Univerzita Palackého v Olomouci
Přírodovědecká fakulta
Katedra biofyziky

DIPLOMOVÁ PRÁCE

Příspěvek elektrostatických interakcí ke stabilitě komplexů proteinů s DNA

Vypracovala: Bc. Anna Jelinková
Studijní obor: Molekulární biofyzika
Vedoucí diplomové práce: doc. RNDr. Petr Jurečka, Ph.D.

Olomouc 2023

Palacký University Olomouc
Faculty of science
Department of Biophysics

MASTER THESIS

Contribution of electrostatic interactions to the stability of protein-DNA complexes

Author: Bc. Anna Jelinková
Field of Study: Molecular Biophysics
Supervisor: doc. RNDr. Petr Jurečka, Ph.D.

Olomouc 2023

Bibliografická identifikace

Jméno a příjmení autora	Bc. Anna Jelinková
Název práce	Příspěvek elektrostatických interakcí ke stabilitě komplexů proteinů s DNA
Typ práce	Diplomová
Pracoviště	Katedra biofyziky a Katedra fyzikální chemie
Vedoucí práce	doc. RNDr. Petr Jurečka, Ph.D.
Rok obhajoby práce	2023
Abstrakt	Ačkoliv silová pole (empirické potenciály) poskytují vysokostandardní simulace, parametry popisující elektrostatické interakce mezi nabitými rezidui, jako fosfáty a kationty, jsou nadhodnocené. Tato práce byla zaměřena na studium elektrostatických interakcí v rámci komplexů proteinů s DNA pomocí simulací molekulové dynamiky, prováděné v silovém poli OL21 na setu vybraných komplexů. Byly zkoumány efekty modifikace CUFIX a fosfátové modi-fikace, přičemž použití CUFIX-u mělo podle očekávání za následek zeslabení elektrostatických interakcí. Nicméně, fosfátová modifikace ze skupiny profesora Case se neukázala jako vhodná alternativa k CUFIX-u, jelikož její efekty byly zanedbatelné nebo dokonce opačného charakteru. Systematická alternativa nezávislá na modelu rozpouštědla je tudíž stále potřebná.
Klíčová slova	molekulová dynamika, molekulová mechanika, silové pole, empirický potenciál, protein, DNA, solný můstek, interakce, OL21, AMBER, CUFIX, fosfátová modifikace
Počet stran	x + 42
Počet příloh	0
Jazyk	anglický

Bibliographical identification

Autor's first name and surname	Bc. Anna Jelinková
Title	Contribution of electrostatic interactions to the stability of protein-DNA complexes
Type of thesis	Master
Department	Department of Biophysics and Department of Physical Chemistry
Supervisor	doc. RNDr. Petr Jurečka, Ph.D.
The year of presentation	2023
Abstract	While various force fields (empirical potentials) provide high-standard simulations, the parameters describing electrostatic interactions between charged residues, such as phosphates and cations, are known to be overestimated. This thesis was focused on studying electrostatic interactions within protein-DNA complexes by performing molecular dynamics simulation in OL21 force field on set of such complexes. The effects of CUFIX and phosphate modifications were analysed and compared, whereas CUFIX weakened the electrostatic interactions as expected. However, this thesis has shown that the phosphate modification by Case group is not a viable alternative to CUFIX, as its effects were negligible or even of the opposite nature. Therefore, an alternative that would be a systematical modification of force field parameters, independent of solvent model, is required.
Keywords	molecular dynamics, molecular mechanics, force field, empirical potential, protein, DNA, salt bridge, interaction, OL21, AMBER, CUFIX, phosphate modification
Number of pages	x + 42
Number of appendices	0
Language	English

I would like to thank my supervisor, doc. RNDr. Petra Jurečka, Ph.D., for his guidance all the way from choosing an interesting topic to completing it, along with professional supervision, valuable advice and patience throughout the completion of this thesis.

Contents

List of abbreviations

DNA	Deoxyribonucleic Acid
Hox	Homeobox
PDB	Protein Data Bank
HMG-box	High Mobility Group box
SRY	Sex-determining Region Y
hSRY(HMG)	Human male Sex-determining Region Y
HJ	Holliday junction
DAI	DNA-dependent Activator of IFN-regulatory factors
MD	Molecular Dynamics
QM	Quantum Mechanics
MM	Molecular Mechanics
WFT	Wave Function Theory
DFT	Density Function Theory
AMBER	Assisted Model Building with Energy Refinement
CHARMM	Chemistry at Harvard Macromolecular Dynamics
GROMACS	Groningen Machine for Chemical Simulation
OPLS	Optimized Potentials for Liquid Simulations
SPC	Simple Point-Charge
TIP3P	Transferable Intermolecular Potential Three Point
OL	Olomouc
RESP	Restrained Electrostatic Potential
LJ	Lennard-Jones
NBFIX	Non-Bonded Fix
XRD	X-ray diffraction
NMR	Nulear Magnetic Resonance
His	Histidine
PMEMD	Particle Mesh Ewald Molecular Dynamics
VMD	Visual Molecular Dynamics
RMSD	Root-Mean-Square Deviation

List of Figures

List of Tables

Introduction

Among the most important molecules for life, undoubtedly, belong DNA and proteins. Each of these biomacromolecules is involved in essential tasks to fulfill the quest of life. Besides their individual contribution to the biological systems, complexes of DNA and proteins are inevitable components of said systems. The understanding of the stability and the interactions of these complexes can be crucial for many sectors in today's society. The studying of said matters, however, cannot be easily done due to the microscopic scale on which processes involving DNA and proteins occur. Still, there are certain limitations of what can be experimentally studied. As addition to applied studies of molecule structure and behavior, theoretical approaches may provide the view needed. Molecular dynamics simulations are one example of such possibilities.

Current parameters of empirical force fields used in the molecular dynamic simulation process have been observed to inaccurately describe the electrostatic interactions, particulary in relation to cation-anion attraction, which plays important role in non-specific interactions within protein-DNA complexes. Consequently, various approaches to modify the force field parameters emerged, as the computational power of today's technology rises. CUFIX is one of these modification, which is based on correction of Lennard-Jones parameters of electrostatic interaction and seems to have significant effect on the simulation. However, this modification is optimized with certain experimental data, moreover, it is designed to work with specific water model TIP3P, which nowadays is less preferred than the SPC/E water model. Therefore, a systematic modification is needed. Another alternative is phosphate modification, which is based on increase of van der Waals radii. This thesis focuses on testing these modifications on suitable set of protein-DNA complexes in OL21 force field, which is currently recommended force field for protein and DNA molecular dynamic simulations, and analysing their relevance in use for these simulations.

Chapter 1

Theory

1.1 DNA and proteins

Understanding the structure and interactions of DNA and proteins in biological systems is one of the most important tasks of ongoing research. Since the determination of DNA structure by Watson and Crick in 1953, humankind was able to solve many health, environmental or food-shortage issues with DNA and proteins involved. As it will be put to see later, negatively charged phosphate in DNA backbone largely interacts with positively charged residues of amino acids in protein backbone. To discuss the intermolecular interactions of DNA and proteins, it is fundamental to describe their structures.

1.1.1 DNA

Deoxyribonucleic acid (DNA) is a biopolymer structure composed of two polynucleotide chains coiled around each other to form a double helix (Watson and Crick 1953). Each nucleotide is composed of one of the four nitrogenous base (adenine, guanine, cytosine or thymine), a sugar called deoxyribose, and a phosphate group (Fig. 1.1).

The character of the bond between nucleotides in the polynucleotide chain is covalent, called phosphodiester linkage. This bond is formed between the sugar of one nucleotide and the phosphate group of the next nucleotide, thus forming sugar-phosphate backbone. Two separate polynucleotide chains are bonded through hydrogen bonds between nitrogenous bases, following Watson-Crick base-pairing: adenine - thymine, guanine - cytosine, forming two, and three hydrogen bonds between them, respectively (Fig. 1.2).

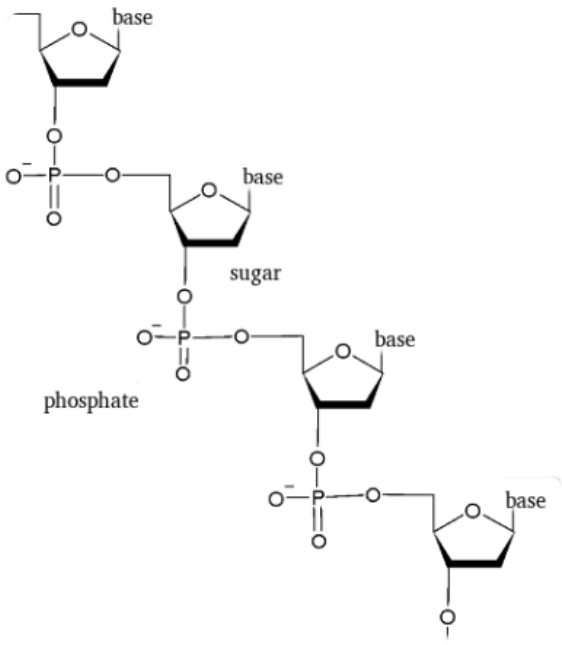
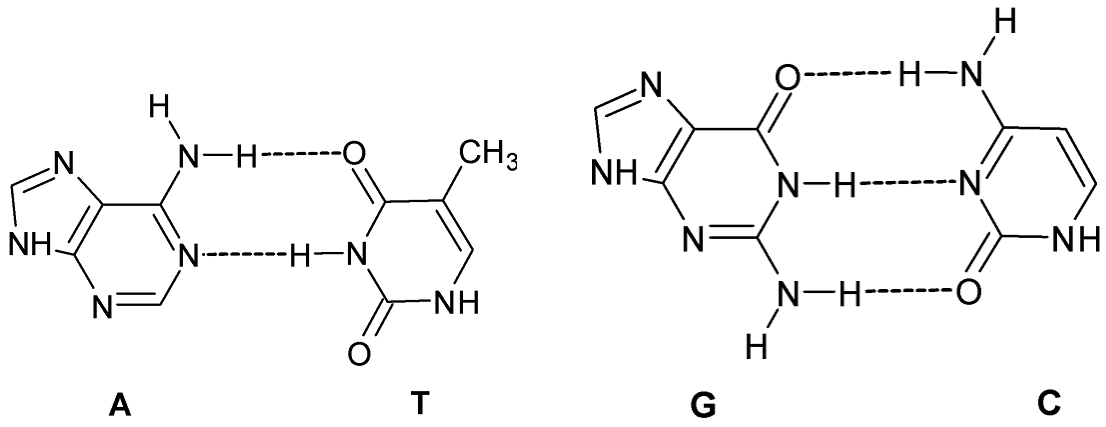


Figure 1.1: DNA backbone composed of nitrogenous base (adenine, guanine, cytosine or thymine), sugar (deoxyribose) and a phosphate group. *(Created in ChemSketch.)*



(a) (b)

Figure 1.2: Watson-Crick base-pairing of DNA nitrogenous bases. a) adenine - thymine, forming two hydrogen bonds between them. b) guanine - cytosine, forming three hydrogen bonds between them. *(Created in ChemSketch.)*

The stability of the DNA double helix is mainly due to the hydrogen bonds formed between base pairs, the base-stacking interactions (also known as π-stacking) that occur between aromatic nucleobases, and the hydrophobic effect (Yakovchuk et al. 2006)(Fig. 1.3, Kawai and Majima 2002).

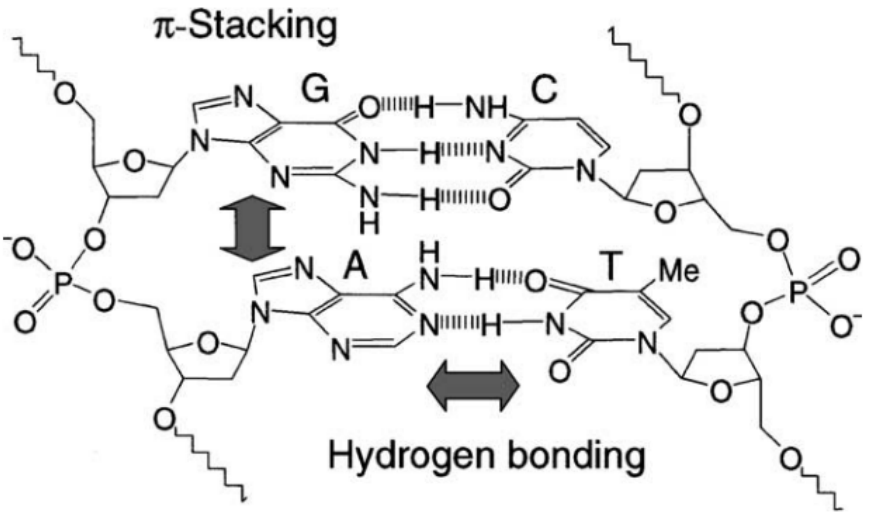


Figure 1.3: Base-stacking interactions and hydrogen bonds between base pairs as a main stabilization force of DNA molecule (Kawai and Majima 2002).

There are many possible conformation of the DNA molecule, while only two forms have been observed in functional organisms: B-DNA and Z-DNA. The most common form of DNA found in cells, is the canonical B-DNA structure. The hydration level, DNA sequence, amount and direction of supercoiling, chemical modifications of the bases, type and concentration of metal ions, and the presence of polyamines in solution are all factors that influence the conformation adopted by DNA.
The A-DNA and B-DNA are both right-handed helixes, with a difference being in shallow, wide minor groove and more narrow, deeper major groove in A-DNA form. Under non-physiological conditions, the A form of DNA can arise in partially dehydrated samples. However, within the cell, it may originate through hybrid pairings of DNA and RNA strands or in enzyme-DNA complexes.

