UNIVERZITA KARLOVA V PRAZE

Přírodovědecká fakulta

Studijní program: Chemie

Studijní obor: Biofyzikální chemie

Bc. Paulína Božíková

Bioinformatická analýza interakcí mezi proteiny a DNA

Bioinformatic analysis of protein/DNA interactions

Diplomová práce

Vedoucí závěrečné práce: doc. Ing. Bohdan Schneider, CSc.

Praha 2015

Prohlášení:

Prohlašuji, že jsem závěrečnou práci zpracovala samostatně a že jsem uvedla všechny použité informační zdroje a literaturu. Tato práce ani její podstatná část nebyla předložena k získání jiného nebo stejného akademického titulu.

V Praze, 14. 8. 2015

Poděkování:

Ráda bych poděkovala doc. Bohdanovi Schneiderovi za vedení, ochotu a trpělivost, kterou se mnou měl při zpracování diplomové práce. Taky bych chtěla poděkovat Jiřímu Černému za pomoc, kterou významně přispěl ke vzniku této práce.

Abstrakt:

V této práci jsme analyzovali lokální strukturu DNA páteře v DNA vázané v komplexech s proteiny a volné DNA, a závislost lokální struktury na typu párování a na sekvenci. Za tím účelem jsme analyzovali přibližně 1400 krystalových struktur DNA v komplexech s proteiny a více než 400 krystalových struktur DNA nevázané v komplexech s proteiny. Lokální konformace DNA byly klasifikovány do 38 tříd dinukleotidových konformerů ntC popsaných dříve (Svozil et al. Nucleic Acids Res. 2008), které byly dále sdruženy do 16 klastrů strukturní abecedy ntA, abychom redukovali počet analyzovaných proměnných. Strukturní třídy ntA dinukleotidů tvořících páry bazí v dvoušroubovicích DNA byly uspořádány do tzv. Asociačních matic tak, že řádky a sloupce matic jsou označeny třídami ntA právě vázaných dinukleotidů. Analyzovali jsme tři základní matice. Dvě pro dinukleotidy vázané pouze do Watson-Crickových párů v protein/DNA komplexech, respektive v samotných DNA. Třetí matice byla sestrojena pro dinukleotidy vázané i jinými než Watson-Crickovými páry. Asociační matice jsme rovněž zkoumali v závislosti na sekvenci přispívajících dinukleotidů. Provedené analýzy ukazují rozdíly ve strukturním chování různých strukturních tříd ntA a jejich sekvenční závislosti.

Klíčová slova

molekulární interakce, specificita, protein, DNA, krystalografie, databáze

Abstract:

In this thesis, we focused on local structural features of the DNA backbone in protein-complexed DNA and non-complexed (naked) DNA, and its dependence on types of a base pairing in DNA, and on the base sequence. To reach this goal we analyzed about 1,400 crystal structures of DNA in complexes with proteins and more than 400 crystal structures of naked DNA. DNA local conformations were structurally classified into 38 dinucleotide conformers ntCs, which were described previously (Svozil et al. Nucleic Acids Res. 2008). The ntC were further clustered into 16 structural alphabet classes ntA to reduce the number of analyzed variables. We assembled base-paired dinucleotides from double helical DNA structures according to their assigned structural alphabet classes into so called Association matrices. Three basic Association matrices were analyzed; two compare ntA/ntA associations between dinucleotides forming only Watson-Crick base pairs in protein/DNA complexes and in naked DNA, respectively; the third one ntA/ntA associations between dinucleotides base-paired also by non-Watson-Crick pairs. We also analyzed Association matrices of dinucleotides as a function of their sequences. The analyzes revealed differences in structural behavior of various ntA and their dependence on dinucleotide sequences.

Keywords:

molecular interaction, specificity, protein, DNA, crystallography, databases

Contents

Chapter 1

DNA Architecture

1.1 DNA Basics

DNA (Deoxyribonucleic Acid) is a polymeric molecule built from monomeric units, nucleotides (Figure 1.1), which form long strands. Nucleotides are composed of two parts, phosphate group and nucleoside, bound together by phosphodiester bond. Nucleosides can be further decomposed into two fragments, which are ribose or deoxyribose sugar unit and one of four nitrogenous bases. Thus, four possible types of nucleotides are present in DNA. Sugar in DNA is a pentose deoxyribose; RNA contains ribose.

Phosphodiester bond P-O links together the deoxyribose with phosphate group via -OH groups. Positions of these functional groups are shown in Figure 1.1 The bases are four types of aromatic nitrogenous heterocycles with NH_2 or keto groups as functional groups. Nucleic acids contain two types of bases, purines (adenine and guanine) and pyrimidines (thymine and in RNA uracil, and cytosine).

Sugar and a base are bound together by glycosidic bond between C1' atom of the sugar ring and a nitrogen of the base; N1 in pyrimidines and N9 in purines. The glycosidic bond in natural nucleic acids is always in β conformation which means that a base is on the same side of the sugar as the phosphate group attached to the 5' position of the sugar above the plane of the sugar (Figure 1.1).

Phosphate groups together with sugars attached by the phosphodiester bonds, are called sugar-phosphate backbone of DNA or RNA strands.

A double-stranded helix is a common structural motif of DNA. It is characterized

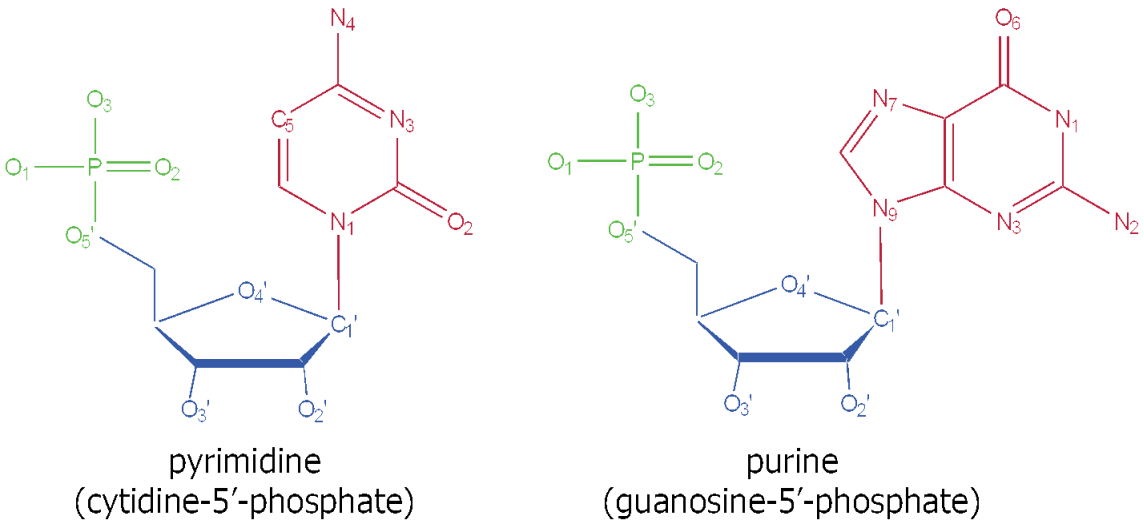


Figure 1.1: **Two examples of the base types.** On the left, cytosine (in red color), is attached through its N1 atom to the sugar (in blue color) by glycosidic bond. The glycosidic bond is between N1 atom of pyrimidine and C1' atom of the sugar. The phosphate group is depicted in green color. On the right, purine base guanine is attached to the sugar through its N9 atom.

by two strands running in opposite directions and stabilized by hydrogen bonds between the complementary bases. The nucleic acid helix has hydrogen bonded bases facing toward the center of the helix while the sugar-phosphate backbone is on the outside of the helix fully exposed to solvent. The backbone is typically negatively charged because one hydroxyl of each phosphate group is under normal, "physiological", pH ionized so that each phosphate group carries one negative charge. This property therefore gives the backbone and also the DNA the ability to interact with positively charged residues or ions [1].

1.2 DNA Building Blocks

1.2.1 Base Pairing

Two strands of DNA are held together by hydrogen bonds between the bases. The term base pairing refers to a connection between bases via hydrogen bonds. There are several types of base pairing differing in combinations of connected bases, a number of hydrogen bonds between bases, and their mutual orientation. The bonds are provided by nitrogens of heterocycles of bases and functional groups attached

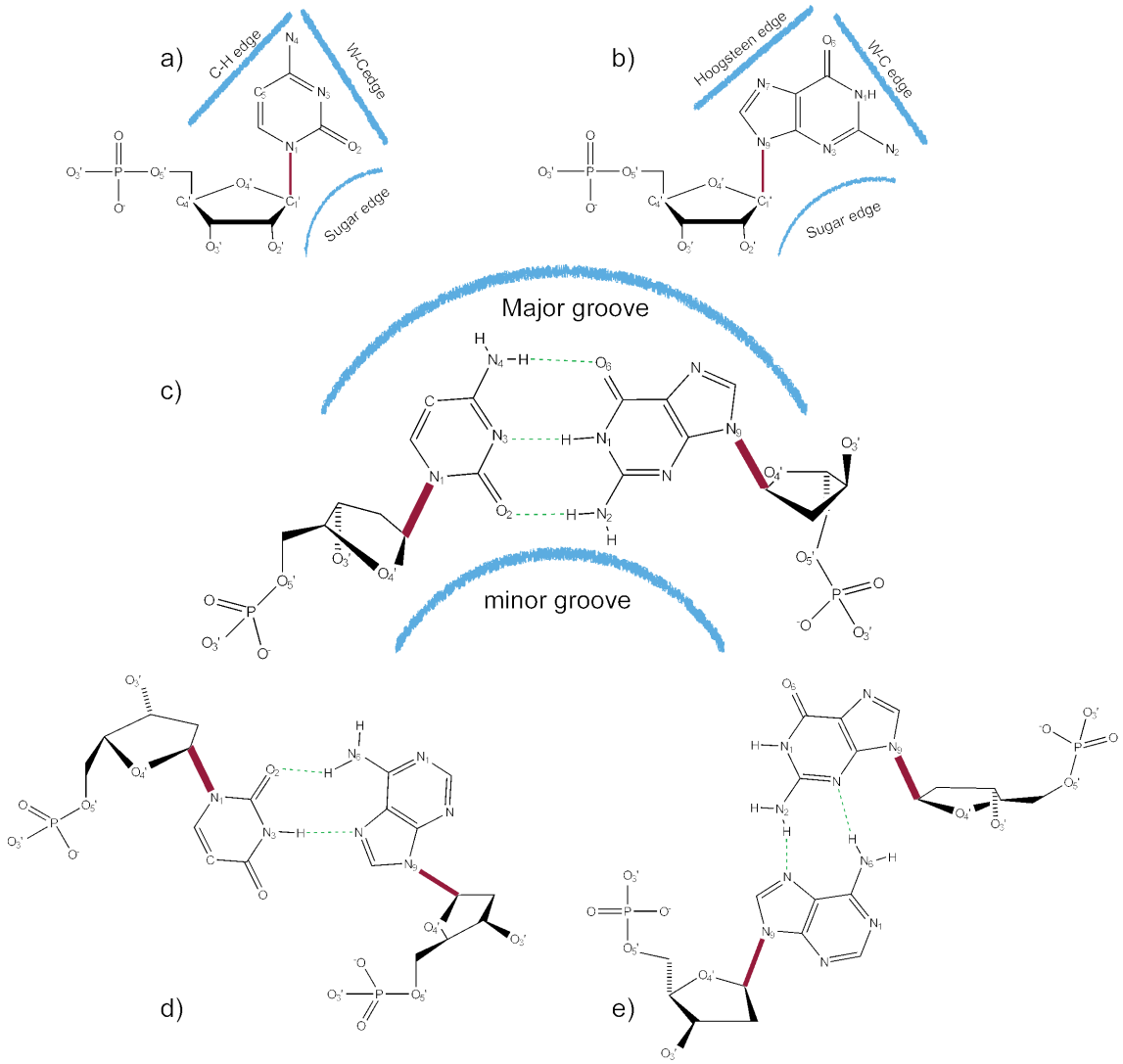


Figure 1.2: **Base pairs patterns.** a) and b) show the base edges via which they can interact with other molecules, c) shows Watson-Crick base pair between cytosine and guanine and two structural features, the major and the minor grooves. d) and e) show non-Watson-Crick base pairs.

to the heterocycles. One type of base pairing is called according to Watson and Crick Watson-Crick (W-C), or sometimes canonical pairing [2]. In a W-C pair, a pyrimidine base always stands opposite to a purine base: thymine (pyrimidine) is bound via two hydrogen bonds with adenine (purine), and cytosine (pyrimidine) is bound via three hydrogen bods with guanine (purine) as it is seen in Figure 1.2 c). Thus, thymine is complementary to adenine and cytosine to guanine.

Other base pairing patterns are also possible although they are not common in DNA; their role in RNA is however important as they stabilize RNA 3D folds. For example, purine-purine base pairs and pyrimidine-pyrimidine base pairs occur.

Several examples of base pairing are visible in Figure 1.2. A list containing 28 base pair types was created by Saenger [3]. In this list, each base pair type is assigned with a number from I to XXVIII. The Watson-Crick base pair between cytosine and guanine has number XIX, and the Watson-Crick base pair between adenine and thymine number XX.

1.2.2 Sugar Puckering

As mentioned before, nitrogenous bases are planar aromatic moieties. In contrast, both nucleic acid sugar units, ribose and deoxyribose, are highly non planar and this non-planarity is called puckering. There are two principle types of puckering, an envelope and a twist conformation. As for the envelope puckering, four atoms of the sugar ring are in a single plane, while the remaining atom is out of this plane. In the case of the twist conformation, two atoms are out of the plane of the other three. Moreover, one of these two atoms has a larger deviation from the plane than the other.

