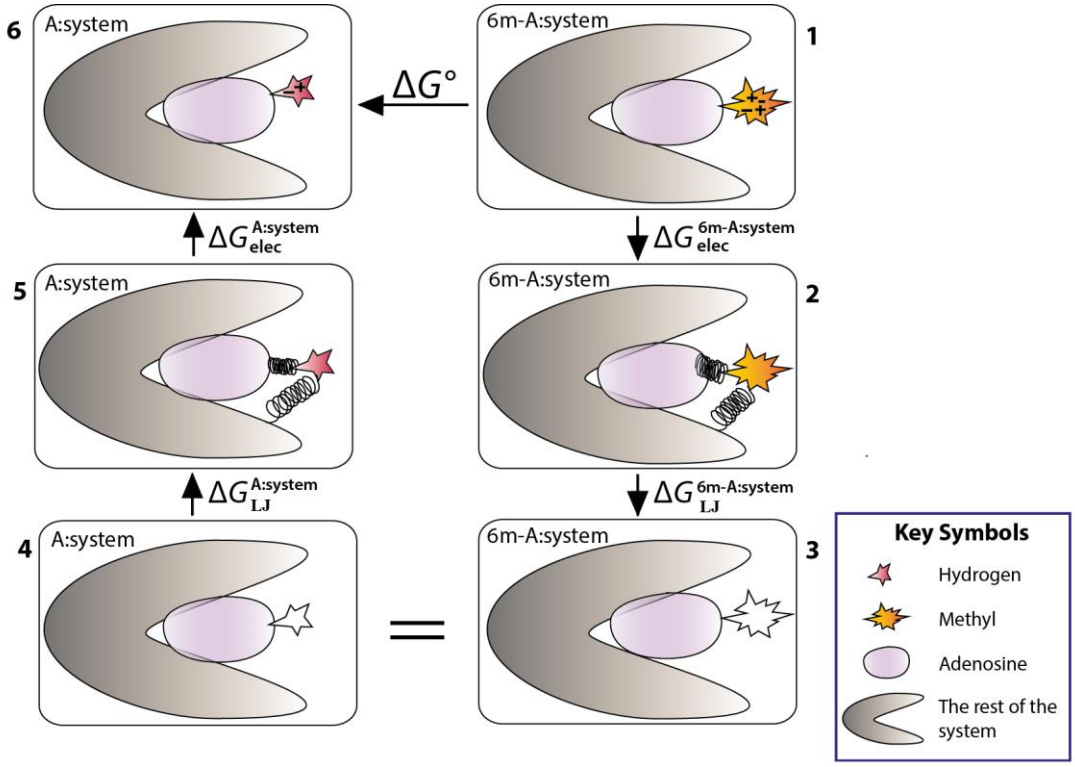


Figure 5.2 The free energy cycle describing the perturbation of a single methyl group (1) in the system, and replacing it by a hydrogen; first through discharging it (2), followed by switching off its LJ interactions (3); and mutation of the non-interacting dummy atoms (4); LJ interactions of the corresponding hydrogen are turned on (5), finally the Coulombic interactions of the hydrogen are turned on (6).

5.2.3 MM-PBSA energy calculations

In the MM-PBSA approach (136), the enthalpic contributions to the free energy of binding are calculated via:

$$H = E_{bonded} + E_{vdW} + E_{elec} + E_{PB} + E_{SA} \quad (5.1)$$

where *bonded* stands for the bonded energy terms (bond lengths, bond angles and torsion angles), *vdW* stands for the van der Waals interactions, and *elec* for the Coulombic interactions. The three terms represent altogether the molecular mechanics terms and are computed in the gas phase. *SA* refers to the surface area contribution. PB stands for the solvation free energy computed here with the Adaptive Poisson Boltzmann solver (137). The net change in enthalpy upon

methylation can be calculated for a single biological entity as follows (with protein contribution $H_{protein}$ cancelling out):

$$\Delta\Delta H = \Delta H_{met.complex} - \Delta H_{nm.complex} = H_{met.complex} - (H_{protein} + H_{met.DNA}) - (H_{nm.complex} - (H_{protein} + H_{nm.DNA})) \quad (5.2a)$$

Here, $H_{met.complex}$ belongs to the complex of methylated DNA and protein, $H_{protein}$ to the unbound protein, and $H_{met.DNA}$ (or $H_{nm.DNA}$) to unbound methylated (or non-methylated) DNA. Since $H_{protein}$ cancels out, this simplifies into:

$$\Delta\Delta H = H_{met.complex} - H_{nm.complex} - (H_{met.DNA} + H_{nmDNA}) \quad (5.2b)$$

We note that this sum is not purely a sum of enthalpic terms, since the PB/SA terms are parameterized as solvation free energies. Enthalpy decomposition was performed using the amber MMPBSA.py tool (138). The original set of MD simulations was performed using GROMACS (as explained before in the MD simulation section), the trajectories of several replicates were merged, and the snapshots were superimposed on the starting structure. Then we used the package AMBERTOOLS to compute the contributions of the individual components to the free energy of binding. For consistency, the CHARMM27 force field was also used in the MM-PBSA calculations. For this, the topology files were generated with the CHAMBER package available in AMBERTOOLS, and applied to the snapshots of the GROMACS MD simulations generated with the same CHARMM27 force field (139). Protein structure files (PSF) in CHARMM format were provided as an input to the CHAMBER package. 10000 fitted PDB snapshots per simulation type were fed as an input to the CPPTRAJ utility in AMBERTOOLS, in order to generate the MDCRD trajectories required for the Amber package. Both structure file types (PDB and PSF) were processed with the PSFGEN plugin in VMD (140,141). Patches and parameters for the nonstandard residue 6MA were added to the input topology and parameter files.

5.2.4 Configurational entropy of DNA and protein

Characterizing the configurational entropy of flexible solute molecules as studied here by molecular simulations is known to require lengthy MD simulations. Thus, we performed simulations in several replicates for the *apo* and the *holo* forms of DNA started from various starting configurations, and then merged the sampled conformations to achieve better and faster convergence of the configurational entropy. For the B-DNA form of all unbound systems, as well as for the crystal forms of the DNA bound to the winged-helix domain and to the catalytic domain, 10 short simulations (30 ns each) were merged with 2 longer simulations of 100 ns each. To

check for the convergence, we followed the strategy introduced by Domene and co-workers (142). Snapshots were collected each 5 ps. For the MeCP2 system, we used snapshots from two replicate simulations of 100 ns length each.

The configurational entropy of the complexes, protein and DNA in the methylated and non-methylated forms was quantified by the approach of Schlitter (82). Entropy differences due to methylation were computed as follows:

$$\Delta\Delta S=\Delta S_{met.complex} -\Delta S_{nm.complex} = S_{met.complex}-S_{nm.complex} - (S_{protein}+S_{met.DNA} - (S_{protein}+S_{nmDNA})) \quad (5.3a)$$

$$\Delta\Delta S=\Delta S_{met.complex} -\Delta S_{nm.complex} = S_{met.complex}-S_{nm.complex} - (S_{met.DNA} - S_{nmDNA}) \quad (5.3b)$$

These contributions were computed for the individual proteins and DNA as well as for the formed complexes.

5.3 Results

In this study, we present results from molecular dynamics simulations to determine structural and energetic characteristics that mediate the specific binding of the bacterial R.DpnI and mammalian MeCP2 proteins to DNA strands carrying either methylated or non-methylated adenine or cytosine bases. For the three studied systems, we found that the DNA structure was well maintained in all simulations with an upper RMSD of 3 Å and an average RMSD of 1.5 Å during MD simulations of 100 ns duration.

5.3.1 Structural adaptation of DNA upon binding

The width of the major groove of DNA is known to have an important effect on the specificity of protein:DNA binding. . Figure 5.3A shows the major groove width in the dinucleotide step at the binding interface for the R.DpnI simulations (both in the complexes and in the unbound DNA); figure 5.3B that for the MeCP2 simulations. Both unmethylated DNA strains (solid blue lines in 2A and 2B) gave a clear peak around 15.5 Å. The major groove width with methylated cytosine is a bit narrower (16.0 Å, solid red, 2B) than with the methylated adenosine (17.0 Å; solid red- 2A). The protein-DNA complexes always showed an opening of the major groove. This tendency was stronger in the winged helix domain of R.DpnI (22.0 Å) than in the catalytic domain (19.0 Å). Methylated DNA bound to proteins generally gave narrower-peaked distance profiles compared to the more 'floppy' non-methylated form. MeCP2-bound

DNA also showed a very clear opening transition. We did not observe contacts between prote-in and the DNA minor groove.

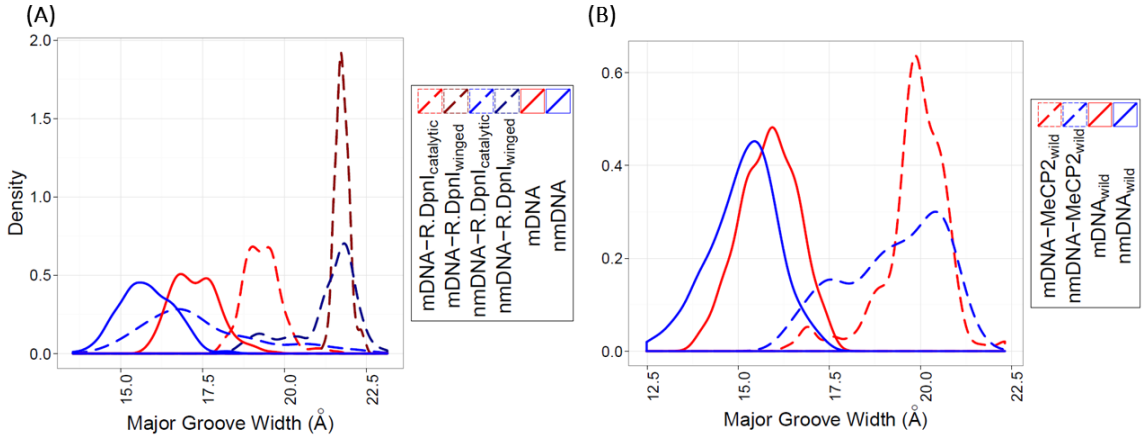


Figure 5.3 Width of the major groove in (A) an ensemble from 500 ns of simulations (merged trajectories) of the R.DpnI system; methylated and non-methylated; and (B) in an ensemble from 200 ns of simulations (merged trajectories) for the MeCP2 system. For this figure, figures 5.4 and 5.11, the same ensembles were used to compute the frequency distributions.

Next, we checked the BI ratio of the DNA strands (see figure 5.4). In the BI conformation, the difference between the two torsion angles ε (C4'-C3'-03'-P) and ζ (C3'-03'-P-05') is about -90º, and is about +90º in the BII form (34). Importantly, the phosphate position in the BI conformation is symmetric with respect to the minor and the major grooves, whereas the BII conformation shifts the phosphates to the minor groove. We found that the unmethylated R.DpnI-binding sequence adopted a slightly smaller BI ratio in solution (0.88) than the MeCP2-binding sequence (0.92). In both cases, methylation of adenosine or cytosine hardly induced any changes in solution. In the complexes, the BI/BII equilibrium was shifted to smaller values indicating a higher proportion of the BII conformation with better accessible nucleic bases (see introduction). Upon binding to MeCP2, the change is rather small (0.02-0.03). Upon binding to R.DpnI, bimodal distributions were observed with peaks near 0.88 and near 0.84. When complexed either to the catalytic or to the winged-helix domains, N6-methylated adenine increased the occupancy of the 0.84 peak.

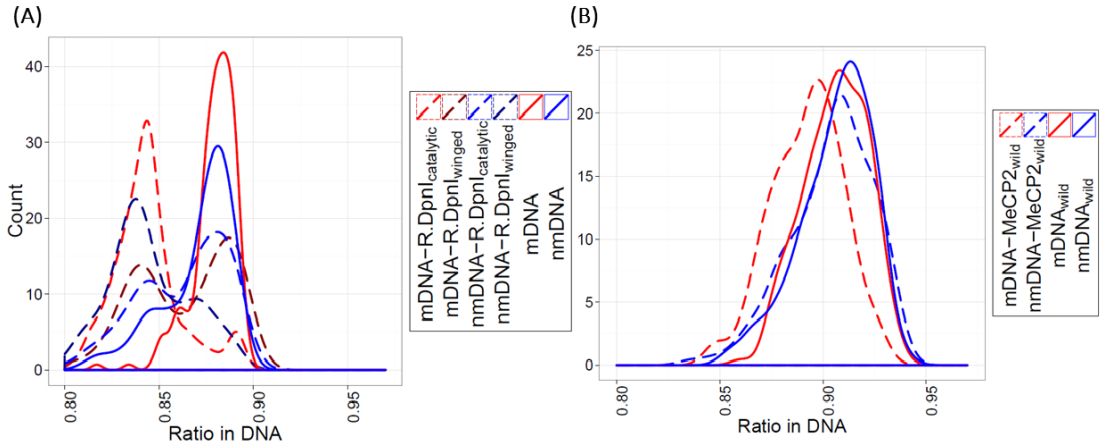


Figure 5.4 BI ratio in (A) the bound and unbound forms of DNA in the R.DpnI systems and in (B) the bound and unbound forms of DNA in the MBD:DNA system.

Investigating the DNA structure in the bound and unbound conformations (Table 5.1) in terms of the well-known basepair steps (rise, roll, shift, slide, tilt, and twist) revealed a certain structural strain in the DNA bound to the catalytic domain of R.DpnI and when bound to the MBD protein. These deformations were smaller in complexes with the winged-helix domain of R.DpnI.

Table 5.1 Basepair step parameters for the whole DNA. Statistically significant changes between the bound and the unbound forms are marked in bold (t-test p-value<0.05)

(A) Bound	Methylation status	Rise (Å)	Roll (°)	Shift (Å)	Slide (Å)	Tilt (°)	Twist (°)
R. dpnI$_{cat}$	Yes	3.45 (0.15)	**4.8 (2.6)**	**-0.10 (0.25)**	**-0.20 (0.6)**	**-0.9 (2.3)**	**31.2 (2.5)**
	No	3.45 (0.2)	**3.1 (2.5)**	**0.18 (0.18)**	**-0.05 (0.5)**	0.2 (1.9)	**32.8 (3.0)**
R.dpnI$_{winged}$	Yes	3.42 (0.05)	**2.8 (2.0)**	0.01 (0.12)	**-0.3 (0.25)**	-0.1 (1.1)	33.1 (1.2)
	No	3.42 (0.15)	**2.8 (1.9)**	-0.1 (0.18)	-0.25 (0.25)	**-0.6 (1.5)**	**32.8 (2.8)**
MeCP2	Yes	3.36 (0.05)	**5.0 (1.5)**	0.05 (0.08)	**-0.1 (0.25)**	0.25 (0.75)	34.4 (2.1)
	No	3.35 (0.05)	5.3 (1.5)	0.05 (0.08)	**-0.1 (0.25)**	0.25 (0.75)	34.3 (2.0)

(B) Unbound	Methylation state	Rise (Å)	Roll (°)	Shift (Å)	Slide (Å)	Tilt (°)	Twist (°)
R. dpnI$_{B\text{-}DNA}$	Yes	3.34 (0.05)	5.7 (2.1)	0.01 (0.12)	-0.13 (0.3)	0.0 (1.1)	33.8 (1.2)

	No	3.38 (0.05)	5.8 (2.1)	-0.01 (0.12)	-0.25 (0.13)	0.0 (1.1)	33.9 (1.3)
$MeCP2_{B\text{-}DNA}$	Yes	3.33 (0.03)	5.5 (1.5)	0.05 (0.07)	-0.02 (0.15)	0.25 (0.75)	34.8 (1.2)
	No	3.32 (0.03)	5.5 (1.5)	0.05 (0.07)	-0.02 (0.15)	0.25 (0.75)	34.8 (1.3)

5.3.2 Methyl-methyl proximity effects on methylation specific binding

DNA is fully methylated in the crystal structure of the R.DpnI enzyme so that the two N6-adenine methyl groups are tightly packed against each other. According to figure 5.5, the methyl-methyl distance is only slightly larger than the sum of their van der Waals radii (2.0 Å + 2.0 Å = 4.0 Å; (143) in the X-ray structure of the catalytic domain of R.DpnI and in a methylated perfect B-DNA fiber, generated for the GATC sequence with the 3DNA program and default parameters (36). When bound to the winged-helix domain, this distance is about 4.5 Å in the X-ray structure. We reasoned that interactions between the methyl groups could on one hand promote a 'shielding effect' of the hydrophobic methyl group (see below). On the other hand, close contacts could reduce the conformational flexibility of DNA in solution, and thus lower its conformational entropy (134). To find out whether this is the case, we computationally grafted trans methyl groups on the N6 atoms of the adenines in the A:T dinucleotide steps in the MD snapshots (fitted to the conformation of the methylated form of the crystal DNA). For unbound DNA simulated in the non-methylated form, grafting showed that the methyl groups were indeed too close (3.4 Å). However, grafts applied to MD snapshots of nonmethylated DNA bound to the winged-helix domain of DNA gave fluctuations around an average distance of 4.0 Å, what is an acceptable value (see above). For the nonmethylated DNA bound to the catalytic domain, however, this was not the case. Here, the grafts applied to DNA showed a minimum distance of 3.0 Å and equilibrated around this value. For the MBD:meDNA complex, the distance between the two methyl groups in the C5-methylcytosines in the methylated complex (8.0 Å) was much larger than the sum of the van der Waals radii. Therefore, there appears no risk of entropic restraints in terms of the binding specificity of 5mC.

