U n i v e r z i t a P a l a c k é h o v O l o m o u c i

Přírodovědecká fakulta

Katedra fyzikální chemie

Stabilita přechodných konformací mezi A- a B-DNA formou v komplexech proteinů s DNA

Bakalařská práce

Autor:	Jan Salomon
Školitel:	doc. RNDr. Petr Jurečka, Ph.D.
Studijní program:	Chemie
Studijní odbor:	Aplikovaná chemie
Forma studia:	Prezenční

Olomouc 2022

P a l a c k y U n i v e r s i t y O l o m o u c

Faculty of Science

Department of Physical Chemistry

Stability of intermediate conformations between A- and B-DNA form in protein-DNA complexes

Bachelor Thesis

Author:	Jan Salomon
Supervisor:	doc. RNDr. Petr Jurečka, Ph.D.
Study programme:	Chemistry
Curriculum:	Applied Chemistry
Study form:	Daily

Olomouc 2022

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně pod vedením doc. Petra Jurečky, Ph.D. a všechny použité literární zdroje uvedl v seznamu použité literatury.

V Olomouci _____

Jan Salomon

I would like to thank my supervisor, doc. RNDr. Petr Jurečka, Ph.D., who offered me to take up this topic and guided me throughout the writing of this thesis. I am also grateful to my grandparents and my mother, who supported me for the last four years of my studies.

Bibliografická identifikace

Autor	Jan Salomon
Název práce	Stabilita přechodných konformací mezi A- a B-DNA formou v komplexech proteinů s DNA
Typ práce	Bakalářská
Pracoviště	Katedra fyzikální chemie, Přírodovědecká fakulta Univerzity Palackého
Vedoucí práce	doc. RNDr. Petr Jurečka, Ph.D.
Rok obhajoby	2022
Abstrakt	Molekulově dynamické simulace se používají k pochopení mechanismů funkce DNA a jejích interakcí. Jedním z důležitých mechanismů rozpoznávání DNA je přechod mezi konformacemi A a B. Aby bylo možné získat spolehlivé informace, musí být tato rovnováha popsána co nejpřesněji. Současná silová pole AMBER vykazují problémy se stabilizací A-formy a přechodných konformací. Analýza v této práci zkoumá stabilitu přechodných konformací a pokouší se najít příčiny těchto nedostatků ve dvou současných silových polích, OL15 a BSC1. K dosažení tohoto cíle byly v obou silových polích simulovány tři komplexy proteinu s DNA.
Klíčová slova	DNA, molekulově dynamické simulace, silové pole, pucker, glykosidický úhel
Počet stran	50
Jazyk	anglický

Bibliografic information

Author	Jan Salomon
Title	Stability of intermediate conformations between A- and B-DNA form in protein-DNA complexes
Type of thesis	Bachelor
Department	Department of Physical Chemistry, Palacký University
Supervisor	doc. RNDr. Petr Jurečka, Ph.D.
Year of presentation	2022
Abstract	Molecular dynamic simulations are used to understand the mechanisms of DNA function and its interactions. One important mechanism of DNA recognition is the transition between the A and B conformations. To produce reliable information, this equilibrium must be described precisely. Current AMBER force fields struggle to stabilize the A-form and the intermediate conformations. The analysis in this work investigates the stability of the intermediate conformation and attempts to find the causes of these shortcomings in two current force fields, OL15 and BSC1. To achieve this goal, three protein-DNA complexes were simulated in both force fields.
Keywords	DNA, molecular dynamic simulation, force field, pucker, glycosidic angle
Number of pages	50
Language	English

Table of contents

1 Introduction

Deoxyribose nucleic acid is one of the most important molecules on our planet. It stores the building instructions for all life on our planet. Inside cells, it interacts with proteins, which read, copy, or repair DNA. These interactions are not solely dependent on the sequence of nucleobases; conformational properties play a major role as well.

To understand the mechanisms of DNA more deeply, X-ray crystallography and NMR analysis is used to obtain the structure of complexes, where a protein binds to a strand of DNA. These methods generate snapshots of the molecules frozen in time. Currently, no method exists, which can precisely determine the structure of a given complex in real time. Molecular dynamic simulations are used to fill this gap, because they allow us to see the given molecules in solution and under conditions like the ones inside cells.

One important aspect in protein-DNA interactions is the transition between two types of DNA conformation: A-DNA and B-DNA. The two conformers differ in the puckering of the sugar and the conformation about the glycosidic bond χ. Those conformations are highly correlated and dependent on each other.

The models used must describe the behaviour of the molecules precisely to produce meaningful information. Since the introduction of the AMBER (Assisted Model Building with Energy Refinement) simulation package[1], different groups strive to improve the modelling of DNA. The most recent developments resulted in two force fields released in 2015, OL15 and BSC1.[2,3] Both adjust the parameters for the glycosidic bond χ, and BSC1 also modifies sugar parameters. This work attempts to assess, whether the modifications succeeded and what changes need to be made to improve the description of the A-B equilibrium.

For this purpose, three protein-DNA complexes were chosen to be simulated using the previously mentioned force fields. The DNA in each complex contains intermediate states between the pure A and B forms.

2 Theoretic

2.1 Structure of DNA

2.1.1 Nucleobases

DNA stores genetic information via sequencing of just four main nucleobases: adenine, thymine, guanine, and cytosine. Adenine and guanine derive from the structure of purine, thymine and cytosine derive from pyrimidine. They are paired either through two or three hydrogen bonds, though there are different possibilities for pairing. The most abundant form is the so-called canonical Watson-Crick pairing, also being the first one discovered. Adenine forms a pair with its complementary base thymine, guanine forms a pair with cytosine. A ball-and-stick model of the pairing is shown in Fig. 1.

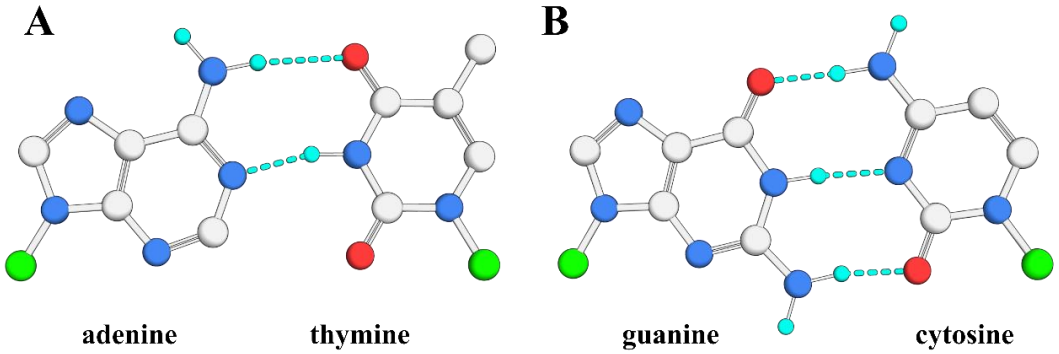

Fig. 1: Watson-Crick base pairing for A) adenine/thymine and B) guanine/cytosine pairs (polar hydrogen atoms and hydrogen bonds marked in cyan, green atoms represent the C1' of the sugar, created in PyMOL[4])

Fig. 2 shows how hydrogen bonds can form between non-complementary bases. These pairs are called mismatches and cause mutations, if not repaired. To repair the incorrectly bonded bases, some proteins specifically bind to such mismatches.[5]

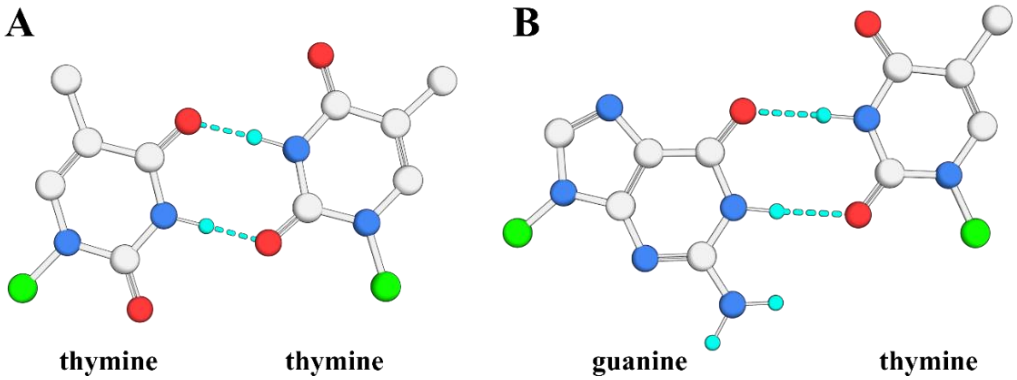

Fig. 2: Mismatched base pairing for A) thymine /thymine and B) guanine/thymine pairs (polar hydrogen atoms and hydrogen bonds marked in cyan, green atoms represent the C1' of the sugar)

Bases can also be unpaired. The nucleotides at the end of DNA strands are more flexible than those inside the strand. Hydrogen bonds between the paired bases can break and the bases become unpaired. This is called fraying. Less common are unpaired bases inside a strand, which happens in some protein-DNA complexes. Bases may either stay stacked with adjacent bases or flip out of the DNA helix similarly to fraying bases.

2.1.2 Backbone

The backbone of DNA consists of chained sequences of deoxyribose and a negatively charged phosphate. Links are established between O5’ of one sugar to the O3’ of another by the phosphate group. The conformation and relative position of each step of the nucleic acid is characterised by nine dihedral angles, whose definitions are shown in Fig. 3, denoted by Greek letters. They can change due to the thermal rotation of a covalent single bond, although they are sterically confined by the surroundings of the DNA to a small section of the full 360°. For each angle there exists a typical range, in which the angle is most of the time.

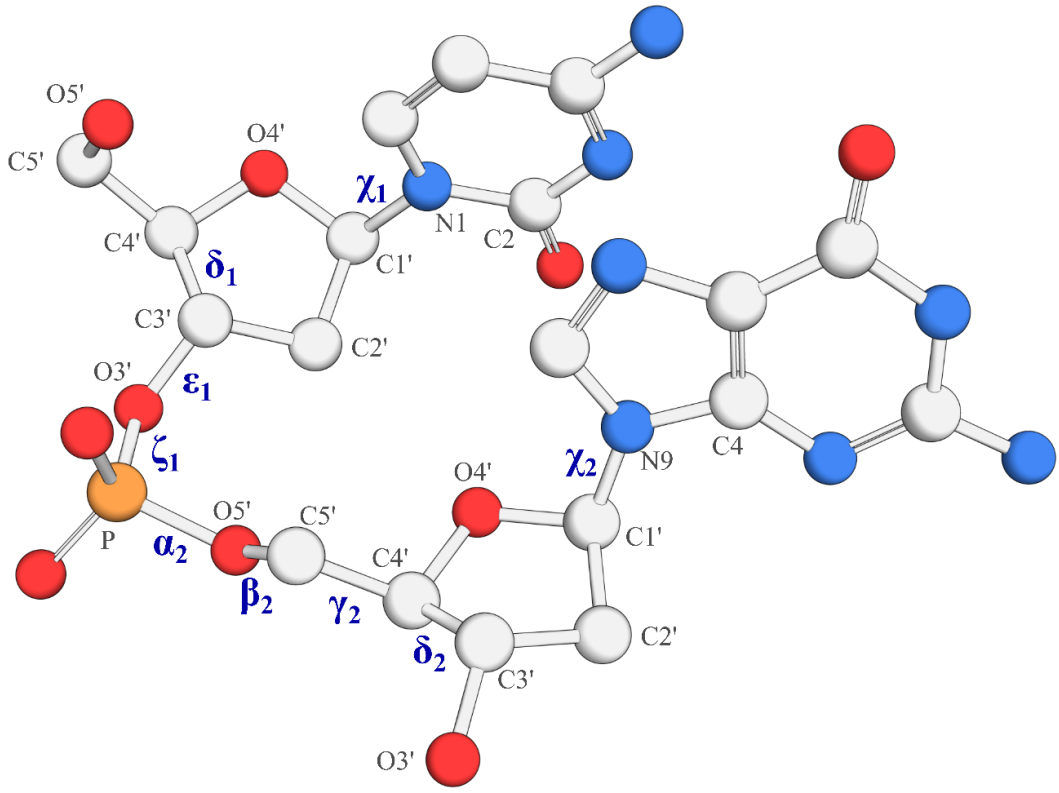

Fig. 3: Atom numbering (grey) and dihedral angles (blue) characterizing a dinucleotide. Dihedral angles of the backbone are defined by the backbone atoms (hydrogen atoms have been omitted for clarity, created in PyMOL[4])

Although the glycosidic angle χ is not a backbone angle, it is important for DNA conformation. The glycosidic angles are defined by O4' to C4 for purines and O4' to C2 for pyrimidines. These angles are constrained to three regions: *syn*, *anti* and high-*anti*. In conformations, where χ is between 0 ° and 120 °, the base is flipped up from its usual position. This occurs in Z-DNA and very rarely in structures, where bases are free to move, like at the fraying ends of double helices. The other two conformations are far more common: *Anti* is typical for A-DNA (~200 °) and high-*anti* for B-DNA (~250 °).

2.1.3 Sugar puckering

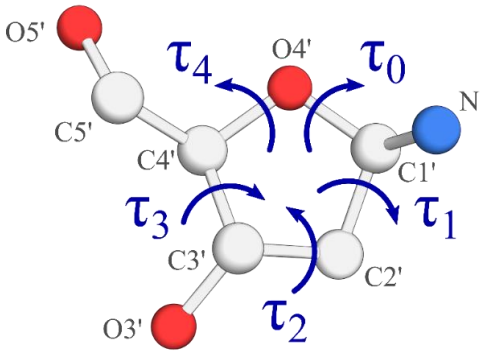

Fig. 4: Deoxyribose in its furanose form with a substituted nitrogen on the C1' carbon (torsions and atoms labelled, created in PyMOL[4])

The deoxyribose is present in its furanose form, which contains a five-membered ring. It is formed by four carbon atoms and one oxygen atom. Due to the usual tetrahedral geometry of sp^3 hybridized carbon, it is impossible for this ring to be planar. Four atoms of the ring can be in one plane at a time with one bulging out of it. The direction of the bulging atom is defined by the C5' atom: if the bulge is on the same side, the conformation is called *endo*, in the other case it is called *exo*. The most common cases are the C3'-*endo* and the C2'-*endo*, characterising the A and the B form respectively. These conformations are described quantitatively by the pucker and the torsion angle delta. The pucker combines the dihedral angles of all five bonds in the ring (shown in Fig. 4) into two descriptors: P is the angle of the pseudorotation (Eq. 1), τ_m is the maximum amplitude (Eq. 2).[6]

$$\tan P = \frac{(\tau_4 + \tau_1) - (\tau_3 + \tau_0)}{2 \cdot \tau_2 \cdot (\sin 36° + \sin 72°)} \qquad \text{Eq. 1}$$

$$\tau_m = \frac{\tau_2}{\cos P} \qquad \text{Eq. 2}$$

The pseudorotation angle has theoretical values from 0 ° to 360 °, in deoxyribose sugars it is found in two typical ranges: north (*N*, 0 °- 30 °), corresponding to C3'-*endo* and the A form, and south (*S*, 150 °- 180 °), corresponding to C2'-*endo* and the B form.

