GÖTEBORGS UNIVERSITET

Unravelling Mechanistic Insights of Eukaryotic Transcriptional Control

Johanna Hörberg

Department of Chemistry and Molecular Biology
Gothenburg 2022

Thesis for degree of Doctor of Philosophy in Natural Science, Chemistry

Unravelling Mechanistic insights of Eukaryotic Transcriptional Control

Department of Chemistry and Molecular Biology
University of Gothenburg SE-405 30, Göteborg, Sweden

ISBN: 978-91-8009-993-6 (PRINT)
ISBN: 978-91-8009-994-3 (PDF)

Printed by Stema Specialtryck AB, Borås, Sweden, 2022

Abstract

DNA transcription involves the association of multiple proteins with DNA to convert the genetic information into messenger-RNA transcripts, which later will be translated into proteins. Controlled regulation of transcription is imperative for preserving the healthy and homeostatic state of cells. In eukaryotes, where the accessibility of DNA is limited by the chromatin packaging, transcriptional control occurs primarily at the initiation stage. The initiation of transcription is regulated by a class of proteins termed transcription factors, which recognise and bind with high specificity short non-coding DNA fragments, so-called response elements, to modulate the recruitment of the transcription machinery. The mapping of how transcription factors achieve binding specificity for their genomic target sites, shielded by the massive amount of non-specific DNA, remains incomplete despite decades of dedicated research. Over the years, traditional structure determination techniques have provided multiple insights into specificity of transcription factors-DNA interactions, though from a more static perspective. We still lack information on why transcription factors bind related sequences with different affinity and unrelated sequences with similar affinity. Furthermore, the response elements occur more frequently in the genome than the actual genes regulated by the transcription factors. To understand these aspects of transcriptional control, it is imperative also to study the dynamic aspects of the transcription factors-DNA interactions. Until relatively recently, DNA has been looked at as a rather passive component in the recognition process. However, DNA is highly polymorphic and its flexibility is sequence specific, allowing DNA to adapt in response to various conditions. In addition, DNA state constantly changes during the life time of a cell as a consequence of DNA supercoiling, the predominant regulatory force of transcriptional control. DNA supercoiling, introduced by molecular motors, can propagate along the chromatin fibre, altering the accessibility of the genetic code, which in turn could affect the binding of transcription factors. Taken together, reveals DNA as the key driver of many biological processes. In this thesis, I use computational methods, to unravel mechanistic details of eukaryotic transcriptional control. I study naked DNA and DNA interactions with various transcription factors, in particular BZIP- and BHTH-factors. In my quest, I have uncovered the mechanism of how DNA epigenetic modifications and transcription factors binding modulate the torsional rigidity of DNA; and how these properties can be utilised in the regulation of transcription. I have also contributed to the understanding of how transcription factors exploit the local sequence specific plasticity of DNA in recognition of their genomic target sites, and how small alterations in DNA flexibility due to DNA methylation could make their targets sites unrecognisable. The major conclusion from this thesis work, is that the enigma of selective binding of transcription factors has its solution in the dynamics of transcription factor-DNA complexation combined with sequence specific DNA flexibility of response elements but also their flanking sites. The flanking sites modulate the conformational adaptability of the response elements, making them more shape-complementary for the transcription factor to form specific contacts, which increases the stability of the complex. The level of conformational adaptability of the DNA response elements to the binding by transcription factor, could modulate the degree of torsional rigidity of DNA to undergo supercoiling transitions, which in turn may modulate the recruitment of different collaborative transcription factors control how long gene promoters stay open and consequently being transcribed.

Keywords: Transcriptional control, gene expression regulation, DNA supercoiling, DNA methylation, DNA-protein recognition, basic-leucine-zipper transcription factors, basic-helix-loop-helix transcription factors, sequence specific DNA plasticity, transcription factor cooperativity, cancer, molecular modelling, molecular dynamics.

Sammanfattning

DNA transkription involverar en association av flertal proteiner med DNA för att översätta genetisk information till budbärar-RNA (mRNA) som sedan används som mall för proteinsyntes. Kontrollerad transkriptionsreglering är essentiellt för bevarandet av cellers hälsosamma och homeostatiska tillstånd. Då DNA:s tillgänglighet begränsas i eukaryoter av dess packning till kromatin sker transkriptionsreglering främst vid initieringsfasen. Transkriptionsinitiering regleras av transkriptionsfaktorer, en klass proteiner, som genom att binda med hög specificitet till korta icke-kodande DNA fragment, så-kallade svarselement, modulerar tillkallandet av transkriptionsmaskineriet. Trots årtionden av dedikerad forskning är kartläggningen för hur transkriptionsfaktorer uppnår specificitet för sina genomiska bindningsställen, gömda av den enorma mängden icke-specifikt DNA, fortfarande ofullständig. Genom åren har traditionella strukturbestämningstekniker bidragit med flertal mekanistiska svar på interaktionsspecificitet för transkriptionsfaktor-DNA komplex – däremot från ett statiskt perspektiv. Vi saknar fortfarande svar på hur transkriptionsfaktorer kan binda orelaterade sekvenser med liknande affiniteteter men relaterade sekvenser med olika affiniteter. Dessutom så förekommer svarselementen mer frekvent i genomet än de gener som faktiskt regleras av en specifik transkriptionsfaktor. För att förstår dessa regleringsaspekter av transkription behöver vi också studera dynamiska faktorer för transkriptionsfaktor-DNA-interaktioner. Tills relativt nyligen har DNA setts som en ganska passiv deltagare i igenkänningsprocessen. Däremot är DNA en mycket polymorf molekyl vars flexibilitet är mycket sekvensspecifik som tillåter DNA att anpassa sig vid olika förhållanden. Dessutom förändas DNA:s tillstånd konstant under en cells livstid som en konsekvens av närvarandet av DNA supercoiling; den dominerande regleringskraften for transkriptionskontroll. DNA supercoiling uppkommer genom verkan av molekylära motorer och kan propagera längst kromatinfibrerna och på så sätt påverka tillgängligheten av den genetiska koden, vilket sin tur kan påverka bindningen av transkriptionsfaktorer. Detta tyder på att DNA kan vara drivkraften för många biologiska processer. I denna avhandlingen använder jag beräkningsmetoder för att bidra med mekanistiska detaljer relaterande transkriptionskontroll i eukaryoter. Jag studerar både naket DNA och DNA interaktioner med olika transkriptionsfaktorer (speciellt BZIP och BHLH-faktorer). Jag har bidragit med mekanismer för hur DNA epigenetiska modifieringar och bindningen av transkriptionsfaktorer modulerar vridstyvheten hos DNA, och hur dessa egenskaper kan användas vid transkriptionsreglering. Jag har också bidragit till förståelsen av hur transkriptionsfaktorer utnyttjar den lokala sekvensspecifika plasticiteten hos DNA för att känna igen deras genomiska bindningsställen, och hur små förändringar i DNA-flexibilitet på grund av DNA-metylering kan göra deras målplatser oigenkännliga. Den viktigaste slutsatsen från detta examensarbete är att selektiv bindning av transkriptionsfaktorer har sitt svar inom dynamiken i transkriptionsfaktor-DNA-komplexbildning kombinerat med sekvensspecifik DNA-flexibilitet hos svarselementen men också deras flankerande ställen. De flankerande platserna modulerar svarselementets anpassningsförmåga genom att göra den mer formkomplementär till transkriptionsfaktorn som då kan bilda gynnsamma specifika kontakter, vilket ökar komplexets stabilitet. Nivån för svarselementets anpassningsförmåga till transkriptionsfaktorn kan modulera granden av vridstyvhet hos DNA för att genomgå supercoiling-övergångar, vilket i sin tur kan modulera rekryteringen av olika kollaborativa transkriptionsfaktorer för att styra hur länge genpromotorer förblir öppna och följaktligen transkriberas.

Nyckelord: Transkriptionskontroll, genuttrycksreglering, DNA-supercoiling, DNA-metylering, DNA-proteinigenkänning, basic-leucine-zipper-transkriptionsfaktorer, basic-helix-loop-helix-transkriptionsfaktorer, sekvensspecifik DNA-plasticitet, transkriptionsfaktorkooperativitet, cancer, molekylärmodellering, molekylärdynamik

List of Publications

Paper I
Hörberg, J., Moreau, K., Tamás, M. J. & Reymer, A. Sequence-specific dynamics of DNA response elements and their flanking sites regulate the recognition by AP-1 transcription factors. *Nucleic Acids Res.* **49**, 9280-9293 (2021).

Contribution: Helped designing the study, located the native Yap1-binding sites, performed the homology modelling and protein-DNA docking. Ran the molecular dynamics simulations and performed the analyses of trajectories and available crystallographic structures. Contributed to the writing of the manuscript.

Paper II
Hörberg, J. & Reymer, A. A sequence environment modulates the impact of methylation on the torsional rigidity of DNA. *Chem. Commun*. **54**, 11885-11888 (2018).

Contribution: Helped designing the study, performed all molecular dynamics simulations and analyses. Contributed to the writing of the manuscript.

Paper III
Hörberg, J. & Reymer, A. Specifically bound BZIP transcription factors modulate DNA supercoiling transitions. *Sci. Rep.* **10**, 1-10 (2020).

Contribution: Helped designing the study, performed all molecular dynamics simulations and analyses. Contributed to the writing of the manuscript.

Paper IV
Hörberg, J., Moreau, K. & Reymer, A. Homologous BHLH transcription factors induce distinct deformations of torsionally-stressed DNA: a potential transcription regulation mechanism. *QRB Discovery,* **3**, e4 (2022).

Contribution: Helped designing the study, performed all molecular dynamics simulations and analyses of trajectories. Contributed to the writing of the manuscript.

Paper V
Hörberg, J., Hallbäck, B., Moreau, K. & Reymer, A. Abnormal methylation in NDUFA13 gene promoter of breast cancer cells breaks the cooperative DNA recognition by transcription factors. *Manuscript*

BioRxiv: **doi:** https://doi.org/10.1101/2022.06.01.494372

Contribution: Helped designing the study. Contributed to the setup of the model systems. Ran the molecular dynamics simulations for the wild-type systems. Contributed to the analyses of trajectories and the writing of the manuscript.

Publications not included in the thesis

Hörberg J. & Reymer A. The glucocorticoid receptor recognizes RNA in a shape specific manner and DNA in a sequence specific manner. *Manuscript in preparation*

Carlesso A., **Hörberg J**., Reymer A. & Eriksson L. A. New insights on human IRE1 tetramer structures based on molecular modeling. *Sci. Rep*. **10,** 1-11 (2020).

Hörberg J. Saenz-Mendez P. & Eriksson L. A. QM/MM studies of Dph5 – a promiscuous methyltransferase in the eukaryotic biosynthetic pathway of diphthamide. *J. Chem. Inf. Model.* **58,** 1406-1414 (2018)

Shankar S. P., Grimsrud K., Lanoue L., Egense A., Willis B., **Hörberg J.** […] Rauen K. A. A novel DPH5-related diphthamide-deficiency syndrome causing embryonic lethality or profound neurodevelopmental disorder. *Genet. Med.* **24.** 1567-1582 (2022).

Maurel M., Obacz J., Avril T., Ding, Y. P., Papadodima O., Treton X., Daniel F., Pilalis E., **Hörberg J.** […] Ogier-Denis E. Control of anterior Gradient 2 (AGR2) dimerization links endoplasmic reticulum proteostasis to inflammation. *EMBO Mol. Med.* **11,** e10120 (2019)

Table of Content

1. Introduction

The central dogma of molecular biology[1] constitutes three fundamental biological processes: DNA replication, DNA transcription and translation (Figure 1), which together allow the genetic information flow between different macromolecules.

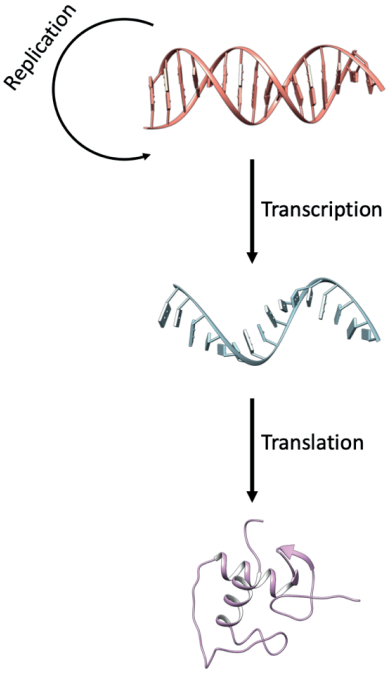


Figure 1: Illustration of the three biological reactions, which together constitute the central dogma of molecular biology.

DNA replication

DNA replication involves the production of two identical DNA molecules from the original DNA molecule.[2] The DNA replication reaction is catalysed by DNA polymerase and ensures the preservation of genetic information upon cell divisions.

DNA transcription

DNA transcription involves the conversion of genetic information stored in DNA into ribosomal-RNA (rRNA), transfer-RNA (tRNA), and messenger-RNA (mRNA)[3]. The transcription reaction is catalysed by RNA polymerase, which together with a number of assisting proteins associate with the core promoter, a region upstream of the transcription starting site of a particular gene, to initiate the RNA synthesis.[3–5]

Translation

Translation involves protein synthesis in the ribosomes, where mRNA is used as a frame.[6] Each three-base codon on the mRNA corresponds to a particular amino acid.

To maintain the healthy state of cells, these processes are highly regulated. However, complete understanding of the regulatory steps involved in gene expression to the synthesis of proteins, still remains uncomplete. My doctoral studies have focused on mechanistic regulatory aspects of eukaryotic transcriptional control. Thus, this first chapter, will cover the general theory of transcriptional regulation from the structure of DNA to its interaction with proteins.

1.1 DNA Structure and Polymorphism

DNA constitutes one of the fundamental biological macromolecules essential for life. Structurally, DNA consists of two polymeric strands that fold into an antiparallel double helix, which stores the genetic information.[7] Each strand includes repeating units of four nucleotides: adenosine- (A), guanosine- (G), thymidine- (T) and cytinidine- (C) monophosphate (Figure 2). In turn, the nucleotides are built of three components, a five-membered furanose ring (deoxyribose), a phosphate group and and a hydrophobic aromatic group (nucleobase): purine (R: A, G) or pyrimidine (Y: T, C). Upon polymerisation, the nucleotides links to each other through phosphodiester bonds to build a DNA strand in the 5'->3' direction.[8]

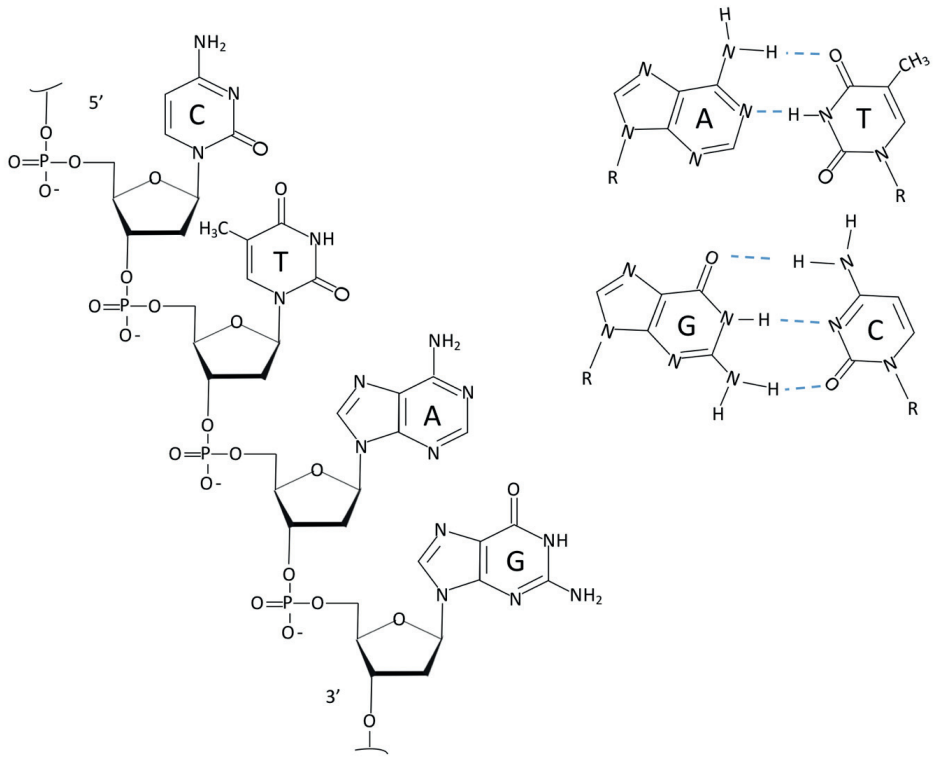


Figure 2: Left side panel: DNA polynucleotide strand in the 5' -> 3' direction. The DNA strand contains the four DNA nucleobases: C: cytosine, T: thymine, A: adenine and G: guanine. Right side panel: complementary purine-pyrimidine base pairs. Hydrogen bonds are illustrated by dashed blue lines. The R-group denotes connecting atoms of the backbone.

The hydrophobic properties of the nucleobases at physiological pH favour the folding and hybridisation of two DNA strands into an antiparallel double helix. This provides an exterior negatively charged sugar-phosphate backbone and a hydrophobic interior core of stacked nucleobases, protected from the aqueous cellular environment. The helix is stabilised intramolecularly by the van der Waals (vdW) forces and hydrophobic contacts between the stacked bases and intermolecularly by hydrogen bonds between complementary base pairs. adenine always pairs with thymine by two hydrogen bonds and guanine always pairs with cytosine by three hydrogen bonds (Figure 2).[7–9] This central property of DNA constitutes the fundamental basis of its function – the storage of the genetic information, its subsequent transfer to RNA, and then into proteins.[8,9]

DNA molecule is highly flexible and polymorphic. DNA exhibits a variety of different conformations, depending on the conditions of the surrounding environment.[8] The most abundant DNA conformation in the cell is the B-form.[7] B-DNA constitutes a right-hand double-stranded helix with the bases oriented nearly perpendicular to the helical axis. Each helical turn includes ~10 base pairs (b.p.) separated by 3.4 Å/b.p. (so-called helical rise), and twisted by ~36°/b.p. The intertwining of the two strands generates repeating voids, called major and minor grooves. These voids serve as key interaction spots for proteins. Two other less abundant but characteristic conformations of DNA include the A- and Z-forms.[10,11] For illustration of the DNA B, A and Z conformations see Figure 3.

