Abstract

Telomeres are complex nucleoprotein assemblies that play a vital role in the maintenance of functional ends of linear chromosomes. Telomeric DNA, composed of tandem repeats of the 5′-TTAGGG-3′ motif, solves the so-called end replication problem: as chromosomes shorten with each cell division, no information is lost, and the telomere can be re-extended. In the cell, many protein factors regulate telomere length, nuclear positioning and conformation in response to cell cycle progression and the cell's proliferative status. Several proteins bind directly to single- or double-stranded telomeric DNA to assemble the main shelterin complex or play accessory roles. However, these interactions will be perturbed when the easily oxidized telomeric DNA is exposed to oxidative stress. In my doctoral work, I used Molecular Dynamics approaches to study the dynamics of protein-DNA complex formation at telomeres on the atomistic level, arriving at the most comprehensive thermodynamic, kinetic and mechanistic description of this process to date, including the first observation of spontaneous complex formation. I then investigated the impact of oxidative lesions on telomeric proteins, showing how base modifications disrupt sequence recognition on telomeric DNA. Finally, I used quantum chemical simulations to assess the feasibility of covalent protein-DNA cross-link formation on telomeres.

Streszczenie

Telomery to struktury nukleoproteinowe odpowiedzialne za utrzymanie funkcjonalnych końców liniowych chromosomów. Telomerowe DNA, zbudowane z tandemowo ułożonych powtórzeń motywu 5′-TTAGGG-3′, rozwiązuje tzw. problem replikacji końca: skracanie telomerów podczas podziału komórki staje się odwracalnym procesem, w trakcie którego nie jest tracona informacja genetyczna. Wiele białek reguluje długość, lokalizację jądrową i konformację telomeru w zależności od fazy cyklu komórkowego i statusu proliferacyjnego komórki; kilka spośród nich bezpośrednio wiąże jedno- lub dwuniciowe DNA telomerowe, inicjując formowanie kompleksu szelteryny lub pełniąc dodatkowe funkcje. Oddziaływania te zostaną jednak zaburzone, gdy podatne na utlenianie DNA telomerowe eksponowane będzie na stres oksydacyjny. W mojej pracy doktorskiej użyłem symulacji molekularnych w celu zbadania dynamiki formowania telomerowych kompleksów białko-DNA z atomową rozdzielczością, otrzymując pierwszy tak szczegółowy mechanistyczny opis termodynamiki i kinetyki tego procesu oraz pierwszą trajektorię opisującą spontaniczne formowanie kompleksu. Następnie zbadałem wpływ uszkodzeń DNA na białka telomerowe, pokazując, w jaki sposób modyfikacje zasad azotowych zaburzają rozpoznanie sekwencji telomerowej. Wreszcie użyłem metod kwantowochemicznych do oszacowania potencjału formowania kowalencyjnych adduktów białko-DNA na telomerach.

Acknowledgements

With an ineffable gratitude to my supervisor, who gave me the tools and space to grow; to the loved ones, for their unconditional support in each and every endeavor I had conceived; and to my friends of the present and past for keeping me striving to grow as a human being.

Contents

List of Abbreviations

AIMD	**A**b **I**nitio **M**olecular **D**ynamics
ALT	**Al**ternative **L**engthening of **T**elomeres
APB	**A**LT-associated **P**romyelocytic leukemia nuclear **B**odies
BAR	**B**ennett **A**cceptance **R**atio
BER	**B**ase **E**xcision **R**epair
CSVR	**C**anonical **S**ampling through **V**elocity **R**escaling
CV	**C**ollective **V**ariable
DBD	**D**NA-**B**inding **D**omain
DDR	**D**NA **D**amage **R**esponse
DFT	**D**ensity **F**unctional **T**heory
DNA	**D**eoxyribo**N**ucleic **A**cid
DSB	**D**ouble **S**trand **B**reak
dsDNA	**d**ouble-**s**tranded **D**eoxyribo**N**ucleic **A**cid
GQ	**G**-**Q**uadruplex
FapyG	2,6-diamino-4-hydroxy-5-**F**orm**a**mido**py**rimidine
FapyA	4,6-diamino-5-**F**orm**a**mido**py**rimidine
Hi-C	**Hi**gh-Throughput Screening **C**hromosome Conformation Capture
FFT	**F**ast **F**ourier **T**ransform
GGA	**G**eneralized **G**radient **A**pproximation
HF	**H**artree-**F**ock
HR	**H**omologous **R**ecombination
LDA	**L**ocal **D**ensity **A**pproximation
LDA	**L**inear **D**iscriminant **A**nalysis
MD	**M**olecular **D**ynamics
MFPT	**M**ean **F**irst **P**assage **T**ime
ML	**M**aximum **L**ikelihood
MM	**M**olecular **M**echanics
MSD	**M**ean **S**quare **D**isplacement
MSM	**M**arkov **S**tate **M**odel
NER	**N**ucleotide **E**xcision **R**epair
NHEJ	**N**on-**H**omologous **E**nd **J**oining
NMA	**N**ormal **M**ode **A**nalysis
PCA	**P**rincipal **C**omponent **A**nalysis
PCCA	**P**erron **C**luster **C**luster **A**nalysis
PES	**P**otential **E**nergy **S**urface
PML	**P**ro**M**yelocytic **L**eukemia nuclear bodies
QM	**Q**uantum **M**echanics
RdRP	**R**NA-**d**ependent **R**NA **P**olymerase
RI	**R**esolution of **I**dentity
RMSD	**R**oot-**M**ean **S**quare **D**isplacement
RNA	**R**ibo**N**ucleic **A**cid
RNS	**R**eactive **N**itrogen **S**pecies

ROS	**R**eactive **O**xygen **S**pecies
SASP	**S**enescence-**A**ssociated **S**ecretory **P**henotype
Sp	2-imino-5,5-**Sp**irodihydantoin
SSB	**S**ingle **S**trand **B**reak
ssDNA	**s**ingle-**s**tranded **D**eoxyribo**N**ucleic **A**cid
ssNA	**s**ingle-**s**tranded **D**eoxyribo**N**ucleic **A**cid
TERRA	**TE**lomeric **R**epeat-containing **R**ibonucleic **A**cid
TAD	**T**opologically **A**ssociated **D**omain
TAF	**T**elomere-**A**ssociated **F**oci
tICA	**t**ime-lagged **I**ndependent **C**omponent **A**nalysis
TI	**T**hermodynamic **I**ntegration
TIF	**T**elomere dysfunction-**I**nduced **F**oci
TPE	**T**elomere-**P**ositioning **E**ffect
TPE-OLD	**T**elomere-**P**ositioning **E**ffect **O**ver **L**arge **D**istance
TRAM	**T**ransition-based **R**eweighting **A**nalysis **M**ethod
US	**U**mbrella**S**ampling
UV	**U**ltra**V**iolet
WHAM	**W**eighted **H**istogram **A**nalysis **M**ethod
WT	**W**ild-**T**ype
8oxoA	7,8-dihydro-**8-oxoA**denine
8oxoG	7,8-dihydro-**8-oxoG**uanine

Chapter 1

Scope and Goals of the Thesis

Already for several decades the topic of telomeres has been drawing the attention of both the scientific community and the general public, and not without a reason. The fundamental link between telomere length and aging, discovered at the cellular level in the early 1960s, sparked hope of there being a simple, one-dimensional biological switch to control the progression of aging and its associated debilitating physical effects. In spite of a decades' worth of promising and ingenious research, though, medicine still hasn't come close to delivering the ultimate youth pill – even less so one that would specifically target telomeres – exemplifying the common experience among biochemists and cell biologists that translating *in vitro* findings into clinical applications is an excruciating endeavor. In the process of finding out the intricate aspects and details that shape cellular processes and give rise to the responses on the physiological level, it is easy to either get lost in the unfathomable complexity of systems shaped by billions of years of haphazard tweaks, or succumb to the temptation of following a simplistic picture that overemphasizes a selected perspective or subset of facts.

While the role of telomeres as markers or modulators of aging remains hotly debated in its finer details, telomeres themselves also emerged as important safeguards against cancer, shutting off cell proliferation after a certain division count has been reached. However, this protection is far from perfect as cases of cancer are not rare in the animal kingdom. The apprehension of the complex pathways involved in suppression of carcinogenesis on telomeres is a prerequisite if one wants to arrive at knowledge-based recommendations regarding treatment and lifestyle choices. One factor that was consistently found to interfere with proper telomere maintenance was chemical stressors – either environmental or endogeneous – found to exert a particularly profound impact in the telomeric region, mainly due to the exceptional sensitivity of telomeric DNA to chemical oxidation. Whether this feature has an adaptive role remains to be elucidated; so far, one can only speculate that telomeres might have evolved to sense the presence of mutagenic factors, causing the cell to enter the non-proliferative stage of senescence to evade progression towards cancer.

In the following thesis, I approach selected aspects of telomere biology from a molecular and computational perspective, focusing on the specific protein-DNA interactions on which the proper telomere maintenance hinges, as well as on mechanistic consequences of their disruption under oxidative conditions. Here, the objective was to quantify and predict selected effects that oxidative stress shall exert on the telomeric protein-DNA complexes. Admittedly, a considerable portion of my research focused on the fundamental and broadly applicable aspects of sequence-specific protein-DNA binding, reflecting my own deep interest in this general phenomenon that governs the orchestration of gene expression in virtually all living organisms. The study could not hence have avoided raising (and, hopefully, answering) certain fundamental questions regarding the predictive power of existing physical models in predicting the behavior of biomacromolecules on the microscale; as a consequence, one of the major objectives became to provide a detailed atomistic description of processes, events and checkpoints involved in the formation of a sequence-specific protein-DNA complex. Although the presented research is exclusively computational, I tried to extensively ground it in the existing experimental literature, connecting and contributing to the collective body of knowledge spanning the fields of telomere biology, chemistry of nucleic acids and molecular biophysics.

Apart from the results of my simulations, discussed in lengths in Chapter 4, the thesis also features a sizeable background section addressing the subjects of telomere biology and protein-DNA interactions (Chapter 2) as well as quantum/classical mechanics, computational biophysics and data science (Chapter 3). More specifically, Chapter 2 provides a broad and synthetic perspective on existing literature related to mechanisms of telomere maintenance and telomere length control, DNA damage on telomeres and its effect on telomeric integrity, and sequence search on DNA as performed by sequence-specific protein domains. In turn, Chapter 3 – besides an inventory of computational methods used and/or implemented in my doctoral work – elaborates on a personal selection of theory fundamentals that form the core of a computational biophysicist's curriculum. I hope this collection of recounts, perspectives and notes will be of use for aspiring biophysicists following a similar path in the future.

Chapter 2

Biological background

2.1 Biology of the Telomere

2.1.1 Telomeres as a Solution to the End-Replication Problem

Telomeres are nucleoprotein complexes that cap the termini of linear chromosomes in all eukaryotic and several prokaryotic species. In vertebrates, the DNA component of a telomere consists of hundreds to thousands of tandem repeats of the hexanucleotide motif 5′-TTAGGG-3′. Telomeric DNA is bound by the heteromeric shelterin complex composed of six distinct proteins – TRF1, TRF2, TIN2, RAP1, TPP1 and POT1 – that confer most functionalities of the telomere [1]. Evolutionarily, telomeres developed as a solution to two main obstacles that linear chromosomes pose. Firstly, the so-called end-replication problem stems directly from the mechanism of DNA replication by replicative polymerases: albeit extremely precise, these molecular machines can only extend a single DNA strand in the 5′→3′ direction, and require a RNA primer to initialize the process of replication. As a result, the 5′-end of the chromosome cannot be properly replicated and becomes progressively shorter with each replicative cycle; if not counteracted, this shortening will prevent further replication once telomeres reach a critical length – typically after 40-60 division cycles – a phenomenon that became known in the 1960s as the Hayflick limit [2]. Secondly, since the cell has to safeguard itself against the deleterious effects of internal DNA damage, unsupervised ends of DNA are quickly recognized by the cellular machinery and elicit a DNA damage signal that leads to further processing [3]. Therefore, chromosomal termini have to be specifically marked as “safe” to prevent interchromosomal fusions that would result in genetic instability.

The end-replication problem is solved in an ingenious way: the presence of telomerase, a nucleoprotein reverse transcriptase with an internal RNA template, allows to extend shortened telomeres without any information loss as the sequence is strictly repetitive. This process has to be strictly regulated, though, as cells immortalized due to improper control of telomerase expression may be prone to carcinogenesis. In fact, the expression of functional human telomerase relies on several stages of regulation, including transcription and splicing of the catalytic hTERT protein subunit as well as polyadenylation and maturation of the RNA counterpart, hTR [4, 5]; the mature holoenzyme can then be inhibited, activated or complemented by other protein and RNA factors [6, 7]. In recent years, new insights into the mechanisms of telomerase activation prompted researchers to consider it a promising target for anticancer therapies, and very recently high-resolution structural data shed new light on the inner workings of the enzyme in humans [8]. However, even when telomerase is lacking or inhibited, a critically short telomere can undergo extension by hijacking the homologous recombination (HR) machinery and using another chromosome's telomere as a template in what is known as alternative lengthening of telomeres (ALT) [9], calling into question the utility of targeting telomerase in cancers.

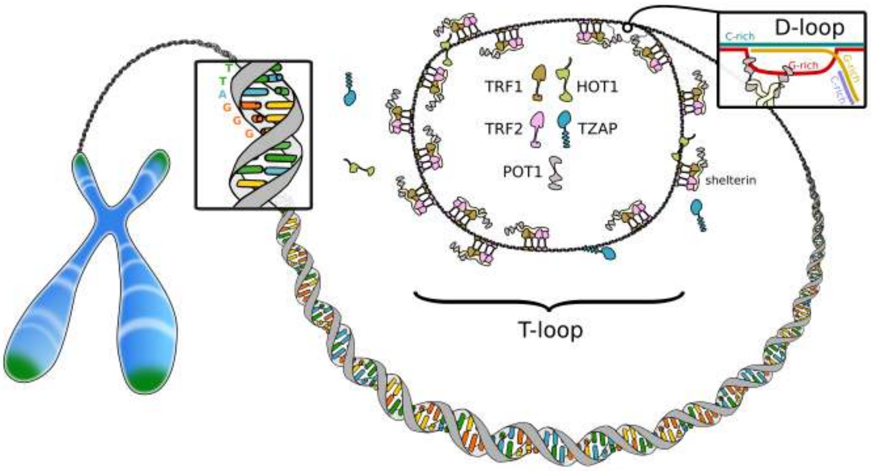


FIGURE 2.1: A simplified picture of telomere organization: the shelterin complex stabilizes the T-loop, whereas POT1 binds the displaced DNA strand of the D-loop. Two extrashelterin proteins, HOT1 and TZAP, are also known to bind to telomeric DNA, as will be discussed below.

Regarding the problem of tagging physiological DNA termini, telomeres are hidden from the DNA damage response machinery thanks to their specific structural feature, the lasso-like loop at the very end of the DNA strand. This so-called T-loop (see Fig. 2.1), whose formation is mainly mediated by the TRF2 protein [10] and involves the creation of a stable Holliday junction [11], maintains its own stability thanks to a single-stranded 3′-terminal overhang that invades an upstream double-stranded region of the telomere. This invasion creates another loop (D-loop) through the displacement of the G-rich strand in the duplex, and the unmatched displaced strand is then presumably bound and stabilized by the the single-stranded DNA (ssDNA) binding protein POT1 [12]. In particular, while the presence of TRF2 suppresses the activation of the ATM kinase signaling pathway, POT1 was found to prevent the activation of the alternative ATR pathway [13]. When telomere shortening or dysfunction leads to a disruption of the T-loop, markers of DNA damage response (DDR) become observable in the nucleus, most notably the γ-H2AX histone variant and 53BP1 foci typically associated with regions surrounding DNA double-strand breaks; on telomeres, they are called the Telomere Dysfunction-Induced Foci (TIFs) [14, 15]. Depending on the strength of the telomeric DDR signal, the cell might then become senescent or trigger apoptosis [16]. A mismanaged DDR signal can, however, lead to the activation of non-homologous end joining (NHEJ), an error-prone alternative to HR, and cause ligase IV-dependent fusion of chromosomes [17, 18]. The resulting genomic instabilities are usually lethal to the cell, but occasionally push the cell towards carcinogenesis as many cancer cells are reportedly characterized by severe chromosomal abnormalities that facilitate further mutation and adaptation to new biological niches [19].

2.1.2 Structure, Maintenance and Organization of Telomeres

The main protein complex involved in telomere maintenance, the shelterin, is a heteromeric assembly that binds both single- and double-stranded DNA (ssDNA and dsDNA). Among its six constituent proteins, TRF1 and TRF2 form homodimers that bind telomeric dsDNA and are joined by the central TIN2 protein; TRF2 also binds RAP1 independently of TIN2 [20]. TPP1 then binds to TIN2, and the ssDNA-binding POT1 protein binds to TPP1, although these two form the "facultative" part of the shelterin, as a quantitative study of relative expression of shelterin components revealed that TPP1 and POT1 are 10-fold less abundant than other telomeric proteins [21]. In fact, the same study found that telomeric populations of TRF1 are also 2–5-fold lower than those of TRF2, TIN2 and RAP1 [21], suggesting that the shelterin does not assume a uniquely defined composition but is rather characterized by an ensemble of stoichometries that depend on the specific cellular conditions.

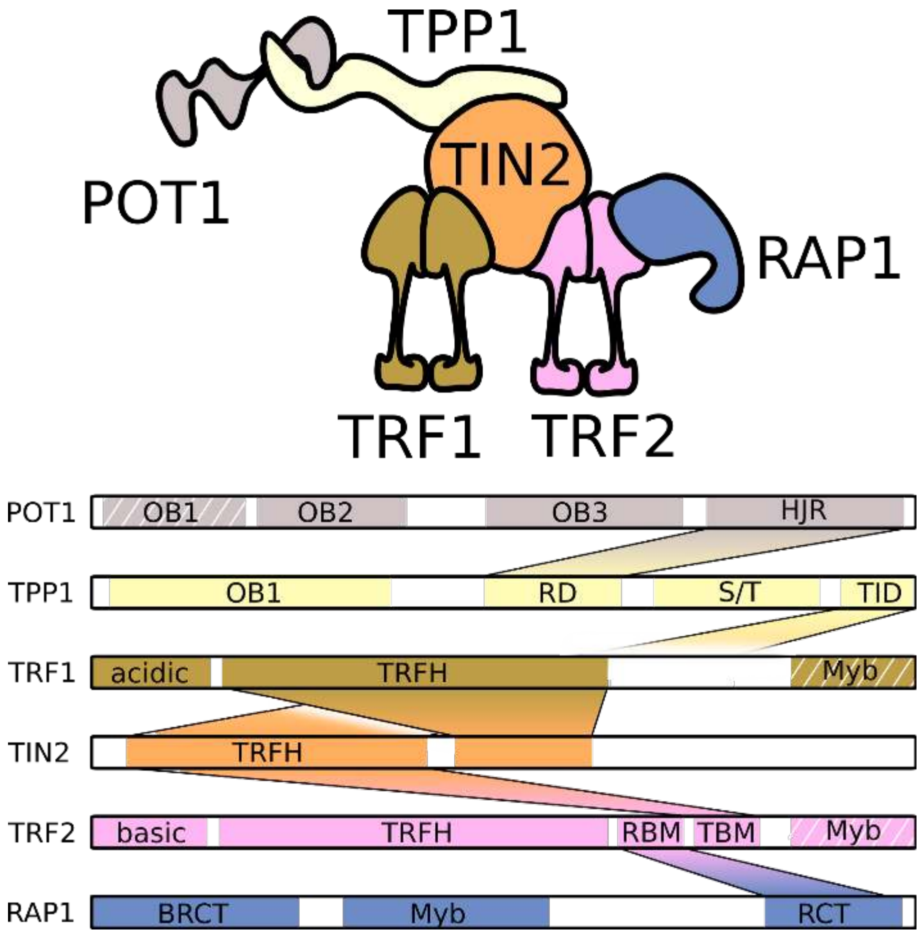


FIGURE 2.2: A schematic representation of the shelterin complex and the interactions between its components. Protein domains or interaction sites are marked with rectangles, and domain-domain interactions are indicated with color gradients. DNA-binding domains are marked with white stripes.

This fact is interesting as one realizes that shelterin components act together to regulate, e.g., the length of telomeres. Shortly following its discovery, TRF1 has been characterized as a negative modulator of telomere length, as its overexpression was observed to induce gradual telomere shortening and a mutant incapable of binding

DNA produced an opposite outcome [22]. This TRF1-induced shortening was initially proposed to result from the mechanical blockage of telomerase access mostly during DNA replication [23], as indicated by the similarities in telomere shortening due to TRF1 overexpression and RNA-dependent telomerase inhibition [24]. It was soon discovered that this connection is more convoluted, likely mediated by POT1 that localizes to telomeres through the TRF1-anchored shelterin and ultimately competes with telomerase for the access to the 3′-overhang [25]. In a separate report, the inhibition of telomerase by TRF1 was linked to the TRF1-dependent recruitment of a telomerase inhibitor PinX1 [26], and it remains unclear whether the two pathways are independent of each other. However, the fact that TRF2, TIN2 and TPP1 were all initially classified as telomerase-dependent negative telomere length regulators [24, 27–29] might indicate that length regulation is primarily achieved by timely recruitment of POT1 to telomeric termini by the shelterin, in accordance with the mechanism suggested above [30, 31].

The role of POT1 in telomere length regulation is somewhat ambiguous due to its involvement in another process related to the modulation of telomerase activity: resolution of G-quadruplex structures. G-quadruplexes (GQ) are non-canonical secondary DNA structures that can form on guanine-rich ssDNA if four runs of at least three consecutive guanines are clustered in close proximity, as is always the case on telomeric ssDNA. In telomeric intramolecular GQs, four guanine residues connected by Hoogsten bonds form a single plane (tetrad), and three tetrads stack upon each other to form a stable but polymorphic structure [32]. The presence of GQs poses a very unique challenge to telomerase due to their slow folding kinetics (with estimates of the folding time constant ranging from seconds to tens of minutes [33, 34]) and kinetic control of formation that yields random positioning patterns [35]. In response, the cell deploys factors such as the RecQ family BLM and WRN helicases and the DEAH family RTEL1 helicase that have the capacity to unwind telomeric quadruplexes by sliding over the single-stranded regions [36, 37]. Notably, deletion of telomeric helicases leaves unresolved GQs on telomeres, slowing down replication fork progression due to mechanical blockage of DNA polymerase and ultimately leading to loss on telomeres on one of the sister chromatids and genomic instabilities [37, 38]. However, this role of telomeric helicases seems to overlap with that of POT1, as POT1 has been found to disrupt GQ formation [39], in particular as a POT1-TPP1 heterodimer that dynamically slides on a ssDNA overhang [40]. This overlapping roles could explain the seemingly counterintuitive positive effect of POT1 deletion on telomere length: it is possible that POT1 that is simultaneously bound to shelterin (via TPP1) and to the ssDNA overhang (via its OB fold domain) has the ability to prevent telomere extension by telomerase *in vivo*, while at the same time in *in vitro* studies shelterin-free POT1 is actually required for telomerase activity due to the absence of alternative GQ-unfolding factors.

Dissecting the role of the 3′-overhang protection is further complicated by the presence of RPA, an abundant ssDNA-binding protein that has at least equal affinity for telomeric ssDNA as the POT1-TPP1 heterodimer, and triggers the ATM DDR signaling pathway when bound to a ssDNA target [41]. It turns out that as the telomeric loop opens during the S phase to allow for replication, RPA likely transiently associates with the ssDNA overhang, but is soon selectively displaced by the hnRNPA1 protein through a yet unknown mechanism [42]. Then, in the G2 phase hnRNPA1 is phosphorylated and released from telomeric ssDNA in a way that allows for POT1 binding, possibly due to local enrichment of POT1 resulting from its interaction with the shelterin [43]. This again highlights the essential role of shelterin integrity in modulating the POT1-dependent signaling, as POT1 devoid of shelterin-binding activity fails to properly localize to telomeres [42].

Interestingly, the described occupancy switch relies on another interesting telomere-associated biomolecule, TERRA [41]. TERRA, or TElomeric Repeat-containing Ribonucleic Acid, is a long non-coding RNA expressed from the subtelomeric region, characterized by a lack of open reading frames, a length of at least 200 bases (hence the name “long non-coding”), and the presence of multiple repeats of exact or degenerate G-rich telomeric subsequences (only the G-rich strand is transcribed onto RNA, with the C-rich strand being used as a template, although very recent data appears to prove otherwise in case telomeres are damaged [44]) [45]. Their tissue-specific transcription is dependent on the main RNA polymerase II, and their stability is modulated by polyadenylation; after transcription, TERRA remains localized to telomeres and inhibits telomerase activity, possibly through direct pairing with the telomeric template RNA [46]. The presence of TERRA also imposes a strict requirement on the RNA-vs-DNA specificity of proteins such as POT1, since the telomere-bound fraction of TERRA can outnumber the native telomeric ssDNA target by orders of magnitude. As a result, POT1 developed means to distinguish its target ssDNA from ssRNA of similar sequence with extremely high accuracy [47].

Diverse roles of the TERRA transcript have been postulated in the literature. Some noted that it is capable of forming intermolecular hybrid DNA-RNA GQs with the 3′-overhang, thereby possibly contributing to its protection during the S-phase [48]. Others point out that TERRA indirectly promotes the compaction of the telomeric chromatin [49], ensuring that telomeres remain inaccessible to the DDR factors, possibly even through the promotion of local phase separation [50]; in that way, TERRA would also autoregulate its own expression as expression levels correlate with chromatin openness, although such claims remain to some extent speculative. It seems certain, though, that at least during some fraction of the cell cycle [51] TERRA localizes to telomeres via direct formation of complementary DNA-RNA hybrids (termed R-loops) [52], and acts as an interaction hub to recruit RNA-binding factors that play further role in telomere maintenance. In this way, TERRA was found to recruit the chromatin remodellers ATRX [53] and PRC2 [54], as well as ORC in conjunction with the N-terminal domain of TRF2 [55], providing partial mechanistic explanations to the functional role of the telomeric non-coding RNA.

The TERRA transcripts originate from the subtelomeric regions, which themselves constitute an important element of the telomere maintenance machinery. Towards the centromere, the fidelity of the telomeric sequence deteriorates and eventually transitions into the non-telomeric region of the chromosome within several to about two hundred kilobase pairs, with the length of the buffering subtelomeric region varying drastically not only between individual chromosomes [56], but also between individual humans as a result of duplications of individual patches and segments [57]. Subtelomeres are transcriptionally active, and several diseases – particularly often associated with mental retardation – have been mapped to aberrations in gene regulation, gene swapping, duplications or translocations [57]; the genetic diversity of subtelomeric regions also rendered them a suitable marker to track ancestry in human populations [58].

The transcriptional activity of subtelomeres, however, seems to be at odds with the high level of chromatin compaction observed consistently in the telomeric and subtelomeric regions. It has long been known – but also long overlooked – that in spite of the high occupancy by shelterin complexes, telomeres are not devoid of nucleosomes [59]. Nucleosomes – the nucleoprotein assemblies that assist in chromatin compaction through wrapping of DNA around the octameric histone core – inherently exhibit some degree of sequence specificity, making certain positions more favorably occupied than others [60]. In case of telomeres, though, the out-of-phase (1 motif per 3/5 of a helix twist) repeats do not exert a positioning effect; on the contrary, telomeric nucleosomes were found to be highly mobile and randomly spaced [61]. Despite the apparently low stability of histones on telomeric DNA, both telomeres and subtelomeres have long been viewed as heterochromatic, i.e., packed tightly and only accessible to most protein factors in the short linker regions [62]. This classification often relied on histone post-translational modifications constituting the so-called histone code, a set of chemical marks that modulate the properties of nucleosomes in a regulated way: for instance, modifications known as H3K9me3 and H4K20me3 (trimethylation of lysine 9 and 20 on histones 3 and 4, respectively) are believed to induce the heterochromatic state as they generate sites for the binding of HP1, the heterochromatin-condensing protein [63]. Some studies, however, reported low levels of heterochromatin markers on human telomeres in selected cell lines, along with the presence of euchromatic PTMs, leading to some confusion [64, 65]. An appealing solution to this conundrum relates to a hypothesized connection between telomere length and compaction, in which long and functional telomeres remain highly compacted and critically short ones shed their heterochromatic markers to become euchromatic [66, 67].

While the detailed models of telomeric structure and organization are becoming increasingly complex, with overlapping and entangled functions of individual factors as well as numerous ties to other processes, many aspects of telomere maintenance remain elusive. One such issue is the physiological role of the shelterin component RAP1 that is solely recruited to the complex by TRF2. It was found not to contribute to telomere length regulation [68], despite controversial claims regarding the direct binding of RAP1 to telomeric DNA and the resulting stabilization of telomere-bound TRF2 [69]. Here, it is possible that both RAP1 and the basic N-terminal domain of TRF2 (i.e. opposite to the canonical DNA-binding domain) bind telomeric DNA in a structure- instead of sequence-specific manner, with moderately high affinity reported for junction regions such as the Holliday junction that forms in the vicinity of the D-loop [11, 69]. The targeting of junctions would be consistent with the reported inhibitory effect of RAP1 on homology-directed repair, a process that can lead to aberrant processing of telomeres of sister chromatids [70]. Nevertheless, most reports point to a largely insignificant role of of RAP1 in telomere maintenance, suggesting instead an evolutionarily conserved, cell type-specific and increasingly redundant role in transcriptional regulation, including transcription from subtelomeric regions [68, 71, 72].

Possibly even more uncertainty resulted from the recent discovery of two novel telomeric dsDNA-binding proteins, HOT1 (previously known as HMBOX1 or TAH1) and TZAP (formerly ZBTB48) [6, 73]. Although research on the biological properties of both proteins is scarce, in the original report HOT1 was found to positively regulate telomere length through direct interaction with both the telomeric dsDNA and the regulatory H/ACA-binding protein subunits of the telomerase holoenzyme. HOT1 was also targeted to Cajal bodies, small membraneless organelles in which telomerase components are assembled [6]. It was independently reported, however, that HOT1 also affects the alternative pathway of telomere lengthening (ALT) through promotion of formation of ALT-associated promyelocytic leukemia nuclear bodies (APBs), but without affecting the mean telomere length [74]. Overall, HOT1 appears to play a protective and signaling role as its depletion results in an increased count of telomeric DNA-damage induced foci [74], and the protein itself is clearly implicated in the regulation of apoptosis in carcinogenesis [75, 76], a link that remains poorly understood in mechanical terms.

In contrast to HOT1, the most recently discovered TZAP was consistently found to act as a negative regulator of telomere length, acting through induction of telomere trimming associated with the formation of the so-called T-circles, circular pieces of extrachromosomal DNA [73, 77]. On the other hand, TZAP overexpression fostered the formation of ALT-associated APBs in a fashion similar to HOT1. As a member of the zinc finger family, TZAP was also implicated in transcriptional regulation at extratelomeric sites; it was indeed shown that out of its 11 zinc-finger domains, only the C-terminal one directly confers specificity to the telomeric sequence [78], allowing for a degree of promiscuity at other chromosomal locations. To date, very little is known about the interactions between TZAP and other telomeric factors, apart for the fact that it does not displace TRF2 from telomeres while being displaced by overexpression of TRF2 itself [73]. In near future, both the relevance of the newly discovered telomeric proteins as well as their mechanism of action will have to be thoroughly verified in independent studies.

2.1.3 Connection to Other Cellular Functions

Although the main evolutionary purpose of telomeres is the protection of chromosome ends from both DDR factors and replicative erosion, evolution often finds alternative usages for its purposefully crafted designs, working across the clear function based distinctions that shape human understanding. As a notable example, telomeres were observed to partake in transcriptional regulation not only within the subtelomeric region (classical telomere-positioning effect or TPE), as was mentioned above, but even over as much as 1 million base pairs (TPE-OLD, TPE over large distances) [79]. While the classical TPE can result from heterochromatin propagation, TPE-OLD is believed to involve chromatin interactions within the so-called topologically associated domains (TADs): the large-scale elements of chromatin organization that can span hundreds of kilobase pairs, and are actively maintained through loop protrusion by two key proteins, CTCF and Cohesin, dictating the intrachromosomal contact patterns [80]. Indeed, a recent chromosome conformation capture/high throughput screening (Hi-C) study reported significant changes in chromatin organization resulting from telomere shortening, even prior to the appearance of DDR signaling at telomeres, suggesting that the impact of telomere length on expression levels as far as 10 megabase pairs away is widespread and direct [81]. This telomere-specific chromosomal looping is supposedly dependent on TRF2 homodimers that can transiently cross-link telomeric and internal genomic 5′-TTAGGG-3′ sequences [82] as well as interact with the nuclear lamin [83] to position telomeres properly in the context of a chromosome and the entire nucleus. Intriguingly, a very recent report suggests that this higher-order looping confers yet another layer of regulation of telomerase expression [84].

Telomerase itself has also been shown to perform certain functions outside of the telomeric region, as it is in fact only recruited to telomeres during replication [85]. Throughout the rest of the cell cycle, it resides in the nucleolus or cytoplasm, and the

TERT component of telomerase was also found to possess a functional mitochondrial targeting sequence [86]. Although individual reports provide drastically conflicting evidence regarding the mitochondrial role of telomerase, most of them link it to the signaling and management of oxidative stress in mitochondria that results from inefficient capture of reactive species emerging during oxidative phosphorylation [87, 88], eventually modulating the onset of apoptosis. Quite surprisingly, mitochondrially targeted TERT was shown to bind a non-canonical RNA component, RMRP, that turns it into a RNA-dependent RNA polymerase (RdRP). The RdRP then stimulates the production of double-stranded RNA that is further processed by the Dicer complex to become siRNA, a modulator of expression that likely interferes with yet undiscovered cellular functions [89]. Even more intriguing is the fact that TERT likely drives reverse transcription using mitochondrial tRNA as a template [90].

After synthesis, cytoplasmic telomerase is targeted to the nucleus in a NF-κB and TNFα-dependent manner [91]. In the nucleus, however, telomerase is not only required to extend telomeres, as its RNA component was found to directly interact with chromatin at multiple genomic sites [92], possibly recruiting TERT and modulating transcription of a range of genes, most notably these involved in the Wnt/β-catenin pathway responsible for embryonic development and cancer progression [93, 94]. Likewise, TRF1 and TRF2 have been shown to bind to interstitial telomeric sequences at extratelomeric sites [82], where they likely contribute to modulation of transcription as most of the bound sites were localized in genic regions. Another layer of TRF2-dependent transcription regulation can be realized through its ability to recruit RAP1, as was already mentioned above. In addition, the truncated, neuron-specific isoform of TRF2 that lacks the DNA-binding ability as well as a nuclear localization signal stabilize the cell fate of differentiated neurons through the sequestration of the major repressor REST [95], while highly expressed full-length TRF2 targets REST to the nuclear PML bodies in order to initialize this differentiation [96].

Perhaps not surprisingly, telomeres have also been implicated in the orchestration of events controlling the cell cycle, in particular mitosis and the S-phase. At both these stages of cell development sister telomere cohesion plays a key role in chromosome pairing and separation. In contrast to the centromeric region, at telomeres TRF1 and TIN2 replace the canonical tripartite ring complex in binding to the SA1 component of cohesin, with the TRF1-SA1 binding providing a mechanical support for the interaction of sister chromatids [97, 98]. This association also requires chromatin condensation dependent on the presence of the heterochromatin-stabilizing protein HP1γ, recruited to telomeres by TIN2 [99]. In mitosis, the eventual resolution of telomere cohesion is required to avoid genomic instabilities. This process relies on ATRX-dependent PARylation of TRF1 by tankyrase, the canonical poly-ADP-ribose polymerase that is able to selectively remove TRF1 (but not TRF2) from telomeres [100]. Subsequently, the cell cycle-dependent ubiquitination of tankyrase restores TRF1 at telomeres, restoring the normal telomeric state after cell division [101].

2.1.4 Telomere Length in Senescence, Therapy and Human Longevity

In the recent two decades, the subject of telomere length has gained significant recognition among the non-scientific public due to popular reports linking this simple aspect of cell biology to the overall human health and expected longevity, with companies offering commercial tests to measure an individual's "biological" or "cellular" age based on the measurement of degree of telomere shortening in somatic cells. It is now widely accepted that the aging of an individual is intimately related to the aging, or senescence, at the cellular level. The notion that critically short or malfunctional telomeres are capable of inducing cellular senescence [102], in conjunction with epidemiological studies that linked psychological and chemical stress factors to telomere shortening [103, 104], resulted in an idea that telomere length can be thought of as a "knob" to accelerate or reverse the effects of aging. This idea has since been challenged by reports claiming that telomere length remains in fact more or less constant throughout life despite large variation on an individual level [105, 106]. It is therefore worthwhile to critically assess the available evidence supporting all such claims.

According to the current scientific consensus on the subject of senescence, it is a generally beneficial process by which damaged cells signal their own disfunctional state to the immune system through secretion of a distinct set of molecular messengers that promote local inflammation and degradation of the extracellular matrix [107]. This so-called senescence-associated secretory phenotype (SASP) is initially triggered by DDR signaling [108], and has recently been shown to depend on a decreased concentration of the ubiquitous high-mobility group protein HMGB2 [109]. Through prolonged secretion of general markers of inflammation such as IL-6 or TNF-α, the senescent phenotype can spread to neighboring cells, resulting in chronic low-level inflammation throughout the affected tissue [110]. With age, the sensitivity of the immune system decreases [111], leading to less senescent cells being cleared and, eventually, a widespread propagation of the senescent phenotype throughout the body. This effect has recently been directly confirmed in rodent models by showing that senolytics, drugs that stimulate the elimination of senescent cells, extend lifespan in mice, and that direct injection of senescent cells results in severe physical dysfunction [112].

The molecular pathways by which SASP is brought about involve the formation of persistent DDR foci through activation of the ATM signaling cascade. The ATM kinase induces a large-scale reorganization of chromatin at the site of damage through phosphorylation of the histone variant H2AX, yielding the self-propagating DNA damage marker γ-H2AX [108, 113]. As discussed above, ATM signaling is also initialized by telomere deprotection, directly linking telomere dysfunction – here viewed broadly as destabilization of the t-loop by any mechanism, regardless of the actual telomere length [114] – with the onset of cellular senescence. In fact, two levels of telomere deprotection were identified that give rise to different outcomes, with more severe impairment of telomere function tipping the cell fate in the direction of apoptosis [115]. The balance between senescence and apoptosis is reportedly achieved through downstream modulation of activation of caspase-3 by p16, a p53-independent signal transducer whose activation is a hallmark of some types of telomeric DDR, in particular those not associated with telomere shortening [116, 117].

An intriguing feature of telomeric DDR signaling is the persistence of DDR foci on telomeres [118, 119], suggesting a "counting" mechanism that quantifies both the strength and the number of sources of the DDR signal to modulate downstream effectors. Details of this counting mechanism likely depend on the source of damage and type of cell, but as few as 2-3 telomere-associated foci (TAFs) were found to be sufficient to induce a senescent phenotype in murine hepato- and enterocytes [120], corresponding to 4-5 persistent TAFs in human fibroblasts [121].

Although telomere deprotection and the ensuing DDR response can be caused by a number of factors, telomere shortening is considered to be the most common of them. Besides the polymerase-dependent mechanism of replicative telomere attrition, telomeres can also shorten abruptly either due to malfunction of the length regulation machinery or as a result of double-strand breaks (DSBs), usually forming when two single-strand breaks (SSBs) coincide in close proximity. In turn, SSBs are often introduced by DNA repair enzymes at sites of damage, so that high incidence of DNA lesions can lead to DSBs through overactivation of the DNA repair machinery [122]. In line with this observation, many studies linked telomere shortening to endogenous generation of reactive oxygen species (ROS) by malfunctioning mitochondria, and showed that the rate of telomere attrition correlates with the innate antioxidant capacity of the cell [123, 124]. Other studies provided direct evidence linking telomere shortening to DDR and gradual induction of senescence [116, 125]. On the other hand, telomere shortening was found not to be necessary for the induction of a senescent phenotype in chronic obstructive pulmonary disease, where DDR in smokers' lung epithelial cells likely resulted from exposure to chemical factors [126]. Yet another study showed that longer telomeres are in fact more prone to accumulation of DSB markers, which would promote senescence were the counting mechanism operative [127].

Quite expectedly, the simple picture that directly linked environmental and metabolic factors to cellular senescence and aging through the "knob" of telomere length turns out to be much more nuanced in reality. However, its general principle has some merit: chemical stressors do affect cell fate and accelerate aging, even though the telomere length itself can remain unaffected in the process. It seems that the absence of chronic "sterile" (i.e., in absence of infectious agents) inflammation is a better predictor of longevity [128], and given the current knowledge it is reasonable to assume that telomeres at most mediate this relationship, as concluded indeed from the analysis of involvement of telomere length in the development of cardiovascular diseases [129]. In addition, a recent long-term epidemiological study found evidence that long telomeres can actually predispose individuals to certain types of cancer, further indicating that one cannot simply equate long telomeres with better health outcomes [130]. Fortunately, the development of senolytics and treatments aimed at the reduction of systemic inflammation still holds promise to extend the human healthy lifespan well beyond today's levels, and trials are underway to bring them to market in the following years.

2.2 Telomeres and Oxidation

2.2.1 Oxidized Telomeres: Active Sensors or Passive Bystanders?

As already mentioned above, it became increasingly clear in the late 1990s and early 2000s that oxidative stress has a significant effect on telomere length and telomeric integrity, and that telomeres are more affected by oxidation than other regions of the genome [131]. Simultaneously, direct *in vitro* evidence confirmed that this effect can be replicated on the level of DNA sequence alone, and that different oxidants – including UV light, single-electron oxidants as well as reactive oxygen and nitrogen species (ROS/RNS) – produce a very distinct oxidation pattern specifically in the region of guanine triplets (the 5′-GGG-3′ subsequence) [132–134]. On the even more fundamental level, this observation was justified by quantum chemical calculations demonstrating that G-triplets are characterized by the lowest ionization potential of all trinucleotides, hence exhibiting the highest susceptibility to oxidation [135].

For the above reasons, it was hypothesized that this susceptibility is actually an adaptive mechanism by which the cell can sense potentially harmful events before genic regions are affected. Indeed, telomeric damage appears to be a good early marker of general DNA damage and potential genomic instabilities, allowing for timely detection of malicious changes that could put the cell on a pathway to carcinogenesis [136]. Such a mechanism is relatively robust as telomere malfunction is signaled by multiple parallel pathways; any alternative mechanism relying on individual signaling molecules would be itself much more prone to disruption by mutations caused by oxidative conditions.

It could also be argued that were telomere oxidation an unwanted effect, cells would have evolved to avoid G-tracts in telomeric DNA. Meanwhile, the opposite is observed, as the presence of 5′-GG-3′ to 5′-GGGG-3′ subsequences in the repetitive motif appears to be a highly conserved property of telomeres even in very distantly related species [137].

Another observation in favor of the "stress sensor" hypothesis is the aberrant, or at least unconventional, behavior of DNA damage signaling and repair factors at telomeres. Although the ability to repress ligase IV-dependent NHEJ as well as HR is required to the core function of telomeres, i.e., protection against spurious activation of DDR response factors, the peculiarity goes beyond that. As noted above, DNA damage signals in form of easily detectable γ-H2AX foci persist on telomeres much longer and are larger than their non-telomeric counterparts [118], even though reportedly do not involve chromatin decompaction [138]. This means that damage is properly recognized, precluding the possibility that the telomeric machinery hides sites of damage from the DNA surveillance factors; simultaneously, little action is undertaken by the repair machinery, suggesting that cells actively suppress the resolution of telomeric DDR signals.

While γ-H2AX foci are markers of the more severe double-strand breaks, the process of DNA damage repair preferentially operates at much earlier stages, i.e., when the original base lesions are produced. Two main pathways are operative in remediating base damage, nucleotide excision repair (NER) and base excision repair (BER) [139, 140]. In NER, when the damage site is detected and bookmarked by the XPC-RAD23B complex, a short patch (ca. 30 bases, corresponding to a single binding site for RPA) of DNA is excised at the damaged strand by XPF and XPG, and the gap is subsequently filled by the non-canonical polymerases $\delta/\varepsilon/\kappa$ and sealed by ligase I/III [140]. In contrast, BER can be initiated by a range of glycosylases that selectively excise the damaged base through cleavage of the N-glycosidic bond, leaving behind an AP (abasic, apuric/apyrimidic) site that is then incised by the APE1 endonuclease or the AP lyase. The resulting gap is then filled by the non-replicative polymerase β in what is called the short-patch repair, or processed in a manner similar to NER in long-patch repair [139]. The repair pathway is chosen based on structural properties of the lesion, with bulky, helix-distorting lesions such as pyrimidine dimers, chemical adducts or covalent cross-links being selectively targeted by XPC for processing *via* NER, and small base modifications such as 8-oxo-7,8-dihydroguanine (8oxoG), thymine glycol, 3-methylpurines or uracil targeted by specific glycosylases and processed *via* BER [139, 140].

At telomeres, conflicting conclusions were drawn regarding the rate of NER-mediated repair of cyclobutane pyrimidine dimers, with one study claiming that NER is dysfunctional at telomeres [141] and another providing evidence for accelerated repair of photoinduced damage [142], a conundrum that surprisingly remains unresolved to this day. In addition, the ERCC1/XPF homodimer implicated in NER-dependent DNA cleavage was found to be physically associated to TRF2 and implicated in the 3'-overhang processing [143], suggesting that telomeric NER-dependent processing is potentially altered. However, NER appears to be operational at telomeres, as revealed by the accelerated telomere attrition in absence of the key XPC-RAD23B recognition complex [144] as well as other components of the pathway [145].

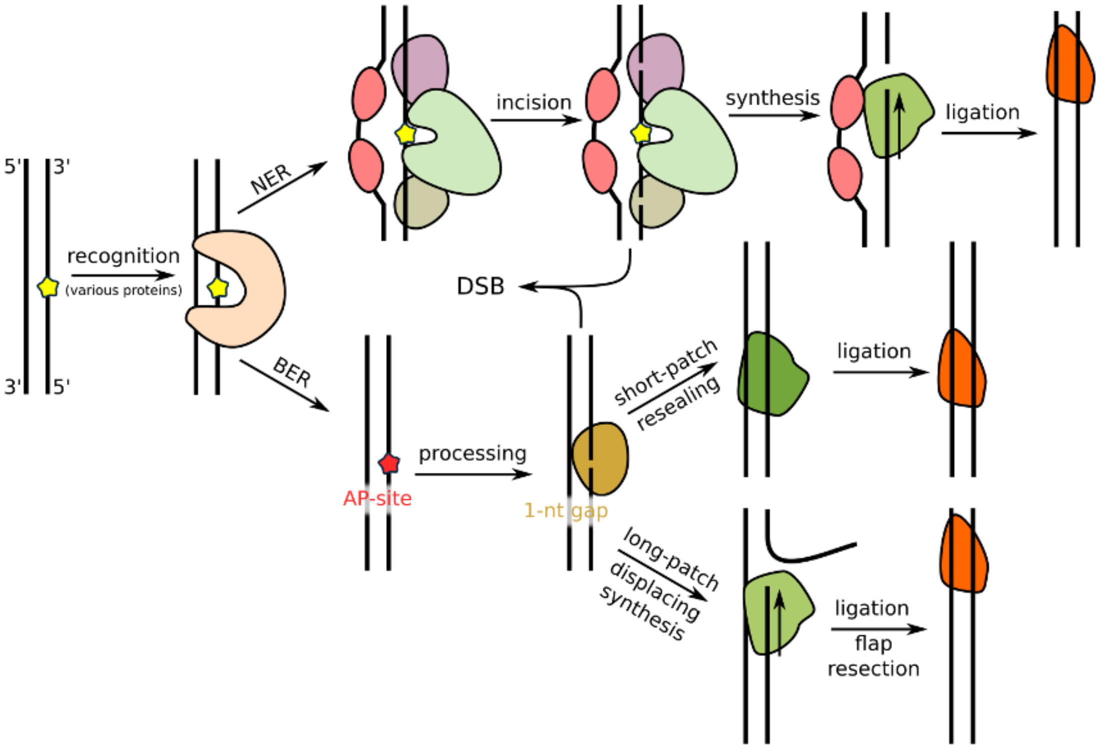


FIGURE 2.3: A simplified overview of NER and BER, the two processes that mediate base damage repair in DNA. Note that both pathways involve the formation of a single-strand break, so that two active NER/BER sites at neighboring locations on opposite strand can give rise to a double-strand break (DSB).

BER is mostly carried out at telomeres by glycosylases specialized in the excision of products of guanine oxidation, OGG1 (8-oxoguanine DNA glycosylase) and the three NEIL glycosylases that recognize formamidopyrimidines and hydantoins [146], although the pyrimidine-specific Nth1 glycosylase has also been shown to affect telomere integrity [147]. While BER was shown to be inhibited in the context of histone-bound chromatin [148], this effect was not replicated on a telomeric DNA substrate [149], possibly due to the aforementioned higher mobility of telomeric nucleosomes [61]. As in the case of NER, BER components were also found to physically associate with shelterin components, simultaneously promoting the long-patch pathway: FEP1 (the flap endonuclease) was shown to bind to TRF1, TRF2 and POT1, and DNA polymerase β remained bound to POT1 [149], possibly facilitating the detection of repair intermediates at telomeres. On the other hand, conclusions might vary between differently designed studies due to the observed cell cycle-dependent modulation of repair proficiency [150]. Overall, in spite of modifications, both NER and BER appear to be functional at telomeres, yet not sufficiently to prevent the accumulation of base lesions in the easily oxidized repeats [151]. Based on the proposed mechanism of DSB formation, one could also note that high activity of BER and NER translates to a higher incidence of DSBs, making it hard to draw conclusions regarding a purposeful sensor-like functionality of telomeres.

It remains a matter of debate whether the effects of oxidative damage induce the downstream DDR signal by (i) a direct deprotection of telomeres, (ii) increased level of DSB signaling, (iii) shortening due to higher incidence of DSBs or (iv) shortening due to misregulation of telomere length homeostasis. On the one hand, persistent DDR signaling was reported to be unrelated to the loss of TRF2, and unremediated by TRF2 overexpression [119]. On the other, base oxidative damage that ultimately triggers DDR reportedly decreases the affinity of TRF2 and TRF1 for telomeric DNA *in vitro* [152], and human cells in which the BER component APE1 is missing have lower levels of telomere-bound TRF2, alongside with an increased telomeric localization of γ-H2AX foci [153]. It appears that all options remain viable with possibly overlapping and redundant functionalities, although more research is needed to properly weight and assess the evidence.

2.2.2 Overview of Oxidative Lesions

Although DNA is considered to be among the most chemically inert molecules in the cell due to its role in information storage, in reality multiple pathways lead to the formation of oxidative lesions in nucleobases. Major causes of oxidative lesions involve ROS/RNS, type I-photosensitizing single-electron oxidants, as well as ionizing radiation energetic enough to eject electrons from molecules [154]. ROS and RNS are a broad group of reactive species including primarily hydroxyl radicals ($OH^{\cdot}$), singlet oxygen ($^1O_2$), superoxide anion radicals ($O_2^{-\cdot}$), peroxynitrites (ONOO) and nitrogen dioxide ($NO_2^{\cdot}$), carbonate radicals ($CO_3^{\cdot}$) as well as hydrogen peroxide (H_2O_2) that can be converted to other oxidants through photoexcitation or Fenton reactions with transition metal ions [155]. Such species can be generated as a result of malfunctional scavenging of oxidation intermediates in the mitochondrial respiratory chain or *in situ*, with the latter being almost impossible to scavenge even in the presence of antioxidants [155]. Type I photosensitizers such as riboflavin or acridine orange [156] promote nucleobase oxidation through direct sequestration of a photoexcited electron, producing a reactive cation radical already within the DNA structure. In contrast, type II photosensitizers such as methylene blue yield a stable excited triplet state that can undergo triplet oxygen quenching in which the ground-state triplet oxygen, $^3O_2$, deexcites the sensitizer to produce the reactive singlet oxygen [156, 157]. Finally, ionizing rays such as γ photons or high energy particles not only dissipate energy along their path through the solvent, leaving behind tracks of radical species (indirect effect), but can also excite an electron out of a molecule (direct effect) [158].