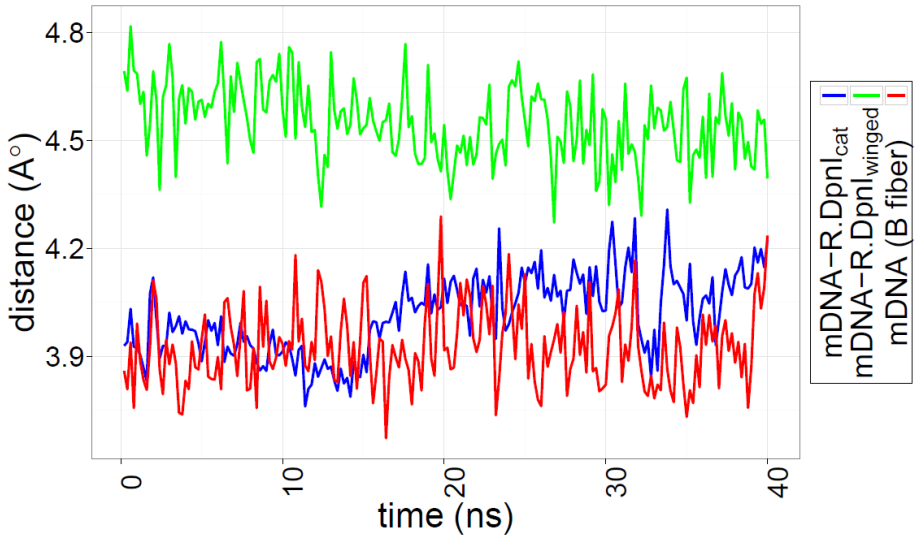


Figure 5.5 Carbon-carbon distance of the methyl groups in the N6-methyladenine of DNA in the free form and in complex with R.DpnI catalytic or winged helix domain during the first 40 ns of the MD simulations.

N6-methyladenosine can adopt both the *"cis"* and the *"trans"* isomers. Here, only the planar *'trans'* conformation was observed in the MD simulation runs for the fully and hemi–methylated Gm6ATC target sequence (figure 5.6). When complexed to R.DpnI, water molecules around the N6 adenine atoms were displaced, both in the case of methylated and the unmethylated DNA (figure 5.7). Water displacement was more pronounced for the winged-helix domain than for the catalytic domain. In contrast, figure 5.8 shows that very few water molecules were displaced with respect to the C5-C5M plane connecting the C5-methylated cytosine.

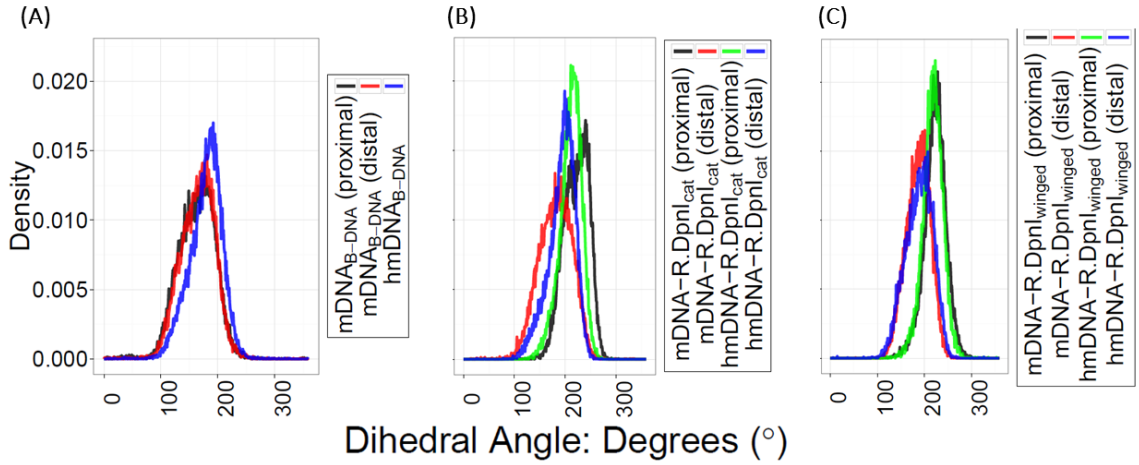


Figure 5.6 Density (frequency) distribution of the torsion angle defined by atoms C1–N6–C6–N1 during the 40 ns long MD simulations of methylated/hemimethylated DNA bound to R.DpnI: (A) unbound and (B-C) bound states (catalytic and winged helix domains, respectively). The 180° angle corresponds to the *trans* conformation of the methyl group.

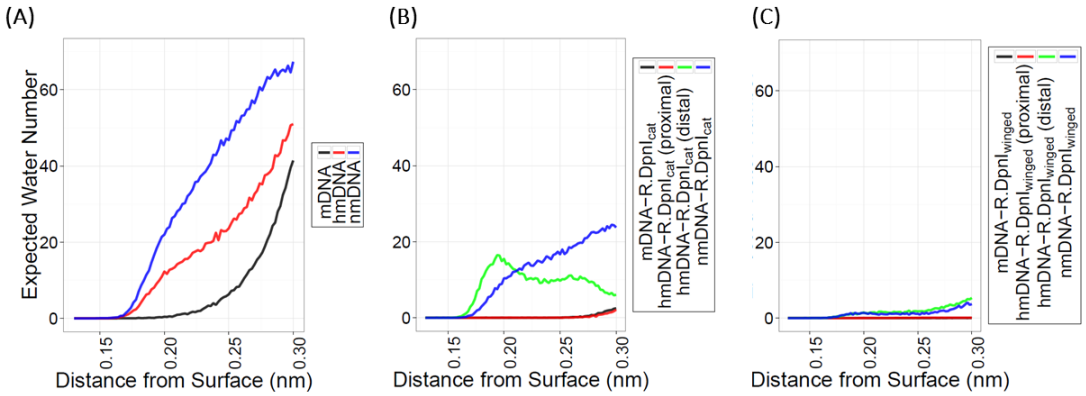


Figure 5.7 Radial distribution plot of water molecules around the hypothetical surface of the two N6-amino/methylamino groups in DNA: (A-C) the expected value (EV) for the distribution of the number of water molecules within 0.13 to 0.3 nm distance from the surface during the 100 ns long MD simulations of methylated, hemimethylated and non-methylated DNA:R.DpnI complexes (B-C); as well as the respective unbound DNAs (A).

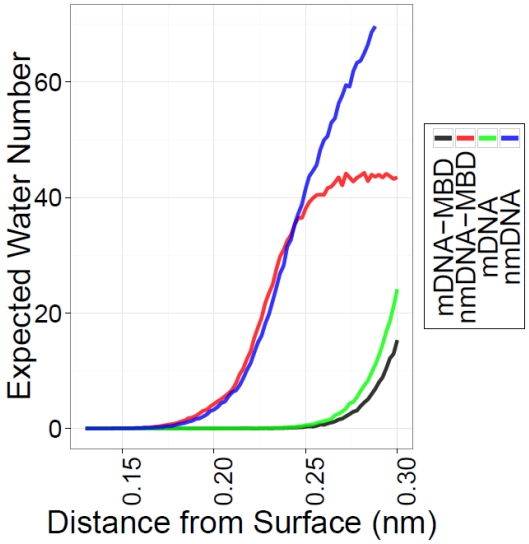


Figure 5.8 The expected value for the distribution of the number of water molecules within 0.13 to 0.3 nm distance from the surface during the 100 ns long MD simulations of methylated and non-methylated DNA:MBD complexes and DNA.

5.3.3 Methyl groups energetically stabilize complexes of DNA with the catalytic and winged-helix domains of R.DpnI

The $\Delta\Delta G$ contribution of DNA methylation to R.DpnI binding was quantified using alchemical free energy perturbation calculations and a thermodynamic cycle (figure 5.9). For this, we transformed the methyl groups on the proximal and distal strands of fully methylated DNA, or the respective hydrogen atoms in the non-methylated form of DNA from a fully interacting state to a dummy state (figure 5.2). This was done in solution as well as in the complexes with the catalytic and the winged-helix domains of R.DpnI.

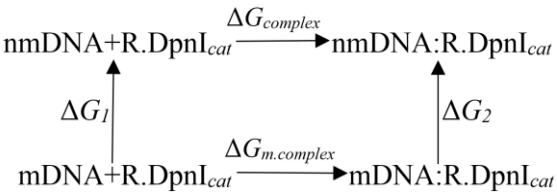


Figure 5.9 Free energy cycle of the 'double' perturbation process, where the process described in figure 5.2 is conducted in two different systems (bound and unbound DNA) to compute the contribution of methylation to protein binding to the modified DNA.

During the free energy calculations for the DNA:R.DpnI complex with DNA either bound to the catalytic or the winged-helix domains, the protein conformation was well maintained. During all stages and windows, the RMSD values between final and starting conformations were between 0.2 and 0.4 nm, which are typical values in MD simulations of proteins. In addition, DNA conformations were well preserved throughout all simulation windows of the discharging step. However, in the last windows of the stages where the LJ potential was turned off (between λ = 0.96 and 1.00, during both the perturbations of the methyl species as well as the replacing hydrogens), the DNA unwound and dissociated from the complex. As the structural transition happened both in the presence and absence of protein, and as most of the free energy change occurs for initial values of λ ≤ 0.8 (see figure 5.10), the computed free energy differences should be almost unaffected. This unexpected computational result is consistent with the biochemically observed low affinity of unmethylated DNA to the catalytic domain (10). Figure 5.10 shows that the cumulative changes in free energy as a function of λ, during the van der Waals elimination step and the discharging step, were quite smooth.

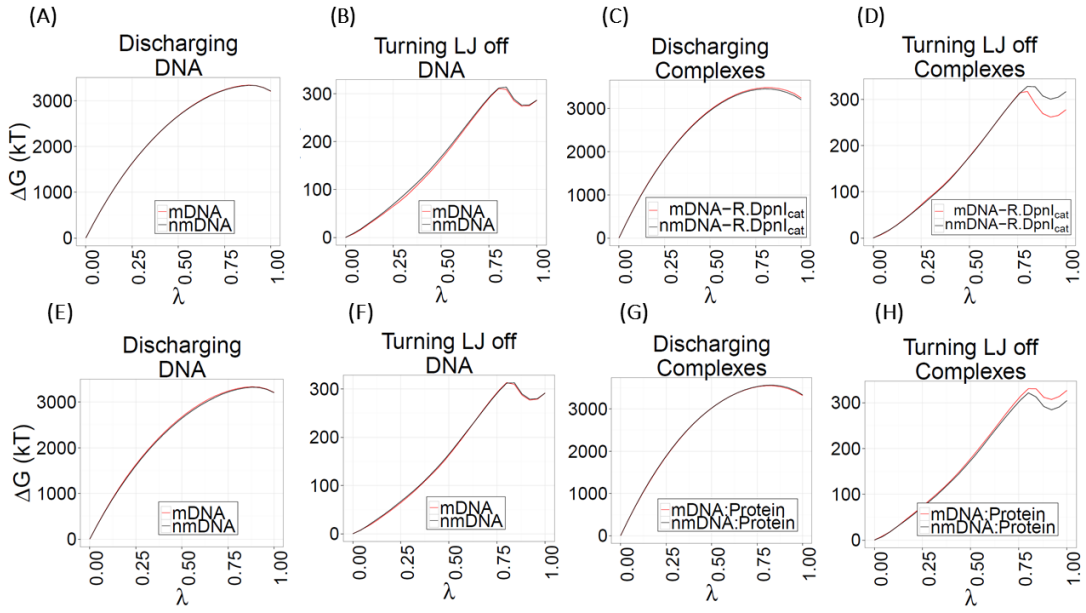


Figure 5.10 Cumulative Δ*G* during the four stages of the free energy perturbation calculations as a function of the coupling parameter λ (A-D) the bound forms in the catalytic domain as well as the corresponding unbound DNA structure, and (E-H) the bound forms in the winged-helix domain as well as the corresponding unbound DNA structure.

Table 5.2 shows the free energy changes (ΔG) with standard error for the individual annihilation processes for the protein-DNA complexes and the unbound DNAs. For unbound DNA started in the conformation extracted from the complex with the catalytic domain of R.DpnI, mutating the N6-adenine-methylated DNA to non-methylated DNA (ΔG_1) gave a favorable free energy change of ΔG_1 of -3.70 (±0.22) kcal/mol (Table 5.2A). Mutating methylated DNA bound to the catalytic domain into its non-methylated counterpart (ΔG_2) gave a slightly unfavorable free energy difference of ΔG_2 of 0.35 (±0.68) kcal/mol. Hence, the total cycle ΔG_1 - ΔG_2 adds up to $\Delta\Delta G$ of -4.05 (±0.71) kcal/mol meaning that N6-adenine-methylated DNA binds more strongly to the catalytic domain of the R.DpnI protein by this amount than unmethylated DNA.

Table 5.2B shows the free energy changes (ΔG) with standard error for the same annihilation processes in the winged-helix domain-DNA complexes and the unbound DNAs. Here, the free energy calculations were started from the final conformations of the protein-DNA complex after 100 ns of plain MD simulations. Mutating the methylated DNA in water to non-methylated DNA (ΔG_1) gave an unfavorable free energy change of ΔG_1 = 2.85 (±0.40) kcal/mol. On the other hand, mutating the methylated DNA bound to the winged-helix domain of R.DpnI into the non-methylated counterpart (ΔG_2) gave an unfavorable free energy difference of ΔG_2 = 4.17 (±0.29) kcal/mol. Hence, the total cycle ΔG_1 - ΔG_2 adds up to $\Delta\Delta G$ = -1.32 (±0.34) kcal/mol suggesting that methylated DNA binds more strongly to the winged-helix domain of the DPNI protein by this amount than unmethylated DNA, given the initial state of the DNA and the protein.

Table 5.2 Thermodynamic cycle to compute the contribution of adenine methylation to the binding free energy of DNA to the catalytic and the winged-helix subunits of the R.DpnI protein. Values are in units of kcal/mol.

(A) Catalytic Domain	$\Delta G_{discharging}$	$\Delta G_{turning\ LJ\ off}$	ΔG_{total}
met. DNA – DANN	-3.47 (±0.11)	-0.23 (±0.18)	-3.70 (±0.22)
met. Complex – Complex	24.25 (±0.58)	-23.9 (±0.36)	0.35 (±0.68)
ΔΔG=			-4.05 (±0.71)

(B) Winged-helix domain	$\Delta G_{discharging}$	$\Delta G_{turning\ LJ\ off}$	ΔG_{total}
met. DNA – DNA	2.76 (±0.44)	0.09 (±0.36)	2.85 (±0.40)
met. Complex - Complex	13.66 (±0.11)	-9.49 (±0.39)	4.17 (±0.29)
ΔΔG=			-1.32 (±0.34)

5.3.4 Enthalpic contribution to the binding free energy

Using the MM-PBSA approach (one trajectory method), we characterized the enthalpic contribution of N6-adenine methylation to the binding affinity toward both domains of the R.DpnI enzyme. For comparison, we also computed the enthalpic contribution of C5-cytosine methylation to the binding to the MeCP2 protein. Using eq. (2), the MM-PBSA calculations with an internal dielectric of 4 showed that adenosine methylation enthalpically favors binding to the winged-helix domain by -11.01 kcal/mol, and by -9.34 kcal/mol to the catalytic domain. In contrast, the calculations showed that C5-cytosine methylation slightly destabilizes the complex with MeCP2 enthalpically by 0.76 kcal/mol (Table 5.3).

Table 5.3 Enthalpic contribution of methylation to the binding energies (Mean (Std. err. of mean) kcal/mol) using the MM-PBSA approach. The internal dielectric constant was set to 4.