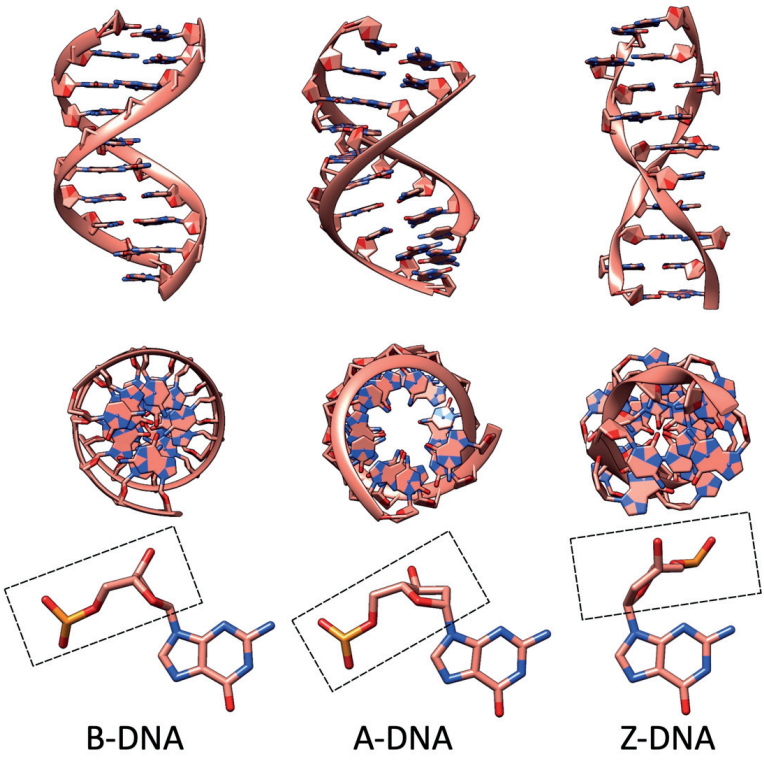


Figure 3: Structural illustration of the three characteristic DNA conformations, B-, A- and Z-DNA. Both B- and A-DNA constitute right-hand helices, whereas Z-DNA constitutes a left-hand helix. The bottom panel shows the changes of the backbone conformation (marked with rectangular box) between the different DNA forms for a guanosine monophosphate.

The polymorphic nature of DNA is coupled to its flexible backbone.[12,13] The non-planar sugar rings together with the phosphodiester bonds include a number of torsional angles, which freely rotate at low energy cost to allow conformational transitions of DNA. The most characteristic backbone conformations are the C2'-endo and C3'-endo sugar puckers[12] and the BI/BII conformations of the phosphate group.[14] The sugar puckers refer to the furanose ring conformation,[12] where one of the atoms is positioned outside the plane to release the strain within the five-membered ring. The backbone torsions can exist in a number of low energy conformations: gauche+ (g+), gauche- (g-) or trans (t).[15] The BI- and BII conformations (Figure 4) arise from changes in the torsional angles, ε and ζ, of the phosphodiester bond. For canonical B-DNA, the BI (ε=t, ζ=g-) conformation is favoured. However, sequence specific effects can result in BI->BII (ε=g-, ζ=t) transitions.[15]

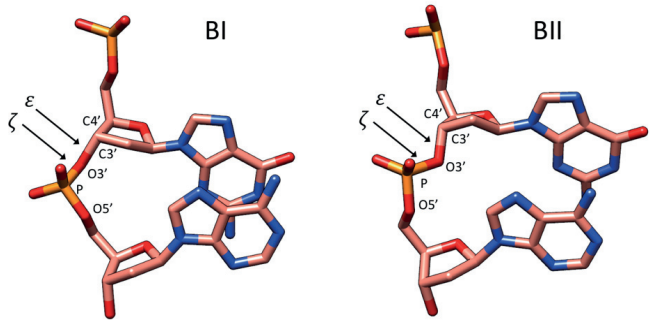


Figure 4: The BI and BII conformations for a GpA dinucleotide step. The BI and BII-conformations involve rotation around the torsional angles, ε (C4'-C3'-O3'-P) and ζ (C3'-O3'-P-O5'). For BI the ε- and ζ-torsions adopt trans and gauche- conformations, respectively. For BII the ε-torsion adopts a gauche- conformation and the ζ-torsion adopts a trans conformation, which puts the O3'-atom closer to the nucleobases.

The flexibility of DNA backbone also contributes to local sequence specific effects in terms of groove parameters (width and depth) and helical parameters (axis-, intra-base and inter-base parameters).[16,17] The axis base pair-parameters include two translational (x- and y-displacement) and two rotational (inclination and tip) parameters. The intra-base pair parameters describe the conformation of base pairs and include three translational (shear, stretch and stagger) and three rotational (buckle, propeller twist and opening) parameters. Likewise, the inter-base parameters describe the geometry of dinucleotide steps and include three translational (shift, slide and rise) and three rotational (twist, tilt and roll) parameters. For illustration of the different helical parameters see Figure 5.

Interestingly, even within a DNA conformational family (B-, A-, Z-DNA), variations in the nucleotide sequence composition can lead to significant changes in helical parameters.[15,18,19] The conformational space of B-DNA has been thoroughly explored. B-DNA exhibits a highly heterogeneous and sequence specific behaviour[15,19,20], linked to the tetranucleotide or even the hexanucleotide environment[21], depending both on the central dinucleotide step and the adjacent flanking nucleotides. Certain dinucleotides (YpR or RpR) in specific flanking environments (Y..R or R..R/Y..Y) oscillate between different conformational substates. This is illustrated as bimodal/non-gaussian-like distributions for several inter-dinucleotide parameters: twist, shift and slide.[15] Together the helical parameters twist, shift and slide exhibit three well-defined conformational substates. (1) The first substate is characteristic for canonical B-DNA and can be occupied by all dinucleotides. (2) The second substate is adopted by RpR steps and involves high twist and negative shift. (3) The third substate is mainly occupied by YpR steps and involves low twist and negative slide. Oscillation between these conformational substates occurs through backbone BI-BII transitions, where the high and low twist substates for RpR and YpR steps, respectively, favour a BII conformation.[15] Additionally, the BII conformation is stabilised by an atypical base-phosphate hydrogen bond (C8-H…O3') between two adjacent purines.[15] It has been further shown that the polymorphism of certain dinucleotide steps arise from sequence specific effects beyond the tetranucleotide level. In particular, the CTAG tetranucleotide with a central TpA step shows significant conformational fluctuations depending on the hexanucleotide environment.[21] This sequence specific conformational flexibility of DNA may play a key regulatory role for modulating of the local DNA deformability (see section 1.4 DNA supercoiling) and DNA interactions with proteins and other ligands.

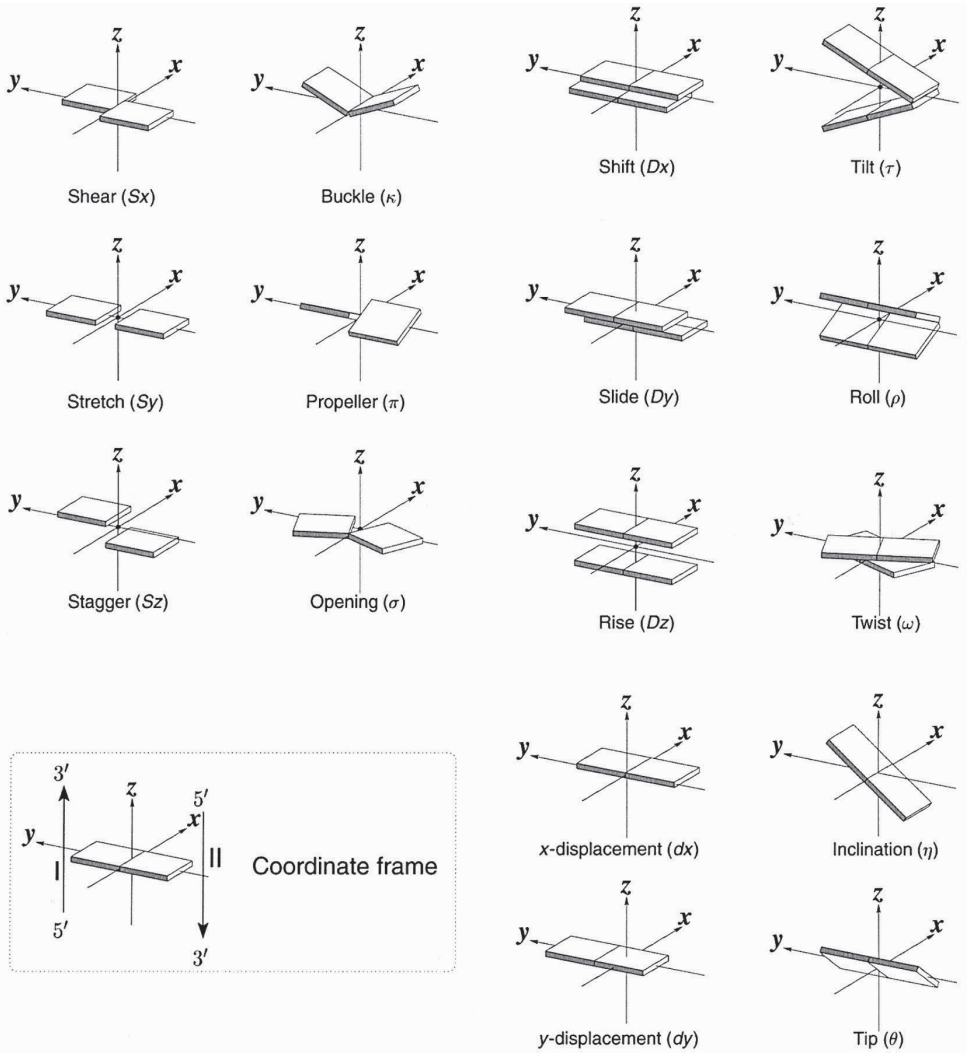


Figure 5: Translational and rotational inter-base pair parameters (top left), intra-base pair parameters (top right) and base pair-axis parameters. The x-axis arrow points towards major groove. The figure was adopted from: http://x3dna.org/highlights/schematic-diagrams-of-base-pair-parameters

1.2 DNA Condensation

Eukaryotic DNA is stored in a small, microscopic organelle, called the cell nucleus (6 µm = 6*10^-6 m). Consider the human genome, which constitutes 6.4 billion base pairs; if one base pair is roughly 3.4 Å, the complete length of linear human DNA estimates to ~2 m – that is 3.4*10^5 times larger than the size of the nucleus. Thus, to fit into the microscopic space of the nucleus, eukaryotic DNA undergoes various levels of compaction (Figure 6).[5,22] DNA is packed into chromosomes, which comprise of chromatin fibres. The chromatin fibres include hetero-chromatin, a densely packed, closed, chromatin fibre (30 nm) associated with transcriptional silencing, and Euchromatin, a lightly packed, open, chromatin fibre (10 nm) associated with higher transcriptional activity.[23] In turn, chromatin constitutes an array of repeating units of nucleosomes. Each nucleosome consists of a ~150 b.p. DNA-segment wrapped 1.7 times around an octamer core of four pairs of histone proteins (H2A, H2B, H3 and H4).[22] The nucleosomes are connected by stretches of linker DNA (~20-90 b.p.).[24] Another histone protein, H1, binds the linker DNA and facilitates the coiling and a zigzag arrangement of nucleosomes into the chromatin fibres.[22,24] The histones contain long random coil N-terminal tails, rich in positively charged residues (lysines and arginines), which can undergo chemical modifications such as methylation and acetylation to impact the compactness state of the chromatin structure.[25]

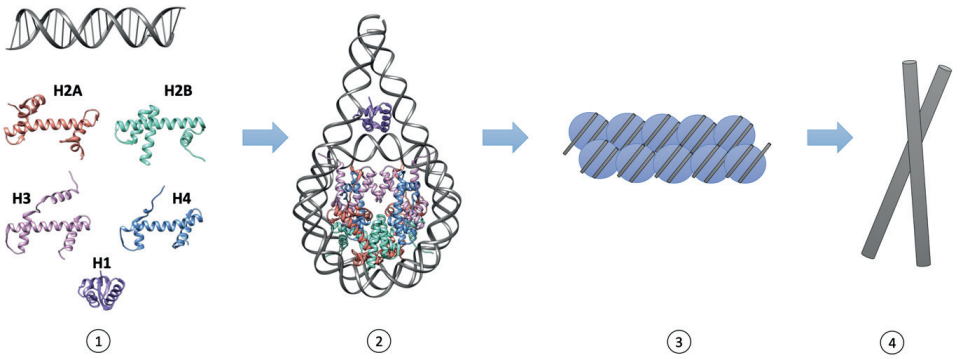


Figure 6: Packing of DNA into chromosomes. (1) A 150 b.p long DNA-segment associates with an octamer core of four pairs of histone proteins (H2A, H2B, H3 and H4) to form a (2) nucleosome core particle. Each nucleosome is followed by stretches of linker DNA (~20-90 b.ps). Another histone protein H1 binds linker DNA to facilitate further coiling of nucleosomes into (3) chromatin fibres of various density. (4) The different levels of compactness of chromatin fibres constitute the structure of a chromosome.

1.3 DNA Methylation

Epigenetics involves reversible regulatory mechanisms that do not alter the DNA sequence, but instead, impact the gene expression. These include for instance histone tails modifications and DNA base modifications. One of the most abundant epigenetic mark, DNA methylation (Figure 7A), involves the covalent attachment of a methyl group (CH3-) to either adenine or cytosine.[26] In eukaryotes, methylation occurs predominantly within CpG dinucleotides (Figure 7B).

Methylated cytosine is susceptible to deamination, resulting in CpG → TpG mutations.[27] Despite being partially mutagenic, CpG methylation has an evolutionary advantage in regulation of gene expression.[28–30] Increased CpG methylation is associated with heterochromatin and silencing of genes. In the human genome, 1-2% of the CpG sites are clustered in regions of high GC content, so-called CpG islands.[31] The majority of CpG islands are unmethylated and located within promoters and enhancers for coded genes.[27] In addition, the methylation of CpG islands has shown to be tissue specific.[32,33] Abnormal changes in DNA methylation can interfere with the healthy state of the cell, which can lead to the onset of various cancers. Hypermethylation (increased methylation) or hypomethylation (reduced methylation) can result in impaired or enhanced transcription, respectively, of certain genes.[34–37] Therefore, it is of great importance to understand transcription regulatory mechanisms involving CpG methylation.

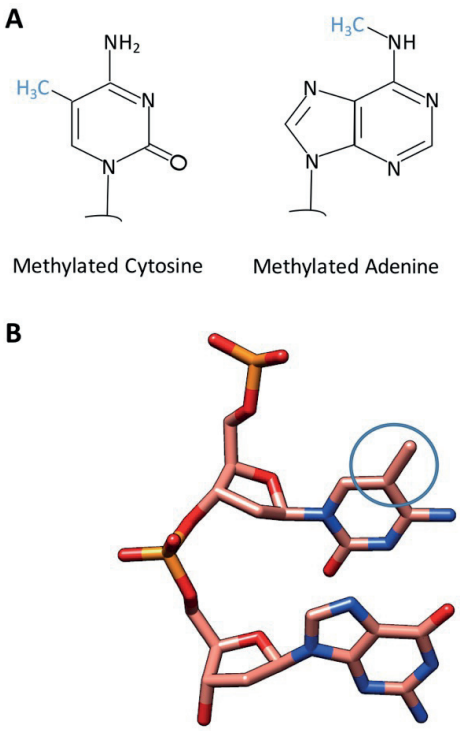


Figure 7: A. DNA methylations of cytosine or adenine. **B.** Methylated CpG dinucleotide step; the methyl group is highlighted with a blue circle

The silencing of genes by CpG methylation has traditionally been described by the means of two mechanisms. (1) The methyl group creates steric hindrance in DNA major groove, which inhibits the interactions of transcription factors with their DNA recognition sites.[38] (2) The methyl group is recognised by a class of proteins, called methyl-CpG-binding proteins (MBP), which recruit histone modification proteins, resulting in remodelling of nucleosomes to provide a more compact and inaccessible chromatin structure.[39,40] Nevertheless, there is also an evidence of a third mechanism, which suggests that the CpG methylation changes the physical properties of DNA:[41,42] The CpG step and adjacent dinucleotide steps become significantly stiffer. Structurally, methylated cytosine does not impact the hydrogen bond interactions within the CG b.p. However, the presence of the methyl group reduces the twist and increases the roll angle of the CpG dinucleotide.[41] The methyl group also creates potential steric clashes with the C2' atom of the sugar phosphate backbone on the adjacent 5'-nucleotide.[40,43] Consequently, the bendability and circularisation efficiency of DNA becomes restricted, which increases the deformation energy for forming stable nucleosomes.[41]

1.4 DNA Supercoiling

DNA supercoiling is a consequence of the torsional stress being introduced by molecular motors such as RNA polymerase. DNA supercoiling constantly changes during the lifetime of a cell and constitutes a key regulatory mechanism of many biological processes including DNA compaction, replication and transcription.[44–48] Mathematically, supercoiling corresponds to the sum of the two interchangeable variables, writhe (loops and knots) and twist (under- and overwinding around the helical axis). Writhe generally accounts for the supercoiling changes on larger scale, whereas twist dominates locally when a DNA fragment <100 b.p. experiences changes in torsional stress.[46,49]

$$DNA\ Supercoiling = Writhe + Twist \qquad (1)$$

On average cellular DNA is torsionally relaxed, however, there are regions of positive and negative supercoiling locally, so-called supercoiling domains, introduced by the transcription machinery.[46,48] Upon transcription, RNA polymerase generates negative supercoiling (undertwisting) upstream and positive supercoiling (overtwisting) downstream of a transcribed gene. The amount of torsional stress generated by the transcription machinery can propagate along the chromatin fibre at speeds modulated by the underlying sequence.[44,48] The propagation impacts the stability of nucleosomes, which by altering the accessibility of the genetic code may regulate transcription of near-located genes and the specific binding of transcription factors.[46,47,50–52]

Supercoiling is not uniformly distributed[48,53,54] and absorbed.[55–57] In fact, the DNA response to torsional stress is highly sequence-specific and heterogeneous.[57] The polymorphic pyrimidine-purine (YpR) but also purine-purine (RpR) dinucleotide steps, in specific sequence environments (see section 1.1 about DNA polymorphism), referred to as twist capacitors, absorb the majority of over- and undertwisting, allowing the rest of the molecule to preserve a B-like conformation. This torsional flexibility arises from the polymorphic nature of the DNA backbone, which allows the dinucleotide steps to oscillate between high and low twist substates, separated by about 20°. The twist transitions are also coupled with changes in helical shift and slide, which makes these dinucleotide steps potential 'hot spots' for the regulation of DNA deformability and specific binding of transcription factors.

1.5 Transcription Factors

The packing of eukaryotic DNA into chromatin fibres function as a general repressor of transcription.[58] Therefore, in eukaryotes from yeast to human, gene expression is primarily regulated at the initiation stage of transcription. The initiation stage is regulated by a class of proteins termed transcription factors.[59] These proteins recognise and bind short DNA sequences, called response elements, within non-coding regions (promoters, enhancers and silencers) in the vicinity of a gene.[60–62] Upon binding to the response element, the transcription factors induce the bioactive DNA conformation, sometimes quite substantial like SRY and TBP (Figure 8) and sometimes small like Zn fingers (Figure 8). The conformational changes may play a role in the requirement of collaborative TFs or the transcription machinery.[59,61,63] Transcription factors are categorised into different families based on their DNA binding domains.[59,64] The most common DNA binding domains, which range from minor groove to major groove binding, are shown in Figure 8.

