Dissertation

Theoretical study of isolated and interacting single-stranded DNA

submitted by

Asmar Nayis

$$\dot{q}_i = \frac{\partial \mathcal{H}}{\partial p_i}, \dot{p}_i = -\frac{\partial \mathcal{H}}{\partial q_i}$$

$$\hat{H}\psi = i\hbar \frac{\partial}{\partial t}\psi$$

$$R_{\mu\nu} = \kappa(T_{\mu\nu} - \tfrac{1}{2}g_{\mu\nu}T)$$

$$1 + e^{i\pi} = 0$$
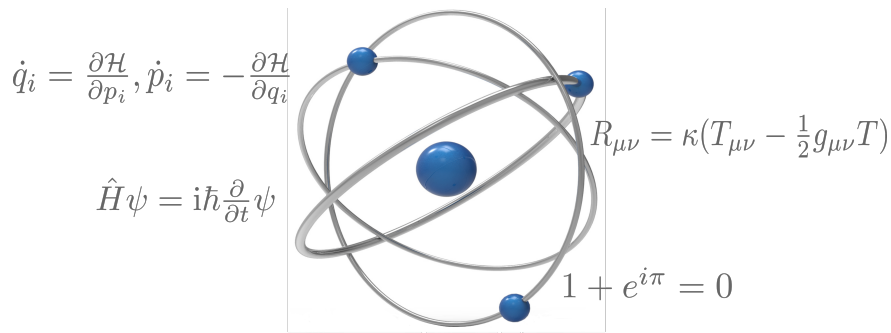
Department of Theoretical Physics

Technische Universität München

Physik Department der Technischen Universität München
Lehrstuhl für Theoretische Biophysik

Theoretical study of isolated and interacting single-stranded DNA

Asmar Nayis

Vollständiger Abdruck der von der Fakultät für Physik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender: Prof. Dr. Friedrich C. Simmel

Prüfer der Dissertation:

      1. Prof. Dr. Martin Zacharias

      2. Prof. Dr. Ulrich Gerland

Die Dissertation wurde am 19.08.2021 bei der Technischen Universität München eingereicht und durch die Fakultät für Physik am 10.11.2021 angenommen.

Contents

1 Introduction

The nature of single-stranded DNA

In the span of one human lifetime more than a light-year of DNA is synthesized. This means over 10^{16} cell divisions! At the same time, the error rate is less than only one mistake per 10^8 nucleotide-additions. The replication fork has to operate at an enormously high speed of ~ 50 nucleotides per second [23]. The high accuracy must be guaranteed since the DNA represents the cellular code and one damage can impair the function of DNA or even induce cell death. The replication process itself is very sophisticated and involves plenty of proteins in almost perfect coordination: Certain proteins trigger local melting in the target DNA sequence into two single strands such that subsequently polymerase proteins find access and start replication of the single strands [107]. But how does this machinery of proteins find and interact with the target DNA? This is a question also concerning the underlying dynamics of the protein and DNA at the same time. The understanding of protein-DNA interactions is a general issue which is an essential part of current research. For instance, specific proteins are capable of finding only the telomeric overhang (i. e. DNA sequences at the ends of chromosomes) to protect it from fusions or lethal DNA-damage response molecules [209, 47, 46, 165]. For telomeres, it is furthermore known that their replication in human cells rarely occurs. Except for few cells like embryonic cells or in some stem cells, missing telomere replication always results in a progressively DNA shortening after each cell division [71, 207]. Cancer cells, in contrast, have an overexpression of telomerases [17]. These are proteins which are perfectly adapted for only replicating the telomeres. Consequently, cancer cells are immortal and their DNA remains fully preserved [145]. For the total US population, the lifetime risk of ever being diagnosed with cancer is approximately 41%. By 2030, it is estimated that 70% of all cancers will occur among adults aged ≥ 65 years [203, 184]. Cancer diseases are always a DNA issue. Recently, certain organometallic drugs gained interest which bind the telomeric sequence of cancer cells with high preference. This inhibits telomerase binding and thus interferes with the rapid growth of long-living cancer cells [145, 6, 64].
The lifetime correlation with telomere lengths was also claimed even a couple of decades ago when the sheep dolly was cloned, for instance. The diminished terminal DNA lengths caused premature aging of the sheep [181, 17].
Single-stranded DNA, however, plays a key role not only in parts of the DNA like the telomeres but also in the entire DNA. Plenty of effects in local DNA regions are mainly independent of the complementary strand. This includes cyclobutane pyrimidine dimer (CPD) damages which are determined by a covalent linkage of two specific neighboring bases induced by UV radiation [61, 174]. Such DNA lesions can block regular DNA replication and thus strongly contribute to mutagenesis in skin cancers [159, 160, 139, 45].

The atomistic fundamentals of DNA form the basis of all of our studies which therefore leads us to the elementary chemical structure of DNA: One strand consists of a backbone and a specific sequence of nucleobases where each is either an adenine, thymine, cytosine, or guanine (see Fig. 1.1). The backbone is composed of a phosphate group and a sugar ring
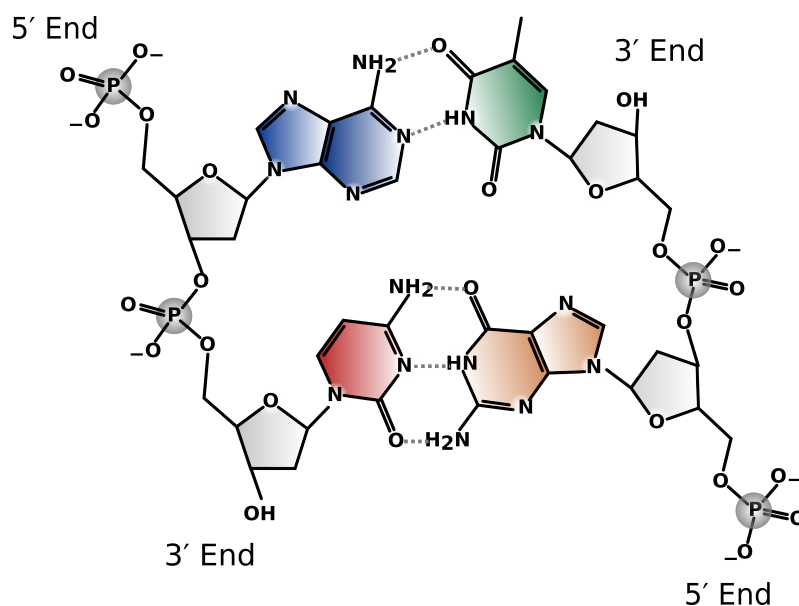
Figure 1.1: *Two base pairs of DNA. Adenine (blue) is paired with thymine (green) and guanine (brown) is paired with cytosine (red). The backbone (grey) consists of a phosphate group and a sugar ring. Source of the figure is Ref [119] (with some modifications).*

which is then linked to the respective nucleobase by a glycosidic bond [3]. The DNA double helix emerges from the hybridization of two single strands. Both strands are polarized in opposite directions. The bases of both strands are paired together in a complementary way where adenine is paired with thymine through two, and cytosine with guanine through three hydrogen bonds [4, 198]. DNA has moreover the property of being negatively charged due to the phosphate group. Strong interaction with other electric and magnetic fields or even with the spin of particles is evidenced [42, 65]. In recent studies, DNA even served as a basic component of nucleic layers to filter equally polarized photoelectrons [65].

However, in experiments about single-stranded DNA much insight was gained in the last couple of decades. Also the microscopes have been progressively improved like the atomic force microscopy [28] or even a quantum-entanglement-based and refined microscopy [32]. But these techniques come with limited insight [39]. Only one microscope can observe the dynamics at the atomic level and femtosecond time resolution: The 'computer-microscope'. Classical Molecular Dynamics (MD) simulations capture atomistic propagations in time by solving Newton's law of motion. Investigation of the flexibility of a system, or the dynamical and conformational changes, for instance, is then easily feasible. When going deeper into the matter to reflect also electric and magnetic effects, a quantum mechanical basis in terms of the Schroedinger equation becomes necessary. This addresses other types of questions. A combination of the classical and quantum mechanical theory seems very promising as we did for our CPD damage research.

In the following, I will share my insights on all projects of my dissertation.

Chapter overview

Motivated by the organometallic drug Au-carbene, which induces cancer cell mortality by inhibition of telomerase access, the first project deals with this interaction of Au-carbene with the cancer quadruplex DNA. It turns out that the Au-carbene undergoes high flexibility upon the quadruplex interface. Nevertheless, very strong binding is determined with an advanced method according to Woo & Roux. Since ingestion of the drug would also affect regular DNA, it is of high importance to have low adverse side effects. Thus, we also computed the binding free energy between the drug and double-stranded DNA which is diminished by half. Further suggestions to reduce the double-stranded DNA affinity are discussed.

Subsequently, we introduce the replication initiator protein (Rep) which marks single-stranded DNA regions in the pre-replication process. As Reps can be multifunctional and thus execute most diverse functions, understanding the binding with single-stranded DNA is essential. So far only fragmented sequences of longer DNAs are found in a binding mode with Rep proteins. We predict, inter alia, how the rest sequences bind Rep proteins. This is a research project along with our collaboration partner Wegrzyn of the University in Gdansk.

The effect of telomere interaction with the Pot1pC protein which protects the telomeres from terminal fusions, for instance, is subject of the next research project. Without DNA accommodation the protein exhibits high flexibility, the capability of opening and closing through a pocket, and the folding of two large connecting loops. This is immediately prevented after fully DNA incorporation which we furthermore quantify with conformational entropy computations. Very promising is the Principal Component Analysis (PCA) approach with some necessary corrections whereby the conformational entropy is obtained from the eigenvalues of the covariance matrix. The DNA incorporation process itself, however, can be subdivided into three main intermediate stages. These stages are discussed and underlined with additionally performed simulations.

The last two research fields are mainly based on quantum mechanical ab initio methods. The first, referred to chapter 6, treats fundamental electric and magnetic properties of single-stranded guanine-rich DNA. We show a strong reliability on the point charge models for different DNA constellations. Furthermore, we calculate the nuclear magnetic shielding of DNA atoms and investigate the effect of electron absorption in terms of dipole moments and polarization.

Finally, we deal with the UV-induced CPD damage. Two major keys are the geometry and the flexibility of two neighboring thymines. It turns out that the distance between the thymines is a dominating factor in this damage. The conformational subspace of sufficient close and appropriately aligned thymines required for CPD damage formation is reached only through thermal fluctuations. This is valid for single-stranded as well as for double-stranded DNA. We quantify these findings and elaborate differences to nucleosomal DNA which exhibits strong structural changes due to the wrapping around the histone.

2 Theoretical Foundations

2.1 Molecular Dynamics Simulations and methods

2.1.1 Integration of Newton's equation of motion

All classical free Molecular Dynamics (MD) simulations of N-particle systems are based on Newton's equation of motion which expresses the propagation in time of each particle and thus the entire system. The main concept is to calculate the potential energy of each particle, derive their forces, and evaluate the new positions and velocities of the particles [75]. The potential energy is of the following form [12]

$$U(\mathbf{r}_1, .., \mathbf{r}_N) = \sum_{\text{bonds}} k_d \cdot (d - d_0)^2 + \sum_{\text{angles}} k_0 \cdot (\theta - \theta_0)^2 +$$
$$+ \sum_{\text{dihedrals}} \frac{V_n}{2} \cdot (1 + \cos(n\phi - \gamma)) + \sum_{i<j} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^6} + \frac{q_i q_j}{\epsilon \cdot r_{ij}} \right), \tag{2.1}$$

what is referred as the so-called 'force field'. The coordinates d, θ, ϕ, and r_{ij} correspond to bond lengths, bond angles, dihedral angles, and inter-atomic distances, respectively. The parameters rely either on empirical data or are numerically computed. Determining all of the parameters has a significant impact on the accuracy of the simulation and is still part of ongoing research. Note that the potential energy contains harmonic oscillators in bonds and angles, a Lennard-Jones- and a Coulomb term but also a dihedral contribution without immediate physical interpretation. It is used to correct for local misdescription [217].

With $\mathbf{F}_i = -\nabla U(\mathbf{r}_1, ..., \mathbf{r}_N)$ and Newton's second law the accelerations can be immediately evaluated. Following the velocity Verlet algorithm [191, 190], first, the new positions are obtained by a Taylor expansion:

$$\mathbf{r}_i(t + \Delta t) \approx \mathbf{r}_i(t) + \Delta t \mathbf{v}_i(t) + \frac{\Delta t^2}{2m_i} \mathbf{F}_i(t) \tag{2.2}$$

for given initial conditions $\mathbf{r}_i(0), \mathbf{v}_i(0)$. The velocities can be derived when evolving the current positions from the one forward time-step position. This finally ends up in

$$\mathbf{v}_i(t + \Delta t) = \mathbf{v}_i(t) + \frac{\Delta t}{2m_i} \big(\mathbf{F}_i(t) + \mathbf{F}_i(t + \Delta t) \big). \tag{2.3}$$

For holonomic, time-independent constraints the time step is set to $\Delta t = 2$ fs [164, 134].

2.1.2 Temperature and pressure coupling

So far the theoretical concept is based on the microcanonical ensemble (NVE) with given energy E. From the physiological point of view, the NVT or NPT ensemble is more appropriate.

In the following, we take into account the two thermodynamic quantities temperature and pressure which we regulate with a thermostat, resp. barostat. One of the primarily used thermostats is the Berendsen thermostat: The system is coupled to a heat bath with a certain temperature, technically accomplished by scaling the velocities $\mathbf{v}_i \mapsto \lambda \cdot \mathbf{v}_i$ with a λ-factor [16, 141] defined as

$$\lambda = \left[1 + \frac{\Delta t}{\tau_T} \cdot \left(\frac{T_0}{T} - 1\right)\right]^{\frac{1}{2}}. \tag{2.4}$$

Here T_0 is the reference temperature, T the current temperature, and τ_T the coupling constant. The latter is recommended to accept values between 0.5 ps and 5.0 ps. Larger values cause a very low kinetic energy fluctuation while lower values risk running into flying ice cube artifacts [69, 33].

In an analog way the pressure coupling can be accomplished by scaling the box size and the coordinates. The scaling factor in $\mathbf{r}_i \mapsto \mu \cdot \mathbf{r}_i$ and $l(t + \Delta t) = \mu l(t)$ is defined as [33]

$$\mu = \left[1 + \frac{\beta \Delta t}{\tau_p} \cdot (P - P_0)\right]^{\frac{1}{3}}. \tag{2.5}$$

Here β is the isothermal compressibility, τ_p the coupling constant, P_0 the reference pressure and P the current pressure.

2.1.3 Periodic boundary conditions

Simulating molecules at physiological conditions has to include explicit solvent molecules such as water molecules and ions. But with an increasing number of atoms in larger sized systems computations get more time-expensive. The apparent way out by setting strict boundaries to the system fails due to artifacts of sticking particles at the boundaries [56]. Instead, a much more satisfying alternative can be achieved by introducing periodic boundary conditions: The simulation system is considered as one unit cell in an infinite lattice of their own copies. Any particle leaving the system, i. e. the unit cell, towards one side has to enter it from the opposite side (periodicity).
This approach leads to the highly relevant consequence that particles from the unit cell interact with those of the periodic image. This side effect can be handled by truncation and shift of the short-range interactions for a predefined cut-off distance (~ 9 Å) [56, 33]. Long-range electrostatic interactions, on the other hand, are treated with the Ewald summation method [56].

2.1.4 Umbrella Sampling

Umbrella Sampling is an advanced sampling method which allows free energy computations along a reaction coordinate ξ [108]. For this purpose, a harmonic biasing potential is added to the system to induce a state change from A to B (see Fig. 2.1). This is typically handled

in several simulation steps (windows). The biasing potential of the i-th umbrella window is of the harmonic form [191]

$$W_i(\xi) = \frac{k}{2}(\xi - \xi_i^{ref})^2, \tag{2.6}$$

whereby only ξ_i^{ref} varies for each window and thus defines the reaction path.

After completed simulations, the corresponding unbiased probability distribution $P_i(\xi)$ can be evaluated from the biased distribution of each umbrella window according to [176, 185]

$$P_i(\xi) = e^{\beta(W_i(\xi)-F_i)} P_i^{bias}(\xi). \tag{2.7}$$

Note that sufficient overlap of the probability distributions of adjacent windows must be
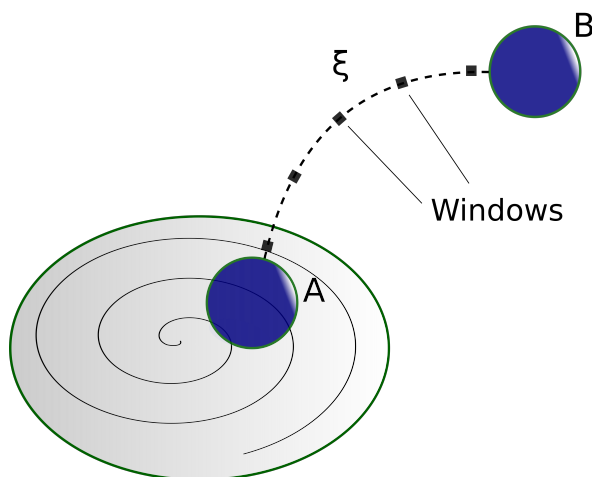
Figure 2.1: *Reaction path from state A to B, separated into distinct windows.*

guaranteed as this is necessary for calculating free energies in terms of the so-called Weighted Histogram Analysis Method (WHAM) [106]. This computes the global unbiased probability distribution $P(\xi)$ of the system which contains the information about the desired free energy difference between state A and B.

$P(\xi)$ is the linear combination of all unbiased distributions from the umbrella windows [185, 191]:

$$P(\xi) = \sum_i c_i(\xi) P_i(\xi). \tag{2.8}$$

The weighting factors c_i are required to be normalized [185]. Utilizing this and minimizing the statistical error by setting $\frac{\partial \sigma^2(P)}{\partial c_i} = 0$ will result in [185, 191, 176]

$$P(\xi) = \sum_i \left(\frac{n_i P_i^{bias}(\xi)}{\sum_j n_j \mathrm{e}^{-\beta(W_j(\xi)-F_j)}} \right), \tag{2.9}$$

whereas n_i is the number of sampling steps in the i-th window.
In order to obtain the free energy difference it is also necessary to have the free energy factors F_i satisfying the relation [191]

$$\mathrm{e}^{-\beta F_i} = \int \mathrm{e}^{-\beta W_i(\xi)} P(\xi) \mathrm{d}\xi. \tag{2.10}$$

Both of the last two equations, (2.9) and (2.10), contain the free energy factors as well as the global probability distribution which itself contains the free energy factors implicitly. Thus, the solution for the free energy difference is found iteratively [191].

2.1.5 Free Energy Perturbation and Thermodynamic Integration

Free energy perturbation and thermodynamical integration are other approaches to calculate the free energy between two arbitrary thermodynamic states A and B. These states differ in potential energies U_A and U_B which yields the free energy difference [191]

$$\Delta F_{A,B} = F_B - F_A = -k_B T \ln\left(\frac{Z_B}{Z_A}\right) \tag{2.11}$$

with the 'configurational' (momentum-free) partition functions Z_A and Z_B as [191]

$$Z_A = \int \mathrm{e}^{-\beta U_A(\mathbf{r}_1,...,\mathbf{r}_N)} \mathrm{d}^N\mathbf{r}, \tag{2.12}$$

$$Z_B = \int \mathrm{e}^{-\beta U_B(\mathbf{r}_1,...,\mathbf{r}_N)} \mathrm{d}^N\mathbf{r}. \tag{2.13}$$

The quotient in the logarithm of (2.11) can be rewritten [191]:

$$\frac{Z_B}{Z_A} = \left\langle \mathrm{e}^{-\beta(U_B(\mathbf{r}_1,...,\mathbf{r}_N)-U_A(\mathbf{r}_1,...,\mathbf{r}_N))} \right\rangle_A. \tag{2.14}$$

Here the bracket $\langle\cdot\rangle_A$ indicates the ensemble average with respect to A. Together with (2.11) we obtain the *free energy perturbation* formula [191]

$$\Delta F_{A,B} = -k_B T \ln\left\langle \mathrm{e}^{-\beta(U_B-U_A)} \right\rangle_A. \tag{2.15}$$

This formula can be interpreted as follows: We take configurations $\{\mathbf{r}_1, ..., \mathbf{r}_N\}$ from the distribution of state A and apply them to state B with the different potential U_B. But this is in fact not sampled which makes it necessary to correct the sampling with the factor $\exp(\beta U_A)$.

However, the approach still exhibits difficulties, particularly, when an overlap of state A and B configurations is poorly underlying. In this case, the potential energy difference $U_B - U_A$ is large and thus the exponential factor becomes negligibly small causing slow free energy convergences. This can be solved by introducing intermediate states defining an optimal transformation path from state A to B. One simple way is to construct a linear transformation by defining the potential function [191]

$$U(\mathbf{r}_1, ..., \mathbf{r}_N, \lambda) := (1 - \lambda)U_A + \lambda U_B, \tag{2.16}$$

whereby λ accepts values between 0 and 1.

Moreover, the free energy difference can be written as (fundamental theorem of calculus)

$$\Delta F_{A,B} = \int_0^1 \frac{\partial F}{\partial \lambda} \mathrm{d}\lambda \tag{2.17}$$

and the derivative of the free energy as the integrand is

$$\frac{\partial F}{\partial \lambda} = -\frac{k_B T}{Z} \frac{\partial Z}{\partial \lambda} = \int \frac{\partial U}{\partial \lambda} \mathrm{e}^{-\beta U(\mathbf{r}_1, ..., \mathbf{r}_N, \lambda)} \mathrm{d}^N \mathbf{r} = \left\langle \frac{\partial U}{\partial \lambda} \right\rangle. \tag{2.18}$$

Z is the partition function with respect to the non-physical potential function U in (2.16). The *thermodynamical integration formula* [191] becomes then

$$\Delta F_{A,B} = \int_0^1 \left\langle \frac{\partial U}{\partial \lambda} \right\rangle \mathrm{d}\lambda \stackrel{\text{linearity}}{=} \int_0^1 \langle U_B - U_A \rangle_\lambda \mathrm{d}\lambda, \tag{2.19}$$

which can be integrated numerically.

2.1.6 Umbrella Sampling including restraints

The advanced sampling method according to Woo & Roux [206] is to some extent a generalization of the umbrella sampling method. For trickier systems it can be appropriate to calculate the free energy along a reaction path in a totally restraint system. Since these restraints bias the real free energy, it will be necessary to correct or reconsider the energetic contributions due to the overall system restrictions.

In the following we refer to [206] when determining the absolute free binding energy for the case of a ligand pulled out from a receptor.

First, the equilibrium binding constant can be written as

$$K_{eq} = \frac{\int_{\text{site}} d\mathcal{L} \int e^{-\beta U} d\mathbf{X}}{\int_{\text{bulk}} \delta(\mathbf{r} - \mathbf{r}^*) d\mathcal{L} \int e^{-\beta U} d\mathbf{X}}. \tag{2.20}$$

Here U is the total potential energy of the system, $\mathcal{L}$ represents the ligand's degrees of freedom, $\mathbf{X}$ represents those of the solvent and receptor, and $\mathbf{r}^*$ is some arbitrary but fixed location of the ligand in the far away bulk region.

The equilibrium binding constant can also be expressed as a product including orientational, conformational, and axial contributions:

$$\begin{aligned}
K_{eq} = &\frac{\int_{\text{site}} d\mathcal{L} \int e^{-\beta U} d\mathbf{X}}{\int_{\text{site}} d\mathcal{L} \int e^{-\beta(U+U_c)} d\mathbf{X}} \times \frac{\int_{\text{site}} d\mathcal{L} \int e^{-\beta(U+U_c)} d\mathbf{X}}{\int_{\text{site}} d\mathcal{L} \int e^{-\beta(U+U_c+U_o)} d\mathbf{X}} \\[2mm]
&\times \frac{\int_{\text{site}} d\mathcal{L} \int e^{-\beta(U+U_c+U_o)} d\mathbf{X}}{\int_{\text{site}} d\mathcal{L} \int e^{-\beta(U+U_c+U_o+U_a)} d\mathbf{X}} \times \frac{\int_{\text{site}} d\mathcal{L} \int e^{-\beta(U+U_c+U_o+U_a)} d\mathbf{X}}{\int_{\text{bulk}} \delta(\mathbf{r} - \mathbf{r}^*) d\mathcal{L} \int e^{-\beta(U+U_c+U_o)} d\mathbf{X}} \\[2mm]
&\times \frac{\int_{\text{bulk}} \delta(\mathbf{r} - \mathbf{r}^*) d\mathcal{L} \int e^{-\beta(U+U_c+U_o)} d\mathbf{X}}{\int_{\text{bulk}} \delta(\mathbf{r} - \mathbf{r}^*) d\mathcal{L} \int e^{-\beta(U+U_c)} d\mathbf{X}} \times \frac{\int_{\text{bulk}} \delta(\mathbf{r} - \mathbf{r}^*) d\mathcal{L} \int e^{-\beta(U+U_c)} d\mathbf{X}}{\int_{\text{bulk}} \delta(\mathbf{r} - \mathbf{r}^*) d\mathcal{L} \int e^{-\beta U} d\mathbf{X}} \\[2mm]
= &\; e^{\beta G_c^{\text{site}}} \cdot e^{\beta G_o^{\text{site}}} \cdot e^{\beta G_a^{\text{site}}} \cdot e^{-\beta G_{\text{bind}}^{\text{restr}}} \cdot e^{-\beta G_o^{\text{bulk}}} \cdot e^{-\beta G_c^{\text{bulk}}}.
\end{aligned} \tag{2.21}$$

U_c, U_o and U_a correspond to the conformational, orientational, and axial restraint potentials, respectively.

The term $e^{-\beta G_{\text{bind}}^{\text{restr}}}$ can be expressed as a product $S \cdot I$ of two integrals concerning the spherical and the radial expansion explicitly. The isotropy of the bulk gives justification to compute the orientational bulk term analytically. All other terms of (2.21) are evaluated with the free energy perturbation method. With C_0 as the standard state concentration of 1 mol/liter the absolute binding free energy becomes

$$G_{\text{bind}} = -\frac{1}{\beta} \ln(K_{eq} C_0) = G_{\text{bind}}^{\text{restr}} + G_o^{\text{bulk}} + G_c^{\text{bulk}} - G_a^{\text{site}} - G_o^{\text{site}} - G_c^{\text{cite}}. \tag{2.22}$$

2.2 Quantum physics

2.2.1 Hartree-Fock method

This method is an approximation of the wave function of a quantum many-body system and its corresponding energy by solving the stationary Schrödinger equation of that system. For our purposes, this is related to the electrons surrounding the nucleus of each atom. As the nucleus is several thousand times heavier than the electrons and consequently clearly slower, an assumption of separated electronic and nuclear motions is reasonable (Born-Oppenheimer

approximation) [158, 88]. Concretely, the Hamiltonian of the N-electron system is [158, 169]

$$\hat{H} = \sum_i^N \underbrace{\left(-\frac{\Delta_i}{2} - \sum_k^{N_k} \frac{Z_k}{|\mathbf{r_i} - \mathbf{R_k}|} \right)}_{\hat{h}_i} + \frac{1}{2} \sum_i^N \sum_{j \neq i}^N \underbrace{\frac{1}{|\mathbf{r_i} - \mathbf{r_j}|}}_{\hat{w}_{ij}} \tag{2.23}$$

with the electron coordinates $\mathbf{r}_i$, fixed nucleus coordinates $\mathbf{R}_k$, electron number N, nucleus number N_k, and nucleus charges Z_k. The last electron-electron interaction term complicates finding the solution Ψ of the Schrödinger equation $\hat{H}\Psi = E\Psi$. In the mean-field ansatz, this term is substituted by an average field each electron experiences and which only depends on the single-electron coordinate $\mathbf{r}_i$. The solution of this approximated Schrödinger equation with the mean-field Hamiltonian $\hat{H}^{mf}$ is a product of all electron wave functions [112, 169]. With the variational principle

$$\partial E = \partial \langle \Psi | \hat{H}^{mf} | \Psi \rangle = 0 \tag{2.24}$$

the energy can be obtained [169, 158].

In terms of the initial Hamiltonian (2.23) the solution has to satisfy both, the indistinguishability of the electrons and the Pauli exclusion principle. Due to the alternating nature of determinants, the latter has to vanish for two identical columns, resp. rows. Thus, we make use of that and choose the ansatz of the so-called Slater determinate [169, 158] for the absolute electronic wave function:

$$\Psi \approx \frac{1}{\sqrt{N!}} \begin{vmatrix} \phi_1(1) & \phi_1(2) & ... & \phi_1(N) \\ \phi_2(1) & \phi_2(2) & ... & \phi_2(N) \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ \phi_N(1) & \phi_N(2) & ... & \phi_N(N) \end{vmatrix}. \tag{2.25}$$

We have $\phi_i(\mathbf{r}, \sigma) = \varphi_i(\mathbf{r}) \cdot \alpha_i(\sigma)$ with the spatial orbitals φ_i and the spin functions α_i [169, 158] which also must be taken into account as the electronic wave function depends on these. In the following, the Dirac notation is used. The energy is given by

$$E = \langle \Psi | \hat{H} | \Psi \rangle. \tag{2.26}$$

E can be understood as a functional of the set of wave functions ϕ_i. With the Lagrange multipliers the functional

$$\mathcal{L} = E[\{\phi_i\}] - \sum_i^N \sum_j^N \epsilon_{ij}(\delta_{ij} - \langle\phi_i|\phi_j\rangle) \tag{2.27}$$

is then minimized by means of the variational principle [169, 158]. This yields the Hartree-Fock equation [112, 169, 158, 88]

$$\hat{F}|\phi_m\rangle := \hat{h}|\phi_m\rangle + \sum_l^N (\langle\phi_l|\hat{w}|\phi_l\rangle|\phi_m\rangle - \langle\phi_l|\hat{w}|\phi_m\rangle|\phi_l\rangle) = \epsilon|\phi_m\rangle. \tag{2.28}$$

$\hat{h}$ contains the kinetic energy as well as the Coulomb interaction with the nuclei. The second term contains the mean Coulomb potential of all the other electrons. Hence, the Hartree-Fock method is a mean-field approach. The solution is determined iteratively with the method of self-consistent fields [169, 158].