When one atom is out of the plane of the remaining four atoms of the sugar ring, it can be on the same side of the ring as the base and the C5' bound phosphate group, the pucker is called *endo*, otherwise, when they are on the opposite side to the base and the C5' phosphate group, the pucker is called *exo* (Figure 1.5). The predominantly occurring sugar puckers are the *endo* forms when the C2' or C3' atoms are out of the sugar ring; the corresponding puckers are called C2'-*endo* or C3'-*endo* sugar puckers, respectively [3, 4].

1.2.3 Conformations of the DNA Backbone

Torsion Angles

An important characteristics of the three-dimensional structure of a molecule is rotation of atoms around bonds. It is measured by torsion angles, that are defined by four atoms, usually connected by chemical bonds. These four atoms specify two planes and their oriented angle defines the value of torsion angle (Figure 1.3).

Nucleotide conformation can be, beside the sugar puckering, characterized by 7 torsion angles including 6 torsions of the backbone proper, α, β, γ, δ, ϵ, ζ, and one

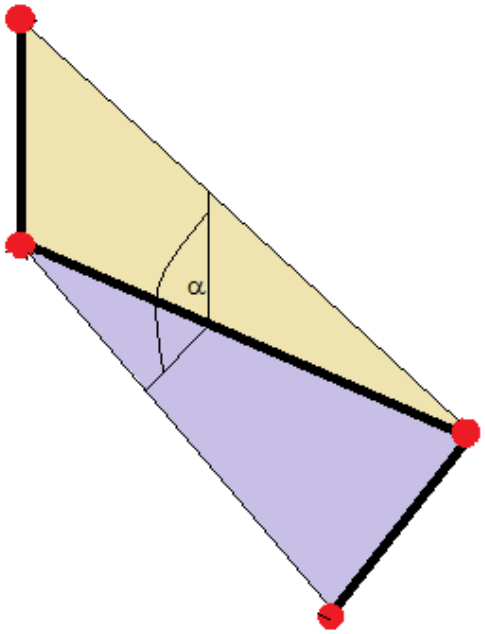


Figure 1.3: **Torsion angle.** Torsion angle is the oriented angle between two planes shown in beige and violet. Each plane is defined by three atoms.

torsion angle χ around the glycosidic bond, which determines the relative orientation between base and sugar [3].

The torsion angle α is specified by four atoms of sugar-phosphate backbone, O3'-P-O5'-C5'. The β angle is specified by P-P5'-C5'-C4' atoms; the γ angle is given by O5'C5'-C4'-C3 atoms; the δ by C5'-C4'-C3'-O3'; the ϵ by C4'-C3'-O3'-P, the ζ angle by C3'-O3'-P-O5' atoms, and finally the χ angle is specified by O4'-C1'-N9-C4 atoms for purines and by O4'-C1'-N1-C2 atoms for pyrimidines (Figure 1.4).

Preferred Values of the Backbone Torsions

Not all values of torsion angles are equally likely and there are certain preferred regions around sterically allowed conformations and not all combinations of torsions are allowed [5, 6]. DNA conformations and torsion angles are influenced by interactions with other molecules, such as proteins, drugs etc. [7–9].

A nucleotide conformation is described by 6 backbone torsion angles plus the χ torsion angle describing rotation around the bond between deoxyribose C1' and a base nitrogen (N1 in pyrimidines, N9 in purines). Important for description of DNA

conformation is torsion δ describing exocyclic rotation around C3'-C4' bond. Due to the fact that this bond is also a part of the deoxyribose ring, torsion δ correlates with the sugar pucker. The sugar pucker mode C2'-*endo*, typical for B-DNA form correlates with δ values around 130°, while the C2'-*endo* typical for A-DNA (and A-RNA) is characterized by δ 80°. The rotation about the exocyclic bond between C4' and C5' atoms, torsion angle γ, describes position of the 5'-oxygen, and therefore the position of the phosphate group attached to the C5' atom, relative to the sugar. The torsion angle ϵ, on the other hand, describes position between the sugar ring and the 3'-oxygen. The highest conformational variability is observed at the phosphodiester bonds P-O3' (torsion ζ) and P-O5' (torsion α). It has been observed that the major difference between A- and B-DNA forms lie in the δ and the χ torsion angles [3].

These and several other correlations between individual torsions have been observed but the complexity of the conformational space of nucleic acids follows from the fact that none of these correlations can describe characteristic features of the main DNA or RNA conformers. These features can be described only by combining torsion values along the polynucleotide backbone. These combinations describing typical DNA and RNA conformers have been described [5, 6, 10].

Glycosidic Torsion Angle

The glycosidic torsion angle, defined around the C1'-N bond between the sugar and a base occurs in two conformational regions, so-called *syn* and *anti*. In the *anti* conformation, the N1 and the C2 atoms of purines and the C2 and the N3 atoms of pyrimidines are directed away from the deoxyribose ring. In this conformation the hydrogen atoms that are attached to the C8 atom of purine and to the C6 atom of pyrimidine are above the sugar ring; the value of the χ is around 180° (so-called *low anti*) and 270° (*high anti*). In the *syn* conformation, the base is rotated approximately by 180° relative to its *anti* orientation (Figure 1.5). Stabilization of this conformation was observed by hydrogen bond formation between the O5' atom and the N3 base atom [4]. When the sugar and a base are in the *syn* conformation, the value of the χ angle is between 0° and 100°. When they are in the *anti* conformation, the value of the χ angle is between 170° and 280° [11].

Values of the χ torsion angle correlate with a base type attached. It means that

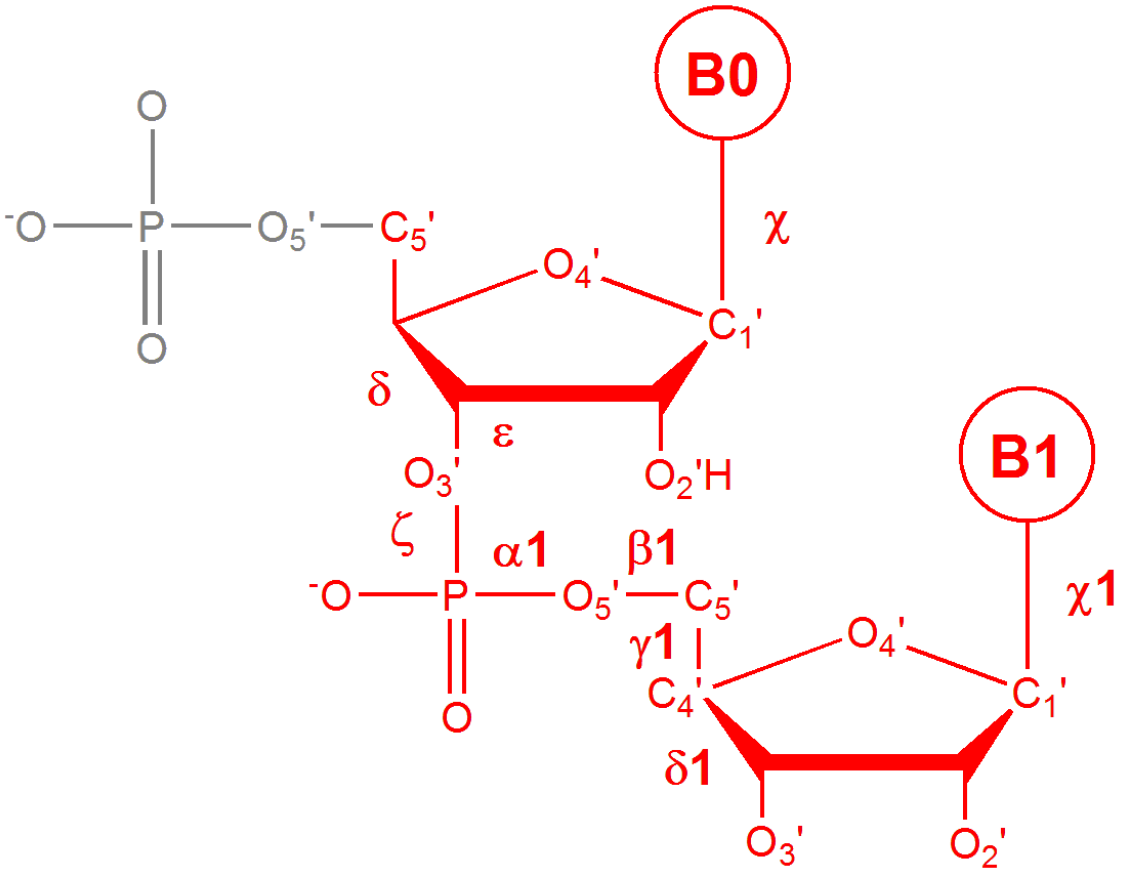


Figure 1.4: **Dinucleotide conformer.** The figure shows two nucleotides which form dinucleotide conformer characterized by nine torsion angles, assigned in the figure. The two of the angles are χ angles about glycosidic bond between the sugar and the nitrogenous base. B0 and B1 represent two nitrogenous bases of the two nucleotides.

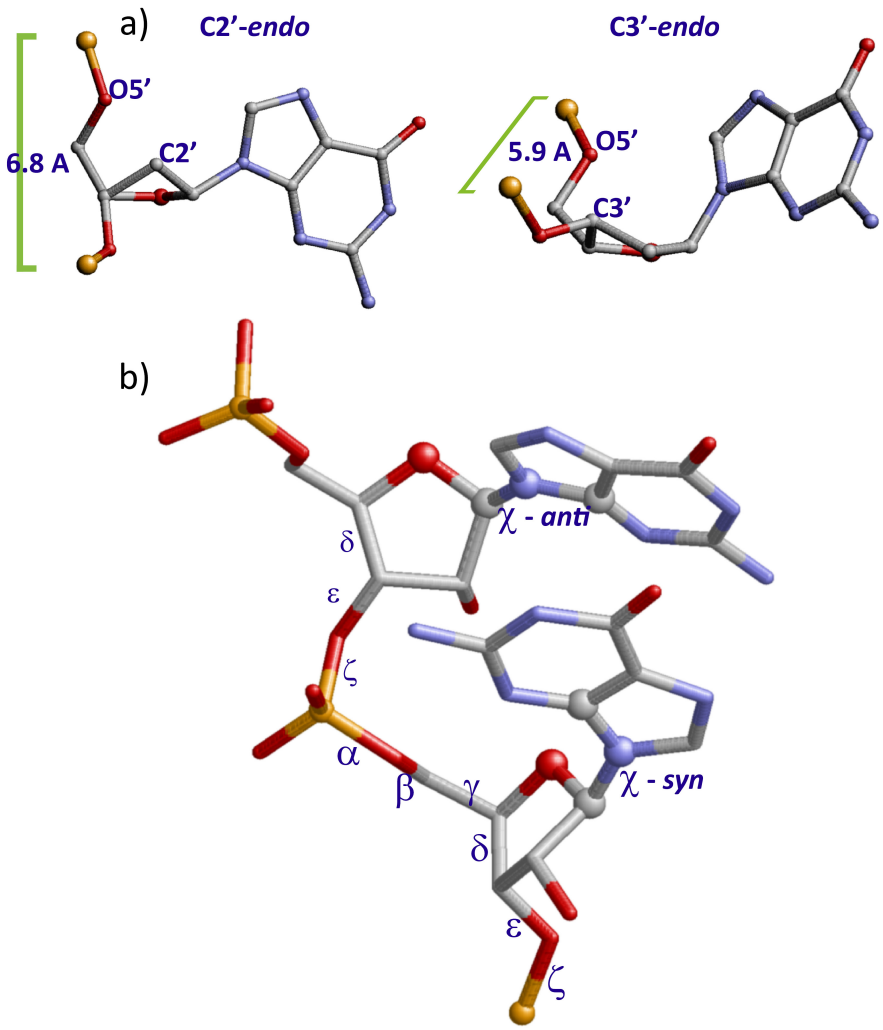


Figure 1.5: **Sugar puckering and *syn/anti* conformations.** The figure shows a) the C2'-(endo) and C3'-(endo) sugar puckering. In C2'-(endo) and C3'-(endo) puckering, C2' atom and C3' atom of the sugar ring, respectively, is placed on the same side of the sugar ring as the nitrogenous base and O5'-phosphate group; and b) shows (anti) and (syn) conformation of the χ angle. In (anti) conformation the nitrogenous base faces away from the sugar ring, while in (syn) conformation it faces toward the sugar ring.

it differs when a base attached by this glycosidic bond is a purine or a pyrimidine. Importantly, it also correlates with the sugar pucker. In the *syn* conformation there is particular steric hindrance between the sugar and a base located over the sugar ring. This hindrance can be compensated when the sugar adopts the C2'-*endo* pucker. No particular steric hindrance occurs in the *anti* conformation [3].

A variant of the *anti* conformation, called a *high anti*, was observed in structures of naturally and chemically modified nucleosides. Almost eclipsed C1'-C2' with N1'-C6 in pyrimidine or with N9-C8 in purine, are characteristic for this conformation. The value of χ is approximately 200°. [5, 12]

The C3'-*endo* puckering is linked to the *anti* orientation of a base and in the case of the C2'-*endo* puckering, the bases can be in both the *anti* and the *syn* orientation. Even though the *anti* conformation is far more common, the *syn* conformation can also occur, for example in G-quadruplex structures or in Z-DNA duplexes [6].