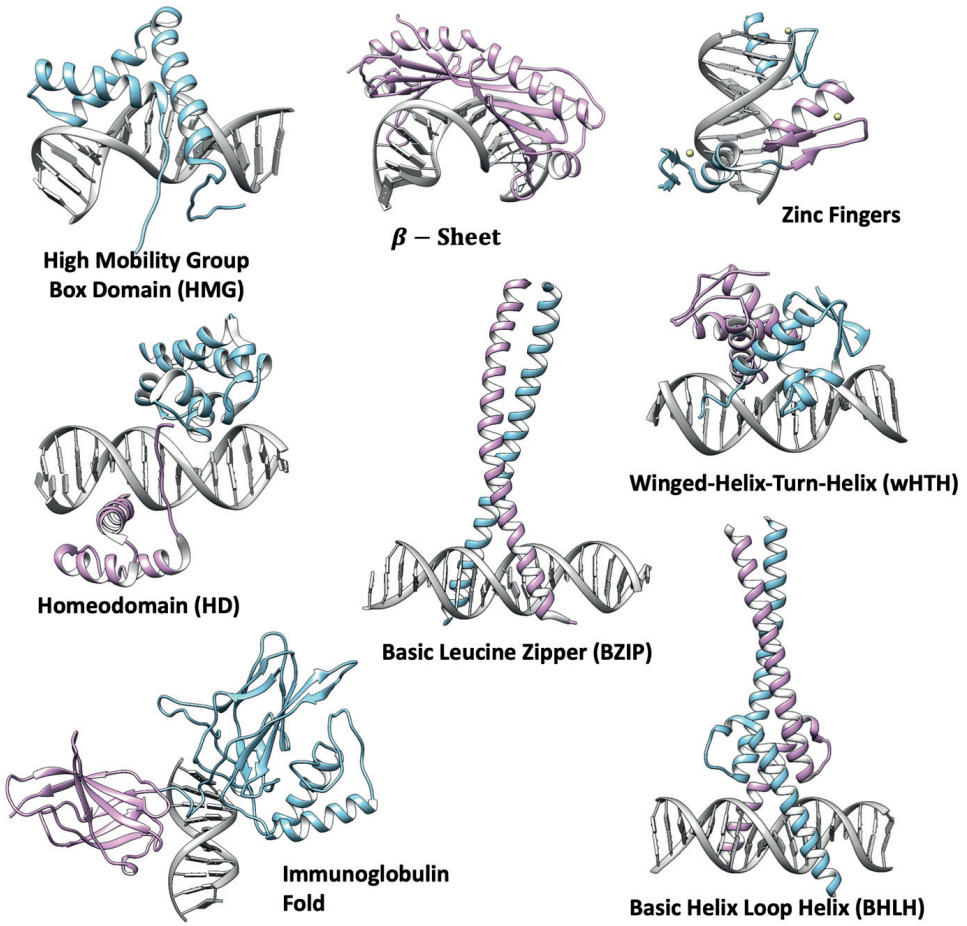


Figure 8: Transcription factor domain families: High mobility group box domain (HMG), e.g. SRY (PDB ID: 1J46).[65] B-sheet, e.g. TATA-box binding protein (PDB ID: 1CDW).[66] Zinc fingers, e.g. Krüppel-like-factor 4 (PDB ID: 4M9E).[67] Homeodomain (HD)/Helix-Turn-Helix (HTH), e.g. Oct4 (PDB ID: 3L1P).[68] Basic-leucine-zipper (BZIP), e.g. FOS-JUN (PDB ID: 1FOS).[69] Winged-Helix-Turn-Helix (wHTH), e.g. E2F4-DP2 (PDB ID: 1CF7).[70] Immunoglobulin fold, e.g. nucleofactor p50 (PDB ID:

1NFK).[71] Basic-helix-loop-helix (BHLH), e.g. Myc-Max (PDB ID: 1NKP).[72] The first two families bind to DNA minor groove, whereas the remaining bind to DNA major groove.

The transcription factor families recognise DNA response elements of various length, ~5-15 b.p.,[73] which are insignificant with respect to the size of eukaryotic genomes. Therefore, to discriminate between the excess number of potential binding sites, transcription factors bind their native regulatory target sites with great specificity.[74,75] There are three mechanisms that have been traditionally used to explain the selective binding of transcription factors to DNA: (1) The direct-readout which involves the formation of specific contacts between protein side chains and DNA bases.[64,76–78] (2) The indirect-readout which involves the protein recognition of a well-defined DNA shape.[64,79,80] (3) The "water-mediated readout" which involves the formation of bridging water contacts between protein residues and DNA bases.[81,82] This thesis, covers studies of transcription factors (TFs) of the basic-leucine zipper (BZIP), the basic-helix-loop-helix (BHLH) and the Winged-helix-turn-helix (wHTH) families, which primarily follow the direct-read-out principle to associate with DNA major groove. The response elements of these families of TFs are typically no longer than 6-8 b.p. Due to the small size of their recognized DNA targets, these sequences may appear in genomes significantly more frequently than there are genes regulated by the TFs, arguing for a more complex and sophisticated recognition mechanism. Recent studies show that local DNA shape and the regions adjacent to the response elements, termed "flanking sites", may play an important role.[83–87] In paper I, we provide atomic level mechanistic details of the recognition process.

1.5.1 Basic-Leucine-Zippers (BZIP)

The BZIP family comprises one of the largest and most evolutionary conserved families of eukaryotic transcription factors.[77] The BZIP factors regulate a broad spectrum of cellular processes including cellular stress responses, cellular development, cellular differentiation, cellular proliferation and apoptosis.[88] In higher eukaryotes, the BZIP proteins can be classified into at least seven subfamilies,[89] e.g. AP-1, CREB, C/EBP, PAR, Maf (See Figure 9 for experimentally derived DNA consensus sequences of each protein subfamily). Nevertheless, the homologs of the mentioned BZIP subfamilies exist also in yeast. Structurally, BZIP TFs are α-helical proteins composed of a leucine-zipper that act as a dimerisation domain for homo- and heterodimerisation within and across different subfamilies, and an adjacent basic region that recognizes and binds their DNA response elements from major groove (Figure 9).[77,90,91] A conserved five residue motif of each BZIP monomer, which varies slightly between the subfamilies and across species, forms specific contacts with DNA bases of the response element's half-site.[77] The dimers are capable of recognizing a diverse set of palindromic, pseudo-palindromic and emergent sites, which contain only a consensus-like half-site.[77,78,91] The repose elements can also vary in length ~7-14 b.p.[78,92] Despite the high degree of homology among the BZIP factors, the proteins target distinct genomic locations to initiate different transcriptional events, and are regulated in distinct manners.[77] In paper I, we outline a molecular mechanism of how sequence specific flexibility of DNA response elements and their flanking sites fine-tune the direct-read out mechanism. In paper III, we show how the binding of BZIP proteins can adapt to DNA conformational changes under torsion, and how the association of a BZIP protein on DNA may contribute to transcriptional control by blocking propagation of torsional stress.

Yeast AP-1	NxxAQxxFR	TTACGTAA or TTACTAA
Human AP-1	NxxAAxxCR	TGACTCA
CREB	NxxAAxxCR	TGACGTCA
CREB-2	NxxAAxxYR	TTACGTAA
C/EBP	NxxAVxxSR	ATTGCGCAAT
PAR	NxxAAxxSR	GTTACGTAAC
Maf	RxxxNxxAAxxSR	TGCTGACGTCATGC or TGCTGACTCATGC

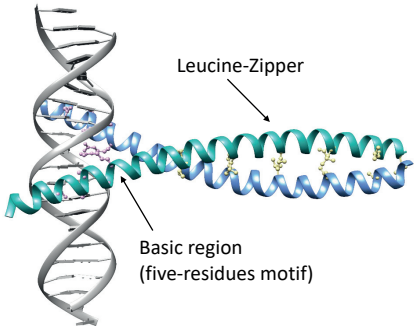


Figure 9: Left panel: Different BZIP subfamilies with their five-residue motifs and their recognized DNA consensus sequences. The Maf-family contains an additional conserved Arg residue (denoted with orange colour), extending the recognition motif to six residues. Right panel: characteristic structure of a BZIP dimer protein with the leucine-zipper and five-residue motif highlighted.

1.5.2 Basic-Helix-Loop-Helix

The BHLH family constitutes another major class of transcription factors (Figure 8), which modulates similar cellular processes as the BZIP factors. The BHLH factors can be divided into six subfamilies (A-F) based on their evolutionary relationship.[93] The BHLH factors bind preferably response elements of the type CANNTG and CACG(A/C)G called the E-Box and the N-box, respectively.[94] Structurally, they are similar to the BZIP proteins, containing a basic region, which include a five-residues motif (**xx**E**xx**R***)[94] that forms specific contacts with DNA bases in the major groove. Following the basic region, they contain a HLH domain which allows homo- and heterodimerisation. In addition, the B-subfamily known as BHLHLZ also includes a leucine zipper domain connected to the HLH domain. In paper IV, we show how torsional stress may contribute to the differential transcriptional response of homologous BHLHLZ factors.

1.5.3 Winged-Helix-Turn-Helix

The Winged-Helix-Turn-Helix (wHTH) family (Figure 8) constitute a subfamily of the HTH family.[95] This is a large and diverse family, composed of a 3-helical bundle and a 3-4 strand beta sheet i.e., a wing domain. One of the helices inserts in DNA major groove.[70,96] In paper V, we provide insights into how DNA methylation impacts the binding of the wHTH heterodimer E2F1-DP1, and how this can be involved in the initiation of breast cancer. The E2F and DP families utilise a conserved four-residues motif (RRxxYD)[70] to interact with their response elements of the type TTT(C/G)GCGC(C/G) to modulate cell-cycle processes etc.

1.6 Transcription Factor Cooperativity

Transcriptional events, especially in higher eukaryotes, is predominantly mediated by the cooperative binding of transcription factors (TFs) to DNA, where one TF enhance the binding of another TF.[59,64,74,97,98] Cooperativity provides additional level of understanding of how TFs discriminate between biologically irrelevant target sites, and how specificity can be achieved among homologous TFs that share identical recognition motifs.

There are different categories of cooperativity,[97] the simplest being TFs that bind DNA as dimers (homo- and hetero-), trimers, or higher order structures. Another type of cooperativity arises through the formation of favourable protein-protein interactions between TFs of different families upon DNA-binding. Cooperativity can also be mediated by structural changes in DNA, termed DNA allostery; binding of one TF to DNA may induce local/distal conformational changes or nucleosome reorganisation to promote binding of additional TFs (Figure 10).[99–104] In fact, DNA-shape is identified as a major component of TF-cooperativity.[98] In paper V, we provide insights into cooperativity between CEBPB and the E2F1-DP1 heterodimer.

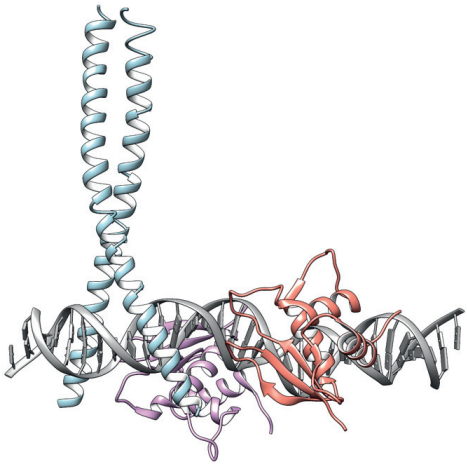


Figure 10: Molecular graphics of a part of the interferon-ß-enhanceosome (PDB ID: 1T2K).[104] Binding of c-Jun-ATF2 (blue) heterodimer to an emergent site, induces allosteric changes in DNA, which facilitates the binding of two IRF3 monomers (plum and salmon colours).

2. Computational Methods and Tools

The computational tools that have been used for this thesis work include unrestrained and restrained molecular dynamics simulations, homology modelling and macromolecular docking, briefly described below.

2.1 Molecular Mechanics and Force Field

One can describe a molecular system computationally by a potential energy functional. This is in molecular mechanics known as a 'force field'. In molecular mechanics, one neglects quantum mechanical aspects such as the existence and movements of electrons. Instead atoms are described as hard spheres with point changes that move and interact with each other according to the laws of Newton's classical mechanics. This allows for more favourable computational cost.

A force field determines the potential energy of a molecular system through the sum of bonded (bonds, angles and torsions) and non-bonded interactions (vdW- and electrostatic interactions) of each atom in the system.[105–107]. To calculate the potential energy, the force field also requires a parameter set used in the energy functional, which assigns each atom type and bond type particular properties. The parameters have been derived using experimental data and through high-level quantum mechanical calculations. In this Thesis work, the two force fields AMBER 14SB[108] and ParmBSC1[109] are used to treat proteins and DNA, respectively. These force fields are part of the AMBER family of force fields,[110] which utilise the energy functional (2) to determine the potential energy.

$$V_{AMBER} = \sum_{bonds} K_l(l - l_0)^2 + \sum_{angles} K_\theta(\theta - \theta_0)^2 + \sum_{all\ torsions} \frac{1}{2} V_n(1 + \cos(n\varphi - \gamma)) + \sum_{i<j} \left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}\right] + \sum_{i<j} \left[\frac{q_i q_j}{4\pi\varepsilon_0 \varepsilon_r r_{ij}}\right] \quad (2)$$

The first three energy terms describe the bonded interactions of the system: bond stretching (l), angle bending (θ) and torsional rotation (φ). The force constants, K_l, K_θ and V_n, constitute the energy cost for changing the bond length or angle from their equilibrium values (l_0 and θ_0), or for the rotation around a dihedral angle. The bond and angle terms are approximated by a harmonic potential. The torsional term involves a periodic rotation around a dihedral angle, defined as a cosine series, where *n* is the periodicity and γ is the phase shift. The torsional energy term considers all torsions, including the improper torsions (atoms that are not connected in sequence) that account for out of bending effects to ensure the retention of planarity of some atomic groups.

The last two energy terms describe the interactions between non-bonded atom pairs (i, j), including repulsion, dispersion, electrostatics. Repulsion and dispersion, i.e., the vdW-interactions, are approximated by a Lennard-Jones 12-6 potential (L-J), where r_{ij} is the distance between atom i and j, and A_{ij} and B_{ij} are specific constants for an atom pair related to the minimum distance at which the interaction shifts from attractive to repulsive. Electrostatic interactions are represented by a Columbic term, where q_i and q_j are the partial charges for atom i and j, and ε_0 is the vacuum permittivity and ε_r is the relative permittivity of the medium.

For large molecular systems, to maintain a desired computational performance, one adds a cut-off distance for non-bonded atomic interactions to treat only the non-bonded interactions within a certain threshold. Complete neglect of the long-range non-bonded interactions, exceeding the threshold, may result in severe artefacts. Therefore, certain corrections are added to treat the long-range non-bonded interactions properly. For long-range Lennard-Jones interactions, a

force switch function is sufficient, which smoothly switches the force to zero at the cut-off distance.[111] For long-range electrostatic interactions, more sophisticated algorithms, based on Fourier transformations, like Ewald summation or particle mesh Ewald (PME) are required.[112] AMBER is a well-established force field family known for its high performance in studies of proteins[108,110] and (now to-date) nucleic acids.[83,109,113] Nevertheless, the development of today's generation of DNA-force fields capable to accurately describe the polymorphism of DNA molecules has been a continuous journey of refinements.[113] The increasing amount of experimental data and QM calculations has provided the reference for the refinement. However, the major improvement of DNA-force fields has emerged from the increase in computational power, which has through microsecond long molecular dynamics simulations detected the structural artefacts in older versions.[109,113] Parm99 showed unreasonable α/γ transitions, which lead to the development of ParmBSC0. Although ParmBSC0 captures the sequence specific properties of DNA and performs well in reproducing experimental results, some further artefacts arise in the microsecond regime, including underestimation of twist, deviations in sugar puckering, biases in BI-BII populations' transitions related to the ε and ζ torsions, and increased base pair fraying. Corrections of these artefacts have resulted in the current version ParmBSC1,[109,113] which in accordance with experiments, captures the polymorphic behaviour of B-DNA both in relaxed form and under torsion.[20,57,83]

2.3 Energy Minimisation

Energy minimisation constitutes algorithms (e.g. Steepest Descent and Conjugate Gradients) that locate stable conformations corresponding to a local or a global energy minima on the potential energy surface (PES) for a given force field.[105,111] Energy minima are stationary points where the negative derivative of the potential energy functional with respect to the atomistic position i.e. the force (3) for each atom *i* in a molecular system is zero, and the second derivative matrix i.e. the Hessian matrix (4), that describes the curvature of PES, is positive (convex) for all pairs *i* and *j*.

$$F_i = -\frac{\partial V}{\partial r_i} = 0 \qquad (3)$$

$$H_{i,j} = -\frac{\partial^2 V}{\partial r_i \partial r_j} > 0 \qquad (4)$$

Typically, in energy minimisation (i) one calculates the force on each atom, (ii) if each force is less than a certain threshold, the minimisation stops, otherwise (iii) one updates the atomic positions and repeats step i.

A widely used algorithm for energy minimisation is Steepest Descent,[111] which locates an energy minimum by moving at steps proportional to the negative of the energy gradient. New atomic positions (r_{n+1}) are calculated through (5), where r_n and F_n are the current position and force, h_n is the maximum displacement (initially 0.01 nm) and max($|F_n|$) is the largest scalar force on any atom in the system.

$$r_{n+1} = r_n + \frac{F_n}{\max(|F_n|)} h_n \qquad (5)$$

If the updated positions decrease the potential energy, then the new positions are accepted and the new displacement, h_{n+1}, is updated to $1.2h_n$. However, if the potential energy is larger than in the previous iteration, the new positions are rejected, and the displacement is changed to $0.2h_n$.

2.4 Molecular Dynamics

Proteins and nucleic acids constitute dynamic and flexible biological macromolecules in constant movement. At a given temperature, there is generally an ensemble of stable conformations present,[111] see the discussed polymorphism of B-DNA (section 1.1), where certain dinucleotides fluctuate between different conformational substates. To capture these dynamic properties of macromolecules one can utilise a computational simulation technique called Molecular Dynamics (MD).[105,111,114] Several software packages are available for MD simulations, including AMBER,[115] CHARMM,[116] GROMACS[117] and NAMD.[118] In this Thesis work, I have used GROMACS.

In MD simulations,[111] one follows the time evolution of a single copy of a system through integration of Newton's second law of motion (6), where m_i and r_i are the mass and position ($r_i = x_i, y_i, z_i$, in a 3D Cartesian space) of atom i, and F_i is the force that acts on atom i in the system. The force is calculated by (7), as the negative gradient of the potential energy function (V) with respect to the atomic position (r_i).

$$F_i = \frac{\partial^2 r_i}{\partial t^2} m_i \,, \; i = 1, 2, 3, \ldots, N \qquad (6)$$

$$F_i = -\frac{\partial V}{\partial r_i} \qquad (7)$$

Running an MD simulation starts with assigning initial positions and velocities to all atoms in the system. The initial coordinates of the system are obtained from high resolution experimentally determined structures (e.g.: X-Ray, NMR) or from molecular modelling (e.g.: homology modelling). In turn, to ensure a stable geometry in the initial state, the system is energy minimised. The initial velocities are randomly assigned based on the Maxwell-Boltzmann distribution. Subsequently, the forces are calculated, which carry the information of the direction of the atoms' momentum. With these instructions together with a pre-defined time step (Δt), Newton's equations are solved numerically for each time step to continuously update the positions and velocities of atoms. This provides a trajectory that pictures the evolution of a molecular system over a specific time.