In DNA, singlet oxygen preferentially reacts with guanine bases to yield the [4+2] cycloadduct in which the oxygen molecule bridges the C8 and C4 atoms of guanine. The endoperoxide ring then opens to yield 8-hydroperoxyguanine that is reduced to 8-hydroxyguanine, a tautomer of 8oxoG, or follows a cycle of de- and rehydration that results in the formation of spiroiminodihydantoin (Sp) [159]. While due to spatial restraints the cycloaddition reaction is significantly more favorable in ssDNA and in free nucleotides, it still occurs in dsDNA with good yields; for the same reason, though, the distribution of products is biased in favor of the planar 8oxoG rather than the bulky Sp [155].

In case of radical oxidants, and in particular the hydroxyl radical, evidence exists that the yield of 8-oxopurine products is considerably higher in case of guanine than adenine [160, 161]. However, as the direct addition of the hydroxyl radical to the C8 atom was shown to be barrierless and hence dictated by the diffusion rate, the proportion of product yields should be similar [162, 163]. This conundrum has been elegantly explained through neighbor effects, in which, e.g., pyrimidine peroxide radicals formed at adjacent sites in the presence of O_2 can either react with the purine or abstract an electron given an advantageous sequence context [161]. In the study, isotope labeling provided evidence that only a small fraction of 8-oxopurines is formed by means of direct addition of the radical moiety, and that the indirect mechanism produces clustered lesions that are less effectively repaired by the respective glycosylases. Besides the most thoroughly studied C8 position, hydroxyl radicals were also found to attach barrierlessly to C4 and C5 carbons, albeit often with lower affinity [163, 164], likely reflecting the low physical accessibility of the site in dsDNA. These adducts are believed to leave behind a guanyl or adenyl radical that can react with oxygen, giving rise to secondary products such as imidazolone and oxazolone derivatives with a central 5-membered ring [163].

In pyrimidines, hydroxyl radicals most often add to the double bond between C5 and C6 (predominantly at the C5 position due to higher electron density), producing a new radical center on the other carbon atom. This intermediate tends to react with water to yield thymine/cytosine glycol, or add oxygen to produce a reactive peroxyl radical. A relatively minor pathway corresponds to the hydrogen abstraction from the methyl group of thymine, yielding 5-hydroxymethyluracil and 5-formyluracil [163].

An interesting feature of radical reactions is that the consecutive intermediates are still reactive radical moieties, so that more than one position can be affected in a single oxidation event. Since DNA is known to mediate charge transfer over several bases, it is not surprising that the secondary reactions can involve electron abstraction from even remote neighbors [161, 165]; however, the complexity involved in the modelling of sequence effects in charge hopping has long remained prohibitive and only recently these effects are being addressed in a more systematic manner [166].

In contrast to radical reactions, single-electron oxidation usually occurs at sites characterized by low ionization potentials, such as runs of consecutive guanine residues. Guanine alone quenches the excited triplet state of riboflavin more than 3-fold faster than adenine, and two orders of magnitude faster than pyrimidine bases [167]. The resulting guanidyl cation radical deprotonates rapidly by transferring its N1 proton to the Watson-Crick-paired cytosine [168], and thus stabilized radical will often add the immediately available nucleophile – typically a water molecule – to yield an intermediate identical to that involved in direct addition of a hydroxyl radical to the C8 carbon atom. Depending on the redox properties in the immediate environment, this intermediate can then undergo oxidation to form 8oxoG, or reduction associated with the opening of the 5-membered ring to form formamidopyrimidine (FapyG); analogously, 8oxoA and FapyA will be produced if adenine is the original reactant [169]. Under specific conditions (γ-irradiation of neoplasmic monocytes), FapyG and FapyA were detected in an HPLC-MS/MS assay at ca. 2-fold higher concentrations than their 8-oxopurine counterparts, showing that 8-oxopurines do not always constitute the predominant group of lesions [160].

Covalent protein-DNA cross-links constitute yet another severely understudied class of oxidative lesions. It was found that in presence of 8oxoG, an easily oxidized lesion, further photosensitized oxidation yields cross-links with a unique lysine residue in the complex of MutY with dsDNA [170]. Efficient production of cross-links was also observed in a minimal model, using KKK oligopeptides and 5'-TGT-3' oligonucleotides [171]. Despite recent progress [172], little is known about the detailed mechanism of lysine-guanine cross-link formation, as well as the prerequisites that render a given protein-DNA complex susceptible to oxidative cross-linking. An NMR and MS/MS study showed, though, that the cross-linking involves either the C8 or C5 atoms of guanine, yielding a product similar to Sp [173].

Throughout the years, 8oxoG received much more attention than other lesions for two major reasons: (a) the multitude of pathways (singlet oxygen cycloaddidion, single electron oxidation, radical addition) that selectively yield this product [155], some of them highly selective with respect to guanine, and (b) the documented mutagenic potential of 8oxoG that assumes a non-canonical *syn-* conformation in dsDNA to pair with adenine, thus yielding G→T transversions [174]. In recent years, however, evidence emerged for high mutagenicity of 8oxoA [175], contradicting previous statements [176]. Also, selected aspects of past reports merit additional discussion. In recent years, concerns were raised that many DNA isolation as well as damage quantification protocols were flawed, either generating new lesions in the act or miscategorizing others [155], and that results revised using the gold standard ESI-MS/MS approach often bring more conservative conclusions. Similarly, certain studies conveniently use free nucleotides, nucleosides or even nucleobases, reporting damage at sites that remain inaccessible in dsDNA, while the geometrical constraints of dsDNA likely impair the formation of bulky products of multiple oxidation events, both due to steric factors and decreased solvent accessibility. Finally, the absolute number of lesions reported also needs be put in perspective, with lesions called "abundant" if they occur with a probability of ca. 10^{-6} per 1 Gy of ionizing radiation [155], even if the probabilities can increase by another order of magnitude at telomeres [177]. Therefore, any mechanisms involving DNA damage has to consider the relative rarity of its occurrence in an actual cellular setting – a reminder that after all, DNA evolved to be resistant to chemical damage.

Remarkably, recent years saw a considerable shift in how oxidative damage is viewed by cell biologists, with novel interest in treating base lesions as epigenetic rather than solely destructive factors. While this idea is not new in the case of telomeric signaling, the concept that DNA oxidation can modulate gene expression in a transient and non-genic yet coordinated way finds more and more support [178, 179]. It remains to be determined, though, whether the mechanisms proposed – modulation of GQ and DDR factors occupancy, effect on direct protein binding – constitute functional standalone regulation pathways, or are rather a random or redundant effects with little impact on the cell's response to external stimuli.

2.3 General Aspects of Protein-DNA Interactions

2.3.1 Sequence specificity in Protein-DNA Interactions

The central dogma of molecular biology, as defined by Crick in 1957, represents a simple and deterministic model in which the information stored in DNA is first transcribed onto RNA using rules of base complementarity, and then – based on the codon readout code – is translated into a protein sequence in ribosomes [180]. This scheme, however, can only be directly applied to the most simple biological entities such as certain viruses, as the plentiful information stored in genomes of virtually all organisms is too complex to be expressed in a simultaneous and uncoordinated manner. The organization and coordination of readout and compaction of DNA in living cells is therefore mostly determined by protein-DNA interactions. While histones and histone-like proteins – indispensable in the global maintenance of chromatin structure – mostly bind DNA independent of sequence (or with a weak sequence positioning effect [181]), the dynamic orchestration of cell's response to a wide range of stimuli depends on sequence-specific DNA binders, i.e. proteins that locate and bind DNA at a defined nucleotide sequence.

This sequence specificity is definitely not a binary property: proteins are characterized by a whole spectrum of specificity, from barely detectable effects linked to local DNA deformability to a distinct thermodynamic preference of several kcal/mol. On top of that there also exists a degree of degeneracy in the target sequence, making the precise mapping of interactions in the genome extremely difficult to predict without experimental input. This protein-DNA interaction code that depends on sequence, elastic properties and availability of the DNA strand can be therefore thought of as a convoluted layer of modulation overlaid on the simple rules of gene expression that allowed evolution to fine-tune the dynamic coupling between processes at the cellular level, at the same hampering human understanding of this interplay.

Typically, the readout of nucleotide sequence by sequence-specific proteins is classified as either direct or indirect [182]. Direct readout relies on the ability of individual amino acids to form unique patterns of contacts with the minor- or major-groove-exposed surface of DNA bases, typically based on the hydrogen bonding properties, water-bridged interactions and hydrophobic contacts (in case of methyl groups of thymine or olefin group of cytosine). Other properties used to decode sequence are collectively referred to as indirect (or "shape readout"), including sequence-dependent groove widths, the local harmonic ellasticity of dsDNA, the propensity to form kinks or to allow for intercalation [183]. Some other proteins bind non-canonical DNA structures (Z-DNA, GQ [184, 185]) whose formation is also strongly sequence-dependent (e.g. Z-DNA requires repetitive 5'-CG-3' runs to form). Clearly, the relative contribution of direct and indirect readout varies considerably between individual proteins.

Somewhat more nuanced is the question of sequence recognition in single-stranded nucleic acids (ssNA). ssDNA is a relatively rare species that mostly functions as an intermediate in dynamic processes such as transcription or replication, and for this reason is mostly bound by non-specific proteins, e.g. RPA [186]. Indeed, the telomeric POT1 is one of the few sequence-specific binders in humans as it targets telomeric DNA [187]. At the same time, many sequence-specific proteins bind to ssRNA, sharing similar strategies for sequence determination [188], and many ssNA binders require robust means of differentiating between RNA and DNA, with POT1 again serving as a perfect example [47]. As ssNA do not exhibit Watson-Crick base pairing, the repertoire of interactions used for binding and sequence determination can be augmented by π-π stacking between nucleobases and side chains of aromatic amino acids and arginine, as well as by extended hydrogen bonding to heteroatoms normally engaged in Watson-Crick pairing [188].

Although the structural aspects of sequence specificity is being debated since the first resolved structures of protein-DNA complexes [189], the abundance of case-specific effects and observations might make the accumulated knowledge appear scattered and inconclusive. However, certain properties are well-documented and come up repeatedly in structural studies. For instance, the shape complementarity between an α-helix and the DNA major groove was suggested to mediate direct binding already in 1959 [190], and this structural property is found to be key for sequence recognition in most instances. Most DNA-binding domains (DBDs) are also modular, with a small number of folds universally adapted for sequence recognition and tethered to other functional domains with flexible linkers that modulate selectivity and binding thermodynamics [191]. Regarding more specific details, the common role of arginine minor groove insertions in the recognition of AT pairs has been broadly described [183], as well as the allosteric effect of protein occupancy at neighboring positions along the DNA strand [192]; universal trends in sequence-dependent DNA elasticity were also characterized e.g. to justify the trends observed in nucleosome positioning [60] and transcription factor binding [193].

The fact that cells use a limited repertoire of structural templates to achieve sequence-specific binding is neatly exemplified at telomeres, with the DBDs of three of the five known DNA-binding proteins – TRF1, TRF2 and HOT1 – assuming the three-helical-bundle homeodomain fold, commonly found in homeobox proteins associated with embryonic morphogenesis [194]. The TZAP protein binds to DNA using an array of zinc-finger domains, representing another extremely popular group of folds stabilized by a central zinc ion tetracoordinated by histidine and cysteine side chains [194]. Finally, the two DBDs of POT1 assume the popular OB-fold, deriving its name from its capability to bind extended oligosaccharide structures but equally frequently used to bind ssNA [195].

2.3.2 Dynamics of Protein Diffusion and Target Search on DNA

In the nucleus of a human cell, a single sequence-specific DNA-binding molecule is faced with a seemingly complex task: search the several billion bases contained in chromosomes in order to locate one of its target sequences. Even though on average, the consensus sequences of eukaryotic transcription factors are roughly only half as short as in prokaryotes (12.5 bp in *Drosophila* vs 24.5 bp in *E.coli* [196]) and hence more abundant in the genome, the problem of sequence search dynamics is far from trivial. To account for this complexity, sequence search on DNA has long been described in terms of so-called facilitated diffusion, i.e. a combination of sliding in a 1-dimensional fashion and hopping or "regular" 3-dimensional diffusion that reportedly allows to significantly shorten the amount of time required to scan the DNA sequence, sometimes referred to as the antenna effect [197, 198]. However, there seems to be little consensus regarding the actual gains in terms of shortened search time, with some models showing that the mixture of 1D sliding and 3D hopping actually slows down the search compared to pure 3D diffusion [199], and researchers noting that excessive (i.e. longer than 50 bp) 1D random walks on DNA would constitute a highly inefficient search method due to its redundancy, as individual positions would be visited multiple times before detachment to a previously unseen region [200]. On the other hand, simulations based on a multi-state kinetic model indicate that sliding lengths of ca. 10 bp yield optimal search times [201].

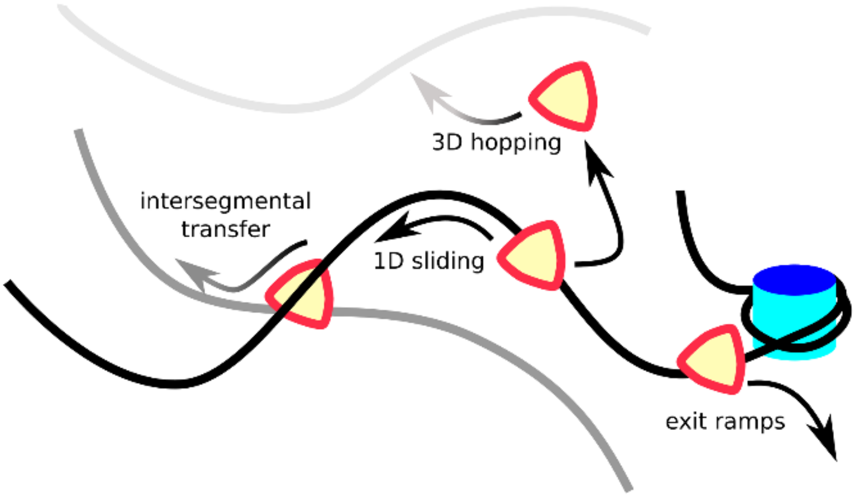


FIGURE 2.4: Possible modes of diffusion employed by DNA-binding proteins (sliding and hopping), and events that affect diffusion (intersegmental transfer, encountering exit ramps). The relative probabilities of individual components can be fine-tuned in the course of evolution to optimize search times in complex genomes.

Regardless of whether proteins actually overcome the diffusion limit, they certainly optimize the search time by finding a balance between tight binding at the target and loose binding in the non-specific mode: on the one hand, it would be wasteful to detach from the eventually located target site too easily, while on the other, tight association at an off-target site would drastically slow down diffusion overall [202]. Another interesting question is that of maintaining a defined orientation during the search. Again, two problems need to be solved simultaneously here: the protein needs to scan both orientations on the DNA strand since the target sequence can be positioned in either orientation due to the pseudo-symmetry of DNA (assuming the target is not palindromic), but only a single face of the protein surface exposes residues that partake in the recognition and should be oriented towards the DNA grooves. It was determined experimentally that proteins indeed couple 1D sliding with rotation about the DNA axis, indicating that they indeed slide on a helical pathway along the DNA grooves in a defined orientation with respect to the DNA [203].

Importantly, if this "processive" sliding length is smaller than the length of a single period of 1D scanning, the protein will be able to switch orientations and scan the sequences in both directions before detachment, so that it will not overlook the target as would happen every other time in case of fully processive rotation-coupled scanning.

In the non-specific scanning mode, a jump between neighboring positions along the sequence was found to be associated with small free energy barriers, averaging 0.66 kcal/mol and varying little between the several proteins that were studied [203]. This low roughness of the free energy profile facilitates rapid search along short stretches of DNA, but it was postulated that occasionally the protein will encounter exit ramps – sequences with a considerably less favorable affinity that terminate the 1D search [204]. As the process has not yet been studied thoroughly in the context of actual protein-DNA systems, it is hard to state whether such an effect is evolutionarily optimized, and if so, whether it is the proteins or the DNA sequence that underwent optimization. Arguably, a similar role can also be played by mechanical obstacles on DNA such as nucleosomes [205]. Yet another example of tuning the proportion between sliding and jumps is the occurrence of intersegmental transfers through the "monkey bar" mechanism [206], recently confirmed to be operational e.g. in the DDR protein PARP1 [207]. Here, a DNA-binding protein that is composed of at least two basic domains or functions as an oligomer can simultaneously contact two DNA segments, effectively performing a 3D hop while minimizing the unproductive time spent as a free-floating species. This behavior can also potentially explain the fact that most DBDs are connected by long unstructured linkers, in particular given that the number and distribution of charges in the linker region can modulate the extent to which the monkey bar mechanism is functional [206].

One could ask whether there are any long-range interactions by which proteins find a direction in which to search for targets. Interestingly, the fine-tuning of neighboring DNA sequences to speed up target search has been observed in what is called the funnel effect. Specifically, it was found that the neighborhood of particular targets is enriched in sequences that yield higher-than-average binding affinities for the protein in question [208, 209]. While it was claimed that such an effect would only provide a marginal speedup since it would only affect the final stage of the process [197], in reality not every near-target encounter results in the formation of a sequence-specific complex [210], and hence more time spent in the vicinity of the target increases the probability of successful binding, rationalizing the evolutionary development of such a mechanism.

On telomeres, the diffusion along the DNA constitutes a unique case since each tandem repeat represents both a target and a free energy barrier to diffusion. For this reason, telomeric proteins do jump between adjacent binding sites on telomeres, but also do so in a much slower manner than on non-telomeric DNA substrate, combining features of the search mode and the sequence-specific bound mode. A recent single-molecule study showed that the roughness of the free energy profile along the helical path is ca. 1.9 kcal/mol higher on the telomeric than non-telomeric sequence [211]. In the specific case of the telomere, such a slowdown was hypothesized to be advantageous as it allowed for the formation of TRF1 and TRF2 homodimers, and ultimately also the assembly of TIN2-stabilized shelterin. This hypothesis, though, overlooks the presence of histones that would significantly restrict the ability of individual units to encounter other components of the complex [212]. Shelterin was also found to assemble *in vitro* independently of the presence of DNA [212], although this does not preclude the competing mechanism of *in situ* assembly, particularly given that telomeric TRF1 is selectively depleted in a cell cycle-dependent manner without affecting the occupancy of TRF2 [23] and has to be reintroduced on telomeres later on.

The last issue that merits additional discussion is that of anomalous diffusion in protein-DNA complexes. If the search process can be modelled as a regular random walk, i.e. when the trajectory evolves due to random displacements sampled from a fixed and position-independent distribution, the mean square displacement (MSD) is directly proportional to time of the walk. However, a study of TRF1 diffusion along non-telomeric DNA found that the dynamics of TRF1 was subdiffusive on λ-DNA, with MSD proportional to $t^{0.72}$ [211]. In this *in vitro* study, the anomalous diffusion was attributed to pausing e.g. at randomly positioned high-affinity sequences. However, in a more general setting other factors exist that can render the search's dynamic subdiffusive, including macromolecular crowding [213] that create fractal-like structures full of dead-end paths [214]. One should also note that the transitions between 1D sliding and 3D hopping drastically affect the local diffusion coefficient [215], possibly contributing to the anomalous diffusion similarly to the effect of kinetic traps described above.

Chapter 3

Computational Methods

3.1 Quantum Mechanics: the Foundation

3.1.1 Wavefunction: Approximations and Connection to Classical Mechanics

In the study of structural, energetic and dynamic molecular processes, quantum mechanics provides the most complete and convenient theoretical framework to accurately describe phenomena on the atomistic scale. Pioneered by the work of Planck and Einstein in the first years of the XX century, and then developed by Heisenberg, Bohr, Schrödinger, Pauli and others in the 1920s, quantum theory was soon applied to molecular multi-body systems composed of multiple nuclei and electrons. Although all but the simplest cases require that some level of approximation be introduced in the mathematical formalism, the resulting emergence and proliferation of quantum chemical models were vital for the success of modern simulational and computational chemistry we benefit from today. In the following section, I will provide a conceptual background for the subsequent discussion of realistic simulations and intuitive descriptions of atomistic systems.

The Born-Oppenheimer Approximation

The starting point for further considerations and approximations is the time-dependent Schrödinger equation, which defines the molecular wavefunction – a central quantity of the quantum theory that encapsulates all stationary and dynamic information about the system of interest. In chemical physics, this equation is often written in the following form:

$$i\hbar\frac{\partial}{\partial t}\Psi(\mathbf{r},\mathbf{R},t) = \hat{\mathcal{H}}\Psi(\mathbf{r},\mathbf{R},t) \tag{3.1}$$

with $\mathbf{r}$ used to denote electronic, and $\mathbf{R}$ nuclear Cartesian coordinates. The key element of this equation, the Hamilton operator (Hamiltonian), corresponds to its classical counterpart – the operator of total energy – and is constructed by analogy to include both the kinetic and potential component:

$$\begin{aligned}
\hat{\mathcal{H}} \equiv & -\sum_{I} \frac{\hbar^2}{2M_I}\nabla_I^2 - \sum_{i} \frac{\hbar^2}{2m_i}\nabla_i^2 \\
& + \frac{1}{4\pi\varepsilon_0}\left(\sum_{i>j} \frac{e^2}{|r_i - r_j|} - \sum_{i,I} \frac{eZ_I}{|r_i - r_I|} + \sum_{I>J} \frac{Z_I Z_J}{|R_I - R_J|}\right) \\
= & -\sum_{I} \frac{\hbar^2}{2M_I}\nabla_I^2 + \hat{\mathcal{H}}_e = \hat{\mathcal{H}}_n + \hat{\mathcal{H}}_e
\end{aligned} \tag{3.2}$$

Here, the kinetic energy operator (included independently for nuclei and electrons, in respective order) is written as $-\frac{\hbar}{2m}\nabla^2$ in analogy to the classical expression $\frac{p^2}{2m}$, with the quantum momentum operator $\hat{p} = -i\hbar\nabla$ used instead of the classical momentum p. The remaining three terms are identical to the classical electrostatic (Coulomb) energy of interaction between electrons, electron-nucleus pairs and nuclei, respectively, with e and Z corresponding to electronic and nuclear charges, and m and M to electronic and nuclear masses. It is also useful to define the electronic Hamiltonian $\hat{\mathcal{H}}_e$ that corresponds to the so-called clamped-nuclei part, i.e. all terms that do not vanish in a fictitious system where the positions of the nuclei are fixed and known to arbitrary precision. Indeed, by writing the stationary eigenvalue problem for the electronic Hamiltonian and including the nuclear coordinates dependence parametrically:

$$\hat{\mathcal{H}}_e \psi(\mathbf{r};\mathbf{R}) = E_e(\mathbf{R})\psi(\mathbf{r};\mathbf{R}) \tag{3.3}$$

we can then formally expand the general wavefunction $\Psi(\mathbf{r},\mathbf{R};t)$ in an infinite series using eigenfunctions of the electronic Hamiltonian, $\psi(\mathbf{r};\mathbf{R})$:

$$\Psi(\mathbf{r},\mathbf{R};t) = \sum_{i} \psi_i(\mathbf{r};\mathbf{R})\chi_i(\mathbf{R};t) \tag{3.4}$$

so that the last term can be thought of as an infinite set of time-dependent expansion coefficients. It is also convenient to choose a set of eigenfunctions $\psi_i(\mathbf{r};\mathbf{R})$ that are mutually orthonormal, i.e. satisfy the condition $\langle\psi_i|\psi_j\rangle = \delta_{ij}$.

When modeling the dynamics of a multiatomic system, one is typically concerned with the evolution of nuclear coordinates and hence it is reasonable to integrate out fast electronic degrees of freedom. When eqn. 3.4 is inserted into eqn. 3.1 and integrated from the left-hand side with an arbitrarily selected eigenfunction $\langle\psi_k|$, one gets an infinite set of coupled differential equations (one for each choice of k) [216]:

$$\begin{aligned}
i\hbar\frac{\partial}{\partial t}\sum_{i}\langle\psi_k|\psi_i\rangle\chi_i &= \sum_{i}\langle\psi_k|(\hat{\mathcal{H}}_e + \hat{\mathcal{H}}_n)|\psi_i\rangle\chi_i \\
i\hbar\frac{\partial}{\partial t}\chi_k &= E_{ek}\chi_k + \sum_{i}\langle\psi_k|\hat{\mathcal{H}}_n|\psi_i\rangle\chi_i \\
i\hbar\frac{\partial}{\partial t}\chi_k &= E_{ek}\chi_k + \hat{\mathcal{H}}_n\chi_k + \sum_{i} C_{ik}\chi_i
\end{aligned} \tag{3.5}$$

where the orthonormality assumption was used to simplify the result. The last transformation follows from the properties of the Laplacian operator acting on a product: $\nabla^2(f)g = \nabla^2(fg) + 2\nabla(f)\nabla(g) + f\nabla^2(g)$. Accordingly, the matrix element C_{ik} is defined as:

$$C_{ik} \equiv \langle \psi_k | \hat{\mathcal{H}}_n | \psi_i \rangle + \sum_I \frac{1}{M_I} \langle \psi_k | \hat{p}_I | \psi_i \rangle \, \hat{p}_I \tag{3.6}$$

with $\hat{p}_I$ being the nuclear momentum operator (see above). the off-diagonal terms ($i \neq k$) couple individual equations, precluding the introduction of an independent set of electronic states. Setting these off-diagonal terms to zero constitutes the so-called adiabatic approximation in which electronic states are separable and have well-defined energies E_{ek}, but coupling between electronic and nuclear degrees of freedom still exists due to the diagonal entries C_{kk}. This means that electronic states are not uniquely specified by positions of the nuclei, but also by their momenta. Finally, by discarding the diagonal entries C_{kk} we arrive at the famous Born-Oppenheimer approximation in which the nuclear time-dependent Schrödinger equation can be written in terms of separate electronic and nuclear kinetic energy terms:

$$i\hbar \frac{\partial}{\partial t} \chi_k = E_{ek} \chi_k + \hat{\mathcal{H}}_n \chi_k \tag{3.7}$$

where now the electronic energy acts as a potential guiding nuclear motion. It is noteworthy that this approximation provides a surprisingly robust description in most everyday applications of quantum chemistry, such as thermochemical calculations or optimization of molecular geometries. One has to point out, however, that it can break down badly in cases that violate key assumptions, including energetic proximity of adjacent electronic states (so-called conical intersections, changes in wavefunction symmetry) or fast-moving nuclei (e.g. bombardment with high-energy α-particles). When this is the case, methods that avoid the Born-Oppenheimer approximation exist that rely on Feynman path-integral formulations [217], grid-based [218] or orbital descriptions of the nuclei [219].

Temporal evolution. The Hellmann-Feynman Theorem

While the above result provides us with a relatively well-behaved (although still second-order differential) equation, it is rarely preferable to employ a quantum-physical, wavefunction-based description of nuclei. Indeed, it is the most convenient, and often satisfactory, to treat the nucleus as a classical charged point particle. If one now introduces an arbitrary complex function $\chi_k = A_k(\mathbf{R}, t) \exp(iS_k(\mathbf{R}, t)/\hbar)$, substitutes it into eqn. 3.7, and divides both sides by χ_k, the real part of the resulting equation will become [216]:

$$-\frac{\partial S_k}{\partial t} = E_{ek} + \sum_I \left(\frac{1}{2M_I} (\nabla_I S_k)^2 - \frac{\hbar^2}{2A_k M_I} \nabla_I^2 A_k \right) \tag{3.8}$$

If one goes to the classical limit in which energy spectra become continuous ($\hbar \to 0$), the sole $\hbar^2$ term will disappear, and the resulting equation:

$$-\frac{\partial S_k}{\partial t} = E_{ek}(\mathbf{R}) + \sum_I \frac{1}{2M_I}(\nabla_I S_k(\mathbf{R}, t))^2 \tag{3.9}$$

exactly mirrors the classical mechanical formulation of the Hamilton-Jacobi equation:

$$-\frac{\partial S}{\partial t} = \mathcal{H}\left(\mathbf{R}, \nabla S, t\right) \tag{3.10}$$

in which the nuclei move in a potential defined by the electronic term E_{ek}, with momentum defined as ∇S_k. (Note the exact correspondence between S_k, the phase factor of χ_k, and the classical action S – the core idea behind Feynman's concept of path integrals.)

Although the above derivation is notationally consistent, it is worth noting that one could similarly re-derive classical mechanics by an *ad-hoc* application of the Hellmann-Feynman theorem (derived briefly below) to a semi-classical system of point nuclei guided by instantaneously adjusting electron densities. The dynamics of such a system would be governed by classical forces acting on the nuclei, $\mathbf{F} = -\nabla_I E$. The corresponding quantum mechanical observable would then be the (negative) nuclear-coordinate gradient of the expected value of the potential energy operator:

$$\begin{aligned}\mathbf{F} = -\nabla_I E_p &= -\nabla_I \langle\psi|\hat{\mathcal{H}}_p|\psi\rangle \\ &= -\langle\nabla_I\psi|\hat{\mathcal{H}}_p|\psi\rangle - \langle\psi|\nabla_I\hat{\mathcal{H}}|\psi\rangle - \langle\psi|\hat{\mathcal{H}}_p|\nabla_I\psi\rangle \\ &= -E_p\langle\nabla_I\psi|\psi\rangle - \langle\psi|\nabla_I\hat{\mathcal{H}}|\psi\rangle - E_p\langle\psi|\nabla_I\psi\rangle \\ &= -E_p\nabla_I\langle\psi|\psi\rangle - \langle\psi|\nabla_I\hat{\mathcal{H}}|\psi\rangle \\ &= -\langle\psi|\nabla_I\hat{\mathcal{H}}|\psi\rangle\end{aligned} \tag{3.11}$$

It is now evident that once the Hamiltonian can be analytically differentiated with respect to nuclear positions, correct semi-classical forces can be derived trivially from a quantum-mechanical description of the system. This approach indeed remains at the core of most software packages and interfaces that allow to perform the so-called *ab initio* Molecular Dynamics (AIMD) in which the real dynamics of an atomistic system is approximated through iterative solution of the stationary Schrödinger equation (or its electron density-based equivalents).

3.1.2 Density Functional Theory

All above considerations rely on the concept of a wavefunction, the cornerstone of quantum mechanics. Throughout the years, wavefunction theory has proven extremely successful in accurately predicting thermochemical, structural and spectroscopic data for a wide range of small- to medium-sized molecules: according to the Copenhagen interpretation, the square of the wavefunction's modulus corresponds to measurable probabilities, while expected values calculated using wavefunctions predict the corresponding expected values of experimentally measured quantities. An early and spectacular result was the highly precise calculation of atomization energy of the hydrogen molecule by Kołos and Wolniewicz in the 1960s: while the computational prediction did not agree with the best experimental estimate of that time, the discrepancy was later shown to result from experimental errors, proving that the accuracy of quantum chemical calculations can indeed exceed that of experimental measurements.

In large multi-body systems such as molecular assemblies, though, manipulating wavefunctions quickly becomes tedious due to their inherently high-dimensional nature – a wavefunction is a function of the positions of all constituent particles, that is both nuclei and electrons – and convoluted functional forms introduced to satisfy the fundamental requirement of antisymmetry. Indeed, the first well-behaved and sufficiently general mathematical model of the N-electronic wavefunction was the Slater determinant, a product of multiple one-electron spatial functions (spinorbitals) antisymmetrized to yield all possible permutations of electronic variables with respect to an ordered list of spinorbitals [220]:

$$\Psi_{sl} = \frac{1}{\sqrt{N!}} \sum_i (-1)^{p_i} \hat{\mathcal{P}}_i \phi_1(\overrightarrow{r_1}) \phi_2(\overrightarrow{r_2}) ... \phi_N(\overrightarrow{r_N}) \tag{3.12}$$

where $\hat{\mathcal{P}}_i$ is one of the $N!$ conceivable permutation operators that swap electronic coordinates, and p_i is the number of "primitive" swaps as would be performed sequentially by the operator.

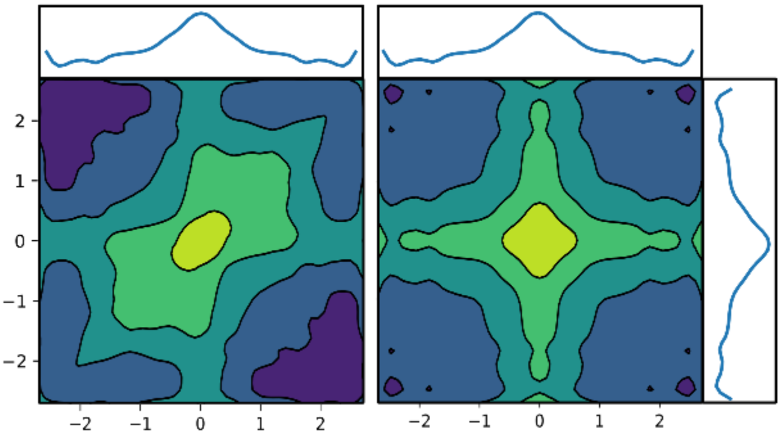


FIGURE 3.1: A comparison between a correlated 2D distribution (top panel) and an uncorrelated one (bottom panel), produced as a product of the marginal distributions. Although the marginal distributions are identical in both cases, all information about correlation between the two variables is lost in the bottom plot.

Such a function (or rather function template, as the functional form of spinorbitals is not yet specified here) has all necessary properties of an all-fermion wavefunction – changes sign when two particles are swapped, vanishes when two particles occupy the same quantum state, and is properly normalized (given the normalization of spinorbitals). It can be shown that any reasonable (i.e., physically meaningful) wavefunction is representable as an infinite sum of Slater determinants, a property employed in high-precision quantum chemical methods such as full configuration interaction (full CI).

Such convoluted mathematical objects, however, produce an enormous amount of complexity when applied in practice, with high-precision multi-determinant methods characterized by a formal scaling of $\mathcal{O}(N^{10})$ or even $\mathcal{O}(N!)$ with respect to the number of spinorbitals used. They also lose any intuitive interpretability, hampering the development of simpler approaches. Finally, a single Slater determinant cannot provide a correct description of dynamic electron correlation: it attempts to model an arbitrarily complex joint distribution through a product of marginal distributions, a problem illustrated by Fig. 3.1. Even though in practice the correlational contribution to electronic energy only accounts for ca. 1% of the total energy, it can easily dominate the calculations of (usually much smaller) energy differences. As a result, the neglect of correlation has a catastrophic effect on the accuracy of single-determinant methods in application to chemical reactions: although Hartree-Fock results are routinely included in computational benchmarks, they serve more as a "whipping boy" than an actual reference accuracy threshold.

The Universal Functional

For the above reasons, physicists had long been searching for simpler workarounds. Soon after the Schrödinger equation was first introduced, Thomas and Fermi independently proposed a model electronic description of multi-electron systems based on *electron density* alone, using the exact properties of homogeneous electron gas as a physical principle. While conceptually interesting, the model remained a curiosity rather than a tool due to its failure to reproduce features as fundamental as chemical bonding. In fact, it was not until 1964 that a seminal paper by Hohenberg and Kohn again spurred interest in electron density-based descriptions [221]. In the article, two brilliantly simple, half-page long proofs were presented: that (i) there exists an exact one-to-one mapping between the ground state of a non-degenerate electronic wavefunction of a molecule and its electron density in $\mathbb{R}^3$, so that there exists a functional that maps electron density exactly to electronic energy (as well as other molecular properties); and that (ii) such a functional follows a variational principle, i.e., it attains the lowest possible value for the true ground state density, so that the density can be optimized – for instance using a self-consistent iterative procedure.

The first property is shown by *reductio ad absurdum*. If one constructs an electronic Hamiltonian composed of electron kinetic, electron-electron repulsive and electron-nuclear attractive potential energy operators, then the first two terms only depend on the number of electrons N, and it is sufficient to specify N and the nuclear (external) potential $v(r)$ in which the electrons move to recover the stationary electronic Schrödinger equation and, consequently, the true wavefunction. Now, assuming there exist two external potentials $v(r)$ and $v'(r)$ with ground states Ψ and Ψ' that produce the same electron density $n(r)$, the original variational principle of quantum mechanics states that the two relationships need be simultaneously satisfied:

$$
\begin{aligned}
E' &= \langle \Psi' | \hat{\mathcal{H}}' | \Psi' \rangle < \langle \Psi | \hat{\mathcal{H}}' | \Psi \rangle = \langle \Psi | (\hat{\mathcal{H}} + (\hat{\mathcal{V}}' - \hat{\mathcal{V}})) | \Psi \rangle \\
&= E + \langle \Psi | \hat{\mathcal{V}}' - \hat{\mathcal{V}} | \Psi \rangle \\
E &= \langle \Psi | \hat{\mathcal{H}} | \Psi \rangle < \langle \Psi' | \hat{\mathcal{H}} | \Psi' \rangle = \langle \Psi' | (\hat{\mathcal{H}}' - (\hat{\mathcal{V}}' - \hat{\mathcal{V}})) | \Psi' \rangle \\
&= E' - \langle \Psi' | \hat{\mathcal{V}}' - \hat{\mathcal{V}} | \Psi' \rangle
\end{aligned}
\tag{3.13}
$$

The main assumption of the proof states that both $\langle\Psi'|\hat{\mathcal{V}}' - \hat{\mathcal{V}}|\Psi'\rangle$ and $\langle\Psi|\hat{\mathcal{V}}' - \hat{\mathcal{V}}|\Psi\rangle$ can be written as $\int v(r)n(r)dr$, so that the above conditions can be rewritten as:

$$\begin{aligned} E' - E &< \int v(r)n(r)dr \\ E' - E &> \int v(r)n(r)dr \end{aligned} \tag{3.14}$$

which is clearly a contradiction. It should be noted in passing that for this result to hold, $v(r)$ and $v'(r)$ have to differ by more than an additive constant.

The one-to-one correspondence between non-degenerate ground state densities and ground state wavefunctions implies that there exists a universal (system-agnostic) functional $F[n]$ that, when evaluated on a density $n(r)$, yields the electron-electron interaction and kinetic energy, so that the total electronic energy can be evaluated from the true density as:

$$E[n(r)] = F[n(r)] + \int v(r)n(r)dr \tag{3.15}$$

Defining $n(r)$ and Ψ as the true N-particle density and wavefunction corresponding to $v(r)$, and $n'(r)$ and Ψ' as the true density and wavefunction corresponding to some other $v'(r)$ (i.e., *v-representable* properties), the wavefunction variational principle requires that:

$$E[n(r)] = F[n(r)] + \int v(r)n(r)dr < E[n'(r)] = F[n'(r)] + \int v(r)n'(r)dr \tag{3.16}$$

which – by the virtue of being true for any $v'(x)$ differing from $v(x)$ by more than an additive constant – proves the density-based variational principle. The above reasoning has two important consequences: firstly, if one is able to deduce the form of the universal functional, it is then sufficient to vary $n(r)$ so as to minimize the resultant energy in order to obtain the true electron density. Secondly, although such a functional exists, there is no deterministic strategy to determine its functional form, and no proof that it can be written down analytically. Unfortunately, this also means that the variational principle might not be valid for approximate functionals.

The Kohn-Sham Scheme and Alternatives

One path to construct an approximate functional that was eventually followed by Kohn and Sham in their 1965 article is to look at individual contributions to the total energy and attempt to rewrite them as density functionals, and finally lump all unknown terms into a separate category that will be parametrized later on [222]. From this principle, one can write the total energy as

$$E = \int v(r)n(r)dr + \frac{1}{2}\int\int \frac{n(r)n(r')}{|r-r'|}drdr' + T[n(r)] + E_{xc}[n(r)] \tag{3.17}$$

The evaluation of the first two terms is straightforward provided that appropriate numerical procedures or analytic expressions are available. The electronic kinetic energy term is considerably more troublesome: although an exact kinetic energy operator, $-\frac{\hbar}{2m}\nabla^2$, exists in the wavefunction theory, it cannot be translated to a corresponding functional acting on the electron density. In some early approaches, a kinetic energy functional based on the electron gas model was used that implied that the local kinetic energy density be proportional to $n^{\frac{5}{3}}(r)$; however, this simplistic model was quickly abandoned due to its very low accuracy.

The approach of Kohn and Sham circumvented this problem in a smart way. By reintroducing single-electron molecular orbitals into the density-based framework, they were able to reuse the exact kinetic energy operator from wavefunction theory. Such orbitals would have a different mathematical interpretation – instead of providing building blocks for the multi-electron wavefunction, here they are just components of the total electron density, the latter defined as a sum of squares of moduli of all occupied orbitals. There was, however, one major caveat: when single-electron orbitals are used, the equation corresponds to a system with no electron correlation, i.e. individual electrons moving in the mean field produced by all other electrons. This required that the final term in the equation – the so-called exchange-correlation functional E_{xc} – contain all information required to reproduce the density of a fully interacting system. As a result, in the Kohn-Sham scheme one models a system of non-interacting (uncorrelated) electrons that behave as if they were correlated. As before, this is formally correct, but does not immediately show how to construct the E_{xc} functional; with an improved treatment of kinetic energy, however, the quest for the universal functional became significantly easier.

In relation to other approaches, and as reflected in its name, E_{xc} can be thought of as composed of two separate parts: (1) the purely quantum-mechanical exchange energy as defined within the Hartree-Fock scheme, pertinent to the indistinguishability of particles and antisymmetry of the wavefunction, and (2) the correlation energy, formally defined as the difference between the exact solution of the Schrödinger equation and the exact expectation value of the energy of the optimal Hartree-Fock wavefunction, correcting for both the kinetic and electron-electron repulsive terms in the uncorrelated scheme.

Although the core procedure initially proposed by Kohn and Sham was essentially identical to that of solving the Hartree-Fock-Roothan equations, this was by no means strictly required by the scheme. In density functional theory, the matrix-based formalism of linear expansion of single-electron determinants constructed from contracted Gaussian-type primitive basis functions was convenient because it closely resembled existing implementations, and hence was easy to include in new software releases. Its formal scaling of $\mathcal{O}(N^4)$ due to the standard calculation of electron-electron terms in the orbital-based scheme was obviously a computational bottleneck that could have been avoided, given that (a) the kinetic energy term scales as $\mathcal{O}(N^2)$, (b) the nuclear attraction and E_{xc} terms only require single integration over the three spatial coordinates, and (c) the Coulombic term requires double integration over space.

Now, different numerical methods can be used for integration. The most naive approach would be to map the density on a grid and integrate point by point, for which the Coulombic term would have a scaling of $\mathcal{O}(N^6)$ with respect to the single-dimension grid size; this is clearly unfeasible. A much more reasonable idea is to introduce an auxiliary basis that facilitates certain calculations, and project the density back and forth between the original and auxiliary basis sets whenever one or the other is required. A common example of this approach is the so-called resolution-of-identity (RI) that employs the well-known algebraic identity $\mathbb{1} = \sum_i |x_i\rangle \langle x_i|$, proven easily by noting that:

$$\phi = \sum_i |\chi_i\rangle \langle \chi_i|\phi\rangle \tag{3.18}$$

given that ϕ lives in the space spanned by orthonormal χs; in fact, $\langle \chi_i|\phi\rangle$ are simply coordinates of ϕ in the basis defined by χs. The standard calculation of the Coulombic and exchange terms requires the computation and storage of N^4 terms $\langle ij|kl\rangle$ for every i, j, k and l, while in the RI approach only N^2M terms (with M being the size of the auxiliary basis) in the form of $\langle ij|\nu\rangle$ need be computed and stored for each i, j and ν, and any term can be later reconstructed as [223]:

$$\langle ij|kl\rangle \approx \sum_{\nu\mu} \langle ij|\nu\rangle \langle \mu|kl\rangle = \sum_{\nu\mu} C_{ij}^{\nu} \langle \nu|\mu\rangle C_{kl}^{\mu} \tag{3.19}$$

where the approximation is introduced if (as is often the case) the auxiliary basis set does not span a subspace spanned by the original basis set. As a result, the calculation of N^4 terms can be reduced to N^2M and M^2 matrix element evaluations. If this seems unimpressive, one has to bear in mind that even for modestly sized molecular systems the number of basis functions can easily exceed several hundreds.

Researchers in solid state simulations routinely choose a yet another path to faster AIMD, utilizing the so-called plane wave auxiliary basis sets. Here the speedup can be achieved by leveraging the algorithmic robustness of Fourier transform-based numerical schemes, with the $\mathcal{O}(N\log(N))$ scaling of the FFT (Fast Fourier Transform) algorithm often called "linear scaling" since $N\log(N) < N^{1+\varepsilon}$ if N is sufficiently large for any positive ε. The Fourier transform enables an efficient calculation of the Ewald sum, in which one treats an infinite periodic system with a given charge distribution as a convolution of a single image of the system with the infinite lattice function. In the reciprocal space, convolutions transform into products, facilitating the solution of the Poisson equation.

Even more importantly, a plane wave-based calculation of the electrostatic energy only involves a single summation over all plane waves forming the basis set, and the size of the basis set is controlled with a single cut-off parameter that defines the maximum frequency in the Fourier expansion, yielding – in principle – linear dependence of execution time on the system size. In the DFT framework, this cut-off corresponds to the maximal local variation in the density, so that with an inappropriately chosen cut-off the projected density will be lacking in the most variable regions, ones that are typically associated with the highest energy density. Fortunately, these regions also usually coincide spatially with the core electron shells that are usually considered to be of little relevance from a chemical point of view. It is therefore customary to combine the use of plane-wave auxiliary basis sets with that of pseudopotentials [224].

The role of pseudopotentials is to represent the potential produced by the ionic core (i.e., the nucleus and core electronic shells) as sensed by the valence electrons in a smoother and easier to process way. In this way, the auxiliary basis set only has to describe the slowly varying valence electron density and a smaller basis set can be used. After a subset of atom's electrons is defined as the core, the requirements for the functional form of a "good", transferable and norm-conserving pseudopotentials is to (a) produce identical eigenvalues as a fully atomistic description; (b) yield identical radial wavefunction profiles beyond a chosen cut-off radius R_{co}; (c) contain identical amounts of charge and pseudo-charge (charge density that would produce the pseudopotential) within a sphere of radius R_{co} centered at the nucleus; (d) both logarithmic and regular first derivatives of the pseudo- and actual wave function need to agree at $r = R_{co}$ [225].

Recent and future developments

Although the search for more and more accurate DFT functionals continues, many further developments overturn the computational advantage provided by locality of the functional for the sake of higher accuracy. So far, all density functionals only relied on the local values of density, $n(r)$ (local density approximation, LDA), or, in more advanced cases, also on the density gradient $\nabla n(r)$ (generalized gradient approximation, GGA). However, in the early 90s first so-called hybrid functionals were introduced that contained an admixture of exact Hartree-Fock exchange in the E_{xc} term, as prompted by the concept of adiabatic connection. By considering a continuous transition between the Kohn-Sham system, i.e. a system of uncorrelated electrons reproducing the true density ($\lambda = 0$) and a fully interacting real system ($\lambda = 1$), it was shown that combining DFT exchange with HF exchange in some unspecified proportion can lead to more realistic results [226]. Indeed, one of the first functionals designed with this principle in mind – B3LYP – remains a paragon of compromise between accuracy and simplicity to this day: compare its three empirically determined parameters with as much as 40 in the much more modern MN12L functional [227, 228]. There was, however, a price to pay for this gain: the calculation of exact HF exchange requires the full $\mathcal{O}(N^3)$ calculation of orbital-based terms even when RI is used. This has relatively little impact on routine geometry optimizations and singe-point energy calculations, but becomes prohibitive for AIMD simulations, where the calculation of the density needs be repeated tens of thousands of times. For this reason, simulations that employ the plane-wave auxiliary basis still routinely employ local (mostly GGA) functionals, with the assumption that the gains from dynamic sampling of molecular ensembles counterweight any losses due to lower accuracy of the functionals. For similar reasons, the field enjoyed little benefit from more recent developments such as Coulomb-attenuated or double-hybrid functionals.

To finish this section, I will enumerate several challenges and breakthroughs that are expected to bring significant gains for the quantum chemical community in the following years.

1. When considering hardware and software developments, the transition from CPU- to GPU-based computation appears to be the most significant drive for change, as reflected in enormous speedups reported for computational suites such as Terachem (routinely interfaced with the Amber suite to perform mixed quantum and classical, i.e. QM/MM simulations) [229].
2. A more conceptual prospect is that of multireference DFT, a recently developing scheme that would allow for a more consistent treatment of notoriously

difficult cases such as e.g. the excited states of polyenes, transition metals and transition states [230].

3. The recent and ongoing development of the DLPNO scheme in high-precision coupled-cluster wavefunction-based calculations brought down the cost of chemically accurate calculations by orders of magnitude, yielding linear scaling of computation time with system size in systems as large as protein molecules and holding promise for further improvement [231].
4. The inclusion of artificial intelligence (AI) into quantum chemical computational workflows has the potential to significantly transform the field, as AI models allow almost unparalleled flexibility in selection of nonlinear functional forms. As quantum chemistry largely relies on curated and standardized data sets, training data are plentiful, facilitating the application of e.g. deep learning models. To little surprise, such approaches have already been proposed [232] while other developments in the field are underway.

3.2 Classical Mechanics: the Framework

3.2.1 Principles of Dynamics in Multi-Body Systems

Despite the inherently quantum mechanical nature of phenomena on the atomistic scale, it is often most convenient to represent molecular systems as interacting point particles in which one is solely concerned with the position of nuclei, effectively integrating out electronic degrees of freedom. In the previous section, I reviewed formally consistent means to justify such a description, explicitly enumerating all approximations introduced along the way. These considerations set the stage for the concept of a potential energy surface (PES), a simple yet useful idea that aims to picture the energetic landscape in which the molecular system evolves in time: the potential energy in such a system can be defined only as a function of the positions of nuclei, $E = E(\mathbf{r})$ (note the conventional – and convenient – switch from $\mathbf{R}$ to $\mathbf{r}$ when denoting nuclear coordinates; throughout this chapter, the dot notation will also be used to denote time derivatives, so that $\dot{\mathbf{r}}$ is equivalent to $\frac{d\mathbf{r}}{dt}$).

Were one to visualize the PES for a N-atomic system, it would have to be plotted in a $3N + 1$-dimensional space (or $3N - 5$ if the system's energy is fully translationally and rotationally invariant). For this reason, usually only simplistic systems of e.g. two particles moving in 1D are used in didactic examples, often misshaping our comprehension of basic concepts in higher-dimensional problems. More powerful frameworks are needed to properly address this issue, and the following sections should provide grounds for such developments.

Newtonian Mechanics

Historically, the treatment of systems of point particles was first formalized by Newton, and to this day his formalism remains in many regards perplexingly useful and accurate. Taught as part of high-school curricula, Newtonian mechanics relies on a set of second-order differential equations valid for each particle i:

$$\ddot{\mathbf{r}}_i = \frac{\mathbf{F}_i}{m_i} \tag{3.20}$$

and if all forces acting on the particles are conservative (which is not a requirement here), one can replace the force acting on particle i, $\mathbf{F}_i$, with the negative gradient of the potential energy, $-\nabla_i E$. As solution to the Newton's equation of motion, we obtain trajectories – curves in the $3N$-dimensional space corresponding to the positions of all particles, also called the configuration space, parametrized by time t. The derivative of the parametric curve with respect to the parameter itself yields the set of all particles' velocities, or momenta if one multiplies them by the respective particles' masses. The resulting $6N$-dimensional space that jointly describes instantaneous positions and momenta in the N-particle system, the phase space, is another constantly recurring concept in both classical mechanics and statistical thermodynamics that sets the ground for numerous further developments.

It has to be noted that Newton's equations are only uncoupled if they describe non-interacting particles in an external field, and become coupled if $\mathbf{F}_i$ is explicitly or implicitly dependent on the positions of other particles, as is typically the case in modelling of physical phenomena. It is widely known that already for three point particles interacting *via* a gravity-like potential, i.e. one with an inverse dependence on the radial distance, there is no analytic solution to the resulting set of equations (the so-called three-body problem). It is therefore customary in any real-world applications to integrate Newton's equations of motion by means of numerical algorithms, an effort that has been – in practice – contingent on the accessibility of computing power and efficient algorithms. A concise description of the latter will be provided later in this chapter.

Lagrangian Formulation of Classical Mechanics

The main drawback encountered when applying the Newtonian formulation of mechanics to molecular systems – and in fact to many real-life mechanical systems mathematicians modelled throughout the ages – is that Newtonian mechanics is best suited to work with Cartesian coordinates, and often the problem at hand requires that more natural coordinate systems be employed to reduce the mathematical complexity of the equations. Moreover, specific physical systems frequently involve mechanical constraints, and in Newtonian mechanics the forces required to satisfy these constraints need be explicitly included in the equations of motions. Therefore for classical mechanics to become a standard engineering tool, a more unified and general framework was needed in which equations of motion could be generated – preferentially from a single central quantity – in a standardized and reliable manner.