Different combinations of the remaining angles define the associated subconformation. The respective conformation can be altered by external influence by ions in solution or in crystals. In biological systems, proteins are especially useful for this task, changing the conformation for easier access by enzymes for example.

2.2 Conformations

Stacking base steps results in oligomers, whose structures are completely defined by the backbone angles. There exist multiple systems of categorizing states, which vary in precision and scope. The most basic categorization for DNA structure is into A-, B- and Z-DNA. The A and B forms are by far the most important and occurrent.

2.2.1 B-DNA

B-DNA is the most common form of DNA. Under physiological conditions with high water activity, it is the most stable form. It has a large major groove, which gives access to the nucleobases for proteins, and the minor groove is narrow.[7] This is shown in Fig. 5B.

This conformation is characterized by the ribose being in a C2'-*endo* state and χ being at around 270 ° (also denoted as high-*anti*). The C2'-*endo* conformation results in a south pucker angle of around 160 °.

2.2.2 A-DNA

This conformer is more compact compared to the B form. This results in a narrow major groove. It occurs in protein-DNA complexes and plays an important role due to its large, shallow minor groove, presenting little difficulty for proteins to bind.[7] Its double helix structure is shown in Fig. 5A.

A-DNA is defined mainly by the delta angle being around 80°, which means sugar pucker around 18 ° and C3'-*endo* conformation. Its χ angles are in the region around 200 °, often called the *anti*-region. These two values being low result in a dense packing of the bases and a relatively short rise per base pair.

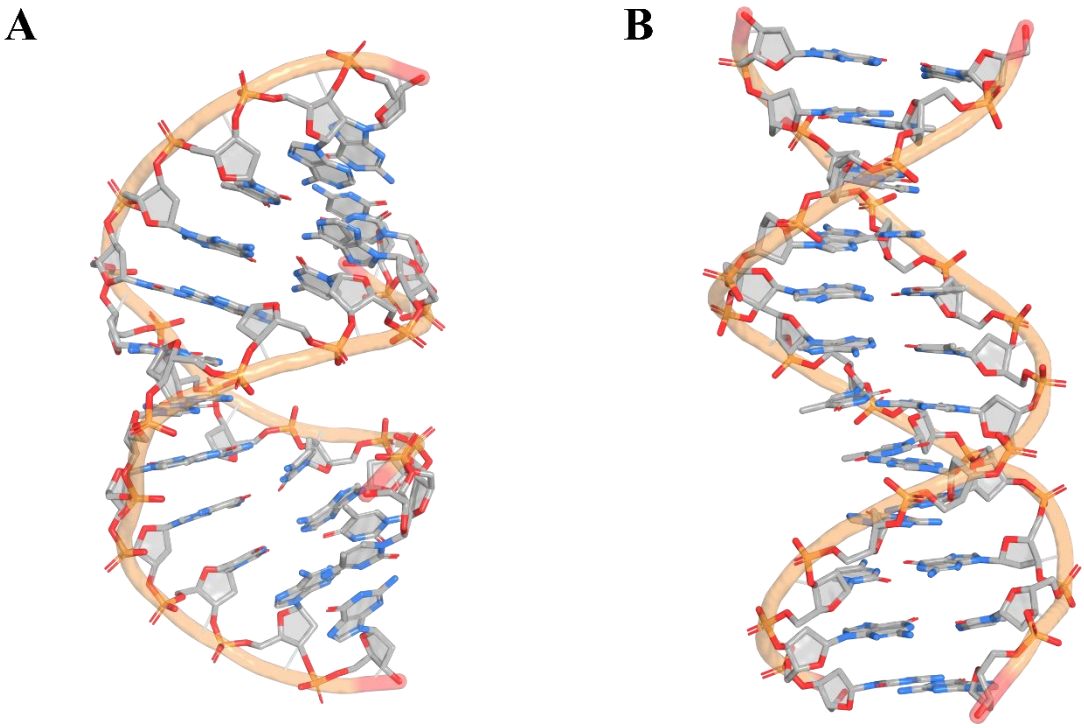

Fig. 5: Dodecamers (5MVK, 1BNA) exemplifying the main types of DNA helix structures: A) A-DNA and B) B-DNA (created in PyMOL[4])

2.2.3 Classification of conformations

More precise categorization has been described by Černý et al. with the NtC (Nucleotide Conformers) and CANA (Conformational Alphabet of Nucleic Acids) system. It categorizes DNA and RNA dinucleotides into these eight subgroups:

- AA and BB – A-DNA and B-DNA structures and similar substates
- AB and BA – transition states between the A and B form
- IC – structures with intercalated bases or residues
- OP – structures with open or distant bases
- SYN – states with unusual *syn* arranged χ angles on one nucleotide
- ZZ – rare Z-DNA structures with substates

The division into A- and B-DNA still applies in this system, most aspects are improved in precision. The sugar conformation decides the classification as the A or B form, the other backbone angles then decide the substate. AA and BB conformations have both puckers in the same C2'/C3'-*endo* state, AB and BA conformations have different sugar conformations with the first named being on the C5'-end of the chain. The following conformations have a less strict structure and can have different puckers.

In this work, two conformers, named AA02 and BB16, were selected for an investigation of their stability in two force fields. Those states are interesting because of their unusual combination of the pucker and χ angles. They are intermediate structures, where the pucker is in one state (either A or B) and χ corresponds to the other state. The precise definitions are shown in Tab. 1.

Tab. 1: Definitions of the most common A state (AA00), the canonical B state (BB00) and the two intermediate states (AA02 and BB16) according to Schneider et al.[8] *The pucker pseudorotation values, which are not specified in the original reference tables, have been calculated from the reference structures. The names of the angles match the backbone angles in Fig. 3.*

State	δ1	ε1	ζ1	α2	β2	γ2	δ2	χ1	χ2	P1	P2
AA00	82	206	288	293	173	55	82	199	200	7.1	13.3
BB00	138	183	258	304	180	44	138	253	258	159.6	153.9
AA02	88	202	274	293	161	54	88	245	246	18.7	28.4
BB16	138	221	282	284	173	48	140	204	270	169.8	157.5

The alphabet by Schneider et al. classifies dinucleotides. In the following analysis the focus will be on mononucleotides, where P and χ belong to different states. This will be denoted by the P/χ notation, where A/A and B/B stand for typical A- or B-DNA, B/A stands for a state with a B-like pucker and A-like χ and vice versa. The definition of each state is explained in Section 3.4. Examples of the four possible states (A/A, A/B, B/A and B/B) are depicted in Fig. 6.

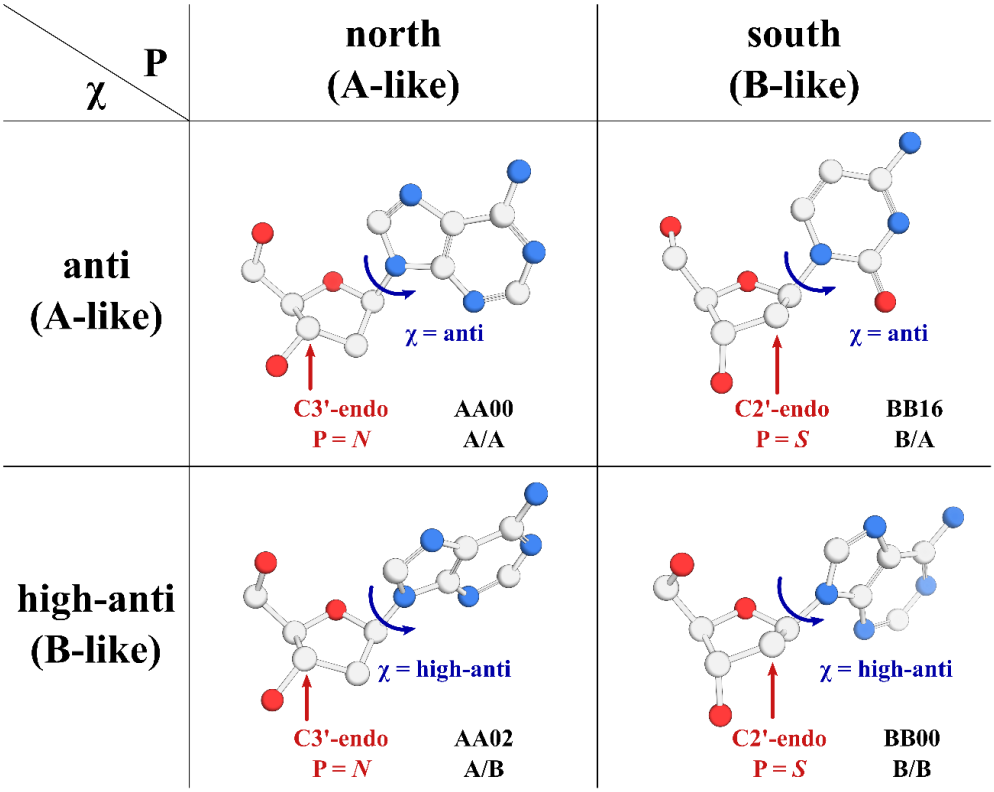

Fig. 6: Comparison of canonical structures of A- and B-DNA with non-canonical intermediate conformations (created in PyMOL[4])

2.3 A – B equilibrium

A- and B-DNA differ mainly in the conformation of the sugar and the glycosidic angle as described previously in Section 2.2.

The equilibrium between the A-form and the B-form is important in protein-DNA interactions. It plays a role in different biological mechanisms, like DNA recognition, protection from damage[9] and structural changes for translation and replication[10]. To understand these processes better, theoretical models must model the stability of the two forms well. There are many factors influencing this equilibrium.

The main reason for B- to A- transitions is a decrease in water activity. The A-form becomes more stable in environments with low humidity, high ion strength or in presence of organic solvents. Water binds strongly to the charged phosphate backbone and polar groups of the nucleobases. The arrangement of water molecules is different in the two conformations.[11]

The transition of B-DNA to A-DNA is also dependent on the base sequence. Sequences rich in cytosine:guanine pairs convert to the A-form more readily than adenine:thymine rich sequences.[11]

2.4 Proteins

There exists an extremely large variety of proteins with many different biological functions. Enzymes are types of protein, which catalyse reactions in organisms whereas structural proteins serve the function of stabilizing an organism. Even though they serve many different purposes, they largely consist of the same elementary units, alpha-amino acids. On Earth, natural amino acids occur in the L-stereoisomer.

The primary structure of proteins in nature is a polypeptide chain of just 20 different amino acids. Depending on the sequence of those, the secondary structure of the polypeptide can take three different forms: the alpha-helix, the beta-pleated sheet, or a simple, unstructured loop. Those elements combine to the tertiary structure of the protein, giving it its properties. It is impossible for the protein to function properly if its tertiary structure isn't upheld.

Alpha-Amino acids used by nature in the synthesis of proteins have the same basic structure with different side chains giving them their structural properties. For our purposes it is useful to divide them into charged, polar, neutral, and hydrophobic.

Hydrophobic amino acids are found inside proteins due to their tendency to interact with each other through van-der-Waals forces to form the protein. Amino acids with polar or charged residues ensure the interaction of the protein surface with aqueous environments.

2.4.1 Protein-DNA interactions

Amino acids with positively charged sidechains like lysine, arginine and histidine provide electrostatic interactions with the negatively charged phosphates of the backbone. Other polar amino acids can interact with the nucleobases via hydrogen bonding to nitrogen or oxygen atoms.

However, there are other amino acids, which are neither charged nor polar, that can interact with DNA, nonetheless. Those with large residues or those including rings can intercalate between the bases of the nucleic acid and induce a bend in the helix. This is typically observed with phenylalanine, proline, or cysteine, which does not have a ring in its structure, but a sulfur atom, whose van-der-Waals radius is comparatively large.[10,12,13]

Proper function of protein-DNA complexes relies on precise fitting of the nucleic acid to the protein's binding site. For this reason, the complexed proteins bound

to DNA often induces a conformational change using the described interactions. These interactions can be non-specific and specific.[7]

Non-specifically binding proteins do not distinguish sequences of base pairs. They mostly bind to the negatively charged phosphate of the backbone using the positively charged amino acids lysine and arginine.

Specifically binding proteins have more complex binding modes to DNA. They recognize sequences of base pairs either by base or shape readout. Base readout occurs at the major or minor groove through hydrogen bonds to the donors and acceptors on the base pairs or through stacking with hydrophobic residues. The hydrogen bonding is mediated through residues, which can accept and donate hydrogen bonds, for example glutamine, asparagine, serine, histidine, lysine, and arginine. The combination of different residues gives different specificity of the protein to a base pair sequence. Shape readout is used with sequences, which have specific deformability. An example is the TATA-box, whose conformation is more variable due to the lack of a third hydrogen bond between bases.[14]

There are different motifs of DNA-binding proteins like zippers, helix-turn-helix, zinc finger or β-sheet types. Helices can insert into the major groove of the DNA. The proteins examined in this work all share the β-sheet motif at the binding site. Binding of β-sheet motifs occurs mostly in the minor groove of the DNA, which often deforms DNA. The resulting structures have widened minor grooves, which increase the probability of A-like conformations.[7]

2.5 Examined protein-DNA complexes

2.5.1 HU architectural factor complex (1P71)

The HU architectural factor plays a role replication and transcription of DNA in prokaryotes by stabilizing bent DNA and straightening compacted DNA.[15] It binds non-specifically to DNA with mismatched bases 9 base pairs apart, which may be important in DNA repair. The complex is represented as a cartoon model in Fig. 7. The structure is available under the PDB ID 1P71 on rcsb.org.[16]

The DNA is double-stranded with three mismatched T:T pairs. The strands have the same sequence of 20 base pairs The two cytosine bases at the ends are unpaired, as well as four thymine bases. Two are facing out of the helix and two are stacked inside it, but unpaired. These bases are important in the crystallization of the complex because they interact with neighbouring complexes in the crystal via hydrogen bonding.