2.2.2 Orbital basis functions

Any molecular orbital Ψ_i is expressed as a linear combination of N underlying basis functions Φ_j [99]

$$\Psi_i = \sum_{j=1}^N c_{ij}\Phi_j. \tag{2.29}$$

The size of this set of basis functions is of high relevance as on the one hand, the choice of a basis at a minimum size likely yields inaccuracies and on the other hand, larger basis sets are accompanied by high computational effort. The type of the basis functions also plays a major role whereby mainly two orbital types are distinguished: The *Slater-type orbitals* (STO) and the *Gaussian-type orbitals* (GTO).
STOs resemble the eigenfunctions of the hydrogen atom and have the general form [99]

$$\Phi_{sto} = const \cdot r^{n-1}\exp(-\xi r) \cdot Y_{lm}(\theta,\varphi) \tag{2.30}$$

where the radial part including the first three terms depends on the radius r from the origin of the basis functions, usually the nucleus position, with parameter ξ and principal quantum number n. The last spherical function depends on the two spherical angles with the angular quantum number l and magnetic quantum number m.
On the downside of the STOs, evaluation of corresponding integrals gets very computationally expensive. GTOs, in contrast, benefit from their special analytical form since the product of two GTOs can be written as a linear combination of GTOs.

The general form is [99]

$$\Phi_{gto} = const \cdot x^l y^m z^n \exp(-\alpha r^2) \tag{2.31}$$

with cartesian coordinates x, y, z and constants l, m, n, α.

2.2.3 Density functional theory

From the density functional theory (DFT) a different, more elegant solution of a many-electron system can be determined. In contrast to Hartree-Fock, the underlying functionals which depend on electron-densities form the basis. It turns out that introducing electron-densities and thus reducing the many-body problem to a single-body problem is accompanied by significantly less computational cost.
The Hamiltonian has the general form [8, 73]

$$\hat{H}\Psi = (\hat{T} + \hat{V} + \hat{U})\Psi = E\Psi \tag{2.32}$$

where $\hat{T}$ is the kinetic energy operator, $\hat{V}$ is the potential energy operator from the external field due to the nuclei, and $\hat{U}$ is the electron-electron interaction energy operator. While $\hat{T}$ and $\hat{U}$ are the same for any N-electron system, $\hat{V}$ varies due to its system-dependence.
The key variable in DFT is the electron density $n(\mathbf{r})$ given by [8, 73]

$$n(\mathbf{r}) = N \int \Psi^*(\mathbf{r}, \mathbf{r}_2, \dots, \mathbf{r}_N)\Psi(\mathbf{r}, \mathbf{r}_2, \dots, \mathbf{r}_N) \, \mathrm{d}^3\mathbf{r}_2 \dots \mathrm{d}^3\mathbf{r}_N \tag{2.33}$$

for a normalized Ψ. With this density given, reversely, Ψ can also be interpreted as a functional depending of $n(\mathbf{r})$. Thus, also any operator is referred to $n(\mathbf{r})$. The external potential energy $\langle \Psi[n]|\hat{V}|\Psi[n]\rangle$ can be written as [8, 73]

$$V[n] = \int V(\mathbf{r})n(\mathbf{r}) \, \mathrm{d}^3\mathbf{r}. \tag{2.34}$$

The energy functional [8, 73, 101]

$$E[n] = T[n] + U[n] + \int V(\mathbf{r})n(\mathbf{r}) \, \mathrm{d}^3\mathbf{r} \tag{2.35}$$

can be then minimized which yields the ground state electron density $n_0(\mathbf{r})$. The minimization is a variational problem which is solved with Lagrangian multipliers [101]. At first, an explicit electron–electron interaction is omitted and the functional

$$E_s[n] = \langle \Psi_{\mathrm{s}}[n]|\hat{T} + \hat{V}_{\mathrm{s}}|\Psi_{\mathrm{s}}[n]\rangle \tag{2.36}$$

is varied where $\hat{V}_{\text{s}}$ is an effective potential which the electrons experience.

This yields the Kohn–Sham equations [101]

$$\left(-\frac{1}{2}\nabla^2 + V_{\text{s}}(\mathbf{r}) \right) \varphi_i(\mathbf{r}) = \varepsilon_i \varphi_i(\mathbf{r}). \tag{2.37}$$

Solving this equation ends up in an expression for the orbitals φ_i and hence for the electron density [101]

$$n(\mathbf{r}) = \sum_{i=1}^{N} \left| \varphi_i(\mathbf{r}) \right|^2. \tag{2.38}$$

The effective potential $V_{\text{s}}(\mathbf{r})$ has the form [8, 101]

$$V_{\text{s}}(\mathbf{r}) = V(\mathbf{r}) + \int \frac{n(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} \, \text{d}^3\mathbf{r}' + V_{\text{xc}}[n](\mathbf{r}). \tag{2.39}$$

Here $V(\mathbf{r})$ describes the external potential, the second term is the electron–electron repulsion potential and the last term $V_{\text{xc}}[n](\mathbf{r})$ is the exchange and correlation potential per electron [8, 101]. The solution is found iteratively as the second and third term depend on $n(\mathbf{r})$, thus on $\varphi_i(\mathbf{r})$ which again depends on $V_{\text{s}}(\mathbf{r})$ (2.37). In general, after starting with an initial guess for $n(\mathbf{r})$ one computes $V_{\text{s}}(\mathbf{r})$ and subsequently solves the Kohn–Sham equations for the $\varphi_i(\mathbf{r})$. This yields a new electron density. The same steps are repeated until convergence is reached.

2.2.4 Time-dependent Kohn-Sham theory

The analog to the stationary based DFT is the time-dependent Kohn-Sham formalism where the Kohn-Sham equations have to be satisfied according to [129, 157]

$$\left(-\frac{1}{2}\nabla^2 + V_{ks}(\mathbf{r}, t) \right) \varphi_i(\mathbf{r}, t) = i\frac{\partial}{\partial t} \varphi_i(\mathbf{r}, t), \quad \varphi_i(\mathbf{r}, 0) = \varphi_i(\mathbf{r}) \tag{2.40}$$

for a system with non-interacting electrons. V_{ks} depends on the electron density defined by [129]

$$n(\mathbf{r}, t) = \sum_{i=1}^{N} \left| \varphi_i(\mathbf{r}, t) \right|^2 \tag{2.41}$$

in consensus with the real system.

Analogously to the previous paragraph we obtain [129, 157]

$$V_{\text{ks}}(\mathbf{r}, t) = V_{\text{ext}}(\mathbf{r}, t) + \int \frac{n(\mathbf{r}', t)}{|\mathbf{r} - \mathbf{r}'|} \, \mathrm{d}^3 \mathbf{r}' + V_{\text{xc}}(\mathbf{r}, t). \tag{2.42}$$

Note that in principle it is possible to find more than one wavefunction for different initial wavefunctions and thus producing different V_{ks} potentials [129, 157]. This would make this theory practically useless. But for most cases, the initial-state dependence can be avoided when having an initial state being in a non-degenerate ground state. Then the Hohenberg-Kohn theorem [100] guarantees that the wavefunction is a functional depending on the ground-state density alone.

3 Targeting Telomeres: Binding of Cancer-Candidate Gold-Carbene to DNA [1]

3.1 Introduction

Telomeres are repetitive noncoding DNA sequences at the ends of chromosomes which can form G-quadruplexes for tandem repeats of guanine nucleotides [117, 37, 24]. These quadruplexes are guanine-rich nucleic acid structures where the guanines form a square plane with several possible levels. The structural motifs are stabilized through Hoogsten hydrogen bonds and monovalent cations bound at a central cavity.
Recently, the stabilization of G-quadruplexes has gained increased attention in the field of drug design, as it may interfere with the growth of cancer cells [168, 6, 67, 79, 218, 148]. That is, drug induced stabilization of DNA quadruplexes prevents elongation of telomeres. Telomere elongation is carried out by a reverse transcriptase as part of the telomerase complex [7], which requires unfolding of the DNA quadruplex structures prior to elongation [173, 170]. Consequently, stabilization of the DNA quadruplex structures may inhibit telomerase binding and thus telomere maintenance. Shortening of the telomeres, in turn, can induce cell apoptosis which is suppressed in cancer cells due to telomerase overexpression. Thus, down-regulation of telomerase fidelity through quadruplex stabilization may represent a promising anti-cancer strategy [180, 94, 53].

In this regard, Gold(Au)-carbene is a promising candidate in the new field of metal complexes. Carbon-gold bonds significantly reduce the risk of non-specific damage in non-cancer cells as a consequence of their regulated redox properties [21, 19, 18]. Bertrand and coworkers have successfully synthesized a class of Au-carbene complexes that allow for efficient uptake by cells and show strong anti-proliferative activity against cancer cells. In the past, crystal structures formed by DNA quadruplexes and Au-carbenes have been solved [13].

Further improvement of the affinity and specificity of Au-carbene binding to quadruplexes requires an understanding of the binding mechanism in molecular detail. Thus, we employed molecular dynamics and free energy simulations to analyze Au(9-methylcaffein-8-ylidene)$_2$+ (abbr. Au-carbene) binding to DNA. These simulations indicate enormous differences in the mobilities of the ligand at the three observed binding sites (see Fig. 3.1). Whereas very restricted mobility of two Au-carbene ligands was found upon binding as a doublet to one side of the G-quadruplex, a much larger translational and orientational mobility was observed for the single Au-carbene binding at the other G-quadruplex surface. The binding of Au-carbenes also reaches an increased DNA-quadruplex melting temperature and thus selective inhibition of telomerase as it is experimentally observed [154, 64, 21, 18, 19, 13]. Furthermore, simulations on duplex DNA exposed in an Au-carbene filled box indicate a strong preference for minor groove binding. Such binding to regular double-stranded DNA (dsDNA) in healthy cells is an important issue for two reasons: First, Au-carbene as a

---

drug is aimed to be minimally cytotoxic and harmful for healthy cells with regular DNA, and second, the insight of minor groove binding is highly valuable to propose possible drug modifications to pursue increased quadruplex over duplex selectivity. Hence, we raise the important question of how strong Au-carbenes bind dsDNA minor grooves and how much it varies from quadruplex binding. The answer is given by comparative free energy simulations in a subsequent study: Here we found a significantly lower affinity for dsDNA minor groove binding. Moreover, dsDNA binding can be successively weakened when stepping towards physiological or higher salt concentrations.
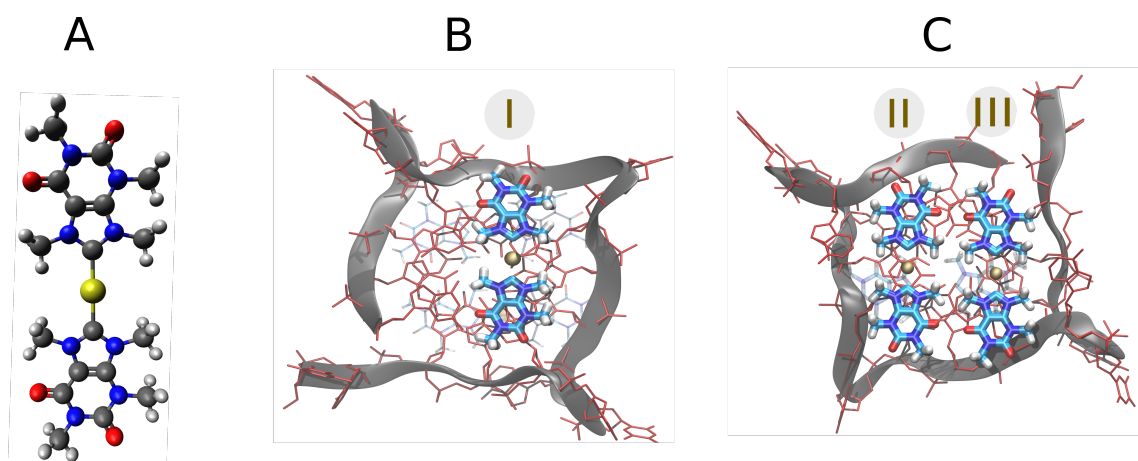
Figure 3.1: *A) Structure of the Au-carbene with a central gold atom. B) Crystal structure of Au-carbene in complex with the telomeric G-quadruplex viewed from the upside (singlet state or mode I) and C) from the downside with two additional bound Au-carbenes (doublet state, mode II and III).*

3.2 Methods

3.2.1 Simulations Details

For all simulations the Amber16 Molecular Dynamics MD Package [33], in combination with the bsc1 force field [82] for nucleic acids, was applied (for force field parameterization of the ligand, see below). The starting structure (ProteinData Bank, PDB: 5CCW) was solvated in explicit water (TIP3P) [91] with rectangular boxes and a 10 Å minimal distance between complex and box boundary. Additional sodium counterions were added to neutralize the system. After energy minimization (10000 steps), the system was heated up to 300 K within three stages (each 0.5 ns) and with a restraint (force constant: 2.388 kcal/(mol · Å^2)) to the starting positions (on all non-hydrogen atoms). Subsequently, the positional restraining was reduced to 0.6 kcal/(mol · Å^2), followed by a short equilibration without restraints (1 ns) under constant pressure of 1 bar. After this equilibration step, the data-gathering MD simulations were carried out with the pmemd.cuda module at a temperature of 300 K and in NVT ensemble for up to 2 μs.

3.2.2 Force field parameterization of Au-carbene

Derivation of the force field parameters for the Au-carbene was performed with the MCPB.py module of Amber16 [33] in combination with the Gaussian09 package [114, 57]. This includes geometry optimization, frequency calculation and Merz-Kollmann population analysis on the B3LYP/6-31G* level of theory [15]. For the gold atom, we used the LANL2DZ basis set [70]. Partial charges were assigned with the RESP module of Amber16, whereby we constrained identical atoms in the symmetric carbene dimer to have the same charge. Bond and angle parameters were obtained with the Seminario method, which excludes dihedral parameters around the gold atom [179]. A Lennard-Jones radius of 1.6465 Å and depth of 0.0390 was assigned to the gold atom.

3.2.3 Scheme for absolute binding free energy calculations

For the absolute energy calculations we use the advanced sampling method developed by Woo and Roux [206]. The main idea behind this method is to restrain the overall complex, consecutively pull the stacked Au-carbene apart from the DNA and finally correct or reconsider the energetic contributions due to the employed restrictions.
The restraints are defined between three virtual sites (VS 1-3) in the quadruplex, resp. dsDNA, and three virtual sites in the Au-carbene molecule (VS 4-6), with each virtual site being defined as the geometrical center-of-mass. The PMF is calculated along the distance vector r. Axial orientations are controlled by the angle θ and the dihedral angle ϕ while the rotational orientation is described by three Euler angles: α, β, and γ. The internal Au-carbene ring rotation is controlled by the dihedral angle ζ_{Dih} (see Fig. 3.2 and Table 3.1). Obtaining a standard binding free energy ΔG_{bind} requires several in-between steps: First, the restrained PMF along the center-of-mass distance coordinate r is calculated via Umbrella Sampling by pulling the Au-carbene molecule out of an instantaneous binding position from the DNA. The simulations cover a distance range from 4.0 to 12.8 Å in case of Au-carbene-quadruplex binding and a distance range from 9.1 to 17.6 Å in case of Au-carbene-B-DNA binding. Each Umbrella window is simulated for 10 ns with a distance-restraint force constant of 10 $\frac{\text{kcal}}{\text{mol} \cdot \text{Å}^2}$ and a force constant of 240 $\frac{\text{kcal}}{\text{mol} \cdot \text{rad}^2}$. This yields $\Delta G_{\text{bind,restr}}$. Next, the underlying axial and orientational restraints are corrected via Free Energy Perturbation (FEP) and the Zwanzig

Table 3.1: *All Woo-Roux restraints and types listed*

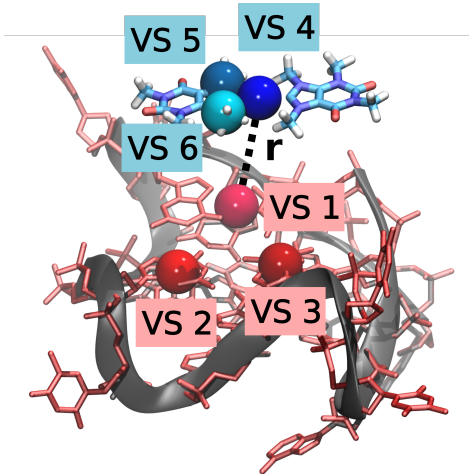| Restraint	Definition	Type
r	VS1-VS4	distance
θ	$\measuredangle$(VS2-VS1-VS4)	axial angle
ϕ	$\measuredangle$(VS3-VS2-VS1-VS4)	axial dihedral
α	$\measuredangle$(VS1-VS4-VS5)	orientational angle
β	$\measuredangle$(VS2-VS1-VS4-VS5)	orientational dihedral
γ	$\measuredangle$(VS1-VS4-VS5-VS6)	orientational dihedral
ζ_{Dih}	$\measuredangle$(N1-C1-C10-N5)	internal dihedral

Figure 3.2: *Virtual sites (VS) for the definition of restraints between the quadruplex structure (VS1-3) and the Au-carbene molecule (VS4-6). The distance vector r for PMF calculation is defined between VS1 and VS4. The axial orientation is controlled by the angle θ and the dihedral angle φ, the rotational orientation is described by three Euler angles: α, β and γ.*

equation [216] at the binding site ($\Delta G_{\text{a,site}}$, $\Delta G_{\text{o,site}}$) and numeric calculation in the bulk ($\Delta G_{\text{o,bulk}}$, see equations [206]). In this manner the axial and orientational restraint force constants are gradually reduced from 240 to 0 $\frac{\text{kcal}}{\text{mol·rad}^2}$ in steps of 24 $\frac{\text{kcal}}{\text{mol·rad}^2}$ and each window is simulated for 10 ns. Note that $\Delta G_{\text{a,bulk}}$ is corrected for implicit within $\Delta G_{\text{bind,restr}}$ [206]. Finally, the free energy associated with the conformational dihedral restraint ζ_{Dih} is obtained via Umbrella Sampling when rotating the internal Au-carbene to ± 3.14 rad in steps of 0.31 rad, with 10 ns simulation length per window and a dihedral restraint force constant of 100 $\frac{\text{kcal}}{\text{mol·rad}^2}$. Boltzmann-weighted integration of the PMF determines $\Delta G_{\text{c,site}}$ and $\Delta G_{\text{c,bulk}}$.

Summa summarum, this yields

$$\Delta G_{\text{bind}} = \Delta G_{\text{bind,restr}} + (\Delta G_{\text{o,bulk}} + \Delta G_{\text{c,bulk}}) - (\Delta G_{\text{a,site}} + \Delta G_{\text{o,site}} + \Delta G_{\text{c,site}}). \quad (3.1)$$

3.2.4 Molecular mechanics Poisson-Boltzmann binding energy calculations

For calculation of mean interaction energies, the MMPBSA (molecular mechanics Poisson-Boltzmann surface area) method was employed [188, 138, 81, 196] using the MMPBSA.py module of Amber16. In the MMPBSA approach, a trajectory generated in explicit solvent is re-evaluated with the same force field but the solvent replaced by a dielectric continuum (with a dielectric constant, $\epsilon = 80$ for solvent and $\epsilon = 1$ for the solute). The electrostatic reaction field is calculated using the finite difference Poisson-Boltzmann approach (as implemented in Amber 16), and nonpolar solvation contributions are obtained from the solvent accessible

surface area [196]. For most calculations, a salt concentration of 150 mM was used. The internal K_+ ions of the quadruplex were included explicitly in the calculations. For the application of Au-carbene binding to the G-quadruplex, the average of over 4000 regularly spaced simulation snapshots were taken from 2 μs simulations.

3.3 Results and Discussion

3.3.1 High mobility for the Au-carbene in singlet state

The crystal structure of a G-quadruplex with a human telomeric DNA sequence (23 nucleotides with sequence: 5'-TAGGGTTAGGGTTAGGGTTAGGG) in complex with Au(9-methylcaffein-8-ylidene)$_2$ (Au-carbene) ligands forms the basis of our Molecular Dynamics (MD) simulations. The structure includes three Au-carbene molecules that bind noncovalently to the two planar surfaces on opposite sides of the G-quadruplex (see mode I-III in Fig. 3.1). One side (the 'top') of the G-quadruplex accommodates one single Au-carbene (singlet state), whereas on the other side (the 'bottom'), two Au-carbenes are bound. In the crystal structure, the single Au-carbene prefers an off-center location due to stronger stacking overlap with guanines. In addition, the G-quadruplex DNA formation is stabilized by two potassium ions at the center between the planes formed by four guanines.
The MD simulation systems are categorized into one with the quadruplex without ligands (apo), with a single accommodated Au-carbene (the singlet state), and the quadruplex simultaneously interacting with all three Au-carbenes on both planar surfaces (triplet state). The root-mean-square deviation (RMSD) of the DNA referred to the crystal structure increased at the beginning significantly to approximately 5-7 Å (Fig. 3.3). The RMSD of the core guanine nucleotides that form the elementary construct of the quadruplex remained in all cases close to the start structure (2-3 Å) indicating that the rearrangement of the connecting loops and terminal nucleotides is responsible for the increased total RMSD. The Root-mean-square fluctuations (RMSF) also reveals an additional stabilization feature of the accommodated Au-carbene(s) that results in a reduced conformational fluctuation of the connecting loops, particularly residue 16.
Interestingly, the Au-carbene bound as singlet on the top surface undergoes significant motions with an RMSD fluctuating between 2-10 Å (Fig. 3.4). The Au-carbene is even capable of rotating and translating upon the overall unbounded neighborhood (Fig. 3.4B). This behavior is seen in both, the simulations with one bound Au-carbene and in the simulation with three Au-carbenes. The simulations indicate a variety of accessible placements of the single Au-carbene and three main arrangements (B1-B3) can be distinguished. In contrast, the two Au-carbenes on the bottom side plane (mode II and III) are much less mobile due to larger restrictions in free space and therefore no rotation or translation was observed.

Note: For qualitative investigations of the stability, the G-quadruplex-Au-carbene complex was also simulated at elevated temperatures of 350 K. At this temperature no dissociation of the ligands and a similar pattern of RMSD and RMSF was observed. This agrees qualitatively with the experimental result of a stable complex at elevated temperatures [21].
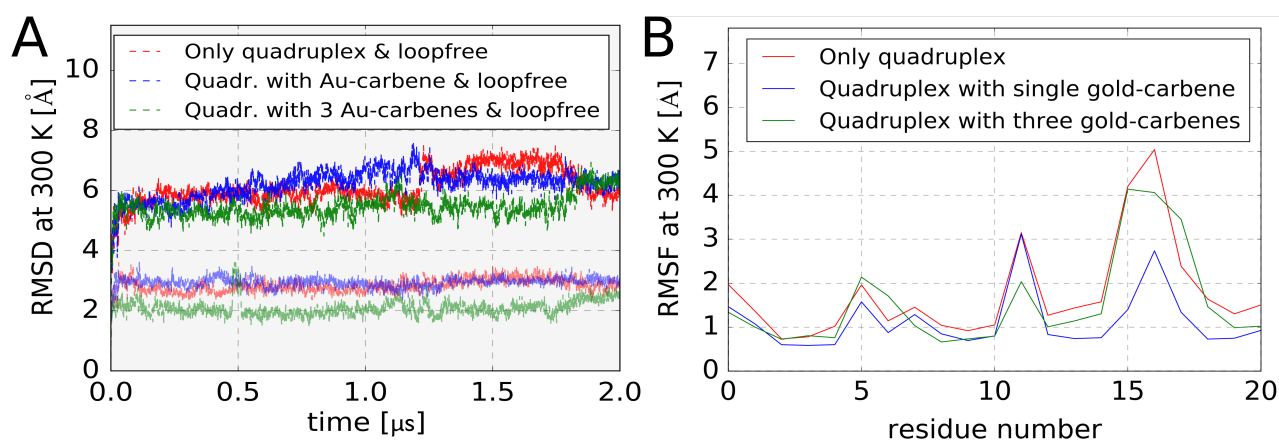
Figure 3.3: *(A) Root-mean-square deviation (RMSD) of all non-hydrogen DNA atoms with respect to the start structure vs. simulation time for MD simulations (300 K) of only the G-quadruplex (dark red), G-quadruplex with one stacked Au-carbene (dark blue), and G-quadruplex with 3 Au-carbenes as observed in the crystal structure (dark green). The RMSD was also computed only for the guanosine nucleotides involved in the quadruplex (without connecting loops and terminal nucleotides) in corresponding transparent colors. (B) Root-mean-square fluctuations (RMSF) of the respective MD-simulations vs. nucleotide (same color scheme as in (A)). The RMSF represents averages over all non-hydrogen atoms of each nucleotide.*
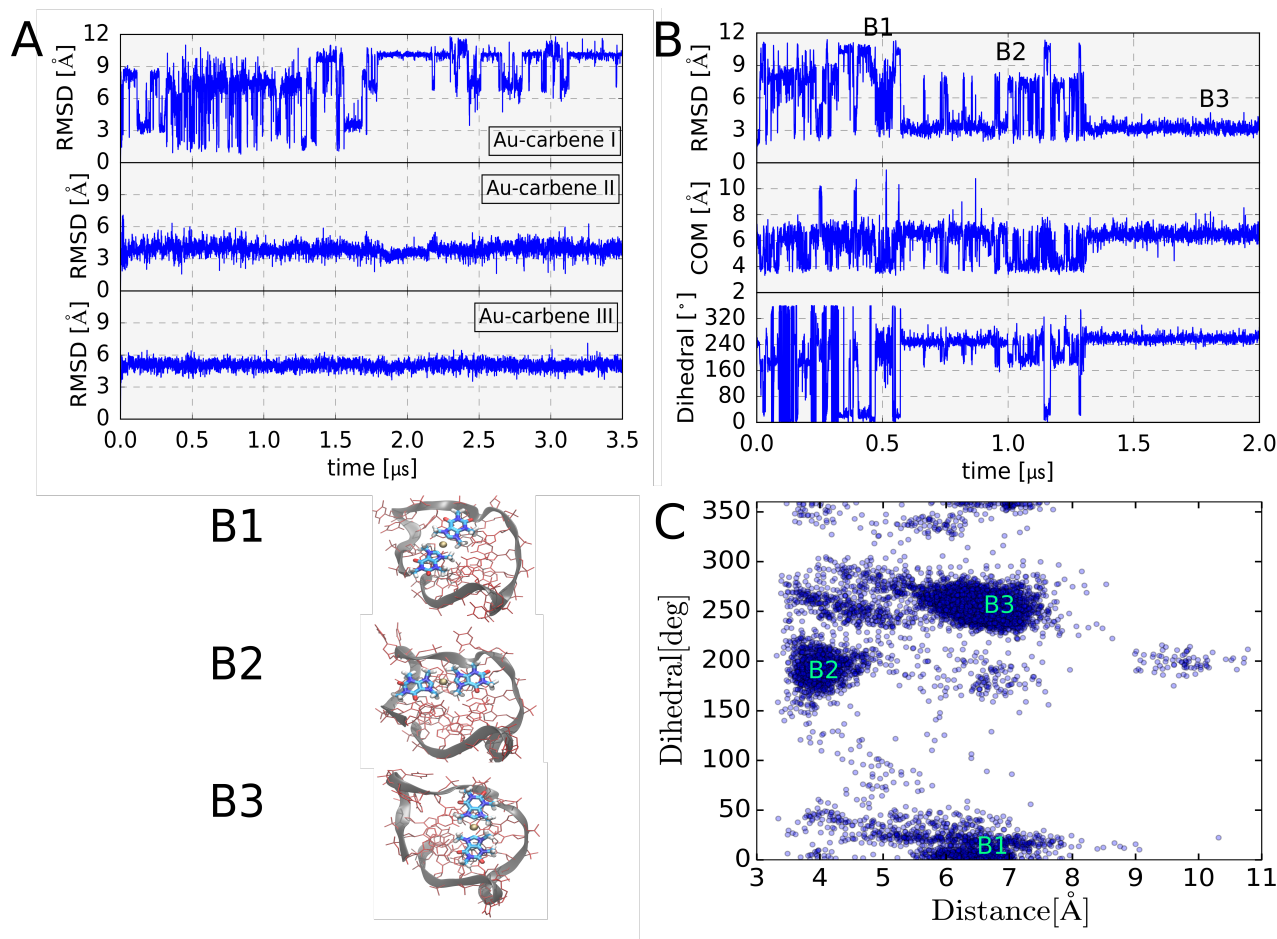
Figure 3.4: *(A) RMSD vs. simulation time (300 K) of the Au-carbene ligands after best superposition of the G-quadruplex on the starting structure. (B) RMSD of the Au-carbene for the G-quadruplex with a single bound Au-carbene. In addition, the center-of-mass (com) distance between ligand and G-quadruplex is evident, as well as the dihedral angle describing the orientation of the Au-carbene relative to the G-quadruplex. The dihedral angle is defined by each of the two carbenes and two top-planed neighboring nucleotides. (C) Sampled distribution of Au-carbene on the side with a single bound ligand in terms of com-distance and angular orientation upon the G-quadruplex surface.*

3.3.2 Selective minor groove double-stranded DNA binding

We have studied the binding of Au-carbene to duplex-DNA by exposing the DNA targets to a high concentration of several randomly placed Au-carbene ligands. The MD simulations reveal clear minor-groove binding behavior of the Au-carbene ligand. We observed binding locations at several minor groove segments and for the entire microsecond-long MD simulation time. In contrast, no stacking was observed for the major groove. As we see later, the electrostatic attraction is significantly stronger compared to binding with the quadruplex. This is due to the negative charge density in the DNA minor groove attracting the positively charged Au-carbene cation. Furthermore, hydrogen bonds between an oxygen of the drug and a hydrogen atom of guanine's amino group have a stabilizing effect (Fig 3.5).