	DNA:R.DpnI$_{cat}$	DNA:R.DpnI$_{winged}$	DNA-MBD
nmDNA			
$\mathbf{H_{VDWAALS}}$	-134.90(0.10)	-103.08 (0.10)	-77.92 (0.11)
$\mathbf{H_{EEL}}$	-748.37(0.30)	-685.60 (0.22)	-986.21 (0.51)
$\mathbf{E_{PB}}$	752.39(0.25)	662.19 (0.21)	942.41 (0.48)
$\mathbf{E_{NPOLAR}}$	-92.03(0.06)	-67.72 (0.05)	-61.70 (0.07)
$\mathbf{E_{DISPER}}$	225.55(0.09)	167.47 (0.12)	150.59 (0.15)
$\mathbf{\Delta H_{gas}}$	-883.27(0.31)	-788.68 (0.24)	-1064.13 (0.57)
$\mathbf{\Delta H_{solv}}$	885.91(0.26)	761.93 (0.24)	1031.31 (0.53)
$\mathbf{\Delta H_{TOTAL}}$	2.64 (0.12)	-26.74 (0.06)	-32.82 (0.07)
mDNA			
$\mathbf{H_{VDWAALS}}$	-137.40(0.10)	-104.72 (0.12)	-74.49 (0.08)
$\mathbf{H_{EEL}}$	-760.59(0.27)	-697.71 (0.33)	-931.54 (0.40)
$\mathbf{E_{PB}}$	757.76(0.24)	670.52 (0.31)	891.30 (0.36)
$\mathbf{E_{NPOLAR}}$	-91.86 (0.06)	-66.75 (0.07)	-59.36 (0.05)
$\mathbf{E_{DISPER}}$	225.40 (0.11)	160.91 (0.16)	141.94 (0.10)
$\mathbf{\Delta H_{gas}}$	-898.00 (0.28)	-802.43 (0.36)	-1006.03 (0.44)
$\mathbf{\Delta H_{solv}}$	891.30 (0.27)	764.68 (0.33)	973.8731 (0.40)
$\mathbf{\Delta H_{TOTAL}}$	-6.70 (0.09)	-37.75 (0.06)	-32.16 (0.07)
$\boldsymbol{\Delta\Delta H_{mDNA\text{-}nmDNA}}$	-9.34 (0.11)	-11.01 (0.06)	0.66 (0.07)

5.3.5 Entropic contribution to the free energy of binding

The configurational entropy of DNA was quantified from MD snapshots of the concatenated sets of 0.5 μs long simulations (R.DpnI) or 0.2 μs (MeCP2), respectively, using the Schlitter method implemented in GROMACS (Schlitter, 1993). The simulations of unbound methylated and non-methylated DNA sequences were

started from perfect B-DNA conformations. The computed configurational entropies are listed in table 5.4 (see also figure 5.11). When considering both protein and DNA, N6-methylation of adenine gave an unfavorable entropic contribution of -17.33 kcal/mol for binding to the catalytic subunit of R.DpnI. When separately considering DNA and protein, both gave a negative contribution each (-5.95, -13.9 respectively). The situation is very different for binding to the winged-helix domain of R.DpnI. Here, we noticed a strong entropic decrease in the DNA, but this was completely offset by a positive contribution of the protein. Overall, methylation was predicted to give a slightly favorable entropic contribution of 2.46 kcal/mol. The case is again different for the 5C-meDNA:MBD system. Here, almost no change is found in DNA alone, whereas the protein shows a clearly lowered entropy when bound to methylated DNA. Interestingly, the entropy computed for the full protein:DNA system is higher for methylated DNA than for non-methylated DNA (9.54 kcal/mol). This reflects the important role of the relative mobility of protein and DNA that is not considered when treating the binding partners individually.

Table 5.4 Partial entropic contribution of methylation to the binding energies (TS 'kcal/mol') using the Schlitter formula at a temperature of 310 K. The entropy was calculated for (A) protein and DNA, (B) DNA alone, (C) proteins alone (only in the bound form)

(A)	**CATALYTIC DOMAIN OF R.DPNI**			**WINGED-HELIX DOMAIN OF R.DPNI**			**MBD:DNA**		
TOTAL ENTROPY	DNA free	DNA-Protein	Difference	DNA free	DNA-Protein	Difference	DNA free	DNA-Protein	Difference
nmDNA	237.80	1940.32	1702.52	237.80	1916.75	1678.95	369.79	1051.44	681.85
mDNA	236.01	1921.20	1685.19	236.01	1917.42	1681.41	368.05	1059.44	691.39
$T\Delta S_{MDNA-NMDNA}$			-17.33			2.46			9.54

(B)	**CATALYTIC DOMAIN OF R.DPNI**			**WINGED-HELIX DOMAIN OF R.DPNI**			**MBD:DNA**		
DNA Contribution	DNA free	DNA in Complex	Difference	DNA free	DNA in Complex	Difference	DNA free	DNA in Complex	Difference
nmDNA	237.80	244.01	6.21	237.80	222.27	-15.53	369.79	365.79	-4.00
mDNA	236.01	236.27	0.26	236.01	198.05	-37.96	368.05	364.52	-3.53
$T\Delta S_{mDNA-nmDNA}$			-5.95			-22.43			0.47

(C)	***CATALYTIC DOMAIN OF R.DPNI***	***WINGED-HELIX DOMAIN OF R.DPNI***	***MBD:DNA***
PROTEIN CONTRIBUTION			
Protein bound to nmDNA	1718.04	1690.33	480.19
Protein bound to mDNA	1704.11	1714.19	464.78
$T\Delta S_{mDNA-nmDNA}$	-13.93	23.86	-15.41

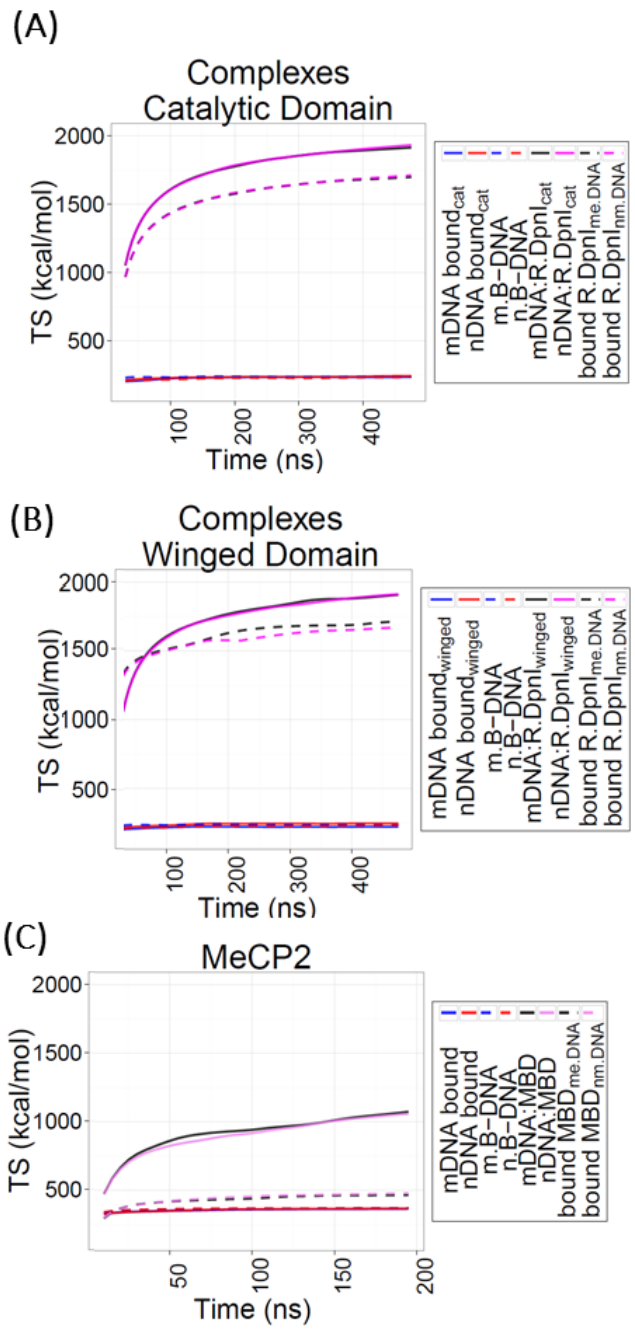


Figure 5.11 Convergence of the configurational entropy (TS) computed with the Schlitter method from the merged trajectories, (A) in the catalytic domain, and (B) in the winged-helix domain of the R.DpnI:DNA system, (C) in the DNA:MBD system.

5.4 Discussion

Previous authors have pointed out the challenge in explaining binding selectivity of proteins to methylated DNA (123,134). In fact, there may exist alternative mechanisms (134) that apply to different proteins as well as DNA methylation states, involving e.g. CH···O hydrogen-bonding interactions (135), cation–π interactions, e.g.

between Arginine residues and cytosine bases (123) and solvation/desolvation effects (28,144). In this study, we employed conventional MD simulations and free energy perturbation to unravel structural and energetic parameters that may explain the advantageous specific binding of methylated DNA to the two domains of the R.DpnI protein and to the MBD domain of the MeCP2 protein over the non-methylated DNA forms.

As is commonly observed upon protein:DNA association (145), both binding of methylated and non-methylated DNA to proteins was accompanied by an increased width of the major groove and a shift of the BI/BII equilibrium to smaller values.

Due to the hydrophobic nature of the methyl groups, one may suspect that desolvation of N6-methyladenosine is energetically favored over desolvation of adenosine. In a recent computational study (91), we found, however, that a single unmethylated N base has a very similar solvation free energy as a single N6-methylated adenine base. In contrast, the C5-methylated form of cytosine is strongly favored over the unmethylated form in water (reflected by a 26.7 kcal/mol difference in solvation free energy). Note that these numbers refer to single bases.

In the context of the DNA strand, steric shielding reduced the solvent accessible surface of a methyl group attached to adenine from an average of 65Å^2 to a range of 28-38Å^2 (see figure 5.12). Upon protein binding, the total solvent-exposed surface area of the methyl groups was reduced to only about 3Å^2 (~11% of the exposed surface for free DNA). The process of water displacement involved more waters for the winged-helix domain, so that their release into the bulk solvent should be entropically favorable, as previously reported (80). The SASA contribution was also calculated for the C5-methylated cytosine upon integration of the C5-methylated cytosine in the BDNF promoter, and further in the BDNF:MBD complex (figure 5.13). The methyl group in free C5-methylcytosine had a total solvent-exposed area of 55Å^2, what reduced to about 35-40 Å^2 when integrated in the DNA strand. In the complex with MBD, the SASA surface was about 10-15 Å^2 (~25-30% of the exposed surface for free DNA). This suggests that the shielding of C5-methylated cytosine likely plays a smaller role for the binding specificity than for N6-methylated adenine.

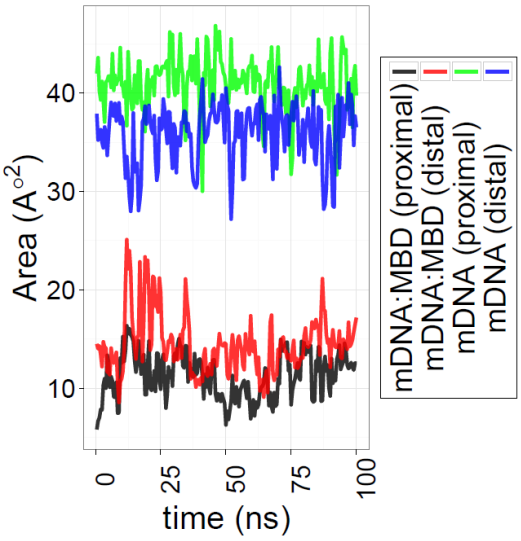


Figure 5.12 Solvent Accessible Surface Area (SASA) for the methyl group of the 6-methyladenine during the first 40 ns long MD simulations. The plot shows the methyl group; (A) in the fully methylated or hemimethylated DNA in the unbound DNA sequence of the R.DpnI bound form; and (B) in the respective complex with R.DpnI catalytic domain or (C) winged-helix domain SASA.

According to a previous study of Zou et al for the MBD:DNA system, the C5-methylation of cytosine contributes about -1.2 kcal/mol of preferential binding free energy to the interaction between methylated DNA and the MBD domain (123). Similar free energy perturbation performed by us showed that N6-adenine-methylated DNA favored binding to the catalytic domain of the R.DpnI protein by a slightly larger amount (-4.05±0.7 kcal/mol). These results are consistent with experimental findings that non-methylated DNA shows at most weak binding to the catalytic subunit (10). N6-methylation of adenine also gave a favourable contribution for binding to the winged-helix domain (-1.32±0.34 kcal/mol).

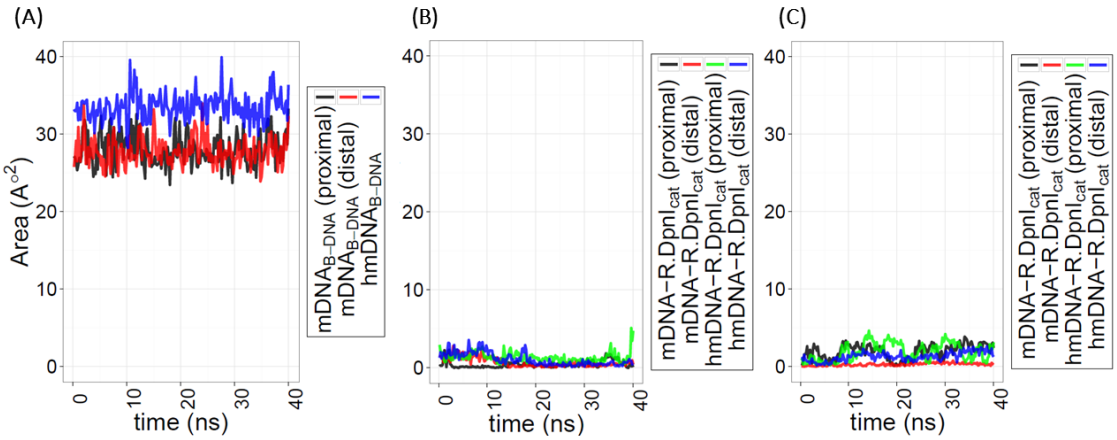


Figure 5.13 Solvent Accessible Surface Area (SASA) for the methyl group of the C5-methylcytosine in the MBD:DNA complex (100 ns simulations) and in the unbound state.

Computational modelling is also able to address individual contributions to the binding free energy. Here, we employed MM-PBSA calculations to compute the differential effect of adenine and cytosine methylation to the binding enthalpy. We

found that N6-adenine-methylation favors binding to the R.dpnI winged-helix domain by -11.01 kcal/mol, and to the R.dpnI catalytic domain by -9.34 kcal/mol. This comparably large favorable enthalpic contribution cannot be attributed alone to non-bonded interactions between the two N6-adenine methyl-groups and protein residues. An important role is likely also played by the different conformational adaptation by DNA to the protein domains between methylated and non-methylated DNA. In addition, one needs to remember that we refer to differences in binding enthalpies between the solvated and bound states, so that one always needs to consider the unbound state as well. Zou and colleagues have discussed for the Mecp2 system that classical force-field are likely not able to capture well the characteristic cation-pi interactions formed between arginine residues of the MBD domain and nucleic basis (123). Omission of these apparently important effects may explain why the difference in binding enthalpy computed was slightly unfavorable in our study.

Next, the Schlitter formula was used to extract configurational entropies of protein and DNA from the co-variances observed in plain MD simulations. In all the species studied here, free DNA had lower entropies in the methylated form than in the non-methylated form.

An experimental study by Jen-Jacobson and colleagues characterized the enthalpic and entropic contributions to the binding of several protein:DNA complexes (80). The authors observed an *isothermal* entropic-enthalpic compensation for the different systems. A similar case was observed here for DNA binding to the winged-helix domain of R.DpnI where a strong decrease in the conformational entropy of DNA was fully compensated by a corresponding gain in conformational entropy of the protein. On the other hand, N6-adenine methylation disfavored binding to the catalytic subunit of R.DpnI entropically, whereas C5-cytosine methylation had an entropically favorable effect on the interaction of the MBD:DNA complex.

5.5 Conclusion

DNA methylation of specific DNA regions is a targeting signal for particular proteins such as e.g. transcription factors. According to our findings, specific binding to N6-adenine-methylated or C5-cytosine-methylated DNA is achieved through structural adaptation in DNA and protein on the one hand, and through the combined effects of more favorable binding enthalpies and modulation of the conformational entropy, on the other hand. For the R.DpnI system, a favorable enthalpic contribution seems to plays a major role in favoring binding of methylated DNA over the non-methylated DNA. In contrast, specific binding of C5-cytosine-methylated DNA to the MBD domain

of the Mecp2 protein appears to be predominately stabilized by a favorable entropic contribution due to the concerted dynamics of protein and DNA. It remains to be studied whether these characteristics are intrinsic properties of the systems investigated here or whether they transfer to other systems and are, thus, general principles of N6-adenine vs. C5-cytosine methylation.

Chapter 6

Cross-talk between intragenic epigenetic modifications and exon usage across developmental stages of human cells *

*The results of this chapter were jointly obtained by Mr. Ahmad Barghash and the author. The main contribution of Mr. Barghash was late normalization and preparing data in tables, calculation of correlation and production of most figures. The main contribution of the author was data retrieval, establishing early calculations and data preprocessing, calculation of read count on the exon level, annotating genes according to exon count and early normalization stages. Results were jointly analyzed.

Abstract

Differential exon usage has been reported to affect the large majority of genes in mammalian genomes. It has been shown that different splice forms sometimes have distinctly different protein function. Here, we present an analysis of the Human Epigenome Atlas (version 8) to connect the differential usage of exons in various developmental stages of human cells/tissues to differential epigenetic modifications at the exon level. We found that the differential incidence of protein isoforms across developmental stages is often associated with changes in histone marks as well as changes in DNA methylation in the gene body or the promoter region. Many of the genes that are differentially regulated at the exon level were found to be associated with development and metabolism.

Summary

Differential exon usage is a mechanism used by complex organisms to increase the usability of the gene-coding regions, so that several different proteins are expressed from the same chromosomal position. Epigenetics studies inherited modifications to genes that do not belong to the raw DNA sequence, but nevertheless modulate gene expression. Epigenetics is well-associated with alternative splicing in the gene body, but the connection to distinct developmental stages has not been addressed so far. Here we show that a sizeable number of genes that are essential for development show strong associations between differential exon usage and epigenetic modifications.