The protein is a dimer of identical chains, which are 94 amino acids long. It binds to DNA by β-sheet ribbons through the minor groove. At the end of each ribbon a proline residue, which intercalates between the bases, stabilizes a sharp kink in the DNA helix. The protein detects its substrate by the way the DNA distorts at the binding site. At the binding site, charged residues like arginine and lysine are prevalent., which means that the surface is positively charged, which allows the protein to bind to the DNA backbone electrostatically. [13]

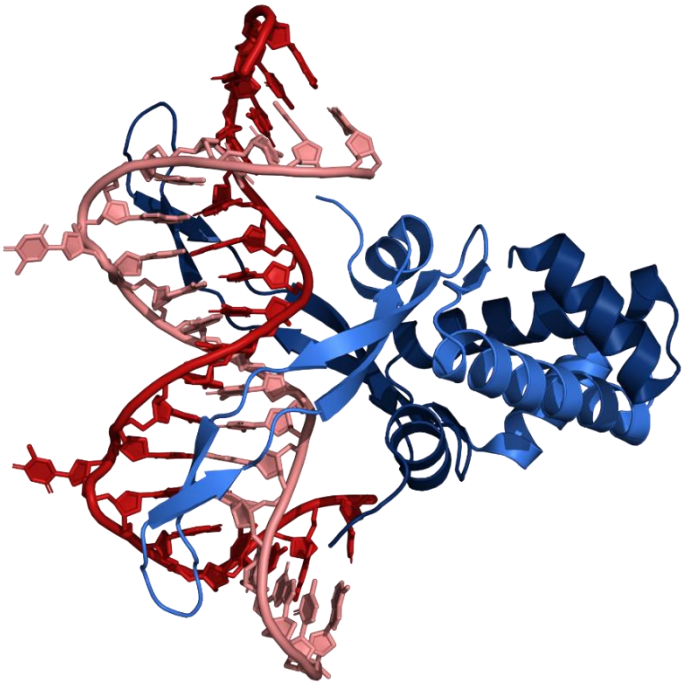

Fig. 7: Representation of the HU architectural factor complex (created in PyMOL[4])

2.5.2 Sso7d chromatin protein complex (1C8C)

This protein found in thermophilic archaeobacteria is believed to have a similar function like histone proteins in eucaryotes. It binds to DNA non-specifically and increases the thermal stability of DNA. Fig. 8 shows a cartoon representation of the complex. The PDB ID of this molecule is 1C8C.[16]

The DNA contained is a double stranded helix, where both strands have the same sequence of 8 nucleobases. The second base pair from each end is a T:G mismatch.

The protein consists of 64 amino acids. Its binding site is formed by a three-stranded β-sheet. On one end of the DNA helix, the side chains of valine and methionine intercalate between bases and induce a sharp kink. Five bases in total interact with the minor groove.[12]

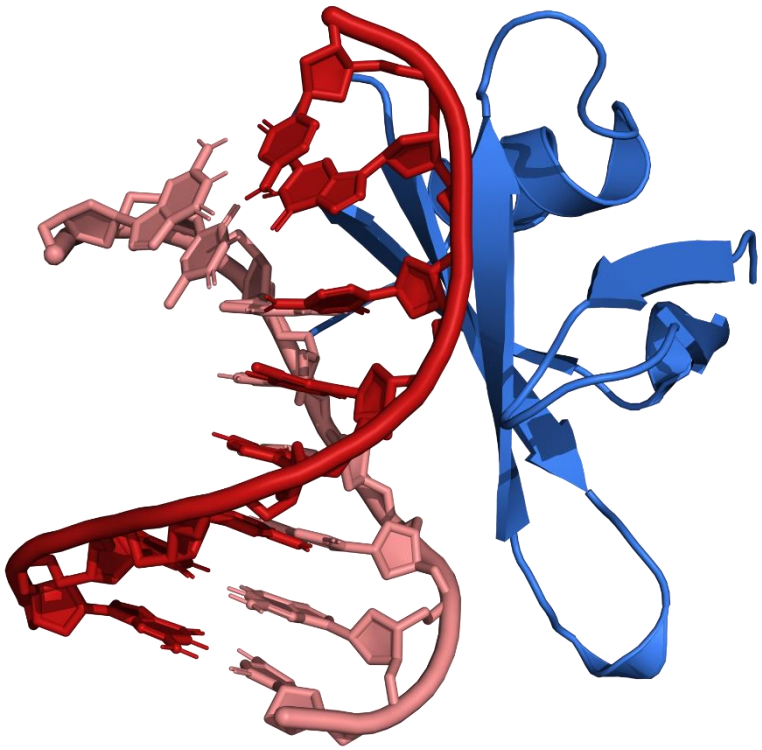

Fig. 8: Representation of the Sso7d chromatin protein complex (created in PyMOL[4])

2.5.3 TATA-box-binding protein complex (1QNE)

The TATA-box-binding protein (TBP) plays an important role in gene transcription in procaryotes as well as eucaryotes. It is selective in binding to DNA sequences of consecutive adenosine and thymine pairs due to the unique flexibility of those polymers.[10,14] In Fig. 9, the complex is depicted in cartoon representation. It is available on rcsb.org under the PDB ID 1QNE.[16]

This complex consists of a wild-type TATA-box-binding protein (TBP) binding complementary double-stranded DNA 14 base pairs in length. The DNA contains the sequence of the Adenovirus major-late promoter TATA-box (TATAAAAG).

The binding is mediated by a 10-stranded β-sheet with charged residues, which interact with the phosphate backbone, and hydrophobic residues, which interact with the nucleobases. The β-sheet interacts with the minor groove. Two pairs of phenylalanine residues intercalate between bases on both ends of the DNA, inducing a kink into the helix.[10]

Fig. 9: Representation of the TATA-box-binding protein complex (created in PyMOL[4])

2.6 Empirical potentials and molecular dynamics

Molecular dynamic simulations are conducted using software packages, which contain so called force fields. They define the potential energy of a molecular system. The AMBER (Assisted Model Building with Energy Refinement) package[1] was developed at the University of California in San Francisco by the Kollman group and is continually improved upon by other groups. It defines the energy of a system as the sum of the bond lengths, the bond angles, the dihedral torsions and the non-covalent interactions of van-der-Waals and electrostatic forces[17].

$$E = \sum_{bonds} K_r\left(r - r_{eq}\right)^2 + \sum_{angles} K_\theta\left(\theta - \theta_{eq}\right)^2 + \sum_{dihedrals} \sum_{n} \frac{V_n}{2} 1 + \cos(n\phi - \gamma) + \sum_{i<j} \left[\frac{A_{ij}}{\left(R_{ij}\right)^{12}} - \frac{B_{ij}}{\left(R_{ij}\right)^{6}} + \frac{q_i q_j}{\epsilon R_{ij}}\right] \quad \text{Eq. 3}$$

The first term defines the bond energy as the sum of the energies of all bonds in the structure. The variable r is the distance between nuclei, r_{eq} is the equilibrium distance. K_r is a parameter, which assigns an energy to a change in distance. This parameter and equilibrium distance is defined for many atom pairings in different hybridizations.

The second term is very similar to the first one. It describes the sum of the energy of all angles. The angle θ is the current bond angle, θ_{eq} is the equilibrium angle and K_θ is the parameter assigning the energy to a change in angle.

The third term formulates the energy of the dihedral torsions. It is the sum of n terms of the Pitzer potential of the torsion. For the majority of four atom combinations AMBER uses two terms, except for atom moieties containing two electronegative atoms on neighbouring carbon atoms, where it uses three. The parameter V_n adjusts the potential energy for any arrangement.

The last term combines the non-covalent interactions. Its first part is the Lennard-Jones potential for van-der-Waals interactions and the second part is the term for coulombic forces of electrostatic interactions. The variables A_{ij} and B_{ij} are parameters for every atom pairing, R_{ij} is the equilibrium distance, where the potential

energy is at its minimum, q_i and q_j are the charges or partial charges on the atoms and ε is a parameter directly proportional to the electric permittivity of the medium.

In this work, two AMBER force fields, which have differences in the parametrization of certain dihedral angles, were used. OL15 and BSC1 are improvements on the BSC0 force field, derived from parm99. They have different approaches in improving the description of DNA. Their development from the parm99 force field is shown in Tab. 2.[3,18,19]

Tab. 2: Development and differences in the examined force fields

Force field	Year	α/γ	β	ε/ζ	χ	P
parm99	2000	parm99	parm99	parm99	parm99	parm99
BSC0	2007	bsc0	parm99	parm99	parm99	parm99
OL15	2015	bsc0	βOL1	ε/ζOL1	χOL4	parm99
BSC1	2015	bsc0	parm99	bsc1 ε/ζ	bsc1 χ	bsc1 P

The most important differences that affect the A-B equilibrium are the descriptions of the pucker and χ. In OL15, there was only a modification to χ and none to the pucker and BSC1 force field introduced modifications to both parameters. These changes were made to better describe the A-B equilibrium. The following work will assess whether these changes succeeded.

The force fields are used in molecular dynamics, a computational method, which studies the movement of the atoms in time, resulting in trajectories. The force fields do not assign forces to the particles, but energies. The derivative of energy with respect to the position yields the force acting on the particle. Through Newton's second law of motion, it is possible to calculate acceleration, when the mass of the particle is a known quantity. The integration of acceleration results in the velocity. The position of the particle after the time step is calculated by multiplying the velocity with the time step used. In the next time step, energies are calculated once more, and the process is repeated over the desired timeframe.

There exist different approaches to simulating solvents in molecular dynamics. The computationally less expensive option is to use the so-called implicit solvent, where the space surrounding the molecule is filled uniformly with a medium of the average properties of the real solvent. Explicit models simulate every molecule of the solvent separately, which increases the computational cost, but also the precision. There are

different generations of models. The major differences are the number of atoms or pseudoatoms used. 3-site models include of the three atoms, 4-site models include the three atoms of water and one pseudoatom for the two lone electron pair, and in 5-site models each lone pair is represented separately. A higher number of sites again increases cost and precision. The use of a water model must be adjusted to the demands of the simulation. In the simulations analysed in this work, the SPC/E model was used. The water molecule in this model has ideal tetrahedral geometry, so the bond angle H-O-H is 109.47 °. It is an improvement on the SPC model, which is very similar, but lacks the polarizability correction.[20]

3 Experimental

3.1 Aim of research

This thesis examines the stability of intermediate states between A- and B-DNA in two AMBER force fields, OL15 and BSC1, which have different parameters for the backbone angles of nucleic acids. There are known shortcomings in the current description of the A-B equilibrium, because of the overestimated stability of the B form.

To improve the description of the equilibrium, the force fields must be adjusted. The analysis focuses on the pucker and χ parametrization and their influence on the stability of the intermediate states to evaluate the impact of each of those angles on the inaccurate description of the stability. With this in mind, a recommendation for future modifications of the force fields is presented.

3.2 Methods

The protein-DNA complexes were selected because of the unusual AA02 and BB16 conformations in the DNA, which are intermediate states between canonical A- and B-DNA, as described in Section 2.2.3. Using the RCSB database and the DNATCO tool by Schneider et al., structures containing these conformers were obtained. A resolution cut-off of 2 Å yielded well-defined molecules, and structures with a high 'confal' score had the best fit with the reference dinucleotides of DNATCO.[8] Three protein-DNA complexes, which had these conformations in proximity of the protein were selected.

Molecular dynamic simulations were conducted with structures of protein DNA complexes obtained via the RCSB Protein Data Bank (URL: rcsb.org), which stores molecular structures in the .pdb format.[16]

The crystallographic structures were solvated in the centre of a truncated octahedral box, whose edges were no less than 10 Å from the solvated molecule. SPC/E water model molecules were used to fill the box. To meet the electroneutrality condition, the residual charges of the molecules were neutralised by adding either potassium or chloride ions[21,22]. Additionally, potassium chloride was added to increase the concentration of potassium cations to 150 $mmol \cdot dm^{-3}$, the physiological value inside cells. The box is used to ease the handling of boundary conditions of the container. Boxes used in molecular dynamics tile the 3D space by repeating infinitely in every direction. This approach is called periodic boundary conditions.

To avoid the separation of the complex or other undesirable outcomes, the entire box had to be equilibrated from a temperature of effectively 0 K to a temperature of 300 K, close to physiological values. The strain in the molecule would tear it apart if a velocity was assigned to every atom randomly. By initially restraining the movement of atoms and hydrogen mass repartitioning, where part of the masses of heavy atoms bonded to hydrogen is partitioned to the bonded hydrogens, the temperature can be raised gradually to the desired value. The result is a file that contains the solvated molecule without strain, which can be used for further simulations.

Each protein-DNA complex was simulated once in each force field. The force field used for proteins in all six simulations was the ff14SB force field.[23] The solvent was simulated with the SPC/E[20] model with ion parameters by Joung et al.[21,22]. The difference between simulations was the two force fields modelling the DNA strands.

The simulations had a duration of 1 µs with a timestep of 4 fs. The temperature was held at 300 K using a Langevin thermostat and the pressure was held constant by the Monte-Carlo barostat at 1 bar. The data was stored every 10 ps and a restart file was stored every 10 ns for backup in case the simulation failed at some point. The engine used was Particle Mesh Ewald Molecular Dynamics (PMEMD), which runs on graphics processing units, which allows it to run faster than its predecessor, Sander. Particle Mesh Ewald summation algorithms are necessary for the accurate simulation of electrostatic interactions in periodic systems. Electrostatics are particularly important in nucleic acid simulation, due to its negatively charged backbone.[24]

The relevant dihedral angles of the DNA backbone were obtained using cpptraj[25] analysis of the trajectory file.

3.3 General stability of complexes

To assess the stability of the complexes, the root-mean-square-deviation (RMSD) of atomic positions is calculated for every timestep. It is the average distance of atoms to a reference structure, in this case the crystal structure in the beginning of the simulation. Due to the thermal movement, it does rise in the beginning of the simulation, but it should stabilize. If the RMSD does not stabilize and rises throughout the simulation, the simulation is unstable, because the complex does not find an equilibrium position. RMSD is also used to detect major conformational changes in the complex, but it loses its usefulness when comparing finer changes like A-to-B transitions.[26] Terminal nucleotides are usually excluded from the RMSD calculation, because they increase RMSD by fraying.

3.3.1 HU architectural factor complex

The development of RMSD in the simulation of the HU architectural factor complex is shown in Fig. 10. It is relatively high compared to other complexes of similar size, but in an acceptable range, because the complex was not expected to be stable in the initial binding mode. The increased RMSD is explained by the nature of the crystal packing of the complex. It interacts with neighbouring complexes by stacking, which induces a much sharper bend in the DNA than would be stable solution. Due to the strain, the DNA contained in the complex attempts to straighten.[13] One further reason is the comparatively loose binding of the termini of the DNA double helix to the protein. They can bind to different amino acids on the protein surface.

RMSD of the DNA is generally higher in OL15 than in BSC1, with spikes to about 3.5 Å. However, it remains stable at that value. In general, the simulations are stable, and the protein binding changes in expected ways.