Dinucleotide Conformers

Svozil et al. [6] studied dinucleotide conformations in crystal structures containing DNA molecules by analyzing distributions of their torsion angles. Dinucleotide steps exhibit a substantial flexibility in a sequence-dependent manner [6]. Based on this study was defined an original DNA structural alphabet [13], which characterizes local conformations of dinucleotide units (Figure 1.4). Using the system of dinucleotide conformers organized into so called structural alphabet ntA (see Methods section), we are able to describe DNA structure by means of symbols characterizing its local structural features. Therefore, we are capable of tracking distinct conformations of free as well as complexed DNA.

In [6], each dinucleotide was characterized by 9 torsion angles. The first nucleotide at the 5'-end was described by torsions δ, ϵ and ζ, the second one was described by torsions α+1, β+1, γ+1, and δ+1. Orientations of the bases relative to the deoxyriboses are described by the χ of the first and the χ+1 of the second base.

The distributions of the torsion angles of the analyzed dinucleotides were clustered into groups (classes) where the nucleotides were considered to belong to the same cluster based on the similarity of values of their torsion angles. Some of the

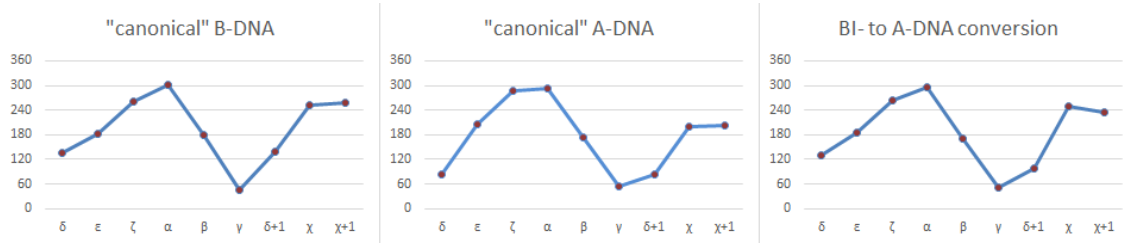


Figure 1.6: **A graphical depiction of the torsion angle values for three selected important DNA conformers.** .

resulting clusters are listed in Table 2.3 with the average values of their nine torsion angles, which characterize each group. The table shows different variants or conformers of the B-DNA form can be characterized. The most important variants shown in the table are called "canonical" BI-DNA, another BI, and BII-DNA variants.

One way of depicting nine torsion angles characterizing dinucleotide conformers are shown in Figure 1.6. Here we show torsion angles of three DNA conformers, the most common "canonical" B-form of DNA, A-form, and one conformer converting BI-form to A-form of DNA. The torsion angles used for this figure are average values of torsion angles for these dinucleotide conformers; the conformers are listed in Table 2.3. The projection shows that for all three groups almost the same values of the angles α, β and γ are characteristic but they more or less differ in other seven torsion angles.

The values of torsion angles can characterize several classes of A-form of DNA. The main one is the "canonical" A-DNA. One of these A-forms is characterized by A-like torsion angles except for the angle χ which is BI-like.

Mixed A/B types were classified as well. Clusters of the mixed forms have one nucleotide in the A-, the other in the B-form. The conversion from the B- to A-form nicely can be demonstrated by the group 32 where the O4'-*endo* sugar puckering is characterized for the conversion from one form of DNA to the other [11, 14].

Z-form variants differ according to the base type. Letter Y, in Table 2.3, is for pyrimidine base (mostly C). Letter R is for purine (mostly G). Distinct ZI and ZII conformers were described. The Z-DNA variants were described with respect to the order of the bases, concretely purine/pyrimidine step for ZI and ZII variant, or pyrimidine/purine step for Z.

Several other clusters are characteristic for other DNA conformations such as conformations with mismatched base pairs and for other DNA structures such as G-quadruplexes or four-way junctions [6, 13].

Assignment of Dinucleotide Conformers

Čech et al. (2013) provided a computer algorithm for automatic classification and assignment of the dinucleotide conformers [13]. The classification program is set up with predefined allowed values of deviations from the mean values of torsion angles. The method of k-nearest neighbor (k-NN) is used for classification. This assignment is based on the class of the nearest neighbors of the dinucleotide to be assigned. This means that the program tries to find a conformer from a set of conformers listed in Tables 2.3 and 2.4 with torsion values as close to the values of the torsion angles of the not yet assigned dinucleotide as possible. These "k nearest neighbors" are weighted by the reciprocal value of the distance of the neighbors from the dinucleotide to be assigned so that it says how "relevant" the nearest neighbors are. A dinucleotide is assigned according to the highest sum of the weighted reciprocal squared deviations which is compared to a predefined value. When all the sums are lower than this value, the dinucleotide is not assigned to any class of dinucleotide conformers. [13].

The assignment of the step conformations used in this work was substantially improved compared to the just described original method by Čech et al. [13]. Most importantly, the standard set of dinucleotides against which all assignments are done has been critically evaluated and it now contains only steps that provide self-consistent assignment. While previously used standard set used different allowed deviations for different torsions, the cleaned up standard set uses 25.0 °, uniformly for all torsions. If the difference between any torsion angle of an unclassified dinucleotide and the corresponding torsion angles in the standard for assignment was greater than the 25.0 °, the dinucleotide was unassigned to any class. Also, physically impossible values of the δ torsion angle, less than 55.0 ° or more than 185.0 °, are left unclassified.

1.3 Double Helical DNA

Overall Architecture of the DNA Double Helix

When looking on a double helical DNA molecule in its B-form, we can notice an important structural feature, its minor and major grooves resulting from the asymmetry of the glycosidic bonds of base pairs (Figure 1.7) and Figure 1.2), the minor groove is on the side of hydrogen-bonded bases which are closer to the sugar, on this side of helix, backbones of the two strands are closer together. The major groove, on the other hand, is on the side of hydrogen-bound bases farther from the sugar, where the backbones of the helix are farther apart.

The properties of the minor and major grooves, mainly their dimensions and distribution of partial charges, are different, and their properties strongly influence the way DNA interacts with other molecules. The groove width contributes to the accessibility to the base edges that display sequence-dependent pattern of hydrogen bonding acceptors and donors and potentially thus influencing recognition of specific DNA sequences by proteins [15].

The Main Structural Forms of Double Helical DNA

Three basic forms of double-stranded DNA naturally occur according to the DNA environmental conditions. These three forms are called A-DNA, B-DNA and Z-DNA, and the first two of them are depicted in Figure 1.8 [3,4]. The basic structural features of these DNA forms are characterized by the direction of the helix rotation, the angle between base pairs and the helical axis, and the pitch of the whole turn of the helix characterized by the number of base pairs per turn. Other characteristics are the groove dimensions, preferred sugar puckers, and the helical diameter. Typical parameters of A-, B- and Z-form of DNA are described briefly below and summarized in Table 1.1.

The most common form from the three mentioned is the right-handed B-form DNA, B-DNA. It has about 10 base pairs per turn which makes it, with a distance of 3.4 Å between two base pairs (also called rise), about 34 Å per whole turn. Base pairs are almost perpendicular to the axis of the helix. It has 20 Å in diameter. Parameters of the grooves differ; the minor groove of B-DNA is narrow and relatively

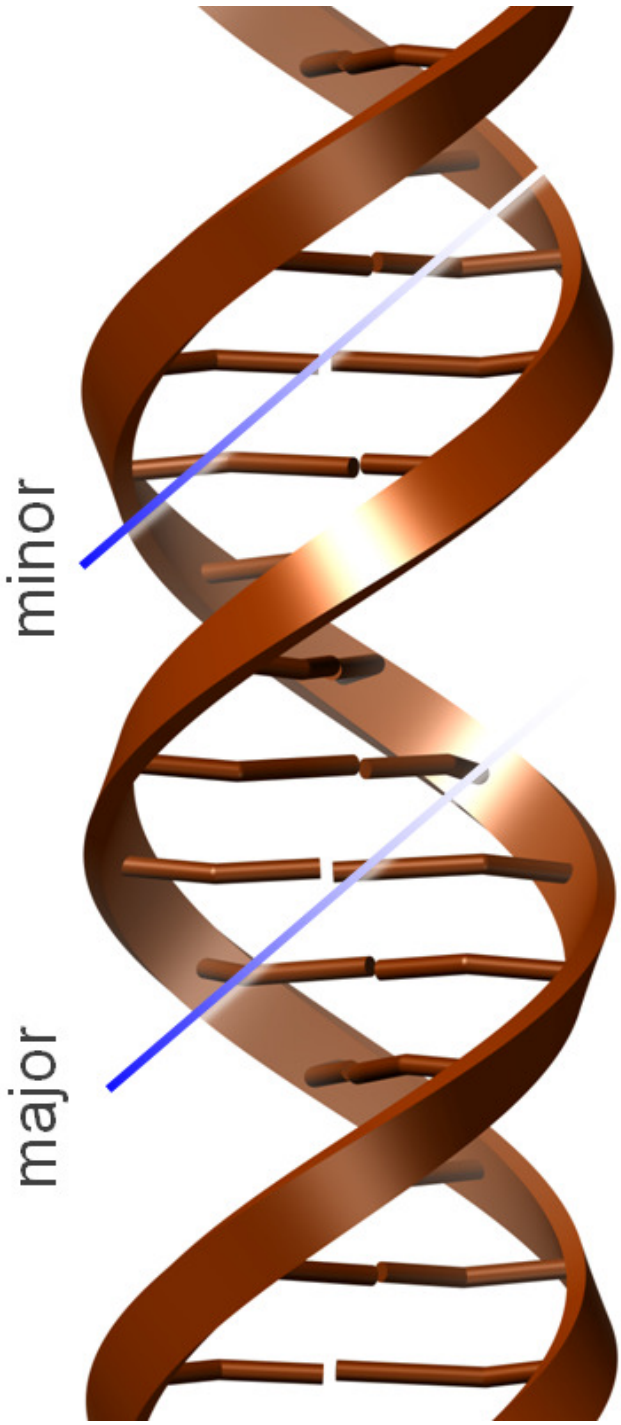


Figure 1.7: **Minor and Major Grooves.** Double helical DNA with its main structural features: the backbone rendered as ribbon, sugars and bases as linked rods, and the minor and major grooves highlighted by blue lines.

Figure 1.8: **Two main DNA forms, A- and B- DNA.** The side and the top view of the A-form of DNA (on the left) and the B-form of DNA (on the right.)

Table 1.1: Table of the basic parameters of DNA forms.

	A-DNA	**B-DNA**	**Z-DNA**
Helix sense	right-handed	right-handed	left-handed
Number of bp per turn	11	10	12
Distance between bp (Å)	2.55	3.4	3.8
Distance per turn (Å)	28	34	45
Diameter of helix (Å)	23	20	18
Major groove	wide, deep	narrow, deep	flat
Minor groove	narrow, shallow	wide, shallow	narrow, deep

deep while the major groove is wide and deep [16].

A-DNA is also the right-handed helix. It has 11 base pairs per turn and the distance between two neighboring base pairs is 2.55 Å, which means that the rise of the A-form DNA is lower than in B-DNA. Base pairs are tilted relative to the helical axis by as much as 20°, and they are shifted to the helix periphery. This creates a hole in the center of helix. A-DNA has 28 Å per whole turn and the diameter 23 Å shows it is wider than B-DNA. The minor groove of A-DNA is broad and shallow and the major groove is narrow and deep [17]. Regarding the sugar puckering, sugar puckers are in the C3'-*endo* conformation in A-DNA and in the C2'-*endo* conformation in B-DNA.

Unlike the A- and B-forms of DNA, the Z-form is the only left-handed double helix [18]. The Z letter refers to the zig-zag pattern of this form of DNA. It has more base pairs per turn than A-DNA, 12. The rise is about 3.8 Å and the distance per turn is 45 Å. A base tilt with respect to the helix axis is about 7° and a diameter is about 18 Å. The minor groove is narrow and deep and the major one is barely apparent [17, 18].

Structure of a DNA duplex depends on its sequence and hydration or relative humidity. While the most common DNA form, B, is stable at naturally high humidity, A-DNA form can be induced in certain sequences by lowering humidity [19]. For the C-G dinucleotide repetitive sequences, low humidity and presence of magnesium or polyvalent cationic amines can induce the Z-form. It is now known that A-DNA form also can be locally induced by interacting with proteins [15].

The data characterizing the mentioned DNA forms are summarized in the Table 1.1.

1.4 Introduction to Protein/DNA Interactions

Protein-DNA recognition is critical for the correct function of key biological processes such as DNA replication during the cell division, DNA transcription into RNA which is further translated into sequence of amino acids in proteins. Another process is gene regulation which includes DNA methylation or binding of regulation proteins, chromosome packaging, so the whole nuclear DNA can fit into the cell nucleus or a DNA repair when DNA is damaged and it would otherwise lead to unwanted processes or even cell's death.

Early hypotheses assumed that the protein/DNA recognition might follow the same straightforward rules as the well-known self-recognition of DNA strands by Watson-Crick base pairing. However, it was shown that such a simple "code of recognition" between DNA and proteins is extremely unlikely [20]. The reason for the lack of this simple rule of recognition is suggested to result from many degrees of freedom of interaction between these structurally complicated molecules [21]. Rapid recognition of the target sequence may be located by so-called "facilitated diffusion" when protein non-specifically binds to DNA and then slides along the DNA until it finds its target sequence. In this case, the original 3D space search for the target sequence reduces to 1D [22–24].