However, no crystal structures of the complex between duplex-DNA and Au-carbene have been reported. It is known that Au-carbenes bind double-stranded DNA. Our prediction of the binding complex specifically in the minor groove is hence novel. This insight might be useful in the ongoing research field of drug design which gained much interest recently. Constructing an Au-carbene-like molecule which is slightly enlarged in size might reduce minor groove attraction but in contrast, is still energetically favored of binding G-quadruplexes. In the previous paragraph, we have even seen two Au-carbene simultaneously binding on one site. This doublet state is also highly favorable and experimentally observed [21, 19]. The design of a covalent linkage between two Au-carbenes may hence be a route to increase affinity to quadruplexes while binding to the DNA minor grooves is strongly weakened.

The most long-lived complex was rapidly formed within a few ns and remained bound for the
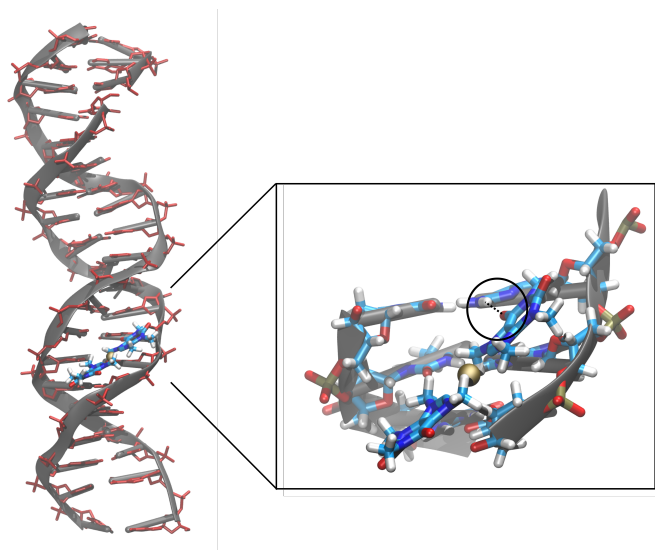

Figure 3.5: *Simulated binding complex between Au-carbene and double-stranded DNA. Besides electrostatic attraction, the binding into DNA's minor groove is stabilized by a hydrogen bond between an Au-carbene-oxygen and a hydrogen atom of guanine's amino group.*

entire simulation. It was used for subsequent comparative binding free energy simulations.

3.3.3 Binding free energy disparities between quadruplex and duplex DNA

The MD simulations of Au-carbene binding to the G-quadruplex and the B-DNA minor groove give qualitative insights into the stability of the complexes. We used a restrained Umbrella Sampling approach to calculate the standard binding free energy ΔG_{bind} according to a scheme by Woo and Roux [206]. It allows for a quantitative comparison of Au-carbene binding to the quadruplex structure vs. binding in the B-DNA minor groove. The approach employs geometrical and conformational restraints during the calculation of a potential of mean force (PMF) along a distance coordinate to separate the binding partners. The inclusion of restraints allows for rapid convergence of the calculated free energy along the separation coordinate. The contributions to the binding due to the release of restraints on axial rotation, orientation between quadruplex/B-DNA and Au-carbene as well as relaxation of Au-carbene conformational restraints can be calculated separately (see Fig. 3.6 and Methods section).

In case of quadruplex binding, the calculated absolute binding free energy ΔG_{bind} amounts to $-10.4 \frac{kcal}{mol}$ which corresponds to a twice as high binding affinity compared to $-5.2 \frac{kcal}{mol}$ for binding to B-DNA (see Table 3.2 and 3.3). For G-quadruplex binding, an affinity in the nanomolar regime was found experimentally [13] which agrees qualitatively with the present result (-10.4 $\frac{kcal}{mol}$ corresponds to a K_d=0.03 μM). The main difference between quadruplex and B-DNA binding results from the restrained Potential of Mean Force (PMF) upon dissociation,

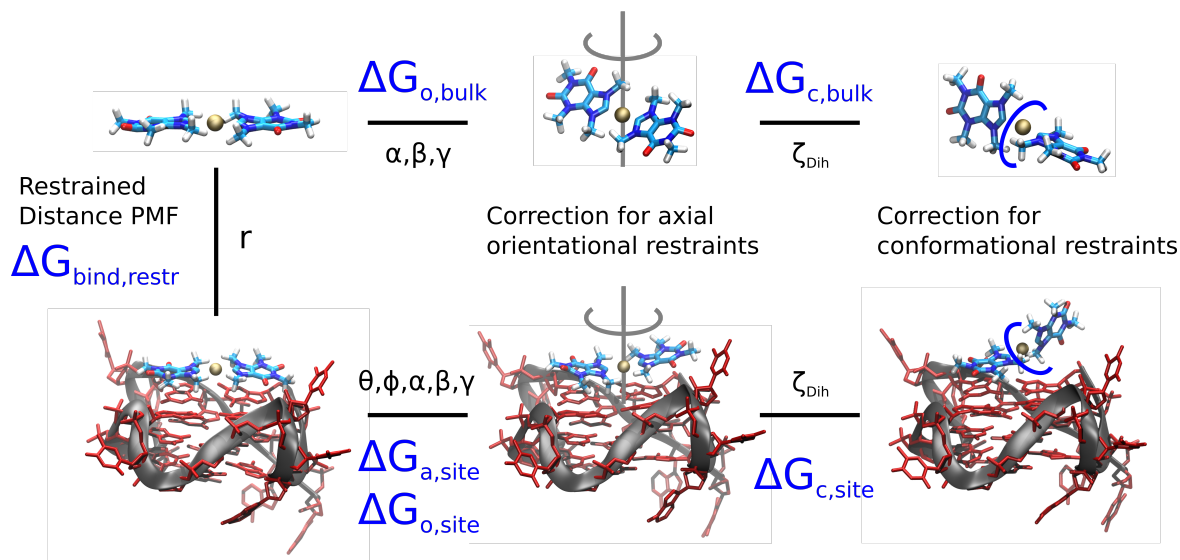**Figure 3.6:** *Schematic illustration of all contributions for absolute binding free energy calculations in the quadruplex case.*

Table 3.2: *Woo-Roux binding free energy for Au-carbene-G-quadruplex binding*

Contribution	Result $[\frac{\text{kcal}}{\text{mol}}]$	Std. error $[\frac{\text{kcal}}{\text{mol}}]$
$\Delta G_{\text{bind,restr}}$	-14.9	$\pm$ 0.7
$\Delta G_{\text{o,bulk}}$	6.4	—
$\Delta G_{\text{c,bulk}}$	3.6	$\pm$ 0.1
$-\Delta G_{\text{a,site}}$	-0.6	$\pm$ 0.2
$-\Delta G_{\text{o,site}}$	-3.1	$\pm$ 0.6
$-\Delta G_{\text{c,site}}$	-1.8	$\pm$ 0.1
ΔG_{bind}	-10.4	$\pm$ 1.0

Table 3.3: *Woo-Roux binding free energy for Au-carbene-dsDNA binding*

Contribution	Result $[\frac{\text{kcal}}{\text{mol}}]$	Std. error $[\frac{\text{kcal}}{\text{mol}}]$
$\Delta G_{\text{bind,restr}}$	-10.9	$\pm$ 0.9
$\Delta G_{\text{o,bulk}}$	6.4	—
$\Delta G_{\text{c,bulk}}$	3.7	$\pm$ 0.1
$-\Delta G_{\text{a,site}}$	-0.3	$\pm$ 0.1
$-\Delta G_{\text{o,site}}$	-1.5	$\pm$ 0.7
$-\Delta G_{\text{c,site}}$	-2.6	$\pm$ 0.1
ΔG_{bind}	-5.2	$\pm$ 1.2

which yields a free energy difference between the binding site and the bulk of $-14.9 \frac{\text{kcal}}{\text{mol}}$ for quadruplex- and $-10.9 \frac{\text{kcal}}{\text{mol}}$ for dsDNA-binding (see Figure 3.7 A,B). This free energy difference is obtained between the bound state and the converged PMF plateau for large distances. The orientational free energy contribution $\Delta G_{\text{o,bulk}}$ of Au-carbene in bulk amounts to around $6.4 \frac{\text{kcal}}{\text{mol}}$ in both cases due to bulk isotropy and independence of the receptor surface. From the PMF along the Au-carbene-inherent dihedral angle ζ_{Dih} in the bound state it turns out that slightly tilted Au-carbene rings being most favorable. In contrast to unbound Au-carbene, the perpendicular ring conformation is considerably less favorable in the bound state, which can be explained by the loss of ring stacking interactions with DNA and especially the quadruplex surface. The latter leads to a reduced conformational flexibility and hence a smaller value of $\Delta G_{\text{c,site}} = 1.8 \frac{\text{kcal}}{\text{mol}}$ compared to $2.6 \frac{\text{kcal}}{\text{mol}}$ in case of B-DNA binding. In addition, the large orientational mobility of the Au-carbene in mode I bound to the quadruplex ($\Delta G_{\text{o,site}}$) results in a significant stabilization compared to B-DNA minor groove binding (Table 3.2 and 3.3).

However, in the case of quadruplex binding, the PMF shows a significant local minimum at around 7.6 Å. We can explain this minimum by a strong increase in the number of water molecules in the first hydration layer around the Au-carbene which occurs while dissociation.
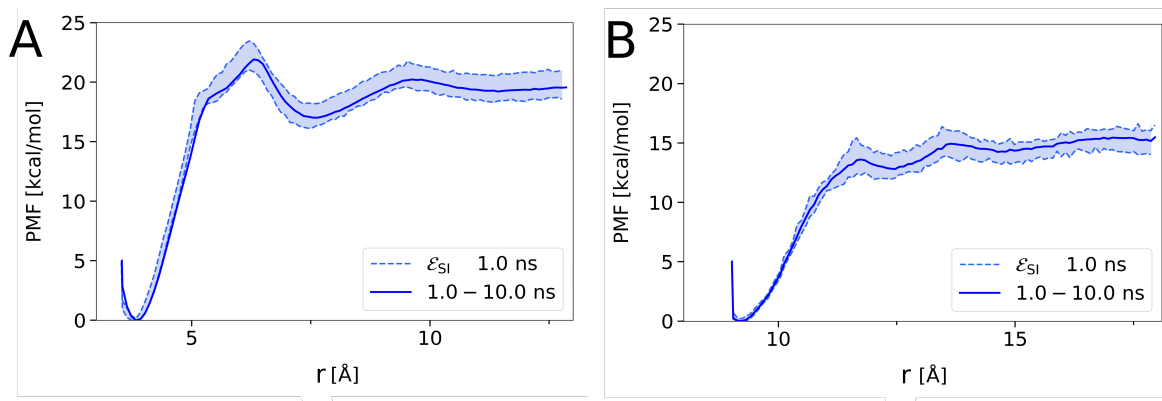
Figure 3.7: *Calculated free energy change along the Au-carbene-DNA separation distance r for restrained complexes. A) PMF for Au-carbene separation from the singlet binding site on the quadruplex. The uncertainty of the PMF ϵ_{SI} is estimated by splitting the data set into subintervals of 1 ns (dotted curves and blue shadow). B) PMF for Au-carbene dissociation with respect to the minor groove binding site of the duplex DNA.*

3.3.4 MMPBSA interaction energy calculations

To investigate the energetic contributions to Au-carbene-DNA binding we employed the Molecular Mechanics Poisson Boltzmann surface area (MMPBSA) method to re-analyse the MD-trajectories based on a continuum solvent model (see Methods). It also allows us to estimate the interaction energy of Au-carbenes at the bottom side relative to the top side of the G-quadruplex. The total MMPBSA interaction energies of the Au-carbene molecule binding in mode I amount to -17.3 $\frac{kcal}{mol}$ for G-quadruplex DNA and -38.7 $\frac{kcal}{mol}$ for dsDNA at a physiological salt concentration of 150 mM (Table 3.4). Note that the calculated interaction energy by means of MMPPBSA does not account for the re-organisation of the DNA due to ligand association, it does not account for conformational entropy contributions and it employs an implicit solvent description. Hence, it is not comparable with the calculated binding free energies of the previous paragraph. The calculated mean interaction energies of the two Au-carbenes binding to the opposite side yield -30.5 $\frac{kcal}{mol}$ and -38.4 $\frac{kcal}{mol}$ for placements II and III, respectively, which indicates strong interactions of both Au-carbenes. When considering both Au-carbenes II and III as one unit, the mean interaction energy to the G-quadruplex turns out to be -64.4 $\frac{kcal}{mol}$. This is less than the sum of the binding energies of II and III, which is -68.9 $\frac{kcal}{mol}$. Hence, there is a mean attraction between both ligands at the bottom side further stabilizing the complex. Indeed, experiments indicate positive cooperation for Au-carbene binding to quadruplexes [21]. The calculated mean Coulomb contribution of the Au-carbene interaction with the B-DNA is 66% higher compared to the Au-carbene interaction with the G-quadruplex. The unfavorable polar solvation energy is always higher than the absolute Coulomb attraction energy. The total electrostatic contribution of mode I binding amounts to 29.5 $\frac{kcal}{mol}$, for mode II and III, respectively, to 32.0 $\frac{kcal}{mol}$ and 24.8 $\frac{kcal}{mol}$. Binding to B-DNA, on the other hand, gives only 18.3 $\frac{kcal}{mol}$ of total electrostatic repulsion. Accord-

Table 3.4: *MMPBSA Binding energy contributions for Au-carbene-DNA binding $[\frac{kcal}{mol}]$*

binding mode	$< E_{bind} >$	$< E_{vdW} >$	$< E_{Coulomb} >$	$< E_{pol-solv} >$	$< E_{nonpol} >$	$\frac{\Delta < E_{bind} >}{\Delta ln(I)}$
quadruplex I	-17.3 ± 4.0	-43.7 ± 3.6	-453.2 ± 12.3	482.7 ± 13.2	-2.9 ± 0.2	-0.14
quadruplex II	-30.5 ± 4.7	-58.8 ± 3.2	-446.7 ± 10.8	478.7 ± 12.1	-3.7 ± 0.1	-0.02
quadruplex III	-38.4 ± 4.2	-59.0 ± 3.9	-459.5 ± 11.3	484.3 ± 12.8	-4.2 ± 0.2	-0.05
quadruplex II+III	-64.4 ± 5.7	-110.7 ± 4.7	-956.8 ± 17.0	1010 ± 18.0	-6.9 ± 0.2	–
dsDNA	-38.7 ± 4.4	-53.2 ± 5.2	-750.5 ± 18.2	768.8 ± 20.5	-3.9 ± 0.3	-1.81
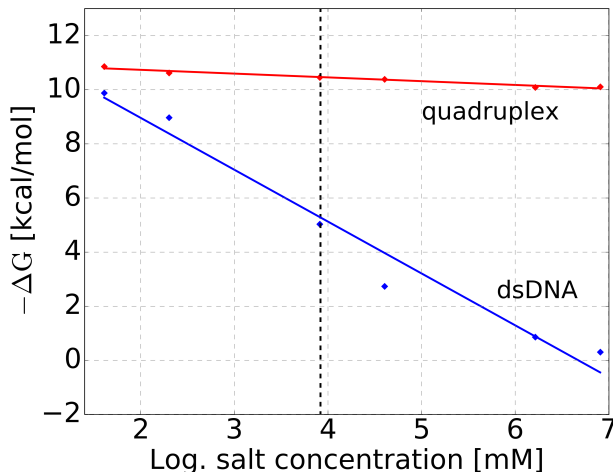
Figure 3.8: *Calculated relative binding free energy of Au-carbene-DNA binding vs. salt concentration. Increased salt concentrations significantly reduce the binding strength of Au-carbene to dsDNA but have minimal effect on binding to the G-quadruplex (mode I binding). The vertical dashed line represents approximately the salt concentration used in the binding free energy calculations.*

ing to the MMPBSA analysis, the main driving force for binding is due to Van-der-Waals interactions and nonpolar solvation contributions in all cases.

3.3.5 Impact to DNA for increased salt-concentration

The MMPBSA method can be further used to predict the change in binding affinity for increasing salt concentrations. Only a very modest salt dependence of free binding energy is obtained for the Au-carbene in complex with the G-quadruplex, in contrast to a significant binding affinity reduction for the binding to B-DNA (Table 3.4, Fig. 3.8). At around 150 mM (appr. physiological salt concentration) the binding free energy for quadruplex-Au-carbene mode I is still ≈-10 $\frac{kcal}{mol}$ (K_d=0.03 μM) while the free binding energy of the dsDNA-Au-carbene complex is reduced to ≈-3 $\frac{kcal}{mol}$ (K_d=6.7 mM). It agrees qualitatively with the experimental results which demonstrate specific binding of Au-carbenes to G-quadruplex under physiological conditions and no significant competition of binding due to 15 fold excess

of B-DNA [21].

3.4 Conclusion

Gold-carbenes as a new class of organometallic drugs can bind and stabilize G-quadruplex structures formed in telomere sequences at chromosome termini [53, 67]. Keeping the termini of chromosomes intact is essential for replication. The Au-carbene molecule generally manifests meaningful anticancer properties and G-quadruplex binding properties correlated with antitelomerase activity [123]. The stabilization of G-quadruplexes and inhibition of telomerase binding is a possible route to interfere with the rapid growth of long-living cancer cells [6, 64].

Understanding the molecular details of the Au-carbene interaction with G-quadruplexes and competing interactions with B-DNA is a prerequisite for the design of Au-carbenes to improve affinity and specificity. Our present free energy simulations for the quadruplex exhibit stronger binding of both Au-carbenes to the bottom side compared to the top-sided single bound Au-carbene. Surprisingly, a much higher mobility of the Au-carbene in the mode I binding was observed which allowed large translation of the ligand and complete rotation on the top surface of the G-quadruplex. The clear minor groove dsDNA binding was observed when arbitrary B-DNA was exposed to a box filled with randomly placed Au-carbenes. Several minor groove activities of Au-carbene were detected but never in association with the major groove. After Woo-Roux binding free calculations we confirmed the experimental observation [13] of specific strong Au-carbene binding to G-quadruplexes compared to unspecific binding in B-DNA minor groove. Binding to the quadruplex turned out to be twice as strong compared to dsDNA binding. In addition, with increased salt concentration this result could be further enhanced as the quadruplex is mainly unaffected for salt concentration alternations.

With the above-mentioned we discussed designing a covalent linkage between the two Au-carbenes which could be a route for strongly increased binding affinity. This construction would also immediately refuse binding to the detected minor groove binding in B-DNA. From a similar point of view, the unusual large mobility - demonstrating accessible sterical space - can be exploited by attaching additional chemical groups to the carbene compounds to increase non-polar interactions with the G-quadruplex. This modification to slightly larger Au-carbenes could also reduce the binding to the narrow B-DNA minor groove.

However, for the application as drugs, one should keep in mind that not only the specificity and affinity of Au-carbenes is of critical importance but also possible side effects and toxicity of these metal containing complexes [89].

4 Prediction of single-stranded DNA binding domains of plasmid Rep proteins

4.1 Introduction

DNA replication is a vital process to transmit genetic information in all organisms. In the standard replication procedure [83], initiation of DNA replication requires a DNA sequence as the replicator and an initiator sequence which binds to the replicator (origin of replication) [68, 60]. The involved DNA region DUE (DNA unwinding element) is a sequence rich in adenine and thymine present within the origin region [199, 25]. In most plasmid replicons, initiation requires a plasmid-encoded initiator protein, named Rep [68]. These Reps bind dsDNA containing iterons which results in DUE melting. Hence, plasmid Reps bind iteron dsDNA and DUE ssDNA simultaneously [211, 199]. Although Reps have been extensively investigated, so far it is not known how specific Reps bind ssDNA precisely. This has a high biological relevance as Reps can be multifunctional and trigger various functions [68, 30]. Biochemical analyses revealed that the sequence or, possibly, the secondary DNA structure is critical for both, the replicating activity of a replicon and crucial multiprotein interactions [166, 199]. The investigation of the nucleoprotein complexes formed at the AT-rich regions of the replication origins is essential for further understanding of the regulation and the molecular dynamical mechanisms involved in the initial origin opening and subsequent processes of replication initiation [166].

In this context, two proteins of the Rep group, RepE and TrfA, are part of the following discussion. For RepE a crystal structure with ssDNA exists but only for a short crystallized sequence of a widely longer known ssDNA (unpublished). With several unrestrained simulations it was possible to predict binding sites of a nine-bases elongated ssDNA. This is treated initially. Subsequently, we deal with the other Rep group protein TrfA. For the latter, no crystal structure with ssDNA exists so far. Some key information about interacting domains are known including an attractive interaction of a certain amino acid. Free simulations of a short ssDNA strand in explicit solvent together with TrfA revealed one major binding configuration which shows many parallels with the RepE-ssDNA binding domain.

This is a research project in collaboration with Dr. Wegrzyn, University in Gdansk.

4.2 Methods

4.2.1 Simulations Details

For all simulations the Amber16 Molecular Dynamics Package [33] with the bsc1 force field [82] for nucleic acids and ff14SB [122] for proteins was applied. The elongation of ssDNA from the given crystal structure was accomplished manually with nab [33] of the Amber package. In terms of the TrfA protein, the single-stranded DNA was positioned close to the expected binding domain of the protein. The starting structures were solvated in TIP3P octahedron boxes and an 18 Å minimum distance between complex and box boundary. In

explicit solvent (TIP3P) [91] additional sodium counter-ions neutralized the system (100 mM). After energy minimization (10000 steps) the system was heated up to 300 K within three stages and with positional restraints (force constant: 2.388 kcal/(mol · Å^2)) to the starting positions. Subsequently, these restraints were reduced gradually to 0 in three stages where each stage was run for 250 ps. After a 1 ns equilibration the MD simulations were carried out with the pmemd.cuda module at a temperature of 300 K and in NVT ensemble.

4.3 Results and Discussion

4.3.1 Elongated single-stranded DNA binding domain of RepE

The ssDNA DUE region, appearing during DNA replication initiation, is far longer with approximately 70 nt (5' CCACGGGGATA**TTTTTATA**ATTATTTTTT... 3' with repetitive sequence in bold) than it was possible to use in crystallography analysis. The crystal structure with accommodated single-stranded DNA shows only a fraction of the much longer ssDNA interacting with RepE. The aim of in silico assays is to predict how and where longer ssDNA interacts with RepE.

For this occasion, the short ssDNA sequence (5' ATTTTTA 3') from the crystal structure was extended according to the known DNA structure by nine bases (5' ATTTTTA + TAAT-TATTT 3'). The elongated DNA appendix so far resides in solvent unbound to the protein. Several free simulations with random starting velocities are run. After 4 μs for each trajectory several binding states of the elongated DNA have consolidated while the most frequent, stable structure is shown in Fig. 4.1. The interacting domain with the elongated ssDNA
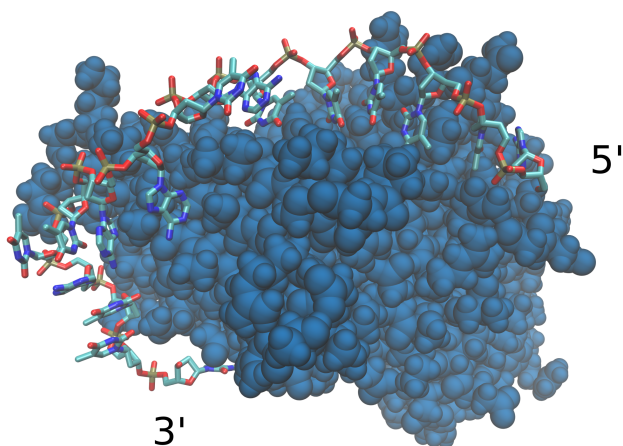

Figure 4.1: *The most sampled and stable binding state of the elongated ssDNA in complex with the RepE. The additional nine bases are attached to the 3' direction (all nucleic acids depicted without hydrogens). The first adenine of the extended DNA tract is accommodated in a minor local cavity which seems to improve complex stability.*

segment consists of many participating amino acids in the following sense: The first adenine is accommodated in a minor local pocket composed of two isoleucines, two serines, a valine, and a tyrosine. This binding location seems highly favorable and complex stabilizing. The middle part of the elongated ssDNA section is accentuated by its backbone stacking, mainly to arginine and lysine. A possible rearrangement may lead to a different binding state. The last three thymines interact mainly with the four amino acids asparagine, glutamine, valine, and arginine. This interaction may be weak and the binding unstable. It is possible that the length of the DNA sequence which forms a complex with RepE is exceeded.

Note: Additionally performed simulations of an elongated ssDNA to the 5'-ending showed no protein binding with the elongated sequence. It seems that a bound conformation state is likely unfavored.

4.3.2 Localized single-stranded DNA binding domain of TrfA

TrfA proteins bind double-stranded DNA as well as single-stranded DNA. For the complex with dsDNA, a crystal structure already exists and is provided by the group of Wegrzyn (collaboration). On the other hand, a crystal structure with TrfA and single-stranded DNA is not known. Since the secondary structure of TrfA and RepE is primarily congruent, except for the N-terminal region where TrfA has an additional domain, the binding site for ssDNA is expected to be approximately at the same position as in RepE. Moreover, at this binding site arginine R230 is found which is known to interact with ssDNA (collaboration: binding assays and mass spectrometry). Besides, from the crystal structures (RepE-ssDNA, [144]) and from in vitro experiments with mutants (collaboration) it is also known that the binding sites of single-stranded DNA and double-stranded DNA are different.

With this information equipped, the aim is to locate the binding site of ssDNA bound to TrfA. A simulation was started where an appropriate short ssDNA strand (5' ATTGG 3') was located close to the presumed TrfA binding domain. This short strand is a part of the longer ssDNA which naturally interacts with TrfA (5' AAAACAAGGTTTATAAATATTG-GTTTAAAAGCAG... 3', collaboration). That is, the simulation includes now an additional short ssDNA together with the TrfA and dsDNA from the crystal structure. After 4 μs simulation time a stable complex of TrfA and ssDNA (and dsDNA) was formed which is depicted in Fig. 4.2. First, it is evident that both nucleic acids are accommodated simultaneously and single-stranded and double-stranded DNA have distinct binding sites. So far it is uncertain whether missing dsDNA occupation causes a rearrangement of the protein, particularly of the top-sided regions (see Fig. 4.2) which may lead to different conformation states of the ssDNA or even to refusal of ssDNA stacking. Second, we find that several nucleic acids are in interaction with the demanded arginine R230 (Fig. 4.2, marked in dark grey). Presumably, the underlying π-stacking is strongly favored and contributes significantly to the complex stabilization. An additional stabilizing effect is captured by the adenine which resides in a local minor pocket. The pocket is formed of a serine, three phenylalanines, a histidine, and an arginine. Altogether, there is a high binding site overlap when comparing this interacting short ssDNA with the (particularly non-elongated) ssDNA bound to RepE.

However, the collaboration partner also detected a binding configuration of TrfA only with the single-stranded DNA (not published). For this purpose, we try to mimic the binding state by removing the double-stranded DNA. After geometry optimization and equilibration, several simulations with different starting velocities were run for 4 μs. It turns out that none of the trajectories show a stable complex. The single-strand dissociates in each case. From the simulation point of view, this suggested complex can be not verified.
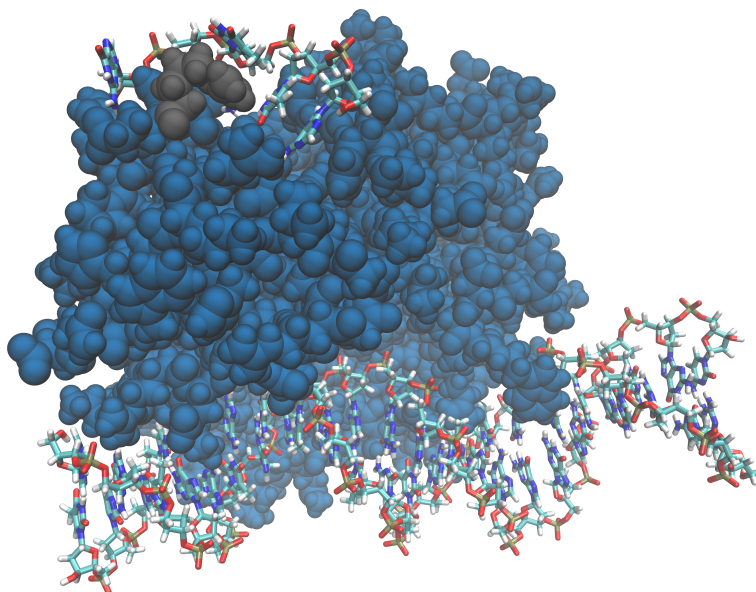
Figure 4.2: *Complex of TrfA (blue), dsDNA, and short ssDNA. The binding of dsDNA and TrfA keeps close to the crystal structure during the simulation. Simultaneous binding of the short ssDNA shows the expected different top-sided binding domain. The ssDNA-interacting arginine R230 is marked in dark grey.*

4.4 Conclusion

Reps, such as RepE or TrfA, exhibit favored binding to AT-rich DNA which is essential for assigned replication initiation. In particular, the further ability of binding single-stranded DNA plays a major role in the entire systematic initiation process [199]. Since there are no or only fragmented crystal structures of single-stranded DNA in complex with RepE, respectively TrfA, it is of interest to predict such complexes. Particularly, as Reps can be multifunctional and thus execute most diverse functions [68].

For RepE, the ssDNA sequence was elongated by nine bases and subsequently simulated for 4 μs. The most stable binding mode is depicted in Fig. 4.1. A major stabilizing factor of the complex is the promoted accommodation of the second base, an adenine, in a local cavity. A similar binding domain was found for a short ssDNA segment in complex with TrfA,

elaborated in the subsequent paragraph. The binding site differs from the double-stranded one and is predominantly stabilized by two factors: First, the adenine fit in the local minor cavity, and second, the attractive base interactions with arginine R230 (Fig. 4.2). On the other hand, a binding configuration of TrfA only with the single-stranded DNA could not be observed.

5 Conformational changes of Pot1pC protein induced by telomere accommodation

5.1 Introduction

Many biological processes require specific binding of proteins to single-stranded DNA, like telomere-end protection proteins (TEP), DNA repairing proteins or DNA polymerases and telomerases [151]. Pot1pC is a protein of the Pot1 TEP family responsible for protecting the single-stranded telomere in most eukaryotic cells.