The Z form of DNA has a distorted structure with alternating purine and pyrimidine bases, which causes the backbone of the DNA to twist in a zigzag pattern. Compared to B-DNA, the major groove of Z-DNA is more narrow and more elongated and the minor groove is wider and shallower. The Z-DNA is found in vivo under specific circumstances, such as in regions of DNA with high GC content, or when DNA undergoes torsional strain or negative supercoiling.

1.1.2 Proteins

Proteins are complex biological macromolecules that can be described at various levels: primary, secondary, tertiary, and quarternary. The primary structure of protein consist of linear sequence of amino acids linked together by peptide bond, thus forming a polypeptide chain. This sequence is determined by the genetic information stored in the DNA sequence.

There are 20 different amino acids that can be used as a building blocks for proteins. Amino acids are categorized into different groups based on their physical and chemical characteristics, such as the polarity, charge, or even size and shape. Non-polar amino acids have their side chains composed predominantly of carbon and hydrogen atoms, making them also hydrophobic, apart from nonpolar (alanine, valine, leucine, isoleucine). Polar, uncharged amino acids are composed in their side chains of polar functional groups, such as hydroxyl or amide groups (serine, threonine, asparagine). Positively charged amino acids have side chains that are positively charged at physiological pH (lysine, arginine, histidine, Fig. 1.4). Negatively charged amino acids have, intuitively, side chains negatively charged at the physiological pH (glutamate, aspartate). Aromatic amino acids have aromatic side chains that contribute to the unique structure and function of proteins (phenylalanine, tyrosine, tryptophan). The order and number of amino acids is a unique for each protein, both parameters being important in determination of three-dimensional structure of protein, consequently defining its function and interactions within their biological context.

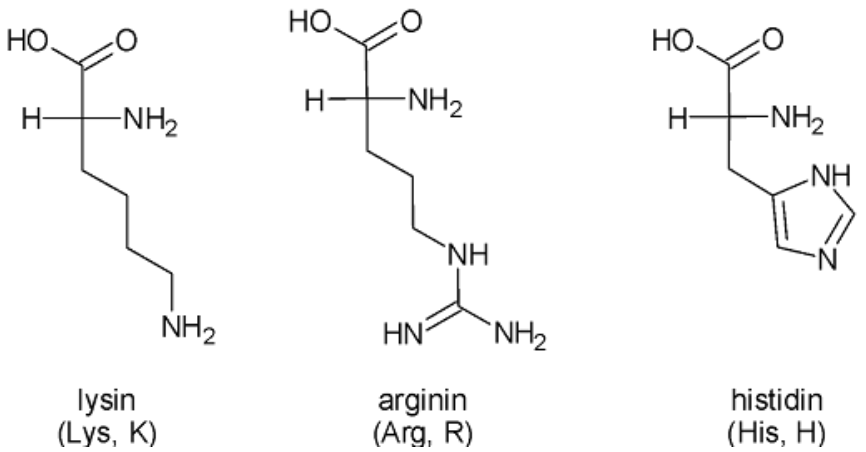


Figure 1.4: Structure of positively charged amino acids: lysine, arginine, histidine, respectively. *(Created in ChemSketch.)*

The secondary structure of a protein refers to the local folding patterns of a single polypeptide chain, which are stabilized by hydrogen bonds between the backbone

atoms of the amino acids. The most common types of secondary structure are β-pleated sheets and α-helices. The main difference is that β-sheet is a flat structure, whereas α-helix is rod-like structure with a spiral shape. β-sheets are formed by hyrogen bonds between the polypeptide chains running in parallel or antiparallel direction to each other. This sheet-like structure can be further classified as either β-strands or β-turns. α-helices are formed by a right-handed coil of the polypeptide chain, where the hydrogen bonds are formed between the amide hydrogen and the carbonyl oxygen of the fourth amino acid ahead of the chain (Fig. 1.5; Hasic et al. 2017). Whether it is α-helix or β-sheet, the secondary structure of protein holds an important contribution to the overall stability and function of a protein, as it also provides the foundation for the higher levels of protein structures.

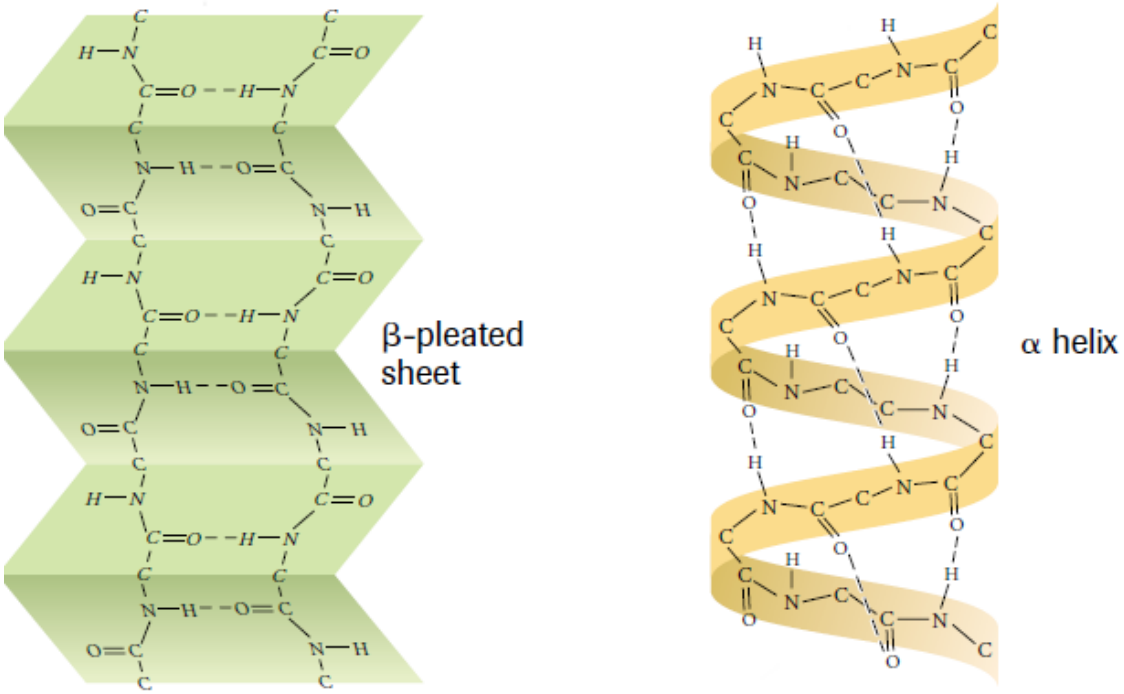


Figure 1.5: Secondary structure of a protein. a) β-pleated sheet. b) α-helix. (Hasic et al. 2017)

The folding of a protein to the tertiary, three-dimensional structure is a critical step in determining its final function. The arrangement of the polypeptide chain, including any folding or bending, occurs due to the interactions between different amino acid side chains. The tertiary structure is stabilized by a variety of interactions, such as a hydrogen bonds, disulfide bonds, van der Waals forces, and hydrophobic interactions. There are two forms of folded proteins: globural and fibrous proteins. Globural proteins are compact, roughly spherical in shape, with their hydrophobic regions buried within the interior of the protein and the hydrophilic

regions exposed to the solvent. Thus, globural proteins are typically soluble in water and play a wide variety of roles in biological processes, such as enzymes, transport proteins, and antibodies. Fibrous proteins are elongated and typically have repeating secondary structures that form long fibres or filaments. Fibrous proteins are thus typically insoluble in water and have structural roles in the body, such as providing support and strength to tissues, for example collagen and keratin.

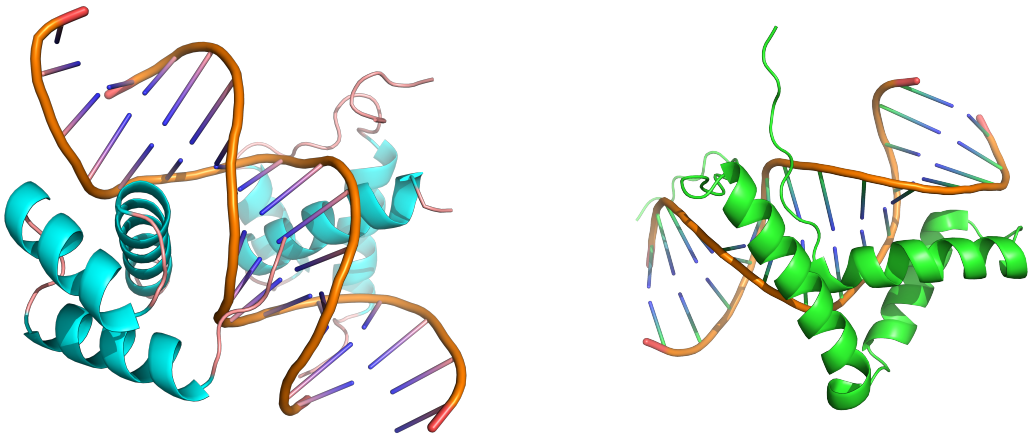
The quaternary structure of protein refers to the arrangement of multiple protein subunits into a larger, functional protein complex. The individual subunits are now not linked by peptide bond, but are held together by various types of non-covalent interactions, such as hydrogen bonds, van der Waals forces, and hydrophobic interactions. Different subunits of protein complex are called heterodimers, equal subunits are homodimers.

1.1.3 Protein-DNA complexes

Complexes formed between proteins and DNA molecules play important roles in a wide range of biological processes, such as gene regulation, DNA replication and DNA repair (Luscombe et al. 2000). There are two types of protein-DNA interactions: specific and non-specific. Specific interactions are based on the ability of certain proteins recognize a specific sequence of nucleotides, allowing them to selectively interact with particular regions of DNA (Rohs et al. 2009). These interactions can lead to changes in DNA conformation, the recruitment of additional proteins, or the modulation of DNA function. Non-specific protein-DNA interactions are based mostly on electrostatic interactions between positively charged amino acid residues in the protein and negatively charged phosphate groups in the DNA backbone, as well as hydrophobic interactions between non-polar amino acid residues and the DNA base pairs. Non-specific interactions can play important role in the overall binding affinity and specificity of a protein for DNA. However, in general, non-specific protein-DNA interactions are weaker then specific interactions.
In this thesis, various protein-DNA complexes were studied.

1B8I

Protein-DNA complex under Protein Data Bank (PDB) code 1B8I, is DNA-bound Ultrabithorax-Extradenticle homeodomain complex (Passner et al. 1999). Homeotic (Hox) genes code important transcription factors in animal development, which govern the choice between alternative developmental pathways along the anterior-posterior axis. Hox proteins have low DNA-biding specificity by themselves but this increases with biding together with the homeoprotein Extradenticle. Thus, complex used in this thesis is that of cooperative heterodimer (Fig. 1.6a).



(a) (b)

Figure 1.6: a) DNA-bound Ultrabithorax-Extradenticle homeodomain complex, PDB: 1B8I b) M9I mutant of the HMG-box domain of the human male sex determining factor SRY complexed to DNA, PDB: 1J47. *(Images generated in PyMol.)*

1J47

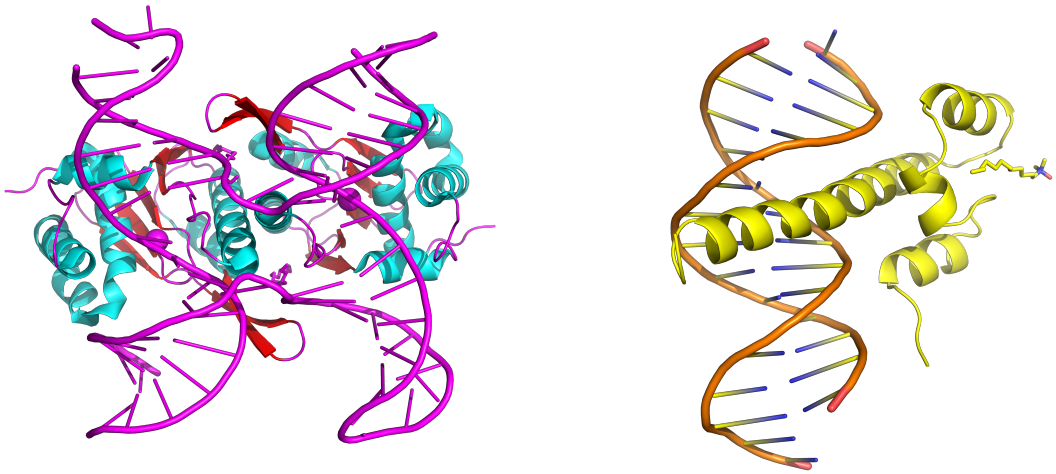
Another simulated complex is that of PDB code 1J47, which is the structure of M9I mutant of the High Mobility Group box (HMG-box) domain of the human male sex-determining region Y (SRY), hSRY(HMG), complexed to DNA (Murphy et al. 2001). The hSRY(HMG) recognizes sequence-specific DNA and binds in the minor groove, resulting in substantial DNA bending. It is shown that the majority of point mutations resulting in 46X, Y sex reversal are located within this domain (Fig. 1.6b).

6IS8

Huge protein-DNA complex, PDB code 6IS8, is a sequence-specific Holliday junction cleavage by MOC1 (Lin et al. 2019). Holliday junction (HJ) plays a critical role as an intermediate during the process of homologous recombination and DNA double-strand break repair. The timely resolution of HJ by resolvases is of utmost importance to maintain the stability of the genome (Fig. 1.7a).

