Long Range Communication In Macromolecular Systems

im Fachbereich Physik der Freien Universität Berlin eingereichte Dissertation

zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften (Dr. rer. nat.)

vorgelegt von
Mahdi Bagherpoor Helabad, Iran

Berlin: 2020

Erster Gutachter: Prof. Dr. Petra Imhof

Zweiter Gutachter: Prof. Dr. Roland Netz

Tag der Disputation: 01.09.2020

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Molecular Modeling: A Method for Studying Complex Systems

Physical and chemical processes take place on various time and length scales. Thus, a single cell, which is a fundamental unit of life, is quite complex. As an example, it is estimated that each μm^3 unit volume of cells include roughly about $\sim 10^6$ proteins [1] in which each protein involves multiple intra- and inter-molecular (with other proteins, ligands, DNA, etc.) interactions. To tackle this complexity, the efforts of different disciplines of studies, i.e. Physics, Biology, Mathematics, Chemistry, and Computer science are required; in both experimental and theoretical perspectives. Molecular modeling, a rather young and strong scientific discipline [2] intends to study different aspects of molecular structures and functions by using computational methods. Depending on the system/question of interest, different methods can be used. These methods can vary from the level of quantum descriptions to molecular dynamics simulations and course grain descriptions or continuum models. The level of accuracy (resolution) and computational cost (CPU time plus storage requirement) are the factors that depend on the system's (biomolecule) characteristic size and the time scales (Figure 1.1) that need to be considered. The present thesis employs molecular dynamic (MD) simulations in atomic detail [3, 4], as an appropriate method for studying the nanoscale dynamical properties of macromolecules.

In order to function properly, different parts/sub-domains of macromolecules need to cooperate [5]. This cooperation takes place through the spacial arrangement of the (groups of) atoms, which is referred to as conformational changes [6]. These conformational changes are critical in biological processes that take place through different courses of actions such as allostery, protein folding, protein-DNA interaction, ligand binding, and enzymatic catalysis.

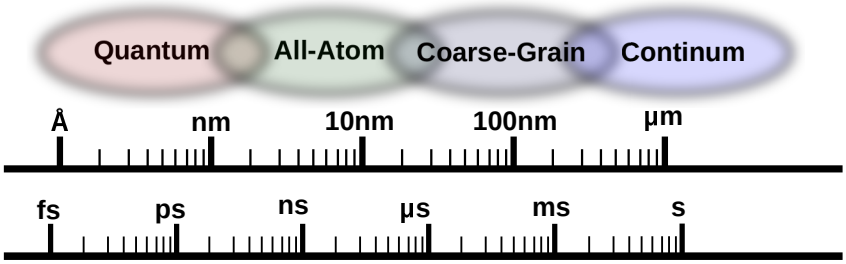


Figure 1.1: computational level of the resolution

Throughout the decades, experimental efforts such as X-ray crystallography [7, 8], nuclear magnetic resonance spectroscopy (NMR) [9], and more recently Cryo-electron microscopy [10] augmented our understanding on the biomolecular atomic-level structures and functions. Nevertheless, the dynamics of conformational changes in biological macromolecules are difficult to capture by these methods. Molecular dynamics (MD) simulations nowadays are a strong and appropriate technique that allows us to fill the aforementioned gap in understanding of biomolecular action. One of the main challenges in studying dynamical and or conformational properties of the biomolecular systems is the presence of long-range communicationin macromolecules [11–13]. Therefore, it is vital to characterize the conformational impacts that such allosteric communication can have.

1.2 Studied Systems

In this thesis, we mainly focus on two different macromolecular systems. Each of these systems belongs to a member of protein families that are crucial in cell processes. The first belongs to a group of proteins, termed transcription factors (TFs), which comprise about one tenth of the proteins, which are produced by the human genome [14]. These proteins interact with optimal local DNA sequences, called specific DNA sequence or response elements and thereby form suitable DNA-protein complexes that regulate gene expression processes in cell [15]. The second system belongs to another kind of proteins that interact with or are part of the biological cell, termed membrane proteins, which comprise about 30% of the proteins in eukaryotic cells [16]. Membrane proteins play various functional roles in the cells such as charge transfer in and through the membrane, ion or molecule transport, signal transduction, and many more [17, 18].

DNA

The discovery of the Deoxyribonucleic acid (DNA) structure by Watson and Crick in 1953 significantly illuminated our knowledge about the essence of life [19]. A DNA molecule is a double-stranded and directed molecule that contains biological instructions, and/or information, which are essential in growth and development of living organisms [20, 21]. The normal DNA structure is a ladder-like flexible molecule that is wrapped around a central imaginary axis, called helical axis. The DNA molecule includes four unique building blocks, called nucleotides. Each nucleotide is formed of three individual groups: phosphate, sugar, and base. The phosphate and sugar groups both constitute the DNA backbone while the base groups make steps of the DNA ladder. DNA strands are anti-parallel, which means the $5'$ end (refers to the $5'$ carbon on the sugar) of one strand pairs with the $3'$ end (refers to the $3'$ carbon on the sugar) of the other strand and vice versa. The bases pair in a complementary manner, i.e. Adenine (A) base always pairs with Thymine (T) and Guanine (G) always pairs with Cytosine (C). The A-T and G-C pairs, which are referred to as base pairs (bps), are held together via two and three suitable hydrogen bond interactions, respectively. The diameter and successive bps distances in a DNA molecule are 2 *nm* and 0.34 *nm*, respectively. The pair of strands in a DNA molecule forms a minor groove and a major groove in its surface, which are essential in DNA-protein interactions (Fig.1.2).

Proteins

To properly function, a cell needs an important class of molecules, called "protein" [20]. All the proteins in nature are constructed from a set of twenty unique building-blocks, named amino acids. Depending on the amino acid sequence, proteins adopt a specific three-dimensional structure that significantly directs their function. It is worth mentioning that only a tiny range of possible amino acid sequences actually forms a protein [22]. In general, an amino acid consists of an amino group, a carboxylate group, a side chain, and a hydrogen atom, shown in Fig.1.3. The side chain group differing among amino acids determines its specific physical and chemical characteristic properties. Proteins can interact with other proteins or with other macromolecules to form a complex system. For instance, the gene transcription process is facilitated by the assembly of proteins that interact among themselves and with short DNA sequences in the promoter region in order to initiate transcription.

1.2.1 Cytochrome c oxidase (CcO)

Dioxygen (O_2) is a key component in nature. The majority of O_2 consumption in the living organisms is catalyzed by a family of enzymes, termed heme-copper oxidases [23]. Cytochrome

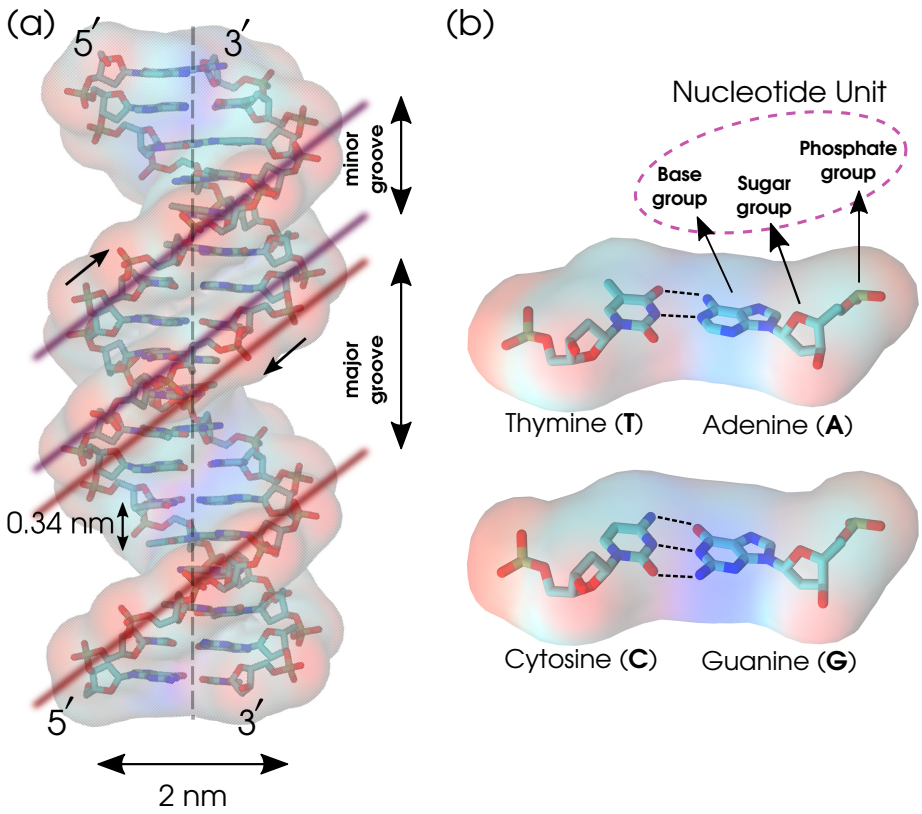


Figure 1.2: Double strand structure of DNA. (a) Double stand wraps around an imagine central helical axis line, shown as dashed line. (b) A-T and G-C pair-nucleotides form unique base pairs (bps).

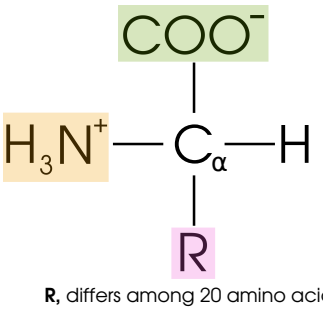


Figure 1.3: General amino acid structure. Each amino acid holds a unique side chain group R.

c Oxidase (CcO) is the terminal enzyme in the respiratory chain of mitochondria and other aerobic prokaryotics that reduce molecular oxygen (O_2) to water in a reaction that is coupled to a proton transfer process [24]. In the course of this reaction, four electrons, which are taken up from the inter-membrane space (called p-side) and four protons taken from the inner membrane side (called N-side) transfer to the binuclear center, where an oxygen molecule is bound, and there reduce the oxygen molecule to water. In addition, four protons are transferred across the membrane, from N-side to P-side, resulting in an electrochemical gradient that drives the ATP synthesis [25, 26]. Therefore, the overall chemical reaction catalyzed by CcO (A-type) can be written as:

$$O_2 + 4e_P^- + 8H_N^+ \rightarrow 2H_2O + 4H_P^+$$

In this chemical equation, e^- and H^+ mean electron and proton, respectively. Index N and P on the left-hand side indicate the phase of electron and proton from N-side and P-side, respectively, whereas index P on the right-hand side shows that protons are pumped across the membrane.

Until now, many crystal structures of CcO have been resolved [27–31]. These structures reveal two proton-conducting channels, termed the D-channel and K-channel, shown in Fig. 1.4a. The D-channel is responsible for transferring both the chemical and pumped protons, whereas the K-channel transfers only chemical protons. It is suggested that the K-channel is used to transfer one or two protons, whereas six or seven protons are transferred through the D-channel [32]. It is important to note that the proton transfer process is strongly coupled to the electron transfer process. This can obviously be seen in the catalytic cycle of the CcO Fig. 1.4b [33, 34]. In this thesis we mainly discuss the $P_R \rightarrow F$ state. This is the first step in a catalytic cycle that requires proton transfer from the bulk at the N-side to the BNC. Although many studies have been made on the various functional and structural aspects of CcO (see review articles [24, 35]), the mechanism of the proton transfer process and how the different steps in the catalytic cycle are regulated is not yet fully understood. Therefore, obtaining information on the molecular level is importantly required.

1.2.2 Steroid Receptors

Steroid receptors (SRs) are ligand-activated transcription factors that are a subfamily of the nuclear receptor superfamily and play a crucial role for a range of physiological and developmental processes as well as the development of various diseases such as cancer, metabolic diseases, and genetic disorders [36–38]. Members of the SRs family are estrogen receptor (ER), androgen receptors (AR), glucocorticoid receptor (GR), progesterone receptor (PR), and mineralocorticoid receptor (MR) [36, 39]. A schematic primary structure of the SRs is shown

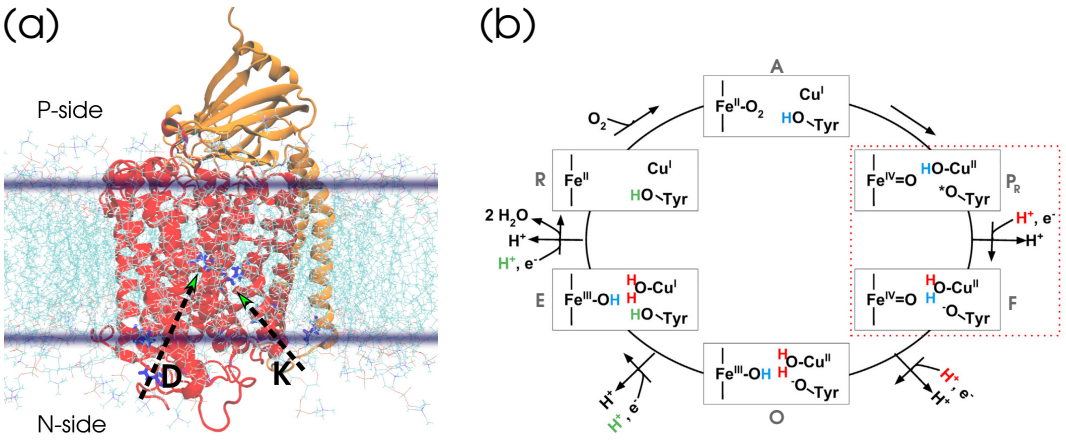


Figure 1.4: (a) Structure of CcO. Arrows in dashed line indicate the D- and K-channel pathways. (b) Catalytic cycle of CcO. In this study we mainly focus on $P_R \rightarrow F$ state, shown in the dashed red box.

in Fig. 1.5a. SRs are structurally composed of three major domains: an amino-terminal domain (NTD), an almost strictly conserved zinc-finger DNA-binding domain (DBD), carboxyl-terminal ligand-binding domain (LBD), and a flexible hinge region that connect DBD to LBD [39, 40]. Until now, the separate DBD and LBD structures of the SRs have been resolved, Fig. 1.5a, but the entire three-dimensional atomistic structures of the proteins are not yet resolved. SRs bind DNA as a homodimer [36]. The DBD, which includes about 70 amino acid residues, functionally contains two vital subdomains, each identified with a zinc ion that is coordinated by four Cysteine residues. The first subdomain includes an α-helix, which is responsible for protein-DNA major groove interaction, while the second subdomain holds a loop domain, termed Dim, which is responsible for protein-protein dimerization [41, 42]. This homodimer structure is shown schematically in Fig. 1.5b. A flexible loop, named "lever arm" (Fig. 1.5b, colored in yellow) connects these subdomains to each other. The core DNA sequence, which is specifically recognized by SRs-DBD includes 15 precise base pairs (bps) composed of two hexameric half-sites (HS1 and HS2, respectively) separated by three bps called spacer (Fig. 1.5b,c). This spacer region allows the protein helices sufficient conformational freedom to make high affinity major groove interactions while the dimerization DBD is preserved intact upon DNA binding [43]. The members of the SRs bind to a group of identical response elements, named classical response elements (CREs), which resemble an inverted repeat (IR) of hexamer AGAACA; i.e. AGAACA*nnn*TGTTCT [44]. The letters *nnn* indicate variable spacer sequences. Next to CREs, there is another kind of elements, named androgen response elements (AREs), which are merely recognized by the AR. The elements of the AREs are organized as a direct repeat (DR); i.e. AGAACA*nnn*AGAACA (Fig. 1.5c) [45]. The question why despite similar structure of SRs-DBD do not form stable complex with CREs is still a matter

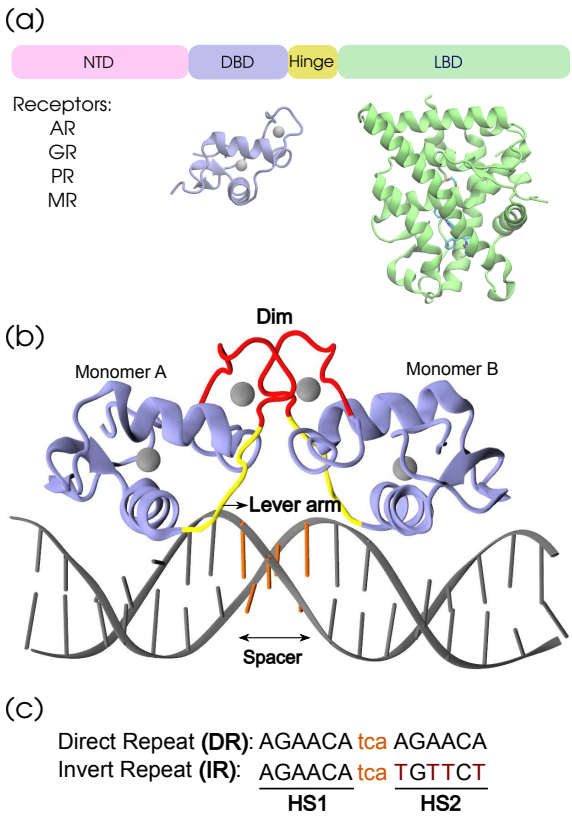


Figure 1.5: (a) Schematic primary structure of steroid receptors. Three-dimensional structure of DBD and LBD, which are crystallized for the GR are shown. (b) The 3D structure of the GR-DBD/DNA complex (pdb ID: 1R4R). A similar structure exists for other SRs-DBD/DNA complexes. The lever arm and dimerization domain (Dim) are shown in yellow and red, respectively. The spacer region of the DNA is colored with orange. (c) DNA sequences for direct (DR) and inverted repeats (IR) are shown. The non-capital letters are the spacer base-pairs, colored in orange.

of debate. It is important that SRs have different functions in vivo, which could be because of differential expression of SRs-interacting proteins or due to different ligand metabolism [46].

Recently, a chimeric AR, termed SPARKI, has been generated by transgenic knock-in mouse model where the second DBD zinc binding subdomain is replaced with that of the GR [47]. It is shown that transgenic SPARKI mouse developes smaller reproductive organs with subfertility and a different gene expression profile than wild-type AR littermates [47, 48]. Studies of this chimeric protein model provide valuable information about differential DNA recognition of SPARKI-AR with AR and GR and therefore between AR and GR.

1.3 Approaches and motivations

The dynamical events at the molecular level are crucial and have a predominant role in biomolecule's function and regulation. Characterizing such dynamical events not only broadens our understanding of the intrinsic conformational nature of the biomolecules but also allows us to identify their possible interactions and communication with other biomolecules. The experimental methods which have been used for characterizing the structure of biomolecules such as X-ray crystallography, however, provide only a static picture of the biomolecules and no dynamical description. NMR on the other hand provides dynamical information, but only with limited time resolution. These structures also give indirect information about the system of interest and therefore the interpretation of such information is possible only in the light of their underlying theory. Molecular dynamics (MD) simulations, a numerical method for (N-body) simulating of the molecules or atoms, is a powerful tool to investigate the biomolecular dynamics and thereby their underlying mechanisms of action.

One of the significant issues in studying the dynamics of the macromolecules is the presence of long-range effects. In other words, the effects that appear between relatively far positioned functional sites of macromolecular systems that can not be described just via interactions that fall into the short-range regime. In this study, we examine the role of long-range effects on the conformational dynamics of two different macromolecular systems.

The first part of this thesis, in chapter three, aims to understand the dynamical behavior of the CcO complex in presence of different protonation states of the key residues in CcO's proton transfer channels, i.e. D- and K-channel, by performing equilibrium molecular dynamics simulations. In this chapter, by analyzing the MD simulations output, we investigate the communication within and between protein residues of the D- and K-channel to elucidate whether and how a change of protonation state of the key residues is sensed by the other key residues. This allows us to capture the interplay among different sub-domains in CcO that determine the proton transfer probability.

In the second part of this thesis (including chapters four, five, and six) the specific protein-DNA interactions of two important members of steroid receptors, i.e. AR and GR, in complex with their specific DNA sequence will be studied. In chapter four, we explore the factors that govern the preferential DNA recognition of the AR/GR DBD bound state, especially addressing the question why despite the structural similarity of the DBDs in AR and GR, the latter is not able to reach a stable bound state with AREs (DR response element). In this regard, the protein-protein and protein-DNA interactions, as well as the long-range communication role in the recognition and specificity of these complexes, will be investigated. In chapter five, we examine the chimeric protein, i.e. SPARKI, and its interactions in complex with both IR and DR response elements. Similar to the GR, the SPARKI also lacks affinity to the DR

response element [47, 49]. Therefore, determining the importance of the AR-like and GR-like segments of this chimeric protein in protein-DNA surface and dimer interface of complex, respectively, enables us to characterize the factors that destabilize both the GR- and SPARKI-DR complexes formation. In chapter six, we study the influence of the flanking nucleotides of the core DNA sequences on the GR DBD-DNA complex. The experimental result show that the structure, as well as the activity of GR complex bound to the specific DNA sequence, significantly depend on the DNA nucleotides flanking the core site [50]. However, the dynamics of DNA-induced conformational changes in this complex are largely unknown. By using MD simulationto study the influence of flanking core DNA sequence on the conformation of the GR DBD-DNA complex we gain an understanding of the molecular basis that gives rise to conformational changes of the individual monomers due to flank influences.

Chapter 2

Methods

As outlined in the motivation and introduction, our interest is to simulate the biological systems at an atomistic level close to conditions in real systems. As a result of current computational limitations, in both time and size scale, it is not feasible for most of the cellular processes to deal with atomistic simulation. However, narrowing down to the nano-world resolution, there are many processes in macromolecular systems whose dynamics of action can be studied by using atomistic scale simulations [2]. In order to make the simulations of our systems more realistic, we need to take into account the thermodynamic properties such as temperature (T), volume (V), and pressure (P), in our MD simulation. These variables describe the macroscopic state of a system and are essentially connected to the system's microscopic quantities via the rules of statistical mechanics [51]. Therefore, simulations are carried out modeling the system at room temperature and (in a box of) constant volume or at constant pressure.

Furthermore, it is essential to treat solvent effects in the simulation by adding a number of explicit solvent molecules (typically water) surrounding the molecule. In the next step, for the purpose of studying the dynamics, the time evolution of the system, including all interacting particles, can be considered by solving the classical mechanics equations of motions with an appropriate force field acting on the particles. This is the foundation of molecular dynamics (MD) simulations. In the following, we describe the statistical mechanics description of an ensemble of particles and then the molecular dynamics simulation method for a liquid phase is explained.

2.1 Statistical Mechanics

Statistical mechanics can be considered as a bridge between macroscopic and microscopic worlds, two important scales in describing the physical state of a system [52, 53]. Statistical

mechanics rely on a probabilistic description of the configurations of a system instead of concentrating on the deterministic motion of the particles. In fact, this probabilistic perspective is the main concept that connects the atomistic motion of particles, i.e. molecular dynamics, to the thermodynamic properties of a system constituted by those particles. The fundamental equation is [54, 55]:

$$S = k_B \ln \Omega \tag{2.1}$$

wherein the entropy S is directly defined by the number of possible microstates of the system, i.e. Ω. Here, k_B is Boltzmann's constant, equal to $1.38065 \times 10^{-23} J/K$ [56]. In equilibrium, based on different characteristic properties of the system's macrostate, which can be described by a limited number of thermodynamic state functions such as energy (E), temperature (T), volume (V), and pressure (P), the microscopic configuration (Ω) can be obtained.

Depending on the thermodynamic state variables, four main ensembles exists in biomolecular simulations: microcanonical ensemble or constant NVE, canonical ensemble or constant NVT, isothermal-isobaric ensemble or constant NPT, and grand-canonical ensemble or constant μVT (μ is the chemical potential). As the grand-canonical ensemble is beyond the scope of this study we neglect to explain it in the following.

2.1.1 Microcanonical or NVE ensemble

In a microcanonical ensemble, the thermodynamic state of a system is described by a set of constant number of particles, i.e. N, a fixed volume, i.e. V, and a fixed total energy, i.e. E. Therefore, the system is isolated from the outer surroundings world (no heat exchange, no particle exchange). Since all the microstates have the same amount of energy of E, the probability of observing each microstate is equal to:

$$\rho_i = \frac{1}{\Omega(N,V,E)}, \quad and \quad \sum_{i=1}^{\Omega} \rho_i = 1 \tag{2.2}$$

while the $\Omega(N,V,E)$ indicates the total number of the system's available microstates or in another word, system's phase space. In a system of N particles, the probability distribution of microstate $x = (q,p)$, can be written as:

$$\int \rho(q,p) d^{3N}q d^{3N}p = 1 \tag{2.3}$$

which is the continuum form of Eq. 2.2. In this equation, q and p indicate the position and momentum of particles, respectively. Therefore, an ensemble average of any quantity $f(p,q)$

can be calculated as:

$$\langle f(q,p)\rangle = \int f(q,p)\rho(q,p)d^{3N}qd^{3N}p \tag{2.4}$$

In the microcanonical ensemble, the conservation of energy means:

$$\mathscr{H}(x_i) = E \tag{2.5}$$

where $\mathscr{H}$ is the Hamiltonian of the system, that is a function of microstate x, which in equilibrium condition obeys the Liouville equation:

$$\{\rho(x), \mathscr{H}(x)\} = 0 \tag{2.6}$$

It is convenient to define the particles energy in a range $(E - \frac{\triangle}{2}) \leq \mathscr{H} \leq (E + \frac{\triangle}{2})$, where $\triangle \ll 1$, rather than a sharp value E. Therefore for an ensemble of particles with $x = (q,p)$ one can write:

$$\Omega(N,V,E) = \alpha \int d^N p \int d^N q \delta(\mathscr{H}(q,p) - E), \tag{2.7}$$

where α is a normalization factor and δ is Dirac's delta. Therefore, the entropy of a microcanonical ensemble is:

$$S(N,V,E) = k_B \ln\left(\alpha \int d^N p \int d^N q \delta(\mathscr{H}(q,p) - E),\right) \tag{2.8}$$

It is important to note that in equilibrium condition, all possible microstates of a system are equally probable, called the fundamental postulate of statistical mechanic [53, 57], as already stated in Eq. 2.2.

2.1.2 Canonical ensemble or NVT ensemble

A canonical ensemble represents a system that is in contact with a heat bath, or reservoir R, at fixed temperature (T), fixed volume (V), and fixed number of particles (N). Therefore, instead of constant energy as in a microcanonical ensemble, there is an energy exchange between the system and the heat bath in the canonical ensemble. Since an ensemble is actually many copies of a system, one can also regard this as an energy exchange between systems. Of course, since the heat bath is considered to be very large in comparison to the given system, any realistic value of the system's energy (E_{Sys}) would be very small with respect to the heat bath energy

(E_R).Therefore, for the total energy (E) of these two interacting systems one can write:

$$E = E_{Sys} + E_R, \qquad and \qquad \frac{E_{Sys}}{E} = 1 - \frac{E_R}{E} \ll 1 \tag{2.9}$$

Hence, the probability of finding a given system in microstate E_i is equal to finding system R in state $E - E_{Sys}$:

$$\rho_i = \frac{1}{\Omega_{Sys}(E_i)} = \frac{1}{\Omega_R(E - E_i)} \tag{2.10}$$

according to the Eq. 2.8, the entropy S, by use of a Taylor series can be written:

$$S = k_B \ln \Omega_S(E_i) = k_B \ln \Omega_R(E - E_i) \approx k_B \ln \Omega_R(E) - \frac{\partial}{\partial E}\left[k_B \ln \Omega_R(E)\right] E_i + \mathscr{O}^{\mathscr{N}}, \quad \mathscr{N} \geq 2 \tag{2.11}$$

now, from thermodynamic rules we know that $\frac{\partial S_R}{\partial E}|_{V,N} = \frac{1}{T}|_{V,N}$, therefore

$$\ln \Omega_R(E - E_i) \approx k_B \ln \Omega_R(E) - \frac{E_i}{T}, \qquad \Rightarrow \qquad \Omega_R(E - E_i) \approx \Omega_R(E) \exp(-\frac{E_i}{k_B T}) \tag{2.12}$$

Since E is constant, with using Eq. 2.10 we conclude that

$$\rho_i \propto \exp(-\frac{E_i}{k_B T}) \tag{2.13}$$

which by considering the normalization property, the probability distribution of microstate i in canonical ensemble, i.e. NVT, would be:

$$\rho_i = \frac{\exp(-\frac{E_i}{k_B T})}{\sum_i \exp(-\frac{E_i}{k_B T})} \tag{2.14}$$

where the denominator, as the fundamental result of the NVT ensemble is:

$$\mathscr{Q}(N, V, T) = \sum_i \exp(-\frac{E_i}{k_B T}) \tag{2.15}$$

this quantity is known as the partition function that represents the sum over all microstates in the canonical ensemble, weighted by the Boltzmann factor. As the energy level of physical systems may include degeneracy g_i, i.e. groups of microstates belonging to the same energy, Eq. 2.15 can be rewritten as:

$$\mathscr{Q}(N, V, T) = \sum_i g_i \exp(-\frac{E_i}{k_B T}) \tag{2.16}$$

in the continuum form, for a system of particles with $x = (q, p)$, partition function can be written as:

$$\mathcal{Q}(N,V,T) = \int g(E)\exp(-\beta E)dE \tag{2.17}$$

where $\beta = \frac{1}{k_B T}$. Here, the factor $exp(-\beta E)$, termed *Boltzmann factor*, represents the distribution of microstates in a canonical ensemble. Using the Eqs. (4,17) for each physical quantity f, the thermodynamic average is:

$$\langle f \rangle = \frac{\int f(E)g(E)\exp(-\beta E)dE}{\int g(E)\exp(-\beta E)dE} \tag{2.18}$$

The Helmholtz free energy $A(= U - TS)$ in a canonical ensemble is:

$$A(N,V,T) = -k_B T \ln \mathcal{Q}(N,V,T) \tag{2.19}$$

This is the fundamental outcome of the canonical ensemble. Also for the ensemble average of microstates energy we have:

$$U \equiv \langle E \rangle = \frac{\int Eg(E)\exp(-\beta E)dE}{\int g(E)\exp(-\beta E)dE} = -\frac{\partial}{\partial \beta}\ln\{\int g(E)\exp(-\beta E)dE\} = -\frac{\partial}{\partial \beta}\mathcal{Q}(N,V,T) \tag{2.20}$$

We conclude that computing the partition function, a quantity that is directly connected to the system's microstates, allows us to describe the macroscopic properties of a system in a canonical ensemble. $\langle \cdots \rangle$ denotes an expectation value.

In the canonical system, one may ask how much indeed the energy fluctuates in a canonical system. From Eq. 2.18, for second moment of energy we can write:

$$\langle E^2 \rangle = \frac{\int E^2 g(E)\exp(-\beta E)dE}{\int g(E)\exp(-\beta E)dE} = \frac{1}{\mathcal{Q}(N,V,T)}\frac{\partial^2}{\partial \beta^2}\mathcal{Q}(N,V,T) \tag{2.21}$$

Hence, for the energy fluctuation we will have:

$$\langle E^2\rangle - \langle E\rangle^2 = \frac{1}{\mathcal{Q}(N,V,T)}\frac{\partial^2}{\partial\beta^2}\mathcal{Q}(N,V,T) - (-\frac{\partial}{\partial\beta}\mathcal{Q}(N,V,T))^2 = \frac{\partial^2}{\partial\beta^2}\ln\mathcal{Q}(N,V,T) \tag{2.22}$$

By knowing that heat capacity in constant volume is $C_V = (\frac{\partial U}{\partial T})_{V,N}$, and using Eq. 2.17 we can rewrite upper formula as:

$$\langle E^2\rangle - \langle E\rangle^2 = k_B T^2 C_V \tag{2.23}$$

As we know from thermodynamic, both energy E and heat capacity C_V are extensive quantities, i.e. $E \propto N$ and $C_V \propto N$, therefore, the root mean square fluctuation of E will be:

$$\frac{\sqrt{\langle E^2\rangle - \langle E\rangle^2}}{\langle E\rangle} = \frac{\sqrt{k_B T^2 C_V}}{\langle E\rangle} \propto \frac{\sqrt{N}}{N} = \frac{1}{\sqrt{N}} \tag{2.24}$$

importantly, this relation indicates that for a large number N (i.e. number of system's particles) the energy fluctuation is quite negligible and in the thermodynamic limit the canonical ensemble is equivalent with microcanonical ensemble.

2.1.3 Isothermal-isobaric or NPT ensemble

The isothermal-isobaric, i.e. NPT, ensemble is very similar to the canonical NVT ensemble. The only difference is instead of keeping the volume V constant, the pressure P is kept fixed. Moreover, the N, in this study, is a typical number of particles in the simulation box which is still far from the thermodynamic limit, i.e. $N \to \infty$, and hence different ensembles are not equivalent. Therefore one needs to choose a proper setup, e.g. NVT or NPT, for the simulation. It is important to note that most experiments are performed under NPT condition. The appropriate free energy description for this ensemble is the Gibbs free energy G, which is defined as:

$$G(N,P,T) = U - TS + PV(P) = A(N,V(P),T) + PV(P) \tag{2.25}$$

here A is the Helmholtz free energy described for the canonical NVT ensemble, last section. In order to control the pressure, the system needs to do work by changing its volume. Therefore, by using of the Laplace transformation [53] from canonical ensemble to isothermal-

isobaric ensemble one can write:

$$\triangle(N,P,T) = \frac{1}{V_0}\int_0^\infty dV \exp(-\beta PV)\mathcal{Q}(N,V,T) \tag{2.26}$$

The major differences between NVT and NPT ensembles is the weighting factor for the microstates that becomes $\exp(-\beta(E+PV))$ in the NPT ensemble. Therefore, analogue to the canonical NVT ensemble's Helmholtz free energy (Eq. 2.19), the Gibbs free energy of an NPT ensemble is:

$$G(N,P,T) = -k_B T \ln \triangle(N,P,T) \tag{2.27}$$

2.2 Molecular dynamics simulations

Molecular dynamics (MD) simulations are a powerful tool for studying the equilibrium properties and dynamics of a system including interacting particles such as biological macromolecules. In other words, MD simulations are a computational approach for the description of the statistical mechanics of a system [2] based on solving the Newtonian equation of motion:

$$F = ma \tag{2.28}$$

demonstrating that the acceleration a of a body with mass m is proportional to the force F acting on that body. This equation is known as Newton's second law. In MD simulations, the acting force on each atom, which is arisen by its interaction with other atoms in the system, is defined as a mathematical function, called force field (see also force field section).

In connection to statistical ensembles, the MD simulation relies on an important hypothesis, called *Ergodic hypothesis* that states: for many-particles system, the ensemble average[1] of a physical quantity $\langle f \rangle_{ensemble}$ converges to the long-time average $\langle f \rangle_t$:

$$\langle f(q,p) \rangle_{ensemble} = \lim_{t\to\infty} \langle f(q(t),p(t)) \rangle_t \tag{2.29}$$

It must be noted that because of the stochastic nature of MD simulations, sometimes the system falls into large probability regions, called metastable states in phase space, and therefore requires a very long time to jump to another metastable state. In this case, there is no guarantee for the system under MD simulation to reach the ergodic limit, refereed as quasi nonergodicity, i.e. the system seems nonergodic on the simulation time scale.

[1] By ensemble average we mean an average over a large number of systems at one time which all include identical thermodynamic properties but are different on the molecular level.

2.2.1 Newton's Equations Numerical Integration

In order to obtain the trajectory (dynamics) in phase space of a system, the Newton equation of motion needs to be numerically solved in MD simulation. There are several integration methods in the field. In here, we explain the velocity verlet algorithm, a most frequently used integration method in MD simulation.

2.2.1.1 Velocity Verlet Integrator

In order to obtain a numerical integration scheme, the Taylor expansion is a naive approach to use, in which the third term and after are truncated.

$$r(t+\triangle t) = r(t) + v(t)\triangle t + \frac{F(t)}{2m}\triangle t^2 + \mathscr{O}(\triangle t^n), \qquad n \geq 3 \tag{2.30}$$

where $F(t) \equiv F(r_1(t), r_2(t), ..., r_N(t))$ indicate the force acting on (N) particles and $v(t)$ is velocity of particles at time t, $\mathbf{r}$ are the atomic positions. The similar expansion can be written for $r(t-\triangle t)$:

$$r(t-\triangle t) = r(t) - v(t)\triangle t + \frac{F(t)}{2m}\triangle t^2 + \mathscr{O}(\triangle t^n), \qquad n \geq 3 \tag{2.31}$$

Adding the Eqs. (1) and (2) results:

$$r(t+\triangle t) + r(t-\triangle t) = 2r(t) + \frac{F(t)}{m}\triangle t^2 \tag{2.32}$$

which, after rearrangement, becomes

$$r(t+\triangle t) = 2r(t) - r(t-\triangle t) + \frac{F(t)}{m}\triangle t^2 \tag{2.33}$$

The Eq. 2.33 is known as Verlet integrator. This algorithm computes the update of the particles positions at two proceeding time steps. By substantiation of the Eqs. (1) and (2), one can compute the velocity of particles at time t with respect to times $t-\triangle t$ and $t+\triangle t$ as follows:

$$v(t) = \frac{r(t+\triangle t) - r(t-\triangle t)}{2\triangle t} \tag{2.34}$$

This algorithm obviously indicates that for defining the velocity, two initial subsequent steps are needed. As the phase space is composed of both positions and velocities of particles, a new scheme is needed to explicitly evolve the velocity in the integration. This can be achieved by constructing different variant of verlet algorithm, known as velocity Verlet

algorithm. In this scheme, the time step of Eq. 2.31 with a $\triangle t$-shift in time can be written as:

$$r(t) = r(t+\triangle t) - v(t+\triangle t)\triangle t + \frac{F(t+\triangle t)}{2m}\triangle t^2 \tag{2.35}$$

Now, if we replace the first term of the right hand side, i.e. $r(t+\triangle t)$, with Eq. 2.30 we can obtain:

$$v(t+\triangle t) = v(t) + \frac{\triangle t}{2m}\left[F(t+\triangle t) + F(t)\right] \tag{2.36}$$

Eq. 2.36 together with Eq. 2.30 is a velocity Verlet scheme that is used to evolve both the positions and velocities of particles simultaneously in MD simulation. The velocity Verlet algorithm satisfies the time-reversibility condition in MD simulation, which should to be preserved in integrator as a fundamental symmetry of Hamilton's equations.

2.2.1.2 Simulation Time Step

One of the important conceptual issues in MD simulations is determining the proper time step ($\triangle t$). The consensus time step, which is related to the chemical bond vibration scale, is on the order of femtoseconds. It is shown that selecting a big time scale leads to unstable motion due to a very big error in MD integration. On the other hand, selecting a very small time scale decreases the computational efficiency due to a very long calculation time. Moreover, the time step should be lower than the highest frequency vibration in the system, e.g. in water, the stretch frequencies are about $\sim$100 THz (10 fs per cycle) [58]. In regard to these issues, studies reveal a time steps 1-2 fs, as an appropriate time step because of producing accurate and stable enough result, in biomolecules MD simulations.

2.2.2 Thermostats

Thermostats are tailored to sample the correct ensemble (NVT, NPT) of the system by adjusting the temperature toward the desired constant, i.e. fluctuations around the target temperature value, in a simulation. But, one may argue what is the mean of temperature in molecular dynamic. In fact, in simulation, the temperature of the system in each instance is nothing else than its contained kinetic energy; determined by the equipartition theorem:

$$\sum_{i=1}^{N} \frac{\left|p_i^2\right|}{2m_i} = \frac{k_B T}{2}(3N - N_c) \tag{2.37}$$

where $3N - N_c$ is the total number of degrees of freedom of the system. Here, N_c is the number of defined constrains. Therefore, this theorem declares that the system's temperature

is related to its average energy. There are a variety of thermostats that remove boundaries effects of energy by coupling the system to a heat bath so as to keep its temperature constant. The widely used thermostats in simulations are Berendsen thermostat [59], Velocity-rescaling thermostat [60], Nosé-Hoover thermostat [61, 62], and Langevin thermostat [63]. The latter has been used in our MD simulation.

2.2.2.1 Langevin thermostat

The Langevin thermostat maintains the temperature of the system by a modified version of the Newton's equation of motion, i.e Langevin equation:

$$\begin{aligned} \frac{d}{dt}q_i &= \frac{1}{m_i}p_i \\ \frac{d}{dt}p_i &= -\nabla U\left(q_i\right) - \gamma_i p_i + f_i(t)\Big|_{i=1,2,\ldots,N} \end{aligned} \tag{2.38}$$

where γ_i is the friction coefficient with units of τ^{-1}, and f_i is a random force or noise with $\left\langle f_i\left(t\right) f_j\left(t+s\right)\right\rangle = 2m_i\gamma_i k_B T\,\delta\left(s\right)\delta_{ij}$; mean the noise is uncorrelated both in time and across particles. In MD simulations using this thermostat, the equation of motion is changed in each time step so that the momentum change is:

$$\Delta p_i = \left(-\nabla U\left(q_i\right) - \gamma_i p_i + \delta f\right)\Delta t \tag{2.39}$$

in which the δf is a Gaussian distributed noise with probability:

$$\rho(\delta f) = \frac{1}{\sqrt{2\pi\sigma}}\exp\left(-\frac{\left|\delta f^2\right|}{2\sigma^2}\right) \tag{2.40}$$

where the $\sigma^2 = 2m_i\gamma_i k_B T$ is noise standard deviation; this is known as the fluctuation-dissipation theorem in which the friction force is related to the random force.

2.2.3 Barostats

Constant temperature and pressure, as present in the NPT ensemble, are conditions that many experimental data are measured at. Like thermostats that are used to keep the temperature constant, there are several algorithms, called barostats, that aim to keep the system's pressure constant during the simulation. Some of the most known barostats are: Berendsen barostat, Parrinello-Rahman barostat, and Nosé-Hoover Langevin piston barostat. The latter thermostat

that has been used in our MD simulation is a combination of the Nose-Hoover thermostat with Langevin fluctuation piston method, which is implemented in the NAMD simulation package.

2.2.4 Force Field

MD simulation fundamentally requires a set of potential energy functions that define the energy of each atom with respect to the other atoms in the system. The sum of these potential functions define the total potential of system, which is called force field. It is important to note that, for many systems in which chemical reactions are not studied, classical force fields are enough to model the the system using MD simulation. Until now, several families of force fields have been developed for biomolecular systems. For example, one of the most known force fields is CHARMM (Chemistry at Harvard Macromolecular Mechanics) [64] that is very well updated over time and different version of this force field are available. Other popular force fields are AMBER (Assisted Model Building with Energy Refinement) [65], OPLS (Optimized Potentials for Liquid Simulations) [66], and GROMOS (GROningen MOlecular Simulation) [67]. The basic function of a force filed includes bonded and nonbonded terms or interactions:

$$U_{totat} = U_{bonded} + U_{nonbonded} \tag{2.41}$$

where the bonded interactions consist of short-range bond stretch, angle, dihedral, and improper terms, while the nonbonded interactions include short-range Lennard-Jones and long-range electrostatic interactions. The standard function of force fields for biomolecular systems is:

$$\begin{aligned} U_{total} \equiv U(r) = & \sum_{bonds} K_{ij}^{bond} \left(b_{ij} - b_{ij}^{0}\right)^2 + \sum_{angles} K_{ijk}^{angle} \left(\theta_{ijk} - \theta_{ijk}^{0}\right)^2 \\ & + \sum_{dihedrals} K_{ijkl}^{dihedral} \left[1 + \cos\left(n\chi_{ijkl} - \chi_{ijkl}^{0}\right)\right] + \sum_{impropers} K_{ijkl}^{improper} \left(\phi_{ijkl} - \phi_{ijkl}^{0}\right)^2 \\ & + \sum_{i} \sum_{j \neq i} \left\{ 4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6} \right] + \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \right\} \end{aligned} \tag{2.42}$$

where $U(r)$ is the potential energy that is a function of the system's coordinates. The first four terms on the right hand side of Eq. (38) are bonded terms in which the atoms covalently interact with each other (schematically sketched in Fig. 2.1) while the last term expresses the atomic nonbonded interactions.

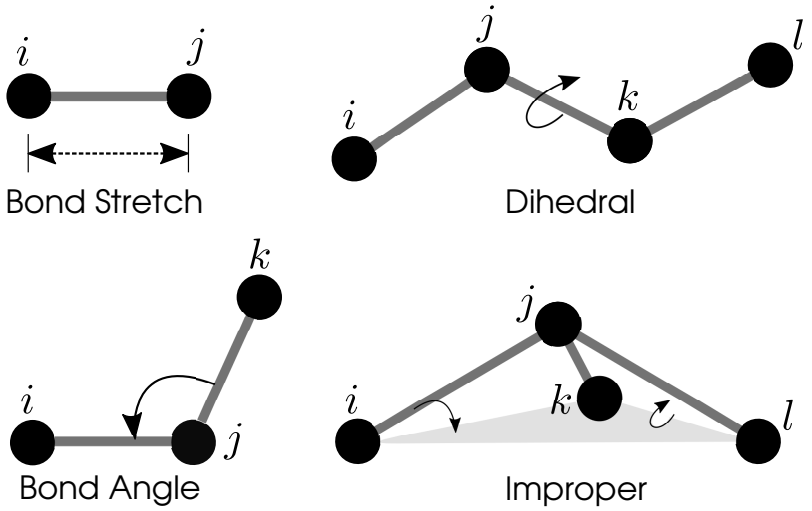


Figure 2.1: Bonded interactions. Bond stretch (upper left), bond angle (down left), dihedral torsion angle (upper right), and improper torsion (down right) that used to define biomolecular force field.

2.2.4.1 Bond Stretching Term

The bond stretching potential function describes the interaction between each two atoms that are bonded covalently. The examples for this potential term are C-N, C-C,C=C, C-O, H-N, and so on. The widely used potential function for bond stretching for MD simulation purpose is a harmonic potential that is based on Hooke's law:

$$U_{bond-stretching} = \sum_{bonds} K_{ij}^{bond} \left(b_{ij} - b_{ij}^{0}\right)^{2} \tag{2.43}$$

where $U_{bond-stretching}$ is the sum of all covalent bonds in the system. In this equation, K_{ij}^{bond} and b_{ij}^{0} are force constant and bond length equilibrium value, respectively.The bond length b_{ij} varies with each time step of the simulation,according its vibrational motion.

2.2.4.2 Bond Angle Term

The angle bend potential function describes the angle made of three atoms that are covalently bonded. Same as the bond stretching function, a harmonic potential is widely used to describe this bond angle:

$$U_{angle} = \sum_{angles} K_{ijk}^{angle} \left(\theta_{ijk} - \theta_{ijk}^{0}\right)^{2} \tag{2.44}$$

where the parameters K_{ijl}^{angle} and θ_{ijk}^{0} are force constant and equilibrium value of the angle,

respectively. The examples for this angle are C-N-C, C-C-C, C-O-C, C-C=O, H-C-H, and so on.

2.2.4.3 Proper Dihedral Angle Term

The proper dihedral angle potential function describes the rotation angle between two planes that are made of four sequential atoms that are covalently bonded. Change of the energy because of this rotation is defined by the potential:

$$U_{dihedral} = \sum_{dihedrals} K_{ijkl}^{dihedral} \left[1 + \cos\left(n\chi_{ijkl} - \chi_{ijkl}^{0}\right)\right] \tag{2.45}$$

where χ_{ijkl} is the dihedral angle defined between the two planes ijk and jkl that are spanned by atoms $ijkl$ (upper right, Fig. 2.1). The parameters $K_{ijkl}^{dihedral}$ and χ_{ijkl}^{0} are force constant and equilibrium value of this dihedral angle. Number n is the periodicity factor that defines the number of a dihedral angle's minima and maxima. The examples for this dihedral angle are sequential bonds C-C-C-O, H-C-C-H, C-C-C-C, H-C-O-H, and so on.

2.2.4.4 Improper torsion Angle Term

The improper torsion angle potential function describes the torsion angle defined by four atoms $ijkl$ so that the central atom j is covalently bonded to atoms i, l, and k. In this case, the improper angle is an angle formed between the i-j bond and the i-l-k plane (down right, Fig. 2.1), i.e ϕ_{ijkl}. The appropriate potential function is a harmonic potential:

$$U_{improper} = \sum_{impropers} K_{ijkl}^{improper} \left(\phi_{ijkl} - \phi_{ijkl}^{0}\right)^2 \tag{2.46}$$

the parameters $K_{ijkl}^{improper}$ and ϕ_{ijkl}^{0} are force constant and equilibrium value of this improper angle. The examples for this improper angle are N-C$^{\alpha}$-C=O, O=C-N-H$_2$, and so on.

2.2.4.5 Van der Waals Interaction Term

The Van der Waals (VdW) potential describe the nonbonded interaction between atom pairs ij, with distance r_{ij}. The VdW potential function mimics the short-range attractive and repulsive forces between the atoms by using a common Lennard-Jones potential:

$$U_{LJ} = \sum_{i} \sum_{j \neq i} 4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6} \right] \tag{2.47}$$

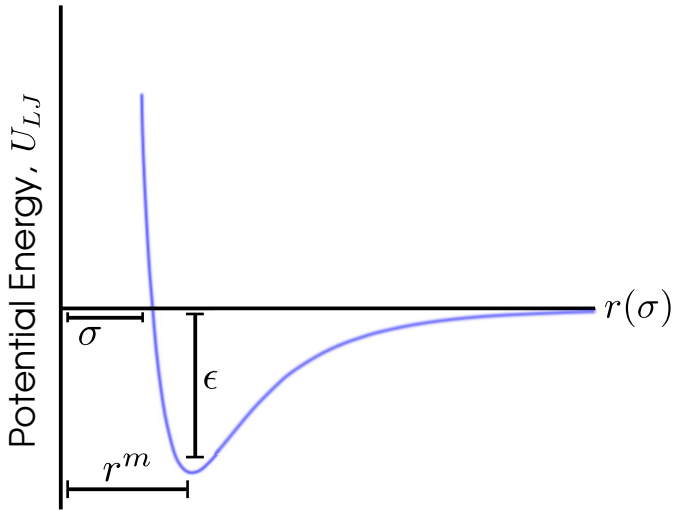


Figure 2.2: Lenard Jones potential mimic the repulsive and attraction energy between atoms. r^m is a distance in which the potential energy is minimum. σ and ε are zero-potential distance and potential wall depth, respectively.

where U_{LJ} is the sum of all pair interactions of atoms in the biomolecular system. The parameter ε_{ij} is the potential well depth between atoms i and j while parameter σ_{ij} is a distance in which the pair interaction potential of atom i and j is zero. The σ_{ij} is related to the distance in which the potential is minimum, i.e. r_{ij}^m (Fig. 2.2), as $\sigma_{ij} = \left(2^{1/6}\right)^{-1} r_{ij}^m$. The first term in Eq. 2.47, proportional to inverse distance with power 12, is the repulsive interaction due to Pauli exclusion, whereas the second term, proportional to inverse distance with power 6, is attractive interaction due to short-range dipole-dipole interaction. The VdW interaction energy is short-ranged and decays rapidly with the increase of distance (it falls with power 6 and 12 of distance). The VdW interaction for Lenard-Jones potential is sketched in Fig. 2.2.

2.2.4.6 Electrostatic or Coulomb Term

The electrostatics interaction energy between two charged atoms i and j, results from the Coulomb potential:

$$U_{elec} = \sum_i \sum_{j \neq i} \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \tag{2.48}$$

where the q_i, q_j indicate the partial charge of atoms i and j, respectively. The ε_0 is the vacuum permittivity. In contrast to the VdW interaction, the electrostatic interaction is long-range interaction; because coulomb potential decays slowly by distance (with power 1).

2.2.5 Periodic Boundary Conditions (PBC)

Periodic boundary conditions (PBC) are an appropriate method for minimizing edge effects, via creating infinite images of a simulation unit cell (box) in space. By use of PBC, infinite copies of the unit cell are sequentially repeated in all directions of the cell, in form of a lattice. This method is widely used in the field of molecular simulations. Fig. 2.3 schematically illustrates the concept of the PBC in a 2D-lattice. The coordinate of the particles in images (cells around the central primary cell) are exactly related to the coordinate of the original cell particles. Once the particles quit the cell, the image particles enter at the opposite site; compare the particles A, C, and D with their opposite images, i.e. A′, C′, and D′, in Fig. 2.3. Therefore, the number of the cell particles is conserved. In practice, because of computational efficiency, most MD simulations use cutoff schemes in order to evaluate the potentials. In regard to those cutoff schemes, each particle in a unit cell interacts only with its nearest periodic image of the other N-1 particles, named minimum-image convention, or with those of particles, in minimum images, that reside inside a sphere with radius R_{cutoff} originated at the particle. Introducing a cutoff scheme in PBC is an appropriate strategy for evaluating the short-range truncated Lennard-Jones potential in a simulation but it leads to significant artifacts and errors in computing the electrostatic potential because of its long-range nature. To overcome this issue, the Particle Mesh Ewald (PME) method [68] is used in MD simulations, nowadays. Based on the PME method, the electrostatic potential is re-written as a sum of two rapidly converging terms, one in real space and the other one in reciprocal space, respectively. The real space term is short-ranged with singularity at the origin and the reciprocal space term is long-ranged and calculated by use of the fast Fourier transform (FFT) method.

2.2.6 Molecular Dynamics Simulations - Initial Setting

In order to perform a MD simulation, several steps are needed to be done before the equilibrium and MD production steps that are summarized below in order:

Initialization. The first step is to prepare the simulation box including the structure of system of interest or solute, possible ions and solvent. In general, the initial coordinates of solute molecules such as proteins and other macromolecules are adopted from experimentally (e.g. NMR spectroscopy, X-ray crystallography) resolved structures that normally are available in the Protein Data Bank (PDB). In some cases, the available PDB structures can be used to model unavailable, yet structurally similar, molecules by introducing appropriate mutations in silico. Also in some other cases such as unfolded proteins, B-DNA model, or membrane bilayer, the coordinates can be modeled from their structurally defined properties. After finding

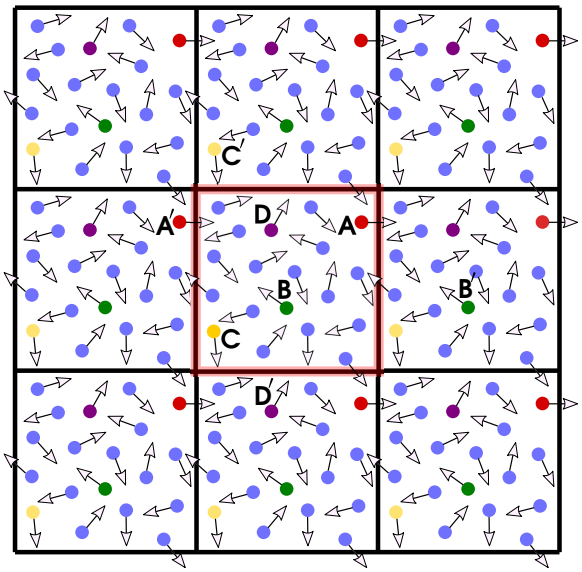


Figure 2.3: Schematic picture of Periodic Boundary Condition (PBC) is drawn for two dimensional system. The arrows shows the random direction of particles. The original unit cell is in the center.

the initial coordinates of the system, a sufficient number of ions and solvent molecules (in our case, water molecules) is added to the system in order to mimic the physiological conditions, including solvent effect. All the modeled system is placed in an unit cell, called simulation box. Because of electrostatic interactions (see section force field), the box size should be adjusted such that the minimum distance of the box edge to the solute is below the short-range cutoff. In order to remove the edge effects as well as appropriately calculate the electrostatic interactions, that is long-ranged interactions, periodic boundary conditions (PBC) is used.

Minimization. In order to remove energetically unfavorable interactions in the system (explained below), which could lead to a large overall energy, it is important to perform an energy minimization so as to find a low-energy/favorable conformation. For example, two atoms very close to each other in space lead to a large penalty in the Van der Waals repulsion energy. Or some artifact from either the experimental structure or the computational modeling can considerably enlarge the potential energy of the system. Energy minimization aims to reduce the net force on the atoms and therefore prepares an excellent starting point for the MD simulation.

Heating. Normally, the MD simulations are performed in 300K (about room temperature). In order to reach this temperature, the system is slowly increased from 0K to 300K. At each increasing step the new velocities are reassigned to the atoms by using the Maxwell–Boltzmann distribution for velocities.

Equilibrium. Once the system is set up by initial preparation steps, as explained above, an equilibrium simulation is needed to perform before the production phase of the simulation. In this step, in order to include the real dynamics, the kinetic and potential energies of the system are exchanged so that the solvent and solute equilibration is satisfied. Over the equilibrium period, both kinetic and potential as well as the total energy need to converge to fluctuations around a certain mean value.

2.3 MD Trajectory Analysis

2.3.1 Root-Mean-Square Deviation/Fluctuation

Root-mean-square deviation (RMSD) is a commonly used method for measuring the deviation of conformations of biomolecules during the simulation with respect to their initial structures. Before the RMSD calculation, the MD trajectories need to be aligned to a reference structure, e.g. the initial structure, thus removing translational and rotational motions of the system. Normally, for proteins the backbone atoms of the amino acids residues are chosen for this calculation. The RMSD formula is:

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N} d_i^2} \tag{2.49}$$

while N is number of particles and d_i is distance of particle i's position in each MD frame with respect to that particle's position in the initial frame.

Root-mean-square fluctuation (RMSF) measures the fluctuation of biomolecules conformation during the simulation time with respect to their mean conformation. The RMSF of atom a is given by:

$$RMSF_i = \frac{1}{T}\sum_{t=1}^{T} \left(r_t^a - \overline{r^a}\right)^2 \tag{2.50}$$

where T is the total number of trajectory frames, r_t^a is the position of atom a at time t and $\overline{r^a}$ is the mean position of atom a during the simulation.

Median Structure The median structure of simulation trajectory was determined as the snapshot that has minimum root mean-squared deviation (RMSD) with respect to the averaged structure of the trajectory.

2.3.2 Time-Averaged Mean Square Displacement (TAMSD)

The diffusion pattern of a single trajectory $r(t)$, can be defined via the time-averaged mean square displacement (TAMSD) [69] as:

$$\overline{\delta^2(\Delta;t)} \equiv \frac{1}{t-\Delta}\int_0^{t-\Delta} dt' \left[r(t'+\Delta)-r(t')\right]^2 \tag{2.51}$$

where Δ is the lag-time, t is the total observation time, i.e. simulation time, and $r(t')$ is the time series of the dynamical variable, respectively.

2.3.3 Direct and Water Mediated Hydrogen Bonds

Hydrogen bonds were analyzed based on geometric criteria, i.e. a maximal distance of 3.2 Å between donor and acceptor atom and a donor--hydrogen atom–acceptor angle that deviates maximally by 42° from linear. Water-mediated hydrogen-bonds between two groups of atoms were identified as a water molecule that simultaneously forms a hydrogen-bond with the first group (i.e. the protein) and the other one (i.e. with the DNA).

2.3.4 Correlation Analysis

To investigate the dynamical dependency of fluctuating variables in the simulations we used two kinds of correlation, the Pearson correlation and generalized correlation, that is defined based on mutual information theory.