Structural differences of the protein conformations in the protein-ligand interface explain the differing DNA requirements between species [49]. Binding is achieved by nonspecific stacking and hydrophobic interactions or interactions with the DNA phosphate group [167, 43, 135]. Between species with different nucleotide sequences orthologous TEP proteins use varying mechanisms to recognize and bind single-stranded DNA [113]. TEP has to bind the telomeric DNA with high affinity to insulate the DNA from damage response molecules and also to regulate access by telomerases. Absence of TEP can activate DNA-damage response mechanisms which may cause senescence and cell death [209, 47, 46, 165]. The immediate conformational and dynamical changes for pre-and post-incorporation of DNA are still unknown. The Pot1pC protein consists of a pocket with two loops, in which the DNA is accommodated, and two further outlying connecting loops. These loops build the dominant dynamics of the protein.

In a first step, the higher flexibility of the unoccupied protein is elaborated. The pocket is capable of opening and closing and the outlying loops form various folding structures. The integration of DNA reduces the overall flexibility enormously and inhibits opening and closing of the responsible pocket loop. This can be also quantified with entropy calculations while two approaches are discussed: The dihedral entropy, which is proportional to the expected value of the logarithmic backbone-dihedral distributions, and the conformational entropy obtained from the eigenvalues of the covariance matrix in the Principal Component Analysis (PCA). While the dihedral entropy gives valuable insight into the regional dynamics, on the downside it overestimates the real entropy by neglecting correlating higher order terms, particularly between neighboring residues [103]. The PCA approach also overestimates the entropy due to both, anharmonicities predominantly in the leading principal components and missing mutual higher order correlations [150]. The nearest-neighbor approach, which is based on the initial work of Kozachenko and Leonenko [102], provides an error estimation for the classical PCA-entropy and it is accurate and independent of the particular distribution. The method has been generalized for not only the nearest neighbor but for the k-nearest [182]. It is also widely used as a basis for mutual information estimators [104, 163, 156]. Besides, the eigenvalues from the PCA approach can be also used to define frequences of a harmonic oscillator and thus determine the conformational entropy quantum mechanically [161, 189]. This results in larger overestimations since Gaussian distributions yield the maximum entropy for a given variance [150, 38]. Another alternative approach is a clustering of the PCA distributions to overcome underlying anharmonicities [50, 52]. But also here a biasing effect is undeniable since clusters can mutually overlap, particularly in modes with

lower eigenvalues.

However, the intermediate stages of the protein consolidation are determined while DNA is progressively incorporated. Inter alia, it turns out that the middle part of the DNA plays a major role in two ways: It is the initiator for protein binding, and at the same time it strongly promotes binding due to an energetically favored cavity.

5.2 Methods

5.2.1 MD Simulation

For all simulations the Amber16 Molecular Dynamics Package [33] with the bsc1 force field [82] for nucleic acids and ff14SB [122] force field for proteins was applied. The starting structure (ProteinData Bank, PDB: 4HIK) was solvated in explicit water (TIP3P) [91] with rectangular boxes and an 8 Å minimum distance between complex and cubic box boundary. Additional sodium counter-ions neutralized the system. After energy minimization (10000 steps) the system was heated up to 300 K within three stages and with a restraint (force constant: $2.388 \, \text{kcal/(mol·Å}^2)$) to the starting positions. Subsequently, this restraint was first reduced to 0.5 and finally to 0 in a 500 ps simulation, respectively. After this equilibration step the MD simulations were carried out for 40 μs with the pmemd.cuda module at a temperature of 300 K and in NVT ensemble. The same procedure was followed for the apoprotein.

5.2.2 Principal Component Analysis

Principal Component Analysis (PCA) of molecular dynamics simulations is a method to account for the essential dynamics of the underlying system on a dimension-reduced energy landscape [183]. For this purpose, the overall translational and rotational motions are eliminated first with a trajectory alignment to the average geometrical center. The configurational space of the trajectory can be then translated into a cartesian 3 N × 3 N covariance matrix [66] where N is the number of all CA atoms (only backbone consideration). After diagonalization the set of eigenvectors (principal components) with corresponding eigenvalues is computed. These give a vectorial description of the motion with the eigenvalues as the variances [90]. Since the few principal components with the highest eigenvalues comprise the dominant and often anharmonic motion, the PCA also serves as a dimensionality reduction technique.

5.2.3 Conformational entropy calculations

Conformational entropy determination can be addressed in several ways. One method computes the dihedral entropy with the python package xentropy [103] by solving the integral

$$S = -R \cdot \int p(x) \ln(p(x)) \, \mathrm{d}x, \tag{5.1}$$

where p is the probability distribution of the dihedral observable x, and R is the gas constant. In our case, the two backbone dihedrals ϕ and ψ are of focus. The difficulty in this integral is to find this dihedral probability distribution. Thus, a kernel density estimation (KDE) with Gaussian functions as the kernel is applied. The KDE has the general form [103]

$$f(x, \lambda) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\sqrt{2\pi\lambda}} e^{-(x-X_i)^2/2\lambda}, \tag{5.2}$$

where n is the number or bins of the data points, X_i is the mean of the distribution and λ the bandwidth of the Gaussian kernel (variance of the distribution). Solving the above-mentioned integral with this estimation yields the 1D dihedral entropy. The sum of all local entropies approximates the total conformational entropy whereby correlations between the dihedrals are neglected.

However, the entropy contribution can be also addressed by a Principal Component Analysis (PCA). In general, the conformational entropy is given via [189, 92]

$$S_{gauss} = const + \frac{1}{2}k_B \ln(\det(\Sigma)), \tag{5.3}$$

where Σ is the covariance matrix based on the x, y, and z atom positions. The determinant of the covariance matrix is the product of all eigenvalues [22], i. e. of the variances in each dimension. The approach requires a multivariate normal distribution. This is often not the case, particularly not for the PCA modes with the largest eigenvalues which are the major contributors to the entropy. Each principal component displaying anharmonicity will overestimate the entropy [38]. Furthermore, for small variances this classical approximation as a limit of the quantum mechanical entropy breaks down in the quantum region [150]. Thus, a subset of 30 leading PCA modes is chosen which account for almost 90% of the motion. A more severe overestimation of the entropy than the anharmonicity is caused by a missing higher order dependence beyond the linear correlation. This *supralinear* dependence can be quantified and considered by mutual information (MI) computations [150].
The total amount of corrections in the entropy which have to be subtracted is [150]

$$S_{corr} = \sum_{i=1}^{d} S_{anh,(i)} + \sum_{i=1}^{d} \sum_{j=i+1}^{d} MI_{(i,j)} \tag{5.4}$$

for d modes. The first term concerns the correction due to the anharmonicity. For the 1-dimensional entropy $S_{non-Gauss,(i)}$, which is estimated by a non-Gaussian numerical method in PCA mode i, this first term is obtained by [150]

$$S_{anh,(i)} = S_{class,Gauss,(i)} - S_{non-Gauss,(i)} \tag{5.5}$$

while $S_{class,Gauss,(i)}$ is the entropy from the classical PCA approach in the i-th PCA mode. The second term describes the supralinear dependence correction. For the 2-dimensional entropy $S_{non-Gauss,(i,j)}$, which is estimated by a non-Gaussian numerical method in PCA mode (i,j), the second term is obtained by [150]

$$MI_{(i,j)} = S_{non-Gauss,(i)} + S_{non-Gauss,(j)} - S_{non-Gauss,(i,j)} \tag{5.6}$$

The only missing quantity now is the non-Gaussian entropy which we will calculate by the k-nearest neighbor algorithm [182]. This non-parametric method does not assume any particular distribution at all [72]. It provides a non-parametric estimate of the probability density for any PCA distribution. The conformational density for each simulation frame is assessed by its distance to its k-nearest neighbor frame. The entropy estimation is then [72]

$$S_{knn,k} = k_B \frac{d}{n} \sum_{i=1}^{n} \ln(R_{i,k}) + c_{k,d} \tag{5.7}$$

where d is three times the particle number, n is the number of total frames, $R_{i,k} = \|z_i - z_{i,k}\|_2$ is the radius of a d-dimensional hyper-sphere with the i-th frame z_i and its k-nearest neighbor $z_{i,k}$, and $c_{k,d}$ is a constant depending on the parameter k and d. This entropy estimation is derived from the mean of the logarithmic probability density in each PC mode (more precisely with a minus sign) [72]. The probability density depends on the volume of the d-dimensional sphere determined by the radius $R_{i,k}$. For the knn-algorithm, we followed a python script according to [194]. The parameter k is set equal to 4 as this is recommended from other studies [150, 72].

All errors are determined by successively calculating the entropy for the trajectory at 70%, 80%, and 90% of the simulation time.

5.3 Results and Discussion

5.3.1 Conformational flexibility reduction

We studied the telomere protecting Schizosaccharomyces pombe Pot1pC protein bound to a nine bases long single-stranded DNA (5'-GGTTACGGT-3') (see Fig. 5.1). The DNA is accommodated in a larger pocket (P1 and P2 region) and is surrounded by the left and right loops (O1 and O2 region). The pocket and these loops characterize the main internal protein dynamics. One way to demonstrate this is by computing the principal components (PC) of the system. The first 10 PCs of the apoprotein trajectory, which account for 67% of the dynamics, describe the pocket as well as the loop regions with high preference. The trajectory coordinates from both, the protein and the DNA-protein complex, can be then projected onto the two leading principal components based on the apoprotein trajectory. The result is a two-dimensional distribution of the protein (blue) and the complex (green) (see Fig. 5.2 A). The first two PCs mainly represent the O1 and O2 loop motion. The three
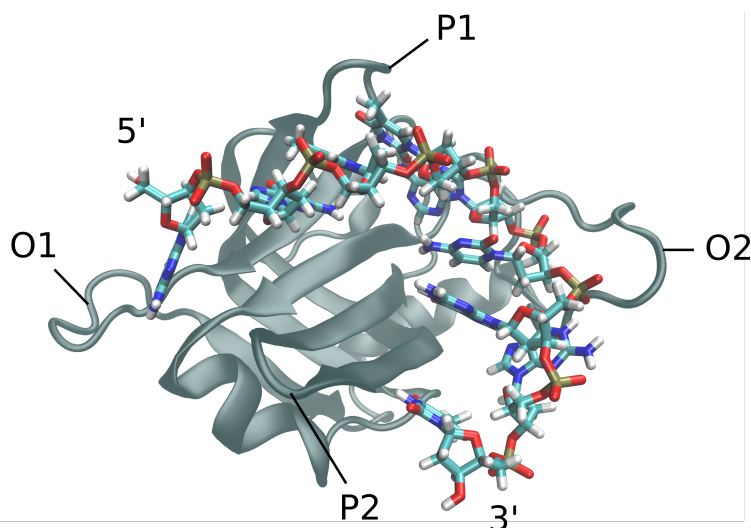
Figure 5.1: *Crystal structure of the Pot1pC protein (grey) accommodated by a nine bases long single-stranded DNA. The internal protein motion is predominantly determined by the P1, P2 pocket flaps and the O1, O2 connecting loops.*

snapshots are representative structures of the corresponding cluster which is computed by a kmeans analysis in 10 dimensions. The left apoprotein cluster is associated with the initial part of the trajectory in which only slight rearrangements occur after DNA dissociation (Fig. 5.2 B). These protein conformations resemble the holoprotein of the crystal structure the most. As it is apparent, only a marginal part of the apoprotein trajectory remains in this cluster. The apoprotein changes its conformations such that the O2 loop gets shifted and slightly opens to a rather ring-shaped formation (Fig. 5.2 A, B). This loop seems flexible and several folding processes occurred. The larger size may be the decisive cause. The difference between the right upper and lower clusters is mainly due to different folding constellations of the O1 loop. The pocket contribution, on the other hand, is underweighted in the two leading PCs. In higher dimensions, the configurational information of several regions is compressed and thus complicates the determination of the exact pocket contribution. For this reason, we elaborate the exclusive pocket analysis later in more detail. Direct comparison between apo- and holoprotein reveals essential insight: The complex distribution is not fully a subset of the protein distribution. Only a slight overlap can be observed which can be explained by a unitary and translational relocation of the holoprotein pocket and loops after DNA incorporation. Furthermore, the protein dynamics undergo much higher flexibility without DNA occupation since the associated distribution exceeds the complex one in size. The apoprotein dynamics are clearly anharmonic. While the complex remains in a conformation similar to the crystal structure for the entire simulation, the protein is highly flexible and allows DNA access. The restriction of the dynamics in the O2 region is caused by the telomeric backbone interacting with this O2 loop and the formation of a restraining H-bonding network. These effects are clearly weaker for the O1 region. The observation of the flexibility reduction is additionally strengthened by the larger eigenvalues (variances) of
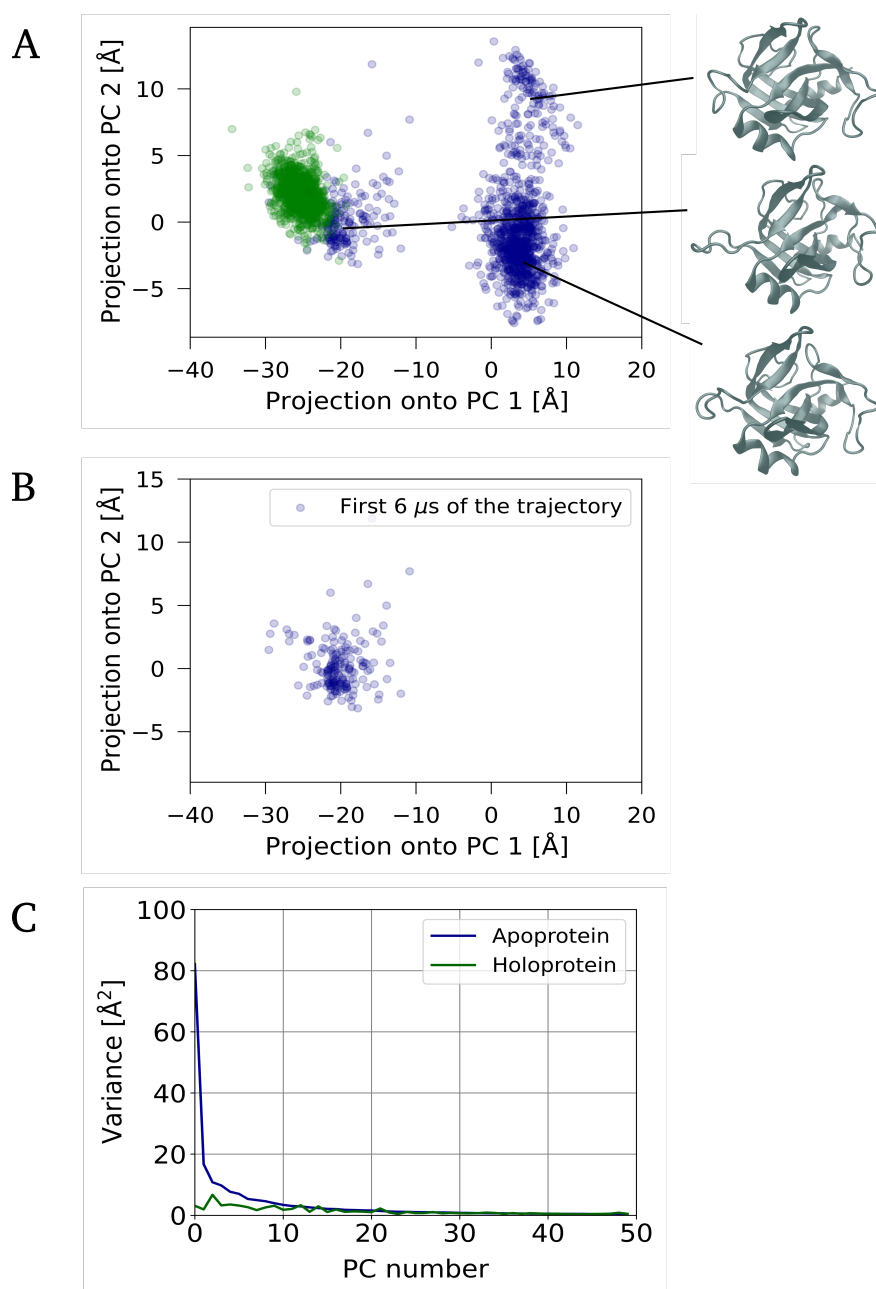
Figure 5.2: *(A) Projection of the apoprotein trajectory (blue) and the holoprotein trajectory (green) onto the two leading principal components based on the apoprotein trajectory. The three snapshots are representative structures of the respective cluster. (B) Apoprotein distribution referred to the first 6 µs of the trajectory. (C) Variances of the respective distributions vs. principal component number of the apo- and holoprotein.*
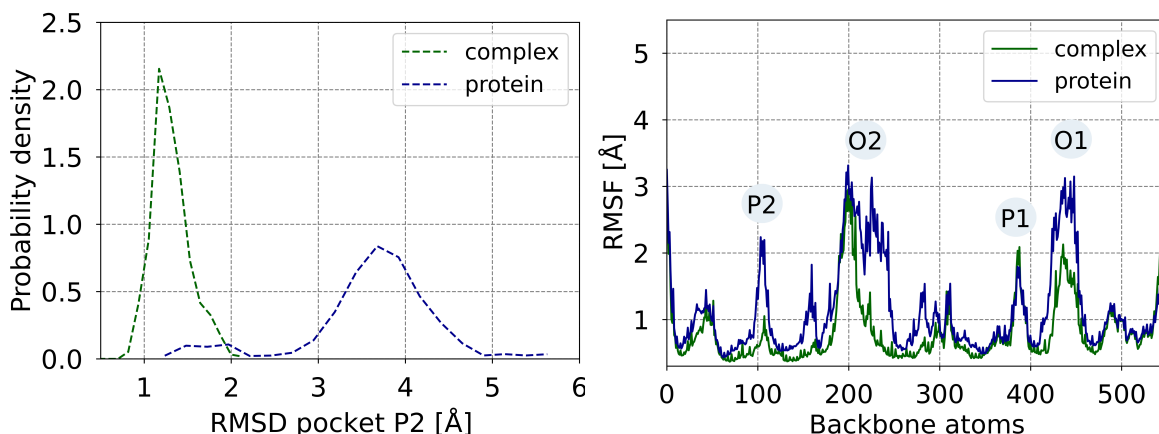
Figure 5.3: *Probability distribution of the P2 pocket region and RMSF of each backbone atom for the protein and complex, respectively.*

the leading PC distributions of the apoprotein (Fig. 5.2 C). While for the decisive principal components a large difference between protein and complex variance is detected, for higher modes, which merely describe Gaussian fluctuations, the two variance curves converge to zero.

The RMSD probability distribution of the pocket P2 region and the RMSF of all backbone atoms is computed after best superposition on the complex crystal structure (Fig. 5.3). The P2 loop exhibits a right shift in RMSD between holo- and apoprotein. Contrary to our expectations, the pocket is not contracted to establish a bound state with the DNA but gets even overstretched such that a spatial occupation is feasible. This is accompanied by a high energetic effort as this exceeds the regular dynamics of the unrestrained apoprotein. Only the lower part of the pocket is capable of opening and closing. Under stress due to the telomere induced overstretch the P2 loop loses substantial flexibility which is noticeable in the narrower distribution (Fig. 5.3). In addition, the RMSF (Fig. 5.3) condenses several aforementioned aspects. The RMSF results for the P2, O2, and O1 regions are within our expectations and clearly accentuate the opening and closing as well as the folding mechanisms. Also notable is the smaller O2-connecting loop at approximately atom 160 with a local peak and the alpha-helix at approximately atom 290 with a higher fluctuation in the apoprotein. The P1 pocket part behaves similarly in both simulations which is interesting. This means that the P1 loop, which is in closer contact with DNA than the lower P2 loop, keeps nevertheless unaffected by the DNA presence.

5.3.2 Conformational entropy contributions

The aforementioned conformational changes and the flexibility reduction can be quantitatively captured by relative conformational entropy calculations. The ϕ and ψ dihedral distributions in each amino acid backbone of the O1, O2, P1, P2 regions, respectively, form the basis for the dihedral entropy determination. The details of this method can be found in

Table 5.1: *Relative dihedral entropy part of the energy between apoprotein and holoprotein in the L1, L2, O1, O2 regions.*

Resp. region	P1	P2	O1	O2	total
Energy $[\frac{kcal}{mol}]$	-0.7 ± 0.2	3.3 ± 0.3	4.3 ± 1.2	0.7 ± 1.4	7.6 ± 3.1

Table 5.2: *Relative conformational entropy part of the energy between apoprotein and holoprotein for different estimation types.*

Estimation type	PCA	PCA-Anh	PCA-Anh-MI
Energy $[\frac{kcal}{mol}]$	5.9 ± 0.2	5.2 ± 0.3	3.6 ± 0.6

the methods section. The energy results in terms of the relative dihedral entropies are listed in Table 5.1. The P1 pocket loop surprisingly experiences a negative relative entropy part of the energy. Hence, this loop undergoes moderately higher fluctuation in the holoprotein compared to the apoprotein. For the P2 counterpart, the relative entropy part of the energy amounts to 3.3 kcal/mol. The strong widening effect of the DNA opens this loop beyond its regular apo dynamics and reduces its flexibility under tension. For the outstanding connecting loop O1 also a significant entropy contribution of the energy is found of approximately 4.3 kcal/mol. Folding processes and loop retracts which regularly occur in the apoprotein are rare after DNA incorporation. The O2 loop is also capable of folding and retracting but in a weaker manner. The total relative dihedral entropy contribution of the energy accounts for 7.6 kcal/mol. The higher error occurs mainly from the O1 and O2 region where specific folding mechanisms can emerge after many microseconds for the first time. The dihedral entropy neglects contributions from certain angle, position or bond-length alternations, and also only the most important residues are considered. Moreover, addition of the local entropies requires independence of the mutual regions. As shown by Polyansky et al. [161] the underlying 1D entropy as an approximation of the real entropy ignores correlating cross terms between the residues, for instance. This counted error can be substantially reduced when considering relative entropies instead of absolute [161]. Still, the result for the dihedral entropy is likely overestimated. Also in the work of Polyansky et al. only good correlations with $R^2 > 0.9$ between the computed and the real relative entropy are reported.

However, the entropy contribution can be also derived from the PCA distribution by means of the corresponding variances. The conformational entropy is proportional to the determinant of the covariance matrix (see methods). Implementation of this entropy access results in an energy difference of 5.9 kcal/mol between apo- and holoprotein. The PCA approach is a rough measure for conformational entropies as it overestimates the entropies due to anharmonicities in the distributions and missing mutual correlations. To correct both mentioned errors in the PCA approach, the k-nearest neighbor approach (see methods) is applied. The error caused by the anharmonicity turns out to be 0.7 kcal/mol for which the leading principal component is predominantly responsible. When considering the mutual information to second order, we find again a lower upper boundary for the conformational entropy estimation of the energy of approximately 3.6 kcal/mol, a deviation of 2.3 kcal/mol to the

PCA approach without any corrections. Consideration of higher order correlations is more severe than the anharmonicity [150]. The error from the anharmonicity is noticeably smaller than the error from neglecting mutual dependencies. The ratio between the PCA and the corrections-included PCA approach is in good agreement with other studies [150, 172].

5.3.3 DNA incorporation and protein consolidation process

Prior to the DNA incorporation, the apoprotein exhibits high flexibility in several regions and can enter in an open-pocket state or form various folding structures in the outlying loops. In an illustrative scheme (see Fig. 5.4) the pre-accommodation state is depicted first (Fig. 5.4 A). This will change as far as the telomeric overhang approaches and partially binds the protein. In order to qualitatively investigate which part of the DNA bind initially and how it is performed, two unrestrained, 5 μs long simulations were conducted: Both simulations include a shortened DNA after truncation of either four bases to the 3'-end (right DNA side) or five bases to the 5'-end (left DNA side). The simulations also reveal instructive insight into conformational consequences of the presence of certain DNA bases. Truncation of the four right terminal bases and maintaining the five left bases results in high instability of the complex and disentanglement of all bases except the adenine (Fig. 5.5 A, A1). The latter base is accommodated in a local cavity which exhibits impressively strong binding and prevents the dissociation of the DNA strand. While it is known that a base substitution from adenine to thymine, for instance, immediately results in a lower binding energy [49], the adenine also seems to have a substantial absolute binding energy. Furthermore, an open state of the P2 loop is observed (Fig. 5.5 A1). This is likely induced by the adenine stack but it could also be an outcome of a too short simulation in which this loop had not enough time to develop its intrinsic dynamics. On the other hand, truncation of the five left bases is accompanied by an enormous instability of the complex. After a temporary cytosine shift into the cavity, a telomeric dissociation is the consequence (Fig. 5.5 B, B1). Since the
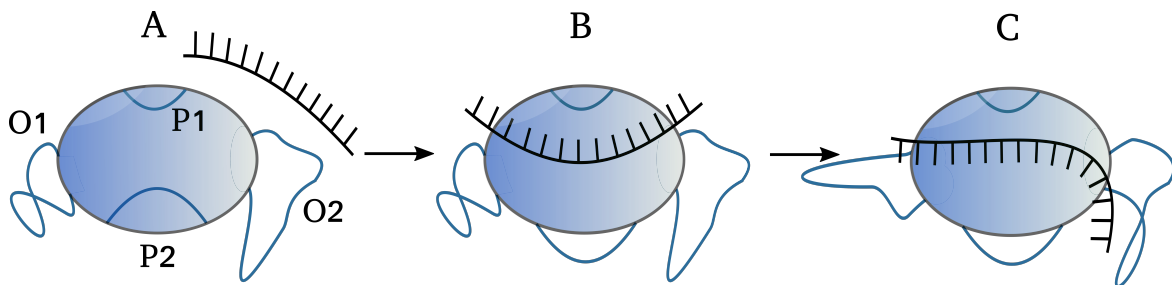
Figure 5.4: *A) DNA separation enables higher flexibility of the apoprotein. The O1 and O2 loops form diverse folded structures and the P2 pocket loop is predominantly closed. B) After integration of the inner DNA section the P2 loop is instantaneously widely opened over its natural apo dynamics. The O1 and O2 regions seem to be not affected. C) Final binding state in close agreement to the crystal structure with different O1 and O2 folded structures after complete DNA incorporation.*
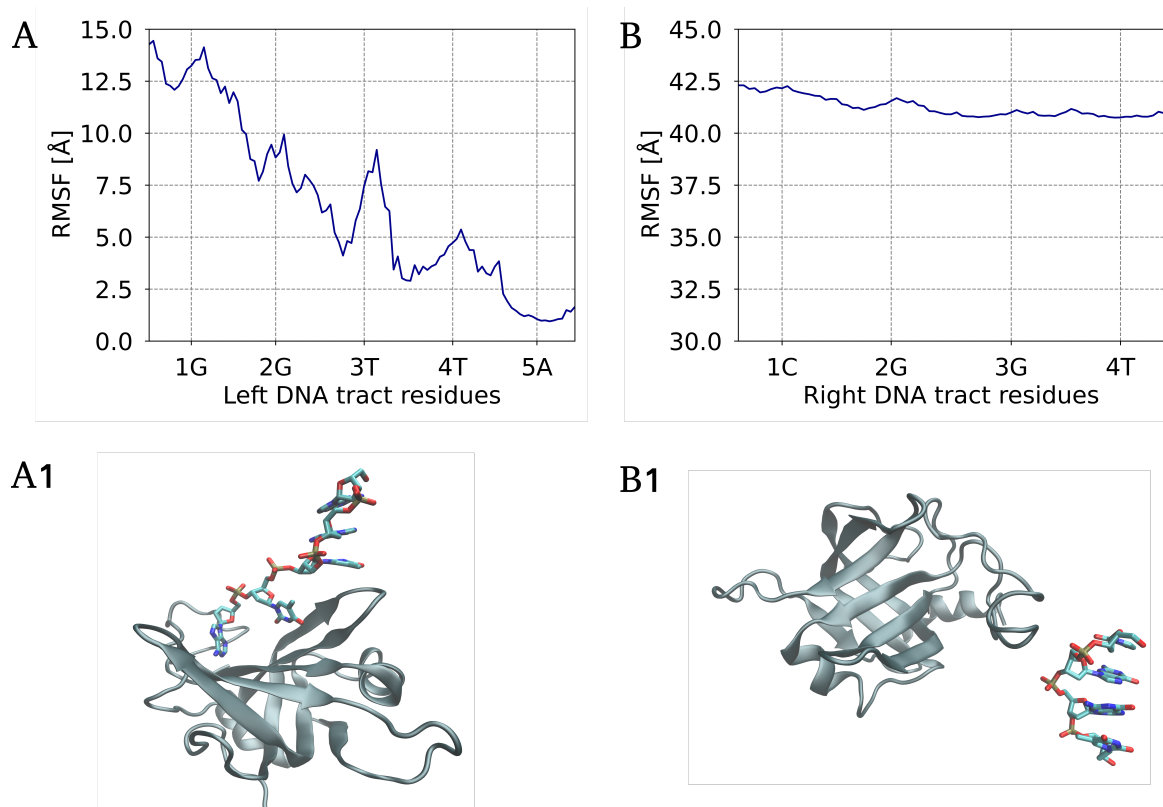
Figure 5.5: *A) RMSF of the remaining five DNA residues after truncation of the right DNA tract to the 3'-end. The adenine stays strongly bound in the favored local cavity. B) RMSF of the remaining four DNA residues after truncation of the left DNA tract to the 5'-end. A complete dissociation is evident. A1) Snapshot of the protein with the left DNA tract and B1) with the right DNA tract.*

truncation of both, the disjoint left and right DNA tract, always results in instability or even DNA separation, it seems that the complex stabilization is only guaranteed by a co-interaction of the left and right DNA tracts.

Altogether, these simulations emphasize a high likelihood for initial binding of the inner DNA bases to the protein (Fig. 5.4 B), followed by the left and the right terminal DNA tracts. The final binding state of the complex is in close conformity with the crystal structure.