6.1 Introduction

Differential exon usage is reported to occur in 90-95% of all human multi-exon genes (146),(147). Different splice variants of a gene may lead to different protein products that exert different functions. As a result, differential exon usage leads to a strong expansion of the eukaryotic proteome (148). An example for this is the well-known *Nanog* gene; where alternative splicing results in two variants of the *Nanog* protein with different capabilities for self-renewal and pluripotency in embryonic stem (ES) cells (149). An alternative scenario takes place when genes coding for different proteins occupy the same position on a chromosome. In such cases, differential exon usage even controls the expression of different proteins. A well-charecterized example for this case are the overlapping imprinted genes PEG3/ZIM2 that are exclusively expressed from the paternal allele (150),(151). However, the notion of alternative splicing across tissues should not be considered as an exclusive either/or mechanism. Thanks to recent advances in RNA-seq technology (44), it is now possible to study the expression of genes at the level of single exons. The granulity of exon usage can thus be increased from the basic classification of a one-or-none expression per gene (alternative splicing) to fine-tuned quantitative read counts that can be accounted for per individual exon.

A recent study reported that differential exon usage in primates shows more profound differences across species than on the intra-species level. It was targeted at adult tissues (brain, cerebellum, heart, kidney, and liver), and did not analyze the effect of differential usage of exons in terms of organismal development (152). Only few studies have so far related alternative splicing events with epigenetic modifications. Zhou *et al.* studied the relationship between alternative splicing and histone marks (153). Another study by Schwartz *et al.* addressed the interplay between chromatin structure and the exon-intron architecture. They showed that histone modifications within the gene body are more pronounced in exon regions than in intron regions, and thus may serve to define the exon-intron boundaries (8).

The notion of splicing was thought to work by transcribing either the full (constitutive) genes or alternatively spliced forms. Until recently, the relation between exon usage and transcript abundance has been scarcely analyzed. To our knowledge, there has been so far no attempt to study the relationship between differential usage of exons and various types of epigenetic modifications and to connect this with different developmental stages of human. This is precisely the aim of this study. Based on data for human development across different stages from the Human Epigenome Atlas (53),(154),(155),(156), we show a correlation between

differential exon usage and several epigenetic modifications at the exon/intron/promoter level, namely DNA methylation and several histone marks. The correlation is significant for both the constitutive genes and for gene clusters. Furthermore, we could associate the occurance of differential exon usage with functional annotations that, indeed, often relate to regulation of signalling and developmental processes.

6.2 Methods

6.2.1 Data Preparation

Data for this study was retrieved from the Human Epigenome Atlas (up to release 8) that is part of the Roadmap Epigenomics project (53),(154),(155),(156). Table 6.1 introduces the assays and the epigenetic modifications analyzed in our study. The aim of this study was to find the link between the differential usage of exons to specific epigenetic marks and how this correlates to different stages of human development. Thus, we only studied sample types for which release 8 provided complete data sets according to table 6.1. These stages included stem cells, early developmental stages, induced differentiated cells, fetus, and adult tissues. Figure 6.1 lists the studied tissues.

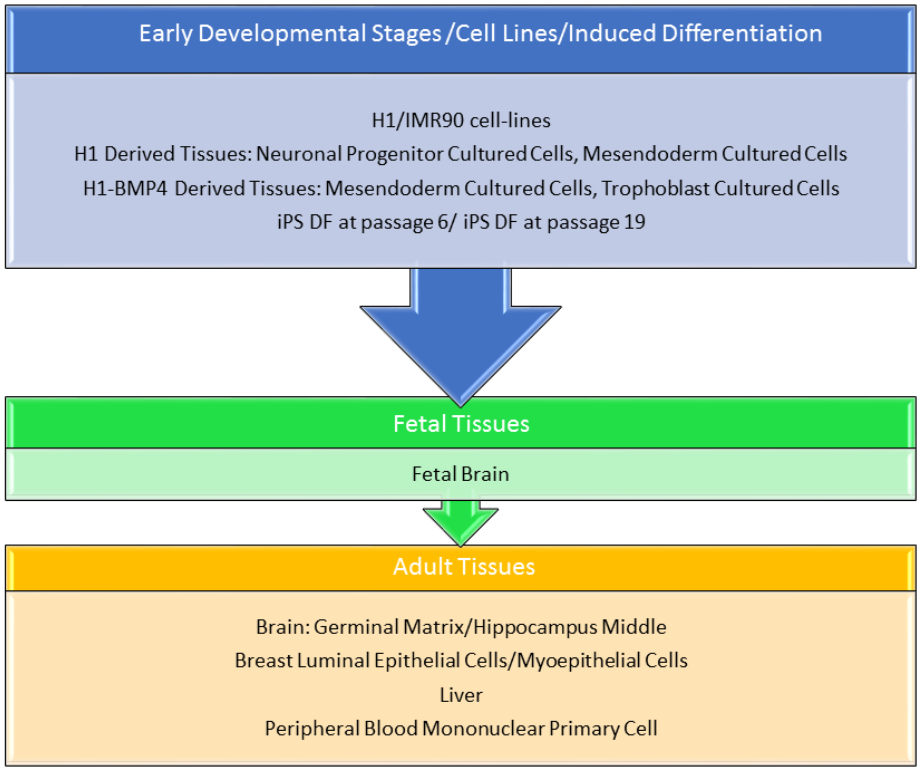


Figure 6.1 The 14 different tissues that were investigated in this study belong to the three different main developmental stages.

Table 6.1 Assays used in this study to evaluate the levels of expression, chromatin organization and DNA methylation in the human genome during different developmental stages.

Expression	Chromatin Organization	DNA methylation
mRNA-Seq siRNA-Seq	ChIP-Seq Input DNAse hypersensitivity H3K27ac/H3K27me3 H3K36me3 K3K4me1/H3K4me3 H3K9ac/H3K9me3	Bisulfite-Seq/RRBS MeDIPS-Seq

We downloaded the human UCSC hg19 reference genome, retrieved the exons of each gene, and prepared them in a temporary annotation file. To account for possible ambiguity, each gene should only be mapped to one genomic region. As a result, we dropped a small set of ~100 genes spanning more than one genomic region from our analysis. Furthermore, we clustered genes that mapped to the same genomic region into one gene cluster to prevent redundancy in mapping, see figure 6.2. Following the strategy of Anders et al, we sorted the group of exons belonging to the genes of one gene cluster, and extracted the unique exons (157). If any two exons from different genes mapped to the same genomic region, we rearranged them and assigned them to a new non-overlapping classification of exons that mapped to the same region, see figure 6.2 for illustration. After that, we mapped introns and promoters accordingly. We defined the promoter region as the region between -2000 bp upstream of the transcriptional start site and 0 bp of the gene/gene cluster region.

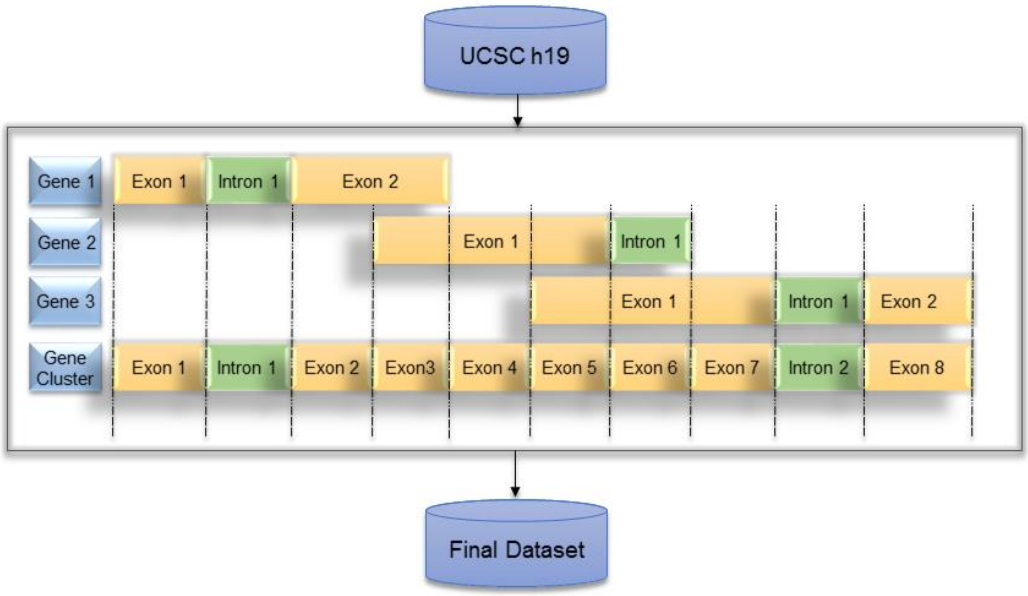


Figure 6.2 A schematic representation of the exon architecture of three exemplary genes that show partial overlap. The virtual gene cluster shown in the bottom row consists of shorter exons 2-7 in order to resolve the overlapping issue. Also shown is how Exon 6 is assigned to resolve a conflict of the overlapping Gene 2:Intron 1 and Gene 3:Exon 1 case. See the main text for further explanation.

For data processing of the ChIP-Seq assays, we called the peaks associated with the reads in the retrieved bed files. To this aim, we used ChIP-Seq Input assay to check for the background effect. Peak calling of the different histone marks was performed using MACS (158).

6.2.2 Data Normalization

In order to account for putative technical noise in the data and to check for differential read usage, we performed pair-wise comparisons of the reads in different tissues. To this aim, we modeled read counts using regression analysis to detect noise in tissues in a pairwise manner.

The peak calls in ChIP-Seq data were normalized using MAnorm, where linear regression analysis is performed (86). Normalization of MeDIP-Seq data was done with the MeDIPS Bioconductor Package that uses negative binomial regression (159). The methylation datasets from the bisulfite-seq and RRBS datasets were normalized using the Biocondutor package methylKit (85). For each basepair position, logistic regression was applied to check for differential methylation per base. These results were processed to obtain the mean methylation ratio per exon. To normalize the mRNA and smRNA data, we first obtained the transcript and exon abundance. We generated SAM files from the supplier's BED files via BedTools and SamTools (160,161) and sorted the SAM files lexicographically. Read counts of genes and exons were prepared from the SAM files using the HTSeq package (162) and used as an input for the Bioconductor DEXSeq package (157) to reduce noise in the data.

Data annotation for the normalized ChIP-Seq and methylation data was performed using BedTools (160). Expression data were already annotated by the HTSeq package (162). After that, we mapped the whole set of normalized reads, including the read numbers for expression, the different histone marks, and the methylation status for each exon in a gene/gene cluster per tissue into a final table per read type. The table consisted of one read value per tissue per exon. If for a read type different read numbers mapped to the same exon, we averaged them. After that, we normalized all read numbers for a single gene to a final range of log values between -1 to +1.

6.2.3 Differential usage of exons

Differential usage of exons was analyzed using the strategy described in figure 6.3. We aimed at identifying genes for which differential usage of their exons across developmental stages in terms of exon expression is associated with clear differences in epigentic marks. To achieve this, we followed two different strategies to examine

correlations between different epigenetic marks and the expression levels of exons. Both marks needed to map to the same exon, to a directly adjacent intron, or to the promoter region for the genes/gene clusters that we defined.

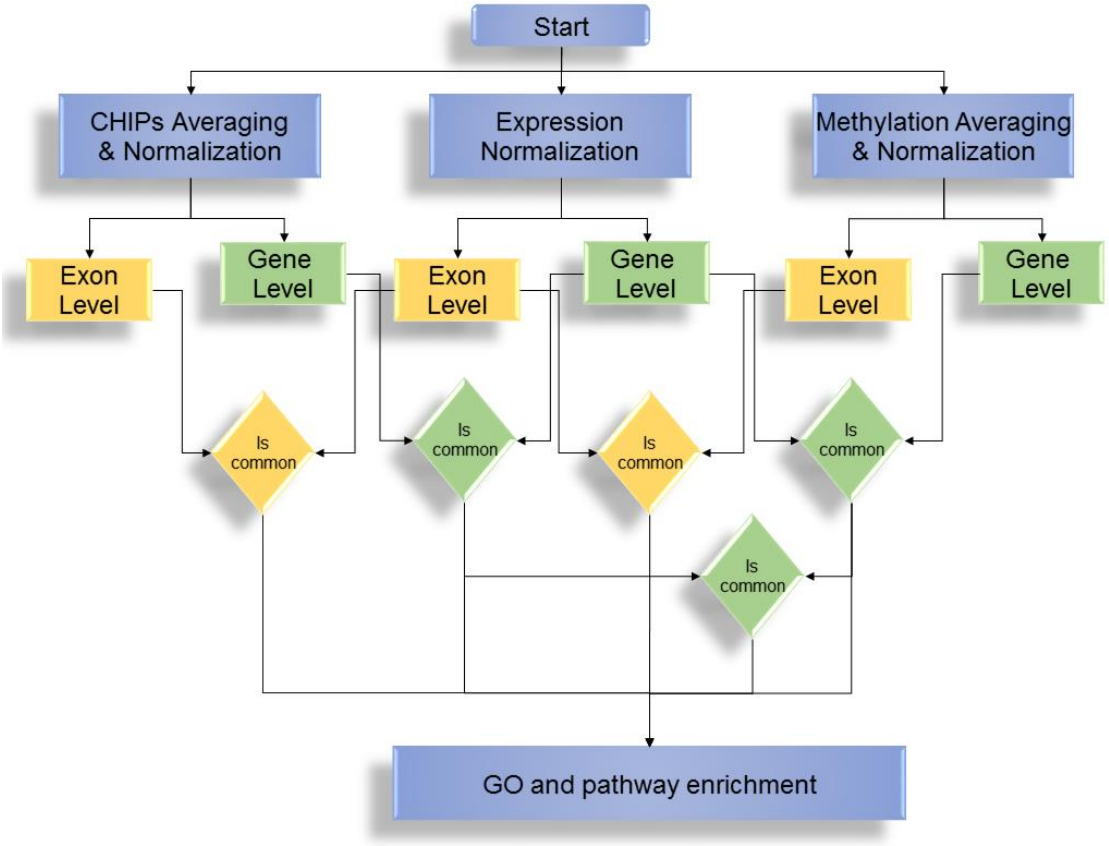


Figure 6.3 A schema for the pipeline of studying the gene- and exon- levels of differential exon usage across developmental stages and correlating this to the differential epigenetic marks. See main text for further explanation.

The first strategy checks for anticorrelations in read counts on the gene level. We calculated these anticorrelations for exons that belong to a single gene/gene cluster in a pair-wise manner between tissues. To this aim, we explored all genes with ≥ 4 exons in all possible pairwise combinations among the 14 tissues studied here. We set the threshold of the Pearson correlation coefficient (PCC) to a tight bound of ≤ -0.7. We followed this strategy for read counts of mRNA, different histone marks, DNAse hypersensitivity, siRNA, or methylation levels. Additionally, we applied the same strategy to examine anticorrelations on the intron level. This test yielded lists of genes with anticorrelated levels of exon expression and one or more of the associated epigenetic marks, or epi-spliced genes (see figure 6.4). Furthermore, if the anticorrelation of expression coincided with both anticorrelations in histone marks and methylation, this was documented as well.

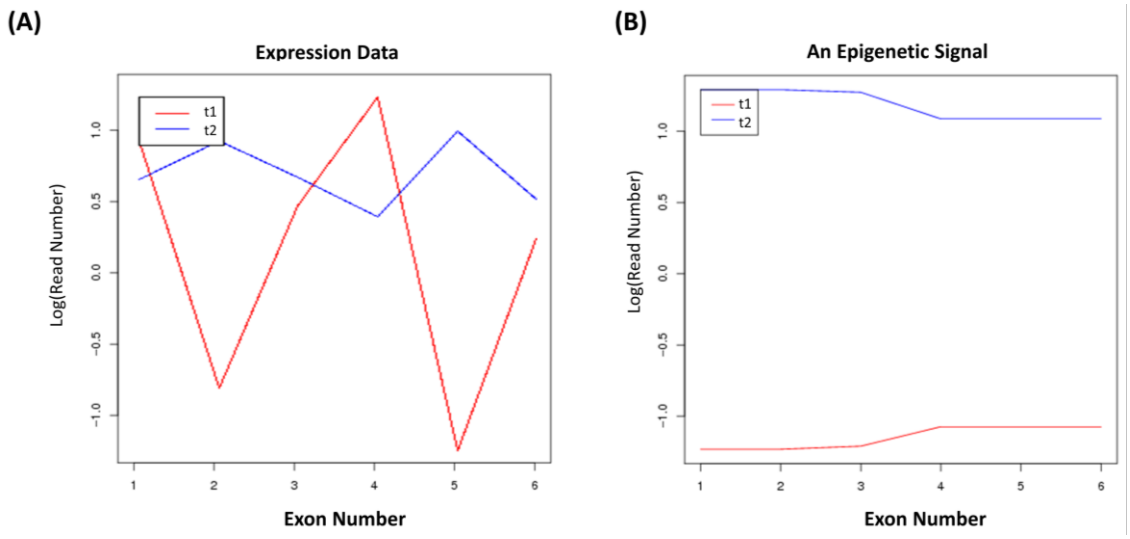


Figure 6.4 A model example of an epi-spliced gene (CCS) that shows negative (anti-) correlations at the exon level in gene expression (A) and in an epigenetic mark, namely H3K36me3 (B).