At the local level, there are two conceptual ways how protein can find its target sequence on DNA called direct and indirect readout [25]. The direct readout is also called a base readout and the indirect is called a shape readout [26]. As it is probably clear from their names, the direct, or base, readout, recognizes the sequence of nitrogenous bases in DNA. The direct readout is mainly considered to occur in the major groove [25] The specificity is suggested to be due to specific hydrogen bonds between the protein and the edges of the bases of the DNA [27]. Indirect, or shape, readout recognizes the shape, the conformation of the DNA molecule including its sugar-phosphate backbone, rather than the base sequence. Examples of these DNA-protein complexes, where proteins are interacting with DNA mainly by the shape readout, is the CAP-DNA complex [26]. The shape recognition can be further divided to a local and a global shape recognition. The local shape recognition refers for example to the recognition when a region of the minor groove is narrow. The global shape recognition is for example when the DNA helix exhibits an overall

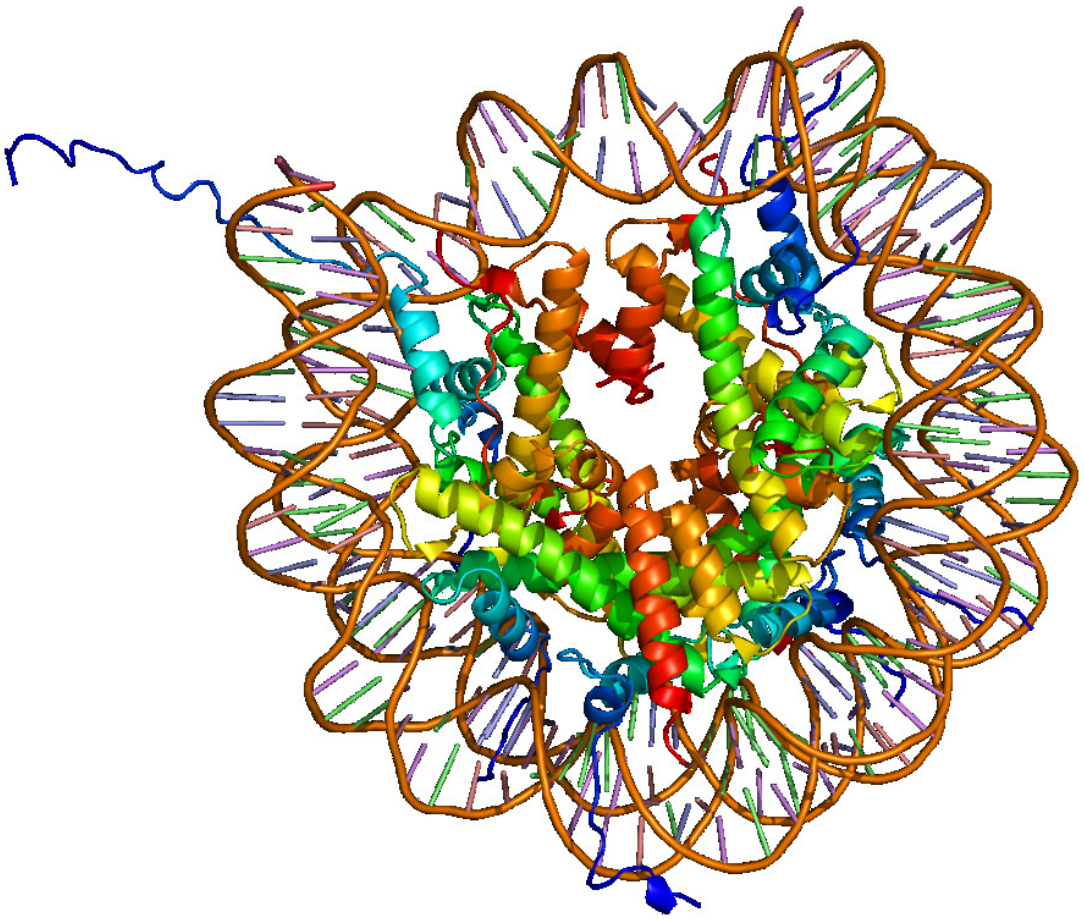


Figure 1.9: **Histone core particle complexed with DNA double helix of 147 base pairs.** The crystal structure of a histone core particle consisting of tetramer of dimers of histone proteins wrapped around by a DNA duplex consisting of almost 150 base pairs. The PDB ID of this structure is 1AOI.

bend [25].

The idea of the direct and the indirect readout is appealing but it is an oversimplification, because DNA-binding proteins combine multiple strategies to recognize their target [25]. The participation of the shape readout indirectly results from the fact that the protein regions binding DNA possess more frequently an intrinsic conformational flexibility [28] than the regions not interacting with DNA. DNA molecules also undergo conformational changes under protein binding [29].

1.4.1 Typical Protein Motifs Involved in Protein/DNA Recognition

There are several structural protein motifs that bind DNA. One of the best described DNA-binding protein motifs are helix-turn-helix motif [30, 31], zinc finger [32–34], leucin zipper [35]. Protein motifs are shown in Figure 1.10.

Helix-turn-helix motif is composed of two alpha-helices joined by a short turn. Function of one helix is to bind to DNA major groove, the other helix has function to stabilize the interaction of first helix with protein. The binding of a protein containing helix-turn-helix motif is influenced by the rest of the protein structure. It is one of the most common DNA-binding motifs [25].

Zinc-finger protein motifs are zinc-coordinated DNA-binding motifs. The name of the zinc-finger motif refers to its appearance. The zinc-finger motif is the most common protein motif found in eukaryotic transcription factors. Naturally occurring zinc-finger proteins recognize a wide variety of different DNA sequences. Many DNA contacts are mediated by more than one zinc finger domain. Zinc finger contacts are usually made to the bases in major groove. Zinc-finger motif bind to DNA in sequence specific manner [21].

In leucine zipper motif, two helices, where one helix is from each monomer, are held together by interactions between hydrophobic amino acids side chains. This amino acids are often leucines. Side chains of these two α-helices interact with DNA through its major groove [35].

It is known that proteins interact with DNA mostly by amino acid residues arginine and lysine, followed by other polar and/or charged amino acids, which is a consequence of the negatively charged DNA surface [36, 37] caused by negatively charged oxygens in phosphate groups in nucleotides. Lipophilic amino acids have low occurrence at the protein/DNA interface [9].

1.4.2 Importance of Solvation for Protein/DNA Recognition

Several researches studied the influence of water on interactions between DNA and proteins [27, 38, 39]. It was discovered that each double helical type has its own

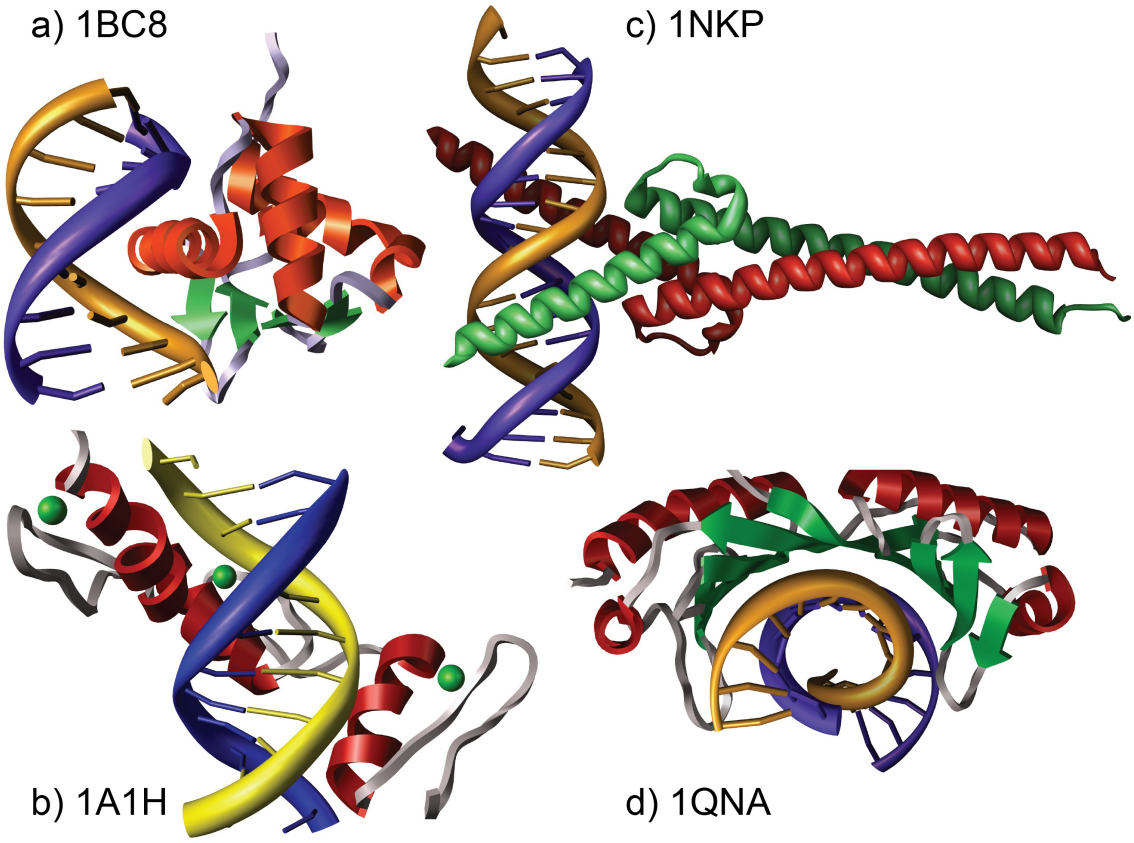


Figure 1.10: **Protein/DNA motifs.** Different types of protein/DNA contacts. The individual motifs are assigned with the PDB codes of the structures. The different structural modifications according to the most commonly occurring B-form of DNA are visible in the figure.

specific hydration pattern and that hydration sites depend on base types, sequence and DNA conformation [40, 41]. It was also proposed that proteins that bind to DNA occupy positions, which were formerly occupied by water molecules in unbound DNA [27]. Later CAP-DNA complexes showed that the protein contacts to the DNA bases agree with the hydration of DNA [39]. Therefore it was showed that water plays an important role in protein/DNA interactions. Water molecules also mediate interactions between proteins and DNA as it is in the case of Trp repressor-operator complex [38]. Water-mediated contacts were observed in many protein/DNA complexes [9].

1.4.3 Structural Properties of DNA during Protein/DNA Interaction

Binding of ligands and proteins to DNA induces conversion from B to A form [15]. This conformational conversion from common B-form to A-form includes transition of sugar puckering from the 2'-*endo* of the B-form to the C3'-*endo* characteristic for the A-form. This transition comprise the sugar puckering O4'-*endo*. Therefore, the O4'-*endo* puckering is the puckering of transitions state between the two DNA forms [9, 11, 14]. The transition between B- and A-form of DNA facilitates DNA bending, and it provides a mechanism to control the width of the minor and major grooves and may therefore facilitate access to base pair edges in the minor groove [15].

Two mentioned structural features of the double helical DNA, its minor and major grooves, participate in interactions between proteins and DNA. The common DNA-binding motifs helix-turn-helix, zinc-finger or leucine zipper, interact with the DNA major groove by means of the base readout, a minority of proteins binds also to the DNA minor groove [42–44] . It was proposed that binding in the DNA minor groove is less specific than binding to the DNA major groove because there are fewer ways to uniquely distinguish among the hydrogen-bond donors and acceptors on the edges of bases [27].

DNA can, after binding of a protein, undergo bending of the sugar-phosphate backbone. It is facilitated by the mentioned conversion from B-to-A conformation which makes e.g. the minor groove more accessible to a protein. The bending can be as much as 90 ° [45]. The bending of DNA occurs e.g. in the complex of DNA

with the TATA-binding protein.

To explore interactions between proteins and DNA, many approaches were developed. Among these approaches, we can find methods studying these interactions in vivo, in vitro or in silico. Experimental methods include different types of methods e.g. the X-ray crystallography [46], protein binding assays [47], footprinting [48, 49], immunoprecipitation [50–52] or microscopy [53, 54]. Along with the experimental methods, theoretical methods were used as well. These include approach from the point of view of energy of binding, molecular simulations and bioinformatics approaches [9, 55, 56]. Due to a large volume of crystallographic data available in databases, bioinformatics and statistical approaches are useful for analysis of protein/DNA interactions according to different criteria such as groove-binding, protein type, meaning of water molecules for protein binding, conformational changes during interactions or sequence dependence of protein binding.

Chapter 2

Methods

2.1 Structure Selection

2.1.1 Protein Data Bank

All analyzed data were queried in the Protein Data Bank (PDB) [57]. PDB was established in 1971 at Brookhaven National Laboratories and is now operated by a consortium of three organizations, so-called World Wide PDB [58]. It contains experimentally determined three-dimensional structures of proteins, nucleic acids and other biological macromolecules. Now, the database contains over 100 000 entries, structures solved by means of X-ray crystallography are about 90% of all entries.

2.1.2 Retrieving Structures from PDB

We retrieved 2406 crystallographic structures of protein/nucleic acid complexes from Protein Data Bank, the release date 2014-04-09. These structures contained proteins, DNA, and did not contain any RNA or hybrids. Therefore, only the structures of DNA alone or with proteins were selected. The crystallographic resolution was chosen to be 3.0 Å.

We also selected crystallographic structures of DNA not complexed with proteins and not containing RNA or DNA/RNA hybrids; there were 882 structures of these "naked" DNA structures retrieved.

Short DNA and Protein Chains

After the selection of the structures from PDB, structures that contained only DNA sequences shorter than 6 nucleotides were excluded because of their low information content. This way, 89 structures of protein/DNA complexes and 149 naked DNA structures were excluded from the analyzed set.

A similar selection was then made based on the protein sequence length. Structures were excluded from the set when the protein sequence was shorter or equal to 20 amino acids; there were 12 such protein/DNA structures.