5.4 Conclusion

In this chapter, we analyzed the conformational consequences of DNA accommodation upon the telomere-protecting Pot1pC protein. On a quantitative level, we outlined the dominant regions mainly responsible for the overall dynamics and elaborated differences between the apo and holo form. While the holoprotein represents a rather rigid conformation, the apoprotein undergoes high flexibility and forms various folding structures in its outlying O1 and

O2 loops. In further exploration, the probability distribution and also the overall RMSF of the P2 pocket loop uncover the high flexibility and the capability of opening and closing in the apo- but not in the holo-state. Only the lower part of the pocket is capable of opening and closing. After DNA association this loop is stretched over its natural apo dynamics which is accomplished with a high energetic cost. In contrast, the P1 pocket loop is mainly unaffected by DNA accommodation.

These conformational changes can be quantitatively captured by relative conformational entropy calculations which are performed in two ways: First, the dihedral probability distributions of the respective protein residuals determine the relative dihedral entropy. The corresponding energy accounts for 7.6 kcal/mol in total. The major contribution results from the P2 pocket restriction and the O1 flexibility reduction. The other method is a PCA approach which determines the conformational entropy in terms of the eigenvalues (variances). The weakness in this approach is a requirement for a multivariate normal distribution and a disregard of higher order correlations. Thus, it needs to be furthermore refined by correcting these errors which finally yields 3.6 kcal/mol. The PCA approach without corrections generally overestimates the real conformational entropy significantly. This applies even more for the 1-dimensional dihedral entropy which neglects correlating higher order terms between the residues. Nevertheless, the latter gives valuable insight into the regional entropy and exhibits the correct relative behavior between the relevant regions.

Finally, the DNA incorporation and protein consolidation are discussed. The understanding of these precise binding mechanisms between DNA and proteins is crucial and part of current research. Unrestrained simulations of shortened DNA reveal a strong absolute nucleotide binding in a local cavity and indicate an initial integration of the inner DNA tract. This is accompanied by an energy-expensive opening process of the pocket. Moreover, it seems that only a co-interaction of the left and right DNA tracts guarantees the complex stabilization. In the last phase with a fully incorporated DNA, the protein loops reorganize to a conformation which rarely occurs in the apo form.

6 Ab initio quantum mechanical phenomena in guanine-rich single-stranded DNA

6.1 Introduction

DNA analysis on a quantum mechanical level allows us to give answers to various types of questions addressing natural processes on a sub-molecular basis. One research field of high relevance concerns the accurate parametrization of the DNA force field. So far, the bsc1 force field is generally used but it exhibits, for instance, overstabilized base-to-base stacking [58, 105, 121, 10, 26] and B-DNA stiffness [217]. One main issue of this force field is the inaccurate partial charge distribution which could be accurately determined with ab initio quantum mechanical techniques.

Furthermore, since DNA is diamagnetic, external magnetic fields induce a magnetic field in DNA. For strong external magnetic fields as in MRI body scanners, NMR spectroscopy, or crystal studies all nuclei with a non-vanishing magnetic moment and angular momentum are perturbed by a weak oscillating field [210, 95, 76]. The surrounding electrons, which are charged and have non-vanishing spin, generate a magnetic field which opposes the applied magnetic field and thus shield the nuclei. Further experiments, where the magnetic behavior is also highly relevant, treat assembled 2D DNA layer which is exposed to an emitted photoelectron flux [143, 93, 65, 215]. A spin-selective transmission through the layer and thus a spin-filtering effect is reported. Interestingly, this spin-polarization effect is only observed for double-stranded DNA and not for single-stranded or mutated DNA [143, 65].

Magnetic properties of DNA thus play a major role in such sub-molecular processes. The electric properties of DNA are also substantial, particularly, for electron transport in photolyase repair mechanisms [212, 214], electron ejections by UV radiation [128, 137] or when applying an electrical field to DNA [54, 162, 136].

In this study, we focus on guanine-rich DNA since it was underlying in some of the above-mentioned experiments. First, an accurate charge distribution is calculated for a single guanine with backbone, a base pair of a guanosine and a cytosine nucleotide and a four guanine single-stranded DNA in native and in deformed conformation. We elaborate the geometry dependence and the neighboring base influence. Our results indicate a strong reliability on an accurately determined point charge model for the different DNA constellations. Subsequently, detailed insight into the nuclear magnetic shieldings of the heavy atoms is given. The phosphate group is highly responsible for shielding. But also other nuclei play a substantial role whereas large differences between several atoms can be observed. The electromagnetic consequences of additional unpaired electrons in G-DNA are treated in the last paragraph. The dipole moment, the increased isotropic polarizability, and the higher g-values in such excited DNA are quantified and also compared with the double-stranded correspondent.

6.2 Methods

6.2.1 Partial charge computations

All DNA structures were generated with the nab module of the Amber16 package [33]. Derivation of the partial charges was performed with the Gaussian09 package [114, 57]. This includes geometry optimization, frequency calculation and Merz-Kollmann population analysis on a B3LYP/6-31G* level of theory [15]. Partial charges were assigned with the RESP module of Amber16. The partial charges of one nucleotide sum up to -1.

6.2.2 Nuclear magnetic resonance (NMR) computations

All NMR calculations were carried out with the ORCA 4.2.1 package [147]. All reference molecule structures are provided from the PubChem databank.

After geometry optimization, the calculations were started at M06L/def2-TZVP [213, 202, 201] level of theory with dispersion correction D3zero [62] and approximation RIJCOSX [171]. The auxiliary basis set def2/JK is required as the algorithm uses Gauge-Independent Atomic Orbitals (GIAOs) [51, 205]. The calculation was run with the implicit water model CPCM [11].

The shielding tensor σ is defined as the second derivative of the total energy with respect to the magnetic moment and the magnetic field [192]:

$$\sigma_{ij} = \frac{\partial^2 E}{\partial B_i \partial \mu_j}\bigg|_{\mu=B=0} \tag{6.1}$$

It is written as a second rank tensor which can be split into a diamagnetic and paramagnetic component [74]. Although the multiplicity of the system is one (no unpaired electrons) the shielding includes paramagnetic effects which are caused by electron excitations due to the external magnetic field.

For practical reasons, the focus is on the chemical shift [74]

$$\delta = \frac{\sigma_{ref} - \sigma}{1 - \sigma_{ref}} \approx \sigma_{ref} - \sigma \tag{6.2}$$

which is a quantity independent of the external magnetic field strength. The principle behind the chemical shift is the magnetic susceptibility of the surrounding electrons covering the nucleus. This results in a partially shielding of the nucleus from the external magnetic field.

6.2.3 Dipole moment, polarizabilty and g-tensor computations

The dipole moment, polarizabilty and g-tensor calculations were carried out with the ORCA 4.2.1 package [147] at a M06L/def2-TZVP [213, 202, 201] level of theory with dispersion correction D3zero [62], approximation RIJCOSX [171] and in implicit water (model CPCM

[11]). The dipole moment operator is defined according to [59]

$$\vec{m} = \sum_N Z_N \vec{R}_N - \sum_i \vec{r}_i, \tag{6.3}$$

where N denotes summation of nuclei with positions $\vec{R}_N$ and charges Z_N, and i denotes summation of electrons at positions $\vec{r}_i$.

The polarizability is calculated analytically with the coupled-perturbed SCF method as this is the most efficient and accurate way [115].

The g-tensor is determined from the second derivative of the energy with respect to the external magnetic field and the effective electron spin, which generally depends on the free electron g-value of approximately 2.002319, the total spin, the spin density matrix, a spatial one-electron spin-orbit term, and the magnetic field [178]. Thus, the g-tensor is gauge-dependent. The gauge origin is the center of electronic charge. For practical purposes, particularly when using large basis sets, the deviations from different gauge origins can be kept negligible small.

6.3 Results and Discussion

6.3.1 Geometry dependence of charge distributions

At first, we focus on guanine-rich DNA in different constellations: A guanine nucleotide, a Watson-Crick coupled base pair of a guanosine and a cytosine nucleotide, a single-stranded G-DNA with four bases in native and in deformed conformation (see also Fig. 6.1). At least one sugar group is considered in any case. The calculated charge distributions are listed in Table 6.1. Comparison of the single guanosine with the coupled guanosine shows that the
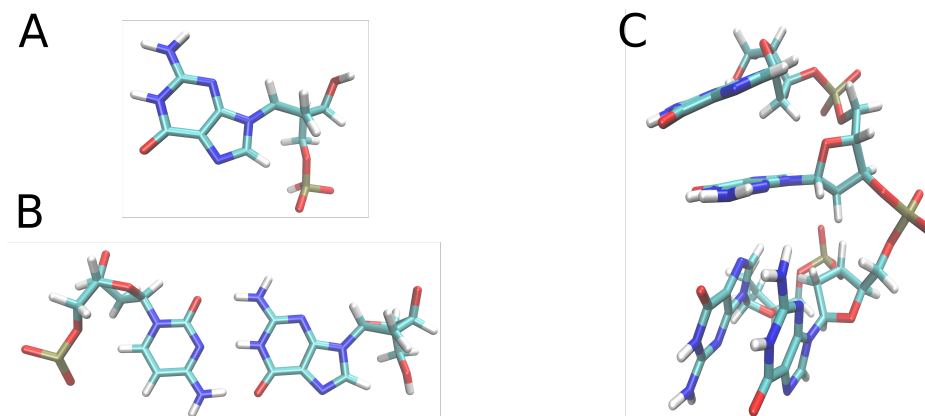

Figure 6.1: *Different G-DNA constellations. (A) Guanine nucleotide with charge -1. (B) Base pair of guanine and cytosine and parts of the backbone. The total charge is equal to -1. (C) Four bases G-DNA with charge -3 in a strongly deformed state.*

Table 6.1: *B3LYP atomic charges of different guanine-DNA constellations. Guanine is abbreviated as G and guanosine as G^+.*

Atom type	G^+	G^+ coupled	G (Sponer)	ssDNA	ssDNA deformed	bsc1
P				1.296	1.275	1.166
OP1				-0.799	-0.792	-0.776
OP2				-0.785	-0.785	-0.776
O5'				-0.480	-0.541	-0.495
C5'	0.140	0.034		-0.165	0.027	-0.007
H5'	0.033	0.021		0.082	0.044	0.075
H5"	0.054	0.006		0.118	0.114	0.075
C4'	-0.080	0.476		0.382	0.359	0.163
H4'	0.111	-0.057		0.018	0.074	0.118
O4'	-0.393	-0.593		-0.541	-0.524	-0.369
C1'	-0.244	0.493		0.185	0.096	0.036
H1'	0.084	0.061		0.207	0.154	0.175
N9	-0.238	-0.201	-0.602	0.015	0.086	0.058
C8	0.308	0.241	0.302	0.216	0.101	0.074
H8	0.109	0.171	0.105	0.156	0.145	0.200
N7	-0.587	-0.649	-0.593	-0.594	-0.534	-0.573
C5	-0.037	0.074	0.194	0.040	-0.095	0.199
C6	0.759	0.659	0.524	0.746	0.733	0.492
O6	-0.621	-0.661	-0.500	-0.634	-0.690	-0.570
N1	-0.871	-0.792	-0.710	-0.861	-0.764	-0.505
H1	0.435	0.452	0.393	0.416	0.417	0.352
C2	0.962	0.932	0.856	1.038	0.960	0.743
N2	-0.973	-1.020	-1.053	-1.143	-0.991	-0.923
H21	0.385	0.470	0.458	0.443	0.398	0.424
H22	0.404	0.417	0.444	0.504	0.427	0.424
N3	-0.758	-0.728	-0.713	-0.627	-0.665	-0.664
C4	0.434	0.325	0.480	0.140	0.351	0.181
C3'	0.582	0.360		0.175	0.208	0.071
H3'	-0.086	-0.061		0.133	0.109	0.099
C2'	-0.033	-0.265		-0.208	-0.352	-0.085
H2'	0.019	0.043		0.102	0.114	0.072
H2"	-0.001	0.058		0.112	0.178	0.072
O3'	-0.777	-0.947		-0.571	-0.544	-0.523

Table 6.2: *Deviations between the B3LYP atomic charges of the respective guanine-DNA constellations and the bsc1 charges. Guanine is abbreviated as G and guanosine as G^+.*

Atom type	ΔG^+	ΔG^+ coupled	ΔG (Sponer)	ΔssDNA	ΔssDNA deformed
P				0.130	0.109
OP1				0.023	0.016
OP2				0.009	0.009
O5'				0.015	0.046
C5'	0.147	0.041		0.158	0.034
H5'	0.042	0.054		0.007	0.031
H5"	0.021	0.069		0.043	0.039
C4'	0.243	0.313		0.219	0.196
H4'	0.007	0.175		0.100	0.044
O4'	0.024	0.224		0.172	0.155
C1'	0.280	0.457		0.149	0.060
H1'	0.091	0.114		0.032	0.021
N9	0.296	0.259	0.660	0.043	0.028
C8	0.234	0.167	0.228	0.142	0.027
H8	0.091	0.029	0.095	0.044	0.055
N7	0.014	0.076	0.020	0.021	0.039
C5	0.236	0.125	0.005	0.159	0.294
C6	0.267	0.167	0.032	0.254	0.241
O6	0.051	0.091	0.070	0.064	0.120
N1	0.366	0.287	0.205	0.356	0.259
H1	0.083	0.100	0.041	0.064	0.065
C2	0.219	0.189	0.113	0.295	0.217
N2	0.050	0.097	0.130	0.220	0.068
H21	0.039	0.046	0.034	0.019	0.026
H22	0.020	0.007	0.020	0.080	0.003
N3	0.094	0.064	0.049	0.037	0.001
C4	0.253	0.144	0.299	0.041	0.170
C3'	0.511	0.289		0.104	0.137
H3'	0.185	0.160		0.034	0.010
C2'	0.052	0.180		0.123	0.267
H2'	0.053	0.029		0.030	0.042
H2"	0.073	0.014		0.040	0.106
O3'	0.254	0.424		0.048	0.021

charge distributions for both guanines (without sugar group) are strongly congruent with a deviation lower than 0.1, except for C4 only. For the sugar group in the backbone, on the other hand, it looks more different. Here the backbone charges exhibit more volatility and are more susceptible to geometry alternations or coupling. For some atoms the partial charges are nevertheless in good agreement. For the less sensible guanine in the nucleoside we also found a good agreement with the charge distribution calculated by Sponer and coworker [219]. The charge of atom C6 is slightly underestimated by the group of Sponer using older MEP-based methods. The N9 atom, which is glycosidically bound to the backbone, deviates significantly in charge but this can be addressed to the missing backbone for the Sponer charges.

Note that for adenine-tracts (and probably other bases) in DNA, underestimation of base-pairing interactions in the current force field has been reported [87, 10]. With more accurately determined partial charges beyond the RESP model this underestimation was successfully eliminated [87].

From the single-stranded 4-guanine DNA and its deformed equivalent (see Fig. 6.1) the average residue structure is considered as the reference system. Comparison of both ssDNA structures shows generally small differences (Table 6.1). The two backbone atoms C5' and C2' and the three guanine atoms C8, N2, and C4 deviate significantly. At least for some of these atoms this probably may be a result of computational inaccuracy. It seems that deformation in DNA keeps the charge distribution mainly unaffected.

This structural deformation, however, also interrupts base-to-base stacking within one strand and thus reduces the polarization between two neighboring bases. Since the charge deviation between deformed and native DNA on average was under 0.1, the altered polarization effect seems to be also small.

Furthermore, in comparison to the bsc1 charges (Table 6.2), the majority of partial charges of the native and deformed single strands are in good agreement. On the other hand, also certain deviations are undeniable. The nitrogen N1 charge shows the largest deviation. Also noticeably many carbons like C6, C5, C4' and C2 deviate above average. Although the deviations seem small, it is known that even small changes in the partial charges may significantly affect the overall conformational dynamics [87].

Altogether, the bsc1 charge model seems satisfying but it is still behind being perfect.

While incorporation of polarization effects or consideration of more accurately computed partial charges may represent a refinement to current DNA force fields, our results indicate a strong reliability on the partial charges between different DNA constellations.

6.3.2 How nuclear magnetic shielding in G-DNA is sustained

We investigate nuclear magnetic resonance (NMR) behavior for single-stranded guanine-rich DNA as an application of the NMR module in ORCA (see methods). Presence of an external magnetic field results in a shielding of the nucleus. The absolute shielding of each nucleus is composed of an isotropic and an anisotropic contribution. The outcome is depicted in Fig. 6.2. These are the nucleus values for the second residue of a four bases long ssDNA which deviates minimally from the average nucleus of the entire single strand. This second
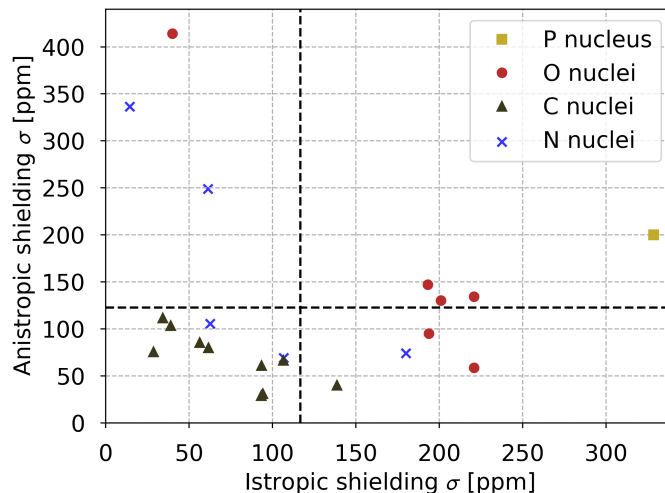
Figure 6.2: *Absolute isotropic vs. absolute anisotropic shielding of each nucleus type of a guanine nucleotide. Different colors/markers characterize different nucleus types. Black vertical, resp. horizontal, line represents the mean value of all isotropic, resp. anisotropic, shielded nuclei.*

residue also has the advantage over terminal residues that terminal biasing effects are excluded. The mean of all nuclei in terms of isotropic/anisotropic shielding is visualized by a vertical/horizontal line. The phosphate at 328.4 ppm shows by far the strongest reaction in terms of isotropic shielding. This is followed by all five backbone oxygen nuclei where the intra-guanine oxygen exhibits low isotropic contribution. Thus, the backbone, particularly the phosphate group, is highly responsible for isotropic shielding. This is also expected as here higher electron densities are underlying. For anisotropic effects the phosphate group is also of importance. Here only the aforementioned oxygen O6 and two further nitrogen nuclei in the guanine show more impact. This intra-guanine oxygen is characterized by a double bond with a carbon resulting in a low isotropic effect. In contrast, the oxygen in the sugar group forming a chemical compound with two carbons experiences multiple higher shielding. Its diamagnetic contribution of -374.8 ppm is noticeably smaller than the all-oxygen average of (-233.9 ± 64.0) ppm. Its paramagnetic counterpart overlaps with the average of (412.1 ± 13.0) ppm. For the phosphate we obtained the highest (in absolute values) paramagnetic and diamagnetic values of (958.5 ± 2.49) ppm, resp. (-629.5 ± 7.16) ppm. The nitrogen shielding, which is discussed later in more detail, is composed of the paramagnetic component of (343.7 ± 14.6) ppm and the diamagnetic of (-258.7 ± 66.7) ppm. Carbons, in contrast, show less relevance in any shielding.

However, comparative experimental results together with the calculated absolute shielding and the calculated relative chemical shift are shown in Table 6.3. The corresponding reference molecules are selected such that accurate computations are possible. These are also often considered as the standard (for ^{15}N also CH_3NO_2 is used). The target molecules are a rough approximation of the real chemical compound of the corresponding nucleus in DNA.

Table 6.3: *Averaged isotropic shielding of respective nuclei in single-stranded DNA compared with experimental data. All σ are referred to the absolute shielding and all δ to the chemical shift with respect to the reference molecule. All values are given in ppm.*

Nuclei	Reference	Target	Exp. σ_{ref}	Exp. σ_{tar}	Calc. σ_{ref}	Calc. σ	δ_{exp}	δ_{calc}
^{31}P	PH$_3$	H$_3$PO$_4$	594.5 [85]	328.4 [41]	621.0	328.4	266.1	292.6
^{17}O	H$_2$O	H$_3$PO$_4$	344.0 [2]	255.0 [5]	346.5	178.2	89.0	168.3
^{13}C	C$_4$H$_{12}$Si	HCN	188.1 [84]	82.1 [84]	191.4	74.6	106.0	116.8
^{15}N	NH$_3$	-	264.5 [86]	-	283.7	85.0	-	198.7

For ^{15}N no appropriate target molecule was found but an experimental relation is referenced later. The calculated absolute shielding of the reference structure is in good qualitative agreement with the experiments for any nucleus. The calculated shielding for the phosphate is in exact agreement with the experimental absolute shielding. Deviations in the chemical shift arise from inaccuracies in the reference NMR calculations. The outlier O6 oxygen biases the calculated chemical shift by 27.7 ppm as it is likely not accurately reflected by the oxygen of the target molecule. More sources of error can arise from computational and experimental inaccuracies, indeed, but also from different solvents in the experiments like D$_2$O. Qualitative good agreement is also found for the carbons. Furthermore, DNA or DNA related NMR studies have been also published: Our NMR results for nitrogen absolutely agree with the experimental data [127, 204] although ^{15}N chemical shifts are known to be very sensitive to structural changes [127]. Our computations reveal a nitrogen-dominating isotropic shielding for N2 (179.9 ppm) in the guanines amino group and an almost negligible N7 (14.4 ppm) shielding. Interestingly, the two nitrogens N7 and N3 have the same chemical compound but differ in isotropic shielding by 46.8 ppm and in anisotropic even by 87.6 ppm. We confirm that the underlying structures play a major role and showed the scope it can reach. The calculated result of the averaged oxygens is moreover strengthened by DNA studies according to Wu and coworkers [208].

Comparative analysis with double-stranded DNA reveals an analogous behavior: The averaged phosphate deviates only by 6% in isotropic shielding compared to the ssDNA case. Furthermore, the absolute shielding of the oxygens is not underestimated as in the single strand. It amounts to 220.7 ppm which much better agrees with the experimental observations (see Table 6.3), particularly when suppressing the outlier oxygen which also occurs here. Nitrogen nuclei exhibit almost no deviation from single-stranded DNA in shielding. Nevertheless, also for double-stranded DNA each nitrogen undergoes high variation due to their sensitivity over structural alternations. The carbons are again the least relevant atoms in this context.

Altogether, we observed that double-stranded G-DNA remains rather unchanged for nuclear magnetic shieldings over guanine-rich single strands. This is particularly true for the atoms close to the Watson-Crick coupled complementary base.
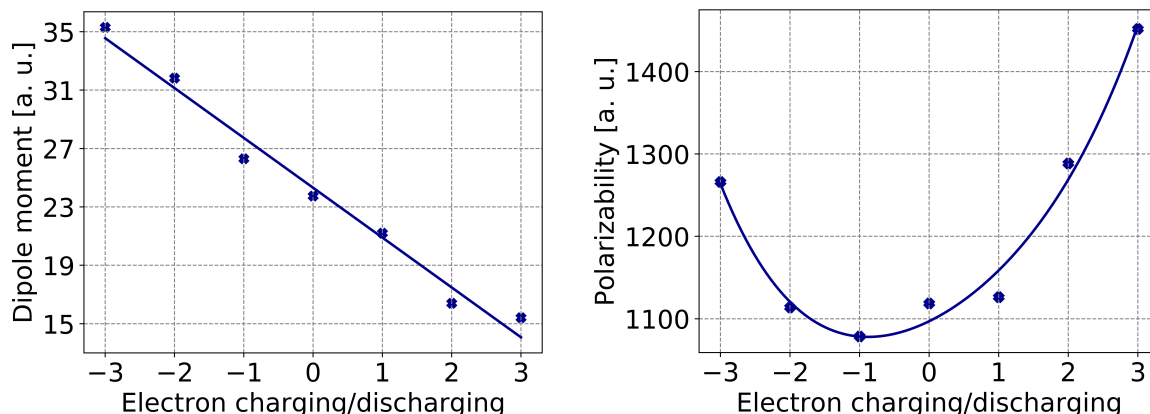
Figure 6.3: *Dipole moment and isotropic polarizability of single-stranded DNA with additional unpaired electrons*

6.3.3 Electric and magnetic consequences of additional unpaired electrons

Additional unpaired electrons have a direct impact on the electric and magnetic properties of DNA. This can occur after a base-to-base electron transfer or after absorption from a photoelectron flux, for instance. In this excited state DNA properties change noticeably. Here a single-stranded G-DNA in an excited state is under study. With up to three additional unpaired electrons we first analyzed the affected electric properties. The electric dipole moment decreases linearly in charging as it is visible in Fig. 6.3. The emission of three additional electrons causes a more than 50% higher dipole moment. The same holds for the native DNA compared to DNA with three absorbed electrons. The linear relation can be also shown for double-stranded DNA.

Moreover, the polarizability is increased for several emitted or absorbed electrons (Fig. 6.3). For DNA, including double-stranded DNA, the induced dipole moments due to an electric field increase strongly for higher charging, respective higher discharging of DNA. A closer look reveals why emitting electrons also cause an increase in polarizability: For any further missing electron another electron is considered here as an unpaired electron leading to a higher multiplicity. In this context emitting or absorbing is related and thus exerts similar polarizability effects. The addition of three unpaired electrons results in substantially, more than 30% higher polarizability.

A solvent dependence of the polarizability is also detected. The DNA calculations were carried out in implicit water. But in vacuo, the polarizability plunges to 767.4 a.u. from 1118.5 a.u., a decline of 31.4%. Consequently, this phenomenon shows that water strengthens DNA polarizability by this amount due to more induced dipole moments in DNA.

However, a magnetic moment in an external magnetic field has a potential energy. Here the magnetic moment depends on the spin but also on a specific constant, the g-factor. This is a case-dependent dimensionless quantity which particularly characterises the magnetic moment. Hence, a lot of information is restored in this quantity. For each excited ssDNA state

the isotropic g-values are listed in Table 6.4. All of these values are higher than the single electron g-value of 2.002319. The deviations mainly originate from additional paramagnetic spin-orbit effects in DNA due to the unpaired electrons. It turns out that the paramagnetic spin-orbit contributions dominate the respective diamagnetic effects.

The higher g-values have an immediate impact on the magnetic moment and consequently, on the interaction of the excited DNA with external magnetic fields.

Table 6.4: *Isotropic g-values of charged/discharged single-stranded DNA.*

Charging state	-3	-2	-1	1	2	3
g-value	2.00446	2.00442	2.00430	2.00362	2.0038	2.00373

6.4 Conclusion

We studied fundamental electric and magnetic properties of single-stranded guanine-rich DNA with quantum mechanical ab initio methods.

First, we raised the question of the geometry dependence of the DNA charge distribution. It turned out that the deformation of DNA has only a slight impact on the charge distribution, even for the more geometry-dependent backbone. Only small differences in the charge distribution were also observed between single-stranded and double-stranded DNA. Our results indicate a strong reliability on the point charge models for different DNA constellations.

We also calculated the nuclear magnetic shielding of the DNA atoms. The dominating isotropic shielding occurs from the phosphate groups with higher electron densities. This includes (in absolute values) strong paramagnetic and diamagnetic effects. Also mentionable are two points: First, the intra-guanine oxygen which is characterized by a double bond with a carbon shows a high anisotropic contribution. Second, nitrogens undergo high variation due to their sensitivity over structural alternations. Only negligible differences between single-stranded and double-stranded DNA were found.

In the last paragraph, we analyzed the impact of additional unpaired electrons in DNA. An expected linear relation of the dipole moments in the number of emitted/absorbed electrons and a parabolic relation of the polarizability was detected. The polarizability is strongly solvent-dependent. Moreover, the g-factor of DNA excited with up to three unpaired electrons turned out to be always greater than the reference g-value of a single electron. This deviation can be explained by dominating paramagnetic spin-orbit contributions due to the additional unpaired electrons.

7 DNA flexibility as a key factor in UV-induced CPD damages

7.1 Introduction

UV light has significant carcinogenic effects, as it can lead to local changes in the chemical structure of DNA. The most frequent UV light damages are cyclobutane pyrimidine dimers (CPD), which result from an electronic HOMO $\rightarrow$ LUMO transition [131, 29, 116]. In vivo, CPD damages are repaired via the nucleotide excision repair (NER) machinery. Mutations in NER and hence reduced or eliminated NER activity therefore cause life-threatening diseases, such as Xeroderma Pigmentosum [130, 98, 142]. If unrepaired, these DNA lesions block regular DNA replication and thus strongly contribute to mutagenesis in skin cancers [159, 160, 139, 45].

However, sufficient activity of repair proteins is just one component in the stability of our hereditary information. Previous studies have addressed and given answers to many essential questions about CPD damages. By means of NMR and X-ray diffraction measurements of DNA structures containing CPD damages have been resolved for both, unbound DNA and DNA in complex with a repair enzyme [111, 27, 133, 155]. These experiments showed that CPD damages lead to substantial structural changes, such as bending and widening of the minor groove, which is also confirmed by gel electrophoresis studies [80, 197]. Moreover, solved structures have been used for multiple MD studies, demonstrating that CPD damages cause not only substantial changes in structure but also in the deformability of the DNA [96, 97, 120]. The altered deformability in turn might serve as a selection criterion for NER proteins, a view supported by combined laser temperature-jump and spectroscopic experiments by Velmurugu et al. [193]. Another essential question is how CPD damages are associated with chromosome folding. Mao and coworkers used sequencing techniques to show that CPD damages are not distributed uniformly along the genome, but rather exhibit an oscillatory pattern originating from the rotational setting of nucleosomal DNA. Particularly, T-rich sequences are mostly found at nucleosomal sites with a relatively low CPD damage level [125, 126]. These T-rich sequences (predominantly TpT steps) are most vulnerable to CPD damage formation [110, 118].