Such a central quantity, called the Lagrangian, was introduced in the late XVIII century (almost exactly a century after Newton published his *Principia Mathematica*) by Joseph-Louis Lagrange. The definition of the Lagrangian is simple yet counter-intuitive: it is the difference between the kinetic and potential energy, $T - V$, as opposed to the seemingly more meaningful quantity, the total energy $T + V$. If one now assumes that all forces operating in the system are conservative or non-dissipative, i.e. preserve the total energy, the kinetic term becomes solely dependent on particle velocities, and potential energy on particle positions, so that differentiation of the Lagrangian yields forces and momenta:

$$\begin{aligned}
\mathcal{L} &= T(\dot{\mathbf{r}}) - V(\mathbf{r}) \\
\frac{\partial \mathcal{L}}{\partial r_i} &= -\frac{\partial V}{\partial r_i} = F_i \\
\frac{\partial \mathcal{L}}{\partial \dot{r}_i} &= \frac{\partial T}{\partial \dot{r}} = p_i
\end{aligned} \tag{3.21}$$

By noting that $F_i = \dot{p}_i$, one arrives at the Euler-Lagrange equations:

$$\frac{\partial \mathcal{L}}{\partial r_i} - \frac{d}{dt}\frac{\partial \mathcal{L}}{\partial \dot{r}_i} = 0 \tag{3.22}$$

that allow to recover the classical Newtonian equations of motion, but are also valid in any generalized coordinate system in which r and $\dot{r}$ are substituted by q and $\dot{q}$. It is however worth pointing out that, while the functional form of $V(\mathbf{q})$ might be much simpler and more explicit than $V(\mathbf{r})$, in general cases the term $T(\dot{\mathbf{q}})$ has to be calculated using the chain rule [233]:

$$\begin{aligned}
T &= \sum_i \frac{1}{2} m_i \dot{\mathbf{r}}_i \cdot \dot{\mathbf{r}}_i \\
\dot{\mathbf{r}}_i &= \sum_\alpha \frac{\partial \mathbf{r}_i}{\partial q_\alpha} \dot{q}_\alpha \\
T &= \sum_\alpha \sum_\beta \left[\sum_i \frac{1}{2} m_i \frac{\partial \mathbf{r}_i}{\partial q_\alpha} \cdot \frac{\partial \mathbf{r}_i}{\partial q_\beta} \right] \dot{q}_\alpha \dot{q}_\beta
\end{aligned} \tag{3.23}$$

where the term in the square bracket is an element of the position-dependent mass metric matrix, $G_{\alpha\beta}$.

In many cases one wants to constrain certain degrees of freedom during the system's evolution in time. In situations typically encountered in atomistic simulations, the purpose of this might be to e.g. eliminate the fastest-changing degrees of freedom in order to extend the time step, or to calculate the mean force acting along a selected coordinate, as will be discussed in subsequent sections. Most often, such constraints can be classified as *holonomic*: if a constraint is generally defined through an equation $f(\dot{\mathbf{r}}, \mathbf{r}, t) = 0$, holonomic constraints are solely dependent on particle positions and time. If this is the case, any number of mutually compatible constraints $f_j(\mathbf{r}, t)$ can be straightforwardly incorporated into the Lagrangian equations by the inclusion of so-called Lagrangian multipliers (one multiplier λ_j per constraint):

$$\frac{\partial \mathcal{L}}{\partial r_i} - \frac{d}{dt}\frac{\partial \mathcal{L}}{\partial \dot{r}_i} + \sum_j \lambda_j \frac{\partial f_j}{\partial r_i} = 0 \tag{3.24}$$

The additional term has a simple intuitive interpretation. I have shown above that the first term, $\frac{\partial \mathcal{L}}{\partial r_i}$, corresponds to the position-dependent force, and the second term ensures that the instantaneous change in momentum indeed matches this force. Hence the third term can be viewed as the exact force that needs to be added so that the constraint is satisfied, i.e. that system remains on the hypersurface cut out from the configuration space by the constraining condition. Indeed, in the configuration space the vector ∇f_j is by construction orthogonal to the $f_j = \text{const}$ hypersurface, so that the smallest force required to satisfy the constraint has to be parallel to ∇f_j. Note also that if f_j has the dimension of distance, λ_j can actually be thought of as the reaction force of the constraint.

Although the Lagrangian formulation might appear as a technicality to the application-oriented biophysicists, it actually found its use in many aspects of molecular simulations. One major advantage of the Lagrangian mechanics relates to the use of the extended Lagrangian scheme. In this approach, virtual degrees of freedom can be added to the system to represent or constrain certain dynamic quantities, and equations of motions can be easily obtained that couple them to the motion of real particles. As an instructive example, the Andersen barostat relies on the extended Lagrangian to maintain pressure in a simulated fluid composed of N identical particles [234]. It is introduced by denoting the box volume as Q and scaling the original variables by the box size so that $\rho_i = Q^{-\frac{1}{3}} r_i$; then, Q is treated as a dynamic variable coupled to the motion of particles, so that the new Lagrangian can be written as $\mathcal{L}(\boldsymbol{\rho}, \dot{\boldsymbol{\rho}}, Q, \dot{Q})$. By analogies, the new variable can be viewed (with minor inconsistencies) as representing the state of a virtual "isotropic" piston, with Q being the volume enclosed by the piston and $\frac{1}{2} M \dot{Q}^2$ as the kinetic energy of the piston. One can then proceed to derive the resulting equations of motion in which constant average pressure is maintained without the use of "hard" external walls, with correct function averages over the simulated trajectories except for a small error inversely proportional to the number of particles. This idea was then developed and generalized by Nosé and Hoover to correct for the minor inconsistencies of Andersen's approach.

Another notable application of the extended Lagrangian scheme is that of extended adaptive biased force, where the additional degree of freedom is coupled to a selected generalized coordinate by a rigid spring so that the two essentially move together, and the mean force acting on that coordinate can be calculated on-the-fly from the forces acting on the fictitious particle [235]. Finally, this idea also underlies the most successful implementations of polarizable force fields, i.e. models of molecular energetics that are capable of explicitly including the mutual polarization of neighboring atoms; here, the Lagrangian is augmented with positions and momenta of the so-called Drude oscillators, auxiliary point charge particles attached to the positions of nuclei with harmonic strings, that allow to avoid the costly self-consistent calculation that would be required otherwise [236].

Hamiltonian and the Hamiltonian Principle

As noted above, the Lagrangian and Newtonian formulations of mechanics are in fact two formal ways to view the same central idea, and given a specific problem neither has any advantage other than convenience of notation. Nevertheless, the Lagrangian formalism emphasizes certain properties and symmetries of the system, allowing for more profound insights and generalizations than would be obtained from the Newtonian description. The same is true for the arguably most developed framework in which classical mechanics had been formulated, the Hamiltonian mechanics: it is perfectly compatible with its predecessors, but makes it easier to make generalizations about the system of interest. Notably, its highly formalized mathematical structure was to a large extent carried over to quantum mechanics when the latter was being developed in the 1920s.

To properly introduce the Hamiltonian mechanism, it is first necessary to define the concept of a Legendre transform. In single-variable calculus, each point $(x_0, f(x_0))$ along the plot of a smooth differentiable function can be assigned a tangent line, specified by the slope $a(x_0) = f'(x_0)$ and the y-intercept $b(x_0)$; if the function is convex (or concave), the mapping between x and a should be invertible, i.e., both functions $a = g(x)$ and $x = g^{-1}(a)$ should exist. Using this fact, we can rewrite the expression for $f(x_0)$ in terms of the slope and intercept so that it yields the intercept in terms of the slope and the mappings f and g:

$$\begin{aligned} f(x_0) &= f'(x_0)x_0 + b(x_0) \\ b(x_0) &= f(x_0) - f'(x_0)x_0 \\ b(g^{-1}(a_0)) &= f(g^{-1}(a_0)) - (a_0)g^{-1}(a_0) \\ \tilde{f}(a) &= b(a) = f(a) - ag^{-1}(a) \end{aligned} \tag{3.25}$$

In higher dimensions, this is easily generalized by substituting the 1D slope a with partial derivatives $(a_1, a_2, ...) = (\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, ...)$ and the last term $ag^{-1}(a)$ with a sum of terms $\sum_i a_i g^{-1}(a_1, a_2, ...)$. It is also possible to Legendre-transform a multivariate function only in a subset of dimensions.

The idea is now to use the Legendre transform to obtain a function that contains the same information as the classical Lagrangian, albeit with a change of coordinates. Due to the generally unspecified functional form of the potential energy, the Lagrangian is only known to be convex in the velocity subspace; it was also shown in eqn. 3.21 that the slope of the Lagrangian with respect to velocity of particle i is the momentum of this particle. Hence performing the actual transformation should yield a function of positions and momenta:

$$\begin{aligned} \tilde{\mathcal{L}}(\mathbf{r}, \mathbf{p}) &= \mathcal{L}(\mathbf{r}, \dot{\mathbf{r}}(\mathbf{p})) - \sum_i \mathbf{p}_i \dot{\mathbf{r}}_i \\ &= \sum_i \frac{m_i}{2} \frac{\mathbf{p}_i}{m_i} \cdot \frac{\mathbf{p}_i}{m_i} - V(\mathbf{r}) - \sum_i \mathbf{p}_i \cdot \frac{\mathbf{p}_i}{m_i} \\ &= -V(\mathbf{r}) - \sum_i \frac{\mathbf{p}_i^2}{2m_i} \\ &= -V(\mathbf{r}) - T(\mathbf{p}) \end{aligned} \tag{3.26}$$

which happens to correspond to the negative of the total energy. The negative partial Legendre transform of the Lagrangian, or the function of the total energy of the system, is now given the name Hamiltonian after the Irish mathematician William Hamilton, and denoted with the symbol $\mathcal{H}$. This function is a close relative of the quantum mechanical Hamiltonian encountered already in previous sections; the latter is however an operator acting on a wavefunction, and not a function defined over a $6N$-dimensional phase space, so they should not be confused.

One marked difference in which the Hamiltonian and Lagrangian formulations differ from each other is the way in which both generate their respective equations of motion. Due to the time derivative of the velocity derivative in the Euler-Lagrange equations, one usually arrives at a set of $3N$ coupled second-order equations, similar to the Newton's 2^{nd} law. On the other hand, Hamilton's equations of motion form a set of $6N$ first-order differential equations, following the famous relation (here I use an arbitrary set of coordinates q to highlight the generality of this scheme):

$$\begin{aligned} \dot{p}_i &= -\frac{\partial \mathcal{H}}{\partial q_i} \\ \dot{q}_i &= \frac{\partial \mathcal{H}}{\partial p_i} \end{aligned} \tag{3.27}$$

from which the conservation of the total energy (identical with the Hamiltonian) follows immediately:

$$\dot{\mathcal{H}} = \sum_i \left(\frac{\partial \mathcal{H}}{\partial q_i} \dot{q}_i + \frac{\partial \mathcal{H}}{\partial p_i} \dot{p}_i \right) = \sum_i \left(\frac{\partial \mathcal{H}}{\partial q_i} \frac{\partial \mathcal{H}}{\partial p_i} - \frac{\partial \mathcal{H}}{\partial p_i} \frac{\partial \mathcal{H}}{\partial q_i} \right) = 0 \tag{3.28}$$

The above equation is often written in a shorthand notation $\{\mathcal{H}, \mathcal{H}\}$ known as the Poisson bracket. It makes it trivial to evaluate the time dependence of any function of coordinates and momenta as $\dot{f} = \{f, \mathcal{H}\}$. Notably, this property facilitated the discovery of a formal link between symmetries of the Hamiltonian and conservation laws, known as Noether's theorem.

Another important property that is easily derived from the Hamiltonian formalism is the incompressibility of the phase space: if one cuts out a $6N$-dimensional volume element of the phase space and follows the time evolution of all points contained in that region of phase space, after any given time they will correspond to the same volume. This can be shown formally as the divergence of the velocity field in phase space being equal to zero:

$$\nabla \cdot \dot{\mathbf{x}} = \sum_i \left(\frac{\partial}{\partial q_i} \dot{q}_i + \frac{\partial}{\partial p_i} \dot{p}_i \right) = \sum_i \left(\frac{\partial}{\partial q_i} \frac{\partial \mathcal{H}}{\partial p_i} - \frac{\partial}{\partial p_i} \frac{\partial \mathcal{H}}{\partial q_i} \right) = 0 \tag{3.29}$$

leveraging the fact that p and q are independent variables. The same fact can be shown by considering the Jacobian of the transformation from coordinates at $t = 0$ to coordinates at $t = \Delta t$ where $x(\Delta t) = x(0) + \dot{x}\Delta t + \mathcal{O}(\Delta t^2)$, and dropping the $\mathcal{O}(\Delta t^2)$ terms:

$$\mathcal{J} = \begin{vmatrix} \frac{\partial q(\Delta t)}{\partial q(0)} & \frac{\partial q(\Delta t)}{\partial p(0)} \\ \frac{\partial p(\Delta t)}{\partial q(0)} & \frac{\partial p(\Delta t)}{\partial p(0)} \end{vmatrix} = \begin{vmatrix} 1 + \frac{\partial^2 \mathcal{H}}{\partial q \partial p} & \frac{\partial^2 \mathcal{H}}{\partial p^2} \\ \frac{\partial^2 \mathcal{H}}{\partial q^2} & 1 - \frac{\partial^2 \mathcal{H}}{\partial q \partial p} \end{vmatrix} = 1 \tag{3.30}$$

and observing that an infinitesimal transformation preserves the volume of the phase space; this test is indeed useful in determining the Hamiltonian-preserving properties of various algorithms used to numerically propagate the equations of motion in Hamiltonian systems.

Finally, the remarkable power, flexibility and unity of Hamiltonian and Lagrangian formulations of classical mechanics are neatly illustrated by the Hamilton's principle – also known as the principle of stationary action – which states that particles' trajectories in the phase space evolve so as to make action stationary, i.e., to satisfy the condition

$$\delta \int_{t_1}^{t_2} \mathcal{L}(\mathbf{q}, \dot{\mathbf{q}}, t) dt = 0 \tag{3.31}$$

which is the functional equivalent of the extremum condition in single-variable calculus. This means that infinitesimally small changes to the system's trajectory do not change the action functional $S = \int \mathcal{L} dt$. It is fascinating to find that from this principle one can also easily derive the Euler-Lagrange equations, showing that the two approaches to dynamics of mechanical systems – differential or causal, based on the idea of local interactions incrementally drive global changes, and integral or global, in which the system "considers" all possible trajectories and chooses to follow the optimal one – are in fact equivalent ways to view the same physical principle. This inspiring concept went on to inspire many fruitful efforts in the most fundamental aspects of quantum mechanics and quantum field theory.

3.2.2 Algorithmic Realizations

Force fields

Modern software packages for atomistic simulation rely on many advanced and ingenious algorithmic developments that make it possible to numerically propagate the equations of motion in large molecular systems in an efficient and accurate manner. However, an equally important component is the so-called force field, or the energy function that maps particle coordinates to energies. Although different classes and instances of force fields are being used in the field of molecular simulations, most of them have a similarly modular structure and can be broken down into well-defined terms. Here I will only briefly outline relevant concepts as specific implementations differ vastly between individual software suites.

The first component of the force field is the bonded terms, i.e. energies of interaction between covalently linked atoms. Typically bond, angle and dihedral components are included here, the former two modelled as harmonic potentials $\frac{1}{2}k(x - x_0)^2$. The use of harmonic potentials is justified when deviations from the potential energy minima are small, as is often the case for chemically stable molecules in ambient conditions; however, in such a framework molecule topologies are fixed, i.e. bonds cannot be broken or created, and the quantum chemical zero-point energy effects are not reproduced properly, slightly distorting the picture of atomistic dynamics. Nevertheless, these approximations are robust enough to yield realistic behavior of molecules even when bond lengths are completely constrained, so that except for very specific cases their use is customary.

The dihedral term, corresponding to a rotation of two bonds ij and kl about the central bond jk, is often modelled as a sum of cosine terms $\sum_{i=1}^{6} k_i \cos(n\theta + \theta_0)$. The rationale for such a functional form is that the potential should be periodic with a period of 2π as a corresponding rotation produces an identical structure, and many chemical moieties often have a 1-, 2-, 3- or 6-fold symmetry with respect to the rotation about the bond. Moreover, so-called improper dihedral terms with 2-fold symmetry are frequently introduced to enforce the planarity of predominantly sp^2-hybrydized moieties, in which the definition of the dihedral involves the four atoms that are intended to remain coplanar.

The remaining force field terms are dubbed non-bonded as they describe interactions through space rather than through chemical bonds. The two non-bonded components used in most modern force fields are the electrostatic interaction, modelled using the Coulomb formula for the electrostatic interaction energy between two point charges, $\frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}}$, and the Lennard-Jones term, intended to represent all van der Waals (i.e., related to induced dipoles) energy terms through a "6-12" potential $\varepsilon_{ij}[(\frac{\sigma_{ij}}{r_{ij}})^{12} - (\frac{\sigma_{ij}}{r_{ij}})^6]$.

These general functional forms can be then modified or overridden with correction terms. For instance, some popular force fields such as Amber scale the charge-charge and Lennard-Jones interactions between the 1-4 atom pairs (i.e., atoms *i* and *l* in a *ijkl* sequence connected by three consecutive bonds) by a specific factor (here 0.8333 and 0.5, respectively) to account for their spatial proximity. Other force fields, such as CHARMM, use explicitly modified σ_{ij} and ε_{ij} parameters for such pairs, as well as to override the standard combination rule $\sigma_{ij} = \frac{1}{2}(\sigma_i + \sigma_j)$ and $\varepsilon_{ij} = \sqrt{\varepsilon_i \varepsilon_j}$ for selected atomtype pairs. Yet another promising approach was the introduction of the CMAP correction in CHARMM22 [237]: here, the idea was to explicitly calculate the correlational contribution to the dihedral energy between adjacent backbone ψ and ϕ dihedral angles, that is, the difference between the 2-dimensional energy map and sum of 1-dimensional profiles.

There remains, however, a deep divide among the scientists on whether one should let such force field-specific corrections proliferate or should we rather converge towards more unified, physics-based models. While it is clear that the corrected models introduced an enormous improvement in the accuracy of molecular simulations, some argue that this is not the proper way to address the underlying shortcomings. These systemic improvements, however, are being steadily delivered through, most notably, the recent (and well anticipated) developments of polarizable force fields and constant-pH molecular dynamics. It is widely believed that the convergent improvement of hardware, algorithms and force fields will soon give us truly predictive power in tackling complex biophysical problems.

Integration of the Equations of Motion

Much less attention is drawn nowadays to the perhaps most central algorithmic aspect of implementation of classical mechanics in simulation suites: the integration of equations of motion. This reflects the fact that most developments here happened already in the past, and a set of agreed upon general rules have been widely accepted. The integration itself is also subject to very little variation between individual codes, and even though a degree of inaccuracy is always present in numerical approaches, other schemes such as thermostats exist that correct for e.g. energy drift during the calculations.

As mentioned above while discussing eqn. 3.30, the integration of Hamiltonian equations of motion require that the algorithm is symplectic, which – for all practical purposes – relates to the preservation of the phase space volume (the Jacobian of transformation from $x(0)$ to $x(\Delta t)$ being equal to 1) and conservation of a shadow Hamiltonian. The term "shadow" here indicates that what is actually conserved is not the exact energy of the system, but the energy of a closely related one. Notably, not all common integration schemes share this property: while it is somewhat expected that the simplistic Euler method fails to meet these criteria, it might be surprising to learn that the same is the case for the whole set of highly precise Runge-Kutta methods.

Fortunately, a deterministic route exists to construct symplectic algorithms of desired accuracy, and the overwhelmingly popular Verlet algorithm can be indeed derived using this set of rules. The algorithm only propagates positions, and can be derived by considering the numerical formula for a second order derivative using the central difference approximation:

$$\begin{aligned} a(0) &= \frac{\Delta^2 x}{\Delta t^2} = \frac{\Delta v}{\Delta t} = \frac{v(\frac{\Delta t}{2}) - v(-\frac{\Delta t}{2})}{\Delta t} \\ &= \frac{\frac{x(\Delta t)-x(0)}{\Delta t} - \frac{x(0)-x(-\Delta t)}{\Delta t}}{\Delta t} = \frac{x(\Delta t) - 2x(0) + x(-\Delta t)}{\Delta t^2} \\ x(\Delta t) &= 2x(0) - x(-\Delta t) + a(0)\Delta t^2 \end{aligned} \tag{3.32}$$

When velocities are required, they can always be recovered *post-hoc* from positions, but sometimes it is more convenient to use the velocity Verlet algorithm – a variant equivalent in terms of precision (both yield the single-step error of $\mathcal{O}(\Delta t^4)$ and a cumulative error of $\mathcal{O}(\Delta t^2)$), but one that explicitly keeps track of velocities; hence its name. In velocity Verlet, we first update the positions of the particles from the simple relation $x(\Delta t) = x(0) + v(0)\Delta t + \frac{1}{2}a(0)\Delta t^2$, and then update the velocities using the mean acceleration between times 0 and Δt, i.e. $v(\Delta t) = v(0) + \frac{1}{2}(a(0) + a(\Delta t))$.

Finally, the family of popular $\mathcal{O}(\Delta t^2)$ symplectic integrators is completed by the somewhat curious leapfrog algorithm, one that evaluates velocities at timestamps shifted by $\frac{1}{2}\Delta t$ with respect to the velocities. This is exactly equivalent to the velocity Verlet algorithm in which velocities are evaluated at intermediate steps, and hence the procedure is very similar: one first updates the positions as $x(\Delta t) = x(0) + v(\frac{1}{2}\Delta t)\Delta t$, and then velocities as $v(\frac{1}{2}\Delta t) = v(-\frac{1}{2}\Delta t) + a(0)\Delta t$.

Importantly, the Verlet, velocity Verlet and leapfrog algorithms share one more fundamental property: they are all time-reversible, which means that they can be first propagated forward and then backward in time to arrive again at the starting point. While not strictly a requirement for symplecticity, time-reversibility is actually a prerequisite for some computational methods such as milestoning, and is one of the fundamental properties of Nature (except for the rare cases in e.g. physics of the weak force where only the more general CPT symmetry holds).

One curious and noteworthy modification of the standard integrators is the RESPA scheme, a multi-step algorithm based on Trotter decomposition proposed by Tuckerman. Here, the assumption is that the forces in the system can be divided into fast- and slowly-varying, and that the slowly-varying (and potentially costly) component only needs be evaluated in several time steps' strides. This promising approach can be used e.g. to fix the calculations that use fast but highly approximate energy functions by performing a more accurate calculation step once in a while, and using the mean difference between the two as the slowly-varying force.

Thermostats and Barostats

The somewhat less strict requirements regarding exact energy conservation are in part due to the fact that in typical simulations of physically relevant systems, one is not interested in conservation of the total energy, but instead wants to maintain constant (yet fluctuating) temperature. This means that when potential energy drops due to e.g. a highly exoenergetic binding event, the excess kinetic energy released in the event has to be taken out from the system by a coupled heat bath for the simulation to be realistic. This heat bath-like behavior is emulated by separately designed algorithms – thermostats – whose sole purpose is to modify the distribution of kinetic energy so that it corresponds to a specified temperature.

Conceptually, perhaps the simplest thermostat is the one introduced by Andersen [234]: it ensures that the velocity distribution is correct by repeatedly drawing the velocity of a randomly selected particle from the correct distribution itself. However, due to the random drawing events, correlations in the system are often lost rapidly, leading to the overestimation of natural timescales for individual events. For this reason, it founds little use in modern simulations.

Soon after Andersen proposed his thermostat, Berendsen came up with what turned out to be one of the most popular thermostatting schemes, the Berendsen (weak coupling) thermostat. Here, the simple idea was to uniformly rescale all velocities in the system so that at each step, the system's temperature T tends towards the specified value T_0 with first order kinetics: $\frac{dT}{dt} = \tau^{-1}(T_0 - T)$. When discretized, this equation produces the rescaling factor in the form of $[1 + \frac{\Delta t}{\tau}(\frac{T_0}{T(-\Delta t)} - 1)]^{\frac{1}{2}}$. Unfortunately, this algorithm yields overdamped temperature fluctuations and hence does not sample the correct thermodynamic ensemble; this also affects properties such as heat capacity, connected to the variance of the total energy through a linear relationship. Regardless of such issues, the thermostat was broadly used due to its simplicity and intuitive behavior.

Nowadays, the corrected scheme of Berendsen introduced recently by Bussi has gained great popularity thanks to the ingenuity of the key correction [238]. The thermostat, often referred to as CSVR (Canonical Sampling through Velocity Rescaling, abbreviation of the seminal article's title), modifies the idea of Berendsen by requiring that – instead of a fixed temperature T_0 – the system tend to a randomly fluctuating temperature T_k, drawn each time from the Maxwell-Boltzmann distribution of kinetic energy. This modification leads to a correcting term in kinetic energy, $2\sqrt{\frac{KK_0}{N_{dof}}}\frac{dW}{\tau}$, in which K and K_0 are the instantaneous and reference kinetic energy, N_{dof} the number of degrees of freedom in the system, and dW a stochastic Wiener process (essentially a random walk). In this form, CSVR finds itself currently among the most often used thermostats.

The only other contender to this title is the popular Nosé-Hoover thermostat, one built on the idea of an extended Lagrangian introduced by Andersen in his early barostat, as discussed above in the section on Lagrangian mechanics. The main similarity is that in the scheme due to Nosé, the thermal bath parameter s is introduced as a scaling factor for momenta, and a fictitious mass-like term is attached to the thermostat to account for the inertia of the coupling. However, this scheme produced a system that was no longer Hamiltonian due to a non-canonical transformation of time, and the distribution of the thermal bath variable itself was incorrect. To alleviate this problem, Hoover introduced a "chain" of extra variables, each designed in such a way to ensure the correct distribution of the previous one. This complicated algorithm has been excellently reviewed in a book by Tuckerman [233] and will not be covered in great detail here; however, it is worth pointing out that while other thermostats thermalize the system in an asymptotic manner, the Nosé-Hoover thermostat produces oscillatory relaxation, and for this reason is not always a viable option for the initial thermalization of the system.

Many designs of barostats mirror those of thermostats, and both are collectively referred to as weak coupling algorithms. The similarities between the Andersen barostat and the Nosé thermostat were already highlighted above; the popular Parinello-Rahman barostat also builds on the idea of Andersen, introducing improvements that allow not only for a change in volume but also in shape [239]. Simultaneously, the Berendsen barostat resembles the corresponding thermostatting scheme in its reliance on the asymptotic rescaling of cell volume according to the instantaneous difference between the current and reference pressures. Notably, though, barostats are often the first algorithms to break down in case of numerical instabilities, as the virial term used to calculate pressures contains a sum of all $F_{ij}r_{ij}$ terms and can cause near-infinite surges in pressure in case singularities are encountered, or harmonic bonds are stretched indefinitely.

Nowadays, even modest simulation engines rely on many more advanced algorithms. To name a few, the use of neighbor lists allows for near-linear scaling calculation of interactions that would otherwise have a quadratic ($\mathcal{O}(N^2)$) dependence on the system size; the separation of electrostatic interactions into short- and long-range components and the treatment of the latter with particle-mesh Ewald summation in the Fourier space also contributes largely to this speedup. As the field progresses towards petascale computing with massive amount of computational resources being allocated for simulations, however, some other algorithmic flaws are becoming clearer, such as the ($\mathcal{O}(M^2)$) timing of interprocess communication or suboptimal allocation of tasks between individual processes. Such flaws do indeed drive further development in the field, resulting in the conception of yet more efficient and precise approaches.

3.2.3 Statistical Thermodynamics and Free Energy

Ensembles and the Boltzmann Distribution

So far, the discussion was mostly concerned with adequately simulating the temporal evolution of molecular systems based on either quantum or classical mechanical laws of motion. However, in making predictions regarding experimentally measurable quantities, one is often less interested in the exact finite-length trajectory of a single realization of a system, and more in the macroscopic properties that often arise from complex behaviors on the atomistic scale. Statistical thermodynamics provides a unifying framework for such efforts by introducing the concept of ensembles – statistical distributions of all possible states of the system, or, viewed alternatively, multiple realizations of the system weighted according to the probability of their occurrence. In thermodynamic equilibrium, i.e., in absence of net energy flows, such ensembles “produce” static properties as weighted averages over all possible configurations and initial conditions, and accurate estimation of these experimentally relevant properties – all while retaining atomistic resolution of the system at hand – is often viewed as the holy grail of molecular simulation.

Let’s first consider the simplest case of the so-called NVE ensemble, the three-letter name referring to quantities that are kept constant throughout the simulation: here, it is the number of particles N, the box volume V and total energy E. If these quantities are specified, along with an agreed upon Hamiltonian, one can (somewhat trivially) state that the dynamics of such a system is restricted in the phase space to a hypersurface defined by the condition $\delta(\mathcal{H}(\mathbf{x}) - E)$, here $\delta(\mathbf{x})$ being the Dirac delta distribution. At this point, the assumption of equal probability has to be introduced: one needs to assume that out of every possible realization of the system along an indefinitely long trajectory on the hypersurface, all microstates have the same probability of occurring since none is distinguished in any way.

In this case, the phase-space volume corresponding to this hypersurface is called the density of states [240] or the microcanonical partition function [233]:

$$\Omega(N, V, E) \propto \int d\mathbf{p} \int d\mathbf{q} \delta(\mathcal{H}(\mathbf{p}, \mathbf{q}) - E) \tag{3.33}$$

and is one of the central quantities of statistical thermodynamics, allowing to relate ensembles of microstates (defined as points in phase space) to experimentally relevant macrostates (defined as probability densities in phase space). If all N particles were indistinguishable, the density of states should be divided by $N!$ to account for all possible permutations that yield the same microstate; however, as long as one is only concerned with its relation to energy, i.e., $\Omega(E; N, V)$, Ω only needs be defined up to a multiplicative constant. (For this reason, it is also convenient to use the proportionality sign to escape the issue of dimensionality: “real” partition functions should be dimensionless, and typically an additional constant factor is introduced to cancel the dimensions of the integral.)

With the definitions at hand, it is now straightforward to write down equations for the ensemble average of any function of positions and momenta $a(\mathbf{x})$:

$$\langle a \rangle = \frac{1}{\Omega} \int d\mathbf{p} \int d\mathbf{q} a(\mathbf{p}, \mathbf{q}) \delta(\mathcal{H}(\mathbf{p}, \mathbf{q}) - E) \tag{3.34}$$

It is also useful to define a quantity called entropy, a measure of the number of microstates available to the system in a given macrostate:

$$S = k_B \log[\Omega(N, V, E)] \tag{3.35}$$

that has the property of additivity, i.e., if for two uncoupled systems A and B the number of jointly available microstates is the product $\Omega_A \Omega_B$, then the total entropy of the two systems is given by $k_B \log(\Omega_A \Omega_B) = S_A + S_B$; here k_B refers to the Boltzmann constant.

For a simulation to be a valid tool for the exploration and calculation of the partition function, one has to assume the validity of the ergodic hypothesis, namely that the distribution sampled in the phase space during an indefinitely long period of temporal evolution is essentially identical irrespective of the boundary condition (i.e., starting point). In practice, the timescale implied by this hypothesis often exceeds the timescales available to actual computations by orders of magnitude.

Although simple, the microcanonical ensemble is rarely realistic as it only correctly describes thermodynamically isolated systems contained in a fixed volume. In practice, experimentally relevant conditions typically involve some sort of coupling to a heat bath, so that the system maintains a constant temperature over time; if one lets the volume remain fixed, what results is the canonical ensemble or the NVT ensemble.

In NVT conditions, the internal energy of the system E_s can fluctuate as the system exchanges heat with the heat bath; for simplicity, the system and the bath can be assumed to be perfectly isolated from the exterior with a total (conserved) energy E_0, and the heat bath to be much larger than the system, so that in thermal equilibrium their energies obey $E_0 \gg E_s$, and the energy of the heat bath is just $E_0 - E_s$. If one now picks a single microstate i of the system with energy ϵ_i, and wishes to find its probability $p(\epsilon_i)$, this probability is proportional to the number of all microstates of the isolated system that are compatible with the selected setup, i.e. one times the number of microstates of the bath with energy $E_0 - \epsilon_i$:

$$p(\epsilon_i) = \Omega(E_0 - \epsilon_i) = \exp\left[\frac{S(E_0 - \epsilon_i)}{k_B}\right] \tag{3.36}$$

Because ϵ_i is very small compared to E_0, one can Taylor expand $S(E_0 - \epsilon_i)$ to get:

$$S(E_0 - \epsilon_i) = S(E_0) - \epsilon_i \frac{\partial S}{\partial E} + ... \tag{3.37}$$

Noting that the exact value of E_0 is constant and arbitrary, 3.36 can be rewritten as:

$$p(\epsilon_i) = \exp\left[\frac{1}{k_B}\left(S(E_0) - \epsilon_i \frac{\partial S}{\partial E}\right)\right] \propto \exp\left(\frac{-\epsilon_i}{k_B}\frac{\partial S}{\partial E}\right) \tag{3.38}$$

Since $\frac{\partial S}{\partial E}$ is the actual statistical mechanical definition of the inverse temperature $\frac{1}{T}$, this relation can be written in form of the easily recognizable Boltzmann distribution:

$$p(\epsilon_i) = \frac{1}{Z}\exp\left(\frac{-\epsilon_i}{k_B T}\right) \tag{3.39}$$

in which the normalizing factor Z becomes simultaneously the new partition function for the canonical ensemble:

$$Z \propto \sum_i \exp\left(\frac{-\epsilon_i}{k_B T}\right) = \int d\epsilon \int d\mathbf{p} \int d\mathbf{q} \exp\left(\frac{-\epsilon}{k_B T}\right)\delta(\mathcal{H} - \epsilon) \tag{3.40}$$

Here, the summation over all discrete microstates i was expressed as explicit integration over the phase space volume corresponding to energy ϵ, and considering all possible values of ϵ; in fact, this new quantity could be easily expressed in terms of the microcanonical energy function:

$$Z(N,V,T) \propto \int d\epsilon \exp\left(\frac{-\epsilon}{k_B T}\right)\Omega(N,V,\epsilon) \tag{3.41}$$

By analogy to eqn. 3.34, the new weighting factor also allows to write down a modified equation for the weighted ensemble average:

$$\langle a \rangle = \frac{1}{Z}\int d\mathbf{p}\int d\mathbf{q} a(\mathbf{p},\mathbf{q}) \exp\left(\frac{-E(\mathbf{p},\mathbf{q})}{k_B T}\right) \tag{3.42}$$

Since the Boltzmann factor is a strictly monotonous function, one could intuitively assume that low-energy microstates will dominate the distribution, essentially freezing the system in an ordered and static configuration. This is, however, not the case thanks to the Dirac delta term: the number of high-energy microstates (the density of states) grows roughly polynomially as more energy is available, balancing the exponential decrease of the Boltzmann factor and usually yielding a sharp, unimodal (and system size-dependent) distribution of microstate energies.

Ultimately, to connect to experimental conditions one would also like to lift the constraint of constant volume, replacing it with constant pressure. In analogy to the above discussion, coupling the system to a "pressure reservoir" by means of a virtual piston produces a Boltzmann factor of $\exp\left(\frac{-E(V)}{k_B T}\right) = \exp\left(\frac{-PV}{k_B T}\right)$. Now, following eqn. 3.41, the new partition function reads:

$$Z_p(N,P,T) \propto \int_0^\infty dV \exp\left(\frac{-PV}{k_B T}\right) Z(N,V,T) \tag{3.43}$$

In aqueous condensed-phase systems under ambient conditions that are typically of interest for biologists, however, the two ensembles – NVT and NPT – typically produce very similar results due to the incompressibility of liquid water: under constant pressure, only a very small range of volumes corresponds to plausible physical outcomes.

Free Energy as the Potential of Mean Force

The statistical interpretation of the partition function – the "generalized number of microstates" available to a system under the given conditions – implies that on average, the system in question will tend to maximize this quantity (or, equivalently, its logarithm). In the microcanonical ensemble, the quantity to be maximized was hence the entropy; as a consequence, if an isolated system is constrained to a subset of possible microstates, the release of the constraint will result in the system expanding to explore the previously unavailable regions of the phase space in the isoenergetic subspace, and its entropy will increase. This tendency to evolve in an unconstrained manner can also be thought of as an entropic force that counteracts the constraining agent (e.g. a barrier that keeps all particles within one half of the box).

Correspondingly, in a given temperature the canonical and NPT ensembles will tend to maximize their respective partition functions, or minimize their negative logarithms (often with a pre-factor), the Helmholtz and Gibbs free energies; here the term "free energy" will be used in a liberal manner to refer to both without distinction. Analogously to the microcanonical ensemble case, this tendency to maximize the partition function will manifest itself as a force acting on the constraint. Now, if one imagines a constraint given by $\xi = \text{const}$ – with $\xi(\mathbf{x})$ being an arbitrary generalized coordinate – it is possible to show that the ensemble average of the force acting on the constraining agent is exactly the (negative) derivative of the free energy, G, along the generalized coordinate:

$$\begin{aligned}\frac{\partial G}{\partial \xi} &= \frac{\partial}{\partial \xi}\left[-k_B T \log \int d\mathbf{x} \exp\left(\frac{-E(\mathbf{x})}{k_B T}\right)\right] = -k_B T \frac{\frac{\partial}{\partial \xi} \int d\mathbf{x} \exp\left(\frac{-E(\mathbf{x})}{k_B T}\right)}{\int d\mathbf{x} \exp\left(\frac{-E(\mathbf{x})}{k_B T}\right)} \\ &= -k_B T \frac{-k_B T \int d\mathbf{x} \frac{\partial E}{\partial \xi} \exp\left(\frac{-E(\mathbf{x})}{k_B T}\right)}{\int d\mathbf{x} \exp\left(\frac{-E(\mathbf{x})}{k_B T}\right)} = \left\langle \frac{\partial E}{\partial \xi} \right\rangle = -\langle F_\xi \rangle\end{aligned} \tag{3.44}$$

Hence, the slope of $G(\xi)$ indicates the direction in which the system will predominantly evolve. The function itself is typically referred to as the free energy profile:

$$G(\xi) = -k_B T \log\left(\int d\mathbf{x} \exp\left(\frac{-E(\mathbf{x})}{k_B T}\right) \delta(\xi(\mathbf{x}) - \xi)\right) \tag{3.45}$$

and is directly tied to the probability density of finding the system at any given value of the generalized coordinate (also called collective variable, CV):

$$\rho(\xi) \propto \exp\left(\frac{-G(\xi)}{k_B T}\right) \tag{3.46}$$

This connection between microscopic forces and ensemble statistics provides an invaluable tool to interpret the results of experiments, and recent efforts in the field gradually tighten the gap between experimental and simulational data. An obvious caveat, though, is that the selection of the CV has to take into consideration the experimental setup, i.e. an adequate forward model of the process of interest has to be available. While in some cases trivial, the construction of proper CVs that simultaneously ensure rapid convergence of the simulation statistics, correspond to experimental observables and account for all nuances of the process can be an art in itself.

Jacobian Contribution to Entropy

Admittedly, some confusion persists in the community regarding the exact meaning of the term "free energy" and its connection to mean force [241, 242]. When used interchangeably with the term "potential of mean force", it supposedly corresponds to the integral of the quantity in eqn. 3.44. However, the term E in this equation refers to the total energy of the system, $E = U + K$, and hence one should not equate the effective force F_ξ with the typical notion of force as the gradient of potential energy,

$\nabla_\xi U$. In fact, in the general case these quantities are connected by the Jacobian term [240]:

$$\left\langle \frac{\partial E}{\partial \xi} \right\rangle_{\xi_0} = \left\langle \frac{\partial U}{\partial \xi} - k_B T \frac{\partial \log |\mathcal{J}|}{\partial \xi} \right\rangle_{\xi_0} \tag{3.47}$$

The meaning of this term can be better understood if one considers a system of non-interacting particles in a large periodic box, i.e. given by $U(\mathbf{q}) = 0$ and, consequently, $\frac{\partial U}{\partial \xi} = 0$. By randomly picking two particles and calculating the distribution of their distances up to some threshold, i.e., $p(r_{12})$, one should find that it is much less probable to observe the particles at a very small distance than at a large one. This is a consequence of the fact that for the distance between the two particles to fall within the range $(r_{12}, r_{12} + \delta r)$, particle 2 has to be within a spherical shell with radius r_{12} and thickness δr. Since the volume of the shell, $4\pi r_{12}^2 \delta r$, grows as r_{12}^2, this is more probable when r_{12} is large; in fact, $p(r_{12}) \propto r_{12}^2$ and, consequently, $G(r_{12}) = -k_B T \log(r_{12}^2) = -2k_B T \log(r_{12})$ (valid up to an additive constant); the Jacobian term can be therefore equated with r_{12}^2, as could be inferred from the radial component of the spherical coordinates' Jacobian. If one requires that the free energy profile only reflect energetic and not geometric properties of the system, i.e. asymptotically approach a selected value (typically chosen to be zero) in the non-interacting large distance limit, this term has to be subtracted from the free energy profile obtained in most conventional manners.

The problem becomes mathematically more easily tractable if eqn. 3.45 is rewritten using a coordinate transformation $\{q_1 ... q_{3N}, p_1 ... p_{3N}\} \rightarrow \{\xi, u_2 ... u_{3N}, p_\xi, \pi_2 ... \pi_{3N}\}$ so as to avoid the use of a Dirac delta:

$$G(\xi) = -k_B T \log \left[\int du_2 ... du_{3N} \int dp_\xi d\pi_2 ... d\pi_{3N} |\mathcal{J}| \exp\left(\frac{-E(\mathbf{x})}{k_B T} \right) \right] \tag{3.48}$$

Now all degrees of freedom except for ξ are integrated out, and the determinant of the transformation's Jacobian explicitly appears in the equation due to the change of coordinates. This is consistent with the pictorial description above, as the natural interpretation of the Jacobian is a measure of compression of the original space due to the transformation: for example, at large values of r_{12} more Cartesian space is compressed into a single point on the ξ axis than at small r_{12}. Also, in principle the choice of all remaining coordinates is arbitrary as long as they are well-behaved, since any effect they have on the Jacobian will vanish during integration.

Free Energy Methods: Umbrella Sampling

Thanks to the direct connection between free energies and probability densities, in principle – if arbitrarily long trajectories were available – one could calculate free energy profiles by the so-called Boltzmann inversion, or multiplication of the log-probability by $-k_B T$. In practice, though, this logarithmic relationship is exactly what complicates the issue: at an ambient temperature, a free energy difference of 10 kcal/mol translates to an almost 20 million-fold ratio of probabilities, meaning that to obtain a single sample from the high-energy state, tens of millions of independent samples would have to be drawn that belong to the low-energy one. For a process with a relatively short autocorrelation time of 1 ps, this yields a mean first-passage time of tens of microseconds. For a realistic moderately-sized system, that would require at least half million CPU hours – an already resource-intensive calculation. Yet, free energy barriers and differences are rarely known *a priori*, so that a more general and robust approach is needed to quantify thermodynamic preferences.

Umbrella sampling, one of the most popular free energy methods, is based on the premise that one can modify the original potential energy surface with a biasing external potential V – in practice, most often explicitly dependent on a collective variable ξ along which the free energy profile is determined – and then recover the (relative) unbiased free energy profile F_0 by subtracting the biasing potential from the biased free energy profile F, obtained from Boltzmann inversion of the biased probability histograms [243]:

$$F_0(\xi) = F(\xi) - V(\xi) - k_B T \log\left(\frac{Q_0}{Q}\right) \tag{3.49}$$

with the generally unknown factor $k_B T \log\left(\frac{Q_0}{Q}\right)$ – depending on the ratio of partition functions of the unbiased and biased ensembles – shifting the free energy profile vertically. This means that if one could *a priori* guess such a biasing potential that the probability of sampling high-energy intervals along the CV would increase considerably, less computational time would be spent on sampling the low-energy regions. Again, however, one usually only has vague clues about the shape of the free energy profile prior to actually running the calculation, and hence trying to guess an appropriate biasing potential could easily become unfeasible. Also, one hidden assumption – and one the most violated – is that all remaining degrees of freedom orthogonal to the selected CV have to be sampled properly, so that the biased free energy profile has actually converged.

Fortunately, 15 years after the original idea was put forward it was realized that in fact, multiple such umbrella sampling runs can be performed in parallel as long as one can then "stitch" them together through statistical estimation of $\frac{Q_0}{Q}$. The reasoning behind the procedure proposed in the seminal paper [244] is as follows: (1) one wishes to minimize the statistical error of the global unbiased probability density; (2) the global unbiased probability density is a weighted sum of local (i.e. estimated in a single simulation, or "window") unbiased probability densities; (3) hence the error of the global unbiased density is a corresponding weighted sum of errors in local unbiased densities; (4) in a discrete setup, the error in a local unbiased density is proportional to the error in independent counts in the histogram; (5) the squared error in histogram counts is proportional to the expected value of histogram counts [245]. By combining the notions 1-5 with the requirement that the weights sum to 1, one arrives at the WHAM (weighted histogram analysis method) equations for the global binned probabilities (a discretized counterpart of the continuous density) π and the relative free energies ($F_i = -k_B T \log\left(\frac{Q_0}{Q_i}\right)$):

$$
\begin{aligned}
\pi(\xi_j) &= \frac{\sum_i n_i(\xi_j)}{\sum_i N_i \exp\left[(F_i - V_i(\xi_j))/k_B T\right]} \\
F_i &= -k_B T \log\left(\sum_j \pi(\xi_j) \exp\left[V_i(\xi_j)/k_B T\right]\right)
\end{aligned}
\tag{3.50}
$$

where i runs over individual simulations (windows) and j over bins of the histogram, while $n_i(\xi_j)$ and N_i refer to the histogram count in a bin centered at ξ_j and the total frame count in simulation i, respectively. Since these equations mutually depend on each other, they are iteratively solved to self-consistency, starting e.g. with a uniform probability distribution. Note also that by definition, the relative free energies F_i are actually expected values (or ensemble averages) of the biasing potential in the i-th window.

In practice, once seeding frames along the pathway of interest are generated (e.g. using steered MD), US runs are performed in parallel using regularly spaced harmonic potentials, and the resulting free energy profiles are retrieved using WHAM. Note that due to the choice of optimization method, histograms along ξ need to overlap between neighboring windows in order to yield meaningful relative free energies F_i. On the other hand, the biasing potential needs not be harmonic, even though most existing implementations of WHAM make that assumption; for some specific applications, I wrote my own Python implementation that is free of this limitation (available at `https://gitlab.com/KomBioMol/wham`).

A useful feature of the combined US/WHAM approach is that it provides statistical weights for individual frames, allowing for easy recalculation of ensemble averages or free energy profiles in other CVs, provided that they were adequately sampled. While in an equilibrium MD simulation all frames have equal statistical weights, frames derived from US runs have weights equal to $\exp[(U_i(\xi) - F_i)/k_B T]$. As a result, US can often be set up using a CV that is convenient to choose (e.g. thanks to a closed-form formula for the corresponding gradient), and then the free energy profile can be recalculated in a more complex CV (e.g. one that requires a post-hoc analysis) through Boltzmann inversion of the WHAM-derived probabilities.

Finally, in recent years US runs have been routinely augmented with Hamiltonian replica exchange, a combination referred to as H-REUS. In line with the famous Metropolis criterion, once every n-th MD step a Monte Carlo move is attempted, in which atomic configurations are swapped between the neighboring windows i and j with a probability equal to [246]:

$$
p_{swap} = \min\left[1, \exp\left(\frac{V_i(\xi_i) + V_j(\xi_j) - V_i(\xi_j) - V_j(\xi_i)}{k_B T}\right)\right] \tag{3.51}
$$

where ξ_i is the value of the CV in window i. With this modifications, slowly relaxing systems stuck in local free energy minima along the orthogonal degrees of freedom can explore the configuration space in an accelerated manner, yet still preserve detailed balance. A minor drawback of this scheme is, however, the loss of most data regarding local kinetics; this is somewhat less of an issue as kinetics in biased simulations is generally perturbed in an unknown manner, and strong assumptions have to be made to model unbiased rates from biased simulations [247].

Free Energy Methods: Metadynamics

The other broadly used family of free energy methods relate to the general notion of "flat histogram" methods, all connected by a common goal to adjust state weights so that the coordinate of interest is sampled according to a uniform distribution [240]. In other words, an external potential has to be first determined and then used to bias the system's potential energy so as to render all values of the collective variable equiprobable; once that is the case, the negative biasing potential is exactly the free energy. This feat is often realized in an iterative way, where the biasing potential is deposited on-the-fly to discourage re-visiting the same intervals along the CV, resulting in an accelerated escape from free energy minima. In the Wang-Landau scheme [248], a Monte Carlo simulation was performed to calculate the entropy as a function of the total energy, and after each Monte Carlo move the estimate of the entropy associated with the current state's energy was increased by a small fixed value δ:

$$\begin{aligned} p_{i\to j} &= \min[1, \exp\left(\mathcal{S}(U_i) - \mathcal{S}(U_j)\right)] \\ \mathcal{S}(U_{curr}) &\to \mathcal{S}(U_{curr}) + \delta \end{aligned} \tag{3.52}$$

(in practice, a discrete histogram was used to store the estimates of $\mathcal{S}(U)$). By this virtue, the initially frequent moves to high-entropy states decline over time as the corresponding estimates of $\mathcal{S}(U)$ increase.

A very similar idea materialized itself in the concept of conformational flooding introduced by Grubmüller [249], who suggested adding a biasing potential in the form of (multivariate) Gaussians to accelerate the escape from free energy wells; this time-dependent bias would then affect the dynamics of the system just as any other external potential in e.g. steered MD runs. Eventually, the method resurfaced again under the name of metadynamics thanks to Laio and Parinello [250], and has gained publicity ever since.

Among subsequent developments, the well-tempered variant seems to be the most commonly used [251]. It was introduced to solve common shortcomings of the basic scheme: (a) "oscillating" convergence due to slow response of the system to bias piling up in one region of the CV, as well as (b) pushing the system out of realistic energy ranges due to bias accumulating indefinitely. In the well-tempered scheme, height of the deposited Gaussians decreases in relation to the locally accumulated bias, so that the total bias converges to the negative free energy profile scaled by a constant factor. As a result, in the limit of infinite time the histogram of counts does not actually become flat, but all free energy barriers are scaled down by a pre-defined factor, set by the user depending on the expected roughness of the free energy profile.

Metadynamics is often used as a method of choice to explore higher-dimensional free energy profiles, as in such cases US suffers strongly from the combinatorial explosion (curse of dimensionality). In contrast, the extension of metadynamics to higher dimensions is straightforward, as multi-dimensional Gaussian biases are used instead of single-dimensional ones. It is also convenient to augment metadynamics with so-called multiple walkers – a set of parallel runs that all "feel" and contribute to a mutual biasing potential – to simultaneously explore multiple free energy minima that can be encountered in higher dimensions. Finally, metadynamics coupled with replica exchange can be used as an enhanced sampling tool, with a set of arbitrarily biased simulations exploring the conformational space in an accelerated manner, and a single unbiased replica coupled by exchanges accepting the conformations according to – and hence recovering – the regular Boltzmann distribution [252].

3.2.4 Markovian Modelling of Dynamic Phenomena

The Markovian Property

The term "Markovian" is being used heavily in the field in recent years, reflecting the proliferation of tools that enable researchers to analyze their trajectories by the discretization of configuration space. But the real meaning of this word refers to a simple property of memorylessness, meaning that the future evolution of the system is only dependent on the present state and not its history. In light of this statement, classical mechanics is also Markovian: given the laws of physics, the knowledge of positions and momenta of particles is sufficient to predict how the system will evolve in time. What is usually referred to as a Markov process in molecular simulations is, however, a stochastic process in a discrete set of states rather than a deterministic one in a continuous phase space: the jump from the current state to a future state is hence performed based on a predetermined set of conditional probabilities, $T_{ij} = p(j_{t+\tau}|i_t)$, that determine the chance of arriving at state j after a time interval τ from starting in state i. (I will not discuss continuous-time Markov models here, as they are rarely used in practice.) The associated stochastic matrix $T(\tau)$ is the central representation of the model, and given a starting distribution $\mathbf{p}_t$ is capable of generating future distributions through matrix multiplication, $\mathbf{p}_{t+\tau} = T\mathbf{p}_t$.