Modified Nucleotides

In this work, all modified nucleotides were excluded from the analysis. This means that all altered residues found in the sequence of DNA were not used to characterize the dinucleotide conformers by the "structural alphabet". Therefore, all dinucleotides in the DNA sequence that contained an altered residue were excluded and only dinucleotides with unaltered residues were used for further analysis.

MolProbity

MolProbity [59,60] is a web service that provides quality validation for 3D structures of proteins, nucleic acids and their complexes. Detailed all-atom contact analysis is provided by this service. In the process of validation and calculation all hydrogen atoms are added and fully optimized, both polar and non polar. The result of the calculations can be obtained in several forms including e.g. overall numeric score, various lists or graphic files. For the purposes of our analysis, we used a numeric score, so-called MolProbity Score, for comparing the quality of structures to select better refined structure when needed.

2.1.3 Sequence Redundancy

We wanted to avoid as many of the redundancies among the selected crystal structures as possible so that the data would be statistically relevant. Therefore, we decided to compare structures sequentially. The first level of the selection was based on the protein sequences. Using ClustalX [61], a program for multiple sequence align-

ment, the protein sequences were divided into two groups. The first group contained sequentially redundant structures, which means that they contained sequences which were 100% identical in more than 90 % of their sequences. This complicated algorithm was selected to disregard the N- and C-terminal protein regions that may be highly variable without contributing to different ways of interaction with DNA. The second group was assigned as sequentially unique protein sequences if they were identical in less than 90% of their sequences.

In the next step, we investigated the DNA sequences of proteins that were previously marked as sequentially redundant. This step was included because we were interested not in sequentially unique protein sequences but rather in sequentially unique interfaces between proteins and DNA. Therefore, sequentially not unique proteins still can be part of unique interfaces when interacting with sequentially different DNA.

DNA sequences were termed as sequentially unique when they shared less than 90 % of their sequences. This means that DNA sequences were marked to be similar if the sequences differed at least by 2 nucleotides for sequences shorter than 24 nucleotides, the sequences longer than 24 nucleotides sequences had to differ at least by 3 nucleotides, this does not include histones with their sequence long about 150 nucleotides (Figure 1.9).

Among the structures with redundant sequences we selected the structure with the highest crystallographic resolution. In the cases when the crystallographic resolutions of two or more similar structures were comparable (differed by 0.2 Å or less) or equal, we selected the structure with the best value of MolProbity Score. MolProbity Score was calculated with a locally installed MolProbity suite [59, 60].

After the selections, we obtained a set of 413 structures of non-complexed "naked" DNA and 1389 structures of protein/DNA complexes. The process of selection with the numbers of retrieved, excluded and selected structures is summarized in Table 2.1.

Table 2.1: Numbers of retrieved, excluded and accepted structures

	Complexes	**Naked**
Retrieved	2,405	879
Excluded short DNA	89	149
Excluded short protein	12	-
Excluded redundant	1,017	466
Finally accepted	1,389	413

2.2 Assignment of Base Pairing Type

The information about base pairing of nitrogenous bases was classified according to the already mentioned Seanger's schema of base pairing patterns as stored in the so-called mmCIF structure files in the PDB archive. In the case that the information about the base pairing pattern was not included in an mmCIF file, the particular base pair was classified as non-Watson-Crick base pair. The base pairing information was extracted to be used in our analysis of the dependence of DNA backbone conformations on base pairing (see below).

mmCIF Format

Macromolecular Crystallographic Information File (mmCIF) [62] is a type of an archive file, which is human and machine readable, and can be edited by any text editor. It is designed to facilitate electronic transmission of crystallographic data between laboratories, journals and databases.

These files contain data names and information about the structures. The data in CIF files is organized hierarchically and arranged in categories. The list of data names and their descriptions are summarized in the publically accessible mmCIF Dictionary [63].

2.3 Assignment of Dinucleotide Conformers

Our dataset of 1,389 complexes and 413 naked DNAs contains together 58,483 dinucleotides to be classified by dinucleotide conformer classes. We used an improved algorithm of the already existed classification by [13]. Together 48,006 dinucleotides

Table 2.2: Numbers of all steps, steps with conformationally assigned dinucleotides, and steps accepted after exclusion of modified residues listed separately for protein/DNA complexes and naked DNA.

	Complexes	**Naked**
All steps	51,055	7,428
Assigned steps	42,067	5,939
Not assigned steps	8,988	1,489
Assigned without modified residues	41,822	5,749

were classified into one of the classes and 10,477 dinucleotides were not classified into any class because they did not satisfy the rules set to classify dinucleotides (the 25° deviation from the torsion angles, not acceptable values of the δ angle). Numbers of dinucleotide conformers for naked DNA and protein/DNA complexes, respectively, are summarized in the Table 2.2.

We did not analyze dinucleotides in one oligonucleotide strand but two base-paired dinucleotides from two strands of a double helical DNA. Therefore, we analyzed pairs of dinucleotides. It resulted in a reduction of the number of dinucleotide pairs approximately to a half. We decided not to analyze steps containing modified residues. It led to a further reduction of the data set. If a modified residue occurred at the end of the DNA sequence, one pair of dinucleotide conformers was excluded. When it occurred in the middle of the sequence, two pairs of dinucleotide conformers (four dinucleotide conformers) were excluded.

The Table 2.2 shows the numbers of assigned dinucleotides for complexed and non-complexed DNA and final number of dinucleotides after the steps containing modified residues were excluded.

Grouping of Dinucleotide Conformers into Structural Alphabet

To reduce the number of analyzed dinucleotide conformational classes (ntC), we performed their grouping according to their main structural features. The resulting structural alphabet hereafter referred to as ntA provided an easier way to analyze the data afterward because it contains fewer groups.

Conformers 113, 114, 121, 122, 119 and 202 are all B-like conformers that contain

Table 2.3: The dinucleotide conformers. Brief annotation of their main structural features and the average values of the backbone torsion angles characterizing the conformers. a) Conformers identified originally by Svozil et al. [6]

ntC	**Description**	δ	ε	ζ	$\alpha+1$	$\beta+1$	$\gamma+1$	$\delta+1$	χ	$\chi+1$
8	"canonical" A-DNA	83	205	287	294	174	54	83	199	203
13	A-DNA, BI-like χ	89	201	275	294	162	54	89	244	244
19	A-DNA, $\alpha + 1/\gamma + 1$ crank	84	194	290	149	192	192	88	205	188
41	A- to-B, δ¿C3'-, $\delta + 1$ C2'-*endo*	90	196	280	299	179	55	142	222	256
32	BI-to-A, $\delta + 1$ O4'-*endo*	129	186	264	295	170	52	98	247	233
35	BI-to-A, $\beta + 1$ in g+, $\alpha + 1/\gamma + 1$ crank (high), *anti/low anti*	136	199	288	253	73	168	87	264	187
109	BII-to-A, $\delta + 1$¿C3'-*endo*	142	213	181	297	139	52	90	273	207
110	as 109 plus $\alpha + 1/\gamma + 1$ crank, high $\beta + 1$	146	257	186	60	224	196	90	260	200
50	BI variant	129	181	265	300	177	50	123	246	245
54	"canonical" BI	136	183	259	303	181	44	138	252	259
116	BI, $\alpha + 1/\gamma + 1$ crank, α/γ normal	140	194	247	31	197	296	150	253	253
115	BI-DNA, high ϵ, *anti/low anti*	140	275	280	300	189	61	148	265	208
117	BI-DNA, $\beta + 1$ in g+, $\alpha + 1/\gamma + 1$ crank (high), *anti/low anti*	139	196	286	249	73	172	145	263	211
86	BII variant	140	201	216	314	154	46	140	262	153
96	BII variant	143	245	170	297	141	46	141	271	257
97	BII-DNA, $\alpha + 1/\gamma + 1$ crank (g+/t), *anti/lowanti*	142	294	110	149	198	55	151	260	185
113	BI-DNA, ϵ/ζ in t/g+, $\alpha + 1/\gamma + 1$ crank (g+/t), *anti/syn*	143	206	61	82	204	192	146	242	68
114	BI-DNA, $\alpha + 1/\gamma + 1$ crank (g-/g-), high $\beta + 1$, *anti/syn*	141	201	282	307	258	304	151	236	65
119	BI, 5'-mismatch, χ *syn*, α/γ crank	144	189	266	303	167	53	138	70	259
121	3'-mismatch, $\chi + 1$ *syn*, δ O4'-*endo*	99	209	278	295	174	54	128	243	67
122	mismatches, B, $\alpha + 1/\gamma + 1$ crank	137	196	225	33	187	295	145	257	70
123	Z, Y/R step	147	264	76	66	186	179	95	205	61
124	Z, R/Y step, ZI	96	242	294	209	231	55	144	63	205
126	Z, R/Y step, ZII	95	187	62	169	162	44	144	58	213

Table 2.4: Newly characterized dinucleotide conformers [64]

ntC	**Description**	$\boldsymbol{\delta}$	$\boldsymbol{\varepsilon}$	$\boldsymbol{\zeta}$	$\boldsymbol{\alpha+1}$	$\boldsymbol{\beta+1}$	$\boldsymbol{\gamma+1}$	$\boldsymbol{\delta+1}$	$\boldsymbol{\chi}$	$\boldsymbol{\chi+1}$
209	A/B O4'-*endo*, extremely low ε	116	106	305	218	257	79	129	260	266
210	A/B O4'-*endo*, extremely low ε	93	63	60	211	181	62	127	238	251
219	BI/BII-to-A, $\alpha+1$ - 120, $\delta+1$ O4'-*endo*, $\chi+1$ *low anti*	139	200	212	115	226	190	111	255	207
214	BII, $\alpha+1$ low	144	234	199	67	230	206	111	267	209
216	BII,$\delta+1$ O4'-*endo*	147	254	178	294	131	42	95	271	234
217	BI/BII, χ *low anti*	140	231	285	289	172	51	140	189	263
211	BI, $\alpha+1$ 330, $\gamma+1$ 0	146	188	258	338	191	358	150	249	262
220	BI, $\beta+1$ low , $\gamma+1$ trans	143	187	292	217	102	160	145	252	218
221	BI, $\alpha+1/\ \gamma+1$ crank (*trans/trans*)	142	177	276	168	163	171	144	239	231
201	BII, ζ, α +1 - 60	154	242	77	63	177	64	137	237	249
207	BII , ζ, α +1 - 60 + $\gamma+1$ trans	145	224	68	75	189	191	137	263	259
218	BII, $\alpha+1/\gamma+1$ extremely low	138	195	194	22	107	19	127	258	257
202	5'-mismatches, BI, α/γ crank	145	189	274	293	175	50	134	62	260

Table 2.5: DNA structural alphabet ntA as defined by conformational classes ntC.

Description of ntA	**Three letter abb.**	**ntC**
A forms	**AAA**	8, 19
A form with B-like values of both χ	**AcB**	13
A-to-B	**A-B**	41
A-to-B with extremely low ϵ	**AeB**	209, 210
BI-to-A	**B1A**	32
B-to-A	**B-A**	35, 109, 110, 214, 216, 219
less populated BI form	**2B1**	50
the most populated, "canonical" B form	**1B1**	54
	wB1	116
various minor B forms	**miB**	97, 115, 117, 201, 207, 211, 217, 220, 221
BI-to-BII form	**B12**	86
BII form	**1B2**	96
"weird" BII forms	**wB2**	218
B-like conformers with at least one *syn* base	**Bcs**	113, 114, 212, 122, 119, 202
All Z-forms	**ZZZ**	123, 124, 126, 127

at least one *syn* base and therefore are in one group. Various minor B forms, classes 97, 115, 117, 201, 207, 211, 217, 220 and 221, were also put together into the group assigned as miB. Classes 35, 109, 110, 214, 216 and 219 were merged together as they refer to a transition from the B to A conformation. Classes 209 and 210 further refer to transition from B to A conformers but moreover they have extremely low value of the ε torsion angle. Class numbers 8 and 19 are both A-DNA forms. In analogy, all Z-DNA conformers (ntCs number 123, 124, 126 and 127) were put into one group named ZZZ. The abbreviations of nucleotide conformer groups and classes merged to form these groups are listed in the Table 2.5.

Chapter 3

Aims of the Thesis

The aims of this thesis were to:

- select the crystal structures with crystallographic resolution 3.0 Å or better with unique DNA sequences for naked DNA or unique protein/DNA interfaces for their complexes

- identify double helical segments and base pair types, and assign dinucleotide conformers to their dinucleotides in the selected structures.

- determine correlations between dinucleotide conformers across the double helical segments and analyze their dependence on dinucleotide sequences and base pairing type

Chapter 4

Results

4.1 Overview of the Results

The results section is divided into four parts. The first two parts discuss matrices showing correlations between dinucleotide conformers across the double helix for Watson-Crick base pairing in naked DNA structures, and in protein/DNA complexes, respectively. The third section presents correlations between dinucleotide conformers for non-Watson-Crick base pairs. The last, fourth part, examines whether the correlations between dinucleotide conformers across the double helix are sequence dependent.

Table 4.1 shows how many of the selected protein/DNA and naked structures were classified as double stranded DNA, which contain mostly Watson-Crick base pairs. Structures labeled as "other" are triple helices, quadruple helices and other types of structures that do not contain base pairs. Despite the fact that there are many more protein/DNA complexes than structures of naked DNA, the numbers of structures other than double stranded is similar for both sets. There are therefore proportionally more non-double helical naked DNA structures than complexes. The number of double helical structures among the complexed DNA is about three times higher than for those among the naked DNA.