1SKN

The DNA-binding domain of Skn-1, a developmental transcription factor that specifies mesoderm in *C. elegans.*, under the PDB code 1SKN (Rupert et al. 1998), is one of the three complexes simulated by Tomáš Nesvadba in his thesis (2022), the one which this thesis is following up. At the C-terminus, a helix extends from the domain to occupy the major groove of DNA (Fig. 1.7b).



(a) (b)

Figure 1.7: a) Sequence-specific Holliday junction cleavage by MOC1, PDB: 6IS8. b) The DNA-binding domain of Skn-1, a developmental transcription factor that specifies mesoderm in *C. elegans.*, PDB: 1SKN. *(Images generated in PyMol.)*

1MNN

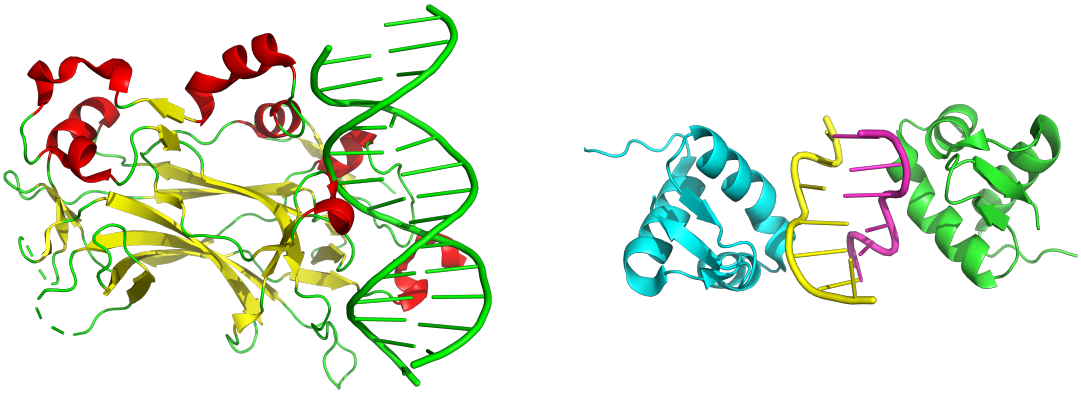
Complex under the PDB code 1MNN, is the complex of sporulation-specific transcription factor Ndt80 bound to DNA (Lamoureux et al. 2022). This protein-DNA complex is activated after successful completion of meiotic recombination in *Saccharomyces cerevisiae* (Fig. 1.8a).

3EYI

The penultimate complex with PDB code 3EYI, is mammalian DNA-dependent activator of IFN-regulatory factors (DAI), which is an activator of the innate immune response (Ha et al. 2008). Two identical protein residues are bind to DNA in Z form (Fig. 1.8b).

1OSL

The last protein-DNA complex is a dimeric lactose DNA-binding domain complexed to a nonspecific DNA sequence (PDB: 1OSL; Kalodimos et al. 2004)). As it was mentioned before, non-specific interactions are mostly of electrostatic character, which makes this complex favorable for studying these electrostatic interactions (Fig. 1.9).



(a) (b)

Figure 1.8: a) Sporulation-specific transcription factor Ndt80 bound to DNA, PDB: 1MNN. b) Mammalian DAI (DNA-dependent activator of IFN-regulatory factors), in a form of two identical protein residues bonded to Z-DNA, PDB: EYI. *(Images generated in PyMol.)*

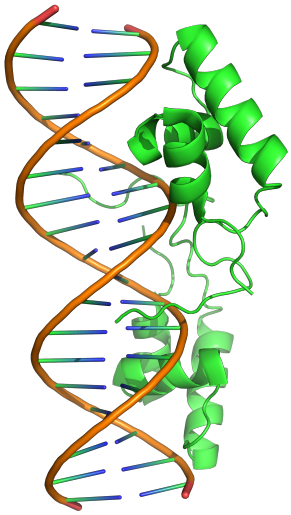


Figure 1.9: Dimeric lactose DNA-binding domain complexed to a nonspecific DNA sequence, PDB: 1OSL. *(Image generated in PyMol.)*

1.2 Molecular Dynamics

Molecular dynamics (MD) is a computational approach that investigates the temporal evolution of atomic and molecular systems by numerically integrating the classical equations of motion. The method is based on the principle that the dynamics of a system can be described by the interactions between its constituent particles and the forces that affect their interactions. The interactions between the particles are represented by a potential energy function, which is derived from quantum mechanical or empirical models. The simulation process initiates with the specification of the initial positions and velocities of the particles. The integration of the equations of motion is performed using numerical algorithms. The output obtained is the time-dependent trajectory of the particles, meanwhile the change of the system's total energy should be negligible. MD simulations are widely applied in various scientific domains, including materials science, biophysics, and chemistry, to gain a deeper understanding of complex systems and to assist in the design of new materials and drugs (Piana et al. 2014, Aranda-Garcia et al. 2022).

1.2.1 Quantum mechanics and molecular mechanics approach

Quantum mechanics (QM) and molecular mechanics (MM) are two distinct but complementary approaches for studying the dynamics of molecules and their interactions. In QM, the behavior of a molecule is described either by the Wave Function Theory (WFT), which is based on Schrödinger wave equation, a partial differential equation describing the evolution of wave function over time, where the information within this wave function includes quantum-mechanical properties of the molecule, such as its energy, spin and probability distribution; or by the Density Functional Theory (DFT), using functionals (function of another function) of the spatially dependent electron density of the system.

In contrast, molecular mechanics uses classical mechanics principles to describe the motion of a molecule. The energy of the system is calculated as a function of the nuclear coordinates (following Born-Oppenheimer approximation of Schrödinger wave equation), ignoring the motion of electrons. Interaction forces between atoms are described by interatomic potentials, often derived from experimental data or ab initio calculations.

The main difference between these two approaches lies in the level of detail and accuracy they provide. Quantum mechanics provides a more accurate description of molecular behavior, however, it is computationally expensive, especially for large

molecules. Molecular mechanics, on the other hand, is computationally less expensive, making it suitable of large-scale simulations. In practice, combination of QM/MM is most effective method for studying chemical processes in solution and in proteins. In this hybrid approach, quantum mechanics is used to describe electronic structure or a small specific site (e.g. enzyme active site), while molecular mechanics is used to describe mechanical behaviour for the rest of the system.

1.3 Force field

A force field is a mathematical representation of interactions between atoms within a molecule and also between molecules, consisting of a functional form and sets of parameters used to calculate the potential energy of the system. The parameters in the force field equations are typically derived from experimental data, such as crystal structures or spectroscopic data, or from quantum mechanical calculations, or both. A common type of force field used in molecular mechanics is the empirical force field, which is based on experimental data and empirical fits to that data. The potential energy of the system is described by a sum of terms, each representing a specific type of interaction between the atoms.

The functional form of a potential energy in molecular mechanics consist of two types of terms describing interactions between atoms: bonded and nonbonded terms.

$$E_{total} = E_{bonded} + E_{nonbonded} \tag{1.1}$$

Bonded terms describe the interactions between atoms that are linked by covalent bonds, and capture the energy associated with changes in bond lengths (bond stretching), angles, and dihedrals (torsions):

$$E_{bonds} = \sum_{bonds} \frac{k_l}{2}(l - l_0)^2, \tag{1.2}$$

where l is the bond length, l_0 is the equilibrium bond length, and k_l is the bond force constant,

$$E_{angles} = \sum_{angles} \frac{k_\theta}{2}(\theta - \theta_0)^2, \tag{1.3}$$

where θ is the angle between the atoms, θ_0 is the equilibrium angle, and k_θ is the angle force constant,

$$E_{dihedrals} = \sum_{dihedrals} \frac{E_n}{2}[1 + \cos(n\Phi - \Phi_1)], \tag{1.4}$$

expressed as Fourier series, where E_n is the height of the energetic barrier, n is the

multiplicity of the torsional term, Φ is dihedral angle and Φ_1 is the phase shift. It is also possible to use other functional forms for the dihedral potential energy, such as polynomial or cosine expansions, depending on the specific requirements of the simulation and the type of system being studied.

Nonbonded terms describe the long-range interactions between atoms, including electrostatic and van der Waals forces. These noncovalent interactions are computationally most intensive. Lennard-Jones potential is often used as a model for van der Waals interactions and electrostatic term is usually computed with Coulomb's law:

$$E_{vdW} = \sum_{i<j} \epsilon \left[\left(\frac{r_{ij}^0}{r_{ij}} \right)^{12} - \left(\frac{r_{ij}^0}{r_{ij}} \right)^{6} \right], \tag{1.5}$$

where ϵ is a well depth, r_{ij}^0 is the equilibrium distance of atoms i and j, and r_{ij} is the distance between given atoms,

$$E_c = \sum_{i<j} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}}, \tag{1.6}$$

where q_i and q_j are atomic charges of atom i and j, respectively, ϵ_0 is vacuum permittivity and r_{ij} is the distance between given atoms.

It is important to note that different force fields may be used for different types of systems, and that different force fields may have different strengths and weaknesses. The choice of a specific force field will depend on the specific requirements of the simulation and the type of system being studied.

1.3.1 Empirical force fields

As stated before, parameter sets for force fields are often empirical. In some cases, the extensive fitting terms are difficult to assign to a physical interpretation. Moreover, force fields use concept of various atom types to address certain properties (geometrical, interaction properties) of the atom. For example, oxygen atom in water molecule is classified as a different force field than the oxygen atom in carbonyl functional group. Thus, the starting point of building a force field, shall be the selection of required atom types. Constants used in equations (1.2) to (1.6) are then acquired from quantum mechanical calculations or from experimental data, such as crystallographic, spectroscopic, or other. Currently, most force fields use a fixed-charge model, which consists of assigning one value for the atomic charge that is not affected by the local electrostatic environment.

Biomacromolecular parameters, such as for proteins, DNA or RNA, were often

derived from understanding the behaviour of small organic molecules, because of accessibility to experimental studies and computative less expensive quantum calculations. However, such approximations bring multiple issues: data from small molecules may not be transferable for larger molecules in terms of atomic charges, polymeric structure, difference between the behaviour of organic molecules in gas phase and the condensed phase, dissimilar experimental conditions, etc. As a consequence, divergent force field parameters have been brought up for biomolecules, including enthalpy of vaporization and sublimation, dipole moments and various spectroscopic parameters (Cornell et al. 1995). Constant room temperature and atmospheric pressure are also one of the chosen parameters to overcome inconsistencies (Lippert).

Empirical force fields have limitations, such as their inability to account for some types of interactions (such as hydrogen bonding) and their dependence on the quality of the parameterization data. Therefore, their accuracy in predicting the properties of a given system depends on the quality of the parameterization and the degree of similarity between the system being studied and the systems used in the parameterization process. The most widely used empirical force fields are the AMBER (Assisted Model Building with Energy Refinement), CHARMM (Chemistry at Harvard Macromolecular Dynamics), GROMACS (GROningen MAchine for Chemical Simulation) and OPLS (Optimized Potentials for Liquid Simulations) force fields.

Another important category in force field parameterization is a water model, since water is an important solvent, but is characterized by its unusual properties. Several water models have been proposed, among which Simple Point-Charge (SPC) and Transferable Intermolecular Potential Three Point (TIP3P) models are a few instances.

In this thesis, AMBER force field was used.

1.3.2 AMBER

AMBER family of force fields was developed for biomacromolecular MD by Peter Kollman's group at the University of California (Case et al. 2005). As mentioned before, AMBER belongs to a group of force fields with the potential energy defined by a functional form. In case of AMBER, the potential energy of the system is

expressed by following functional form:

$$
\begin{aligned}
V(r^N) = & \sum_{i \in bonds} k_{li}(l_i - l_{0i})^2 + \sum_{i \in angles} k_{\theta i}(\theta_i - \theta_{0i})^2 \\
& + \sum_{i \in dihedrals} \sum_n \frac{1}{2} V_i^n [1 + \cos(n\omega_i - \gamma_i)] \\
& + \sum_{j=1}^{N-1} \sum_{i=j+1}^{N} \left\{ \epsilon_{ij} \left[\left(\frac{r_{ij}^0}{r_{ij}} \right)^{12} - 2 \left(\frac{r_{ij}^0}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right\}
\end{aligned}
\tag{1.7}
$$

The first term (sum over bonds) is the representation of the energy between covalently bonded atoms. This energy is modeled by a harmonic (ideal spring) force, which is a sufficient approximation when atoms are near their equilibrium bond length. However, as the distance between atoms increases, this approximation becomes less and less reliable.

The second term (sum over angles) is the representation of the energy resulting from the geometry of electron orbitals that participate in covalent bonding.

The third term (sum over dihedrals or torsions) is the representation of the energy required to twist a bond, which is influenced by bond order (such as in double bonds) as well as neighboring bonds or lone pairs of electrons. It's possible for one bond to have multiple dihedral terms, which results in the total dihedral energy being expressed as a Fourier series.

The fourth term (double sum over i and j) is the representation of the non-bonded energy between all atom pairs. This energy can be separated into two components: van der Waals energy (first term of summation) and electrostatic energy (second term of summation).

The van der Waals energy is determined by using the fourth term that contains the equilibrium distance (r_0) and well depth (ϵ). In some cases, the energy equation is rewritten using $\sigma = \left(\frac{r_0}{2}\right)^{\frac{1}{6}}$, which is used in softcore potential implementations.

The electrostatic energy equation assumes that the charges from an atom's protons and electrons can be simplified to a single point charge, or a small number of point charges in the case of parameter sets that include lone pairs. Point charges, which are some of the most important force field parameters, can be obtained using various procedures. In AMBER, For instance, the Restrained Electrostatic Potential (RESP) methodology is a straightforward and reproducible approach that has been shown to provide well-behaved atomic partial charges and is therefore used in AMBER (Bayly et al. 1993).