2.3.4.1 Pearson Correlation

The Pearson correlation [70] that defines the linear dependency of random variables is calculated by the covariance matrix of variables, normalized by the corresponding variances:

$$C_{ij} = \frac{cov(r_i, r_j)}{\sigma_{r_i}\sigma_{r_j}}, \quad cov(r_i, r_j) = \left\langle r_i r_j \right\rangle - \left\langle r_i \right\rangle \left\langle r_j \right\rangle, \quad \sigma = cov(r,r) = \left\langle r^2 \right\rangle - \left\langle r \right\rangle^2 \tag{2.52}$$

which r_i and r_j are the positional fluctuation of atoms i and j, respectively. $cov(r_i, r_j)$ is covariance and σ_r is variance.

To calculate the linear correlation between two sets of angles, for instance, side chain and backbone dihedral angles of a protein, we use the circular statics approach by Jammalamadaka

and Sengupta. The linear correlation between two angles α and β is defined by:

$$C_{\alpha\beta}^{circular} = \frac{\sum_{i=1}^{N} sin(\alpha_i - \bar{\alpha}).sin(\beta_i - \bar{\beta})}{\sqrt{\sum_{i=1}^{N} sin(\alpha_i - \bar{\alpha})^2 . \sum_{i=1}^{N} sin(\beta_i - \bar{\beta})^2}} \tag{2.53}$$

where the mean is defined by quadrant-specific inverse of the tangent as:

$$\bar{\alpha}^{circular} = arctan\left(\left[\sum_{i=1}^{N} sin(\alpha_i)\right], \left[\sum_{i=1}^{N} cos(\alpha_i)\right]\right) \tag{2.54}$$

2.3.4.2 Generalized Correlation

The main disadvantage of Pearson correlation is that it only measures linear dependencies between two random variables- The nonlinear dependencies however can be estimated by using the concept of mutual information (MI). The MI between two random variable densities of x and y is defined as:

$$I(x,y) = H(x) + H(y) + H(x,y) \tag{2.55}$$

here, $H(x)$ is the information content or entropy of random variable x. $H(x,y)$ is the joint entropy. By knowing the relation of entropy with probability density $H(x) = -\int \mu(x) \ln \mu(x) dx$, another form of MI is:

$$I(x,y) = \int\int dxdy\mu(x,y) log \frac{\mu(x,y)}{\mu_X(x)\mu_Y(x)}, \quad \mu_X(x) = \int dx\mu(x,y), \quad \mu_Y(y) = \int dy\mu(x,y), \tag{2.56}$$

where $\mu(x,y)$, $\mu_X(x)$, and $\mu_Y(y)$ are joint probability density, marginal densities of X, and marginal densities of Y, respectively. If the distribution of the random variable is Gaussian, i.e.

$$\mu(x) = \left[(2\pi)^n \det(\sigma_X)\right]^{-1/2} \exp\left[-\frac{1}{2}(x-\bar{x})^{\dagger} {\sigma_X}^{-1} (x-\bar{x})\right] \tag{2.57}$$

where σ_X is $n \times n$ covariance matrix, the entropy of this variable will be:

$$H(x) = \frac{1}{2} \log\left[(2\pi e)^n \det(\sigma_X)\right] \tag{2.58}$$

while e is Napier's constant. Therefore, for two Gaussian-distributed random variables x

and y, one can obtain:

$$I(x,y) = \frac{1}{2}\log\left(\frac{\det(\sigma_X)\det(\sigma_Y)}{\det(\sigma)}\right) \tag{2.59}$$

The unsubscribed matrix σ is the covariance matrix of joint variables (x,y). In case of bivariate Gaussian with covariance:

$$\sigma_{xy} = \begin{pmatrix} \sigma_x^2 & c\sigma_x\sigma_y \\ c\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix} \tag{2.60}$$

where c is the correlation coefficient. Therefore, one can obtain for MI:

$$I(x,y) = -\frac{d}{2}\log(1-c^2) \tag{2.61}$$

where d is the dimension of random variables x and y.

Therefore, one can define the generalized correlation [71], C^{MI}, as:

$$C^{MI} = \left[1 - exp(-\frac{2I(x,y)}{d})\right]^{1/2} \tag{2.62}$$

2.3.4.3 Mutual Information Estimation

There are many methods for the estimation of MI. One of the most leading approach is a k-nearest neighbor estimator described by Kraskow et al [72]. In this approach, for each point in joint space, the Kth euclidean neighbor distance is used for statistics. In this study, the parameter k=6 is selected for calculating the MI.

2.3.4.4 Linear Correlation Score Function

Correlations between all pairs of fluctuating atom positions were calculated as Pearson correlation. The Pearson correlation, is defined by the normalized covariance matrix [73]:

$$r_{ki} = \frac{cov(x_k, x_i)}{\sigma_{x_k}\sigma_{x_i}} \tag{2.63}$$

where x_k and x_i are the fluctuations of random variable k and i, respectively.

The correlation score function is a measure of the intensity of correlation for each variable k (here, the position of the Cα atoms of the protein residues), defined as [74]:

$$CS_k = \frac{1}{N-1}\sum_{i}^{N-1} r_{ki} \tag{2.64}$$

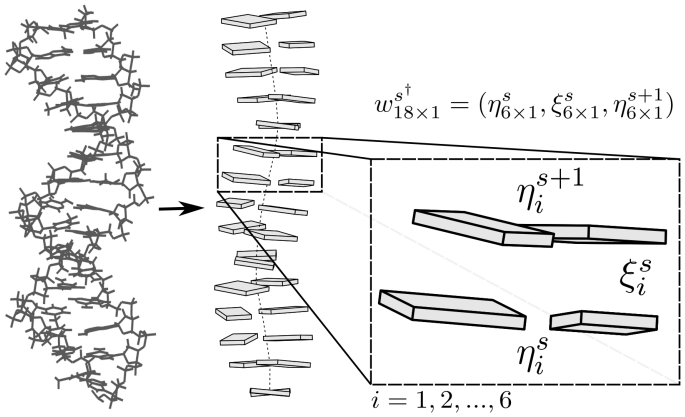


Figure 2.4: Coarse-grained DNA rigid-base model. Each base is modeled as a rigid plane.

Here, the correlation score function is normalized. In order to remove the trivial and non-important correlations only pairs with a of $r_{ki} \geq 0.4$ were considered.

2.3.4.5 Entropy estimation

The configurational entropy of the protein is estimated based on the mass weighted covariance matrix of atomic fluctuations via two well established methods, one proposed by Schlitter [75] and another one by Andricioaei [76].

For computation of the protein entropy we used the fluctuations of the backbone Cα atom. All the calculations are done via Grcarma software, a Task-Oriented Interface for the Analysis of MD trajectories [77].

2.3.5 DNA Conformation

The local DNA conformation was analyzed using Curves+, a program for analyzing the coarse-grained geometry of DNA [78]. These parameters include local geometrical inter and intra base-pair (bp) parameters, groove parameters, and DNA bending parameters.

2.3.5.1 DNA Elastic model

In order to evaluate the elastic response of the DNA upon protein binding, we used the rigid-base coarse-grained model of DNA up to the nearest-neighbor level. Based on this model, each dinucleotide is defined as a unique building block, which is described by a vector including 18 intra and inter base-pair parameters (the model is schematically shown in Figure 2.4).

The coarse-grained free energy can be written in quadratic form as:

$$U = \frac{1}{2}(w_i - \overline{w}_i)A_{ij}(w_j - \overline{w}_j) \tag{2.65}$$

where index i and j indicate the bps parameters and $w^{s^\dagger}_{18\times1} = (\eta^s_{6\times1}, \xi^s_{6\times1}, \eta^{s+1}_{6\times1})$ is the vector of internal configuration coordinates, η, ξ, with the ground state $\overline{w}$ (here, index s indicates the base-pair x_i and dinucleotide $x_i x_{i+1}$). A_{ij} is a symmetric, positive-definite stiffness matrix. As a result of the central limit theorem, the parameters, which are the result of many random variables, are in good approximation Gaussian distributed. Therefore, the stiffness matrix can be calculated by the second moment of a Gaussian distribution:

$$< \Delta w \otimes \Delta w >= k_B T (A_{ij})^{-1} \tag{2.66}$$

where $\Delta w = w - \overline{w}$, k_B is the Boltzmann constant and T is temperature. It is noteworthy here that the coordinates are non-dimensionalized with 1 Å and 10.6 for length and angle scales, respectively (see also ref [79]).

To estimate the protein-induced deformation energy, the fluctuation difference of the complex's DNA variables (w^c_i) with respect to their means in the free DNA model ($\overline{w}^f_i$) are calculated. Therefore this energy can be written as:

$$U_d = \frac{1}{2}(w^c_i - \overline{w}^f_i)A^c_{ij}(w^c_j - \overline{w}^f_j) \tag{2.67}$$

whereA^c_{ij} is the stiffness matrix of the DNA in the complex systems. In equilibrium, each dinucleotide in the free DNA, with 18 degrees of freedom (18 base pair parameters), contributes, as a consequence of the equipartition theorem, $9k_BT$ of energy [80].

The relative entropy of bound state dinucleotides with respect to their unbound state was calculated as follows [81]. By definition, the relative entropy of a density $\rho_2(w)$ with respect to a density $\rho_1(w)$ is defined as:

$$S_{rel}(\rho_2||\rho_1) = -\int \rho_2(w) \ln\left[\frac{\rho_1(w)}{\rho_2(w)}\right] dw \tag{2.68}$$

In the case of given Gaussian densities ρ_c and ρ_f, with stiffnesses A^c and A^f, and means $\overline{w}^c$ and $\overline{w}^f$, the relative entropy can be reformulated as:

$$S_{rel}(\rho_c||\rho_f) = -\frac{1}{2}[\ln\left(detA^f/detA^c\right) - (A^c)^{-1} : A^f + I : I] + \frac{1}{2}(\overline{w}^c - \overline{w}^f).A^f(\overline{w}^c - \overline{w}^f) \tag{2.69}$$

Here, the colon signifies the standard Frobenius inner product, and I is the identity matrix

with dimension of 18, i.e. number of degrees of freedom for each dinucleotide.

Chapter 3

Protonation-State-Dependent Communication in Cytochrome c Oxidase

This chapter is based on the publication:

Bagherpoor Helabad, M., Ghane, T., Reidelbach, M., Woelke, A. L., Knapp, E. W., and Imhof, P. Biophysical J. 113 (2017) 817-828.DOI

Proton transfer in the Cytochrome C Oxidase (CcO), known as complex IV, can occur through two distinct pathways, the D- and K-channels [82, 83]. In total, eight protons are transferred through these channels. The four protons called chemical protons are transferred to and consumed in the binuclear redox center (BNC) in order to reduce the molecular oxygen to water and four other protons are pumped across the membrane, leading to an electrochemical gradient that drives ATP synthesis. So as to carry out its proper function, both the transfer of "chemical" and "pumped" protons in CcO must be highly regulated. The proposed catalytic cycle is shown in Fig. 1.4(b). The first proton transfer in the cyclic process occurs in the $P_R \rightarrow F$ intermediate state, in the oxidative part of the reaction. The D-channel, which is suggested to be the predominant proton transfer pathway [84], starts at residue D132 and ends with residue E286, located about ~ 10 Å away from the BNC while the K-channel spans from residue E101 to Y288. The residue E286 (in D-channel) is required in the reductive phase of the oxidase reaction to compensate for the electrons being transferred to heme *a* and Y288 is discussed to be the residue that provides the necessary reaction protons in heme *a3*. A key residue in the K-channel is K362 that has an essential active role in the proton conducting process through the channel rather than just providing the positive charge [33]. An amount of water molecules is present in both channels, and a number of polar residues in these channels can further facilitate the proton transfer by forming an appropriate hydrogen-bond network. In contrast to the consensus that all the pumped protons are transferred through the D-channel,

it is not clearly known in which order and number the "chemical" protons are transferred through the D- and/or K-channels.

Crystal structures indicate residue K362 pointing down toward the N-side of the membrane and not towards the BNC side. And in the available structure insufficient water molecules are resolved to bridge this gap, rendering proton transfer along a hydrogen-bonded chain of water molecules and protein residues unlikely [27, 85, 86]. Molecular dynamics simulations of protonated K362 reveal the pointing up of this residue toward BNC whereas neutral K362 is unlikely to do so [33, 87]. The pKa calculation in different redox states has revealed that in the $P_R \rightarrow$ F state, the K362 protonation is likely [33]. The K362 conformational rearrangement in the reductive phase has shown to facilitate the transfer of (further) protons in the K-channel during the oxygen reduction [88]. Simulations of CcO with protonated K362 demonstrate an increased number of water molecules inside the K-channel, thus enhancing the likelihood of proton transfer via water molecules [33]. It is also shown that an up-orientation of K362 may facilitate the formation of the $P_R \rightarrow$ F state by providing the positive charge close to BNC, thus compensating the additional negative charge in BNC. However, no further transfer of protons is assumed by K362 [89].

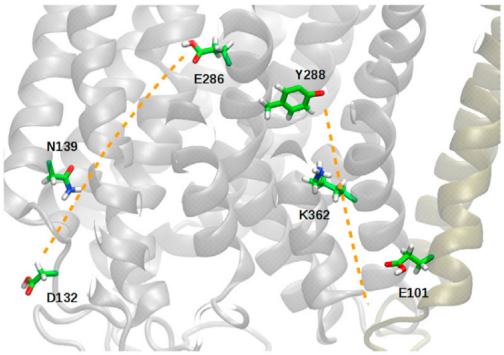
Here, we consider a situation in which the first electron has arrived at the BNC and the first proton is transferred from bulk to the protein interior; represented by $P_R \rightarrow$ F intermediate step. In this case, the residue Y288 is negatively charged.

In order to mimic proton transfer through the channels, different protonation states of the key residues are considered. In the D-channel, both the D132 and E286 as an entrance and terminal residue, respectively, are considered as key residues. In the K-channel, the residue E101, which is discussed as a possible entry point, and K362, are considered as key residues [88, 90] that may also affect the transfer of protons through D-channel.

In order to understand how different protonation states of the key residues in CcO's proton transfer channels alter the dynamical behavior of the complex, we have performed molecular dynamics simulations of CcO in the $P_R \rightarrow$ F state, in different protonation states of those key residues. The key residues and their possible protonation states combinations are shown in Figure 3.1. Using the simulations output, communication within and between the D- and K-channels, as an important aspect in understanding better the mechanism of proton transfer is analyzed.

3.1 Molecular dynamics simulations: systems and protocols

The model setup follows the same protocol as described in [33]. For protein residues, the CHARMM22 force field [91] was applied, whereas the parameters for the cofactors are based



D132	E286	K362	E101
0	0	0	0
0	0	0	1
0	1	0	0
0	1	0	1
1	0	0	0
1	0	0	1
1	1	0	0
1	1	0	1
0	0	1	0
0	0	1	1
0	1	1	0
0	1	1	1
1	0	1	0
1	0	1	1
1	1	1	0
1	1	1	1

Figure 3.1: (*Left*) Lower part of cytochrome c oxidase highlighting important residues of the D- and K-channel, respectively. The proton transfer channels are indicated by dashed lines. For residues D132, E286, K362, and E101, the protonation states have been varied in this study. The figure shows the 1111 model with all four residues protonated. (*Right*) Simulated protonation states of cytochrome c oxidase. The number "1" refers to a proton at the respective residue, leading to a protonated lysine, or neutral aspartate and glutamate residues, respectively. The number "0" indicates no such proton, corresponding to neutral lysine, or negatively charged aspartate or glutamate, respectively.

on quantum chemically derived atomic partial charges and optimized cofactor geometry by Woelke et al. [33]. The core enzyme complex was embedded in a lipid bilayer of phosphatidylcholines and solvated in TIP3 water [92] (115,000 atoms in total). Parameters for the lipid bilayer were obtained from the CHARMM36 extension for lipids [93]. Heme a3 was modeled with Fe(IV) bound to one oxygen atom (i.e., half an oxygen molecule) and the copper(II)B-ion was modeled bound to a hydroxyl ion. Y288 was modeled in the deprotonated state. This setup as well as charges and bonds to ligands is the same as described in [33]. Na^+ counterions were added by random substitution of water molecules to neutralize the charge of the system.

The protonation states for the key residues of the two channels, D132, E286, K362, and E101, were varied in all possible combinations, resulting in a total of 16 models (see Figure 3.1). We refer to the differently protonated models by a binary number indicating the protonation state of the four key residues as "1" if protonated and as "0" if not. Model 1010, for example, has D132 and K362 protonated, but E286 and E101 unprotonated. Simulations were performed using periodic boundary conditions in a tetragonal box of size (x = y = 96 Å, z = 124 Å). Long-range electrostatic interactions were treated using the particle mesh Ewald method [94] on a 96 $\times$ 96 $\times$ 128 charge grid. A nonbonded cutoff of 12 Å was applied. The short-range electrostatics and van der Waals interactions were truncated at 12 Å , using a switch function starting at 10 Å . The solvated structures were minimized using 5000 steps of steepest descent, then gradually heated for 30 ps to 300 K with 1 K temperature steps with harmonic restraints on the solute atoms. The systems were equilibrated in three different stages with the numbers of particles, pressure (1 bar), and temperature (300 K) kept constant (NPT ensemble) during 75 ps. Pressure control was introduced using the Nose-Hoover Langevin piston with a decay period of 500 fs. After equilibration, for all 16 models, the restraints were lifted and four individual NPT production runs, started with different initial velocities, were performed for 100 ns each. This way, a total simulation time for all models and all MD runs of 6.4 μs was obtained. The integration time step was 2 fs and coordinates were saved with a sampling interval of 2 ps. All covalent bond lengths involving hydrogen atoms were fixed using the SHAKE algorithm [95]. For all analyses, only the last 60 ns of each simulation run were considered.

3.2 Results

3.2.1 Conformational analysis

The conformational dynamics of key D- and K-channel residues in CcO have been analyzed in terms of their side-chain dihedral angles and distances between pairs of residues. The histogram of side chain distances between key residues, i.e. D132, N139, E286 (in D-channel), E101, and K362 (in K-channle) are plotted in Appendix A, Figures A.1-A.11. As can be seen in these figures, depending on the different protonation states, each residue can adopt different distance and dihedral angle values. The most populated conformational states of key residues are summarized in Figure 3.2.

The side-chain conformation of D132 is directly influenced by the protonation state of this residue. Unprotonated D132 populates more conformations than protonated D132, this is particularly obvious in the side-chain dihedral angle χ_1^{D132}. Interestingly, the conformation of this residue is significantly effected by the protonation state of the residue E286. Models with both unprotonated D132 and E286, i.e. 0000, 0001, 0010, and 0011, exhibit an additional side-chain angle population with $\chi_1^{D132} \approx +60^\circ/\chi_2^{D132} \approx \pm 120^\circ$, whereas the combination of unprotonated D132 with protonated state of E286 results in observing states with $\chi_1^{D132} \approx 180^\circ$. In contrast, the protonation states of the K-channel residues, i.e. E101 and K362, do not show any effect on the conformational states of D132. Both the $\chi_1^{D132} \approx -60^\circ/\chi_2^{D132} \approx -60^\circ$ and $\chi_1^{D132} \approx -60^\circ/\chi_2^{D132} \approx 120^\circ$ conformations are populated in all different models.

Residue E286 exhibits rather restricted conformational dynamics. As shown in Figure 3.2, just three out of nine possible combinations of $\chi_2^{E286}/\chi_3^{E286}$ are observed. The dihedral angle $\chi_1^{E286} \approx 180^\circ$ is reached by all models in a situation in which protonated E286 is limited to $\chi_2^{E286} \approx 60^\circ/\chi_3^{E286} \approx 0^\circ$; a position that corresponds to E286 pointing down into the D-channel. The only exception is model 0111 in which the residue E286 points upward; this is a model in which K-channel residues E101 and K362 are protonated. Models with unprotonated D132 and E286 only take the conformation determined by $\chi_2^{E286} \approx 60^\circ/\chi_3^{E286} \approx -120^\circ$.

K362 side-chain conformation is largely flexible with a preference for values of $\chi_1^{K362} \approx -60^\circ/\chi_2^{K362} \approx 180^\circ$, in all models. Models 0111 and 1110 are the only models that include an additional side chain angle $\chi_1^{K362} \approx 60^\circ/\chi_2^{K362} \approx 180^\circ$. $\chi_3^{K362} \approx 60^\circ$ conformations are observed for all models with neutral K362. In contrast, the $\chi_3^{K362} \approx -60^\circ$ can be occupied just by models with protonated K362. The results do not show any protonation-state dependent conformation for dihedral angle χ_4^{K362} (Figure 3.2).

Residue E101's conformation is little dependent on its protonation state. Two conformations with side chain angles $\chi_1^{E101} \approx 180^\circ/\chi_2^{E101} \approx 60^\circ$ and $\chi_1^{E101} \approx -60^\circ/\chi_2^{E101} \approx 60^\circ$ are only observed with six models with protonated E101. However, for models with protonated

E101, the difference between the side chain dihedral angle's histogram for individual simulation runs are considerably larger than those of histograms seen for models with unprotonated E101 (compare Appendix A Figures A.12-A.26). Also, the protonation states of residue K362, in the K-channel, and or residues D132 and E286, in the D-channel, do not exhibit a significant impact on E101's conformation.

N139 is not a titratable residue itself. Our results show that the conformation of this residue, N139, is considerably dependent on the protonation states of the other residues. The protonation state of residue E286 seems to have a considerable role in N139's conformational flexibility. Indeed, the models with unprotonated E286 exhibit more dihedral states of N139 than those with protonated E286. The models with protonated E286 only occupy the state $\chi_1^{N139} \approx 180^\circ / \chi_2^{N139} \approx 0^\circ$. In contrast, the models with protonated D132 and unprotonated E286 do not visit the state $\chi_2^{N139} \approx 0^\circ$. A conformation with $\chi_1^{N139} \approx 60^\circ$ is not observed in any of the 16 different models.

3.2.2 Communication Analysis via Generalized Correlation

In order to see how CcO's key residues communicate with each other, a communication analysis was performed using generalized correlation as a metric for communication strength (see method chapter). To do this, we have used our own Java-written code, which is based on the JGraph library as well as our written python/Matlab codes. The generalized correlation matrices of other models are plotted in Appendix A Figure A.27. Figure 3.3 shows the generalized correlation matrices for models with the most non-negligible entries and the strongest communication between key residues, i.e. 0011, 0111, 1110, and 1010. Table 3.1 shows the shortest paths lengths between the key residues in CcO. As seen in this table, these four models have strongest communication in all the pathways analyzed. Appendix A Table A.28 shows the shortest paths lengths between the key residues, for all models. Characteristic path length (CPL) by definition is the average of the length of the shortest paths between two considered residues (nodes in the graph). Models 0111 and 0010 are the models that possess the shortest characteristic path lengths, indicating that the average connections in these models are stronger than in the other models.

For models with protonated E286, strongest communication is observed within the D-channel. Residues D132 and N139, in all models, communicate strongly with each other, likely because of their close distance to each other while the communication between the residues N139 and E286, which have larger distance, are moderate in almost all the models. The shortest communication paths between these residues, i.e. N139 and E286, are observed for models 0010, 0111, and 1110, where the latter two models have E286 protonated. Another communication path is between D-channel terminal residues, i.e D132 and E286, in which the

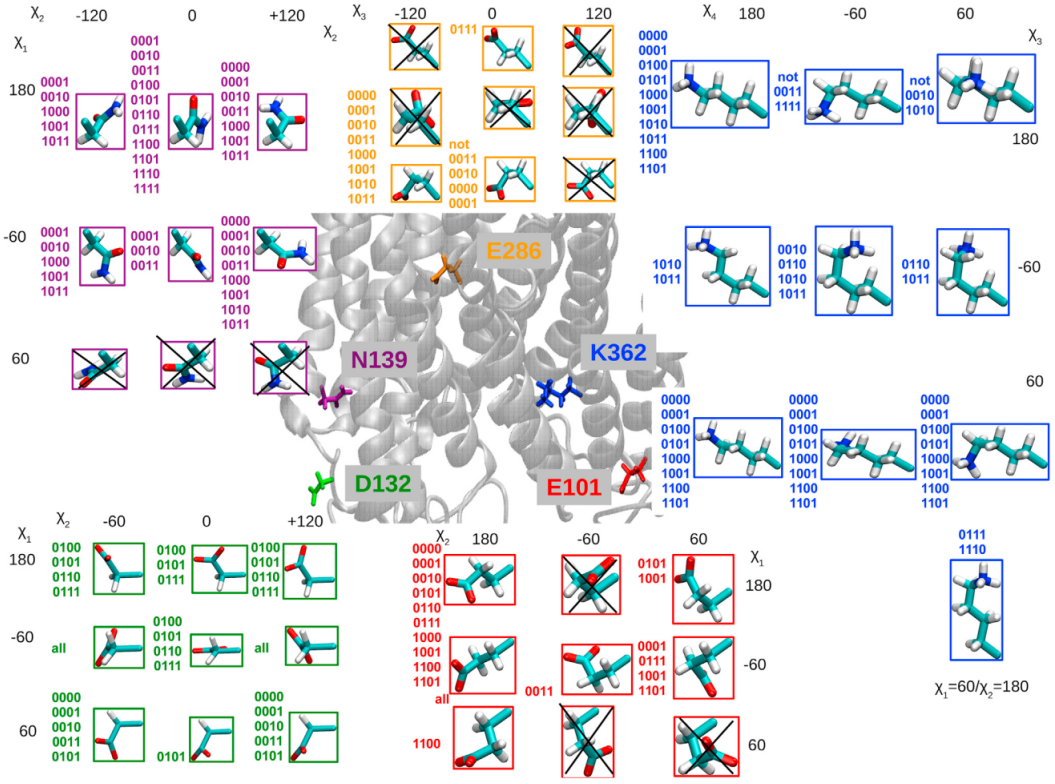


Figure 3.2: Dihedral conformational states of residues D132 (green), N139 (purple), E286 (orange), K362 (blue), and E101 (red). The orientation shown for each residue corresponds to the one viewed from the same perspective as the one for the protein cartoon in the center. Next to each state those protonation models for which the state has been observed in at least two of the simulations are listed. Crossed-out states have not been observed in any model. For simplicity, side-chain dihedral χ_1 of E286 ($\chi_1^{E286} \approx 180°$) and χ_3 of E101 ($\chi_3^{E101} \approx 120°$) are shown in their most probable orientations for all models. K362 is shown with a conformation $\chi_1^{K362} \approx 180°/\chi_2^{K362} \approx 180°$ that is occupied in all models. Models 0111 and 1110 adopt an additional $\chi_1^{K362} \approx 60°/\chi_2^{K362} \approx 180°$ conformation, which is shown separately. The protonation models are labeled by "1" for protonated and "0" for unprotonated D132, E286, K362, and E101, respectively.

Path	Models (Shortest Path Length (Cost))					
E286–D132	1100	(16.9 ± 2.7)	1111	(17.1 ± 1.8)		
N139–D132	1100	(7.7 ± 1.7)	0101	(8.1 ± 1.3)		
E286–N139	1111	(10.8 ± 1.9)	1100	(11.0 ± 1.8)	0100(11.1 ± 0.7)	0110(11.2 ± 1.3)
Y288–E101	0000	(12.1 ± 2.8)	0100	(12.8 ± 3.5)		
K362–Y288	1000	(3.5 ± 0.2)				
K362–E101	0000	(7.1 ± 1.1)	0100	(7.7 ± 1.3)		
K362–D132				—		
K362–N139	0100	(14.8 ± 2.0)	1001	(15.0 ± 1.7)	1101(15.3 ± 3.0)	
K362–E286	0100	(7.4 ± 1.9)	1000	(7.6 ± 0.3)		
Y288–D132	1100	(17.4 ± 4.0)				
D132–E101				—		
N139–E101				—		
E286–E101	0100	(15.2 ± 2.9)	0000	(16.7 ± 3.4)		

Table 3.1: Only the models with lowest path costs are listed. The cost or length of a shortest path $P(v_1, v_2, ..., v_n)$ is the sum of the edge weights along that path, here $\sum_{k=1}^{n} -\ln\left(R_{k,k+1}\right)$, where $R_{k,k+1}$ is the generalized correlation. For N139–E101, no shortest path with a cost below the characteristic path length could be obtained. The protonation models are labeled by "1" for protonated and "0" for unprotonated D132, E286, K362, and E101, respectively.

only substantial path is observed for model 1110.

In the K-channel, communication between residue K362 and the channel terminal residues, i.e. E101 and Y288, are strong in almost all the models, which can be due to their relatively short respective distances. However, the communication between residues E101 and Y288 is only significant for model 0111.

Significant interchannel communications are observed between the residue K362, in the K-channel, and D-channel residues. Interestingly, for the models that show the strongest communication paths between D-channel terminal residues, i.e. models 1110, 0111, 1010, and 1011, a strong communication is also observed between the residues K362 and E286. Except for the model 1110, which has a weak communication, there are a strong communication between the residues K362 and N139 for all models. Also, the only models that hold weak communication between residues K362 and D132, are models 1110, 0110, and 1010. Moreover, except for the model 0111 that has a weak communication between the channels' entrance residues, i.e. E101 and D132, no significant communication is observed between these residues. However, for models 0111 and 1110, which are the most talkative models by the mean of the communication (with generalized correlation metric), there is a significant communication between residues E101 and E286, despite their rather large distance. Communication between Y288 and D132 is moderate only in model 1110.

It is noteworthy to note that in almost all models that show a strong communication via generalized correlation the residue K362 is protonated and therefore protonation of this residue seems to result in a considerable increase of the communication strength.

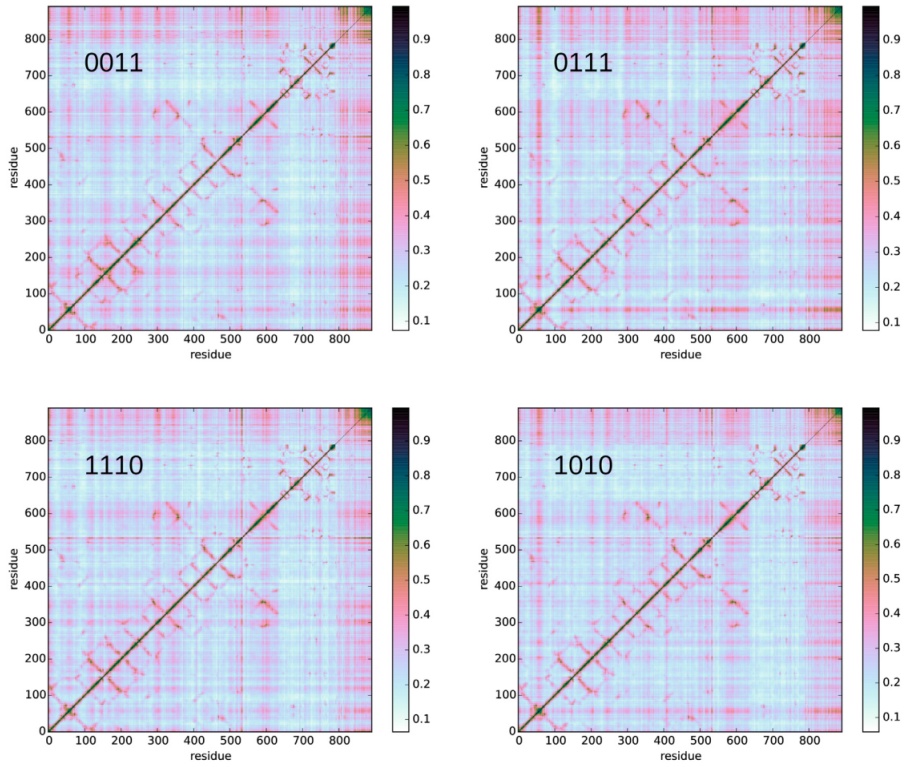


Figure 3.3: Generalized residue-residue correlation matrix of cytochrome c oxidase in the protonation state models that show largest correlations. For the other models, see Appendix A Figure A.27. The protonation models are labeled by "1" for protonated and "0" for unprotonated D132, E286, K362, and E101, respectively. The values of the matrix elements, i.e., the generalized correlations coefficients $r_{MI}\left[x_i, x_j\right]$, are color-coded as indicated by the color bars plotted next to the matrices.

3.2.3 Hydrogen Bond Interactions

In order to understand how the dynamics of CcO are protonation state dependent, the communication via hydrogen bond connections is analyzed. Here, instead of generalized correlation metric, which is described in the previous section, the hydrogen bond connection metric is used. Also, the hydrogen bond dynamics between the key residues and water molecule are analyzed via calculating the hydrogen bond lifetimes.

It is important to mention that the result of this section is analyzed by other authors of our published article, mentioned in first page of this chapter. Therefore, as we need the result of this section in the discussion part, I will give a short summary of the hydrogen bond interaction analysis in this section. For details about the method and result, the readers can refer to the original article.

3.2.4 Communication via hydrogen-bond interactions

Intra- and inter-channel communication are calculated based on a hydrogen bond connection matrix. In the D-channel, the significant communication is between N139 and E286, which are strong for models with protonated E286. In the K-channel, communication of K362 is strong with both Y288 and E101. Communication between terminal residues, i.e. Y288 and E101 is moderate in almost all models, but strongest in models that show strong communication between K362 and E101. Interestingly, models with unprotonated K362 exhibit strong communication between the residues E286 and K362. Also, for models with unprotonated K362, there is a moderate communication between the N139 and K362. In general, the communication patterns via hydrogen bonds are different than communication observed via generalized correlation.

3.2.5 Hydrogen-bond dynamics

Hydrogen-bond dynamics between key residues and water molecules were calculated. The results show that models with protonated E286 exhibit shorter hydrogen-bond lifetimes than those with unprotonated E286. Also the hydrogen-bond dynamics of K362 do not show a clear dependence on this residue's protonation state. The most interesting result has been observed for the hydrogen-bond dynamics of residue N139, which is affected by a combination of other key residues protonation states. Models with protonated E286 show shorter hydrogen-bond lifetime for residue N139 than those with unprotonated E286. The hydrogen-bond lifetime of N139 is significantly altered in models with protonated K362, with respect to models with unprotonated K362. Particularly, in cases with D132 protonated, hydrogen-bonds of N139

indicate longer lifetimes. Interestingly, in models which have instead of D132 the residue E286 protonated, the opposite effect is observed and hydrogen-bond lifetimes of N139 are reduced upon K362 protonation.

3.3 Discussion

The analysis done in this study reveals the dependence of the key residues not only by their own protonation state but also the protonation states of other residues.

Within the K-channel, between the residues K362 and E101, a communication is observed along the correlation while both of these residues are protonated. In contrast, communication by hydrogen-bond connection is observed only if both of these residues, i.e. K362 and E101, are unprotonated. This may be because of a preferable pointing up of the protonated K362, away from E101, and consequently having less impact on the hydrogen-bond network of E101.

Within the D-channel, different protonation states of the key residues alter their mutual impacts. The conformation of residue E286, pointing down toward the D-channel or up toward the BNC, has been reported to depend on its protonation state, the redox state, and the water cavities hydration level. In CcO P_M state, where heme *a* and BNC are reduced and oxidized, respectively, there is no conformational preference for E286 to point down or up. But, as soon as an electron is transferred from heme *a* to BNC (called, P_R state), residue E286 prefers a conformation that points down into the D-channel [96]. Considerably, our MD simulations result for this state, i.e. $P_R \rightarrow F$, show that independent of E286's protonation state, this residue prefers the down-conformation. The up-conformation is only observed for model 0111, in which the E286 as well as the K-channel residues E101 and K362 are protonated. MD simulations of the bovine heart CcO also show a preference for a down conformation of protonated E286 [96].

Although the protonation state of the D-channel residues, i.e. D132 and E286, are mainly dependent on their own protonation state, our results show that, in addition, protonation of residue E286 leads to residue D132 exploring a larger conformational space, preferably toward the D-channel interior. In turn, protonation of D132 leads to an additional dihedral angle of E286 being populated. This slight conformational preference of two residues toward each other is in agreement with the weak communication (hydrogen-bond or correlation based) between D132 and E286, which is observed only for models in which these residues are protonated.

Our results clearly indicate that the conformational dynamics of the nontitratable residue N139 are influenced by E286, when it is protonated. The hydrogen-bond communication between these two residues, i.e. N139 and E286, is increased for models that have one or

both D-channel terminal residues protonated. On the other hand, these models show a weak hydrogen-bond communication between residues D132 and E286. Moreover, single protonation of the D132 does not show any particular effect on N139.

Our results of the conformational dynamics of the residue N139 is in agreement with free energy study by Henry et al., in which two distinct states are observed for this residue; a closed state with conformation $\chi_1^{N139} \approx -165^\circ/\chi_2^{N139} \approx 41^\circ$, which would not allow water nor protons to pass, and an open state with conformation $\chi_1^{N139} \approx -75^\circ/\chi_2^{N139} \approx -70^\circ$, which facilitates water and proton transfer. In our simulation, a conformation $\chi_1^{N139} \approx 180^\circ/\chi_2^{N139} \approx 0^\circ$ corresponds to the closed state and conformation $\chi_1^{N139} \approx -60^\circ/\chi_2^{N139} \approx 120^\circ$ corresponds to the open state, respectively. The open state in our study, however, is different from the one seen by Henry et al. [84] by a swap between the atoms OD and NH2 of the N139, but more similar to the rotamer observed in the crystal structure. Our study also shows a conformational dependence of the residue N139 on the protonated E286. All models with protonated E286 significantly favor a state with N139 closed, with only models 1100 and 1110 showing a low population of the open state in more than one of the MD runs. In contrast, models with unprotonated E286 exhibit a higher conformational flexibility in N139 such that the open state is observed significantly often. This open state is essential for forming hydrogen-bonded water chains that facilitate the proton transfer in a Grotthuss mechanism [97]. Moreover, models with protonated E286 lead to a shorter hydrogen-bond lifetime between residue N139 and water molecules. Therefore, the residue N139 indeed have a crucial role in the D-channel, such that it acts as a gate that allows protons to transfer through the D-channel whilst the residue E286 is unprotonated, thus preventing an additional, excess proton before releasing of E286's proton in its appropriate time. Thus, an auto-regulated proton transfer through the D-channel can be observed [98].

Interestingly, the hydrogen-bond dynamics of N139 is also dependent on the protonation state of K362. Our study indicates that in models with both D132 and K362 protonated, the hydrogen-bond lifetime is longer for N139 in comparison to those models with unprotonated K362, but for models with E286 protonated (and D132 is not) and K362 protonated, the N139 lifetime is shorter in comparison to those models with unprotonated K362.

For models with unprotonated E286, in which N139 samples "open" conformations, the slower hydrogen-bond dynamics of N139, in models with protonated K362, can be regarded as an increased probability of proton transfer from the protonated D132 through the D-channel via forming an appropriate hydrogen-bond network with an increased and thus sufficiently long lifetime. On the other hand, for models with protonated E286, in which N139 samples closed conformations with shorter lifetime of N139–water hydrogen bonds, proton transfer through the D-channel is prevented. Therefore, the release of proton from E286, to the protein

loading site or to BNC, likely occurs before a new proton uptake through the D-channel.

Our result interestingly exhibits that the communication between K362 and N139 is rather weak. There is only one model, i.e. 1110, in which the communication via correlation between these two residues is noteworthy. Communication via hydrogen-bonds, however, is stronger in models with unprotonated K362 than in those with protonated K362. Therefore, the long-range effects between residues K362 and N139 must follow other pathways. Interestingly, there is a rather considerable communication between K362 and E286, both via hydrogen-bond connections and correlation. Consequently, it is conceivable that the effect of the protonated K362 on N139 is assisted by E286. Our results also indicate a longer and shorter distance of protonated K362 with N139 and E286, respectively. Although the shorter distance between K362 and E286 is still rather long, likely this decrease in distance affects the conformation and interaction of other residues, i.e., water, in between.

Results of this study clearly show that the communication between the key residues are protonation-state dependent. Models with protonated K362 have strongest communication pathways via generalized correlation. The four models 0111, 1110, 0010, and 1010, all with K362 is protonated, show significant generalized correlation (Figure 3.3). Hence, these are the models that have the shortest paths between residues at the lowest cost. In all of these models, the number of water molecules inside the K-channel are increased which presumably leads to a stronger communication between residues. Also, in these models, the communication via generalized correlation not only is observed within the channels but also between residues in different channels. Models with protonated K362 occupy an additional up-conformation that is not observed for models with unprotonated K362, thus likely impacting a greater number of different residues, which enhances the mean of communication between them. Thereby, our result support the idea that conformational flexibility strengthens communication, e.g. the additional side-chain angle $\chi_1 \approx 60^\circ$ can be observed in at least one of the individual MD runs for models 0111 and 1110, the models that have strongest communication via correlation. In contrast, the most talkative models via hydrogen-bond, i.e. 0100 and 1100, have K362 unprotonated.

Two models, 0111 and 1110, are the systems that have most pronounced communication between the two proton transfer channels, i.e. D- and K-channels, in CcO. These are the models in which a proton is ready to leave either channel toward the proton loading site (PLS) or BNC, respectively. However, in order to properly function, i.e., act as oxidase and as a proton pump, the proton needs to be transferred to the PLS, from the D-channel, and later on to the P-side. Hence, as long as E286 is not yet protonated, being protonated of K362 result in elongated N139 hydrogen-bond lifetimes compared to those models with unprotonated K362, therefby facilitating the proton transfer through the D-channel. This effect is further supported

by the conformational preference for open states of N139 with unprotonated E286. One can therefore assume that a protonated K362 supports the proton transfer through the D-channel.

3.4 Conclusion

In this study, we used MD simulations in order to better understand how different protonation states of the key residues within the two proton conducting channels, i.e. D- and K-channels, alter the conformational dynamics of the CcO.Sixteen different combinations of key residues protonation states, i.e. D132, E286 (D-channel), E101, and K362 (K-channel), have been simulated. The results of our simulations show that the conformational dynamics, hydrogen-bond dynamics, and communication of the key residues in both channels of CcO are not only influenced by the protonation state of a single residue but mainly by the combination of these key residues protonation states. In particular, communication within each channel is mainly effected by the protonation states of all the key residues of that channel. The conformational dynamics of non-titratable residue N139 is directly influenced by the protonation state of both terminal residues of the D-channel, i.e. D132 and E286, resulting in a gating role of N139. Significantly, a considerable impact of the K-channel on the D-channel is observed between residues K362 and N139. Within each channel, the main communication is via correlated motion, indicating an interplay of channel residues and water motion with hydrogen-bond lifetime. Our results suggest that the proton transfer probability of the D-channel is likely affected by K362's protonation state via changing the hydrogen-bond dynamics of N139, which, in turn, is coordinated by D-channel terminal residues protonation states.

Our results also indicate that the communication between different residues is protonation-state dependent. For instance, communication changes from mainly correlated motion for protonated K362 to hydrogen-bonded connections-based communication for unprotonated K362.

The states that have E286 protonated can be understood as states ready to release protons. Reaching such a state is more likely with protonated than with unprotonated K362. The protonation-state dependent communication between the two channels may thus also regulate proton release from the D-channel and the completion of the $P_R \rightarrow F$ transition [34].

Chapter 4

Specificity of Androgen and Glucocorticoid Receptor DNA Binding Domains for Direct and Inverted Repeat Response Elements

The DBD of steroid receptors (SRs), which includes about 70 amino acid residues, functionally contains two vital subdomains, each identified with a zinc ion that is coordinated by four Cysteine residues (Figure 4.1(a)). The first subdomain includes an α-helix, termed H1, which is responsible for protein-DNA major groove interaction, while the second subdomain holds a loop domain, termed Dim, which is responsible for protein-protein dimerization [41, 42]. A flexible loop, named "lever arm" (Figure 4.1(b)) connects these subdomains to each other. The core DNA sequence, which is specifically recognized by SRs-DBD, includes 15 precise base pairs (bps) composed of two hexameric half-sites (HS1 and HS2, respectively) separated by three bps called "spacer" (Figure 4.1(c)). This spacer region allows the protein helices sufficient conformational freedom to make high affinity major groove interactions while the dimerization DBD is preserved intact upon DNA binding [43]. The members of the SRs bind to a group of identical response elements, named classical response elements (CREs), which resemble an inverted repeat (IR) of hexamer AGAACA; i.e. AGAACA*nnn*TGTTCT [44]. The letters *nnn* indicate variable spacer sequences. Next to CREs, there is another kind of elements, named androgen response elements (AREs), which are merely recognized by the AR. The elements of the AREs are organized as a direct repeat (DR); i.e. AGAACA*nnn*AGAACA [45] (Figure 4.1(c)).

Yet, several variant forms of the DNA sequences are bound by the DBDs of GR and AR [43, 99]. These variable sequences hold distinct signals that control the receptor's structure

and activity [100–102]. The first half-site (HS1) on the 5′-upstream side is almost identical in both CREs and AREs whereas the largest sequences variability mainly occurs in the second half-site (HS2), leading to imperfect palindromic sequences. This suggests a higher affinity of the first monomer, A, of the dimeric protein than the second monomer, B, in binding to the first, HS1, and second half site, HS2, respectively [45, 103]. A recent study has furthermore shown that the conformation of the GR protein bound to an imperfect HS2 is also highly dependent on the sequence flanking the response element [50].

The conformational states of the TFs bound to DNA are indeed sequence dependent [104]. Hence, studies show that AR interacts differently upon binding CREs or AREs [45, 105]. As shown by transcription assays, AREs result in higher AR responses than CREs [105]. The only crystallized structure of the AR(DBD)-DNA, done by Paul L. Shaffer, et al. in 2004 [45], revealed an unexpected heat-to-head conformation of the AR dimer in complex with a perfect ARE (which we call from hereon AR-DR). This structure reveals additional hydrogen-bond interactions of the dimer interface, introduced by amino acid S580, which is not present in the GR, as responsible for the stabilization of the unexpected head-to-head complex arrangement [45, 106]. It has been shown that disruption of the dimerization interface, Dim, has different impact on the AR activity, depending on the response element in question. For instance, R581D mutation in the dimerization domain of AR-DBD enhances AR's activity on CREs but has less effect on AREs. On the other hand, the A579T mutation shows reduced activity on AREs but not on CREs [105]. In contrast, mutations at points that differ between the AR and GR Dim, i.e. S580G and T585I, in the AR, and G478S and I483T, in the GR, do not show much effect on DNA binding affinity and activity of these receptors [107]. These mutation data indicate that less of the AR-DR binding specificity can be attributed to the Dim interface than suggested by the crystal structure. On the other hand, replacing the second zinc-binding motif residues of AR-DBD (K573-G610) by those of GR indicate a loss of AR's high binding affinity to AREs, while its binding affinity to CREs is unchanged or even increased [47]. Also, it is shown that the changes in AR activity due to the loss of Dim interactions strongly depends on the engaged DNA response element [108]. Therefore, other parts of DBDs than the dimerization interface presumably play a role in DNA binding specificity, especially since the Dim region is too far (about 18 Å) from the DNA surface to build direct interaction [43].

In contrast to AR, no affinity of GR for binding AREs has been observed [109–111]. This is speculated to be due to the weaker dimerization interaction of GR in comparison with AR [45, 46]. How the dimer interface or other interactions and/or conformational changes could inhibit GR to form stable complexes with AREs is, however, still unclear.

Several resolved crystal structures of the GR, in complex with different variants of CREs, reveal that the GR activity and conformation is considerably influenced by the HS2 bps ele-

ments [42, 43]. In fact, it has been shown that the major contribution of the sequence specific gene regulation is carried by DNA binding energetics [112]. Besides the hexameric half-sites, which provide most of the binding energy by forming appropriate major groove contacts with the protein, it was shown that the spacer region also significantly contributes to the DBDs conformation and their resulting binding affinity [100, 111]. Most importantly, Watson, et al. showed that the lever arm conformation strictly depends on the spacer sequence and therefore suggested as an allosteric modulator that not only connects the H1 to the Dim (see Figure 1), but also associates the DNA response sequence to its respective dimer partner [100]. Furthermore, a recent NMR study combined with molecular dynamics (MD) simulations has revealed that the major conformational differences between the free and the bound state of the GR DBD are situated at the lever arm, accentuating the functional importance of this region [113]. Mutational studies in the lever arm of both, AR and GR, show a remarkable change in the respective receptor's activity [43, 114, 115]. Importantly, Meijsing et al. showed that the mutations H472R and H474A dramatically decrease and increase the GR activity, respectively. In the AR, mutation of the respective residue, i.e. Q574A or Q574D as well as mutations Y576A and Y576D effectually inhibit AR's transcriptional activity [115], showing a similar allosteric role of the lever arm in AR, as suggested for GR.

A recent study on DNA-binding preferences of AR and GR has revealed that the AR binding to DNA is more energized by the enthalpy, while the GR is directed more by entropy during the binding process [116]. However, to capture the modality of interactions between the DBDs and the core response elements, a detailed understanding about the dynamics of these systems is needed.

In this study, by employing all-atom molecular dynamics simulations, we explore the factors that govern the preferential DNA recognition of the AR/GR DBD bound state, especially addressing the question why despite the similar structure of the DBDs in AR and GR, the latter is not able to reach a stable bound state with AREs. In this regard, we simulated four protein-DNA complexes consisting of the DNA binding domains of AR and GR, each bound to a DNA sequence with IR and DR, (termed AR-IR and AR-DR, and GR-IR and GR-IR, respectively). In addition, two mutant systems GR(H472R) in complex with both IR/DR sequences were simulated.

Our MD simulations allowed us to determine the significant dynamics of these receptors' DBD-DNA interface. We show that strong hydrogen bond interactions of residue H472 with the spacer sequence, which is facilitated by a conformational change of the adjacent lever arm residues, can be a key factor in destabilizing the GR-DR complex. Furthermore, we demonstrate that the AR DBD's dimer interface as well as the AR DBD's-DNA interactions vary, depending on the bound response elements. Analysis of the communication network

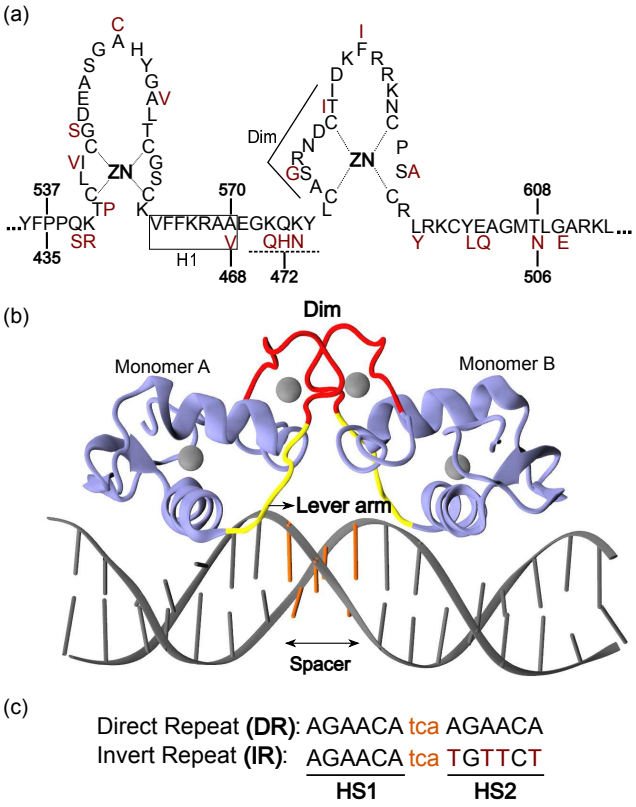


Figure 4.1: **(a)** Schematic overview of the AR/GR-DBD and DNA sequences. The amino acids colored in dark red are those elements of the GR-DBD that differ from the AR-DBD sequence. The other amino acids are the same in the AR and GR DBD. **(b)** The 3D structure of the GR-DBD/DNA complex (pdb ID: 1R4R). A similar structure exists for the AR-DBD/DNA complex (pdb ID: 1R4I). The lever arm and dimerization domain (Dim) are shown in yellow and red, respectively. The spacer region of the DNA is colored with orange. **(c)** DNA sequences for direct (DR) and inverted repeats (IR) are shown. The non-capital letters are the spacer base-pairs, colored in orange.

available within and between different domains of the protein-DNA complexes shows that the communication between two half-sites is not only connected via the Dim region but also via the protein-DNA interface. Furthermore, we show that the lever arm-spacer interface plays a predominant role in this allosteric communication between the dimerization domain and the protein monomers with their respective DNA hexamers, thus controlling stable complex formation.

4.1 Molecular dynamics simulations: systems and protocols

Classical all-atom MD simulations of the DNA binding domain of the AR/GR in complex with the IR/DR response elements were carried out using the NAMD program [117] along with the CHARMM27 force field [118, 119]. The initial structures of the AR and the GR were prepared based on the crystal structures with pdb IDs **1R4I** and **1R4R**, respectively. The IR and the DR response element sequences are 5′-CC**AGAACAtcaTGTTCT**GA-3′ and 5′-CC**AGAACAtcaAGAACA**GA-3′ respectively. The residues listed in bold are the core response elements including the two half sites, HS1 and HS2, and spacer (small letters). In addition, two simulations of the GR DBDs with H472R mutations (in both protein monomers) in complex with both IR/DR elements were carried out. Simulations based on free (uncomplexed) IR/DR B-DNA sequences were also performed. The free canonical B-DNA models were generated with Chimera, an interactive molecular visualization program [120].

The systems were solvated by using the TIP3P water model parameters [92] in a cubic box with side lengths 90 Å. Sodium ions were added to the systems so as to obtain a zero net charge. Counter-ions were placed randomly, while a minimum distance 10.5 Å from the solute and at least 5 Å distance between the ions were set. The simulation box involved ∼69,000 atoms in total. To treat the electrostatic interactions, the particle-mesh Ewald method [68] with 1 Å grid space and periodic boundary conditions were employed. The short range electrostatic and the Lennard–Jones interactions were truncated at 12 Å. The SHAKE algorithm [95] was used to constrain all covalent bond lengths, for both solute and water molecules. To refine the systems, energy minimization were performed with the conjugate gradient method. Initially, water molecules and ions of each system were relaxed in a series of 5000 steps energy minimization, then all atoms (including the solute) were minimized for another 5000 steps. The systems were then heated up to 300 K during 30 ps in an NVT ensemble simulation (time step 1 fs) following by 500 ps of relaxation time at 300 K (time step 1 fs) in an NPT ensemble. After the equilibration phase, three 100 ns MD replicas (with different initial velocities) for each system were carried out (time step of 2 fs). From those, one run per system was chosen for longer simulation, based on the calculated root mean-squared deviation (RMSD), see Appendix C Figure C.1. Finally, 0.5 μs MD production (the chosen run is called as *run0*, here) were performed for each system at constant temperature (300 K) and pressure (1 bar) with a 2 fs time step. To further evaluate the distorted GR-DR conformation (see result section), another 1.4 μs long MD production of this system, termed *run1*, was carried out (This is another run out of other two 100 ns replicas of the GR-DR, explained above). Temperature and pressure were controlled by using the Nosé Hoover Langevin thermostat [121] and the Langevin dynamics barostat [122], respectively. To prevent boundary effects on the DNA oligomers, the terminal base pairs of the DNA fragments were restrained harmonically.

4.2 Results

4.2.1 DNA conformation

4.2.1.1 Intrinsic local DNA conformation

To study the impact of the receptors DBD on their respective DNA structure, the local geometrical parameters of DNA, i.e. inter- and intra-bp parameters (Appendix B Figure B.1), minor- and major-groove (Figure 4.2(a)), and helical axis bending (Figure 4.2(b)) were calculated for the last 100 ns of the trajectories. Comparing these parameters between the bound and the unbound state of DNA elucidates the significant influence of the protein on the local geometry of the DNA. In particular, widened major- and narrowed minor-grooves in the bound state, with respect to the unbound state, can be interpreted as a potential which allows the protein to make favorable electrostatic interactions with the DNA [123]. Regarding the groove results, the complexes with an inverted repeat (IR) sequence, i.e. AR-IR and GR-IR, show similar major and minor groove widths, but for the complexes with a direct repeat DNA sequence, i.e. the AR-DR and the GR-DR, this is not the case. Groove parameters of the spacer elements, i.e. tca (Figure 4.1(c)), are remarkably different between the AR-DR and the GR-DR complex. In GR-DR, bigger T_7 major groove- and C_8A_9 minor groove-widths can be observed with respect to AR-DR. Also, a narrow C_5 major groove can be seen for GR-DR, in comparison to the respective base pair in AR-DR. It is interesting to note that both AR/GR-DR complexes show a narrower minor groove of position T_7, than their respective counterparts, AR/GR-IR. The local bend of the DNA helical axis shows rather similar values in the unbound IR/DR models and the bound IR models. In contrast, the bound DR models show different local bend values in the AR and GR complexes as well as in comparison to both bound IR models, AR-IR and AR-GR (Figure 4.2(b)). For the AR models, the local bend differences are located in the spacer and HS2, whereas for the GR models, GR-DR exhibits a significantly different local bending of the DNA (in HS1, spacer, and HS2) with respect to GR-IR. This distinct conformation of the DNA in the GR-DR model is also manifested in the inter base-pair parameters slide, rise, and twist as well as in the intra base-pair parameters opening and buckle (see Appendix B figure B.1).

4.2.1.2 Influence of the protein on the local DNA deformation

The local deformation responses of both IR/DR DNA due to complexation with AR/GR DBDs are shown in Figure 4.3(a). In equilibrium, each of the dinucleotide steps contributes on average 9 k_BT energy (red line, see the supplementary information for details on the elastic DNA model), which is used as a reference. As Figure 4.3(a) shows, dinucleotides AC_4– AT_6 in

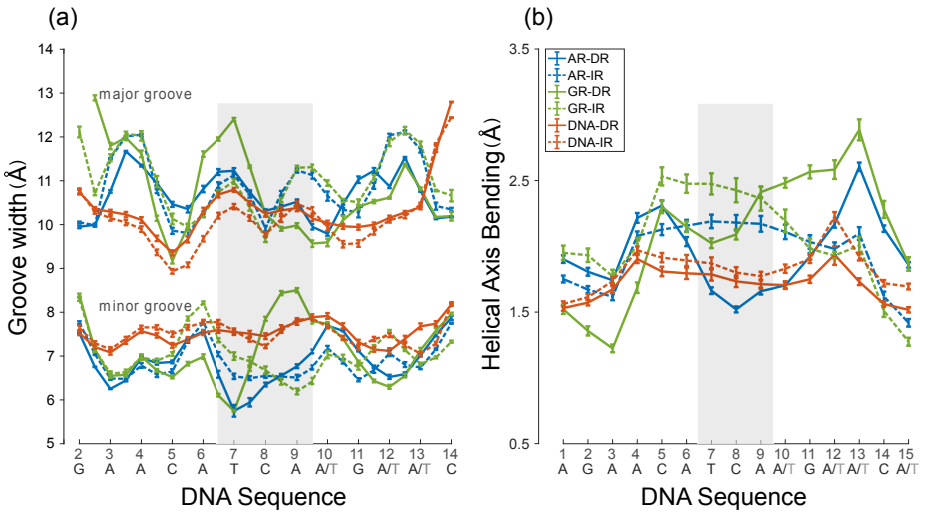


Figure 4.2: Comparison of the DNA conformation with direct (DR) and inverted repeat (IR) complexed to AR and GR and in free form, respectively. **(a)** Major and minor groove width **(b)** DNA helical axis bending.