For classification of the base pair types in DNA we used the data stored in the mmCIF files under the category _ndb_struct_na_base_pair.hbond_type_28. In the Saenger's classification schema, Watson-Crick pairs are labeled XX (A-T) and XIX (C-G). Therefore, Table 4.2 shows the numbers of occurrences of Watson-Crick,

Table 4.1: Classification of the selected structures

	Complexed DNA	**Naked DNA**
Double helical	1,269	373
Other	42	39

Table 4.2: Numbers of analyzed base-paired dinucleotide steps in double helical protein/DNA complexes and naked DNA.

	Double helical DNA	
	protein/DNA complexes	**Naked DNA**
All step pairs	21,712	3,003
Step pairs without modif. residues	20,911	2,897
W−C step pairs	19,682	2,823
Non−W − C step pairs	2,030	179
W−C unassigned	4,047	722
Non−W − C unassigned	992	127
Step pairs containing modif. residues	2,030	106

non-Watson-Crick, and base pairs which were not assigned with any base pairing pattern from the list of Saenger's patterns. It is visible from the table that numbers of non-Watson-Crick base pairs are low compared to the Watson-Crick base pairs and also that there is a relatively high number of unassigned base pairs.

4.2 Association Matrices

For further analysis, we introduce so-called Association matrices as a useful tool to study correlation of two characteristics, here structural between two dinucleotide conformers base-paired across the DNA double helix. In these matrices, both dinucleotide conformers were described by the DNA structural alphabet ntA; the assignment of t he alphabet classes was briefly introduced in the Methods section. Elements of the Association matrix show how many times is a conformer A from one strand connected to a conformer B from the other strand of a duplex.

The Association matrices were analyzed separately for Watson-Crick base paired steps in protein-DNA complexes (Figure 4.2), in naked DNA (Figure 4.1), and for non-Watson-Crick pairs in all queried structures (Figure 4.3).

The Association matrices were also assembled for occurrences of dinucleotide conformers of different sequences, e.g. for the steps with sequences AA, AC, AG, AT, etc. These sequence-dependent Association matrices were assembled only for Watson-Crick base paired steps in protein/DNA complexes; there are too few of the other base pair types to be analyzed sequentially. Therefore, there are ten sequence-dependent Association matrices, one for each unique dinucleotide sequence. There are six matrices identical for a dinucleotide and its Watson-Crick paired equivalent dinucleotide: AA/TT, GG/CC, GA/TC, AG/CT, GT/AC, and TG/CA. Four matrices comprise data for palindromic dinucleotides whose sequence is identical in both strands: AT, TA, GC, and CG.

We therefore analyzed 10 matrices depicting occurrences of dinucleotide conformers hydrogen-bonded across strands of the DNA double helix in the protein/DNA complexes. Structural bias and relatively small number of steps in the naked DNA structures led to a decision not to analyze these steps as a function of the sequence.

4.2.1 The Design of Association Matrices

The Association matrices as presented below (Figures 4.1, 4.2, 4.3) contain information about the absolute numbers of occurrences of ntA/ntA combinations in their upper right triangle, and a statistical measure of these numbers in the lower left triangle. Note that these matrices have two diagonals, one for absolute numbers, one for their statistical measures. Statistical analysis was performed by a χ^2 test described below.

Statistical Analysis of the Data by χ^2 Goodness-of-Fit Test

The χ^2 test is commonly used to compare observed data with their expected values. The principle of χ^2 test is described by equation No. 4.1, where x_o refers to the observed occurrence of a particular ntA/ntA combination, and x_e, refers to their expected occurrence,

$$\chi^2 = \frac{(x_o - x_e)^2}{x_e}. \tag{4.1}$$

In the case the observed value was bigger than the expected one, the resulting value of χ^2 was assigned the plus sign, in case that the observed value is smaller

than the expected value, the χ^2 value was multiplied by -1. Positive χ^2 values thus indicate over-representation, while negative ones under-representation of a particular ntA/ntA combination.

Degrees of Freedom of the χ^2 Tests. The number of degrees of freedom was assigned according to a contingency table for the χ^2 test as one.

Probability Values of the χ^2 Tests. For one degree of freedom, the value of $\chi^2 = 3.84$ equals to probability value 0.05, value 6.64 equals probability value 0.01, and 10.83 equals probability value 0.001. The values with minus sign were colored gradually using blue shades, values with plus sign indicating over-representation were colored in the shades of red.

Two Types of χ^2 Goodness-of-Fit Tests

There is no unique way to determine the expected number of occurrences of any particular ntA/ntA combination. In addition, these expected numbers may be different for the number of occurrences of all sequences and for individual dinucleotide sequences. Different tests were applied to estimate significance of the overall ntA/ntA associations and the sequence-dependent ones.

χ^2 Tests for the Overall ntA/ntA Associations. The null hypothesis for the χ^2 test for the overall, sequence-independent, associations between two ntA conformers was that the observed associations are the same as their overall frequencies in the sample. Therefore, the expected number of associations between the alphabet classes ntA_i and ntA_j was estimated from their total frequencies: When the total numbers of alphabet classes ntA_i and ntA_j in the sample are NTA_i and NTA_j, respectively, and the total number of steps is NTA, then the expected number of contacting conformers $A_{ij,\mathrm{exp}}$ can be calculated as

$$A_{ij,\mathrm{exp}} = (\mathrm{NTA}_i * \mathrm{NTA}_j)/\mathrm{NTA}. \tag{4.2}$$

The corresponding χ^2 value is then

$$\chi^2 = \frac{(A_{ij,\mathrm{obs}} - A_{ij,\mathrm{exp}})^2}{A_{ij,\mathrm{exp}}}. \tag{4.3}$$

If the number of occurrences of ntA AAA (NTA_i) is 1,097 in the whole set

(NTA = 39,364) and the number of nTA 1B1 (NTA_j) is 16,415, then the number of expected occurrences of the AAA/1B1 combinations is:

$$\begin{aligned} A_{ij,\mathrm{exp}} &= \frac{\mathrm{NTA}_i}{\mathrm{NTA}} \times \frac{\mathrm{NTA}_j}{\mathrm{NTA}} \times \mathrm{NTA} \\ A_{ij,\mathrm{exp}} &= \frac{1,097}{39,364} \times \frac{16,415}{39,364} \times 39,364 \\ A_{ij,\mathrm{exp}} &= 0.0279 \times 0.417 \times 39,364 \\ A_{ij,\mathrm{exp}} &= 0.0116 \times 39,364 \\ A_{ij,\mathrm{exp}} &= 457 \end{aligned}$$

We can now calculate the χ^2 value of the AAA/1B1 combination as $\mathrm{NTA}_{ij,\mathrm{obs}} = 168$ and the calculated expected value $\mathrm{NTA}_{ij,\mathrm{exp}} = 457$:

$$\begin{aligned} \chi^2 &= \frac{(A_{ij,\mathrm{obs}} - A_{ij,\mathrm{exp}})^2}{A_{ij,\mathrm{exp}}} \\ \chi^2 &= \frac{(168 - 457)^2}{457} \\ \chi^2 &= 183 \end{aligned}$$

Because the number of the observed occurrences of this combination of the ntAs is lower than the number of the expected occurrences, we multiplied the calculated χ^2 value by (-1) and the resulting value is therefore -183.

χ^2 Tests for the Sequence-dependent ntA/ntA Associations. The null hypothesis for the sequence-dependent Association matrices was that the observed number of ntA/ntA associations in a particular sequence does not differ from the number of the same associations observed in the other nine sequences. The corresponding expected number of occurrences was estimated as proportion of the sequence in the whole set times the occurrence of the particular pair of conformers in the whole set,

$$\mathrm{NTA}_{ij,\mathrm{eseq}} = \frac{\mathrm{NTA}_{ij,\mathrm{seq}}}{\mathrm{NTA}_{ij}}. \tag{4.4}$$

The χ^2 value for the sequence dependency is calculated as follows:

$$\chi^2 = \frac{(\mathrm{NTA}_{ij,\mathrm{oseq}} - \mathrm{NTA}_{ij,\mathrm{eseq}})^2}{\mathrm{NTA}_{ij,\mathrm{eseq}}} + \frac{((\mathrm{NTA}_{ij} - \mathrm{NTA}_{ij,\mathrm{oseq}}) - (\mathrm{NTA}_{ij} - \mathrm{NTA}_{ij,\mathrm{eseq}}))^2}{\mathrm{NTA}_{ij} - \mathrm{NTA}_{ij,\mathrm{eseq}}} \tag{4.5}$$

In equations 4.4 and 4.5, $\mathrm{NTA}_{ij,\mathrm{seq}}$ is the proportion of a particular sequence in the whole data set NTA, the $\mathrm{NTA}_{ij,\mathrm{eseq}}$ represents expected values of occurrences of ntA combinations ntA_i and ntA_j in the particular sequence and $\mathrm{NTA}_{ij,\mathrm{oseq}}$ is the value of the observed occurrences of the mentioned combination $\mathrm{ntA}_i/\mathrm{ntA}_j$.

As discussed earlier, there are ten unique dinucleotide sequences. If these sequences were distributed evenly, each would represent 1/10 of all dinucleotides in the data set. However, this is not the case and e.g. the sequence AA forms 13 % from the whole data set. We then assume that any combination of two conformers for the dinucleotide sequence AA will also form 13 % from the same combination of conformers in the remaining nine sequences. The numbers of steps and their fractions observed for the ten dinucleotide sequences are listed in Table 4.3 for complexed and naked DNA.

The number of occurrences of the dinuclotides with the sequence AA for protein/DNA complexes is 2,549 which is approximately $\mathrm{NTA}_{ij,\mathrm{perc}} \sim$ 13% of the whole set set of dinucleotide conformer pairs (19,682). The observed number of occurrences of the ntA/ntA combination AAA/1B1 in the whole set of protein/DNA complexes (NTA_{ij}) is 168 and its occurrence ($A_{ij,\mathrm{obs}}$) in the sequence AA-TT is 10. The expected number of occurrences is calculated as the percentage of the sequence in the whole set times number of occurrences of particular ntA/ntA combination:

$$\begin{aligned} A_{ij,\mathrm{exp}} &= \mathrm{NTA}_{ij} \times \mathrm{NTA}_{ij,\mathrm{seq}} \\ A_{ij,\mathrm{exp}} &= 168 \times 0.13 \\ A_{ij,\mathrm{exp}} &= 22 \end{aligned}$$

We wanted to calculate χ^2 for each combination of ntA/ntA in a particular sequence in comparison (as a relation) to the same combination of ntA/ntA in all the rest sequences. The observed value for combination AAA/1B1 in the other nine sequences was calculated as:

$$\begin{aligned} A_{ij,\mathrm{obs9}} &= \mathrm{NTA}_{ij} - \mathrm{NTA}_{ij,\mathrm{obs}} \\ A_{ij,\mathrm{obs9}} &= 168 - 10 \\ A_{ij,\mathrm{obs9}} &= 158 \end{aligned}$$

Expected value of occurrences for this combination of ntA/ntA is then:

$$\begin{aligned} A_{ij,\text{exp9}} &= \text{NTA}_{ij} - \text{NTA}_{ij,\text{exp}} \\ A_{ij,\text{exp9}} &= 168 - 22 \\ A_{ij,\text{exp9}} &= 146 \end{aligned}$$

The value of χ^2 was calculated as:

$$\begin{aligned} \chi^2 &= \frac{(A_{ij,\text{obs}} - A_{ij,\text{exp}})^2}{A_{ij,\text{exp}}} + \frac{((A_{ij,\text{obs9}}) - (A_{ij,\text{exp9}}))^2}{A_{ij,\text{exp9}}} \\ \chi^2 &= \frac{(10-22)^2}{22} + \frac{(158-146)^2}{146} \\ \chi^2 &= 6.545 \times 0.986 \\ \chi^2 &= 6.5 \end{aligned}$$

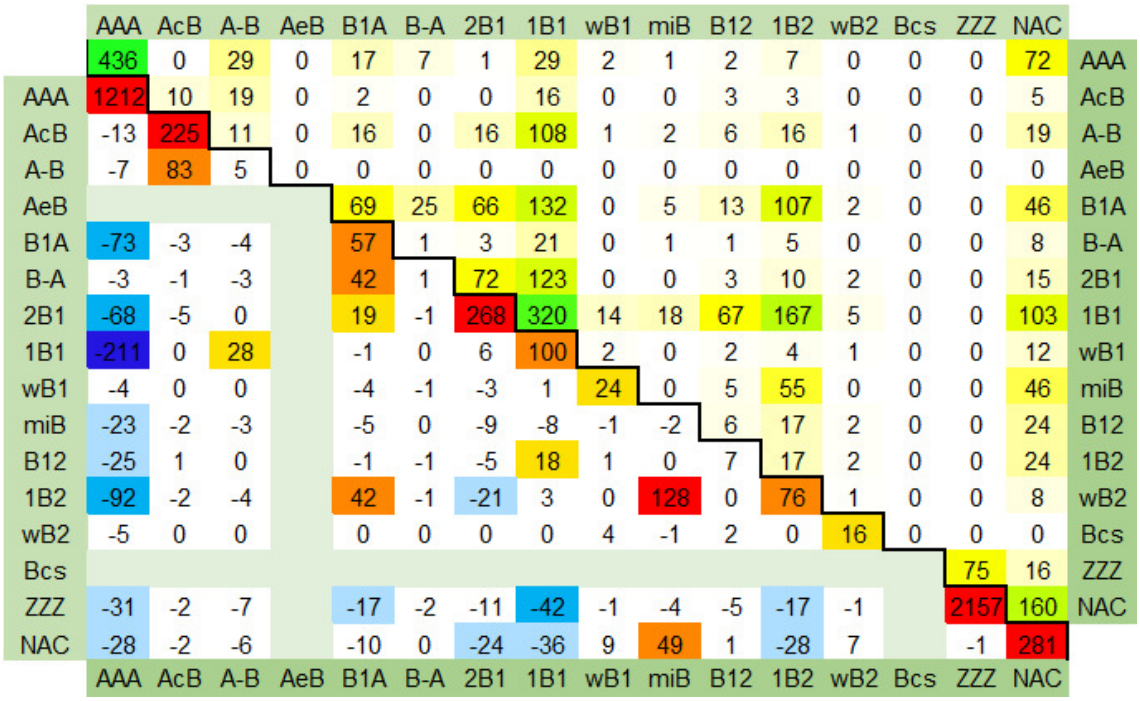
The χ^2 value of AAA/1B1 combination in the particular sequence is approximately 6.5. The number of the observed occurrences of this combination is higher than the expected value, therefore the χ^2 value is positive to show over-representation of the AAA/1B1 combination in the sequence AA.