For AMBER force field, various sets of parameters exist for certain types of

molecules. In this thesis, the parameter set named ff14SB was used for proteins and OL21 for nucleic acid simulations (Maier et al. 2015).

Since the initial release in 2002, many corrections and modifications to parameter sets were developed. For instance, canonical B-DNA is described relatively accurately, but when it comes to noncanonical structures, such as ones found in protein-DNA complexes, these are often described incorrectly. It is shown that dihedral angle parameters α/γ are crucial for description of the conformational equilibria involving nucleic acids. One of these modifications is known as OL21, named after the city of Olomouc (Zgarbová et al. 2021). It improves the stability of native α/γ Z-DNA substates while the canonical DNA description is kept unchanged. OL21 force field is derived from previous version OL15 (Zgarbová et al. 2015), which is based on top of ff99+bsc0 force field. It contains refinement of glycosidic dihedral (χ_{OL4}), epsilon/zeta modification (ϵ/ζ_{OL1}) and beta dihedral (β_{OL1}) for DNA simulations.

In current days, OL21 and parmbsc1 are two force fields commonly used in MD simulations of biomolecules. They vary in several parameters, such as description of atomic charges: while OL21 uses atomic charges based on RESP fitting, parmbsc1 uses atomic charges based on a high-level quantum mechanical data (Ivani et al. 2016).

1.3.3 Deficiencies of current empirical potentials

In latest years, parameters used in MD simulations of multi-component protein, nucleic acid and lipid systems were observed to be overestimated, specially in relation to cation-anion attraction (Yoo and Aksimentiev 2012). An artificial aggregation of simulated biological systems can be seen as a result of these overestimated attractive interactions between charged and hydrophobic groups. Denaturated conformations are particularly affected by the improper parameterization of ion pairs, as a consequence of force fields being calibrated to reproduce the properties of folded biomolecules.

There are currently various approaches to adjust this issue. One of the alternatives is correction of pair-specific Lennard-Jones (LJ) parameters with reference to the experimental data while parameters for solute-water interactions remain intact (You et al. 2020). This modification of non-bonded parameters is referred to as NBFIX (Non-Bonded FIX) (Yoo and Aksimentiev 2016). NBFIX adjusts all selected pairwise LJ interactions of the atom types in the force field by surpassing Lorentz-Berthelot combining rules. This correction may have an impact on the behaviour of water molecules nearby solutes, even though it does not explicitly modify the solute-water interactions. The proportion of contact ion pairs is controlled by

the ability of water molecules to mediate the interactions between solutes. The solute-solute interactions are calibrated using osmotic pressure experimental data. In this thesis, a specific variant of NBFIX modification called CUFIX was tested. It is a systematic refinement of LJ parameters describing amine-carboxylate, amine-phosphate and aliphatic carbon-carbon interactions. This refinement improves the accuracy of MD simulations of proteins, nucleic acids and lipid, resulting in notably improved agreement with experiments. However, CUFIX is optimized with the TIP3P water model, which nowadays is not considered the best option within water models (Jorgensen et al. 1983), which may complicate its wider use for MD simulations (Yoo and Aksimentiev 2018).

Another option within modification alternatives is to modify bio-organic phosphates. To obtain better balanced electrostatic interactions between water and the phosphate oxygen (solvation energy), Case group modified (increased) the van der Waals phosphate oxygen radii (Steinbrecher et al. 2012). The magnified radii indirectly reduce the electrostatic interaction with the cations, because the average interaction distance in the Coulomb formula is increased. However, it is not straightforward to assume that this weakening of the phosphate-cation interaction will translate to reduced association strength of phosphates with cations in solution, because also the phosphate-water interaction is influenced. Therefore, the effect of modified phosphate radii on protein-DNA interactions requires further investigation.

Chapter 2

Aim of the thesis

The aim of this thesis was to assemble a set of protein-DNA complexes suitable for testing non-covalent interactions described in the empirical potentials. Additionally, the aim was also to test the original parameters versus the modified versions, such as CUFIX or phosphate modification. Ultimately, it was also important to evaluate the assembled set of complexes and the method and its relevance for testing the accuracy of the description of intermolecular electrostatic interactions within various force field variants.

Chapter 3

Material and methods

Protein-DNA complexes that were suitable for MD simulation were chosen from the Protein DataBank (PDB, RCSB), using specific criteria. Decision regarding selection of complexes were based on having a high amount of electrostatic contacts (salt bridges) between protein and DNA molecule, high resolution structure (less than 2.5 Å) and the absence of excess ions or atoms in experimental setup that could not be described using the atom types of the force field being used, nor could be easily removed from the structure. Based on the said criteria, complexes with PDB codes 1B8I, 1J47, 1MNN, 1OSL, 1SKN, 3EYI and 6IS8 were chosen for the MD simulations.

Original PDB files were adjusted as needed. Structures of chosen complexes were obtained either by X-ray crystallography (X-ray diffraction, XRD) or Nuclear Magnetic Resonance (NMR) spectroscopy, therefore unnecessary atoms such as those of crystallization agents or redundant atoms present in the structure were eliminated for the simulation purposes. One of the first moves was the inspection of histidine (His) form in the crystal structure. Namely, His residues in proteins possess the ability to adopt three different protonation states depending on pH, which presents an ongoing challenge when adding protons to a protein crystal structure for MD simulations. AMBER distinguishes three His forms (illustrated in the Figure 3.1):

- HID - hydrogen atom located on the δ nitrogen atom of imidazole group,
- HIE - hydrogen atom located on the ϵ nitrogen atom,
- HIP - hydrogen atoms located on both nitrogen atoms - positively charged residue.

To determine the protonation state of His residues, the H++ internet server which provides outputs compatible with AMBER format can be used in combination with visual inspection of the geometry in PyMOL (Schrödinger et al. 2020). Specific form of His residues was determined for complexes 1B8I, 1MNN and 6IS8. Individual pH

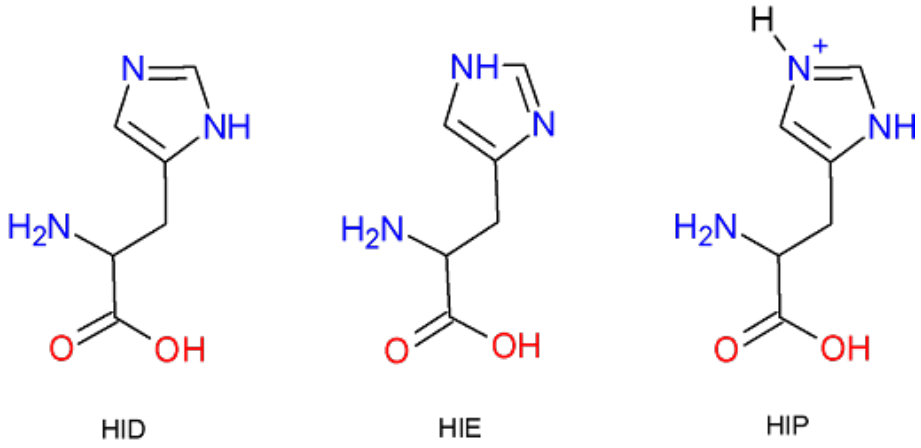


Figure 3.1: Protonation state of histidine residues depending on specific pH. *(Created in ChemSketch.)*

values for each complex were acquired from the experimental data corresponding to those crystal structures in PDB.

Another step in preparation of the structures for simulation was to neutralize charge of DNA backbone and protein by adding certain amount of K^+/Na^+ and Cl^- ions in such a manner that the concentration of surrounding environment was 0.15 M which roughly corresponds to the cell environment. Joung-Cheatham parameters were used for monovalent ions. Experimental values, such as pH, method for obtaining the structure of complexes, resolution of the structures and the type of added ions (solution) are presented in the Table 3.1.

Table 3.1: **Experimental pH values and methods by which the PDB structures of simulated complexes were obtained, and the ions added (solution) for the simulation purposes.**

	1B8I	1J47	1MNN	1OSL	1SKN	3EYI	6IS8
pH	4.5	-	7.0	-	6.0	8.5	7.5
method	XRD	NMR	XRD	NMR	XRD	XRD	XRD
solution	KCl	KCl	KCl	KCl	KCl	NaCl	NaCl
resolution [Å]	2.40	-	1.40	-	2.50	1.45	1.68

Special treatment was received by the complex 1OSL and 6IS8. The PDB structure of 1OSL was in need of alteration, resulting from the fact that it consisted of separated models, therefore it was necessary to establish a disulfide link between those models. Complication of different sort was encountered with the structure 6IS8, as it contained magnesium atom in its crystal structure. For this reason additional parameters defining magnesium atom type interactions were loaded into the LEaP script (discussed further later in the following text).

The accuracy of MD simulations performed on biological systems is yielded also by the implementation of isothermal-isobaric ensemble, which is referred to as NPT ensemble. It is a statistical mechanical ensemble that maintains constant number of particles, as well as constant temperature and pressure applied. Room temperature (298 K, 25 °C) together with atmospheric pressure (1 atm, 101.325 kPa) are typical values used in MD.

Before running the simulation, the AMBER input files (top and crd) were prepared using tLEaP program starting form adjusted PDB files of each complex. The input of tLEaP was customized for each complex by specifically determining solute ions, that were added based on the experimental setup provided by the database information, as it was mentioned before. To allow larger time step (4 fs) to be used, hydrogen mass repartitioning (HMR) was performed using the parmed package. HMR provides redistribution of mass from heavy atoms that are connected to hydrogen atoms into the hydrogen bonds, which enables the accuracy of the simulation to be preserved for longer time steps, so that it would not encounter stability-related errors caused by high-frequency hydrogen motion.The solvation of the system is ensured by using a truncated octahedral box of a size such that the peripheral atoms of the system are at a distance of at least 10 Å from the box edge. The box is filled with water described by the SPC/E water model. The output of the LEaP script is topology and coordinate file. The generated coordinates reflect the given structure after clearing any incomplete residues or constructing assigned tasks (such as aforementioned creation of disulfide link). The corresponding topology provides a comprehensive description of the system's behaviour. Nevertheless, due to the complexity of the system, it is unlikely that analytic solutions can be obtained, thus simulation is still a necessary step. Prior to taking that action, energy minimization and equilibration of the box with the included system take place.

The very process of MD simulations was performed in AMBER using ff14SB force field for protein and OL21 force field, which in the present day is a recommended force field for protein and nucleic acid simulations. The primary engine for running equilibration and MD simulation in AMBER is Particle Mesh Ewald Molecular Dynamics (PMEMD), which uses GPU and the trajectory is processed and analyzed by the program CPPTRAJ. Average time span of simulation running was 4 days. Visual analyses was performed in the program Visual Molecular Dynamics (VMD) (Humphrey et al. 1996). The storing of the coordinates was done every 10 ps and the total length of the simulation was 1 μs. The water molecules were then removed from the trajectory (stripped) and after that the resulting trajectory was pruned taking every hundredth sample, and thus making the final time step of 1 ns. The original and the pruned trajectories were used for different stages of the analysis

process.

To achieve the objectives of this thesis, each complex was simulated using the modified force field known as CUFIX. The CUFIX modified force field was downloaded from The Aksimentiev Group as amber14sb_OL15_cufix.ff file. The parameters files had to be modified to include OL21 corrections. As the CUFIX modification is designed for use with the TIP3P water model, it was necessary to perform comparative simulations of at least one complex using this water model with OL21 simulation without CUFIX. Additionally, the effect of phosphate radii modification was also investigated in this thesis on each complex, in combination with OL21 force field.

To study the electrostatic interactions in the protein-DNA complexes, it was necessary to determine the types of contacts that should be taken into account. The distance within which it would be considered as the contact between DNA phosphate and the arginine or lysine residue of the protein, was set to 7.2 Å, which included water mediated contacts. As native contacts, the contacts between the phosphate atoms of DNA backbone labeled OP1 or OP2, and the arginine nitrogen NE, NH1, NH2 or CZ, as well as the lysine nitrogen NZ, were considered.

In-house scripts were developed to sort, measure and process the contacts. Firstly, the contacts were categorized into unique directories based on the specific phosphate group involved in the interaction. Then, subdirectories were created for each phosphate group, based on the cationic residue that interacted with that particular phosphate. Initially, there may have been one or more cations interacting with each phosphate in the initial structure, but it was important to consider the possibility that these initial contacts may be lost during the MD simulation and new contacts may form with different cations. In the final step, the distances between all polar non-hydrogen atoms of the phosphate group (i.e. OP1, OP2, O5', O3') and either lysine (NZ) or arginine (NE, NH1, NH2 or CZ) were measured throughout the MD simulation for each phosphate-cation contact. Since the interacting polar groups were highly mobile and the geometry of their interaction changed during the simulation, the closest contact between the two residues was identified for each frame of the MD simulation. These shortest contacts were then processed as a histogram, both for each residue separately and for the entire complex (with individual histograms merged into one that represented all phosphate-cation pairs in that complex, separately for arginine and lysine). The discussion in the following section focuses on these histograms.

Chapter 4

Results and discussion

4.1 Stability of simulated complexes

Stability of MD simulations can be judged from Root-Mean-Square Deviation (RMSD), which is a metric for measuring the position of atoms in simulated complexes against the original structure (in this case crystallographic and NMR structure). RMSD was calculated for each complex as a whole, as well as for its individual components, DNA and protein, in order to obtain better understanding of the stability of each part. Lower RMSD values stand for more stable complex (higher agreement with the original structure). RMSD values of complexes simulated in the OL21 force field without CUFIX nor phosphate modifications are presented in following text.

Due to the high RMSD for the protein and the overall calculation of complex of the 1OSL (Fig. 4.1), which indicated a significant disagreement with the original NMR structure, it was decided to abandon further analysis of this complex.
The rest of the simulated complexes were reasonably stable during MD simulation, as judged by their RMSD values.