HS1 and AA_9– GA_{11} in HS2 undergo the largest deformation upon protein binding, in all the systems. However, the terminal core dinucleotides (AC_{13} and CA_{14}) of the AR-DR experience more deformation energy than those of the other systems. Additionally, apart from the GR-IR, which shows similar energy in both half sites, the HS1 of the other systems encounter higher deformation energy than the respective HS2 (Again, the only exception are the AR-DR terminal residues). In the following, in order to evaluate the entropic effects on the DNA upon protein binding, the relative entropy S_{rel} of individual dinucleotides in their bound state with respect to the corresponding free state have been estimated, see Figure 4.3(b). This figure reveals considerable entropic effects upon protein binding on residues AC_4– AT_6 and AA_9– GA_{11}, in HS1 and HS2, respectively. These effects, however, are larger for the dinucleotides AC_4– AT_6 than for AA_9– GA_{11}. Moreover, a huge shift in the S_{rel} of the GR-DR system can be observed with respect to the other systems. Interestingly, the computed S_{rel} of the second half-site, HS2, reveals larger entropic effects for GR-IR than for AR-IR. Also the S_{rel} of dinucleotides AC_4– AT_6 are higher in the GR-IR than in the AR-IR model. Considerable differences in S_{rel} can be observed in both the HS1 and HS2 of the AR-IR system with respect to the AR-DR: the terminal dinucleotides of the AR-DR model show higher (absolute) S_{rel} values than the AR-IR, whereas the mid-elements, i,e. AC_4– AT_6 and AA_9– GA of the AR-IR show higher S_{rel} values than their counterparts in the AR-DR complex.

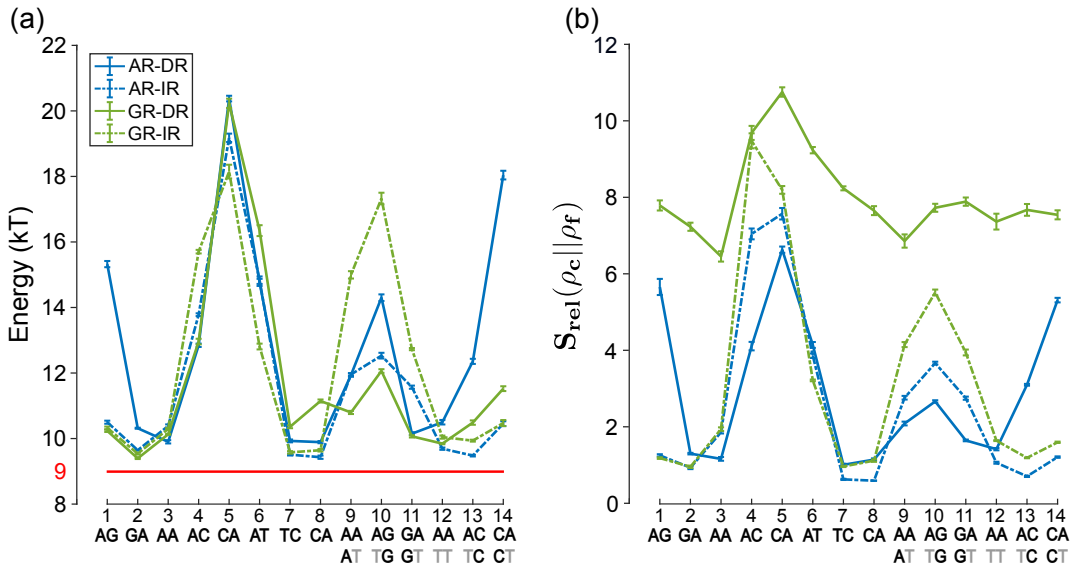


Figure 4.3: **(a)** Protein-induced deformation energy of core DNA dinucleotides, calculated for all AR/GR models. The red line shows the deformation energy of a free DNA model. **(b)** Relative entropy (S_{rel}) of core DNA dinucleotides, calculated for all AR/GR models. The inset shows schematically the rigid planes that represent a dinucleotide.

4.2.2 Protein-DNA interactions

To better analyze the interaction strength, the hydrogen bond probabilities between protein and DNA have been calculated. Therefore, both direct and indirect (mediated by water molecules) hydrogen bonds were categorized into strong and moderate, corresponding to high and medium occupancies (Figure 4.4(a)). The protein-DNA interactions of the AR/GR-DBDs with IR/DR sequences are shown in Figure 4.4(a). For each DNA hexamer, i.e. HS1 and HS2, there are four sites whose hydrogen bond interactions with the protein are conserved among all four systems. These are s1A1, s1G2, s2G5, and s2T6 in HS1 and s1A10, s1G11, s2T15, and s2G14 in HS2. The guanine residues at positions s1G11 and s2G5 are the predominant residues that form strong, i.e. highly probable, hydrogen bond interactions with the protein in all systems. In particular, the residue R568 in the helix H1 of the AR-DBD, and residues R466 in helix H1 of the GR-DBD form base-specific hydrogen bonds with guanine residues s1G11 and s2G5, respectively. Comparison of the hydrogen bond patterns between the two AR systems indicates that the AR-DR complex involves more hydrogen-bonded protein-DNA interactions than the AR-IR complex. Moreover, hydrogen bonds of residues s1G2 and s2G14 with K563 and K567, respectively, and also those of residues s2A7 and s2T6 (in the spacer) with Y576 are stronger in the AR-DR complex than in the AR-IR complex.

However, in the AR-IR complex, the hydrogen bond interactions of monomer A with hexamer HS1 and monomer B with hexamer HS2 are almost symmetric. As an example,

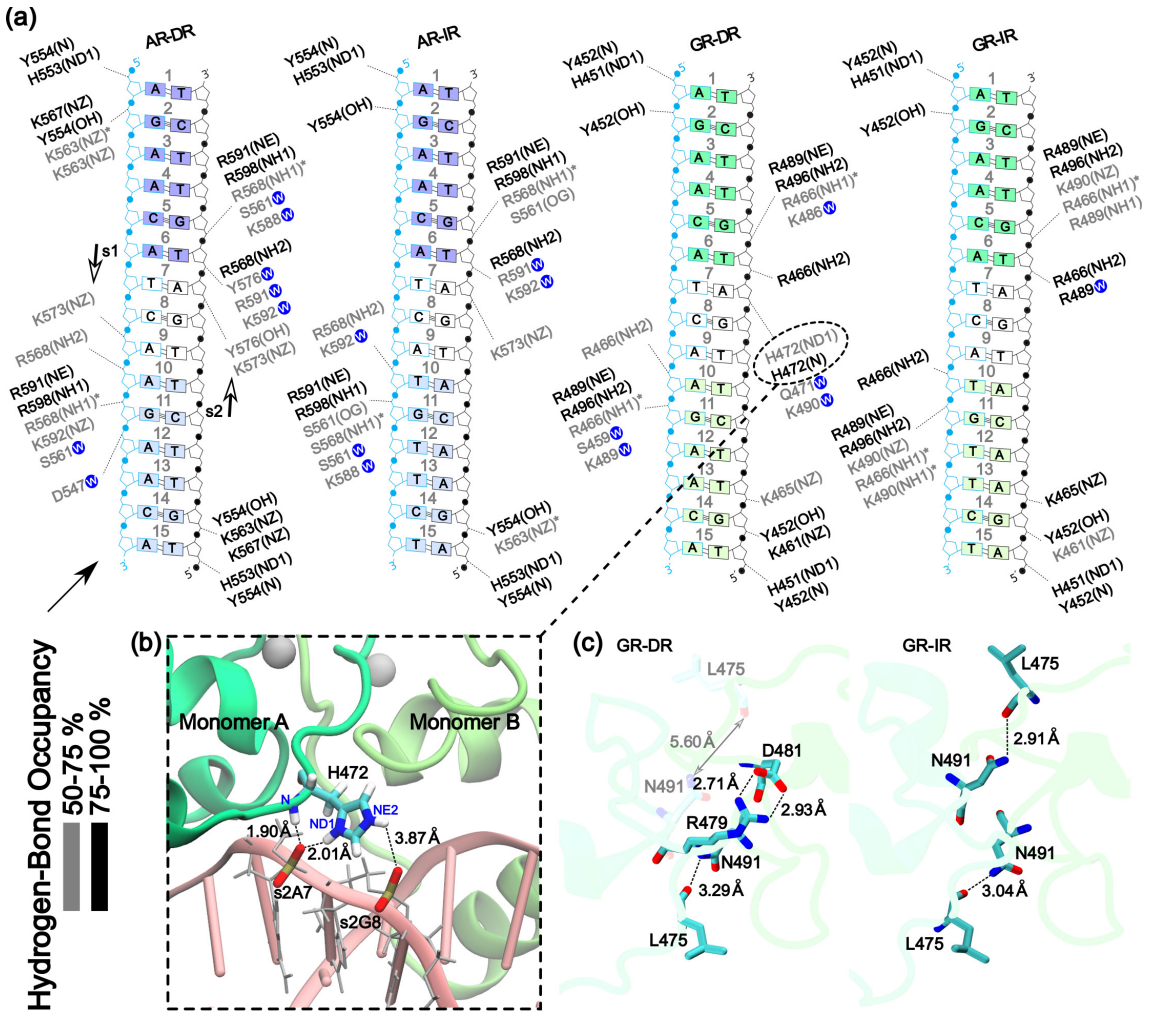


Figure 4.4: (a) Diagram of protein-DNA hydrogen-bond interactions. The nucleotides of core 15 bps DNA sequence are numbered from 5′-Hexamer (numbers: 1 to 6) to 3′-Hexamer (numbers: 10 to 15). The spacer region is highlighted with gray-colored boxes around the numbers of the bases (numbers: 7-9). The hydrogen bonds are categorized based on their occupancy, 50-75% (gray), and 75-100% (black). The water mediated hydrogen bonds are shown with blue circles. (b) Hydrogen-bond interaction of HSP472 with the phosphate group of nucleotide S2A7 in the GR-DR complex. (c) Hydrogen bonds in the dimerization domain of GR-DR and GR-IR.

two extra direct hydrogen bonds, s1G11 and s2G5, with residues S561 (in monomers A and B) are symmetrically formed in the AR-IR complex whereas no such direct hydrogen bonds can be observed in the AR-DR complex. This symmetry can be explained by the symmetric head-to-head conformation of the AR-DBDs with the palindromic IR sequence.

Comparing the hydrogen bond patterns between the GR systems shows that the GR-IR complex has stronger hydrogen bond interactions than the GR-DR. In particular, for the GR-IR model two more hydrogen bonds than in the GR-DR complex can be observed for each specific guanine residue, i.e. s1G11 and s2G5. Two further residues, s1A10 and s2T6, exhibit stronger hydrogen-bonded interactions with the protein in the GR-IR than in GR-DR complex.

In the GR-DR complex a significant exception is residue s2A7 in the spacer that forms strong interactions with residue H472 (in the monomer A) and also with residues Q471 and K490 (in the monomer A) mediated by water molecules. Interestingly, the interaction with residue H472 takes place through both its backbone and its side chain atoms, i.e. N and ND1. As shown in Figure 4.4(b), a weak hydrogen bond interaction between H472's side chain atom ND2 and s2G8's phosphate group atom O1P can be observed (together with a stationary $\sim 3.89 \pm 1.04$ Å distance between these atoms throughout the 500 ns of the simulation). The strong hydrogen bond interactions of the residue H472 with the spacer, in the GR-DR complex, inhibit the residues of that region to move freely as manifested by the root mean square fluctuation (RMSF) of the lever arm region that is considerably lower in the GR-DR complex than in the GR-IR complex (Figure B.2(a), shaded area in monomer A).

4.2.3 Protein-protein interactions

The hydrogen bond interactions between protein subunits are listed in Table 4.1. Our results indicate that the dimer interface of the AR-IR system forms different hydrogen bond patterns than those seen in the AR-DR. In particular, the inter-subunit hydrogen bond $S580_A$-$S580_B$, which is crucial for tight dimerization of the AR-DR complex [45], is not present in the AR-IR dimer interface and instead another head-on hydrogen bond $N582_A$-$N582_B$ is formed (Appendix B Figure B.13). GR which has a glycine, G478, instead of serine S580 in the AR, does not form the same inter-subunit hydrogen bond interaction. Neither does GR exhibit a hydrogen bond between its N481 residues (corresponding to the N582 in AR). The dimerization interactions of the GR-DR model are quite different from those in the other models. As listed in Table 4.1, the hydrogen bond interactions L577-N593(AR)/L475-N491(GR), which are conserved among the other systems, are missing in the GR-DR dimer interface. Instead, two strong hydrogen bonds, C476-R488 and R479-D481, from monomer A (as a donor) to monomer B (as an acceptor) are formed in the GR-DR dimer interface which are significantly weaker and only unidirectional in the GR-IR complex, see also Figure 4.4(c).

Table 4.1: Protein-protein hydrogen bond interactions. The star indicates that more than one hydrogen bond is formed simultaneously. *AB* and *BA* refer to the monomer A as donor and monomer B as acceptor and vice versa.

AR-DR	*AB*	*BA*	**AR-IR**	*AB*	*BA*
L577-N593	62%	44%	L577-N593	42%	50%
A579-T585	89%	93%	A579-T585	95%	85%
C578-R590	-	47%	C578-R590	-	66%
R581-D583	100%*	100%*	R581-D583	50%	100%*
S580-S580	80%	80%	N582-N582	76%	76%
S580-D583	-	48%	R581-T585	73%	-
GR-DR	*AB*	*BA*	**GR-IR**	*AB*	*BA*
L475-N491	-	-	L475-N491	59%	74%
A477-I483	89%	81%	A477-I483	91%	85%
C476-R488	86%	94%	C476-R488	-	66%
R479-D481	100%*	100%*	R479-D481	-	59%*

4.2.4 Complex conformation

4.2.4.1 Protein subunits

In order to estimate the structural change of each complex during the simulation, the median structures representing the first 20 ns and last 50 ns (of the total of 500 ns simulation time), respectively, were aligned with respect to each other and compared. As can be seen in Figure 4.5, the lever arm is the most variable domain whereas the initial and final conformations of the remainder of the systems are similar. A remarkable exception is the monomer A of the GR-DR model which exhibits a considerable tilt, $\sim 8.5^\circ$ with respect to the monomer B as well as the DNA. This tilt takes place not only in the lever arm but also in the upper α-helix of the monomer, H1, the Dim, and also the second zinc ion and its coordination sphere change their positions.

To evaluate the impact of the tilt of monomer A in the GR-DR model, we have determined the number of water molecules that are confined between the protein and the DNA spacer (schematically depicted in Figure 4.6(a). As shown in Figure 4.6(b), the number of confined water molecules is in both AR models higher than in the GR-IR complex, by ~ 8 water molecules. Similarly, in the GR-DR model, the number of confined water molecules is also significantly increased, by ~ 8 water molecules, compared to the GR-IR model. This can be understood as a consequence of the notable conformational change of monomer A, which forms a cavity in which extra water molecules can move into. This conformational change is further reflected in the dihedral angle formed by the four zinc ions in the dimer interface which is $\sim 4.4^\circ$ larger in the GR-DR model than in the GR-IR complex whereas the difference be-

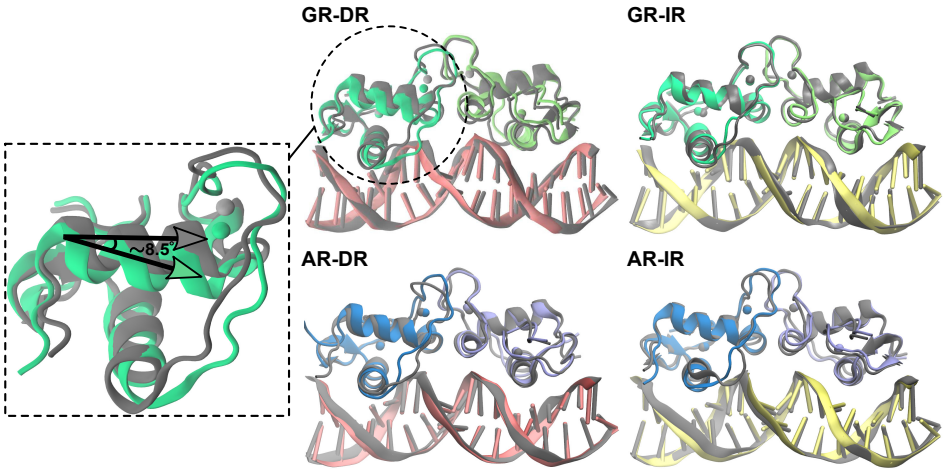


Figure 4.5: The 3D median structures of the complexes. In each system, the median structure of last 50 ns (colored) are aligned to its first 20 ns median structure (gray). The zoom in the left shows monomer A of GR-DR, highlighting the tilt angle between the median structures of the initial and final part of the simulation.

tween the AR models is $\sim 3.2^{\circ}$ (see Figure 4.6(c)). Associated with the larger dihedral angle is the distance between the second zinc ions (Zn2A and Zn2B) in the dimer interface which is significantly larger (by $\sim 2.7 \pm 0.2$ Å) in the GR-DR complex than in the GR-IR model. In the AR-IR model this distance is only larger by ~ 1 Å than in the AR-DR model (see Figure 4.6(d)).

The diffusion trend of the dimer interface, as quantified by the time series of the distances between the zinc ions, i.e. Zn2A and Zn2B, as well as the centers of mass (COM) of the dimerization loop domains (residues 474-496 and 576-598, for GR and AR, respectively) of all monomers, was calculated as time-averaged mean square displacement (TAMSD), see Appendix B Figure B.3. While this trend is flat (over the increasing lag time) for both AR-IR/DR complexes as well as for the GR-IR complex, the GR-DR complex exhibits an increasing diffusion mode for the Zinc ions distances as well as for the COMs of the dimerization loop domain (solid green line, Appendix B Figure B.3). Our results indicate that these dramatic increases of TAMSDs in the GR-DR are mainly due to the displacement of monomer A. These trends of diffusion suggest that the conformational change of monomer A in GR-DR and a response by monomer B are likely to continue on longer time scales.

Analysis of a second, longer (1.4 μs) MD simulation of GR-DR (named *run1*) shows that the distance between the COMs of the protein subunits gradually increases (see Appendix B Figure B.4(a)). Moreover, the monomer A-HS1 distance (red line in Appendix B Figure B.4(b)) suddenly increases and subsequently the monomer B-HS2 distance (blue line in Appendix B Figure B.4(b)) decreases. Similar to the conformational changes observed in the

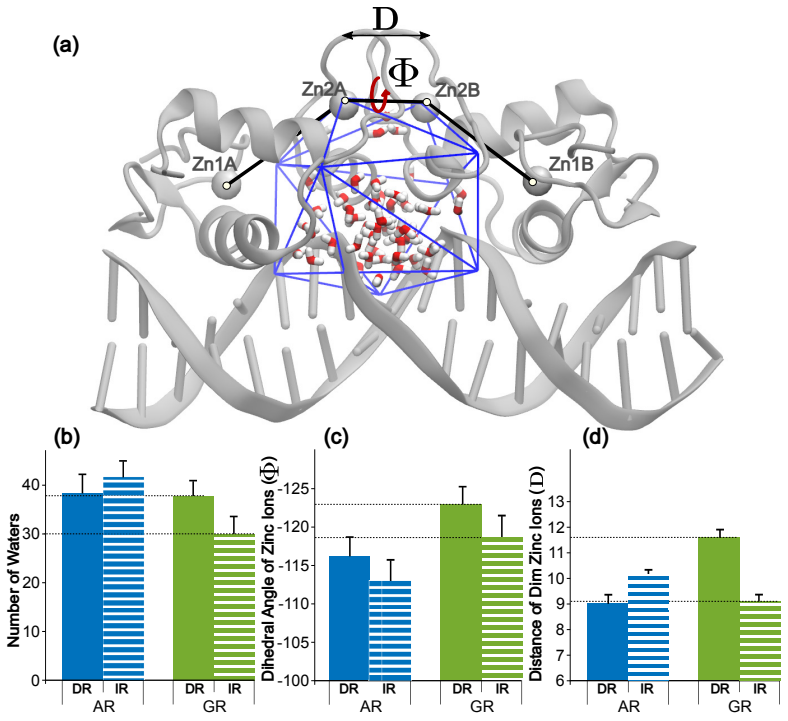


Figure 4.6: **(a)** Schematic structure of the polyhedron chosen to count the number of water molecules that reside between the protein and DNA in the spacer region. **(b)** Average number of water molecules inside the polyhedron. **(c)** Distance between zinc atoms in the dimerization region. **(d)** Dihedral angle defined by the four zinc ions in the dimerization domain.

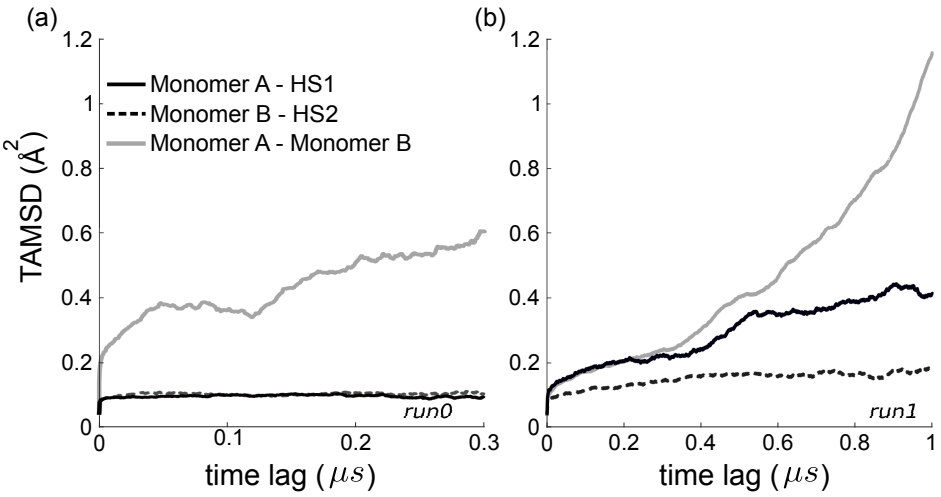


Figure 4.7: Time averaged mean square displacement (TAMSD) of monomer A-HS1 (solid black color), monomer B-HS2 (dashed black color), and monomer A-monomer B (solid gray color) distances of the GR-DR complex. **(a)** TAMSD's of *run0*, calculated within time-lag 0.3 μs.**(b)** TAMSD's of *run1*, calculated within time-lag 1 μs. Note that in *run0* simulation, for the calculation of the monomer A-monomer B distance, only those residues are considered that are involved in the dimerization interfaces, i.e. residues 474-496 (AR) and 576-598 (GR). In simulation *run1*, the COMs of all residues of monomer A and monomer B, respectively, are considered.

first MD simulation (*run*0), a distinct displacement of monomer A is reached after 1.4 μs in *run1* (see Figure B.4(c)), albeit with a different tilt axis. The TAMSD of the GR-DR subunits distances for both GR-DR simulations, i.e. *run*0 and *run1* are sketched in Figure 4.7. The TAMSD patterns clearly reveal that the diffusion of the monomer A-monomer B distances, in both runs, as well as the monomer A-HS1 distances in *run1* move far from a plateau during the examined simulation time.

4.2.4.2 Lever arm of GR

Our results further demonstrate that the lever arm of GR-DR switches during the simulation from its initial conformational state to a final state in which H472 is hydrogen-bonded to the spacer nucleotides (see Appendix B Figure B.5). This conformational change takes place via an intermediate state, which is characterized by the conformation of residue Y474. As detailed in Figure 4.8, at first, the residue Y474 leaves its initial conformation, I, and assumes an intermediate conformation, II. This step is accompanied with a flip of residue H472 towards the DNA. In this step, the hydroxyl group of Y474 (OH) forms a hydrogen bond interaction with s2A7(O1P). The pairwise distance of atoms of interest for these transitions are shown in Figure 4.8. Ultimately, H472 finds a conformation in which it can interact simultaneously

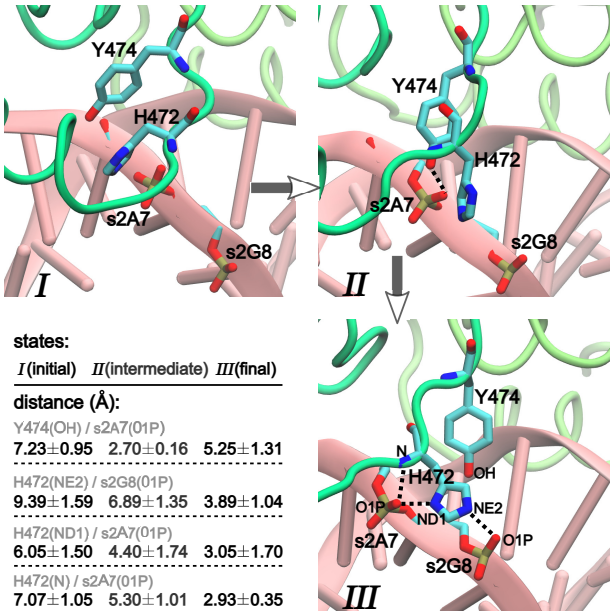


Figure 4.8: The three conformational states of the lever arm as observed in the simulation of GR-DR and corresponding atomic distances of the important residues.

with s2A7(O1P) and s2G8(O1P), while Y474 reaches its final conformation, resulting in state III (Figure 4.8).

The results of the longer simulation, *run1*, in contrast to the shorter one, *run0*, do not show any hydrogen bond interaction of residue H472 with the DNA spacer. Although this final state III of the lever arm conformation, observed in the *run0* GR-DR simulation, is not reached in the *run1* simulation, still in the last 330 ns a conformational change of the lever arm takes place that reaches a state similar to the previously observed intermediate state II. Figure 4.9 illustrates the different conformational states of Y474 for both simulations, *run0* and *run1*. Residue Y474 apparently leaves its initial state I and forms a hydrogen bond with atom O1P of s2A7. This intermediate state II lives about 5 ns in *run0*, followed by a rapid jump of this residue to its final state. Furthermore, the backbone dihedral angle ψ of residue H472 in the last 330 ns of *run1* shifts toward an angle that resembles that of the final state in *run0* (Appendix B Figure B.6). In addition, the ϕ angle of residue G470 transitions to the same value as in the final state of the *run0* while G470's ψ angle shifts to the value of the intermediate state II (Appendix B Figure B.7). Our result furthermore show that at simulation times 370 ns and 650 ns, for about $\sim$10 ns residue Y474 and to some extent also residue H472 reach the conformation of the final state (indicated by star symbols in Figure 4.9).

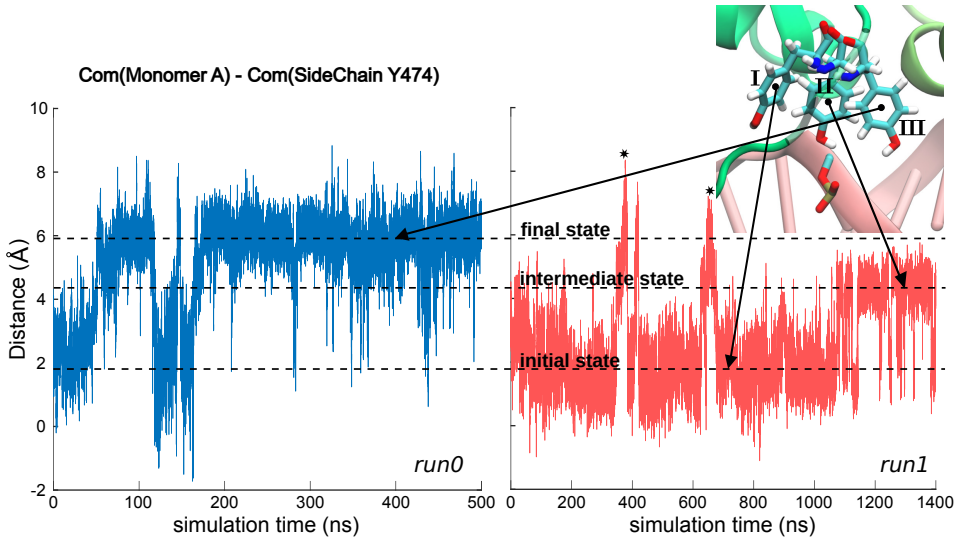


Figure 4.9: Three conformational states of Y474 as expressed by the distance of the COM of Y474 with respect to the COM of monomer A in the two simulations of GR-DR (left: *run0,* and right: *run1*). The stars in the time series of run1 indicate the short time periods in which Y474 reaches the conformation of final state III.

4.2.5 H472R mutant of GR

In order to further explore the importance of the residue H472 in GR DBD-DNA binding specificity, simulations of the mutant GR(H472R) in complex with IR and DR DNA elements were performed and analyzed.

Comparison of the median structures of the first 20 ns and last 50 ns of the total MD simulation reveals a conformational change similar to the one observed in the wild-type GR-DR complex, i.e. a tilt of monomer A (compare Figure 4.10(a) and Figure 4.5(GR-DR)). Analysis of the hydrogen bonds between the lever arm and the DNA in the mutant GR(H472R)-DR system furthermore shows that residue R472 (in monomer A) forms very similar interactions with s2A7 as well as with s2G8 (Figure 4.10(b)), as observed for residue H472 in the wild-type GR-DR system. Also the increase of the Zn2A-Zn2B distance in the GR(H472R)-DR to 11.51±0.26 Å, is close to the corresponding increase in Zn2A-Zn2B distance observed in the wild-type GR-DR system (11.61±0.29 Å). As Figure 4.10(c) indicates, a distorted conformation can also be observed at the Dim interface of the GR(H472R)-IR system. This distortion is a result of the residue R472 in monomer A whose side chain is pointing toward the DNA backbone and forms a strong hydrogen bond interaction with nucleotide s2T6 (see Figure 4.10(d)). Furthermore, the conformation of the lever arm, as sketched in the Ramachandran plots of residues G470 to Y474 (Appendix B Figures B.8-B.12), is very similar in the wild-type GR-DR and mutant GR(H472R)-DR and GR(H472R)-IR systems whereas the conformation of

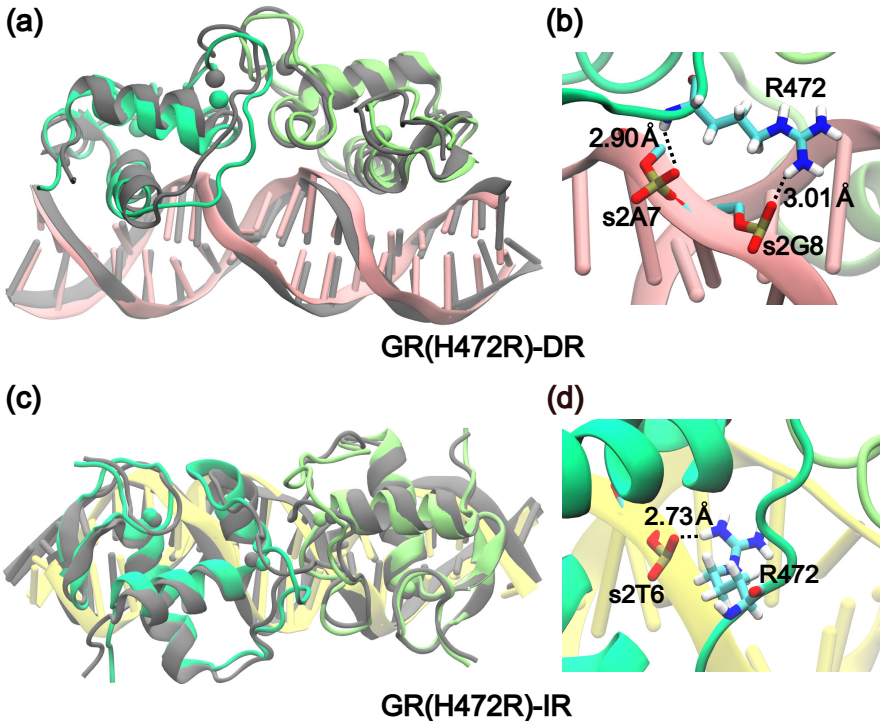


Figure 4.10: Conformational changes of the H472R mutant of **(a)** GR-DR and **(c)** GR-IR complexes illustrated by comparing the median structures obtained from the first 20 ns (gray color) and last 50 ns (colored structure) of the 300 ns MD simulation, respectively. Interaction of R472 with DNA backbone, in the **(b)** GR-DR and the **(d)** GR-IR complexes. The dashed line indicates the hydrogen bond connection.

the lever arm is rather different in the wild-type GR-IR system.

R472 thus affects the conformations of GR(H472R)-IR and GR(H472R)-DR complexes as does H472 in the wild-type GR-DR complex, that is formation of a strong hydrogen-bonded interaction between the positively charged residue 472 and the DNA resulting in a distorted protein dimer and ultimately in an inactive complex.

4.3 Discussion

Both the AR and GR are capable to bind the CREs. However, despite the high resemblance of the DNA elements, their genome-wide distributions are distinct [116, 124] and only for AR a significant affinity for DR sequences (AREs) has been reported [45, 106].

AR binds more strongly to DR than to IR

AR-DBDs interaction mode depends on its bound elements, i.e. IR or DR [105]. Our result show that the hydrogen bond interactions of the AR bound to the DR sequence are significantly stronger than those in the AR-IR system (Figures 4.4(a)). Considerably, these stronger interactions in the AR-DR complex not only occur in the HS1 but also in the HS2 (that is the "repeat") and the spacer, suggesting the AR binds more tightly to a DNA with a DR sequence. This is in agreement with the higher response of AR to AREs (that is DR) than to CREs (i.e. IR) [105].

Our data furthermore show different interaction patterns of the AR's dimerization interface, depending on the response element bound (Table 4.1 and Appendix B Figure B.13). In particular, we observe that cross-subunit interaction S580-S580, which is discussed as an important factor in head-to-head AR-DR dimer stability [45, 108], is broken in the AR-IR and another N582-N582 hydrogen bond interaction is formed instead. This formation of an alternative inter-subunit hydrogen bond suggests a conformational flexibility in the two AR protein monomers with respect to each other, allowing it to adapt when binding either DR or IR and remain an intact dimer. This is in agreement with AR's experimentally observed ability to bind and act on DR and IR elements [106].

AR and GR differ in entropic effect upon IR binding

Comparison of the AR/GR-IR systems shows almost similar direct hydrogen bonding patterns. However, the water mediated hydrogen bond interactions in the AR-IR protein-DNA interface are stronger than those in the GR-IR complex (Figure 4.4(a)). Moreover, the number of water molecules which are confined in the protein monomers-spacer region of AR-IR is substantially larger than in the GR-IR (Figure 4.6(a,b)). Hence, the stronger/additional water-mediated interactions in the AR-IR complex are counteracted by a loss in entropy since more water molecules are confined. This is in agreement with the recently published finding [116]

that the DNA binding affinity of the AR is more enthalpy-driven whereas the GR binding affinity is more entropy-driven. Indeed, our simulation data show less confined water molecules in the GR-IR than the AR-IR system. Moreover, the DNA conformation in the GR complex is more different (as manifested by a higher deformation energy, Figure 4.3) from free DNA than it is in the AR complex. Since the interactions are, however, stronger in AR-IR than in GR-IR the free energy required to reach such a conformation is provided by entropic effects.

Binding to DR sequence destabilizes GR dimer

The most striking finding in this study is the relatively unstable conformation of the GR dimer when bound to the DR sequence, in agreement with experimental evidences that show no affinity of the GR to bind to the DR [109–111]. The hydrogen bond analysis shows that the GR-DR complex lacks a number of interactions in both hexameric sites that are, however, present in the GR-IR complex, (Figure 4.4(a)). Instead, in the GR-DR complex, a very strong hydrogen bond is formed between the residue H472, located at the lever arm of monomer A, and the DNA. This intersection is achieved in a three-step process, which is associated with conformational changes of Y474, G470, and H472 (Figure 4.8) that ultimately lead to strong hydrogen bonds of H472 with spacer nucleotides s2A7 and s2G8 (Figure 4.4(b)). Associated with this three-step binding event is the conformational change of monomer A as exhibited by a tilt of the entire subunit and an increase of the distance between the two dimerization domains (Figure 4.5), resulting in an abnormal conformation in the DNA as well (see Figure 4.2(a,b) and Figure 4.3(b)).

The diffusion pattern of the monomer A (Figure 4.7) furthermore suggests that the rotational displacement of this monomer in the GR-DR complex has not reached its dynamical equilibrium in the course of the simulations. Rather, on longer time-scales further dislocation and hence destabilization of the complex can be envisaged.

The conformational movement of the GR-DR toward an unsteady configuration is confirmed by a second MD simulation of the GR-DR (*run1*), clearly showing that subsequent to the dissociation of the protein monomers, the orientation of monomer A is considerably changed with respect to that of monomer B and DNA (Figure 4.5). In this second simulation, the monomer A is tilted around a different axis than in the other run, and the fully deformed state in which H472 forms two strong hydrogen bonds to the spacer DNA is not quite reached.

Nevertheless, the conformational changes of the lever arm in GR-DR obviously show that the system, in the last 330 ns of *run1*, jumps to a conformation (of residues G470, H472, and Y474) similar to the intermediate state of the GR-DR lever arm observed in the simulation *run0*. Moreover, short visits ($\sim$10 ns long) to the final, deformed-state conformation are observed in this simulation (at 370 ns and 650 ns simulation time, see Figure 4.9, star signs). It is therefore conceivable that the lever arm eventually, i.e. on a longer time scale, assumes

the conformation of the final, fully deformed state.

H472R mutation stresses importance of residue 472 in GR binding to DNA

The importance of H472 stabilizing the deformed conformation by strong hydrogen bonds to the spacer DNA and of the lever arm's dynamics in reaching such a state is further emphasized by our observations for the H472R mutant. A similar conformational change of monomer A as observed in the wild type GR-DR complex occurs also in the GR(H472R)-DR mutant complex (Figure 4.5, top left and Figure 4.10a). During the course of the simulation the lever arm of this mutant complex rapidly (within $\sim$ 10 ns) adopts a conformation which is quite similar to the final state of the lever arm (state III) seen in the wild-type GR-DR simulation, also with strong hydrogen bonds of R472 to s2G8 and s2A7 that stabilize the deformed conformation (Figures 4.4(b), 4.10(b), and Appendix B Figure B.8-B.12).

Non-favorable conformational changes are observed also in the simulations of GR's H472R mutant complexed to the IR sequence (GR(H472R)-IR). In this complex the Dim interface is significantly altered with respect to the GR-IR wild-type complex (Figure 4.10(c)). This distorted conformation is in agreement with the considerably diminished GR activity of the H472R mutant to its (CRE) binding sequences [43]. The longer arginine residue R472, compared to H472, allows the formation of strong hydrogen bonds to the DNA (phosphate group of s2T6, Figure 4.10(d)). In the mutant GR-IR complex not only the conformation of R472 itself, but also of other residues in the lever arm (G470 and Y474) are different compared to the wild-type GR-IR complex.

In AR, a non-charged residue (Q574) is located at the position corresponding to H472, but flanked by two lysine residues (K573 and K575). In both complexes, AR-DR and AR-IR, the side chain of K573 reaches the DNA spacer without a conformational change of the lever arm or the protein subunits, forming moderately strong hydrogen bonds (see Figure 4.4(a)). In contrast to GR, these distinct lever arm-DNA contacts do not provide a means of sequence discrimination for AR.

Role of the lever arm

The reconfiguration of the lever arm of monomer A observed in the H472R mutant complexes of GR (Appendix B Figures B.8-B.12) together with the conformational dynamics observed in the wild-type GR-DR complex is clearly associated with the displacement of monomer A. The lever arm can thus be understood as a modulator of GR–DNA binding, providing a link between the bound DNA sequence and the Dim domain of the protein, fully supporting the idea that the lever arm plays an essential role in DNA-Dim allosteric communication [43, 100, 111, 113].

It is interesting to note that all the conformational changes take place in the lever arm region of the monomer A and thus affect only monomer A, although the differences in repeat

sequence are located in the second half-site that is bound to monomer B.

Analysis of the DNA geometry (grooves widths and base pair parameters) in the GR complexes shows that the different sequences in HS2, IR or DR, result in different conformations in the spacer region of the DNA. Since the spacer region is the direct binding site of the lever arm, conformational differences in this region are directly conveyed to monomer A and ultimately to the dimerization domain. When bound to the DR sequence, the lever arm of monomer A, that is the subunit interacting with a cognate sequence, can be regarded as being closer to the DNA, including the spacer, than that of monomer B. Thus, changes in the second half-site are communicated to monomer A via the spacer region and the lever arm.

4.4 Conclusion

In this work, we applied all-atom MD simulations to analyze the DNA-binding domains of the hormone receptor proteins AR and GR in complex with DR and IR DNA sequences, respectively. This allowed us to observe individual dynamical events and their relation to multiple interactions between protein and DNA as well as between the protein monomers.

Our data show that the AR DBD-DNA interactions vary depending on binding IR or DR DNA elements, in agreement with experiment [105]. Although AR has affinity to both IR and DR sequences with similar conformation, it interacts more strongly with DR. In contrast, GR does not form a stable complex with a DR sequence explaining the lack of affinity of GR on such response elements. Our study furthermore suggests that reconfiguration of the lever arm towards the DNA spacer, through three states of Y474 and mainly H472, and consequently formation of a strong hydrogen bond interaction of residue H472 with the spacer, significantly impairs the GR-DR dimerization interface as well as leading to a repositioning of monomer A towards an unstable state. The same behavior is observed by H472R mutation in the GR-DR stressing the importance of direct contacts of a positively charged residue 472 with the DNA backbone in competition with protein-protein interaction in the dimerization domain. Also in the H472R mutant complex of GR-IR these strong interactions of R472 with the DNA result in a distorted protein dimer, in agreement with the significantly reduced activity compared to the wild-type GR [43]. The lever arm, on which residue H/R472 is located, connects the DNA binding helix to the dimerization domain. Conformational changes of the lever arm, upon directly binding to the DNA are thus directly communicated to the dimerization domain.

Direct contacts of GR's lever arm to the DNA spacer are observed only in the deformed, and thus likely inactive conformation of the GR-DR complex. In contrast, in AR, the cooperative conformational changes observed in the GR-DR complex, are not necessary for AR's lever arm to form moderately strong contacts with both IR and DR response elements in stable

complexes. Taken together, the different DNA binding modes of AR and GR are responsible for their different specificity for direct and inverted repeat response elements.

Chapter 5

Molecular Dynamics Simulations of a Chimeric Androgen Receptor Protein (SPARKI) confirm the Importance of the Dimerization Domain on DNA Binding Specificity

This chapter materials are published in the article:

"Bagherpoor Helabad M, Volkenandt S and Imhof P (2020) Molecular Dynamics Simulations of a Chimeric Androgen Receptor Protein (SPARKI) Confirm the Importance of the Dimerization Domain on DNA Binding Specificity. Front. Mol. Biosci. 7:4."DOI

In 2007, an *in vivo* study done by Kris Schauwaers et al. generated a chimeric receptor, termed SPARKI (SPecificity-affecting AR KnockIn), in which 12 amino acids of AR in its second zinc-binding domain were replaced by those of GR (Figure 5.1(a,c)) [47]. *In vitro* studies have shown that swapping this second zinc-binding motif between the AR and GR leads to the loss of affinity of this chimeric receptor with a DR-like motif [110, 125]. Consistently, the *in vivo* experiment exhibited a reduced affinity of the SPARKI receptor for DR-like elements whereas for IR-like elements it showed similar or even better binding affinity than AR [47]. The lack of the SPARKI system's ability to bind to DR-like response elements was also confirmed by a later *in vivo* study, done by Biswajyoti Sahu, et al. at 2014 [49]. Interestingly, this study shows that for DR-like elements, which were selectively enriched by wild-type AR, there is a well conserved 5′-hexamer (HS1, Figure 5.1(b)) but not a strin-

gent 3′-hexamer (HS2) sequence conservation. In contrast, binding of both wild-type AR and SPARKI to IR-like elements requires a specific HS2 sequence [49]. Moreover, *in vitro* assays show the high-affinity of AR and GR receptors to HS1, due to its highly conserved sequences [103]. It is speculated that due to the high-affinity of the two subunits in the AR dimer, this receptor could bind to a more diverse HS2 than the GR could. For instance, it is shown that the thymine (T) next to guanine (G) in HS2 of the IR elements is a highly conserved base in the response elements of SRs. This specific T is not required for AR, allowing this receptor to bind to DR-like elements which have an adenine (A) in that position [49, 126–130]. However, it is not yet clear how the high affinity of AR-DBD to DR-like response elements, which leads to strong interactions in the protein's dimerization interface, is influenced by (more diverse) HS2 elements. Moreover, the distinct binding of AR(DBD)-DR (or IR) and GR(DBD)-IR is still not well understood. The SPARKI is an outstanding model that could explain the distinct regulation of AR-specific responses with respect to those which can be regulated by GR as well.

In this study, by employing all-atom molecular dynamics simulations, we investigate the factors that lead to a different binding of AR and GR receptors to DNA response elements. In this regard, we simulated six protein-DNA complexes consisting of the DNA binding domains of wild type AR and GR, bound to a DNA sequence with IR and DR, respectively, and SPARKI models (with both IR and DR elements) made by AR and GR mutation. Our MD simulations allowed us to determine the significant dynamics of these receptor's DBD-DNA interface. These results suggest a loss of affinity of the chimeric proteins, i.e. SPARKI, to DR sequences and a strong affinity for IR sequences. Furthermore, our data reveal that the "weaker" dimerization interface interactions in the IR complexes, compared to the AR-DR complex, allows those dimeric proteins to be properly accommodated on IR sequences.

5.1 Molecular dynamics simulations: systems and protocols

5.1.1 Structural Models

We have constructed two atomic models of the SPARKI receptor, one based on the structure of the AR-DNA complex (1R4I) and one on the structure of the GR-DNA complex (1R4R) (see also chapter 4, section 4.1). In the AR-based model, termed SpAR, residues in the second zinc-binding motif of AR that differ from GR (highlighted in green in Figure 5.1(a)), were replaced with the corresponding residues of the GR protein, as in the experimental mutation [47]. These residues are located at the dimerization interface (see Figure 5.1(a) and (c)). The second model, termed SpGR, is based on the GR protein in which the residues of the first zinc-

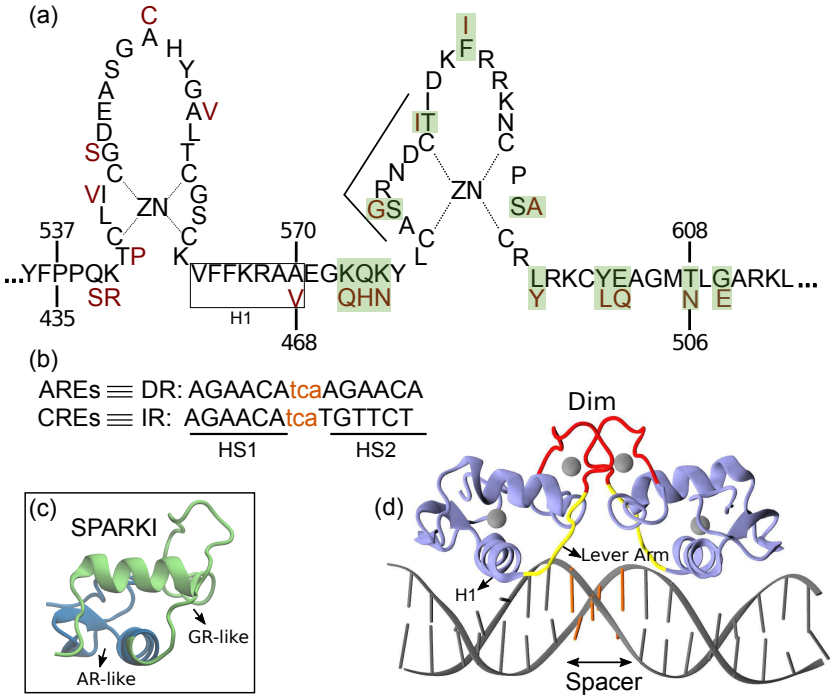


Figure 5.1: (a) Schematic overview of the DNA binding domain (DBD) sequences in the androgen receptor (AR) and glucocorticoid receptor (GR) protein with corresponding residue numbers above and below, respectively. The amino acids colored in dark red are those elements of the GR-DBD that differ from the AR-DBD sequence. The other amino acids are the same in the AR- and GR-DBD. The amino acids shown with green shadow are those elements in AR that are replaced with residues from GR in order to make Sparki [47]. (b) DNA sequences for direct (DR) and inverted repeats (IR). The non-capital letters are the spacer base-pairs, colored in orange. (c) Schematic 3D structure of one monomer of Sparki-DBD, regions colored in green and blue are those subdomains that are GR- and AR-like, respectively. (d) The 3D structure of the GR- DBD/DNA complex (pdb ID: 1R4R). A similar structure exists for the AR-DBD/DNA complex (pdb ID: 1R4I). The lever arm and dimerization domain (Dim) are shown in yellow and red, respectively. The spacer region of the DNA is colored with orange.

binding motif of GR that differ from AR, which are part of the DNA-binding interface, were mutated to those of AR. The resulting sequence of the proteins in both Sparki models, SpAR and SpGR is thus identical, however, their initial structures differ, since these are based on two different crystal structures. Both SPARKI models were furthermore modeled in complex with both DNA sequences, DR and IR, respectively. Therefore, a total of 4 models, i.e. SpAR-DR, SpAR-IR, SpGR-DR, and SpGR-IR have been simulated.

5.1.2 Molecular Dynamics Simulations

The systems were solvated with ∼23000 water molecules in a cubic box of ∼ 90×90×90 $Å^3$ and a number of sodium ions were added to neutralize the systems. The CHARMM-27 force field [91, 131] and the TIP3 water model were used in the simulations [132]. Long-range electrostatic interactions were treated by the particle mesh Ewald method via a switch function with a cutoff of 14-12 Å and employing periodic boundary conditions [94]. The systems were energy minimized for 5000 steps (conjugate gradient with an energy tolerance of 10^{-4} kcal/mol), followed by a molecular dynamics (MD) simulation of 30 ps (time step of 1 fs) to heat the system by velocity scaling (with harmonic constraint on all heavy atoms, by force constant 10 kcal·mol^{-1}·$Å^{-2}$). Then, 100 ps of MD relaxation (in NPT ensemble) at target temperature (300 K) and time step 1 fs were computed. Langevin dynamics with a damping factor of 1 ps^{-1} have been used for temperature control [63]. The Nosé–Hoover Langevin pressure control, with piston period of 200 fs and a damping time of 100 fs, have been used in order to maintain the pressure at 1 bar [133]. After the equilibration phase, three 100 ns MD replicas (with different initial velocities) for each Sparki systems were carried out (time step of 2 fs). From those, one run per system was chosen for longer simulation, based on the calculated root mean-squared deviation (RMSD), see Appendix C Figure C.1 (this is same procedure that we did for the wild type AR and GR simulations, explained in chapter 4). These longer MD simulations were carried out for 900 ns for the SPARKI systems and for 500 ns for AR-DR and GR-IR, respectively, and saved at 2 ps intervals. In all simulations, the terminal DNA base pairs were restrained (centered around 3 Å between the centers of mass of the respective bases) by a harmonic potential with a force constant of 20 kcal/mol in order to decrease the edge effects. The MD simulations were run using version 2.10 of NAMD [117].

5.2 Results

The results are organized to first present a comparison of the overall structure of the complexes. This is followed by an analysis of the proteins, first, in terms of flexibility and an estimate of

their entropies in the different complexes. Then, the protein-protein interactions between the two subdomains are investigated. Subsequently, the conformation of the two DNA sequences in the different complexes is analyzed. Finally, the hydrogen-bond interactions between the proteins and the DNA are reported.

In this chapter, in order to make a comparison between the SPARKI systems and wild type AR and GR systems, some of the MD simulations results of the systems AR-DR and GR-IR (exhaustively explained in chapter 4) are discussed here.

5.2.1 Median Structure

In order to estimate the overall structural change of each complex during the simulation, the median structures representing the first 100 ns and last 100 ns (of the total of 900 ns simulation time for SPARKI models), respectively, were aligned with respect to each other and compared. As can be seen in Figure 5.2, the lever arm is the most variable domain whereas the initial and final conformations of the remainder of the systems are similar. Remarkable exceptions are the monomer A, located at the first half-site, and the Dim interface of the SpGR-DR model, which exhibit a considerable distortion. In this model, a conformational change takes place not only in the lever arm but also in both zinc-binding subdomains where the zinc ions, together with their coordinating ligands, change positions. The distances between different domains/subdomains of protein-DNA complexes are listed in Appendix C Table C.1. The SpGR-DR system exhibits a larger distance between the receptor's dimer interfaces as well as between the respective zinc ions of the two subunits, than the other systems. The simulations of the SpAR-DR model, which represent the same system but were started from a different initial structure, in contrast, do not exhibit a distortion of the Dim interface, Accordingly, the distance between the two monomeric subunits in this model are shorter than in the SpGR-DR model.

5.2.2 Root mean square fluctuations (RMSF)

Figure 5.3 shows the per-residue root mean square fluctuations (RMSF) of protein monomers for all the systems. As can be seen in this figure, the lever arm corresponding to residues 571-576(AR, SpAR)/469-474(GR, SpGR) is the most fluctuating region in all models. Comparison of fluctuations between monomer A and monomer B shows almost similar fluctuations of the protein residues in all systems, except for SpGR-DR. The IR complexes, thought, exhibit higher flexibility than the DR complexes in lever arm region, i.e. residues 469–474 or 571–576 in GR or AR numbering, respectively. SpGR-DR exhibits particularly high fluctuations of the protein residues, especially in monomer A; higher than the fluctuations of monomer A

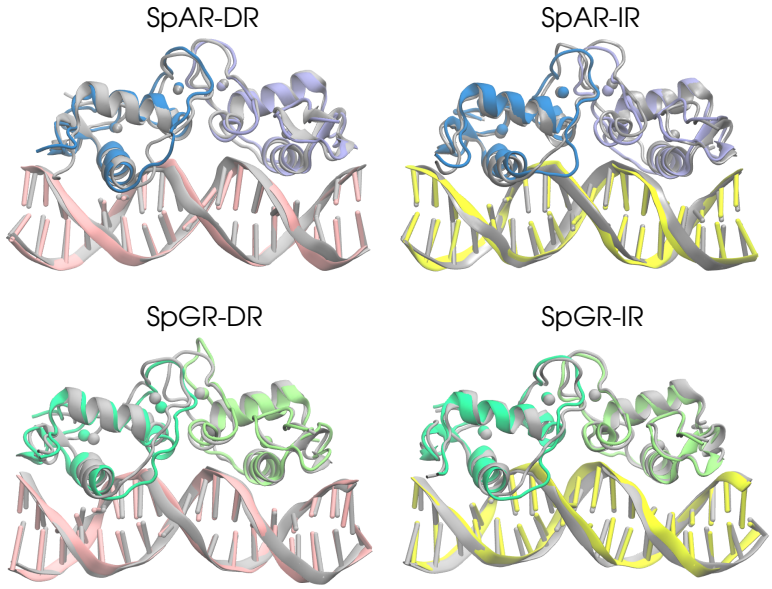


Figure 5.2: The 3D median structures of the complexes. In each system, the median structure of the last 100 ns of simulation (colored) is aligned to the median structure of the first 100 ns simulation (gray).

in any of the other systems. Monomer B of SpGR-DR, however, shows larger fluctuations than the other systems only for the residues situated in dimer interface, i.e. 576-581(AR, SpAR)/474-479(GR, SpGR). Of note, in the SpGR models, residues in the dimer interface are directly modeled, that is without in silico mutation, from the crystal structure of the wild-type GR protein and may therefore represent a GR-like conformation.

5.2.3 Entropy estimation

As can be seen from Figure 5.4, the estimated entropy of SpGR-DR and SpGR-IR are higher than those computed for SpAR-DR and SpAR-IR, respectively. This is the case for both entropy estimation methods. Both AR-DR and GR-IR exhibit rather similar values in entropy, although the two proteins are in complex with different DNA sequences. Comparison of only DR or IR complexes, respectively, shows higher entropy values for the Sparki models than for the respective wild-type complexes. Among the chimeric Sparki models, SpAR does not exhibit a significant difference in entropy when complexed to DR or IR sequence, whereas SpGR shows a significantly higher entropy in the DR complex compared to the IR complex.

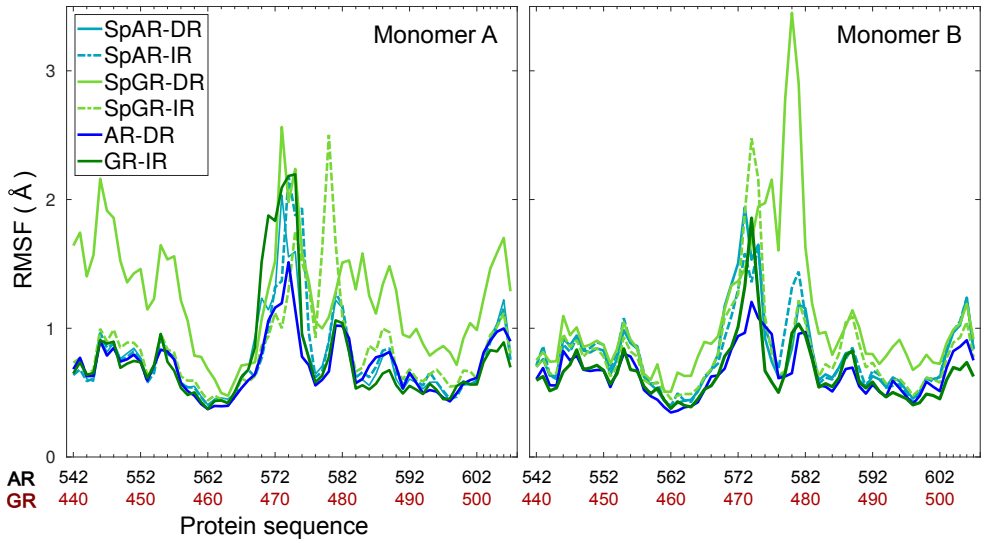


Figure 5.3: Per-residue root mean square fluctuations of Cα atoms of the protein (monomer A&B) for all systems.