The second strategy identified changes of the read number on the exon level in all possible genes across developmental stages. The results were then correlated with the changes in read counts for the different epigenetic marks described above. We set the Pearson correlation coefficient to a tight bound with an absolute value of at least 0.7.

After that, we checked for the enrichment in GO terms using the GOSim package (89). We examined functional similarity in two sets of genes. The first set included genes that were identified in the same pairwise tissue comparison. For this analysis, we only considered tissue pairs that have at least 10 genes that are both differentially expressed and show differential epigenetic marks between those two tissues. In the second set, we grouped the genes that showed correlated changes of expression for a single exon and one epigenetic modification across tissues.

For completeness, we also analyzed positive correlations in read counts on the gene level. To identify cases of constitutive gene expression, we calculated these correlations for exons in a single gene/gene cluster in a pair-wise manner between tissues. As before, we explored all genes in all possible combinations of the 14 tissues we studied. Again, we set the threshold of Pearson correlation coefficient (PCC) to a tight bound of ≥ 0.7.

6.3 Results/Discussion

Our analysis considered 7960 constitutive genes and 14668 gene clusters. Figure 6.5 shows results on the gene level for the number of genes/gene clusters with differential exon usage that was negatively correlated with epigenetic marks between each pair of tissues (PCC < -0.7, see methods for definition of differential exon usage and for the definition of gene clusters). Figure 6.5a shows the dissimilarity of exon expression. The largest differences were found between trophoblast cultured cells and mesendoderm cultured cells, as well as iPS passage 19 (dark blue). Differences between later developmental stages were rather small in comparison. Figure 6.5b shows the dissimilarity of DNA methylation for all genes. Notably, trophoblast cultured cells were the least similar to all other tissues. A seemingly peculiar similarity was found between fetal brain and all other tissues as well as for the two breast tissues (straight light grey bars). This could be traced back to the fact that very few genes showed differential methylation of their exons for these three tissues.

Figure 6.5c shows the dissimilarity of H3K36me3 as an example of the respective histone analysis. H3K36me3 was selected for this because it showed the largest number of histone marks in the gene body, as has been reported before (163). In contrast to figures 6.5 (a) and (b), rather balanced differences were found between all tissues.

Figure 6.5d shows the results from an integrated analysis, where the set of genes showing differential exon usage (measured by expression, see Fig. 6.5a) was intersected with the set of genes showing either anticorrelation (PCC ≤ -0.7) in DNA methylation (Fig. 6.5b) or in histone marks (Fig. 6.5c and other histone marks that are not shown). We will refer to such genes as “epi-spliced” genes. Dissimilarity was measured as the ratio of genes in the intersection set over the max number of intersections in all studied tissues. Clearly, trophoblast cells were the most different from all other tissues. Two different passages of iPS cells showed very similar combined expression/epigenetic marks, whereas ESCs are more distant to the iPS cells. Fetal brain and adult brain also showed similarity. As expected, we found that mesendoderm cultured cells and trophoblast cultured cells exhibited large differences in their epi-spliced genes. Based on the data from figure 6.5d, we generated a cluster dendogram by average-linkage hierarchical clustering, see figure 6.6. Trophoblast cells showed by far the largest dissimilarity to all other tissues (note that fig. 6.6 shows the logarithm of the distance). As expected, breast tissues showed high similarity in terms of associations with epigenetic marks. This was also the case

for some of the brain tissues, various stem cell-like stages, and for the two passages of the iPS cells.

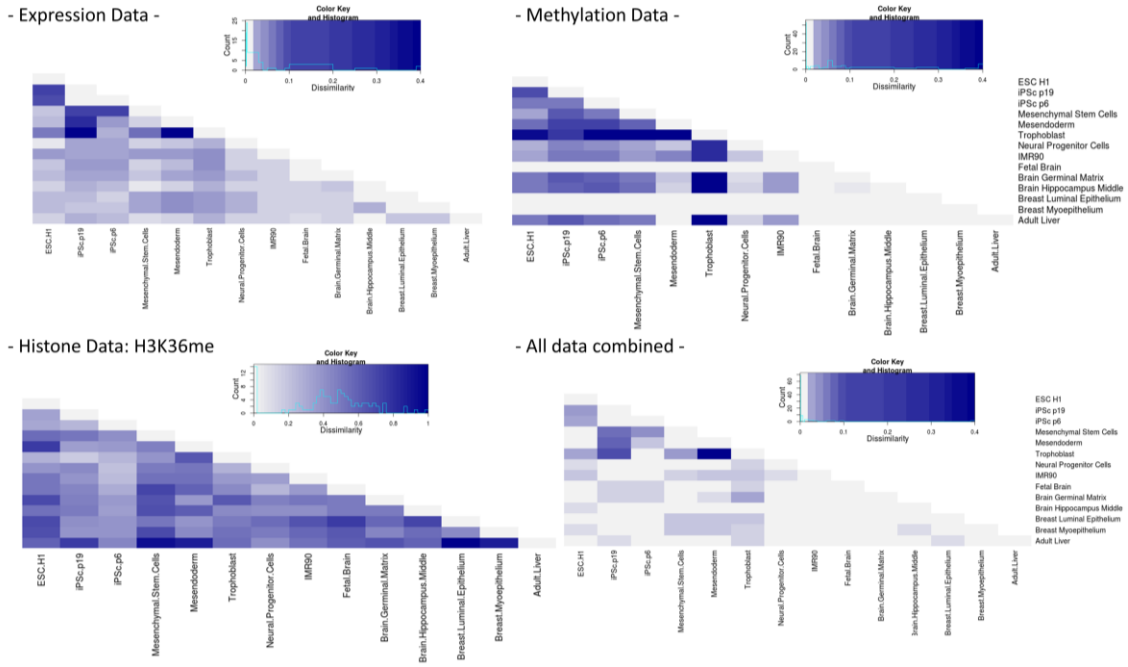


Figure 6.5 Heatmaps of the number of the resulting pairwise negative correlations for (a) expression data, (b) methylation data, (c) histone modifications, here H3K36me3, (d) the above mentioned union.

For a better representation, we selected mutual negative correlations in the pairwise tissue comparisons with at least 10 epi-spliced genes (see Methods section). Within the constitutive genes, we found a total of 1529 epi-spliced genes. From this list, only 81 genes/gene clusters showed common modulation at the level of histone modification and methylation at the same time.

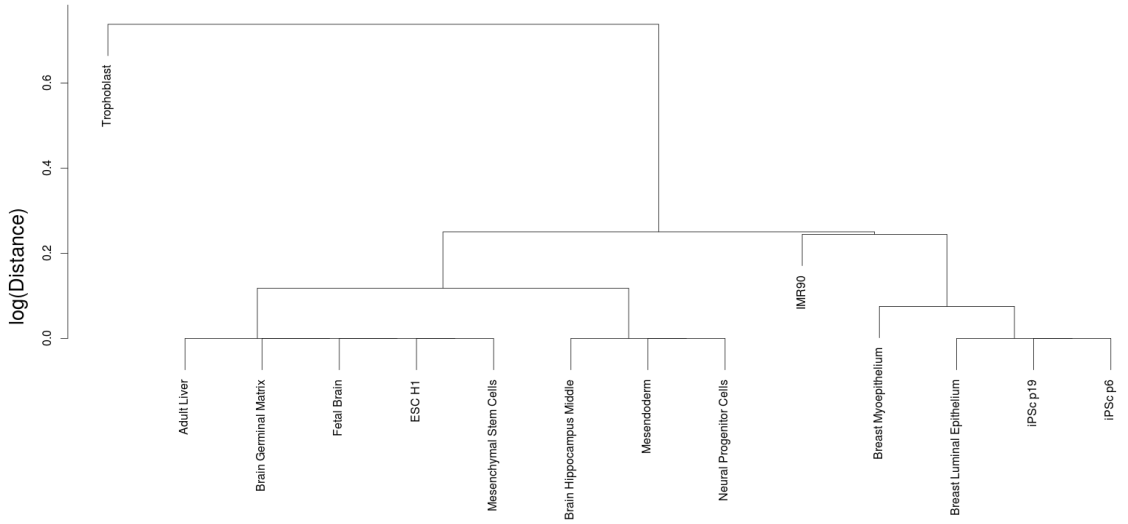


Figure 6.6 Hierarchical clustering for the set of genes that were analyzed in figure 6.5d.

We further investigated the list of exons where changes in expression were associated with changes in epigenetic marks across the studied developmental stages. We only considered epigenetic changes in the gene body of the same exon or in the promoter. Such changes can help to assign an effect of a single histone mark, methylation state or siRNA regulation to human development by investigating crucial genes/gene clusters which are convolute with a single epigenetic modification. Table 6.2 lists the number of exons showing high correlations ($r \geq 0.7$) between an epigenetic modification and expression as well as the number of genes containing these exons. We did not account for putative exons that can be both positively and negatively associated with the same epigenetic mark, but only identified those showing either one of the two trends. However, the same gene can contain exons that are either positively or negatively associated with the same epigenetic mark.

Table 6.2 Number of exons/genes with significant correlation of exon-level expression and an epigenetic mark.

Epigenetic modification	Chromatin Accessibility	H3K27ac	H3K27me3	H3K36me3	H3K4me1
Exon	2673	9081	568	3990	2267
Gene	725	452	187	1416	886
Epigenetic modification	H3K4me3	H3K9ac	H3K9me3	Methylation	siRNA
Exon	3519	319	121	318	0
Gene	942	145	44	122	0

6.3.1 Functional Classification

Based on the result of the correlation analysis, we identified enriched GO terms for the resulting gene sets, both on the gene and exon levels. In doing so, we ignored the fact that different splice variants of a gene may sometimes promote very different functions (164). We first analyzed the results for the negative associations on the gene level in a pair-wise manner, and considered enriched gene groups in terms of pair-wise tissue allocation. For this, we identified genes where changes of a histone mark as well as in the DNA methylation state coincide significantly with differential exon usage for the same gene. Such cases were exclusively found for combinations between the trophoblast cultured cells, mesendoderm cultured cells and induced pluripotent stem cells. GO terms associated with epi-spliced genes in those stages were associated with chromatin organization, (e.g.; the introduction of the heterochromatin and telomere structuring; growth of the ovarian follicle, oocyte, etc; transport processes of organic and inorganic molecules;) with metabolism; with transcriptional/translational and post translational regulation (e.g., K48- or K63-linked deubiquitination) and with homeostasis by regulation of embryonic hormones, interferons and Rac GTPase gene. The Rac protein has a role in growth and epithelial tissue differentiation and also a well established role in cancer. One further enriched GO term was H3K4 methylation.

We also analyzed GO terms for groups of genes with differential exon usage showing significant common changes of either histone marks or the DNA methylation state. We grouped genes according to the differential pairwise tissue allocation. The majority of significant intersections in histone modifications show early in development. Apart from the trophoblast cultured cells, the mesendoderm cells and the iPS cells, we also found significant changes between the trophoblast cultured cells and any of mesenchymal stem cells, the H1 embryonic stem cells, and the brain germinal matrix. iPS cells also display significant differences from the H1-derived mesenchymal stem cells. Apparently, using different passages (passage 6 and passage 19) of iPS- cells results in significant differences.

We then grouped the GO terms into seven broad functional categories, see figure 6.7, namely development, DNA and chromatin organization, regulation of transcription and translation, signaling pathways, metabolism, regulation, and others. Epi-spliced genes were overrepresented in developmental processes associated with the following tissues: blood vessels, chondrocytes, cytotoxic T cell, keratocyte, oogenesis, organelle assembly, and several others. The biological processes related to chromatin organization that involve epi-spliced genes include processes associated with M-

phase of the cell cycle, and several preparatory processes of the G1/S/G2 phases of the cell cycle. The category of transcription and translation involved many regulatory processes at the level of transcription, translation and post-translational modifications. The identified metabolic processes were associated with sugar metabolism, e.g., fructose 6-phosphate and fructose 1,6-phosphate metabolism, with phosphate metabolism, fatty acid metabolism, growth factors production, etc. Regulation included Ras GTPase activity, neuron migration, keratenocytes migration, and several others. Signaling cascades associated with epi-spliced genes included for instance the regulation of the MAPK cascade, bone morphogenic protein (BMP) signaling, signal transduction, involving Rac and Rho proteins and nerve growth factor receptor signaling pathways, as well as SMAD proteins. Rac and Rho proteins belong to the Ras family and regulate important cellular processes as cytoskeleton remodelling, gene expression, cell proliferation and organelle development (165),(166). SMADs are involved in TGF-β signalling from the cell membrane to the nucleus (167).

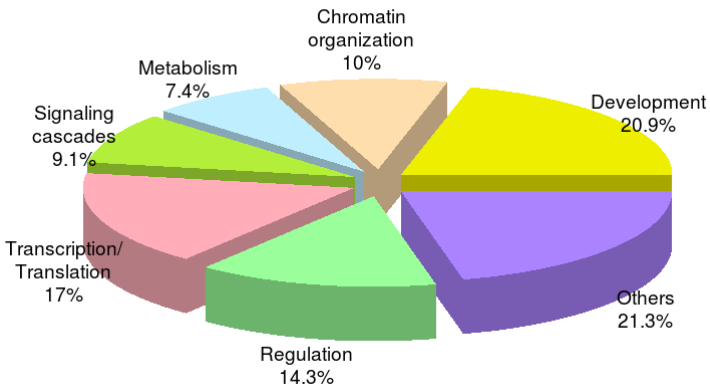


Figure 6.7 Frequency of Gene Ontology terms belonging to epi-spliced genes to seven according to seven manually defined biological categories. Epi-spliced genes are those showing a common negative correlation on the expression/epigenetic modification level across pair-wise tissue comparisons.

We then identified GO terms of epi-spliced genes that were significantly linked to individual epigenetic modifications. This grouping was based on significant correlations at the single exon level across developmental stages. Figure 6.8 illustrates the set of biological processes and their modulation via epigenetic signals. Overall, more GO terms were associated with differential histone marks than with differential DNA methylation. Additionally, H3K36me3 showed the strongest association with regulation processes related to transcription/translation/post-translational modification, chromatin modeling, and development. Whereas several histone modifications showed strong effects on genes of the given biological categories, others exhibited weak correlations in the same context, namely H3K9me3 and H3K27me3.

6.3.2 Association of epi-splicing with developmental stages

GO terms associated with development and growth were enriched in genes that showed correlation between exon usage and DNA methylation, see Figure 6.8. The most pronounced developmental effects related to DNA methylation were associated with nervous system development. Interestingly, we observed that genes for which high methylation levels of their exons were correlated with their expression had an important effect on DNA conformation, what is a well known effect documented from experiments (168). We also noticed that differential methylation was associated with crucial regulatory processes, including the regulation of protein phosphatase 2B, GTP catabolism, and Rho protein signal transduction. Next, we examined the biological processes enriched in epi-spliced genes that are associated with DNAse hypersensitivity. In this context, we found that this assay targets genes enriched with GO terms of open/closed chromatin organization. A striking example for this effect is the H3K4 histone methylation level.

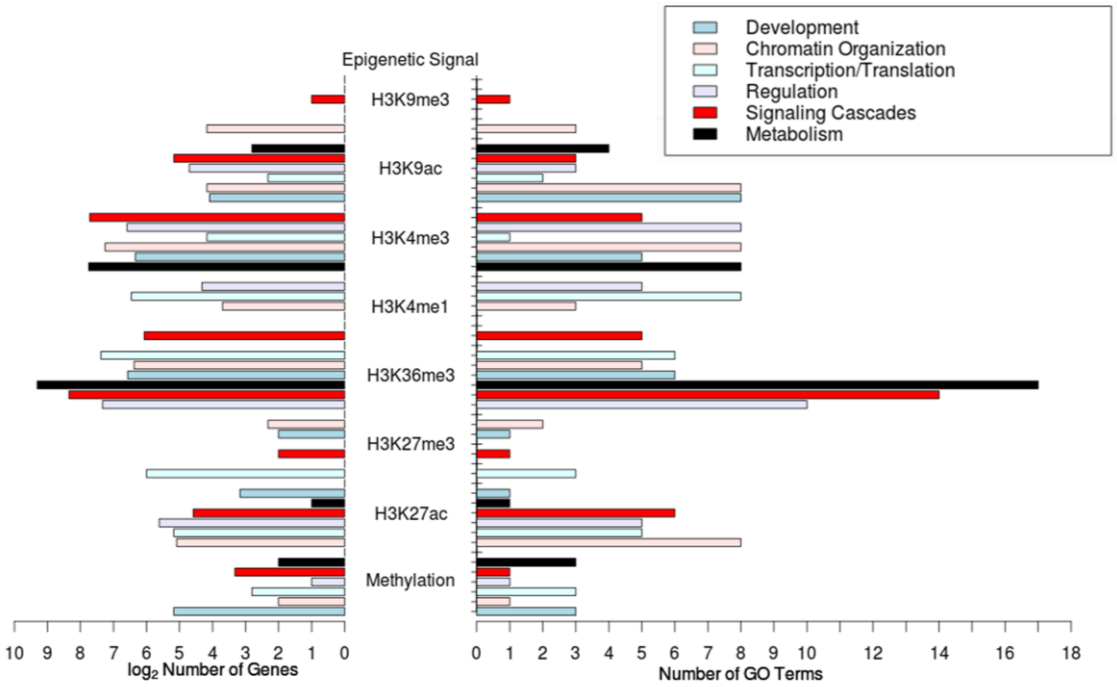


Figure 6.8 The numbers of selected GO trms belonging to genes showing differential regulation of the exon level for different types of epigenetic modifications.