There are several caveats associated with the construction of a Markov state model (MSM): firstly, how does one discretize the phase space? Intuitively, through "infinite" discretization one recovers the strictly Markovian continuous case; on the other hand, a coarse discretization provides more robust statistics when model parameters such as transition frequencies are estimated. An optimal choice has therefore to balance these two sources of error in a way that minimizes their sum. The same is true for the selection of a lag time τ – with a small time step one will likely violate Markovianity because the system will fail to completely decorrelate from its history within the discrete state (i.e. might be stuck within a "substate" not resolved by the discretization), but an infinitely large τ will yield transition probabilities equal to the equilibrium distribution since the system will have the time to equilibrate in between t and $t + \tau$. Here, the common practice is to use τ as small as possible to preserve the Markovian property, but not smaller. The method of discretization is also an important component: standard workflows involve the extraction of structural descriptors (features) from trajectories, then dimensionality reduction (projection) into a lower-dimensional subspace that retains most of the original variance, and finally clustering to obtain discrete states. However, each stage can be realized in many different ways (different sets of features, different methods and parameters of dimensionality reduction, different clustering algorithms) so that the parameter space in which the procedure could be optimized becomes prohibitively large for direct optimization. In recent years an exact variational principle for hyperparameter optimization and cross-validation was developed [253], but the cross-validation procedure still can become computationally demanding, and in case of MSMs not all data sets can be conveniently split into a teaching and training set. As a result, individual models are often optimized and designed based on intuition and heuristics, in particular when training is time- and memory-consuming; this is often the case as the construction of an N-state model involves the diagonalization of a N×N stochastic matrix, whereas the time required for diagonalization scales as $\mathcal{O}(N^3)$.

The Algebra and Estimation of Stochastic Matrices

Several unique properties of stochastic matrices make them stand out among other matrices. Due to their physical interpretation, they need to be row-normalized (the i-th row lists all probabilities of getting to any final state j from i, including $i = j$), so that $\sum_j T_{ij} = 1$, and have strictly non-negative entries. As a result, all well-behaved stochastic matrices have a single largest eigenvalue of 1, and all remaining eigenvalues fall within the interval (0, 1). (The term well-behaved as used here is a technicality, however it is possible to write down stochastic matrices that do not converge to a stationary distribution; one such matrix would be the antidiagonal 2×2 unit matrix that represents a system constantly jumping between its two states.) The first eigenvector, i.e. the one corresponding to $\lambda_1 = 1$, represents the relaxation of the system in the limit of infinite time, and hence describes the equilibrium distribution π; it can be seen immediately by writing the eigenvalue problem as $T\boldsymbol{\pi} = \boldsymbol{\pi}$. The subsequent eigenvectors can be interpreted as orthogonal transitions between subsets of states, characterized by relaxation (also called implied) timescales given by $-\frac{\tau}{\log(\lambda_i)}$. This formula implies that the closer the eigenvalue approaches unity, the slower the relaxation timescale (hence the limit of infinite relaxation timescale for $\lambda_1 = 1$).

In recent years, the emphasis in the proper construction of MSMs shifted from maximizing the eigenvalues of the stochastic matrix (and, correspondingly, the implied timescales) to finding a good approximation of the continuous transfer operator [254], i.e. the exact classical propagator that calculates the probability density of finding the system in a region of the configuration space y at a time $t + \tau$ [255]:

$$\rho_{t+\tau}(\mathbf{y}) = \mathcal{Q}(\tau)\rho_t(\mathbf{y}) = \int d\mathbf{x} p(\mathbf{y}, t+\tau|\mathbf{x}, t)\rho_t(\mathbf{x}) \tag{3.53}$$

Since in the MSM framework the fast modes of motion are of little relevance, it is now widely assumed the model should attempt to provide a discrete approximation to the projection of the transfer operator on the m selected slowest eigenvectors; observations were hence made on model, analytically tractable systems in order to gain insights into recommendable practices in MSM construction [254]. Because discretization is most efficient in coordinate subspaces that already correspond to slowly decorrelating degrees of freedom, the introduction of time lagged-independent component analysis (tICA)-based dimensionality reduction as an intermediate step in model construction significantly improved the quality of produced MSMs.

A separate issue is, however, that of MSM estimation from simulation data. In naive way, one could iterate over all discretized trajectories and directly count all $i \rightarrow j$ transitions using a time step τ, and then row-normalize the obtained count matrix C to obtain a valid stochastic matrix T. (Notably, if the dynamics of the system is time reversible and global equilibrium can be assumed, the count matrix C should be symmetric, as the observation of all transition events in the reverse order is just as likely; due to the normalization, this does not imply the symmetry of T.) Nevertheless, due to finite sampling this method would yield stochastic matrices that do not preserve the detailed balance condition, $\pi_i T_{ij} = \pi_j T_{ji}$, crucial to define local equilibrium between states. For this reason, most approaches to MSM estimation now rely on maximum likelihood (ML) methods, in which the formula for the likelihood of generating data given a stochastic matrix, $p(\mathbf{X}|T)$, is written down and used to maximize the (log-) likelihood of the model. It was shown that in order to obtain an MSM that satisfies detailed balance, it suffices (even though is not the most computationally efficient) to write down the likelihood as [256]:

$$\mathcal{L} = \prod_{i,j} \left(\frac{X_{ij}}{X_i} \right)^{C_{ij}} \tag{3.54}$$

where the matrix X is a new estimate of the count matrix C, and the term X_i is the (unnormalized) probability of being in state i, $X_i = \sum_j X_{ij}$.

Since then, several ML schemes have been proposed that build upon this idea to integrate additional data into MSM estimation. For instance, DHAMed incorporates both the detailed balance condition and state counts to arrive at the likelihood in the following form [247]:

$$\mathcal{L} = \prod_{i<j} (T_{ij})^{C_{ij}} \left(\frac{T_{ij}\pi_j}{\pi_i} \right)^{C_{ij}} \prod_k (T_{kk})^{C_{kk}} \tag{3.55}$$

while the reportedly superior TRAM approach introduced recently by Noé allows to combine data from multiple (e.g. biased) ensembles into a single multiensemble model by casting the likelihood in the form [257]:

$$\mathcal{L} = \prod_k \left(\prod_{i,j} (T^k_{ij})^{C^k_{ij}} \right) \left(\prod_i \prod_x \mu(x) \exp\left(f^k_i - b^k(x) \right) \right) \tag{3.56}$$

where k enumerates all ensembles in which simulations were performed, $\mu(x)$ are frame weights in the unbiased (reference) ensemble, and f^k_i and $b^k(x)$ are free energies and bias energies related to the biasing potential in biased simulation k, respectively. This development of integrative approaches that attempt to simultaneously account for multiple sources of data is representative of a broader trend that can recently be observed in the literature [258, 259]

Insights from Markov State Models

MSMs are exceptional tools in that they provide both a thermodynamic and kinetic description of the process in question: equilibrium probabilities allow to assign a free energy to each state, while specific formulas enable the calculation of e.g. mean first passage times (MFPTs) between sets or states. MFPT from state i to j (also known as the inverse rate constant $(k_{ij})^{-1}$) is obtained from a set of coupled equations that consider all pathways that can be initialized from i, weighting them by the probability of the first step:

$$\mathrm{MFPT}_{ij} = \sum_k T_{ik}(\tau + \mathrm{MFPT}_{kj}) \tag{3.57}$$

If individual states can be resolved structurally, additional algorithms such as the Perron Cluster Cluster Analysis (PCCA) and its variants [260] are available to aid in visualization of the results, providing robust tools that render the model easily comprehensible, even if prone to oversimplification.

These features of MSMs have earned them a considerable amount of popularity in recent years, fueled in parallel by the rapid development of two automatized Python-based platforms for the construction and analysis of models from any type of simulation data, PyEMMA and MSMBuilder [261, 262]. In this way, the numerous recent theoretical developments are quickly implemented and disseminated within the field, accounting for a large part of the scheme's popularity. Notably, MSMs have been widely used to analyze protein folding [263] and ligand binding [257] – multipathway processes that are notoriously difficult to describe with a single reaction coordinate – as well as less complex problems such as ion translocation through a membrane protein [264].

Within the framework of Markov models it is also possible to analyze so-called hidden processes, i.e. random processes that are not observed directly but whose time evolution is implied based on other observables. The fundamental idea here is that the hidden process, defined over the hidden states $X_1, X_2, ...$, is Markovian, and that every hidden state emits observables according to its own probability distribution, $p(y|X_n)$. It is then possible to write the total probability of the realization of a single process in terms of a product of probabilities of transitions between the hidden states and the corresponding emission probabilities, and use e.g. the maximum likelihood approach to assign the hidden states to specific time intervals of the simulation. This scheme, while more often used in experimental settings such as FRET experiments, can also be used to analyze molecular simulations in order to e.g. infer the existence or quantify populations of conformational states within a molecule [265].

3.2.5 Alchemical Transformations

Linear Interpolation between Chemical States

A well-known shortcoming of classical molecular dynamics simulations is that they require a fixed topology, i.e. are not suitable for the modelling of chemical reactions that typically involve the formation or breaking of chemical bonds. While the reactive force fields are being developed to address this issue within the classical framework [266], they are rarely used in practice due to the necessity of empirical parametrization, as well as limited number of implementations. Apart from a fully QM or QM/MM treatment, the problem of morphing chemically distinct species into each other is hence addressed through so-called alchemical transformations.

In alchemical transformations, two different molecule topologies – generically referred to as "state A" and "state B" – are simultaneously defined, giving rise to two distinct Hamiltonians, $\mathcal{H}_A$ and $\mathcal{H}_B$. To connect the two species in a continuous way, the Hamiltonian at any intermediate point of the transformation denoted by $\lambda \in (0, 1)$ is constructed as a linear interpolation:

$$\mathcal{H}(\lambda) = (1-\lambda)\mathcal{H}_A + \lambda\mathcal{H}_B \tag{3.58}$$

As a result, all properties of the molecule – bond lengths, reference angles and dihedrals as well as their force constants, particle masses, charges and effective radii – are interpolated linearly (which is not a strict requirement, but rather a common practice) from their respective values in states A and B. Often, though, it is not sufficient to simply change the properties of individual atoms to mutate one residue into another, and for this reason alchemical transformations typically involve the creation or deletion of atoms. This can prove burdensome as such disappearing atoms are morphed into virtually non-interacting dummies that might introduce numerical instabilities near the endpoints, i.e. when the atom almost ceases to exist. The reason for that is that numerical integration relies on finite-length steps, and within one such step the algorithm can easily go from a region where forces are negligible to a near-singularity, particularly when charges are still present on the disappearing particle.

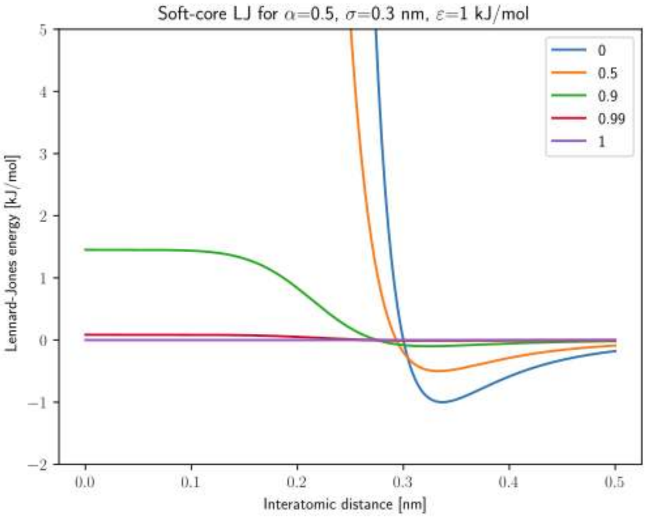


FIGURE 3.2: The soft-core Lennard-Jones potential for a disappearing particle at different values of λ.

Traditionally, this issue was addressed by splitting the alchemical process into stages in which first the charges on disappearing atoms are scaled to zero, then the steric properties are modified, and finally charges reappear on newly formed atoms. However, a more convenient means to circumvent this problem is the introduction of soft-core potentials – modified Lennard-Jones and Coulombic interaction terms that avoid the singularity at integer values of λ. Their functional form is designed so as to make the r_{12}-dependence of the interaction term converge to a small fixed value for $r_{12} \to 0$ at λ close to the boundary value, and get rid of the r_{12} dependence altogether at integer λ; the design of such schemes is facilitated by the requirement that only the end-point, chemically relevant states be described with a realistic potential. Correspondingly, for an atom bound to disappear at $\lambda = 1$ the general modified Lennard-Jones term is usually defined as follows ([267], see Fig. 3.2):

$$V_{LJ} = 4\varepsilon(1-\lambda)\left[\left(\alpha_{sc}\lambda^2 + \left(\frac{r_{12}^6}{\sigma_{ij}^6}\right)^{-2}\right) - \left(\alpha_{sc}\lambda^2 + \left(\frac{r_{12}^6}{\sigma_{ij}^6}\right)^{-1}\right)\right] \tag{3.59}$$

A separate issue is that of formation and disappearance of bonds, in particular when new rings are created or opened. In principle, such transformations can be carried out if one can afford adequate sampling of the intermediate states, with only minor issues resulting from the ill-defined 1-4 scaling in such cases. However, in such cases a popular choice is the alternative dual-topology approach in which entire molecules or residues disappear or come into existence, the respective appearing/disappearing parts do not interact with each other and are often connected by harmonic bonds to ensure that they sample similar regions of the conformational space.

Free Energies from Alchemistry

Since a force field provides a complete analytic formula for $\mathcal{H}_A$ and $\mathcal{H}_B$, propagating the system at any $\mathcal{H}(\lambda)$ is trivial. Even more importantly, the properties of the Hamiltonian can be easily connected to free energies in a way analogous to eqn. 3.44:

$$\begin{aligned}\frac{\partial G}{\partial \lambda} &= \frac{\partial}{\partial \lambda}\left[-k_BT\log\int d\mathbf{x}\exp\left(\frac{-\mathcal{H}(\mathbf{x},\lambda)}{k_BT}\right)\right] \\ &= -k_BT\frac{\frac{\partial}{\partial\lambda}\int d\mathbf{x}\exp\left(\frac{-\mathcal{H}(\mathbf{x},\lambda)}{k_BT}\right)}{\int d\mathbf{x}\exp\left(\frac{-\mathcal{H}(\mathbf{x},\lambda)}{k_BT}\right)} \\ &= -k_BT\frac{-k_BT\int d\mathbf{x}\frac{\partial\mathcal{H}}{\partial\lambda}\exp\left(\frac{-\mathcal{H}(\mathbf{x},\lambda)}{k_BT}\right)}{\int d\mathbf{x}\exp\left(\frac{-\mathcal{H}(\mathbf{x},\lambda)}{k_BT}\right)} = \left\langle\frac{\partial\mathcal{H}}{\partial\lambda}\right\rangle\end{aligned} \tag{3.60}$$

so that the change in free energy, ΔG, can be obtained through estimation of the ensemble averages of $\frac{\partial\mathcal{H}}{\partial\lambda}$ at constant values of λ, and subsequent numerical integration. The following equation:

$$\Delta G = \int_0^1 d\lambda' \left\langle\frac{\partial\mathcal{H}}{\partial\lambda}\right\rangle_{\lambda=\lambda'} \tag{3.61}$$

defines the thermodynamic integration (TI) approach to alchemical transformations. This quantity, however, can be used directly in a limited number of cases, as it mixes several contributions pertinent to solvation, environment effects and internal degrees of freedom, at the same time mistreating the chemical energy term associated with purely quantum electronic effects such as bonding or aromaticity. Most often only relative free energies are of interest: a simple case would be the determination of pK$_a$ shift resulting from a change of environment. If an amino acid's pK$_a$ was measured in an aqueous phase, the calculated difference in protonation free energy between water and the chosen environment (e.g. protein interior or lipid bilayer) can be used to offset the experimentally measured value as a reasonable estimate of the new pK$_a$. In more complex cases, it is possible to indirectly calculate e.g. the difference in binding affinity between two ligands or two receptor conformations through the construction of an appropriate thermodynamic cycle (see Fig. 3.3). While such protocols require that two alchemical systems be simulated in parallel, the computational and technical (e.g. related to charge neutrality) burden of these simulations can often be alleviated by performing two alchemical transitions in opposite directions in a single simulation box.

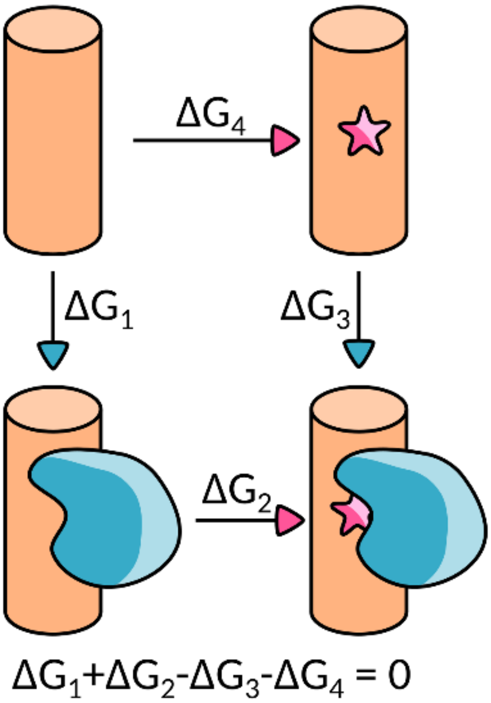


FIGURE 3.3: The concept of a thermodynamic cycle illustrated using a model protein-DNA system: by leveraging the state function property (i.e. the fact that values along a cycle sum up to 0), one can estimate the change in binding affinity ($\Delta\Delta G = \Delta G_3 - \Delta G_1$) by calculating two seemingly unrelated quantities, ΔG_2 and ΔG_4.

Regarding the estimation of ΔG, an often-used alternative to the TI scheme is the Bennett acceptance ratio (BAR [268]) or its more recent multistate version (mBAR [269]), both based on the fundamental free energy perturbation equation due to Zwanzig [270]:

$$\Delta G_{A \to B} = -k_B T \log \left\langle \exp\left(-\frac{\mathcal{H}_B - \mathcal{H}_A}{k_B T} \right) \right\rangle_A \tag{3.62}$$

In BAR, the ensemble-averaged differences between the total energies for swapped configurations at two neighboring λ values is used to estimate the difference in free energy between the corresponding ensembles, and then the resulting free energy differences are added together to yield the total ΔG of the transformation. In contrast, mBAR aggregates data about the total energy of each configuration in each ensemble, yielding per-frame weights that can be used to reconstruct free energies in the manner of choice.

3.3 Data Analysis Tools

3.3.1 Dimensionality Reduction in Molecular Simulations

A significant drawback of the phase space-based frameworks of classical mechanics is that they invariably rely on extremely high-dimensional spaces that human minds cannot explore but through the eye of mathematical formalisms. Nevertheless, key concepts of statistical mechanics are often taught using simplified two- or one-dimensional representations, and many mental images used by scientists to interpret or analyze results necessarily conform to this tendency. For this reason, methods used routinely to project high-dimensional data onto lower-dimensional, visually interpretable subspaces comprise an important tool in a computational biophysicist's toolbox. Below I briefly recapitulate the key ideas used to construct such tools, and key caveats or issues associated with their application.

Principal Component Analysis

Perhaps the most popular and widespread approach to dimensionality reduction in the analysis of high-dimensional data is the principal component analysis (PCA). The underlying concept is brilliantly simple: based on the distribution of the original n-dimensional data, one performs a rotation of the coordinate system and ranks the coordinates so as to obtain new orthogonal directions (termed principal components) along which the distribution is characterized by the highest variance. As a result, by keeping only a 2-dimensional subspace of the original space, we obtain an interpretable data set while retaining as much variability as possible in a linear framework (i.e. one in which the new coordinates are just linear combinations of the original ones).

From a technical standpoint, the selection of new coordinates in PCA is equivalent to the algebraic procedure of diagonalization of the covariance matrix. In turn, covariance between two variables X_i and X_j can be estimated from values these variables assume:

$$cov(X_i, X_j) = \frac{1}{N}\sum_{k=1}^{N}(x_{ik} - \bar{x}_i)(x_{jk} - \bar{y}_j) \tag{3.63}$$

reducing to regular variance, $cov(X_i, X_i) = \sigma_i^2$, at the diagonal of the matrix. Covariance is bound from below and above by $-\sigma_i\sigma_j$ and $\sigma_i\sigma_j$, values that would correspond to perfect correlation and anti-correlation, respectively, and is indeed related to the determination coefficient R^2 through normalization by the factor $\sigma_i\sigma_j$. Due to the definition of covariance, the covariance matrix can be easily obtained by matrix multiplication of the matrix of residuals with its own transpose and division by N.

Once the covariance matrix is assembled and diagonalized, its eigenvectors are the new basis vectors expressed in the original coordinate system, and its eigenvalues report the variances along the corresponding new coordinates. Since variances of uncorrelated random variables are additive, by restricting the data to k highest-ranking dimensions one recovers $\frac{\sigma_1^2+...+\sigma_k^2}{\sigma_1^2+...+\sigma_n^2}$ of the total variance, which provides a good diagnostics of how realistic the projection is.

In the new coordinate system spanned by the k best (normalized) eigenvectors $\mathbf{u}_1, ..., \mathbf{u}_k$, the coordinates of an original data point $\mathbf{c}$ is given by $\mathbf{c}' = (\mathbf{c} \cdot \mathbf{u}_1, ..., \mathbf{c} \cdot \mathbf{u}_k)$, where the dot denotes a scalar product between the vectors. In this way, the original data can be projected on the new coordinates, hopefully providing physically meaningful insights into the correlated large-scale changes that occur in the system of interest.

It is instructive to compare PCA to the somewhat similar normal mode analysis (NMA). While PCA is data-driven, identifying high-variance dimensions in multidimensional data, NMA relies on the diagonalization of the Hessian matrix (containing second-order energy derivatives, $\frac{\partial^2 E}{\partial x_i \partial x_j}$) and hence provides insight into locally accessible modes of motion surrounding a potential energy minimum. While PCA and NMA will tend to identify similar modes in case of Gaussian-shaped unimodal distributions, PCA will outperform NMA in case of multiple free energy basins, although at a cost of additional sampling of the conformational space.

Time-lagged Independent Component Analysis

In the analysis of molecular simulations, PCA will occasionally fail to identify the most relevant modes e.g. due to the presence of flexible elements that results in large correlated fluctuations preferentially picked by the algorithm. For this reason, a modification of this scheme was proposed recently with the introduction of the time-lagged independent component analysis (tICA) [271, 272]. Here, instead of maximizing the variance one looks for a set of orthogonal coordinates that yield the highest normalized autocovariance at a chosen time τ, approximating the eigenvectors of the transfer operator similarly to the case of Markov state models. The eigenvectors of tICA are obtained by solving the generalized eigenvalue problem:

$$C(\tau)\mathbf{u} = C(0)\mathbf{u}\lambda \tag{3.64}$$

where $C(\tau)$ is the time-lagged covariance matrix formed from ensemble-averaged products of residuals, $C_{ij}(\tau) = \langle x_i(t) x_j(t+\tau) \rangle$.

The tICA approach became instantaneously popular within the MSM community, as it was shown to provide an optimal approximation to the transfer operator in the linear regime, thus yielding low-dimensional data well-suited for further processing within the MSM framework. Correspondingly, it fared better than other pre-processing methods when applied to a real-life system such as conformational dynamics of a small protein [272], also proving robust to the selection of the hyperparameter τ. On the other hand, the most obvious limitation of tICA is its linear character: the new components cannot follow non-linear trends in original data unless non-linearity is introduced explicitly by adding extra components to the data set. In many cases involving non-trivial symmetries, such as the helical pseudo-symmetry of dsDNA, it is also unclear how to select input features so as to preserve these symmetries in the resulting model.

Linear Discriminant Analysis

While PCA and tICA are often the tools of choice in the analysis of single systems, they are not always well-suited to analyze and visualize categorical data, i.e. data sets that are split into subclasses based on arbitrary criteria. Here the advantages of yet another space-transforming method, the so-called linear discriminant analysis (LDA), can be leveraged. As a dimensionality reduction tool, LDA attempts to maximize inter-group separation to optimally resolve the subclasses in a lower-dimensional subspace. Although reliant on a strongly simplifying assumption of equal within-class covariance matrices, LDA has been shown to fare well in dimensionality reduction tasks even when it failed to perform well as a classification tool.

Algebraically, LDA is also performed by solving an eigenvalue problem [273]. Prior to that, however, the within-class scatter matrix S_W and the between-class scatter matrix S_B have to be computed. The former is merely a class population-weighted sum of covariance matrices calculated for individual classes; the latter is a similarly weighted product of a vector of mean residuals (the differences between the within-class means and the overall means) with its own transpose. The resulting matrix $S_W^{-1} S_B$ is then diagonalized to yield eigenvalues – providing information about the quality of class separation under the assumptions listed above – and eigenvectors that give the new components. The analysis then follows as outlined previously for PCA and tICA.

Deep Learning Models

Whereas the recent deep learning revolution seems to have had relatively little impact on the field of molecular simulations so far, dimensionality reduction is one area that drew more attention in that regard. Although machine learning concepts such as self-organizing maps or stochastic neighbor embeddings have been around for a while, the growing popularity of deep autoencoders sparked novel interest among computational biophysicists – in part due to their novelty itself, in part as a result of their real advantages. Indeed, while all three approaches discussed in detail above only rely on linear transformations, deep learning models are inherently non-linear and thus offer much greater elasticity, possibly encoding more information in the resulting two-dimensional plots [274]. Concerns regarding the interpretability of the models have also been largely addressed with the introduction of saliency maps [275]. In line with the previously described developments in the field of linear transformations, time-lagged autoencoders have also been introduced to aid in data processing aimed at the construction of MSMs [276]. Perhaps the most appealing feature of autoencoders is the ease with which any desired property of the embedding can be preserved through appropriate design of the loss function, as compared with the tedious derivations of the respective linear transformations. Taken altogether, these unique advantages ensure that deep learning models will experience a rapid development in coming years.

3.3.2 Information Theory

Many regression- or correlation-based approaches to data analysis suffer from their inherent assumptions about the linear or multilinear relationships between individual data features, as well as the normality (or at least unimodality) of feature distributions. In fact, non-linearity and multimodality are often introduced into statistical models in a *post-hoc* fashion, frequently leading to overfitting if models are not properly cross-validated on separate test data sets.

This problem can be partially alleviated by the use of information theory-based statistical tools, as they usually do not rely on particular functional forms of the underlying processes but rather on the idea of *information flow*. Information theory builds on the notion of an abstract encoder-decoder system, with the former attempting to convey a message through a (possibly lossy) channel and the latter attempting to faithfully reconstruct the original message from the received data. The theory provides basic tools to quantify, e.g., data redundancy, information loss/gain or information transfer, randomness and statistical independence.

Information Entropy

The central quantity of information theory is entropy. Often dubbed "the measure of randomness", it is actually better understood as a measure of information: each incoming signal resolves a certain degree of uncertainty, providing us with new data about the process in question. Limiting the didactic example to a simple case, we can note that a coin toss is most "informative" – in the framework of the theory – if we use a perfectly balanced coin: each result provides us with new data point that could not have been foreseen in advance. At the same time, tossing a perfectly biased coin (one that always lands on the same side) does not yield any information: the result is entirely predictable even before the signal arrives.

This intuitively means that the less probable an event x is, the more informative its occurrence becomes. One can call the quantity $-\log_2 p(X = x)$ the *information content* of event x, and then the expectation value of information content (given a set of possible events $\mathbb{X}$) can be written as:

$$H = \mathbb{E}[-\log_2 p(X = x)] = -\sum_{x \in \mathbb{X}} p(X = x) \log_2 p(X = x) \tag{3.65}$$

which is exactly the information-theoretical definition of (Shannon) *entropy*.

(Note on notation: by convention, capital letters correspond to random variables or random processes, while minuscules represent individual outcomes of the process. Also, in further derivations the base of the logarithm will be dropped, as the choice of individual bases only changes the units of entropy/information: bits for base 2, nats for base e, dits for base 10.)

A quick sanity check reveals that for the coin toss example, setting $p(heads) = p(tails) = \frac{1}{2}$ yields $H = 1$ (high entropy), while $p(heads) = 0$ and $p(tails) = 1$ result in $H = 0$ (low entropy). It is also easy to show that H is maximized when both state probabilities are equal, and that this property in fact generalizes to arbitrarily many states; it is sufficient to show that for any two non-equal state probabilities p_1

and p_2 chosen so that $p_2 > p_1$, decreasing p_2 by an arbitrarily small ε and increasing p_1 by the same amount always results in an increase in entropy:

$$\begin{aligned}\Delta H &= -p_1 \log \frac{p_1+\varepsilon}{p_1} - p_2 \log \frac{p_2-\varepsilon}{p_2} + \varepsilon[\log(p_2-\varepsilon) - \log(p_1+\varepsilon)] \\ &\approx -p_1 \frac{\varepsilon}{p_1} + p_2 \frac{\varepsilon}{p_2} + \varepsilon \left(-\frac{\varepsilon}{p_2} + \log p_2 - \frac{\varepsilon}{p_1} - \log p_1 \right) \\ &= \varepsilon \log \frac{p_2}{p_1} + \mathcal{O}(\varepsilon^2) > 0 \text{ for small } \varepsilon\end{aligned} \tag{3.66}$$

The transition between the first and second line was made using the fact that for small q, $\log(1+q) \approx q$.

Entropy-based Measures of Pairwise Correlation

However, in real applications one hardly ever deals with single random variables, and it is in the analysis of multi-dimensional data where information theory begins to provide a real advantage. A natural question to ask is whether two random processes X and Y are correlated in some (not necessarily linear) way, and, more specifically, how much uncertainty remains in the result of Y if one already knows the value of X. To reframe the question it in mathematical terms – what is the expected information gain from Y if X is known to be equal to x_0:

$$\begin{aligned}H(Y|X=x_0) &= \mathbb{E}[-\log p(Y|X=x_0)] \\ &= -\sum_{y\in\mathbb{Y}} p(Y=y|X=x_0) \log p(Y=y|X=x_0)\end{aligned} \tag{3.67}$$

By averaging over all possible values of the conditioning variable X and using the definition of conditionality $p(X=x)p(Y=y|X=x) = p(X=x, Y=y)$, we get the expression for *conditional entropy* $H(Y|X)$:

$$\begin{aligned}H(Y|X) &= \sum_{x_0\in\mathbb{X}} H(Y|X=x_0) \\ &= -\sum_{x_0\in\mathbb{X}} \sum_{y\in\mathbb{Y}} p(X=x_0)p(Y=y|X=x_0) \log p(Y=y|X=x_0) \\ &= -\sum_{x\in\mathbb{X}} \sum_{y\in\mathbb{Y}} p(x,y) \log \frac{p(x,y)}{p(x)}\end{aligned} \tag{3.68}$$

where the shorthand notation $p(x)$ was used for brevity to denote $p(X=x)$.

A similar but even more useful quantity is the *mutual information* (I) between two variables. It quantifies how much additional information is captured in the joint distribution, $p(x,y)$, as compared to a combination of marginal distributions, $p(x)$ and $p(y)$. If two variables are independent, by definition $p(x,y) = p(x)p(y)$ and hence the information content is identical in both cases, as reflected by a $I(X;Y)$ of 0. A positive value of I is indicative of an interdependence between variables;

note that I cannot be negative, as two marginal distributions cannot contain more information that a joint distribution (the former can be trivially obtained from the latter).

The above definition – the amount of information co-encoded simultaneously in the two variables – makes it possible to view mutual information as the difference between the estimated information content of X, the entropy $H(X)$, and the excess information encoded in X when Y is known, the conditional entropy $H(X|Y)$:

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= -\sum_{x\in\mathbb{X}} p(x)\log p(x) + \sum_{x\in\mathbb{X}}\sum_{y\in\mathbb{Y}} p(x,y)\log\frac{p(x,y)}{p(y)} \\ &= \sum_{x\in\mathbb{X}}\sum_{y\in\mathbb{Y}} p(x,y)\log\frac{1}{p(x)} + \sum_{x\in\mathbb{X}}\sum_{y\in\mathbb{Y}} p(x,y)\log\frac{p(x,y)}{p(y)} \\ &= \sum_{x\in\mathbb{X}}\sum_{y\in\mathbb{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \end{aligned} \tag{3.69}$$

The formula reveals the permutational symmetry of mutual information that was already evident from the above description: $I(X;Y) = I(Y;X)$. It also becomes evident that mutual information can be viewed as the Kullback-Leibler (KL) divergence – a popular statistical measure of dissimilarity – between the joint distribution and the product of marginal distributions. It should be re-emphasized that the quantity makes no assumptions regarding the functional form of the correlation between X and Y, and if one is strictly a function of the other, this fact will be reflected in $I(X;Y)$ being equal to $H(X)$. Here, the main source of error lies in the representation of the marginal and joint probability distributions: they can either be represented as (sums of) analytic functions with parameters fitted to empirical data, or discretized in the form of 1D and 2D histograms. In some fields such as bioinformatics or digital image analysis, though, data is inherently discrete and can be used with little to no processing. For the above reasons, mutual information is often preferentially employed to analyze genomic or systems biology data [277, 278].

Transfer Entropy

Conceptually, all the quantities introduced above can be related to simple set operations such as intersections and differences, as illustrated in 3.4, and hence it is natural to use them as building blocks for the design of more complex descriptors tailored to specific needs.

One such need, poorly addressed by the existing methodology, pertains to the description of dynamic phenomena in terms of causal relationships. Mutual information does not address this issue as (a) it typically only relies on instantaneous correlations and (b) its symmetry does not allow to distinguish causes from effects. Since the idea of causality is difficult to define formally in dynamic multi-body systems, a useful proxy is the time-delayed correlation. It is however important to note that two things that correlate in a time-delayed manner need not be linked by a causal relationship, as both can have a common cause that manifests itself with different delays.

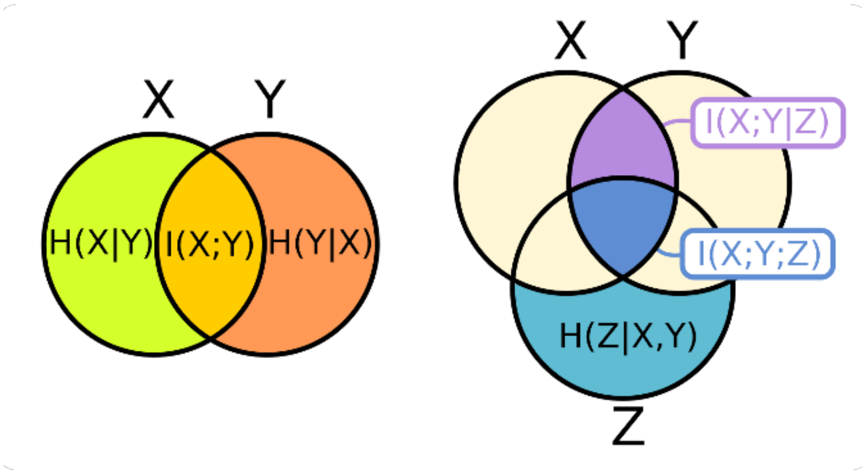


FIGURE 3.4: A Venn diagram conceptually illustrating information theory-based measures of correlation described in this section: conditional entropy $H(X|Y)$, mutual information $I(X;Y)$ and conditional mutual information $I(X;Y|Z)$.

To proceed, we first need to define *conditional mutual information* in the spirit of definitions provided above: such a quantity, $I(X;Y|Z)$ indicates how much additional information about X can be extracted from Y if all information stored in Z is simultaneously available. Derivation proceeds analogously to eqns. 3.68 and 3.69, and the final formula reads:

$$I(X;Y|Z) = \sum_{x\in\mathbb{X}} \sum_{y\in\mathbb{Y}} \sum_{z\in\mathbb{Z}} p(x,y,z) \log \frac{p(x|y,z)}{p(x|z)} \tag{3.70}$$

Finally, one can use conditional mutual information to quantify how much additional information about the *future evolution* of X can be extracted from the history of Y if all information about the history of X is available. This quantity, introduced only recently to the field of molecular simulations [279], $T_{Y\to X}$, is called the *transfer entropy* from Y to X, and is defined as:

$$T_{Y\to X} = I(X_n; Y_{n-k:n-1}|X_{n-k:n-1}) \tag{3.71}$$

where the notation $X_{n-k:n-1}$ corresponds to a history of values of X with a (variable) history length of k, and X_n is the current (as of n-th sampling event) value of X.

Transfer entropy is clearly an extensive variable, i.e. its value increases as more data is gathered, making it difficult to directly compare values obtained from individual experiments. However, it attains a maximum value if all information about the future of X missing from the history of X is encoded in the history of Y instead; in other words, the upper limit for the values of transfer entropy is the conditional entropy $H(X_n, X_{n-k:n-1})$. It is hence useful to define the normalized transfer entropy $\bar{T}_{Y\to X}$ as the quotient of transfer entropy from Y to X and conditional entropy of X. Furthermore, to facilitate the identification of direction of information transfer, an antisymmetric pairwise quantity – the *directional index* – can be defined as the difference of normalized transfer entropies in both directions:

$$D_{xy} = \frac{T_{Y \to X}}{H(X_n, X_{n-k:n-1})} - \frac{T_{X \to Y}}{H(Y_n, Y_{n-k:n-1})} \tag{3.72}$$

Several caveats need to be addressed regarding the newly defined quantity. Firstly, values of D_{xy} fall within the range $[-1;1]$ and can be used as a proxy for causality in an intuitive way: values will be positive if Y is the driving force behind changes in X, and negative if the converse is true. However, it is possible for both transfer entropies to be high and D_{xy} to be zero. Secondly, specific properties of the marginal distributions can cause transfer entropies to be relatively high even in absence of an actual time-lagged correlation; to correct for this effect in an *ad-hoc* way, it is customary to scramble the history record of the supposedly causal variable, re-calculate transfer entropy using the scrambled data and use this value (or an average of several such values) as a correcting factor to unbias our result. Moreover, while the history length k is a free parameter, it is almost unheard of to use k larger than 1 in practice: for instance, if one uses histogramming to estimate the joint distribution $p(X_n, Y_{n-k:n-1}, X_{n-k:n-1})$, the extension of history length by one increases the dimensionality of the histogram by 2 (one extra dimension per X and Y). In effect, very large data sets and/or very few discrete states would be required to obtain meaningful statistics. Finally, one can look at different time-lagged correlations by varying the time step τ that separates the $n-1$-th and n-th step; in practice, however, at long timescales the dynamics becomes more and more Markovian and the sought-after effect of inter-variable correlation might vanish.

3.4 Simulational methods used in the study

Setup of systems used to study recognition and binding of TRF1

All simulated models involving the TRF1 DBD (residues 379–430, capped on both termini) were based on the X-ray structure of the DBD bound to telomeric double-stranded (ds) DNA found in PDB entry 1W0T. All fully atomistic simulations employed a periodic, effectively infinite dsDNA model built using the ideal B-DNA parameters as implemented in the X3DNA package [280], with 20 base pairs corresponding to two full turns of B-DNA double helix. Such an approach has been successfully used by several groups so far [281–283], allowing to bypass common problems associated with the behavior of DNA termini and excessive elasticity of short DNA oligomers complexed with proteins. Due to a mismatch between the periodicity of telomeric 5-TTAGGG-3 tandem repeats and the helical pitch (10–10.5 bp), the periodic sequence (5-GGTTAGGGTTAGGGTTAGGG-3) consisted of three tandem repeats and two additional GC pairs. A native structure of the specific TRF1–DNA complex was obtained by superimposing phosphorus atoms in the X-ray structure with the artificially created 20-bp periodic model.

Details of simulations involving TRF1

For all free energy simulations, a cubic 6.62 nm × 6.62 nm × 6.62 nm box was used in which the protein–DNA complex was solvated with 8695 TIP3P water molecules. For spontaneous binding simulations, we employed a rectangular 6.5 nm 6.5 nm 6.62 nm box containing the protein, DNA and 8217 TIP3P water molecules. The number of K+ and Cl ions was adjusted to maintain a physiological salt concentration of 0.154 M and neutralize the net charge of the system. All simulations were performed in Gromacs 4.5 (free energy) or 5.0.4 (spontaneous binding) [284]. The Amber99sb-parmbsc0 force field was used [285], and temperature was maintained at 300 K using the stochastic velocity rescaling thermostat with a time constant of 0.1 ps. In order to use the z-coordinate as the reaction coordinate, in free energy simulations the z axis vector length was constrained to a fixed value using the semi-isotropic coupling scheme; besides that, pressure was maintained at 1 bar using the Berendsen barostat with a time constant of 2.0 ps. Particle Mesh Ewald (PME) summation was used for the calculation of electrostatic interactions, and van der Waals interactions were cut off at 1.0 nm.

DNA-binding affinity of TRF1 mutants

The umbrella sampling/WHAM approach was used for the calculation of free energy profiles in the radial direction, in analogy to our previous work [286]. The distance between DNA phosphorus atoms and core residues of the protein (12 residues closest to the protein COM during an equilibrium simulation) projected onto the XY-plane (r-distance) was used as the reaction coordinate. Initial frames for individual windows were generated from a 1-μs steered MD simulation in which the center of the restraining potential was changed at a constant velocity in the radial direction from the starting value of 1.55 up to 3.0 nm, with a force constant of 2500 kJ/mol nm^2. From this trajectory, 30 frames were extracted that corresponded to geometries in 0.05-nm intervals along the reaction coordinate. These geometries were then used to assess the effect of single amino acid mutations on the thermodynamics of specific and non-specific TRF1-DNA binding.

In the simulations, an inverse telomeric sequence (5-CCCTAA-3 repeats) was used as a model non-specific target, and in this case initial geometries for umbrella sampling were obtained by mutating all 40 DNA bases in the original 30 frames (extracted from steered MD trajectories) using the X3DNA package, as described below. Overall, a total of 12 free energy profiles were obtained for the wild-type protein and five mutants (R380A, V418A, K421A, D422A and R425A) with respect to the standard (5-TTAGGG-3 repeats) and inverse (5-CCCTAA-3 repeats) telomeric sequence. Amino acid mutations were introduced by simple deletion/renaming of existing atoms. The number of ions was then adjusted to ensure charge neutrality. All modifications described above were followed by energy minimization, and 500 ns simulations were carried out in each US window, yielding a total of 180 μs. 100 ns was discarded at the beginning of each trajectory in individual US windows to allow the systems to adjust to any introduced changes. Importantly, the use of a single steered MD trajectory results in desirable error cancellation, allowing us to capture relatively minor changes in the behavior of all systems considered with high sensitivity.

Free energy along the DNA major groove

The free energy along the major DNA groove (i.e. in close vicinity to the DNA) was calculated using the umbrella sampling (US)/weighted histogram analysis (WHAM) method [243, 244]. To generate initial frames for individual US windows along the DNA helix, a rotation-translation matrix was used to propagate the protein in 69 steps along a helical path about the main axis of the DNA helix, as defined by standard B-DNA geometry. This approach is different from the one used in the recent study by Marklund et al., where helical movement along the major groove was enforced by pulling in the helical direction [283], but similar to that of Furini et al. [281]. DNA bases in frames generated along the standard telomeric sequence (target, 5-GGGTTAGGGTTAGGGTTAGG-3) were then mutated using X3DNA to create a corresponding set of frames along the inverse telomeric sequence (model off-target, 5-CCCTAACCCTAACCCTAACC-3). After energy minimization, the PLUMED plugin [287] was used to restrain the protein in its initial position along the Z-axis with a force constant of 200 kJ/mol nm^2. This Z-coordinate was defined with respect to a single base pair not involved in protein binding (1.6 nm below the lowest US window) whose position in space was restrained in the Z-direction. In addition, one-sided harmonic potentials were added to prevent the COM of DNA from diffusing away in the XY-plane, in order to avoid periodic boundary artifacts. To ensure that the obtained free energy profile captures the effect of DNA sequence, spontaneous dissociation from non-native interfaces was prevented by adding a one-sided harmonic potential with a force constant of 500 kJ/mol nm^2 at protein–DNA COM XY-distance of 1.55 nm. For the purpose of subsequent analyses, a proper equilibrium distribution was recovered using a weighting factor of $\exp\left(\frac{U(r,z)F_i}{k_B T}\right)$, where $U(r,z)$ is the applied biasing potential and F_i is the free energy associated with the constraint in i-th window as calculated by the WHAM algorithm.

For both DNA orientations, a set of 750-ns simulations in each umbrella sampling window was ran. For the standard orientation, additional data from 1000-ns simulations performed with a larger force constant (500 kJ/mol nm2) that did not yield proper histogram overlap were also included in the construction of free energy maps and subsequent calculations. Hence, the total simulation time used to construct the profiles along the DNA was greater than 170 μs.

Spontaneous binding and spawning

To study spontaneous binding of TRF1 to telomeric DNA, 50 systems have been prepared in which the protein was placed randomly in the simulation box containing a periodic DNA molecule. All systems were solvated with identical number of ions and water molecules and, after energy minimization, 50 equilibrium simulations were ran from thus obtained geometries. 20 trajectories have been propagated for 4 μs each, and another 30 for 2 μs each, yielding a total of 140 μs. From the resulting trajectories, sampled at each 25 ns, a subset of 77 frames has been identified that captured geometries close to the native protein–DNA complex, and additional seventy seven 500-ns long simulations were ran starting from these frames (later referred to as 'spawning' simulations). Geometries were chosen based on an mRMSD criterion. The mRMSD parameter was defined so as to take into account the relative position of 10 phosphate atoms from the DNA backbone (5 bp at the protein–DNA interface) and 15 Cα atoms from the DNA-binding helix, indicative of the overall geometry of the native complex. Then, mRMSD was calculated as the lowest RMSD value for this subset of atoms with respect to any consecutive chain of five phosphate pairs among 20 possible alignments (in geometry corresponding to the reference 1W0T X-ray structure), since there are 20 possible sites at which the protein–DNA complex can be formed, or 40 if both orientations are possible. If mRMSD was lower than 0.175 nm, the respective frame was accepted as a starting point for the spawning simulations. Since the procedure was aimed at generating trajectories that bind in a sequence-specific manner, only the standard orientation of the DNA duplex (5-TTAGGG-3) and not the inverse sequence (5-CCCTAA-3) was considered when applying the criterion. By this virtue, the original 50 trajectories have equal *a priori* probabilities of binding in either orientation, while the spawning trajectories are strongly biased towards the standard one.

Prevalence of acidic residues in base readout

In the Protein Data Bank (PDB), 3891 protein-DNA complexes have been identified and downloaded. Using custom scripts and the MDTraj Python library [288], we selected structures containing amino acid side chains h-bonded to individual nucleobases in duplex DNA. Then, statistics were recovered by simply binning the identified residue pairs in a histogram.

Principal component analysis of h-bonding patterns

In all umbrella sampling windows in the axial direction, per-amino acid hydrogen bond counts were calculated with respect to (a) the DNA backbone, (b) DNA nucleobases and (c) other amino acids, producing a vector of 3×51 = 153 values. These vectors were bin-averaged along the z coordinate in 0.05 nm bins, with the unbiasing factor $\exp\left(\frac{U_i(r)-F_i}{k_B T}\right)$ used as a weighing function, where U and F are defined as above. The resulting 60×153 data matrix was used to compute the 153×153 correlation matrix, which was then diagonalized to calculate the eigenvectors. Original data was projected onto individual eigenvectors to provide an intuitive interpretation of the results.

Free energy of protein-DNA interactions approximated at large intermolecular distances

From 3.0 to 4.5 nm, the umbrella sampling-derived profiles were continuously extended with a free energy profile obtained using the Debye-Huckel approximation. Here, the free energy was calculated as the Debye-Huckel interaction energy minus the entropic term that accounted for cylindrical symmetry, $RT\log(2\pi r)$, where T is the temperature and r the radial distance in the XY plane. To accurately calculate the Debye-Huckel interaction energy in a system composed of two molecules with nontrivial charge distributions, we ran another six 100-ns steered MD simulations (three forward, from 3.0 to 4.5 nm, and three backwards, from 4.5 to 3.0 nm) in a larger simulation box and used Plumed to compute the respective energies. These were then averaged over all six trajectories, and values were reported $\pm\sigma$. This approach allowed us to rescale the DH potential by a constant factor ($\frac{q_{mut}}{q_{WT}}$) in cases where the WT overall charge of the mutant (q_{mut}) was different than that of the wild-type protein (q_{WT}).

Calculation of the binding free energy

The reported binding free energy of 9.0 kcal/mol was calculated using the formula $\Delta G = -\frac{k_B T}{V_{std}} \int_0^b e^{\frac{-F(r)}{k_B T}} dr$ where b is the upper boundary of the bound state (here the upper boundary was assumed at the inflection point of the free energy profile, i.e., at 2.2 nm), and V_{std} is the volume corresponding to the standard concentration of 1 M (1.66 nm^3), consistently with the infinite dilution approximation.

Effect of starting geometries and force field corrections on the free energy profiles

While the unchanged Amber99sb-parmbsc0 2 force field was used in almost all simulations, we performed an additional set of US simulations employing the CUFIX (Champaign-Urbana NBFIX) correction designed to reduce the excessive stability of lysine-carboxyl and lysine-phosphate salt bridges [289], as well as using the updated parmbsc1 parameters designed to improve the structural properties of DNA [290], in order to estimate whether the obtained free energy profiles display strong dependence on the particular force field variant. Three umbrella sampling simulations (with the original Amber99sb-parmbsc0 force field, with Amber99sb-parmbsc1 and with the CUFIX correction) were performed using frames extracted from the spontaneous association trajectory in which full reconstitution of the native complex was observed. In this case, 45 windows (equally spaced in the 1.40–2.00 nm range in 0.025-nm intervals, and 2.00–3.00 nm in 0.05-nm intervals) were simulated for 250 ns, with initial 50 ns discarded to allow for equilibration, producing furhter 52.5 s of simulations.

Brownian dynamics simulations

Langevin dynamics on telomeric (target) and non-telomeric (off-target) DNA was simulated using a custom Python script. In the simulations, the protein was represented by a point corresponding to its center of mass (COM), and moved in 3D space in an effective potential dictated by the periodicized 2D free energy profiles $V(r,z)$ according to the Langevin equation in the overdamped limit: $x_i(t+\Delta t) = x_i(t) - \frac{D(r)}{k_B T}\frac{\partial V(r,z)}{\partial x_i}\Delta t + \sqrt{2D(r)\Delta}\zeta$ where x_i is the i-th spatial coordinate, $D(r)$ is the position-dependent diffusion coefficient, ζ is the normalized Gaussian random variable, k_B is the Boltzmann constant and T the temperature. The DNA strand was fixed and rigid and aligned along the z axis. To emulate the effect of helical geometry of DNA, we calculated the distribution of the protein COM position with respect to the major/minor groove (see Fig. 4.11A) and used it to define rigid reflective walls at r = 1.3 nm in the "groove" region and 2.0 nm elsewhere, as illustrated in Fig. 4.11B. To correct for the entropic penalty associated with restricted access to angular positions outside the $\frac{\pi}{2}$-wide "groove", an additional term equal to $RT \log \frac{1}{4}$ was added to the 2D free energy profiles at radial distances below 2.0 nm. The experimental diffusion coefficient [211] was employed (assuming isotropy) when the protein was in close vicinity to the DNA, and at distances where most protein-DNA contacts disappeared (3.0 nm) this diffusion coefficient was smoothly switched to a value typical for free diffusion of similarly sized proteins (see Fig. 4.11C).

Merging of free energy data into 2D profiles

The 1D profiles in the radial direction (at larger protein-DNA distances) derived from the spontaneous binding trajectory were merged with the 2D profile along the major groove (at small protein-DNA distances) to produce a full 2D free energy profile. To merge the two profiles at a specific radial protein-DNA distance, a switching function was used, in the general form $\frac{1}{\pi}(\text{atan}(a(r-b))+\frac{\pi}{2})$. Here, b determines the midpoint and a the abruptness of the switching transition; in actual calculations values of 1.8 nm and 75 were used, respectively. In addition, as the insertion of Arg380 in the minor groove confers a degree of sequence-specificity even at protein-DNA distances up to 2.5 nm, we evaluated how the fraction of time spent by Arg380 in the minor groove f varies along the DNA in US simulations, and merged the two profiles – with Arg380 bound and dissociated from the minor groove – as a linear combination of the respective probability densities, $G(r,z) = RT\log[f(z)\rho_{bound}(r,z) + (1-f(z))\rho_{unbound}(r,z)]$. For this reason at large distances, the 2D profiles are not uniform along the DNA axis, as would be the case in a naive approach ignoring long-distance specificity. To obtain the free energy profile with Arg380 dissociated from the minor groove, the number of direct contacts between the guanidinium moiety and nucleobases (with a threshold of 0.4 nm) was continuously decreased to 0 in a set of 45 10-ns simulations, with the DNA and bulk part of the protein kept positionally restrained. In actual US simulations, the biasing potential was kept to prevent R380 from re-insertion, while allowing the residue to form direct contacts with the backbone phosphates.