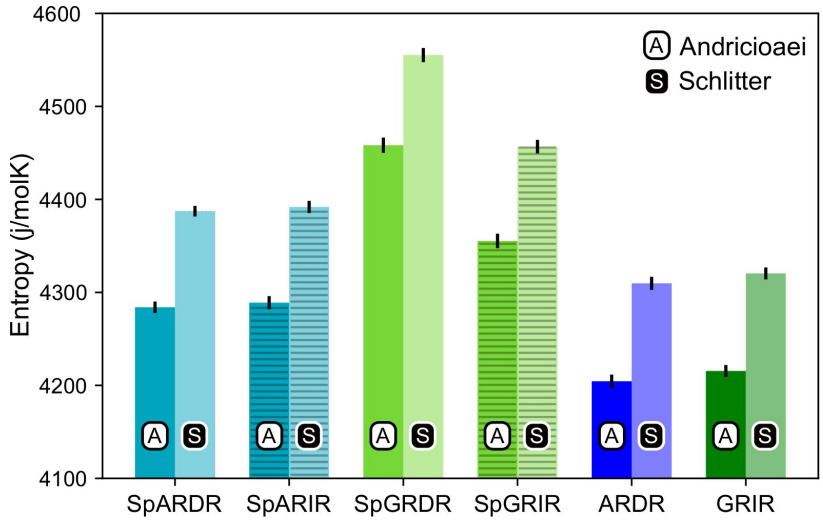


Figure 5.4: Entropy estimates for the proteins of all complexes. The first and second columns, shown with black **A** and white **S** are the entropy values estimated with the Andricioaei and Schlitter models, respectively.

Table 5.1: Protein-protein hydrogen-bond interactions. The star indicates that more than one hydrogen bond is formed simultaneously. AB and BA refer to the monomer A as donor and monomer B as acceptor and vice versa. Here, the hydrogen-bond interaction occupancies below 40% are considered as weak interactions and are therefore not listed.

	AR-DR		**SpAR-DR**		**SpAR-IR**	
	AB	BA	AB	BA	AB	BA
L577 - N593	62%	44%	72%	70%	51%	55%
A579 - I585	89%	93%	95%	91%	95%	88%
C578 - R590	-	47%	60%	-	71%	52%
R581 - D583	100%*	100%*	-	-	-	-
S580 - S580	80%	80%	-	-	-	-
S580 - D583	-	48%	-	-	-	-
	GR-IR		**SpGR-DR**		**SpGR-IR**	
	AB	BA	AB	BA	AB	BA
L475 - N491	59%	74%	-	90%	71%	51%
A477 - I483	91%	85%	81%	-	92%	66%
C476 - R488	-	66%	-	-	-	-
R479 - D481	-	59%*	-	-	-	-

5.2.4 Protein-protein hydrogen bond interactions

The hydrogen bond interactions between the protein subunits are listed in Table 5.1. Note that these results for the AR-DR and GR-IR systems have been already given in chapter 4, however, in order to facilitate comparison, we have listed them here again. Our results indicate that the dimer interface of the AR-DR system forms more strong hydrogen-bond interactions than those seen in the SPARKI systems and in the GR-IR. In particular, the inter-subunit hydrogen bond $S580_A$-$S580_B$, which has been discussed to be crucial for tight dimerization of the AR-DR complex [45], is not present in the other systems. Furthermore, a strong interaction of R581-D583 can also be seen in AR-DR, but not in the other systems. Two interactions, L577-N593(AR, SpAR)/L475-N491(GR, SpGR) and A579-I585(AR, SpAR)/A477-I483(GR, SpGR), exist in all the systems, in both directions, that is from monomer A to monomer B (AB) and vice versa (BA). However, in the SpGR-DR, only a one-sided of these interactions is formed, indicating a weaker dimer interface interaction of the SpGR-DR than in the other systems. Moreover, the dimer interfaces of the SpAR complexes exhibit stronger hydrogen-bond interactions than the SpGR models. An extra interaction of C578-R590 can be seen in SpARs that is not present in SpGRs. This extra interaction is also observed in the AR-DR complex, based upon which the SpAR-DR model has been built. The dimerization interactions of the GR-IR model also exhibit two moderate and one-way (BA-side) hydrogen-bond interactions C476-R488 and R479-D481 that are not present in SpGR models.

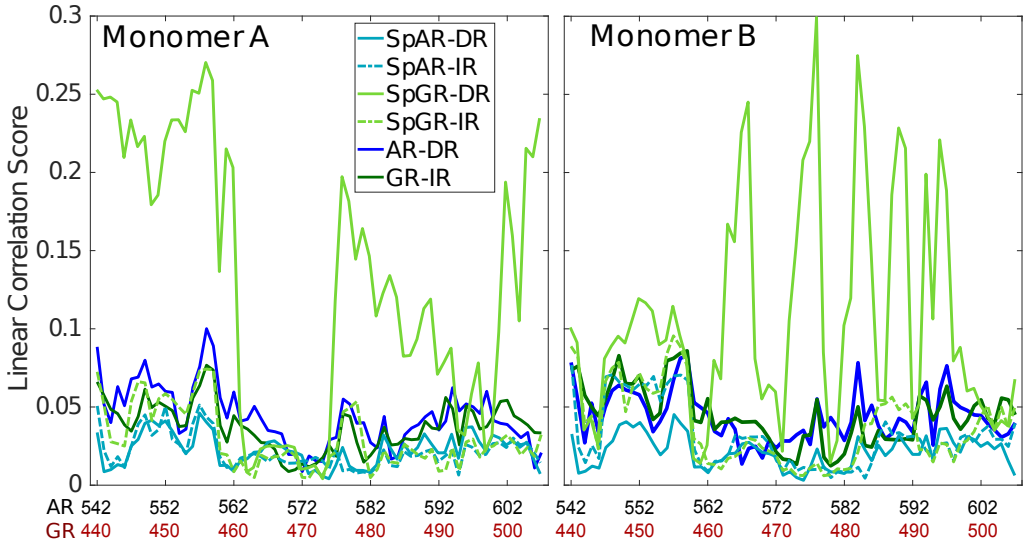


Figure 5.5: Correlation score per residue, computed for intra-domain correlations with $r_{ki} \geq 0.4$.

5.2.5 Linear correlation score

In order to capture how the protein residues in each monomer are influenced by other residues of that monomer, the linear correlation score has been calculated for all the systems. (Linear correlation scores calculated for the first 100 ns and middle 100 ns of trajectories of the SPARKI systems are shown as supplementary material, see Appendix C Figure C.13). As can be seen in Figure 5.5, almost all the residues show a similar magnitude of correlation score in all the systems, except for SpGR-DR. This model exhibits considerably higher correlation score values, in both protein monomers, than any of the other models. This indicates that the fluctuating motion of each residue is highly dependent on the rest of the residues in that protein. Any local conformational change, as observed for the lever arm and the Dim of SpGR-DR, as visualized by the median structures (see above), does not only affect the neighboring residues but also distal domains of the protein and thus has a more global effect. Moreover, for SpGR-DR the correlation score increases during the simulation, corresponding to an increase in conformational change of the monomers in this model, see Appendix C Figure C.13.

5.2.6 DNA conformation

To study the impact of the DBD of the receptors on their respective DNA structure, the local geometrical parameters of DNA, i.e. inter- and intra-bp parameters (Appendix C Figures C.5 to C.10), major- and minor-groove widths (Figures 5.6), and helical axis bending (Figure C.4) were calculated for the last 100 ns of the trajectories. For the SPARKI systems, the changes of

these parameters in the course of the simulations were also considered (Appendix C Figures C.2 and C.3) and are discussed.

The DNA grooves of the IR complexes differ from those of DRs. Interestingly, these differences can not only be observed in the second hexamer, which is expected due to the different DNA sequence, but also in the spacer and in the first hexamer in the IR complexes, see Figure 5.6. For instance, the major groove at position C_8, in the spacer region, is narrower in the IRs than in DRs. Also, a narrower major groove at positions C_5-A_6 (in HS1) can be observed in Sp(AR/GR)-IR compared to SpAR-DR or AR-DR. The DNA of both SPARKI-IR systems exhibits very similar conformations. This can be seen in almost all DNA parameters, see Appendix C Figures C.5 to C.10.

The DNA parameters in both SPARKI-IR complexes show some differences from the GR-IR parameters. The minor groove of Sp(AR/GR)-IR at positions between A_4-T_7 (in HS1) is narrower than that in the GR-IR, see Figure 5.6. Also, the DNA of the GR-IR complex shows higher bending than the DNA of the Sp(AR/GR)-IR complexes, Appendix C Figure C.4(b,d). Since the DNA sequence is the same in all IR complexes, the observed differences in the DNA conformation can be attributed to the interaction with the different proteins.

In contrast to the two SPARKI-IR complexes, all DNA parameters of the SpAR-DR complex and the SpGR-DR complex represent conformations that are considerably different from AR-DR, see Figures 5.6 and Appendix C Figures C.4 to C.10. SpAR-DR and SpGR-DR, moreover, show differences between some of their DNA parameters. For instance, in SpGR-DR the HS2 has a wider major and narrower minor groove and HS1 has a considerably wider minor groove than in SpAR-DR. Furthermore, the DNA helical axis bending is higher in SpGR-DR than in SpAR-DR (Figure C.4). In the two SPARKI-DR models not only the DNA sequence is the same, but also the residues of the protein. The different DNA conformations may also be attributed to different interactions with the (same) proteins, representing different (metastable) binding modes due to different initial starting conformations.

In the SPARKI-IR systems, the first hexamer exhibits a narrower major groove than the second hexamer whereas the opposite is observed for the SpAR-DR and AR-DR systems, see Figure 5.6. Interestingly, the position T_{12}, in the second hexamer, seems to have an important role in the IR complexes. For most IR complexes the dinucleotide $G_{11}T_{12}$ shows an extreme value which is not the case in the DNA parameters of the DRs with $G_{11}A_{12}$ at this position (see Appendix C Figures C.5, C.6, and C.9). Also the intra base pair parameters exhibit at position G_{11} more extreme values in the IR complexes than in those with DR (Appendix C Figures C.7, C.8, and C.10), which may be an effect of the neighboring residue being Thymines at positions T_{10} and T_{12} in IRs, instead of adenine residues in DRs.

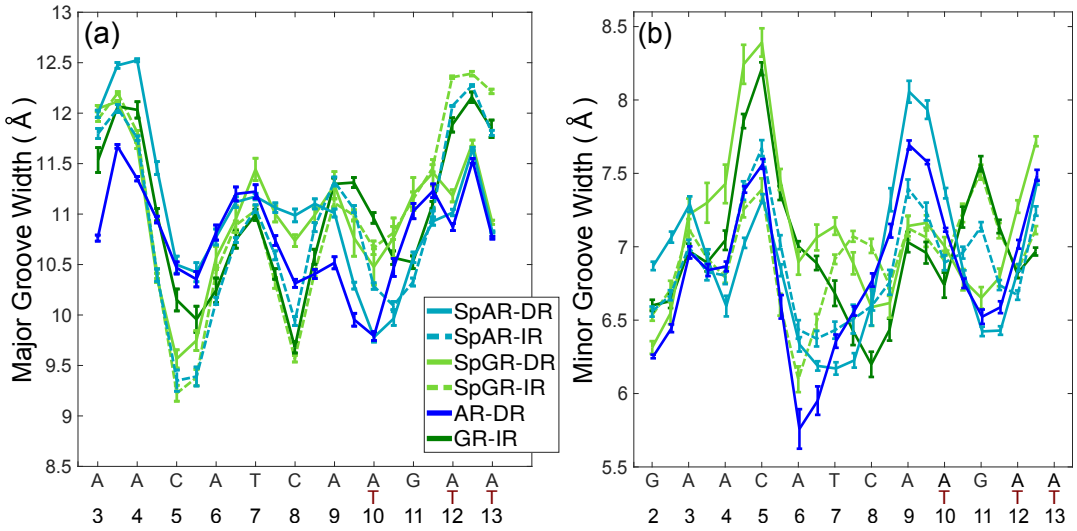


Figure 5.6: The DNA (a) major groove and (b) minor groove widths for all systems.

5.2.7 Protein-DNA hydrogen-bond interactions

In order to analyze the interaction strengths, probabilities of direct and indirect (mediated by water molecules) hydrogen bonds between protein and DNA have been calculated. Figures 5.7, and 5.8 show the hydrogen bond interactions of all studied systems, calculated from the last 100 ns of the simulations. Also, the hydrogen bond interactions of the middle 100 ns (W2 interval) were also calculated, see Appendix C Figures C.11 and C.12. According to these figures, differences in protein-DNA interactions between W2 and W3 intervals in SpARs can be seen only in the first hexamer, HS1, (Appendix C Figure C.11), whereas for SpGRs such differences exist in both DNA hexamers (Appendix C Figure C.12).

For each DNA hexamer, i.e. HS1 and HS2, there are four sites whose hydrogen bond interactions with the protein are conserved among all the systems. These are s1A1, s1G2, s2G5, and s2T6 in HS1 and s1A10, s1G11, s2T15, and s2G14 in HS2. The guanine residues at positions s1G11 and s2G5 are the predominant residues that form strong, i.e. highly probable, hydrogen-bond interactions with the protein in all systems.

Comparison of the hydrogen-bond patterns between the SpAR systems shows that the SpAR-IR complex has more strong and moderate hydrogen-bond interactions than the SpAR-DR complex. In particular, residues s1T10 and s2G5 are more strongly hydrogen-bonded in the SpAR-IR model than in the SpAR-DR complex, see Figure 5.7. The two SpGR systems show rather similar protein-DNA hydrogen-bond interactions, see Figure 5.8. However, comparing the hydrogen-bond interactions between the SpAR-IR and SpGR-IR shows that the SpAR-IR includes more and stronger hydrogen interactions than the SpGR-IR. In particular, for the SpAR-IR model more hydrogen bonds than in the SpGR-IR complex can be

observed for each specific guanine residue, i.e. s1G11 and s2G5. One further residue, i.e. s1T10, forms stronger hydrogen-bonded interactions with the protein in the SpAR-IR than in SpGR-IR complex. There is also a strong interaction in residue s2A7 of SpGR-IR which is not present in SpAR-IR. These differences in the protein-DNA interaction between the SpAR-IR and SpGR-IR complexes, that is two models of the same system, may represent two slightly different binding modes, as a consequence of different initial conformations used in the simulations.

On the other hand, our results show that both the Sp(AR/GR)-IR complexes exhibit stronger hydrogen-bond interactions than the GR-IR (see chapter 4, Figure 4.4) complex (compare residues s1G2 and s1G3, between Sp(AR/GR)-IR and GR-IR, residue s1T10 between SpAR-IR and GR-IR, and residue s2T6 and s2A7 between SpGR-IR and GR-IR). Furthermore, the AR-DR complex (see chapter 4, Figure 4.4) exhibits slightly stronger hydrogen-bond interactions than observed in the SpGR-DR but considerably stronger than observed in SpAR-DR. Interestingly, those interactions, present in AR-DR but not in SpAR-DR, are mostly formed with the HS1 and the spacer. Moreover, there are more water-mediated interactions in SpAR-IR than in SpGR-IR.

5.3 Discussion

All the protein-DNA complexes modeled in this work, represent states in which the DNA is bound by the respective DBD. The interaction strengths within the complexes, as manifested by hydrogen bond interactions between protein and DNA, as well as between the protein subunits, and conformational flexibility, however, varies between the different systems.

Of all the protein-DNA systems, including those shown in chapter 4, the AR-DR complex exhibits the strongest interactions between protein and the DNA via direct and water-mediated hydrogen bonds.

The mutations in the SPARKI systems, which transform an AR into the chimeric protein, are mainly located in one loop that constitutes the dimerization interface. The protein-protein interactions in all the SPARKI systems are weaker than in the AR-DR and comparable to (or even weaker than) those in the GR-IR system (see also chapter 4). This suggests that the dimerization interface of SPARKI is indeed GR-like, as would be expected based on its constituting sequence.

A significant conformational distortion can be seen in monomer A and the dimer interface of SpGR-DR, that is not observed in the SpGR-IR. In addition, the dimer interface of SpGR-DR has two hydrogen bonds fewer than the SpGR-IR. The SpGR-DR model, moreover, exhibits the largest Zn-Zn distances and the largest distance between the loops of the

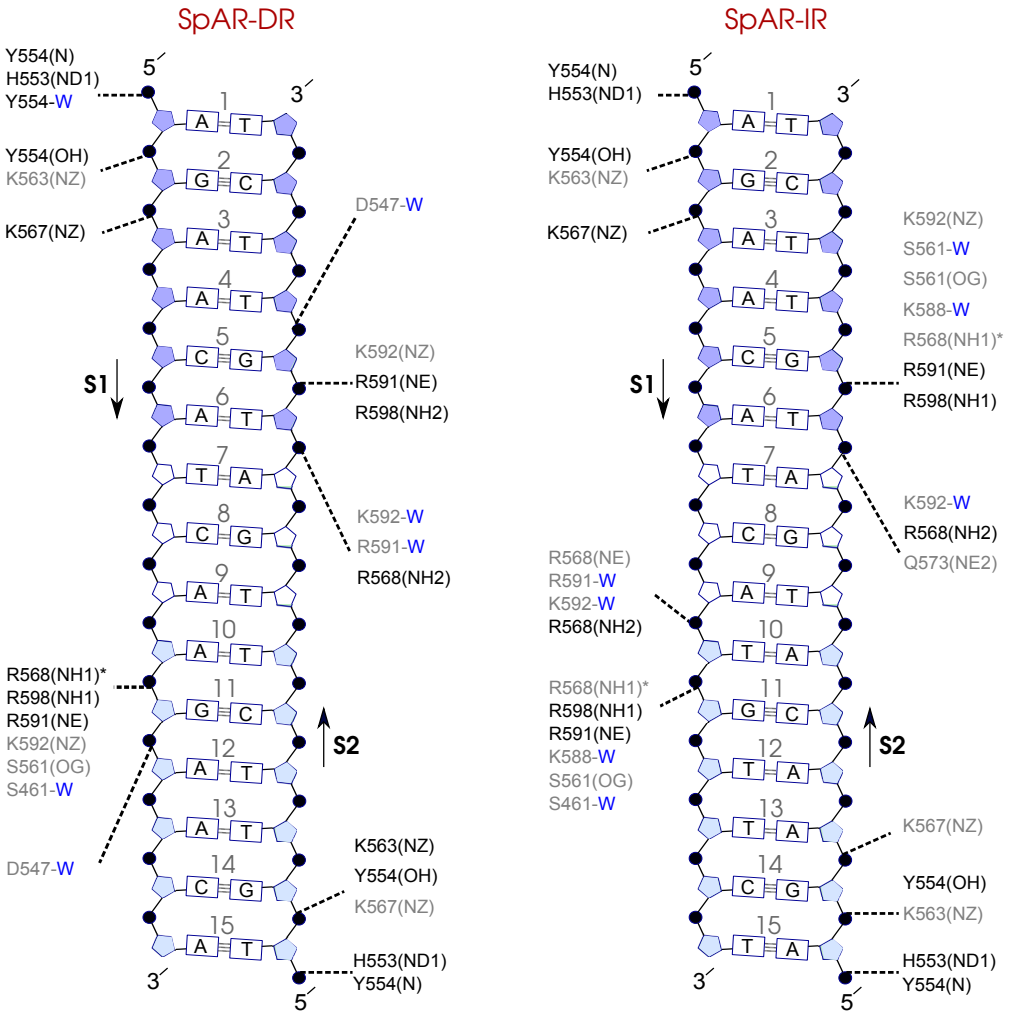


Figure 5.7: Diagram of protein-DNA hydrogen-bond interactions for (*left*) SpAR-DR and (*right*) SpAR-IR. The nucleotides of the 15 bps core DNA sequence are numbered from HS1 (numbers: 1 to 6) to HS2 (numbers: 10 to 15). The spacer region is highlighted with non-colored boxes around the numbers of the bases (numbers: 7-9). The hydrogen bonds are categorized based on their occupancy, 50-75% (gray), and 75-100% (black). The water mediated hydrogen bonds are shown with a blue letter "W". The residues shown with star sign form base-specific hydrogen-bond interactions while the other residues interact with the backbone of the DNA.

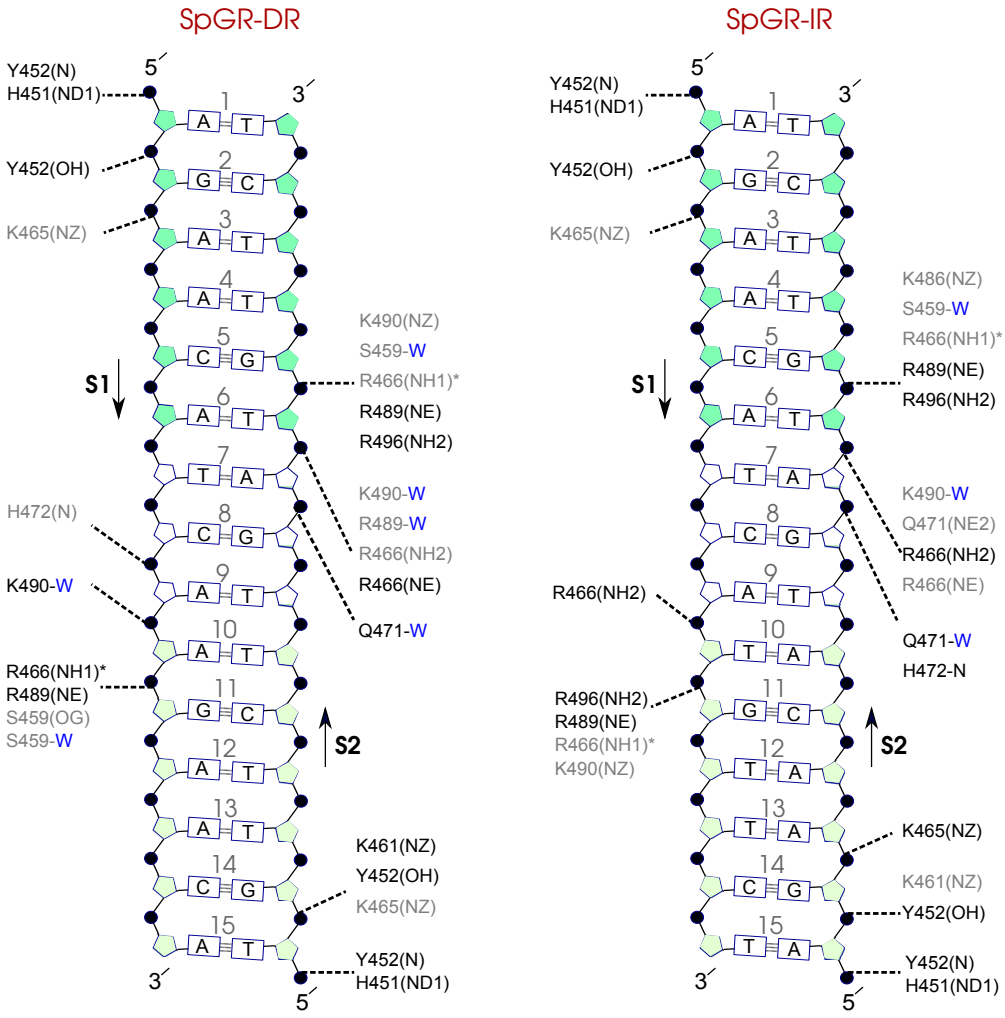


Figure 5.8: Diagram of protein-DNA hydrogen-bond interactions for (*left*) SpGR-DR and (*right*) SpGR-IR. The nucleotides of the 15 bps core DNA sequence are numbered from HS1 (numbers: 1 to 6) to HS2 (numbers: 10 to 15). The spacer region is highlighted with non-colored boxes around the numbers of the bases (numbers: 7-9). The hydrogen bonds are categorized based on their occupancy, 50-75% (gray), and 75-100% (black). The water mediated hydrogen bonds are shown with a blue letter "W". The residues shown with star sign form base-specific hydrogen-bond interactions while the other residues interact with the backbone of the DNA.

dimerization interface of all the models investigated in this work. These findings suggest that in the SpGR model, accommodation of the DR sequence, and interactions with the protein comparable to a IR sequence, can be achieved only at the expense of a distortion of the dimerization interface.

The deformation of monomer A and the dimerization interface observed in the SpGR-DR model is not observed in the SpAR-DR model, that is the complex that has been modeled from the crystal structure of the AR-DR. We attribute this difference to the different starting points for the simulations, AR-DR and GR-IR, respectively. In the SpAR models, the residues which have been in silico mutated (second zinc-binding motif) are located at the dimerization interface, whereas in the SpGR models these residues (first zinc-binding motif) are part of the DNA-binding interface. Furthermore, in the SpGR-DR model the DNA sequence has been changed from IR to DR in silico.

In the SpAR-DR model, the monomers of SpAR are tightly bound in the AR-like starting conformation. The modified dimerization interface leads to a weaker protein-protein interaction as manifested by the longer distance and fewer hydrogen bonds between the two subunits. The protein, on the other hand, does not "reach" the DNA as good as in the other models as can be seen by SpAR-DR showing the longest, though not by much, protein-DNA distances of all the complexes. Moreover, the number of hydrogen bonds between protein and DNA is smaller than in the wild-type AR-DR (see chapter 4), in particular in HS1, pointing towards a loser complex in the chimeric model. This is in agreement, albeit does not fully explain the experimentally observed low affinity of SPARKI for DR elements [47, 49, 125].

In the SpGR-DR model the dimerization interface is GR-like, that is weak to start with. In addition the protein is not properly oriented on the DR sequence. In the course of the simulation, the protein undergoes conformational changes in the dimerization interface, considerably weakening the protein-protein interaction. The distortion, weakened interactions in the dimerization interface, results in a reoriented monomer A and a deformed monomer B. That means that monomer B in SpGR does not manage to fully adjust onto the direct repeat on HS2 to form strong contacts. The observed conformational change in the Dim regions and the monomer A may be regarded as an attempt by the system to make favorable contacts in other pars of the complex. Indeed in the SpGR-DR model, more contacts, that is hydrogen-bonds between protein and DNA, are observed than in the SpAR-DR model. However, these contacts are with the HS1. Strong interactions with only one hexamer and a distorted protein-protein interface suggest a low affinity, or a rather unstable Sp(GR)-DR complex. The SpGR model is, by construction a GR-like SPARKI. Also GR lacks affinity for DR sequences, possibly because no stable complexes can be formed between GR and DR (as suggested by the simulations of GR-DR, see chapter 4). A deformed conformation in the dimerization interface of SpGR-DR

may thus point towards a loss of stability in that wild-type GR-DR complex.

Analysis of the DNA parameters around T12 exhibits extreme values in the neighboring G11 (intra bp) as well as extreme inter base pair parameters in the GT step that are not present in the GA step of the direct repeat. The affected G11 has strong interactions with the protein and is therefore an important residue for binding. This interplay may explain why T12 is essential for specific DNA recognition by GR [49] as has been shown by in vivo experiments.

The sequence and conformation in the HS2, moreover, affect the spacer region. In this region, a narrower major groove has been observed for the IR sequence than for the DR sequence. Such a DNA conformation, though not quite a kink in the DNA spacer, requires the protein to "follow" the DNA conformation so as to form favorable contacts. This is achieved by a lever arm that is more flexible in the IR-bound systems, i.e. GR and SPARKI (see Figure 5.3), and the two protein subunits being slightly further apart, as manifested by longer monomer-monomer distances in GR-IR compared to AR-DR, while the distances of the protein subunits to their respective half site on the DNA are similar. Among the complexes with an IR sequence, both SPARKI models, SpAR-IR and SpGR-IR, reveal stronger protein-DNA interactions, especially with the HS1, than the other wild-type complex, GR-IR, in agreement with experiments that show similar or higher affinity of SPARKI systems for the IR elements or classical response element, i.e. CREs [47].

The higher affinity of the SpAR/GR complexes to the IR sequence, compared to that of GR-IR (see also chapter 4), can thus be explained by the chimeric systems having both properties, the AR-like ability to strongly interact with the DNA and the GR-like "softness", that is weaker interactions, of the dimerization interface, that allows the protein to flexibly accommodate to the binding on the DNA. Qualitatively, the higher flexibility in the dimerization interface and lever arm region of the SPARKI-IR systems can be understood as entropically favorable. Indeed, the SPARKI models show a higher entropy than the wild-type complexes. Additionally, the stronger protein-DNA interactions can be understood as an increased enthalpic contribution. An increased binding affinity of SPARKI compared to GR can thus be attributed to favorable enthalpic and entropic contributions.

The AR-DR complex (see also chapter 4), in contrast, is more enthalpically stabilized by the contribution of both, protein-protein and protein-DNA hydrogen-bond interactions. In the DR-DNA the minor groove is ~ 1Å narrower at the GA step than at the corresponding GT in an inverted repeat DNA. This narrower minor groove is associated with the phosphate groups of the DNA backbone being closer to each other, and thus providing a higher negative charge density. Electrostatic interactions of the positively charged Arg (and Lys at other positions) residues with the DNA is therefore strengthened, as manifested by the larger number of strong hydrogen bonds in the AR-DR system.

The protein-DNA complexes studied in this work are characteristic for, a competition between the protein-protein interactions and protein-DNA interactions, that is, a stable dimerization interface versus specific contacts to the DNA. A balance to the former or the latter thus decides about specificity, or at least preference, for direct or inverted repeat DNA, respectively.

5.4 Conclusion

Our simulations of the chimeric SPARKI protein, complexed to inverted and direct repeat sequences, reveal a higher affinity of this model protein for IR than for DR sequences. In fact, binding to a DR results in a loose complex, eventually even with a distorted protein conformation, a possible explanation for the experimentally observed weak affinity for such a sequence [47, 49, 125].

Since AR, GR (exhaustively explained in chapter 4), and the SPARKI models can in principle all form the same contacts with specific residues of the DNA, IR or DR, the ability to accommodate the protein on the DNA is important for specificity. The required flexibility is observed in those systems with a "weaker" dimerization interface, that is GR and the GR-like SPARKI, which can thus be considered to have more entropy driven specificity. The interactions in the dimerization interface and protein-DNA interactions are balanced to allow proper accommodation of the protein on the DNA and formation of specific contacts, tuning the enthalpic contribution to specific complex formation. In this competition, the stability of the dimerization interface is important and to a large extend determines the preferred response element.

The starting point, that is the crystal structure used for model building, has, even after rather long simulation time, still an effect on the protein conformation in the complex. SPARKI models initiated from the structure of the GR-IR complex are not capable of forming strong interactions in the dimerization domain. In contrast, SPARKI models started from an AR-DR complex structure maintain a rather stable dimerization interface, despite the mutation of some residues in this domain to those of GR. Still, this interface is weaker than in the wild-type AR-DR complex,. Moreover, the chimeric SPARKI protein shows fewer interactions with DR than observed in AR-DR, rendering its specificity GR-like.

All together, this study reveals the importance of the dimerization domain on distinct specificity of AR and GR, bound to DR and IR response elements, respectively.

Chapter 6

Impact of the sequence flanking the core-binding site on the structure of the glucocorticoid receptor

Some results of this chapter are published in the article:

"Schöne S., Jurk M., Helabad M.B., Dror I., Lebars I., Kieffer B., Imhof P., Rohs R., Vingron M., Thomas-Chollier M.et al. Sequences flanking the core-binding site modulate glucocorticoid receptor structure and activity. Nat. Commun. 2016; 7:12621."DOI

Glucocorticoid receptor (GR) activity is modulated by the sequence of its DNA-binding site, called GR-binding sequence (GBS) [111, 134]. The GBS sequence resembles a consensus classical response element that is organized as an inverted repeat (IR) of hexamer "AGAACA", separated by three base-pairs (bps) called spacer [42, 111]. Depending on different GBS sequences, the direction of regulation might be influenced such that it activates or represses gene transcription [36, 135, 136]. Moreover, it is shown that the magnitude of transcription activation directly depends on the exact core 15-bps, mentioned above [111]. In order for the GR DNA binding domain (DBD) to have a an intact protein-protein dimerization domain, the DBD-alpha helix H1 (introduction chapter, Figure 1.5) of each monomer binds specifically to the major groove of the DNA. However, the specific GBS sequence, with its different sequence compositions, which all are IR-like, is not the only factor modulating GR activity [100, 112]. The shape of the DNA can be read out by GR, through non-specific interactions with the phosphate backbone and the spacer, and can modulate the protein-DNA interactions [104] and thus the GR activity as well [43, 100]. Outside of the core sequence, GR-DBD also contacts the minor groove [43]. The flanking sequence context may therefore have a role in DNA-protein recognition. In fact, recent studies have demonstrated that the presence of

the core specific DNA sequence is not sufficient to explain the high affinity protein-DNA interactions. Rather, immediate flanking nucleotides of the core sequence are discussed to be a determining factor. For instance, a recent, high-throughput in vitro and in silico study has revealed that stability, geometry, and flexibility of the sequence flanking the core DNA sequence are important factors in the modulation of the proteins binding to their core specific sequences [116, 137].

A recent experimental study, by Dr. Sebastian Meijsing and his group at the Max Planck institute for Molecular Genetics, with whom we collaborated in this project, has shown that different flanking nucleotides of core GBS sequence not only alter the three dimensional structure of GR-DBD and the quaternary structure of the dimeric complex, but also the GR binding activity [50]. It is interesting to note that despite the remarkable impact of the sequence flanking the core GBS on the GR structure and activity, no apparent influences on the binding affinity could be observed.

In general, the sequence specific of protein-DNA interactions is highly dependent on both DNA and the properties of the protein binding to it [111, 138]. In the previous chapter, we mainly focused on the core DNA elements and showed how an exchange of the core DNA sequence from inverted repeat (IR) to the direct repeat (DR) distorts the GR DBD-DNA complex, in agreement with experimental evidences (for details see chapters 4 and 5). In this chapter, we study the influence of bases flanking the core DNA sequence on the conformation of the GR DBD-DNA complex via performing molecular dynamics (MD) simulations. Our results show that depending on the immediate flanking sequence, or "proximal flank" of the core GBS, the conformation of the GR DBD-DNA complex changes. Thanks to our collaborator, i.e. Dr. Sebastian Meijsing and his group, we were also able to compare the results of our MD simulations with their experimental outputs. Our results point to a distinct conformational change of the GR-DBD, in presence of different flanking sequences, as seen in the experimental results.

6.1 Molecular dynamics simulations: systems and protocols

6.1.1 Molecular systems

Classical MD simulations were carried out for A/T and G/C flank variants of the Cgt GBS. The initial structure was prepared based on a crystal structure of the GR DNA-binding site in complex with the Cgt-binding site (PDB ID 3FYL [43]), including the A/T flanking nucleotides. Position +5 was mutated in silico (C to A). Five and four nucleotides per strand in a perfect B-form were added to the $5'$ and $3'$ side of the DNA fragment, respectively, resulting

in DNA fragments with 24 nucleotides length: 5′-CACCAAGAACATTTTGTACGTCTC-3′ and 5′-CACCGAGAACATTTTGTACGCCTC-3′ for the A/T and G/C Cgt flank variant, respectively.

6.1.2 Molecular dynamics simulations

The simulations were performed with the program package NAMD 2.10 [117] using the CHARMM27 force field [118, 119]. The DNA fragments of the initial structures were energy minimized (3,000 steps of conjugate gradient) to remove energetically unfavorable conformations resulting from the addition of the additional nucleotides. The systems were solvated in TIP3P water [92] and a total of 35 sodium ions were placed randomly within a minimum distance of 10.5 Å from the solute and 5 Å between sodium ions to ensure a zero net charge for the solute–solvent–counterion complex. The systems contained ∼127,000 atoms. The final complexes were equilibrated by 5,000 steps of energy minimization, followed by a 30 ps MD simulation (time step 1 fs) to heat up the system to 300 K by velocity scaling. Next, a relaxation of 200 ps (time step 1 fs) was performed for an NPT ensemble. Periodic boundary conditions were implemented with the particle-mesh Ewald method [68] for electrostatic interactions with cut-off distance 14 Å. Lennard–Jones interactions were truncated at 14 Å. The SHAKE algorithm was applied to constrain all bonds involving hydrogen atoms [95]. Three independent, 100-ns-long MD simulations were performed in constant pressure (1 bar) and constant temperature (300 K) with a 2 fs time step for each A/T and G/C flanking the GBS. During these simulations, pressure and temperature were maintained constant using a Langevin dynamics barostat and Nosé Hoover Langevin thermostat [121, 122]. The terminal base pairs of the DNA fragments were restrained harmonically. One simulation run for each of the models with A/T- and G/C-flanking nucleotides, respectively, was further prolonged to 500 ns.

6.2 Result

6.2.1 DNA Conformation

DNA parameters, including DNA minor- and major-groove widths, DNA helical axis bending, and inter- and intra-bps parameters are calculated for both A/T and G/C Cgt model of DNA, for the last 100 ns of the simulations trajectories. Comparing these parameters between the two systems elucidates a significant influence of the flanking bases on the local geometry of the DNA. Figure 6.1 shows the groove parameters of DNA. For both, minor and major grooves, the major effect of flanking bps are present at the proximal flanking bps (shown in Figure 6.1)

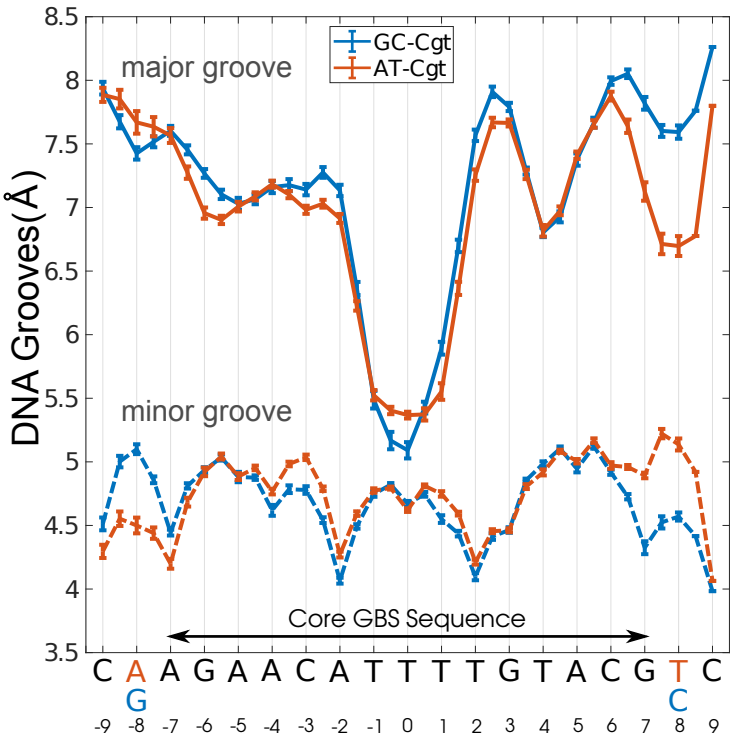


Figure 6.1: Major and minor groove widths of the GR-bound DNA with AT-Cgt and GC-Cgt sequences (indicated by red and blue letters, respectively)

as well as their adjacent bps, i.e. $(A_{-8}/G_{-8})A_{-7}$, $G_7(T_8/C_8)$. However, a slight difference between the two systems can also be observed for the grooves of the core GBS elements, i.e. bps G_{-6} to C_6. Depending on the flanking bps elements, the minor groove width of the DNAs show opposite effects in the two half-sites. Dinucleotide $A_{-8}A_{-7}$ in the AT-Cgt system shows a narrowed minor groove compared to the $G_{-8}A_{-7}$ dinucleotide in the GC-Cgt system, while dinucleotide G_7T_8 in the AT-Cgt system shows a wider minor groove width in comparison to the G_7C_8 dinucleotide in the GC-Cgt system. Interestingly, the effect of mutating the flanking bps on the major groove width is only observed in the second half-site. The major groove of dinucleotide G_7T_8 in AT-Cgt is narrower than that of dinucleotide G_7C_8 in GC-Cgt whereas the major groove widths of the first half-sites are about the same in both systems, AT-Cgt and GC-Cgt.

A significant result has been observed for the DNA helical axis bending parameters. As shown in the Figure 6.2, the DNA helical axis exhibits a considerable bend in the GC-Cgt system compared to the AT-Cgt system. Interestingly, the predominant bending occurs in the second DNA half-site, i.e. residues T_4 to C_9, such that a remarkable increase of bending in this half-site can be observed for the GC-Cgt system with respect to the AT-Cgt system. However, we observe a slight decrease of bending, at the first DNA half-site and spacer (residues C_{-9} to

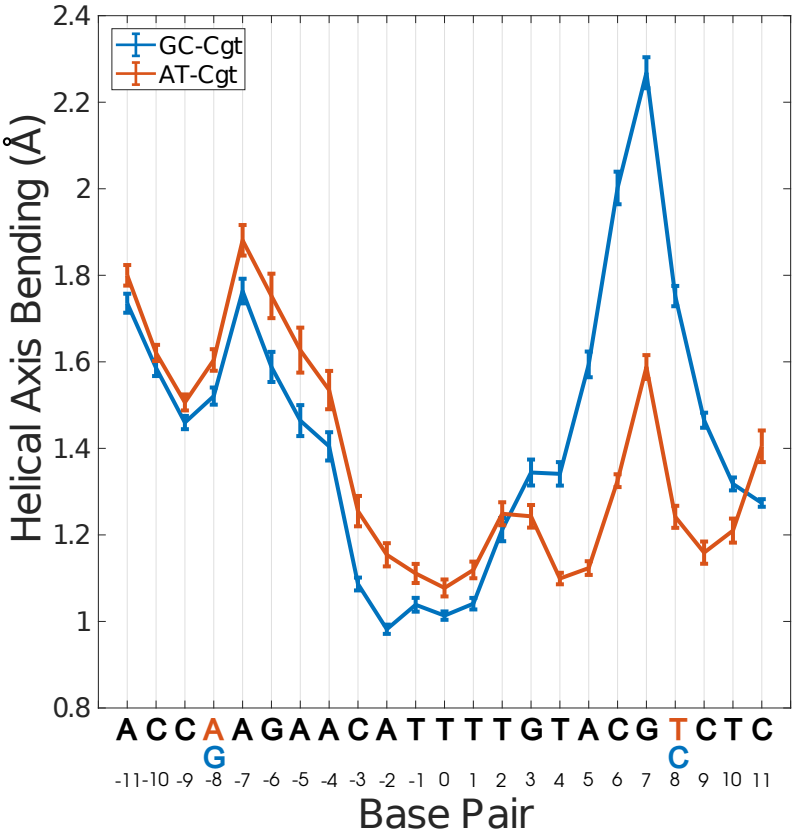


Figure 6.2: DNA helical axis bending of the GR-bound DNA with AT-Cgt and GC-Cgt sequences (indicated by red and blue letters, respectively).

T_1), for the GC-Cgt system in comparison to the AT-Cgt system.

Moreover, analysis of the inter- and intra-bps parameters indicates that a sequence exchange of the flanking regions alters the local geometry of their nearby core GBS bps. Inter-bps parameters of DNA for both systems are depicted in Figure 6.3. As shown in this figure, dominant differences are observed in the second half-site, and smaller ones in the first half-site. For instance, except for the slide parameter, which exhibits considerable differences between two systems in the first half-site, the other dominant dissimilarity is found for base pair parameters in the second half-site, e.g. shift, rise, roll, and twist parameters. In contrast, for intra bps parameters, shown in Figure 6.4, a conformational deviation of the DNA due to mutation of the flanking bps occurs in both half-sites.

6.2.2 Conformational Fluctuation

In order to estimate the structural differences of the two complexes, the median structures representing the last 50 ns (of the total of 500 ns simulation time) of both complexes were aligned

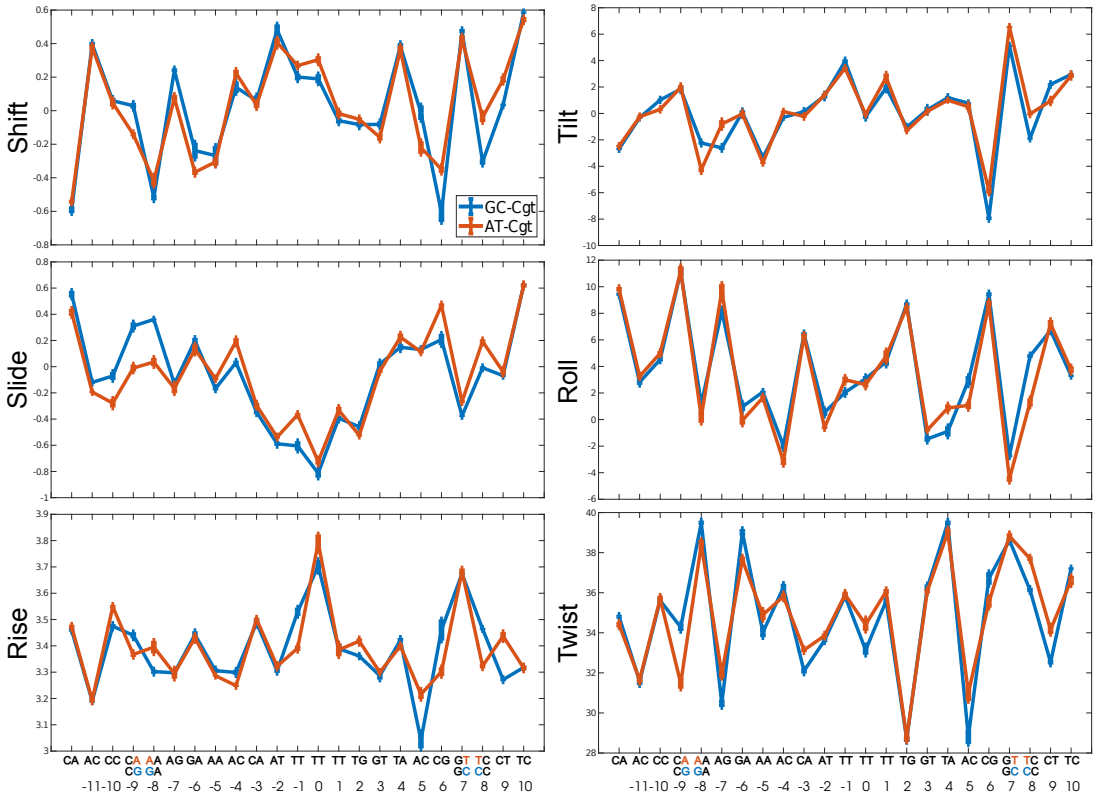


Figure 6.3: Inter bps parameters of the GR-bound DNA with AT-Cgt and GC-Cgt sequences (indicated by red and blue letters, respectively).

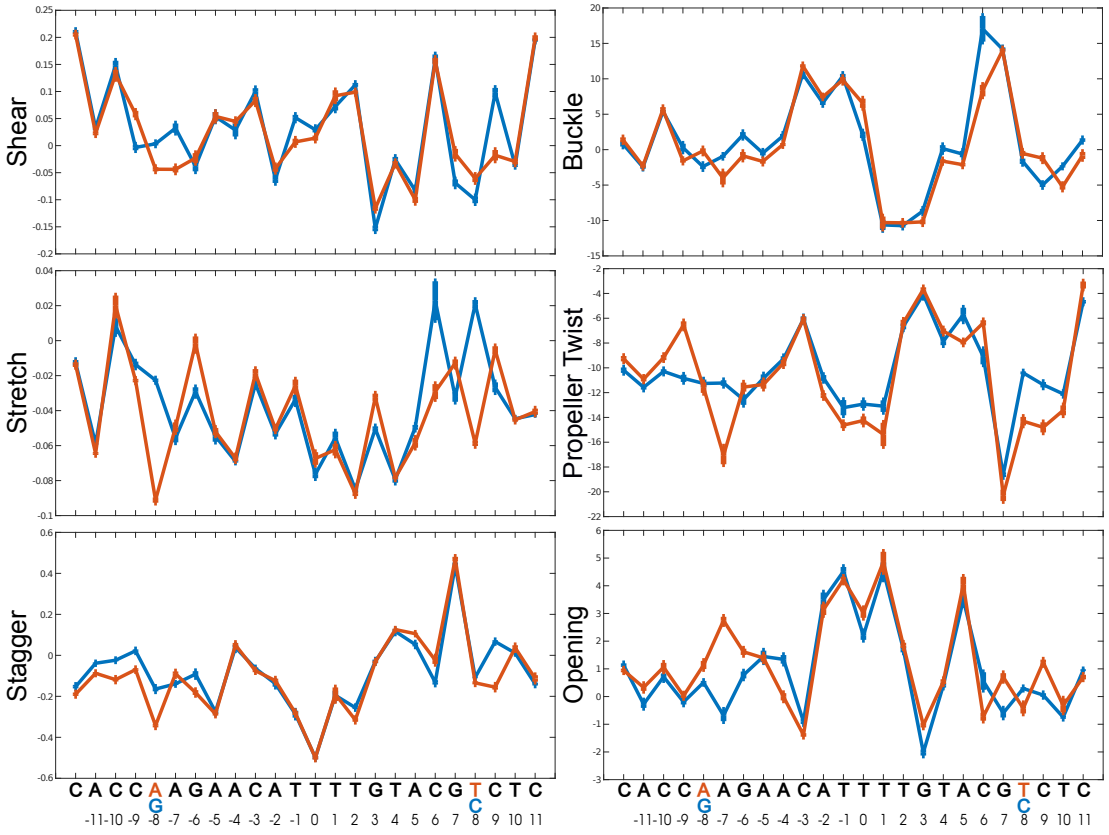


Figure 6.4: Intra bps parameters of the GR-bound DNA with AT-Cgt and GC-Cgt sequences (indicated by red and blue letters, respectively).

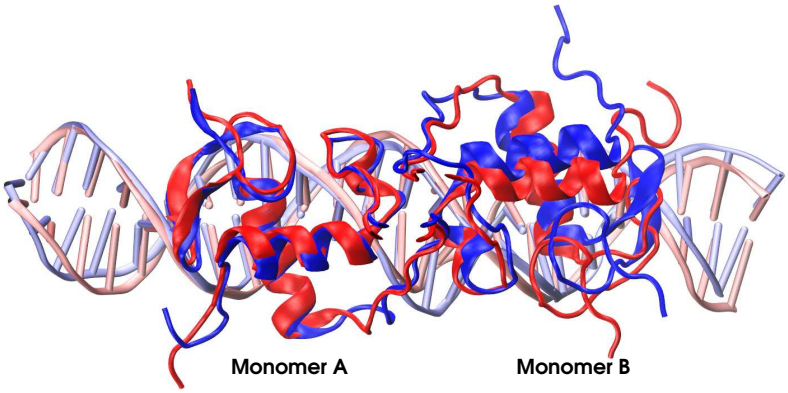


Figure 6.5: Comparison of GR median structure bound to A/T (red) and G/C (blue) flanked Cgt in MD simulations over last 50 ns reveals relative repositioning of GR monomers.

with respect to each other and compared. As can be seen in Figure 6.5, we observed that due to different flanking bps, the relative positioning of the protein's monomer B with respect to the second half-site is considerably differend in the GC-Cgt system (colored blue in Figure 6.5) from the position in the AT-Cgt system (colored red in Figure 6.5) while no such conformational change has been observed for monomer A. In order to compare the flexibility of the GR DBD between A/T and G/C flanked sequences, the root mean square fluctuation (RMSF) of protein residues of both systems are calculated, shown in Figure 6.6. Our results exhibit rather identical fluctuations of monomer A residues for both GC-Cgt and AT-Cgt systems (Figure 6.6, top). There are only small differences in the lever arm and in the dimerization region of monomer A. In contrast, we observe considerably higher fluctuations of monomer B residues of the GC-Cgt system with respect to those residues of the monomer B of AT-Cgt system; residues ALA447 to ARG466 and CYS500 to ASN506 have higher RMSF values in the GC-Cgt system than in the AT-Cgt system.

6.3 Discussion

Proteins recognize a specific DNA sequence by both base readout and shape readout mechanisms [104]. In addition, the sequence content of the specific DNA plays an important role in gene expression such that the transcriptional activity and also affinity of binding are affected

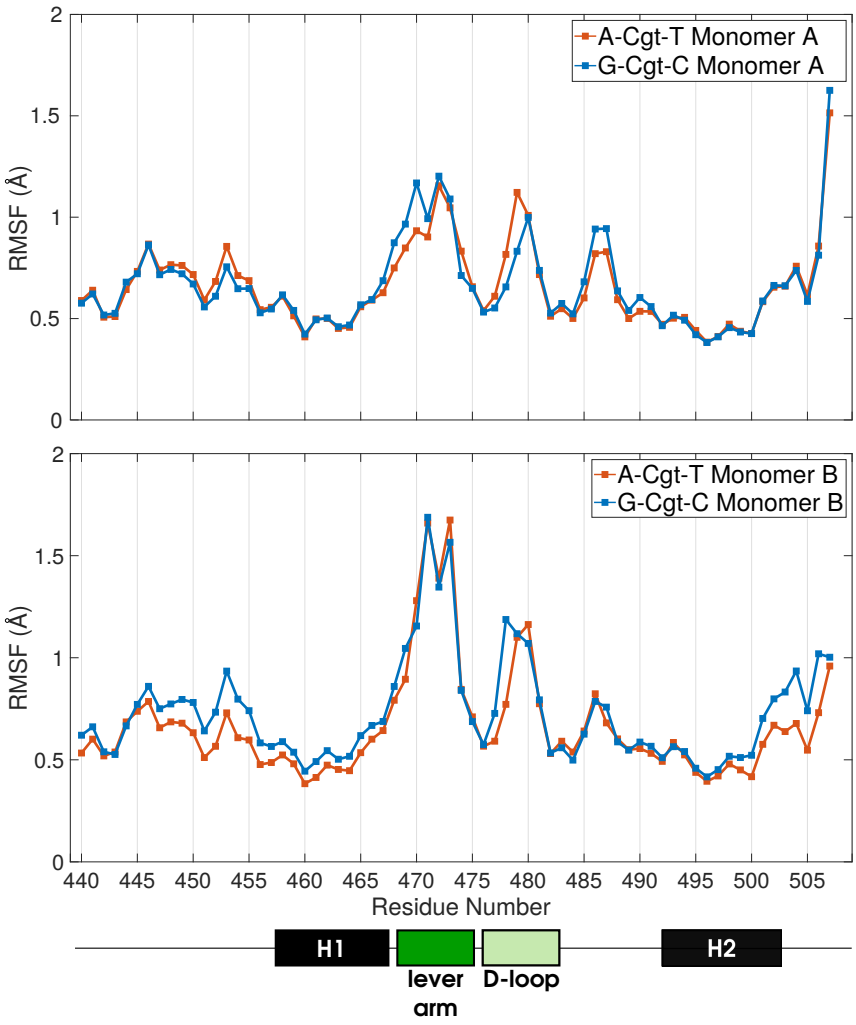


Figure 6.6: RMSF for each amino acid as indicated for (top) monomer A and (bottom) monomer B of GR over the last 100 ns MD simulation for GR bound to A/T flanked Cgt (red) and GR bound to G/C flanked Cgt (blue).

[111]. Our simulations demonstrate that the GR binding to the core specific DNA sequence is modulated by residues flanking the core GBS. This is in agreement with experimental results that indicate the significant flank sequence-dependent activity of GR bound to GBS. Structural analysis by NMR suggested this different modulation of GR activity might be due to a change in the relative positioning of the dimeric partners as a consequence of distinct flank elements [50]. Interestingly, our results show such conformational change of protein monomers in G/C flanked Cgt with respect to the A/T flanked Cgt.

Furthermore, comparison of the median structures of both AT-Cgt and GC-Cgt complexes (Figure 6.5) demonstrates that the change of A/T- to G/C-flank sequences mainly leads to a repositioning of monomer B and not monomer A, i.e. there is no considerable change in the position of monomer A. This is consistent with experimental data on these complexes that show a considerable change in monomer B of the G/C flanked system with respect to the A/T flanked [50]. This repositioning of the monomer B with respect to the monomer A, due to flanking sequence effects, also confirms the experimental results that suggest the higher affinity of monomer A to DNA in comparison to the monomer B [49, 103].

The protein-DNA hydrogen-bond interaction analysis of A/T flanked and G/C flanked of the Cgt systems shows similar interaction patterns for both complexes. This is in line with experimental data that show similar affinity of GR bound to A/T and G/C flanked DNA elements [50]. In contrast, the activity of GR differs, depending on the different flanking elements, i.e. A/T or G/C. Experimental results shows that changing of the A/T flanking sequence to G/C sequence considerably decreases the GR activity [50]. As the similar affinity of binding between these two different flanked response elements is observed, the conformational change of the dimer partners in the G/C flanked system is suggested as a main factor in its loss of activity. In our study, we observe that besides the positional change of monomer B in the GC-Cgt sytesm (Figure 6.5), the residues of this monomer show also higher fluctuations in motion with respect to monomer A as well as the monomer B of AT-Cgt system, see Figure 6.6.

Our results also show the same hydrogen-bond interactions of protein residues with flanking sequences and their adjacent base-pairs. In experiment, the mutational analysis on protein residues that have direct contact with flanking nucleotides have shown that flank effects are still present; suggesting that direct contacts with the flanking nucleotides are not responsible for the flank effect [50]. Instead, the DNA shape likely plays an important role. There are several studies that emphasize the importance of DNA shape in protein-DNA interaction [104, 123, 139]. Interestingly, our results show that the DNA geometry due to different flanking nucleotides is considerably altered. Significantly, the main differences can be observed in the second half-site of the DNA and less in the first half-site. This shows that a change of

flanking nucleotides in the second half-site has a larger influence on the DNA shape than the nucleotides flanking the first half-site, in agreement with experiment [50]. These results suggest that the altered conformation of the second half-site of the DNA, that is the DNA shape, modulates the positioning of the GR-DBD on the DNA and therefore can alter the GR activity.

6.4 Conclusion

Our simulation of the GR-DBD protein, complexed to classical response element with A/T (wild type) and G/C (mutant) flanked sequences, reveal a considerable change in the relative positioning of the dimer protein in the comlex with a G/C flanked response element compare to the complex with an A/T flanked sequence. Our results show that this relative positioning of monomers is mainly due to repositioning of monomer B with respect to the monomer A, in agreement with experimental result, which can be understood as a response to a considerably altered DNA shape. Mutation of the wild type flanking sequence, i.e. A/T flank, to G/C sequence results in a change in DNA shape that is predominently observed in its second half-site, where the monomer B resides on. However, the hydrogen bond interactions pattern are almost similar in both complexes, in agreement with the experimentally observed similar binding affinities of GR on the two sequences. Therefore, in agreement with experimental result, we suggest that the diminished activity of GR when bound to a G/C flanked sequence is due to the conformational change of monomer B which is coupled to the altered -second DNA half-site. To conclude base pairs flanking the core DNA response element of GR can significantly alter the complex structure and therefore its activity.

Chapter 7

Conclusion

Understanding long-range effects is essential for characterizing the conformation-dependent function of macromolecules. The underlying challenge is to identify sites that are located away from each other but functionally hold substantial communications. An enhanced evaluation of these long-range effects requires a dynamical description of the system under study, which is rarely achievable by experiment. To achieve such dynamical description in atomistic detail, molecular dynamics (MD) is an effective and accurate tool to simulate macromolecules in a broad range of time and size scales. Both systems considered in this dissertation, i.e. inner membrane Cytochrome c oxidase (CcO) protein and steroid protein-DNA complexes, are outstanding models for studying the long-range communications and interactions. In the first system, i.e. CcO protein, our results exhibit significant presence of long-range communication between the residues playing a key role in the transfer of protons through the proton-conducting channels. In this regard, we observe that the mutual impact of key residues are protonation-state dependent. For instance, our results show that depending on the protonation state of the K-channel residue K362, the means of communication change considerably for the different key residues. Moreover, a gating role of the D-channel residue N139 depending on the protonation-states of the terminal residues of the D-channel, i.e. D132 and E286, has been observed in our study. This, together with the depdence of D-channel residue N139's conformational and hydrogen-bond dynamics on the K-channel residue K362 suggest the long-range communication to be a determinant factor in dynamical regulation of the CcO conformation and therefore its function. These results indicate that mutations of the functionally key residues can exert a long-range influence that may result in a drastic change in dynamical properties of the other distantly positioned key elements of the protein [140].