Next, epi-spliced genes at the exon level were studied across several developmental stages and analyzed for GO terms enriched in these genes. The inhibitory/activating marks of H3K4 methylation/tri-methylation are associated on the developmental level with notochord regression, neuron projection regeneration and several

morphogenetic processes. We also found that these histone modifications are associated with female pregnancy and hippo signaling pathways that are prominent in the regulation of cell proliferation and apoptosis (169). Hippo signaling also serves the organisms to stop growth at a specific point, thus aiding in size control (170). H3K4 methylation/tri-methylation have a relevant effect on the levels of two proteins that act as a heterodimer, namely TLR1 and TLR2 that have roles in immune response (171). Differential H3K4 methylation correlated to exon usage is associated with the signaling cascade of the oncogene smoothened protein, with the bone morphogenic protein signaling cascade, with cascades including the SMAD protein and the transforming growth factor proteins, and with the nerve growth factor signaling cascade, including the well known BDNF protein that is also controlled through DNA methylation (172),(173). In terms of post-translational modifications, we found that these histone modifications also control the phosphoprotein phosphatase activity.

With respect to modifications of H3K9 associated with differential exon usage, we found that H3K9 acetylation is strongly connected to DNA and chromatin organization, cell cycle events, bone morphogenesis and differentiation, and with post-translational modifications, including for example Hsp90 chaperon acetylation. H3K9 tri-methylation, on the other hand, is mainly associated with nervous system development. Acetylation of H3K27 is directly associated with GO terms related to histone acetylation, suggesting a possible negative/positive feedback effect. It is also associated with the developmental control on the level of embryonic heart muscles and with processes related to chromatin organization. On the other hand, H3K27 trimethylation is associated with nervous system development and platelet-derived growth factor (PDGF) receptor signaling pathways. Lastly, we found that H3K36 trimethylation is associated with DNA replication and repair, chromatin organization, cell cycle events (e.g. G2/M phase checkpoints and mitotic cell division), regulation of transcription, and several developmental stages. We found that H3K36 trimethylation also modulates signaling cascades together with H3K4 methylation.

The biological processes just discussed only involved associations of epi-spliced genes and individual epigenetic marks. Next, we identified biological categories that are directly or indirectly related to several synchronously changing epigenetic marks. We performed this analysis for epi-genes both on the gene level in pairwise tissue comparisons and on the exon level across developmental stages. The identified categories included several imprinted genes, chromatin remodellers, protein kinases, and transcription factors and cofactors. For example, we found that four different paternally expressed genes vary their exon usage in a common manner due to a synchronous change of histone modifications and DNA methylation between

mesendoderm cultured cells and trophoblast cultured cells. These genes are PEG3/ZIM2 and SNRPN/SNURF that are all known to undergo alternative splicing (150,151,174). Another exciting example of a chromatin remodeller gene that varies the expression of its isomers in the same manner and in the same tissues is DNMT3L methyltransferase that is well known to recruit chromatin remodellers, especially histone deacetylases (175,176). This enzyme is also known to have crucial roles in early developmental stages, especially in the establishment of imprints together with *de novo* methyltransferases (177). A splice variant has been introduced for this gene in Ref-Seq genes, 2012.

Furthermore, we explored the list of genes for which exon expression is associated with a specific epigenetic mark. Interestingly, exon expression of a few imprinted genes changes across developmental stages, and this expression was modulated by several epigenetic marks. For example, the maternally expressed gene SLC22A18 changed its expression according to the padding at the chromatin structure and is modulated by H3K4 mono-methylation. This gene has been linked to alternative splicing events before (178). Transcription factors are another example for genes with documented modulation at the exon level. Here, we found two well-known transcription factors, ZFP42 and NANOG, that regulate pluripotency and differentiation in the embryonic stem cells (179). For example, ZFP42, on the exon level, changes its chromatin organization (DNAse hypersensitivity) and is modulated by H3K4 trimethylation. This gene was shown to undergo changes on the exon level in early development (180,181) . Moreover, NANOG, a regulating transcription factor of the ZFP42 gene (182) which is also known to undergo alternative splicing (149), is also modulated by the same epigenetic mark, H3K4me3, on the exon level.

To complete our analysis, we finally searched for common positive correlations in the expression level of exons across tissues with PCC ≥0.7. Interestingly, we found that such constitutively expressed genes were usually not ubiquitously expressed across tissues. Additionally and as expected, coexpression was predominately found for highly similar tissues (lower left half of figure 6.9b), thus arguing against an important role of constitutive genes in development. The two genes that showed the largest number of abundant constitutive expression, CA2 and FOXO4, also showed the highest abundance in alternative splicing. Figure 6.9a-b shows a comparison of the numbers and allocation of positive and negative correlations of gene expression. The range of the number of genes in both matrices is similar on average. However, the anticorrelations involve mostly genes/gene clusters from early developmental stages. Figures 6.9c and 6.9d show the normalized expression of the exons contained in the set of genes that show anticorrelations and correlations in at least 26

combinations of tissues, respectively. In general, where changes do occur for exons in the anticorrelated genes, they do not occur at the level of the full genes. Rather specific exons are responsible for the variation, and the other exons of these genes are more constitutively expressed.

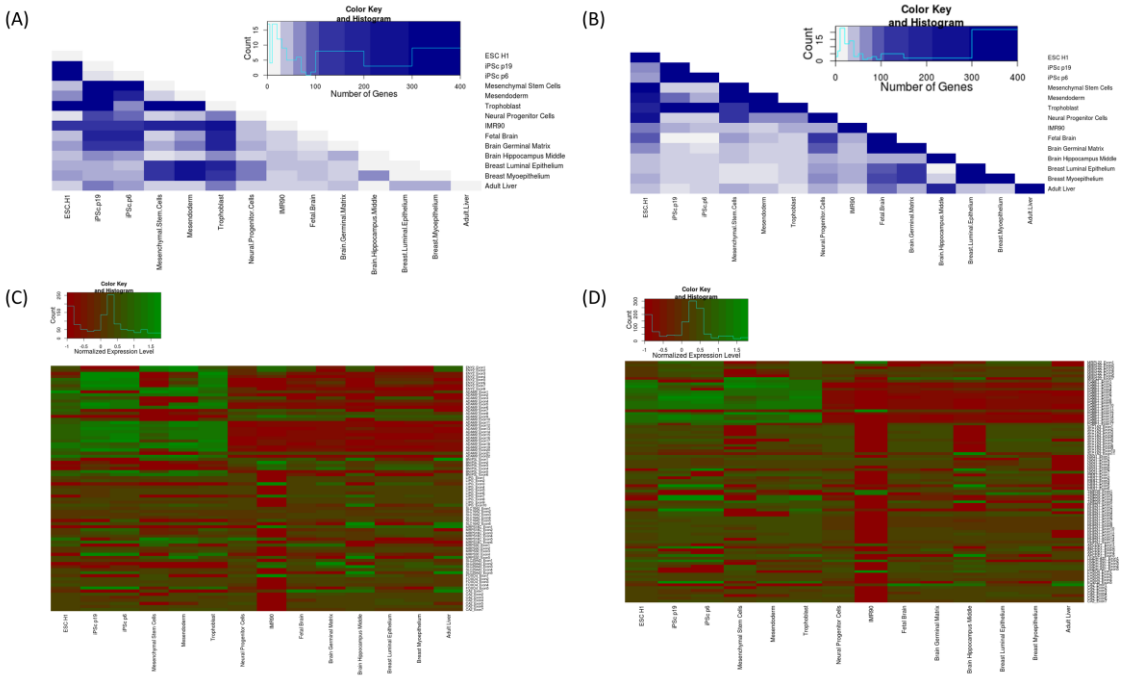


Figure 6.9 Heatmaps for the expression levels and gene numbers in pair-wise tissue correlations for (a) alternatively spliced genes, (b) constitutively expressed genes, (c) expression levels of individual exons belonging to 11 selected genes that show strong anticorrelation, and (d) expression levels of individual exons belonging to 10 selected genes that show strong positive correlation.

In conclusion, exon-intron boundaries set by histones/epigenetic marks are not only used to define the ends of the elements for the mRNA transcript to be expressed. Rather, they can also be considered as a part of a machinery for regulating and controlling the relative abundance of the several transcripts or protein isoforms that map to the same chromosomal region across tissues. This relationship seems to be most prominent in early developmental stages, and this suggests differential regulation across developmental stages, brought about by the distinct epi-genes. Additionally, exon-body epigenetic effect is more pronounced than that of intronic or promoter effects.

Chapter 7

Summary and Outlook

In this work, we aimed at understanding the role that DNA methylation, namely C5-cytosine and N6-adenine methylation, plays in the context of sequence stability, specificity and the differential stabilities of the different fiber forms of DNA in terms of biologically relevant sequences. Additionally, several reader proteins of the DNA methylation were considered. A global picture of the correlation between expression at one end and methylation/epigenetic marks at the other end was met.

First, we performed free energy calculations to investigate the effect of specific sequence context on the stability of different methylated/non-methylated DNA sequences. We found that the free energy of demethylation depends strongly on the sequence content that leads to small, but distinct structural variations of DNA and the conformation of coordinating ionic water. Methylation and DNA sequence content altogether seem to have substantial effects on the properties of water surrounding DNA, so that the specific sequence code appears to be tightly coordinated with its respective methylation status. We also found that specific DNA sequences and methylation forms have higher propensities to be transformed to the Z-DNA form.

After that, an investigation of the structural properties of the MeCP2-bound DNA was held. The effects of methylation, specific sequence context and salt concentration were studied and compared altogether. Conclusions were drawn in correlation with the expression results for the first reported human epigenetic switch in bacterial cell-free extract. This system provided a robust technique for a better understanding of the epigenetic modifications caused by DNA-bound protein.

Next, we extended structural and energetic calculations to investigate sequence specificity of binding for different proteins to sequences with the mentioned methylation forms. According to our results, specific binding to N6-adenine-methylated or C5-cytosine-methylated DNA is achieved through structural adaptation in DNA and protein on the one hand, and through the combined effects of more favorable binding enthalpies and modulation of the conformational entropy, on the other hand. For the R.DpnI system, a favorable enthalpic contribution seems to plays a major role in favoring binding of methylated DNA over the non-methylated DNA. Specific binding of C5-cytosine-methylated DNA to the MBD domain of the

Mecp2 protein, in contrast, appears to be predominately become stable by a favorable entropic contribution; which is the result of the concerted dynamics of protein and DNA.

Finally, we aimed at investigating the role differential methylation as well as the differential usage of the several histone marks play in terms of development. Our results showed that exon-intron boundaries set by methylation/histones/epigenetic marks not only define the ends of the elements for the mRNA transcript to be expressed, but they can also be considered as a part of a tackle that regulate the relative copiousness of the numerous transcripts/protein isoforms that map to the same chromosomal regions and across tissues. This relationship seems to be most prominent in early developmental stages. This also suggests differential regulation across developmental stages, brought about by the distinct epi-genes. Furthermore, we found that exon-body epigenetic effect is more conspicuous than that of intronic or promoter effects.

The work described in this thesis provides a further insight to understand the impact of DNA methylation as an epigenetic mark. As methylation targets specific sequence context, this comprehension can be further extended to understand why there lie some hot spots for targeting DNA methylation; while on the other hand some sequences are 'resistent' to the methylation marks. Structural transition in DNA fiber has been already encountered, but the effect of reader proteins to further enhance this transition was not accounted for. Further experimental setup, e.g., CD spectroscopy can help further understand this. Different methylation forms of DNA were found to play the energetics of specific binding differently in terms of the enthalpic and entropic contributions. This can be also extended to understand whether this is a general stamp. In addition, the global picture of differential exon usage has proven to be correlated with epigenetic marks across developmental stages. This can be further extended to cancer research such that more control on cancer can be established.

Finally, MD simulations and machine learning provide robust techniques to unravel some of Nature's exquisite mysteries, e.g., secrets of DNA methylation, when experiments fail to predict them. Lab work can be then linked to such *in silico* work to transfer knowledge to the real life.

References

1. Bird, A. (2007) Perceptions of epigenetics. *Nature*, **447**, 396-398.

2. Jones, P.A., Archer, T.K., Baylin, S.B., Beck, S., Berger, S., Bernstein, B.E., Carpten, J.D., Clark, S.J., Costello, J.F., Doerge, R.W. *et al.* (2008) Moving AHEAD with an international human epigenome project. *Nature*, **454**, 711-715.

3. Feingold, E.A., Good, P.J., Guyer, M.S., Kamholz, S., Liefer, L., Wetterstrand, K., Collins, F.S., Gingeras, T.R., Kampa, D., Sekinger, E.A. *et al.* (2004) The ENCODE (ENCyclopedia of DNA elements) Project. *Science*, **306**, 636-640.

4. Amaral, P.P. and Mattick, J.S. (2008) Noncoding RNA in development. *Mammalian Genome*, **19**, 454-492.

5. Whitehead, J., Pandey, G.K. and Kanduri, C. (2009) Regulation of the mammalian epigenome by long noncoding RNAs. *Biochimica Et Biophysica Acta-General Subjects*, **1790**, 936-947.

6. Zheng, B., Wang, Z., Li, S., Yu, B., Liu, J.-Y. and Chen, X. (2009) Intergenic transcription by RNA Polymerase II coordinates Pol IV and Pol V in siRNA-directed transcriptional gene silencing in Arabidopsis. *Genes & Development*, **23**, 2850-2860.

7. Teif, V.B. and Rippe, K. (2009) Predicting nucleosome positions on the DNA: combining intrinsic sequence preferences and remodeler activities. *Nucleic Acids Research*, **37**, 5641-5655.

8. Schwartz, S., Meshorer, E. and Ast, G. (2009) Chromatin organization marks exon-intron structure. *Nature Structural & Molecular Biology*, **16**, 990-U117.

9. Ratel, D., Ravanat, J.L., Berger, F. and Wion, D. (2006) N6-methyladenine: the other methylated base of DNA. *Bioessays*, **28**, 309-315.

10. Siwek, W., Czapinska, H., Bochtler, M., Bujnicki, J.M. and Skowronek, K. (2012) Crystal structure and mechanism of action of the N6-methyladenine-dependent type IIM restriction endonuclease R.DpnI. *Nucleic Acids Research*, **40**, 7563-7572.

11. Ashapkin, V.V., Kutueva, L.I. and Vanyushin, B.F. (2002) The gene for domains rearranged methyltransferase (DRM2) in Arabidopsis thaliana plants is methylated at both cytosine and adenine residues. *Febs Letters*, **532**, 367-372.

12. Pintortoro, J.A. (1987) Adenine methylation in zein genes. *Biochemical and Biophysical Research Communications*, **147**, 1082-1087.

13. Cummings, D.J., Tait, A. and Goddard, J.M. (1974) METHYLATED BASES IN DNA FROM PARAMECIUM-AURELIA. *Biochimica Et Biophysica Acta*, **374**, 1-11.

14. Aik, W., Scotti, J.S., Choi, H., Gong, L., Demetriades, M., Schofield, C.J. and McDonough, M.A. (2014) Structure of human RNA N-6-methyladenine demethylase ALKBH5 provides insights into its mechanisms of nucleic acid recognition and demethylation. *Nucleic Acids Research*, **42**, 4741-4754.

15. Meyer, K.D. and Jaffrey, S.R. (2014) The dynamic epitranscriptome: N-6-methyladenosine and gene expression control. *Nature Reviews Molecular Cell Biology*, **15**, 313-326.

16. Capuano, F., Muelleder, M., Kok, R., Blom, H.J. and Ralser, M. (2014) Cytosine DNA Methylation Is Found in Drosophila melanogaster but Absent in Saccharomyces cerevisiae, Schizosaccharomyces pombe, and Other Yeast Species. *Analytical Chemistry*, **86**, 3697-3702.

17. Seisenberger, S., Peat, J.R., Hore, T.A., Santos, F., Dean, W. and Reik, W. (2013) Reprogramming DNA methylation in the mammalian life cycle: building and breaking epigenetic barriers. *Philosophical Transactions of the Royal Society B-Biological Sciences*, **368**.

18. Iqbal, K., Jin, S.-G., Pfeifer, G.P. and Szabo, P.E. (2011) Reprogramming of the paternal genome upon fertilization involves genome-wide oxidation of 5-methylcytosine. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 3642-3647.

19. Okano, M., Bell, D.W., Haber, D.A. and Li, E. (1999) DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell*, **99**, 247-257.

20. Takeshita, K., Suetake, I., Yamashita, E., Suga, M., Narita, H., Nakagawa, A. and Tajima, S. (2011) Structural insight into maintenance methylation by mouse DNA methyltransferase 1 (Dnmt1). *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 9055-9059.

21. Curradi, M., Izzo, A., Badaracco, G. and Landsberger, N. (2002) Molecular mechanisms of gene silencing mediated by DNA methylation. *Molecular and Cellular Biology*, **22**, 3157-3173.