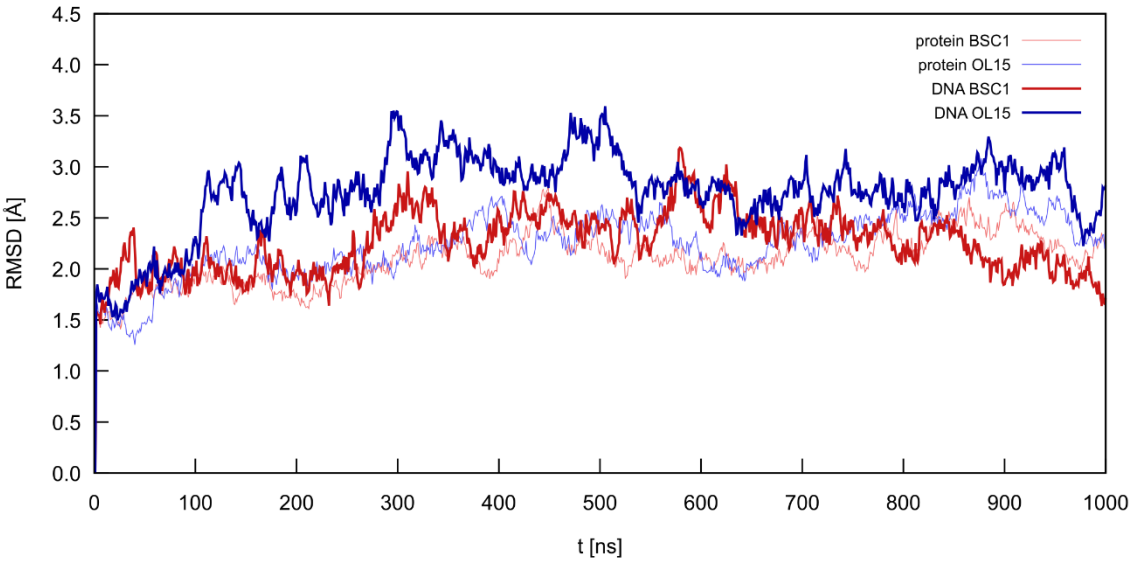

Fig. 10: RMSD development for the DNA and the protein during the simulation of the HU architectural factor complex (the terminal nucleotides have been excluded)

3.3.2 Sso7d chromatin protein complex

Fig. 11 shows the development of RMSD of the chromatin protein complex. The RMSD of the protein is high in both force fields. This is mainly caused by a very mobile loop at the end of the protein, which does not bind to the DNA. For the DNA, RMSD stays at about 1 Å for the whole simulation with some spikes to 1.5 Å. The DNA is only 8 base pairs long and does not extend beyond the binding site, which means that major reconformations are impossible.

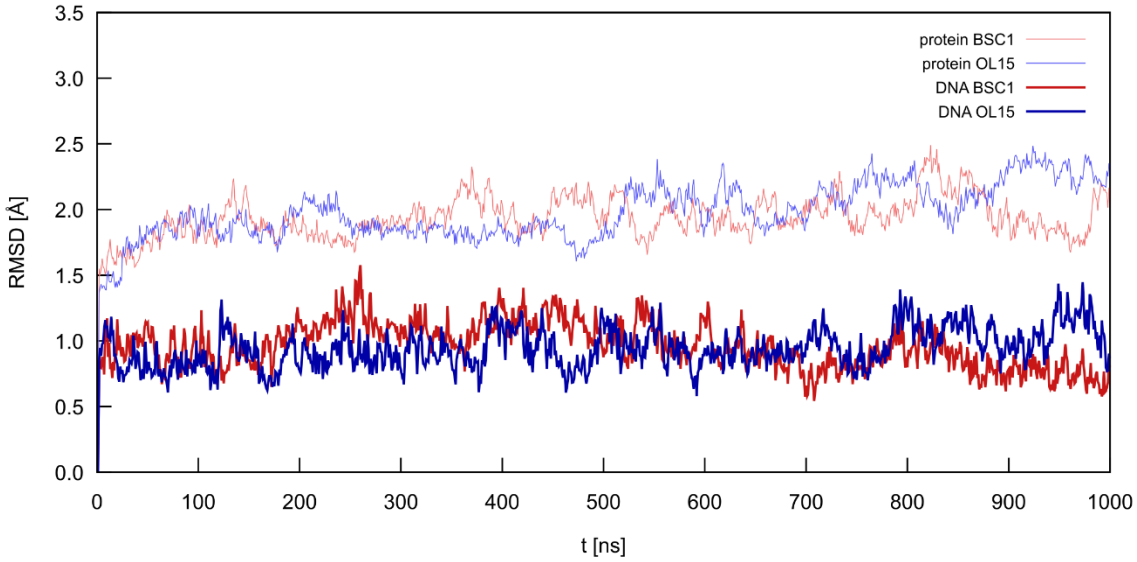

Fig. 11: RMSD development for the DNA and the protein during the simulation of the chromatin protein complex (the terminal nucleotides have been excluded)

3.3.3 TATA-box-binding protein complex

In Fig. 12, the development of the RMSD of the TATA-box-binding protein complex is shown. RMSD of the DNA remains at about 1.5 Å for the whole duration of the simulations in both force fields, sometimes spiking above 2 Å in. The RMSD shows that no major change in conformation occurred, because the protein is strongly bound to the DNA.

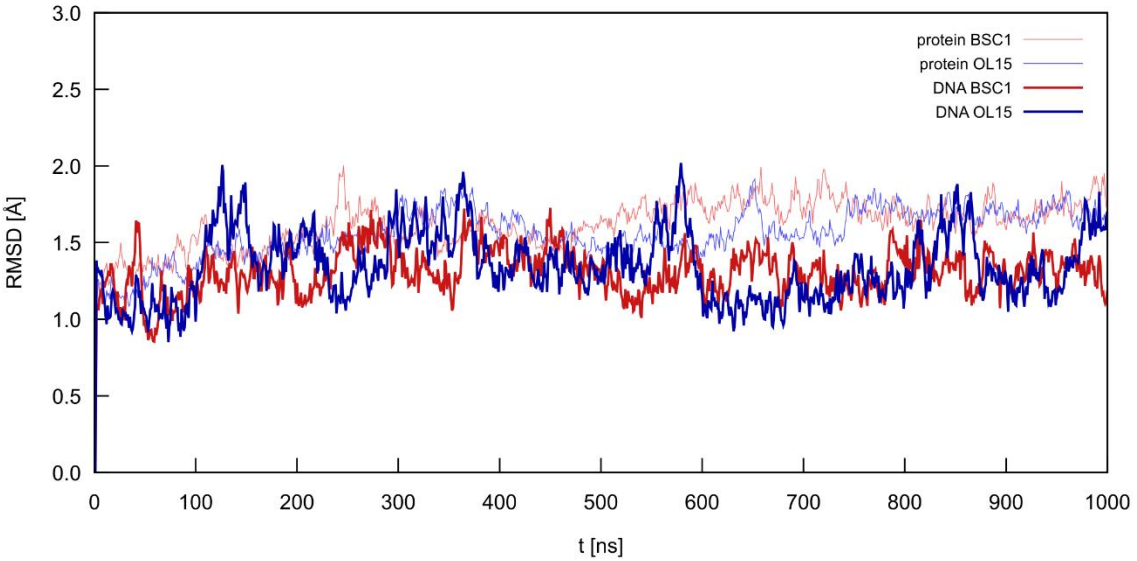

Fig. 12: RMSD development for the DNA and the protein during the simulation of the TATA-box-binding protein complex (the terminal nucleotides have been excluded)

3.4 P and χ distributions

Using P and χ as the sole descriptors of a nucleotide as a simplified model, a method for assigning structures had to be devised. The first approach was the assigning by shortest distance to a reference state in the P/χ phase space. Using only four reference states, assigning structures by shortest distance similarly to the analysis of NtC conformers in Section 3.6 results in the Voronoi diagram as shown in Fig. 13. This leads to incorrect assigning of structures because structures with χ at around 200 °and the pucker P at 30 °should be assigned to the A/A state for example. This forced us to use an alternative, simplified method of assignment, based on thresholds for P and χ to divide the P/χ states. The thresholds are described in the following.

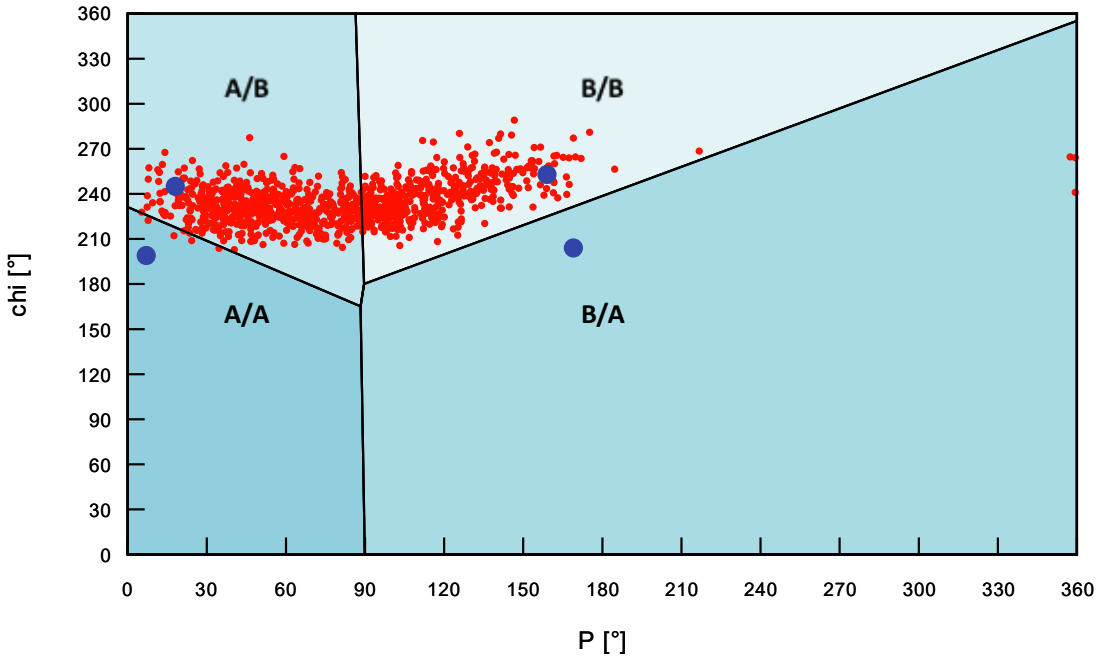

Fig. 13: Voronoi tiling of the P/χ phase space using the points marked in blue as reference states. Red points mark states, in which residue 4 in the HU architectural factor complex occurs in a simulation as an example

The distribution of the pucker angles for nucleotides representative for each of the states studied (A/A, A/B, B/A and B/B) is shown in Fig. 14. The states with P between 60 ° and 120 ° as well as 210 ° to 300 ° have no significant population in ideal states. This serves as a divider between the A and B states of the pucker

Cutoffs between the north and south pucker at 65 ° and 295 ° have been applied in literature to unbound DNA, where the frequency of the north pucker drops off sharply until about 30 °.[26] In simulations of the DNA in protein-DNA complexes the

region above 30 ° had significant population and a portion of those states would not have been assigned to the A form. Therefore, cutoffs were selected at 90 ° and 270 °, where populations are at their minimum.

The graph of the pucker distribution for the nucleotides that started in the A/A state has a small peak at 160 °, which is due to the instability of this state in the OL15 force field and consequently a small population of the south pucker in the chosen nucleotide (red curve in Fig. 14). These peaks occur with almost all nucleotides starting from the A/A state, because of the underestimated stability of the north pucker in the force fields. In contrast to the A/A pucker, the A/B pucker usually has a wide peak, which is shifted away from the typical north pucker (yellow curve in Fig. 14). In some nucleotides, peaks may even be shifted towards 330 °, corresponding to a C2'-*exo* state.

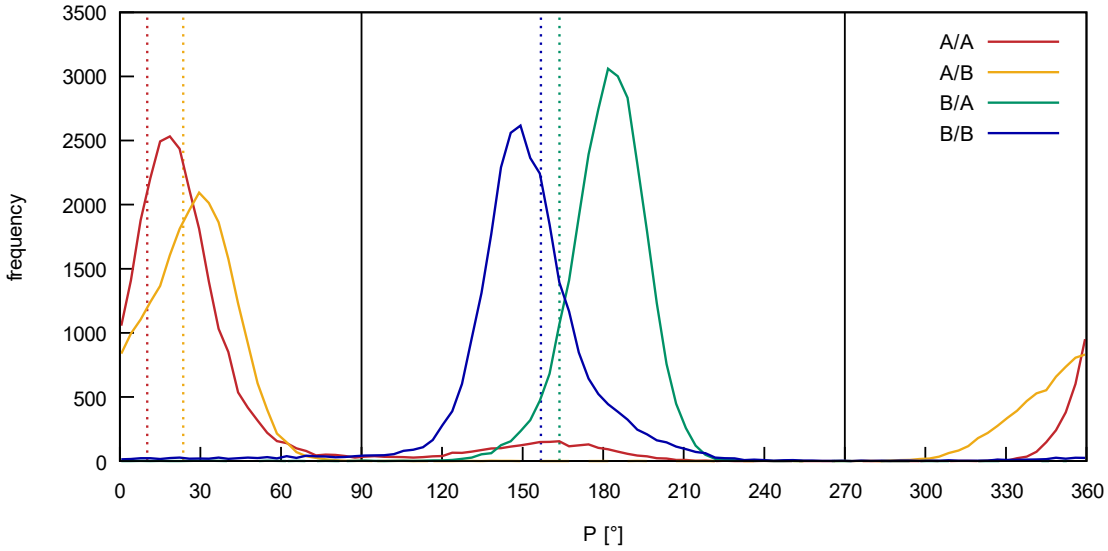

Fig. 14: Distribution of P for typical nucleotides during a simulation: A/A is represented by 1QNE_C_204_A /BSC1, A/B by 1QNE_C_207_A/OL15, B/A by 1P71_D_8_C/BSC1 and B/B by 1QNE_D_226_A/OL15. Dotted lines are reference values from crystallographic data, solid lines are cutoffs between conformations.

In Fig. 15, the distribution of the glycosidic angles of representative states are shown. The distinction between the two conformations is harder to define than with the pucker, due to the significant overlap between the regions. The cutoff selected in this work is the arithmetic mean between the crystallographic values in canonical A- and B-DNA at 227.5 °. The *anti* conformers have sharper peaks close to the crystallographic value. The peaks in the high-*anti* region are flatter and spread above 300 °.

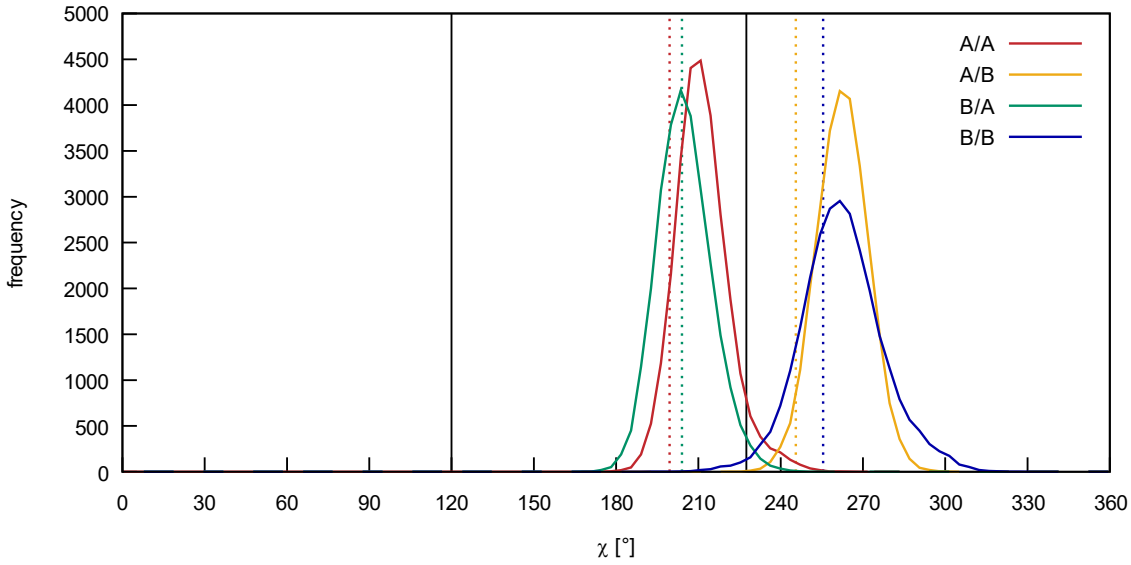

Fig. 15: Distribution of χ for typical nucleotides during a simulation: A/A is represented by 1QNE_A_204/BSC1, A/B by 1QNE_A_207/OL15, B/A by 1P71_C_8/BSC1 and B/B by 1QNE_A_226/OL15. Dotted lines are reference values from crystallographic data, solid lines are cutoffs between conformations.