Besides the significance of the intramolecular communication in macromolecules, such as in case of the CcO, it is moreover crucial to identify the role of the long-range effects upon intermolecular communications. This is what we have studied in the second system, i.e steroid

protein-DNA complexes. In fact, the main question is whether the local interactions can solely modulate the stable formation of the specific protein-DNA complexes or not. Our results show that although the presence of such local interactions is vital, the communication between different sites in the complex, which are distantly far positioned, significantly mediates the proper binding in the protein-DNA complex. For instance, our results clearly show that the steroid receptors dimerization interfaces are strongly influenced by their binding to the DNA response elements which are about ~ 18 Å distanced from each other. In agreement with experiments [43, 100, 111], we show that such allosteric communication is modulated by the "lever arm", a flexible loop that links the DNA-reading helix of the protein to the dimerization interface. On the other hand, we observe that the conformation of the lever arm highly depends on the shape of the 3 bps "spacer", which is, itself, correlated with the content of the adjacent hexameric sequence. As a surprising result, we observed that the mutation of the DNA's second hexamer in the glucocorticoid receptor-DNA complex, from consensus inverted repeat to a direct repeat-like sequence, significantly destabilizes the conformation of its non-counterpart protein monomer, i.e. the monomer that does not involve any direct interaction with this but the other hexamer (Figure 1.5 b,c). Our results show that such interaction is most likely mediated via the cooperative impacts of the "spacer" and the "lever arm". These findings furthermore show that the coupling of domains inside specific protein-DNA complexes are essential for their stability and therefore biological function. However, these couplings are not just limited to the sequence inside the core specific site. In our study, we observe that mutation of the sequence flanking the core specific DNA sequence considerably changes the relative positioning of protein monomers with respect to each other, in agreement with experimental results [50]. In fact, such long-range effect between the protein and flanking element on the DNA is mediated through dynamical variation of the DNA shape as a consequence of flanking site mutation. Altogether, in this study, we conclude that the overall conformational stability of the macromolecules are significantly modulated by the course of long-range effects that allow proper communication between their different functional sites.

Appendix A

Supplementary Information for: Protonation-State-Dependent Communication in Cytochrome c Oxidase

Conformational Analysis

This section lists the distribution of distances between side chains of key residues D132, N139, E286, K362, and E101. For the distances atoms C_γ, C_γ, C_δ, N_ζ, and C_δ have been considered for D132, N139, E286, K362, and E101, respectively.

Side chain distances

Dihedral conformations

This section lists the distribution of side chain dihedral angles of key residues D132, N139, E286, K362, and E101, respectively, obtained from the MD simulations of the different protonation models.

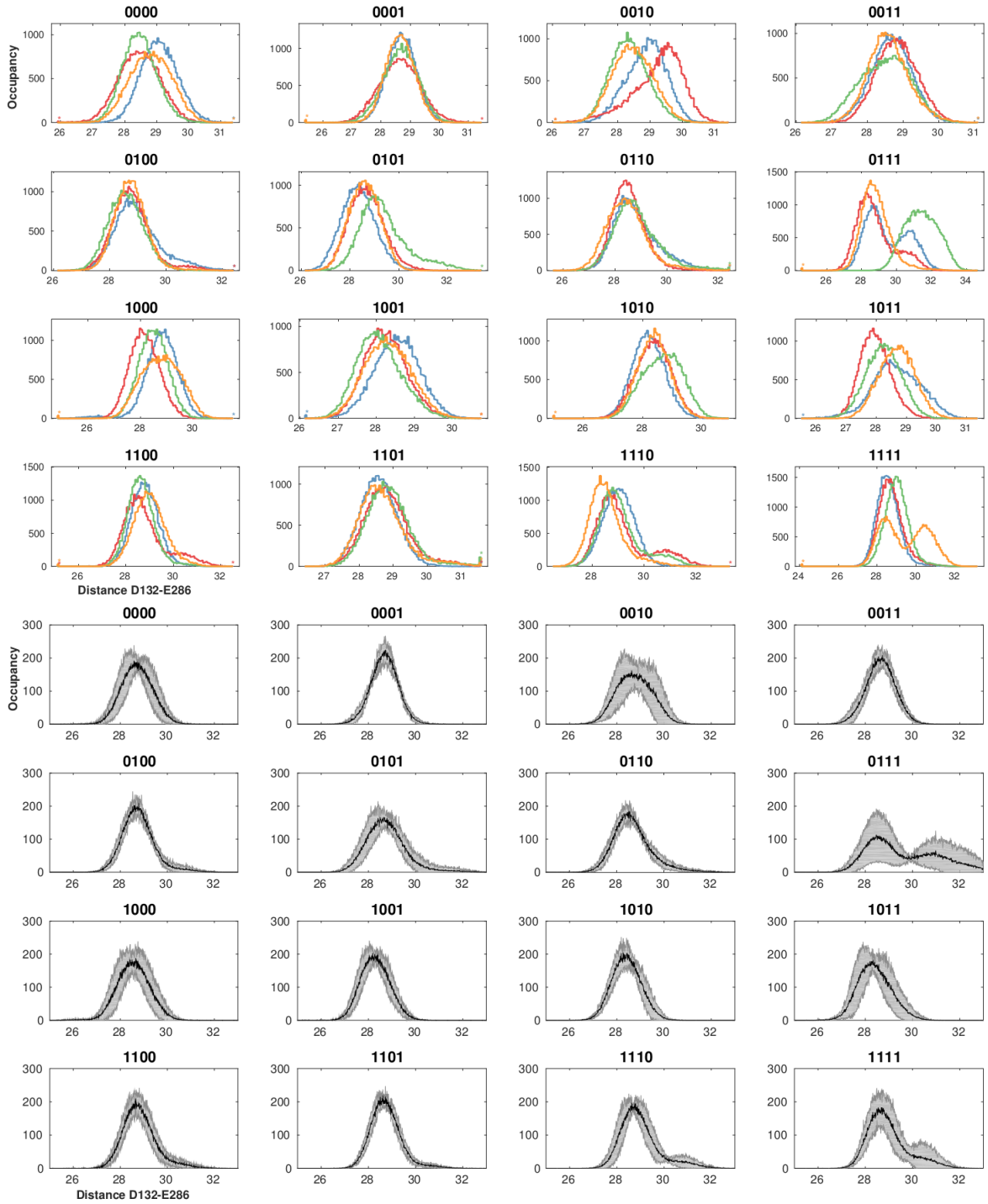


Figure A.1: Distribution of distances between D132 and E286 in different protonation states of Cytochrome c oxidase from (*up*) individual MD simulations and (*down*) averaged over the different runs. The protonation models are labeled by "1" for protonated and "0" for unprotonated D132, E286, K362, and E101, respectively.

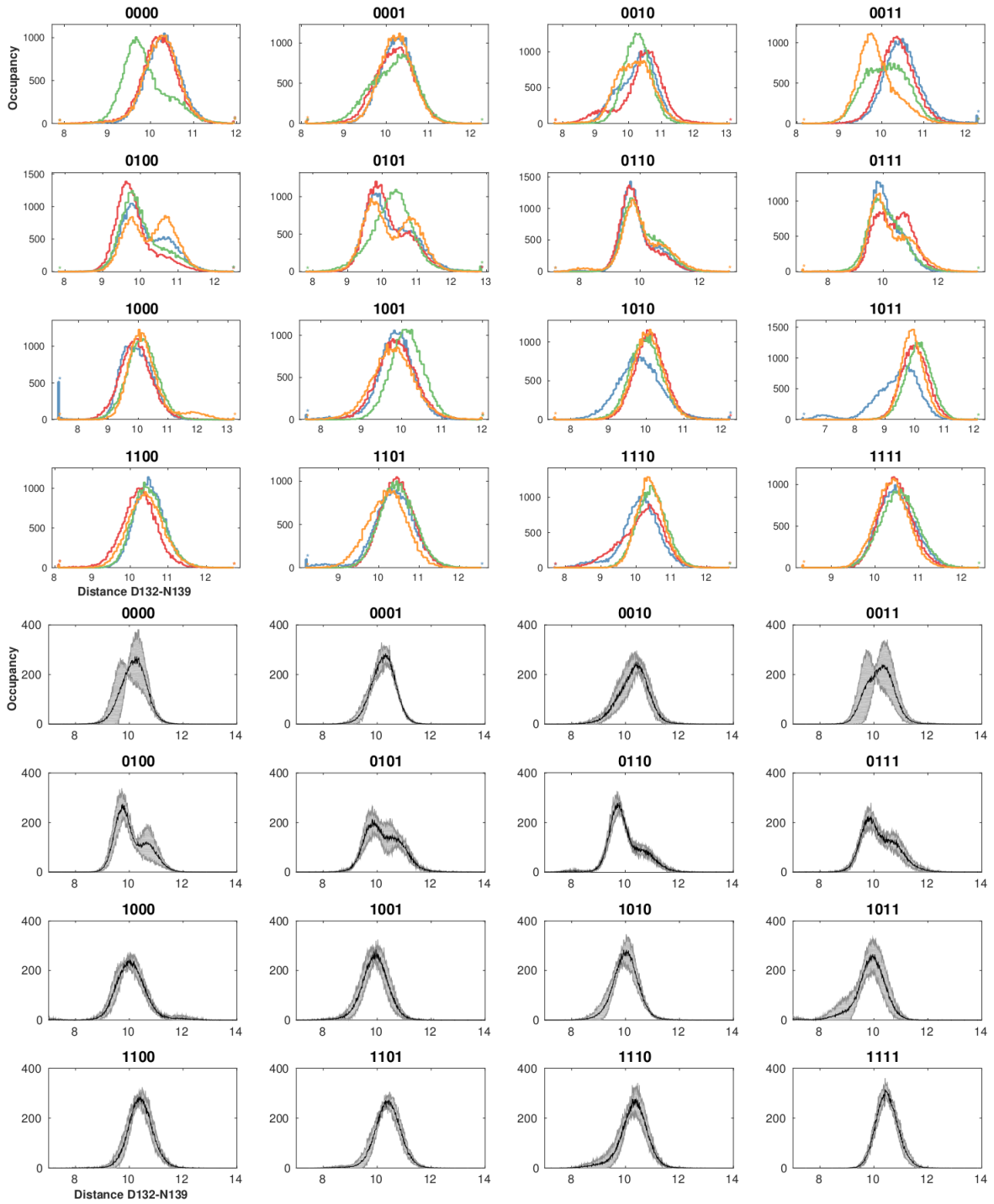


Figure A.2: Distribution of distances between D132 and N139 in different protonation states of Cytochrome c oxidase from (*up*) individual MD simulations and (*down*) averaged over the different runs. The protonation models are labeled by "1" for protonated and "0" for unprotonated D132, E286, K362, and E101, respectively.

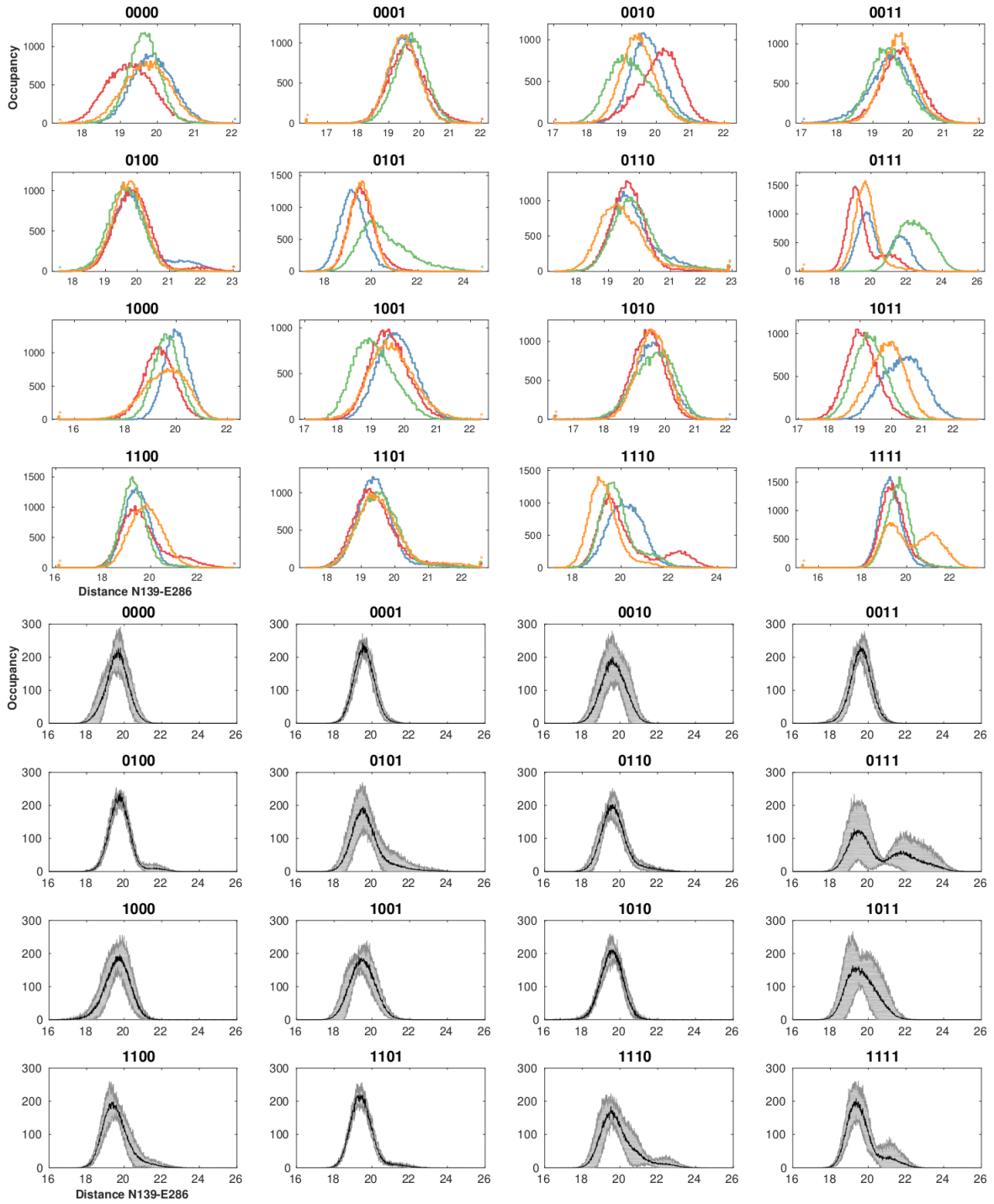


Figure A.3: Distribution of distances between N139 and E286 in different protonation states of Cytochrome c oxidase from (*up*) individual MD simulations and (*down*) averaged over the different runs. The protonation models are labeled by "1" for protonated and "0" for unprotonated D132, E286, K362, and E101, respectively.

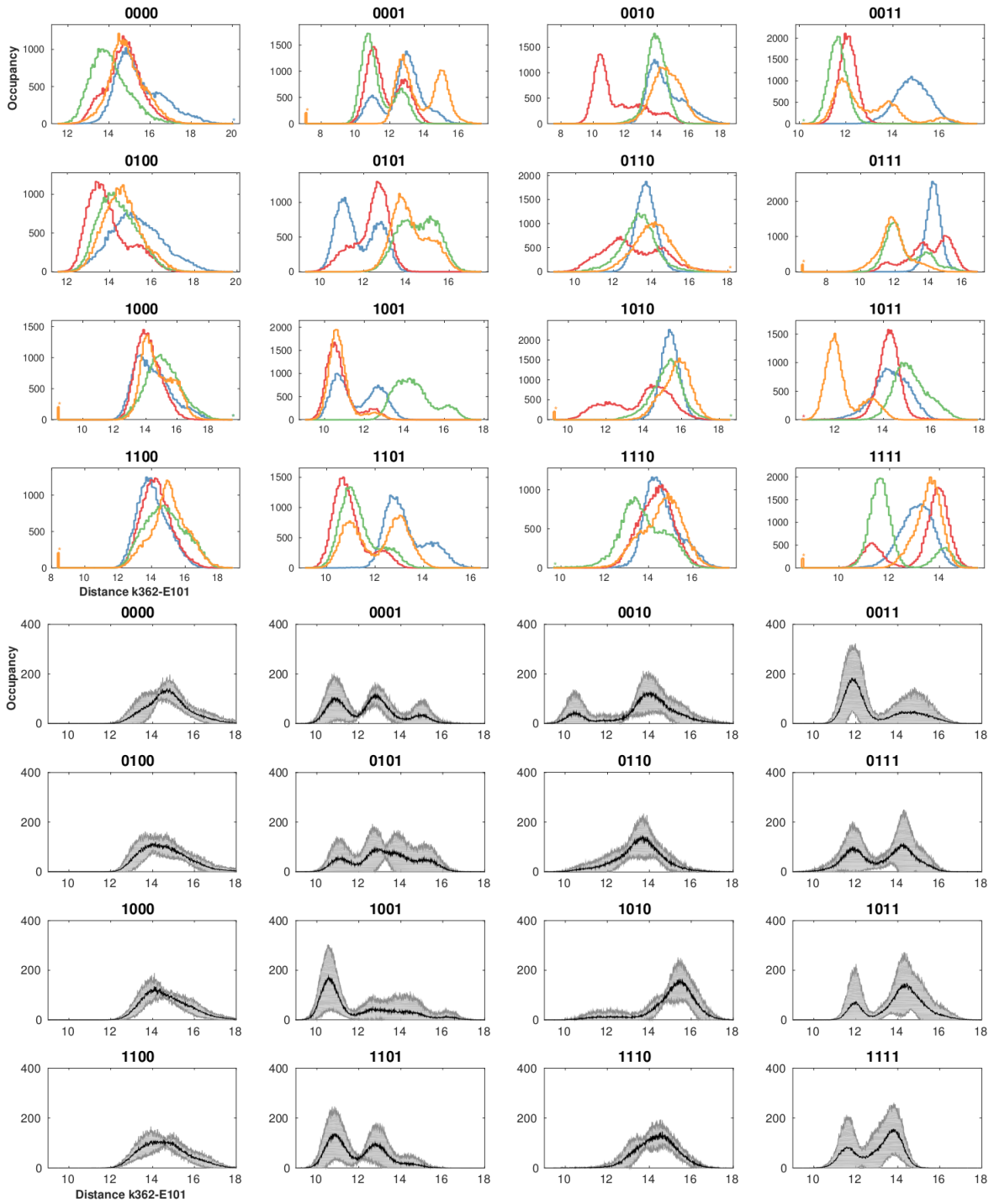


Figure A.4: Distribution of distances between K362 and E101 in different protonation states of Cytochrome c oxidase from (*up*) individual MD simulations and (*down*) averaged over the different runs. The protonation models are labeled by "1" for protonated and "0" for unprotonated D132, E286, K362, and E101, respectively.

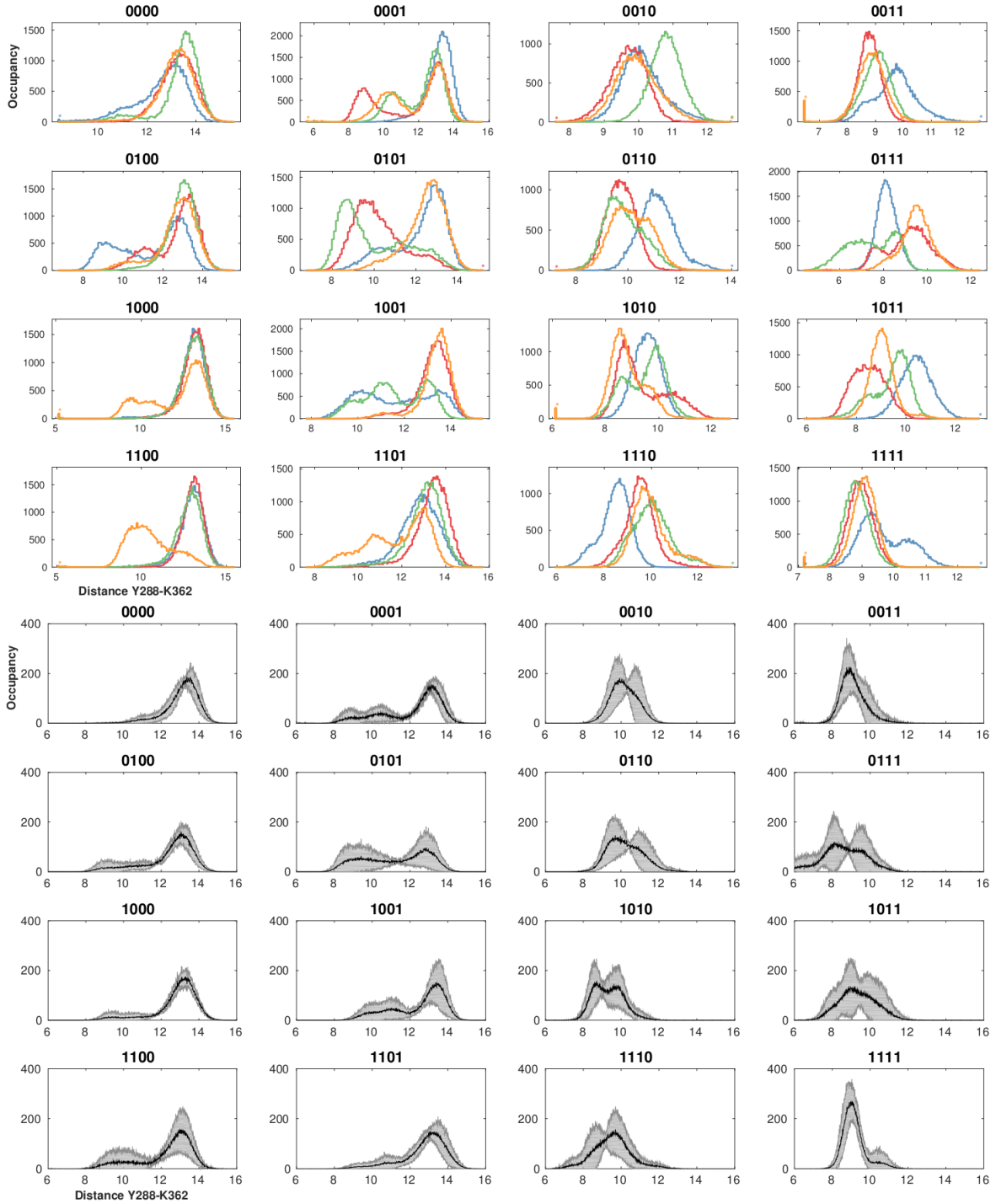


Figure A.5: Distribution of distances between Y288 and K362 in different protonation states of Cytochrome c oxidase from (*up*) individual MD simulations and (*down*) averaged over the different runs. The protonation models are labeled by "1" for protonated and "0" for unprotonated D132, E286, K362, and E101, respectively.

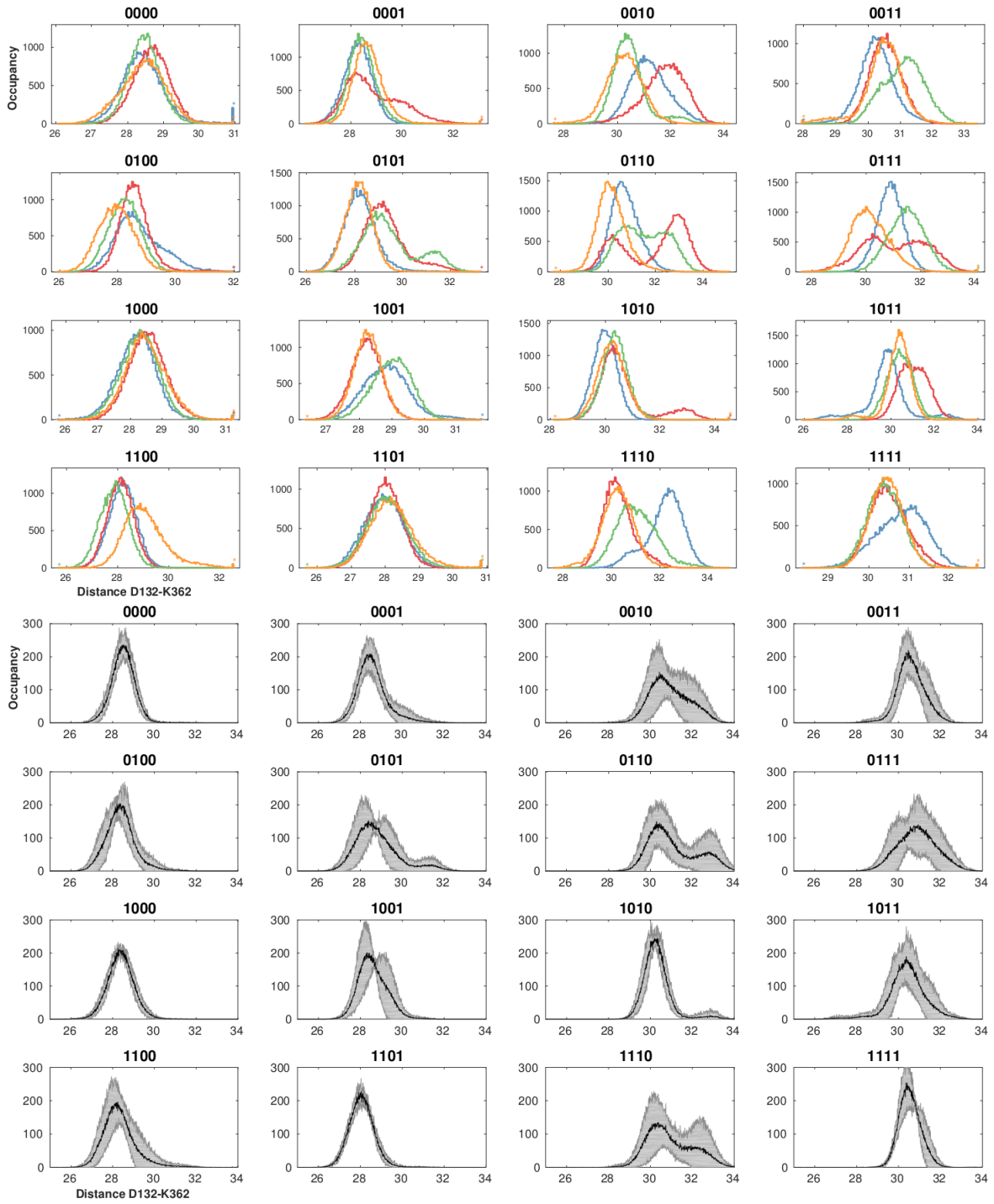


Figure A.6: Distribution of distances between D132 and K362 in different protonation states of Cytochrome c oxidase from (*up*) individual MD simulations and (*down*) averaged over the different runs. The protonation models are labeled by "1" for protonated and "0" for unprotonated D132, E286, K362, and E101, respectively.

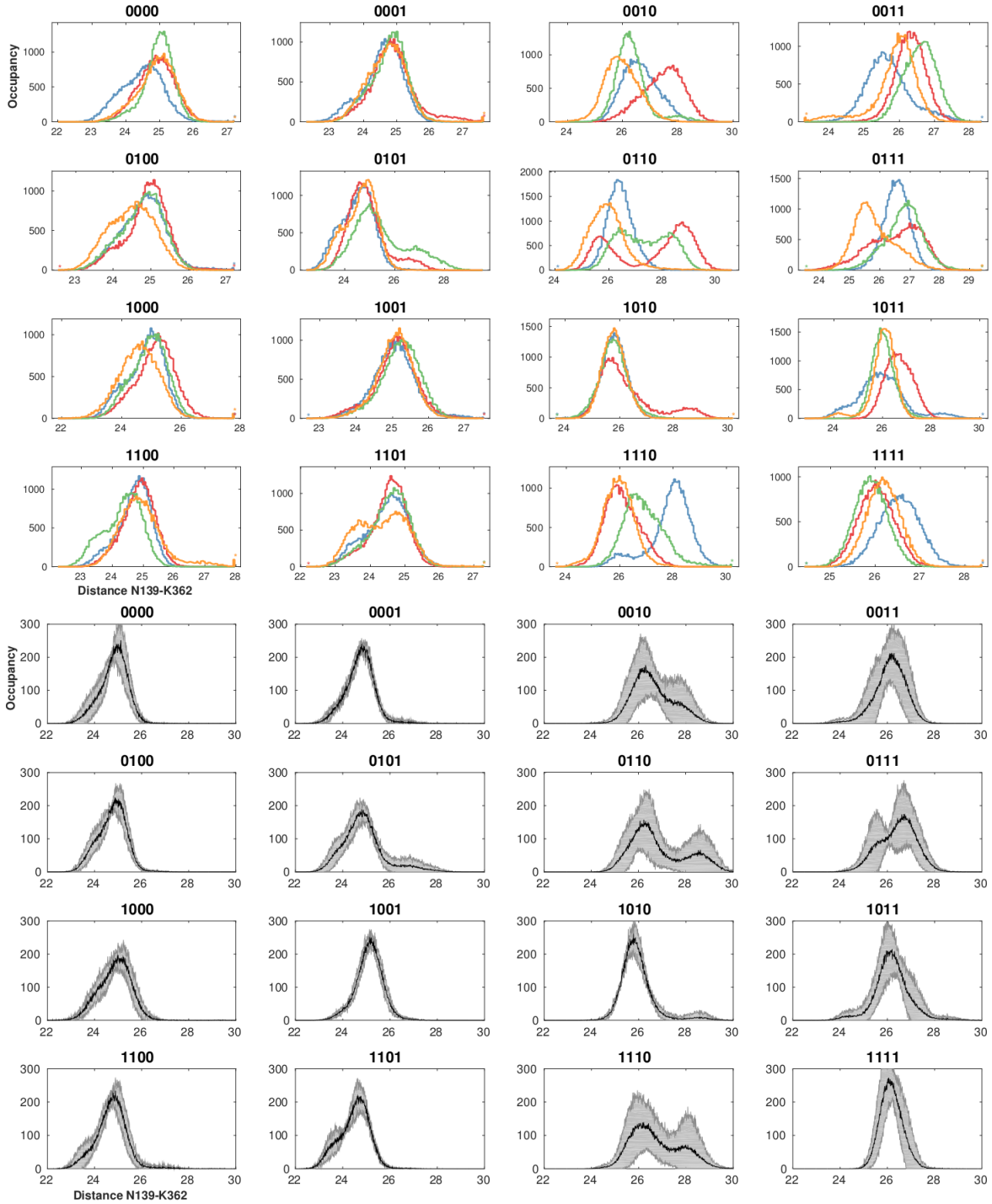


Figure A.7: Distribution of distances between N139 and K362 in different protonation states of Cytochrome c oxidase from (*up*) individual MD simulations and (*down*) averaged over the different runs. The protonation models are labeled by "1" for protonated and "0" for unprotonated D132, E286, K362, and E101, respectively.

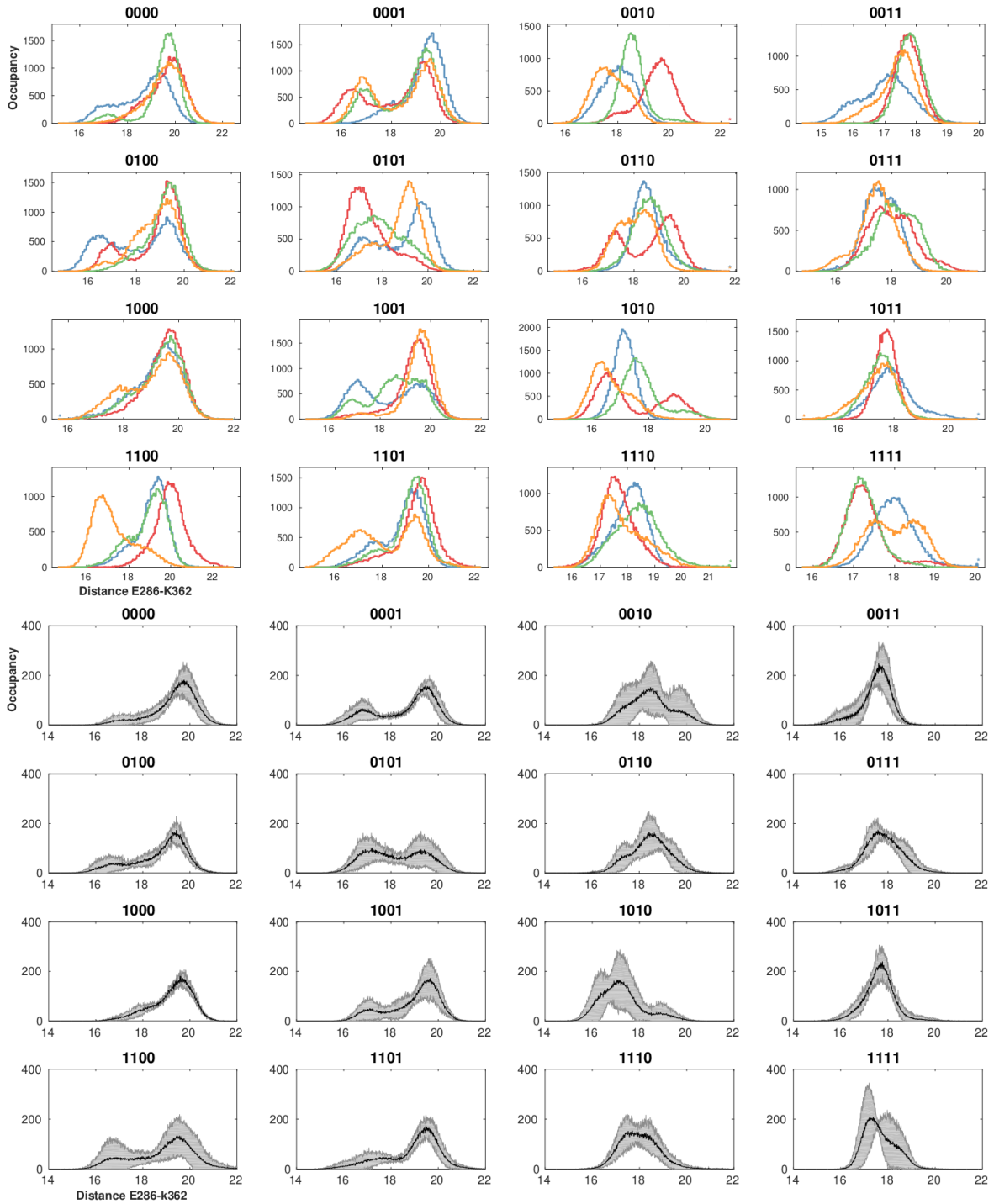


Figure A.8: Distribution of distances between E286 and K362 in different protonation states of Cytochrome c oxidase from (*up*) individual MD simulations and (*down*) averaged over the different runs. The protonation models are labeled by "1" for protonated and "0" for unprotonated D132, E286, K362, and E101, respectively.

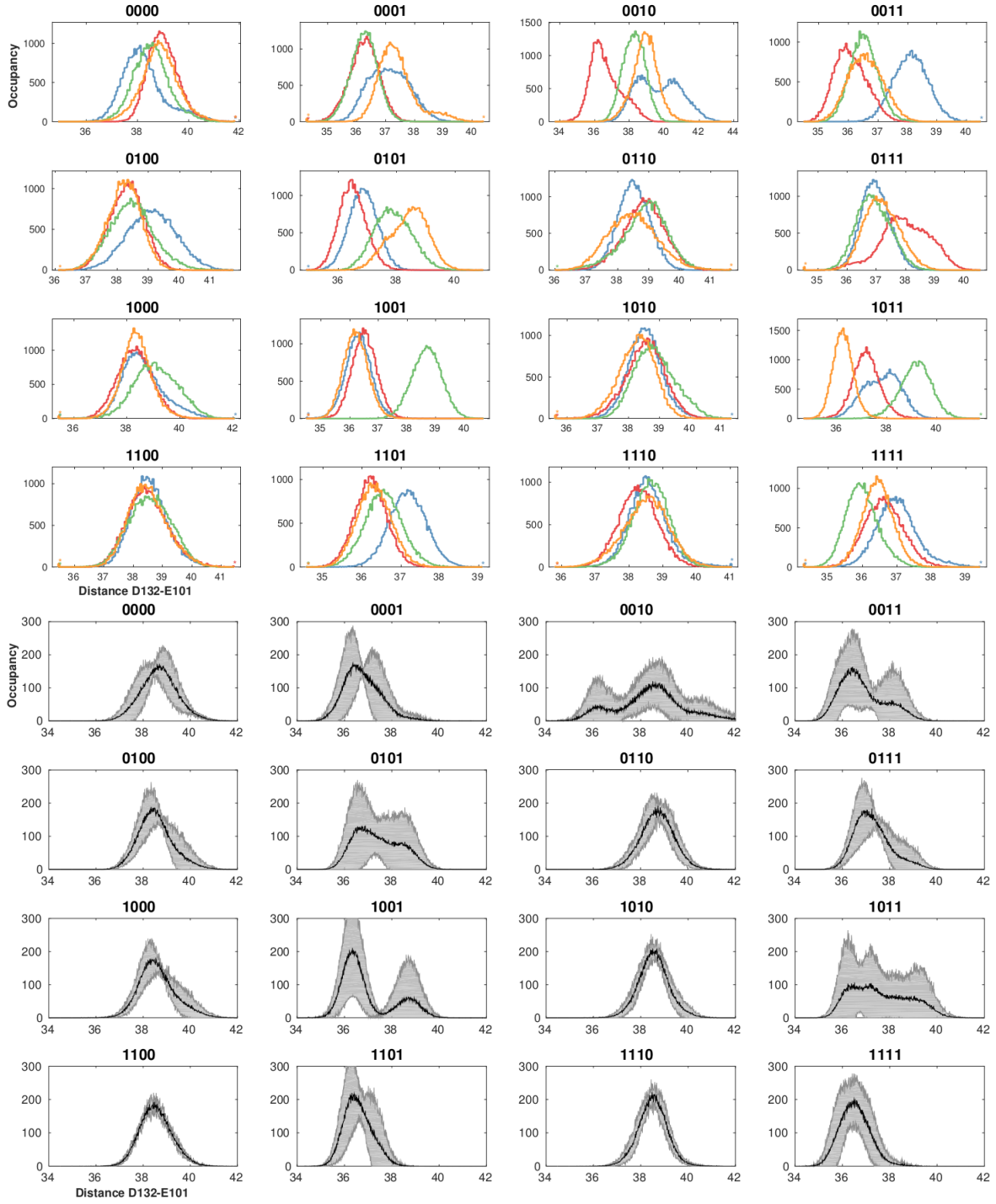


Figure A.9: Distribution of distances between D132 and E101 in different protonation states of Cytochrome c oxidase from (*up*) individual MD simulations and (*down*) averaged over the different runs. The protonation models are labeled by "1" for protonated and "0" for unprotonated D132, E286, K362, and E101, respectively.

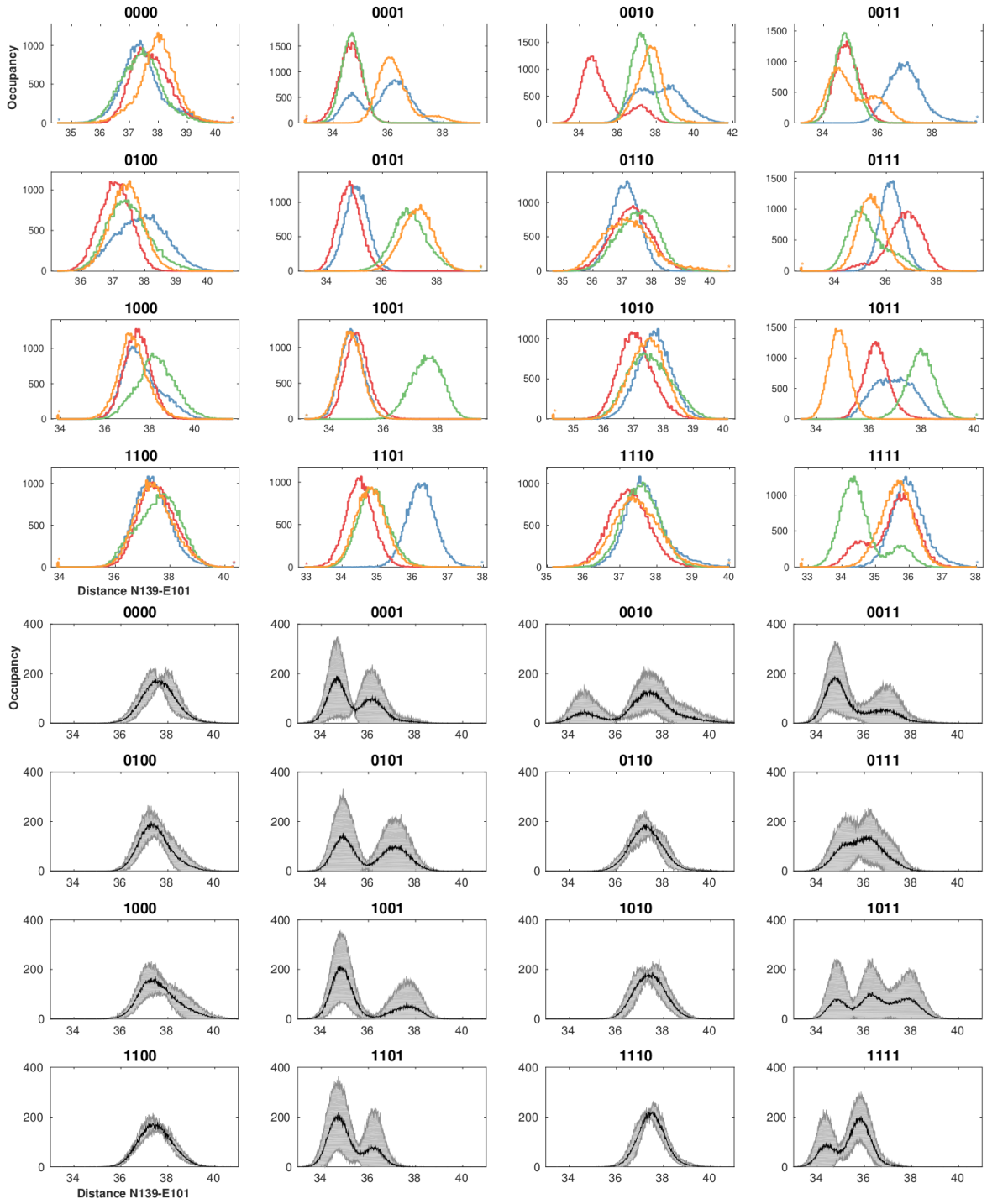


Figure A.10: Distribution of distances between N139 and E101 in different protonation states of Cytochrome c oxidase from (*up*) individual MD simulations and (*down*) averaged over the different runs. The protonation models are labeled by "1" for protonated and "0" for unprotonated D132, E286, K362, and E101, respectively.

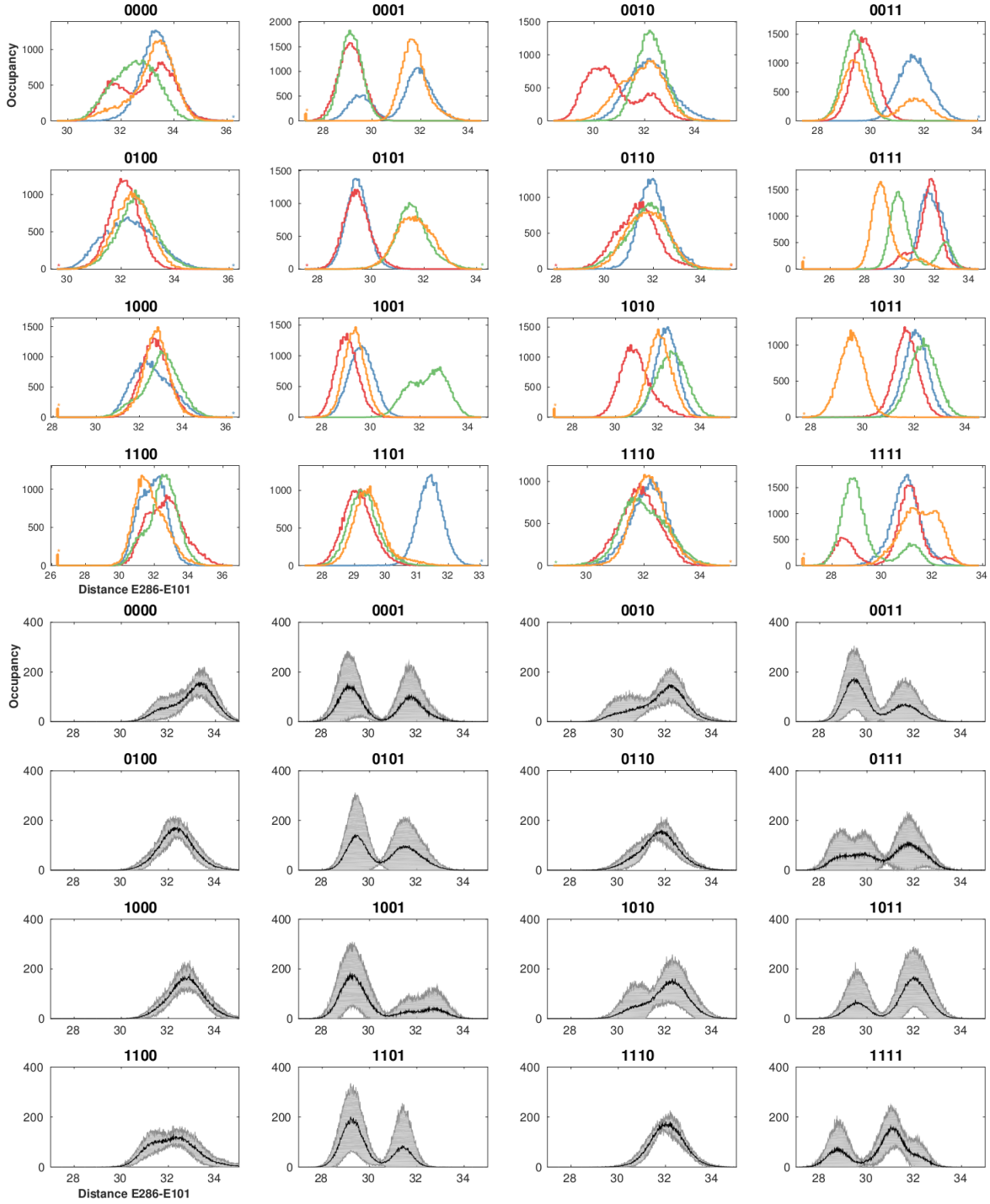


Figure A.11: Distribution of distances between E286 and E101 in different protonation states of Cytochrome c oxidase from (*up*) individual MD simulations and (*down*) averaged over the different runs. The protonation models are labeled by "1" for protonated and "0" for unprotonated D132, E286, K362, and E101, respectively.

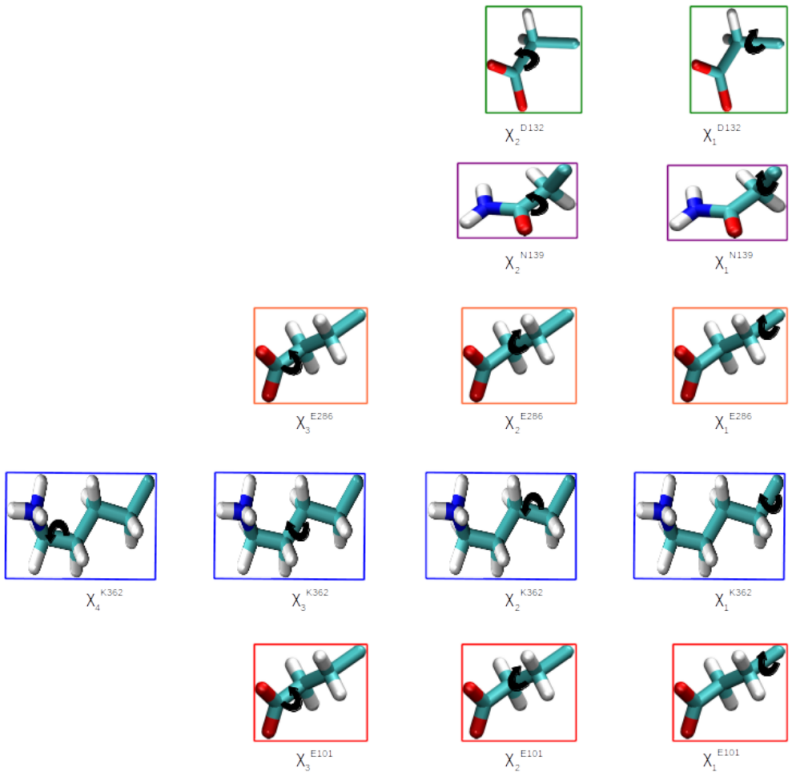


Figure A.12: Side chain dihedral angles of key residues D132, N139, E286, K362, and E101.

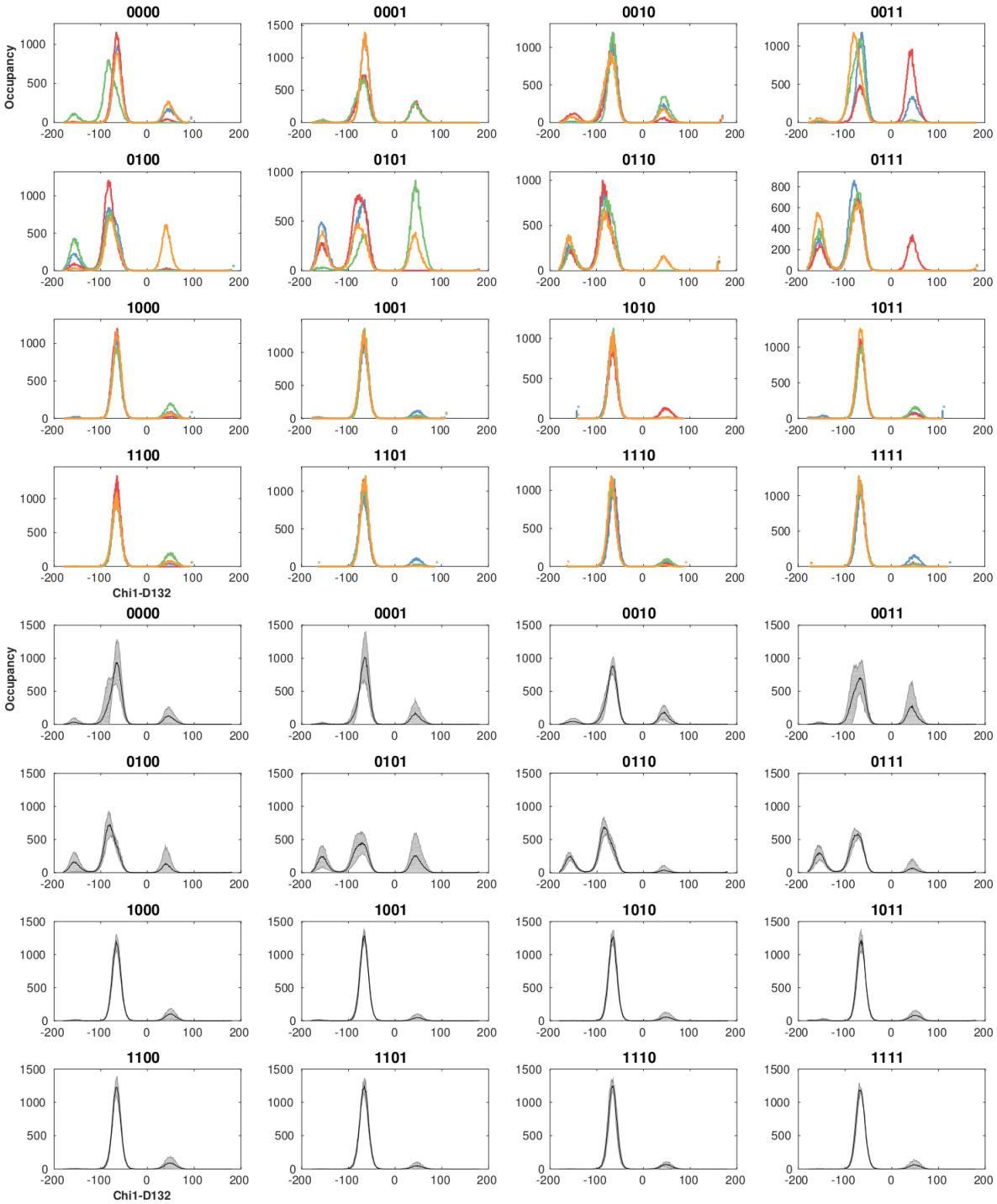


Figure A.13: Distribution of side chain dihedral angle $\chi 1$ of residue D132 in different protonation states of Cytochrome c oxidase from (*up*) individual MD simulations and (*down*) averaged over the different runs.The protonation models are labeled by "1" for protonated and "0" for unprotonated D132, E286, K362, and E101, respectively.

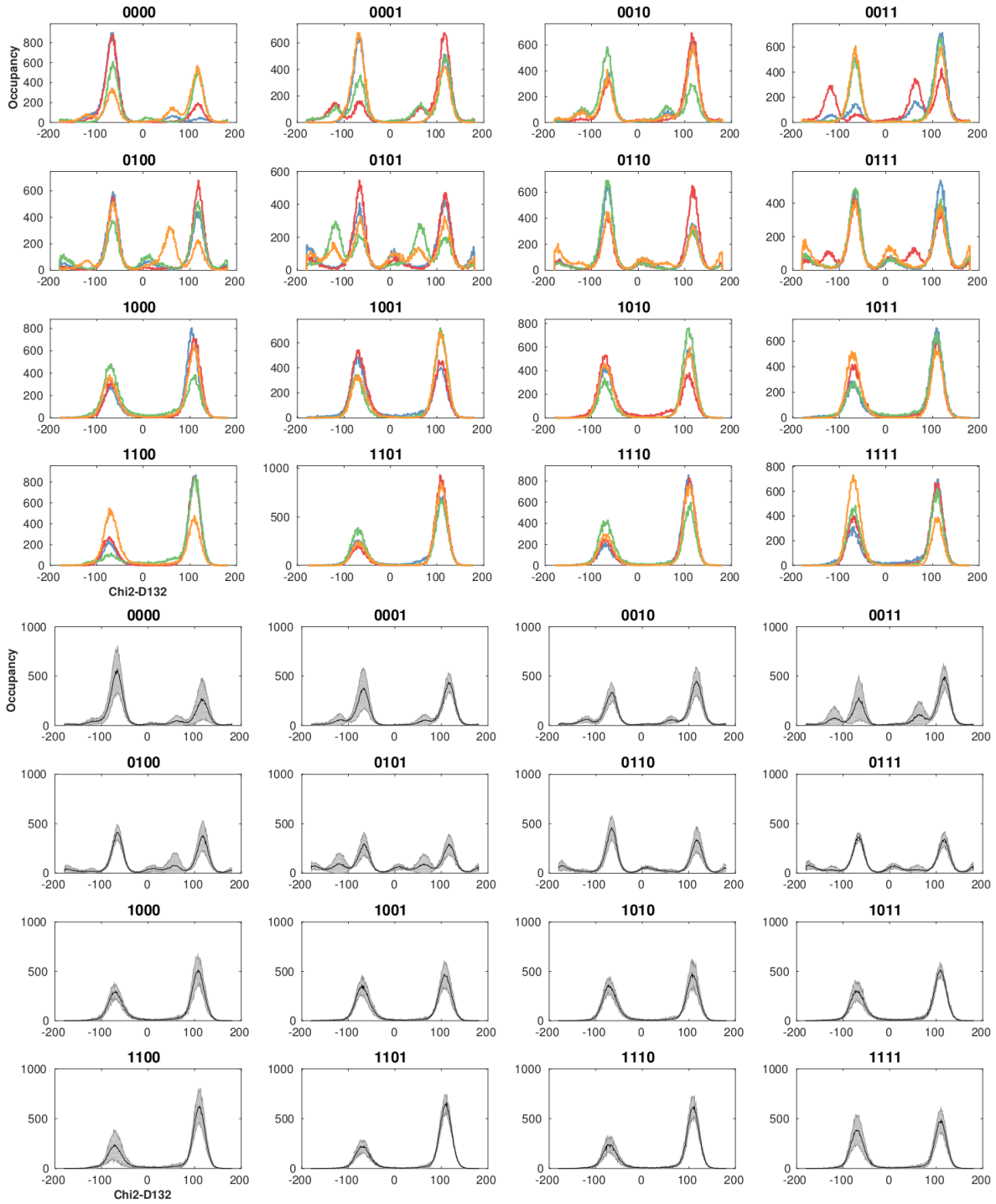


Figure A.14: Distribution of side chain dihedral angle $\chi 2$ of residue D132 in different protonation states of Cytochrome c oxidase from (*up*) individual MD simulations and (*down*) averaged over the different runs.The protonation models are labeled by "1" for protonated and "0" for unprotonated D132, E286, K362, and E101, respectively.