22. Ray, B.K., Dhar, S., Henry, C., Rich, A. and Ray, A. (2013) Epigenetic Regulation by Z-DNA Silencer Function Controls Cancer-Associated ADAM-12 Expression in Breast Cancer: Cross-talk between MeCP2 and NF1 Transcription Factor Family. *Cancer Research*, **73**, 736-744.

23. Bergman, Y. and Cedar, H. (2013) DNA methylation dynamics in health and disease (vol 20, pg 274, 2013). *Nature Structural & Molecular Biology*, **20**, 1236-1236.

24. Vaissiere, T., Sawan, C. and Herceg, Z. (2008) Epigenetic interplay between histone modifications and DNA methylation in gene silencing. *Mutation Research-Reviews in Mutation Research*, **659**, 40-48.

25. Bickle, T.A. and Kruger, D.H. (1993) Biology of DNA restriction. *Microbiological Reviews*, **57**, 434-450.

26. Hemavathy, K.C. and Nagaraja, V. (1995) DNA methylation in mycobacteria-absence of methylation at GATC (dam) and CCA/TGG (dcm) sequences. *Fems Immunology and Medical Microbiology*, **11**, 291-296.

27. Pettigrew, M.M., Gent, J.F., Revai, K., Patel, J.A. and Chonmaitree, T. (2008) Microbial interactions during upper respiratory tract infections. *Emerging Infectious Diseases*, **14**, 1584-1591.

28. Ho, K.L., McNae, L.W., Schmiedeberg, L., Klose, R.J., Bird, A.P. and Walkinshaw, M.D. (2008) MeCP2 binding to DNA depends upon hydration at methyl-CpG. *Molecular Cell*, **29**, 525-531.

29. Franklin, R.E. and Gosling, R.G. (1953) Evidence for 2-chain helix in crystalline structure of sodium deoxyribonucleate. *Nature*, **172**, 156-157.

30. Watson, J.D. and Crick, F.H.C. (1953) Molecular structure of nucleic acids- a structure for deoxyribose nucleic acid. *Nature*, **171**, 737-738.

31. Behe, M. and Felsenfeld, G. (1981) Effects of methylation on a synthetic polynucleotide- The B-Z transition in poly(DG-m5DC).poly(DG-m5DC). *Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences*, **78**, 1619-1623.

32. Nordheim, A. and Rich, A. (1983) Negative supercoiled simian virus-40 DNA contains Z-DNA segments within transcriptional enhancer sequences. *Nature*, **303**, 674-679.

33. Galindo-Murillo, R., Bergonzo, C. and Cheatham, T.E., 3rd. (2013) Molecular modeling of nucleic acid structure. *Current protocols in nucleic acid chemistry / edited by Serge L. Beaucage ... [et al.]*, **54**, Unit 7.5.-Unit 7.5.

34. Hartmann, B., Piazzola, D. and Lavery, R. (1993) BI-BII transitions in B-DNA. *Nucleic Acids Research*, **21**, 561-568.

35. Wecker, K., Bonnet, M.C., Meurs, E.F. and Delepierre, M. (2002) The role of the phosphorus BI-BII transition in protein-DNA recognition: the NF-kappa B complex. *Nucleic Acids Research*, **30**, 4452-4459.

36. Lu, X.J. and Olson, W.K. (2003) 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Research*, **31**, 5108-5121.

37. Alder, B.J. and Wainwright, T.E. (1959) Studies in molecular dynamics. 1. General method. *Journal of Chemical Physics*, **31**, 459-466.

38. Rahman, A. (1964) Correlations in motion of atoms in liquid Argon. *Physical Review a-General Physics*, **136**, A405-&.

39. Stilling.Fh and Rahman, A. (1974) Improved simulation of liquid water by molecular dynamics. *Journal of Chemical Physics*, **60**, 1545-1557.

40. McCammon, J.A., Gelin, B.R. and Karplus, M. (1977) Dynamics of folded proteins. *Nature*, **267**, 585-590.

41. Fayyad, U., PiatetskyShapiro, G. and Smyth, P. (1996) From data mining to knowledge discovery in databases. *Ai Magazine*, **17**, 37-54.

42. Schuster, S.C. (2008) Next-generation sequencing transforms today's biology. *Nature Methods*, **5**, 16-18.

43. Blankenberg, D., Gordon, A., Von Kuster, G., Coraor, N., Taylor, J., Nekrutenko, A. and Galaxy, T. (2010) Manipulation of FASTQ data with Galaxy. *Bioinformatics*, **26**, 1783-1785.

44. Wang, Z., Gerstein, M. and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, **10**, 57-63.

45. Velculescu, V.E., Vogelstein, B. and Kinzler, K.W. (2000) Analysing uncharted transcriptomes with SAGE. *Trends in Genetics*, **16**, 423-425.

46. Park, P.J. (2009) ChIP-seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, **10**, 669-680.

47. Boyle, A.P., Davis, S., Shulha, H.P., Meltzer, P., Margulies, E.H., Weng, Z., Furey, T.S. and Crawford, G.E. (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell*, **132**, 311-322.

48. Deng, J., Shoemaker, R., Xie, B., Gore, A., LeProust, E.M., Antosiewicz-Bourget, J., Egli, D., Maherali, N., Park, I.-H., Yu, J. *et al.* (2009) Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. *Nature Biotechnology*, **27**, 353-360.

49. Dupont, J.M., Tost, J., Jammes, H. and Gut, N.G. (2004) De novo quantitative bisulfite sequencing using the pyrosequencing technology. *Analytical Biochemistry*, **333**, 119-127.

50. Meissner, A., Gnirke, A., Bell, G.W., Ramsahoye, B., Lander, E.S. and Jaenisch, R. (2005) Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Research*, **33**, 5868-5877.

51. Taiwo, O., Wilson, G.A., Morris, T., Seisenberger, S., Reik, W., Pearce, D., Beck, S. and Butcher, L.M. (2012) Methylome analysis using MeDIP-seq with low DNA concentrations. *Nature Protocols*, **7**, 617-636.

52. Shapiro, R., Braverma.B, Louis, J.B. and Servis, R.E. (1973) Nucleic acid reactivity and conformations. 2. Reaction of cytosine and uracil with sodium bisulfite. *Journal of Biological Chemistry*, **248**, 4060-4064.

53. Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M.A., Beaudet, A.L., Ecker, J.R. *et al.* (2010) The NIH Roadmap Epigenomics Mapping Consortium. *Nature Biotechnology*, **28**, 1045-1048.

54. Zhou, X., Maricque, B., Xie, M., Li, D., Sundaram, V., Martin, E.A., Koebbe, B.C., Nielsen, C., Hirst, M., Farnham, P. *et al.* (2011) The Human Epigenome Browser at Washington University. *Nature Methods*, **8**, 989-990.

55. Zoghbi, H.Y., Amir, R.E., Wan, M., Lee, S.S., Van den Veyver, I.B., Tran, C.Q., Malicki, D., Schanen, N.C. and Francke, U. (2000) Rett syndrome is caused by mutations in the X-linked MECP2 gene encoding methyl-CpG-binding protein. *American Journal of Human Genetics*, **66**, 1723-1723.

56. Essen, H. (1977) The Physics of the Born-Oppenheimer Approximation. *International Journal of Quantum Chemistry*, **12**, 721-735.

57. Ponder, J.W. and Case, D.A. (2003) Force fields for protein simulations. *Protein Simulations*, **66**, 27-+.

58. Vanommeslaeghe, K., Hatcher, E., Acharya, C., Kundu, S., Zhong, S., Shim, J., Darian, E., Guvench, O., Lopes, P., Vorobyov, I. *et al.* (2010) CHARMM General Force Field: A Force Field for Drug-Like Molecules Compatible with the CHARMM All-Atom Additive Biological Force Fields. *Journal of Computational Chemistry*, **31**, 671-690.

59. Steiner, D. and van Gunsteren, W.F. (2010) Force-field development for computer simulation of biomolecular systems: The GROMOS case. *Abstracts of Papers of the American Chemical Society*, **240**.

60. Jiang, F., Zhou, C.-Y. and Wu, Y.-D. (2014) Residue-Specific Force Field Based on the Protein Coil Library. RSFF1: Modification of OPLS-AA/L. *The journal of physical chemistry. B*, **118**, 6983-6998.

61. Verlet, L. (1967) Computer experiments on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules. *Physical Review*, **159**, 98-&.

62. Swope, W.C., Andersen, H.C., Berens, P.H. and Wilson, K.R. (1982) A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules- application to small water clusters. *Journal of Chemical Physics*, **76**, 637-649.

63. Skeel, R.D. (1993) Variable step size destabilizes the stormer Leapfrog-Verlet method. *Bit*, **33**, 172-175.

64. Hess, B. (2008) P-LINCS: A parallel linear constraint solver for molecular simulation. *Journal of Chemical Theory and Computation*, **4**, 116-122.

65. Ryckaert, J.P., Ciccotti, G. and Berendsen, H.J.C. (1977) Numerical integration of cartesian equations of motion of a system with constraints- Molecular dynamics of n-alkanes. *Journal of Computational Physics*, **23**, 327-341.

66. Steinbach, P.J. and Brooks, B.R. (1994) New spherical cutoff methods for long-range forces in macromolecular simulation. *Journal of Computational Chemistry*, **15**, 667-683.

67. Nose, S. (2002) A molecular dynamics method for simulations in the canonical ensemble (Reprinted from Molecular Physics, vol 52, pg 255, 1984). *Molecular Physics*, **100**, 191-198.

68. Andersen, H.C. (1980) Molecular-dynamics simulations at constant pressure and-or temperature. *Journal of Chemical Physics*, **72**, 2384-2393.

69. Parrinello, M. and Rahman, A. (1980) Crystal structure and pair potentials- a molecular dynamics study. *Physical Review Letters*, **45**, 1196-1199.

70. Hoover, W.G., Ladd, A.J.C. and Moran, B. (1982) High-strain-rate plastic-flow studies via non-equilibrium molecular dynamics. *Physical Review Letters*, **48**, 1818-1820.

71. Zwanzig, R.W. (1954) High-temperature equation of state by a perturbation method. 1. nonpolar gases. *Journal of Chemical Physics*, **22**, 1420-1426.

72. Tembe, B.L. and McCammon, J.A. (1984) Ligand-receptor interactions. *Computers & Chemistry*, **8**, 281-283.

73. Straatsma, T.P. and McCammon, J.A. (1991) Multiconfiguration thermodynamic integration. *Journal of Chemical Physics*, **95**, 1175-1188.

74. Jarzynski, C. (1997) Equilibrium free-energy differences from nonequilibrium measurements: A master-equation approach. *Physical Review E*, **56**, 5018-5035.

75. Bennett, C.H. (1976) Efficient estimation of free-energy differences from Monte-Carlo data. *Journal of Computational Physics*, **22**, 245-268.

76. Beutler, T.C., Mark, A.E., Vanschaik, R.C., Gerber, P.R. and Vangunsteren, W.F. (1994) Avoiding singularities and numerical instabilities in free-energy calculations based on molecular simulations. *Chemical Physics Letters*, **222**, 529-539.

77. Gapsys, V., Seeliger, D. and de Groot, B.L. (2012) New Soft-Core Potential Function for Molecular Dynamics Based Alchemical Free Energy Calculations. *Journal of Chemical Theory and Computation*, **8**, 2373-2382.

78. Darden, T., York, D. and Pedersen, L. (1993) Particle Mesh Ewald- an n.log(n) method for Ewald sums in large systems. *Journal of Chemical Physics*, **98**, 10089-10092.

79. Gilson, M.K., Given, J.A., Bush, B.L. and McCammon, J.A. (1997) The statistical-thermodynamic basis for computation of binding affinities: A critical review. *Biophysical Journal*, **72**, 1047-1069.

80. Jen-Jacobson, L., Engler, L.E. and Jacobson, L.A. (2000) Structural and thermodynamic strategies for site-specific DNA binding proteins (vol 8, pg 1015, 2000). *Structure*, **8**, U7-U7.

81. Dunitz, J.D. (1995) WIN SOME, LOSE SOME - ENTHALPY-ENTROPY COMPENSATION IN WEAK INTERMOLECULAR INTERACTIONS. *Chemistry & Biology*, **2**, 709-712.

82. Schlitter, J. (1993) Estimation of absolute and relative entropies of macromolecules using the covariance matrix. *Chemical Physics Letters*, **215**, 617-621.

83. Numata, J., Wan, M. and Knapp, E.-W. (2007) Conformational entropy of biomolecules: beyond the quasi-harmonic approximation. *Genome informatics. International Conference on Genome Informatics*, **18**, 192-205.

84. Anders, S. and Huber, W. (2010) Differential expression analysis for sequence count data. *Genome Biology*, **11**.

85. Akalin, A., Kormaksson, M., Li, S., Garrett-Bakelman, F.E., Figueroa, M.E., Melnick, A. and Mason, C.E. (2012) methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biology*, **13**.

86. Shao, Z., Zhang, Y., Yuan, G.-C., Orkin, S.H. and Waxman, D.J. (2012) MAnorm: a robust model for quantitative comparison of ChIP-Seq data sets. *Genome Biology*, **13**.

87. Frigui, H. and Krishnapuram, R. (1997) Clustering by competitive agglomeration. *Pattern Recognition*, **30**, 1109-1119.

88. Olson, C.F. (1995) Parallel algorithms for hierarchical clustering. *Parallel Computing*, **21**, 1313-1325.

89. Frohlich, H., Speer, N., Poustka, A. and Beissarth, T. (2007) GOSim - an R-package for computation of information theoretic GO similarities between terms and gene products. *Bmc Bioinformatics*, **8**.

90. Couto, F.M. and Silva, M.J. (2011) Disjunctive shared information between ontology concepts: application to Gene Ontology. *Journal of biomedical semantics*, **2**, 5-5.

91. Shanak, S. and Helms, V. (2014). J Chem Phys, Vol. 141, pp. 22D512.

92. Fatemi, M., Pao, M.M., Jeong, S., Gal-Yam, E.N., Egger, G., Weisenberger, D.J. and Jones, P.A. (2005) Footprinting of mammalian promoters: use of a CpG DNA methyltransferase revealing nucleosome positions at a single molecule level. *Nucleic Acids Research*, **33**.

93. Wilson, G.G. (1991) Organization of restriction-modification systems. *Nucleic Acids Research*, **19**, 2539-2566.

94. Lacks, S.A., Mannarelli, B.M., Springhorn, S.S. and Greenberg, B. (1986) Genetic-bases of the complementary DpnI and DpnII restriction systems of Streptococcus pneumonia- an intercellular cassette mechanism. *Cell*, **46**, 993-1000.

95. Liu, J., Yue, Y., Han, D., Wang, X., Fu, Y., Zhang, L., Jia, G., Yu, M., Lu, Z., Deng, X. *et al.* (2014) A METTL3-METTL14 complex mediates mammalian nuclear RNA N-6-adenosine methylation. *Nature Chemical Biology*, **10**, 93-95.

96. Kypr, J., Kejnovska, I., Renciuk, D. and Vorlickova, M. (2009) Circular dichroism and conformational polymorphism of DNA. *Nucleic Acids Research*, **37**, 1713-1725.

97. Liu, H., Mulholland, N., Fu, H.Q. and Zhao, K. (2006) Cooperative activity of BRG1 and Z-DNA formation in chromatin remodeling. *Molecular and Cellular Biology*, **26**, 2550-2559.

98. Braaten, B.A., Nou, X.W., Kaltenbach, L.S. and Low, D.A. (1994) Methylation patterns in Pap regulatory DNA control pyelonephritis-associated pili phase variation in Escherichia-coli. *Cell*, **76**, 577-588.

99. Chen, W.G., Chang, Q., Lin, Y.X., Meissner, A., West, A.E., Griffith, E.C., Jaenisch, R. and Greenberg, M.E. (2003) Derepression of BDNF transcription involves calcium-dependent phosphorylation of MeCP2. *Science*, **302**, 885-889.

100. Pruunsild, P., Kazantseva, A., Aid, T., Palm, K. and Timmusk, T. (2007) Dissecting the human BDNF locus: Bidirectional transcription, complex splicing, and multiple promoters. *Genomics*, **90**, 397-406.

101. Furmanchuk, A., Shishkin, O.V., Isayev, O., Gorb, L. and Leszczynski, J. (2010) New insight on structural properties of hydrated nucleic acid bases from ab initio molecular dynamics. *Physical Chemistry Chemical Physics*, **12**, 9945-9954.

102. Mayer-Jung, C., Moras, D. and Timsit, Y. (1998) Hydration and recognition of methylated CpG steps in DNA. *Embo Journal*, **17**, 2709-2718.

103. Hess, B., Kutzner, C., van der Spoel, D. and Lindahl, E. (2008) GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *Journal of Chemical Theory and Computation*, **4**, 435-447.

104. Foloppe, N. and MacKerell, A.D. (2000) All-atom empirical force field for nucleic acids: I. Parameter optimization based on small molecule and condensed phase macromolecular target data. *Journal of Computational Chemistry*, **21**, 86-104.

105. Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W. and Klein, M.L. (1983) Comparison of simple potential functions for simulating liquid water. *Journal of Chemical Physics*, **79**, 926-935.

106. Frisch, M.J., Trucks, G.W., Schlegel, H.B., Scuseria, G.E., Robb, M.A., Cheeseman, J.R., Montgomery, J., J. A., Vreven, T., Kudin, K.N., Burant, J.C. *et al.* (2004).