The previously described thresholds correspond to a partitioning of the P/χ phase space shown in Fig. 16. The problematic conformations are now correctly assigned to the A/A state.

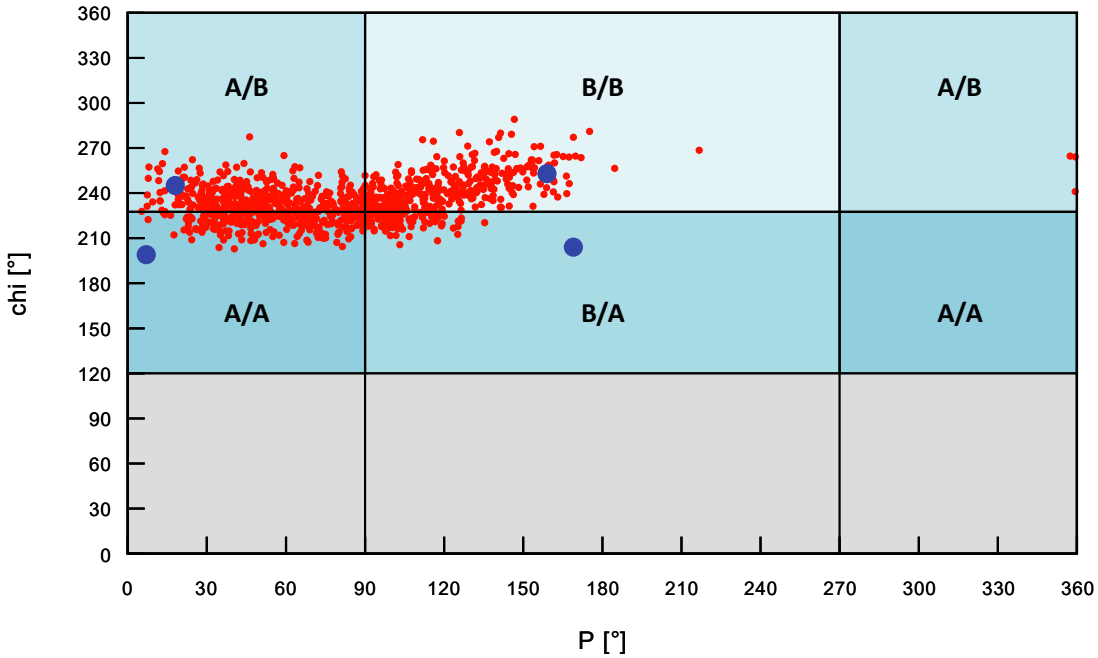

Fig. 16: Simplified partitioning of the P/χ phase space using the described thresholds for the same mononucleotide as in Fig. 13. The grey area contains syn states of χ.

Fig. 17 shows the distribution of the pucker conformation in OL15 and BSC1 for all simulated nucleotides. It is apparent that the behaviour of the B-like south pucker is similar in both force fields. The peaks are at about 150 °, only 4 ° apart, which is very close to the crystallographic value. The force fields describe the west region, corresponding to O4'-*exo* states similarly with virtually no population due to the steric effects.

Differences arise between 0 ° and 120 ° and between 300 ° and 360 °. BSC1 has a wide peak at 35 ° and OL15 at 47 °. The regions between the A and the B state behaves differently as well. In OL15, the population of the east pucker is high in comparison to BSC1. OL15 also has an inflection point at 0 ° degrees, where the population decreases more slowly than in BSC1 towards the west conformation. This is due to some nucleotides in A/B states, whose population maximum is in the region of C2'-*exo* states.

Generally, intermediate states between the *N* and *S* pucker are more stable in OL15. It has a broader distribution of states. BSC1 has comparatively sharp distributions closer to the crystallographic values.

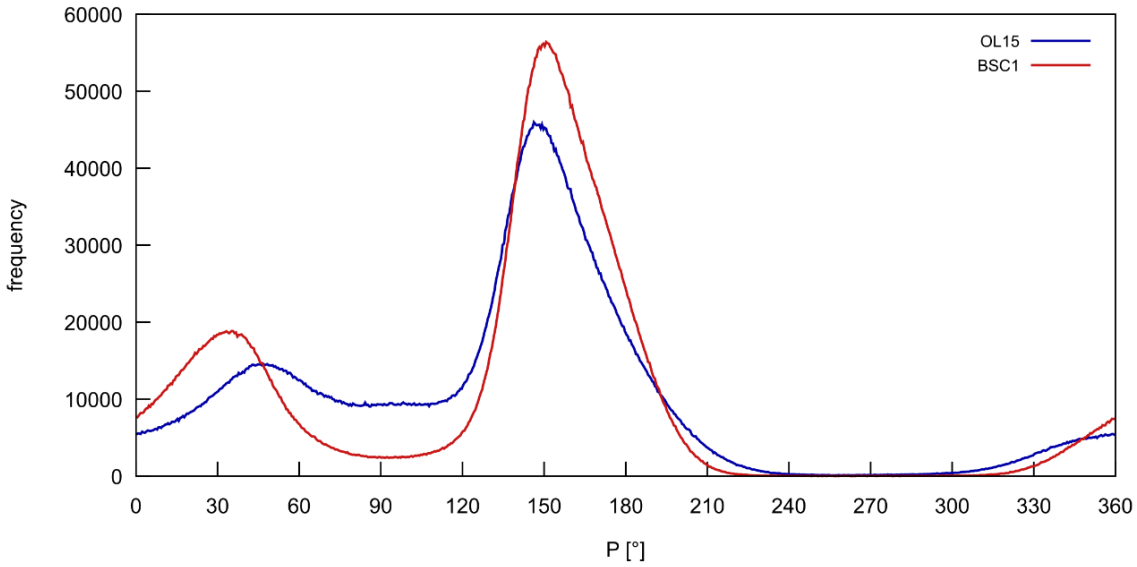

Fig. 17: Distribution of the pucker states in both force fields for all nucleotides in the simulated structures (excluding terminal nucleotides)

Fig. 18 shows the distribution of the glycosidic angle χ for the same set of nucleotides as used in Fig. 17. The *anti* region is the most populated. The maximum of the population is shifted towards high-*anti* in OL15 by 7 °. BSC1 has a broader distribution, but the beginning and ends of the peaks are the same.

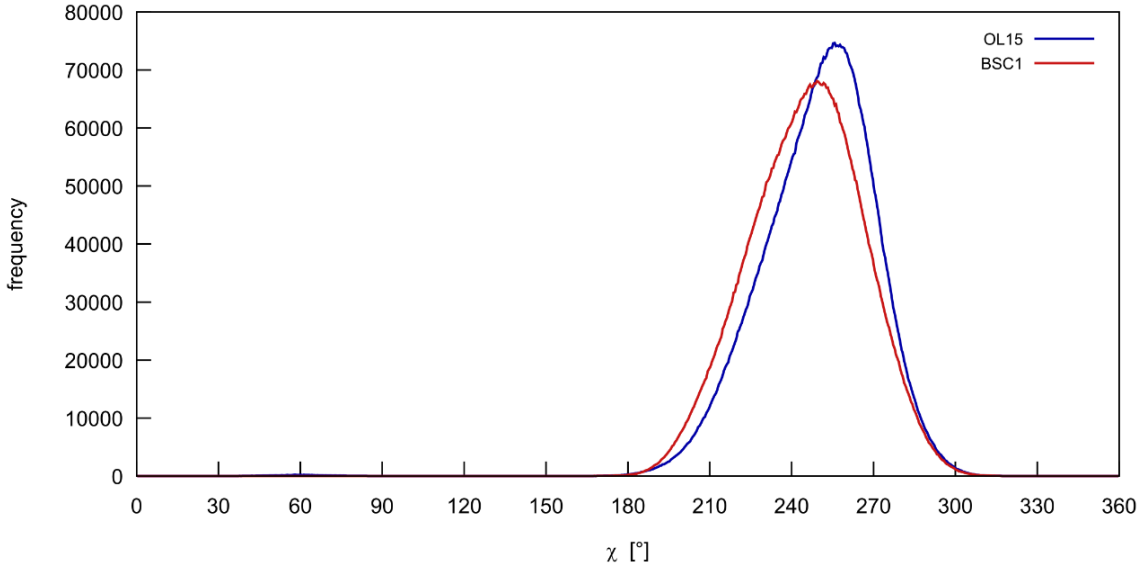

Fig. 18: Distribution of the states of the glycosidic angle χ in both force fields for all nucleotides in the simulated structures (excluding terminal nucleotides)

3.5 Stability of P/χ states

Because the force fields OL15 and BSC1 have similar parametrizations, they describe the behaviour of nucleotides very similarly for most states. The differences become clear in uncommon states. Most of the states remain stable in their initial conformation, and some irreversibly change conformation or establish equilibria between states.

The angles of stable states remain near the crystallographic value in the beginning of the simulation. Although thermal movement always occurs, those angles never visit a distant region. Unstable states leave their initial conformation close to the beginning of the simulation and do not return to that conformation. Some nucleotides establish equilibria between states. In this case, at least one of the angles oscillates between regions corresponding to different states.

The graphs of the angles against time were constructed from data sampled every 5 frames, which corresponds to sampling once every 50 picoseconds.

3.5.1 A/A conformations

In the examined complexes, only three non-terminal nucleotides initially were in the A/A state according to the definition explained in Section 3.4. None of them were particularly stable. Both relevant angles in nucleotide 2 on chain D in the HU architectural factor complex leave the initial values immediately and the conformation

changes to B/B for the duration of simulations with both force fields, shown in Fig. 19. The angles fluctuate considerably, but never return to the starting conformation.

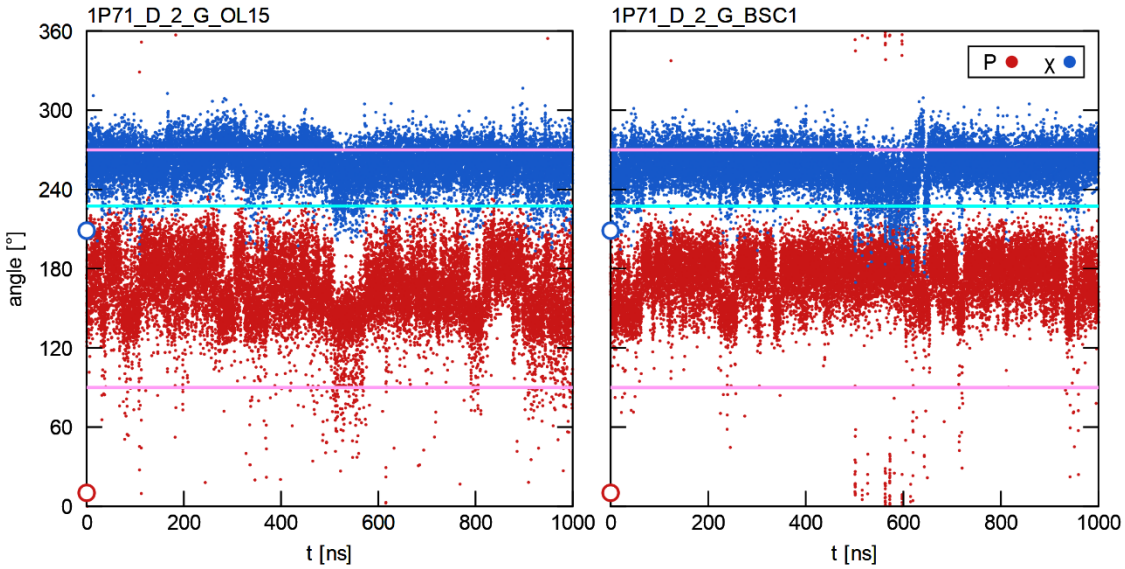

Fig. 19: Comparison of the development of the χ and pucker for residue 2 on chain D (guanine) in the HU architectural factor complex in both force fields (OL15 left, BSC1 right). Division lines between N and S pucker are pink, the division line between anti and high-anti χ is in cyan, points on the y-axis mark the initial values

The other two nucleotides in the A/A state behave differently. Fig. 20 shows an example of the development of the states in both force fields. The χ angle remains stable between the two regions, but the pucker oscillates between *N* and *S*. Both force fields describe the conformation as an equilibrium between those states, but differences in the behaviour are apparent. In OL15, the sugar conformation changes on a very small timescale with the intermediate east pucker being highly populated. In BSC1 on the other hand, states are stable for longer periods at a time. There is also a clear gap between the bands in the *N* and *S* conformation, which means the east pucker is almost unpopulated.

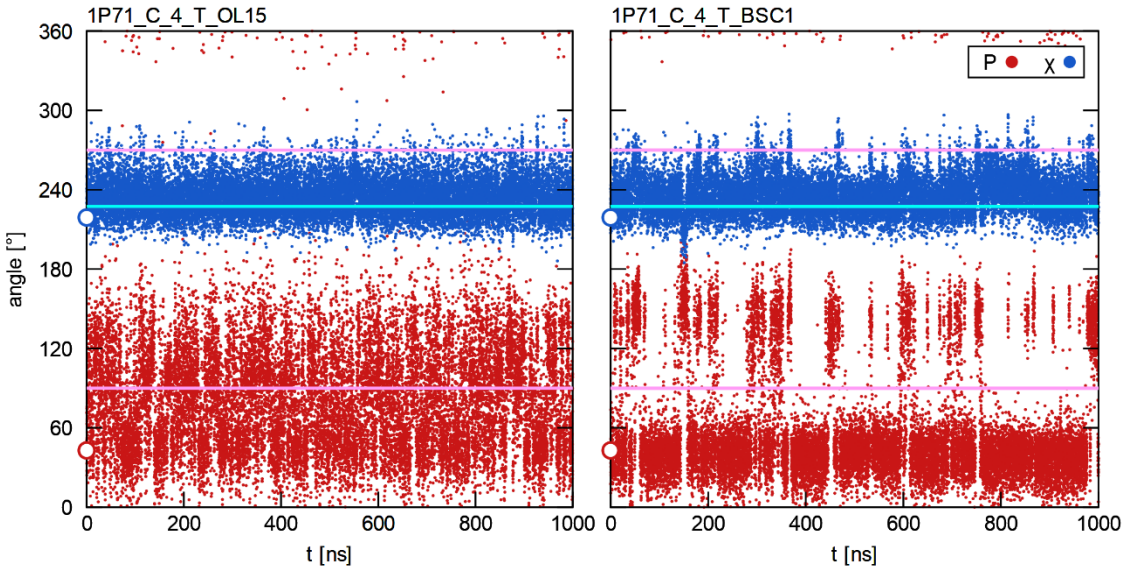

Fig. 20: Comparison of the development of the χ and pucker for residue 4 on chain C (thymine) in the HU architectural factor complex in both force fields (OL15 left, BSC1 right). Division lines between N and S pucker are pink, the division line between anti and high-anti χ is in cyan, points on the y-axis mark the initial values

3.5.2 A/B conformations

Of all examined non-terminal nucleotides, 16 were in this conformation initially. The behaviour of these states was largely identical. In the TATA-box-binding protein complex, each strand of the double helix contained a stretch of five consecutive bases the A/B state, which all were stable and did not form an equilibrium with another state. An example of one of these stable states is shown in Fig. 21. The angles persist close to the crystallographic values and have almost no outliers in other states.