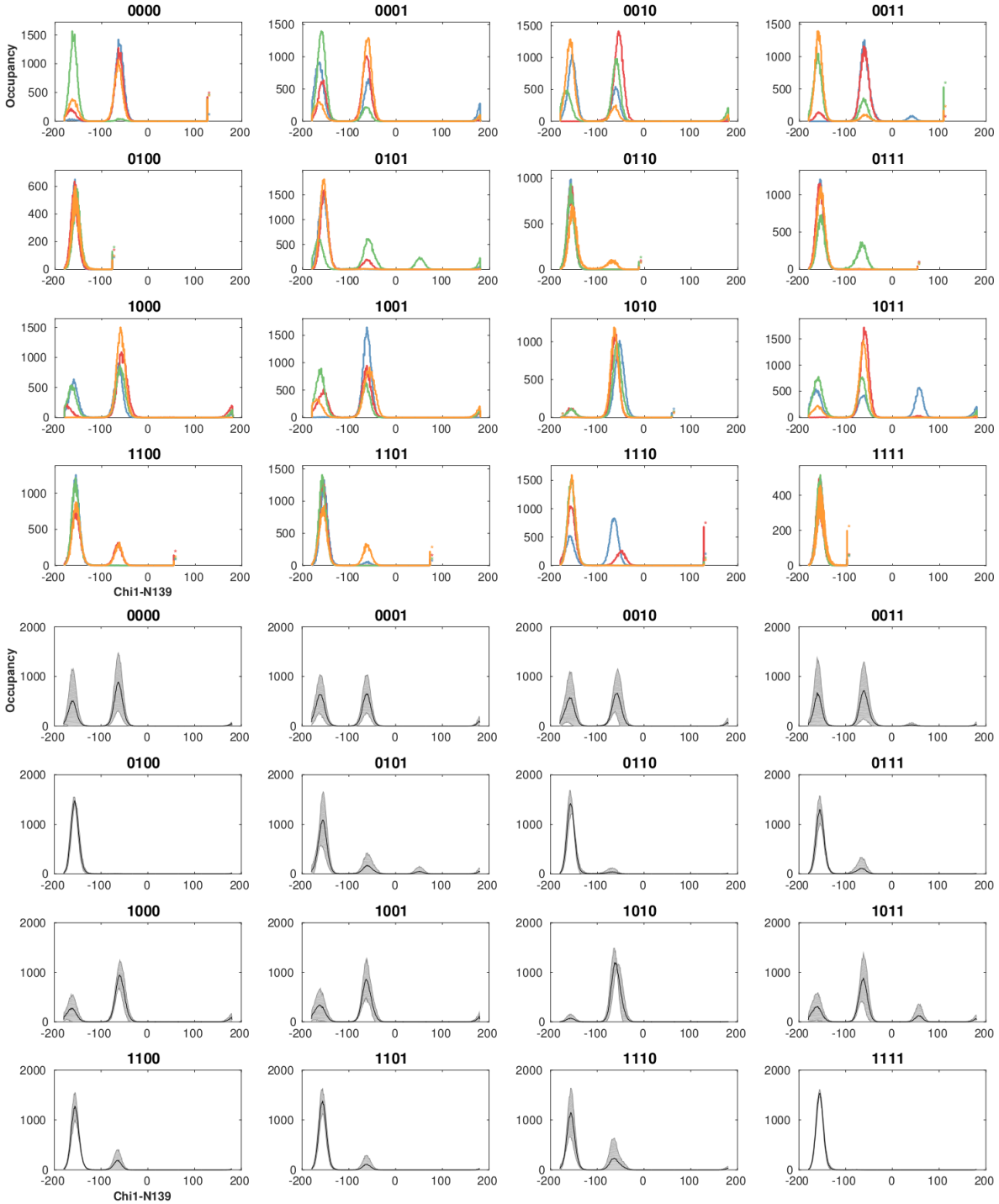


Figure A.15: Distribution of side chain dihedral angle $\chi 1$ of residue N139 in different protonation states of Cytochrome c oxidase from (*up*) individual MD simulations and (*down*) averaged over the different runs.The protonation models are labeled by "1" for protonated and "0" for unprotonated D132, E286, K362, and E101, respectively.

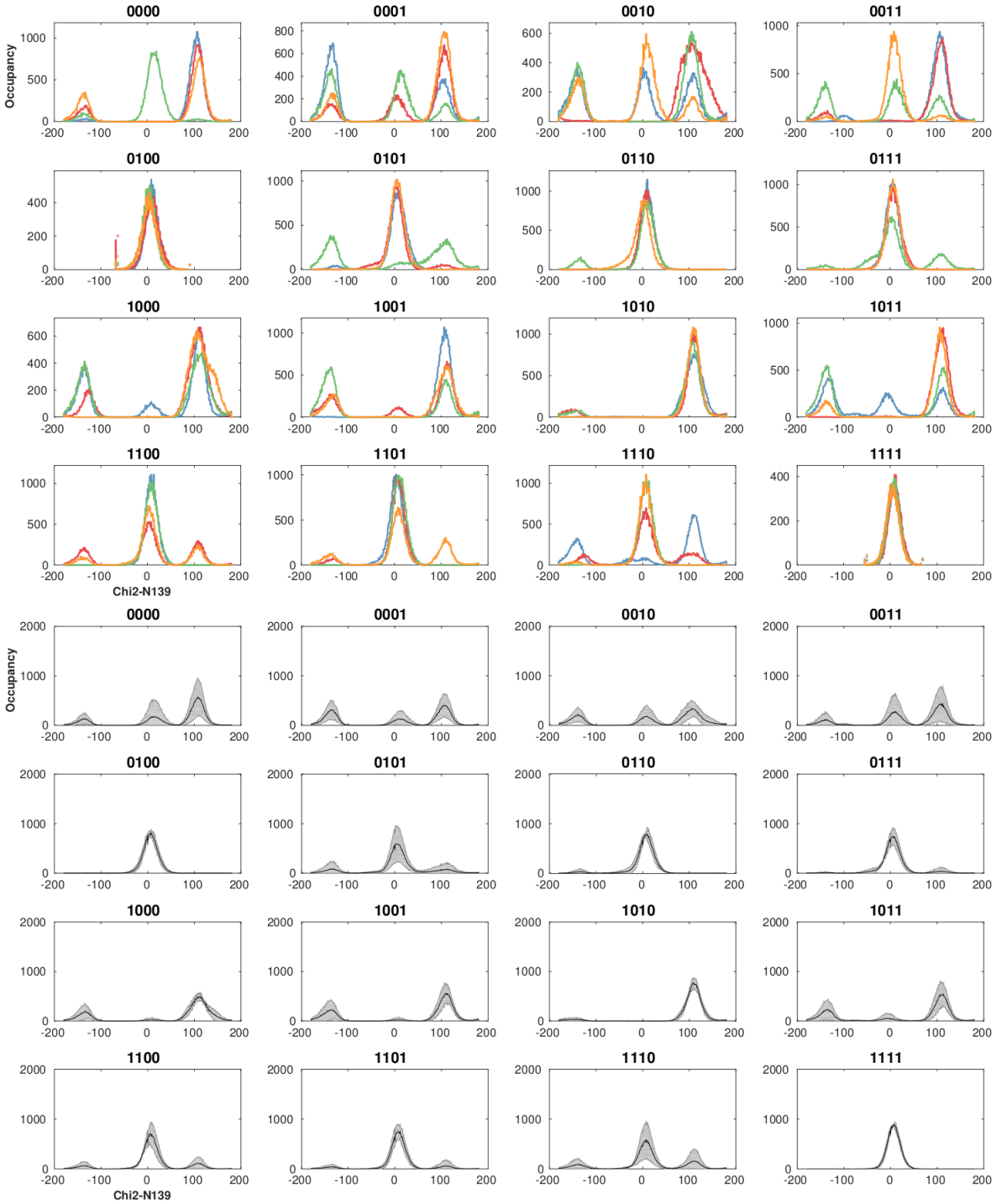


Figure A.16: Distribution of side chain dihedral angle $\chi 2$ of residue N139 in different protonation states of Cytochrome c oxidase from (*up*) individual MD simulations and (*down*) averaged over the different runs.The protonation models are labeled by "1" for protonated and "0" for unprotonated D132, E286, K362, and E101, respectively.

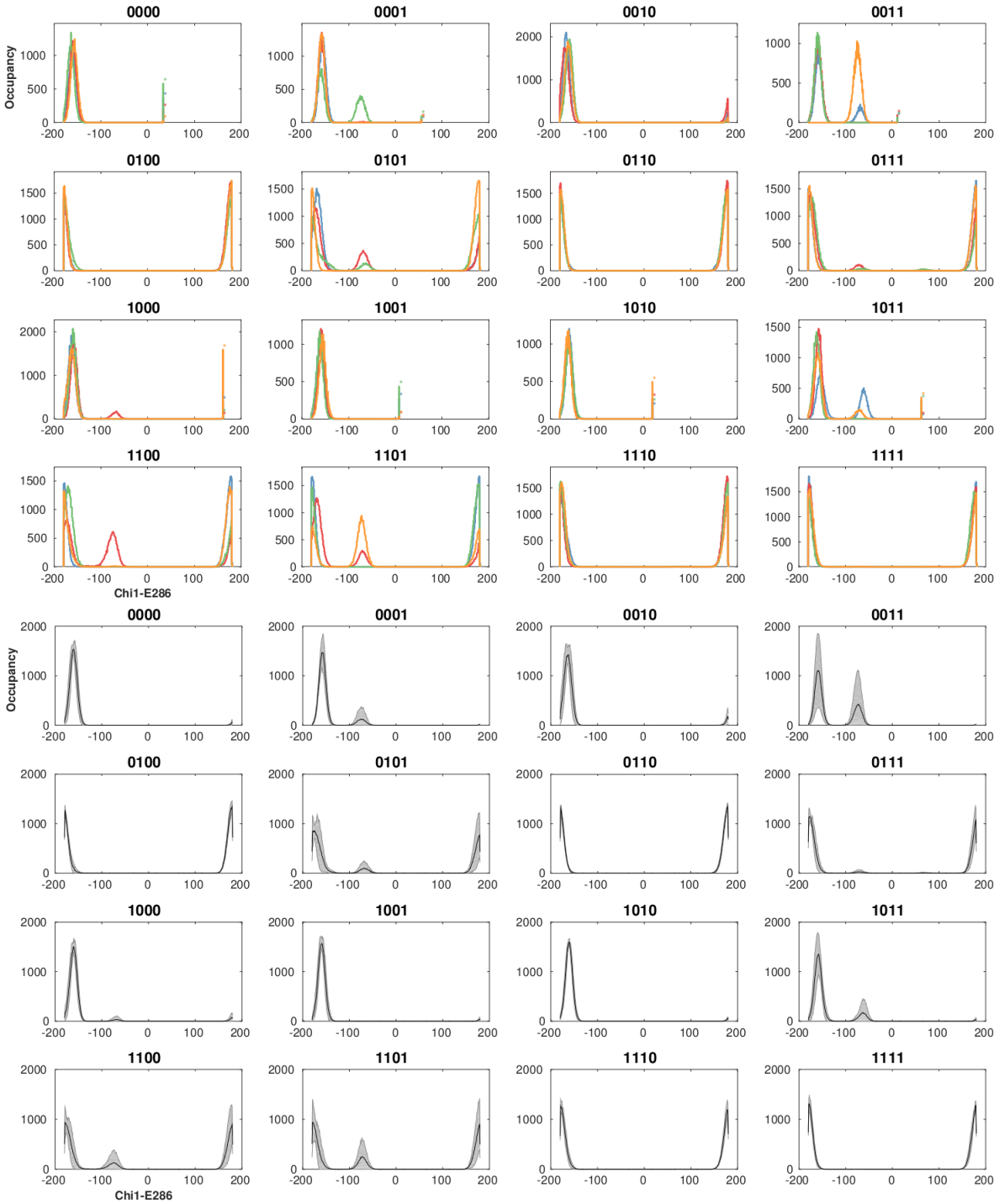


Figure A.17: Distribution of side chain dihedral angle $\chi 1$ of residue E286 in different protonation states of Cytochrome c oxidase from (*up*) individual MD simulations and (*down*) averaged over the different runs.The protonation models are labeled by "1" for protonated and "0" for unprotonated D132, E286, K362, and E101, respectively.

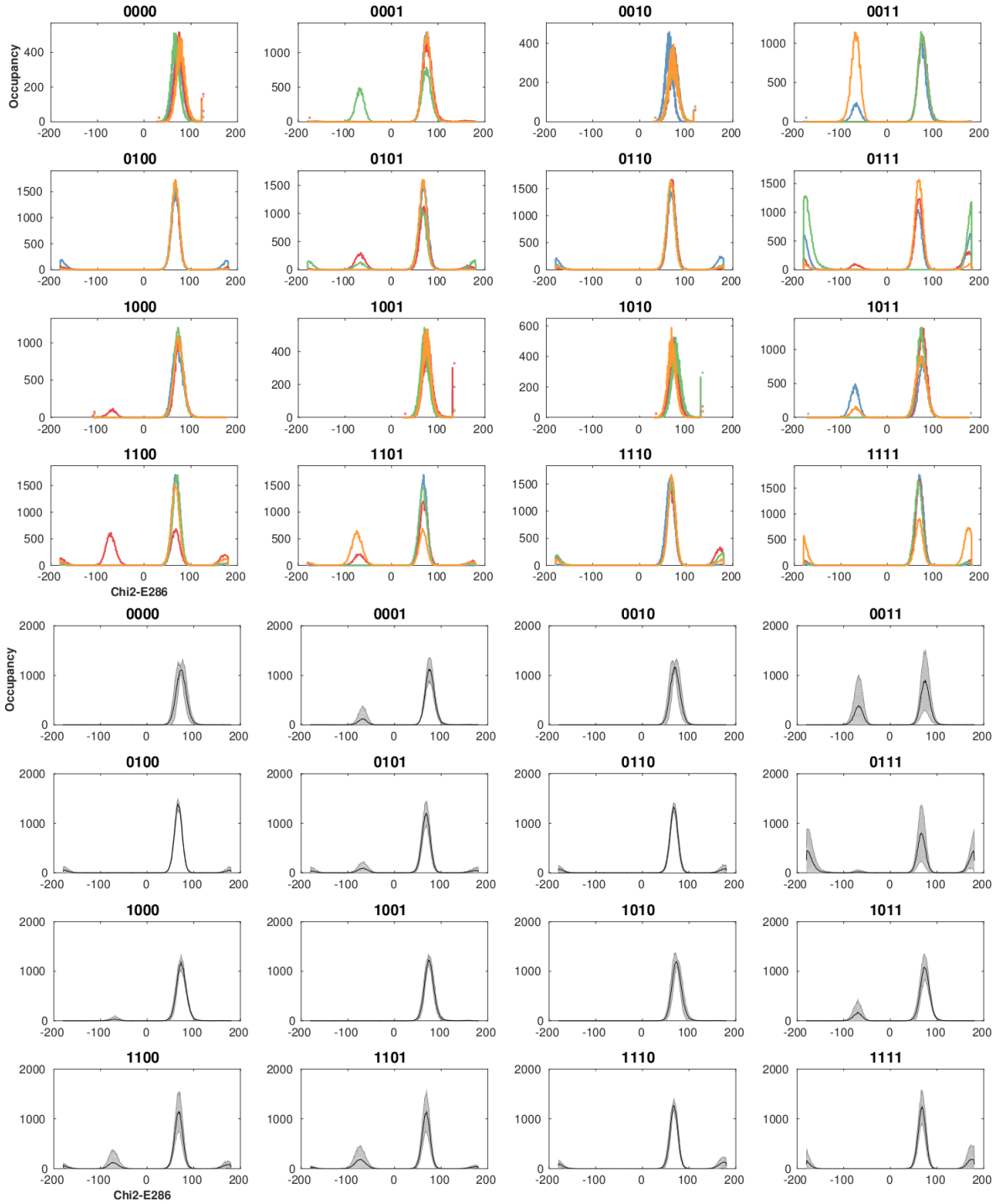


Figure A.18: Distribution of side chain dihedral angle $\chi 2$ of residue E286 in different protonation states of Cytochrome c oxidase from (*up*) individual MD simulations and (*down*) averaged over the different runs.The protonation models are labeled by "1" for protonated and "0" for unprotonated D132, E286, K362, and E101, respectively.

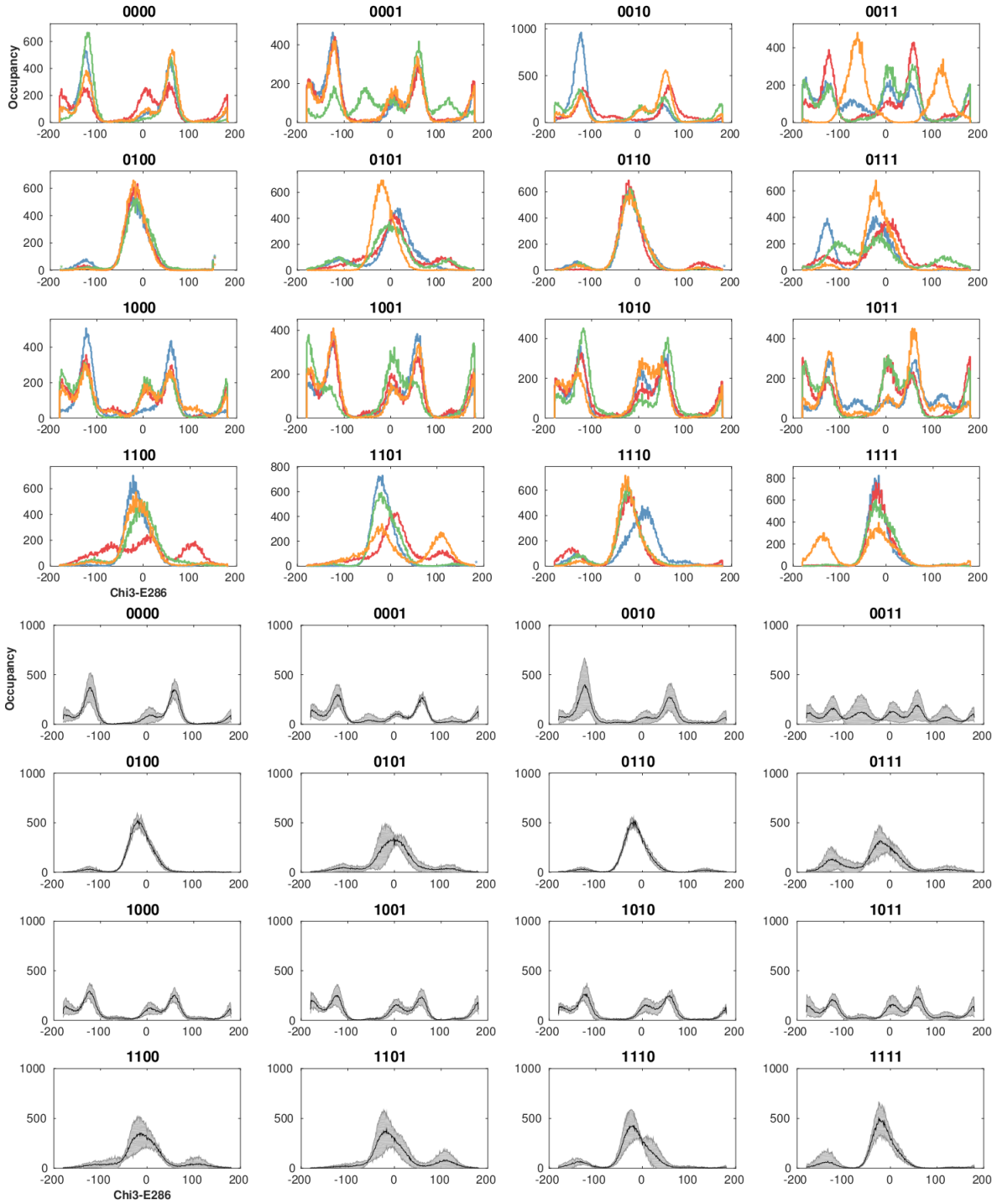


Figure A.19: Distribution of side chain dihedral angle $\chi 3$ of residue E286 in different protonation states of Cytochrome c oxidase from (*up*) individual MD simulations and (*down*) averaged over the different runs.The protonation models are labeled by "1" for protonated and "0" for unprotonated D132, E286, K362, and E101, respectively.

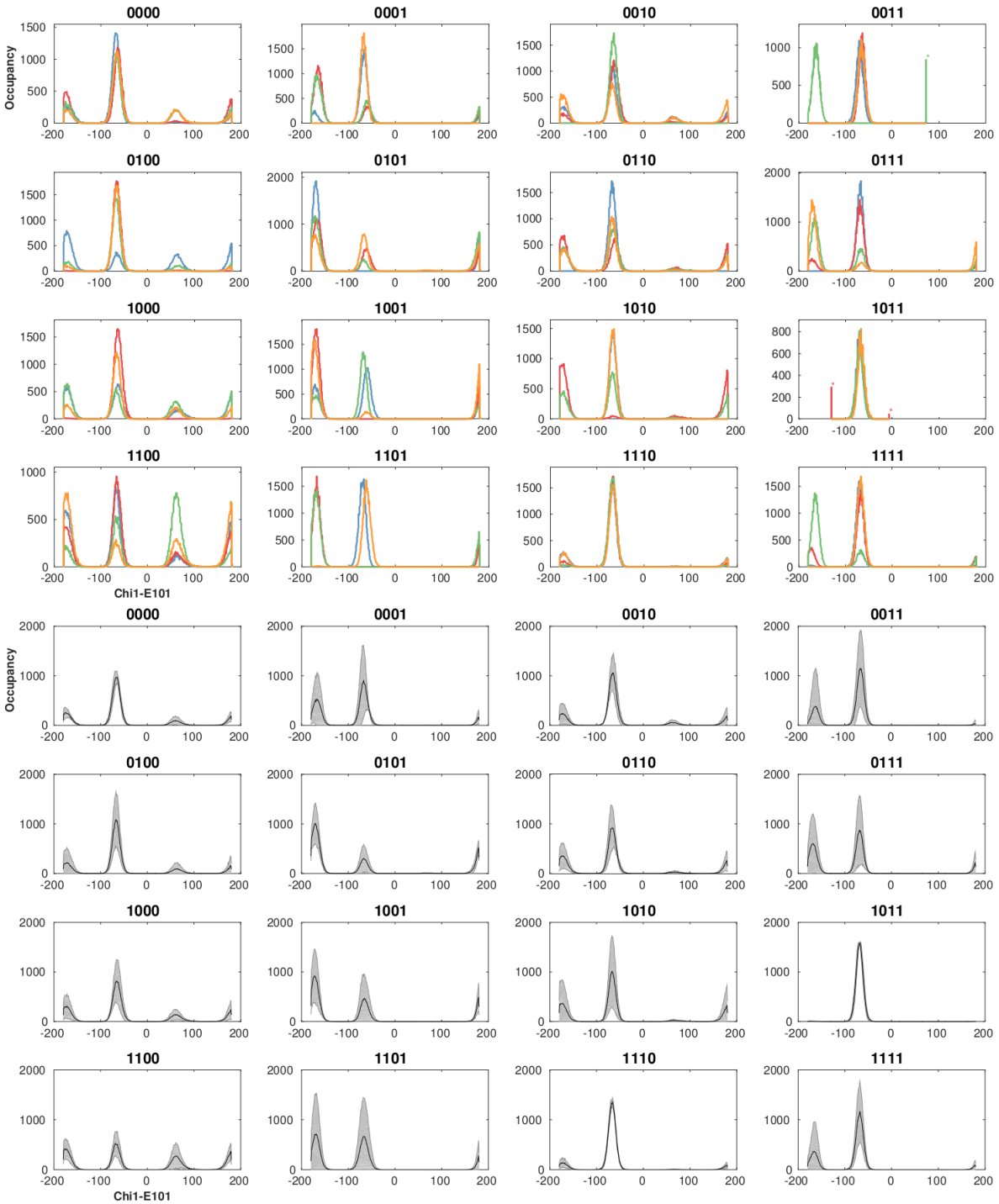


Figure A.20: Distribution of side chain dihedral angle $\chi 1$ of residue E101 in different protonation states of Cytochrome c oxidase from (*up*) individual MD simulations and (*down*) averaged over the different runs.The protonation models are labeled by "1" for protonated and "0" for unprotonated D132, E286, K362, and E101, respectively.

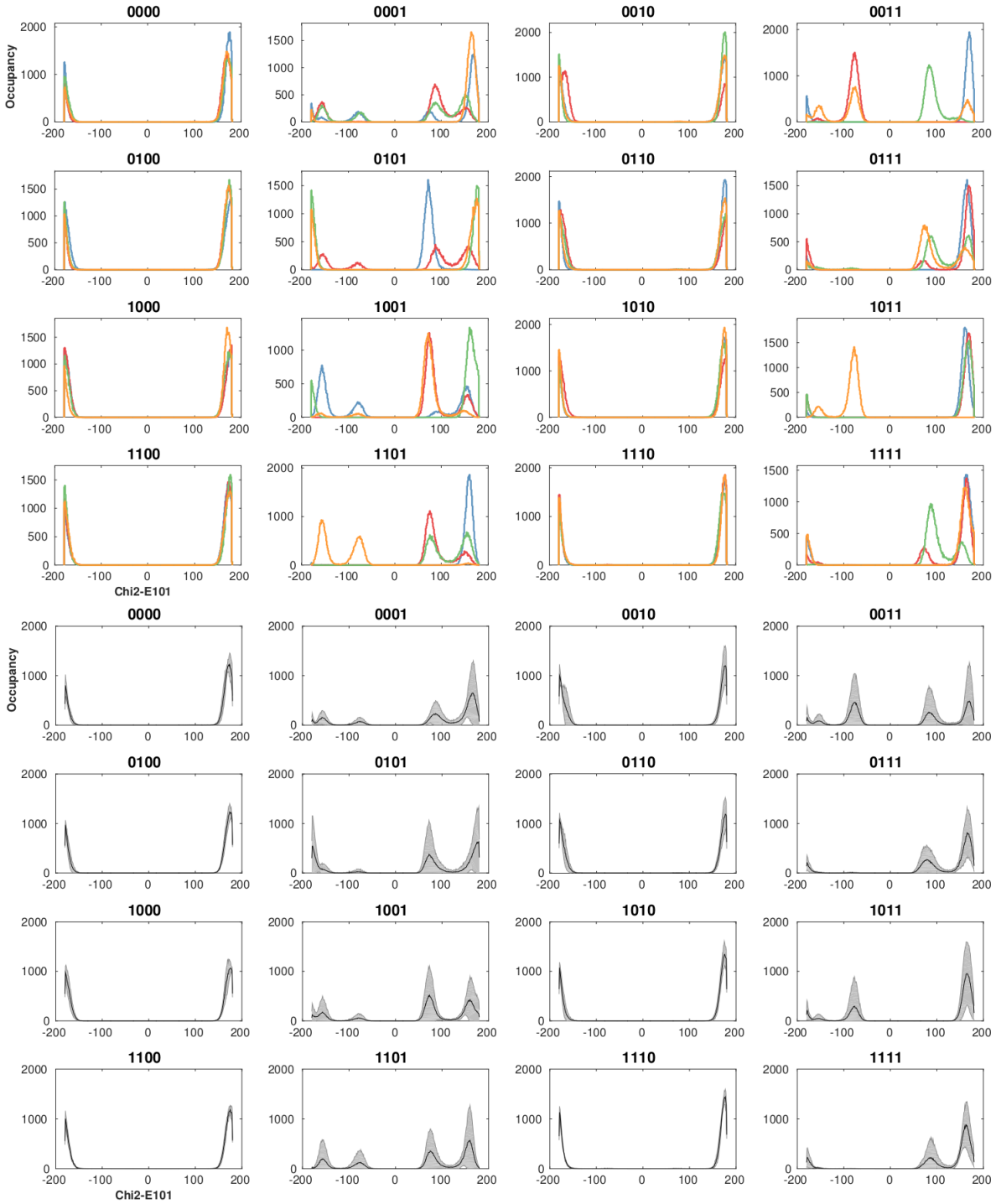


Figure A.21: Distribution of side chain dihedral angle $\chi 2$ of residue E101 in different protonation states of Cytochrome c oxidase from (*up*) individual MD simulations and (*down*) averaged over the different runs.The protonation models are labeled by "1" for protonated and "0" for unprotonated D132, E286, K362, and E101, respectively.

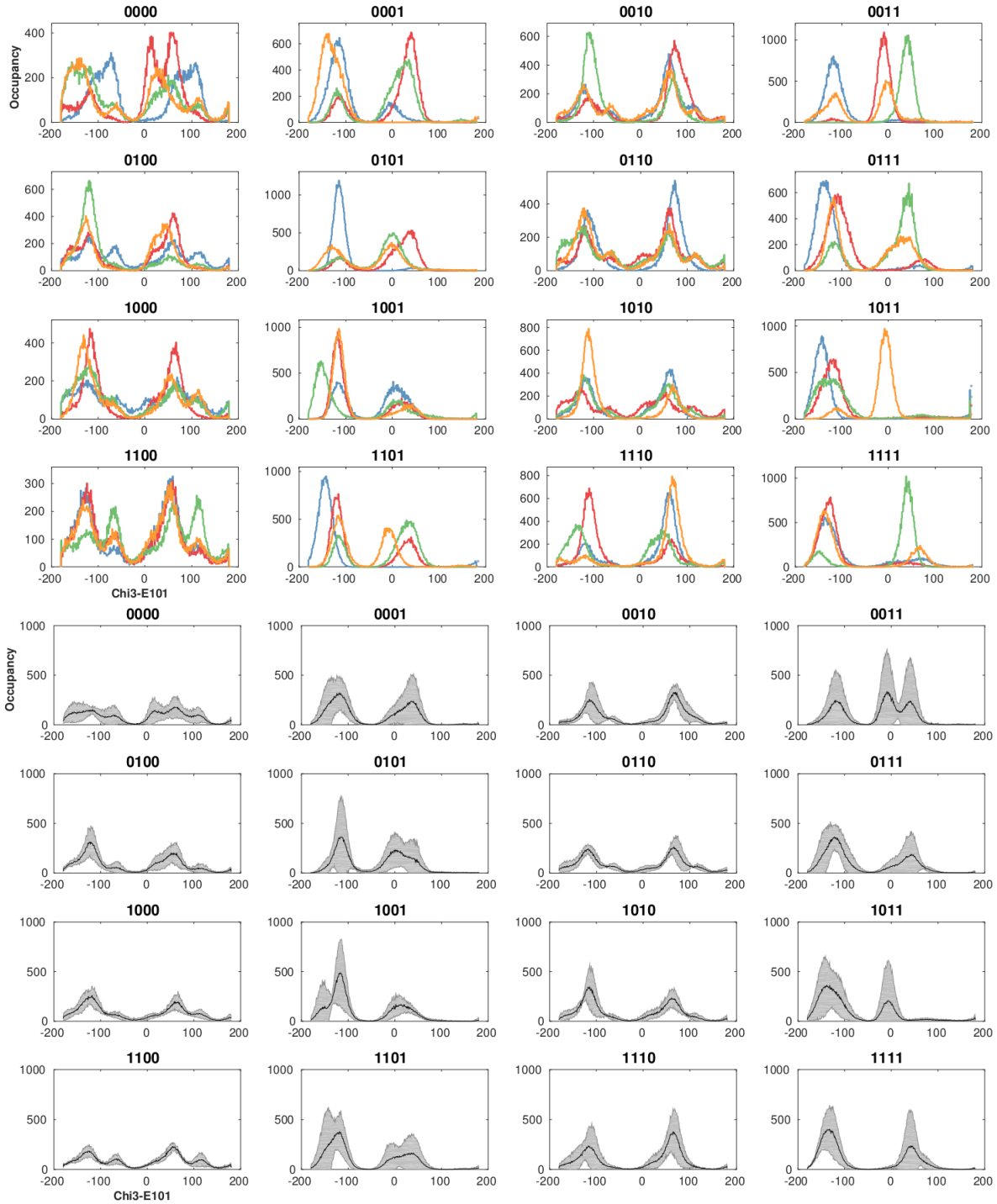


Figure A.22: Distribution of side chain dihedral angle $\chi 3$ of residue E101 in different protonation states of Cytochrome c oxidase from (*up*) individual MD simulations and (*down*) averaged over the different runs.The protonation models are labeled by "1" for protonated and "0" for unprotonated D132, E286, K362, and E101, respectively.

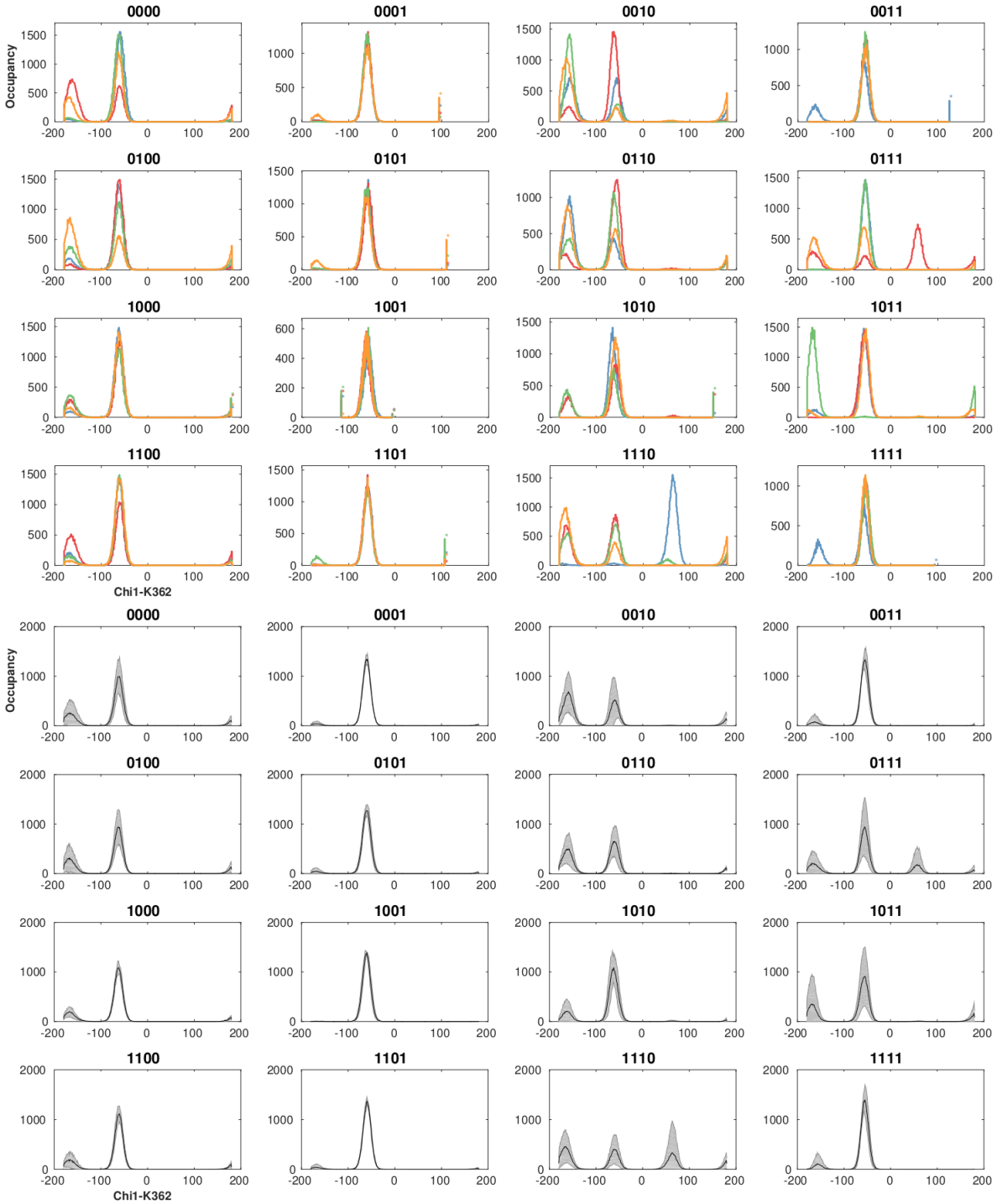


Figure A.23: Distribution of side chain dihedral angle $\chi 1$ of residue K362 in different protonation states of Cytochrome c oxidase from (*up*) individual MD simulations and (*down*) averaged over the different runs.The protonation models are labeled by "1" for protonated and "0" for unprotonated D132, E286, K362, and E101, respectively.

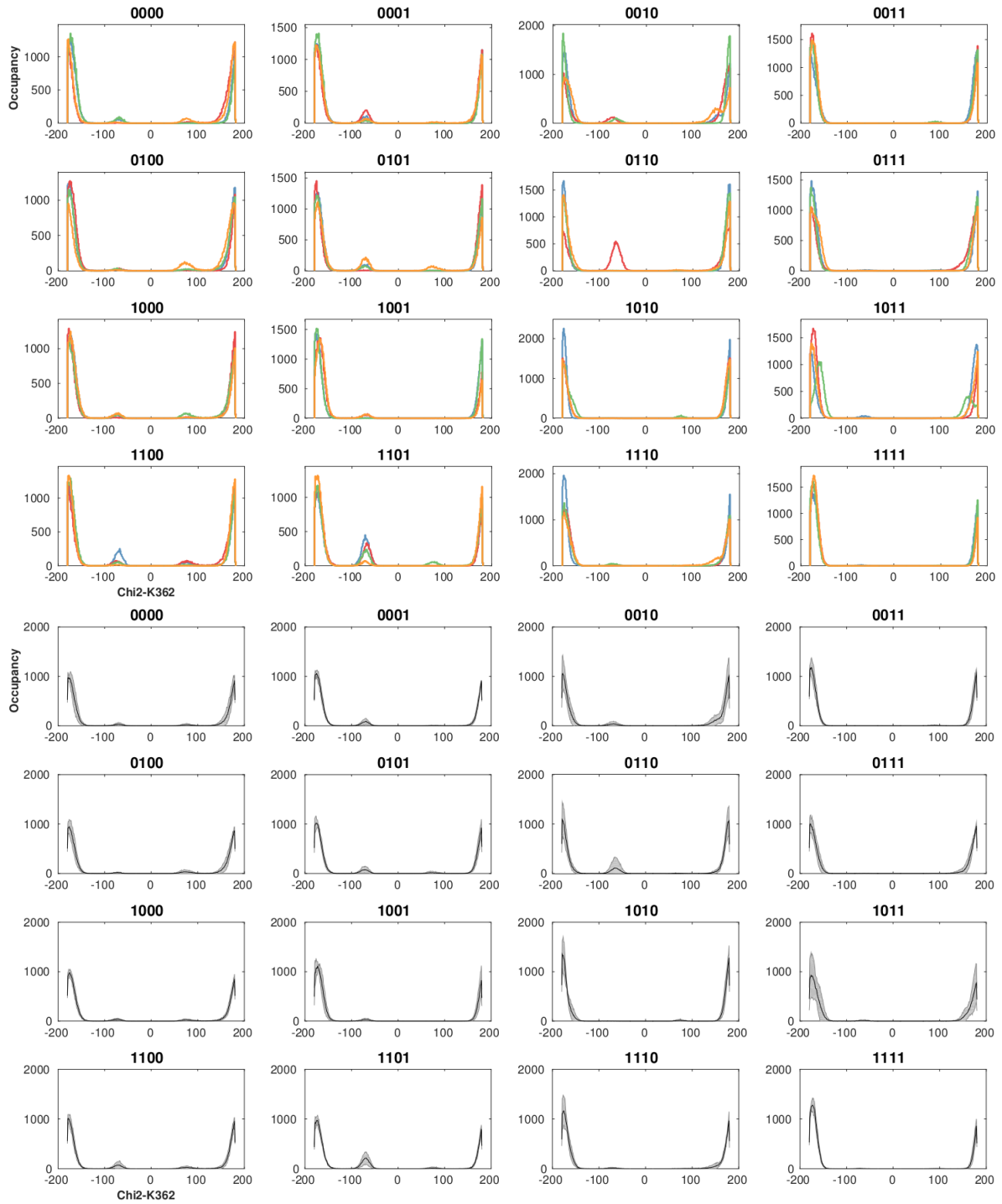


Figure A.24: Distribution of side chain dihedral angle $\chi 2$ of residue K362 in different protonation states of Cytochrome c oxidase from (*up*) individual MD simulations and (*down*) averaged over the different runs.The protonation models are labeled by "1" for protonated and "0" for unprotonated D132, E286, K362, and E101, respectively.

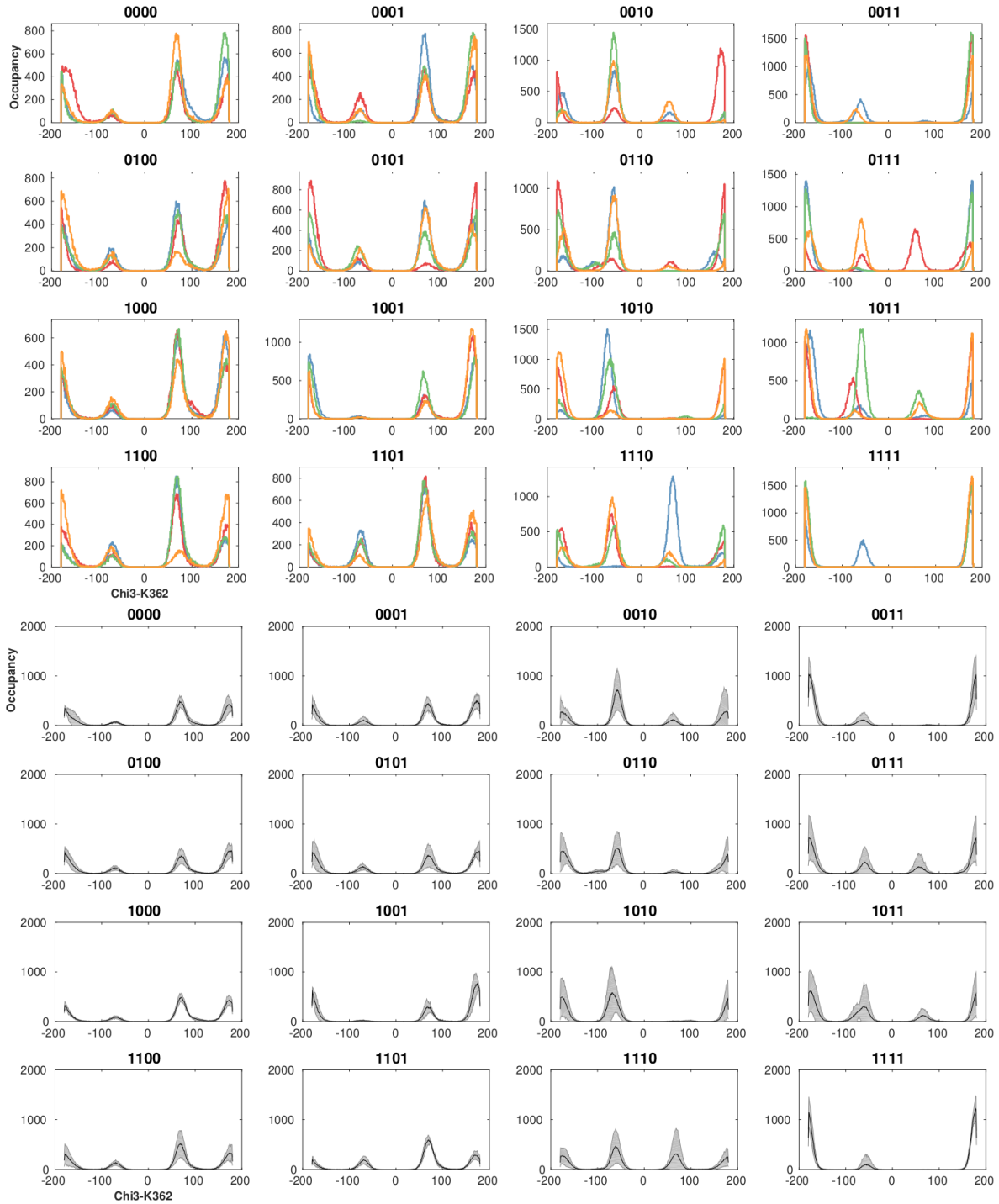


Figure A.25: Distribution of side chain dihedral angle $\chi 3$ of residue K362 in different protonation states of Cytochrome c oxidase from (*up*) individual MD simulations and (*down*) averaged over the different runs.The protonation models are labeled by "1" for protonated and "0" for unprotonated D132, E286, K362, and E101, respectively.

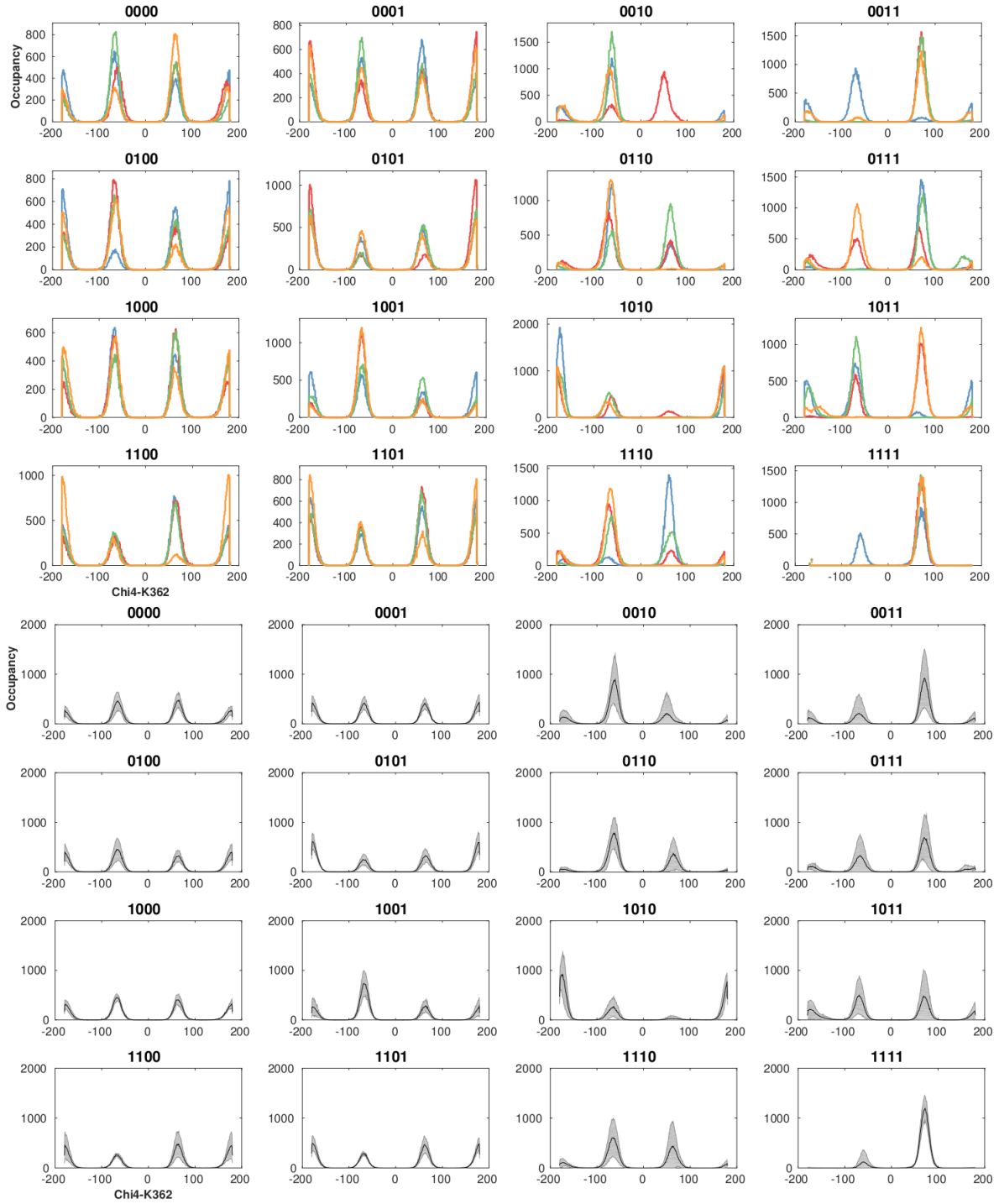


Figure A.26: Distribution of side chain dihedral angle $\chi 4$ of residue K362 in different protonation states of Cytochrome c oxidase from (*up*) individual MD simulations and (*down*) averaged over the different runs.The protonation models are labeled by "1" for protonated and "0" for unprotonated D132, E286, K362, and E101, respectively.

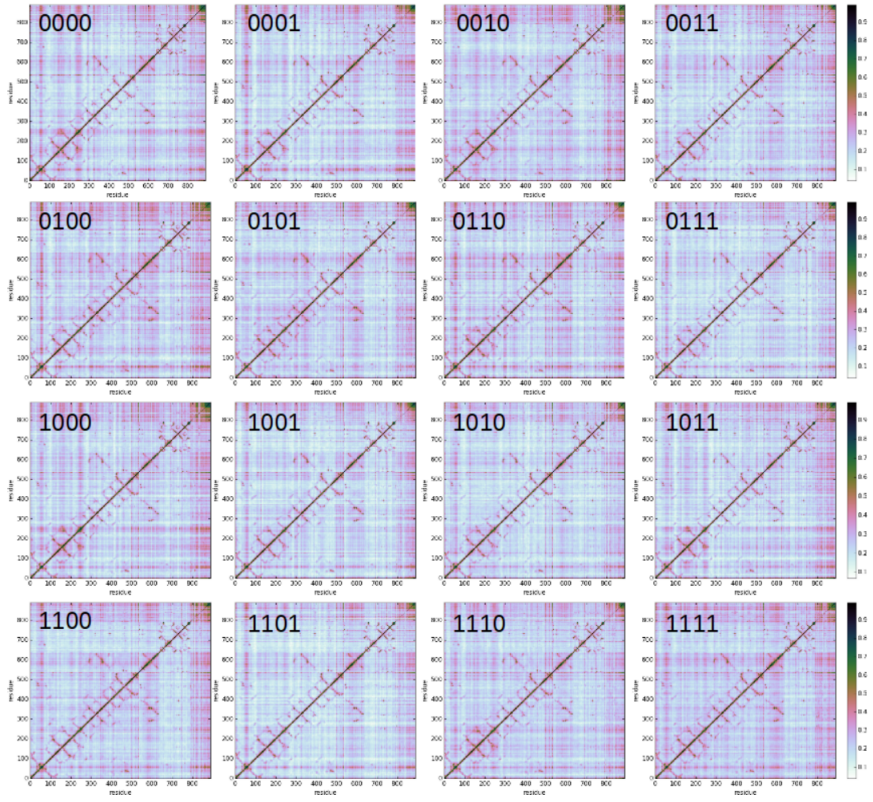


Figure A.27: Generalized correlation matrices of Cytochrome C Oxidase in the different protonation state models labeled by "1" for protonated and "0" for unprotonated D132, E286, K362, and E101, respectively.

Path	0000	0001	0010	0011	0100	0101	0110	0111
E286-D132	2.4±0.3	2.7±0.2	2.4±0.22	2.5±0.2	2.7±0.1	2.6±0.5	2.4±0.37	2.5±0.3
N139-D132	0.8±0.1	0.8±0.1	0.8±0.14	0.8±0.1	0.9±0.1	0.9±0.1	0.8±0.08	0.8±0.1
E286-N139	1.8±0.2	2.1±0.2	**1.7±0.15**	1.9±0.2	2.1±0.2	2.1±0.4	1.9±0.14	**1.7±0.2**
Y288-E101	2.3±0.2	2.5±0.3	**1.8±0.40**	2.1±0.4	2.5±0.3	2.1±0.1	1.9±0.29	**1.6±0.7**
K362-Y288	1.9±0.3	1.9±0.3	1.2±0.15	1.2±0.3	1.6±0.2	1.4±0.3	1.3±0.33	**1.1±0.3**
K362-E101	1.4±0.1	1.0±0.2	1.1±0.37	1.0±0.2	1.2±0.2	1.2±0.2	1.0±0.11	**0.8±0.3**
K362-D132	3.0±0.2	2.9±0.4	2.6±0.30	2.7±0.3	2.8±0.2	2.8±0.2	**2.4±0.25**	2.5±0.4
K362-N139	3.2±0.2	3.1±0.3	3.0±0.36	2.8±0.3	3.2±0.1	3.2±0.3	2.7±0.41	2.9±0.5
K362-E286	2.3±0.2	2.4±0.2	**1.6±0.17**	1.7±0.3	2.2±0.2	1.9±0.4	1.8±0.38	**1.5±0.5**
Y288-D132	2.7±0.2	3.0±0.1	2.7±0.31	2.6±0.1	2.8±0.1	2.6±0.2	2.4±0.32	2.5±0.3
D132-E101	3.0±0.2	3.1±0.4	2.8±0.46	2.9±0.2	3.1±0.2	2.9±0.1	2.6±0.23	**2.5±0.4**
N139-E101	3.4±0.2	3.4±0.4	3.3±0.55	3.2±0.2	3.6±0.3	3.5±0.2	3.0±0.43	3.0±0.5
E286-E101	2.7±0.2	3.1±0.3	2.3±0.40	2.6±0.4	3.0±0.3	2.6±0.3	2.5±0.38	**2.0±0.8**
CPL	3.1±0.2	3.2±0.1	2.9±0.05	3.1±0.2	3.3±0.1	3.1±0.3	3.0±0.22	2.8±0.4

Path	1000	1001	1010	1011	1100	1101	1110	1111
E286-D132	2.5±0.3	2.6±0.4	2.5±0.41	2.4±0.1	2.5±0.4	2.7±0.2	**2.1±0.36**	2.5±0.2
N139-D132	0.9±0.1	0.8±0.2	0.9±0.20	0.8±0.1	0.9±0.1	0.9±0.1	**0.7±0.07**	0.8±0.1
E286-N139	1.8±0.3	1.9±0.4	2.0±0.25	1.9±0.2	1.9±0.4	2.1±0.3	**1.7±0.24**	2.0±0.3
Y288-E101	2.4±0.4	2.4±0.1	2.1±0.56	2.1±0.3	2.4±0.2	2.7±0.1	**1.8±0.18**	2.1±0.3
K362-Y288	1.6±0.4	1.6±0.4	1.2±0.54	1.2±0.4	1.7±0.2	2.0±0.3	**1.1±0.08**	1.4±0.3
K362-E101	1.3±0.2	1.1±0.1	1.1±0.27	1.0±0.2	1.3±0.0	1.1±0.1	1.1±0.19	1.0±0.1
K362-D132	2.6±0.2	2.9±0.4	**2.4±0.41**	2.5±0.3	2.8±0.2	3.1±0.1	**2.4±0.18**	2.8±0.2
K362-N139	3.0±0.4	2.9±0.4	2.6±0.57	2.7±0.2	3.2±0.2	3.4±0.2	**2.5±0.26**	2.9±0.3
K362-E286	2.1±0.4	2.0±0.3	1.6±0.57	**1.6±0.4**	2.2±0.2	2.5±0.3	**1.5±0.09**	1.9±0.3
Y288-D132	2.7±0.3	2.9±0.3	2.6±0.42	2.5±0.2	2.7±0.4	2.8±0.2	**2.2±0.33**	2.6±0.1
D132-E101	3.0±0.3	3.1±0.5	2.9±0.43	2.8±0.5	3.2±0.1	3.3±0.1	2.7±0.17	2.9±0.1
N139-E101	3.5±0.3	3.5±0.5	3.4±0.53	3.3±0.4	3.6±0.1	3.7±0.1	3.0±0.32	3.2±0.1
E286-E101	2.9±0.5	2.9±0.2	2.6±0.62	2.5±0.4	3.0±0.3	3.2±0.1	**2.1±0.30**	2.6±0.4
CPL	3.0±0.2	3.0±0.1	3.0±0.30	3.1±0.1	3.2±0.3	3.3±0.2	3.0±0.19	3.2±0.2

Figure A.28: Shortest correlation paths lengths. CPL is the characteristic path length. The protonation models are labeled by "1" for protonated and "0" for unprotonated D132, E286, K362, and E101, respectively. Values in bold are those shown in the main text.

Appendix B

Supplementary Information for: Specificity of Androgen and Glucocorticoid Receptor DNA Binding Domains for Direct and Inverted Repeat Response Elements

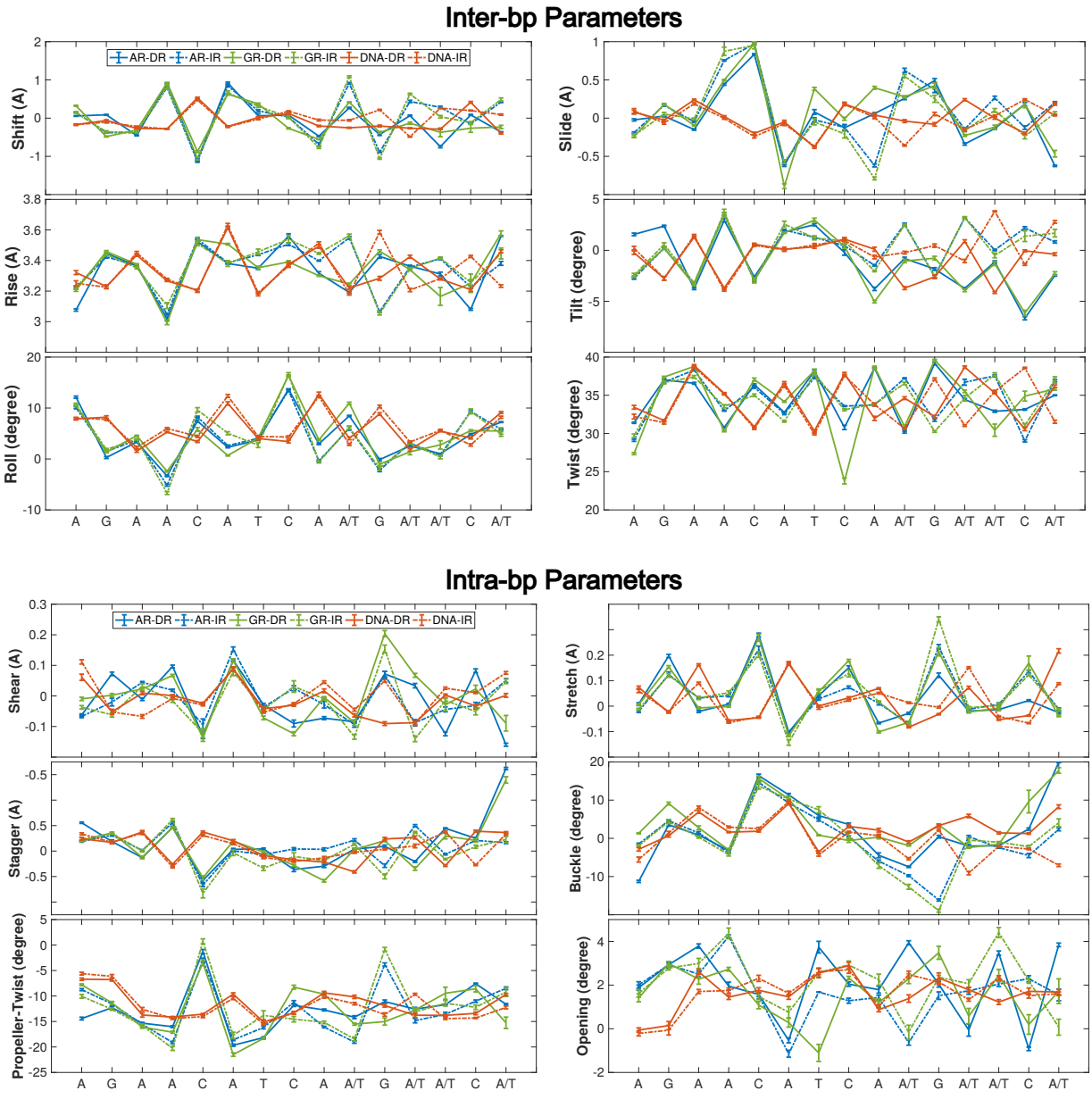


Figure B.1: DNA inter and intra bps parameters, calculated for both bound and unbound DNA.