Transfer entropy calculations

Transfer entropy between two observables i and j, defined in the section above, is a positive quantity that measures the amount of information that the knowledge of previous value of j (in general, arbitrarily long history of values of j) provides about the future evolution of i. In practice, observables i and j are discretized and the respective joint and conditional probabilities are calculated as 3- or 2-dimensional matrices based on observational data. To explicitly infer causal relationships from observational data, the normalized directional index is calculated as $D_{j\to i} = \frac{T_{j\to i}}{h(I_{k+1}|I_k)} - \frac{T_{i\to j}}{h(J_{k+1}|J_k)}$ which ensures that the matrix D is antisymmetric. In this work, the normalized directional index was calculated using a custom Python script (available at `https://gitlab.com/KomBioMol/transfer_entropy`), and i and j were binary observables (1s and 0s) that corresponded to the existence of a given h-bond at frame k. The implementation followed the methodology outlined in the article by Kamberaj et al. [279], to which the reader is referred for a thorough discussion. In addition, since numerical noise and correlations can bias of the outcome, the time series was then randomly scrambled, yielding a dataset in which causal relationships no longer exist. This scrambling was repeated three times, and the mean transfer entropy obtained from the scrambled time series was subtracted from the original result to yield the corrected directional index.

Markov state model

To depict the binding process in a more intuitive manner, we started by using a high-dimensional representation of the spontaneous binding trajectories encompassing a set of generalized coordinates. These involved: the radial distance r; mRMSD with respect to the standard and inverse orientation of the DNA duplex; dihedral angles between the main axes of two major helices of the DBD and the Z-axis (or, equivalently, DNA main axis; both sine and cosine of these angles were used to avoid incontinuities); and the binary (0 or 1) descriptors of the existence of most frequently occurring h-bonds (see Fig. 4.10 for a precise description). These generalized coordinates were subjected to dimensionality reduction using PCA to produce a projection onto an 8-dimensional subspace of largest variability. Thus obtained data was then clustered into microstates using the batch K-medoids algorithm to produce discrete-state trajectories used for the construction of the MSM, and model parameters were chosen based on the optimization of the GMRQ score, as proposed by McGibbon and Pande [291]. In spectral clustering with PCCA+, macrostates were chosen based on the relaxation timescales criterion. The analysis was performed using the MSM-Builder library [292].

Preparation of the alchemical systems

The complexes of TRF1, TRF2, HOT1 and POT1 with their cognate DNA fragments were extracted from PDB entries 1W0T, 1W0U, 4J19 and 3KJP, respectively. In case of TZAP, a homology model was built based on the canonical Zif268 zinc-finger protein (PDB entry 1P47), with Modeller [293] used for amino acid and X3DNA [280] for nucleobase substitutions. Parameters for bonded interactions with zinc atoms were taken from ZAFF [294], and three consecutive zinc-finger domains were modelled to ensure proper affinity. Two replicas of the TZAP/DNA complex were prepared that differed in the relative location of the C-terminal domain with respect to the G-triplet in DNA. During 1-μs equilibration runs, one of the replicas failed to create sequence-specific contacts, while the other eventually formed stable h-bonds with DNA bases identical to those found later in the crystal structure (see Fig. 3.5).

The preparation of oxidatively modified systems was automatized with a custom Python script that converts a regular Gromacs topology file into a single-topology input for alchemical simulations. The script is available at https://gitlab.com/KomBioMol/guanine_lesions.

The molecular systems contained either a solvated DNA molecule or a solvated protein-DNA complex. They were contained in dodecahedron boxes, and solvated with 8531, 8386, 11634, 12811 and 15091 (DNA only) or 8254, 8101, 10926, 12393 and 14607 (protein-DNA) TIP3P molecules in case of TRF1, TRF2, POT1, HOT1 and TZAP, respectively. K+ and Cl- ions were added to ensure charge neutrality and maintain a physiological ion concentration of 0.154 M.

Parametrization of lesions

The dual topology of T (a G/C $\rightarrow$ T/A double mutation) was obtained *via* the pmx webserver. 8oxoG and O8oxoG were parametrized following a standard procedure of molecule parametrization in Amber force fields, i.e., charges were recalculated using Gaussian 09 [295] at the HF/6-31G* level of theory, and atom types were matched to the existing parametrization (parmbsc1) by analogy.

In case of Sp and FapyG, the above procedure was followed by fitting dihedral parameters to QM-derived energy profiles (see Fig. 3.6). For Sp alone, internal angles

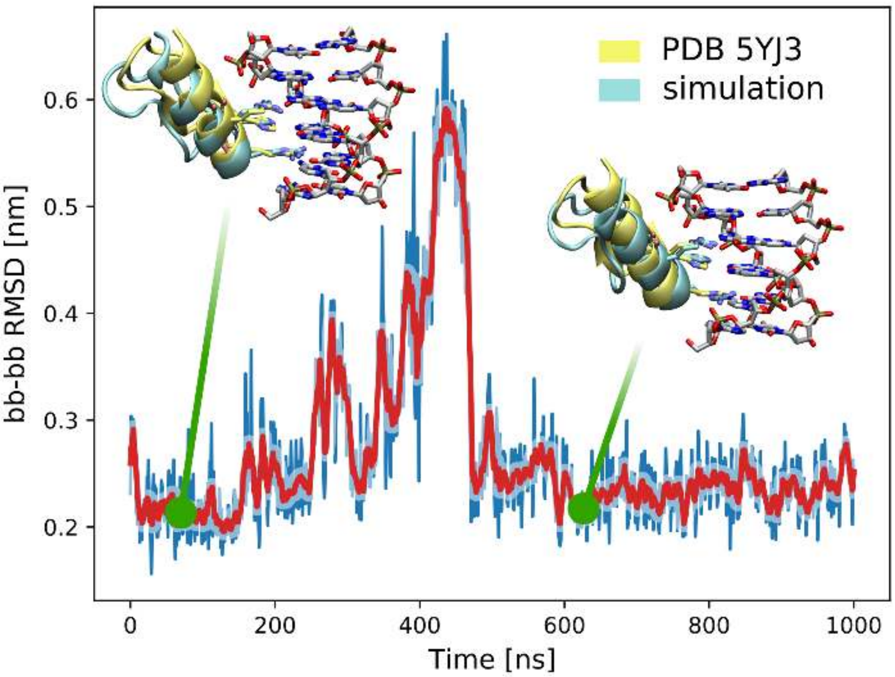


FIGURE 3.5: Evolution of RMSD between the homology model of the TZAP-DNA complex and the recently deposited PDB entry 5YJ3 during the 1 μs refinement. At ca. 500 ns the shift/rebinding event led to the formation of a stable complex, with the three residues key for sequence specificity cooperatively bound to the guanine triplet. For the RMSD calculation, only heavy backbone atoms were considered.

in both 5-membered rings were additionally reparametrized based on the QM Hessian using the modified Seminario method [296]. All QM runs other than charge calculations were performed using the MP2 method and the 6-31G** basis set.

All modified parameters are listed in the script freely available on our GitLab page.

Parameters of the alchemical simulations

All simulations were run using Gromacs 5.1 and Plumed 2.3 [284, 287], employing the Amber99-parmbsc1 force field widely used for DNA systems [290]. Temperature was kept at 300 K using the CSVR thermostat with a time constant of 0.1 ps, and pressure was maintained at 1 bar using the isotropic Berendsen barostat. Particle-Mesh Ewald summation was used to calculate long-range electrostatic interactions. To avoid singularities, the default soft-core potential was used in free energy simulations. SETTLE was used to constrain the geometry of water molecules, and LINCS to constrain the length of all bonds (in case of T, which suffered from poor stability due to the choice of a dual-topology approach) or h-bonds only (in all remaining cases).

Convergence of λ-values

To arrive at an optimal set of λ-values for alchemical transformations, we created a script that iteratively restarts short replica exchange runs and modifies λ-intervals

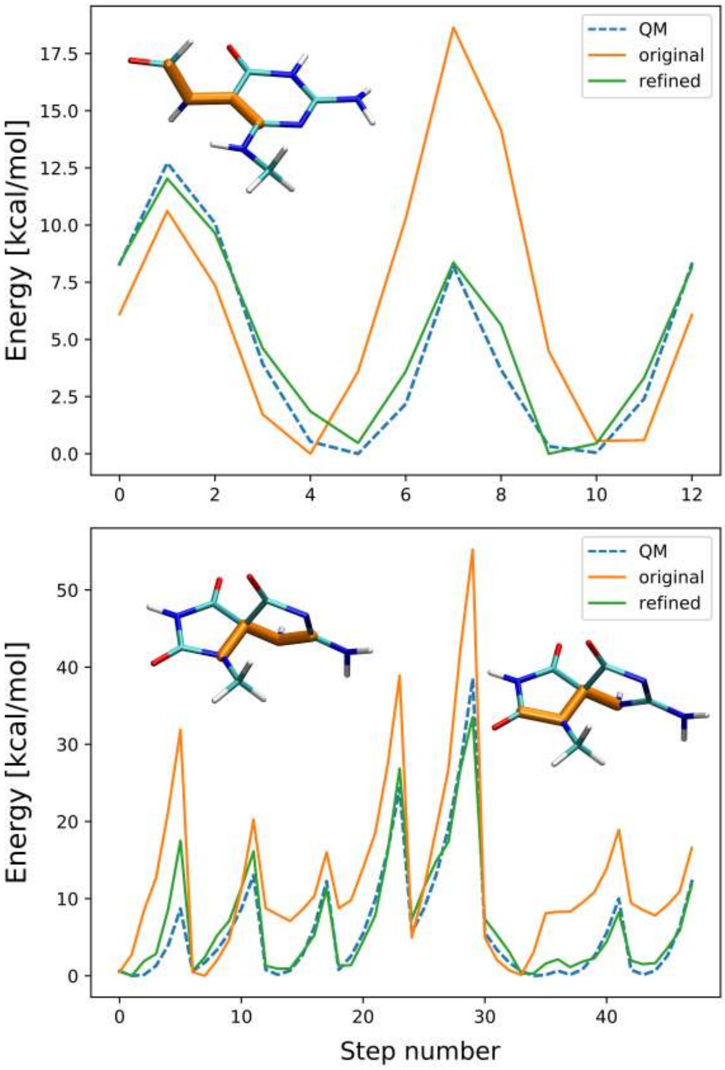


FIGURE 3.6: A comparison of the initial MM energies (orange), reference QM values (dashed) and refined MM energies (green) for three dihedral angles that were explicitly parametrized. The bottom panel refers to the C4-C5-N7-C8 dihedral in FapyG, while the bottom one corresponds to dihedrals N9-C4-N3-C2 and C8-N9-C4-N2 in Sp (marked in orange in the schematic pictures).

to produce equal exchange probabilities between windows. Each next iteration of simulations is slightly longer to provide better sampling, and a short-term memory buffer averages previous results to avoid strong fluctuations; the script itself is available at `https://gitlab.com/KomBioMol/converge_lambdas`. This approach allowed us to automatize λ optimization and perform it separately for each system. The number of λ-points was chosen as a multiple of 8 that allowed exchanges to be accepted with a probability of at least 10%, which yielded 16 replicas for 8oxoG and O8oxoG, 24 for T and FapyG and 32 for Sp. Example convergence plots obtained using the script are shown in Fig. 3.7.

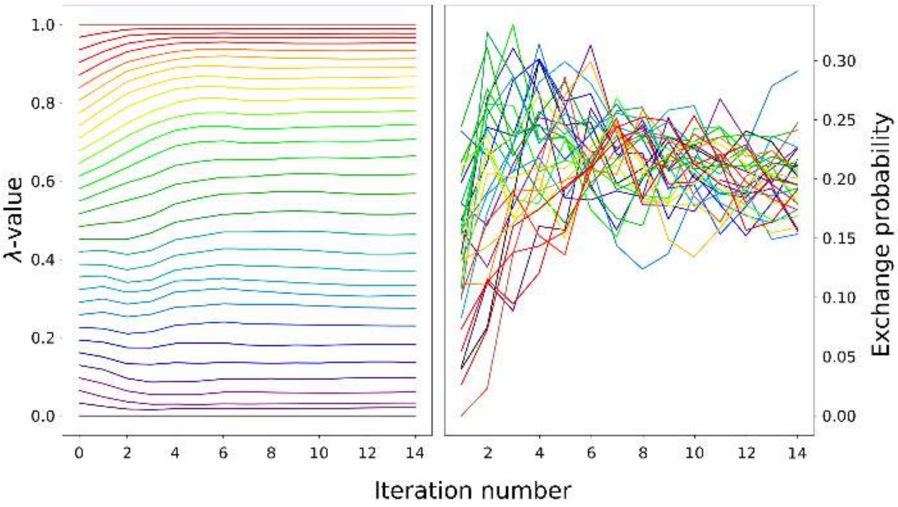


FIGURE 3.7: A sample convergence plot of the λ values (left) and exchange probabilities (right) yielded by the Python script in 14 iterations of the custom algorithm.

Free energy calculations and error analysis

Free energy changes were calculated from the alchemical simulations using the Bennett acceptance ratio (BAR) using the gmx bar utility in Gromacs, and an appropriate thermodynamic cycle, in which the lesion was separately introduced in free solvated DNA and in a protein-DNA complex. Convergence of the free energy changes are shown in Fig. 3.8.

For consistency, the respective mutations in free DNA have been carried out multiple times in any given sequence context (4 for dsDNA – with DNA taken from systems initially containing TRF1, TRF2, TZAP and HOT1 – and once for ssDNA, with DNA taken from the POT1 system). We therefore utilized this redundancy to estimate the magnitude of error introduced in a single calculation with lesion i as a standard deviation of the obtained ΔG_i values, σ_G^i. Moreover, in case of free dsDNA systems, we could use the averaged values as an estimate of the true ΔG, with the error now estimated as $\frac{\sigma_G^i}{\sqrt{4}} = \frac{\sigma_G^i}{2}$. Assuming that one can also use σ_G^i as an error estimate for protein-DNA systems, we estimated the error in $\Delta\Delta G$ (being a difference of two ΔG values, one for free DNA and another for protein-DNA complex) as (a) $\sqrt{\frac{3}{2}}\sigma_G^i$ in case of TRF1, TRF2, HOT1 and TZAP, and (b) $\sqrt{2}\sigma_G^i$ in case of POT1 (due to lack of redundancy in free ssDNA calculations).

Estimation of specific affinities

To provide a best guess of the specific affinities of individual proteins (i.e. the difference in binding free energies between the binding to a telomeric sequence and an off-target site), we searched the literature for experimental estimates of binding constants. For TRF1 and TRF2, Lin *et al.* calculated this affinity difference as equal to 2.0 and 0.7 kcal/mol, respectively, using telomeric and λ-DNA [211]. For POT1, an affinity difference of 2.0 kcal/mol was observed for the shelterin complex harboring a single POT1 molecule depending on whether the ssDNA overhang contained

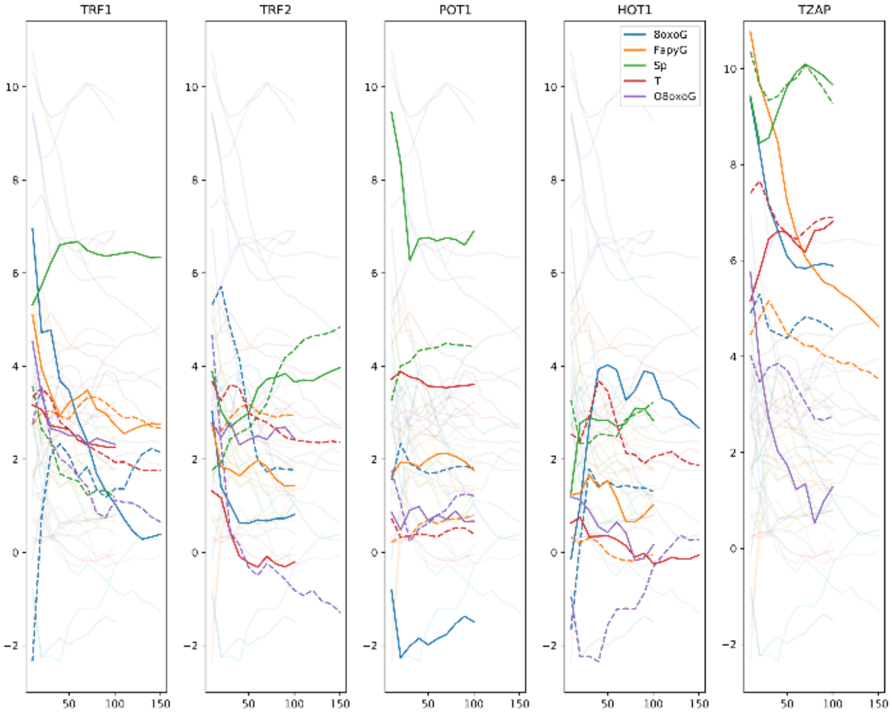


FIGURE 3.8: Convergence plot of the differential free energies (ΔΔ Gs) in all systems considered in the study. In a number of systems that failed to converge within 100 ns (chosen based on the min-max difference within the last 40 ns), the simulations were additionally prolonged to reach 150 ns. Note that individual values might differ from those in the histogram due to averaging of ΔGs in chemically identical naked DNA systems.

a telomeric or random sequence [297]. For TZAP, no direct comparison with a randomized sequence was found, but affinity differences of 2 kcal/mol were reported when the binding interface was disrupted by mutation of a single amino acid to alanine, and the affinity was reduced by ca. 0.6 kcal/mol upon single nucleotide mutation (G→A in the central G-triplet), so that 2.0 kcal/mol is a reasonable upper bound for the specific affinity [78]. For HOT1, no reliable quantitative data was found.

Feature selection and clustering

To describe the properties of protein-DNA complexes using a consistent set of interpretable parameters, we calculated a set of 48 diverse descriptors for each frame in our alchemical trajectories. 30 descriptors/features corresponded to local DNA structure (including relative positions of the two bases immediately preceding (-1) and following (+1) the lesion, backbone dihedrals at the -1/0/+1 position, as well as sugar puckering parameters at the -1/0/+1 position. The other 18 descriptors/features corresponded to counts of intermolecular hydrogen bonds between (a) protein residues highlighted in Fig. 4.12, (b) other protein residues, (c) DNA base pairs, (d) DNA backbone and (e) water molecules, with donor/acceptor and acceptor/donor pairs considered separately. Hydrogen bonds were identified using the gmx hbond utility of Gromacs.

The importance-based feature selection allowed us to extract subsets of these features that correlated with free energy changes at all λ-values in the alchemical simulations. More specifically, the overlap between feature distribution in unmodified base ($\lambda = 0$) with the corresponding distribution in a modified base ($0 < \lambda_1 \leq 1$) was measured as the Bhattacharyya distance between the two distributions, and the correlation coefficient between the log-Bhattacharyya distance and $\Delta\Delta G_{0 \to \lambda_1}$ was computed. Absolute correlation coefficients were averaged for all simulations involving individual proteins. In this manner, the 1/3 of top scoring features was selected for further analysis.

Using these selected features, a matrix of absolute values of correlation coefficients *between individual features* was calculated and used as similarity matrix for spectral clustering with a pre-defined threshold. If several features exhibited high degree of correlation, they were replaced with the projection onto the highest-variance principal component from PCA performed on this subset of features.

LDA analysis

LDA was applied to time-series data representing the existence of individual protein-DNA hydrogen bonds at the residue level. The analysis was performed separately for each protein, and data was categorized according to the the presence of individual lesions ($\lambda = 1$) as well as the unmodified dG ($\lambda = 0$). From each simulation, 51 equally spaced time frames corresponding to the interval 50-100 ns were selected for the analysis.

Free energy decomposition

In order to dissect contributions to $\Delta\Delta G$ coming from individual parts of the systems, we employed the additive property of force fields and split the trajectories into subsystems consisting of (a) the residue that was modified; (b) the DNA molecule, including (a); (c) DNA and the solvent, including (b); (d) the entire protein-DNA complex and the solvent (where relevant), including (c). Reruns were performed on each subtrajectory to calculate $\frac{\partial H}{\partial \lambda}$, and $\Delta\Delta G$ values were obtained for each subsystem using the TI equation. In that way, subsystem-specific contributions to $\Delta\Delta G$ could be obtained by simple subtraction: the "intrinsic" contribution from part (a), the contribution of DNA environment from (b)$-$(a) etc. Note that the contribution from the solvent in the free DNA system corresponds to the contribution from the solvent and the protein in the protein-DNA system, so that these were treated together as "environment".

Structural analysis of DNA

The X3DNA *analyze* utility was used to calculate the structural properties of DNA [280]. For the modified base/base pair, sugar puckering angle and the BI/BII conformational ratio were calculated in this way, based on a set of representative conformations sampled with a 1 ns stride as described above. BI/BII conformations were calculated for the backbone segment immediately preceding and following the modified base, and population averages are calculated over this pair of values. For energetic reasons, the mean population of BI virtually always exceeds 50% as two BII backbone segments almost never coexist adjacent to each other.

For the PCA/LDA analysis, the full set of rigid body descriptors was calculated in a similar manner for the base pairs immediately adjacent to the modified base (i.e. at positions N-1 and N+1). In this way, a common plane-defining subset of atoms could be found for any combination of lesion and position, as either A-T/G-C or G-C/G-C pairs contain the same purine/pyrimidine templates.

Preparation of seeding frames for QM/MM calculations

For the purpose of running QM/MM MD simulations, classical MD simulations of a solvated TRF1-DNA complex were first performed. To this end, a cubic box (a=7.23 nm) containing the complex as deposited in the PDB entry 1W0T (with K421 deprotonated), 11739 TIP3P water molecules as well as 50 K+ and 35 Cl- ions was prepared in several versions: (i) G cation radical (prior to proton transfer to C); (ii) G radical (after proton transfer to C); (iii) 8oxoG; (iv) 8oxoG radical (after proton transfer to C); (v) O8oxoG with protonated C. Gromacs 5.1 was used to perform the classical simulations, and the standard Amber99sb force field with the bsc1 correction was employed. Each system was simulated for at least 500 ns, and initial frames were selected that were characterized by both (a) low Nζ-C5 distance and (b) the presence of at least 5 water molecules in the vicinity of the reactive site. Topology files in the appropriate format were converted from Gromacs using the ParmEd library.

Parametrization of modified nucleobases

The parametrization procedure followed the standard workflow recommended for the Amber force field, i.e. calculation of *in vacuo* ESP-fitted charges at the HF/6-31G* level of theory and reassignment of atom types. A total of 3 modified nucleobases were specifically parametrized for the purpose of running equilibrium MD simulations: the cytosyl cation, guanyl radical and the 8-oxoguanyl radical. Simulations containing 8oxoG and O8oxoG employed parameters obtained as described above.

QM/MM MD simulation setup

All QM/MM MD simulations were performed in cp2k 4.x, employing the plane-wave auxillary basis to achieve a considerable speedup and using a plane wave cut-off of 300 Ry. The QM subsystem was cubic with a box vector of 2.1 nm, and the standard 0.5 fs timestep was employed to integrate the equations of motion. To avoid the costly calculation of HF exchange, a local M06-L DFT functional implemented in LibXC was used for all simulations, except for the metadynamics runs that employed BLYP for technical reasons. Norm-conserving Goedecker-Teter-Hutter pseudopotentials were used along with a TZVP basis set and Grimme's D3 correction aimed to improve the description of noncovalent interactions. The parametrization of the pseudopotentials followed the report I co-authored recently [265].

QM/MM Umbrella Sampling simulations

All QM/MM systems were first equilibrated for at least 1 ps, and then a steered MD run was performed during at least 10 ps over which the C-N distance was gradually decreased from ca. 3.5 to 1.3 using a force constant of 1 atomic unit (1 Hartree/bohr^2). For umbrella sampling, 12 parallel simulations were run in which the C-N distance was restrained at values from 1.3 to 3.5 in 0.2 intervals using a force constant of 500 kcal/mol ^{2}. Free energies were extracted using WHAM.

QM/MM metadynamics

For the QM/MM MD metadynamics, 3 collective variables were chosen: (1) the C-N distance, (2) the difference between the H-Nζ and H-N7 distances, and (3) the difference between C5-C6 and C4-C6 distances. To enhance sampling, a total of 6 walkers were being run in parallel, exchanging data every 5 MD steps. A well-tempered variant of metadynamics was used [251] with a bias factor of 35. Hills with a height of 0.5 kcal/mol and width of 0.1, 0.4 and 0.3 (in respective order of the CVs) were added every 5 MD steps. In addition, restraints were added that prevented the C8-N7, N7-C9 and N7-C5 bond lengths from exceeding 1.55 . Summation of hills was performed using the fes utility of cp2k.

Chapter 4

Results and Discussion

4.1 Sequence Recognition in the Binding of TRF1 to DNA

The initial scientific objective I set in my doctoral work was the elucidation of the role of amino acids located on the TRF1-DNA interface that mediate direct DNA sequence readout. To this end, I first calculated free energy profiles for the association of the wild-type (WT) protein to telomeric DNA in two orientations – standard, corresponding to a specific interaction with the target 5′-TTAGGG-3′ sequence, and inverted by a 180° rotation about the protein-DNA axis, corresponding to non-specific binding at an off-target 5′-CCCTTA-3′ sequence. This data served as a reference for the subsequently calculated set of analogous free energy profiles for mutant proteins in which interfacial amino acids were substituted with alanines. Five amino acid substitutions were chosen – R380A, V418A, K421A, D422A and R425A – based on direct side chain-nucleobase contacts in the native complex. The results of the calculations are shown in Fig. 4.1.

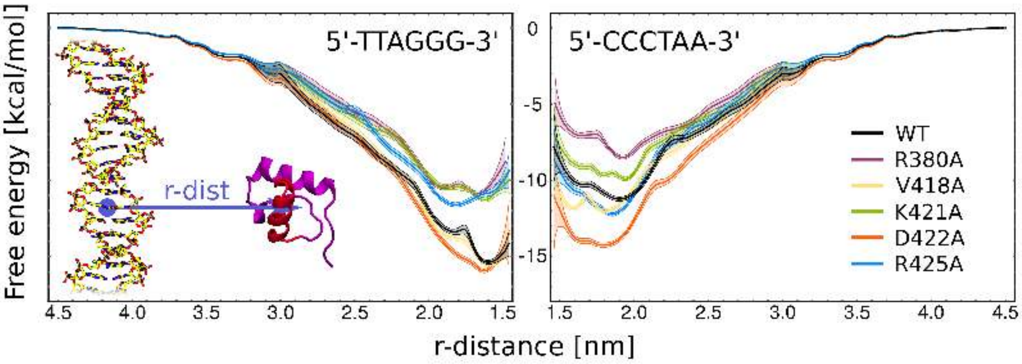


FIGURE 4.1: Radial free energy profiles for a set of TRF1 variants (including the wild-type domain and single amino acid mutants) on two dsDNA substrates: the target telomeric sequence (left) and the inverse telomeric sequence that models an off-target site (right). Standard WHAM errors are marked as semi-transparent. At distances larger than 3.0 nm, the free energy profile obtained from US/WHAM is extended with an entropy-corrected Debye-Huckel energy profile.

While the WT protein expectedly shows a much higher affinity to the native than non-specific sequence – as judged from the depth of free energy minima marked by the black line – it is surprising to note that the D422A mutation largely (by ca. 4 kcal/mol) abolishes this preference; a similar but less pronounced effect can also be seen for V418A. At the same time, both D422A and V418A contribute almost nothing to the affinity for the target sequence: the corresponding red and tan curves on the left panel of Fig. 4.1 overlap with the black one to a large extent.

A simple explanation for this fact is that in the native complex, the existence of a formal negative charge of D422 on the interface with a negatively charged DNA molecule is neutralized by the side chain's interaction with the two major groove-exposed amino groups of cytosines in the 5′-CCC-3′ run complementary to the G triplet. On a non-specific sequence, this charge-charge repulsion is not neutralized and hence results in a largely unfavorable thermodynamic effect in the WT protein that vanishes when D422 is mutated to alanine. The case of V418 can be explained similarly: the hydrophobic interaction between the side chain of valine and the small patch formed by C5/C6 atoms of pyrimidines does not contribute strongly to binding, but when valine faces the polar major groove-exposed surface of purines a significant penalty is produced.

In contrast, the more "well-behaved" residues such as R380, K421 and R425 act mostly by increasing the affinity for the target sequence, as illustrated by the affinity decreasing significantly (by 4-5 kcal/mol) when the residues are mutated to alanines; the corresponding affinities for the off-target sequence change only slightly in either direction. Based on the two distinct behaviors, one can define two groups of chemical moieties involved in specific recognition of any kind: (a) positive selectors, i.e. structural elements that provide sufficient affinity for the correct binding partner, such as R380, K421 and R425; and (b) negative selectors, i.e. moieties whose primary purpose is to prevent the binding of non-target competitors, such as D422 and V418. In molecular recognition, the former residues are typically involved in plain charge complementarity- or hydrogen bond-based interactions, while the latter can take advantage of any structural peculiarities such as large dipole moments or rare shape patterns. While seemingly trivial, these guiding principles need to and should be harnessed in the future design of specific and selective molecular interactions.

In the free energy profiles, it is also worth noticing that the binding of TRF1 and DNA is clearly bimodal, with two distinct free energy minima located at the intermolecular distances of approx. 1.55 and 1.85 nm. This small barrier vanishes in the red (D422A) and blue (R425A) profiles, suggesting that the cooperative binding of D422 and R425 – the two amino acids implicated in the core protein-nucleobase interface (see Fig. 4.2) – contributes to the "lock-in" mechanism that kinetically stabilizes the complex in the bound state.

To estimate how frequently the aspartate-based negative selection mechanism might be actually employed by nature, I carried out a structural bioinformatic analysis of all protein-DNA complexes deposited in the Protein Data Bank (PDB). By identification of all direct nucleobase-amino acid contacts in the analyzed structures, I was then able to produce a histogram of nucleobase-amino acid pairs, shown in Fig. 4.3. From the figure, it can be seen that the carboxyl-cytosine amino group constitutes a popular motif in base recognition, accounting for ca. 8% of all instances. In fact, such an interaction can also be found in the prototypical c-Myb homeodomain. Only four other residues – arginine, lysine, glutamine and asparagine – are more frequently used in DNA sequence recognition than the negatively charged aspartate and glutamate, with the bidentate arginine-guanine interaction prevailing in the statistics by a large margin. Similarly, a preference for the bidentate aspartate/glutamate-cytosine

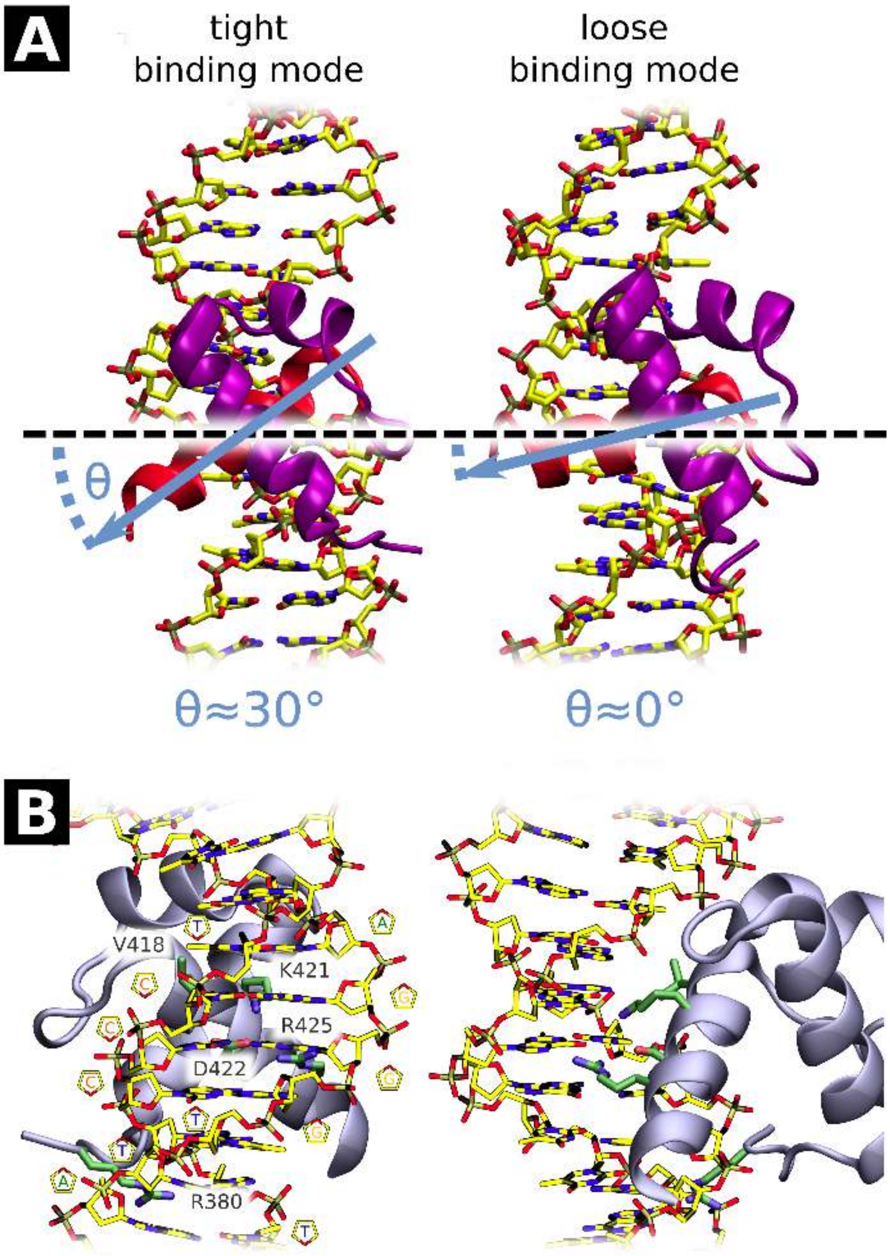


FIGURE 4.2: (A) A comparison of two binding modes distinguishable in the free energy profiles. (B) A view of the binding interface. Side chains involved in sequence recognition are shown explicitly in green.

recognition mode can be found: if there is a cytosine in the central D/E-bound position, in 61% of cases another cytosine can be found at a 3′- or 5′-neighboring position, up from the conditional probability of $\frac{7}{16} = 44\%$ that corresponds to a purely random distribution.

As a means of additional validation of the computational model, I also recalculated the free energy profiles while varying two of the model's key components: the force field and the set of seeding frames. While all other simulations in this chapter relied on the older amber99/parmbsc0 force field, here the free energy profiles were recalculated using (a) the CUFIX correction by Yoo and Aksimantiev introduced to fix the

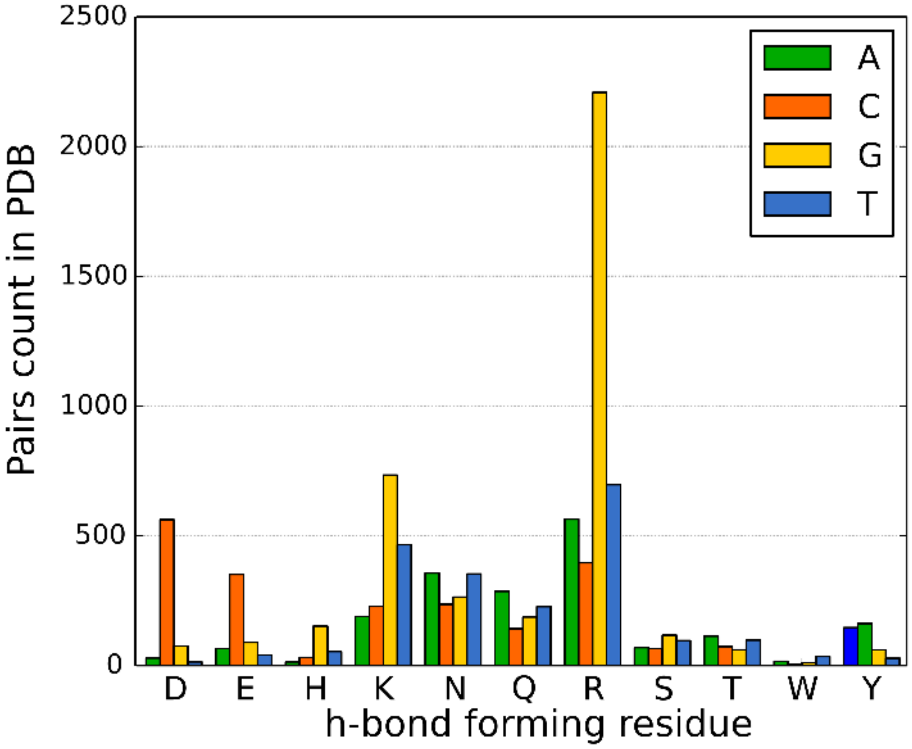


FIGURE 4.3: Histogram of hydrogen-bonded amino acid-nucleobase pairs found in the PDB database. Using a distance criterion of 0.35 nm and a custom Python/MDTraj script, 10168 such pairs were identified in the 3891 batch-downloaded structures.

overstabilization of lysine-phosphate and lysine-carboxyl salt bridges [289] and (b) the updated parmbsc1 correction that provides a systematic improvement over its predecessor [290]. Also, two sets of seeding frames were used in free energy simulations: (a) one generated from a slow steered MD simulation in which the molecules in the complex were pulled apart, and (b) one obtained from an unbiased simulation of spontaneous complex formation that will be described below. How both choices affect the shapes of the estimated free energy profiles can be seen in Fig. 4.4.

The plots in panel A illustrate that the choice of seeding frames can have a profound impact on free energy estimates from regular umbrella sampling simulations, with seeds chosen from spontaneous binding simulations yielding a much lower affinity estimate than ones selected from an enforced dissociation run. This is most likely indicative of a hysteresis effect in which certain non-covalent bonds are only stretched but not ruptured along the unbinding pathway, producing a spring-like harmonic force that prevents further dissociation. On the other hand, when the protein binds spontaneously, there is no external force to drive this effect and one recovers the mean force exerted on one binding partner by the other with much better accuracy. This is indeed revealed by the excellent agreement between the binding free energy estimated from the spontaneous binding seeds (-9.0 kcal/mol) and the value determined experimentally for the DNA-binding domain (-9.2 kcal/mol) [298]. It is also worth noting that the bimodal characteristics of TRF1 binding to DNA is to some extent preserved in the binding-based profile. As could be expected, pooling all data together (labelled as pull/bind) produces a curve located roughly halfway between the ones corresponding to separate data sets.

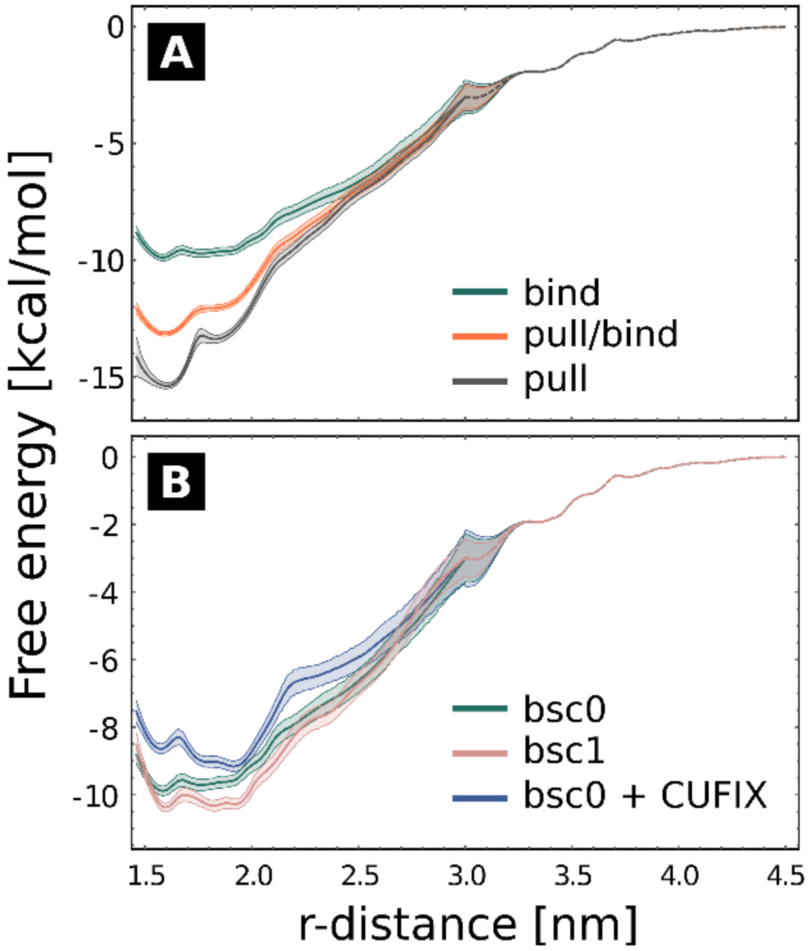


FIGURE 4.4: The effect of (A) choice of seeding frames and (B) force field corrections on the free energy profiles as obtained in the US/WHAM procedure. The profiles labeled as "bind" were obtained using starting frames from a trajectory in which we observed spontaneous reconstitution of the native complex. The profiles labeled as "pull", on the contrary, were obtained using starting frames from a steered MD trajectory in which the protein was pulled away from the DNA with external force.

The effect of force field corrections, shown in panel B of the figure, is clearly more subtle – perhaps not surprisingly as each correction only affects a small number of interactions. CUFIX yields a slightly (-1.5 kcal/mol) destabilized tightly bound mode and a steeper profile about the 2.0 nm mark, lowering the overall estimate of the binding affinity by 1.2 kcal/mol. The parmbsc1 correction, on the other hand, introduces only minor changes to the binding profile, increasing the estimate of affinity by 0.1 kcal/mol.

4.2 Free Energy Maps of Specific and Non-Specific Interaction Between TRF1 and DNA

A natural extension of the 1-dimensional free energy analysis is to picture the free energy landscape in higher dimensions. Here, my objective was to obtain a comprehensive description of the thermodynamics of TRF1 on the telomeric repeats in both the radial (i.e. away from the DNA axis) and axial (along the DNA axis) coordinates, here denoted as r and z. This was achieved through a combination of two independent sets of free energy calculations, namely (1) umbrella sampling simulations along the radial coordinate that used seeds from unbiased binding, and (2) umbrella sampling simulations along the axial coordinate in which the protein was allowed to sample the radial coordinate within the limits of the bound state. Then, data was merged together so that the energetics of the bound state was generated directly from properly reweighted set (2), and in the unbound and intermediate state (1.8 nm and up) the set (1) was used to describe long-range interactions. Long-range sequence dependence was additionally introduced by considering two profiles in the radial dimension – one with R380 tucked in the minor groove and one with R380 unbound – and merging them in proportion to the fraction of R380 bound to DNA at a given z-coordinate. This was motivated by the observation that R380 is capable of mediating sequence-specific interactions through the N-terminal basic tail of the DNA binding domain.

The resulting two-dimensional free energy landscape, illustrated in Fig. 4.5, depicts the free energy as sensed by the domain diffusing in the vicinity of target (left panel) and off-target (right panel) sequence. As both sequences are periodic, the 2D free energy function has a periodicity of ca. 2.0 nm (roughly $\frac{6}{10}$ of the B-DNA helix pitch of 3.6 nm) in the z dimension, so that the profile effectively describes the energetics of the interaction on an entire telomere. Noteworthily, at large radial distances, the two profiles correspond to the same physical situation – a protein loosely interacting with the DNA through its extended N-terminal tail. As the radial distance decreases and TRF1 binds in the native binding pose, the protein has to assume one of the two possible orientations, encountering either the native telomeric or the off-target inverse sequence, as visualized in Fig. 4.5D.

A comparison of the two 1D profiles in the radial direction (upper panel of Fig. 4.5) reveals that the estimated difference in binding free energy between the specific and non-specific binding, here equal to 1.7 kcal/mol, very well matches the experimentally determined difference of 2.0 kcal/mol reported recently based on single-molecule measurements [211]. Combined with the accurate determination of the absolute binding affinity, as described above, this reinforces the view that properly set up free energy calculations can serve as a very sensitive and powerful tool in the prediction of interactions between biomolecules.

From the main free energy surfaces depicted in Fig. 4.5A, one can identify two distinct free energy basins of similar depth (ca. -8.5 kcal/mol) at low protein-DNA radial distances and z-coordinate of 1.95 and 2.2-2.8 nm. This surprising feature suggests that the binding between TRF1 and telomeric DNA is feasible not only in the crystallographically determined pose (z = 1.9 nm), but also on the neighboring sites as long as the key contacts are formed. Indeed, a structural inspection reveals that at these locations the cooperatively stabilized interface formed by D422 and R425 faces a guanine-cytosine pair, K421 can switch from a guanine to a neighboring adenine,

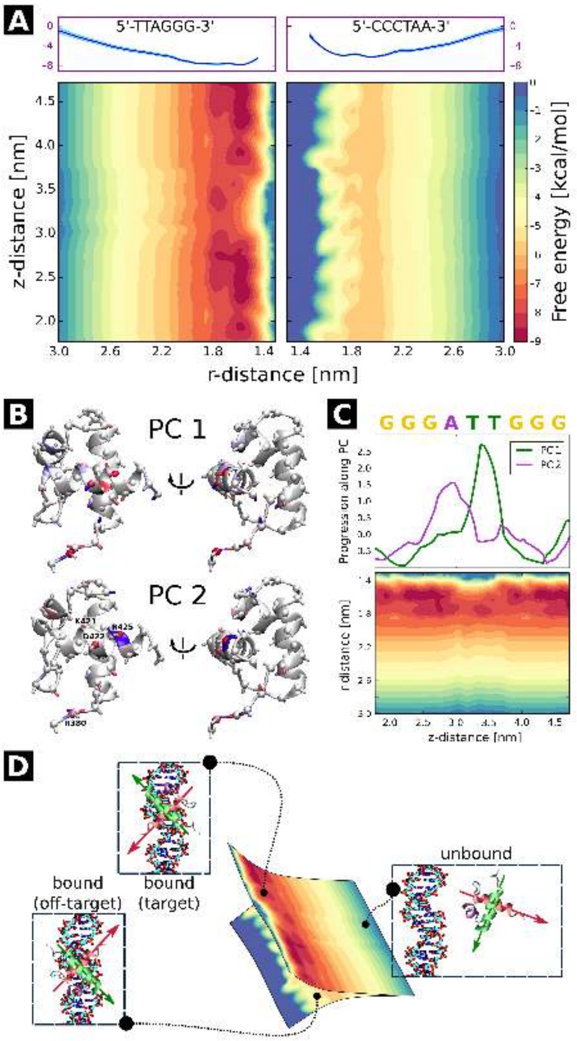


FIGURE 4.5: (A) Two-dimensional free energy maps describing the thermodynamics of TRF1-DNA interaction along the telomeric sequence (left) and the inverse telomeric sequence, modelling an off-target site (right). The 2D maps are reflected so that the process of binding proceeds from the sides to the center of the figure. (B, C) Results of PCA of hydrogen bonding along the target telomeric sequence. Structures in panel B depict the first two eigenvectors of the covariance matrix, while panel C illustrates the mean projection of the data on these two eigenvectors, corresponding to correlated and large changes in hydrogen bonding patterns. The free energy map and the DNA sequence are aligned to aid in data interpretation. (D) A schematic picture illustrating the use of the inverse telomeric sequence as a model off-target site: as the protein domain approaches the DNA strand, it assumes one of the two possible orientations, effectively binding to either the "standard" or the "inverse" sequence. At large distances, both profiles correspond to the same physical situation.

and R380 interacts with hydrogen bond acceptors of the adenine-thymine pair in the major groove. Conversely, in the free energy barrier region (z within the range of 3.0-3.7 nm) D422 and R425 encounter adenines and thymines unfit to serve as dual hydrogen bond acceptors for the arginine and donor for the aspartate. At the same time, the minor grove-bound R380 dissociates when AT pairs are exchanged for CG ones: while in the minor groove an AT pair exposes two hydrogen bond acceptors that perfectly accommodate the bond-donating guanidyl group, a CG pair exposes an acceptor-donor-acceptor pattern that disrupts the interaction.

In order to quantify the connection between the observed free energy minima and hydrogen bonding patterns, I used PCA to observe correlated changes in h-bonding that occur as the protein progressed along the axial coordinate. Here, mean numbers of h-bonds formed by each of the 51 amino acids with (a) nucleobases, (b) DNA backbone and (c) all other amino acids were considered as separate features, so that every z-value was represented as a 153-component vector. The results are shown in Fig. 4.5B and C, illustrating both the dominant change in h-bonding pattern associated with the two principal components and the projections of input data on both components aligned with the free energy maps. The projections show that the first component corresponds to the G/C→T/A transition, with a peak centered at ca. 3.4 nm, while the second describes moving from the G/C stretch to the A/T pair. Interestingly, the peaks and wells in the top of panel C clearly coincide with free energy barriers and minima in the bottom, providing strong support for the claim that direct h-bonds strongly contribute to the positioning of TRF1 on telomeric tracts. From panel B, one can identify amino acids primarily contributing to the first and second eigenvectors of the covariance matrix, with coloring of the ribbon corresponding to h-bonds formed with the DNA backbone, coloring of the sphere – with DNA bases, and of the stick – with other amino acids. It is then evident that in both cases (i.e., PC1 and PC2), moving away from the native interface results in a decrease in the population of sequence-specific h-bonds formed by D422, R425 and R380 with DNA bases (red spheres; to some extent also applicable to K421). On a non-native interface, these amino acids then begin to interact with the DNA backbone (blue ribbons/tubes, R380 and R425 in PC2) or other amino acids (blue stick, D422 in PC1). As both R425 and D422 here primarily move away from binding to DNA bases, the affinity is locally reduced, giving rise to the observed free energy barrier when z is in the range of 3.0-3.7 nm.

When TRF1 approaches DNA in the opposite orientation, encountering the inverse telomeric sequence (right panel in Fig. 4.5), a free energy barrier prevents it from reaching radial protein-DNA distances smaller than 1.6 nm, a distance at which most sequence-specific amino acid-base contacts form. As a result, the free energy is largely invariant to translation in the axial direction, making the binding affinity independent of sequence. This observed smoothing of the axial free energy profile, corresponding to a difference in roughness of ca. 1.5 kcal/mol, is in line with both previous experimental observations and theoretical predictions [197, 203], as well as the value of 1.7 kcal/mol measured specifically for TRF1 in single-molecule experiments [211]. Even though there is very little variation in free energy between individual binding sites along the sequence, the ripple-like patterns noticeable at z-distances of ca. 1.6 nm hint at the existence of the aforementioned "lock-in" effect related to the alignment of the protein's interfacial h-bond donors and basic residues with backbone phosphates. To provide a more visual description of this effect, I plotted the average normalized XY-positions of the protein's center of mass with respect to a standardized orientation of the DNA double helix in all umbrella sampling windows, color-coding them according to the progression along the axial coordinate (z). The resulting scatter plot, shown in Fig. 4.6, depicts the clustering of primary protein locations in discrete positions, with a 10-fold pseudosymmetry reflecting neatly the DNA helix pitch corresponding to 10-10.5 base pairs. Although the clustering is markedly more evident at the native telomeric sequence, a residual effect can also be seen at the inverse sequence, sufficient to explain the ripple-like patterns seen in the free energy maps.

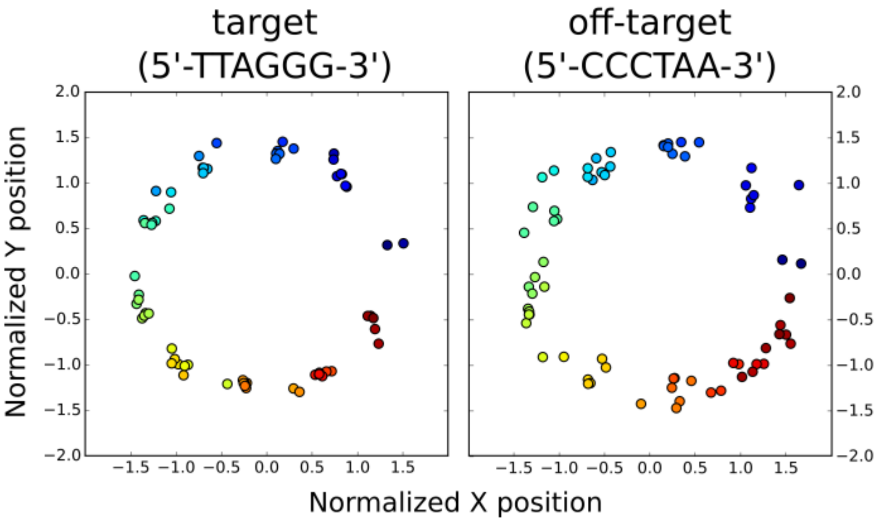


FIGURE 4.6: The distribution of centers of mass of TRF1 in individual umbrella sampling windows distributed along both target and off-target sequence. The lock-in mechanism is visible as clustering of neighboring points in a 10-fold symmetric pattern, characteristic of the DNA double helix.