The two A/B states in the HU architectural factor complex also remain stable, although the pucker is shifted towards the east region and a significant population of *N* pucker states.

In the chromatin protein complex the behaviour in the two force fields is different. An example of a nucleotide is shown in Fig. 22. The pucker angle remains around the starting value in OL15 in the east conformation for the entire duration of the simulation, in BSC1 it oscillates between the north and south, and it settles in the north conformation eventually. Like the pucker angle before, the glycosidic angle is also steady in OL15. In BSC1, χ deviates coupled with the pucker. With a south pucker, χ changes to a high-*anti* value, and with a change to the north pucker, it shifts towards the *anti* region.

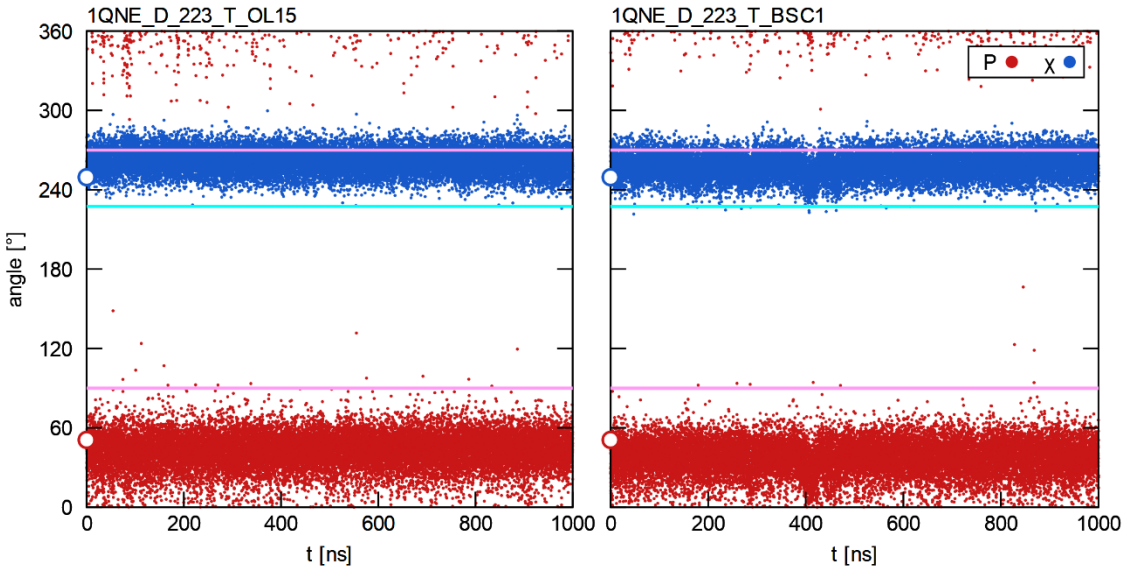

Fig. 21: Comparison of the development of the χ and pucker for residue 223 (thymine) in the TATA-box-binding protein complex in both force fields (OL15 left, BSC1 right). Division lines between N and S pucker are pink, the division line between anti and high-anti χ is in cyan, points on the y-axis mark the initial values

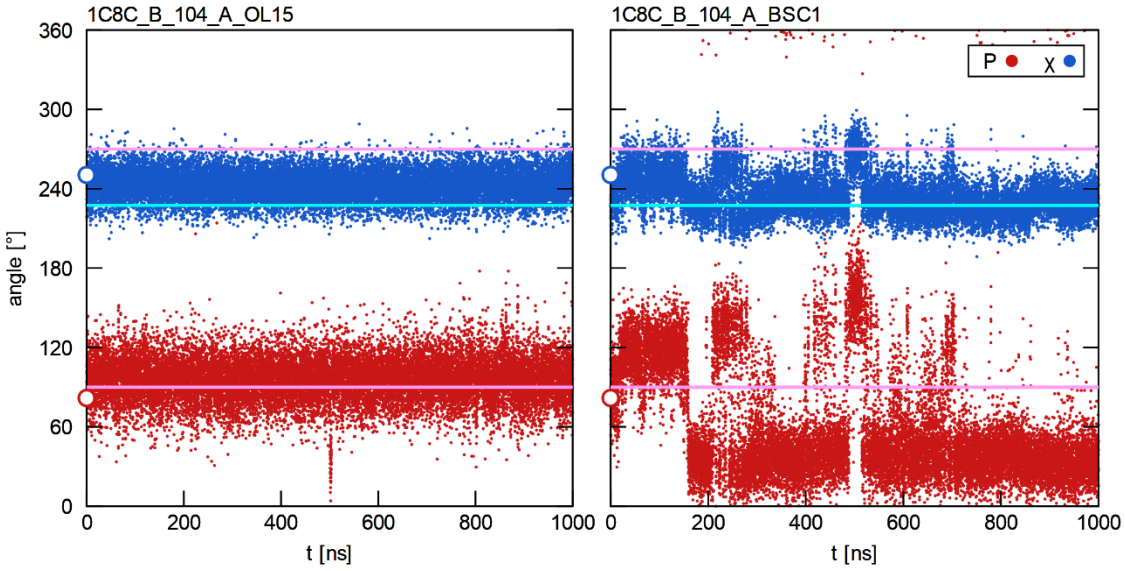

Fig. 22: Comparison of the development of the χ and pucker for residue 104 (adenine) in the chromatin protein complex in both force fields (OL15 left, BSC1 right). Division lines between N and S pucker are pink, the division line between anti and high-anti χ is in cyan, points on the y-axis mark the initial values

In the cytosine nucleotide 113 on chain C in the chromatin protein complex, which starts in an A/B state, the stability is different in both forcefields. Fig. 23 shows the development of χ and P in both forcefields. In OL15, the state is relatively stable with the pucker being mostly between north and east and χ being stable in the high-*anti* region. BSC1 describes the glycosidic angle similarly, the pucker, however, changes conformation to the south region without converting back or establishing an equilibrium. This is an example for a nucleotide, where the decreased stability of the east pucker in BSC1 affects the resulting conformation.

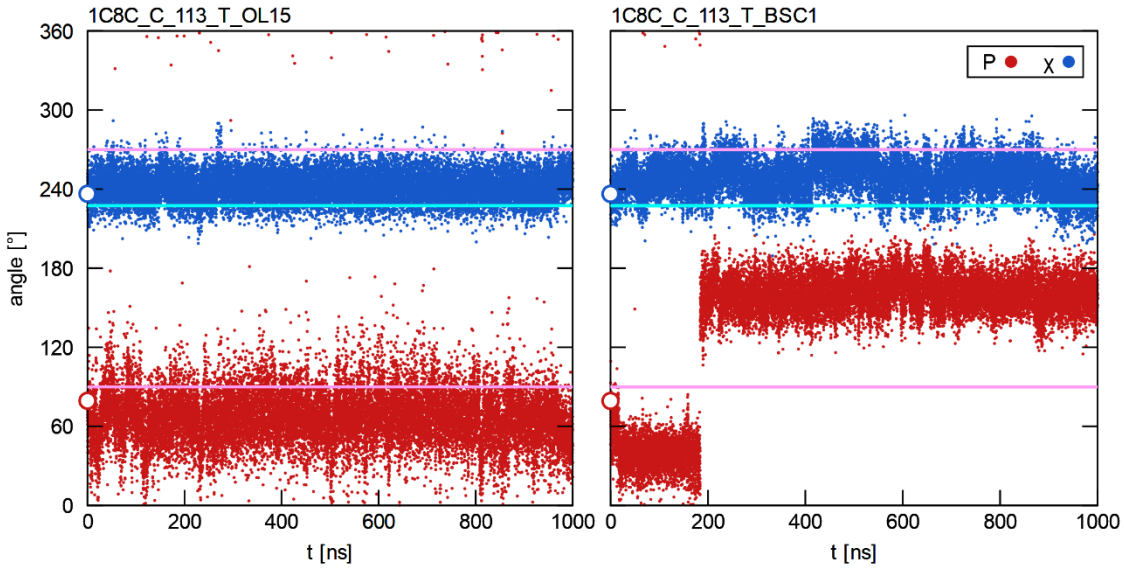

Fig. 23: Comparison of the development of the χ and pucker for residue 113 (cytosine) in the chromatin protein complex in both force fields (OL15 left, BSC1 right). Division lines between N and S pucker are pink, the division line between anti and high-anti χ is in cyan, points on the y-axis mark the initial values

From the distribution of the pucker in the 16 nucleotides, which were in the A/B conformation in the beginning of the simulation, it is apparent, that the force fields describe this state very differently (Fig. 24). The distribution in BSC1 has two peaks close to the crystallographic values of the pure A/A and B/B states. The population has a minimum at about 110 °, whereas in OL15, the population of the east region is significantly higher.

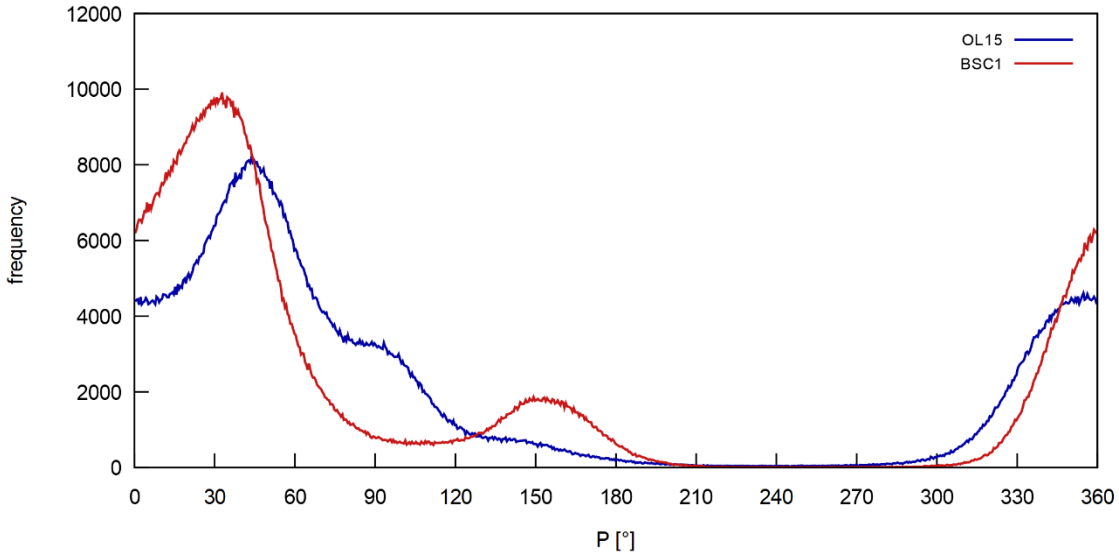

Fig. 24: Distribution of pucker angles for nucleotides, which were in the A/B state initially.

3.5.3 B/A conformations

The B/A conformation occurred in three non-terminal nucleotides in the HU architectural factor complex. This molecule was selected because of these nucleotides, which were hypothesized to be stable due to the strong protein-DNA interaction. The symmetry of the molecule also allowed to test if symmetrical nucleotides behaved identically. Fig. 25 shows one of the symmetrical nucleotides. The state is very stable in the BSC1 force field with almost no deviation from the initial values. In OL15, the glycosidic angle deviates from the initial *anti* conformation only in the beginning. The sugar conformation is not as stable as in BSC1, with the pucker visiting states closer to the east conformation, but it could still be considered stable, because it never leaves the B/A region for extended periods. The symmetrical residue behaves similarly, but in OL15, the glycosidic angle changes to the high-*anti* conformation for a portion of the simulation.

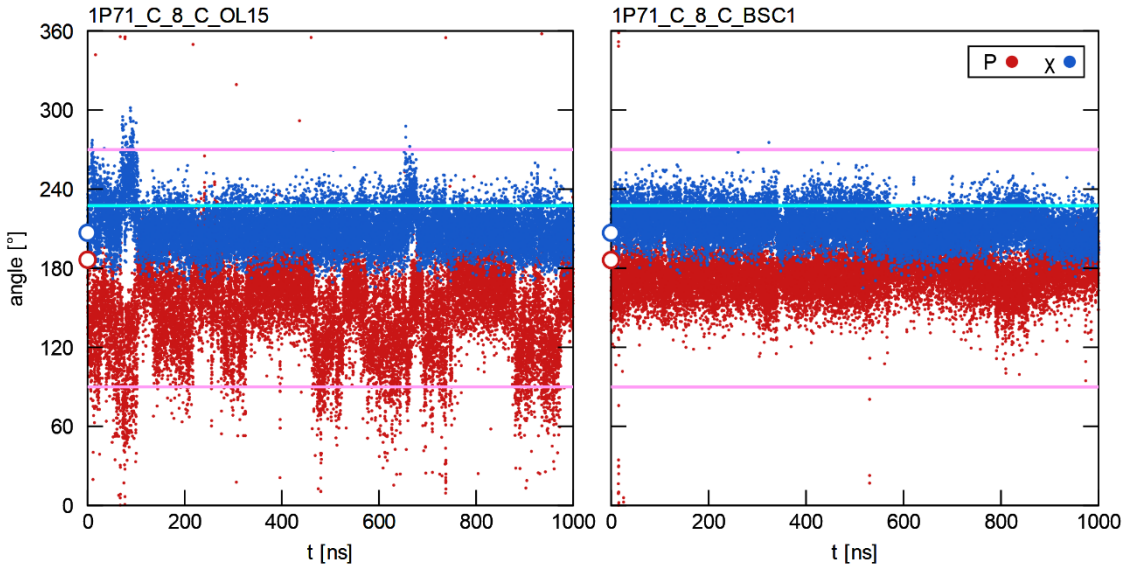

Fig. 25: Comparison of the development of the χ and pucker for residue 8 on chain C (cytosine) in the HU architectural factor complex in both force fields (OL15 left, BSC1 right). Division lines between N and S pucker are pink, the division line between anti and high-anti χ is in cyan, points on the y-axis mark the initial values

B/A states can occur in equilibria with other states, mainly B/B. Fig. 26 shows the glycosidic angle and the pucker for two terminal nucleotides, which have been excluded from the preceding analyses, because they reconform easily, which usually leads to less reliable results. In these cases, however, an equilibrium between the B/A and the B/B state is established in the BSC1 force field. It is apparent, that in the B/A state the pucker shifts to higher values than in the B/B state with less fluctuation. This is unexpected, because the pucker of the B/A state is not between the pucker of the A/A and the B/B state, but even higher than the canonical B/B pucker. The shift occurs with all B/A states in other nucleotides, the ones shown in Fig. 26 exemplify this trend most clearly. In OL15, these two nucleotides are either in the B/B state or start to fray and flip to the *syn* conformation. The shift to higher pucker values does not appear as distinctly, because the stability of east puckers is significantly higher in this force field.