There are many algorithms, based on the Taylor expansions, available for the integration of Newton's equations. One includes the Verlet algorithm[119] (8), which calculates the new position of atom i by using the atomic position (r_i) and acceleration (a_i) at time t together with the previous determined position $r_i(t-\Delta t)$.

$$r_i\,(t + \Delta t) = \; 2r_i(t) - r_i\,(t - \Delta t) \; + \; \Delta t^2 a_i(t) \qquad (8)$$

The limitation with this approach is that the Verlet algorithm does not explicitly store atomic velocities. Therefore, another more accurate integration algorithm, leap-frog,[120] is preferably used, which updates and stores positions and forces at times $t_1, t_2, t_3, .., t_n$ and velocities at times $t_{1/2}, t_{3/2}, t_{5/2}, .., t_{n-1/2}$ according to the relations (9) and (10):

$$r_i\,(t + \Delta t) = \; r_i(t) + \; v_i \Delta t (t + \tfrac{1}{2}\Delta t) \qquad (9)$$

$$v_i\left(t + \tfrac{1}{2}\Delta t\right) = \; v_i\left(t - \tfrac{1}{2}\Delta t\right) + \tfrac{\Delta t}{m} F_i(t) \qquad (10)$$

For accuracy, the time step,[111,114] Δt, utilised for the integration, should be smaller than the fastest motion within the system. This corresponds to the vibrations of bonds between hydrogen atoms and heavy atoms occurring at the time frame of 10 fs. To account for those vibrations, a

time step of 0.5-1 fs is necessary. This limits the performance and makes MD simulations computationally expensive. The efficiency can be slightly improved by the use of certain algorithms (LINCS[121] or SHAKE[122]) that put constraints on the fastest bond vibrations, allowing a time step of 2 fs. With the performance of today's computers, the state-of-the-art simulation time is in the microsecond range.

Another important aspect to consider is the simulation conditions. MD simulations are preferably run under conditions mimicking experiments, to allow for comparison with experimentally determined values. Most biological macromolecules exist in aqueous solutions. To mimic the bulk conditions at an affordable computational level, MD simulations are performed under periodic boundary conditions (PBC). A single copy of the solute is centred in a periodic 3D box and solvated by explicit solvent. Subsequently, by applying PBC, the system is multiplied infinitely in all directions. The system's replicas move in parallel: i.e., if one copy of the solute jumps out of its cell, an image of the solute enters from the opposite side. The distance of the solute to each wall of the periodic box is set to account for the cut-off distance of non-bonded interactions to ensure that the solute does not interact with its periodic image, which might create artefacts.[111,114]

MD simulations can be performed in one of three thermodynamic ensembles: the NEV (microcanonical), the NVT (canonical) or the NPT (isobaric-isothermal) ensemble, where the abbreviations constitute the thermodynamic quantities (N: number of atoms, E: energy, V: volume, T: temperature and P: pressure) that are kept fixed during simulations. The NPT ensemble is generally utilised as it resembles the conditions of biological processes and experiments that proceed under constant pressure and temperature. The temperature and pressure are maintained constant through coupling of a thermostat and a barostat, respectively, to the system.[111]

2.5 Umbrella Sampling and Free Energy Calculations

Classic Molecular Dynamics is often limited by insufficient sampling around minima states separated by small energetic barriers.[123] If the simulations are run long enough, the barriers will eventually be overcome and more sufficient sampling around conformational space is obtained. However, the simulation time required to traverse the barriers is typically unreachable by todays computational performance. Therefore, several enhanced sampling techniques have been developed, including Replica-Exchange MD,[124] Steered MD,[125] Metadynamics[126] and Umbrella Sampling.[127] These methods improve the sampling by introducing a bias, allowing the system to cross the energy barriers to explore the conformational landscape broader. In this thesis work, the Umbrella Sampling (US) approach has been used to investigate the impact of sequence specific response of DNA to torsional stress in transcriptional control. US, firstly proposed by Torrie and Valleau,[128,129] utilises a bias potential to sample around a specific value of the reaction coordinate of interest, also known as a collective variable (CV). The idea is then to extract the free energy associated with the coordinate pathway. The form of the bias potential is commonly a harmonic restraint potential (11) where the collective variable ξ is restrained to a desired value ξ_0 with the force K_{CV}. In this Thesis work, I use a collective variable that constitutes the total twist of a DNA fragment.

$$V_{CV} = K_{CV}(\xi - \xi_0)^2 \qquad (11)$$

The US protocol comprises of a series of individual simulations "windows" along the reaction coordinate. The reaction coordinate is progressively modified in small steps between the consecutive windows to obtain a smooth transition along the reaction pathway. For the work presented in this thesis I run umbrella sampling in a cascade fashion, where the final frame of the current window is used as a starting state for the consecutive window; alternatively, all windows

can be run in parallel. With sufficient sampling of each US window, one can reweight and combine the probability distributions for the different states along the reaction coordinate to derive the potential mean force (PMF), i.e., a free energy profile of the reaction path with respect to the collective variable. Reweighting is generally done with the Weighted Histogram Analysis Method (WHAM).[130]

In papers II-IV as a collective variable, we used a "DNA-twist" variable that describes the total twist of a DNA fragment, applied between desired base pairs, e.g., N and M. The "DNA-twist" variable is introduced to the energy function in analogous way with a quadratic function $K_{tw}(twist - twist_{ref})^2$ analogously to (11), where the force constant K_{tw} is set to 0.06 kcal mol^{-1} deg^{-2}. The value of K_{tw} is derived in a series of computational experiments as the value that allows to maintain the desired twist without producing any structural artefacts. The total twist of the DNA fragment is calculated similarly as the approach reported by Curves+.[17] The algorithm of the DNA twist collective variable is described in Reymer et al.[57]

2.6 Homology Modelling

If a structure of a protein of interest is unknown, one can utilise a modelling approach called homology modelling.[131,132] In homology modelling the unknown structure is built from its amino acid sequence using as templates, the biologically related so-called homologous proteins with experimentally solved structures. The approach relies on the concept that the 3D structure is evolutionary more conserved than the amino acid sequence since the function of a protein is mainly determined by its 3D fold. Therefore, proteins with similar functions are likely to have similar 3D structures even though their amino acid sequences may differ. Homology modelling consists of five steps listed in Figure 11.

Figure 11: The five steps of homology modelling.

The first step, template selections, involves identification of homologous proteins that possess the desirable level of sequence identity (the relative number of identical residues in the sequence alignment) and sequence similarity (the relative number of aligned residues with the same physicochemical properties) to the target protein. This is generally done by using algorithms like BLAST (Basic Local Alignment Search Tool) or PSI-BLAST (Position Specific Iterative Basic Local Alignment Search Tool), which through sequence alignment compare the sequences of the target protein to proteins deposited in Protein Data Bank (PDB).[133] The second step involves secondary structure predictions and the use of more sophisticated multiple sequences alignment algorithms to provide a more optimal alignment between the target and the template(s). This step is critical to assure i) that hydrophobic residues are buried in the protein core, ii) that the minimum number of gaps and/or deletions are introduced, and iii) that they are positioned within loop regions, outside secondary structure elements. The third step involves the model building, where the structure of the target protein is built by copying the coordinates of the template(s) for the aligned residues. For those aligned residues that differ between the target and template(s), one only copies the template backbone coordinates. The placement of side chains is modelled using rotamer libraries combined with energy scoring functions. In unaligned target-template regions, loops are introduced. The loop-modelling step contributes to the most errors in homology modelling, especially when the loops are longer than 10 residues.

Loops are generally modelled by *ab initio* methods or knowledge-based approaches that extract loop conformations, with similar amino acids sequence, from PDB. The fourth step involves the refinement, generally, through energy minimisation. The fifth step involves the quality evaluation.[131,134] The accuracy of homology modelling drops with decreasing sequence identity between the target and template(s) proteins as a result of inevitable errors being introduced at the target-template alignment stage. This increases the number of gaps (loops), which in turn may cause miss-prediction in the fold of certain regions. Therefore, as a rule of thumb, homology modelling should not be used if the sequence identity falls below 25%.

There is a large number of software packages available for homology modelling, including SWISS_MODEL,[135] MODELLER,[136,137] ROSETTA[138] and YASARA.[134] In paper I, and paper V, YASARA, is used to generate a model of Yap1 protein and the E2F1- DP1 heterodimer, respectively.

2.8 Macromolecular Docking

Most biological events involve the association of biological macromolecules into complexes e.g. DNA transcription. Although, the number of available protein-DNA complexes in PDB has increased considerably over the years, they only account for a small fraction of the complexes occurring within the cell. Therefore, to study a particular protein-DNA complex that has not yet been solved experimentally, one can use a modelling approach known as macromolecular docking,[139,140] which predicts the structure of a complex from the individual structures of the interacting partners. Docking is a dual process consisting of the sampling and scoring processes.[141,142] The sampling generates all possible poses (binding modes) between the interacting partners, using the information from a scoring function that ranks the poses. Docking of small organic molecules to a specific target protein is a well-established technique in drug discovery. However, the reconstruction of biomacromolecular complexes, where the individual macromolecules provide many potential interaction surfaces and often may undergo considerable conformational changes (induced fits) upon association, makes macromolecular docking a computationally demanding process and decreases the probability of identifying the correct binding mode. Therefore, additional structural information, e.g. knowledge of key residues at the interaction interface, is commonly necessary to guide the docking towards the biological relevant pose.[139,141,143]

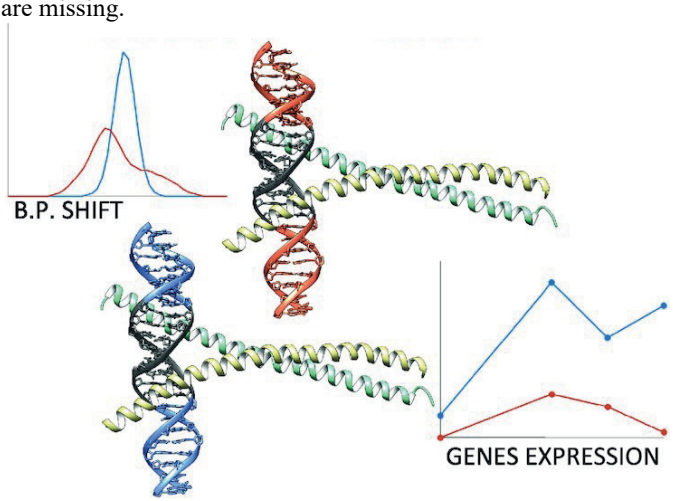
A number of web-servers and software packages are available for performing macromolecular docking, including HADDOCK,[144] AutoDock Vina,[145] RosettaDock server,[146] SwamDock,[147] ZDOCK server[148] and HDOCK.[139] In paper I, and paper V, the web-server HDOCK is used to derive the Yap1-DNA complexes and E2F1-DP1-CEBPB complexes. HDOCK uses a hybrid algorithm of template-based modelling, searching through PDB for homologous complexes, and *ab initio,* Fast Fourier Transform (FFT), free docking to perform rigid macromolecular docking.

3. Results

In this chapter, the main results and conclusions for Papers I-V are presented.

3.3 Paper I. Sequence-specific Dynamics of DNA Response Elements and their Flanking Sites Regulate the Recognition by AP-1 Transcription Factors

In paper I, we were motivated to address the collaborative impact of sequence specific DNA flexibility of transcription factors response elements and their flanking sites on the Activator proteins 1 (AP-1)-DNA recognition process. The AP-1 transcription factors represent one of the largest and most evolutionary conserved families of eukaryotic transcription factors,[149,150] The AP-1 factors have traditionally been thought to achieve their regulatory specificity for their genomic target sites, AP-1 DNA response elements (ARE), through the direct-readout mechanism[77], where the conserved five-residue motif of the proteins forms base-specific contacts with DNA bases from DNA major groove. Though, the AP-1 factors are highly homologous, they modulate different transcriptional pathways. In addition, the AREs, varying in sequence length of 7-14 b.ps. occur more frequently in the genome than there are genes modulated by the specific AP-1 factors. This suggests that there are still mechanistic details in the AP-1 recognition process that are missing.



	ATR1	OYE2
YRE1	5'-TATAGTGA**TTACTAA**TGGAATGG-3' 3'-ATATCACT**AATGATT**ACCTTACC-5'	5'-GTTTTGCT**TTACTAA**GCACACGA-3' 3'-CAAAACGA**AATGATT**CGTGTCGT-5'
YRE2	5'-GCCACAGA**TTACGTAA**GCGATTT-3' 3'-CGGTGTCT**AATGCATT**CGCTAAA-5'	5'-GAAATATC**TTACGTAA**TGAACTT-3' 3'-CTTTATAG**AATGCATT**ACTTGAA-5'
YRE3	5'-TGATTATA**TGACAAA**GTTGAGGG-3' 3'-ACTAATAT**ACTGTTT**CAACTCCC-5'	5'-GCTAGCGA**TGACAAA**ATGTCTCC-3' 3'-CGATCGCT**ACTGTTT**TACAGAGG-5'

Figure 12: Homology model of Yap1 (monomer 1: yellow, monomer 2: aquamarine) bound to DNA, where YRE is highlighted with dark grey, and the genomic flanking environment is coloured blue for ATR1 and orange for OYE2 genes. The bottom panel table shows the different sequences studied. We see a potential relationship between DNA conformational flexibility within the response element, which contributes to stronger Yap1-DNA association, and the gene expression levels for the genes. Reproduced with permission from Oxford University Press[151].

For our computational study, we selected Yap1, a member of Saccharomyces cerevisiae AP-1 family, involved in regulation of oxidative stress responses and cell detoxification.[152] The five-residue motif of Yap1 basic region constitutes a particular interesting case as it differs from the AP-1 factors of other eukaryotic organisms in terms of two residues; Glu and Phe instead of Ala and Cys respectively (**N**xxAQxx**F**R).[77,152] The Phe residue increase the preference for a 5'-TT dinucleotide instead of the 5'-TG dinucleotide at the extremities of the AP-1 consensus sequence 'TGACTCA'. Nevertheless, the response elements recognised by Yap1 remain highly diverse: TTACTAA, TTACGTAA, TGA(C/G)TAA and T(T/G)ACAAA.[152] For our study, we selected three different Yap1 response elements (YREs) (Figure 12), representing three categories of response elements recognised by BZIP factors: a pseudo-palindromic 'TTACTAA' (YRE1), a palindromic 'TTACGTAA' (YRE2) and an emergent site 'TGACAAA' (YRE3), from promoter regions of two known Yap1 regulated genes (ATR1 and OYE2),[153] to account for a variation in the flanking sequences. To provide mechanistic insights into the Yap1-DNA recognition, we performed microsecond long all-atomistic molecular dynamics simulations for unbound DNA and Yap1-bound DNA, followed by a thorough analysis of helical parameters and the contact network between Yap1 and the different YREs.

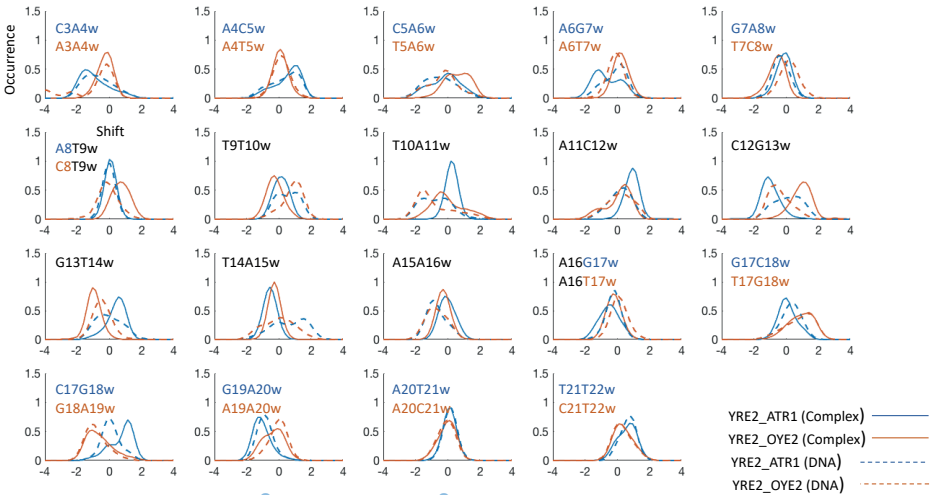


Figure 13: Normalised shift distributions for unbound (dashed lines) and Yap1-bound (thick lines) YRE2-DNA. The ATR1-environment is blue-marked, and the OYE2-environment is orange-marked. Reproduced with permission from Oxford University Press[151].

The results showed that local sequence-specific DNA flexibility facilitates the direct-readout mechanisms. Adjustments in the helical parameters shift, and to a lesser extend slide and twist within the response element create a favourable environment in the major groove to allow for stable and strong specific protein-DNA contacts. Helical shift, defines how much a b.p. is displaced towards major (positive shift) or minor groove (negative shift) relative to the 5'-adjacent b.p, Therefore, the displacement of b.ps. in the grooves of the RE may either facilitate or hinder specific contacts. In addition, our computational results showed that shift-flexibility, and consequently, the contact network, depend significantly on the flanking environment. The 4-6 adjacent flanking nucleotides fine-tune the conformational adaptability of the RE for Yap1. As a results, we observed varying contact networks for Yap1 bound to the same YRE in two different genomic environments (example for YRE2 in Figure 13-14). Unfavourable flanking sites lead to broad shift distributions within the RE sequences (Figure 13), which either cause a reduction or a rearrangement in the contacts (Figure 14). Further analysis of available crystal structures

of BZIP-DNA complexes suggests that the described mechanism is universal for the BZIP family, and potentially is employed also by other families of transcription factors.

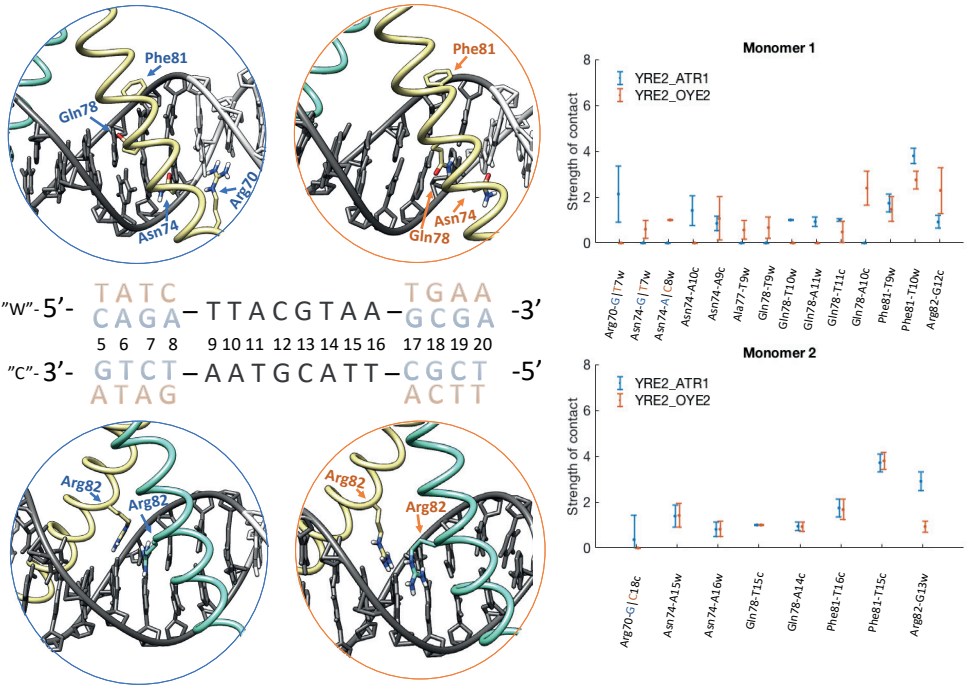


Figure 14: Specific contacts between Yap1 recognition motif (RxxxNxxAQxxFR) and YRE2 in two genomic environments, ATR1 and OYE2. For the Yap1-DNA complexes: Yap1 monomer 1 is yellow, Yap1 monomer 2 is in turquoise, and DNA is in grey, where YRE is denoted with dark grey. The ATR1-environment is blue-marked, and the OYE2-environment is orange-marked. The plots show the strength of specific contacts exploited by Yap1 monomers in two genomic environments. We define a contact strength by pairs of residues, i.e., for each pair of protein-DNA residues we sum all the contacts involving the protein residue side chains and DNA bases. Reproduced with permission from Oxford University Press[151].