107. Lu, X.-J. and Olson, W.K. (2008) 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. *Nature Protocols*, **3**, 1213-1227.

108. Van Gunsteren, W.F. and Berendsen, H.J.C. (1988) A leap-frog algorithm for stochastic dynamics. *Molecular Simulation*, **1**, 173-185.

109. Pohorille, A., Jarzynski, C. and Chipot, C. (2010) Good Practices in Free-Energy Calculations. *Journal of Physical Chemistry B*, **114**, 10235-10253.

110. Mobley, D.L., Chodera, J.D. and Dill, K.A. (2006) On the use of orientational restraints and symmetry corrections in alchemical free energy calculations. *Journal of Chemical Physics*, **125**.

111. Hornak, V. and Simmerling, C. (2004) Development of softcore potential functions for overcoming steric barriers in molecular dynamics simulations. *Journal of Molecular Graphics & Modelling*, **22**, 405-413.

112. Eisenhaber, F., Lijnzaad, P., Argos, P., Sander, C. and Scharf, M. (1995) The double cubic lattice method- efficient approaches to numerical-integration of surface area and volume and to dot surface contouring of molecular assemblies. *Journal of Computational Chemistry*, **16**, 273-284.

113. Kovacs, H., Mark, A.E. and vanGunsteren, W.F. (1997) Solvent structure at a hydrophobic protein surface. *Proteins-Structure Function and Genetics*, **27**, 395-404.

114. Kumar, P., Buldyrev, S.V. and Stanley, H.E. (2009) A tetrahedral entropy for water. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 22130-22134.

115. Jana, B., Pal, S. and Bagchi, B. (2010) Enhanced Tetrahedral Ordering of Water Molecules in Minor Grooves of DNA: Relative Role of DNA Rigidity, Nanoconfinement, and Surface Specific Interactions. *Journal of Physical Chemistry B*, **114**, 3633-3638.

116. Schenkelberger, M. (2014) *Cooperative Biomolecular Binding: High specificity of competitive single stranded DNA hybridization, influence of DNA methylation on the duplexstability, and epigenetic regulation of in vitro gene switches*. PhD Thesis. Saarland University: Germany.

117. Shin, J. and Noireaux, V. (2010) Efficient cell-free expression with the endogenous E. Coli RNA polymerase and sigma factor 70. *Journal of biological engineering*, **4**, 8-8.

118. Kahramanoglou, C., Prieto, A.I., Khedkar, S., Haase, B., Gupta, A., Benes, V., Fraser, G.M., Luscombe, N.M. and Seshasayee, A.S.N. (2012) Genomics of DNA cytosine methylation in Escherichia coli reveals its role in stationary phase transcription. *Nature Communications*, **3**.

119. Khrapunov, S., Warren, C., Cheng, H., Berko, E.R., Greally, J.M. and Brenowitz, M. (2014) Unusual Characteristics of the DNA Binding Domain of Epigenetic Regulatory Protein MeCP2 Determine Its Binding Specificity. *Biochemistry*, **53**, 3379-3391.

120. Klose, R.J., Sarraf, S.A., Schmiedeberg, L., McDermott, S.M., Stancheva, I. and Bird, A.P. (2005) DNA binding selectivity of MeCP2 due to a requirement for A/T sequences adjacent to methyl-CpG. *Molecular Cell*, **19**, 667-678.

121. Lewis, J.D., Meehan, R.R., Henzel, W.J., Maurerfogy, I., Jeppesen, P., Klein, F. and Bird, A. (1992) Purification, sequences, and cellular localization of a novel chromosomal protein that binds to methylated DNA. *Cell*, **69**, 905-914.

122. Lopez-Serra, L., Ballestar, E., Fraga, M.F., Alaminos, M., Setien, F. and Esteller, M. (2006) A profile of methyl-CpG binding domain protein occupancy of hypermethylated promoter CpG islands of tumor suppressor genes in human cancer. *Cancer Research*, **66**, 8342-8346.

123. Zou, X., Ma, W., Solov'yov, I.A., Chipot, C. and Schulten, K. (2012) Recognition of methylated DNA through methyl-CpG binding domain proteins. *Nucleic Acids Research*, **40**, 2747-2758.

124. Kigawa, T., Yabuki, T., Matsuda, N., Matsuda, T., Nakajima, R., Tanaka, A. and Yokoyama, S. (2004) Preparation of Escherichia coli cell extract for highly productive cell-free protein expression. *Journal of structural and functional genomics*, **5**, 63-68.

125. Zheng, G., Lu, X.-J. and Olson, W.K. (2009) Web 3DNA-a web server for the analysis, reconstruction, and visualization of three-dimensional nucleic-acid structures. *Nucleic Acids Research*, **37**, W240-W246.

126. Moradi, M., Babin, V., Roland, C. and Sagui, C. (2013) Reaction path ensemble of the B-Z-DNA transition: a comprehensive atomistic study. *Nucleic Acids Research*, **41**, 33-43.

127. Pabo, C.O. and Sauer, R.T. (1984) Protein-DNA recognition. *Annual Review of Biochemistry*, **53**, 293-321.

128. Gowher, H., Leismann, O. and Jeltsch, A. (2000) DNA of Drosophila melanogaster contains 5-methylcytosine. *Embo Journal*, **19**, 6918-6923.

129. Cheng, Y.H., Korolev, N. and Nordenskiold, L. (2006) Similarities and differences in interaction of K+ and Na+ with condensed ordered DNA. A molecular dynamics computer simulation study. *Nucleic Acids Research*, **34**, 686-696.

130. Terry, C.A., Fernandez, M.-J., Gude, L., Lorente, A. and Grant, K.B. (2011) Physiologically Relevant Concentrations of NaCl and KCl Increase DNA Photocleavage by an N-Substituted 9-Aminomethylanthracene Dye. *Biochemistry*, **50**, 10375-10389.

131. Moore, R.D. and Morrill, G.A. (1976) Possible mechanism for concentrating sodium and potassium in cell-nucleus. *Biophysical Journal*, **16**, 527-533.

132. Low, D.A., Weyand, N.J. and Mahan, M.J. (2001) Roles of DNA adenine methylation in regulating bacterial gene expression and virulence. *Infection and Immunity*, **69**, 7197-7204.

133. Delacampa, A.G., Springhorn, S.S., Kale, P. and Lacks, S.A. (1988) Proteins encoded by DpnI restriction gene cassette- hyperproduction and characterization of the DpnI endonuclease. *Journal of Biological Chemistry*, **263**, 14696-14702.

134. Mierzejewska, K., Siwek, W., Czapinska, H., Skowronek, K., Bujnicki, J. and Bochtler, M. (2014) Structural basis of the methylation specificity of R.DpnI. *Nucl. Acids Res.*, **42**, 8745-8754.

135. Buck-Koehntop, B.A., Stanfield, R.L., Ekiert, D.C., Martinez-Yamout, M.A., Dyson, H.J., Wilson, I.A. and Wright, P.E. (2012) Molecular basis for recognition of methylated and specific DNA sequences by the zinc finger protein Kaiso. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 15229-15234.

136. Srinivasan, J., Cheatham, T.E., Cieplak, P., Kollman, P.A. and Case, D.A. (1998) Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate - DNA helices. *Journal of the American Chemical Society*, **120**, 9401-9409.

137. Baker, N.A., Sept, D., Holst, M.J. and McCammon, J.A. (2001) The adaptive multilevel finite element solution of the Poisson-Boltzmann equation on massively parallel computers. *Ibm Journal of Research and Development*, **45**, 427-438.

138. Miller, B.R., III, McGee, T.D., Jr., Swails, J.M., Homeyer, N., Gohlke, H. and Roitberg, A.E. (2012) MMPBSA.py: An Efficient Program for End-State Free Energy Calculations. *Journal of Chemical Theory and Computation*, **8**, 3314-3321.

139. Crowley, M.F., Williamson, M.J. and Walker, R.C. (2009) CHAMBER: Comprehensive Support for CHARMM Force Fields Within the AMBER Software. *International Journal of Quantum Chemistry*, **109**, 3767-3772.

140. Humphrey, W., Dalke, A. and Schulten, K. (1996) VMD: Visual molecular dynamics. *Journal of Molecular Graphics & Modelling*, **14**, 33-38.

141. Roe, D.R. and Cheatham, T.E., III. (2013) PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *Journal of Chemical Theory and Computation*, **9**, 3084-3095.

142. Furini, S., Barbini, P. and Domene, C. (2013) DNA-recognition process described by MD simulations of the lactose repressor protein on a specific and a non-specific DNA sequence. *Nucleic Acids Research*, **41**, 3963-3972.

143. Pauling, L. (1992) The nature of chemical bond. *Journal of Chemical Education*, **69**, 519-521.

144. Liu, Y., Toh, H., Sasaki, H., Zhang, X. and Cheng, X. (2012) An atomic model of Zfp57 recognition of CpG methylation within a specific DNA sequence. *Genes & Development*, **26**, 2374-2379.

145. Rohs, R., Jin, X., West, S.M., Joshi, R., Honig, B. and Mann, R.S. (2010) Origins of Specificity in Protein-DNA Recognition. *Annual Review of Biochemistry, Vol 79*, **79**, 233-269.

146. Mironov, A.A., Fickett, J.W. and Gelfand, M.S. (1999) Frequent alternative splicing of human genes. *Genome Research*, **9**, 1288-1293.

147. Koscielny, G., Le Texier, V., Gopalakrishnan, C., Kumanduri, V., Riethoven, J.-J., Nardone, F., Stanley, E., Fallsehr, C., Hofmann, O., Kull, M. *et al.* (2009) ASTD: The Alternative Splicing and Transcript Diversity database. *Genomics*, **93**, 213-220.

148. Nilsen, T.W. and Graveley, B.R. (2010) Expansion of the eukaryotic proteome by alternative splicing. *Nature*, **463**, 457-463.

149. Das, S., Jena, S. and Levasseur, D.N. (2011) Alternative Splicing Produces Nanog Protein Variants with Different Capacities for Self-renewal and Pluripotency in Embryonic Stem Cells. *Journal of Biological Chemistry*, **286**, 42690-42703.

150. Kim, J., Noskov, V.N., Li, X.C., Bergmann, A., Ren, X.J., Warth, T., Richardson, P., Kouprina, N. and Stubbs, L. (2000) Discovery of a novel, paternally expressed ubiquitin-specific processing protease gene through comparative analysis of an imprinted region of mouse chromosome 7 and human chromosome 19q13.4. *Genome Research*, **10**, 1138-1147.

151. Kim, J., Bergmann, A., Lucas, S., Stone, R. and Stubbs, L. (2004) Lineage-specific imprinting and evolution of the zinc-finger gene ZIM2. *Genomics*, **84**, 47-58.

152. Reyes, A., Anders, S., Weatheritt, R.J., Gibson, T.J., Steinmetz, L.M. and Huber, W. (2013) Drift and conservation of differential exon usage across tissues in primate species. *Proceedings of the National Academy of Sciences of the United States of America*, **110**, 15377-15382.

153. Zhou, H.-L., Luo, G., Wise, J.A. and Lou, H. (2014) Regulation of alternative splicing by local histone modifications: potential roles for RNA-guided mechanisms. *Nucleic Acids Research*, **42**, 701-713.

154. Harris, R.A., Wang, T., Coarfa, C., Nagarajan, R.P., Hong, C., Downey, S.L., Johnson, B.E., Fouse, S.D., Delaney, A., Zhao, Y. *et al.* (2010) Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nature Biotechnology*, **28**, 1097-U1194.

155. Milosavljevic, A. (2010) Putting epigenome comparison into practice. *Nature Biotechnology*, **28**, 1053-1056.

156. Milosavljevic, A. (2011) Emerging patterns of epigenomic variation. *Trends in Genetics*, **27**, 242-250.

157. Anders, S., Reyes, A. and Huber, W. (2012) Detecting differential usage of exons from RNA-seq data. *Genome Research*, **22**, 2008-2017.

158. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nussbaum, C., Myers, R.M., Brown, M., Li, W. *et al.* (2008) Model-based Analysis of ChIP-Seq (MACS). *Genome Biology*, **9**.

159. Lienhard, M., Grimm, C., Morkel, M., Herwig, R. and Chavez, L. (2014) MEDIPS: genome-wide differential coverage analysis of sequencing data derived from DNA enrichment experiments. *Bioinformatics*, **30**, 284-286.

160. Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841-842.

161. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and Genome Project Data, P. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078-2079.

162. Chandramohan, R., Wu, P.-Y., Phan, J.H. and Wang, M.D. (2013) Benchmarking RNA-Seq quantification tools. *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference*, **2013**, 647-650.

163. Strahl, B.D., Grant, P.A., Briggs, S.D., Sun, Z.W., Bone, J.R., Caldwell, J.A., Mollah, S., Cook, R.G., Shabanowitz, J., Hunt, D.F. *et al.* (2002) Set2 is a nucleosomal histone H3-selective methyltransferase that mediates transcriptional repression. *Molecular and Cellular Biology*, **22**, 1298-1306.

164. Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P. and Burge, C.B. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470-476.

165. Ridley, A.J. (2006) Rho GTPases and actin dynamics in membrane protrusions and vesicle trafficking. *Trends in Cell Biology*, **16**, 522-529.

166. Ellenbroek, S.I.J. and Collard, J.G. (2007) Rho GTPases: functions and association with cancer. *Clinical & Experimental Metastasis*, **24**, 657-672.

167. Heldin, C.H., Miyazono, K. and tenDijke, P. (1997) TGF-beta signalling from cell membrane to nucleus through SMAD proteins. *Nature*, **390**, 465-471.

168. Gupta, G., Bansal, M. and Sasisekharan, V. (1980) Conformational flexibility of DNA- Polymorphism and handedness. *Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences*, **77**, 6486-6490.

169. Luk, J.M. and Guan, K.-L. (2014) An alternative DNA damage pathway to apoptosis in hematological cancers. *Nature Medicine*, **20**, 587-588.

170. Pan, D. (2007) Hippo signaling in organ size control. *Genes & Development*, **21**, 886-897.

171. Takeuchi, O., Sato, S., Horiuchi, T., Hoshino, K., Takeda, K., Dong, Z.Y., Modlin, R.L. and Akira, S. (2002) Cutting edge: Role of Toll-like receptor 1 in mediating immune response to microbial lipoproteins. *Journal of Immunology*, **169**, 10-14.

172. Karpova, N.N. (2014) Role of BDNF epigenetics in activity-dependent neuronal plasticity. *Neuropharmacology*, **76**, 709-718.

173. Ikegame, T., Bundo, M., Sunaga, F., Asai, T., Nishimura, F., Yoshikawa, A., Kawamura, Y., Hibino, H., Tochigi, M., Kakiuchi, C. *et al.* (2013) DNA methylation analysis of BDNF gene promoters in peripheral blood cells of schizophrenia patients. *Neuroscience Research*, **77**, 208-214.

174. Gray, T.A., Saitoh, S. and Nicholls, R.D. (1999) An imprinted, mammalian bicistronic transcript encodes two independent proteins. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 5616-5621.

175. Burgers, W.A., Fuks, F. and Kouzarides, T. (2002) DNA methyltransferases get connected to chromatin. *Trends in Genetics*, **18**, 275-277.

176. Aapola, U., Liiv, I. and Peterson, P. (2002) Imprinting regulator DNMT3L is a transcriptional repressor associated with histone deacetylase activity. *Nucleic Acids Research*, **30**, 3602-3608.

177. Hata, K., Okano, M., Lei, H. and Li, E. (2002) Dnmt3L cooperates with the Dnmt3 family of de novo DNA methyltransferases to establish maternal imprints in mice. *Development*, **129**, 1983-1993.

178. Simonis, M., Atanur, S.S., Linsen, S., Guryev, V., Ruzius, F.-P., Game, L., Lansu, N., de Bruijn, E., van Heesch, S., Jones, S.J.M. *et al.* (2012) Genetic basis of transcriptome differences between the founder strains of the rat HXB/BXH recombinant inbred panel. *Genome Biology*, **13**.

179. Kashyap, V., Rezende, N.C., Scotland, K.B., Shaffer, S.M., Persson, J.L., Gudas, L.J. and Mongan, N.P. (2009) Regulation of Stem Cell Pluripotency and Differentiation Involves a Mutual Regulatory Circuit of the Nanog, OCT4, and SOX2 Pluripotency Transcription Factors With Polycomb Repressive Complexes and Stem Cell microRNAs. *Stem Cells and Development*, **18**, 1093-1108.

180. Cloonan, N., Forrest, A.R.R., Kolle, G., Gardiner, B.B.A., Faulkner, G.J., Brown, M.K., Taylor, D.F., Steptoe, A.L., Wani, S., Bethel, G. *et al.* (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nature Methods*, **5**, 613-619.

181. Yeo, G.W., Xu, X., Liang, T.Y., Muotri, A.R., Carson, C.T., Coufal, N.G. and Gage, F.H. (2007) Alternative splicing events identified in human embryonic stem cells and neural progenitors. *Plos Computational Biology*, **3**, 1951-1967.

182. Shi, W., Wang, H., Pan, G., Geng, Y., Guo, Y. and Pei, D. (2006) Regulation of the pluripotency marker Rex-1 by Nanog and Sox2. *Journal of Biological Chemistry*, **281**, 23319-23325.