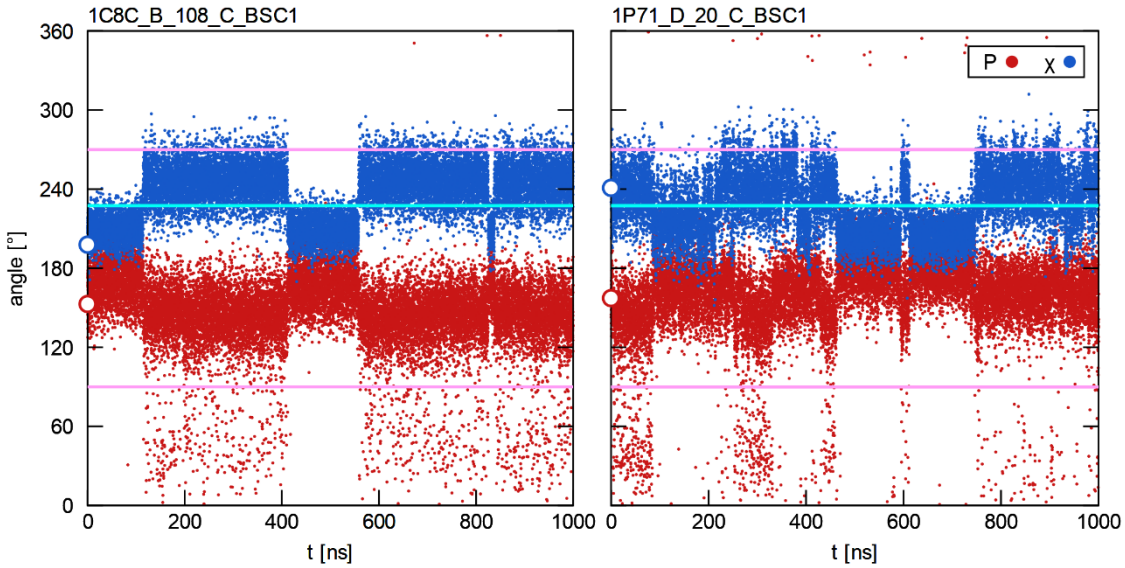

Fig. 26: Development of χ and the pucker for two terminal cytosine nucleotides in the BSC1 force field, which have stable portions of the B/A conformation. Division lines between N and S pucker are pink, the division line between anti and high-anti χ is in cyan, points on the y-axis mark the initial values

3.5.4 B/B conformations

The canonical B/B conformation is the most prevalent state of DNA. In total, 50 of the 72 examined non-terminal nucleotides were in the B/B state initially. Because of the relatively large sample size, there was a wider variety in behaviour. Most of them remained close to the crystallographic values, as shown in Fig. 27. The stable B/B states were described very similarly in OL15 and BSC1.

Unstable B/B states were rare in the simulations. One example of such a case is shown in Fig. 28 for the nucleotide 11 on chain D in the HU architectural factor complex. The pucker in the BSC1 simulation switches to the north state almost immediately and stays there for the duration of the simulation. The stability of the pucker on the symmetrical nucleotide 11 on chain C in the complex is reversed. This behaviour is unusual, but it seems to have a structural reason. A possibility is allosteric conformational change, where the protein randomly forces one nucleotide into one conformation and does not allow the same change in another place in the DNA.

Fig. 29 shows a nucleotide behaving similarly in both force fields. The glycosidic angle remains around the starting value between *anti* and high-*anti*. The pucker is in an equilibrium between north and south, but the east pucker is less populated in the BSC1 force field.

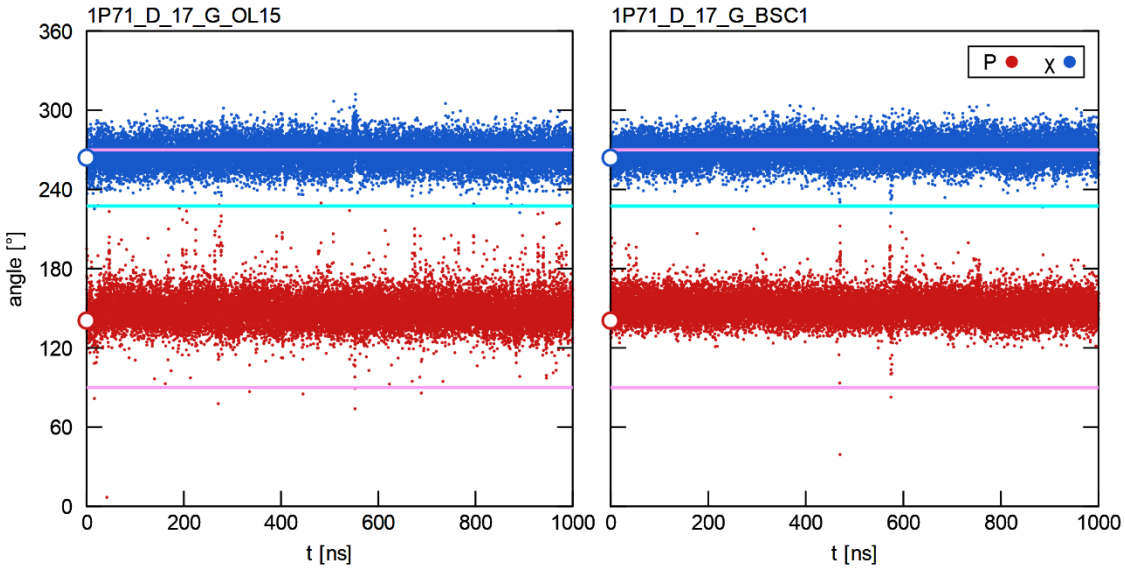

Fig. 27: Comparison of the development of the χ and pucker for residue 17 on chain D (guanine) in the HU architectural factor complex in both force fields (OL15 left, BSC1 right). Division lines between N and S pucker are pink, the division line between anti and high-anti χ is in cyan, points on the y-axis mark the initial values

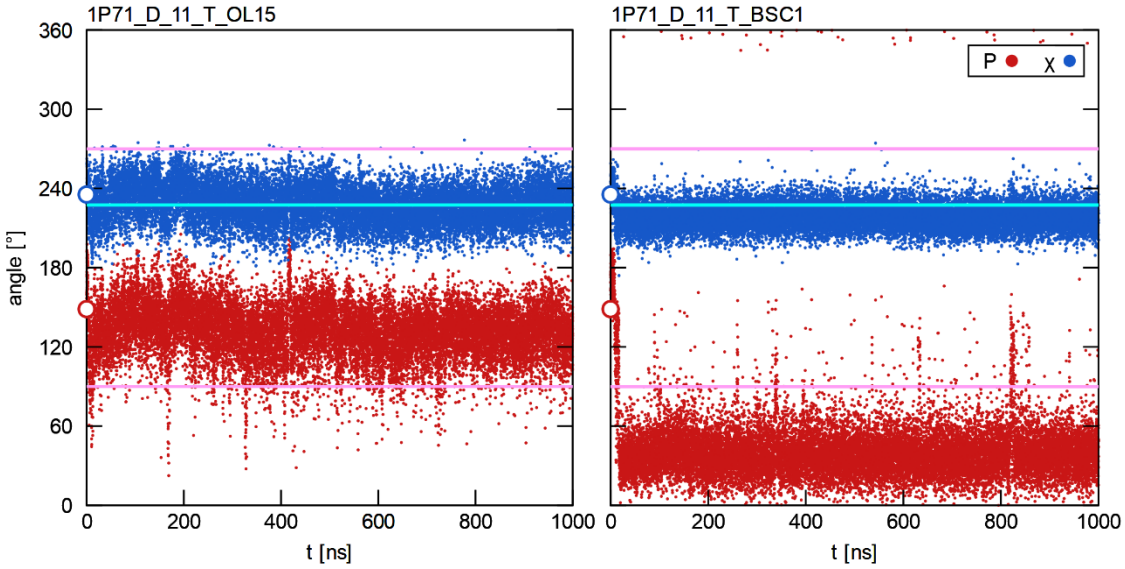

Fig. 28: Comparison of the development of the χ and pucker for residue 11 on chain D (thymine) in the HU architectural factor complex in both force fields (OL15 left, BSC1 right). Division lines between N and S pucker are pink, the division line between anti and high-anti χ is in cyan, points on the y-axis mark the initial values

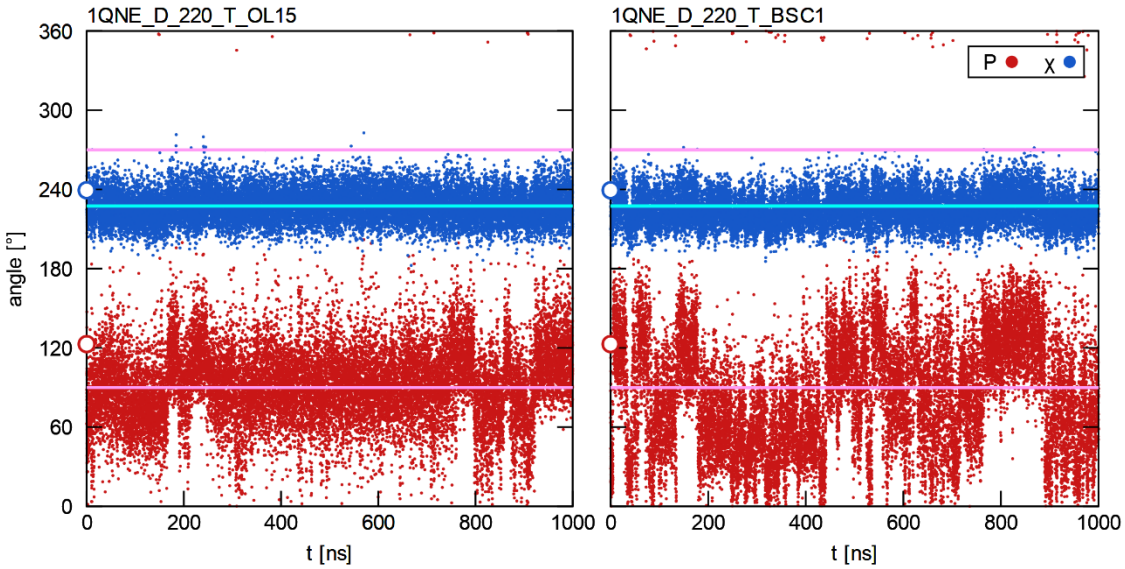

Fig. 29: Comparison of the development of the χ and pucker for residue 220 on chain D (thymine) in the TATA-box-binding protein complex in both force fields (OL15 left, BSC1 right). Division lines between N and S pucker are pink, the division line between anti and high-anti χ is in cyan, points on the y-axis mark the initial values

3.5.5 Comparison of stability in both force fields

The P/χ states were assigned for all non-terminal nucleotides in both force fields using the thresholds described in Section 3.4. The bar chart in Fig. 30 shows the stability of the four initial states. The A/A conformation is the least stable of all, with an almost 50 % transition to B/B in OL15 and almost 40 % in BSC1. The pure A conformation is slightly more stable in BSC1 than in OL15. The A/B conformation is more stable in OL15, but the differences are only minor. The behaviour of the B/A state is significantly different in the force fields. In OL15, the state transitions to the pure B form for about half of the frames and remains in the B/A domain for the other half. The B/A state was almost twice as stable in BSC1 when the initial conformation was B/A. The pure B conformation behaves very similarly in both force fields and is expectedly very stable. It remains in its initial conformation for about 80 % of the time in both force fields.

The sample size of nucleotides is relatively small (A/A=3, A/B=16, B/A=3, B/B=50), which may affect the statistical significance of the results when transferred to another set of nucleotides. The analysis is adequate for the comparison of the fore fields because it was conducted on the same nucleotides.

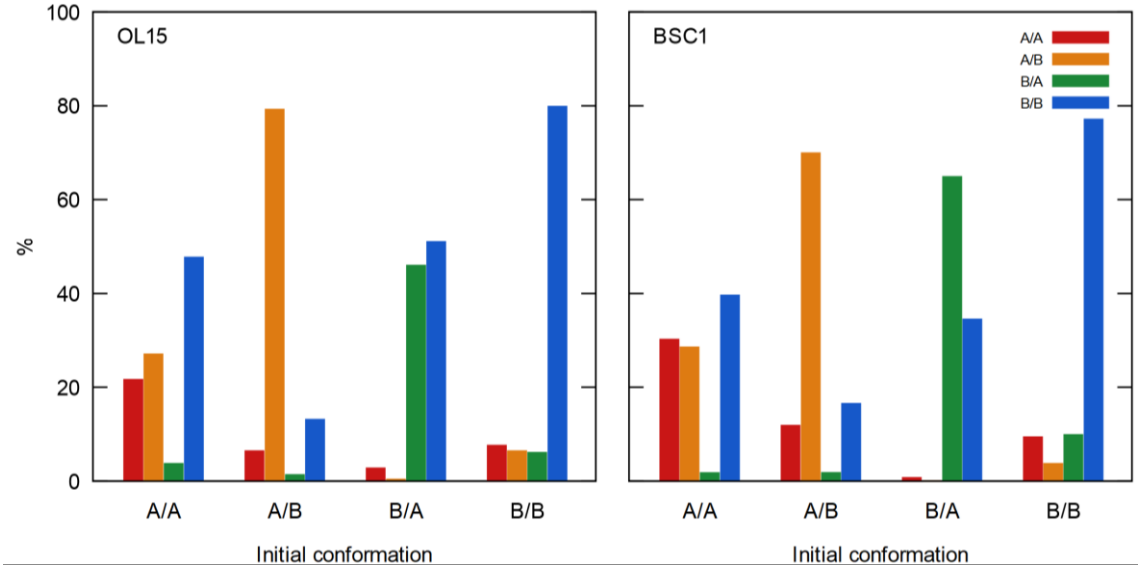

Fig. 30: Bar chart of the states, in which nucleotides occur, divided up by initial states in both force fields. All non-terminal nucleotides from all three complexes are included.

The instability of B/A states is most likely caused by the χ parametrization. BSC1 stabilizes the *anti* conformation of the glycosidic angle in the combination with the south pucker more than OL15. The pucker angle remains south for B/A states in both force fields.

The same effect causes the higher stability of A/B states in OL15. A/A states are more likely to transition to the A/B form than in BSC1. The role of the glycosidic bond, however, is minor in comparison to the effect pucker conformation has on the A/A instability. The majority of A/A states transition to the B/B state for most of the simulation time, which is unexpected, because both P and χ change conformation.

The explanation for this is the different distribution of pucker angles in the two A/χ states. A/A has a very narrow stable region close to the canonical value of about 18 °. In A/B, the pucker angle can be very variable with usually wide regions of stability at higher values closer to the east puckers. For this reason, the description of the pucker in A/B states is more reliable than in A/A states.