Changes in shift translate into changes in twist – a key modulator of local DNA supercoiling transitions and, by extension, transcriptional control. Thus, the degree to which shift of b.p. steps within REs can adjust and be restrained by the binding of TFs, we believe will not only modulate the proteins binding affinity, but also regulate the torsional rigidity of DNA, i.e., the energetic cost of DNA supercoiling transitions. This observation together with bioinformatic analysis of differential expression of the studied genes (ATR1 and OYE2)[154] allow us to hypothesise about a mechanism how transcription factors contribute to the regulation of promoters firing potential. The DNA conformational flexibility of the RE, impacted by the flanking environment, modulate the strength of the TF-DNA association, which in turn affects the gene expression levels.

3.1 Paper II. A Sequence Environment Modulates the Impact of Methylation on the Torsional Rigidity of DNA

In paper II we were motivated to address the impact of CpG methylation on the response of DNA towards torsional stress. The methylation of CpG sites changes the physical and mechanical properties of DNA,[41] making DNA stiffer. Methylated CpG (MpG) steps experience a reduction in twist and an increase in roll angle. Since twist is involved in the regulation of local changes of supercoiling,[46,53] a key modulator of transcriptional control,[44] we hypothesised that CpG methylation may impact the DNA torsional rigidity. This could be a contributing mechanism, explaining the association of CpG methylation with transcriptional silencing.

Thus, using molecular dynamics simulations together with the twist restraint[57] that controls the total twist of a DNA fragment, we addressed the impact of CpG methylation on DNA response to torsional stress relative to unmethylated DNA. In the study, to account for sequence specific effects, we selected three repetitive sequences of tetranucleotides, that covers possible tetranucleotide compositions with a central CpG step (YCGR: TCGA; YCGY/RCGR: CCGC and RCGY: ACGT). We choose the tetranucleotides that correspond to the average- (ACGT), most- (CCGC) and least (TCGA) occurring CpG tetranucleotide motifs in the human genome.[155] For each tetranucleotide, we also studied the methylated variant, hence, six systems in total. The six systems constituted the 20 b.p. oligomers: GG-$(ACGT)_4$-GG, GG-$(TCGA)_4$-GG, GG-$(CCGC)_4$-GG, GG-ACGT-$(ACGT)_2$-ACGT-GG, GG-TCGA-$(TCGA)_2$-TCGA-GG, GG-CCGC-$(CCGC)_2$-CCGC-GG. We applied the twist restraint to the central 8 b.p. fragment of each oligomer, where DNA was gradually over- and underwound by ±0.5°/b.p./window until a maximum of ±6°/b.p.

Our computational experiments showed (Figure 15A) that CpG methylation, in a sequence specific manner, asymmetrically impacts the energy cost for DNA twisting, making DNA more torsionally rigid. For the most abundant tetranucleotide, CCGC, methylation hinders overtwisting but not undertwisting. For the least abundant tetranucleotide, TCGA, methylation exhibits a more symmetric response, and hinders both under- and over twisting. For the average abundant tetranucleotide, ACGT, methylation makes undertwisting more unfavourable but has minor effects on overtwisting. These results suggest that CpG methylation can contribute to transcriptional regulation by modulating the energy cost of supercoiling transitions, which in turn can impact the binding of proteins and the positioning of nucleosomes.

To understand a plausible mechanism for the sequence specific torsional rigidity of CpG methylation, we firstly analysed the individual contribution of each b.p. step to the absorption of twist (Figure 15B). As shown in Figure 15B, the CpG step of the CCGC and TCGA tetranucleotides effectively absorbs both negative and positive torsional stress. However, upon methylation, the MpG step is locked in a low twist state, unable to effectively absorb the applied positive torsional stress. Instead, other b.p. steps (CpC step for CMGC and TpM/GpA steps for TMGA) have to assist in the absorbing of the torsional stress during overwinding. For the ACGT step, due to the sequence environment (see section 1.1), the TpA step contributes the most to the absorption of the applied torsional stress. However, upon methylation, the TpA step is unable to attain a low twist state. Instead, the MpG step has to absorb the negative torsional stress, which costs more energy.

Secondly, we analysed other helical parameters and backbone parameters as well as steric effects of methylated cytosine (Figure 15C). This allowed us to conclude that the sequence specific induced torsional rigidity of CpG methylation can be explained by steric clashes caused by the bulky methyl group, which significantly reduces BI/BII transitions of DNA backbone. For the CMGC and TMGA tetranucleotides, a high twist state is blocked by steric clashes between the methyl group and the sugar C2'-atom of the 5'-adjacent nucleotides. For the TMGA tetranucleotide, during underwinding, another potential steric clash also arises between the methyl group of cytosine and the methyl group of the 5'-adjacent thymine, which increases the

energy cost for DNA underwinding. For the AMGT tetranucleotide the low twist state for the TpA step is prevented as this would create steric clashes between the methyl group of the cytosine (MpG step) and the C2'-atom of the 5'-adjacent nucleotides.

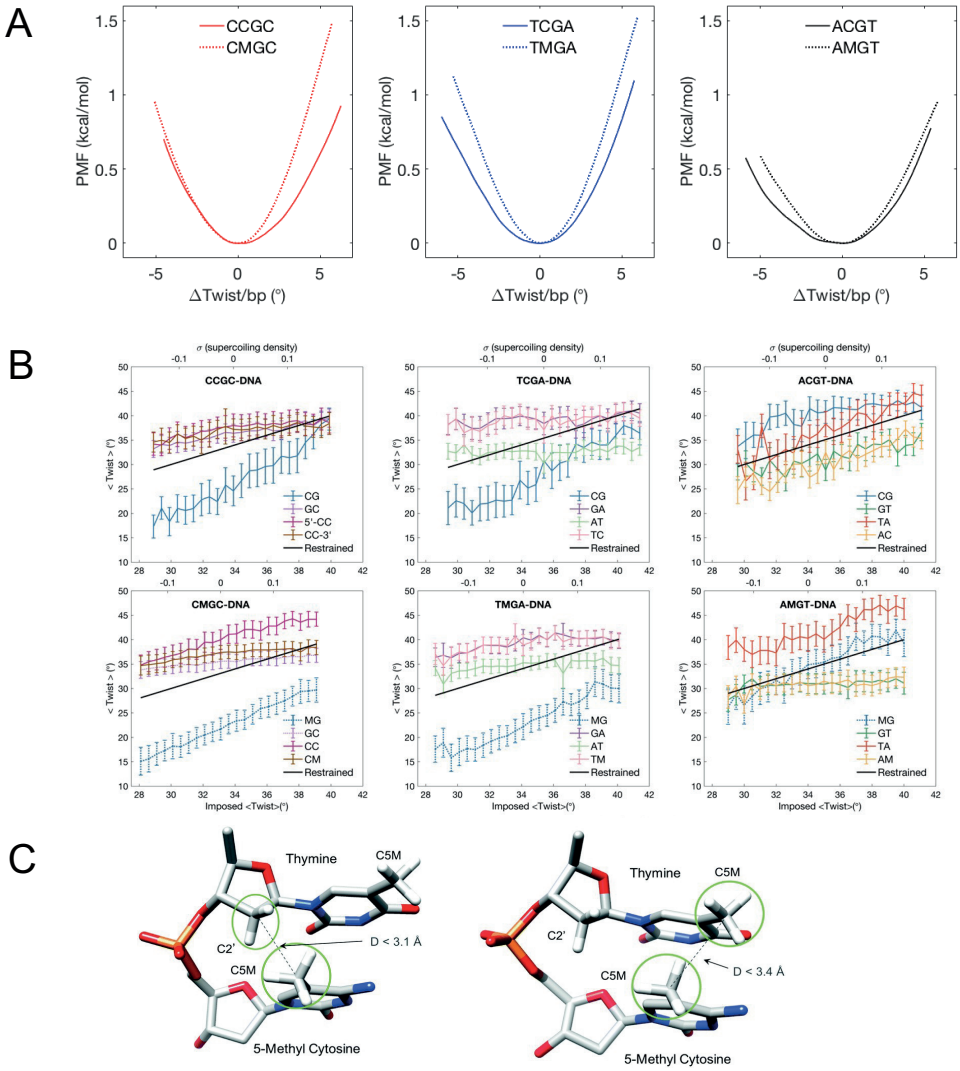


Figure 15: A. PMF plots for the oligomers, showing the energy cost for DNA twisting with respect to the average change of twist/b.p. step relative to the corresponding value derived from the relaxed state MD simulations (CCGC: 33.6°; CMGC: 33.3°; TCGA: 35.5°; TMGA: 34.0°; ACGT: 35.5°; AMGT: 34.1°). **B.** Twist response to the imposed torsion, showing the average twist of individual base pair steps going from undertwisting to overtwisting. **C.** Potential steric clashes: Between the methyl group of 5-methylcytosine and the sugar C2'–atom of the 5'-adjacent nucleotides left-hand panel). Between methyl groups of thymine and 5-methylcytosine (right-hand panel). Reproduced with permission from the Royal Society of Chemistry[156].

To conclude this study contributed with further insights into the regulatory role of CpG methylation, suggesting that CpG methylation may change the distribution of torsional stress along the genome, which could have biological regulatory role. To derive a more detailed picture of the regulatory role of CpG methylation and its impact on DNA response to torsional stress, upcoming studies should focus on longer native genomic sequences, alone as well as bound by proteins.

3.2 Paper III. Specifically Bound BZIP Transcription Factors Modulate DNA Supercoiling Transitions

Both the specific binding of TFs to their genomic target sites and torsional stress play important roles in transcriptional regulation.[44,50,52] However, what impact torsional stress has on TF recognition and TF-DNA complex stability is far from being understood. Thus, in paper III we were motivated to address the impact of torsional stress on TF-DNA complexation. As our model system, we selected human MafB[78,157], a representative of the BZIP family of transcription factors, which does not considerably deform DNA upon association. MafB recognises extended, 13-14 b.p., palindromic and pseudo-palindromic, AP-1 response elements referred to as Maf recognition element (MARE: 'TGCTGAC(G)TCAGCA').[78] By binding to MARE, MafB acts as either an activator or a repressor, modulating the transcription of genes involved in cell development and cell differentiation.[157] The recognition proceeds by the direct readout mechanism, where a six-residue motif of each MafB monomer (**R**xxx**N**xx**YA**xx**CR**) forms specific contacts with the bases from DNA major groove of each MARE half-site 'TGCTGAC'. Characteristic for the Maf family, is the Tyr residue (RxxxNxx**Y**AxxCR), which stabilises the orientation of the Arg and Asn (**R**xxx**N**xxYAxxCR) residues to extend the AP-1 consensus sequence with the 'TGC' flanks.[78]

We studied both unbound and MafB-bound DNA (Figure 16A), containing the DNA sequence 'GGTAAT**TGCTGACGTCAGCA**TTATGG' where we applied the twist restraint[57] to the palindromic MARE motif, in bold. We under- and overtwisted the MARE region gradually from 0°/b.p. to a maximum of ±5°/b.p.

Our computational experiments showed that MafB asymmetrically impacts the free energy for twisting relative to unbound DNA, where overtwisting becomes significantly more unfavourable (Figure 16B). The results also showed that MafB prefers a slightly underwound MARE motif compared to unbound DNA (34° vs 35° in average twist/b.p.). The MafB-induced torsional rigidity suggests that this transcription factor, and potentially other members of the BZIP family, can modulate DNA supercoiling transitions.

Structural analyses of helical- and groove parameters and protein-DNA contacts allowed us to explain the molecular mechanism of the increased torsional rigidity upon MafB binding. The torsionally flexible steps within the MARE region (the TpG and CpA steps, marked blue in Figure 16C) become torsionally rigid when MafB is bound. These dinucleotide steps are involved in specific contacts with the protein, which causes them to shift into the major groove and consequently locks them into a high twist substate. Instead, the induced torsional stress is distributed and absorbed by other less flexible b.p. steps, which results in an asymmetric increase in free energy for DNA twisting.

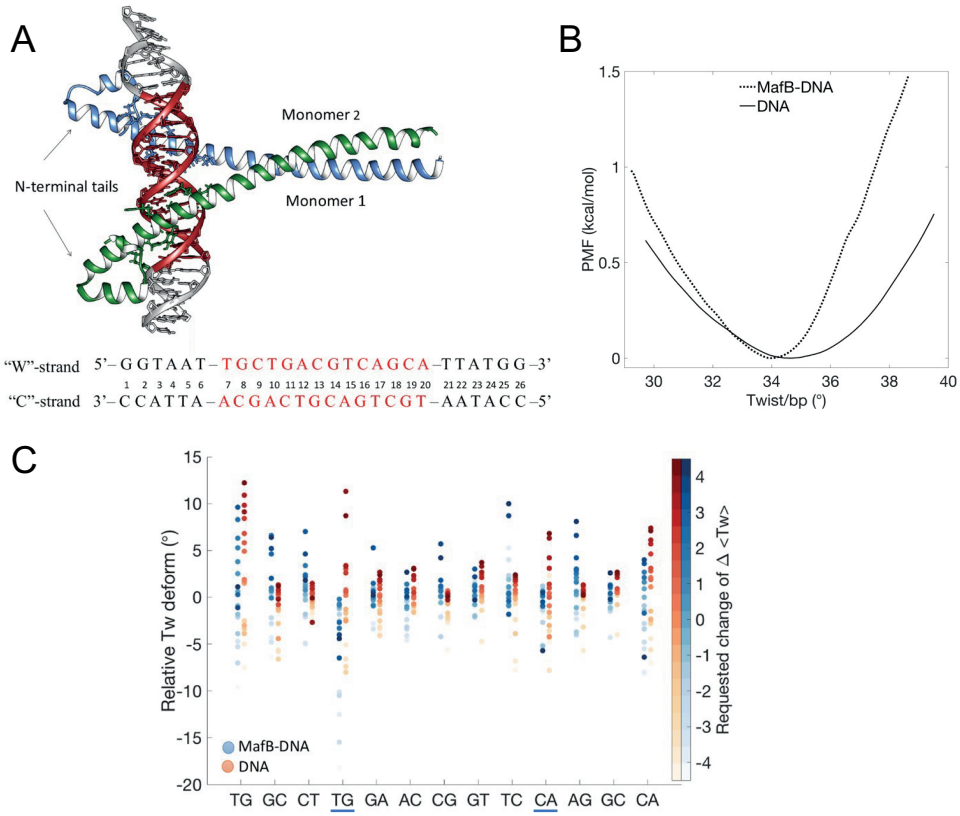


Figure 16: A. Crystal structure of MafB-DNA complex (PDB ID: 4AUW) with MARE sequence in red. **B.** PMF profiles of DNA twisting transitions with respect to average twist per b.p. step in unbound and MafB-bound MARE-DNA. **C.** Changes of twist for the restrained MARE-region in unbound and complexed DNA as a function of the requested average change of twist per b.p. step, indicated by the colorbar to the right. Reproduced with permission from Springer Nature[158].

The computational results also showed that MafB preserves most of the specific contacts with its binding partner by undergoing conformational changes (Figure 17). We believe this conformational adaptability is characteristic for the BZIP family, and explains how certain BZIP factors can function as pioneer factors,[86] binding to nucleosomes that are slightly negatively supercoiled and not fully accessible due to interactions with histones.

Although in the study, we employ only MafB BZIP transcription factor, the derived results could be generally characteristic for specifically bound BZIP factors, as BZIP factors share high level of sequence homology and follow a direct-readout mechanism of DNA recognition. The ability of MafB, and potentially other BZIP factors, to asymmetrically impact the energy cost for DNA twisting, suggests that specifically bound BZIP factors can act as torsional stress insulators, locally adjusting the topological environment by preventing the propagation of supercoiling along the chromatin fibre. This might promote cooperative binding of additional transcription factors and modulate transcription of nearby genes. At the same time, the impact of torsional stress on TF-DNA complexation, we believe, is likely to be considerably different for non-specifically bound BZIP factors, which awaits to be explored in further details.

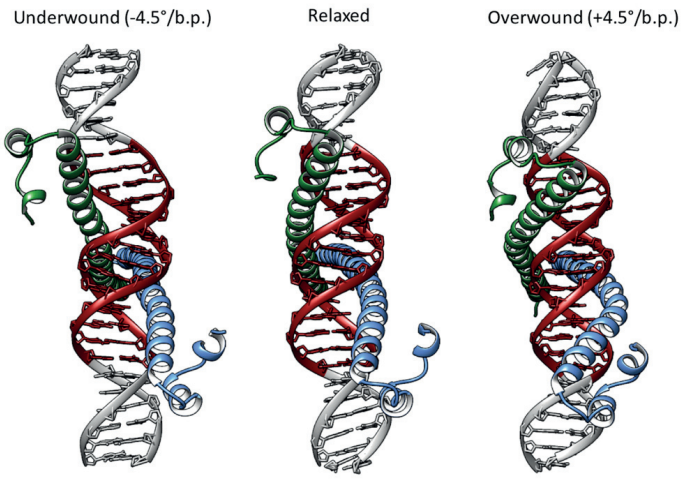


Figure 17: Conformational changes of the BZIP domain of the MafB-dimer at the underwound and overwound DNA states, with respect to the torsionally relaxed state. Reproduced with permission from Springer Nature[158].

3.4 Paper IV. Homologous Basic-Helix–Loop–Helix Transcription Factors Induce Distinct Deformations of Torsionally-stressed DNA: a Potential Transcription Regulation Mechanism

In paper IV we set to further investigate the impact of torsional stress on TFs-DNA complexation, to explore if the response varies among homologous TFs. TFs within the same subfamily often exhibit specificity towards same DNA response elements and, upon association, induce similar bioactive DNA conformations – yet they modulate different transcriptional pathways. For our model system, we selected the homologous homo- and heterodimers MaxMax, MycMax and MadMax, part of the BHLH family, which recognise E-Box (CANNTG) response elements through a conserved five-residues motif (******xx**E**xx**R***: HNxxExxRR).[94,159,160] Together they form the Myc/Max/Mad network, involved in the regulation of many house-keeping genes. The heterodimers MycMax and MadMax act as an activator and a repressor, respectively, through recruitment of different chromatin remodelling proteins. Like MadMax, the Max homodimer, antagonises MycMax, however, since Max lacks a transactivation/transrepression domain, it is considered transcriptionally inert.[159]

We studied unbound and MaxMax- MycMax- and MadMax-bound DNA containing the sequence "GGCGA*GTAG****CACGTG****CTAC*TCGC" (Figure 18A) under torsional stress from 0° to ±5°/b.p. using all-atomistic MD simulations. The restraint[57] was applied to the to the central E-Box sequence and the four adjacent flanking nucleotides.

The simulations showed, consistently to our study of MafB (paper II), that the binding of a BHLH factor also makes DNA more torsionally rigid (Figure 18B). However, the induced torsional rigidity is stronger for MafB. This is explained by the fact that MafB forms specific contacts with a longer response element (14 b.ps) compared to the BHLH factors (6 b.ps). We expect that by increasing the length of the restrained region, the PMF profile for TF-bound DNA will eventually converge towards the same profile as for unbound DNA. Nevertheless, the local TF-induced rigidity suggests that binding of a TF protein may block the propagation of torsional stress.