As it can be seen on the Figure 4.2a, DNA in the 1B8I complex was considerably stable, as its RMSD value was most of the simulation under 1.5 Å. RMSD values of the protein ranged between 2 and 3 Å (with some minor exceptions), as well as the values of the whole complex, which is also considered as a quite stable movement during the simulation. Slightly increased RMSD of protein and also whole complex after 650 ns may be caused due to low definition of some protein parts in original structure.

At first glance, the RMSD values for complex 1J47 may appear alarming, as all components exhibited high values (Fig. 4.2b), but upon closer inspection of the simulation, the reason for these high values became apparent. The protein structure includes a long free end that moved considerably during the simulation, while the main core of the protein-DNA complex remained relatively stable. The DNA RMSD

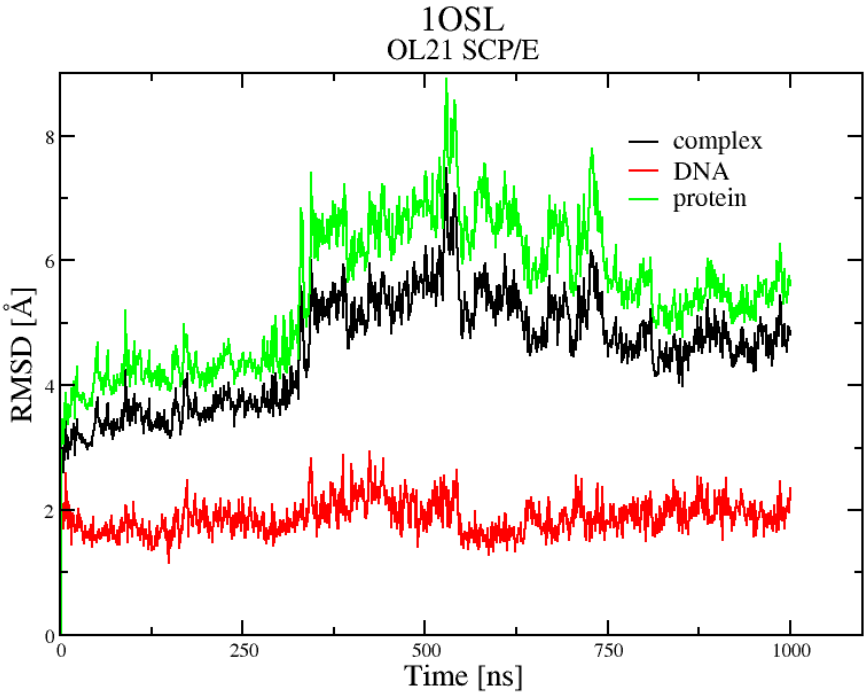
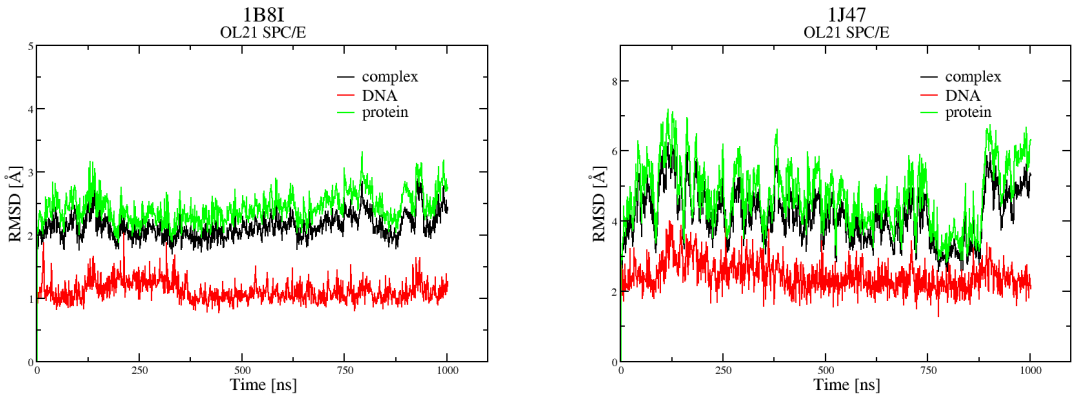


Figure 4.1: RMSD of whole complex 1OSL (black), as well as for individual components: DNA (red) and protein (green).

was reasonably stable, staying under 2.5 Å during the simulation, with occasional increases to over 3 Å at approximately 250 ns, when it made contact with the free protein end. However, it quickly returned to its former state and remained stable until the end of the simulation, which further demonstrates the tolerable stability of the complex.

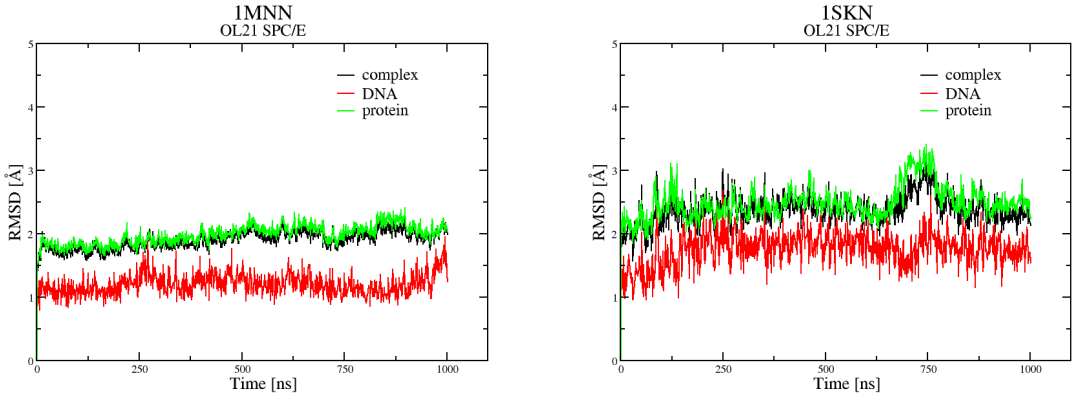
Simulation of the complex 1MNN was one of the most stable simulations conducted for this thesis (presented in Figure 4.3a). The RMSD value for the DNA was stable around 1 Å through the simulation, with very slight increase at the end, but only to 1.5 Å, which is still very respectable score. Similarly, the protein and overall RMSD were stable during the simulation, initially starting above 1.5 Å and then slightly increasing until they both gained a stable state around 2 Å.

The RMSD of complex 1SKN is presented in Figure 4.3b. The RMSD of DNA can be seen slightly increasing from the beginning until approximately 250 ns, from which point it stabilized at between 1.5 and 2 Å. The RMSD of the protein and the complex had a slightly similar trend, starting at 2 Å and holding a stable state under 2.5 Å. A notable peak appeared with its highest point at 750 ns. It was caused by the unpaired base end of the DNA, which at this moment flipped over and made contact with the protein. Overall, the RMSD did not exceed much over 3 Å (and that also did not occur often during the simulation), which is considered as a stable



(a) (b)

Figure 4.2: RMSD of whole complex (black), as well as for individual components: DNA (red) and protein (green) a) 1B8I b) 1J47.



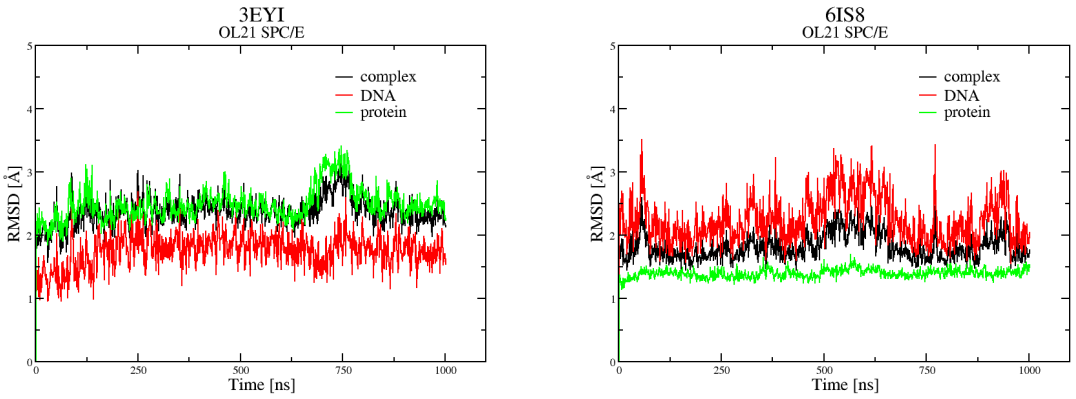
(a) (b)

Figure 4.3: RMSD of whole complex (black), as well as for individual components: DNA (red) and protein (green) a) 1MNN b) 1SKN.

simulation.

Complex 3EYI consists of DNA sandwiched between two protein chains. Each protein chain reacted with the DNA molecule slightly differently, and also had various free ends that moved significantly throughout the simulation, which can be seen in oscillating RMSD values that peaked at around 4 and 5 Å. However, the DNA remained stable, with RMSD ranging between 1 and 1.5 Å (Fig. 4.4a). Protein-DNA complex 6IS8 is a huge and complicated structure. The protein RMSD was very stable during simulation, with an RMSD value under 1.5 Å. Overall complex was stable at the RMSD values between 1.5 and 2 Å, with a slight increase around 550 ns, followed by a decrease to a stable state (Fig. 4.4b). It may seem that the DNA molecule had high RMSD values, but it was due to four end with unpaired bases in the structure, which could freely fluctuate. However, the core of the DNA and also the whole complex was reasonable stable during the simulation.

In the Figure 4.5 are presented RMSD of all complexes simulated in the OL21



(a) (b)

Figure 4.4: RMSD of whole complex (black), as well as for individual components: DNA (red) and protein (green) a) 3EYI b) 6IS8.

force field with CUFIX modification (on the left side) and in the OL21 force field with phosphate modification (on the right side). Overall, simulations in the OL21 force field with phosphate modification appeared to be more stable compared to those in the OL21 force field with CUFIX modification and also to those in OL21 without any further modification.

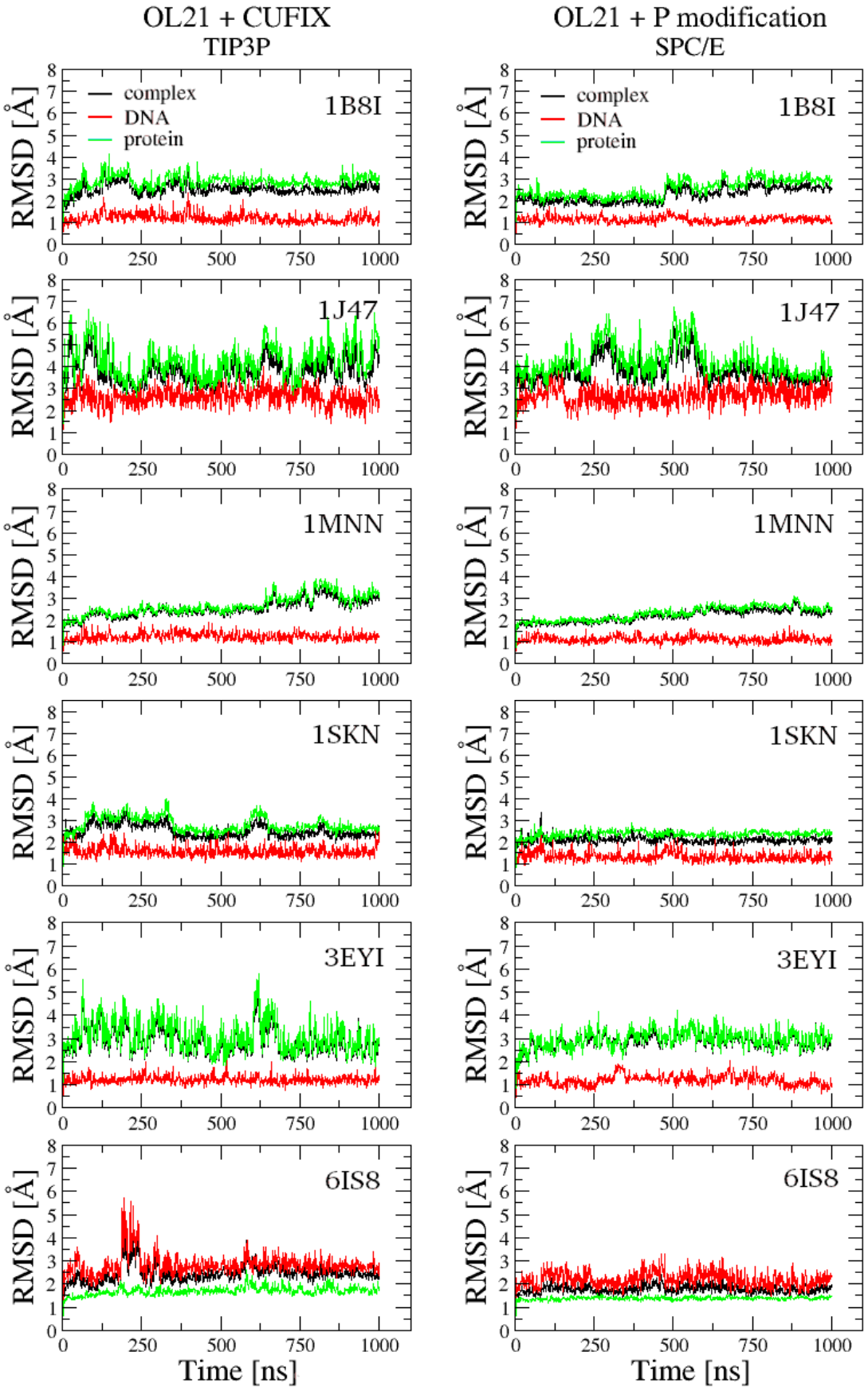


Figure 4.5: RMSD of simulated complexes in OL21 force field with CUFIX modification (left) and in OL21 force field with phosphate modification (right).