3.6 Stability of unusual NtC states

The analysis of the stability of NtC nucleotide states offers a different view on the problem. In this section, the stability of the NtC states AA02 and BB16 will be analysed, using the reference structures presented by Schneider et al.[8]. The states from the simulations were assigned by the smallest distance to the 95 NtC states of dinucleotides using distance of the 9 angles in Euclidian space with sampling every 100 frames, which corresponds to a sampling interval of 1 ns, which is fine enough to detect major conformational changes. In Tab. 3, the results of the analysis are shown. For each step, the two most frequently occurring conformations are shown. The occurrence of the third conformer was lower than 20 % in all cases.

With most dinucleotides, the initial state was visited most frequently. The second most occurring conformation is always a similar state, with either the pucker or χ changing. In four nucleotides, however, this isn't the case.

The AA02 state in dinucleotide 113 on chain C in the chromatin protein complex is unstable in BSC1, because the pucker on the C5'-end changes to the B conformation, as previously shown in Fig. 23. As mentioned before, the pucker of the AA02 state is closer to the east pucker, which is less stable in BSC1 than in OL15.

The dinucleotide 14 on chain C in the HU architectural factor complex is less stable in BSC1. It converts to a state with a south pucker and a high-*anti* χ. In OL15 it establishes an equilibrium between many different states, which is the reason for a comparatively low portion of frames being in the AA02 conformation.

The cytosine dinucleotide 219 in chain D of the TATA-box-binding protein also establishes an equilibrium in both force fields. The C3'-end of the dinucleotide (nucleotide 220, shown in Fig. 29) starts in the south region, but frequently changes between north and south, mostly remaining around the east pucker. The glycosidic angle remains around its initial value between *anti* and high-*anti*.

An unexpected change in conformation is found in nucleotide 204 in the TATA-box-binding protein complex. It starts in AA02 and converts to the pure A conformation AA00 during the simulation because the glycosidic angle decreases on the C5'-nucleotide. AA00 is the most occurrent state in both fore fields, with the stability being different. The north pucker is more stable in BSC1 which leads to increased stability of the pure A-DNA state.

Tab. 3: Comparison of the stability of unusual conformations in both force fields. The second column specifies the number of the C5'-end residue of the dinucleotide. The numbers in in the columns of the forcefields are the percentages of frames, in which the dinucleotide is in this conformation.

Mol.	**C5' end**	**NtC**	**OL15**				**BSC1**			
1C8C	113	AA02	AA02	96.8	AA09	1.1	BA05	48.9	BA01	31.9
1P71	8	BB16	BB16	71.6	IC06	10.4	BB16	95.1	BB00	1.5
1P71	C_14	AA02	AA02	34.9	AB01	15.8	BA17	69.6	BA08	11.7
1QNE	219	BB16	BA05	40.9	BA01	18.3	BA05	36.1	BA01	21.5
1QNE	204	AA02	AA00	47.7	AA02	27.8	AA00	70.2	AA02	13.6
1QNE	205	AA02	AA02	69.7	AB01	26.4	AA02	81.2	AB01	9.6
1QNE	206	AA02	AA02	95.0	AB01	2.0	AA02	97.9	AB01	0.8
1QNE	207	AA02	AA02	95.3	AB01	1.5	AA02	99.5	BB01	0.2
1QNE	208	AA02	AA02	93.5	BA05	4.2	AA02	98.6	BA05	0.9
1QNE	221	AA02	AA02	85.9	AB01	10.1	AA02	81.0	AA00	10.9
1QNE	222	AA02	AA02	91.3	AA09	7.9	AA02	88.8	AA09	10.8
1QNE	223	AA02	AA02	94.3	AB01	4.0	AA02	98.5	AB01	1.2
1QNE	224	AA02	AA02	93.6	BA05	3.2	AA02	97.7	AA09	0.8

4 Summary

Three protein-DNA complexes were simulated in two force fields, OL15 and BSC1, to assess, if the recent modifications improved the description of the intermediate states between A- and B-DNA. The complexes contained nucleotides, which had a combination of a typical B sugar pucker angle and a glycosidic angle typical for A-DNA and vice versa, which represent the two A-B intermediate states.

The analysis of the stability of states shows different stability of the north pucker depending on the conformation about the glycosidic bond χ. States with high-*anti* (B-like) χ conformations had a broader distribution of pucker angle than those combined with χ in the A-like *anti* conformation. The north pucker in the A/B conformations is shifted to the east region in some cases with some nucleotides being stable at that pucker and therefore the east conformation not being a transitional state, but rather a new distinct conformation. Thus, the description of the pucker in protein-DNA complexes should be considered a three-state problem with north, east and south conformations, rather than the frequently assumed two-state problem.

Associated with this is the debatable use of the δ backbone angle as a descriptor of the sugar conformation. The differences between the δ in different conformations are smaller than the corresponding differences in the pucker. The pucker angle is thus a more precise and unambiguous description of the sugar conformation, also because it includes all five dihedral angles in the sugar ring.

From the results in Section 3.5.5, it is evident, that both force fields still require modification in the description of the north pucker. The stability of the pure A/A states is most likely underestimated. Most nucleotides initially in this state convert to the B/B state with the pucker in the south region. A stabilization of the canonical north pucker at about 18 ° would improve the stability of A-DNA conformations. To a lesser extent this also holds true for A/B states, which tend to convert to B/B as well.

The A/A to A/B transition is more prevalent in the OL15 force field, which is caused by the glycosidic angle χ being more stable in the high-*anti* region. A modification to OL15 to stabilize the glycosidic angle in the *anti* state would improve the stability of those states, but also decrease the stability of the high-*anti* state, which could lead to an unexpected transitioning of the B/B state to a B/A state and overestimation of the B/A state's stability. Therefore, any such future modification needs to be carefully tested to ensure the balanced description of the equilibria described above.

The stability of the canonical B/B conformation is well-described in both force fields. It converts to other conformations to a small extent, which is to be expected, because the potentials of the angles are relatively flat and the energy barriers between states are low.

The geometries of all states are described rather well, but the thermodynamic stability of the unusual states previously described are very low compared to the stability of canonical ones.

The stability of NtC states was similar in both force fields with the aforementioned differences, like the increased stability of the north pucker in BSC1 increasing the percentage of frames, in which states similar to A-DNA occur.

In conclusion, both force fields, OL15 and BSC1 are lacking in precision when describing the thermodynamic stability of the A/A state and the B/A state. A stabilization of the north pucker would improve the description of pure A-DNA and a modification to the glycosidic angle could improve the stability of all states in the *anti* region.

These results are part of a manuscript submitted to a peer reviewed journal, which presents a more comprehensive view on the stability of all types of nucleotide states performed on 8 protein-DNA complexes in three force fields. The sample size of all conformations is larger in this manuscript, and in addition a different modification of the OL15 force field to the glycosidic angle, χOL3, has been tested.[27]

5 References

(1) D.A. Case, H.M. Aktulga, K. Belfon, I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, G.A. Cisneros, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, G. Giambasu, M.K. Gilson, H. Gohlke, A.W. Goetz, R. Harris, S. Izadi, S.A. Izmailov, C. Jin, K. Ka, and P. A. K. Amber 2021. University of California, San Francisco.

(2) Zgarbová, M.; Šponer, J.; Otyepka, M.; Cheatham, T. E.; Galindo-Murillo, R.; Jurečka, P. Refinement of the Sugar-Phosphate Backbone Torsion Beta for AMBER Force Fields Improves the Description of Z- and B-DNA. *J. Chem. Theory Comput.* **2015**, *11* (12), 5723–5736. https://doi.org/10.1021/ACS.JCTC.5B00716.

(3) Ivani, I.; Dans, P. D.; Noy, A.; Pérez, A.; Faustino, I.; Hospital, A.; Walther, J.; Andrio, P.; Goñi, R.; Balaceanu, A.; et al. Parmbsc1: A Refined Force Field for DNA Simulations. *Nat. Methods* **2016**, *13* (1), 55–58. https://doi.org/10.1038/NMETH.3658.

(4) The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.

(5) Modrich, P. DNA MISMATCH CORRECTION. *https://doi.org/10.1146/annurev.bi.56.070187.002251* **2003**, *56*, 435–466. https://doi.org/10.1146/ANNUREV.BI.56.070187.002251.

(6) Altona, C.; Sundaralingam, M. Conformational Analysis of the Sugar Ring in Nucleosides and Nucleotides. A New Description Using the Concept of Pseudorotation. *J. Am. Chem. Soc.* **1972**, *94* (23), 8205–8212. https://doi.org/10.1021/JA00778A043.

(7) Rohs, R.; Jin, X.; West, S. M.; Joshi, R.; Honig, B.; Mann, R. S. Origins of Specificity in Protein-DNA Recognition. **2010**. https://doi.org/10.1146/annurev-biochem-060408-091030.

(8) Černý, J.; Božíková, P.; Svoboda, J.; Schneider, B. A Unified Dinucleotide Alphabet Describing Both RNA and DNA Structures. *Nucleic Acids Res.* **2021**, *48* (11), 6367–6381. https://doi.org/10.1093/NAR/GKAA383.

(9) Mohr, S. C.; Sokolov, N. V. H. A.; Chaomei, H.; Setlow, P. Binding of Small Acid-Soluble Spore Proteins from Bacillus Subtilis Changes the Conformation of DNA from B to A. *Proc. Natl. Acad. Sci. U. S. A.* **1991**, *88* (1), 77. https://doi.org/10.1073/PNAS.88.1.77.

(10) Patikoglou, G. A.; Kim, J. L.; Sun, L.; Yang, S. H.; Kodadek, T.; Burley, S. K. TATA Element Recognition by the TATA Box-Binding Protein Has Been Conserved throughout Evolution. *Genes Dev.* **1999**, *13* (24), 3217–3230. https://doi.org/10.1101/gad.13.24.3217.

(11) Feig, M.; Montgomery Pettitt, B. A Molecular Simulation Picture of DNA Hydration around A- And B-DNA. *Biopolymers* **1998**, *48* (4), 199–209.

https://doi.org/10.1002/(SICI)1097-0282(1998)48:4<199::AID-BIP2>3.0.CO;2-5.

(12) Su, S.; Gao, Y. G.; Robinson, H.; Liaw, Y. C.; Edmondson, S. P.; Shriver, J. W.; Wang, A. H. J. Crystal Structures of the Chromosomal Proteins Sso7d/Sac7d Bound to DNA Containing T-G Mismatched Base-Pairs. *J. Mol. Biol.* **2000**, *303* (3), 395–403. https://doi.org/10.1006/jmbi.2000.4112.

(13) Swinger, K. K.; Lemberg, K. M.; Zhang, Y.; Rice, P. A. Flexible DNA Bending in HU-DNA Cocrystal Structures. *EMBO J.* **2003**, *22* (14), 3749–3760. https://doi.org/10.1093/emboj/cdg351.

(14) Juo, Z. S.; Chiu, T. K.; Leiberman, P. M.; Baikalov, I.; Berk, A. J.; Dickerson, R. E. How Proteins Recognize the TATA Box. *J. Mol. Biol.* **1996**, *261* (2), 239–254. https://doi.org/10.1006/JMBI.1996.0456.

(15) Dame, R. T.; Goosen, N. HU: Promoting or Counteracting DNA Compaction? *FEBS Lett.* **2002**, *529* (2–3), 151–156. https://doi.org/10.1016/S0014-5793(02)03363-X.

(16) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235–242. https://doi.org/10.1093/NAR/28.1.235.

(17) Wendy D. Cornell; Piotr Cieplak; Christopher I. Bayly; Ian R. Gould; Kenneth M. Merz, J. .; David M. Ferguson; David C. Spellmeyer; Thomas Fox; James W. Caldwell, and; Kollman*, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules J. Am. Chem. Soc. 1995, 117, 5179−5197. *J. Am. Chem. Soc.* **1996**, *118* (9), 2309–2309. https://doi.org/10.1021/JA955032E.

(18) Galindo-Murillo, R.; Robertson, J. C.; Zgarbová, M.; Šponer, J.; Otyepka, M.; Jurečka, P.; Cheatham, T. E. Assessing the Current State of Amber Force Field Modifications for DNA. *J. Chem. Theory Comput.* **2016**, *12* (8), 4114–4127. https://doi.org/10.1021/acs.jctc.6b00186.

(19) Zgarbová, M.; Šponer, J.; Otyepka, M.; Cheatham, T. E.; Galindo-Murillo, R.; Jurečka, P. Refinement of the Sugar-Phosphate Backbone Torsion Beta for AMBER Force Fields Improves the Description of Z- and B-DNA. *J. Chem. Theory Comput.* **2015**, *11* (12), 5723–5736. https://doi.org/10.1021/ACS.JCTC.5B00716/SUPPL_FILE/CT5B00716_SI_001.PDF.

(20) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. The Missing Term in Effective Pair Potentials. *J. Phys. Chem.* **2002**, *91* (24), 6269–6271. https://doi.org/10.1021/J100308A038.

(21) Joung, I. S.; Cheatham, T. E. Determination of Alkali and Halide Monovalent Ion Parameters for Use in Explicitly Solvated Biomolecular Simulations. *J. Phys. Chem. B* **2008**, *112* (30), 9020–9041.

https://doi.org/10.1021/JP8001614/SUPPL_FILE/JP8001614-FILE003.PDF.

(22) Joung, S.; Cheatham, T. E. Molecular Dynamics Simulations of the Dynamic and Energetic Properties of Alkali and Halide Ions Using Water-Model-Specific Ion Parameters. *J. Phys. Chem. B* **2009**, *113* (40), 13279–13290. https://doi.org/10.1021/JP902584C/SUPPL_FILE/JP902584C_SI_001.PDF.

(23) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. Ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from Ff99SB. *J. Chem. Theory Comput.* **2015**, *11* (8), 3696–3713. https://doi.org/10.1021/ACS.JCTC.5B00255/SUPPL_FILE/CT5B00255_SI_001.PDF.

(24) Darden, T.; Perera, L.; Li, L.; Lee, P. New Tricks for Modelers from the Crystallography Toolkit: The Particle Mesh Ewald Algorithm and Its Use in Nucleic Acid Simulations. *Structure* **1999**, *7* (3), R55–R60. https://doi.org/10.1016/S0969-2126(99)80033-1.

(25) Roe, D. R.; Cheatham, T. E. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J. Chem. Theory Comput.* **2013**, *9* (7), 3084–3095. https://doi.org/10.1021/CT400341P/SUPPL_FILE/CT400341P_SI_001.PDF.

(26) Zgarbová, M.; Jurečka, P.; Šponer, J.; Otyepka, M. A- to B-DNA Transition in AMBER Force Fields and Its Coupling to Sugar Pucker. *J. Chem. Theory Comput.* **2018**, *14* (1), 319–328. https://doi.org/10.1021/acs.jctc.7b00926.

(27) Jurečka, P.; Zgarbová, M.; Černý, F.; Salomon, J. Continuous B- to A-Transition in Protein-DNA Binding - How Well Is It Described by Current AMBER Force Fields? *bioRxiv* **2022**, 2022.01.13.476176. https://doi.org/10.1101/2022.01.13.476176.