4.3 Dynamics of Target Recognition and Binding by TRF1

A major part of my work was the investigation the dynamic properties of sequence-specific complex formation, as opposed to most existing computational studies that only analyze the properties of the bound state or the process of enforced unbinding. Typically, such studies are limited in their predictive capabilities as they only explore a small region of the conformational space, restricted by the short nanosecond timescales often used in the literature. To address this problem, here I ran 50 long equilibrium simulations that totaled 140 μs, seeded from random unbound geometries. In such a way, I was able to explore the conformational space in an unbiased manner, which allowed me to observe the initial stages of protein-DNA complex formation. I then spawned 80 additional unbiased trajectories seeded from frames selected from the initial 50 runs, using a simple minimal RMSD-based criterion for selection (10 phosphorus atoms from the DNA backbone and 15 C_α atoms from the recognition helix were chosen as a reference group, and the minimal RMSD with respect to any position along the DNA was required to be smaller than 0.175 nm). By propagating each new simulation for additional 500 ns, I was able to recover a native-like complex characterized by heavy-atom root-mean square deviation (RMSD) of less than 0.2 nm.

Fig. 4.7A illustrates the RMSD with respect to the structure of the native complex, tracked in time over the single trajectory that resulted in complete reconstitution of the bound state. In panel B, the resulting structure (green) is superimposed onto the crystal structure (yellow) used as a reference. As seen from the evolution of RMSD,

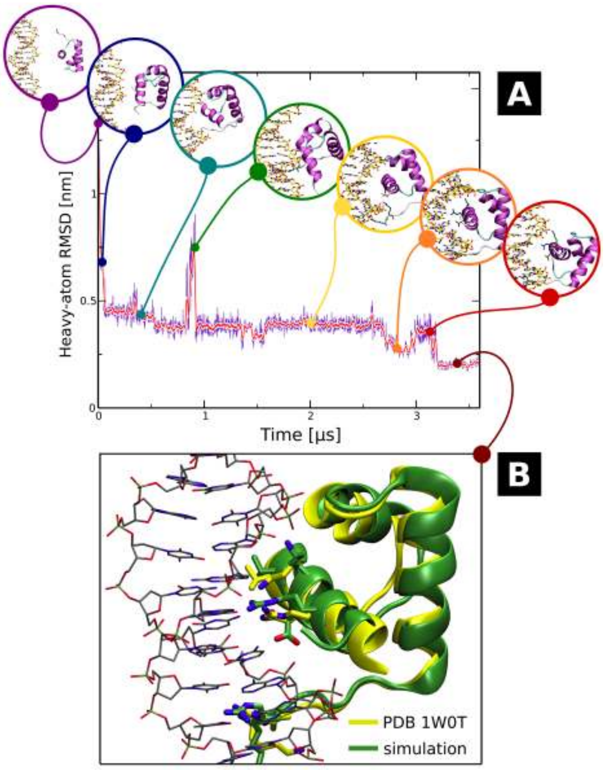


FIGURE 4.7: (A) Time-evolution of heavy-atom RMSD during the 3.6 μs spontaneous binding trajectory, with characteristic milestones along the complex formation pathway illustrated in the circular insets, as discussed in the text. (B) An overlay of the spontaneously formed TRF1-DNA complex (green) atop the structure taken from PDB entry 1W0T (yellow), along with the side chains involved in sequence recognition.

the initial stage of complex formation proceeds very rapidly due to electrostatic attraction, with a steep decrease in RMSD from 1.5 nm to ca. 0.5 nm within less than the initial 100 ns, and first protein-DNA contacts formed within 2 ns. This rapid association is consistent with the slope and lack of barriers in the corresponding region of the free energy profile. However, the initial orientation of the protein was clearly incompatible with the dominant binding mode, as the recognition helix was oriented parallel to the DNA axis and faced the minor groove (dark blue circle in Fig. 4.7A). While in this case the recognition helix required about 50 ns to locate the major groove (as illustrated in the cyan circle), such a fast transition was not always a rule, with many trajectories stuck in a minor groove-bound state for more than 1 μs (not shown).

When TRF1 was bound loosely in the major groove, the basic residues of its unstructured tail – K379 and R380 – were positioned in a way that enabled them to scan the minor groove. As a result, within the next 300 ns, K379 became inserted in the minor groove, effectively anchoring the protein to a potential target site that was already preselected for binding due to the characteristic hydrogen bonding patterns exposed by AT pairs in the minor groove. This pivot-like interaction proved to be key for the subsequent complex formation when the protein spontaneously escaped from the major groove near the 850 ns mark, only remaining anchored *via* K379 (green circle). Even after prompt rebinding, two key sequence-specific residues

– R380 and D422 – remained connected by a bidentate salt bridge, preventing the formation of a proper contact pattern in both the minor and major groove (yellow circle). It took another almost 2 μs for this salt bridge to break, eventually allowing K379 to be swapped by R380 and releasing the sequence-scanning D422 to the major groove. The corresponding tightening of the near-native complex and formation of first sequence-specific amino acid-nucleobase interactions (orange circle) accumulated tension that was released in the final refinement step (red circle). In this event, the domain quickly reestablished both its specific and non-specific contacts with the DNA helix on an immediately adjacent site, forming a relaxed native-like interface. The resulting complex then remained stably bound with an RMSD of 0.2 nm, similar to values seen in simulations seeded from the crystal structure.

The above observations derived from a single realization of the binding process highlight some general features of sequence recognition by homeodomains. Firstly, in addition to the reported involvement in the "monkey bar" mechanism of facilitated intersegmental transfer [299], I found that the basic disordered tails play an instrumental role in sequence preselection and anchoring of the DBD in the minor groove while the domain itself scans the local sequence with its interfacial residues. This finding supports and expands on several previous propositions regarding their functional role [191, 300]. Secondly, a sequence of checkpoints apparently exists in the process of complex formation, with rapid events separated by long dwells during which the DBD can move on if certain conditions regarding local sequence are not met. Finally, certain intermediate states accumulate a degree of structural frustration that might accelerate either productive binding or abandonment of the currently sampled site.

Simultaneously, an inspection of trajectories that failed to yield a native-like complex despite significant progress along the binding pathway could be just as informative. For this reason, I visually analyzed eight other trajectories that later served as seeds for the extended 500-ns runs; the corresponding videos can be found in the supplementary files published along with the original manuscript [301]. In particular, trajectories labeled with numbers 2-6 highlight the fact that the anchoring interaction between R380 and the minor groove is indeed a key preselecting factor: in trajectory 2, R380 slides along the GC-rich stretch of the minor groove only to form a stable contact upon arrival at an AT-rich segment. When this anchoring contact is not present, the protein is capable of switching orientations or jumping to sites several bp away. Notably, such an intermittent random walk can result in subdiffusive motion, with the DBD switching between a slow and fast diffusive mode of motion, establishing at least two modes of non-specific binding. In movies labeled as 3-6, R380 remains bound (mostly) to D422, which prevents it from inserting into the minor groove. While not necessarily functional, this transition from an amino acid-bound to minor groove-bound state appears to be a significant rate-limiting step in sequence recognition and complex formation. When the arginine anchor is not present, the protein-DNA interaction can be transiently stabilized by interactions between the C-terminal portion of the DNA-binding helix and DNA backbone phosphates (e.g. in movie 7), allowing for the R380-minor groove contact to form. In support of the above findings, movies 7 and 8 also highlight the ability of the domain to extensively sample neighboring sequences in presence of the anchoring interaction, with the geometry of the linker region such that the DNA-binding helix preferentially faces the DNA major groove.

To provide a more quantitative insight into the causal relationships that result in the formation of a protein-DNA complex, I used transfer entropy, a quantity reviewed extensively in the previous chapter as well as in a dedicated article [279], to identify information flows between individual hydrogen bonds in the analyzed system. As a reminder, transfer entropy measures how much new information about the future of variable i can be extracted from the knowledge of history of another variable j if the history of i is already known, providing a proxy for causal relationships between time-shifted changes in i and j. Following the suggested practice [279], instead of transfer entropy alone I used the directional index $D_{j \to i}$ – the antisymmetrized and normalized form of transfer entropy: $D_{j \to i} = -D_{i \to j}$ and $-1 \leq D_{j \to i} \leq 1$ – corrected for the mean of several calculations utilizing scrambled values of J. It is worth noting that while a value of 0 suggests that no causal link can be inferred, it does not imply a lack of correlation between the variables in question. The sign of the directional index also only indicates the direction of information transfer, and not whether the correlation itself is positive or negative, so that this information has to be supplied e.g. from an independently computed (time-lagged) correlation matrix.

In Fig. 4.8, the directional index matrix for both intra- (observables 1-33) and intermolecular (observables 34-68) hydrogen bonds is shown as well as mapped on the TRF1 structure. As can be seen, instead of single entries entire columns and rows of large values show up in the matrix, indicating the presence of interactions that initiate coordinated changes in the global interaction patterns, likely corresponding to the formation of a native complex. Selected residues involved in the formation of those "early" hydrogen bonds are mapped onto the sequence with lines as well as explicitly shown in the top panel of Fig. 4.8, colored green if they correspond to intra- and cyan if to intermolecular interactions. The most intense signal corresponds to four residues located on a small patch in the top (in the figure's orientation) portion of the DBD: W403, S404, S417 and K421. These residues are therefore responsible for the initiation of formation of a properly aligned major groove-bound complex, providing an electrostatic anchor opposite to the previously described R380. The time-lagged correlation matrix, shown in the bottom panel of Fig. 4.8, indicates that the binding of these residues is highly cooperative, also correlating with the binding of R380 to DNA, so that the two anchors act together to properly position the protein. On the N-terminal side of the protein, somewhat smaller but still notable contributions stem from the interactions of K379, R380, Q381 and W383 with DNA, also clustered spatially and positively correlated with each other, reflecting the role of the other anchor in complex formation. This is consistent with the above observations from spontaneous binding trajectories, where the two anchoring interactions stabilized each other, thereby contributing to sequence recognition in both minor and major groove.

Several intramolecular contacts also seem to be implicated in the complex formation, including the D422-R425 salt bridge that is present in both the unbound and sequence-specifically bound state, but often competes with the D422-R380 pair in the intermediate state. Similarly, the E387-R415 salt bridge forms or dissociates depending on the availability of DNA backbone, as seen from a strong anticorrelation signal for the R415-DNA contact (the most intense blue matrix element in the bottom left section of the correlation matrix).

After characterizing the initial events leading to the formation of a properly aligned

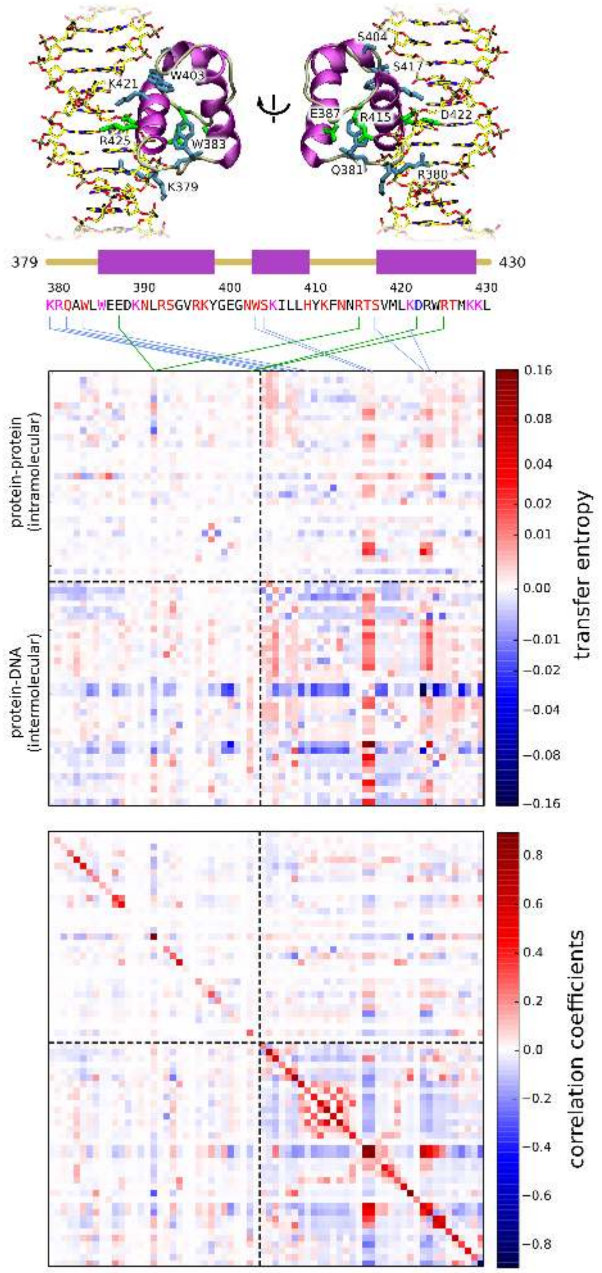
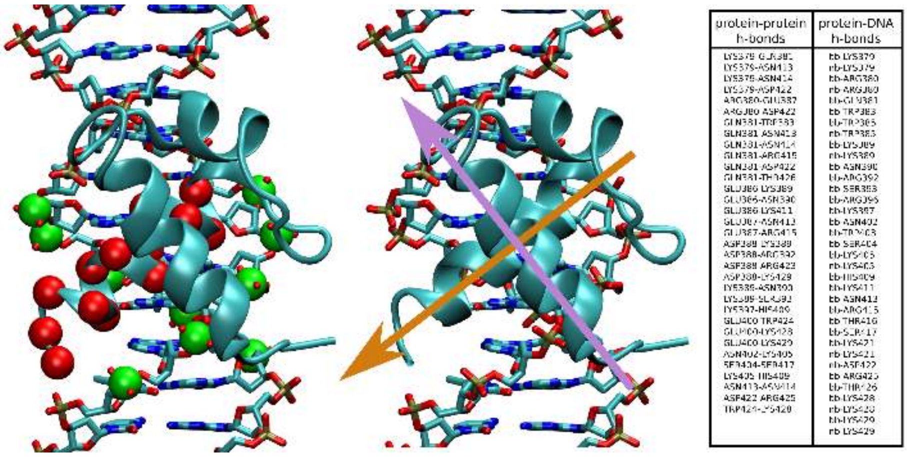


FIGURE 4.8: The directional index matrix (top) with top-scoring columns mapped onto intra- (green) or intermolecular (blue) contacts, using both the TRF1 sequence and structure of the complex. To facilitate interpretation, a time-lagged correlation matrix (bottom) is shown that indicates the direction of the observed correlations.

complex, I then used the MSM formalism to integrate the analysis of both productive and non-productive events into a single model. Due to insufficient statistics regarding association events at individual subsequences, I chose to employ sequence-agnostic descriptors as a basis for the model construction: (i) minimal RMSD with respect to a subset of DNA phosphorus atoms and recognition helix α carbons, (ii) relative orientations of the binding partners expressed as scalar products between protein helix and DNA main axes, and (iii) the most frequently occurring hydrogen bonds, as visualized in Fig. 4.9. The descriptors were autoscaled and subjected to



protein-protein h-bonds	protein-DNA h-bonds
LYS379-GLN381	bb-LYS379
LYS379-ASN413	nb-LYS379
LYS379-ASN414	bb-ARG380
LYS379-ASP422	nb-ARG380
ARG380-GLU387	bb-GLN381
ARG380-ASP422	bb-TRP383
GLN381-TRP383	bb-TRP385
GLN381-ASN413	nb-TRP385
GLN381-ASN414	bb-LYS389
GLN381-ARG415	nb-LYS389
GLN381-ASP422	bb-ASN390
GLN381-THR426	bb-ARG392
GLU386-LYS389	bb-SER393
GLU386-ASN390	bb-ARG395
GLU386-LYS411	bb-LYS397
GLU387-ASN413	bb-ASN402
GLU387-ARG415	bb-TRP403
ASP388-LYS389	bb-SER404
ASP388-ARG392	bb-LYS405
ASP388-ARG423	nb-LYS405
ASP388-LYS429	bb-HIS409
LYS389-ASN390	bb-LYS411
LYS389-SER393	bb-ASN413
LYS397-HIS409	bb-ARG415
GLU400-TRP424	bb-THR416
GLU400-LYS428	bb-SER417
GLU400-LYS429	bb-LYS421
ASN402-LYS405	nb-LYS421
SER404-SER417	nb-ASP422
LYS405-HIS409	bb-ARG425
ASN413-ASN414	bb-THR426
ASP422-ARG425	bb-LYS428
TRP424-LYS428	nb-LYS428
	bb-LYS429
	nb-LYS429

FIGURE 4.9: An illustration of molecular descriptors used to map structures onto the feature space: minimal RMSD between the recognition helix α carbons and DNA phosphates in both standard and inverse orientations (left), spatial orientation of the protein helices' main axes (middle) and a set of commonly formed hydrogen bonds (right).

dimensionality reduction using PCA as a preprocessing step prior to discretization. Then, clustering yielded discrete states, allowing for the conversion of original trajectories to discrete ones as required by the MSM. Finally, fuzzy kinetic clustering with PCCA+ assigned the original states to 10 slowly interconverting macrostates, providing interpretable structural data.

Fig. 4.10 presents the results of the PCCA+ clustering, with MSM states color-coded according to their assigned macrostate, and projected onto the planes spanned by PCA eigenvectors 1 and 2 (left panel) and 1 and 3 (right panel). Additionally, the 8 heavily-populated regions, discernible as yellowish blobs in the 2D plots, are directly visualized using representative structures. In both panels, the rightmost side – corresponding to high values of PC1 – represents the native-like complex in a standard orientation, with the dense population of states resulting from a largely negative free energy associated with the bound state. Within this region, two sub-states can be distinguished, marked as 1 and 7, that roughly correspond to the two binding modes (tight and loose) discussed above. Simultaneously at negative values of PC1, regions marked as 4 and 6 correspond to the DBD bound in the inverse orientation, suggesting an intuitive interpretation of the two extremes of the 2D map. The middle region of the plot is occupied by a pool of transient intermediates, in which either the N-terminal helix is facing the major groove (sample 8), or only tips of the two long DBD helices are oriented towards the major groove (samples 2, 3, 5). Interestingly, intermediates in which helices are bound to minor grooves, such as the one observed transiently in the successful binding trajectory, are apparently too unstable to be picked up in the analysis. However, a closer inspection indicates that the macrostates yielded by PCCA+ are not structurally uniform, so that the sample structures should be thought of more in terms of guiding examples than as definitive assignments. This should also be expected as the number of possible relative protein-DNA orientations is much too high to be strictly resolved using a 2-dimensional projection of the initially very high-dimensional data.

Table 4.1 contains the calculated mean first passage times (MFPTs) between the PCCA+ derived macrostates, obtained using an improved MSM-based

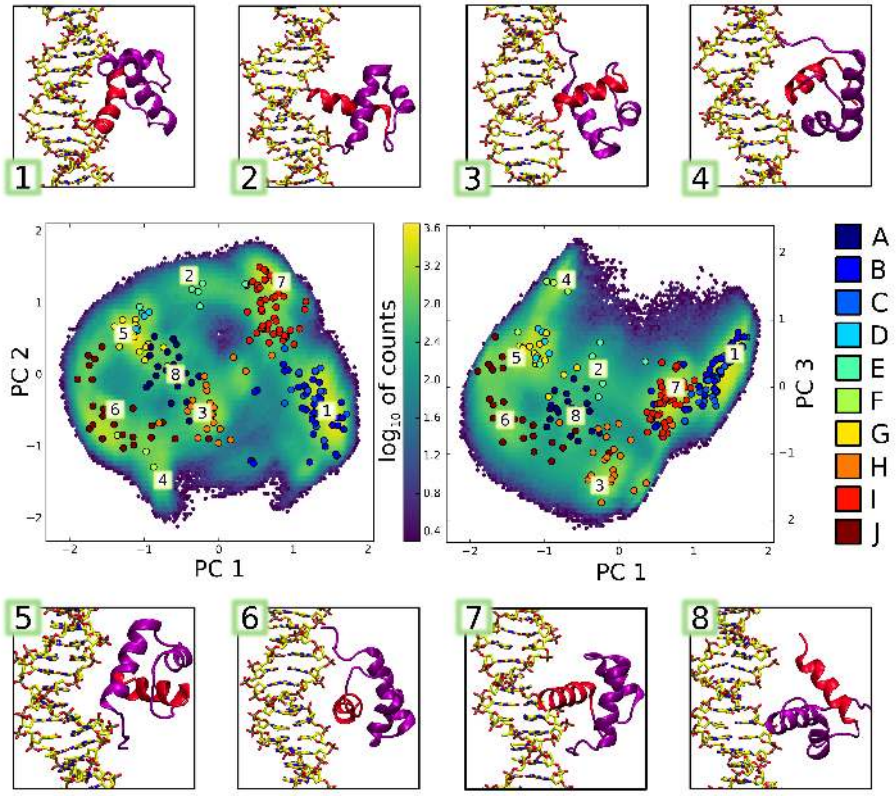


FIGURE 4.10: A depiction of intermediates formed along the binding pathway in the equilibrium simulations, mapped onto principal components 1/2 and 1/3. Dots correspond to individual states selected for the Markov state analysis, and are color-coded as assigned by the PCCA+ method.

method [302]. As indicated by multiple sub-microsecond entries, many closely overlapping macrostates interconvert within as little as several to tens of nanoseconds. However, selected transitions, e.g. starting in the state denoted as G, can take several microseconds to complete, indicating the presence of kinetic traps along the complex formation pathway. It is worth pointing out, however, that the lack of clear structural distinctions between the macrostates might have led to underestimation of individual MFPTs due to a degree of overlap between the supposedly kinetically separated states. I then used the same algorithm to calculate the MFPT corresponding to the formation of a native-like TRF1-DNA complex, defined by the presence of sequence-specific hydrogen bonds formed by R380, D422 and R425, and obtained a value of 34 μs, a value qualitatively consistent with the observation of a single successful binding event in 127 simulations that totaled 180 μs but were mostly (except for the 500 ns extensions) seeded from the unbound state. In the same way, I estimated the MFPTs for domain flipping from the inverse to standard orientation and *vice versa* as equal to 11 and 88 μs, respectively.

4.4 Diffusion of Telomeric Proteins on DNA

To explore the dynamics of DNA-bound TRF1 on a larger spatial and temporal scale, I ran a number of Brownian dynamics simulations using the obtained free energy

TABLE 4.1: Mean first passage times (MFPT$_{i\to j}$) between macrostates identified with the PCCA+ algorithm, as calculated using the Markov+Color algorithm by Suarez *et al.* All times are given in μs.

	A	B	C	D	E	F	G	H	I	J
A	0.0004	0.1406	0.7308	0.027	0.0287	0.0142	0.3677	0.0128	0.0273	0.0063
B	0.149	0.0002	0.0014	0.5234	2.3808	1.0182	4.7412	0.0174	0.005	0.0442
C	0.5651	0.0011	0.0001	0.9854	1.9949	1.772	4.5288	0.2058	0.0057	0.1908
D	0.0143	0.208	0.8093	0.0005	0.0087	0.0361	3.4184	0.9101	0.0075	0.0332
E	0.0159	0.4932	1.5117	0.0099	0.0003	0.0059	0.1782	1.1662	0.0792	0.0034
F	0.0078	0.5488	2.0294	0.0537	0.0062	0.0002	0.5187	0.1996	0.4144	0.0047
G	0.082	0.8654	2.0177	1.4684	0.076	0.2196	0.0002	1.1999	0.6147	0.0049
H	0.0109	0.0176	0.2009	1.1569	1.75	0.3297	3.9873	0.0002	0.0225	0.004
I	0.0262	0.006	0.0106	0.0207	0.1584	1.2038	5.0431	0.0378	0.0001	0.0535
J	0.0115	0.0465	0.294	0.0895	0.0078	0.0114	0.0195	0.0064	0.0473	0.0002

maps. In the simulations, protein dynamics was modelled as that of a point particle moving stochastically in the effective 3D potential generated by either the off-target or telomeric DNA strand, with additional soft and hard boundaries introduced to mimic the presence of the major groove on the otherwise cylindrically shaped ds-DNA. The geometry, with the simplified major groove that allows the protein to enter the r=1.4-2.0 nm range, is illustrated in Fig. 4.11. It has to be noted that such a model is only capable to grasp general trends, e.g. rotation-coupled sliding or residence times; as it turns out, some aspects of protein diffusion along DNA cannot be reproduced in a low-dimensional setting due to averaging of inherently multi-step processes such as formation of individual residue-specific contacts. Such effects will be addressed using a high-dimensional Markov model-based approach that is currently being refined and developed but will not be presented here due to suboptimal sampling.

As shown in Fig. 4.11A, TRF1 predominantly exhibits “processive” rotation-coupled sliding: the trajectories remain mostly restricted to the major groove, diffusing along long stretches of DNA between individual detachment events. On the off-target sequence, though, sliding is less strongly coupled to rotation as now detachment from the groove successfully competes with diffusion along the groove, so that the DBD can easily switch orientations in the search mode, as highlighted previously in theoretical considerations. Nevertheless, the coupling is clearly observable, in line with past reports on the mode of sliding of several DNA-binding proteins [203]. The qualitative picture is supported by the good agreement between the experimental residence times of ca. 15 s on telomeric and 1.8 s on non-specific (λ) DNA [211] and the observation of one detachment event during 12.8 s simulations on the target and three detachment events during 3.2 s simulations on the off-target sequence. On the other hand, the effective diffusion rate along the DNA sequence – and, in particular, the ratio of diffusion rates at a target and non-target sequences – as well as the experimentally determined anomalous diffusion coefficient were not reproduced by the Brownian model, highlighting the need of better structure-based approaches to this problem.

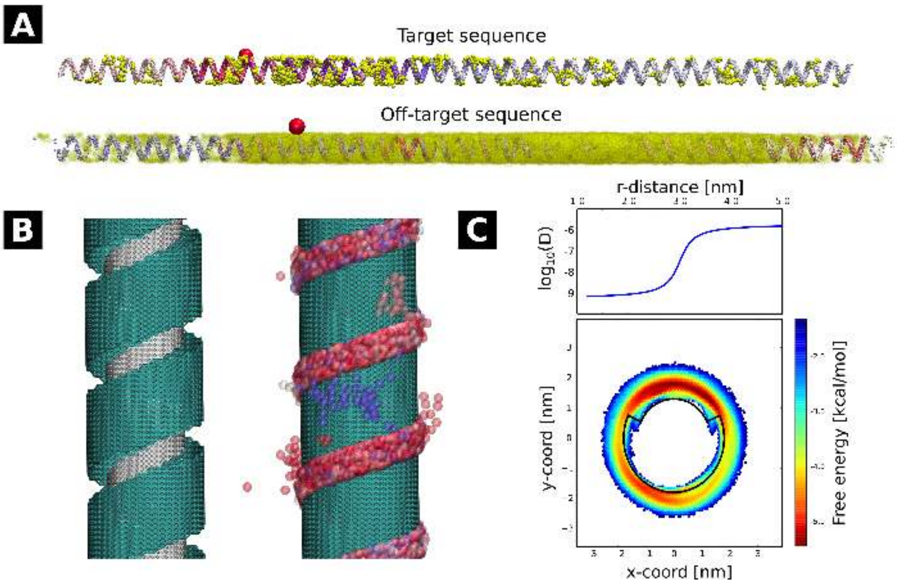


FIGURE 4.11: (A) Sample 100-ms Brownian Dynamics trajectories on a periodic target and off-target sequence, with time progression color-mapped on the trajectory from red through white to blue. The large red dot indicates the starting point, and yellow dots mark frames in which radial distance exceeded 2.0 nm, so that the protein was able to diffuse out of the DNA groove instead of strictly following the helical path. (B) Geometry of the "solid" DNA model used in the simulation, with a zoom-in on a fragment of the trajectory shown in panel A. (C) The distance-dependent diffusion coefficient used in the simulations (top) and the free energy map of protein-DNA interactions in the xy-plane inferred from spontaneous association simulations (bottom). The uppermost basin corresponds to the major groove, and the lowermost one to the minor groove. The thick black line marks the reflective barrier used in BD simulations to enforce the helical geometry of the DNA model depicted in panel B.

4.5 The Impact of DNA Lesions on DNA Binding Affinities of Telomeric Proteins

In the other major part of my doctoral work I focused on a quantitative analysis of the direct impact of DNA oxidation on the occupancy of telomeric proteins on telomeres. Here, the premise is that the accumulation of lesions can at least partially disrupt the binding equilibrium, in particular given that (i) only a fraction of the binding sites on telomeric DNA is exposed for binding (the other being sequestered in inaccessible regions of nucleosomal DNA); (ii) previous reports indicated that such an effect is operational *in vitro* [152]; and (iii) telomeric DNA damage is abundant, and repair after exposure to acute oxidative stress can take multiple hours [119]. I selected five possible oxidative lesions: 8oxoG, its oxidized version O8oxoG, FapyG, Sp, and thymine resulting from a G→T transversion. Using computational alchemistry coupled with replica exchange, I calculated affinity differences resulting from the base substitutions enumerated above in the sequence-specific complexes of the five known telomeric proteins.

To explicitly calculate the affinity differences, I used a thermodynamic cycle in which the lesion was alchemically introduced in both free solvated DNA and in a solvated

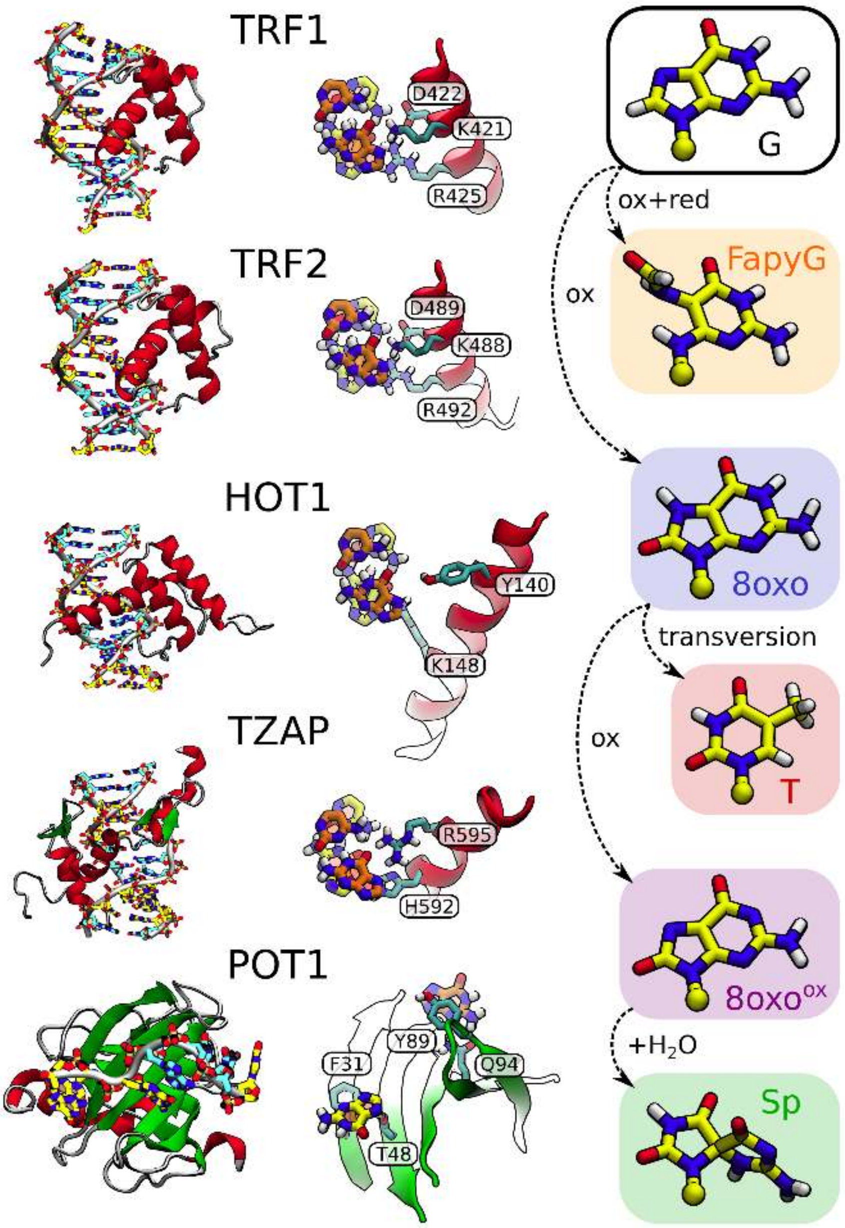


FIGURE 4.12: Known human telomeric DNA-binding proteins. Amino acids contributing to the recognition of the two guanines susceptible to oxidation within the telomeric repeats are shown explicitly and labeled. Alpha-helices are shown in red and beta-sheets in green. In the right panel, all oxidative lesions considered in this study are shown, along with simplified paths that lead to their formation. Color coding of lesions is kept consistent throughout the study.

protein-DNA complex, so that the former value of ΔG served as a reference. This in turn required that accurate high-resolution structures of the respective complexes be known. While crystal structures of four telomeric proteins – TRF1, TRF2, POT1, and HOT1 – had been available at the time this project was started, the structure of TZAP was released only recently [73]. Nevertheless, the available biochemical data allowed to identify the C-terminal zinc-finger domain as the one responsible for sequence specificity at telomeres, with the last 3 domains (9-11) sufficient to reconstitute the full activity of TZAP at telomeres [77]. For this reason, I used the canonical zinc-finger protein ZNF268 as a template for homology modelling, and combined the use of Modeller with extensive MD-based refinement to obtain a well-equilibrated model of the complex of TZAP with telomeric DNA. In the refined model, strong and stable hydrogen bonds were maintained between the protein side chains (R589, H592, R595) and a run of three guanines in a major groove, so that the three planar side chains stacked upon each other. Once the experimental structure became available [78], I verified that the homology-based model indeed represents the bound state with high fidelity, as it was able to stably maintain a backbone heavy-atom RMSD of 0.2 nm when the crystal structure was used as a reference (see Fig. 3.5).

It was shown experimentally that in dsDNA, guanines in the 5′-GGG-3′ triplet are not uniformly prone to oxidation, with the 5′ and central positions being much more frequently oxidized than the 3′ one [134]. To curb the computational cost of the combinatorically exhaustive procedure, the 3′ position was hence excluded for the computation. A full set of combinations involving the remaining two susceptible sites, 5 telomeric proteins and 5 oxidative lesions yield a total of 50 independent estimates of affinity change ($\Delta\Delta G$), reported in Fig. 4.13.

When discussing changes in affinity of sequence-specific proteins, it is convenient to put such values in the context of the thermodynamic binding specificity of the protein, i.e. the difference between the affinity to target and typical off-target sites. For this reason, I estimated the corresponding values from existing literature where possible, noting that in most cases they fall within the range between 1 and 3 kcal/mol. If the estimated affinity changes (shown as bars in the histogram in Fig. 4.13) are higher than the corresponding dashed black line, it should be expected that the lesion-containing target is seen by the protein as an off-target sequence, i.e. the specific binding is abolished.

In the histograms, the mostly positive values of $\Delta\Delta G$ show that the presence of virtually all lesions decrease the affinity for the target sites, even within the broad uncertainty estimates. However, not all proteins are equally sensitive to structural changes on the binding interface. Here, the affinity change is almost always the largest in case of TZAP, reflecting the presence of a densely packed and highly cooperative sequence recognition interface [78], but also highlighting the quality of the homology-based structural model. On the other extreme is the case of HOT1, a protein that makes relatively few sequence-specific contacts with the G-triplet [6]. As a result, the estimated affinity changes are small to non-existent, and despite I found no data in the literature to directly estimate the specific affinity, structurally similar proteins have been consistently found to bind to target sequences preferentially by a 2-3 kcal/mol margin [303, 304]. Were that the case, the only lesion that would consistently disrupt the binding of HOT1 is the bulky Sp. In fact, among all analyzed lesions Sp has the most dramatic effect on binding affinity, with a mean $\Delta\Delta G$ equal to 5.3 kcal/mol, reflecting the large impact it has on the tightly optimized binding interface. Even though the remaining lesions also make binding more unfavorable, the effects of T, FapyG, 8oxoG and O8oxoG are much smaller, averaging 2.6, 2.1, 2.0 and 1.0 kcal/mol in respective order.

Notably, almost all lesions – except for the central O8oxoG (O8oxoG-2) and the 5′-positioned thymine (T-1) – abolish the sequence-specific binding of the two major shelterin components, TRF1 and TRF2, potentially acting as exit ramps for the shelterin diffusing along the DNA strand, according to a mechanism discussed in the previous chapters. Such an affinity decrease agrees well with experimental values of

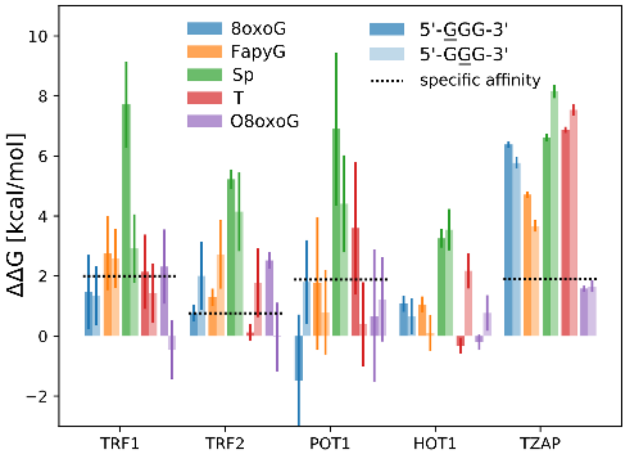


FIGURE 4.13: Changes in affinity of telomeric proteins for telomeric DNA resulting from the presence of oxidative lesions. Bar heights show the calculated $\Delta\Delta G$ values, with solid and semi-transparent bars corresponding to different position of the lesion within a G-triplet. Horizontal dashed lines visualize the estimated difference between the affinity to target and off-target sequences for individual proteins. Error bar heights are calculated based on the standard deviation of independent $\Delta\Delta G$ estimations in several independent calculations of chemically identical systems.

ca. 1 kcal/mol reported for both TRF1 and TRF2 when the central guanine was substituted with 8oxoG [152]. However, the design of the *in vitro* study arguably did not allow to interpret the results on an atomistic level, as potential contributions from multivalent binding and protein dimerization were not controlled for explicitly.

In case of POT1, two processes have to be considered simultaneously to obtain a comprehensive picture, as the binding of POT1 to its telomeric ssDNA target competes with the formation of GQ structures. For this reason, the outcome of any *in vitro* study involving base lesions will be affected by two components: (i) the effect of the lesion on the stability of the GQ itself, and (ii) an analogous effect on the stability of the POT1-ssDNA complex. Regarding the former, it was shown that 8oxoG in the central position strongly destabilizes the telomeric GQ [305], but had a much more modest effect on GQ stability when found in the 5′ position [306]. Correspondingly, the small (< 1 kcal/mol) favorable effect of the central 8oxoG on the binding of POT1 to telomeric ssDNA observed *in vitro* likely resulted from higher accessibility of ssDNA due to a decreased likelihood of GQ formation, offset by a slightly less favorable binding of POT1 to the unfolded ssDNA substrate (see Fig. 4.13). In a similar way, the two small effects of 8oxoG should roughly cancel in the 5′-position, yielding the experimentally observed stabilization of the POT1-ssDNA complex of ca. 1 kcal/mol. Across other lesions, the experimentally observed effect would necessarily depend on their yet unknown effect on the stability of the telomeric GQ, and only Sp and T-1 appear to introduce a sufficient hindrance to binding so as to prevent the formation of a POT1-ssDNA complex in an experimental setting.

In order to present a more intuitive and structure-based interpretation of the observed affinity changes, I subjected the alchemical trajectories to feature selection to identify molecular descriptors whose distribution shifts correlated strongly with $\Delta\Delta G$ estimates. The top panel of Fig. 4.14 illustrates the highest-scoring descriptors, ranked by the respective correlation coefficients. The highest correlation is observed

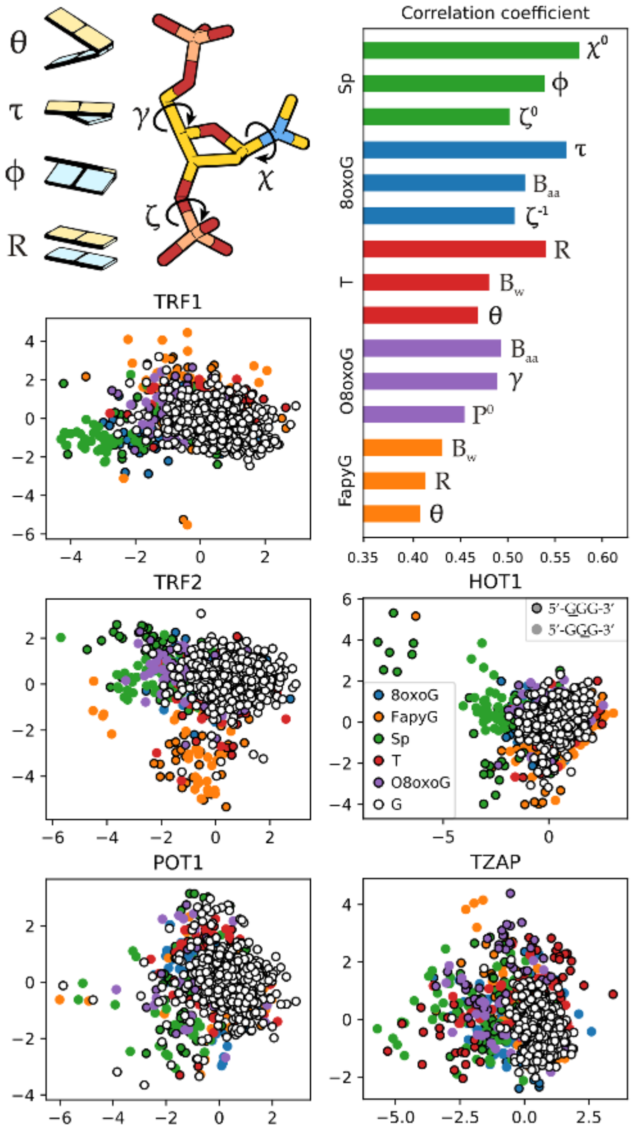


FIGURE 4.14: LDA analysis. Representative points from the alchemical trajectories are projected onto the plane that best separates individual chemical moieties. Original data corresponds to hydrogen bonding patterns in simulations of protein-DNA complexes. White markers correspond to undamaged DNA, while colored ones correspond to lesions in specified positions (see legend).

for the glycosidic angle rotation of Sp, χ, suggesting that the perturbed syn-anti equilibrium of the bulky lesion corresponds significantly to the perturbation of the protein-DNA interface, consistently with previous studies reporting altered conformational preferences of hydantoin derivatives [307]. This change is accompanied by a shift in the roll angle (ϕ) as well as the local backbone conformation (ζ). Interestingly, while the other bulky lesion – FapyG – also induces local changes in the helix parameters, most notably rise (R) and tilt (θ), these changes correlate less with the effect on binding, and are even overshadowed by the change in base solvation (B_w). The observed significant change in sugar puckerings (see Fig. 4.17), though, apparently do not translate onto affinity changes.

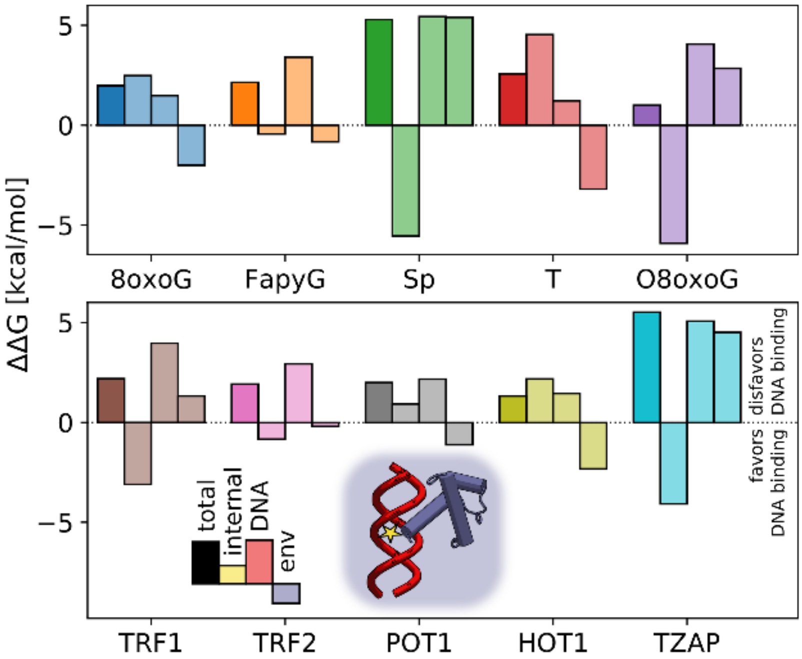


FIGURE 4.15: Free energy decomposition. The histograms illustrate contributions to $\Delta\Delta G$ that originate from individual subsystems, illustrated in the schematic picture in the bottom panel. In the top panel, results are averaged by lesion, and in the bottom panel averaging is performed protein-wise. As noted in the figure, positive contributions are indicative of decreased DNA-binding propensity in the presence of lesions, and *vice versa*.

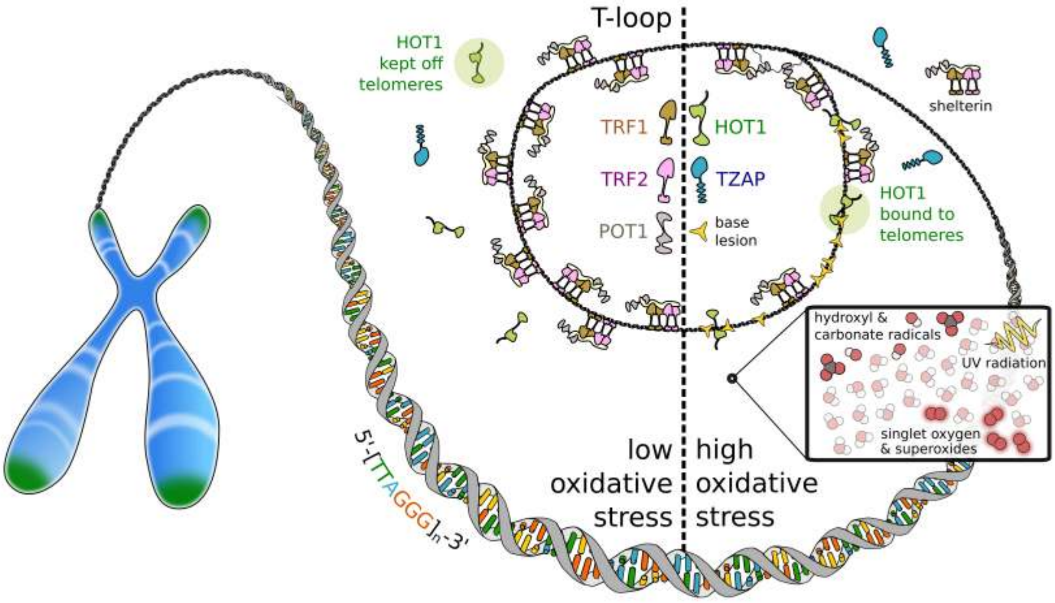


FIGURE 4.16: A simple model for the hypothesized modulation of telomeric DNA damage response by HOT1. When telomeres are intact, high occupancy of the shelterin complex exerts a protective effect. In presence of oxidative damage, shelterin occupancy is reduced, exposing binding sites for HOT1. Note that for simplicity, the presence of telomeric nucleosomes and DNA repair factors is omitted.

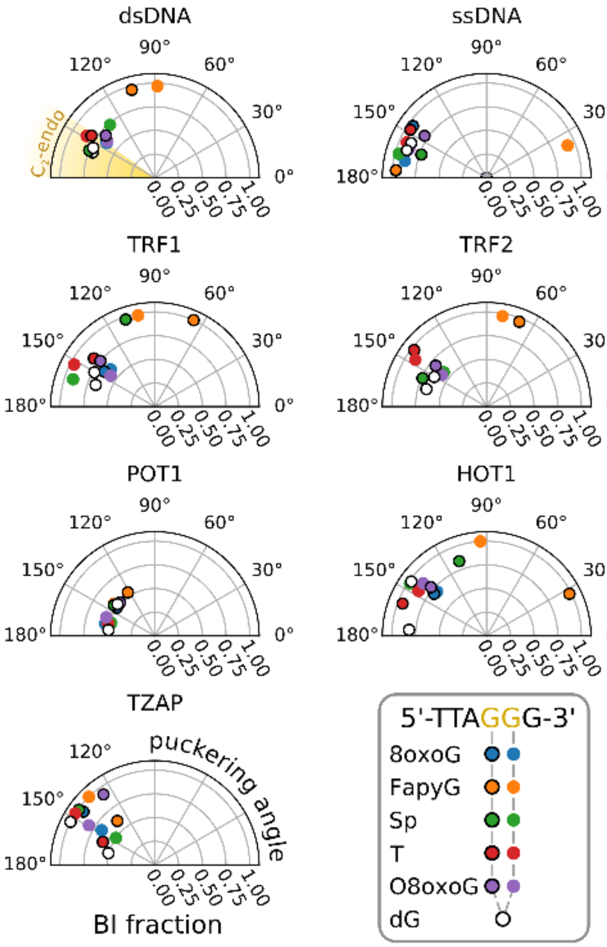


FIGURE 4.17: Mean sugar puckering angles observed in simulations of individual DNA and protein-DNA systems (angular coordinate), along with the mean BI-BII fraction (radial coordinate). The transparent yellow section illustrates the typical range of puckering angles in regular dsDNA.

The planar 8oxoG and O8oxoG are characterized by different hydrogen bond donor/acceptor patterns than guanine itself, and correspondingly affect the binding by altering hydrogen bonding with amino acids; however, 8oxoG also affects protein binding by altering the twist angle τ and the backbone angle ζ that defines the BI/BII ratio in DNA, consistently with previous findings regarding the structural properties of 8oxoG in the context of free dsDNA [308, 309]. Similarly, sugar puckering and the C4′-C5′ torsion appear as important determinants of O8oxoG-dependent affinity changes. Finally, the curious case of thymine – a naturally occurring base – seems to introduce perturbations similar to that of much bulkier FapyG, perturbing the DNA structure itself. This raises an interesting possibility that DNA structure can be perturbed not only when bulky lesions cannot be accommodated in B-DNA, but also through frustrations on the protein-DNA interface when native contacts are not formed.

To visualize the distributions of generalized structural features, I initially used PCA, but the dominant variability turned out to be dominated by large-scale motions unaffected by the presence of lesions, so that PCA did not resolve systems containing individual lesions (see Fig. 4.18). For this reason I then used LDA, a procedure that projects the data on a plane that best separates individual data classes when data is categorical. The scatter plots shown in the bottom of Fig. 4.14 visualize the changes in the respective equilibrium distributions of collective descriptors based on top-scoring features, with Sp (green circles) standing out the most among individual lesions, and TZAP among the studied proteins. In most cases, however, even the optimal LDA projections cannot separate individual distributions, indicating that inter-class variance is typically larger than intra-class variance. I also found that a direct quantification of overlaps (Bhattacharyya distances) between Gaussian-smeared distributions does not strongly correlate with changes in affinity, suggesting that such changes cannot be trivially predicted from perturbation of the DNA structure alone.