4.2 Protein-DNA interactions

After assembling a suitable set of protein-DNA complexes (ultimately based on RMSD), the focus was then shifted to studying electrostatic interactions between DNA phosphate atoms and arginine and lysine residues of the proteins within the used complexes. As stated before, in-house scripts were developed as needed to analyse the data. The number of DNA phosphates interacting with cations of protein residues was counted per each complex, as well as the total number of phosphate contacts made with proteins during the simulation, and the count of contacts made specifically with arginine or lysine residues. The total count of interacting DNA phosphates was 123 phosphate atoms from all 6 analysed complexes. The total count of phosphate contacts made with arginine or lysine residues was 231, with 128 contacts made with arginine residue and 103 contacts made with lysine residue. The results are listed in the Table 4.1.

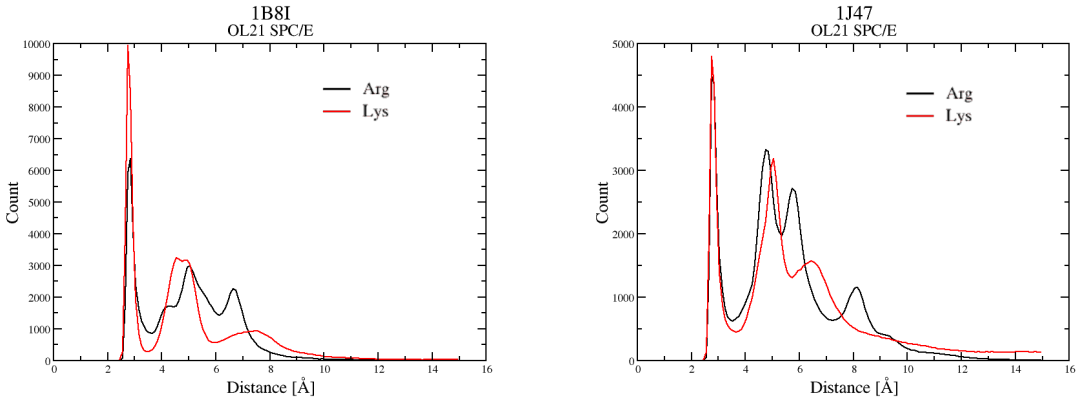
Table 4.1: **Contacts count of all DNA phosphate atoms in each complex interacting with cations of protein residues.** P - phosphate count in the complex interacting with cations, n - all phosphate contacts count, ARG and LYS - contacts count of phosphate interacting with cations of arginine and lysine, respectively.

	P	n	ARG	LYS
1B8I	21	38	23	15
1J47	22	41	24	17
1MNN	14	35	21	14
1SKN	16	35	26	9
3EYI	13	27	11	16
6IS8	37	55	23	32
SUM	123	231	128	103

In the Chapter 3 of this thesis, it was established that only those contacts within a distance of 7.2 Å, which included water-mediated contacts, were considered. The histograms presented in this section illustrate the shortest phosphate contacts with arginine or lysine residues, both at the beginning of the simulation and further during the simulation. However, it should be noted that contacts within the distance longer than 7.2 Å, which are visible in the histograms, are actually the initial short contacts that were tracked during the simulation, but gained a longer distance and were no longer in contact with the phosphate.

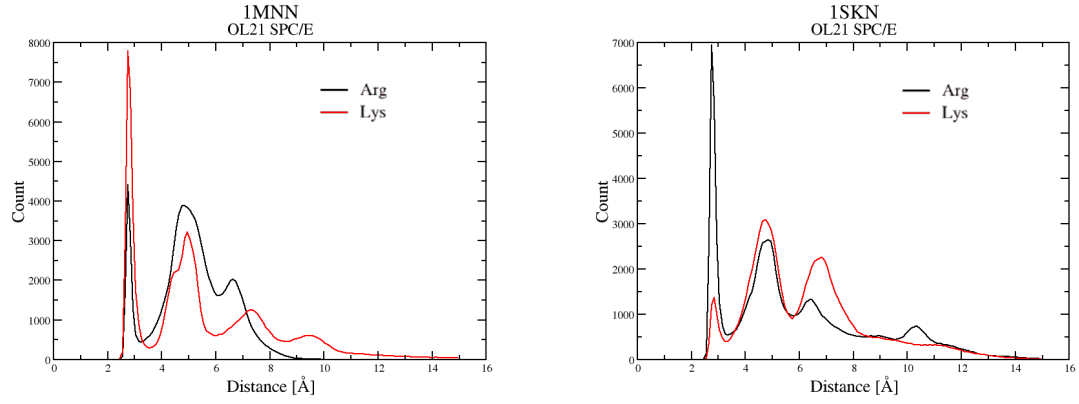
4.3 Simulations in OL21 force field

The first peak in all histograms represents count of direct contacts of phosphates with arginine or lysine residues at around 2.8 Å on average (this distance can also be manually measured in the original PDB structure visualized in PyMOL). In the complex 1B8I (Fig. 4.6a), most water-mediated contacts with arginine were made at the distance of 5 Å. Direct contacts of lysine were preferred over the water-mediated contacts in higher proportion than in the case of arginine in this complex. Also, most of the water-mediated contacts with lysine were made at a slightly shorter distance. In the complex 1J47 (Fig. 4.6b), both arginine and lysine made direct and water-mediated contacts with phosphates in similar ratio, with water-mediated contacts at the distance at around 5 Å.



(a) (b)

Figure 4.6: Histograms of phosphate contacts with arginine and lysine residue in the complex 1B8I (a) and the complex 1J47 (b). The first peak represents direct contacts, whereas water-mediated contacts were made at the distance around 5 Å.

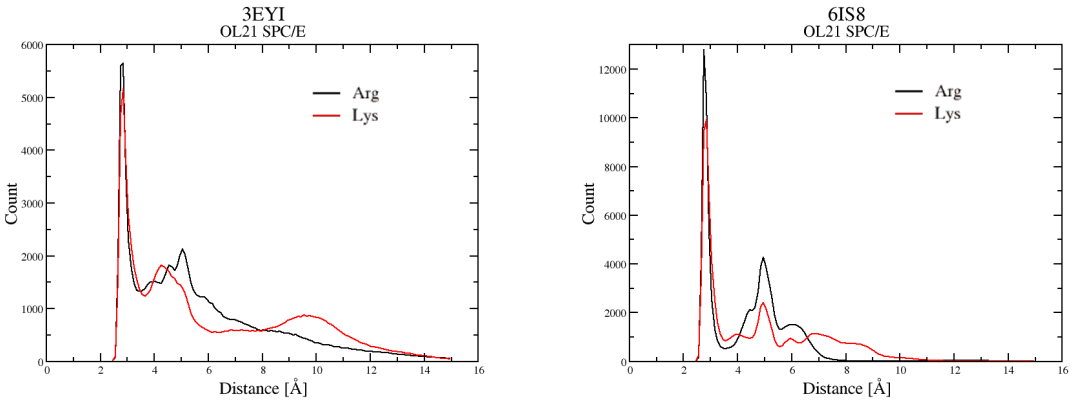


(a) (b)

Figure 4.7: Histograms of phosphate contacts with arginine and lysine residue in the complex 1MNN (a) and the complex 1SKN (b). The first peak represents direct contacts, whereas water-mediated contacts were made at the distance around 5 Å in 1MNN and slightly under 5 Å in 1SKN.

Arginine residue in the complex 1MNN made direct and water-mediated contacts at almost the same ratio. For lysine residue in this complex, the direct contacts were far more preferred than the water-mediated contacts. For both residues, water-mediated contacts were made at around 5 Å (Fig. 4.7a). In the complex 1SKN, the preference for direct contacts over the water-mediated were very obvious for arginine residue (Fig. 4.7b). Interestingly, lysine residue in this complex preferred water-mediated contacts much more over direct contacts. Both residues made most water-mediated contacts at the distance under 5 Å.

In the complex 3EYI, both protein residues preferred direct contacts with phosphate. Most lysine water-mediated contacts were made at the distance under 4.5 Å, whereas arginine water-mediated contacts were preferred at a greater distance (Fig. 4.8a). Both residues in the complex 6IS8 much more preferred direct contacts than the water-mediated, which were then made at around 5 Å (Fig. 4.8b).



(a) (b)

Figure 4.8: Histograms of phosphate contacts with arginine and lysine residue in the complex 3EYI (a) and the complex 6IS8 (b). The first peak represents direct contacts, whereas water-mediated contacts were made at the distance 4-5 Åin 3EYI and at around 5 Åin 6IS8.

It is worth mention that the variations observed in the histograms of the described complexes indicate individual differences in the bonding situations within these complexes. The relatively significant variations suggest the importance of gathering a larger set of complexes, where individual differences could be averaged out and statistical averages could more accurately represent the effects of force field modifications.

4.4 Simulations in OL21 force field with CUFIX modification

The histograms in Figure 4.9 represent the count of contacts between arginine (left) and lysine (right) residues with the phosphate group at various distances during the simulation. The black line represent residues simulated in the OL21 force field with the SPC/E water model, while the red line represents residues simulated in the OL21 force field with the TIP3P water model and CUFIX modification. It is evident from the histograms that the CUFIX modification had a significant impact on all of the simulated complexes. This modification weakens electrostatic interactions, resulting in fewer direct contacts between the phosphate group and arginine or lysine residues. However, the preference for water-mediated contacts remained similar to the simulations without CUFIX modification.

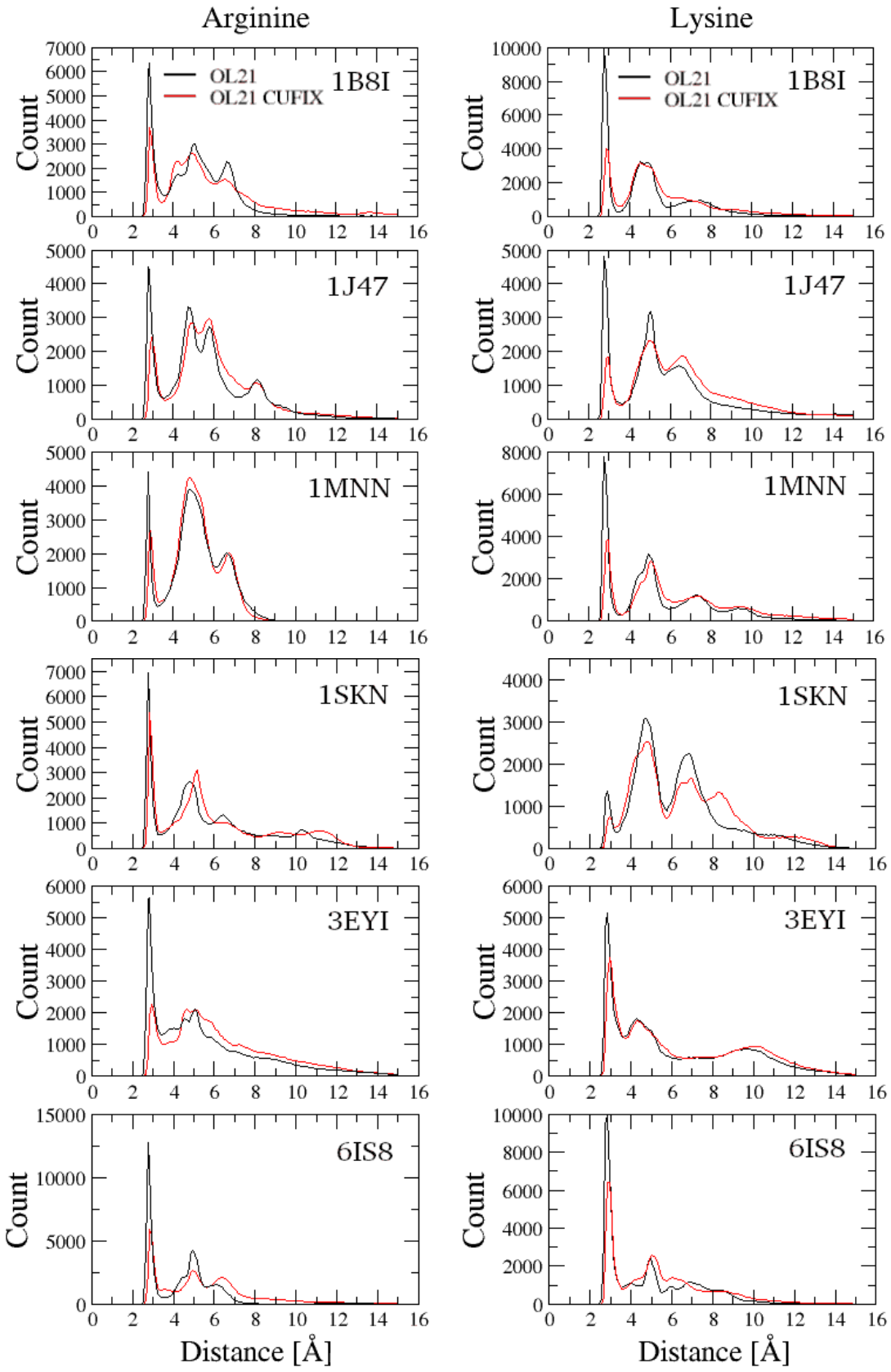


Figure 4.9: Histograms of arginine (left) and lysine (right) counts of contacts with phosphate at the certain distance through the simulation in OL21 force field without any modification (black) and with CUFIX modification (red).