Coloring of the Association Matrices.

The elements of Association matrices were colored to help their interpretation. The upper right part with numbers of associations is proportionally color-coded from the low to high occurrences from white to yellow to green. The lower left part of the matrices, which contains χ^2 values, is colored by probabilities according to the χ^2 values. Blue color is used for under-represented while red color for overrepresented occurrences.

4.2.2 Association Matrices between ntA Conformers in Double Helical DNA Structures

In this section, we present Association matrices for three data sets, Watson-Crick base paired dinucleotides in naked DNA structures and in protein/DNA complexes, and for non-Watson-Crick base paired dinucleotides in all double helices in our data set.



	AAA	AcB	A-B	AeB	B1A	B-A	2B1	1B1	wB1	miB	B12	1B2	wB2	Bcs	ZZZ	NAC	
	436	0	29	0	17	7	1	29	2	1	2	7	0	0	0	72	AAA
AAA	1212	10	19	0	2	0	0	16	0	0	3	3	0	0	0	5	AcB
AcB	-13	225	11	0	16	0	16	108	1	2	6	16	1	0	0	19	A-B
A-B	-7	83	5	0	0	0	0	0	0	0	0	0	0	0	0	0	AeB
AeB					69	25	66	132	0	5	13	107	2	0	0	46	B1A
B1A	-73	-3	-4		57	1	3	21	0	1	1	5	0	0	0	8	B-A
B-A	-3	-1	-3		42	1	72	123	0	0	3	10	2	0	0	15	2B1
2B1	-68	-5	0		19	-1	268	320	14	18	67	167	5	0	0	103	1B1
1B1	-211	0	28		-1	0	6	100	2	0	2	4	1	0	0	12	wB1
wB1	-4	0	0		-4	-1	-3	1	24	0	5	55	0	0	0	46	miB
miB	-23	-2	-3		-5	0	-9	-8	-1	-2	6	17	2	0	0	24	B12
B12	-25	1	0		-1	-1	-5	18	1	0	7	17	2	0	0	24	1B2
1B2	-92	-2	-4		42	-1	-21	3	0	128	0	76	1	0	0	8	wB2
wB2	-5	0	0		0	0	0	0	4	-1	2	0	16	0	0	0	Bcs
Bcs															75	16	ZZZ
ZZZ	-31	-2	-7		-17	-2	-11	-42	-1	-4	-5	-17	-1		2157	160	NAC
NAC	-28	-2	-6		-10	0	-24	-36	9	49	1	-28	7		-1	281	
	AAA	AcB	A-B	AeB	B1A	B-A	2B1	1B1	wB1	miB	B12	1B2	wB2	Bcs	ZZZ	NAC	

Figure 4.1: **Association matrix for naked DNA.**

Steps Linked by Watson-Crick Base Pairing in Naked DNA

Figure 4.1 shows the Association matrix for dinucleotide conformers linked by Watson-Crick base pairing in double stranded naked DNA. In the upper part with the numbers of occurrences, we see several apparent features.

The most populated DNA, "canonical" B-DNA (ntA 1B1), interacts with almost all dinucleotides, but most often with itself. The 1B1 partners are often other members of the BI conformer family, 2B1 and also conformers A-B and B1A of the transition between A and B form. It also often interacts with not assigned conformers (ntA NAC). This behavior shows that the most populated DNA form is also quite flexible in forming W-C pairs across the strands, further increasing its universal role in the DNA architecture.

High occurrences of the A- and Z-type conformers present in the naked DNA double helices are a consequence of the data set of available crystal structures, which contains many "pure" A- and Z-type structures. The left-handed Z-type conformers obviously pair almost exclusively with themselves. Also the A-type conformers AAA pair mostly with themselves. This contrasts with the ability of 1B1 (and other B-type conformers such as 2B1, B12, and 1B2) which pairs with other conformers as

well and agrees with the known flexibility of B and rigidity of A forms.

Important is the observation that the BII form, ntA 1B2, very rarely associates with itself and prefers B1A or 1B1 conformers. Low occurrences of rare B forms such as wB1, miB, and wB2 contrast with their higher frequencies in protein complexes (see below). The absence of conformers with one base in *syn* conformation (ntA Bcs) confirms the fact that a base in *syn* conformation cannot form a W-C pair in a right-handed double helix.

χ^2 statistics (Figure 4.1) indicate two tendencies: i) most conformers tend to associate with themselves (red and orange highlights in the diagonal), and ii) A-form avoids pairing with any other conformation. As we will see in the protein/DNA structures, both these tendencies are only partially general and are more a consequence of the available crystal structures.

Steps Linked by Watson-Crick Base Pairing in Protein/DNA Complexes

The main difference between Association matrices for the naked DNA (Figure 4.1) and the complexes with proteins (Figure 4.2) is in the volume of data they analyze: while the former reports about less than 3 thousand associated steps, the latter analyzes almost 20 thousand of them. Some prominent features are shared by both matrices. These can be summarized as preference for self-association of ntA conformers (highly significant χ^2 statistics highlighted in red on the diagonal), preference of the BII conformer 1B2 to associate with 1B1, and tendency of A and A-B forms not to associate with B forms.

The "canonical" B DNA 1B1 can form W-C pairs with all ntAs except those with bases in the *syn* orientation, ntA Bcs and ZZZ. 1B1 prefers binding to BI forms, i. e. to itself, to 2B1, wB1, and B12. Significant is a high number of association between 1B1 and ntA 1B2 representing BII form. 1B1 can also form W-C pairs with unclassified conformers NAC. An important observation is that 1B1 is compatible with the A-DNA forms, especially AAA and with mixed A-to-B conformers A-B and rare AeB even when the AAA/1B1 association is statistically under-represented in the protein/DNA as well as in the naked DNA.

In protein complexes, the "canonical" B conformer 1B1 is also able to form W-C pairs with untypical B conformers wB1, miB and wB2. These untypical B-

	AAA	AcB	A-B	AeB	B1A	B-A	2B1	1B1	wB1	miB	B12	1B2	wB2	Bcs	ZZZ	NAC	
	214	9	106	0	100	26	46	168	9	11	18	26	11	0	0	139	AAA
AAA	2583	77	15	0	73	2	8	46	5	1	6	7	0	0	0	19	AcB
AcB	0	3769	110	7	218	49	127	559	15	18	47	44	22	0	0	167	A-B
A-B	83	0	179	1	18	3	12	92	4	3	20	10	22	0	0	55	AeB
AeB	-7	-2	-1	0	178	106	318	944	47	58	137	138	52	0	0	278	B1A
B1A	5	93	88	0	55	10	74	345	43	32	47	24	21	0	0	114	B-A
B-A	0	-4	4	-1	25	0	178	987	35	44	109	89	48	0	0	184	2B1
2B1	-7	-8	7	-1	115	6	139	3901	994	521	1287	1175	302	0	0	1193	1B1
1B1	-183	-67	-19	-1	-49	-3	-1	67	276	132	116	59	15	0	0	178	wB1
wB1	-45	-11	-63	-7	-79	-1	-75	6	744	114	68	60	16	0	0	165	miB
miB	-19	-10	-25	-4	-16	0	-19	-4	41	351	118	167	67	0	0	198	B12
B12	-39	-12	-31	1	-11	-2	-14	52	-5	-4	17	127	58	0	0	191	1B2
1B2	-23	-9	-27	-1	-5	-16	-20	48	-38	-5	3	53	48	0	0	133	wB2
wB2	-7	-8	-5	50	-2	0	-1	-9	-23	-6	2	1	157	0	0	0	Bcs
Bcs															75	13	ZZZ
ZZZ	-5	-1	-7	-1	-12	-4	-10	-68	-9	-6	-10	-10	-4			510	NAC
NAC	6	-8	0	34	-1	5	-18	-145	-10	5	-15	-9	22		-1	438	
	AAA	AcB	A-B	AeB	B1A	B-A	2B1	1B1	wB1	miB	B12	1B2	wB2	Bcs	ZZZ	NAC	

Figure 4.2: **Association matrix for Watson-Crick base paired dinucleotides in protein/DNA complexes.**

forms actually associate more often with 1B1 than with themselves; the associations are however of low statistical weight. In the unassigned conformations NAC, self-association is statistically preferred. Despite that, they associate with all conformers, with rare B-forms (wB2) and especially with mixed A/B forms (AeB) actually quite frequently.

Steps Linked by Non-Watson-Crick Base Pairs

In this paragraph, we analyze the ntA/ntA associations of dinucleotides paired by at least one non-Watson-Crick base pair. The data set contains both base pairs explicitly labeled as non-Watson-Crick or those that were not assigned any class from the Saenger's list.

The most prominent characteristics of the Association matrix of dinucleotide steps base paired by non-Watson-Crick pairs (Figure 4.3) is a high occurrence of dinucleotides with unclassified structure, ntA NAC. Associations NAC/NAC, NAC/1B1, and 1B1/1B1 comprise a third of all associations. The non-W-C steps still contain many BI conformers from ntA 1B1 but their dominance is much less pronounced than in the W-C paired segments of double helices. In comparison to

	AAA	AcB	A-B	AeB	B1A	B-A	2B1	1B1	wB1	miB	B12	1B2	wB2	Bcs	ZZZ	NAC	
	29	3	11	1	11	5	2	11	0	1	3	1	0	1	0	37	AAA
AAA	302	0	2	0	0	3	0	3	0	0	0	1	0	0	0	11	AcB
AcB	7	0	7	0	7	3	6	25	3	4	6	2	4	13	0	35	A-B
A-B	10	2	12	1	6	2	3	23	0	2	5	3	8	0	0	13	AeB
AeB	-1	0	-2	0	13	8	9	58	8	4	9	6	14	6	0	59	B1A
B1A	2	-1	0	2	8	4	6	29	5	3	6	7	5	2	0	40	B-A
B-A	0	8	0	0	0	2	4	45	2	2	5	2	3	2	0	25	2B1
2B1	-1	-1	2	1	1	2	4	276	50	45	112	82	47	12	0	242	1B1
1B1	-24	-2	-6	0	-2	-3	2	29	9	8	12	6	3	2	0	29	wB1
wB1	-5	-1	0	-2	0	0	-1	1	18	19	8	8	1	1	0	41	miB
miB	-4	-1	0	0	-2	-1	-1	0	1	82	20	19	13	10	0	58	B12
B12	-5	-2	-1	0	-3	-1	-1	5	0	-1	9	21	6	2	0	46	1B2
1B2	-6	0	-4	0	-3	0	-3	2	0	0	1	37	11	0	0	36	wB2
wB2	-5	-1	0	12	4	0	0	0	-1	-4	0	-1	22	27	0	13	Bcs
Bcs	-2	-1	25	-2	0	-1	0	-15	-1	-3	0	-3	-4	416	11	0	ZZZ
ZZZ	-1	0	-1	0	-1	-1	-1	-7	-1	-1	-2	-1	-1	-1	2214	222	NAC
NAC	0	5	0	-1	0	1	-1	-27	-2	0	-5	-3	-1	-9	-6	44	
	AAA	AcB	A-B	AeB	B1A	B-A	2B1	1B1	wB1	miB	B12	1B2	wB2	Bcs	ZZZ	NAC	

Figure 4.3: **Association matrix for Non-Watson-Crick base pairs.**

the previously discussed matrices, rare B-forms (wB1, wB2, miB) are relatively more represented in the set.

Besides associations of ntA ZZZ, which associate exclusively with itself, all the other ntA can associate with almost any other ntA. ntA Bcs with one base in the *syn* orientation prefers association with itself, and also with the mixed A/B ntA, especially A-B.

4.2.3 Sequence-dependent Association Matrices

In the next section, we analyze the sequence dependency for occurrences of ntA/ntA dinucleotide associations in double stranded segments of protein/DNA complexes. Due to the volume of available data, we decided to analyze only sequences in the protein/DNA complexes, because there was not enough data for analysis of the naked DNA.

Distribution of the ten dinucleotide sequences is not even as shown in Table 4.3. Six sequences, AA, AG, GG, GA, GT, and TG have the highest proportion of 12-13 %, while the four remaining, AT, GC, CG, and TA only 6-7 %. Despite the fact that this distribution is not uniform (10 % for each sequence), even the lowest number

Table 4.3: Numbers and percentages of the ten dinucleotide sequences as observed in the set of the analyzed protein/DNA structures.

	Complexes		**Naked DNA**	
Sequences	**Number**	**Percent**	**Number**	**Percent**
AA	2,549	13	316	11
AG	2,288	12	130	5
GG	2,378	12	412	15
GA	2,418	12	248	9
AT	1,338	7	199	7
GT	2,387	12	246	9
GC	1,381	7	359	13
TA	1,268	6	131	5
TG	2,420	12	188	7
CG	1,255	6	594	21

of W-C paired steps – 1,255 observed for the sequence CG – is acceptable for the analysis.