In this study, we perform QM calculations and MD simulations to reveal the impact of DNA conformational flexibility on CPD damage formation. In the first step, we start from the smallest DNA component, the isolated TpT dinucleotide, where CPD damages are relevant. By carrying out a series of QM calculations, the geometric conditions to form such a damage are determined. In additional computations, the differences between single-stranded and double-stranded DNA are elaborated. It turns out that the presence of the complementary strand has only a marginal impact on the CPD damage formation. Contributing factors to the deviations are the rigidity of duplexes and orbital realignments of the thymidines. Thus, CPD damages are an exquisite example of the importance of single-stranded dinucleotides which play an essential role in the entire duplex DNA. However, we also perform MD simulations on isolated DNA. Projecting the conformational criteria determined from the QM calculations onto the MD trajectories allows us to compute flexibility-dependent probabil-

ities for a CPD damage formation. In this way, it turns out that single-stranded DNA is, for conformational reasons, slightly less vulnerable to CPD damage formation than double-stranded DNA. Nevertheless, double-stranded DNA shows a high degree of robustness. The conformational subspace required for CPD damage formation is reached only through thermal fluctuations, and we quantify the probability for occupying this subspace to only 0.5% for single-stranded and 1.0% to 2.5% for double-stranded DNA. Finally, the conformational subspace for DNA in the nucleosome complex is investigated. We show that stacking conformations due to wrapping around the histone core alone already explain the underlying oscillatory CPD damage pattern. TpT sequences located at rotational inward settings (i.e. minor groove faces the histone) have a clear conformational reluctance for CPD damage formation.

7.2 Methods

7.2.1 QM calculations

All QM calculations were carried out with the ORCA 4.2.1 package [147]. Five structures of two adenine-thymine base pairs were extracted from a nucleosome complex (Protein-Data Bank, PDB: 2CV5). In addition, a DNA structure of two adenine-thymine base pairs was generated with the nab module of the Amber16 package [33]. This structure served to generate 15 more structures by geometry-optimization with distance constraints between the C6 atoms (see Fig. 7.1) of the two thymines from 3.6 Å to 4.2 Å in equidistant 0.05 Å steps. All above-mentioned structures were geometry-optimized at B3LYP/def2-TZVP [14, 15, 202, 200] level of theory with dispersion correction D3zero [62] and approximation RIJCOSX [171] while for the 5 nucleosome structures all dihedral angles were constrained. The optimization was run with the implicit water model CPCM [11].

For the dihedral-dependence analysis, structures with constraint C6-C6 and C5-C5 distances are geometry-optimized while the $\angle(C5, C6, C6, C5)$ dihedral angle is increased from 20° to 45° in 5° steps. For one group, the C6-C6 distance is 3.79 Å and the C5-C5 distance is 3.45 Å, and for another, the C6-C6 distance is 3.91 Å and the C5-C5 distance is 3.57 Å. This provides 6 further structures per group with different dihedral angles but consistent distances.

For all structures time-dependent density functional theory (TDDFT) [177, 63, 36, 35] calculations for the first 10 excited states and at B3LYP/Zora-Def2-TZVP [14, 202, 153, 48] level of theory were performed. Dispersion correction D3zero [62], approximation RIJCOSX, RI-SOMF(1X) [171, 146] and implicit water (CPCM) [11] were considered.

7.2.2 MD simulation

The MD simulations were performed with the pmemd.cuda module of the Amber18 package [34] with the bsc1 force field [82] for nucleic acids and ff14SB [122] for proteins. The isolated DNA molecules were generated with the nab-module, neutralized with sodium ions, and exposed to a water-box (TIP3P model, [91]) with a minimum distance of 10 Å between box boundary and molecule. For the nucleosome systems, the human nucleosome core particle

(ProteinData Bank, PDB: 2CV5) served as the starting structure, and we used 200 mM NaCl salt concentrations and a minimal solvation distance of 15 Å. The starting structure for the T-tract in complex with the histone was built with a script, which arranges the desired mutation of the DNA sequence. Afterwards, the systems were energy-minimized in 5000 steps with the sander module. Subsequently, the systems were equilibrated, which included four stages of temperature increase, turning-on of the barostat, and a gradual reduction of the restraints in four stages (from 25 to 0 kcal/(mol · Å^2)). Overall, the equilibration procedure covers a simulation time of 3.5 ns. Finally, production runs were carried out, with simulation times of 500 ns for the isolated DNA systems and 50 ns for the nucleosome system.

7.3 Results and Discussion

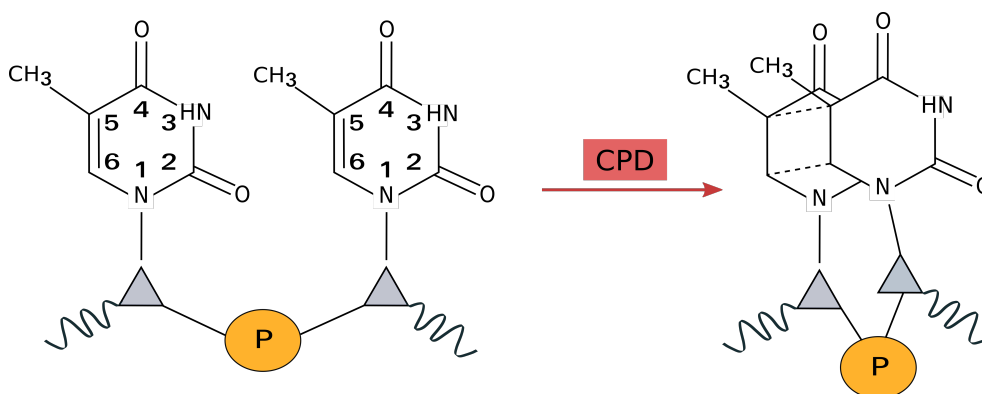### 7.3.1 Geometry dependence of CPD damages

Figure 7.1: *UV radiation-induced CPD damage of two thymines where the two C5 atoms and the two C6 atoms covalently bind in the first excitation state (LUMO) and thus form a cyclobutane compound.*

For cyclobutane pyrimidine dimer (CPD) damages it is known that under exposure to UV radiation the C6 atoms and simultaneously C5 atoms covalently bind in the first excitation state and thus form this damage ([155, 174], Fig. 7.1). The choice of thymine bases is due to the consequences of C5-methylation on the photo-induced reactivity of the thymine-rich DNA [131]. For a better understanding of this CPD effect, particularly with regard to the geometry dependence, several TDDFT calculations are conducted: First, 15 generated structures of two adenine-thymine base pairs are geometry-optimized (see methods). In the geometry optimization step, all atoms need to reorganize to accomplish the closer distance of 3.6 Å - 4.2 Å due to the constraints. In each case this is accompanied by a slight displacement of the phosphor group to the outside (see visualization in Fig. 7.1). Moreover, closer C6 atoms, particularly for a distance of around 3.6 Å, results in a rotation of the two single strands to each other. As we see later, this is also found for some nucleosomal DNA crystal structures.
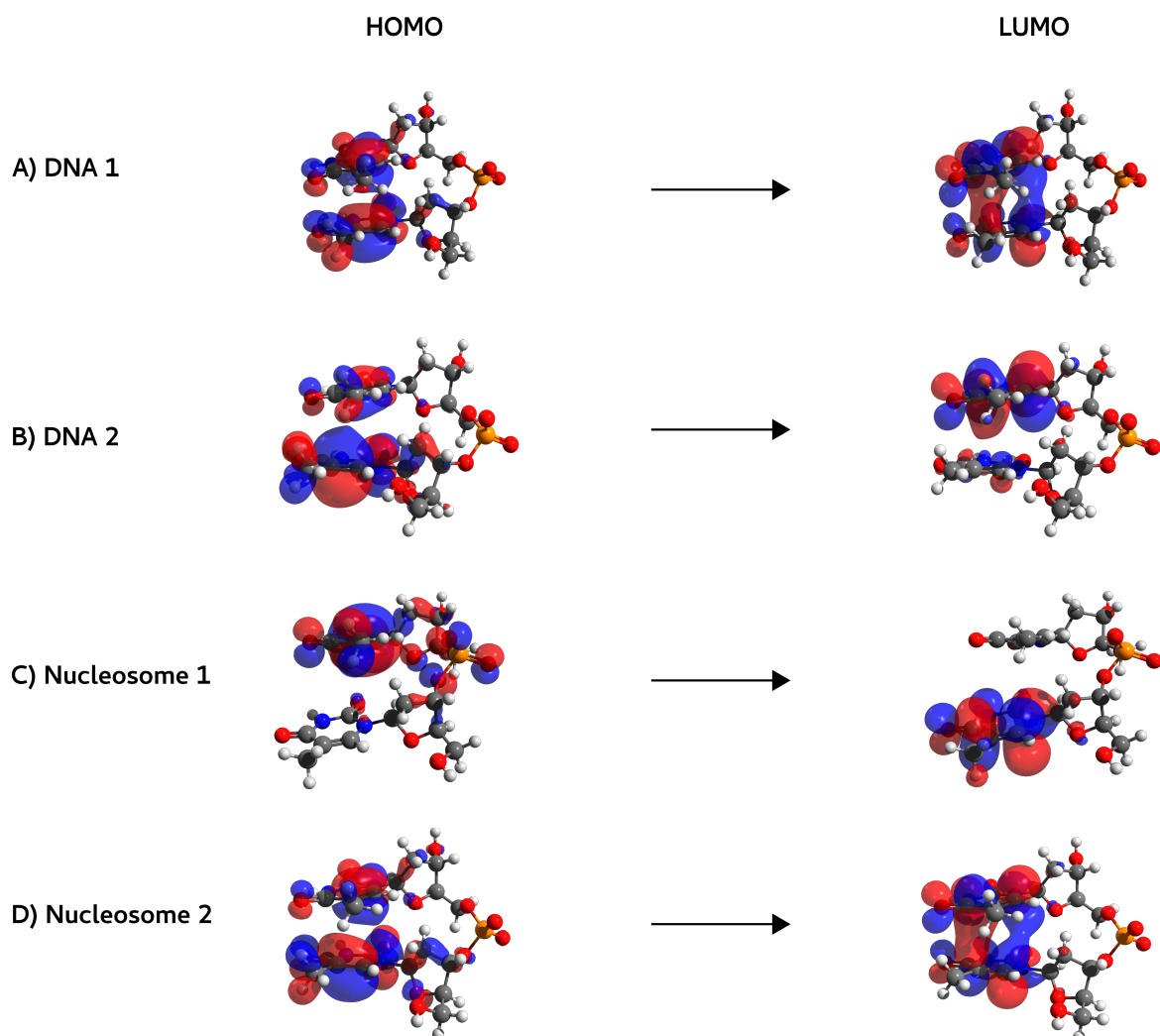
Figure 7.2: *Highest occupied molecular orbit (HOMO) and lowest unoccupied molecular orbit (LUMO) for four representative single-stranded DNA structures. A) Generated DNA with distance 3.65 Å between the C6-atoms shows clear CPD damage formation in the first excitation state. B) Generated DNA with corresponding distance 4.05 Å exhibits no CPD formation. C) Nucleosome structure with corresponding distance 5.99 Å clearly refuses covalent binding. D) Nucleosome structure with corresponding distance 3.72 Å results in a CPD damage formation in LUMO.*
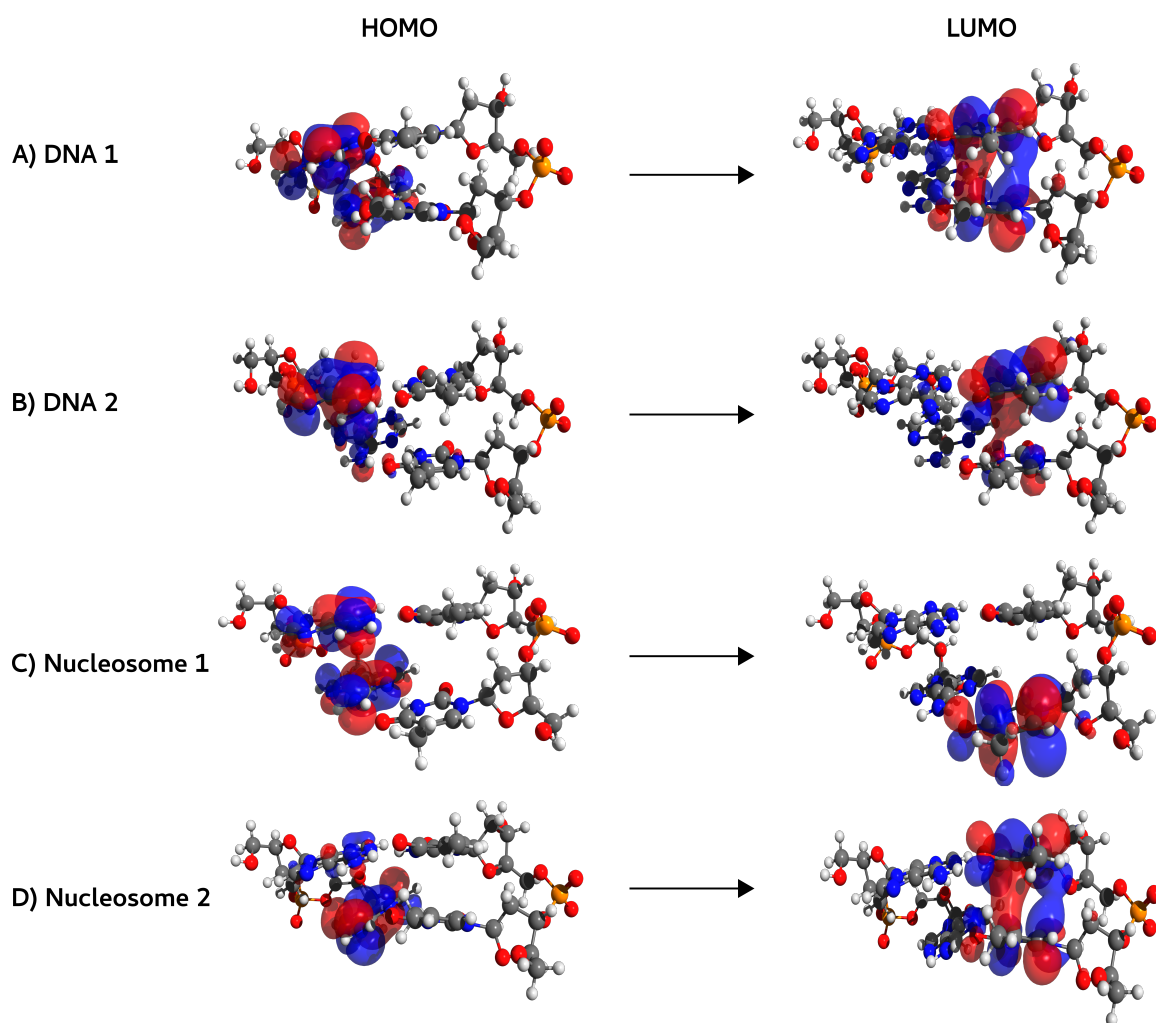
Figure 7.3: *Highest occupied molecular orbit (HOMO) and lowest unoccupied molecular orbit (LUMO) for four representative double-stranded DNA structures. A) Generated DNA with distance 3.65 Å between the C6-atoms shows clear CPD damage formation in the first excitation state. B) Generated DNA with distance 4.05 Å exhibits no CPD formation but a covalent connection between the closer C5 atoms. C) Nucleosome structure with distance 5.99 Å clearly refuses covalent binding. D) Nucleosome structure with distance 3.72 Å results in a CPD damage formation in LUMO.*
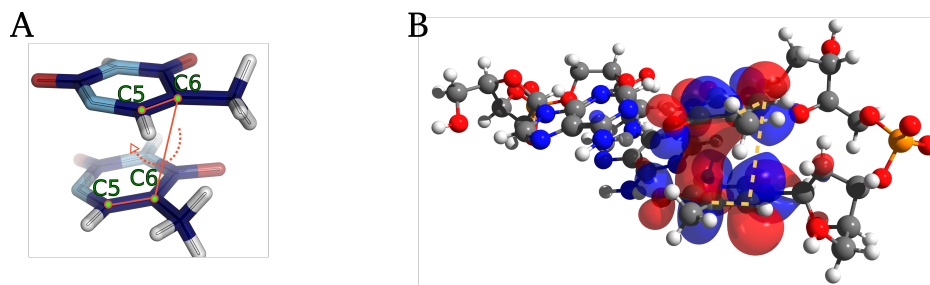
Figure 7.4: *A) Definition of the $\angle(C5,C6,C6,C5)$ dihedral angle between two thymines. B) DNA structure with C6-C6 distance of 3.79 Å and C5-C5 distance of 3.45 Å can not establish a covalent linkage between the C6 atoms for an increased $\angle(C5,C6,C6,C5)$ dihedral angle of 45° (yellow dashed line).*

Distances smaller than 3.6 Å hardly occur as this comes along with significantly higher energy barriers which we elaborate statistically with MD simulations later in more detail. The TDDFT calculations of 15 generated DNA and further 5 nucleosomal DNA structures are performed while the highest occupied molecular orbit (HOMO) and the lowest unoccupied molecular orbit (LUMO) are of focus. The calculations reveal the high relevance of the distance between the C6 (and also C5) atoms as a key role in such damages. The critical distance, i. e. for shorter distances CPD damage formations are observed while for larger distances no damages are detected, can be determined with high accuracy. It turns out that this critical distance is (3.86 ± 0.05) Å. For shorter distances, the CPD damage formation is observed for any analyzed DNA structure while for larger, a CPD damage formation is always suppressed.

Two representative example structures of the generated and two of the nucleosomal DNA in HOMO and LUMO (and as single-strands and double-strands), respectively, are visualized in Fig. 7.2, 7.3. The first DNA structure with a distance of 3.65 Å between the C6-atoms shows a clear CPD damage formation in the first excitation state. The second DNA structure is characterized by a larger distance of 4.05 Å and exhibits no CPD formation. Interestingly, in the double-stranded DNA the two C5 atoms are still covalently bound. The explanation for this phenomenon is given in the next paragraph. The covalent connection between the C5 atoms is not observed for greater distances than 3.79 Å. The last two nucleosomal DNA structures show either no damage (distance 5.99 Å) or a clear damage (distance 3.72 Å).

However, not only the distances between the C6 atoms are relevant but also the orientation of the pyrimidines to each other. The main orientation is captured by the $\angle(C5, C6, C6, C5)$ dihedral angle of the two thymines (Fig. 7.4 A). To study the impact of the dihedral angle, several TDDFT calculations were started concerning structures which only differ in the dihedral angle (see methods). For a C6-C6 distance of 3.91 Å and thus exceeding the critical distance, even a dihedral angle of 20° can not induce a CPD damage formation, either for any other dihedral angle. A shorter distance as 3.79 Å induces a CPD damage formation from 20° to 40° but for a dihedral angle of 45° it can not be maintained (see Fig. 7.4 B). This leads to two immediate consequences: First, the distances as a key in CPD damages dominate the dihedral angle contribution and second, an increase of the dihedral angle for
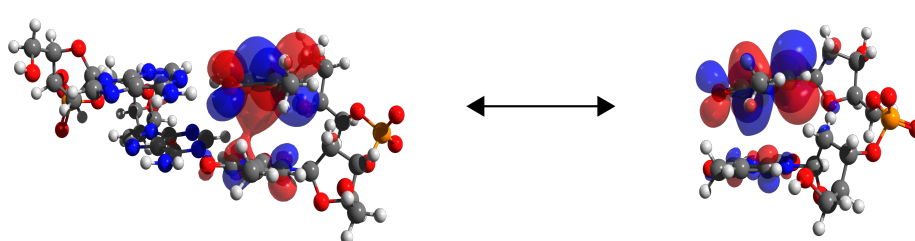
Figure 7.5: *Comparison of the first excited state (LUMO) between double-stranded and single-stranded DNA. The distance between the C5-atoms amounts to 3.76 Å. A covalent linkage between the C5 atoms is only observed for double-stranded DNA.*

fixed distances can disrupt the cyclobutane compound structure with the implication that all heavy-atomic orbitals in the thymine ring are correlated.

7.3.2 Double-strands as a damage promoter

As previously elaborated, base distance is one key factor in CPD damages. In addition to that, double-stranded DNA slightly promotes CPD damage formations at least in two ways: First, the rigidity of duplexes favors excitation energy transfers [131] and second, an electron-orbital realignment, which is independent of the molecular DNA geometry, also promotes CPD damages. In order to show the latter, from all double-stranded DNA structures only the extracted TpT dinucleotides are screened and compared with the original double-stranded DNA including this TpT sequence. The double-stranded DNA structures were geometry-optimized but not the subsequently extracted dinucleotides such that a different TTDFT result can not be a consequence of positional rearrangements. This a priori excludes effects which would originate from duplex stabilization. The TTDFT calculations emphasize that a clear presence or absence of a CPD damage formation in duplex DNA is also observed for single-stranded DNA. This is also in agreement with the experiments [61]. The orbital effect is only apparent in the critical zone. For instance, the duplex DNA structure in Fig. 7.3 B indicates covalently bound C5 atoms in LUMO while in contrast, the single-stranded equivalent shows no C5 connection anymore (Fig. 7.2 B). Another TDDFT calculation of two DNA molecules with a distance of 3.76 Å between the C5 carbons emphasize the orbital differences between single-strands and double-strands (Fig. 7.5): A covalent C5-C5 interaction is only observed for the duplex case and furthermore, the LUMO state of the double-stranded and single-stranded DNA seems slightly different. A larger energy barrier between HOMO and LUMO is found. In the duplex case, an energy of 4.84 eV while for the single-stranded case a higher energy of 5.19 eV is necessary to overcome the energy barrier between HOMO and LUMO to occupy the higher excitation band. The finding of a higher energy barrier is also valid for any other analysed DNA structure. As the energy barrier is higher in single-stranded DNA, it outlines the different wavelengths of UV radiation becoming necessary to induce CPD lesions whereby, in addition, also the flexibility plays a role in this context which we more generally discuss in the following.
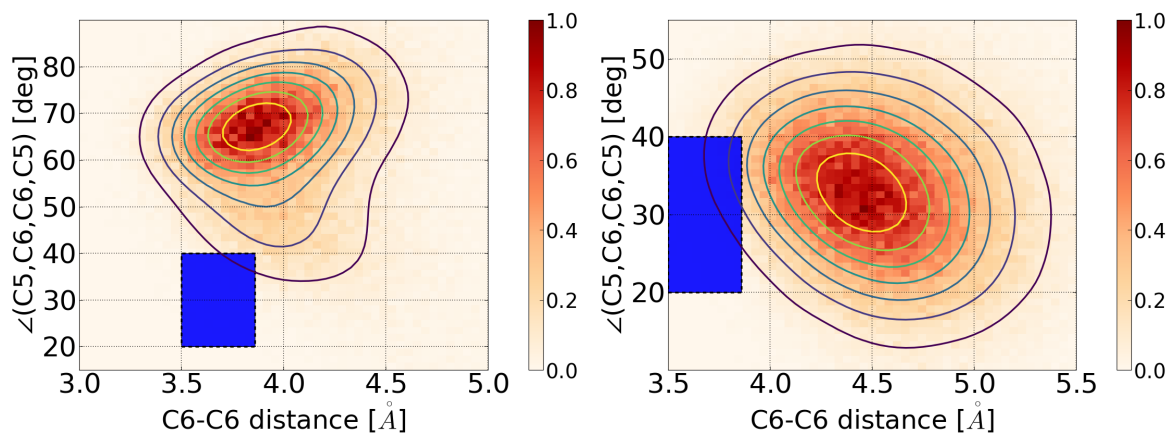
Figure 7.6: *2D probability distribution of the C6-C6 distance and the $\angle(C5,C6,C6,C5)$ dihedral angle, sampled from a MD simulation of single-stranded (left) and double-stranded DNA (right). The blue square defines the conformational space of CPD damage formation.*

7.3.3 Impact of conformational flexibility on CPD damage formation

The conducted QM calculations provide us with a strict criterion to assess the possibility of UV light-induced damage formation: Stacked thymidine nucleotides can undergo CPD damage formation only if the inter-base distance between the C6 atoms is smaller than $3.86\,\mathring{A}$ and the $\angle(C5,C6,C6,C5)$ dihedral angle is approximately in the range of $20°$ to $40°$. This finding shows a clear context between conformational flexibility and CPD damage formation, and hence has far-reaching biological consequences. Deformation of DNA through binding by proteins, for instance, may then also regulate DNA's vulnerability. From two X-ray structures (PDB: 1BNA and 3BSE) we have measured C6-C6 distances from five TpT steps to be in the range of 4.4 Å to 4.6 Å, which largely exceeds the threshold to form a CPD damage. DNA's atomistic structure thus reveals intrinsic robustness, which implies that UV light-induced damage formation must be driven by thermal fluctuations into the regime of C6-C6 proximity.

To validate the latter effect, we have first performed a 0.5 μs long all-atom MD simulation of a 5'-GCTTTCG-3' single-strand and recorded the C6-C6 distances and the $\angle(C5,C6,C6,C5)$ dihedral angle (Fig 7.6). The conformational subspace sampled during the MD simulation confirms the previous notion: The average C6-C6 distances are too high to allow a CPD damage formation (Fig 7.7), and the conformational subspace of CPD damage formation is only marginally occupied through thermal fluctuation. For single-stranded DNA, the probability of entering this subspace is only 0.5%. Note that the chosen sequence does not undergo hairpin folding or a similar structure formation process, which might also impact CPD dam-
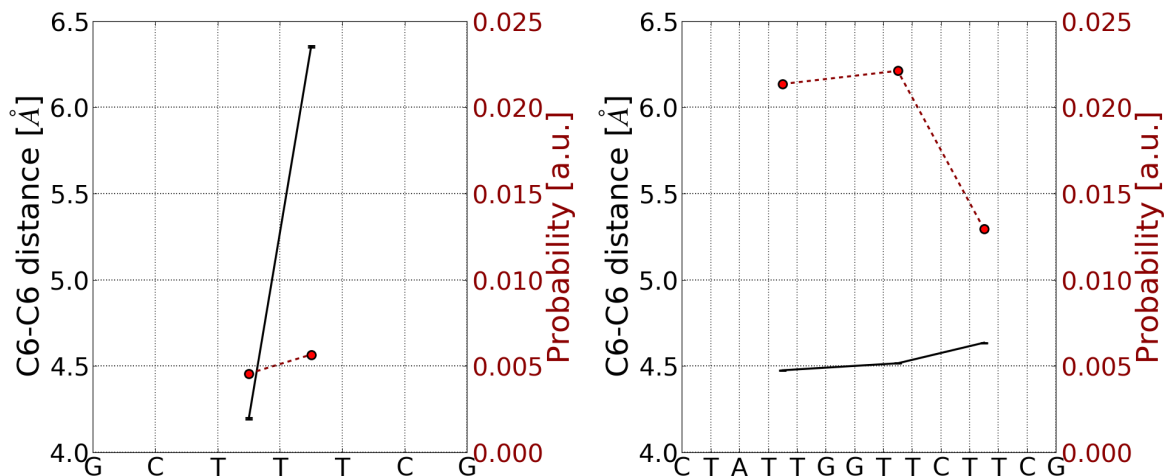
Figure 7.7: *Average C6-C6 distances and related probability of CPD damage formation in single-stranded (left) and double-stranded DNA (right), respectively, measured from MD simulations. The probability is computed by the fraction of samples with C6-C6 distances smaller than 3.86 Å and ∠(C5,C6,C6,C5) dihedral angles in the range of 20° to 40°.*

age formation. Analogously, we have investigated the conformational probability for CPD damage formation for double-stranded DNA by carrying out a 0.5 μs long MD simulation of a 5'-CTATTGGTTCTTCG-3' DNA duplex. The recorded mean C6-C6 distances and corresponding probabilities for the TpT steps are given in Fig 7.7. The probability of entering the respective subspace is $\sim 1.0\%$ to 2.5%, a slight increase over the single-stranded correspondent. As a consequence of the looser stacking of single-stranded DNA, their TpT stacks fluctuate less frequently into the critical regime, hence leading to a lower probability compared to double-stranded DNA. For the latter, a slight dependence on the tetranucleotide is evident (Fig 7.7). A tetranucleotide hierarchy is also experimentally measured according to Lu et al. [118]. The higher damage rate of GTTC over CTTC as obtained in our MD simulations is not reported for a sensitization with acetone. Different solvents, as utilized in the experiments, can largely change the CPD damage rate among different tetranucleotides [118]. This reflects that while the stacking conformation provides clear boundaries, sequence-dependent effects due to the electronic structure may still have an impact on CPD damage formation.