4.5 Effect of water model

At this point, a question arises regarding the effect of using different water models for the simulations. Since the CUFIX modification was designed specifically for use with the TIP3P water model, it was necessary to look into the impact of using different water models on the protein-DNA interactions, specifically into the effect on the phosphate contacts with arginine or lysine residues. To address this question, it was decided to perform another MD simulation of the complex 1B8I, this time using the OL21 force field with the TIP3P water model and without any further modifications.

The comparison of the results of the simulations using different water models started with reviewing RMSD of the complex 1B8I in each simulation. The Figure 4.10 illustrates the comparison of the overall RMSD of the complex 1B8I simulated in the OL21 force field with the SPC/E water model, in the OL21 force field with the TIP3P water model, and in the OL21 force field with the CUFIX modification and the TIP3P water model. As shown in the Figure 4.11, the effect of using different water models on the contacts between the phosphate group and the arginine or lysine residues was found to be at the threshold of statistical significance. The histograms of the contacts count at different distances for the different simulations showed that the direct and water-mediated contacts were both affected by the choice of water model, but the differences were not large enough to be statistically significant. Therefore, it can be concluded that the effect of using different water models on the protein-DNA interactions in this case was minimal, and that the effect of CUFIX is truly significant in terms of modifying the electrostatic interactions between DNA phosphates and protein positively charged residues.

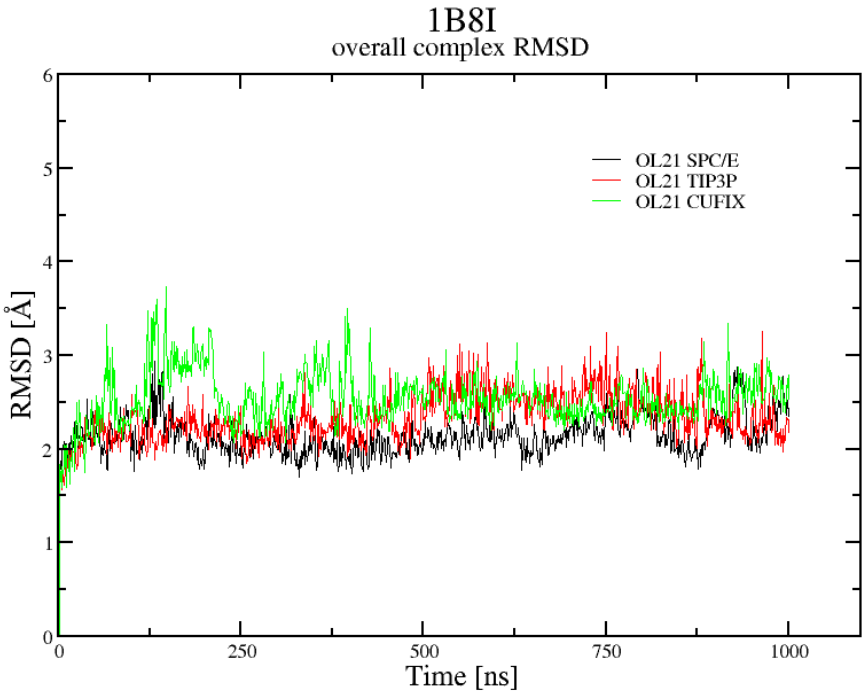
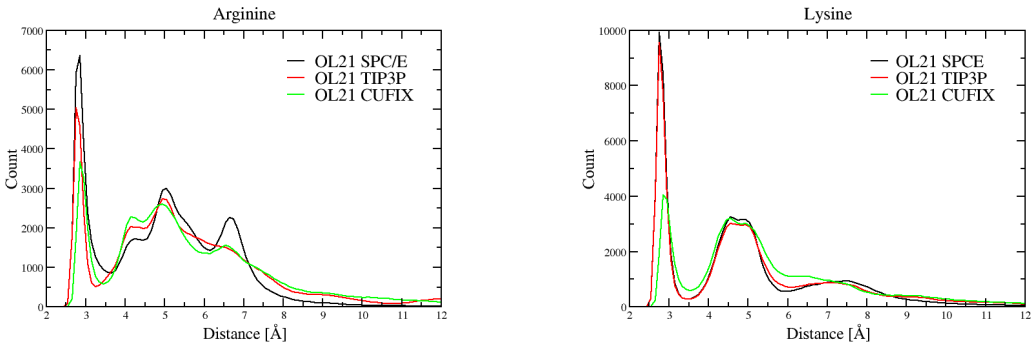


Figure 4.10: Comparison of overall RMSD for the complex 1B8I simulated in OL21 with SPC/E water model (black), in OL21 with TIP3P water model (red) and in OL21 with CUFIX modification with TIP3P water model (green).



(a) (b)

Figure 4.11: Comparison of histograms of arginine and lysine residues derived from simulations with different parameters: in OL21 with SPC/E water model, in OL21 with TIP3P water model and in OL21 with CUFIX modification and TIP3P water model.

4.6 Simulations in OL21 with phosphate modification

As mentioned earlier in this thesis, the CUFIX modification was specifically designed for use with the TIP3P water model, and it represents a step towards greater accuracy in MD simulations by modifying electrostatic interactions in force fields. However, despite its usefulness in weakening electrostatic interactions, in this case resulting in reduced number of direct contacts between phosphate and arginine or lysine residues, there remains a need for further systematic modifications to force fields. To this end, the histograms shown in Figure 4.12 provide a useful tool for visualizing the effects of different modifications on arginine and lysine residues in protein-DNA complexes. Specifically, each complex was simulated using the OL21 force field with no modifications (represented by the black line), with the CUFIX modification (represented by the red line), and with a phosphate modification (represented by the green line). As the histograms clearly demonstrate, the CUFIX modification has a significant impact on direct contacts between the phosphate group and arginine or lysine residues. On the other hand, the phosphate modification appears to have only a small and inconsistent effect on LJ parameters, and it even has the unexpected effect of increasing some contacts rather than weakening them. While the hypothesis behind the phosphate modification was sound in theory, the results suggest that it is not a viable alternative to CUFIX for improving force fields in simulations of protein-DNA interactions.

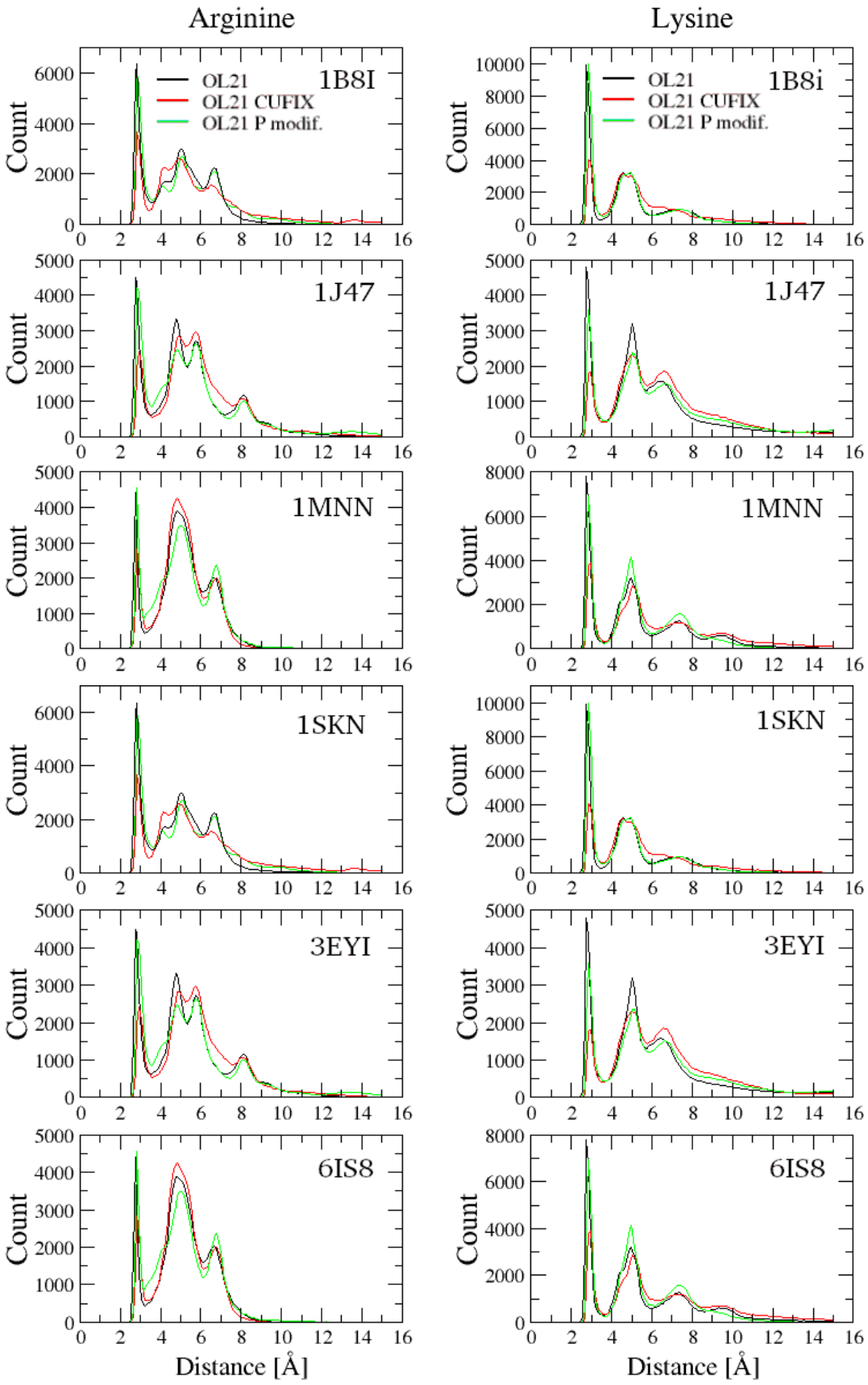


Figure 4.12: Histograms of arginine (left) and lysine (right) counts of contacts with phosphate at the certain distance through the simulation in OL21 force field without any modification, with CUFIX (red) and with phosphate modification (green).

4.7 Summary

The overall effect of phosphate modification on contacts with arginine and lysine residues of all simulated protein-DNA complexes in comparison with CUFIX effect to the force field without further modification can be observed in the Figure 4.13 for arginine and in the Figure 4.14 for lysine. The data derived from all histograms together demonstrate the general preference of direct phosphate contacts with arginine or lysine residues at a distance of around 2.8 Å. Water-mediated contacts appear to be formed at a distance of around 5 Å, which is consistent with the size of a water molecule. The trend of direct versus water-mediated contacts remains consistent across all the simulated complexes, even when phosphate modification or CUFIX is applied. Despite the fact that the phosphate modification consists of alternating the van der Waals radii, which should affect electrostatic interactions in weakened way, the overall histogram shows different result. Nonetheless, it is worth noting a minor shift in the overall effect of the phosphate modification.

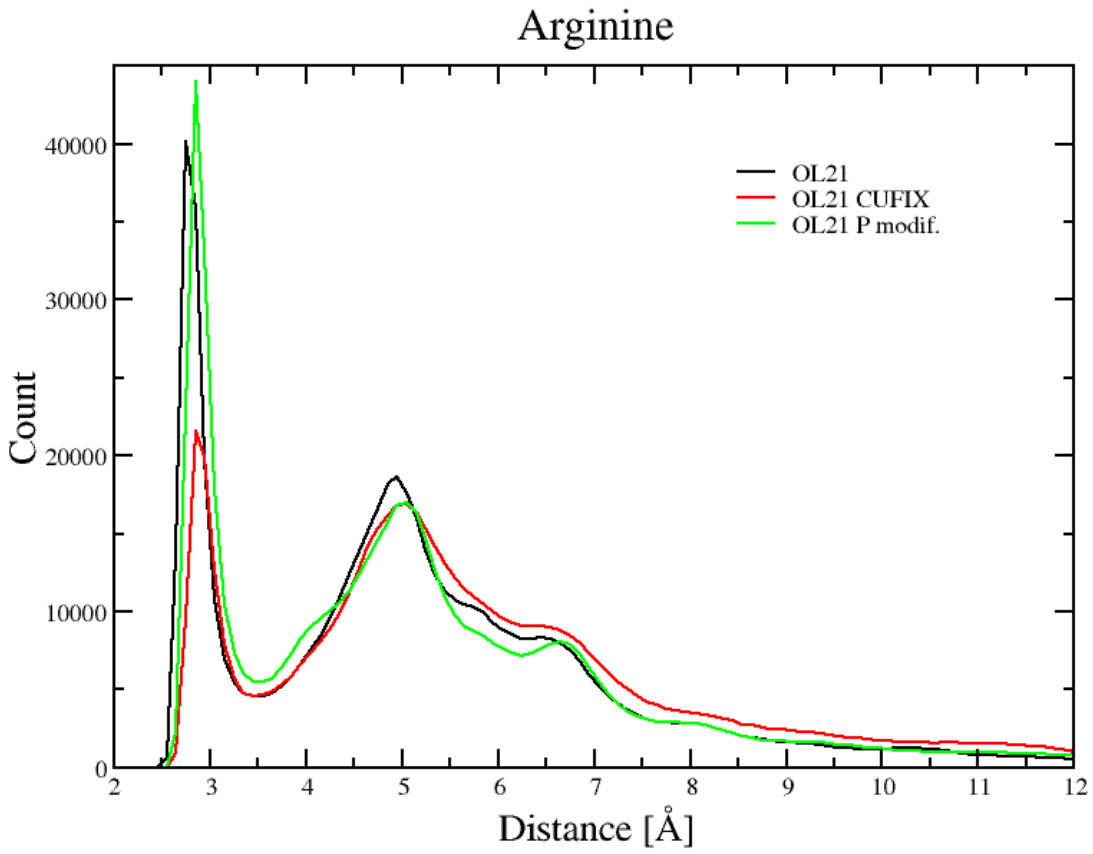


Figure 4.13: Histograms of arginine counts of contacts with phosphate at the certain distance through the simulation in OL21 force field without any modification, with CUFIX (red) and with phosphate modification (green).