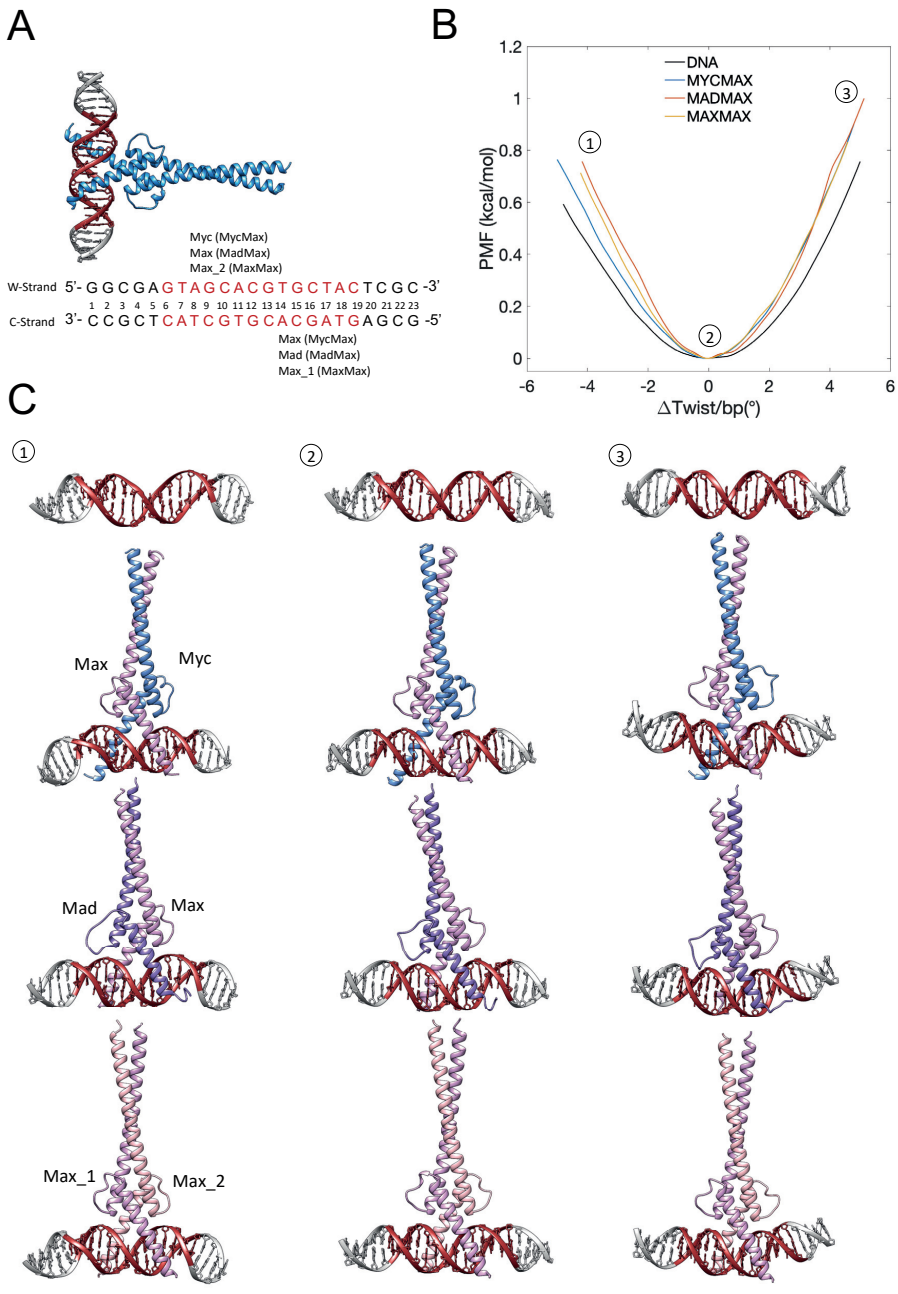


Figure 18: A. A BHLH-DNA complex structure with the studied DNA sequence underneath. The restrained region is marked red, also indicated which protein monomers bound to which side of the DNA sequence. **B.** PMF profiles for the protein-bound and unbound DNA systems, showing the energy cost for DNA twisting with respect to the average change of twist/b.p. step relative to the value of twist from the relaxed state: DNA: 34.5°, MycMax: 34.7°, MadMax: 34.0°, MaxMax:34.4°. **C.** DNA deformations

for unbound and protein-bound DNA are showed for (1) underwound regime (-4.5°/b.p.), (2) relaxed regime and (3) overwound regime (+4.5°/b.p.). Reproduced with permission from Cambridge University press[161].

Analysis of the contributions of individual of b.p. steps to the absorption of the applied torsional stress further supports this theory (Figure 19A). For unbound DNA, the torsional stress is evenly distributed over the entire restrained region, whereas for BHLH-bound DNA, the imposed stress is predominantly accumulated at the flanking sites. The three BHLH dimers exhibit similar DNA torsional rigidity, however, in the presence of torsional stress, in particular during underwinding, the dimers induce distinct DNA deformations (Figure 18C and Figure 1 in Paper IV).

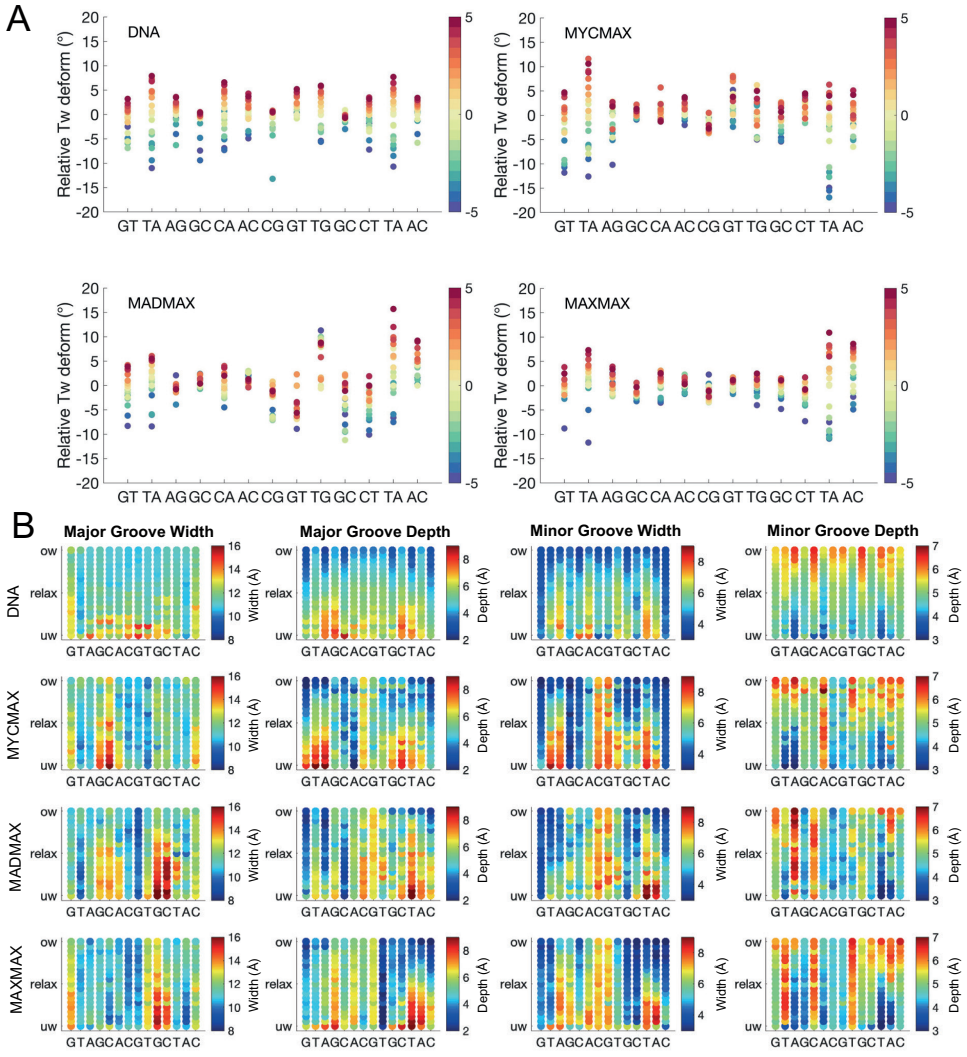


Figure 19: A. Changes of twist for the restrained DNA-region in unbound and complexed DNA as a function of the requested average change of twist per b.p. step, indicated by the colourbar to the right. **B.** Changes in major and minor groove width and depth along the torsional regimes denoted with a colourbar. Reproduced with permission from Cambridge University press[161].

The DNA deformations are characterised by changes in DNA grooves geometry (Figure 19B) and asymmetric bending, which arise due to small differences in the accumulation of imposed torsional stress at the flanking sites and differences in the protein-DNA contact networks for the different BHLH-dimers. The bending deformations, facilitated by changes in roll-angle of flanking TA steps (*G**TA**GCACGTGC**TA**C*T), are more significant for the MycMax and MaxMax dimers. The deformations occur at the Myc and Max_1 side (monomer 1 of MaxMax dimer), as these monomers form more base-specific contacts than their dimer partners, Max and Max_2 (monomer 2 of MaxMax dimer), respectively. Contrary, the Max and Max_2 monomers exhibit more nonspecific contacts with the flanking sites, which act as a steric barrier for bending towards the basic region of the BHLH proteins. MadMax-induced deformations under the underwinding regime involve changes in groove geometries but no significant bending. The loop of the Mad monomer interacts with minor groove of the first TA step of the restrained region, and the basic region of Mad forms several nonspecific contacts with the flanking nucleotides adjacent to Mad. This impacts how the torsional stress is distributed in the flanking sites and prevents bending.

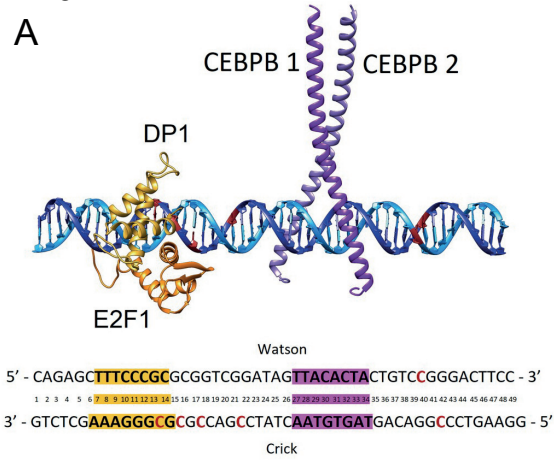
The results presented in paper IV allowed us to hypothesise that torsional stress, together with existing transactivation/transrepression domains, contribute to the execution of differential transcriptional programs of homologous TFs. The presence of torsional stress leads to TF-specific induced DNA deformations, which may allosterically impact the local topological environment or the recruitment of different collaborative TFs. It should be noted that we studied one DNA sequence. It is likely that the observed DNA deformations are also sequence specific. This may also potentially explain different transcriptional responses for different genes. Analysis of available NGS data provided some support to our hypothesis.

3.5 Paper V. Abnormal Methylation in NDUFA13 Gene Promoter of Breast Cancer Cells Breaks the Cooperative DNA Recognition by Transcription Factors

In paper V, we conducted a case study, where we looked at the impact of abnormal CpG methylation on the downregulation of NDUFA13 gene in breast cancer.[162–164] The NDUFA13 gene, also known as GRIM-19, codes a NADH dehydrogenase enzyme, part of the electron transport chain in the mitochondria, that can function as a tumour suppressor. Using NGS data[165], a hypermehylated region in the NDUFA13 promoter ~130 b.ps from transcriptional starting site (TSS) was identified through explorative bioinformatics analysis. The region contained the binding sites for two TFs (Figure 19A); E2F1-DP1 heterodimer and CEBPB homodimer, separated by 13 b.p. The E2F1-DP1 heterodimer is part of the winged-helix-family of TFs, and CEBPB homodimer is part of the BZIP family of TFs. The region contained six methylation marks, four single-methylated cytosines in the binding site of E2F1-DP1 and the linker region, and one double CpG methylation in the flanking region adjacent to the CEBPB response element (Figure 20A). To derive mechanistic insights if methylation could impact the binding of the TFs or their cooperative communication, we performed microsecond long all-atomistic MD simulations for wild type (WT) and methylated (Me) DNA unbound and bound to one or both TF dimers. Our MD simulations allowed us to derive a mechanism for the cooperative binding of the TF dimers and suggest the potential impact of the abnormal methylation on the downregulation of the NDUFA13 gene.

Our results showed an asymmetric cooperative binding order for the two TF dimers, where CEBPB binds first to facilitate the E2F1-DP1-DNA association. Analysis of the protein-DNA contacts and interaction energies (Figure 20B) showed that E2F1-DP1 forms more favourable

DNA contacts in presence of CEBPB (Figure 20C), which leads to significantly stronger interactions energies, whereas the contact network of CEBPB was independent of the E2F1-DP1 binding.



B

	Specific (kcal/mol)	Cumulative (kcal/mol)
CEBPB-DNA	-70.7 ± 11.0 -72.2 ± 11.6	-438.9 ± 63.5 -431.1 ± 65.1
E2F1-DP1-DNA	-29.5 ± 12.2 -50.2 ± 10.1	-284.6 ± 61.8 -320.3 ± 60.7
E2F1-DP1-**CEBPB-DNA**	-60.2 ± 12.5 -74.2 ± 9.8	-429.9 ± 63.7 -444.4 ± 58.7
CEBPB-**E2F1-DP1-DNA**	-68.3 ± 13.2 -81.3 ± 11.4	-383.3 ± 51.1 -396.8 ± 59.9

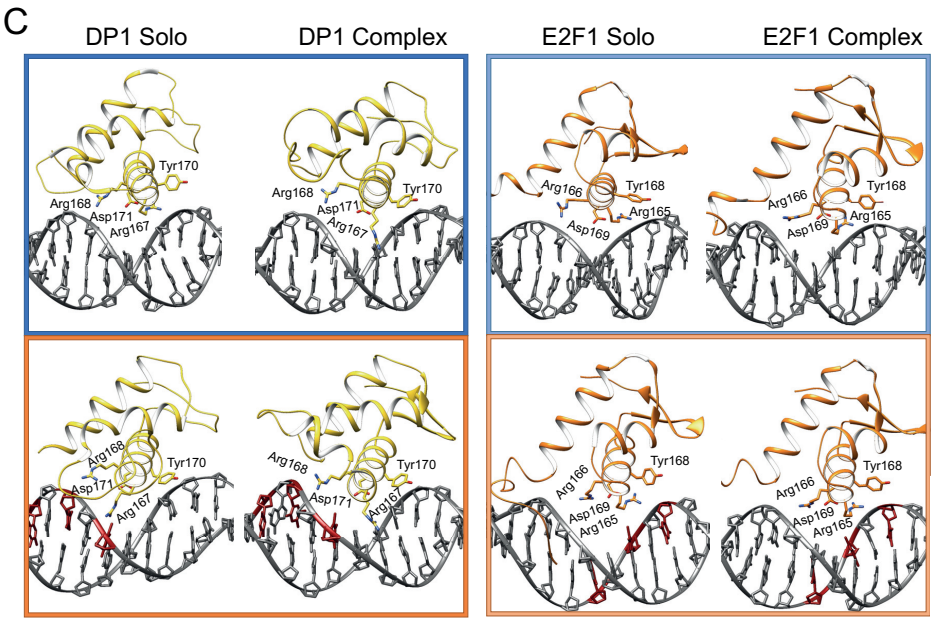


Figure 20: A. Model system for the E2F1-DP1-CEBPB-DNA enhanceosome complex. DNA Watson strand (5'-> 3') in light blue and DNA Crick strand (3'-> 5') in dark blue. Cytosines where methylations occur in cancer are highlighted with red colour. DP1 monomer in yellow, E2F1 monomer in orange, CEBPB monomer 1 in dark purple and CEBPB monomer 2 in light purple. The studied DNA sequence is outlined below, where the TF-dimer response elements are highlighted (E2F1-DP1: yellow, CEBPB: magenta), and positions of methylated sites are in red. **B.** TF-DNA specific (specific contacts) and cumulative (both nonspecific and specific contacts) interaction energies (kcal/mol) including standard deviations. For the E2F1-DP1-CEBPB-DNA trajectories, the provided interaction energies are for the protein dimer in bold. For each table row, the first value is for the wild type (WT) and the second value is

for the methylated (ME) systems. **C:** Cartoon representation of the binding orientations and DNA contacts of the E2F1-DP1 dimer for the WT (blue) and ME (orange) systems when bound alone to DNA (solo) and together with CEBPB (complex).

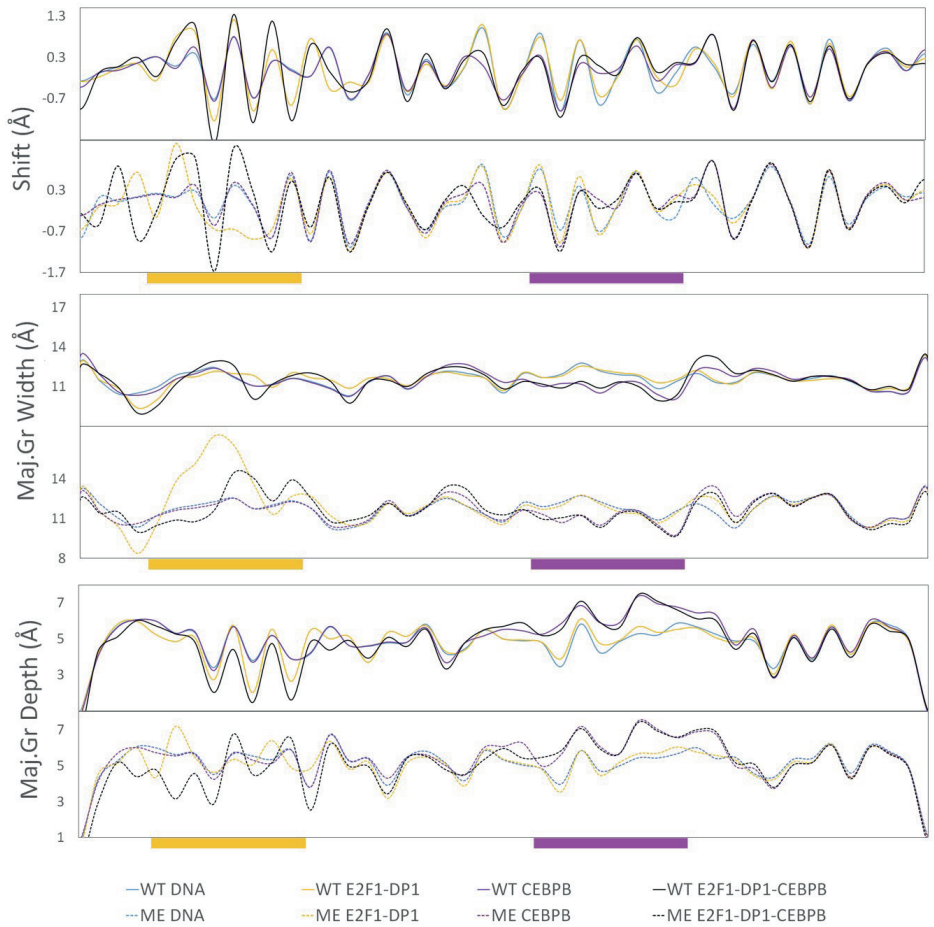


Figure 21: Average values for shift, major groove width and depth (in Å) for WT and ME DNA alone and in complex with one or both TF dimers. WT-values are depicted with bold lines and ME with dashed lines. The E2F1-DP1- and CEBPB- binding sites are denoted with a yellow- and a purple rectangular box respectively.