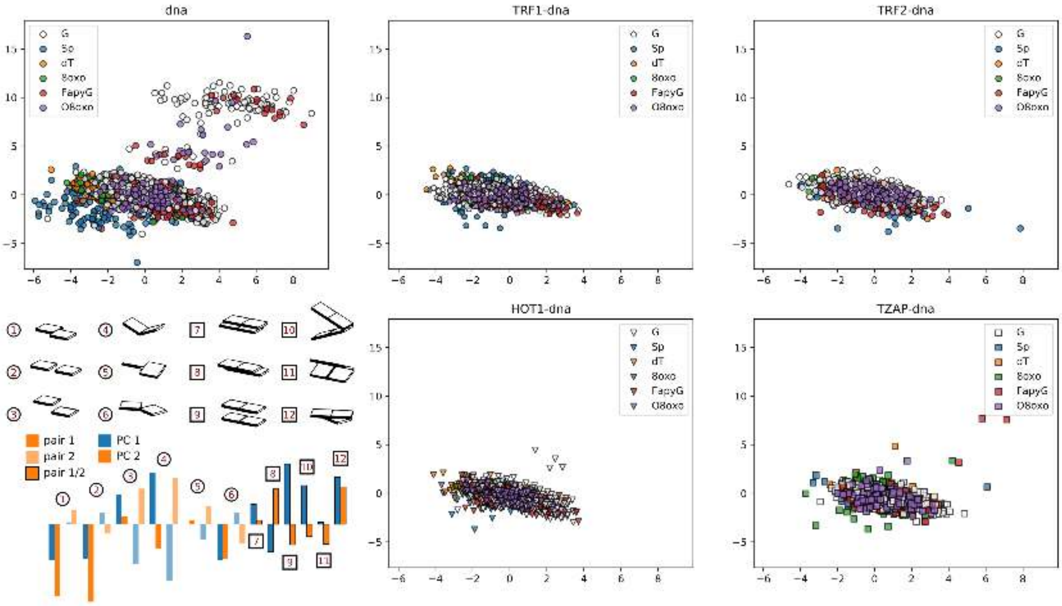


FIGURE 4.18: Results of the PCA performed on a set of X3DNA-derived structural descriptors of the two nucleotides immediately adjacent to the modified base, illustrated schematically in the bottom left panel. The descriptors' contributions to the two principal components are schematically depicted as orange and blue bars.

The lack of robust predictive power indicates that a structure-based description only provides a partial insight into the molecular origin of the estimated changes in affinity. For this reason I complemented it with a free-energy based analysis in which I decomposed the overall $\Delta\Delta G$ values into contributions coming from individual subsystems. Using the additive property of classical force fields, it was possible to reevaluate the alchemical free energies from trajectories that involved subsets of the original molecular system: (a) the modified nucleotide itself, (b) the remaining part of the DNA molecule, and (c) the environment composed of water and protein molecules. The subsystems (a), (b) and (c) are colored yellow, red and blue in the scheme in Fig. 4.15, in respective order. The resulting values should be interpreted as follows: positive values indicate that the interaction between the given subsystem and the lesion prevents the formation of a protein-DNA complex, while negative contributions – conversely – favor protein-DNA binding when the lesion is present, as indicated in the figure. In order to make the discussion more general, values are averaged either by the lesion (top panel) or by the protein (bottom panel), so that only mean contributions are shown.

As seen in the resulting histograms, the contribution from the DNA environment, i.e. the nearby nucleotides, is consistently positive in each case. This indicates that the DNA molecule containing a lesion cannot fully relax when the protein is bound, consistently with a hypothesis that protein binding considerably restricts the local structural flexibility of DNA. Such a conclusion could be drawn independently from the PCA results shown in Fig. 4.18, as the other structural mode detected by PCA was essentially repressed in the presence of proteins. In line with this reasoning, the said contribution is notably smaller in case of lesions that do not perturb $\pi - \pi$ stacking and Watson-Crick pairing, thymine and 8oxoG (1.2 and 1.5 kcal/mol, respectively). Simultaneously, the two other contributions – internal, i.e. coming from the modified nucleotide, and the one due to the protein/solvent environment – roughly cancel each other in most cases, so that their combined magnitude only exceeds 1.5 kcal/mol twice. This in turn can be interpreted by noting that while in the presence of the protein the introduction of the lesion becomes less favorable, the penalty itself will be compensated by a relaxation of the protein-DNA interface, and *vice versa*. Altogether, the observed effect indicates that (a) the protein can locally adjust to local structural changes introduced by the presence of the lesion, while (b) such an adjustment in case of the DNA environment is impossible due to structural restraints of the DNA double helix, additionally rigidified by the binding partner.

Ultimately, the above results should be put in a proper biological context. In the previous chapters, I discussed how the depletion of individual telomeric proteins initiates different downstream signaling patterns. Specifically, the repression of telomeric localization of TRF2 and POT1 activates ATR or ATM, two distinct DDR pathways that can result in senescence or initiate programmed cell death [310]. The depletion of TRF1 does not induce DDR signaling, but was shown to result in abnormal elongation of telomeres as well as telomere fragility [28]. Finally, while the biological function of the extrashelterin factors TZAP and HOT1 is not well established, the articles that reported their discovery claimed that they exert opposite effects on telomere integrity: while HOT1 enhances the recruitment of telomerase to elongate the telomere, the binding of TZAP resulted in telomere trimming and the appearence of extrachromosomal telomeric DNA in the form of C-circles [6, 311].

Given the estimated changes in affinity, the presence of oxidative lesions in telomeric DNA should exert a mostly similar, negative effect on the occupancy of the major shelterin components (TRF1 and TRF2) on telomeres, regardless of the mechanism of oxidation and, consequently, the prevailing lesion. Out of all considered lesions, only a G→T transversion at the 5′-position (T-1) would actually exert a differentiating effect, leaving TRF2 largely unaffected while promoting the detachment of TRF1. There is, however, no known oxidation pathway that would consistently and selectively yield G→T transversions, so that such a situation is extremely unlikely to be observed *in vivo*. In contrast, the impact of lesions on the binding of POT1 is somewhat ambiguous due to effect on the competing GQ formation; however, it should be noted that the telomeric localization of POT1 is largely dependent on the presence of the shelterin, as was discussed previously.

Finally, the relative populations of the extrashelterin proteins TZAP and HOT1 on telomeres would be drastically affected in favor of HOT1 if lesions other than Sp were present in DNA. It could then be postulated that HOT1, a still poorly characterized protein, may become more abundant on oxidized telomeres by occupying sites to which TRF1 and TRF2 cannot bind, potentially playing a signaling or protective function. Given its relatively late discovery and known extratelomeric functions, HOT1 cannot be very abundant on undamaged telomeres, so that such a replacement resulting from the presence of oxidative lesions would increase the telomeric population of HOT1 by a large factor even if the absolute number of lesions was small. This would hence provide a sensitive feedback, constituting a sensor-like mechanism illustrated in Fig. 4.16. It is interesting to note that despite structural similarity, the DBDs of TRF1 and HOT1 evolved independently to bind at slightly shifted sites along the same telomeric sequence, thus rendering HOT1 less susceptible to oxidative changes in the guanine triplet [6]. While this hypothesis has yet to be confirmed or rejected by experimental studies, it finds some support in past reports linking HOT1 with the formation of TIFs as well as the regulation of apoptosis [74–76], highlighting the need for further research in this area.

4.6 Protein-DNA Cross-Links as a Possible Mode of Oxidative Damage on Telomeres

Since the late 1990s, reports have been emerging that show that DNA oxidation not only affects nucleobases but also DNA-bound proteins. Specifically, the presence of single-electron oxidants – such as UV-irradiated riboflavin – has been shown to efficiently generate covalent protein-DNA cross-links between the nucleophilic N_ζ atom of lysine (sometimes also the hydroxyl oxygen of tyrosine [312]) and the electrophilic centers in the guanine moiety, atoms C5 and C8. So far, however, the only study that reported and verified the occurrence of covalent lysine-guanine linking in a realistic protein-dsDNA complex were performed on the MutY protein back in 1999 [170], and protein-specific studies have been scarce ever since, favoring instead model systems consisting of oligonucleotides and oligopeptides [173, 313] or otherwise non-specific DNA binders [314].

Surprisingly, the questions of cross-linking and the exceptional susceptibility of telomeric DNA to oxidative damage have never been investigated, even though three of the four known telomeric dsDNA-binding proteins contain a nucleophilic amino acid in direct contact with the guanine on the 5′ side of the triplet (the conserved K421 in TRF1 and K488 in TRF2, as well as Y327 in HOT1). Although the origins of oxidative damage at telomeres has been characterized experimentally before, these studies often rely on DNA digestion, extraction and mass-spectrometric analysis of the resulting oligonucleotides, or immunological or glycosylase-and-sequencing-based assays [177, 315], so that protein-DNA cross-links could easily evade analysis if not investigated explicitly. For this reason, I used QM/MM free energy methods to investigate the feasibility of cross-link formation in a TRF1-DNA complex using different oxidation states of the nucleobase as a reactant.

In modelling of lysine-guanine cross-links, several issues have to be discussed to understand the complexity of the task. Firstly, with the pKa of lysine side chain of ca. 10.5, the residue remains in its non-nucleophilic protonated state for most of the time. This introduces a free energy penalty of $-RT\ln\left(10^{10.5-7}\right) \approx 4.8$ kcal/mol associated with the initial deprotonation step that makes the reaction feasible. Arguably, the penalty can also be modified by the molecular environment – increased by the presence of the charged DNA backbone, or decreased by the neighboring amino acids, either positively charged or capable of acting as temporary proton acceptors. Secondly, the identity of the actual reactive nucleobase intermediate that undergoes the coupling reaction is not well established, but the initial study that used a mild

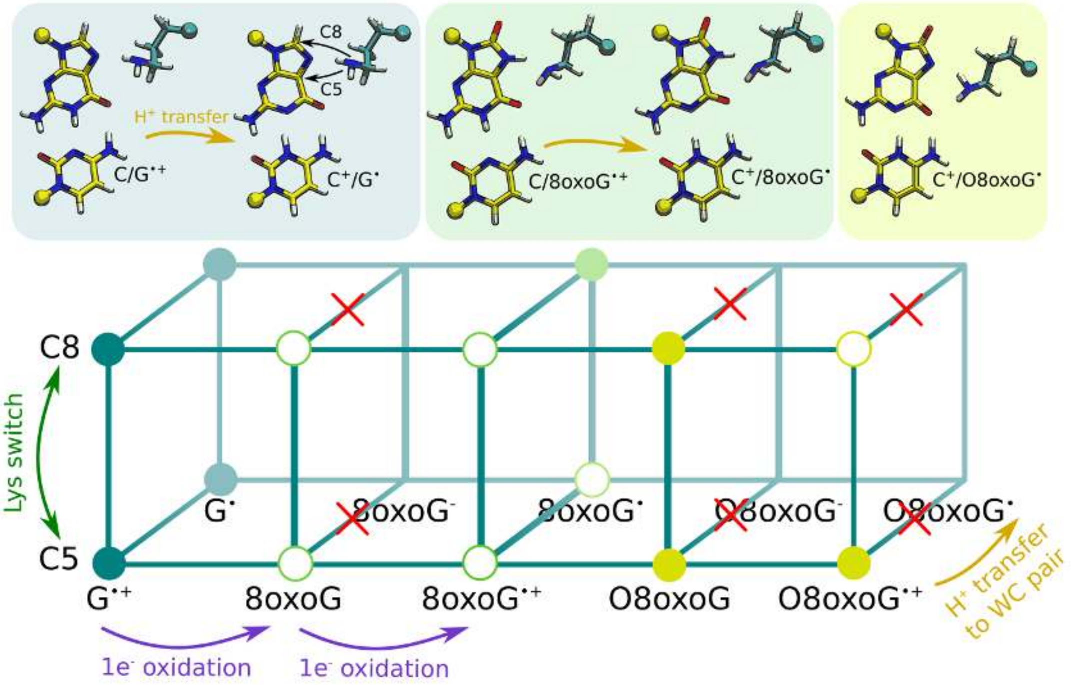


FIGURE 4.19: A graph of possible reactants involved in the committed step of lysine-guanine cross-link formation, along with a pictorial representation of the reactants' structures at individual steps. Red crosses mark paths that were *a priori* considered unfeasible. Full circles mark systems for which free energy profiles were obtained.

iridium-based oxidant indicated that 8oxoG and its oxidation products are involved along one of the feasible pathways [170]. A recent quantum chemical study suggests the key role of two-electron oxidized 8oxoG, O8oxoG, but relies on minimal models that do not take into account the spatial restraints and environment effects of dsDNA [172], as well as does not show which criteria render the cross-linking feasible in actual protein-DNA systems. Finally, a pulse radiolysis study revealed that the guanine cation radical deprotonates rapidly to yield a (8oxo-)G$^{\cdot}$:C+ Watson-Crick pair that slowly releases the excess proton into the solvent [316]. This situation is likely to be observed also in 8oxoG [317], provided that deprotonation at N7 does not outpace that at N1 [318]. To further complicate the issues, the presence of a proton on N7 would facilitate the deprotonation of the major groove-bound lysine, while protonated lysine would promote deprotonation of N7; unfortunately, no data exists currently to support either mechanism in an actual Watson-Crick-paired DNA. The relationship between individual pathways, as well as a scheme of the reactants, is illustrated in Fig. 4.19.

Due to limited computational resources, the QM subsystems employed in my QM/MM simulations were restricted to two nucleotides (5′-guanine and the adjacent adenine), the deprotonated side chain of lysine (starting with the β carbon) and five water molecules located near the protein-DNA interface. Starting with properly parametrized and equilibrated MM models, I selected frames with low Nζ-C5 distances for 1 ps QM/MM equilibration, 10 ps steered QM/MM MD simulations in which the respective N-C distance was decreased using a harmonic restraint, and subsequent >2 ps long umbrella sampling runs. In total, the procedure was repeated for eight individual systems, marked with full circles in the bottom panel of Fig. 4.19.

In turn, empty circles refer to systems that were not investigated after the pathway was deemed unproductive by existing results, with the assumption that the free energy difference between C5/C8 adducts and protonated/deprotonated N1 guanine should be no larger than several kcal/mol.

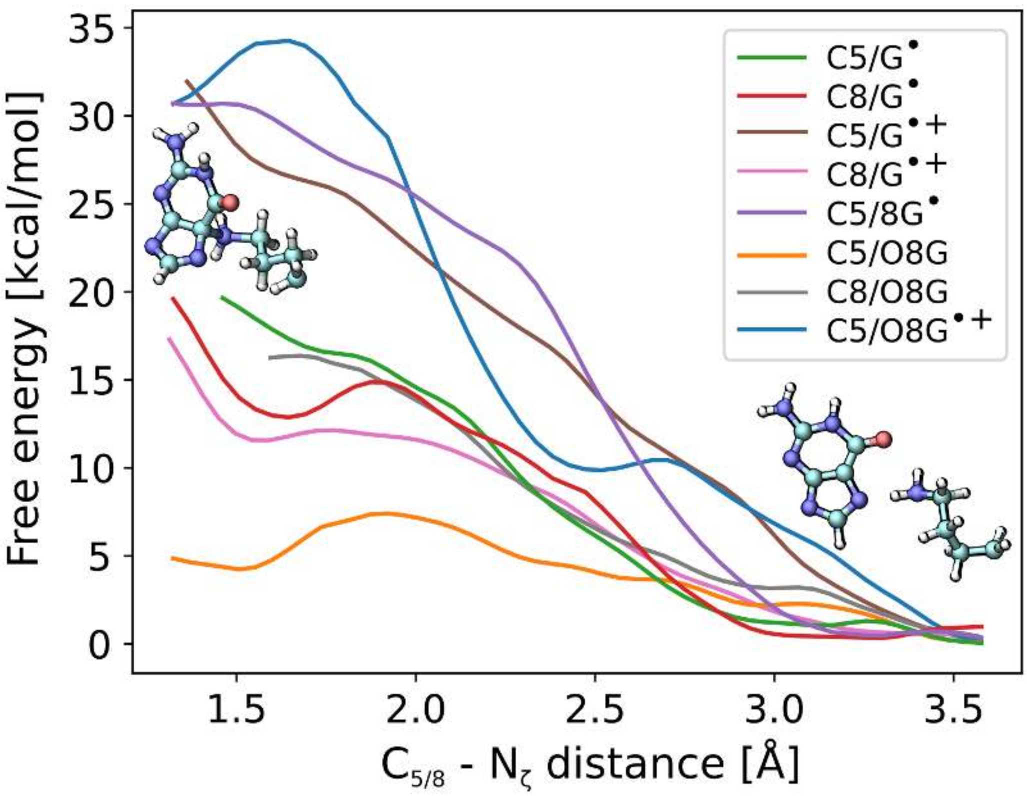


FIGURE 4.20: QM/MM MD/US-derived free energy profiles for the formation of a C-N bond in all systems considered in the study. The schematic atomistic representations mark the initial (unbound) and final (bound) states.

The obtained free energy profiles are shown in Fig. 4.20. In general terms, systems can be visually divided into three classes based on the overall shape and height of the profile: (a) these that require ca. 30 kcal/mol to reach the bound state, mostly without a well-defined minimum (guanine cation radical, 8oxoG radical and O8oxoG cation radical, all reactive at position C5); (b) these that require 12-17 kcal/mol to reach the bound state, where some of the can actually become metastable within a shallow free energy minimum (most notably the guanine cation radical and radical reactive at positions C8); and (c) the single case of neutral O8oxoG, for which the profile exhibits a free energy barrier of 7-8 kcal/mol and a minimum in the bound state at 4 kcal/mol. Minding the ca. 4 kcal/mol penalty associated with Lys deprotonation, one might conclude that the last system is the only feasible intermediate in the committed step of cross-link formation, in agreement with the recent computational study [172].

Upon closer inspection, however, it turned out that the formation of the C-N bond in the C5/O8oxoG system triggers an opening of the 5-membered purine ring, running counter to the postulated mechanism. I hypothesized that this effect results from a delay in proton transfer due to the fast enforced formation of the C-N bond. Hence, to adequately sample all three relevant variables – bond formation, proton transfer and opening of the 6-membered ring – in a controlled manner, I re-performed the free energy calculations using QM/MM-MD multiple walker well-tempered 3D metadynamics, adding restraints on the bond broken in the unproductive pathway. To enhance the sampling along the pathway leading to the final Sp-like product, the bond opening CV was defined as the difference between C5-C6 and C4-C6 distances, yielding positive values when the molecule was planar (C5-C6 bond), and negative in the non-planar state in which the C4 spiro carbon is bound to C6.

Fig. 4.21A illustrates two side views of the 3-dimensional free energy plot, with six isovalues shown simultaneously as semi-transparent surfaces. To facilitate interpretation, panels B-D of the figure show 2-dimensional *conditional* free energy maps, conditioned on the progress along the CV that was integrated out, and split into fields to guide the reader. It should be pointed out that due to order-of-magnitude differences in the free energy estimates, different color scales were used in the rightmost plots of panels B-D. From the rightmost plot in panel B, it can be observed that prior to the proton transfer, the formation of a C-N bond is almost barrierless and favorable, with a free energy difference of ca. 6 kcal/mol between the minima at 2.2 Å and 1.55 Å in the near-planar configuration. If the proton transfer does not take place, ring opening becomes feasible and somewhat (2 kcal/mol) thermodynamically favorable, yet is largely slowed down by a 6 kcal/mol free energy barrier. Proton transfer allows the reaction to proceed to completion (leftmost panel), leading to the low-energy product state (green circle) through a short-lived open-ring intermediate.

The proton transfer itself is thermodynamically favorable – yet with a considerable free energy barrier – when the C-N bond is not yet formed, as shown in the rightmost map in Fig. 4.21C. Consistently, Fig. 4.21D indicates that in the near-planar structure, proton transfer becomes increasingly favorable as the C-N bond forms, even though the barrier is still present. Finally, the Sp-like product state is stable, with a single deep minimum in the left panel of Fig. 4.21D.

I shall note that the surprisingly high estimate of the reaction free energy might represent an artifact of the computational procedure, likely due to an ambiguous identity of states characterized by distance *differences* instead of distances alone. The corresponding DFT estimate of enthalpy of the first joint reaction step (bond formation and proton transfer) in a minimal QM system and with implicit solvent was somewhat similar, on the order of 30 kcal/mol; however, transition to the Sp product was only slightly favorable, by ca. 5 kcal/mol [172]. This quantitative discrepancy will have to be addressed in future investigations; however, the qualitative prediction regarding the overall feasibility of the reaction will soon undergo the most stringent test, i.e. experimental verification in an *in vitro* system.

4.7 Conclusions

In my doctoral work, I used a broad range of Molecular Dynamics-based methods to shed light on the molecular aspects of protein-DNA interactions on telomeres. Through an *in silico* mutagenesis assay, I showed how biomacromolecules use two classes of residues – here termed positive and negative selectors – to not only attain high affinity towards the target, but also avoid spending too much time on off-target sites. A bioinformatic analysis suggested that the surprisingly frequent presence of

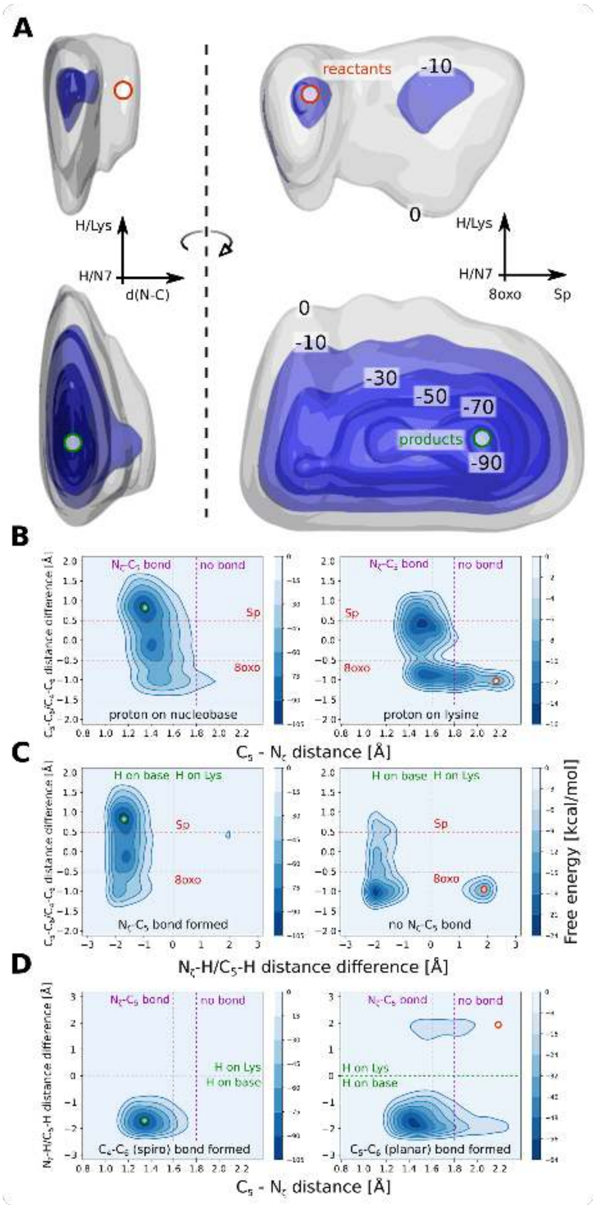


FIGURE 4.21: (A) Two side views of the 3-dimensional free energy map plotted using VMD, with 6 isosurfaces corresponding to individual free energy values. Points corresponding to reactants' and products' structures are marked with red and green circles, respectively. Reaction coordinates are labeled schematically, with 8oxo→Sp corresponding to the transition between a near-planar and a spiro atom-centered bicyclic structure, and H/N7→H/Lys to the proton transfer between the nucleobase and the lysine side chain. (B-D) 2-dimensional projections of slices of the 3D free energy map ("conditional" free energy maps). The black label indicates which part of the 3D volume was selected as the condition (e.g. "Nζ-C5 bond formed" means that only the region in which d(C-N) was lower than 1.6 Å), while the colored labels shall help the reader distinguish between individual regions of the 2D surface.

negatively charged residues on different protein-DNA interfaces can be explained in these terms. Further analysis revealed that a similar functionality is conferred by the positively charged tail residues, only providing a strong anchoring interaction at selected subsequences to foster more precise sampling, and keeping the protein domain mobile when sampling would be unproductive anyway. This results strongly indicate that biomacromolecules employ a multitude of weakly discriminating checkpoints along the native complex formation pathway, having evolved clever strategies to achieve an advantageous compromise between mobility at off-target sites and affinity at the target. Such a notion could also help refine existing simulation-based models of target search on DNA given the existing discrepancies in the theoretical understanding of the issue.

Another key insight comes from the construction of the first complete 2-dimensional free energy maps of protein-DNA interaction along both a repetitive target and off-target sequences, validated against experimental results of single-molecule measurements and in good agreement with theoretical constraints. Along with the accompanying analyses, the map shows how the rotation-coupled sliding features discrete steps, and how the free energy barriers and basins emerge from specific changes in hydrogen bonding patterns. It also explains the greater mobility at off-target DNA in terms of changes in the mode of interaction and lack of pronounced free energy barriers. Interestingly, the map predicts the existence of more free energy minima for TRF1 than the one represented by the crystal structure; while speculative, this result has some merit in structural terms, and could perhaps be experimentally validated in the future. Finally, Brownian dynamics simulations based on the free energy maps reveal marked differences in rotation-translation coupling between the target telomeric and off-target sequences, again highlighting the increased translational and orientational dynamics of the domain in the search mode.

The observation of a complete spontaneous reconstitution of a sequence-specifically bound TRF1-DNA complex was a major milestone in my doctoral work, as it allowed for an unprecedented insight into the actual sequence of events that results in complex formation. This includes such issues as appropriate positioning and orientation at the interface *via* the initial anchoring contacts, sequential formation of nucleobase-specific hydrogen bonds, and major (re-)binding events that relieve the structural stress in improperly formed, pre-bound conformational states. In parallel, the analysis of unproductive simulations helped identify the committed or rate-limiting steps of complex formation, as well as verify that the postulated checkpoints are actually operative when the sequence context is different than desired. I shall also note that the successful re-creation of a crystallographically determined structure in equilibrium MD simulations marks a formidable achievement of force field developers, showing again (and reinforcing the hope) that modern force fields have attained true predictive power, and that the results of MD simulations – while certainly not always to be taken at face value – at least remain credible as long as are supported by extensive sampling (here totaling several hundreds of microseconds) and parallel experimental efforts.

Subsequent analyses of all equilibrium simulations, based on the concept of transfer entropy and the Markov state model framework, largely reaffirmed the above conclusions, revealing clusters of "early binders" that interact strongly with the DNA backbone during sequence search as well as helping identify a set of (sometimes long-lived) intermediates that arise along the complex formation pathway. By combining all trajectories in a single model, I was able to estimate the kinetics of unbiased complex formation, as well as the rates of interconversion between individual orientations of the protein with respect to DNA.

By investigating the effect of presence of oxidative lesions on the affinities of known telomeric proteins for their DNA targets, I confirmed that virtually any nucleobase modification is sufficient to abolish sequence-specific binding to a given DNA site, yet with magnitudes of this affinity changes following well-defined trends. As I found out, this effect partially results from the rigidity of the protein-bound DNA, so that the binding partner does not allow the DNA strand to relax and properly accommodate the lesion in its structure. However, while bulky lesions such as Sp introduce a large thermodynamic penalty for binding, they are also easy to locate and repair due to their easily identifiable impact on DNA structure. Less structure-distorting lesions – such as 8oxoG or the G→T transversion – might be more persistent, especially on nucleosome-bound telomeric DNA, thereby contributing more to any biological effects caused by partial deprotection of telomeres.

Ultimately, the biological relevance of this "epigenetic" oxidative modulation depends on the quantity of nucleobase lesions, known to be rather scarce even in high oxidative stress. I propose that already a small population of telomeric lesions can yield a detectable signal by promoting local binding of HOT1 in place of the shelterin components TRF1 and TRF2. Indeed, my results show that HOT1 – whose telomere-bound population is rather small under normal circumstances – is much less affected by the presence of lesions than other telomeric proteins. While this remains a hypothesis, HOT1 has already been implicated in a range of apoptosis-related functions in past reports, so that this functional connection should be further investigated.

The results of my QM/MM MD free energy simulations seem to independently confirm the mechanism of lysine-guanine cross-link formation postulated recently by the research group of Burrows. Out of eight tested intermediates involving guanine, 8oxoG and O8oxoG as either neutral species, radicals or cation radicals, only neutral O8oxoG was found to be sufficiently susceptible to nucleophilic attack by K421 of TRF1 at the C5 carbon atom to make the reaction thermodynamically and kinetically feasible. Notably, this cross-link formation was simulated for the first time in a realistic model, involving an inhomogeneous reaction environment and the actual structural constraints of DNA. The feasibility of the reaction was further explored using 3D metadynamics, and the results suggest that once multiple structural changes are allowed to occur in concert, the reaction becomes even more favorable in both thermodynamic and kinetic terms, even despite minor artifacts of the method that need to be addressed in further studies. Combined, the results provide a strong indication that the proposed mechanism is operational under physiological conditions. Experimental studies that are already underway will help determine whether (a) covalent lysine-guanine cross-links actually form in a reconstituted TRF1-DNA complex *in vitro* and (b) whether this mode of oxidative damage contributes significantly to the population of DNA lesions on telomeres given the susceptibility of telomeric DNA to oxidation.

Bibliography

[1] T. de Lange, "Shelterin: the protein complex that shapes and safeguards human telomeres.", Genes & development **19**, 2100–10 (2005).

[2] J. W. Shay and W. E. Wright, "Hayflick, his limit, and cellular ageing", Nature Reviews Molecular Cell Biology **1**, 72–76 (2000).

[3] E. Mladenov and G. Iliakis, "Induction and repair of DNA double strand breaks: The increasing spectrum of non-homologous end joining pathways", Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis **711**, 61–72 (2011).

[4] Y.-S. Cong, W. E. Wright, and J. W. Shay, "Human telomerase and its regulation.", Microbiology and molecular biology reviews : MMBR **66**, 407–25 (2002).

[5] D. E. MacNeil, H. J. Bensoussan, and C. Autexier, "Telomerase regulation from beginning to the end", Genes **7** (2016) `10.3390/genes7090064`.

[6] D. Kappei, F. Butter, C. Benda, M. Scheibe, I. Draškovič, M. Stevense, C. L. Novo, C. Basquin, M. Araki, K. Araki, D. B. Krastev, R. Kittler, R. Jessberger, J. A. Londoño-Vallejo, M. Mann, and F. Buchholz, "HOT1 is a mammalian direct telomere repeat-binding protein contributing to telomerase recruitment", EMBO Journal **32**, 1681–1701 (2013).

[7] C. Wang, L. Zhao, and S. Lu, "Role of TERRA in the regulation of telomere length.", International journal of biological sciences **11**, 316–23 (2015).

[8] T. H. D. Nguyen, J. Tam, R. A. Wu, B. J. Greber, D. Toso, E. Nogales, and K. Collins, "Cryo-EM structure of substrate-bound human telomerase holoenzyme", Nature **557**, 190–195 (2018).

[9] N. W. Cho, R. L. Dilley, M. A. Lampson, and R. A. Greenberg, "Interchromosomal homology searches drive directional ALT telomere movement and synapsis.", Cell **159**, 108–121 (2014).

[10] S. H. Yoshimura, H. Maruyama, F. Ishikawa, R. Ohki, and K. Takeyasu, "Molecular mechanisms of DNA end-loop formation by TRF2", Genes to Cells **9**, 205–218 (2004).

[11] G. J. Nora, N. A. Buncher, and P. L. Opresko, "Telomeric protein TRF2 protects Holliday junctions with telomeric arms from displacement by the Werner syndrome helicase", Nucleic Acids Research **38**, 3984–3998 (2010).

[12] C. W. Greider, "Telomeres do D-Loop–T-Loop", Cell **97**, 419–422 (1999).

[13] E. L. Denchi and T. de Lange, "Protection of telomeres through independent control of ATM and ATR by TRF2 and POT1", Nature **448**, 1068–1071 (2007).

[14] F. d. d. Fagagna, P. M. Reaper, L. Clay-Farrace, H. Fiegler, P. Carr, T. von Zglinicki, G. Saretzki, N. P. Carter, and S. P. Jackson, "A DNA damage checkpoint response in telomere-initiated senescence", Nature **426**, 194–198 (2003).

[15] H. Takai, A. Smogorzewska, and T. de Lange, "DNA damage foci at dysfunctional telomeres", Current Biology **13**, 1549–1556 (2003).

[16] A. Lechel, A. Satyanarayana, Z. Ju, R. R. Plentz, S. Schaetzlein, C. Rudolph, L. Wilkens, S. U. Wiemann, G. Saretzki, N. P. Malek, M. P. Manns, J. Buer, and K.

L. Rudolph, “The cellular level of telomere dysfunction determines induction of senescence or apoptosis in vivo.”, EMBO reports **6**, 275–281 (2005).

[17] A. Smogorzewska, J. Karlseder, H. Holtgreve-Grez, A. Jauch, and T. De Lange, “DNA ligase IV-dependent NHEJ of deprotected mammalian telomeres in G1 and G2”, Current Biology **12**, 1635–1644 (2002).

[18] H. Almeida and M. Godinho Ferreira, “Spontaneous telomere to telomere fusions occur in unperturbed fission yeast cells”, Nucleic Acids Research **41**, 3056–3067 (2013).

[19] N. Vargas-Rondón, V. E. Villegas, and M. Rondón-Lagos, “The role of chromosomal instability in cancer and therapeutic responses”, Cancers **10** (2018) `10.3390/cancers10010004`.

[20] R. Diotti and D. Loayza, “Shelterin complex and associated factors at human telomeres.”, Nucleus (Austin, Tex.) **2**, 119–35 (2011).

[21] K. K. Takai, S. Hooper, S. Blackwood, R. Gandhi, and T. de Lange, “In vivo stoichiometry of shelterin components.”, The Journal of biological chemistry **285**, 1457–67 (2010).

[22] B. van Steensel and T. de Lange, “Control of telomere length by the human telomeric protein TRF1”, Nature **385**, 740–743 (1997).

[23] W. Chang, J. N. Dynek, and S. Smith, “TRF1 is degraded by ubiquitin-mediated proteolysis after release from telomeres.”, Genes & development **17**, 1328–33 (2003).

[24] K. Ancelin, M. Brunori, S. Bauwens, C.-E. Koering, C. Brun, M. Ricoul, J.-P. Pommier, L. Sabatier, and E. Gilson, “Targeting assay to study the cis functions of human telomeric proteins: evidence for inhibition of telomerase by TRF1 and for activation of telomere degradation by TRF2.”, Molecular and cellular biology **22**, 3474–87 (2002).

[25] D. Loayza and T. de Lange, “POT1 as a terminal transducer of TRF1 telomere length control”, Nature **423**, 1013–1018 (2003).

[26] C. Y. Soohoo, R. Shi, T. H. Lee, P. Huang, K. P. Lu, and X. Z. Zhou, “Telomerase inhibitor PinX1 provides a link between TRF1 and telomerase to prevent telomere elongation.”, The Journal of biological chemistry **286**, 3894–906 (2011).

[27] S.-h. Kim, P. Kaminker, and J. Campisi, “TIN2, a new regulator of telomere length in human cells”, Nature Genetics **23**, 405–412 (1999).

[28] A. Smogorzewska, B. van Steensel, A. Bianchi, S. Oelmann, M. R. Schaefer, G. Schnapp, and T. d. Lange, “Control of human telomere length by TRF1 and TRF2”, Molecular and Cellular Biology **20**, 1659 (2000).

[29] B. R. Houghtaling, L. Cuttonaro, W. Chang, and S. Smith, “A dynamic molecular link between the telomere length regulator TRF1 and the chromosome end protector TRF2”, Current Biology **14**, 1621–1631 (2004).

[30] M. F. Kendellen, K. S. Barrientos, and C. M. Counter, “POT1 association with TRF2 regulates telomere length.”, Molecular and cellular biology **29**, 5611–9 (2009).

[31] D. Hockemeyer and K. Collins, “Control of telomerase action at human telomeres.”, Nature structural & molecular biology **22**, 848–52 (2015).

[32] J. Dai, M. Carver, and D. Yang, “Polymorphism of human telomeric quadruplex structures”, Biochimie **90**, 1172–1183 (2008).

[33] R. D. Gray, J. O. Trent, and J. B. Chaires, “Folding and unfolding pathways of the human telomeric G-quadruplex.”, Journal of molecular biology **426**, 1629–50 (2014).

[34] J. You, H. Li, X.-M. Lu, W. Li, P.-Y. Wang, S.-X. Dou, and X.-G. Xi, “Effects of monovalent cations on folding kinetics of G-quadruplexes”, Bioscience Reports **37**, BSR20170771 (2017).

[35]J. Abraham Punnoose, Y. Ma, M. E. Hoque, Y. Cui, S. Sasaki, A. H. Guo, K. Nagasawa, and H. Mao, "Random Formation of G-Quadruplexes in the Full-Length Human Telomere Overhangs Leads to a Kinetic Folding Pattern with Targetable Vacant G-Tracts", Biochemistry **57**, 6946–6955 (2018).

[36]E.-J. Uringa, J. L. Youds, K. Lisaingo, P. M. Lansdorp, and S. J. Boulton, "RTEL1: an essential helicase for telomere maintenance and the regulation of homologous recombination", Nucleic Acids Research **39**, 1647 (2011).

[37]O. Mendoza, A. Bourdoncle, J.-B. Boulé, R. M. Brosh, Jr, and J.-L. Mergny, "G-quadruplexes and helicases", Nucleic Acids Research **44**, 1989 (2016).

[38]S. M. Bailey, R. E. Verdun, C. I. Haggblom, and J. Karlseder, "Strand-Specific Postreplicative Processing of Mammalian Telomeres", Science **293**, 2462–2465 (2001).

[39]A. J. Zaug, E. R. Podell, and T. R. Cech, "Human POT1 disrupts telomeric G-quadruplexes allowing telomerase extension in vitro", Proceedings of the National Academy of Sciences **102**, 10864–10869 (2005).

[40]H. Hwang, N. Buncher, P. L. Opresko, and S. Myong, "POT1-TPP1 regulates telomeric overhang structural dynamics.", Structure (London, England : 1993) **20**, 1872–80 (2012).

[41]R. Litman Flynn, S. Chang, and L. Zou, "RPA and POT1: Friends or foes at telomeres?", Cell Cycle **11**, 652–657 (2012).

[42]R. L. Flynn, R. C. Centore, R. J. O'Sullivan, R. Rai, A. Tse, Z. Songyang, S. Chang, J. Karlseder, and L. Zou, "TERRA and hnRNPA1 orchestrate an RPA-to-POT1 switch on telomeric single-stranded DNA", Nature **471**, 532–536 (2011).

[43]J. Sui, Y.-F. Lin, K. Xu, K.-J. Lee, D. Wang, and B. P. Chen, "DNA-PKcs phosphorylates hnRNP-A1 to facilitate the RPA-to-POT1 switch and telomere capping after replication", Nucleic Acids Research **43**, 5971 (2015).

[44]F. Rossiello, J. Aguado, S. Sepe, F. Iannelli, Q. Nguyen, S. Pitchiaya, P. Carninci, and F. d'Adda di Fagagna, "DNA damage response inhibition at dysfunctional telomeres by modulation of telomeric DNA damage response RNAs", Nature Communications **8**, 13980 (2017).

[45]D. Oliva-Rico and L. A. Herrera, "Regulated expression of the lncRNA TERRA and its impact on telomere biology", Mechanisms of Ageing and Development **167**, 16–23 (2017).

[46]S. Schoeftner and M. A. Blasco, "Developmentally regulated transcription of mammalian telomeres by DNA-dependent RNA polymerase II", Nature Cell Biology **10**, 228–236 (2008).

[47]J. Nandakumar, E. R. Podell, and T. R. Cech, "How telomeric protein POT1 avoids RNA to achieve specificity for single-stranded DNA.", Proceedings of the National Academy of Sciences of the United States of America **107**, 651–6 (2010).

[48]Y. Xu, Y. Suzuki, T. Ishizuka, C.-D. Xiao, X. Liu, T. Hayashi, and M. Komiyama, "Finding a human telomere DNA–RNA hybrid G-quadruplex formed by human telomeric 6-mer RNA and 16-mer DNA using click chemistry: A protective structure for telomere end", Bioorganic & Medicinal Chemistry **22**, 4419–4421 (2014).

[49]K. Takahama, A. Takada, S. Tada, M. Shimizu, K. Sayama, R. Kurokawa, and T. Oyoshi, "Regulation of telomere length by G-quadruplex telomere DNA- and TERRA-binding protein TLS/FUS", Chemistry and Biology **20**, 341–350 (2013).

[50]S. J. Tang, "Potential role of phase separation of repetitive DNA in chromosomal organization", Genes **8** (2017) `10.3390/genes8100279`.

[51]M. Graf, D. Bonetti, A. Lockhart, K. Serhal, V. Kellner, A. Maicher, P. Jolivet, M. T. Teixeira, and B. Luke, "Telomere length determines TERRA and R-loop regulation through the cell cycle", Cell **170**, 72–85 (2017).

[52] B. Balk, A. Maicher, M. Dees, J. Klermund, S. Luke-Glaser, K. Bender, and B. Luke, "Telomeric RNA-DNA hybrids affect telomere-length dynamics and senescence", Nature Structural & Molecular Biology **20**, 1199–1205 (2013).

[53] H. P. Chu, C. Cifuentes-Rojas, B. Kesner, E. Aeby, H. g. Lee, C. Wei, H. J. Oh, M. Boukhali, W. Haas, and J. T. Lee, "TERRA RNA antagonizes ATRX and protects telomeres", Cell **170**, 86–101 (2017).

[54] J. J. Montero, I. López-Silanes, D. Megías, M. F. Fraga, Castells-García, and M. A. Blasco, "TERRA recruitment of polycomb to telomeres is essential for histone trymethylation marks at telomeric heterochromatin", Nature Communications **9**, 1548 (2018).

[55] Z. Deng, J. Norseen, A. Wiedmer, H. Riethman, and P. M. Lieberman, "TERRA RNA binding to TRF2 facilitates heterochromatin formation and ORC recruitment at telomeres", Molecular Cell **35**, 403–413 (2009).

[56] H. Riethman, A. Ambrosini, and S. Paul, "Human subtelomere structure and variation", Chromosome Research **13**, 505–515 (2005).

[57] H. C. Mefford and B. J. Trask, "The complex structure and dynamic evolution of human subtelomeres", Nature Reviews Genetics **3**, 91–102 (2002).

[58] B. J. Trask, C. Friedman, A. Martin-Gallardo, L. Rowen, C. Akinbami, J. Blankenship, C. Collins, D. Giorgi, S. Iadonato, F. Johnson, W.-L. Kuo, H. Massa, T. Morrish, S. Naylor, O. T. H. Nguyen, S. Rouquier, T. Smith, D. J. Wong, J. Youngblom, and G. van den Engh, "Members of the olfactory receptor gene family are contained in large blocks of DNA duplicated polymorphically near the ends of human chromosomes", Human Molecular Genetics **7**, 13–26 (1998).

[59] S. Pisano, A. Galati, and S. Cacchione, "Telomeric nucleosomes: Forgotten players at chromosome ends", Cellular and Molecular Life Sciences **65**, 3553–3563 (2008).

[60] K. Struhl and E. Segal, "Determinants of nucleosome positioning.", Nature structural & molecular biology **20**, 267–73 (2013).

[61] S. Pisano, E. Marchioni, A. Galati, R. Mechelli, M. Savino, and S. Cacchione, "Telomeric nucleosomes are intrinsically mobile", Journal of Molecular Biology **369**, 1153–1162 (2007).

[62] S. Schoeftner and M. A. Blasco, "A 'higher order' of telomere regulation: telomere heterochromatin and telomeric RNAs.", The EMBO journal **28**, 2323–36 (2009).

[63] J. M. Zijlmans, U. M. Martens, S. S. Poon, A. K. Raap, H. J. Tanke, R. K. Ward, and P. M. Lansdorp, "Telomeres in the mouse have large inter-chromosomal variations in the number of T2AG3 repeats.", Proceedings of the National Academy of Sciences of the United States of America **94**, 7423–8 (1997).

[64] J. A. Rosenfeld, Z. Wang, D. E. Schones, K. Zhao, R. DeSalle, and M. Q. Zhang, "Determination of enriched histone modifications in non-genic portions of the human genome", BMC Genomics **10**, 143 (2009).

[65] M. D. Cubiles, S. Barroso, M. I. Vaquero-Sedas, A. Enguix, A. Aguilera, and M. A. Vega-Palas, "Epigenetic features of human telomeres", Nucleic Acids Research **46**, 2347–2355 (2018).

[66] R. Benetti, M. García-Cao, and M. A. Blasco, "Telomere length regulates the epigenetic status of mammalian telomeres and subtelomeres", Nature Genetics **39**, 243–250 (2007).

[67] Y. Ichikawa, H. Kurumizaka, Y. Nishimura, and M. Shimizu, "Nucleosome organization and chromatin dynamics in telomeres", BioMol Concepts **6**, 67–75 (2015).

[68] S. Kabir, D. Hockemeyer, and T. de Lange, "TALEN gene knockouts reveal no requirement for the conserved human shelterin protein Rap1 in telomere protection and length regulation.", Cell reports **9**, 1273–80 (2014).

[69]N. Arat and J. D. Griffith, “Human Rap1 interacts directly with telomeric DNA and regulates TRF2 localization at the telomere.”, The Journal of biological chemistry **287**, 41583–94 (2012).

[70]Y. Cai, V. Kandula, R. Kosuru, X. Ye, M. G. Irwin, and Z. Xia, “Decoding telomere protein Rap1: Its telomeric and nontelomeric functions and potential implications in diabetic cardiomyopathy.”, Cell cycle (Georgetown, Tex.) **16**, 1765–1773 (2017).

[71]P. Martinez, M. Thanasoula, A. R. Carlos, G. Gómez-López, A. M. Tejera, S. Schoeftner, O. Dominguez, D. G. Pisano, M. Tarsounas, and M. A. Blasco, “Mammalian Rap1 controls telomere function and gene expression through binding to telomeric and extratelomeric sites”, Nature Cell Biology **12**, 768–780 (2010).

[72]P. Martínez, G. Gómez-López, F. García, E. Mercken, S. Mitchell, J. M. Flores, R. DeCabo, and M. A. Blasco, “RAP1 protects from obesity through its extratelomeric role regulating gene expression”, Cell Reports **3**, 2059–2074 (2013).

[73]J. S. Z. Li, J. M. Fusté, T. Simavorian, C. Bartocci, J. Tsai, J. Karlseder, and E. L. Denchi, “TZAP: A telomere-associated protein involved in telomere length control”, Science **355**, 638–641 (2017).

[74]X. Feng, Z. Luo, S. Jiang, F. Li, X. Han, Y. Hu, D. Wang, Y. Zhao, W. Ma, D. Liu, J. Huang, and Z. Songyang, “The telomere-associated homeobox-containing protein TAH1/HMBOX1 participates in telomere maintenance in ALT cells.”, Journal of cell science **126**, 3982–9 (2013).

[75]H. Ma, L. Su, H. Yue, X. Yin, J. Zhao, S. Zhang, H. Kung, Z. Xu, and J. Miao, “HMBOX1 interacts with MT2A to regulate autophagy and apoptosis in vascular endothelial cells”, Scientific Reports **5**, 15121 (2015).

[76]H. Yu, P. R. Heenan, D. T. Edwards, L. Uyetake, and T. T. Perkins, “Quantifying the initial unfolding of bacteriorhodopsin reveals retinal stabilization”, Angewandte Chemie - International Edition **58**, 1710–1713 (2019).

[77]A. Jahn, G. Rane, M. Paszkowski-Rogacz, S. Sayols, A. Bluhm, C. Han, I. Draškovič, J. A. Londoño-Vallejo, A. P. Kumar, F. Buchholz, F. Butter, and D. Kappei, “ZBTB48 is both a vertebrate telomerebinding protein and a transcriptional activator”, EMBO reports **18**, 929–946 (2017).

[78]Y. Zhao, G. Zhang, C. He, Y. Mei, Y. Shi, and F. Li, “The 11th C2H2 zinc finger and an adjacent C-terminal arm are responsible for TZAP recognition of telomeric DNA”, Cell Research **28**, 130–134 (2018).

[79]Z. Lou, J. Wei, H. Riethman, J. A. Baur, R. Voglauer, J. W. Shay, and W. E. Wright, “Telomere length regulates ISG15 expression in human cells.”, Aging **1**, 608–21 (2009).

[80]J. R. Dixon, D. U. Gorkin, and B. Ren, “Chromatin fomains: The unit of chromosome organization”, Molecular Cell **62**, 668–680 (2016).

[81]J. D. Robin, A. T. Ludlow, K. Batten, F. Magdinier, G. Stadler, K. R. Wagner, J. W. Shay, and W. E. Wright, “Telomere position effect: regulation of gene expression with progressive telomere shortening over long distances.”, Genes & development **28**, 2464–76 (2014).

[82]T. Simonet, L.-E. Zaragosi, C. Philippe, K. Lebrigand, C. Schouteden, A. Augereau, S. Bauwens, J. Ye, M. Santagostino, E. Giulotto, F. Magdinier, B. Horard, P. Barbry, R. Waldmann, and E. Gilson, “The human TTAGGG repeat factors 1 and 2 bind to a subset of interstitial telomeric sequences and satellite repeats”, Cell Research **21**, 1028–1038 (2011).

[83]A. M. Wood, J. M. R. Danielsen, C. A. Lucas, E. L. Rice, D. Scalzo, T. Shimi, R. D. Goldman, E. D. Smith, M. M. Le Beau, and S. T. Kosak, “TRF2 and lamin A/C interact to facilitate the functional organization of chromosome ends”, Nature Communications **5**, 5467 (2014).

[84] W. Kim and J. W. Shay, “Long-range telomere regulation of gene expression: Telomere looping and telomere position effect over long distances (TPE-OLD).”, Differentiation; research in biological diversity **99**, 1–9 (2018).

[85] F. M. Bollmann, “The many faces of telomerase: emerging extratelomeric effects”, BioEssays **30**, 728–732 (2008).

[86] J. H. Santos, J. N. Meyer, M. Skorvaga, L. A. Annab, and B. van Houten, “Mitochondrial hTERT exacerbates free-radical-mediated mtDNA damage”, Aging Cell **3**, 399–411 (2004).

[87] J. H. Santos, J. N. Meyer, and B. van Houten, “Mitochondrial localization of telomerase as a determinant for hydrogen peroxide-induced mitochondrial DNA damage and apoptosis”, Human Molecular Genetics **15**, 1757–1768 (2006).

[88] J. Haendeler, S. Dröse, N. Büchner, S. Jakob, J. Altschmied, C. Goy, I. Spyridopoulos, A. M. Zeiher, U. Brandt, and S. Dimmeler, “Mitochondrial telomerase reverse transcriptase binds to and protects mitochondrial DNA and function from damage”, Arteriosclerosis, Thrombosis, and Vascular Biology **29**, 929–935 (2009).

[89] Y. Maida, M. Yasukawa, M. Furuuchi, T. Lassmann, R. Possemato, N. Okamoto, V. Kasim, Y. Hayashizaki, W. C. Hahn, and K. Masutomi, “An RNA-dependent RNA polymerase formed by TERT and the RMRP RNA.”, Nature **461**, 230–5 (2009).

[90] N. K. Sharma, A. Reyes, P. Green, M. J. Caron, M. G. Bonini, D. M. Gordon, I. J. Holt, and J. H. Santos, “Human telomerase acts as a hTR-independent reverse transcriptase in mitochondria”, Nucleic Acids Research **40**, 712–725 (2012).

[91] M. Akiyama, T. Hideshima, T. Hayashi, Y.-T. Tai, C. S. Mitsiades, N. Mitsiades, D. Chauhan, P. Richardson, N. C. Munshi, and K. C. Anderson, “Nuclear Factor-κB p65 mediates Tumor Necrosis Factor α-induced nuclear translocation of telomerase reverse transcriptase protein”, Cancer Res. **62**, 3876–3882 (2003).

[92] C. Chu, K. Qu, F. L. Zhong, S. E. Artandi, and H. Y. Chang, “Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions”, Molecular Cell **44**, 667–678 (2011).

[93] P. Martínez and M. A. Blasco, “Telomeric and extra-telomeric roles for telomerase and the telomere-binding proteins”, Nature Reviews Cancer **11**, 161–176 (2011).

[94] J. Ye, V. M. Renault, K. Jamet, and E. Gilson, “Transcriptional outcome of telomere signalling”, Nature Reviews Genetics **15**, 491–503 (2014).

[95] I. Grammatikakis, P. Zhang, M. P. Mattson, and M. Gorospe, “The long and the short of TRF2 in neurogenesis.”, Cell cycle (Georgetown, Tex.) **15**, 3026–3032 (2016).

[96] P. Zhang, M. J. Pazin, C. M. Schwartz, K. G. Becker, R. P. Wersto, C. M. Dilley, and M. P. Mattson, “Nontelomeric TRF2-REST interaction modulates neuronal gene silencing and fate of tumor and stem cells”, Current Biology **18**, 1489–1494 (2008).