However, since DNA is mostly present in a complex with histones, we briefly discuss the effects of structure modulation on CPD damage formation for the nucleosome. The protein-induced deformation leads to a dipyrimidine formation which is strongly changed: Experiments in UV damage mapping have revealed an oscillatory pattern with a periodicity of ~ 10 bps [125]. To better understand this effect, we have performed a 50 ns MD simulation of a nucleosome with a DNA T-tract (sequence d(5'...TTTTT..3'), see Fig. 7.8), i.e. every step is a TpT stack which can undergo CPD damage formation. From the trajectory we have inferred the conformational probability for CPD damage formation (see Fig. 7.9). Intrigu-
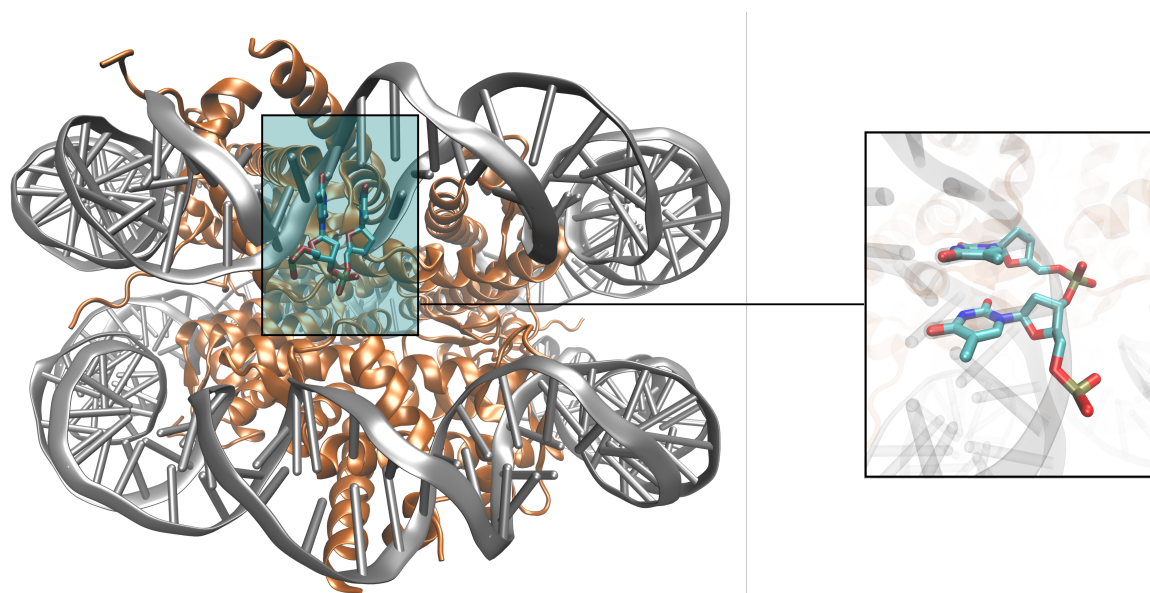
Figure 7.8: *Nucleosome structure including a histone (orange) and a DNA T-tract (grey). The zoomed snapshot shows an example of a dinucleotide at a rotational-inward position. The C6-C6 distance is strongly increased and hence the probability for a damage decreased.*
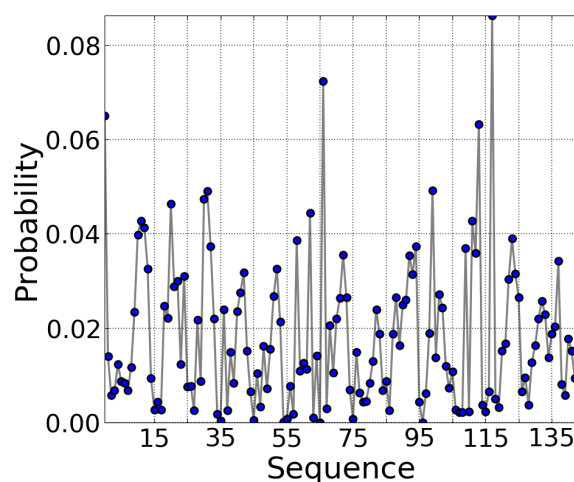

Figure 7.9: *CPD formation depends on nucleosome position: Conformational probability for CPD damage formation in the respective sequence of the DNA T-tract, computed by the fraction of samples with C6-C6 distances smaller than 3.86 Å and $\angle$(C5,C6,C6,C5) dihedral angles in the range of 20° to 40°. The oscillatory profile with a periodicity of $\sim$ 10 base pairs reveals that rotational-inward positions are less prone for CPD damage formation.*

ingly, the resulting CPD probability pattern exhibits a periodicity in very good agreement with the experimental findings [125]. For every 10 base-pair steps ($\sim$ 1 turn of DNA), the possibility for CPD lesions is significantly suppressed. In this case the dinucleotide forms a rotational-inward setting (i.e. the minor groove faces the histone, see Fig. 7.8). We therefore conclude that the non-uniform deformation of DNA in nucleosomes directly translates into a non-uniform UV damage pattern.

This analysis revealed that nucleosomes substantially change CPD damage formations and protect nucleosomal DNA with a rotational-inward setting. The more compressed minor grooves of the A-T–rich DNA regions tend to adopt inward rotational settings to facilitate DNA bending around the histone octamer and to strengthen histone interactions with the DNA minor groove [132, 175].

7.4 Conclusion

In this study we discussed the flexibility dependence of CPD damage formations for diverse DNA T-tracts. For CPD damages in thymine-rich DNA, the C6 and respectively C5 atoms of the dipyrimidines bind covalently after absorption of UV light. With the additional photon energy, the energy barrier between HOMO and LUMO is overcome for occupying the higher LUMO excitation band. This effect depends strongly on the geometric conditions of the dinucleotide. TDDFT calculations revealed a C6-C6 maximal distance of 3.86 Å until a CPD can be observed. When exceeding this distance, a CPD damage formation was suppressed in any case. Moreover, a $\angle(C5, C6, C6, C5)$ dihedral angle dependence was also found. TDDFT calculations indicated a CPD damage formation until 40° and rejected it for 45°. A too high dihedral angle can disrupt the cyclobutane structure and thus the CPD damage formation although the distance of 3.86 Å is undercut. Still, the distance accounts for the dominant root to this damage. The presence of the complementary strand has only a marginal impact on the CPD damage formation but double-strands are nevertheless CPD damage promoters for at least two reasons: Primary, it is the rigidity of duplexes, and furthermore, induced orbital realignments favor excitation energy transfers.

However, we raise the question when DNA reaches or undercuts the sufficient distance of 3.86 Å and simultaneously the dihedral angle of 40° by natural base-to-base oscillations. It turns out that the average C6-C6 distance is too high to allow a CPD damage formation. The conformational subspace of CPD damage formations is only marginally occupied through thermal fluctuation where single-stranded DNA enters this subspace with a probability of only 0.5% and double-stranded DNA with approximately 1.0% to 2.5%. Thus, DNA is protected by its natural flexibility. In nucleosomes, on the other hand, the conformational setting of DNA is changed: The DNA T-tract structure is strongly deformed due to the wrapping around the histone core. It turns out that for approximately every 10 bps the possibility for CPD lesions is significantly suppressed as a more favored histone-DNA interaction leads to a rotational-inward setting of the TpT dinucleotide.

8 Outlook

In my view, the idea behind MD simulations is convincing but also in some points criticizable. We gain a molecular insight into the femtosecond-timescale dynamics which is not accessible on the same level for experimentalists. This endows us with a significant edge. For instance, in the first project, motivated by the telomere-binding gold-carbene which is known to induce mortality in cancer cells, the free binding energy of this drug-quadruplex complex exceeded the one of the comparative complex with the duplex DNA twofold. This example demonstrates the strength of free energy calculations which can be carried over to countless other application cases. Studies of regional dynamics, as in the Pot1pC protein where the protein pocket and outstanding connecting loops changed successively due to DNA incorporation, emphasize the capability of MD to explain or even predict underlying mechanisms. This is also a great strength of MD. On the other hand, this demands a certain accuracy in the simulations and methods which can be problematic in some cases. Methods like MMPBSA should be, to my opinion, more cautiously used as it is notorious for giving inaccurate results. Its calculation of the 'free energy', even for excluding the error-prone entropic part, computes inaccurate solvent interactions or neglects reorganisations of the receptor, for instance. Utilization of the advanced method according to Woo and Roux is more sophisticated and to date, state-of-the-art to compute free binding energies. Another concern is the current DNA force field. The DNA force field seems to be satisfying but it is also in need of improvements: One main issue in the force field parametrization is the decades-old and inaccurate partial charge distribution. Accurately determined partial charges would be reliable for single-stranded, double-stranded DNA or deformed DNA, as we have shown. A refined force field with additional terms considering quantum mechanical effects, like polarization, is also essential. This will be even more prudent when the simulation times get longer which is expected in the future due to progressively better CPUs and graphic cards. So far, ultralong classical MD simulations indicate too much stiffness in B-DNA, for instance [217]. Comparative long QM-based simulations are probably more accurate but unreported to date [217].

However, a combination of quantum mechanical and molecular classical theory seems promising. In the last chapter treating the CPD effect, we computed the threshold between two bases, for which an undercut causes a CPD damage, with high accuracy ($\sim 0.05\,\text{Å}$). Simultaneously, we made use of the MD analysis to elaborate when this conformational subspace is reached through thermal fluctuations. Only this combination of theories made this result possible. In general, it is also reasonable to simulate a major part of a system with classical MD simulations while at the same time a minor part is simulated on a quantum mechanical level. This is basically the QM/MM method for which the Nobel Prize was awarded in 2013. Unfortunately, the computation time for QM simulations is very high and increases with the number of particles. As the performance enhancement of CPUs and graphic cards will likely flatten, alternative technologies become more crucial. Quantum computers gained greater interest in the last couple of years. If this was realized and accessible, it could be a milestone in computational research. An immediate boost in quantum mechanical computations and simulations might be the consequence while it could also mean a role change for MD and QM.

Although for today some restrictions are associated with MD and also QM, both theories with more accurate methods and a refined DNA force field are very promising for the future.

Warum?

Eine simple und zugleich philosophisch hoch relevante Frage. Noch interessanter ist die Frage, warum wir über dieses Warum nachdenken. Im Prinzip ist dies auch die Ursache für unsere menschliche Neugier, die uns dazu streben lässt, nach Lösungen für Unklarheiten über die Natur zu suchen. Schwierig ist es, diesen speziellen Lebensweg in Kategorien wie richtig oder falsch einzuordnen. Gibt es überhaupt einen guten oder schlechten Weg? Wenn wir, wie oft argumentiert wird, nur aus Atomen bestehen und alles im Leben relativ ist, wäre es nicht dann auch unerheblich, ob das Leben gleich endet oder erst in Jahrzehnten? Tatsächlich ist es nicht so, dass wir nur aus Atomen bestehen, sondern auch aus der Anordnung und den Wechselwirkungen zwischen jenen, die im Rahmen physikalischer Gesetze agieren. Unsere ganz spezielle materielle Anordnung, nennen wir sie Biologie, führt zu einer Reihe von logischen Konsequenzen. Wie die physikalischen Gesetze kann die zugrunde liegende Biologie nicht a priori verändert werden. Ebenso verfügt jeder gesunde Mensch über einen Verstand. Wenn wir eine Axiomatik aufbauen wollen, aus der wir alle Handlungen ableiten wollen, dann müssen diese die Grundpfeiler bilden. Etwas, das für alle Menschen gleichermaßen gilt und aus den primitivsten, unveränderlichen Eigenschaften besteht. Aus der Axiomatik können wir Entscheidungen ableiten, die uns zum nächsten Schritt führen, wie klein dieser auch sein mag. Die Entscheidung ist das Prinzip im Leben und sie ist eine Konsequenz der Axiomatik. Aus Letzterem folgt auch der Grundsatz über die Bewahrung der genetischen Information durch Vererbung. Damit wird ein bewusst vorzeitig beendetes Leben rational ausgeschlossen. Unser biologisches Steuerelement ist in der Lage, objektiv und rational Entscheidungen zu treffen. Eine Grundlage nur aufgrund dessen ist jedoch unvollständig, da wir nicht einen anderen komplementären Teil vernachlässigen können, die Emotionen. Eine rein rationale oder rein emotionale Entscheidung kann mit verschiedenen Gegenbeispielen verworfen werden. Demnach muss eine ideale Entscheidung so getroffen werden, dass sie in einer Bilanz aus Rationalität und Emotionalität liegt. Nach Nash sollte die Entscheidung für sich selbst und gleichzeitig für die Gruppe optimal sein. Dies bringt uns zurück zur ursprünglichen Frage und damit auch zum Sinn des Lebens, auf welchen es nur eine Antwort zu geben scheint: Ja, er existiert.

9 Acknowledgments - Danksagung

References

[1] H. Abdi and L. J. Williams. Principal component analysis. *John Wiley & Sons, Inc.WIREs Comp Stat*, 2:433–459, 2010.

[2] R. J. Abraham and W. T. Raynes. *Nuclear Magnetic Resonance*, volume 7. The Royal Society of Chemistry, 1978.

[3] B. Alberts, J. Alexander, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell*, chapter 2, pages 45–124. Cell Chemistry and Biosynthesis. Garland Science, Taylor & Francis Group, 5th edition, 2008.

[4] B. Alberts, J. Alexander, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell*, chapter 4, pages 195–262. DNA, Chromosomes and Genomes. Garland Science, Taylor & Francis Group, 5th edition, 2008.

[5] I. Alkorta and J. Elguero. Ab initio (GIAO) calculations of absolute nuclear shieldings for representative compounds containing 1(2)H, 6(7)Li, 11B, 13C, 14(15)N, 17O, 19F, 29Si, 31P, 33S, and 35Cl nuclei. *Structural Chemistry*, 9(3):187–202, 1998.

[6] A. Arola and R. Vilar. Stabilisation of G-Quadruplex DNA by small molecules. *Current Topics in Medicinal Chemistry*, 8(15):1405–1415, 2008.

[7] C. Autexier and N. F. Lue. The structure and function of telomerase reverse transcriptase. *Annu. Rev. Biochem.*, 75:493–517, 2006.

[8] D. Bagayoko. Understanding density functional theory (DFT) and completing it in practice. *AIP Advances*, 4(12):127104–, Dec. 2014.

[9] A. Bakan, L. M. Meireles, and I. Bahar. ProDy: Protein Dynamics Inferred from Theory and Experiments. *Bioinformatics*, 27(11):1575–1577, June 2011.

[10] P. Banáš, A. Mládek, M. Otyepka, M. Zgarbová, P. Jurečka, D. Svozil, F. Lankaš, and J. Šponer. Can we accurately describe the structure of adenine tracts in B-DNA? Reference quantum-chemical computations reveal overstabilization of stacking by molecular mechanics. *J. Chem. Theory Comput.*, 8(7):2448–2460, July 2012.

[11] V. Barone and M. Cossi. Quantum calculation of molecular energies and energy gradients in solution by a conductor solvent model. *J. Phys. Chem. A*, 102(11):1995–2001, Mar. 1998.

[12] C. Bayly, K. Merz, D. Ferguson, W. Cornell, T. Fox, J. Caldwell, P. Kollman, P. Cieplak, I. Gould, and D. Spellmeyer. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society*, 117(19):5179–5197, May 1995.

[13] C. Bazzicalupi, M. Ferraroni, F. Papi, L. Massai, B. Bertrand, L. Messori, P. Gratteri, and A. Casini. Determinants for tight and selective binding of a medicinal dicarbene

gold(I) complex to a telomeric DNA G-quadruplex: A joint ESI MS and XRD investigation. *Angew. Chem. Int. Ed.*, 55(13):4256–4259, Mar. 2016.

[14] A. D. Becke. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A*, 38(6):3098–3100, Sept. 1988.

[15] A. D. Becke. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.*, 98(7):5648–5652, Apr. 1993.

[16] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, 81(8):3684–3690, Oct. 1984.

[17] B. Bernardes de Jesus and M. A. Blasco. Telomerase at the intersection of cancer and aging. *Trends in Genetics*, 29(9):513–520, 2013.

[18] B. Bertrand, A. Citta, I. L. Franken, M. Picquet, A. Folda, V. Scalcon, M. P. Rigobello, P. Le Gendre, A. Casini, and E. Bodio. Gold(I) NHC-based homo- and heterobimetallic complexes: Synthesis, characterization and evaluation as potential anticancer agents. *JBIC Journal of Biological Inorganic Chemistry*, 20(6):1005–1020, 2015.

[19] B. Bertrand, A. de Almeida, E. P. M. van der Burgt, M. Picquet, A. Citta, A. Folda, M. P. Rigobello, P. Le Gendre, E. Bodio, and A. Casini. New gold(I) organometallic compounds with biological activity in cancer cells. *Eur. J. Inorg. Chem.*, 2014(27):4532–4536, Sept. 2014.

[20] B. Bertrand, J. Fernandez-Cestau, J. Angulo, M. M. D. Cominetti, Z. A. E. Waller, M. Searcey, M. A. O'Connell, and M. Bochmann. Cytotoxicity of pyrazine-based cyclometalated $(C\hat{\ }N^{pz}\hat{\ }C)Au(III)$ carbene complexes: Impact of the nature of the ancillary ligand on the biological properties. *Inorg. Chem.*, 56(10):5728–5740, May 2017.

[21] B. Bertrand, L. Stefan, M. Pirrotta, D. Monchaud, E. Bodio, P. Richard, P. Le Gendre, E. Warmerdam, M. H. de Jager, G. M. Groothuis, M. Picquet, and A. Casini. Caffeine-based gold(I) n-heterocyclic carbenes as possible anticancer agents: Synthesis and biological properties. *Inorg. Chem.*, 53(4):2296–2303, Feb. 2014.

[22] P. J. Bickel and E. Levina. Some theory for fisher's linear discriminant function, 'naive bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, 10(6):989–1010, Dec. 2004.

[23] F. Bleichert, M. R. Botchan, and J. M. Berger. Mechanisms for initiating cellular DNA replication. *Science*, 355(6327):eaah6317–, Feb. 2017.

[24] M. L. Bochman, K. Paeschke, and V. A. Zakian. DNA secondary structures: Stability and function of G-quadruplex structures. *Nature Reviews Genetics*, 13(11):770–780, 2012.

[25] D. Bramhill and A. Kornberg. Duplex opening by dnaA protein at novel sequences in initiation of replication at the origin of the E. coli chromosome. *Cell*, 52(5):743–755, 1988.

[26] R. F. Brown, C. T. Andrews, and A. H. Elcock. Stacking free energies of all DNA and RNA nucleoside pairs and dinucleoside-monophosphates computed using recently revised AMBER parameters and compared with experiment. *J. Chem. Theory Comput.*, 11(5):2315–2328, May 2015.

[27] F. Brueckner, U. Hennecke, T. Carell, and P. Cramer. CPD damage recognition by transcribing RNA polymerase II. *Science*, 315(5813):859, Feb. 2007.

[28] C. Bustamante, S. B. Smith, J. Liphardt, and D. Smith. Single-molecule studies of DNA mechanics. *Current Opinion in Structural Biology*, 10(3):279–285, 2000.

[29] J. Cadet and T. Douki. Formation of UV-induced DNA damage contributing to skin cancer development. *Photochem. Photobiol. Sci.*, 17(12):1816–1841, 2018.

[30] R. Campos-Olivas, J. M. Louis, D. Clérot, B. Gronenborn, and A. M. Gronenborn. The structure of a replication initiator unites diverse aspects of nucleic acid metabolism. *Proc Natl Acad Sci USA*, 99(16):10310–, Aug. 2002.

[31] X. Cang, J. Šponer, and T. E. Cheatham III. Explaining the varied glycosidic conformational, G-tract length and sequence preferences for anti-parallel G-quadruplexes. *Nucleic acids research*, 39(10):4499–4512, 2011.

[32] C. A. Casacio, L. S. Madsen, A. Terrasson, M. Waleed, K. Barnscheidt, B. Hage, M. A. Taylor, and W. P. Bowen. Quantum-enhanced nonlinear microscopy. *Nature*, 594(7862):201–206, 2021.

[33] D. Case, R. B. an D.S. Cerutti, I. T.E. Cheatham, T. Darden, R. Duke, T. Giese, H. Gohlke, A. Goetz, N. Homeyer, S. Izadi, P. Janowski, A. K. J. Kaus, T. Lee, S. LeGrand, P. Li, C. Lin, T. Luchko, R. Luo, B. Madej, D. Mermelstein, K. Merz, G. Monard, H. Nguyen, H. Nguyen, I. Omelyan, A. Onufriev, D. Roe, A. Roitberg, C. Sagui, C. Simmerling, W. Botello-Smith, J. Swails, R. Walker, J. Wang, R. Wolf, X. Wu, L. Xiao, and P. Kollman. *Amber 16.* University of California, San Francisco, 2016.

[34] D. Case, I. Y. Ben-Shalom, S. R. Brozell, D. Cerutti, I. T.E. Cheatham, T. Darden, R. Duke, D. Ghoreish, M. K. Gilson, H. Gohlke, A. Goetz, D. Greene, R. Harris, N. Homeyer, Y. Huang, S. Izadi, A. Kovalenko, T. Kurtzmann, T. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, D. Mermelstein, K. Merz, Y. Miao, G. Monard, C. Nguyen, H. Nguyen, I. Omelyan, A. Onufriev, F. Pan, R. Qi, D. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C. Simmerling, J. Smith, R. Salomon-Ferrer, J. Swails, R. C. Walker, J. Wang, H. Wei, R. Wolf, X. Wu, L. Xiao, D. M. York, and P. Kollman. *Amber 18.* University of California, San Francisco, 2018.

[35] M. E. Casida. Time-dependent density functional response theory for molecules. In *Recent Advances in Computational Chemistry*, volume 1, pages 155–192–. WORLD SCIENTIFIC, Nov. 1995.

[36] M. E. Casida and D. R. Salahub. Asymptotic correction approach to improving approximate exchange-correlation potentials: Time-dependent density-functional theory calculations of molecular excitation spectra. *J. Chem. Phys.*, 113(20):8918–8935, Nov. 2000.

[37] J. B. Chaires. Human telomeric G-quadruplex: Thermodynamic and kinetic studies of telomeric quadruplex stability. *The FEBS journal*, 277(5):1098–1106, 2010.

[38] C.-E. Chang, W. Chen, and M. K. Gilson. Evaluating the accuracy of the quasiharmonic approximation. *J. Chem. Theory Comput.*, 1(5):1017–1028, Sept. 2005.

[39] Z. Chen, Y. Jiang, Y.-T. Shao, M. E. Holtz, M. Odstrčil, M. Guizar-Sicairos, I. Hanke, S. Ganschow, D. G. Schlom, and D. A. Muller. Electron ptychography achieves atomic-resolution limits set by lattice vibrations. *Science*, 372(6544):826–, May 2021.

[40] W. Cheney and D. Kincaid. *Linear Algebra: Theory and Applications*. Jones & Bartlett Publishers, 2012. Google-Books-ID: S0imN2tl1qwC.

[41] D. B. Chesnut. A theoretical study of ^{31}P NMR chemical shielding models for concentrated phosphoric acid solution. *J. Phys. Chem. A*, 109(51):11962–11966, Dec. 2005.

[42] K.-C. Chow and W. L. Tung. Magnetic field exposure enhances DNA repair through the induction of DnaK/J synthesis. *FEBS Letters*, 478(1):133–136, 2000.

[43] A. Cléry, M. Blatter, and F. H.-T. Allain. RNA recognition motifs: Boring? Not quite. *Current Opinion in Structural Biology*, 18(3):290–298, 2008.

[44] J. E. Croy, S. E. Altschuler, N. E. Grimm, and D. S. Wuttke. Nonadditivity in the recognition of single-stranded DNA by the schizosaccharomyces pombe protection of telomeres 1 DNA-binding domain, Pot1-DBD. *Biochemistry*, 48(29):6864–6875, July 2009.

[45] L. Daya-Grosjean and A. Sarasin. The role of UV induced lesions in skin carcinogenesis: An overview of oncogene and tumor suppressor gene modifications in xeroderma pigmentosum skin tumors. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 571(1):43–56, 2005.

[46] T. de Lange. How telomeres solve the end-protection problem. *Science*, 326(5955):948–952, Nov. 2009.

[47] E. L. Denchi and T. de Lange. Protection of telomeres through independent control of ATM and ATR by TRF2 and POT1. *Nature*, 448(7157):1068–1071, Aug. 2007.

[48] F. J. Devlin, J. W. Finley, P. J. Stephens, and M. J. Frisch. Ab initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields: A comparison of local, nonlocal, and hybrid density functionals. *J. Phys. Chem.*, 99(46):16883–16902, Nov. 1995.

[49] T. Dickey, M. McKercher, and D. Wuttke. Nonspecific recognition is achieved in Pot1pC through the use of multiple binding modes. *Structure*, 21(1):121–132, 2013.

[50] C. Ding and X. He. *K-Means Clustering via Principal Component Analysis*. ICML '04. Association for Computing Machinery, New York, NY, USA, 2004.

[51] R. Ditchfield. Self-consistent perturbation theory of diamagnetism. *Molecular Physics*, 27(4):789–807, 1974.

[52] M. Einasto, L. J. Liivamagi, E. Saar, J. Einasto, E. Tempel, E. Tago, and V. J. Martinez. SDSS DR7 superclusters. Principal component analysis. *A&A*, 535:A36, Nov. 2011. arXiv: 1108.4372.

[53] D. M. Feldser and C. W. Greider. Short telomeres limit tumor progression in vivo by inducing senescence. *Cancer Cell*, 11(5):461–469, 2007.

[54] H.-W. Fink and C. Schönenberger. Electrical conduction through DNA molecules. *Nature*, 398(6726):407–410, 1999.

[55] P. S. Freemont, A. N. Lane, and M. R. Sanderson. Structural aspects of protein-DNA recognition. *Biochem. J.*, 278 (Pt 1):1–23, Aug. 1991.

[56] D. Frenkel and B. Smit. Understanding molecular simulation. *Academic Press, 2nd edition*, pages 23–62, 2002.

[57] 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, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, J. Cioslowski, and D. J. Fox. *Gaussian-09 Revision C.01*, 2009. Gaussian Inc. Wallingford CT.

[58] R. Galindo-Murillo, J. C. Robertson, M. Zgarbová, J. Šponer, M. Otyepka, P. Jurečka, and T. E. Cheatham. Assessing the current state of Amber force field modifications for DNA. *J. Chem. Theory Comput.*, 12(8):4114–4127, Aug. 2016.

[59] D. Ganyushin and F. Neese. First-principles calculations of magnetic circular dichroism spectra. *J. Chem. Phys.*, 128(11):114117–, Mar. 2008.

[60] D. M. Gilbert. In search of the holy replicator. *Nature Reviews Molecular Cell Biology*, 5(10):848–855, 2004.

[61] L. K. Gordon and W. A. Haseltine. Quantitation of cyclobutane pyrimidine dimer formation in double- and single-stranded DNA fragments of defined sequence. *Jstor*, 89(1):99–112, 1982.

[62] S. Grimme, J. Antony, S. Ehrlich, and H. Krieg. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.*, 132(15):154104–, Apr. 2010.

[63] E. Gross and W. Kohn. Time-dependent density-functional theory. In P.-O. Löwdin, editor, *Density Functional Theory of Many-Fermion Systems*, volume 21, pages 255–291. Academic Press, 1990.

[64] F. Guarra, T. Marzo, M. Ferraroni, F. Papi, C. Bazzicalupi, P. Gratteri, G. Pescitelli, L. Messori, T. Biver, and C. Gabbiani. Interaction of a gold(I) dicarbene anticancer drug with human telomeric DNA G-quadruplex: Solution and computationally aided X-ray diffraction analysis. *Dalton Trans.*, 47(45):16132–16138, 2018.

[65] B. Göhler, V. Hamelbeck, T. Z. Markus, M. Kettner, G. F. Hanne, Z. Vager, R. Naaman, and H. Zacharias. Spin selectivity in electron transmission through self-assembled monolayers of double-stranded DNA. *Science*, 331(6019):894–, Feb. 2011.

[66] S. Haider, G. N. Parkinson, and S. Neidle. Molecular dynamics and principal components analysis of human telomeric quadruplex multimers. *Biophysical Journal*, 95(1):296–311, 2008.

[67] H. Han and L. H. Hurley. G-quadruplex DNA: A potential target for anti-cancer drug design. *Trends in Pharmacological Sciences*, 21(4):136–142, 2000.

[68] M. Han, M. Yagura, and T. Itoh. Specific interaction between the initiator protein (Rep) and origin of plasmid ColE2-P9. *J. Bacteriol.*, 189(3):1061–, Feb. 2007.

[69] S. C. Harvey, R. K.-Z. Tan, and T. E. Cheatham III. The flying ice cube: Velocity rescaling in molecular dynamics leads to violation of energy equipartition. *J. Comput. Chem.*, 19(7):726–740, May 1998.

[70] P. J. Hay and W. R. Wadt. Ab initio effective core potentials for molecular calculations. Potentials for K to Au including the outermost core orbitals. *J. Chem. Phys.*, 82(1):299–310, Jan. 1985.

[71] E. Hiyama and K. Hiyama. Telomere and telomerase in stem cells. *British journal of cancer*, 96(7):1020–1024, Apr. 2007.

[72] V. Hnizdo, E. Darian, A. Fedorowicz, E. Demchuk, S. Li, and H. Singh. Nearest-neighbor nonparametric method for estimating the configurational entropy of complex molecules. *J. Comput. Chem.*, 28(3):655–668, Feb. 2007.

[73] P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Phys. Rev.*, 136(3B):B864–B871, Nov. 1964.

[74] S. T. Holmes and R. J. Iuliucci. Modeling NMR chemical shift tensors. In *Modern Magnetic Resonance*, pages 959–993. Springer International Publishing, Cham, 2018.

[75] A. Hospital, J. R. Goñi, M. Orozco, and J. L. Gelpí. Molecular dynamics simulations: Advances and applications. *Advances and applications in bioinformatics and chemistry : AABC*, 8:37–47, Nov. 2015.

[76] D. I. Hoult and B. Bhakar. NMR signal reception: Virtual photons and coherent spontaneous emission. *Concepts Magn. Reson.*, 9(5):277–297, Jan. 1997.

[77] D. Hsu, S. M. Kakade, and T. Zhang. A Spectral Algorithm for Learning Hidden Markov Models. *arXiv:0811.4413 [cs]*, Nov. 2008. arXiv: 0811.4413.

[78] W. Humphrey, A. Dalke, and K. Schulten. VMD: Visual molecular dynamics. *J. Molec. Graphics*, 14(1):33–38, 1996.

[79] J. L. Huppert. Four-stranded nucleic acids: Structure, function and targeting of G-quadruplexes. *Chemical Society Reviews*, 37(7):1375–1384, 2008.

[80] I. Husain, J. Griffith, and A. Sancar. Thymine dimers bend DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 85(8):2558–2562, Apr. 1988.

[81] B. Islam, P. Stadlbauer, S. Neidle, S. Haider, and J. Šponer. Can we execute reliable MM-PBSA free energy computations of relative stabilities of different guanine quadruplex folds? *J. Phys. Chem. B*, 120(11):2899–2912, Mar. 2016.

[82] I. Ivani, P. D. Dans, A. Noy, A. Pérez, I. Faustino, A. Hospital, J. Walther, P. Andrio, R. Goñi, and A. Balaceanu. Parmbsc1: A refined force field for DNA simulations. *Nat. Methods*, 13(1):55–58, 2016.

[83] F. Jacob, S. Brenner, and F. Cuzin. On the regulation of DNA replication in bacteria. *Cold Spring Harbor Symposia on Quantitative Biology*, 28:329–348, Jan. 1963.