Interestingly, it had an unexpected effect on arginine contacts, actually strengthening the interactions and resulting in a significant preference for direct contacts

than it was in the force field without any modification (Fig. 4.13). However, in the case of lysine residues, the effect on direct contacts was almost negligible (Fig. 4.14). These results further demonstrate that the phosphate modification cannot be used in MD simulations of proteins and DNA instead of CUFIX, and the search for a systematic adjustment of electrostatic parameters remains an ongoing area of interest in current research.

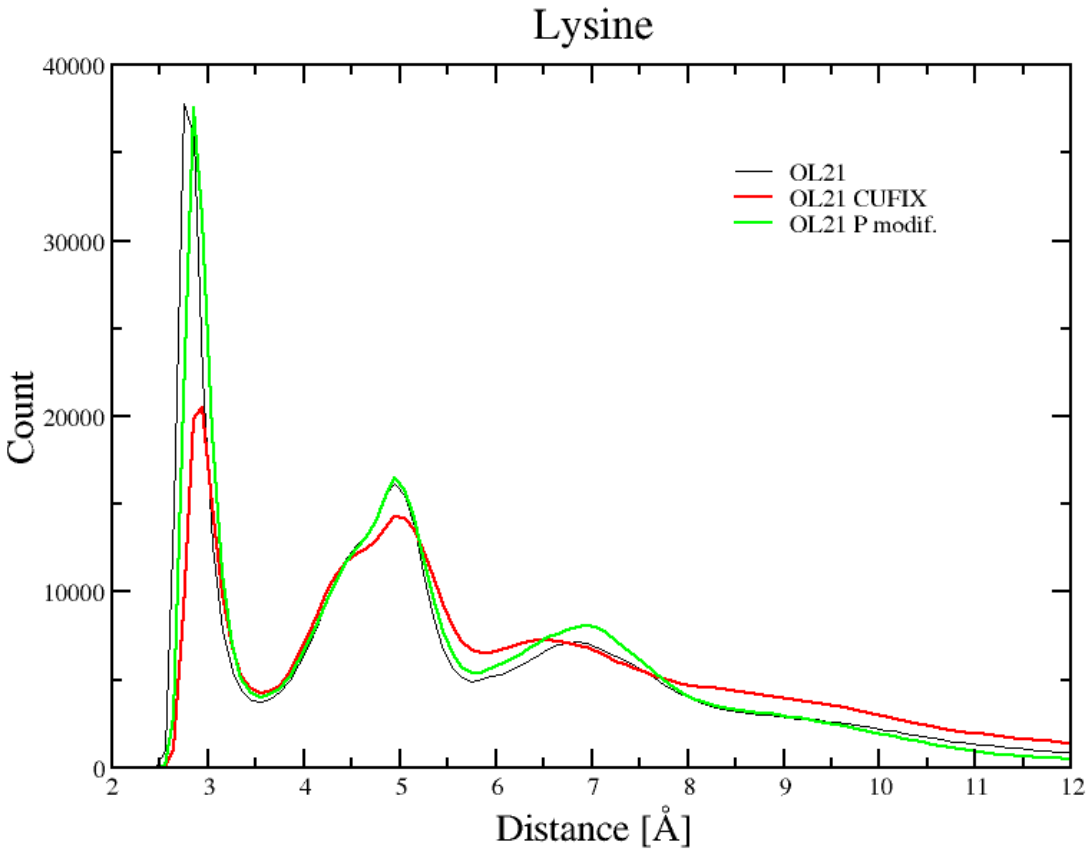


Figure 4.14: Histograms of lysine counts of contacts with phosphate at the certain distance through the simulation in OL21 force field without any modification, with CUFIX (red) and with phosphate modification (green).

Conclusion

Proteins, DNA and their complexes are among the most important biomolecules, therefore it is crucial to understand the nature of their interactions and to be able perform the accurate simulations to obtain relevant results. This thesis was focused on studying electrostatic interactions within protein-DNA complexes by performing molecular dynamics (MD) simulation on set of such complexes. While various force fields (empirical potentials) provide high-standard simulations, the parameters describing electrostatic interactions between charged residues, such as phosphates and cations, are known to be overestimated. In this thesis, the set of suitable protein-DNA complexes was established for further testing of various force field modifications. Complexes were selected based on various criteria, such as the presence of a high amount of salt bridges. Originally, seven complexes were selected from Protein Data Bank: 1B8I, 1J47, 1MNN, 1OSL, 1SKN, 3EYI and 6IS8. The simulations were performed in OL21 force field in AMBER force field family. After the initial MD simulation, all complexes but 1OSL proceeded to the analysis, based on their stability during the simulation process, as measured by their RMSD values. The protein-DNA interactions were assessed by counting direct and water-mediated contacts between DNA phosphate atoms and protein positively charged residues, arginine and lysine, at the distance within 7.2 Å. Total number of interacting phosphates was 123, making in total 231 contacts with either arginine or lysine residue. Overall across the complexes, 128 arginine contacts and 103 lysine contacts were made. Direct contacts were formed at the distance around 2.8 Å across all simulated complexes. Water-mediated contacts were formed at 5 Å on average, which corresponds to the size of water molecule. These contacts are represented in histograms.

CUFIX and phosphate modification were then tested on the selected set to better understand their effects. CUFIX weakens the electrostatic interactions by modifying van der Waals parameters, thereby resulting in reduced direct contacts, as expected. However, CUFIX is optimized based on experimental data and is designed to work with TIP3P water model, which is now considered less accurate than the SPC water model. Therefore, an alternative that would be a systematical modification of force field parameters, independent of solvent model, is required. One of such possibilities is phosphate modification, which increases van der Waals radii and works with the

SPC water model. However, this thesis has shown that the phosphate modification by Case group is not a viable alternative to CUFIX, as its results were negligible or even had an opposite effect.

Bibliography

Aranda-Garcia D., Torrens-Fontanals M., Medel-Lacruz B., Lopez-Balastegui M., Peralta-García A., Dieguez-Eceolaza M., Morales-Pastor A., Sotillo-Núñez D., Abbondandolo D., Stępniewski T. M., Selent J. (2022) Simulating Time-Resolved Dynamics of Biomolecular Systems. *Comprehensive Pharmacology*, 115-134.

Bayly C. I., Cieplak P., Cornell W., Kollman P. A. (1993) A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *The Journal of Physical Chemistry* **97**, 10269-10280.

Cornell W. D., Cieplak P., Bayly C. I., Gould I. R., Merz K. M., Ferguson D. M., Spellmeyer D. C., Fox T., Caldwell J. W., Kollman P. A. (1995) A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society* **117**, 5179-5197.

Ha S. C., Kim D., Hwang H. Y., Rich A., Kim Y. G., Kim K. K. (2008) The crystal structure of the second Z-DNA binding domain of human DAI (ZBP1) in complex with Z-DNA reveals an unusual binding mode to Z-DNA. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 20671-20676.

Hasic H., Buza E., Akagic A. (2017) A hybrid method for prediction of protein secondary structure based on multiple artificial neural networks *40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia,* 1195-1200.

Ivani I., Dans P. D., Noy A., Pérez A., Faustino I., Hospital A., Walther J., Andrio P., Goñi R., Balaceanu A., Portella G., Battistini F., Gelpí J. L., González C., Vendruscolo M., Laughton C. A., Harris S. A., Case D. A., Orozco M. (2016) Parmbsc1: a refined force field for DNA simulations. *Nature methods* **13**, 55–58.

Jorgensen W. L., Chandrasekhar J., Madura J. D., Impey R. W., Klein M. L. (1983) Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* **79**, 926–935. Chem. Phys. 15 July 1983; 79 (2): 926–935

Kalodimos C. G., Bonvin A. M. J. J., Boelens R., Kaptein R. (2004) Structure and Flexibility Adaptation in Nonspecific and Specific Protein-DNA Complexes. *Science* **305**, 386-389.

Kawai K., Majima T., (2002) Effect of hydrogen bonding on the photo-oxidation of DNA. *Journal of Photochemistry and Photobiology C: Photochemistry Reviews* **3**, 53-66.

Lamoureux J. S., Stuart D., Tsang R., Wu C., Glover J. N. (2002) Structure of the sporulation-specific transcription factor Ndt80 bound to DNA. *The EMBO Journal* **21**, 5721–5732.

Lin H., Zhang D., Zuo K., Yuan C., Li J., Huang M., Lin Z. (2019) Structural basis of sequence-specific Holliday junction cleavage by MOC1. *Nature Chemical Biology* **15**, 1241–1248.

Lippert R. A., Predescu C., Ierardi D. J., Mackenzie K. M., Eastwood M. P., Drod R. O., Shaw D. A. (2013) Accurate and efficient integration for molecular dynamics simulations at constant temperature and pressure. *The Journal of Chemical Physics* **139**, 164106

Luscombe N. M., Austin S. E., Berman H. M., Thornton J. M. (2000) An overview of the structures of protein-DNA complexes. *Genome Biology* **1**, 001.1–001.37.

Maier J. A., Martinez C., Kasavajhala K., Wickstrom L., Hauser K. E., Simmerling C. (2015) ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *Journal of Chemical Theory and Computation* **11**, 3696-3713.

Murphy E.C., Zhurkin V. B., Louis J. M., Cornilescu G., Clore G. M. (2001) Structural Basis for SRY-dependent 46-X,Y Sex Reversal: Modulation of DNA Bending by a Naturally Occurring Point Mutation. *Journal of Molecular Biology* **312**, 481-499.

Nesvadba T. (2022) *Mapping polar interactions in protein-DNA complexes.* Bachelor thesis, Palacký University Olomouc, Czech republic.

Passner J., Ryoo H., Shen L., Mann R., Aggarwal A. (rok) Structure of a DNA-bound Ultrabithorax–Extradenticle homeodomain complex. *Nature* **397**, 714–719.

Piana S., Klepeis J. L., Shaw D. E. (2014) ssessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular dynamics simulations. *Current Opinion in Structural Biology* **24**, 98-105.

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

Rupert P. B., Daughdrill G. W., Bowerman B., Matthews B. W. (1998) A new DNA-binding motif in the Skn-1 binding domain–DNA complex. *Nature Structural Biology* **5**, 484–491.

Steinbrecher T., Latzer J., Case D. A. (2012) Revised AMBER Parameters for Bioorganic Phosphates. *Journal of Chemical Theory and Computation* **8**, 4405-4412.

Watson J., Crick F. (1953) Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **171**, 737–738.

Yakovchuk P., Protozanova E., Frank-Kamenetskii M. D.(2006) Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. *Nucleic Acids Res.* **34**, 564-574.

Yoo J., Aksimentiev A. (2012) Improved Parametrization of Li^+, Na^+, K^+, and Mg^{2+} Ions for All-Atom Molecular Dynamics Simulations of Nucleic Acid Systems. *The Journal of Physical Chemistry Letters* **3**, 45-50.

Yoo J., Aksimentiev A. (2016) Improved Parameterization of Amine–Carboxylate and Amine–Phosphate Interactions for Molecular Dynamics Simulations Using the CHARMM and AMBER Force Fields. *Journal of Chemical Theory and Computation* **12**, 430-443.

Yoo J., Aksimentiev A. (2018) New tricks for old dogs: improving the accuracy of biomolecular force fields by pair-specific corrections to non-bonded interactions. *Physical chemistry chemical physics* **13**, 8432–8449.

You S., Lee H. G., Kim K., Yoo J. (2020) Improved Parameterization of Protein-DNA Interactions for Molecular Dynamics Simulations of PCNA Diffusion on DNA. *Journal of chemical theory and computation* **16**, 4006–4013.

Zgarbová M., Šponer J., Otyepka M., Cheatham T. E. III., Galindo-Murillo R., Jurečka P. (2015) Refinement of the Sugar–Phosphate Backbone Torsion Beta for AMBER Force Fields Improves the Description of Z- and B-DNA *Journal of Chemical Theory and Computation* **11**, 5723-5736.

Zgarbová M., Šponer J., Jurečka P. (2021) Z-DNA as a Touchstone for Additive Empirical Force Fields and a Refinement of the Alpha/Gamma DNA Torsions for AMBER *Journal of Chemical Theory and Computation* **17**, 6292-6301.

Others

Case D. A., Cheatham T. E. III., Darden T., Gohlke H., Luo R., Merz K. M. Jr., Onufriev A., Simmerling C., Wang B., Woods R. J. (2005) The Amber biomolecular simulation programs. *Journal of computational chemistry* **26**, 1668–1688.

ChemSketch, Advanced Chemistry Development, Inc. (ACD/Labs), Toronto, ON, Canada, Avaible from www.acdlabs.com.

Gordon J. C., Myers J. B., Folta T., Shoja V., Heath L. S., Onufriev A. (2005) H++: a server for estimating pKas and adding missing hydrogens to macromolecules. *Nucleic acids research* **33**, W368–W371, Avaible from http://newbiophysics.cs.vt.edu/H++/index.php

Humphrey W., Dalke A. and Schulten K. (1996) VMD - Visual Molecular Dynamics *J. Molec. Graphics* **14**, 33-38.

wwPDB consortium (2019) Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Research* **47**, D520–D528, Avaible from https://www.rcsb.org

Schrödinger L., DeLano W. (2020) *PyMOL* **číslo**, Avaible from http://www.pymol.org/pymol