[97] S. Canudas, B. R. Houghtaling, J. Y. Kim, J. N. Dynek, W. G. Chang, and S. Smith, “Protein requirements for sister telomere association in human cells.”, The EMBO journal **26**, 4867–78 (2007).

[98] J. Lin, P. Countryman, H. Chen, H. Pan, Y. Fan, Y. Jiang, P. Kaur, W. Miao, G. Gurgel, C. You, J. Piehler, N. M. Kad, R. Riehn, P. L. Opresko, S. Smith, Y. J. Tao, and H. Wang, “Functional interplay between SA1 and TRF1 in telomeric DNA binding and DNA–DNA pairing”, Nucleic Acids Research **44**, 6363–6376 (2016).

[99] B. R. Houghtaling, S. Canudas, and S. Smith, “A role for sister telomere cohesion in telomere elongation by telomerase.”, Cell cycle (Georgetown, Tex.) **11**, 19–25 (2012).

[100] M. Ramamoorthy and S. Smith, “Loss of ATRX suppresses resolution of telomere cohesion to control recombination in ALT cancer cells”, Cancer Cell **28**, 357–369 (2015).

[101]E. Tripathi and S. Smith, “Cell cycle-regulated ubiquitination of tankyrase 1 by RNF8 and ABRO1/BRCC36 controls the timing of sister telomere resolution.”, The EMBO journal **36**, 503–519 (2017).

[102]J. F. Passos, G. Saretzki, and T. von Zglinicki, “DNA damage in telomeres and mitochondria during cellular senescence: is there a connection?”, Nucleic Acids Research **35**, 7505–7513 (2007).

[103]E. S. Epel, E. H. Blackburn, J. Lin, F. S. Dhabhar, N. E. Adler, J. D. Morrow, and R. M. Cawthon, “Accelerated telomere shortening in response to life stress”, Proceedings of the National Academy of Sciences **101**, 17312–17315 (2004).

[104]J. Huzen, L. S. Wong, D. J. van Veldhuisen, N. J. Samani, A. H. Zwinderman, V. Codd, R. M. Cawthon, G. F. Benus, I. C. van der Horst, G. Navis, S. J. Bakker, R. T. Gansevoort, P. E. de Jong, H. L. Hillege, W. H. van Gilst, R. A. de Boer, and P. Van der Harst, “Telomere length loss due to smoking and metabolic traits”, Journal of Internal Medicine **275**, 155–163 (2014).

[105]B. J. Heidinger, J. D. Blount, W. Boner, K. Griffiths, N. B. Metcalfe, and P. Monaghan, “Telomere length in early life predicts life span”, Obstetrical and Gynecological Survey **67**, 283–284 (2012).

[106]M. Weischer, S. E. Bojesen, and B. G. Nordestgaard, “Telomere shortening unrelated to smoking, body weight, physical activity, and alcohol intake: 4,576 general population individuals with repeat measurements 10 years apart”, PLoS Genetics **10** (2014) `10.1371/journal.pgen.1004191`.

[107]S. Victorelli and J. F. Passos, “Telomeres and cell senescence - size matters not”, EBioMedicine **21**, 14–20 (2017).

[108]F. Rodier, J.-P. Coppé, C. K. Patil, W. A. M. Hoeijmakers, D. P. Muñoz, S. R. Raza, A. Freund, E. Campeau, A. R. Davalos, and J. Campisi, “Persistent DNA damage signalling triggers senescence-associated inflammatory cytokine secretion.”, Nature cell biology **11**, 973–9 (2009).

[109]A. Guerrero and J. Gil, “HMGB2 holds the key to the senescence-associated secretory phenotype.”, The Journal of cell biology **215**, 297–299 (2016).

[110]T. Tchkonia, Y. Zhu, J. van Deursen, J. Campisi, and J. L. Kirkland, “Cellular senescence and the senescent secretory phenotype: therapeutic opportunities.”, The Journal of clinical investigation **123**, 966–72 (2013).

[111]J. Lloberas and A. Celada, “Effect of aging on macrophage function”, Experimental Gerontology **37**, 1325–1331 (2002).

[112]M. Xu, T. Pirtskhalava, J. N. Farr, B. M. Weigand, A. K. Palmer, M. M. Weivoda, C. L. Inman, M. B. Ogrodnik, C. M. Hachfeld, D. G. Fraser, J. L. Onken, K. O. Johnson, G. C. Verzosa, L. G. P. Langhi, M. Weigl, N. Giorgadze, N. K. LeBrasseur, J. D. Miller, D. Jurk, R. J. Singh, D. B. Allison, K. Ejima, G. B. Hubbard, Y. Ikeno, H. Cubro, V. D. Garovic, X. Hou, S. J. Weroha, P. D. Robbins, L. J. Niedernhofer, S. Khosla, T. Tchkonia, and J. L. Kirkland, “Senolytics improve physical function and increase lifespan in old age.”, Nature medicine **24**, 1246–1256 (2018).

[113]P. L. Olive, “Endogenous DNA breaks: gammaH2AX and the role of telomeres.”, Aging **1**, 154–6 (2009).

[114]J. Karlseder, A. Smogorzewska, and T. de Lange, “Senescence induced by altered telomere state, not telomere loss.”, Science (New York, N.Y.) **295**, 2446–2449 (2002).

[115]Y. Zou, S. Misri, J. W. Shay, T. K. Pandita, and W. E. Wright, “Altered states of telomere deprotection and the two-stage mechanism of replicative aging”, Molecular and Cellular Biology **29**, 2390–2397 (2009).

[116]U. Herbig, W. A. Jobling, B. P. C. Chen, D. J. Chen, and J. M. Sedivy, “Telomere shortening triggers senescence of human cells through a pathway involving ATM, p53, and p21(CIP1), but not p16(INK4a).”, Molecular cell **14**, 501–13 (2004).

[117] S. P. Selvam, B. M. Roth, R. Nganga, J. Kim, M. A. Cooley, K. Helke, C. D. Smith, and B. Ogretmen, “Balance between senescence and apoptosis is regulated by telomere damage–induced association between p16 and caspase-3”, Journal of Biological Chemistry **293**, 9784–9800 (2018).

[118] G. Hewitt, D. Jurk, F. D. Marques, C. Correia-Melo, T. Hardy, A. Gackowska, R. Anderson, M. Taschuk, J. Mann, and J. F. Passos, “Telomeres are favoured targets of a persistent DNA damage response in ageing and stress-induced senescence”, Nature Communications **3**, 708 (2012).

[119] M. P. Longhese, F. d’Adda di Fagagna, G. Bucci, V. Matti, D. Cittaro, C. M. Beausejour, M. Dobreva, S. Barozzi, M. Fumagalli, F. Rossiello, M. Clerici, J. M. Kaplunov, and U. Herbig, “Telomeric DNA damage is irreparable and causes persistent DNA-damage-response activation”, Nature Cell Biology **14**, 355–365 (2012).

[120] D. Jurk, C. Wilson, J. F. Passos, F. Oakley, C. Correia-Melo, L. Greaves, G. Saretzki, C. Fox, C. Lawless, R. Anderson, G. Hewitt, S. L. Pender, N. Fullard, G. Nelson, J. Mann, B. van de Sluis, D. A. Mann, and T. von Zglinicki, “Chronic inflammation induces telomere dysfunction and accelerates ageing in mice”, Nature Communications **5**, 4172 (2014).

[121] Z. Kaul, A. J. Cesare, L. I. Huschtscha, A. A. Neumann, and R. R. Reddel, “Five dysfunctional telomeres predict onset of senescence in human cells.”, EMBO reports **13**, 52–9 (2011).

[122] W. J. Cannan and D. S. Pederson, “Mechanisms and consequences of double-strand DNA break formation in chromatin”, Journal of Cellular Physiology **231**, 3–14 (2016).

[123] L. Liu, J. R. Trimarchi, P. J. Smith, and D. L. Keefe, “Mitochondrial dysfunction leads to telomere attrition and genomic instability.”, Aging cell (2002) `10.1046/j.1474-9728.2002.00004.x`.

[124] T. Richter and T. v. Zglinicki, “A continuous correlation between oxidative stress and telomere shortening in fibroblasts”, Experimental Gerontology (2007) `10.1016/j.exger.2007.08.005`.

[125] Y. Zou, A. Sfeir, S. M. Gryaznov, J. W. Shay, and W. E. Wright, “Does a sentinel or a subset of short telomeres determine replicative senescence?”, Molecular biology of the cell **15**, 3709–18 (2004).

[126] J. Birch, R. K. Anderson, C. Correia-Melo, D. Jurk, G. Hewitt, F. M. Marques, N. J. Green, E. Moisey, M. A. Birrell, M. G. Belvisi, F. Black, J. J. Taylor, A. J. Fisher, A. De Soyza, and J. F. Passos, “DNA damage response at telomeres contributes to lung aging and chronic obstructive pulmonary disease”, American Journal of Physiology-Lung Cellular and Molecular Physiology **309**, L1124–L1137 (2015).

[127] J. Fairlie and L. Harrington, “Enforced telomere elongation increases the sensitivity of human tumour cells to ionizing radiation”, DNA Repair **25**, 54–59 (2015).

[128] Y. Arai, C. M. Martin-Ruiz, M. Takayama, Y. Abe, T. Takebayashi, S. Koyasu, M. Suematsu, N. Hirose, and T. von Zglinicki, “Inflammation, but not telomere length, predicts successful ageing at extreme old age: A longitudinal study of semi-supercentenarians”, EBioMedicine **2**, 1549–1558 (2015).

[129] J. K. Yeh and C. Y. Wang, “Telomeres and telomerase in cardiovascular diseases”, Genes **7** (2016) `10.3390/genes7090058`.

[130] J.-M. Yuan, K. B. Beckman, R. Wang, C. Bull, J. Adams-Haduch, J. Y. Huang, A. Jin, P. Opresko, A. B. Newman, Y.-L. Zheng, M. Fenech, and W.-P. Koh, “Leukocyte telomere length in relation to risk of lung adenocarcinoma incidence: Findings from the Singapore Chinese Health Study”, International Journal of Cancer **142**, 2234–2243 (2018).

[131] T. von Zglinicki, “Oxidative stress shortens telomeres”, Trends in Biochemical Sciences **27**, 339–344 (2002).

[132] S. Oikawa and S. Kawanishi, “Site-specific DNA damage at GGG sequence by oxidative stress may accelerate telomere shortening”, FEBS Letters **453**, 365–368 (1999).

[133] S. Oikawa, S. Tada-Oikawa, and S. Kawanishi, “Site-specific DNA damage at the GGG sequence by UVA involves acceleration of telomere shortening”, Biochemistry **40**, 4763–4768 (2001).

[134] S. Kawanishi and S. Oikawa, “Mechanism of telomere shortening by oxidative stress”, Annals of the New York Academy of Sciences **1019**, 278–284 (2004).

[135] I. Saito, T. Nakamura, K. Nakatani, Y. Yoshioka, K. Yamaguchi, and H. Sugiyama, “Mapping of the hot spots for DNA damage by one-electron oxidation: Efficacy of GG doublets and GGG triplets as a trap in long-range hole migration”, Journal of the American Chemical Society **120**, 12686–12687 (1998).

[136] A. K. Koliada, D. S. Krasnenkov, and A. M. Vaiserman, “Telomeric aging: mitotic clock or stress indicator?”, Frontiers in genetics **6**, 82 (2015).

[137] H. Oeseburg, R. A. de Boer, W. H. van Gilst, and P. van der Harst, “Telomere biology in healthy aging and disease”, Pflügers Archiv - European Journal of Physiology **459**, 259–268 (2010).

[138] A. Vancevska, K. M. Douglass, V. Pfeiffer, S. Manley, and J. Lingner, “The telomeric DNA damage response occurs in the absence of chromatin decompaction.”, Genes & development **31**, 567–577 (2017).

[139] H. E. Krokan and M. Bjørås, “Base excision repair.”, Cold Spring Harbor perspectives in biology **5**, a012583 (2013).

[140] O. D. Schärer, “Nucleotide excision repair in eukaryotes.”, Cold Spring Harbor perspectives in biology **5**, a012609 (2013).

[141] P. J. Rochette and D. E. Brash, “Human telomeres are hypersensitive to UV-induced DNA damage and refractory to repair”, PLoS Genetics **6**, edited by J. M. Ford, e1000926 (2010).

[142] D. Parikh, E. Fouquerel, C. T. Murphy, H. Wang, and P. L. Opresko, “Telomeres are partly shielded from ultraviolet-induced damage and proficient for nucleotide excision repair of photoproducts”, Nature Communications **6**, 8214 (2015).

[143] X. D. Zhu, L. Niedernhofer, B. Kuster, M. Mann, J. H. Hoeijmakers, and T. De Lange, “ERCC1/XPF removes the 3 overhang from uncapped telomeres and represses formation of telomeric DNA-containing double minute chromosomes”, Molecular Cell **12**, 1489–1498 (2003).

[144] G. J. Stout and M. A. Blasco, “Telomere length and telomerase activity impact the UV sensitivity syndrome xeroderma pigmentosum C”, Cancer Research **73**, 1844–1854 (2013).

[145] P. Jia, C. Her, and W. Chai, “DNA excision repair at telomeres.”, DNA repair **36**, 137–45 (2015).

[146] J. Zhou, A. M. Fleming, A. M. Averill, C. J. Burrows, and S. S. Wallace, “The NEIL glycosylases remove oxidized guanine lesions from telomeric and promoter quadruplex DNA structures.”, Nucleic acids research **43**, 4039–54 (2015).

[147] H. Vallabhaneni, N. O’Callaghan, J. Sidorova, and Y. Liu, “Defective repair of oxidative base lesions by the DNA glycosylase Nth1 associates with multiple telomere defects”, PLoS Genetics **9**, edited by J.-Q. Zhou, e1003639 (2013).

[148] I. Jagannathan, H. A. Cole, and J. J. Hayes, “Base excision repair in nucleosome substrates”, Chromosome Research **14**, 27–37 (2006).

[149] A. S. Miller, L. Balakrishnan, N. A. Buncher, P. L. Opresko, and R. A. Bambara, "Telomere proteins POT1, TRF1 and TRF2 augment long-patch base excision repair in vitro", Cell Cycle **11**, 998–1007 (2012).

[150] J. Zhou, J. Chan, M. Lambelé, T. Yusufzai, J. Stumpff, P. L. Opresko, M. Thali, and S. S. Wallace, "NEIL3 repairs telomere damage during S phase to secure chromosome segregation at mitosis", Cell Reports **20**, 2044–2056 (2017).

[151] J. Wu, M. McKeague, and S. J. Sturla, "Nucleotide-resolution genome-wide mapping of oxidative DNA damage by click-code-seq", Journal of the American Chemical Society **140**, 9783–9787 (2018).

[152] P. L. Opresko, J. Fan, S. Danzy, D. M. Wilson, and V. A. Bohr, "Oxidative damage in telomeric DNA disrupts recognition by TRF1 and TRF2", Nucleic Acids Research **33**, 1230–1239 (2005).

[153] S. Madlener, T. Strobel, S. Vose, O. Saydam, B. D. Price, B. Demple, and N. Saydam, "Essential role for mammalian apurinic/apyrimidinic (AP) endonuclease Ape1/Ref-1 in telomere maintenance", Proceedings of the National Academy of Sciences **110**, 17844–17849 (2013).

[154] J. Cadet and J. R. Wagner, "DNA base damage by reactive oxygen species, oxidizing agents, and UV radiation.", Cold Spring Harbor perspectives in biology **5** (2013) `10.1101/cshperspect.a012559`.

[155] J. Cadet, K. J. Davies, M. H. Medeiros, P. Di Mascio, and J. R. Wagner, "Formation and repair of oxidatively generated damage in cellular DNA", Free Radical Biology and Medicine **107**, 13–34 (2017).

[156] M. Pflaum, O. Will, H.-C. Mahler, and B. Epe, "DNA oxidation products determined with repair endonucleases in mammalian cells: Types, basal levels and influence of cell proliferation", Free Radical Research **29**, 585–594 (1998).

[157] M. A. Filatov, S. Baluschev, and K. Landfester, "Protection of densely populated excited triplet state ensembles against deactivation by molecular oxygen", Chemical Society Reviews **45**, 4668–4689 (2016).

[158] E. Alizadeh, P. Cloutier, D. Hunting, and L. Sanche, "Soft X-ray and low energy electron-induced damage to DNA under N2 and O2 atmospheres.", The journal of physical chemistry. B **115**, 4523–31 (2011).

[159] J.-L. Ravanat, G. R. Martinez, M. H. Medeiros, P. Di Mascio, and J. Cadet, "Mechanistic aspects of the oxidation of DNA constituents mediated by singlet molecular oxygen", Archives of Biochemistry and Biophysics **423**, 23–30 (2004).

[160] J.-P. Pouget, S. Frelon, J.-L. Ravanat, I. Testard, F. Odin, and J. Cadet, "Formation of modified DNA bases in cells exposed either to gamma radiation or to high-LET particles.", Radiation research **157**, 589–95 (2002).

[161] F. Bergeron, F. Auvre, J. P. Radicella, and J.-L. Ravanat, "HO radicals induce an unexpected high proportion of tandem base lesions refractory to repair by DNA glycosylases", Proceedings of the National Academy of Sciences (2010) `10.1073/pnas.1000193107`.

[162] N. R. Jena, "DNA damage by reactive species: Mechanisms, mutation and repair", Journal of Biosciences **37**, 503–517 (2012).

[163] M. Dizdaroglu, "Oxidatively induced DNA damage: Mechanisms, repair and disease", Cancer Letters **327**, 26–47 (2012).

[164] A. Kumar, V. Pottiboyina, and M. D. Sevilla, "Hydroxyl radical (OH•) reaction with guanine in an aqueous environment: a DFT study.", The journal of physical chemistry. B **115**, 15129–37 (2011).

[165] I. Saito, M. Takayama, H. Sugiyama, K. Nakatani, A. Tsuchida, and M. Yamamoto, "Photoinduced DNA cleavage via electron transfer: Demonstration that guanine

residues located 5′ to guanine are the most electron-donating sites", Journal of the American Chemical Society **117**, 6406–6407 (1995).

[166] L. Cupellini, P. Wityk, B. Mennucci, and J. Rak, "Photoinduced electron transfer in 5-bromouracil labeled DNA. A contrathermodynamic mechanism revisited by electron transfer theories", Physical Chemistry Chemical Physics **21**, 4387–4393 (2019).

[167] D. R. Cardoso, S. H. Libardi, and L. H. Skibsted, "Riboflavin as a photosensitizer. Effects on human health and food quality", Food & Function **3**, 487 (2012).

[168] J. Jie, K. Liu, L. Wu, H. Zhao, D. Song, and H. Su, "Capturing the radical ion-pair intermediate in DNA guanine oxidation", Science Advances **3**, e1700171 (2017).

[169] J Cadet, "Oxidative damage to DNA: formation, measurement and biochemical features", Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis **531**, 5–23 (2003).

[170] R. P. Hickerson, C. L. Chepanoske, S. D. Williams, S. S. David, and C. J. Burrows, "Mechanism-based DNA-protein cross-linking of MutY via oxidation of 8- oxoguanosine", Journal of the American Chemical Society **121**, 9901–9902 (1999).

[171] S. Perrier, J. Hau, D. Gasparutto, J. Cadet, A. Favier, and J. L. Ravanat, "Characterization of lysine-guanine cross-links upon one-electron oxidation of a guanine-containing oligonucleotide in the presence of a trilysine peptide", Journal of the American Chemical Society **128**, 5703–5710 (2006).

[172] B. Thapa, B. H. Munk, C. J. Burrows, and H. B. Schlegel, "Computational Study of Oxidation of Guanine by Singlet Oxygen (1Δg) and Formation of Guanine:Lysine Cross-Links", Chemistry - A European Journal **23**, 5804–5813 (2017).

[173] X. Xu, J. G. Muller, Y. Ye, and C. J. Burrows, "DNA-protein cross-links between guanine and lysine depend on the mechanism of oxidation for formation of C5 vs C8 guanosine adducts", Journal of the American Chemical Society **130**, 703–709 (2008).

[174] S. Bjelland and E. Seeberg, "Mutagenicity, toxicity and repair of DNA base damage induced by oxidation", Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis **531**, 37–80 (2003).

[175] I. Talhaoui, S. Couvé, A. A. Ishchenko, C. Kunz, P. Schär, and M. Saparbaev, "7,8-Dihydro-8-oxoadenine, a highly mutagenic adduct, is repaired by Escherichia coli and human mismatch-specific uracil/thymine-DNA glycosylases.", Nucleic acids research **41**, 912–23 (2013).

[176] M. L. Wood, A Esteve, M. L. Morningstar, G. M. Kuziemko, and J. M. Essigmann, "Genetic effects of oxidative DNA damage: comparative mutagenesis of 7,8-dihydro-8-oxoguanine and 7,8-dihydro-8-oxoadenine in Escherichia coli.", Nucleic acids research **20**, 6023–32 (1992).

[177] W. Wu and J. Kieffer, "New hybrid method for the calculation of the solvation free energy of small molecules in aqueous solutions", Journal of Chemical Theory and Computation **15**, 371–381 (2019).

[178] L. Pan, B. Zhu, W. Hao, X. Zeng, S. A. Vlahopoulos, T. K. Hazra, M. L. Hegde, Z. Radak, A. Bacsi, A. R. Brasier, X. Ba, and I. Boldogh, "Oxidized guanine base lesions function in 8-oxoguanine DNA glycosylase-1-mediated epigenetic regulation of nuclear factor κB-driven gene expression", Journal of Biological Chemistry **291**, 25553–25566 (2016).

[179] A. M. Fleming and C. J. Burrows, "8-Oxo-7,8-dihydroguanine, friend and foe: Epigenetic-like regulator versus initiator of mutagenesis", DNA Repair **56**, 75–83 (2017).

[180] F. Crick, "Central dogma of molecular biology", Nature (1970) 10.1038/227561a0.

[181] A. Travers, E. Hiriart, M. Churcher, M. Caserta, and E. Di Mauro, "The DNA sequence-dependence of nucleosome positioning in vivo and in vitro", Journal of Biomolecular Structure and Dynamics **27**, 713–724 (2010).

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

[183] R. Rohs, S. M. West, A. Sosinsky, P. Liu, R. S. Mann, and B. Honig, "The role of DNA shape in protein-DNA recognition", Nature **461**, 1248–1253 (2009).

[184] T. Schwartz, J. Behlke, K. Lowenhaupt, U. Heinemann, and A. Rich, "Structure of the DLM-1–Z-DNA complex reveals a conserved family of Z-DNA-binding proteins", Nature Structural Biology **8**, 761–765 (2001).

[185] V. González, K. Guo, L. Hurley, and D. Sun, "Identification and characterization of nucleolin as a c-myc G-quadruplex-binding protein.", The Journal of biological chemistry **284**, 23622–35 (2009).

[186] M. S. Wold, "Replication Protein A: A heterotrimeric, single-stranded DNA-binding protein required for eukaryotic DNA metabolism", Annual Review of Biochemistry **66**, 61–92 (2002).

[187] G. Mishra and Y. Levy, "SUPPL: Molecular determinants of the interactions between proteins and ssDNA.", Proceedings of the National Academy of Sciences of the United States of America **112**, 5033–8 (2015).

[188] M. P. Horvath, "Single-stranded Nucleic Acid (SSNA)-binding Proteins", in *Protein–nucleic acid interactions. structural biology*, edited by P. A. Rice and C. C. Correll (Royal Society of Chemistry, 2008), pp. 91–128.

[189] A. Aggarwal, D. Rodgers, M. Drottar, M. Ptashne, and S. Harrison, "Recognition of a DNA operator by the repressor of phage 434: a view at high resolution", Science **242**, 899–907 (1988).

[190] G. Zubay and P. Doty, "The isolation and properties of deoxyribonucleoprotein particles containing single nucleic acid molecules", Journal of Molecular Biology **1**, 1–IN1 (1959).

[191] D. Vuzman and Y. Levy, "Intrinsically disordered regions as affinity tuners in protein-DNA interactions", Molecular BioSystems **8**, 47–57 (2012).

[192] F. Merino, B. Bouvier, and V. Cojocaru, "Cooperative DNA recognition modulated by an interplay between protein-protein interactions and DNA-mediated allostery", PLoS Computational Biology **11**, e1004287 (2015).

[193] L. Etheve, J. Martin, and R. Lavery, "Decomposing protein-DNA binding and recognition using simplified protein models", Nucleic Acids Research **45**, 10270–10283 (2017).

[194] D. S. Latchman, "Families of DNA binding transcription factors", in *Eukaryotic transcription factors* (Academic Press, Jan. 2007), pp. 77–133.

[195] A. Murzin, "OB(oligonucleotide/oligosaccharide binding)-fold: common structural and functional solution for non-homologous sequences.", The EMBO Journal **12**, 861–867 (2018).

[196] Y. Bilu and N. Barkai, "The design of transcription-factor binding sites is affected by combinatorial regulation", Genome Biology **6**, R103 (2005).

[197] M. Slutsky and L. A. Mirny, "Kinetics of protein-DNA interaction: Facilitated target location in sequence-dependent potential", Biophysical Journal **87**, 4021–4035 (2004).

[198] L. Mirny, M. Slutsky, Z. Wunderlich, A. Tafvizi, J. Leith, and A. Kosmrlj, "How a protein searches for its site on DNA: The mechanism of facilitated diffusion", Journal of Physics A: Mathematical and Theoretical **42**, 434013 (2009).

[199] A. M. Florescu and M. Joyeux, “Comparison of kinetic and dynamical models of DNA-protein interaction and facilitated diffusion”, Journal of Physical Chemistry A **114**, 9662–9672 (2010).

[200] S. Halford, “An end to 40 years of mistakes in DNA–protein association kinetics?”, Biochemical Society Transactions **37**, 343–348 (2009).

[201] A. A. Shvets, M. P. Kochugaeva, and A. B. Kolomeisky, *Mechanisms of protein search for targets on DNA: Theoretical insights*, Aug. 2018.

[202] L. Zandarashvili, A. Esadze, D. Vuzman, C. A. Kemme, Y. Levy, and J. Iwahara, “Balancing between affinity and speed in target DNA search by zinc-finger proteins via modulation of dynamic conformational ensemble”, Proceedings of the National Academy of Sciences **112**, E5142–E5149 (2015).

[203] P. C. Blainey, G. Luo, S. C. Kou, W. F. Mangel, G. L. Verdine, B. Bagchi, and X. S. Xie, “Nonspecifically bound proteins spin while diffusing along DNA”, Nature Structural & Molecular Biology **16**, 1224–1229 (2009).

[204] I. Echeverria and G. A. Papoian, “DNA exit eamps are revealed in the binding landscapes obtained from simulations in helical coordinates”, PLoS Computational Biology **11**, edited by J. M. Briggs, e1003980 (2015).

[205] A. Marcovitz and Y. Levy, “Obstacles may facilitate and direct DNA search by proteins”, Biophysical Journal **104**, 2042–2050 (2013).

[206] D. Vuzman and Y. Levy, “The “monkey-bar” mechanism for searching for the DNA target site: The molecular determinants”, Israel Journal of Chemistry **54**, 1374–1381 (2014).

[207] J. Rudolph, J. Mahadevan, P. Dyer, and K. Luger, “Poly(ADP-ribose) polymerase 1 searches DNA via a ‘monkey bar’ mechanism”, eLife **7** (2018) `10.7554/elife.37818`.

[208] M. Bauer, E. S. Rasmussen, M. A. Lomholt, and R. Metzler, “Real sequence effects on the search dynamics of transcription factors on DNA”, Scientific Reports **5**, 10072 (2015).

[209] M. Cencini and S. Pigolotti, “Energetic funnel facilitates facilitated diffusion”, Nucleic Acids Research **46**, 558–567 (2018).

[210] P. Hammar, P. Leroy, A. Mahmutovic, E. G. Marklund, O. G. Berg, and J. Elf, “The lac repressor displays facilitated diffusion in living cells”, Science **336**, 1595–1598 (2012).

[211] J. Lin, P. Countryman, N. Buncher, P. Kaur, E. Longjiang, Y. Zhang, G. Gibson, C. You, S. C. Watkins, J. Piehler, P. L. Opresko, N. M. Kad, and H. Wang, “TRF1 and TRF2 use different mechanisms to find telomeric DNA but share a novel mechanism to search for protein partners at telomeres”, Nucleic Acids Research **42**, 2493–2504 (2014).

[212] F. Erdel, K. Kratz, S. Willcox, J. D. Griffith, E. C. Greene, and T. de Lange, “Telomere Recognition and Assembly Mechanism of Mammalian Shelterin”, Cell Reports **18**, 41–53 (2017).

[213] P. Dey and A. Bhattacherjee, “Disparity in anomalous diffusion of proteins searching for their target DNA sites in a crowded medium is controlled by the size, shape and mobility of macromolecular crowders”, Soft Matter **15**, 1960–1969 (2019).

[214] C. Loverdo, O. Bénichou, R. Voituriez, A. Biebricher, I. Bonnet, and P. Desbiolles, “Quantifying hopping and jumping in facilitated diffusion of DNA-binding proteins”, Physical Review Letters **102**, 188101 (2009).

[215] A. G. Cherstvy, A. B. Kolomeisky, and A. A. Kornyshev, “Protein - DNA interactions: Reaching and recognizing the targets”, Journal of Physical Chemistry B **112**, 4741–4750 (2008).

[216]D. Marx and J. Hutter, *Ab Initio Molecular Dynamics: Basic Theory and Advanced Methods* (Cambridge University Press, New York, 2009).

[217]J. A. Morrone and R. Car, "Nuclear quantum effects in water", Physical Review Letters **101**, 017801 (2008).

[218]P. Goyal, C. A. Schwerdtfeger, A. V. Soudackov, and S. Hammes-Schiffer, "Nonadiabatic dynamics of photoinduced proton-coupled electron transfer in a solvated phenol-amine complex", Journal of Physical Chemistry B **119**, 2758–2768 (2015).

[219]S. P. Webb, T. Iordanov, and S. Hammes-Schiffer, "Multiconfigurational nuclear-electronic orbital approach: Incorporation of nuclear quantum effects in electronic structure calculations", The Journal of Chemical Physics **117**, 4106–4118 (2002).

[220]L. Piela, *Ideas of Quantum Chemistry* (Elsevier, 2007).

[221]P Hohenberg and W Kohn, "The inhomogeneous electron gas", Phys. Rev. **136**, B864 (1964).

[222]W. Kohn and L. J. Sham, "Self-consistent equations including exchange and correlation effects", Physical Review **140**, A1133–A1138 (1965).

[223]X. Ren, P. Rinke, V. Blum, J. Wieferink, A. Tkatchenko, A. Sanfilippo, K. Reuter, and M. Scheffler, "Resolution-of-identity approach to Hartree–Fock, hybrid density functionals, RPA, MP2 and GW with numeric atom-centered orbital basis functions", New Journal of Physics **14**, 053020 (2012).

[224]J. VandeVondele, M. Krack, F. Mohamed, M. Parrinello, T. Chassaing, and J. Hutter, "Quickstep: Fast and accurate density functional calculations using a mixed Gaussian and plane waves approach", Computer Physics Communications **167**, 103–128 (2005).

[225]D. R. Hamann, M. Schlüter, and C. Chiang, "Norm-conserving pseudopotentials", Physical Review Letters **43**, 1494–1497 (1979).

[226]W. Koch and M. C. Holthausen, *A Chemist's Guide to Density Functional Theory*, Second Edi (Wiley-VCH, Weinheim, 2001).

[227]A. D. Becke, "Density-functional thermochemistry. III. The role of exact exchange", The Journal of Chemical Physics **98**, 5648–5652 (1993).

[228]R. Peverati and D. G. Truhlar, "An improved and broadly accurate local approximation to the exchange–correlation density functional: The MN12-L functional for electronic structure calculations in chemistry and physics", Physical Chemistry Chemical Physics **14**, 13171 (2012).

[229]N. Luehr, I. S. Ufimtsev, and T. J. Martínez, "Dynamic precision for electron repulsion integral evaluation on graphical processing units (GPUs)", Journal of Chemical Theory and Computation **7**, 949–954 (2011).

[230]Z. Chen, D. Zhang, Y. Jin, Y. Yang, N. Q. Su, and W. Yang, "Multireference Density Functional Theory with generalized auxiliary systems for ground and excited states", Journal of Physical Chemistry Letters **8**, 4479–4485 (2017).

[231]F. Neese, "Software update: the ORCA program system, version 4.0", Wiley Interdisciplinary Reviews: Computational Molecular Science **8**, e1327 (2018).

[232]F. Brockherde, L. Vogt, L. Li, M. E. Tuckerman, K. Burke, and K. R. Müller, "Bypassing the Kohn-Sham equations with machine learning", Nature Communications **8**, 872 (2017).

[233]M. E. Tuckerman, *Statistical Mechanics: Theory and Molecular Simulation* (New York, 2010).

[234]H. C. Andersen, "Molecular dynamics simulations at constant pressure and/or temperature", The Journal of Chemical Physics **72**, 2384–2393 (1980).

[235]H. Chen, H. Fu, X. Shao, C. Chipot, and W. Cai, "ELF: An extended-Lagrangian free energy calculation module for multiple molecular dynamics engines", Journal of Chemical Information and Modeling **58**, 1315–1318 (2018).

[236] J. A. Lemkul, B. Roux, D. van der Spoel, and A. D. MacKerell, "Implementation of extended Lagrangian dynamics in GROMACS for polarizable simulations using the classical Drude oscillator model", Journal of Computational Chemistry **36**, 1473–1479 (2015).

[237] A. D. Mackerell, "Empirical force fields for biological macromolecules: Overview and issues", Journal of Computational Chemistry **25**, 1584–1604 (2004).

[238] G. Bussi, D. Donadio, and M. Parrinello, "Canonical sampling through velocity rescaling", The Journal of Chemical Physics **126**, 014101 (2007).

[239] M. Parrinello and A. Rahman, "Polymorphic transitions in single crystals: A new molecular dynamics method", Journal of Applied Physics **52**, 7182–7190 (1981).

[240] C. Chipot and A. Pohorille, *Free Energy Calculations : Theory and Applications in Chemistry and Biology* (2012).

[241] K.-Y. Wong and D. M. York, "Exact relation between potential of mean force and free-energy profile", Journal of chemical theory and computation **8**, 3998–4003 (2012).

[242] W. K. den Otter, "Revisiting the exact relation between potential of mean force and free-energy profile", Journal of Chemical Theory and Computation **9**, 3861–3865 (2013).

[243] G. Torrie and J. Valleau, "Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling", Journal of Computational Physics **23**, 187–199 (1977).

[244] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman, "THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method", Journal of Computational Chemistry **13**, 1011–1021 (1992).

[245] A. M. Ferrenberg and R. H. Swendsen, "Optimized Monte Carlo data analysis", Physical Review Letters **63**, 1195–1198 (1989).

[246] G. Bussi, "Hamiltonian replica exchange in GROMACS: A flexible implementation", in Molecular physics, Vol. 112, 3-4 (Feb. 2014), pp. 379–384.

[247] L. S. Stelzl, A. Kells, E. Rosta, and G. Hummer, "Dynamic Histogram Analysis to determine free energies and rates from biased simulations", Journal of Chemical Theory and Computation **13**, 6328–6342 (2017).

[248] F. Wang and D. P. Landau, "Efficient, multiple-range random walk algorithm to calculate the density of states", Physical Review Letters **86**, 2050–2053 (2001).

[249] H. Grubmüller, "Predicting slow structural transitions in macromolecular systems: Conformational flooding", Physical Review E **52**, 2893–2906 (1995).

[250] A. Laio and M. Parrinello, "Escaping free-energy minima.", Proceedings of the National Academy of Sciences of the United States of America **99**, 12562–6 (2002).

[251] A. Barducci, G. Bussi, and M. Parrinello, "Well-tempered metadynamics: A smoothly converging and tunable free-energy method", Physical Review Letters **100**, 020603 (2008).

[252] S. Piana and A. Laio, "A bias-exchange approach to protein folding", Journal of Physical Chemistry B **111**, 4553–4559 (2007).

[253] F. Noé and F. Nüske, "A variational approach to modeling slow processes in stochastic dynamical systems", Multiscale Modeling & Simulation **11**, 635–655 (2012).

[254] J. D. Chodera and F. Noé, "Markov state models of biomolecular conformational dynamics.", Current opinion in structural biology **25**, 135–44 (2014).

[255] J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte, and F. Noé, "Markov models of molecular kinetics: Generation and validation", The Journal of Chemical Physics **134**, 174105 (2011).

[256] G. R. Bowman, K. A. Beauchamp, G. Boxer, and V. S. Pande, "Progress and challenges in the automated construction of Markov state models for full protein systems", The Journal of Chemical Physics **131**, 124101 (2009).

[257] H. Wu, F. Paul, C. Wehmeyer, and F. Noé, "Multiensemble Markov models of molecular thermodynamics and kinetics", Proceedings of the National Academy of Sciences **113**, E3221–E3230 (2016).

[258] S. Olsson, H. Wu, F. Paul, C. Clementi, and F. Noé, "Combining experimental and simulation data of molecular processes via augmented Markov models.", Proceedings of the National Academy of Sciences of the United States of America **114**, 8265–8270 (2017).

[259] A. Cesari, S. Reißer, and G. Bussi, "Using the Maximum Entropy Principle to combine simulations and solution experiments", Computation **6**, 15 (2018).

[260] S. Röblitz and M. Weber, "Fuzzy spectral clustering by PCCA+: application to Markov state models and data classification", Advances in Data Analysis and Classification **7**, 147–179 (2013).

[261] M. K. Scherer, B. Trendelkamp-Schroer, F. Paul, G. Pérez-Hernández, M. Hoffmann, N. Plattner, C. Wehmeyer, J. H. Prinz, and F. Noé, "PyEMMA 2: A software package for estimation, validation, and analysis of Markov models", Journal of Chemical Theory and Computation **11**, 5525–5542 (2015).

[262] M. P. Harrigan, M. M. Sultan, C. X. Hernández, B. E. Husic, P. Eastman, C. R. Schwantes, K. A. Beauchamp, R. T. McGibbon, and V. S. Pande, "MSMBuilder: Statistical models for biomolecular dynamics", Biophysical Journal **112**, 10–15 (2017).

[263] V. S. Pande, *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*, Vol. 797 (2014).

[264] X. Hu, Y. Wang, A. Hunkele, D. Provasi, G. W. Pasternak, and M. Filizola, "Kinetic and thermodynamic insights into sodium ion translocation through the μ-opioid receptor from molecular dynamics and machine learning analysis", PLOS Computational Biology **15**, edited by A. MacKerell, e1006689 (2019).

[265] P. Wityk, M. Wieczór, S. Makurat, L. Chomicz-Mańka, J. Czub, and J. Rak, "Dominant pathways of adenosyl radical-induced DNA damage revealed by QM/MM metadynamics", Journal of Chemical Theory and Computation **13**, 6415–6423 (2017).

[266] K. Chenoweth, A. C. Van Duin, and W. A. Goddard, "ReaxFF reactive force field for molecular dynamics simulations of hydrocarbon oxidation", Journal of Physical Chemistry A **112**, 1040–1053 (2008).

[267] T. Steinbrecher, I. Joung, and D. A. Case, "Soft-core potentials in thermodynamic integration: Comparing one- and two-step transformations", Journal of Computational Chemistry **32**, 3253–3263 (2011).

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

[269] M. R. Shirts and J. D. Chodera, "Statistically optimal analysis of samples from multiple equilibrium states", The Journal of Chemical Physics **129**, 124105 (2008).

[270] R. W. Zwanzig, "High-temperature equation of state by a perturbation method. I. Nonpolar gases", Journal of Chemical Physics **22**, 1420–1426 (1954).

[271] G. Pérez-Hernández, F. Paul, T. Giorgino, G. De Fabritiis, and F. Noé, "Identification of slow molecular order parameters for Markov model construction", The Journal of Chemical Physics **139**, 015102 (2013).

[272] C. R. Schwantes and V. S. Pande, "Improvements in Markov State Model construction reveal many non-native interactions in the folding of NTL9", Journal of Chemical Theory and Computation **9**, 2000–2009 (2013).

[273] A. J. Izenman, “Linear Discriminant Analysis”, in *Modern multivariate statistical techniques* (Springer, New York, NY, 2013), pp. 237–280.

[274] M. Teletin, G. Czibula, M. I. Bocicor, S. Albert, and A. Pandini, “Deep autoencoders for additional insight into protein dynamics”, in *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)*, Vol. 11140 LNCS (Springer, Cham, Oct. 2018), pp. 79–89.

[275] C. X. Hernández, H. K. Wayment-Steele, M. M. Sultan, B. E. Husic, and V. S. Pande, “Variational encoding of complex dynamics”, Physical Review E **97**, 062412 (2018).

[276] C. Wehmeyer and F. Noé, “Time-lagged autoencoders: Deep learning of slow collective variables for molecular kinetics”, The Journal of Chemical Physics **148**, 241703 (2018).

[277] P. E. Meyer, F. Lafitte, and G. Bontempi, “Minet: A R/Bioconductor package for inferring large transcriptional networks using mutual information”, BMC Bioinformatics **9**, 461 (2008).

[278] M. Bauer, S. M. Schuster, and K. Sayood, “The average mutual information profile as a genomic signature”, BMC Bioinformatics **9**, 48 (2008).

[279] H. Kamberaj and A. Van Der Vaart, “Extracting the causality of correlated motions from molecular dynamics simulations”, Biophysical Journal **97**, 1747–1755 (2009).

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

[281] S. Furini, C. Domen, and S. Cavalcanti, “Insights into the sliding movement of the lac repressor nonspecifically bound to DNA”, Journal of Physical Chemistry B **114**, 2238–2245 (2010).

[282] C. Maffeo, R. Schöpflin, H. Brutzer, R. Stehr, A. Aksimentiev, G. Wedemann, and R. Seidel, “DNA-DNA interactions in tight supercoils are described by a small effective charge density”, Physical Review Letters **105** (2010) `10.1103/PhysRevLett.105.158101`.

[283] E. G. Marklund, A Mahmutovic, O. G. Berg, P Hammar, D van der Spoel, D Fange, and J Elf, “Transcription-factor binding and sliding on {DNA} studied using micro- and macroscopic models”, Proceedings of the National Academy of Sciences **110**, 19796–19801 (2013).

[284] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindah, “Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers”, SoftwareX **1-2**, 19–25 (2015).

[285] A. Perez, I. Marchan, D. Svozil, J. Sponer, T. E. Cheatham, C. A. Laughton, and M. Orozco, “Refinement of the Amber force field for nucleic acids: Improving the description of α/γ conformers”, Biophysical Journal **92**, 3817–3829 (2007).

[286] M. Wieczor, A. Tobiszewski, P. Wityk, B. Tomiczek, and J. Czub, “Molecular Recognition in Complexes of TRF Proteins with Telomeric {DNA}”, PLoS ONE **9**, edited by Y. K. Levy, e89460 (2014).

[287] G. A. Tribello, M. Bonomi, D. Branduardi, C. Camilloni, and G. Bussi, “PLUMED 2: New feathers for an old bird”, Computer Physics Communications **185**, 604–613 (2014).

[288] R. T. McGibbon, K. A. Beauchamp, M. P. Harrigan, C. Klein, J. M. Swails, C. X. Hernández, C. R. Schwantes, L.-P. Wang, T. J. Lane, and V. S. Pande, “MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories”, Biophysical Journal **109**, 1528–1532 (2015).

[289] J. Yoo and A. Aksimentiev, “Improved parameterization of amine-carboxylate and amine-phosphate interactions for Molecular Dynamics simulations using the

CHARMM and AMBER force fields", Journal of Chemical Theory and Computation **12**, 430–443 (2016).

[290]I. Ivani, P. D. Dans, A. Noy, A. Pérez, I. Faustino, A. Hospital, J. Walther, P. Andrio, R. Goñi, A. Balaceanu, G. Portella, F. Battistini, J. L. Gelpí, C. González, M. Vendruscolo, C. A. Laughton, S. A. Harris, D. A. Case, and M. Orozco, "Parmbsc1: a refined force field for DNA simulations", Nature Methods (2015) 10.1038/nmeth.3658.

[291]R. T. McGibbon and V. S. Pande, "Variational cross-validation of slow dynamical modes in molecular kinetics", The Journal of Chemical Physics **142**, 124105 (2015).

[292]K. A. Beauchamp, G. R. Bowman, T. J. Lane, L. Maibaum, I. S. Haque, and V. S. Pande, "MSMBuilder2: Modeling conformational dynamics on the picosecond to millisecond scale", Journal of Chemical Theory and Computation **7**, 3412–3419 (2011).

[293]N. Eswar, B. Webb, M. A. Marti-Renom, M. Madhusudhan, D. Eramian, M.-y. Shen, U. Pieper, and A. Sali, "Comparative Protein Structure Modeling Using Modeller", Current Protocols in Bioinformatics **15**, 1–5 (2006).

[294]M. B. Peters, Y. Yang, B. Wang, L. Füsti-Molnár, M. N. Weaver, K. M. Merz, and Jr, "Structural Survey of Zinc Containing Proteins and the Development of the Zinc AMBER Force Field (ZAFF).", Journal of chemical theory and computation **6**, 2935–2947 (2010).

[295]M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G Scalmani, V Barone, B Mennucci, G. A. Petersson, H Nakatsuji, M Caricato, X Li, H. P. Hratchian, A. F. Izmaylov, J Bloino, G Zheng, J. L. Sonnenberg, M Hada, M Ehara, K Toyota, R Fukuda, J Hasegawa, M Ishida, T Nakajima, Y Honda, O Kitao, H Nakai, T Vreven, J. A. Montgomery Jr., J. E. Peralta, F Ogliaro, M Bearpark, J. J. Heyd, E Brothers, K. N. Kudin, V. N. Staroverov, R Kobayashi, J Normand, K Raghavachari, A Rendell, J. C. Burant, S. S. Iyengar, J Tomasi, M Cossi, N Rega, J. M. Millam, M Klene, J. E. Knox, J. B. Cross, V Bakken, C Adamo, J Jaramillo, R Gomperts, R. E. Stratmann, O Yazyev, A. J. Austin, R Cammi, C Pomelli, J. W. Ochterski, R. L. Martin, K Morokuma, V. G. Zakrzewski, G. A. Voth, P Salvador, J. J. Dannenberg, S Dapprich, A. D. Daniels, Farkas, J. B. Foresman, J. V. Ortiz, J Cioslowski, and D. J. Fox, *Gaussian 09, Revision A.02*, 2009.

[296]A. E. A. Allen, M. C. Payne, and D. J. Cole, "Harmonic Force Constants for Molecular Mechanics Force Fields via Hessian Matrix Projection", Journal of Chemical Theory and Computation **14**, 274–281 (2018).

[297]C. J. Lim, A. J. Zaug, H. J. Kim, and T. R. Cech, "Reconstitution of human shelterin complexes reveals unexpected stoichiometry and dual pathways to enhance telomerase processivity", Nature Communications **8**, 1075 (2017).

[298]S. Hanaoka and A. Nagadoi, "Comparison between TRF2 and TRF1 of their telomeric DNA-bound structures and DNA-binding activities", Protein Science, 119–130 (2005).

[299]D. Vuzman, A. Azia, and Y. Levy, "Searching DNA via a "monkey bar" mechanism: The significance of disordered tails", Journal of Molecular Biology **396**, 674–684 (2010).

[300]T. Kophengnavong, A. S. Carroll, and T. K. Blackwell, "The SKN-1 amino-terminal arm is a DNA specificity segment", Molecular and Cellular Biology **19**, 3039–3050 (2015).

[301]M. Wieczór and J. Czub, "How proteins bind to DNA: target discrimination and dynamic sequence search by the telomeric protein TRF1", Nucleic Acids Research **45**, 7643–7654 (2017).

[302] E. Suárez, J. L. Adelman, and D. M. Zuckerman, "Accurate estimation of protein folding and unfolding times: beyond Markov state models", Journal of Chemical Theory and Computation **12**, 3473–3481 (2016).

[303] M Oda, K Furukawa, K Ogata, a Sarai, and H Nakamura, "Thermodynamics of specific and non-specific DNA binding by the c-Myb DNA-binding domain.", Journal of molecular biology **276**, 571–90 (1998).

[304] Y. Zhang, C. A. Larsen, H. S. Stadler, and J. B. Ames, "Structural basis for sequence specific DNA binding and protein dimerization of HOXA13", PLoS ONE **6**, edited by V. N. Uversky, e23069 (2011).

[305] H.-T. Lee, A. Bose, C.-Y. Lee, P. L. Opresko, and S. Myong, "Molecular mechanisms by which oxidative DNA damage promotes telomerase activity", Nucleic acids research **45**, 11752–11765 (2017).

[306] M. Vorlícková, M. Tomasko, A. J. Sagi, K. Bednarova, and J. Sagi, "8-Oxoguanine in a quadruplex of the human telomere DNA sequence", FEBS Journal **279**, 29–39 (2012).

[307] P. Aller, Y. Ye, S. S. Wallace, C. J. Burrows, and S. Doublié, "Crystal structure of a replicative DNA polymerase bound to the oxidized guanine lesion guanidinohydantoin", Biochemistry **49**, 2502–2509 (2010).

[308] T. Dršata, M. Kara, M. Zacharias, and F. Lankaš, "Effect of 8-oxoguanine on DNA structure and deformability", Journal of Physical Chemistry B **117**, 11617–11622 (2013).

[309] G. La Rosa and M. Zacharias, "Global deformation facilitates flipping of damaged 8-oxo-guanine and guanine in DNA.", Nucleic acids research **44**, 9591–9599 (2016).

[310] E. L. Denchi and T. de Lange, "Protection of telomeres through independent control of ATM and ATR by TRF2 and POT1", Nature **448**, 1068–1071 (2007).

[311] J. S. Z. Li, J. M. Fusté, T. Simavorian, C. Bartocci, J. Tsai, J. Karlseder, and E. L. Denchi, "TZAP: A telomere-associated protein involved in telomere length control", Science **355**, 638–641 (2017).

[312] X. Xu, A. M. Fleming, J. G. Muller, and C. J. Burrows, "Formation of tricyclic [4.3.3.0] adducts between 8-oxoguanosine and tyrosine under conditions of oxidative DNA-protein cross-linking", Journal of the American Chemical Society **130**, 10080–10081 (2008).

[313] S. Perrier, J. Hau, D. Gasparutto, J. Cadet, A. Favier, and J. L. Ravanat, "Characterization of lysine-guanine cross-links upon one-electron oxidation of a guanine-containing oligonucleotide in the presence of a trilysine peptide", Journal of the American Chemical Society **128**, 5703–5710 (2006).

[314] M. J. Solivio, D. B. Nemera, L. Sallans, and E. J. Merino, "Biologically relevant oxidants cause bound proteins to readily oxidatively cross-link at guanine", Chemical Research in Toxicology **25**, 326–336 (2012).

[315] A. Collins, J. Brown, M. Bogdanov, J. Cadet, M. Cooke, T. Douki, C. Dunster, J. Eakins, B. Epe, M. Evans, P. Farmer, C. M. Gedik, B. Halliwell, K. Herbert, T. Hofer, R. Hutchinson, A. Jenner, G. D. Jones, H. Kasai, F. Kelly, A. Lloret, S. Loft, J. Lunec, M. McEwan, L. Möller, R. Olinski, I. Podmore, H. Poulsen, J. L. Ravanat, J. E. Rees, F. Reetz, H. Shertzer, B. Spiegelhalder, R. Turesky, R. Tyrrell, J. Viña, D. Vinicombe, A. Weimann, B. de Wergifosse, and S. G. Wood, "Comparison of different methods of measuring 8-oxoguanine as a marker of oxidative DNA damage", Free Radical Research **32**, 333–341 (2000).

[316] K. Kobayashi and S. Tagawa, "Direct observation of guanine radical cation deprotonation in duplex DNA using pulse radiolysis", Journal of the American Chemical Society **125**, 10213–10218 (2003).

[317] B. H. Munk, C. J. Burrows, and H. B. Schlegel, "An exploration of mechanisms for the transformation of 8-oxoguanine to guanidinohydantoin and spiroiminodihydantoin by density functional theory", Journal of the American Chemical Society **130**, 5245–5256 (2008).

[318] J. Llano and L. A. Eriksson, "Oxidation pathways of adenine and guanine in aqueous solution from first principles electrochemistry", Physical Chemistry Chemical Physics **6**, 4707–4713 (2004).