The origin of this could be further understood through analysis of helical and groove parameters (Figure 21). In the presence of CEBPB, b.ps within the E2F1-DP1 binding site become more displaced towards the major groove, which make the groove shallower. This allows the recognition helix of DP1 to be more favourably positioned in the major groove. Small changes in shift and major groove width and depth within the linker region (between the two binding sites) and the E2F1-DP1 binding site, at b.p. 10 and 20 (according to the model numbering, Figure 20A), are also observed for the WT solo CEBPB-DNA system. These changes may allow E2F1-DP1 to better recognise its response element. However, the signal is small, and becomes significantly stronger when both TFs dimers are bound to their corresponding binding sites. This

allowed us to hypothesise that strong cooperativity will arise, once CEBPB is already bound to DNA and E2F1-DP1 is approaching its response element.
In terms of methylation, our results imply a loss of cooperativity between the two TF dimers, either through the loss of binding of E2F1-DP1 or a disturbance in the communication between the two dimers. Analysis of protein-DNA contacts and interaction energies (Figure 20B), showed no significant impact of methylation on the binding of CEBPB. All simulated systems exhibit some differences in the CEBPB-DNA contact-network, however, these differences are due to flickering power of long-chained residues, i.e., the ability of the residues to exploit different conformational substates. In contrast, the methylation has a greater impact on the E2F1-DP1-DNA contact network. As for the WT systems, E2F1-DP1 is dependent on CEBPB to allow more favourable protein-DNA contacts that yield stronger interaction energy. Interestingly, in opposed, the interaction-energies analyses suggest a stronger E2F1-DP1 association to methylated DNA. Nevertheless, the interaction-energies for the WT and ME systems overlap within the standard deviations indicating that there is some uncertainty. In addition, for the ME-system, Arg166 of the four-residues motif of E2F1 (R**R**xYD) loses its interactions with DNA (Figure 20C) to instead interact with Asp169 of the four-residues motif of DP1 (RRxY**D**). This could indicate that the dimer is more loosely anchored to methylated DNA.

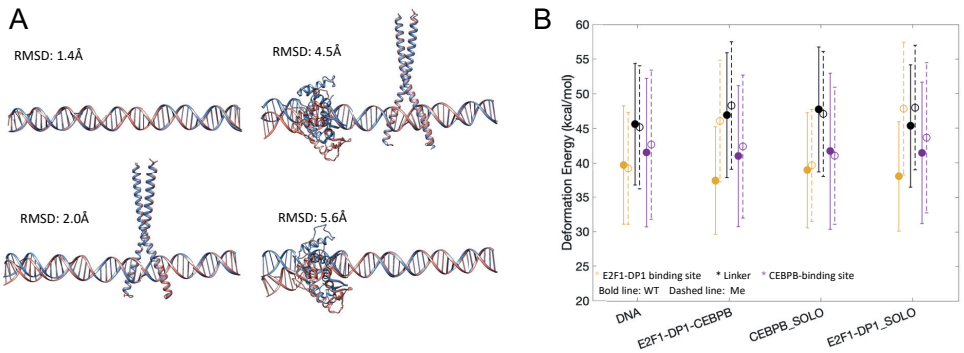


Figure 22: A. Comparison of averages structures after 1 µs simulation of WT (Blue) and Me (orange) systems. **B.** DNA deformation energies calculated with a multivariate Isling model[162] for E2F1-DP1 binding site (b.p. 6-15), linker region (b.p. 15-26) and CEBPB binding site (b.p. 26-35). WT in bold lines and Me in dashed lines.

Analysis of DNA conformational dynamics showed that CpG methylation within the E2F1-DP1 binding site induce different conformational substates in the helical and groove parameters (Figure 21), resulting in a wider and deeper major groove. In addition, increased roll-angles of the methylated CpG steps, bend ME-DNA in another direction compared to WT-DNA (Figure 22A). The DNA bending becomes more significant upon binding of E2F1-DP1. The overall changes allow E2F1-DP1 to maintain interactions with ME-DNA, however, result in a significant increase in deformation energy[166], i.e.: the energy cost for DNA helical parameters to change from their preferred state (Figure 22B). This allowed us to suggest that the E2F1-DP1 dimer will no longer recognise its DNA target as a true binding site. Instead, the dimer would rather "see" the sequence as random. The CpG methylations also result in the loss of bimodality of soft b.p. steps within the linker region between the two TF dimers binding sites. This could disturb the allosteric communication between the two TF dimers. Analysis of long-distance correlations in helical and groove parameters between all b.ps showed a different pattern for TF-bound ME-DNA. The fact that this region is near the transcription starting site, suggests that any disturbance in the binding of the transcription factors to DNA and/or their communi-

cation, would have a potential impact on of the NDUFA13 gene transcription. Our results suggest that the downregulation of the gene could potentially be explained by (1) the loss in the binding of E2F1-DP1 to ME-DNA and/or (2) the disturbance in the cooperative communication between E2F1-DP1 and CEBPB. This study shows the power of computational modelling to derive valuable information which could be further validated experimentally.

4. Concluding Remarks

Correct regulation of gene expression is vital for the healthy state of every cell in every living organism.[167] Gene expression in eukaryotes, from yeast to human, is primarily regulated at the initiation stage of DNA transcription reaction, where transcription factors must unmistakably locate and bind their response elements to facilitate the recruitment of the transcription machinery.[5,60–63] Despite years of intense research the full mapping of regulatory mechanisms of eukaryotic transcriptional control is far from being completed. Each new discovery constitutes another piece of the puzzle. With the scientific work presented in this thesis we contribute with additional pieces.

Firstly, in paper I, we provide further mechanistic insights into the sequence specific binding of BZIP TFs. DNA conformational flexibility of both the response elements and the flanking sites fine-tune the direct-redout mechanism for the BZIP-DNA recognition. Local conformational changes in helical parameters of the flanking sites, impact the conformational adaptability of the response element for the TF. We identify helical shift as a key parameter for the formation of stable specific protein-DNA contacts. In turn, the formation of stable protein-DNA contacts lock DNA in a particular "bioactive" DNA conformation, where helical shift, twist and slide are defined.

Secondly, in papers II and III, we provide mechanistic insights into how DNA modifications and transcription factor binding can modulate DNA response to supercoiling transitions and through this contribute to transcriptional control. Both CpG methylation and MafB (BZIP) binding asymmetrically increase the energy cost for DNA twisting transitions, hindering either over- or undertwisting. For CpG methylation, the torsional rigidity arises from the steric clashes of the methyl group, which is modulated by the nucleotide sequence environment. For MafB-binding, the induced torsional rigidity arises from the specific protein-DNA contacts, blocking twist-flexible dinucleotide steps that can efficiently absorb torsional stress. The study of MafB-DNA complexation also reveals additional insights into unexplored properties of the BZIP family; the BZIP domain is highly flexible and can adjust its conformation to adapt to torsionally deformed DNA to preserve stable specific binding. This can explain how BZIP factors can act as pioneer factors,[86,91] binding to nucleosomes that are slightly negatively supercoiled.

Thirdly, in paper IV, we provide insights into how torsional stress may contribute to the execution of differential transcriptional programs by homologous BHLH factors. The torsional stress is accumulated in the flanking sites for the protein-bound DNA, which further indicates that the binding of a TF may block the propagation of supercoiling. The homologous TFs exhibit similar torsional rigidity, however, the accumulation of torsional stress at the flanking sites leads to TF-specific DNA deformations. This allows us to propose that torsional stress could contribute to the differential transcriptional response of homologous TFs, by through different deformations impact local topological changes or recruitment of different collaborative TFs.

Fourthly, in paper V, we provide further insights into TF-cooperativity, where BZIP binding facilitates the binding of a collaborative TF (E2F1-DP1). The binding of the BZIP factor, reduces DNA fluctuations, and induces local allosterically induced changes in helical parameters and groove parameters within the binding site of E2F1-DP1, which are likely to be sensed by E2F1-DP1 while approaching its target site. We also show how DNA methylation within the

E2F1-DP1 binding site and the linker region, impacts the energy cost for the DNA conformational adaptability towards E2F1-DP1, which could make the binding site unrecognizable, and could be involved in the downregulation of the NDUFA13 gene leading to the onset of cancer. In conclusion, knowledge about the regulatory mechanisms encoded within DNA non-coding regions is valuable for understanding how the genome is expressed, and may also be of significance for clinical applications. It will allow to predict disease causing sequence alterations, which in turn can serve as a base for development of personalised medicines.

5. Acknowledgements

Firstly, I would like to acknowledge my supervisors Dr. Anna Reymer and Prof. Leif Eriksson for giving me the opportunity to do my PhD studies in your team. I am very grateful for having been awarded a position in a field and subject that I find both very interesting and valuable for fundamental science. I further also want to thank Anna for having you as a dear friend and great teacher. You have always encouraged me to participate in the designing of our studies, which have helped me grow as a scientist.
Secondly, I would like to acknowledge Kevin, Björn, Antonio and Isaac for great collaboration and the rest of our group members (both previous and current) for making the work environment special ☺.
Thirdly, I would like to acknowledge Martin and my family for always giving me the support I need.
Finally, I would like to acknowledge Swedish Foundation for Strategic Research SSF and Hasselblad for founding my PhD studies and Swedish National Infrastructure for Computing (SNIC) for the computing resources.

6. Bibliography

1. Crick, F. Central Dogma of Molecular Biology. *Nature* **227**, 561–563 (1970).
2. Alberts, B. DNA replication and recombination. *Nature* **421**, 431–435 (2003).
3. Kornberg, R. D. The molecular basis of eukaryotic transcription. *Proc. Natl. Acad. Sci.* **104**, 12955 – 12961 (2007).
4. Venters, B. J. & Pugh, B. F. How eukaryotic genes are transcribed. *Crit. Rev. Biochem. Mol. Biol.* **44**, 117–141 (2009).
5. Kujirai, T. & Kurumizaka, H. Transcription through the nucleosome. *Curr. Opin. Struct. Biol.* **61**, 42–49 (2020).
6. Jackson, R. J., Hellen, C. U. T. & Pestova, T. V. The mechanism of eukaryotic translation initiation and principles of its regulation. *Nat. Rev. Mol. Cell Biol.* **11**, 113–127 (2010).
7. Watson, J. D. & Crick, F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **171**, 737–738 (1953).
8. Ussery, D. W. DNA Structure: A-, B-and Z-DNA Helix Families. *e LS* (2001).
9. Chargaff, E. Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. *Experientia* **6**, 201–209 (1950).
10. Dickerson, R. E. [5] DNA structure from A to Z. *Methods Enzymol*. **211**, 67–111 (Academic Press, 1992).
11. Rich, A. & Zhang, S. Z-DNA: the long road to biological function. *Nat. Rev. Genet.* **4**, 566–572 (2003).
12. Olson, W. K. & Sussman, J. L. How flexible is the furanose ring? 1. A comparison of experimental and theoretical studies. *J. Am. Chem. Soc.* **104**, 270–278 (1982).
13. Schuerman, G. S. & Van Meervelt, L. Conformational Flexibility of the DNA

Backbone. *J. Am. Chem. Soc.* **122**, 232–240 (2000).
14. Bertrand, H.-O., Fermandjian, S., Ha-Duong, T. & Hartmann, B. Flexibility of the B-DNA backbone: Effects of local and neighbouring sequences on pyrimidine-purine steps. *Nucleic Acids Res.* **26**, 1261–1267 (1998).
15. Pasi, M. *et al.* μABC: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. *Nucleic Acids Res.* **42**, 12272–12283 (2014).
16. Lu, X.-J. & Olson, W. K. 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. *Nat. Protoc.* **3**, 1213–1227 (2008).
17. Lavery, R., Moakher, M., Maddocks, J. H., Petkeviciute, D. & Zakrzewska, K. Conformational analysis of nucleic acids revisited: Curves+. *Nucleic Acids Res.* **37**, 5917–5929 (2009).
18. Olson, W. K., Gorin, A. A., Lu, X.-J. J., Hock, L. M. & Zhurkin, V. B. DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 11163–11168 (1998).
19. Dans, P. D., Pérez, A., Faustino, I., Lavery, R. & Orozco, M. Exploring polymorphisms in B-DNA helical conformations. *Nucleic Acids Res.* **40**, 10668–78 (2012).
20. Dans, P. D. *et al.* The static and dynamic structural heterogeneities of B-DNA: extending Calladine–Dickerson rules. *Nucleic Acids Res.* **47**, 11090–11102 (2019).
21. Balaceanu, A. *et al.* Modulation of the helical properties of DNA: next-to-nearest neighbour effects and beyond. *Nucleic Acids Res.* **47**, 4418–4430 (2019).
22. Annunziato, A. DNA packaging: nucleosomes and chromatin. *Nat. Educ.* **1**, 26 (2008).
23. Murakami, Y. Heterochromatin and Euchromatin. *Encyclopedia of Systems Biology;* Dubitzky, W., Wolkenhauer, O., Cho, K.-H. & Yokota, H. Eds.; Springer New York, 881–884 (2013).
24. Szerlong, H. J. & Hansen, J. C. Nucleosome distribution and linker DNA: connecting nuclear function to dynamic chromatin structure. *Biochem. Cell Biol.* **89**, 24–34 (2011).
25. Bannister, A. J. & Kouzarides, T. Regulation of chromatin by histone modifications. *Cell Res.* **21**, 381–395 (2011).
26. Kumar, S., Chinnusamy, V. & Mohapatra, T. Epigenetics of Modified DNA Bases: 5-Methylcytosine and Beyond. *Front Genet*. **9**, 640 (2018).
27. Greenberg, M. V. C. & Bourc'his, D. The diverse roles of DNA methylation in mammalian development and disease. *Nat. Rev. Mol. Cell Biol.* **20**, 590–607 (2019).
28. Newell-Price, J., Clark, A. J. L. & King, P. DNA Methylation and Silencing of Gene Expression. *Trends Endocrinol. Metab.* **11**, 142–148 (2000).
29. Curradi, M., Izzo, A., Badaracco, G. & Landsberger, N. Molecular mechanisms of gene silencing mediated by DNA methylation. *Mol. Cell. Biol.* **22**, 3157–3173 (2002).
30. Ehrlich, M. & Lacey, M. DNA methylation and differentiation: silencing, upregulation and modulation of gene expression. *Epigenomics* **5**, 553–568 (2013).
31. Strichman-Almashanu, L. Z. *et al.* A genome-wide screen for normally methylated human CpG islands that can identify novel imprinted genes. *Genome Res.* **12**, 543–554 (2002).
32. Ghosh, S. *et al.* Tissue specific DNA methylation of CpG islands in normal human adult somatic tissues distinguishes neural from non-neural tissues. *Epigenetics* **5**, 527–538 (2010).
33. Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* **13**, 484–492 (2012).
34. Jin, B., Li, Y. & Robertson, K. D. DNA methylation: superior or subordinate in the

epigenetic hierarchy? *Genes Cancer* **2**, 607–617 (2011).
35. Ehrlich, M. DNA methylation in cancer: too much, but also too little. *Oncogene* **21**, 5400–5413 (2002).
36. Ehrlich, M. DNA hypomethylation in cancer cells. *Epigenomics* **1**, 239–259 (2009).
37. Suzuki, M. M. & Bird, A. DNA methylation landscapes: provocative insights from epigenomics. *Nat. Rev. Genet.* **9**, 465–476 (2008).
38. Derreumaux, S., Chaoui, M., Tevanian, G. & Fermandjian, S. Impact of CpG methylation on structure, dynamics and solvation of cAMP DNA responsive element. *Nucleic Acids Res.* **29**, 2314–2326 (2001).
39. Klose, R. J. & Bird, A. P. Genomic DNA methylation: the mark and its mediators. *Trends Biochem. Sci.* **31**, 89–97 (2006).
40. Machado, A. C. D. *et al.* Evolving insights on how cytosine methylation affects protein-DNA binding. *Brief. Funct. Genomics* **14**, 61–73 (2015).
41. Pérez, A. *et al.* Impact of methylation on the physical properties of DNA. *Biophys. J.* **102**, 2140–2148 (2012).
42. Buitrago, D. *et al.* Impact of DNA methylation on 3D genome structure. *Nat. Commun.* **12**, 3243 (2021).
43. Liebl, K. & Zacharias, M. How methyl–sugar interactions determine DNA structure and flexibility. *Nucleic Acids Res.* **47**, 1132–1140 (2018).
44. Lavelle, C. DNA torsional stress propagates through chromatin fiber and participates in transcriptional regulation. *Nat. Struct. Mol. Biol.* **15**, 123–5 (2008).
45. Ma, J., Bai, L. & Wang, M. D. Transcription under torsion. *Science* **340**, 1580–3 (2013).
46. Corless, S. & Gilbert, N. Effects of DNA supercoiling on chromatin architecture. *Biophys. Rev.* **8**, 245–258 (2016).
47. Kouzine, F. *et al.* Transcription-dependent dynamic supercoiling is a short-range genomic force. *Nat. Struct. Mol. Biol.* **20**, 396–403 (2013).
48. Naughton, C. *et al.* Transcription forms and remodels supercoiling domains unfolding large-scale chromatin structures. *Nat. Struct. Mol. Biol.* **20**, 387–395 (2013).
49. Irobalieva, R. N. *et al.* Structural diversity of supercoiled DNA. *Nat. Commun.* **6**, 1–10 (2015).
50. Teves, S. S. & Henikoff, S. Transcription-generated torsional stress destabilizes nucleosomes. *Nat. Struct. Mol. Biol.* **21**, 88–94 (2014).
51. Noy, A., Sutthibutpong, T. & A. Harris, S. Protein/DNA interactions in complex DNA topologies: expect the unexpected. *Biophys. Rev.* **8**, 233–243 (2016).
52. Kaczmarczyk, A., Meng, H., Ordu, O., Noort, J. van & Dekker, N. H. Chromatin fibers stabilize nucleosomes under torsional stress. *Nat. Commun.* **11**, 126 (2020).
53. Muskhelishvili, G. & Travers, A. The regulatory role of DNA supercoiling in nucleoprotein complex assembly and genetic activity. *Biophys. Rev.* **8**, 5–22 (2016).
54. Liebl, K. & Zacharias, M. How global DNA unwinding causes non-uniform stress distribution and melting of DNA. *PLoS One* **15**, e0232976 (2020).
55. Randall, G. L., Zechiedrich, L. & Pettitt, B. M. In the absence of writhe, DNA relieves torsional stress with localized, sequence-dependent structural failure to preserve B-form. *Nucleic Acids Res.* **37**, 5568–5577 (2009).
56. Kannan, S., Kohlhoff, K. & Zacharias, M. B-DNA under stress: over- and untwisting of DNA during molecular dynamics simulations. *Biophys. J.* **91**, 2956–2965 (2006).
57. Reymer, A., Zakrzewska, K. & Lavery, R. Sequence-dependent response of DNA to torsional stress: a potential biological regulation mechanism. *Nucleic Acids Res.* **46**, 1684–1694 (2018).
58. Kornberg, R. D. & Lorch, Y. Primary Role of the Nucleosome. *Mol. Cell* **79**, 371–375