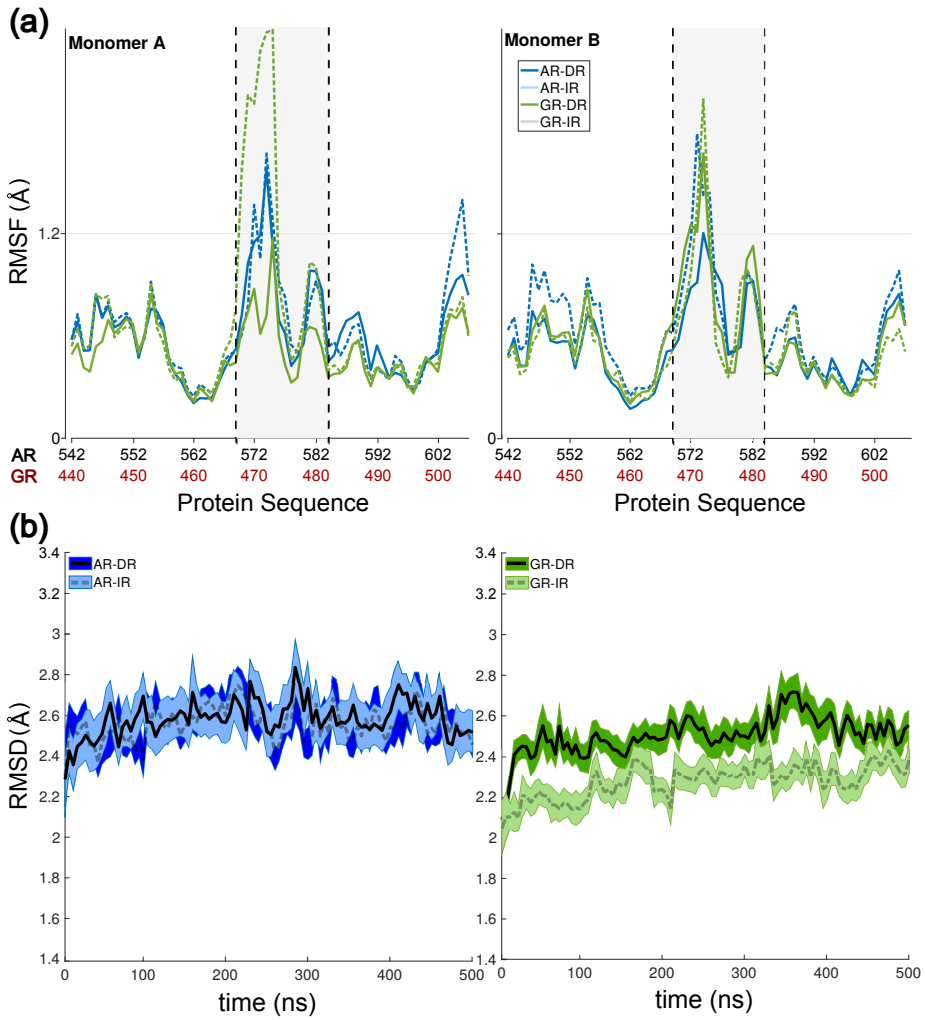


Figure B.2: **(a)** Root Mean Square Fluctuation of CA atoms of protein (monomer A&B) for all systems. Vertical dashed lines indicate the lever arm and Dim regions. **(b)** RMSD values for the AR/GR-DR/IR complexes in the course of the MD simulation.

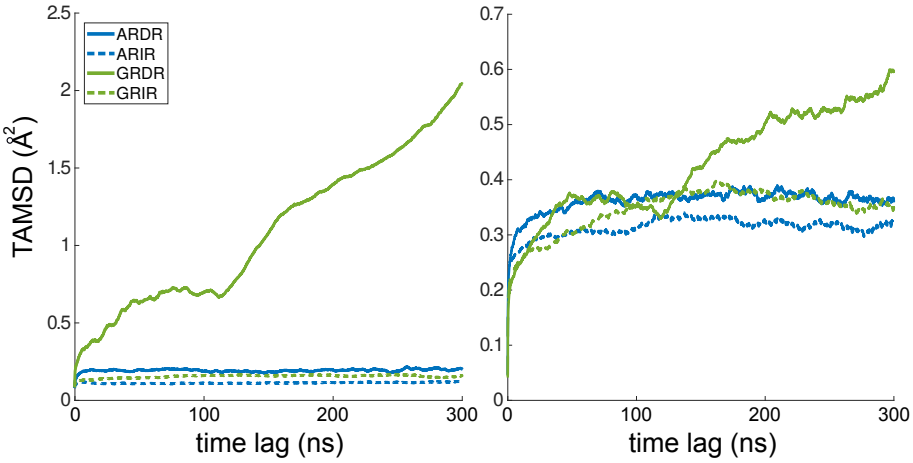


Figure B.3: Time average mean square displacement (TAMSD) of second zinc ions, i.e. Zn2A and ZNsB (left side) and center of mass of loop regions (right side), i.e. residues 474-496 and 576-598 (in the dimerization interface) for all AR/GR models.

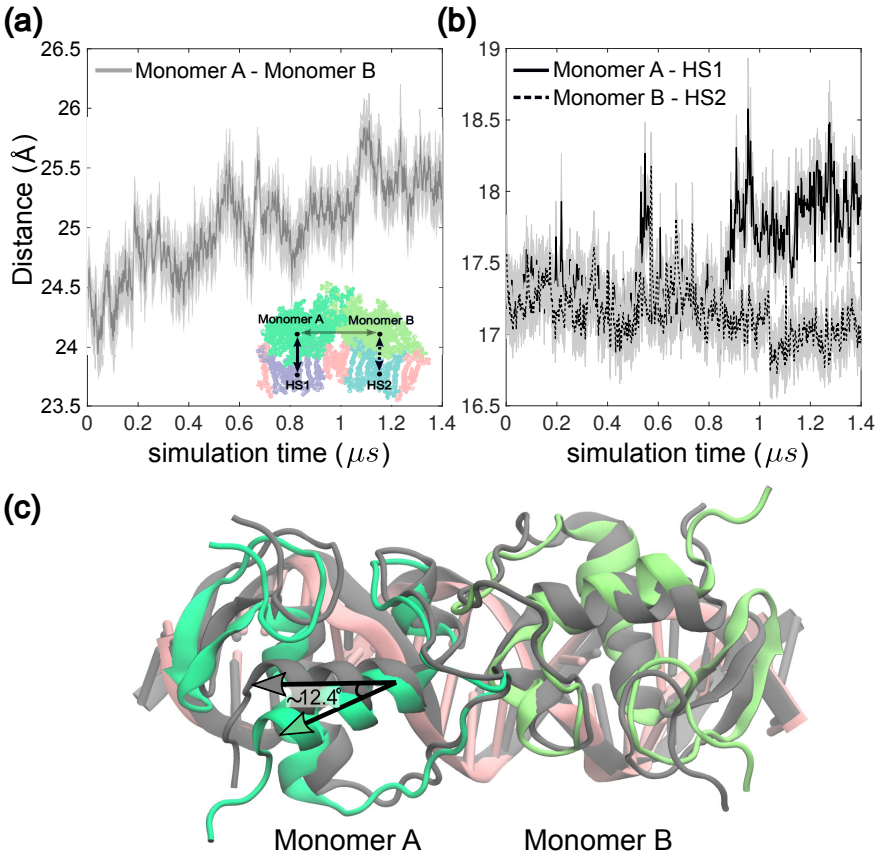


Figure B.4: Extended MD simulation (*run1*) of GR-DR revealing it as unstable due to the separation of the protein monomers. **(a)** Time series of the COM(monomer A)-COM(monomer B) distance. **(b)** Time series of distances COM(monomer A)-COM(HS1) and COM(monomer B)-COM(HS2) colored in red and blue, respectively. **(c)** The median structure of last 50 ns of the 1.4 μs simulation (*run1*) aligned to its first 50 ns median structure (colored and gray, respectively).

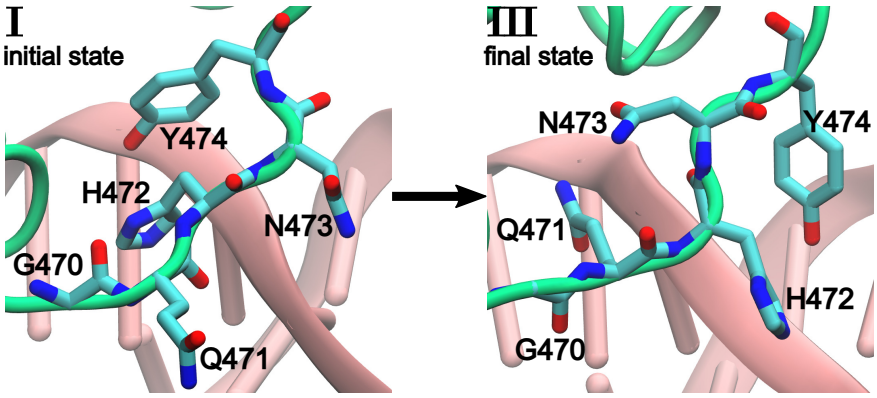


Figure B.5: Initial and final state of the lever in the GR-DR (*run0*).

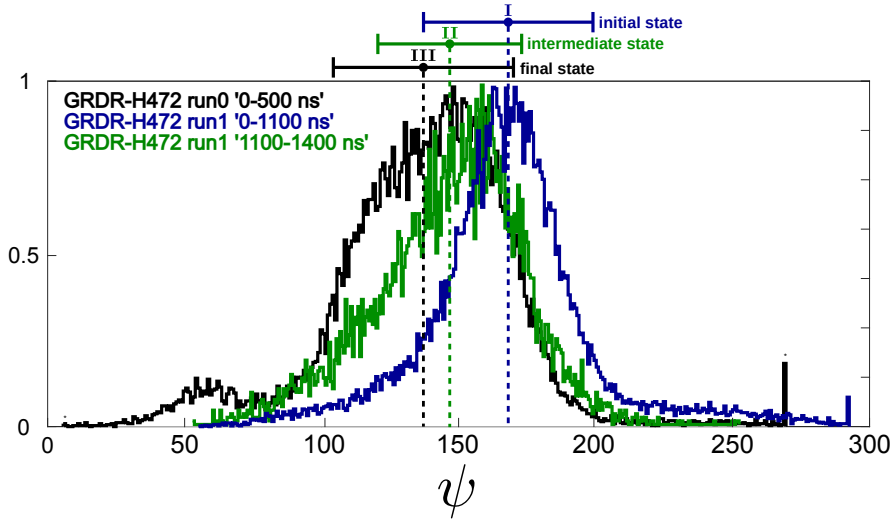


Figure B.6: The ψ backbone angle densities of H472 in both GR-DR simulations (*run0* and *run1*). The conformation of H472 in the last 300 ns of *run1* (green color) approaches the final state of H472 in *run0* (black).

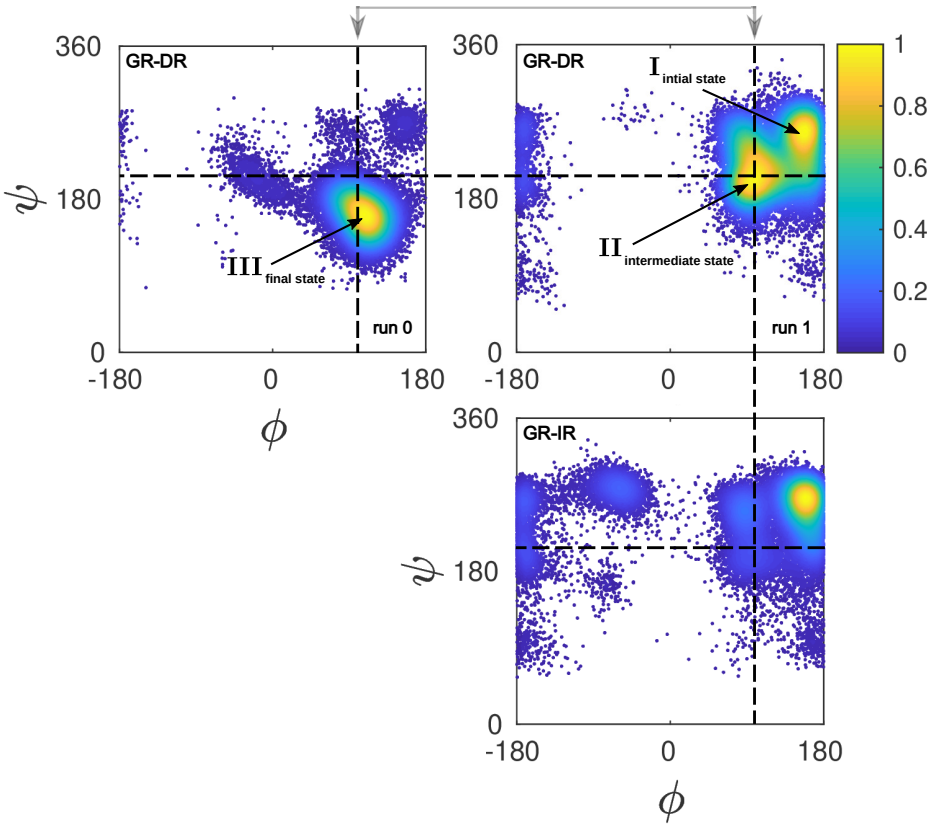


Figure B.7: Comparison of the G470 dihedral angle in the two GR-DR simulations (*run0* & *run1*). An intermediate state of G470 dihedral angles can be observed in the GR-DR (*run1*) simulation.

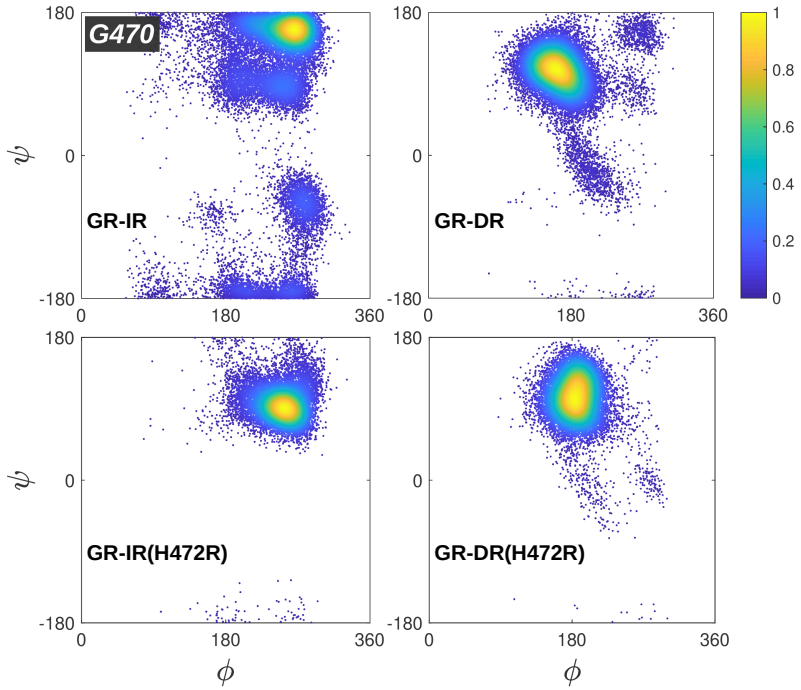


Figure B.8: Comparison of the G470 dihedral angles of the wild type GR-DR(*run0*) and GR-IR with the respective mutant simulations GR(H472R)-DR and GR(H472R)-IR.

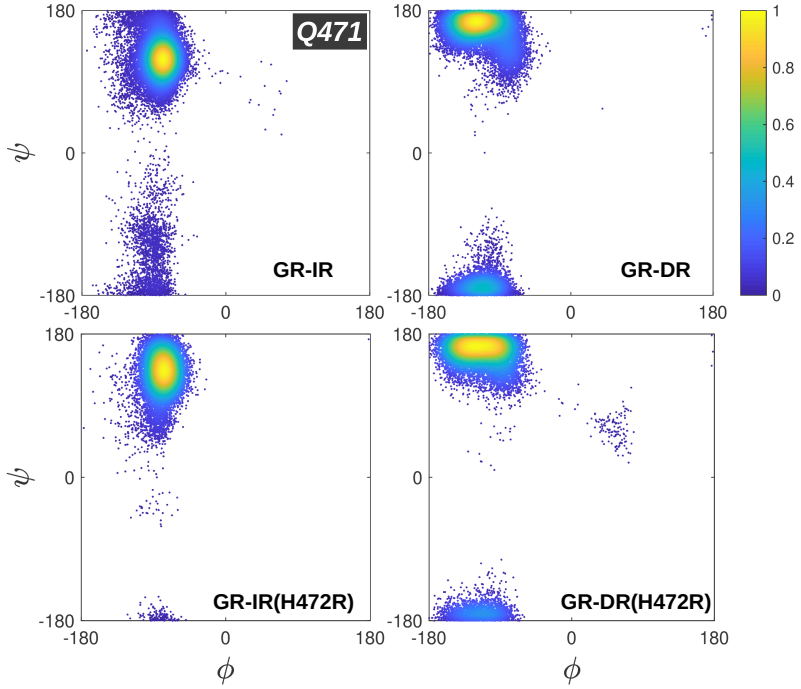


Figure B.9: Comparison of the Q471 dihedral angles of the wild type GR-DR (*run0*) and GR-IR with the respective mutant simulations GR(H472R)-DR and GR(H472R)-IR.

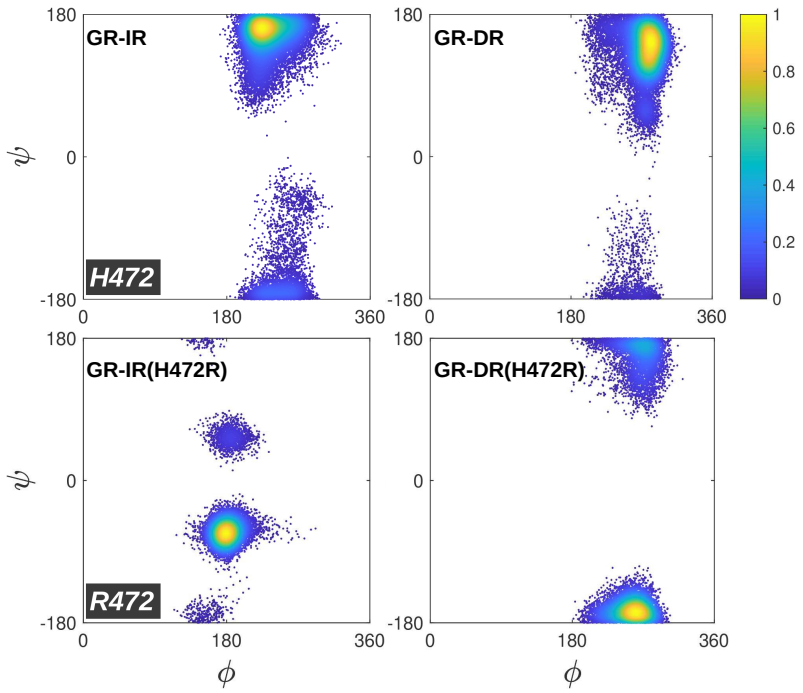


Figure B.10: Comparison of the H472 dihedral angles of the wild type GR-DR (*run0*) and GR-IR with the dihedral angles of residue R472 in the mutant simulations GR(H472R)-DR and GR(H472R)-IR.

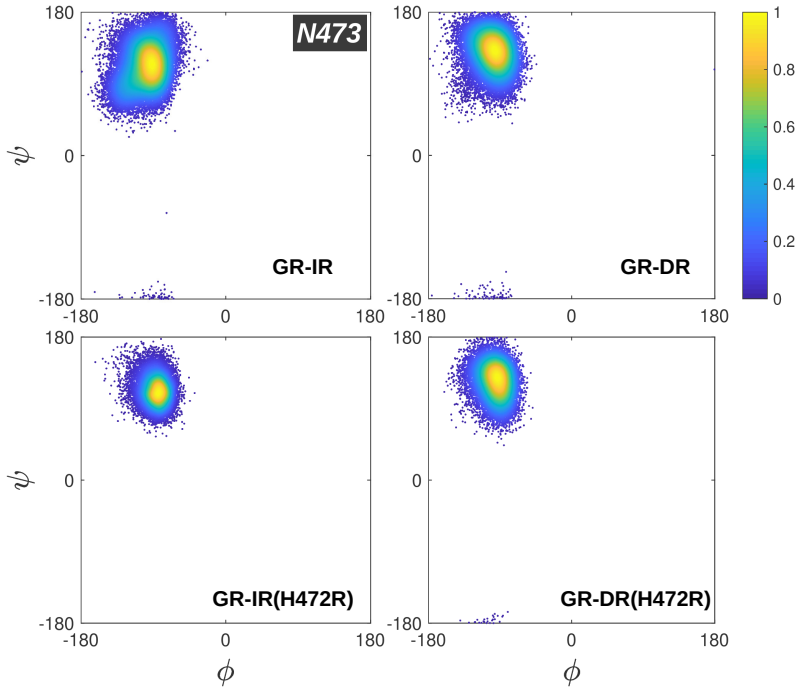


Figure B.11: Comparison of the N473 dihedral angles of the wild type GR-DR (*run0*) and GR-IR with the respective mutant simulations GR(H472R)-DR and GR(H472R)-IR.

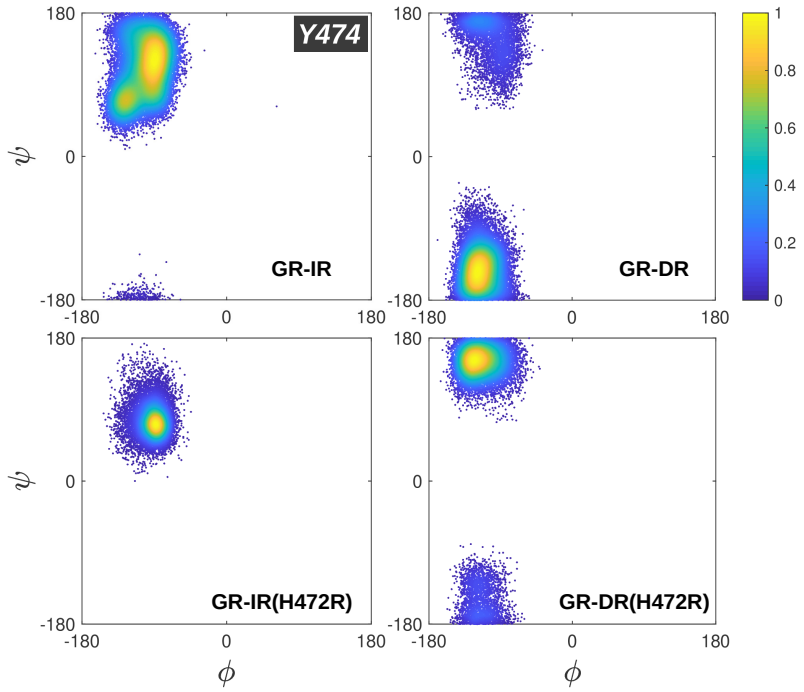


Figure B.12: Comparison of the Y474 dihedral angles of the wild type GR-DR (*run0*) and GR-IR with the respective mutant simulations GR(H472R)-DR and GR(H472R)-IR.

Protein-Protein interactions

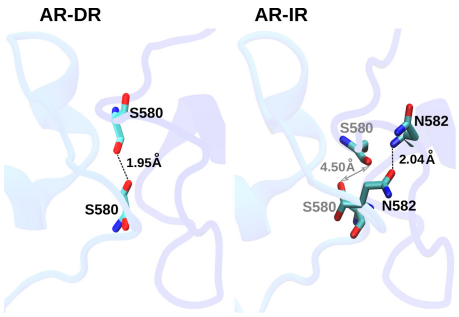


Figure B.13: Hydrogen bond interaction of the AR-DR/IR dimer interface. The interaction of the S580 with its counterpart (in AR-DR) is broken in the AR-IR system and a new hydrogen bond N582-N582 is formed.

Appendix C

Supplementary Information for: Molecular Dynamics Simulations of a Chimeric Androgen Receptor Protein (SPARKI) conrm the Importance of the Dimerization Domain on DNA Binding Specificity

The DNA geometries of the SPARKI systems exhibit different values of parameters in the three time intervals: the first (W1), middle (W2), and last 100 ns (W3) windows of the trajectories. The chimeric SPARKI systems are based on AR-DR and GR-IR systems and are thus anticipated to show a longer relaxation time in the simulations. However, for IR complexes, the results of intervals W2 and W3 show almost the same values in DNA parameters (Figure 5.6 and SI Figures C.3,C.4(b,d)) whereas for DRs, the values in the W3 differ from those observed in the earlier W1 and W2 intervals (Figure 5.6 and SI Figures C.3,C.4(a,c)). An exception is the SpGR-IR model for which we observe in the W3 interval bending parameters different from those in intervals W1 and W2, see SI Figure C.4(d).

Comparison of the DNA parameters of SpGR-DR in the W3 interval with those in the W1 interval shows a considerable change in DNA geometry in the course of the simulation. As our results indicate, the DNA of both SPARKI-DR complexes also has a geometry that is different from the DNA in the AR-DR system. However, these differences are considerably larger between SpGR-DR and AR-DR than between SpAR-DR and AR-DR.

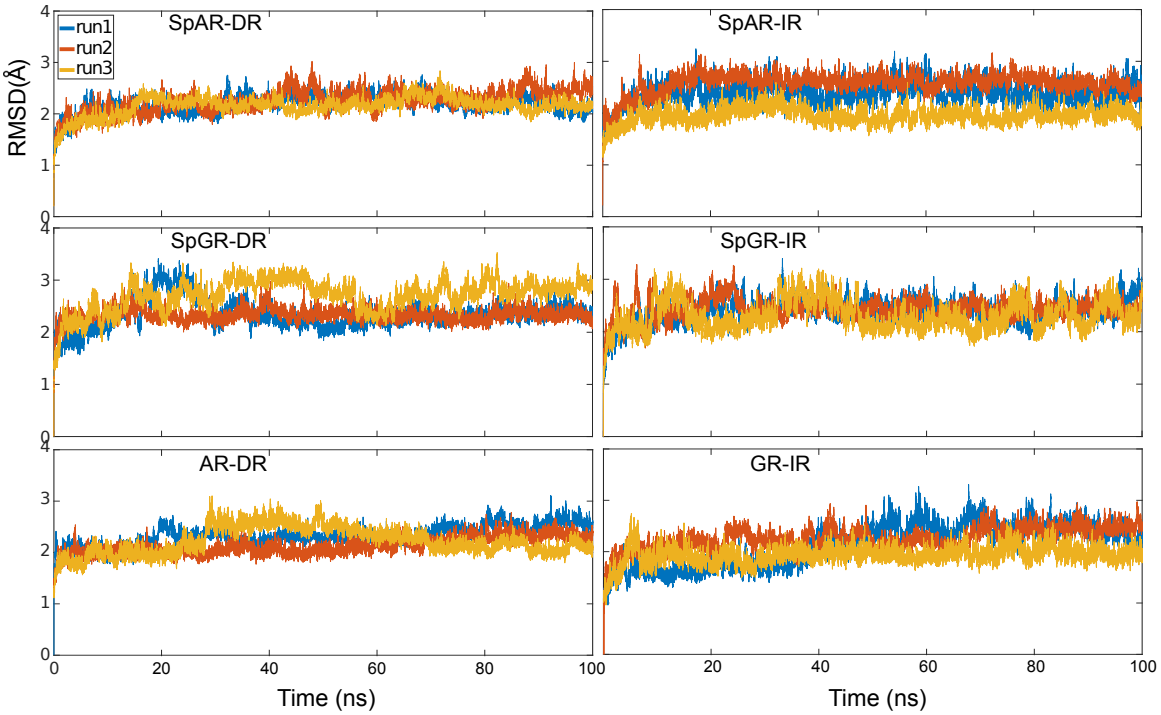


Figure C.1: Root mean square displacement of three 100 ns replicas MD simulations, for each system. For SpAR-DR, SpGR-DR, SpGR-IR, and AR-DR, run 1 is selected for carrying out the longer simulations and for SpAR-IR and GR-IR, run3 is selected for carrying out the longer simulations.

Table C.1: Average distance between different domain/subdomains (as characterized by the distances between the respective centers of mass) of the protein-DNA complexes.

Distance (Å)	**AR-DR**		**SpAR-DR**		**SpAR-IR**	
	mean	std	mean	std	mean	std
Monomer A - HS1	16.89	0.19	17.04	0.21	16.68	0.20
Monomer B - HS1	16.62	0.21	17.08	0.19	16.75	0.18
Monomer A - Monomer B	24.37	0.31	24.08	0.24	23.68	0.25
Dimer interface	9.70	0.26	9.97	0.20	9.69	0.22
ZN_A-ZN_B (Dim)	9.06	0.30	9.35	0.27	9.19	0.30
Distance (Å)	**GR-IR**		**SpGR-DR**		**SpGR-IR**	
	mean	std	mean	std	mean	std
Monomer A - HS1	16.42	0.19	16.78	0.23	16.68	0.20
Monomer B - HS1	16.27	0.17	16.97	0.26	16.75	0.18
Monomer A - Monomer B	25.08	0.20	24.45	0.39	24.04	0.32
Dimer interface	9.88	0.18	10.47	0.33	9.87	0.43
ZN_A-ZN_B (Dim)	9.08	0.24	10.20	0.50	9.33	0.36

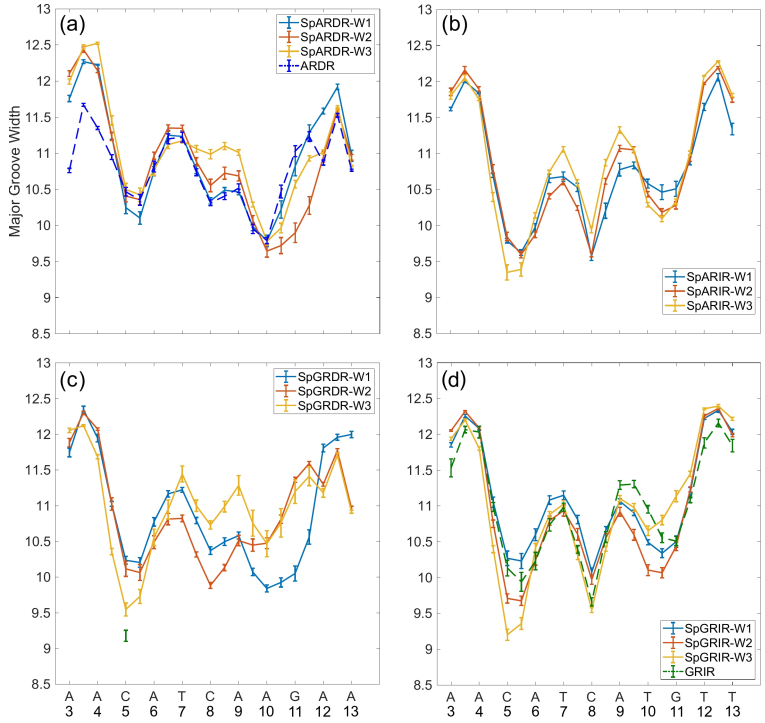


Figure C.2: Comparison of the DNA major groove for (a) SpAR-DR and AR-DR (dashed line colored in blue) (b) SpAR-IR (c) SpGR-DR (d) SpGR-IR and GR-IR (dashed line colored in green). For SPARKI systems, the lines colored in light blue, red, and orange correspond to the major groove analysis for the first 100 ns (W1), middle 100 ns (W2), and last 100 ns (W3) intervals of 900 ns MD simulations, respectively. For AR-DR and GR-IR, the results are calculated for the last 100 ns of the 500 ns MD simulations.

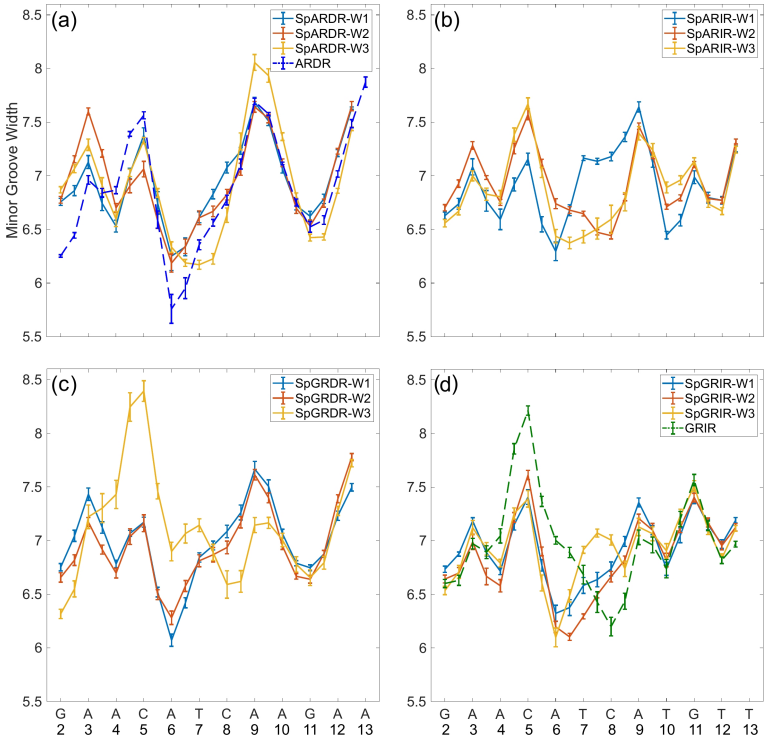


Figure C.3: Comparison of the DNA minor groove for (a) SpAR-DR and AR-DR (dashed line colored in blue) (b) SpAR-IR (c) SpGR-DR (d) SpGR-IR and GR-IR (dashed line colored in green). For the SPARKI systems, the lines colored in light blue, red, and orange correspond to the minor groove analysis for the first 100 ns (W1), middle 100 ns (W2), and last 100 ns (W3) intervals of 900 ns MD simulations, respectively. For AR-DR and GR-IR the results are calculated for last 100 ns of 500 ns MD simulations.

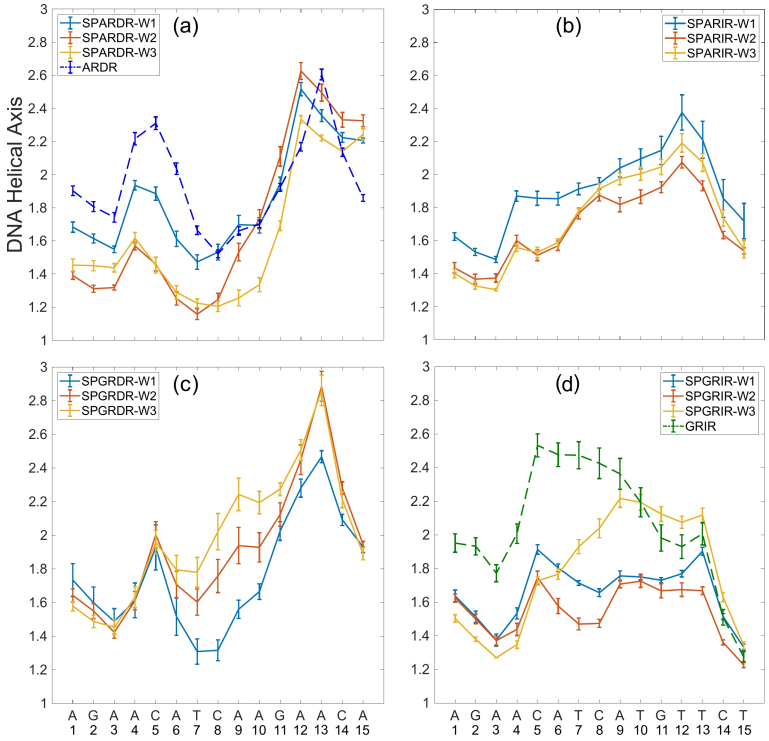


Figure C.4: Comparison of the DNA helical axis bending for (a) SpAR-DR and AR-DR (dashed line colored in blue) (b) SpAR-IR (c) SpGR-DR (d) SpGR-IR and GR-IR (dashed line colored in green). For SPARKI systems, the lines colored in light blue, red, and orange corresponds to DNA helical axis bending analysis for the first 100 ns (W1), middle 100 ns (W2), and last 100 ns (W3) intervals of 900 ns MD simulations, respectively. For AR-DR and GR-IR the results are calculated for last 100 ns of 500 ns MD simulations.

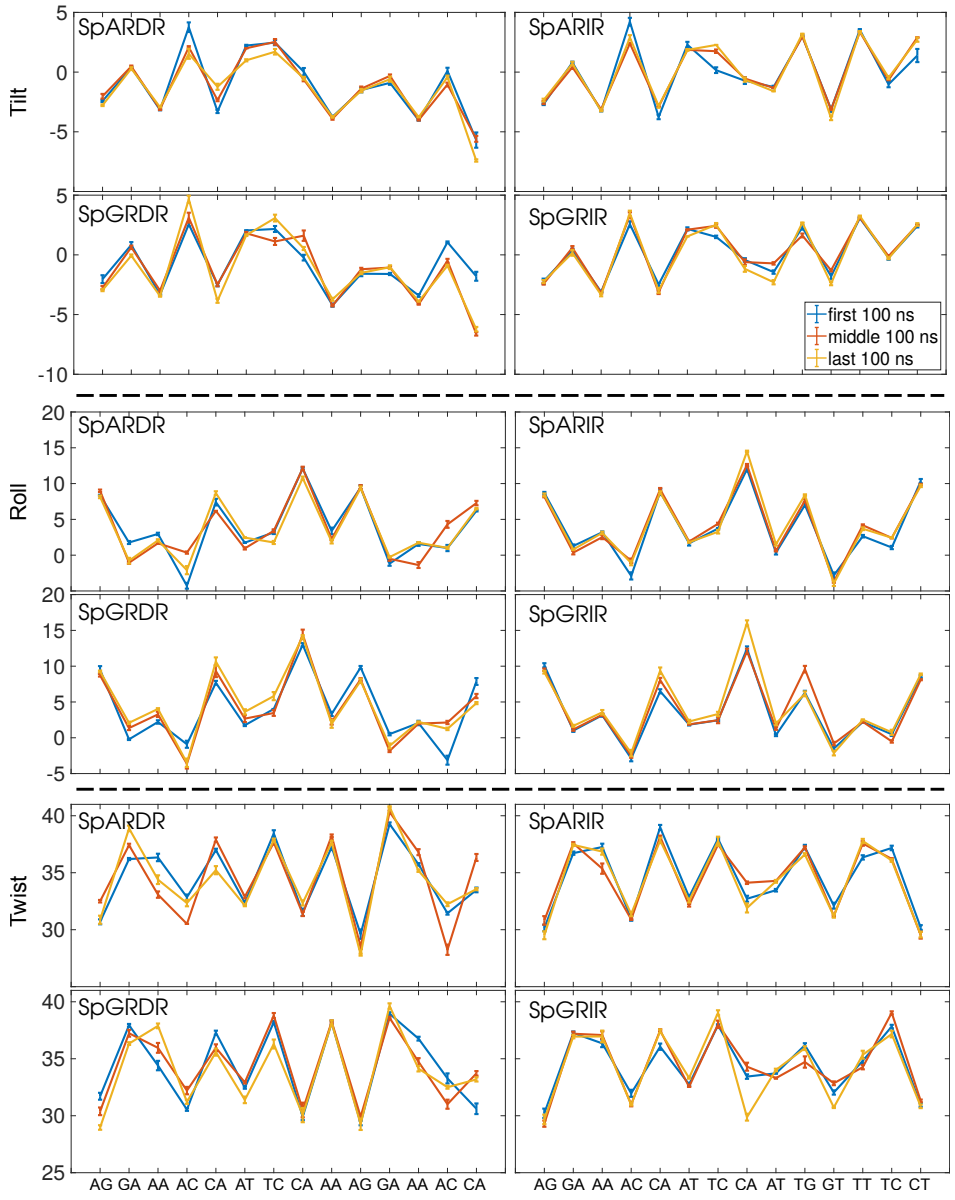


Figure C.5: DNA rotational inter base pair parameters for SPARKI models.

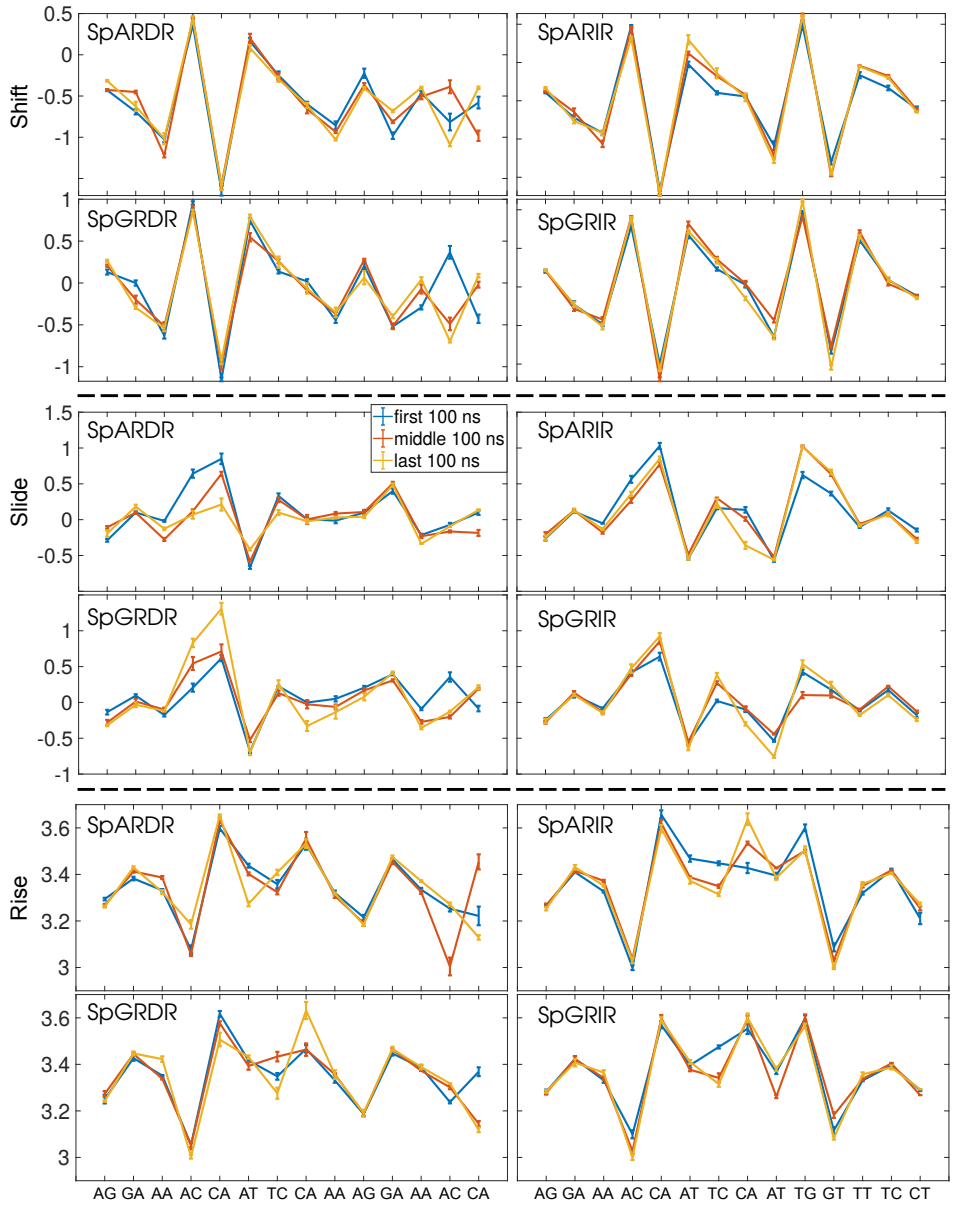


Figure C.6: DNA translational inter base pair parameters for SPARKI models.

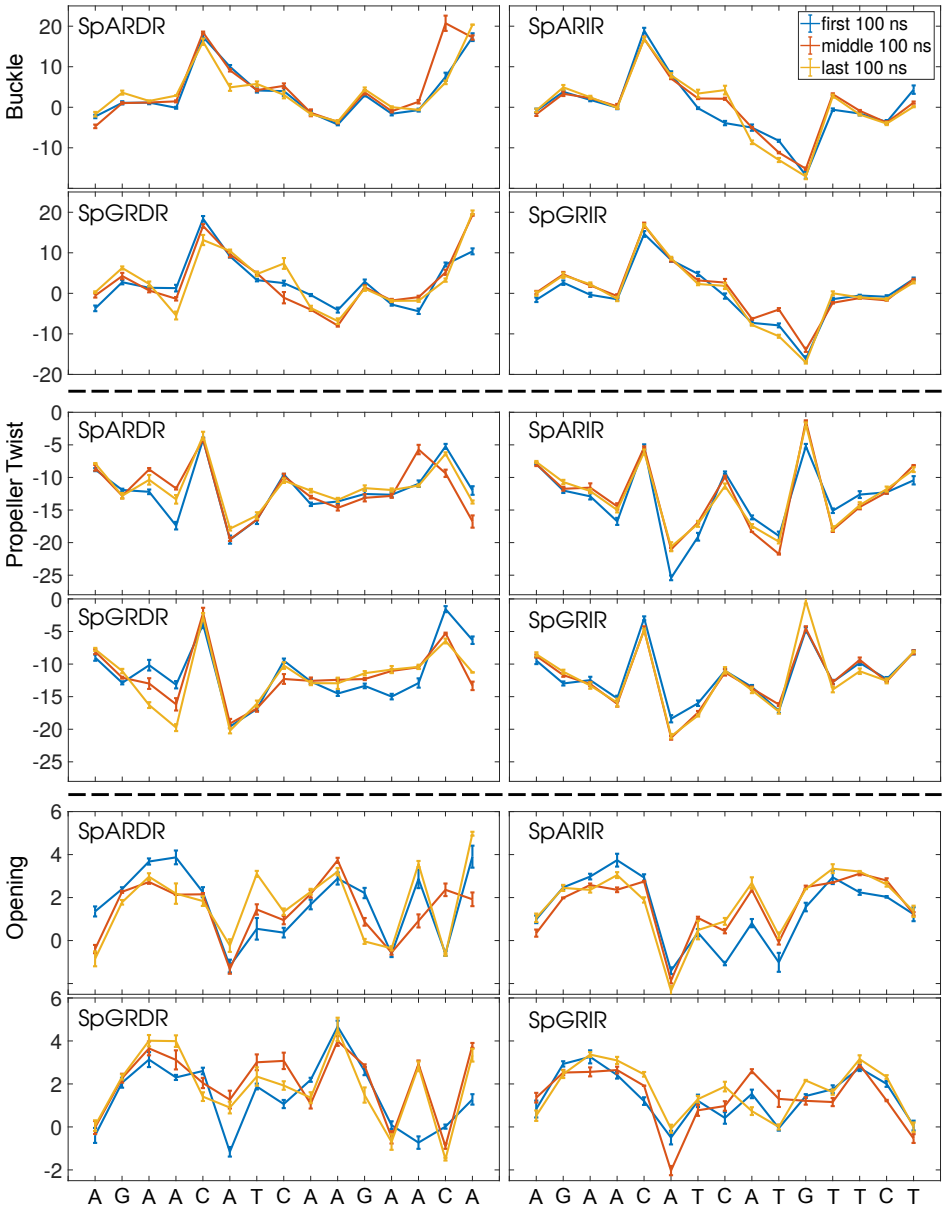


Figure C.7: DNA rotational intra base pair parameter fors SPARKI DNA.

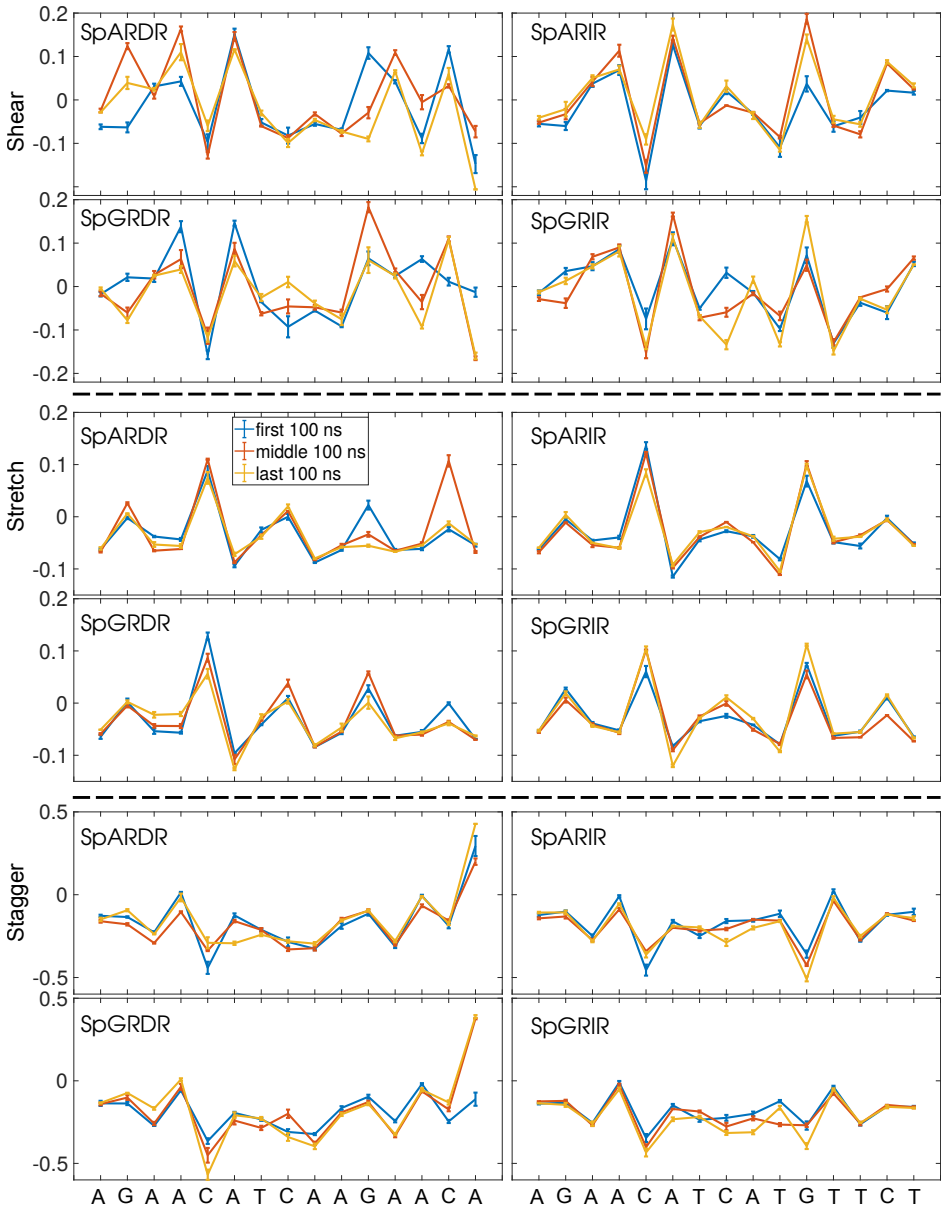


Figure C.8: DNA translational intra base pair parameters for SPARKI models.

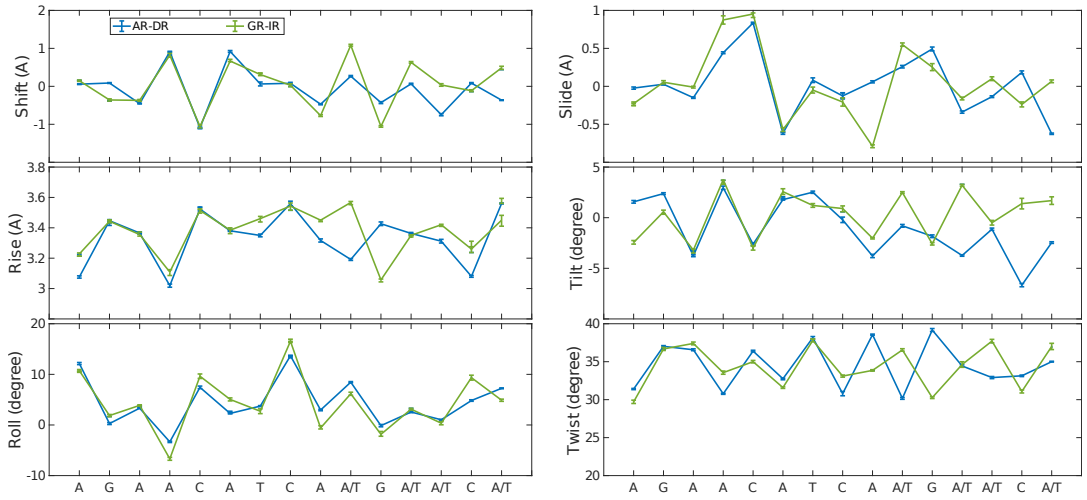


Figure C.9: DNA translational and rotational inter base pair parameter fors AR-DR and GR-IR models.

For SpAR-DR both groove widths in the W3 interval show only small differences with respect to the W1 interval, although the DNA considerably loses its bending at the end of simulation, i.e. in the W3 interval (see SI Figures C.2,C.3, and C.4(a)). In the SpGR-DR complex, the DNA shape is significantly changed from the W1 and W2 intervals to the W3 interval. As our results show, the DNA of both Sp(AR/GR)-DR complexes, also show different conformations with respect to the DNA of the AR-DR system. For SpAR-DR, the most significant change happens in the DNA bending parameters, as can be seen from the comparison with AR-DR, SI Figure C.4(a).

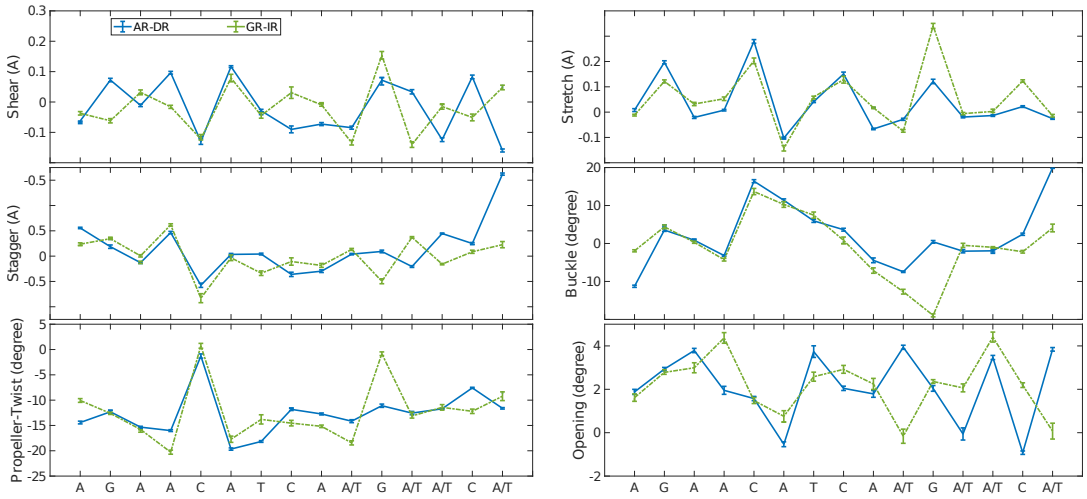


Figure C.10: DNA translational and rotational intra base pair parameters for AR-DR and GR-IR models.

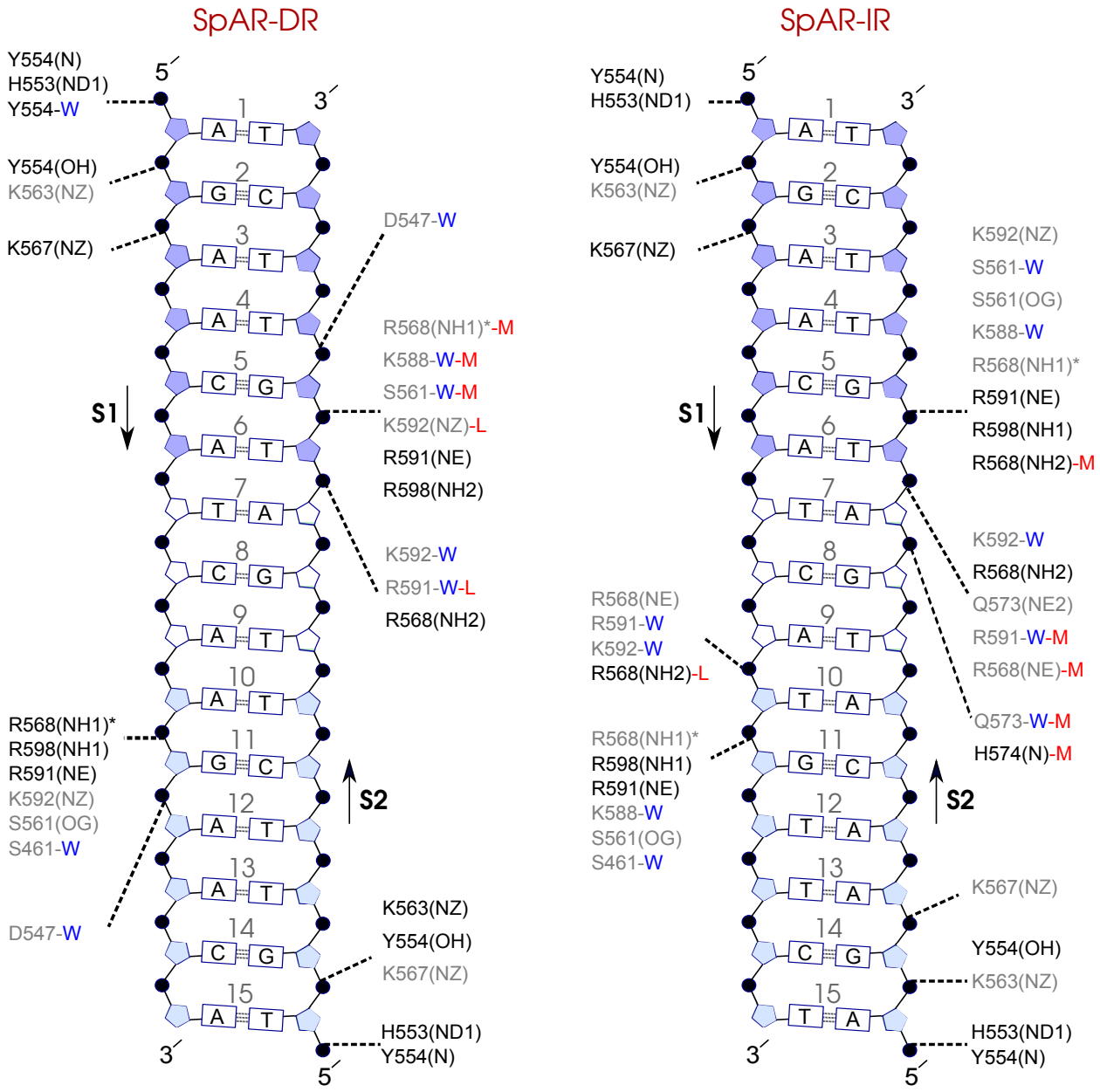


Figure C.11: Diagram of protein-DNA hydrogen-bond interactions for (*left*) SpAR-DR and (*right*) SpAR-IR. The nucleotides of the 15 bps core DNA sequence are numbered from HS1 (numbers: 1 to 6) to HS2 (numbers: 10 to 15). The spacer region is highlighted with non-colored boxes around the numbers of the bases (numbers: 7-9). The hydrogen bonds are categorized based on their occupancy, 50-75% (gray), and 75-100% (black). The water mediated hydrogen bonds are labeled with a blue letter "W". The residues shown with star sign form base-specific hydrogen-bond interactions while the other residues form interactions with the backbone of DNA. The letters "M" and "L" colored in red show the hydrogen-bond interactions only seen in "W2 (middle 100 ns)" and "W3 (last 100 ns)" time interval of trajectories, respectively. All other interactions are observed in both "W2" and "W3" intervals.