[84] A. Jameson and C. J. Jameson. Gas-phase ^{13}C chemical shifts in the zero-pressure limit: Refinements to the absolute shielding scale for ^{13}C. *Chemical Physics Letters*, 134(5):461–466, 1987.

[85] C. J. Jameson and A. De Dios. Absolute shielding scale for ^{31}P from gas-phase NMR studies. *Chemical Physics Letters*, 167(6), 1990.

[86] C. J. Jameson, A. K. Jameson, D. Oppusunggu, S. Wille, P. M. Burrell, and J. Mason. ^{15}N nuclear magnetic shielding scale from gas phase studies. *J. Chem. Phys.*, 74(1):81–88, Jan. 1981.

[87] M. Janeček, P. Kührová, V. Mlýnský, M. Otyepka, J. Šponer, and P. Banáš. W-resp: Well-restrained electrostatic potential-derived charges. Revisiting the charge derivation model. *J. Chem. Theory Comput.*, 17(6):3495–3509, June 2021.

[88] F. Jensen. *Introduction to Computational Chemistry*. John Wiley & Sons, 2nd edition, 2007.

[89] T. C. Johnstone, K. Suntharalingam, and S. J. Lippard. The next generation of platinum drugs: Targeted Pt(II) agents, nanoparticle delivery, and Pt(IV) prodrugs. *Chem. Rev.*, 116(5):3436–3486, Mar. 2016.

[90] I. T. Jolliffe and J. Cadima. Principal component analysis: A review and recent developments. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 374(2065):20150202–20150202, Apr. 2016.

[91] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79(2):926–935, 1983.

[92] M. Karplus and J. N. Kushick. Method for estimating the configurational entropy of macromolecules. *Macromolecules*, 14(2):325–332, Mar. 1981.

[93] M. Kettner, B. Göhler, H. Zacharias, D. Mishra, V. Kiran, R. Naaman, D. H. Waldeck, S. Sęk, J. Pawłowski, and J. Juhaniewicz. Spin filtering in electron transport through chiral oligopeptides. *J. Phys. Chem. C*, 119(26):14542–14547, July 2015.

[94] N. Kim, M. Piatyszek, K. Prowse, C. Harley, M. West, P. Ho, G. Coviello, W. Wright, S. Weinrich, and J. Shay. Specific association of human telomerase activity with immortal cells and cancer. *Science*, 266(5193):2011–2015, Dec. 1994.

[95] R. L. Kleinberg and J. A. Jackson. An introduction to the history of NMR well logging. *Concepts Magn. Reson.*, 13(6):340–342, Jan. 2001.

[96] A. Knips and M. Zacharias. Influence of a cis,syn-cyclobutane pyrimidine dimer damage on DNA conformation studied by molecular dynamics simulations. *Biopolymers*, 103(4):215–222, Apr. 2015.

[97] A. Knips and M. Zacharias. Both DNA global deformation and repair enzyme contacts mediate flipping of thymine dimer damage. *Scientific Reports*, 7(1):41324, Jan. 2017.

[98] S. C. Koch, J. Kuper, K. L. Gasteiger, N. Simon, R. Strasser, D. Eisen, S. Geiger, S. Schneider, C. Kisker, and T. Carell. Structural insights into the recognition of cisplatin and AAF-dG lesion by Rad14 (XPA). *Proceedings of the National Academy of Sciences of the United States of America*, 112:8272–7, Jul 2015.

[99] W. Koch and M. C. Holthausen. *The Basic Machinery of Density Functional Programs*, chapter 7, pages 93–116. John Wiley & Sons, Ltd, 2001.

[100] W. Koch and M. C. Holthausen. *The Hohenberg-Kohn Theorems*, chapter 4, pages 33–40. John Wiley & Sons, Ltd, 2001.

[101] W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Phys. Rev.*, 140(4A):A1133–A1138, Nov. 1965.

[102] L. F. Kozachenko and N. N. Leonenko. Sample estimate of the entropy of a random vector. *Probl. Peredachi Inf.*, 23(9), 1987.

[103] J. Kraml, F. Hofer, P. K. Quoika, A. S. Kamenik, and K. R. Liedl. X-Entropy: A parallelized kernel density estimator with automated bandwidth selection to calculate entropy. *J. Chem. Inf. Model.*, Mar. 2021.

[104] A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information. *Physical Review E*, 69(6), Jun 2004.

[105] H. Kruse, P. Banáš, and J. Šponer. Investigations of stacked DNA base-pair steps: Highly accurate stacking interaction energies, energy decomposition, and many-body stacking effects. *J. Chem. Theory Comput.*, 15(1):95–115, Jan. 2019.

[106] 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. *J. Comput. Chem.*, 13(8):1011–1021, 1992.

[107] T. A. Kunkel and K. Bebenek. DNA replication fidelity. *Annu. Rev. Biochem.*, 69(1):497–529, June 2000.

[108] J. Kästner. Umbrella sampling. *WIREs Comput Mol Sci*, 1(6):932–942, Nov. 2011.

[109] P. Kührová, V. Mlýnský, M. Zgarbová, M. Krepl, G. Bussi, R. B. Best, M. Otyepka, J. Šponer, and P. Banáš. Improving the performance of the Amber RNA force field by tuning the hydrogen-bonding interactions. *J. Chem. Theory Comput.*, 15(5):3288–3305, May 2019.

[110] Y. K. Law, R. A. Forties, X. Liu, M. G. Poirier, and B. Kohler. Sequence-dependent thymine dimer formation and photoreversal rates in double-stranded DNA. *Photochemical & photobiological sciences : Official journal of the European Photochemistry Association and the European Society for Photobiology*, 12(8):1431–1439, Aug. 2013.

[111] J.-H. Lee, C.-J. Park, J.-S. Shin, T. Ikegami, H. Akutsu, and B.-S. Choi. NMR structure of the DNA decamer duplex containing double T*G mismatches of cis-syn cyclobutane pyrimidine dimer: Implications for DNA damage recognition by the XPC-hHR23B complex. *Nucleic acids research*, 32:2474–81, 2004.

[112] I. N. Levine. *Quantum Chemistry*. Prentice Hall, 5th edition, 1999.

[113] K. Lewis and D. Wuttke. Telomerase and telomere-associated proteins: Structural insights into mechanism and evolution. *Structure*, 20(1):28–39, Jan. 2012.

[114] P. Li and K. M. Merz. MCPB.py: A python based metal center parameter builder. *J. Chem. Inf. Model.*, 56(4):599–604, Apr. 2016.

[115] S. Li, L. Hu, L. Peng, W. Yang, and F. L. Gu. Coupled-perturbed SCF approach for calculating static polarizabilities and hyperpolarizabilities with nonorthogonal localized molecular orbitals. *J. Chem. Theory Comput.*, 11(3):923–931, Mar. 2015.

[116] J. A. Lippke, L. K. Gordon, D. E. Brash, and W. A. Haseltine. Distribution of UV light-induced damage in a defined sequence of human DNA: Detection of alkaline-sensitive lesions at pyrimidine nucleoside-cytidine sequences. *Proc Natl Acad Sci USA*, 78(6):3388–, June 1981.

[117] H. J. Lipps and D. Rhodes. G-quadruplex structures: In vivo evidence and function. *Trends in cell biology*, 19(8):414–422, 2009.

[118] C. Lu, N. E. Gutierrez-Bayona, and J.-S. Taylor. The effect of flanking bases on direct and triplet sensitized cyclobutane pyrimidine dimer formation in DNA depends on the dipyrimidine, wavelength and the photosensitizer. *Nucleic Acids Res*, 49(8):4266–4280, May 2021.

[119] Madprime. https://en.wikipedia.org/wiki/dna. Accessed: 2021-05-15.

[120] C. Maffeo, H.-Y. Chou, and A. Aksimentiev. Molecular mechanisms of DNA replication and repair machinery: Insights from microscopic simulations. *Adv. Theory Simul.*, 2(5):1800191, May 2019.

[121] C. Maffeo, B. Luan, and A. Aksimentiev. End-to-end attraction of duplex DNA. *Nucleic Acids Res*, 40(9):3812–3821, May 2012.

[122] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, and C. Simmerling. ff14sb: Improving the accuracy of protein side chain and backbone parameters from ff99sb. *J. Chem. Theory Comput.*, 11(8):3696–3713, Aug. 2015.

[123] B. Maji and S. Bhattacharya. Advances in the molecular design of potential anti-cancer agents via targeting of human telomeric DNA. *Chemical Communications*, 50(49):6422–6438, 2014.

[124] W. Makulski and K. Jackowski. The ^{17}O nuclear magnetic shielding scale from gas-phase measurements. *Journal of Molecular Structure*, 651-653:265–269, 2003.

[125] P. Mao, M. J. Smerdon, S. A. Roberts, and J. J. Wyrick. Chromosomal landscape of UV damage formation and repair at single-nucleotide resolution. *Proc Natl Acad Sci USA*, 113(32):9057, Aug. 2016.

[126] P. Mao, J. J. Wyrick, S. A. Roberts, and M. J. Smerdon. UV-induced DNA damage and mutagenesis in chromatin. *Photochem Photobiol*, 93(1):216–228, Jan. 2017.

[127] R. Marek, J. Brus, J. Toušek, L. Kovács, and D. Hocková. N^7- and N^9-substituted purine derivatives: A ^{15}N NMR study. *Magn. Reson. Chem.*, 40(5):353–360, May 2002.

[128] D. Markovitsi, T. Gustavsson, and A. Banyasz. Absorption of UV radiation by DNA: Spatial and temporal features. *Mutation Research/Reviews in Mutation Research*, 704(1):21–28, 2010.

[129] M. A. L. Marques, C. A. Ullrich, F. Nogueira, A. Rubio, K. Burke, and E. K. U. Gross. *Time-dependent density functional theory.* Lecture notes in physics 706. Springer, 1st edition, 2006.

[130] J. A. Marteijn, H. Lans, W. Vermeulen, and J. H. J. Hoeijmakers. Understanding nucleotide excision repair and its roles in cancer and ageing. *Nature reviews. Molecular cell biology*, 15:465–81, Jul 2014.

[131] L. Martinez-Fernandez, A. Banyasz, L. Esposito, D. Markovitsi, and R. Improta. UV-induced damage to DNA: Effect of cytosine methylation on pyrimidine dimerization. *Signal Transduction and Targeted Therapy*, 2(1):17021–, 2017.

[132] R. K. McGinty and S. Tan. Nucleosome structure and function. *Chem. Rev.*, 115(6):2255–2273, Mar. 2015.

[133] A. Mees, T. Klar, P. Gnau, U. Hennecke, A. P. M. Eker, T. Carell, and L.-O. Essen. Crystal structure of a photolyase bound to a CPD-like DNA lesion after in situ repair. *Science*, 306(5702):1789, Dec. 2004.

[134] J. Meller. *Molecular Dynamics.* American Cancer Society, 2001.

[135] A. C. Messias and M. Sattler. Structural basis of single-stranded RNA recognition. *Acc. Chem. Res.*, 37(5):279–287, May 2004.

[136] K. Michaeli, D. N. Beratan, D. H. Waldeck, and R. Naaman. Voltage-induced long-range coherent electron transfer through organic molecules. *Proc Natl Acad Sci USA*, 116(13):5931–, Mar. 2019.

[137] C. T. Middleton, K. de La Harpe, C. Su, Y. K. Law, C. E. Crespo-Hernández, and B. Kohler. DNA excited-state dynamics: From single bases to the double helix. *Annu. Rev. Phys. Chem.*, 60(1):217–239, Mar. 2009.

[138] B. R. Miller, T. D. McGee, J. M. Swails, N. Homeyer, H. Gohlke, and A. E. Roitberg. MMPBSA.py: An efficient program for end-state free energy calculations. *J. Chem. Theory Comput.*, 8(9):3314–3321, Sept. 2012.

[139] B. Montelone. DNA repair and mutagenesis. *The Quarterly Review of Biology*, 81(3):273–273, Sept. 2006.

[140] F. Moraca, J. Amato, F. Ortuso, A. Artese, B. Pagano, E. Novellino, S. Alcaro, M. Parrinello, and V. Limongelli. Ligand binding to telomeric G-quadruplex DNA investigated by funnel-metadynamics simulations. *Proceedings of the National Academy of Sciences*, 114(11):E2136–E2145, 2017.

[141] T. Morishita. Fluctuation formulas in molecular-dynamics simulations with the weak coupling heat bath. *J. Chem. Phys.*, 113(8):2976–2982, Aug. 2000.

[142] S.-I. Moriwaki and K. H. Kraemer. Xeroderma pigmentosum - bridging a gap between clinic and laboratory. *Photodermatology, Photoimmunology & Photomedicine*, 17(2):47–54, Apr. 2001.

[143] R. Naaman and D. H. Waldeck. Spintronics and chirality: Spin selectivity in electron transport through chiral molecules. *Annu. Rev. Phys. Chem.*, 66(1):263–281, Apr. 2015.

[144] A. Nakamura, C. Wada, and K. Miki. Structural basis for regulation of bifunctional roles in replication initiator protein. *Proc Natl Acad Sci USA*, 104(47):18484–, Nov. 2007.

[145] A. Nayis, K. Liebl, C. V. Frost, and M. Zacharias. Targeting telomeres: Molecular dynamics and free energy simulation of gold-carbene binding to DNA. *Biophysical Journal*, 120(1):101–108, Jan. 2021.

[146] F. Neese. Efficient and accurate approximations to the molecular spin-orbit coupling operator and their use in molecular g-tensor calculations. *J. Chem. Phys.*, 122(3):034107–, Jan. 2005.

[147] F. Neese. The ORCA program system. *WIREs Comput Mol Sci*, 2(1):73–78, Jan. 2012.

[148] S. Neidle. Quadruplex nucleic acids as novel therapeutic targets. *Journal of medicinal chemistry*, 59(13):5987–6011, 2016.

[149] F. Noé, H. Wu, J.-H. Prinz, and N. Plattner. Projected and hidden Markov models for calculating kinetics and metastable states of complex molecules. *J. Chem. Phys.*, 139(18):184114, Nov. 2013.

[150] J. Numata, M. Wan, and E.-W. Knapp. Conformational entropy of biomolecules: Beyond the quasi-harmonic approximation. *Genome Informatics*, 18:192–205, 2007.

[151] D. L. Ollis, P. Brick, R. Hamlin, N. G. Xuong, and T. A. Steitz. Structure of large fragment of Escherichia coli DNA polymerase I complexed with dTMP. *Nature*, 313(6005):762–766, Feb. 1985.

[152] A. Onufriev, D. Bashford, and D. A. Case. Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins*, 55(2):383–394, May 2004.

[153] D. A. Pantazis, X.-Y. Chen, C. R. Landis, and F. Neese. All-electron scalar relativistic basis sets for third-row transition metal atoms. *J. Chem. Theory Comput.*, 4(6):908–919, June 2008.

[154] F. Papi, C. Bazzicalupi, M. Ferraroni, L. Massai, B. Bertrand, P. Gratteri, D. Colangelo, and L. Messori. [Au(9-methylcaffein-8-ylidene)2]+/DNA Tel23 system: Solution, computational, and biological studies. *Chem. Eur. J.*, 23(55):13784–13791, Oct. 2017.

[155] H. Park, K. Zhang, Y. Ren, S. Nadji, N. Sinha, J.-S. Taylor, and C. Kang. Crystal structure of a DNA decamer containing a cis-syn thymine dimer. *Proc Natl Acad Sci USA*, 99(25):15965, Dec. 2002.

[156] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, Aug. 2005.

[157] E. Peter, B. Kieron, and F. Filipp. *Excited states from time-dependent density functional theory*, chapter 3, pages 91–. John Wiley & Sons, 51 edition, 2009.

[158] L. Piela. *Ideas of quantum chemistry*. Elsevier, 1st edition, 2007.

[159] E. D. Pleasance, R. K. Cheetham, P. J. Stephens, D. J. McBride, S. J. Humphray, C. D. Greenman, I. Varela, M.-L. Lin, G. R. Ordóñez, G. R. Bignell, K. Ye, J. Alipaz, M. J. Bauer, D. Beare, A. Butler, R. J. Carter, L. Chen, A. J. Cox, S. Edkins, P. I. Kokko-Gonzales, N. A. Gormley, R. J. Grocock, C. D. Haudenschild, M. M. Hims, T. James, M. Jia, Z. Kingsbury, C. Leroy, J. Marshall, A. Menzies, L. J. Mudie, Z. Ning, T. Royce, O. B. Schulz-Trieglaff, A. Spiridou, L. A. Stebbings, L. Szajkowski, J. Teague, D. Williamson, L. Chin, M. T. Ross, P. J. Campbell, D. R. Bentley, P. A. Futreal, and M. R. Stratton. A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*, 463(7278):191–196, 2010.

[160] P. Polak, M. S. Lawrence, E. Haugen, N. Stoletzki, P. Stojanov, R. E. Thurman, L. A. Garraway, S. Mirkin, G. Getz, J. A. Stamatoyannopoulos, and S. R. Sunyaev. Reduced local mutation density in regulatory DNA of cancer genomes is linked to DNA repair. *Nature biotechnology*, 32(1):71–75, Jan. 2014.

[161] A. A. Polyansky, A. Kuzmanic, M. Hlevnjak, and B. Zagrovic. On the contribution of linear correlations to quasi-harmonic conformational entropy in proteins. *J. Chem. Theory Comput.*, 8(10):3820–3829, Oct. 2012.

[162] D. Porath, A. Bezryadin, S. de Vries, and C. Dekker. Direct measurement of electrical transport through DNA molecules. *Nature*, 403(6770):635–638, 2000.

[163] L. Pronzato, E. Thierry, and E. Wolsztynski. Minimum entropy estimation in semi-parametric models: A candidate for adaptive estimation? In *mODa 7 – Advances in Model-Oriented Design and Analysis*, pages 125–132, Heidelberg, 2004. Physica-Verlag HD.

[164] A. Pérez, M. Luque FJ FAU Orozco, and O. M. Frontiers in molecular dynamics simulations of DNA. *Acc. Chem. Res*, 45(2):196–205, 2012.

[165] R. Rai, H. Zheng, H. He, Y. Luo, A. Multani, P. B. Carpenter, and S. Chang. The function of classical and alternative non-homologous end-joining pathways in the fusion of dysfunctional telomeres. *EMBO J.*, 29(15):2598–2610, Aug. 2010.

[166] M. Rajewska, K. Wegrzyn, and I. Konieczny. At-rich region and repeated sequences - the essential elements of replication origins of bacterial replicons. *FEMS Microbiol Rev*, 36(2):408–434, Mar. 2012.

[167] M. T. Record, T. M. Lohman, and P. d. Haseth. Ion effects on ligand-nucleic acid interactions. *Journal of Molecular Biology*, 107(2):145–158, 1976.

[168] J. E. Reed, A. A. Arnal, S. Neidle, and R. Vilar. Stabilization of G-Quadruplex DNA and inhibition of telomerase activity by square-planar Nickel(II) complexes. *J. Am. Chem. Soc.*, 128(18):5992–5993, May 2006.

[169] J. Reinhold. Quantentheorie der Moleküle. *Teubner*, 2006.

[170] N. Relitti, A. P. Saraswati, S. Federico, T. Khan, M. Brindisi, D. Zisterer, S. Brogi, S. Gemma, S. Butini, and G. Campiani. Telomerase-based cancer therapeutics: A review on their clinical trials. *Current Topics in Medicinal Chemistry*, 20(6):433–457, 2020.

[171] 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(5):053020, 2012.

[172] B. Riccardo. *Biomolecular simulation: Calculation of entropy and free energy, polypeptide and carbopeptoid folding, simplification of the force field for lipid simulations*. ETH Zurich, 2006.

[173] C. M. Roake and S. E. Artandi. Regulation of human telomerase in homeostasis and disease. *Nature Reviews Molecular Cell Biology*, pages 1–14, 2020.

[174] P. J. Rochette, J.-P. Therrien, R. Drouin, D. Perdiz, N. Bastien, E. A. Drobetsky, and E. Sage. UVA-induced cyclobutane pyrimidine dimers form predominantly at thymine-thymine dipyrimidines and correlate with the mutation spectrum in rodent cells. *Nucleic acids research*, 31(11):2786–2794, June 2003.

[175] 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(7268):1248–1253, 2009.

[176] B. Roux. The calculation of the potential of mean force using computer simulations. *Computer Physics Communications*, 91(1):275–282, 1995.

[177] E. Runge and E. K. U. Gross. Density-functional theory for time-dependent systems. *Phys. Rev. Lett.*, 52(12):997–1000, Mar. 1984.

[178] B. Sandhoefer and F. Neese. One-electron contributions to the g-tensor for second-order Douglas-Kroll-Hess theory. *J. Chem. Phys.*, 137(9):094102–, Sept. 2012.

[179] J. M. Seminario. Calculation of intramolecular force fields from second-derivative tensors. *Int. J. Quantum Chem.*, 60(7):1271–1277, Jan. 1996.

[180] J. Shay and S. Bacchetti. A survey of telomerase activity in human cancer. *European Journal of Cancer*, 33(5):787–791, Apr. 1997.

[181] P. G. Shiels, A. J. Kind, K. H. Campbell, I. Wilmut, D. Waddington, A. Colman, and A. E. Schnieke. Analysis of telomere length in dolly, a sheep derived by nuclear transfer. *Cloning*, 1(2):119–125, June 1999.

[182] H. Singh, N. Misra, V. Hnizdo, A. Fedorowicz, and E. Demchuk. Nearest neighbor estimates of entropy. *American Journal of Mathematical and Management Sciences*, 23(3-4):301–321, 2003.

[183] F. Sittel, A. Jain, and G. Stock. Principal component analysis of molecular dynamics: On the use of cartesian vs. internal coordinates. *J. Chem. Phys.*, 141(1):014111–, July 2014.

[184] B. D. Smith, G. L. Smith, A. Hurria, G. N. Hortobagyi, and T. A. Buchholz. Future of cancer incidence in the united states: Burdens upon an aging, changing nation. *JCO*, 27(17):2758–2765, Apr. 2009.

[185] M. Souaille and B. Roux. Extension to the weighted histogram analysis method: Combining umbrella sampling with free energy calculations. *Computer Physics Communications*, 135(1):40–57, 2001.

[186] J. Šponer, G. Bussi, P. Stadlbauer, P. Kührová, P. Banáš, B. Islam, S. Haider, S. Neidle, and M. Otyepka. Folding of guanine quadruplex molecules–funnel-like mechanism or kinetic partitioning? An overview from MD simulation studies. *Biochimica et Biophysica Acta (BBA)-General Subjects*, 1861(5):1246–1263, 2017.

[187] J. Šponer, X. Cang, and T. E. Cheatham III. Molecular dynamics simulations of G-DNA and perspectives on the simulation of nucleic acid structures. *Methods*, 57(1):25–39, 2012.

[188] J. Srinivasan, T. E. Cheatham, P. Cieplak, P. A. Kollman, and D. A. Case. Continuum solvent studies of the stability of DNA, RNA, and Phosphoramidate-DNA helices. *J. Am. Chem. Soc.*, 120(37):9401–9409, Sept. 1998.

[189] E. Suárez, N. Díaz, and D. Suárez. Entropy calculations of single molecules by combining the rigid-rotor and harmonic-oscillator approximations with conformational entropy estimations from molecular dynamics simulations. *J. Chem. Theory Comput.*, 7(8):2638–2653, Aug. 2011.

[190] W. C. Swope, H. C. Andersen, P. H. Berens, and K. R. Wilson. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *J. Chem. Phys.*, 76(1):637–649, Jan. 1982.

[191] M. E. Tuckerman. Statistical mechanics: Theory and molecular simulation. *Oxford University Press*, pages 74–132, 2010.

[192] W. Van den Heuvel and A. Soncini. NMR chemical shift as analytical derivative of the Helmholtz free energy. *J. Chem. Phys.*, 138(5):054113–, Feb. 2013.

[193] Y. Velmurugu, X. Chen, P. Slogoff Sevilla, J.-H. Min, and A. Ansari. Twist-open mechanism of DNA damage recognition by the Rad4/XPC nucleotide excision repair complex. *Proc Natl Acad Sci USA*, 113(16):E2296, Apr. 2016.

[194] G. Ver Steeg. https://github.com/gregversteeg/npeet. Accessed: 2021-06-20.

[195] P. H. von Hippel and O. G. Berg. On the specificity of DNA-protein interactions. *Proc Natl Acad Sci USA*, 83(6):1608–1612, Mar. 1986.

[196] C. Wang, D. Greene, L. Xiao, R. Qi, and R. Luo. Recent developments and applications of the MMPBSA method. *Frontiers in Molecular Biosciences*, 4:87, 2018.

[197] C. I. Wang and J. S. Taylor. In vitro evidence that UV-induced frameshift and substitution mutations at T tracts are the result of misalignment-mediated replication past a specific thymine dimer. *Biochemistry*, 31(14):3671–3681, Apr. 1992.

[198] J. D. WATSON and F. H. C. CRICK. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, 1953.

[199] K. Wegrzyn, M. E. Fuentes-Perez, K. Bury, M. Rajewska, F. Moreno-Herrero, and I. Konieczny. Sequence-specific interactions of Rep proteins with ssDNA in the AT-rich region of the plasmid replication origin. *Nucleic acids research*, 42(12):7807–7818, July 2014.

[200] F. Weigend. Accurate coulomb-fitting basis sets for H to Rn. *Phys. Chem. Chem. Phys.*, 8(9):1057–1065, 2006.

[201] F. Weigend. Hartree-fock exchange fitting basis sets for H to Rn. *J. Comput. Chem.*, 29(2):167–175, Jan. 2008.

[202] F. Weigend and R. Ahlrichs. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.*, 7(18):3297–3305, 2005.

[203] M. C. White, D. M. Holman, J. E. Boehm, L. A. Peipins, M. Grossman, and S. Jane Henley. Age and cancer risk: A potentially modifiable relationship. *American Journal of Preventive Medicine*, 46(3, Supplement 1):S7–S15, 2014.

[204] M. Witanowski, L. Stefaniak, and G. Webb. Nitrogen NMR spectroscopy. In G. Webb, editor, *Annual Reports on NMR Spectroscopy*, volume 18, pages 1–211. Academic Press, 1987.

[205] K. Wolinski, J. F. Hinton, and P. Pulay. Efficient implementation of the gauge-independent atomic orbital method for NMR chemical shift calculations. *J. Am. Chem. Soc.*, 112(23):8251–8260, Nov. 1990.

[206] H.-J. Woo and B. Roux. Calculation of absolute protein-ligand binding free energy from computer simulations. *Proc Natl Acad Sci USA*, 102(19):6825–, May 2005.

[207] W. E. Wright, M. A. Piatyszek, W. E. Rainey, W. Byrd, and J. W. Shay. Telomerase activity in human germline and embryonic tissues and cells. *Dev. Genet.*, 18(2):173–179, Jan. 1996.

[208] G. Wu, S. Dong, R. Ida, and N. Reen. A solid-state ^{17}O Nuclear Magnetic Resonance study of nucleic acid bases. *J. Am. Chem. Soc.*, 124(8):1768–1777, Feb. 2002.

[209] L. Wu, A. S. Multani, H. He, W. Cosme-Blanco, Y. Deng, J. M. Deng, O. Bachilo, S. Pathak, H. Tahara, S. M. Bailey, Y. Deng, R. R. Behringer, and S. Chang. Pot1 deficiency initiates DNA damage checkpoint activation and aberrant homologous recombination at telomeres. *Cell*, 126(1):49–62, July 2006.

[210] Y. Xiang, G. Zheng, Z. Liang, Y. Jin, X. Liu, S. Chen, K. Zhou, J. Zhu, M. Lin, H. He, J. Wan, S. Yu, G. Zhong, R. Fu, Y. Li, and Y. Yang. Visualizing the growth process of sodium microstructures in sodium batteries by in-situ ^{23}Na MRI and NMR spectroscopy. *Nature Nanotechnology*, 15(10):883–890, 2020.

[211] E. Zabrocka, K. Wegrzyn, and I. Konieczny. Two replication initiators - one mechanism for replication origin opening? *Plasmid*, 76:72–78, 2014.

[212] M. Zhang, L. Wang, and D. Zhong. Photolyase: Dynamics and electron-transfer mechanisms of DNA repair. *Archives of Biochemistry and Biophysics*, 632:158–174, 2017.

[213] Y. Zhao and D. G. Truhlar. A new local density functional for main-group thermochemistry, transition metal bonding, thermochemical kinetics, and noncovalent interactions. *J. Chem. Phys.*, 125(19), Nov. 2006.

[214] D. Zhong. Electron transfer mechanisms of DNA repair by photolyase. *Annu. Rev. Phys. Chem.*, 66(1):691–715, Apr. 2015.

[215] T. J. Zwang, S. Hürlimann, M. G. Hill, and J. K. Barton. Helix-dependent spin filtering through the DNA duplex. *J. Am. Chem. Soc.*, 138(48):15551–15554, Dec. 2016.

[216] R. W. Zwanzig. High-temperature equation of state by a perturbation method. II. Polar gases. *The Journal of Chemical Physics*, 23(10):1915–1922, 1955.

[217] J. Šponer, G. Bussi, M. Krepl, P. Banáš, S. Bottaro, R. A. Cunha, A. Gil-Ley, G. Pinamonti, S. Poblete, P. Jurečka, N. G. Walter, and M. Otyepka. RNA structural dynamics as captured by molecular simulations: A comprehensive overview. *Chem. Rev.*, 118(8):4177–4338, Apr. 2018.

[218] J. Šponer, B. Islam, P. Stadlbauer, and S. Haider. Molecular dynamics simulations of G-quadruplexes: The basic principles and their application to folding and ligand binding. *Annual Reports in Medicinal Chemistry*, 54:197–241, 2020.

[219] J. Šponer, J. Leszczyński, and P. Hobza. Nature of nucleic acid-base stacking: Nonempirical ab initio and empirical potential characterization of 10 stacked base dimers. Comparison of stacked and H-bonded base pairs. *J. Phys. Chem.*, 100(13):5590–5596, Jan. 1996.

[220] J. Šponer and N. Špačková. Molecular dynamics simulations and their application to four-stranded DNA. *Methods*, 43(4):278–290, 2007.