(2020).
59. Lambert, S. A. *et al.* The Human Transcription Factors. *Cell* **172**, 650–665 (2018).
60. Biłas, R., Szafran, K., Hnatuszko-Konka, K. & Kononowicz, A. K. Cis-regulatory elements used to control gene expression in plants. *Plant Cell, Tissue Organ Cult.* **127**, 269–287 (2016).
61. Mitsis, T. *et al.* Transcription factors and evolution: An integral part of gene expression (Review). *World Acad Sci J.* **2**, 3–8 (2020).
62. Spitz, F. & Furlong, E. E. M. Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.* **13**, 613–626 (2012).
63. Li, B., Carey, M. & Workman, J. L. The Role of Chromatin during Transcription. *Cell* **128**, 707–719 (2007).
64. Yesudh, D., Batool, M., Anwar, M. A., Panneerselvam, S. & Choi, S. Proteins Recognizing DNA : Structural Uniqueness and Versatility of DNA-Binding Domains in Stem Cell Transcription Factors. *MDPI-Genees.* **8**, 1–22 (2017).
65. Murphy, E. C., Zhurkin, V. B., Louis, J. M., Cornilescu, G. & Clore, G. M. Structural Basis for SRY-dependent 46-X,Y Sex Reversal: Modulation of DNA Bending by a Naturally Occurring Point Mutation. *J. Mol. Biol.* **312**, 481–499 (2001).
66. Nikolov, D. B. *et al.* Crystal structure of a human TATA box-binding protein/TATA element complex. *Proc. Natl. Acad. Sci.* **93**, 4862 – 4867 (1996).
67. Liu, Y. *et al.* Structural basis for Klf4 recognition of methylated DNA. *Nucleic Acids Res.* **42**, 4859–4867 (2014).
68. Esch, D. *et al.* A unique Oct4 interface is crucial for reprogramming to pluripotency. *Nat. Cell Biol.* **15**, 295–301 (2013).
69. Glover, J. N. M. & Harrison, S. C. Crystal structure of the heterodimeric bZIP transcription factor c-Fos–c-Jun bound to DNA. *Nature* **373**, 257–261 (1995).
70. Zheng, N., Fraenkel, E., Pabo, C. O. & Pavletich, N. P. Structural basis of DNA recognition by the heterodimeric cell cycle transcription factor E2F–DP. *Genes Dev.* **13**, 666–674 (1999).
71. Ghosh, G., Duyne, G. Van, Ghosh, S. & Sigler, P. B. Structure of NF-κB p50 homodimer bound to a κB site. *Nature* **373**, 303–310 (1995).
72. Nair, S. K. & Burley, S. K. X-Ray Structures of Myc-Max and Mad-Max Recognizing DNA: Molecular Bases of Regulation by Proto-Oncogenic Transcription Factors. *Cell* **112**, 193–205 (2003).
73. Yu, C.-P., Lin, J.-J. & Li, W.-H. Positional distribution of transcription factor binding sites in Arabidopsis thaliana. *Sci. Rep.* **6**, 25164 (2016).
74. Zakrzewska, K. & Lavery, R. Towards a molecular view of transcriptional control. *Curr. Opin. Struct. Biol.* **22**, 160–167 (2012).
75. Todeschini, A. L., Georges, A. & Veitia, R. A. Transcription factors: Specific DNA binding and specific gene regulation. *Trends Genet.* **30**, 211–219 (2014).
76. Watkins, D., Hsiao, C., Woods, K. K., Koudelka, G. B. & Williams, L. D. P22 c2 repressor-operator complex: Mechanisms of direct and indirect readout. *Biochemistry* **47**, 2325–2338 (2008).
77. Fujii, Y., Shimizu, T., Toda, T., Yanagida, M. & Hakoshima, T. Structural basis for the diversity of DNA recognition by bZIP transcription factors. *Nat. Struct. Biol.* **7**, 889 (2000).
78. Kurokawa, H. *et al.* Structural basis of alternative DNA recognition by Maf transcription factors. *Mol. Cell. Biol.* **29**, 6232–6244 (2009).
79. Dror, I., Zhou, T., Mandel-Gutfreund, Y. & Rohs, R. Covariation between homeodomain transcription factors and the shape of their DNA binding sites. *Nucleic Acids Res.* **42**, 430–441 (2014).

80. Bouvier, B., Zakrzewska, K. & Lavery, R. Protein-DNA recognition triggered by a DNA conformational switch. *Angew. Chem. Int. Ed. Engl.* **50**, 6516–6518 (2011).
81. Yonetani, Y. & Kono, H. Sequence dependencies of DNA deformability and hydration in the minor groove. *Biophys. J.* **97**, 1138–47 (2009).
82. Tsui, V., Radhakrishnan, I., Wright, P. E. & Case, D. A. NMR and molecular dynamics studies of the hydration of a zinc finger-DNA complex11Edited by M. F. Summers. *J. Mol. Biol.* **302**, 1101–1117 (2000).
83. Battistini, F. *et al.* How B-DNA Dynamics Decipher Sequence-Selective Protein Recognition. *J. Mol. Biol.* **431**, 3845–3859 (2019).
84. Dror, I., Rohs, R. & Mandel-Gutfreund, Y. How motif environment influences transcription factor search dynamics: Finding a needle in a haystack. *Bioessays* **38**, 605–612 (2016).
85. Castellanos, M., Mothi, N. & Muñoz, V. Eukaryotic transcription factors can track and control their target genes using DNA antennas. *Nat. Commun.* **11**, 540 (2020).
86. Cohen, D. M., Lim, H. W., Won, K. J. & Steger, D. J. Shared nucleotide flanks confer transcriptional competency to bZip core motifs. *Nucleic Acids Res.* **46**, 8371–8384 (2018).
87. Yella, V. R. *et al.* Flexibility and structure of flanking DNA impact transcription factor affinity for its core motif. *Nucleic Acids Res.* **46**, 11883–11897 (2018).
88. Jindrich, K. & Degnan, B. M. The diversification of the basic leucine zipper family in eukaryotes correlates with the evolution of multicellularity. *BMC Evol. Biol.* **16**, 28 (2016).
89. Yang, Y. & Cvekl, A. Large Maf Transcription Factors: Cousins of AP-1 Proteins and Important Regulators of Cellular Differentiation. *Einstein J. Biol. Med.* **23**, 2–11 (2007).
90. Karin, M., Liu, Z. & Zandi, E. AP-1 function and regulation. *Curr. Opin. Cell Biol.* **9**, 240–246 (1997).
91. Rodríguez-Martínez, J. A., Reinke, A. W., Bhimsaria, D., Keating, A. E. & Ansari, A. Z. Combinatorial bZIP dimers display complex DNA-binding specificity landscapes. *Elife* **6**, e19272 (2017).
92. Maciaszczyk-Dziubinska, E. *et al.* The ancillary N-terminal region of the yeast AP-1 transcription factor Yap8 contributes to its DNA binding specificity. *Nucleic Acids Res.* **48,** 5426–5441 (2020).
93. Atchley, W. R. & Fitch, W. M. A natural classification of the basic helix-loop-helix class of transcription factors. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 5172–5176 (1997).
94. De Masi, F. *et al.* Using a structural and logics systems approach to infer bHLH–DNA binding specificity determinants. *Nucleic Acids Res.* **39**, 4553–4563 (2011).
95. Gajiwala, K. S. & Burley, S. K. Winged helix proteins. *Curr. Opin. Struct. Biol.* **10**, 110–116 (2000).
96. Wei, Z. *et al.* Crystal Structure of Human eIF3k, the First Structure of eIF3 Subunits *. *J. Biol. Chem.* **279**, 34983–34990 (2004).
97. Morgunova, E. & Taipale, J. Structural perspective of cooperative transcription factor binding. *Curr. Opin. Struct. Biol.* **47**, 1–8 (2017).
98. Ibarra, I. L. *et al.* Mechanistic insights into transcription factor cooperativity and its impact on protein-phenotype interactions. *Nat. Commun.* **11**, 124 (2020).
99. Kim, S. *et al.* Probing Allostery Through DNA. *Science.* **339**, 816–819 (2013).
100. Balaceanu, A., Pérez, A., Dans, P. D. & Orozco, M. Allosterism and signal transfer in DNA. *Nucleic Acids Res.* **46**, 7554–7565 (2018).
101. Tan, C. & Takada, S. Nucleosome allostery in pioneer transcription factor binding. *Proc. Natl. Acad. Sci.* **117**, 20586–20596 (2020).

102. Merino, F., Bouvier, B. & Cojocaru, V. Cooperative DNA Recognition Modulated by an Interplay between Protein-Protein Interactions and DNA-Mediated Allostery. *PLOS Comput. Biol.* **11**, e1004287 (2015).
103. Pan, Y. & Nussinov, R. The Role of Response Elements Organization in Transcription Factor Selectivity: The IFN-β Enhanceosome Example. *PLoS Comput. Biol.* **7**, e1002077 (2011).
104. Panne, D., Maniatis, T. & Harrison, S. C. Crystal structure of ATF-2/c-Jun and IRF-3 bound to the interferon-beta enhancer. *EMBO J.* **23**, 4384–4393 (2004).
105. Lewars, E. Computational chemistry. *Introd. to theory Appl. Mol. quantum Mech.* 318 (2003).
106. Jensen, F. *Introduction to computational chemistry*. (John wiley & sons, 2017).
107. Vanommeslaeghe, K. & Guvench, O. Molecular mechanics. *Curr. Pharm. Des.* **20**, 3281–3292 (2014).
108. Maier, J. A. *et al.* ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **11**, 3696–3713 (2015).
109. Ivani, I. *et al.* Parmbsc1: a refined force field for DNA simulations. *Nat. Methods* **13**, 55–58 (2016).
110. Ponder, J. W. & Case, D. A. Force fields for protein simulations. *Adv. Prot. Chem.* **66**, 27–85 (2003).
111. Abraham, M. J., Van Der Spoel, D., Lindahl, E. & Hess, B. GROMACS user manual version 5.0. 4. *Sweden R. Inst. Technol. Uppsala Univ.* (2014).
112. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **98**, 10089–10092 (1993).
113. Dans, P. D. *et al.* How accurate are accurate force-fields for B-DNA? *Nucleic Acids Res.* **45**, 4217–4230 (2017).
114. Braun, E. *et al.* Best Practices for Foundations in Molecular Simulations [Article v1.0]. *Living J. Comput. Mol. Sci.* **1**, 5957 (2019).
115. Case, D. A. *et al.* The Amber biomolecular simulation programs. *J. Comput. Chem.* **26**, 1668–1688 (2005).
116. Brooks, B. R. *et al.* CHARMM: the biomolecular simulation program. *J. Comput. Chem.* **30**, 1545–1614 (2009).
117. Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1**–**2**, 19–25 (2015).
118. Phillips, J. C. *et al.* Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **26**, 1781–1802 (2005).
119. Verlet, L. Computer" experiments" on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules. *Phys. Rev.* **159**, 98 (1967).
120. Van Gunsteren, W. F. & Berendsen, H. J. C. A leap-frog algorithm for stochastic dynamics. *Mol. Simul.* **1**, 173–185 (1988).
121. Hess, B., Bekker, H., Berendsen, H. J. C. & Fraaije, J. G. E. M. LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.* **18**, 1463–1472 (1997).
122. Ryckaert, J.-P., Ciccotti, G. & Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **23**, 327–341 (1977).
123. Bernardi, R. C., Melo, M. C. R. & Schulten, K. Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochim. Biophys. Acta* **1850**, 872–877 (2015).
124. Qi, R., Wei, G., Ma, B. & Nussinov, R. Replica Exchange Molecular Dynamics: A Practical Application Protocol with Solutions to Common Problems and a Peptide Aggregation and Self-Assembly Example. *Methods Mol. Biol.* **1777**, 101–119 (2018).

125. Izrailev, S. *et al.* Steered Molecular Dynamics. *Computational Molecular Dynamics: Challenges, Methods, Ideas;* Deuflhard, P. et al. Eds.; Springer Berlin, 39–65 (1999).
126. Barducci, A., Bonomi, M. & Parrinello, M. Metadynamics. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **1**, 826–843 (2011).
127. Kästner, J. Umbrella sampling. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **1**, 932–942 (2011).
128. Torrie, G. M. & Valleau, J. P. Monte Carlo free energy estimates using non-Boltzmann sampling: Application to the sub-critical Lennard-Jones fluid. *Chem. Phys. Lett.* **28**, 578–581 (1974).
129. Torrie, G. M. & Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.* **23**, 187–199 (1977).
130. Kumar, S., Rosenberg, J. M., Bouzida, D., Swendsen, R. H. & Kollman, P. A. THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comput. Chem.* **13**, 1011–1021 (1992).
131. Mishra, S. Fundamentals of Homology Modeling Steps and Comparison among Important Bioinformatics Tools: An Overview Akansha Saxena Halberg Hospital and Research Center, Civil Lines, Moradabad 244 001, UP, India Rajender Singh Sangwan Central Institute of Medicinal a. *Sci. Int.* **1**, (2013).
132. Wiltgen, M. & Tilz, G. P. Homology modelling: a review about the method on hand of the diabetic antigen GAD 65 structure prediction. *Wiener Medizinische Wochenschrift* **159**, 112–125 (2009).
133. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
134. Land, H. & Humble, M. S. YASARA: A Tool to Obtain Structural Guidance in Biocatalytic Investigations. Methods Mol. Biol. **1685,** 43-67 (2018)
135. Waterhouse, A. *et al.* SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* **46**, W296–W303 (2018).
136. Šali, A. & Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779–815 (1993).
137. Webb, B. & Sali, A. Comparative protein structure modeling using MODELLER. *Curr. Protoc. Bioinforma.* **54**, 5–6 (2016).
138. Song, Y. *et al.* High-resolution comparative modeling with RosettaCM. *Structure* **21**, 1735–1742 (2013).
139. Yan, Y., Zhang, D., Zhou, P., Li, B. & Huang, S.-Y. HDOCK: a web server for protein–protein and protein–DNA/RNA docking based on a hybrid strategy. *Nucleic Acids Res.* **45**, W365–W373 (2017).
140. Dijk, M. & Bonvin, A. A protein–DNA docking benchmark. *Nucleic Acids Res.* **36**, e88 (2008).
141. Huang, S.-Y. Search strategies and evaluation in protein–protein docking: principles, advances and challenges. *Drug Discov. Today* **19**, 1081–1096 (2014).
142. Vajda, S., Hall, D. R. & Kozakov, D. Sampling and scoring: a marriage made in heaven. *Proteins* **81**, 1874–1884 (2013).
143. van Dijk, M. & Bonvin, A. M. J. J. Pushing the limits of what is achievable in protein-DNA docking: benchmarking HADDOCK's performance. *Nucleic Acids Res.* **38**, 5634–5647 (2010).
144. de Vries, S. J., van Dijk, M. & Bonvin, A. M. J. J. The HADDOCK web server for data-driven biomolecular docking. *Nat. Protoc.* **5**, 883–897 (2010).
145. Trott, O. & Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **31**, 455–461 (2010).
146. Lyskov, S. & Gray, J. J. The RosettaDock server for local protein-protein docking.

Nucleic Acids Res. **36**, W233–W238 (2008).
147. Torchala, M., Moal, I. H., Chaleil, R. A. G., Fernandez-Recio, J. & Bates, P. A. SwarmDock: a server for flexible protein–protein docking. *Bioinformatics* **29**, 807–809 (2013).
148. Pierce, B. G. *et al.* ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers. *Bioinformatics* **30**, 1771–1773 (2014).
149. Shaulian, E. & Karin, M. AP-1 as a regulator of cell life and death. *Nat. Cell Biol.* **4**, E131–E136 (2002).
150. Garces de Los Fayos Alonso, I. *et al.* The Role of Activator Protein-1 (AP-1) Family Members in CD30-Positive Lymphomas. *Cancers (Basel).* **10**, 93 (2018).
151. Hörberg, J., Moreau, K., Tamás, M. J. & Reymer, A. Sequence-specific dynamics of DNA response elements and their flanking sites regulate the recognition by AP-1 transcription factors. *Nucleic Acids Res.* **49,** 9280-9293 (2021).
152. Rodrigues-Pousada, C. *et al.* Yeast AP-1 like transcription factors (Yap) and stress response: a current overview. *Microb. cell (Graz, Austria)* **6**, 267–285 (2019).
153. Jelinsky, S. *et al.* A systems approach to delineate functions of paralogous transcription factors: Role of the Yap family in the DNA damage response. *PNAS* **96**, 1486–1491 (2008).
154. Salin, H. *et al.* Structure and properties of transcriptional networks driving selenite stress response in yeasts. *BMC Genomics* **9**, 333 (2008).
155. Dans, P. D. *et al.* Unraveling the sequence-dependent polymorphic behavior of d(CpG) steps in B-DNA. *Nucleic Acids Res.* **42**, 11304–11320 (2014).
156. Hörberg, J. & Reymer, A. A sequence environment modulates the impact of methylation on the torsional rigidity of DNA. *Chem. Commun.* **54**, 11885–11888 (2018).
157. Textor, L. C., Wilmanns, M. & Holton, S. J. Expression, purification, crystallization and preliminary crystallographic analysis of the mouse transcription factor MafB in complex with its DNA-recognition motif Cmare. *Acta Crystallogr. Sect. F. Struct. Biol. Cryst. Commun.* **63**, 657–661 (2007).
158. Hörberg, J. & Reymer, A. Specifically bound BZIP transcription factors modulate DNA supercoiling transitions. *Sci. Rep.* **10**, (2020).
159. Grandori, C., Cowley, S. M., James, L. P. & Eisenman, R. N. The Myc/Max/Mad Network and the Transcriptional Control of Cell Behavior. *Annu. Rev. Cell Dev. Biol.* **16**, 653–699 (2000).
160. Giardino Torchia, M. L. & Ashwell, J. D. Getting MAD at MYC. *Proc. Natl. Acad. Sci.* **115**, 9821 LP – 9823 (2018).
161. Hörberg, J., Moreau, K. & Reymer, A. Homologous basic helix–loop–helix transcription factors induce distinct deformations of torsionally-stressed DNA: a potential transcription regulation mechanism. *QRB Discov.* **3**, e4 (2022).
162. Wang, X. *et al.* Mitochondrial GRIM-19 deficiency facilitates gastric cancer metastasis through oncogenic ROS-NRF2-HO-1 axis via a NRF2-HO-1 loop. *Gastric Cancer* **24**, 117–132 (2021).
163. Zhou, T. *et al.* Down-regulation of GRIM-19 is associated with STAT3 overexpression in breast carcinomas. *Hum. Pathol.* **44**, (2013).
164. Pinto, M. & Máximo, V. NDUFA13 (NADH:ubiquinone oxidoreductase subunit A13). *Atlas Genet. Cytogenet. Oncol. Haematol.* (2018) doi:10.4267/2042/66061.
165. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
166. Liebl, K. & Zacharias, M. Accurate modeling of DNA conformational flexibility by a multivariate Ising model. *Proc. Natl. Acad. Sci.* **118**, e2021263118 (2021).

167. Lee, T. I. & Young, R. A. Transcriptional regulation and its misregulation in disease. *Cell* **152**, 1237–1251 (2013).