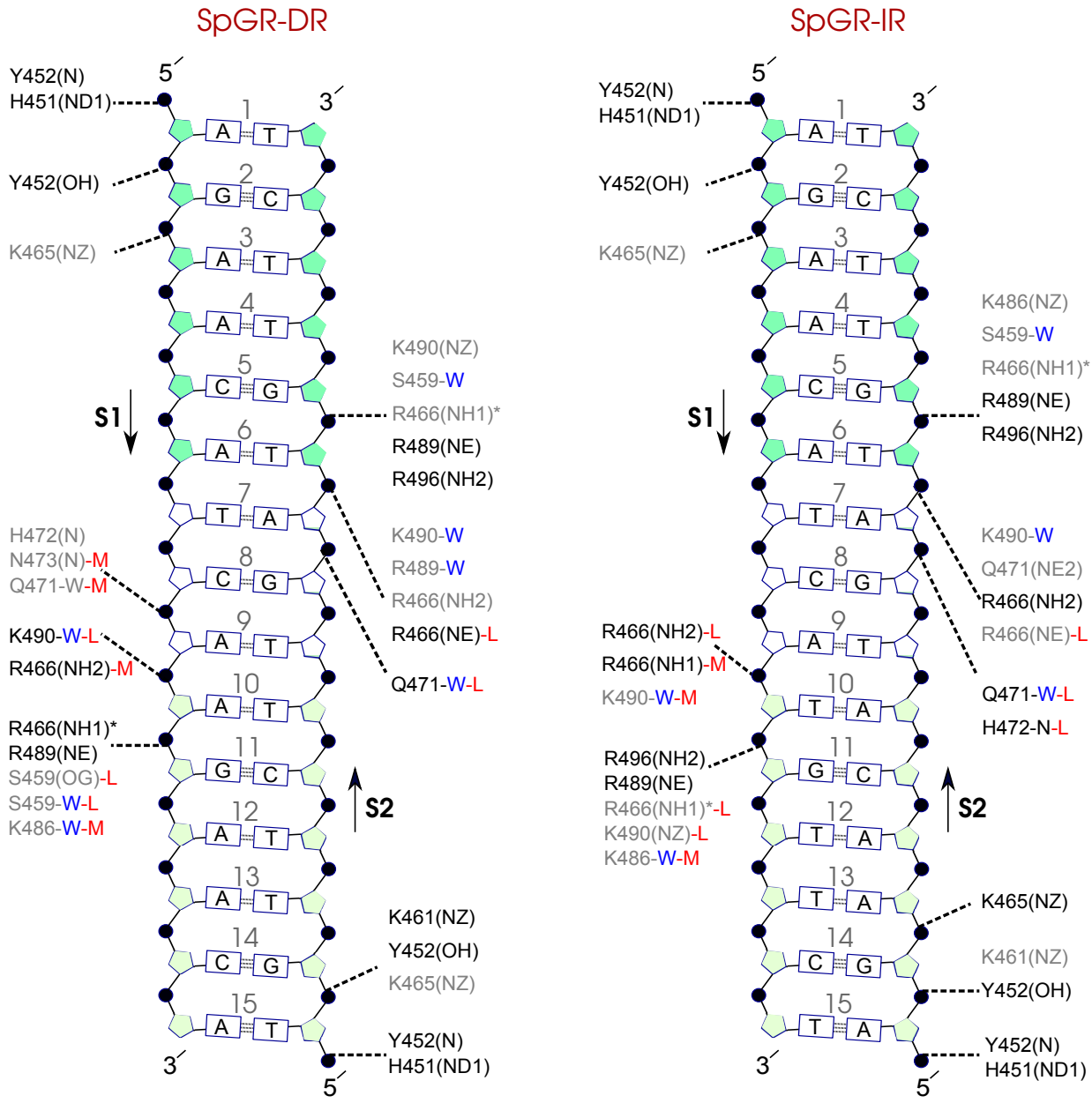


Figure C.12: Diagram of protein-DNA hydrogen-bond interactions for (*left*) SpGR-DR and (*right*) SpGR-IR. The nucleotides of the 15 bps core DNA sequence are numbered from HS1 (numbers: 1 to 6) to HS2 (numbers: 10 to 15). The spacer region is highlighted with non-colored boxes around the numbers of the bases (numbers: 7-9). The hydrogen bonds are categorized based on their occupancy, 50-75% (gray), and 75-100% (black). The water mediated hydrogen bonds are shown with a blue letter "W". The residues shown with star sign form base-specific hydrogen-bond interactions while the other residues make interactions with the backbone of DNA. The letters "M" and "L" colored in red show the hydrogen-bond interactions only seen in "W2 (middle 100 ns)" and "W3 (last 100 ns)" time interval of trajectories, respectively. All other interactions are observed in both "W2" and "W3" intervals.

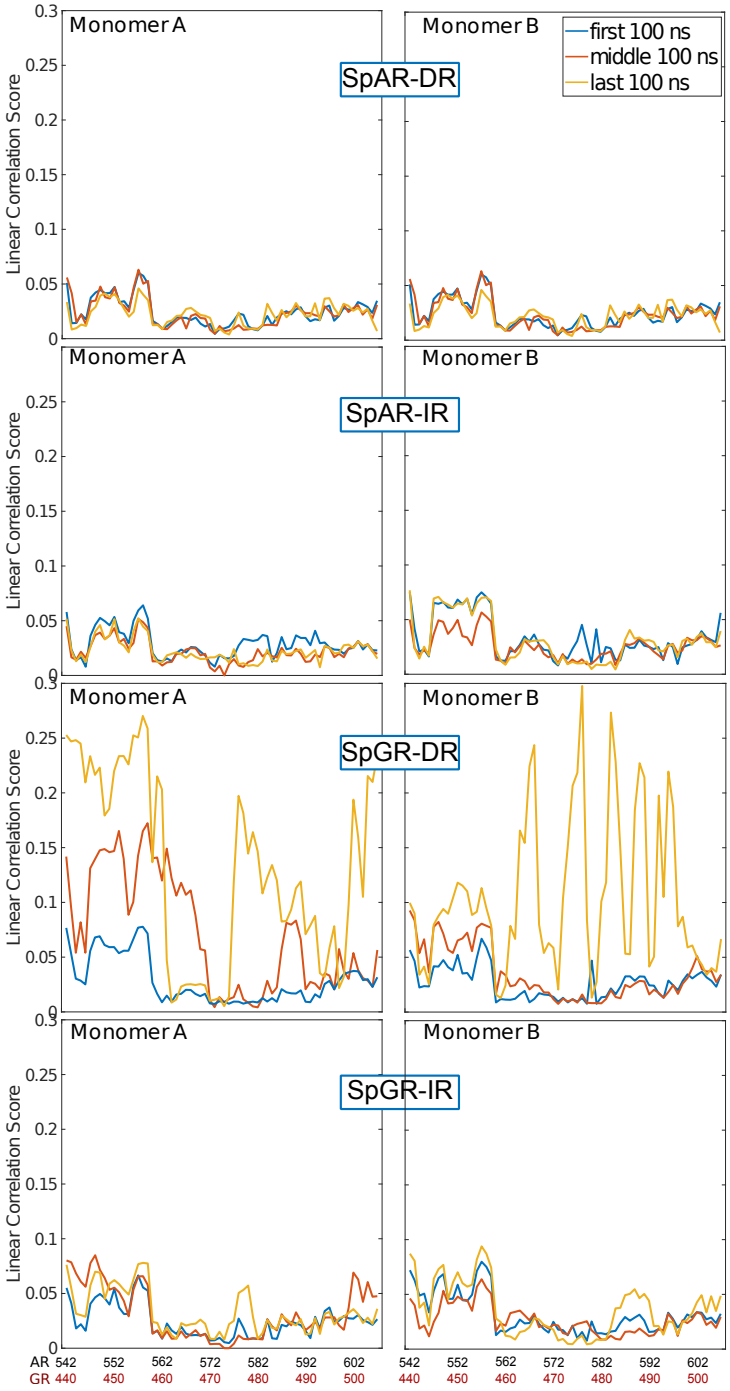


Figure C.13: Correlation score per residue, computed for intra-domain correlations with $r_{ki} \geq 0.4$ for SPARKI system, calculated for first 100 ns ("W1"), middle 100 ns ("W2"), and last 100 ns ("W3") time intervals of 900 ns SPARKI's MD trajectories.

Summary

In this dissertation, we investigated the significance of long-range communication in macromolecules' conformational dynamics by means of molecular simulations. At the molecular level, these cooperative long-range interactions are delineated to be important in the regulation of biological systems [5, 11, 12]. The primary goal was to determine how the transmission of these long-range effects is manifested in conformational stability and characteristic dynamic properties of the studied macromolecular systems, which, in the light of this thesis, are: (1) Cytochrome c oxidase (CcO), a membrane protein in the respiratory chain of mitochondria, (2) Androgen (AR) and Glucocorticoid (GR) receptors DNA binding domain-DNA complexes. To obtain a noteworthy insight about the interactions underlying such intra and inter molecular communication mechanisms, molecular dynamics (MD) simulation of studied systems have been performed and subsequently analyzed.

CcO system

In CcO the transfer of protons towards the binuclear redox center (BNC), for the reduction reaction of dioxygen to water, or towards a proton loading site, for transferring protons through the membrane, occur within two distinct pathways, the D- and K-channels. Yet, despite many experimental and computational studies on CcO [33, 83, 86, 88, 90, 141, 142], the mechanism underlying such proton transfer and protonation-state dependent action of key residues inside channels and their possible inter/intra communication is not yet clearly understood. To this, in our study, MD simulations of CcO (in the first step of the redox cycle, i.e. $P_R \rightarrow$F state) with 16 different combinations of protonation state of the key residues, i.e. D132, E286 (D-channel), E101 and K362 (K-channel), have been performed and analyzed. This allows us to mimic the proton transfer through the channels. The result of our simulations indicate that the conformational dynamics and communication of the channel residues in CcO depend on a combination of their protonation state rather than on individual residue's protonation state. This cooperative behavior is especially seen among the residues within the same channel. In particular, within the D-channel, a substantial conformational dependency of residue N139 on protonation states of the terminal residues, i.e. D132 and E286, can be observed; suggesting a gating role of N139. On the other side, a significant coupling of the hydrogen-bond dynamics of N139 with respect to the protonation state of the residue K362 in K-channel has been observed [34]. Together, the protonation-state dependent communication between these distal-site pairs might explain the regulation of proton release from the D-channel in the $P_R \rightarrow$F transition.

Protein-DNA system

Site-specific binding of short DNA sequences by transcription factor proteins facilitates gene regulation in the cell. AR and GRs are ligand-activated transcription factors which bind as homo-dimers to DNA with a repeated recognition sequence, spaced by a few base pairs. Despite the structural similarity of the two proteins' DNA binding domains, the AR binds to DNA with a direct repeat, whereas the GR fails to do so. Our simulations suggest that long-ranged communication in the protein-protein-DNA complexes plays an important role in recognition and specificity. We could reveal an altered shape of the DNA spacer in the direct repeat sequence. A conformational change in one loop, called lever arm, of the GR's DNA binding domain leads to an erroneous binding to the DNA spacer. The resulting tilt of one protein subunit into a distorted conformation impairs the protein-protein interactions and thus the stability of the complex. The different composition of the respective loop in the androgen receptor prevents such distorting interactions to the DNA spacer, maintaining the conformation of the protein-DNA complex. Both proteins show intact complexes when bound to an inverted repeat DNA sequence.

A chimeric AR protein, termed SPARKI, in which the second zinc-binding motif of AR (dimer interface) is swapped with that of GR fails to recognize direct repeat-like elements. This is unlike AR that makes specific contacts with these sequences but rather like GR that fails to make such contacts [47]. Therefore, the question arises whether the dimer interface is a main factor in the distinct recognition of AR and GR toward a direct repeat sequence. Thus, studying the SPARKI receptor allowed us to better understand the role of the AR-like and GR-like domains in the distinct recognition mechanisms of AR and GR. The results of our simulations show that the competition between a strong (or flexible) dimer interface versus distinct direct protein-DNA is an essential factor for the dimeric protein to allow proper accommodation on DNA. However, the stability of the dimerization interface plays a predominant role as supported by the fact that all complexes exhibit rather similar specific contacts with DNA. A SPARKI model built from the structure of the GR, i.e. SPARKI-GR, shows a considerable distortion in its dimerization domain when complexed to a direct repeat sequence. On the other hand, a SPARKI model based on the AR, i.e. SPARKI-AR, shows significantly fewer protein-protein and protein-DNA hydrogen bond interactions when complexed with direct repeat sequence than with inverted repeat. The diminished interaction of SPARKI-AR with and the instability of SPARKI-GR on direct repeat response elements agree with SPARKI's lack of affinity for these sequences. The more GR-like binding specificity of the chimeric SPARKI protein is further emphasized by both SPARKI models binding even more strongly to inverted repeat elements than observed for the DNA binding domain of the GR.

GR activity modulation depends on DNA sequence of its binding site. Different core specific-sequence composition and the DNA shape, 'read-out' through non-specific DNA contacts, was proved to be key factors in modulating the GR activity. In this regard, a sequence flanking to the core specific sequence might play an important role in GR structure and its interaction with DNA. Our simulations suggest significant long-range communication between the specific protein-DNA core site and its proximal flanking base pairs. In this regard, we could reveal an altered shape of the core DNA sequence due to different flanking elements. Nevertheless, the impact of DNA-shape variation results not in direct protein-DNA interaction (via hydrogen-bonds) but rather repositioning of the dimer GR DNA-binding domains. Interestingly, such altered protein-DNA conformation has been mainly observed in the complex's second half-site, pointing out the predominant role of direction flanking nucleotide in the GR structure and conformation.

Our MD simulations and their outcome results afforded us to understand the detailed interactions at atomic-detail that are governing long-range effects.

Zusammenfassunga

In dieser Doktorarbeit haben wir mit Hilfe von Molekülsimulationen die Bedeutung von langreichweitiger Kommunikation in der Konformationsdynamik von Makromolekülen untersucht. Kooperative langreichweitge Wechselwirkungen sind auf der molekularen Ebene wichtig für die Regulation biologischer Systeme [5, 11, 12]. Das primäre Ziel war es, herauszufinden, wie sich die Transmission dieser Fernwirkungen in der Konformationsstabilität und den charakteristischen dynamischen Eigenschaften der im Rahmen dieser Arbeit untersuchten makromolekularen Systeme niederschlägt. Bei den Systemen handelt sich zum einem um die (1) Cytochrom-c-Oxidase (CcO), ein Membranprotein der mitochondrialen Atmungskette, und zum anderen um die (2) DNA-Bindedomänen der Androgen- (AR) und Glukokortikoid- (GR) Rezeptoren im Komplex mit DNA. Um einen aussagekräftigen Einblick in die Wechselwirkungen zu erhalten, die solchen intra- und intermolekularen Kommunikationsmechanismen zugrunde liegen, wurden Molekulardynamik- (MD) Simulationen der beschriebenen Systeme durchgeführt und anschließend analysiert.

CcO system

Der Protonentransfer innerhalb der CcO erfolgt auf zwei verschiedenen Wegen, dem D- und dem K-Kanal. Die Protonen gelangen so zum binuklearen Redoxzentrum für die Reduktionsreaktion von Dioxygen zu Wasser oder in Richtung einer Protonenladestelle zum Transfer durch die Membran. Trotz einer Vielzahl an experimentellen und computerbasierten Studien zur CcO [33, 83, 86, 88, 90, 141, 142], ist der dem Protonentransfer zugrundeliegende Mechnismus und das Verhalten von Schlüsselresiduen in Abhängigkeit ihres Protonierungszustands sowie deren mögliche Inter- und Intrakommunikation nicht ausreichend verstanden. Dazu wurden in unserer Studie MD-Simulationen der CcO (im ersten Schritt des Redoxzyklus, also im Übergang $P_R \rightarrow F$ state) in 16 verschiedenen Kombinationen von Protonierungszuständen der Schlüsselresiduen durchgeführt und analysiert. Bei den Schlüsselresiduen handelt es sich um D132 und E286 im D-Kanal sowie E101 und K362 im K-Kanal. Die Simulationen ermöglichen es uns, den Protonentransfer durch die Kanäle nachzubilden. Die Ergebnisse unserer Simulationen zeigen, dass die Konformationsdynamik und Kommunikation der Kanalresiduen von einer Kombination an Protonierungszuständen der Schlüsselresiduen abhängen anstatt von deren individuellen Protonierungszuständen. Dieses kooperative Verhalten zeigt sich vor allem bei den Residuen innerhalb desselben Kanals. Insbesonere innerhalb des D-Kanals hängt die Konformation von N139 von den Protonierungszuständen der Anfangs- und Endresiduen, also D132 und E286, ab; das deutet auf eine Gating-Rolle von N139 hin. Außerdem konnte eine signifikante Kopplung der Wasserstoffbrückendynamik von N139 mit dem Protonierungszustand von K362 im K-Kanal beobachtet werden [34]. Zusammengenommen könnte die protonierungszustandsabhängige Kommunikation zwischen diesen voneinander entfernten Stellen die Regulation der Protonenfreisetzung aus dem D-Kanal im $P_R \rightarrow F$ Übergang erklären.

Protein-DNA system

Die ortsspezifische Bindung von kurzen DNA-Sequenzen durch Transkriptionsfaktoren ermöglicht die Genregulation in der Zelle. ARs und GRs sind ligandenaktivierte Transkriptionsfaktoren, die als Homodimere an DNA mit einer sich nach wenigen Basenpaaren (Spacer) wiederholende Erkennungssequenz binden. Trotz der strukturellen Ähnlichkeit der DNA-Bindedomänen beider Proteine kann AR im Gegensatz zu GR an DNA mit einem direct Repeat binden. Unsere Simualtionen deuten darauf hin, dass langreichweitige Kommunikation in den Protein-Protein-DNA-Komplexen eine wichtige Rolle für Erkennung und Spezifität spielt. Wenn GR an eine direct-Repeat-Sequenz bindet, so kann eine Verformung der Spacer-DNA beobachtet werden. Eine Konformationsänderung in einer Loop der DNA-Bindedomäne, dem sogenannten Lever-arm, des GR führt zu einer fehlerhaften Bindung an die Spacer-DNA. Die daraus resultierende Verkippung eines der Monomere in eine verzerrte Konformation beeinträchtigt die Protein-Protein-Wechselwirkungen und damit die Stabilität des gesamten Komplexes. Die davon abweichende Zusammensetzung der entsprechenden Loop im Androgenrezeptor verhindert solcherlei gestörte Wechselwirkungen mit der Spacer-DNA und bewahrt so die Konformation des Protein-DNA-Komplexes. Beide Proteine bilden intakte Komplexe, wenn sie an inverted-Repeat-DNA-Sequenzen gebunden sind.

SPARKI, ein chimäres AR-Protein, bei dem das zweite Zinkfinger-Motiv (Dimer-Interface) des AR durch das des GR ersetzt ist, erkennt keine direct-Repeat-DNA-Sequenzen. SPARKI verhält sich also wie GR, das im Gegensatz zu AR keine spezifischen Kontakte mit diesen Sequenzen herstellt [47]. Somit stellt sich die Frage, ob das Dimer-Interface einen Hauptfaktor in der eindeutigen Erkennung des direct Repeat durch AR und GR darstellt. Der untersuchte SPARKI-Rezeptor hat es uns erlaubt, die Rollen der AR- und GR-artigen Domänen in den Erkennungsmechanismen von AR und GR zu verifizieren. Die Ergebnisse unserer Simulationen zeigen, dass die Konkurrenz zwischen der Flexibilität des Dimer-Interfaces und den direkten Protein-DNA-Wechselwirkungen einen entscheidenen Faktor für die passgenaue Unterbringung des Dimers auf der DNA darstellt. Die herausragende Rolle des Dimer-Interfaces wird zusätzlich dadurch unterstützt, dass alle Komplexe ähnliche spezifische Kontakte mit der DNA aufweisen. Ein SPARKI-Modell, das von einer GR-Kristallstruktur ausgehend gebaut wurde (SPARKI-GR), zeigt eine erhebliche Verformung in seiner Dimerisationsdomäne, wenn es an einen direct Repeat gebunden ist. Ein SPARKI-Modell, das hingegen auf einer AR-Kristallstruktur basiert (SPARKI-AR), zeigt deutlich weniger Protein-Protein- und Protein-DNA-Wasserstoffbrücken im Komplex mit einem direct Repeat im Vergleich zu einem inverted Repeat. Die verminderte Wechselwirkung von SPARKI-AR und die Instabilität von SPARKI-GR bestätigen die mangelnde Affinität von SPARKI zu direct-Repeat-Sequenzen. Die GR-ähnliche Bindungsspezifität des chimären SPARKI-Proteins wird noch dadurch unterstrichen, dass beide SPARKI-Modelle noch stärker an inverted Repeat-Elemente binden als für die DNA-Bindedomäne des GR beobachtet.

Die GR-Aktivitätsmodulation hängt von der DNA-Sequenze seiner Bindungstelle ab. Verschiedene Core-spezifische Sequenzstrukturen und die DNA-Form, die durch nicht-spezifische DNA-Kontakte 'gelesen' wird, haben sich als Schlüsselfaktoren für die Modulation der GR-Aktivität erwiesen. In diesem Kontext könnte die Sequenz, die die Core-spezifische Sequenz flankiert, eine wichtige Rolle für die GR-Struktur und dessen Wechselwirkung mit DNA spielen. Unsere Simulationen weisen auf eine erhebliche langreichweitige Kommunikation zwischen der spezifischen Protein-DNA Kern Sequenz und ihren nahen flankierenden Basenpaaren hin; wir konnten eine veränderte Form der Kern-DNA-Sequenz aufgrund der flankierenden Basenpaare erkennen. Dennoch führt der Einfluss der DNA-Formveränderung nicht zu direkter Protein-DNA-Wechselwirkung (via Wasserstoffbrücken), sondern vielmehr zu einer Repositionierung der dimeren GR-DNA-Bindedomänen. Bemerkenswerterweise wurde eine solche veränderte Protein-DNA-Konformation hauptsächlich in der zweiten Hälfte des Komplexes beobachtet, was auf eine vorherrschende Rolle der -Richtung flankierenden Nukleotide für die GR-Struktur und -Konformation hinweist.

Unsere MD Simulationen und deren Analyseergebnisse haben es uns ermöglicht, diejenigen detaillieren Wechselwirkungen, welche die langreichweitigen Effekte dominieren, auf Atomlevel zu verstehen.

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Dissertationsschrift mit dem Titel

Long Range Communication In Macromolecular Systems

selbständig angefertigt und hierfür keine anderen als die angegebenen Hilfsmittel verwendet habe. Diese Arbeit wurde nicht schon einmal in einem früheren Promotionsverfahren eingereicht.

Berlin, 19. Februar 2020

List of Publications

The present thesis is based on the following manuscripts, which have been published or are in preparation for publication in peer-reviewed journals:

- *Bagherpoor Helabad, M., Ghane, T., Reidelbach, M., Woelke, A. L., Knapp, E. W., and Imhof, P. Biophysical J. 113 (2017) 817-828.*
- *Bagherpoor Helabad, M., Volkenandt, S., and Imhof, P. (2020) Molecular Dynamics Simulations of a Chimeric Androgen Receptor Protein (SPARKI) Confirm the Importance of the Dimerization Domain on DNA Binding Specificity. Front. Mol. Biosci. 7:4.*
- *Schöne S., Jurk M., Bagherpoor Helabad M., Dror I., Lebars I., Kieffer B., Imhof P., Rohs R., Vingron M., Thomas-Chollier M.et al. Sequences flanking the core-binding site modulate glucocorticoid receptor structure and activity. Nat. Commun. 2016; 7:12621.*
- *Bagherpoor Helabad, M. , Volkenandt, S. and Imhof, P. "An Investigation of the Specificity of Androgen and Glucocorticoid Receptor DNA Binding Domains for Direct and Inverted Repeat Response Elements". in preparation.*

The following publication is not discussed in the present thesis:

- *Bagherpoor Helabad, M., Kanaan, N., and Imhof, P. (2014) Base Flip in DNA Studies by Molecular Dynamics Simulation of Differently-Oxidized Forms of Methyl-Cytosine. int. J. Mol. Sci*

Acknowledgements

I must thank my love, Zahra; your unfailing support is a gift, you are my partner in every sense of the word.

I would like to express my special appreciation and thanks to my advisor *Professor Dr. Petra Imhof*, you have been a tremendous mentor for me. I would like to thank you for encouraging my research and for allowing me to grow as a research scientist. Your advice on both research as well as on my career have been priceless. I would also like to thank my committee members, especially professor Dr. Roland Netz for accepting to be my second referee.

I would like to thank my parents, *Golchin Irani* and *Mojtaba Bagherpoor Helabad*, my brothers (*Arsalan*, *Masoud*, *Mabood*, and *Ali*), and my sisters (*Fariba*, *Fatemeh*, *Saryeh*, and *Somayeh*) for unquestioningly supporting and encouraging me, for a long as I can remember in my life.

I express my heart-felt gratitude to my lab mates for their useful comments and discussions. Personal helps, scientific inputs and friendly nature has always made me feel at ease with them.

I thank all people who are working in IT-Service at Physik departement, Freie Universität Berlin for their unfailing support.

References

[1] Ron Milo and Rob Phillips. *Cell biology by the numbers*. Garland Science, 2015.

[2] Tamar Schlick. *Molecular modeling and simulation: an interdisciplinary guide: an interdisciplinary guide*, volume 21. Springer Science & Business Media, 2010.

[3] Martin Karplus and J Andrew McCammon. Molecular dynamics simulations of biomolecules. *Nature Structural and Molecular Biology*, 9(9):646, 2002.

[4] Adam Hospital, Josep Ramon Goñi, Modesto Orozco, and Josep L Gelpí. Molecular dynamics simulations: advances and applications. *Advances and applications in bioinformatics and chemistry: AABC*, 8:37, 2015.

[5] Max F Perutz. Mechanisms of cooperativity and allosteric regulation in proteins. *Quarterly reviews of biophysics*, 22(2):139–237, 1989.

[6] Joel Janin and Shoshanna J Wodak. Structural domains in proteins and their role in the dynamics of protein function. *Progress in biophysics and molecular biology*, 42:21–78, 1983.

[7] John C Kendrew, G Bodo, Howard M Dintzis, RG Parrish, Harold Wyckoff, and David C Phillips. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, 181(4610):662–666, 1958.

[8] Claudine Mayer. X-ray diffraction in biology: How can we see dna and proteins in three dimensions? In *X-ray Scattering*. InTech, 2017.

[9] Kurt Wuthrich. Nmr of proteins and nucleic acids. 1986.

[10] Joachim Frank, Jun Zhu, Pawel Penczek, Yanhong Li, Suman Srivastava, Adriana Verschoor, Michael Radermacher, Robert Grassucci, Ramani K Lata, and Rajendra K Agrawal. A model of protein synthesis based on cryo-electron microscopy of the e. coli ribosome. *Nature*, 376(6539):441, 1995.

[11] Wilfredo Evangelista, Lee-Chuan C Yeh, Aleksandra Gmyrek, J Ching Lee, and John C Lee. Long-range communication network in the type 1b bone morphogenetic protein receptor. *Biochemistry*, 54(48):7079–7088, 2015.

[12] Yan M Chan, Ibrahim M Moustafa, Jamie J Arnold, Craig E Cameron, and David D Boehr. Long-range communication between different functional sites in the picornaviral 3c protein. *Structure*, 24(4):509–517, 2016.

[13] Amit Das, Mahua Ghosh, and J Chakrabarti. Time dependent correlation between dihedral angles as probe for long range communication in proteins. *Chemical Physics Letters*, 645:200–204, 2016.

[14] Chaitanya Rastogi, H Tomas Rube, Judith F Kribelbauer, Justin Crocker, Ryan E Loker, Gabriella D Martini, Oleg Laptenko, William A Freed-Pastor, Carol Prives, David L Stern, et al. Accurate and sensitive quantification of protein-dna binding affinity. *Proceedings of the National Academy of Sciences*, page 201714376, 2018.

[15] Glenn A Maston, Sara K Evans, and Michael R Green. Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet.*, 7:29–59, 2006.

[16] Kathy Liszewski. Dissecting the structure of membrane proteins. *Genetic Engineering & Biotechnology News*, 35(17):1–14, 2015.

[17] Ankita Roy. Membrane preparation and solubilization. In *Methods in enzymology*, volume 557, pages 45–56. Elsevier, 2015.

[18] Jose G Almeida, Antonio J Preto, Panagiotis I Koukos, Alexandre MJJ Bonvin, and Irina S Moreira. Membrane proteins structures: A review on computational modeling tools. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1859(10):2021–2039, 2017.

[19] James D Watson, Francis HC Crick, et al. Molecular structure of nucleic acids. *Nature*, 171(4356):737–738, 1953.

[20] Standard Amino Acid. Lehninger principles of biochemistry. 2004.

[21] Andrew Travers and Georgi Muskhelishvili. Dna structure and function. *The FEBS journal*, 282(12):2279–2295, 2015.

[22] Po-Ssu Huang, Scott E Boyken, and David Baker. The coming of age of de novo protein design. *Nature*, 537(7620):320, 2016.

[23] Manuela M Pereira, Margarida Santana, and Miguel Teixeira. A novel scenario for the evolution of haem–copper oxygen reductases. *Biochimica et Biophysica Acta (BBA)-Bioenergetics*, 1505(2-3):185–208, 2001.

[24] Mårten Wikström, Klaas Krab, and Vivek Sharma. Oxygen activation and energy conservation by cytochrome c oxidase. *Chemical reviews*, 118(5):2469–2490, 2018.

[25] Gerald T Babcock and Mårten Wikström. Oxygen activation and the conservation of energy in cell respiration. *Nature*, 356(6367):301, 1992.

[26] Mårten Wikström. Cytochrome c oxidase: 25 years of the elusive proton pump. *Biochimica et Biophysica Acta (BBA)-Bioenergetics*, 1655:241–247, 2004.

[27] So Iwata, Christian Ostermeier, Bernd Ludwig, and Hartmut Michel. Structure at 2.8 å resolution of cytochrome c oxidase from paracoccus denitrificans. *Nature*, 376(6542):660, 1995.

[28] Tomitake Tsukihara, Hiroshi Aoyama, Eiki Yamashita, Takashi Tomizaki, Hiroshi Yamaguchi, Kyoko Shinzawa-Itoh, Ryosuke Nakashima, Rieko Yaono, and Shinya Yoshikawa. The whole structure of the 13-subunit oxidized cytochrome c oxidase at 2.8 å. *Science*, 272(5265):1136–1144, 1996.

[29] Shinya Yoshikawa, Kyoko Shinzawa-Itoh, Ryosuke Nakashima, Rieko Yaono, Eiki Yamashita, Noriko Inoue, Min Yao, Ming Jie Fei, Clare Peters Libeu, Tsunehiro Mizushima, et al. Redox-coupled crystal structural changes in bovine heart cytochrome c oxidase. *Science*, 280(5370):1723–1729, 1998.

[30] Ling Qin, Carrie Hiser, Anne Mulichak, R Michael Garavito, and Shelagh Ferguson-Miller. Identification of conserved lipid/detergent-binding sites in a high-resolution structure of the membrane protein cytochrome c oxidase. *Proceedings of the National Academy of Sciences*, 103(44):16117–16122, 2006.

[31] Juergen Koepke, Elena Olkhova, Heike Angerer, Hannelore Müller, Guohong Peng, and Hartmut Michel. High resolution crystal structure of paracoccus denitrificans cytochrome c oxidase: new insights into the active site and the proton transfer pathways. *Biochimica et Biophysica Acta (BBA)-Bioenergetics*, 1787(6):635–645, 2009.

[32] Hyun Ju Lee, Emelie Svahn, Jessica MJ Swanson, Håkan Lepp, Gregory A Voth, Peter Brzezinski, and Robert B Gennis. Intricate role of water in proton transport through cytochrome c oxidase. *Journal of the American Chemical Society*, 132(45):16225–16239, 2010.

[33] Anna Lena Woelke, Gegham Galstyan, and Ernst-Walter Knapp. Lysine 362 in cytochrome c oxidase regulates opening of the k-channel via changes in pka and conformation. *Biochimica et Biophysica Acta (BBA)-Bioenergetics*, 1837(12):1998–2003, 2014.

[34] Mahdi Bagherpoor Helabad, Tahereh Ghane, Marco Reidelbach, Anna Lena Woelke, Ernst Walter Knapp, and Petra Imhof. Protonation-state-dependent communication in cytochrome c oxidase. *Biophysical journal*, 113(4):817–828, 2017.

[35] Shinya Yoshikawa and Atsuhiro Shimada. Reaction mechanism of cytochrome c oxidase. *Chemical reviews*, 115(4):1936–1989, 2015.

[36] Chris M Bunce and Moray J Campbell. *Nuclear Receptors: current concepts and future challenges*, volume 8. Springer Science & Business Media, 2010.

[37] John C Achermann, John Schwabe, Louise Fairall, and Krishna Chatterjee. Genetic disorders of nuclear receptors. *The Journal of clinical investigation*, 127(4):1181–1192, 2017.

[38] Valentina Contrò, J Basile, and Patrizia Proia. Sex steroid hormone receptors, their ligands, and nuclear and non-nuclear pathways. *AIMS Molecular Science*, 2, 2015.

[39] Mary Anne Carson-Jurica, William T Schrader, and Bert W O'Malley. Steroid receptor family: structure and functions. *Endocrine reviews*, 11(2):201–220, 1990.

[40] Albert O Brinkmann, PW Faber, HCJ Van Rooij, GGJM Kuiper, C Ris, P Klaassen, JAGM Van der Korput, MM Voorhorst, JH Van Laar, E Mulder, et al. The human androgen receptor: domain structure, genomic organization and regulation of expression. *Journal of steroid biochemistry*, 34(1-6):307–310, 1989.

[41] Raj Kumar and E Brad Thompson. The structure of the nuclear hormone receptors. *Steroids*, 64(5):310–319, 1999.

[42] Bf F Luisi, WX_ Xu, Z Otwinowski, LP Freedman, KR Yamamoto, and PB Sigler. Crystallographic analysis of the interaction of the glucocorticoid receptor with dna. *Nature*, 352(6335):497, 1991.

[43] Sebastiaan H Meijsing, Miles A Pufall, Alex Y So, Darren L Bates, Lin Chen, and Keith R Yamamoto. Dna binding site sequence directs glucocorticoid receptor structure and activity. *Science*, 324(5925):407–410, 2009.

[44] Miguel Beato, Peter Herrlich, and Günther Schütz. Steroid hormone receptors: many actors in search of a plot. *Cell*, 83(6):851–857, 1995.

[45] Paul L Shaffer, Arif Jivan, D Eric Dollins, Frank Claessens, and Daniel T Gewirth. Structural basis of androgen receptor binding to selective androgen response elements. *Proceedings of the National Academy of Sciences*, 101(14):4758–4763, 2004.

[46] Frank Claessens, Sarah Denayer, Nora Van Tilborgh, Stefanie Kerkhofs, Christine Helsen, and Annemie Haelens. Diverse roles of androgen receptor (ar) domains in ar-mediated signaling. *Nuclear receptor signaling*, 6(1):nrs–06008, 2008.

[47] Kris Schauwaers, Karel De Gendt, Philippa TK Saunders, Nina Atanassova, Annemie Haelens, Leen Callewaert, Udo Moehren, Johannes V Swinnen, Guido Verhoeven, Guy Verrijdt, et al. Loss of androgen receptor binding to selective androgen response elements causes a reproductive phenotype in a knockin mouse model. *Proceedings of the National Academy of Sciences*, 104(12):4961–4966, 2007.

[48] Stefanie Kerkhofs, Vanessa Dubois, Karel De Gendt, Christine Helsen, Liesbeth Clinckemalie, Lien Spans, Frans Schuit, Steven Boonen, Dirk Vanderschueren, Philippa TK Saunders, et al. A role for selective androgen response elements in the development of the epididymis and the androgen control of the 5α reductase ii gene. *The FASEB Journal*, 26(10):4360–4372, 2012.

[49] Biswajyoti Sahu, Päivi Pihlajamaa, Vanessa Dubois, Stefanie Kerkhofs, Frank Claessens, and Olli A Jänne. Androgen receptor uses relaxed response element stringency for selective chromatin binding and transcriptional regulation in vivo. *Nucleic acids research*, 42(7):4230–4240, 2014.

[50] Stefanie Schöne, Marcel Jurk, Mahdi Bagherpoor Helabad, Iris Dror, Isabelle Lebars, Bruno Kieffer, Petra Imhof, Remo Rohs, Martin Vingron, Morgane Thomas-Chollier, et al. Sequences flanking the core-binding site modulate glucocorticoid receptor structure and activity. *Nature communications*, 7:12621, 2016.

[51] Mark Tuckerman. *Statistical mechanics: theory and molecular simulation*. Oxford university press, 2010.

[52] Jeff Wereszczynski and J Andrew McCammon. Statistical mechanics and molecular dynamics in evaluating thermodynamic properties of biomolecular recognition. *Quarterly reviews of biophysics*, 45(1):1–25, 2012.

[53] RK Pathria and Paul D Beale. Statistical mechanics, 1996. *Butter worth*, page 32.

[54] Daniel V Schroeder. An introduction to thermal physics, 1999.

[55] Daan Frenkel and Berend Smit. *Understanding molecular simulation: from algorithms to applications*, volume 1. Elsevier, 2001.

[56] J Kenneth Shultis and Richard E Faw. *Fundamentals of Nuclear Science and Engineering Third Edition*. CRC press, 2016.

[57] RC Tolman. The principles of statistical mechanics, oxford. 1938.

[58] Maayke Stomp, Jef Huisman, Lucas J Stal, and Hans CP Matthijs. Colorful niches of phototrophic microorganisms shaped by vibrations of the water molecule. *The ISME journal*, 1(4):271, 2007.

[59] Herman JC Berendsen, JPM van Postma, Wilfred F van Gunsteren, ARHJ DiNola, and Jan R Haak. Molecular dynamics with coupling to an external bath. *The Journal of chemical physics*, 81(8):3684–3690, 1984.

[60] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *The Journal of chemical physics*, 126(1):014101, 2007.

[61] Shuichi Nosé. A unified formulation of the constant temperature molecular dynamics methods. *The Journal of chemical physics*, 81(1):511–519, 1984.

[62] William G Hoover. Canonical dynamics: Equilibrium phase-space distributions. *Physical review A*, 31(3):1695, 1985.

[63] Michael P Allen and Dominic J Tildesley. *Computer simulation of liquids*. Oxford university press, 2017.

[64] Bernard R Brooks, Charles L Brooks III, Alexander D Mackerell Jr, Lennart Nilsson, Robert J Petrella, Benoît Roux, Youngdo Won, Georgios Archontis, Christian Bartels, Stefan Boresch, et al. Charmm: the biomolecular simulation program. *Journal of computational chemistry*, 30(10):1545–1614, 2009.

[65] David A Case, Thomas E Cheatham III, Tom Darden, Holger Gohlke, Ray Luo, Kenneth M Merz Jr, Alexey Onufriev, Carlos Simmerling, Bing Wang, and Robert J Woods. The amber biomolecular simulation programs. *Journal of computational chemistry*, 26(16):1668–1688, 2005.

[66] William L Jorgensen and Julian Tirado-Rives. The opls force field for proteins. energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc*, 110(6):1657–1723, 1988.

[67] HJC Berendsen, D Van der Spoel, E Lindahl, B Hess, G Groenhof, and AE Mark. Gromacs: fast, flexible, and free. *J. Comput. Chem*, 26:1701–1718, 2005.

[68] Ulrich Essmann, Lalith Perera, Max L Berkowitz, Tom Darden, Hsing Lee, and Lee G Pedersen. A smooth particle mesh ewald method. *The Journal of chemical physics*, 103(19):8577–8593, 1995.

[69] Eli Barkai, Yuval Garini, and Ralf Metzler. of single molecules in living cells. *Phys. Today*, 65(8):29, 2012.

[70] PH Hünenberger, AE Mark, and WF Van Gunsteren. Fluctuation and cross-correlation analysis of protein motions observed in nanosecond molecular dynamics simulations. *Journal of molecular biology*, 252(4):492–503, 1995.

[71] Oliver F Lange and Helmut Grubmüller. Generalized correlation for biomolecular dynamics. *Proteins: Structure, Function, and Bioinformatics*, 62(4):1053–1061, 2006.

[72] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical review E*, 69(6):066138, 2004.

[73] Toshiko Ichiye and Martin Karplus. Collective motions in proteins: a covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. *Proteins: Structure, Function, and Bioinformatics*, 11(3):205–217, 1991.

[74] Clarisse G Ricci, Rodrigo L Silveira, Ivan Rivalta, Victor S Batista, and Munir S Skaf. Allosteric pathways in the pparγ-rxrα nuclear receptor complex. *Scientific reports*, 6:19940, 2016.

[75] Jürgen Schlitter. Estimation of absolute and relative entropies of macromolecules using the covariance matrix. *Chemical Physics Letters*, 215(6):617–621, 1993.

[76] Ioan Andricioaei and Martin Karplus. On the calculation of entropy from covariance matrices of the atomic fluctuations. *The Journal of Chemical Physics*, 115(14):6289–6292, 2001.

[77] Panagiotis I Koukos and Nicholas M Glykos. Grcarma: A fully automated task-oriented interface for the analysis of molecular dynamics trajectories. *Journal of computational chemistry*, 34(26):2310–2312, 2013.

[78] R Lavery, M Moakher, JH Maddocks, D Petkeviciute, and K Zakrzewska. Curves+ web server for analyzing and visualizing the helical, backbone and groove parameters of nucleic acid structures. *Nucleic Acids Res*, 37:5917–5929, 2009.

[79] Tomáš Dršata, Naďa Špačková, Petr Jurečka, Marie Zgarbová, Jiří Šponer, and Filip Lankaš. Mechanical properties of symmetric and asymmetric dna a-tracts: implications for looping and nucleosome positioning. *Nucleic acids research*, 42(11):7383–7394, 2014.

[80] Richard Chace Tolman. *The principles of statistical mechanics*. Courier Corporation, 1979.

[81] Oscar Gonzalez, Marco Pasi, Daiva Petkeviciute, Jaroslaw Glowacki, and JH Maddocks. Absolute versus relative entropy parameter estimation in a coarse-grain model of dna. *Multiscale Modeling & Simulation*, 15(3):1073–1107, 2017.

[82] Jonathan P Hosler, Shelagh Ferguson-Miller, and Denise A Mills. Energy transduction: proton transfer through the respiratory complexes. *Annu. Rev. Biochem.*, 75:165–187, 2006.

[83] Ilya Belevich and Michael I Verkhovsky. Molecular mechanism of proton translocation by cytochrome c oxidase. *Antioxidants & redox signaling*, 10(1):1–30, 2008.

[84] Rowan M Henry, Ching-Hsing Yu, Tomas Rodinger, and Régis Pomès. Functional hydration and conformational gating of proton uptake in cytochrome c oxidase. *Journal of molecular biology*, 387(5):1165–1185, 2009.

[85] Andreas Namslauer, Håkan Lepp, Magnus Brändén, Audrius Jasaitis, Michael I Verkhovsky, and Peter Brzezinski. Plasticity of proton pathway structure and water coordination in cytochrome c oxidase. *Journal of Biological Chemistry*, 282(20):15148–15158, 2007.

[86] Christian Ostermeier, Axel Harrenga, Ulrich Ermler, and Hartmut Michel. Structure at 2.7 å resolution of the paracoccus denitrificans two-subunit cytochrome c oxidase complexed with an antibody fv fragment. *Proceedings of the National Academy of Sciences*, 94(20):10547–10553, 1997.

[87] Ivo Hofacker and Klaus Schulten. Oxygen and proton pathways in cytochrome c oxidase. *Proteins: Structure, Function, and Bioinformatics*, 30(1):100–107, 1998.

[88] Magnus Brändén, Håkan Sigurdson, Andreas Namslauer, Robert B Gennis, Pia Ädelroth, and Peter Brzezinski. On the role of the k-proton transfer pathway in cytochrome c oxidase. *Proceedings of the National Academy of Sciences*, 98(9):5013–5018, 2001.

[89] Vivek Sharma and Mårten Wikström. The role of the k-channel and the active-site tyrosine in the catalytic mechanism of cytochrome c oxidase. *Biochimica et Biophysica Acta (BBA)-Bioenergetics*, 1857(8):1111–1115, 2016.

[90] Farol L Tomson, Joel E Morgan, Guoping Gu, Blanca Barquera, TV Vygodina, and Robert B Gennis. Substitutions for glutamate 101 in subunit ii of cytochrome c oxidase from rhodobacter sphaeroides result in blocking the proton-conducting k-channel. *Biochemistry*, 42(6):1711–1717, 2003.

[91] Alex D MacKerell Jr, Donald Bashford, MLDR Bellott, Roland Leslie Dunbrack Jr, Jeffrey D Evanseck, Martin J Field, Stefan Fischer, Jiali Gao, H Guo, Sookhee Ha, et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *The journal of physical chemistry B*, 102(18):3586–3616, 1998.

[92] William L Jorgensen, Jayaraman Chandrasekhar, Jeffry D Madura, Roger W Impey, and Michael L Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of chemical physics*, 79(2):926–935, 1983.

[93] Jeffery B Klauda, Richard M Venable, J Alfredo Freites, Joseph W O'Connor, Douglas J Tobias, Carlos Mondragon-Ramirez, Igor Vorobyov, Alexander D MacKerell Jr, and Richard W Pastor. Update of the charmm all-atom additive force field for lipids: validation on six lipid types. *The journal of physical chemistry B*, 114(23):7830–7843, 2010.

[94] Tom Darden, Darrin York, and Lee Pedersen. Particle mesh ewald: An nlog(n) method for ewald sums in large systems. *The Journal of chemical physics*, 98(12):10089–10092, 1993.

[95] Jean-Paul Ryckaert, Giovanni Ciccotti, and Herman JC Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of computational physics*, 23(3):327–341, 1977.

[96] Ville RI Kaila, Michael I Verkhovsky, Gerhard Hummer, and Mårten Wikström. Glutamic acid 242 is a valve in the proton pump of cytochrome c oxidase. *Proceedings of the National Academy of Sciences*, 105(17):6255–6259, 2008.

[97] CJT De Grotthuss. Memoir on the decomposition of water and of the bodies that it holds in solution by means of galvanic electricity. *Biochimica et Biophysica Acta (BBA)-Bioenergetics*, 1757(8):871–875, 2006.

[98] Tahereh Ghane, Rene F Gorriz, Sandro Wrzalek, Senta Volkenandt, Ferand Dalatieh, Marco Reidelbach, and Petra Imhof. Hydrogen-bonded network and water dynamics in the d-channel of cytochrome c oxidase. *The Journal of membrane biology*, 251(3):299–314, 2018.

[99] Erik Schoenmakers, Guy Verrijdt, Ben Peeters, Guido Verhoeven, Wilfried Rombauts, and Frank Claessens. Differences in dna binding characteristics of the androgen and glucocorticoid receptors can determine hormone-specific responses. *Journal of Biological Chemistry*, 275(16):12290–12297, 2000.

[100] Lisa C Watson, Kristopher M Kuchenbecker, Benjamin J Schiller, John D Gross, Miles A Pufall, and Keith R Yamamoto. The glucocorticoid receptor dimer interface allosterically transmits sequence-specific dna signals. *Nature Structural and Molecular Biology*, 20(7):876, 2013.

[101] Jeffrey A Lefstin and Keith R Yamamoto. Allosteric effects of dna on transcriptional regulators. *Nature*, 392(6679):885, 1998.

[102] Annemie Haelens, Guy Verrijdt, Leen Callewaert, Valerie Christiaens, Kris Schauwaers, Ben Peeters, Wilfried Rombauts, and Frank Claessens. Dna recognition by the androgen receptor: evidence for an alternative dna-dependent dimerization, and an active role of sequences flanking the response element on transactivation. *Biochemical Journal*, 369(1):141–151, 2003.

[103] Guy Verrijdt, Erik Schoenmakers, Annemie Haelens, Ben Peeters, Guido Verhoeven, Wilfried Rombauts, and Frank Claessens. Change of specificity mutations in androgen-selective enhancers evidence for a role of differential dna binding by the androgen receptor. *Journal of Biological Chemistry*, 275(16):12298–12305, 2000.

[104] Remo Rohs, Xiangshu Jin, Sean M West, Rohit Joshi, Barry Honig, and Richard S Mann. Origins of specificity in protein-dna recognition. *Annual review of biochemistry*, 79:233–269, 2010.

[105] Christoph Geserick, Hellmuth-Alexander Meyer, Karina Barbulescu, and Bernard Haendler. Differential modulation of androgen receptor action by deoxyribonucleic acid response elements. *Molecular Endocrinology*, 17(9):1738–1750, 2003.

[106] Guy Verrijdt, Annemie Haelens, and Frank Claessens. Selective dna recognition by the androgen receptor as a mechanism for hormone-specific regulation of gene expression. *Molecular genetics and metabolism*, 78(3):175–185, 2003.

[107] Guy Verrijdt, Tamzin Tanner, U Moehren, Leen Callewaert, Anna Haelens, and Frank Claessens. The androgen receptor dna-binding domain determines androgen selectivity of transcriptional response, 2006.

[108] Martin E van Royen, Wiggert A van Cappellen, Carola de Vos, Adriaan B Houtsmuller, and Jan Trapman. Stepwise androgen receptor dimerization. *J Cell Sci*, 125(8):1970–1979, 2012.

[109] Zhifeng Zhou, Jeffry L Corden, and Terry R Brown. Identification and characterization of a novel androgen response element composed of a direct repeat. *Journal of Biological Chemistry*, 272(13):8227–8235, 1997.

[110] Erik Schoenmakers, ALEN Philippe, Guy Verrijdt, Ben PEETERS, Guido Verhoeven, Wilfried Rombauts, and Frank Claessens. Differential dna binding by the androgen and glucocorticoid receptors involves the second zn-finger and a c-terminal extension of the dna-binding domains. *Biochemical Journal*, 341(3):515–521, 1999.

[111] Emily R Weikum, Matthew T Knuesel, Eric A Ortlund, and Keith R Yamamoto. Glucocorticoid receptor control of transcription: precision and plasticity via allostery. *Nature reviews Molecular cell biology*, 18(3):159, 2017.

[112] David L Bain, Qin Yang, Keith D Connaghan, James P Robblee, Michael T Miura, Gregory D Degala, James R Lambert, and Nasib K Maluf. Glucocorticoid receptor–dna interactions: binding energetics are the primary determinant of sequence-specific transcriptional activity. *Journal of molecular biology*, 422(1):18–32, 2012.

[113] Filipp Frank, C Denise Okafor, and Eric A Ortlund. The first crystal structure of a dna-free nuclear receptor dna binding domain sheds light on dna-driven allostery in the glucocorticoid receptor. *Scientific reports*, 8(1):13497, 2018.

[114] Christine Helsen, Vanessa Dubois, Annelien Verfaillie, Jacques Young, Mieke Trekels, Renée Vancraenenbroeck, Marc De Maeyer, and Frank Claessens. Evidence for dna-binding domain–ligand-binding domain communications in the androgen receptor. *Molecular and cellular biology*, 32(15):3033–3043, 2012.

[115] Kush Dalal, Mani Roshan-Moniri, Aishwariya Sharma, Huifang Li, Fuqiang Ban, Mohamed Hessein, Michael Hsing, Kriti Singh, Eric LeBlanc, Scott Dehm, et al. Selectively targeting the dna-binding domain of the androgen receptor as a prospective therapy for prostate cancer. *Journal of Biological Chemistry*, 289(38):26417–26429, 2014.

[116] Liyang Zhang, Gabriella D Martini, H Tomas Rube, Judith F Kribelbauer, Chaitanya Rastogi, Vincent D FitzPatrick, Jon C Houtman, Harmen J Bussemaker, and Miles A Pufall. Selexglm differentiates androgen and glucocorticoid receptor dna-binding preference over an extended binding site. *Genome research*, 28(1):111–121, 2018.

[117] James C Phillips, Rosemary Braun, Wei Wang, James Gumbart, Emad Tajkhorshid, Elizabeth Villa, Christophe Chipot, Robert D Skeel, Laxmikant Kale, and Klaus Schulten. Scalable molecular dynamics with namd. *Journal of computational chemistry*, 26(16):1781–1802, 2005.

[118] Nicolas Foloppe and Alexander D MacKerell Jr. All-atom empirical force field for nucleic acids: I. parameter optimization based on small molecule and condensed phase macromolecular target data. *Journal of computational chemistry*, 21(2):86–104, 2000.

[119] Alexander D Mackerell and Nilesh K Banavali. All-atom empirical force field for nucleic acids: Ii. application to molecular dynamics simulations of dna and rna in solution. *Journal of Computational Chemistry*, 21(2):105–120, 2000.

[120] Eric F Pettersen, Thomas D Goddard, Conrad C Huang, Gregory S Couch, Daniel M Greenblatt, Elaine C Meng, and Thomas E Ferrin. Ucsf chimera-a visualization system for exploratory research and analysis. *Journal of computational chemistry*, 25(13):1605–1612, 2004.

[121] Nosé Shuichi. Constant temperature molecular dynamics methods. *Progress of Theoretical Physics Supplement*, 103:1–46, 1991.

[122] Scott E Feller, Yuhong Zhang, Richard W Pastor, and Bernard R Brooks. Constant pressure molecular dynamics simulation: the langevin piston method. *The Journal of chemical physics*, 103(11):4613–4621, 1995.

[123] Remo Rohs, Sean M West, Alona Sosinsky, Peng Liu, Richard S Mann, and Barry Honig. The role of dna shape in protein–dna recognition. *Nature*, 461(7268):1248, 2009.

[124] Päivi Pihlajamaa, Biswajyoti Sahu, and Olli A Jänne. Determinants of receptor-and tissue-specific actions in androgen signaling. *Endocrine reviews*, 36(4):357–384, 2015.

[125] Udo Moehren, Sarah Denayer, Michael Podvinec, Guy Verrijdt, and Frank Claessens. Identification of androgen-selective androgen-response elements in the human aquaporin-5 and rad9 genes. *Biochemical Journal*, 411(3):679–686, 2008.

[126] Sam John, Peter J Sabo, Robert E Thurman, Myong-Hee Sung, Simon C Biddie, Thomas A Johnson, Gordon L Hager, and John A Stamatoyannopoulos. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nature genetics*, 43(3):264, 2011.

[127] Biswajyoti Sahu, Marko Laakso, Kristian Ovaska, Tuomas Mirtti, Johan Lundin, Antti Rannikko, Anna Sankila, Juha-Pekka Turunen, Mikael Lundin, Juho Konsti, et al. Dual role of foxa1 in androgen receptor binding to chromatin, androgen signalling and prostate cancer. *The EMBO journal*, 30(19):3962–3976, 2011.

[128] Cecilia Ballaré, Roser Zaurin, Guillermo P Vicent, and Miguel Beato. More help than hindrance: nucleosomes aid transcriptional regulation. *Nucleus*, 4(3):189–194, 2013.

[129] Ping Yin, Damian Roqueiro, Lei Huang, Jonas K Owen, Anna Xie, Antonia Navarro, Diana Monsivais, J Julie Kim, Yang Dai, Serdar E Bulun, et al. Genome-wide progesterone receptor binding: cell type-specific and shared mechanisms in t47d breast cancer cells and primary leiomyoma cells. *PLoS one*, 7(1):e29021, 2012.

[130] Lars Grøntved, Sam John, Songjoon Baek, Ying Liu, John R Buckley, Charles Vinson, Greti Aguilera, and Gordon L Hager. C/ebp maintains chromatin accessibility in liver and facilitates glucocorticoid receptor recruitment to steroid response elements. *The EMBO journal*, 32(11):1568–1583, 2013.

[131] Bernard R Brooks, Robert E Bruccoleri, Barry D Olafson, David J States, S a Swaminathan, and Martin Karplus. Charmm: a program for macromolecular energy, minimization, and dynamics calculations. *Journal of computational chemistry*, 4(2):187–217, 1983.

[132] Michael W Mahoney and William L Jorgensen. A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. *The Journal of Chemical Physics*, 112(20):8910–8922, 2000.

[133] Glenn J Martyna, Douglas J Tobias, and Michael L Klein. Constant pressure molecular dynamics algorithms. *The Journal of chemical physics*, 101(5):4177–4189, 1994.

[134] Magaly Del Monaco, Seana P Covello, Susan H Kennedy, Gwen Gilinger, Gerald Litwack, and Jouni Uitto. Identification of novel glucocorticoid-response elements in human elastin promoter and demonstration of nucleotide sequence specificity of the receptor binding. *Journal of investigative dermatology*, 108(6):938–942, 1997.

[135] Alex Yick-Lun So, Samantha B Cooper, Brian J Feldman, Mitra Manuchehri, and Keith R Yamamoto. Conservation analysis predicts in vivo occupancy of glucocorticoid receptor-binding sequences at glucocorticoid-induced genes. *Proceedings of the National Academy of Sciences*, 105(15):5745–5749, 2008.

[136] Milan Surjit, Krishna Priya Ganti, Atish Mukherji, Tao Ye, Guoqiang Hua, Daniel Metzger, Mei Li, and Pierre Chambon. Widespread negative response elements mediate direct repression by agonist-liganded glucocorticoid receptor. *Cell*, 145(2):224–241, 2011.

[137] Venkata Rajesh Yella, Devesh Bhimsaria, Debostuti Ghoshdastidar, José A Rodríguez-Martínez, Aseem Z Ansari, and Manju Bansal. Flexibility and structure of flanking dna impact transcription factor affinity for its core motif. *Nucleic acids research*, 46(22):11883–11897, 2018.

[138] Gary D Stormo and Yue Zhao. Determining the specificity of protein–dna interactions. *Nature Reviews Genetics*, 11(11):751, 2010.

[139] Dmitrij Golovenko, Bastian Bräuning, Pratik Vyas, Tali E Haran, Haim Rozenberg, and Zippora Shakked. New insights into the role of dna shape on its recognition by p53 proteins. *Structure*, 26(9):1237–1250, 2018.

[140] Ruibin Liang, Jessica MJ Swanson, Mårten Wikström, and Gregory A Voth. Understanding the essential proton-pumping kinetic gates and decoupling mutations in cytochrome c oxidase. *Proceedings of the National Academy of Sciences*, 114(23):5924–5929, 2017.

[141] Magnus Brändén, Farol Tomson, Robert B Gennis, and Peter Brzezinski. The entry point of the k-proton-transfer pathway in cytochrome c oxidase. *Biochemistry*, 41(35):10794–10798, 2002.

[142] Takefumi Yamashita and Gregory A Voth. Insights into the mechanism of proton transport in cytochrome c oxidase. *Journal of the American Chemical Society*, 134(2):1147–1152, 2012.