The sequence-dependent Association matrices are shown in Figure 4.4. It is apparent that each sequence has a distinct pattern of ntA/ntA associations. We can therefore conclude that distribution of conformers of base-paired dinucleotides depends on their sequence.

The Association matrices confirm some well-known facts: Z-DNA conformation (ntA ZZZ) is strongly sequence-dependent so that it occurs almost exclusively in the Association matrices for sequences CG/CG and GC/GC. Z-DNA conformations interact almost exclusively with itself, a few ZZZ/NAC associations are caused by a failure to classify some Z-like dinucleotides correctly. A-DNA conformations are just a slightly less sequentially dependent and also prefer G and C rich sequences, in this case GG/CC, GC/GC, and CG/CG.

Other observed sequence preferences represent several previously unobserved relationships. Perhaps most importantly, the BII ntA 1B2 shows a distinct sequence dependence. Striking is comparison of sequences with extremely low and extremely high 1B1/1B2 associations, AT and GT versus TG, CG, and TA. In general, some purine-pyrimidine steps disfavor base pairing between BI and BII, while the pyrimidine-purine steps (especially those containing thymine) favor forming of these pairs. This sequence sensitivity of the BII form is further emphasized by high

number of 1B2/1B2 self-associations in TG and TA while all other sequences avoid this particular base-pairing pattern.

Sequence dependence is visible also for the ntA B1A, which represents conversion from the BI-DNA conformation to A form. The transition state is characterized by O4'-*endo* sugar puckering. The occurrences of this dinucleotide conformer in pair with other dinucleotide conformer in two strands of DNA are overpopulated in purine/pyrimidine sequences AT/AT, GC/GC, and GT/AC.

Comparison of the CG/CG and GC/GC steps reveals certain complementary pattern of their Association matrices. A part of it follows from the conformational dependency of G/C rich sequences that both induce Z- and A-forms of DNA. There are however other, less obvious observations. The GC sequence induces BI-to-A transitional conformer B1A, while the CG/CG sequence induces the complementary conformer ntA A-B, transforming A- to B-form. Both these conformations of transition states between A- and B-forms interact in the second strand with other transition states, the most populated B-from 1B1, and BII-form 1B2. The B1A conformer in the sequence GC/GC prefers interactions also with the A-form of DNA, ntA AAA.

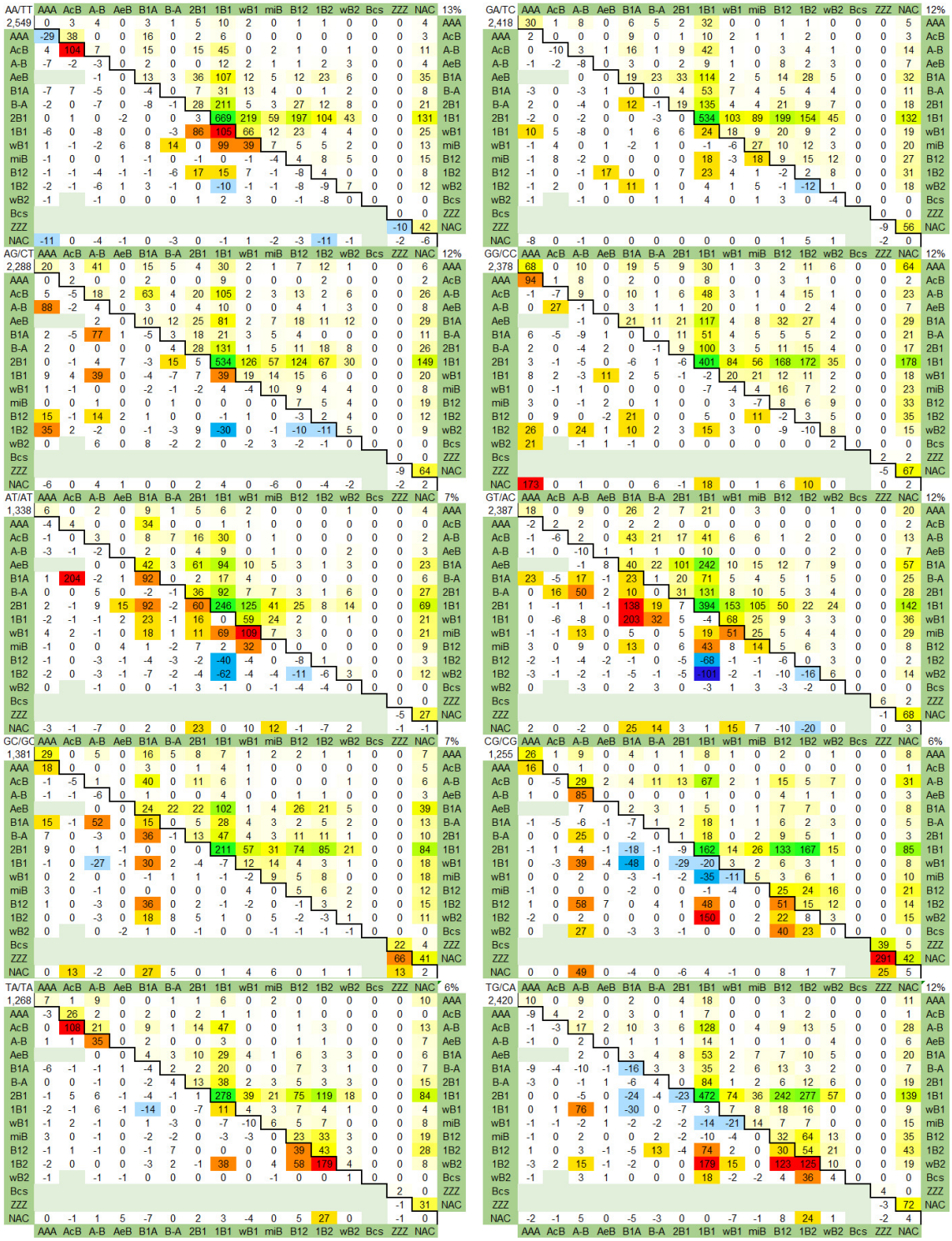


Figure 4.4: **Ten sequence dependent Association matrices of W-C paired steps in protein/DNA complexes.**

Chapter 5

Discussion

This thesis presents an overview of structural preferences of double helical DNA. The analysis takes advantage of a large volume of information available in the Protein Data Bank. The design of the present work uses an approach similar to that taken previously by Schneider et al. (2014) [9], but here we do not focus on the protein/DNA interfaces but on the base paired dinucleotides in double helical DNA structures. The presented analysis is to the best of our knowledge first such analysis.

We selected crystal structures, paying attention to avoid their sequential bias. This was important to allow us to see really existing correlations rather than trivial ones resulting from analysis of repeated sequentially or otherwise similar structures. In the set of released structures, we identified those that form double helices. We analyzed base pairing of all 47,571 dinucleotide steps in the queried structures and sorted them according to their base-pairing pattern to W-C and non-W-C steps.

We assigned dinucleotides with numbers of dinucleotide conformer classes, ntC [6, 13], and grouped them to the structural alphabet (ntA). While there are about 50 ntC and their systematic comparison is complicated and of limited statistical significance even with large numbers of currently available crystal structures, ntA constitute a compact group of conformational families that can be analyzed by statistical tools.

We arranged steps into so-called Association matrices. In the first three parts of our results (with Watson-Crick and non Watson-Crick base paired step conformers), significant differences are visible in patterns of Association matrices between data for Watson-Crick and non-Watson-Crick base pairs. Notable differences can be found

also between naked and protein-complexed DNA within the data for Watson-Crick base pairs.

Association matrices are dependent on the structures available in the Protein Data Bank. It is visible in the Association matrix for W-C base-paired dinucleotide conformers in the naked DNA. In this matrix, even after excluding sequentially redundant structures, A- and Z-conformers occur in naked DNA structures more often than we expected. The "canonical" B-DNA captured by the conformational class 1B1 is the most commonly occurring DNA form. It can associate with almost any other conformer across the double helix.

The "canonical" B-form of DNA may have as a binding partner in double helical DNA almost all other conformers, except for the conformers with *syn* base for Watson-Crick base paired dinucleotides because dinucleotides with a base in the *syn* position do not form Watson-Crick base pairs. NtA 1B1 does not pair with Z-DNA conformations but it is a result of the rigidity of the Z-DNA which prefers as its partner only other Z-conformations or not assigned conformers. The possibility of this B-DNA to bind to almost all conformations support the theory about the flexibility of B-DNA and also probably explains its significant occurrence in DNA structures as a flexible partner for other conformations.

On the other hand, the A-DNA conformers have preferences for their binding partners. They in general prefer other A-DNA conformers or conformers characterized as transition conformers between the A- and B- conformations (B1A and A-B). The matrices therefore confirm certain rigidity of the A-form of DNA. In naked DNA, A-form can be induced by decrease of humidity but this form can also be induced by protein or drug binding [15, 19]. It clearly prefers dinucleotides sequences where numbers of the C and G are higher than those of A and T. This distribution also supports uneven selection of the structures' sequences.

The three matrices show the higher occurrences of associations of BII-DNA conformers with BI-conformers rather than with other BII conformers. This contrasts with BI ability to interact with all conformers, including the other BI.

The most rigid conformer in our data set is ZZZ which contains Z-form DNA conformers. This conformer does not interact with any other assigned conformer than itself.

From the statistical point of view, all three matrices show a similar pattern, over-represented occurrences on their diagonals which suggest preferences of conformers to stand in the double stranded DNA opposite to the same or at least similar conformer.

A surprisingly high number of "canonical" B-DNA conformers in the set of non-W-C base paired dinucleotides may be a consequence of Watson-Crick base pairs incorrectly classified as non-Watson-Crick base pairs due to the lack of the proper information in some mmCIF files. In comparison to the other two Association matrices, dinucleotide conformers with *syn* bases and other minor B-forms of DNA are relatively highly represented.

The analysis of structural features of DNA is influenced by the amount of data available in PDB. Not enough data available for W-C paired dinucleotide conformers in naked DNA and the whole set containing paired dinucleotides via at least one non-W-C base pair allowed us to sequentially analyze only data set of W-C base-paired step conformers in protein/DNA complexes. A sufficient amount of data in the last mentioned set allowed us to assemble Association matrices for all ten sequences possible in the DNA sequences and also to statistically analyze the data in this matrix.

The sequential analysis of the Watson-Crick base paired dinucleotides in protein/DNA complexes revealed different patterns of dinucleotide conformers' combinations in the Association matrices. It suggests that these step conformers' combinations are sequence dependent. They confirm the known fact that C and G rich sequences form A- and Z-DNA conformations [3, 4]. The matrices also show other dependencies of conformer distributions in the Association matrices for different sequences. Several sequences such as AT-AT and TA-TA, and CG-CG and GC-GC show that they are complementary to a certain extent. The complementarity between AA-TT and CG-CG visible in the matrices is rather surprising. Certain patterns and conformer occurrences even suggest the dependency on a general sequence of purine and pyrimidine bases.

Our analyzed data set for the protein/DNA complexes include DNA segments, which are in contact with proteins, but also the segments that are outside the contact. To compare DNA which is really in contact with protein and therefore

influenced by the binding and DNA which is out of contact with protein, we should in the future analyze these two types of DNA duplexes individually. It has been shown in several studies and summarized in [9] that DNA may posses different conformations when free and when bound to a protein. Therefore, an analysis considering the protein binding and also conformation of the bound protein should be performed.

Chapter 6

Conclusions

In this work, 1802 crystallographic structures with crystallographic resolution better than 3.0 Å and with unique DNA sequences (for the naked DNA) or unique interfaces (in the protein/DNA complexes) were selected. Overall, we analyzed 413 naked DNA and 1389 protein/DNA complexes. The dinucleotide steps in the selected structures were conformationally analyzed and their backbone was assigned with number of dinucleotide conformer classes (ntC) and then with three-letter structural alphabet ntA.

So-called Association matrices were assembled to compare step pairs base-paired in double stranded DNA according to a type of two hydrogen bonds via which the steps were connected, Watson-Crick and non-Watson-Crick. The data set for Watson-Crick base pairs was further divided into base pairs in naked DNA or protein/DNA complexes. Therefore, three Association matrices were assembled. Other ten Association matrices were assembled to compare occurrences of hydrogen-bonded bound dinucleotide conformers influenced by the sequence of these step conformers.

In this work we observed that the local DNA conformation depends on a base pairing between the two strands of DNA and also on the DNA sequence.

The Association matrices show that the most common, "canonical", B-DNA interacts mostly with all other DNA conformations while the A-form of DNA interacts mostly with other A-forms. This confirms rigidity of the A-form and contrasts it with highly flexible B-form. The matrices also show interesting behavior of BII-DNA which prefers as a partner in the second strand of DNA BI conformers rather

than other BII conformers while there is a pattern among the conformers that they prefer to be bound in pair with the same conformers as themselves.

List of Abbreviations

DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
W-C	Watson-Crick
k-NN	Method of k-nearest neighbors
bp	Base pair
PDB	Protein Data Bank
mmCIF	Macromolecular Crystallographic Information File
CIF	Crystallographic Information File
ntA	Structural alphabet
ntC	Dinucleotide conformer class
NAT	Total number of alphabet classes
NAC	Not assigned conformers

List of Figures

List of Tables

Bibliography

[1] S. Neidle, B. Schneider, and H. M. Berman. *Fundamentals of DNA and RNA structure.*, volume 44. John Wiley & Sons, 2003.

[2] J. D. Watson and F. H. C. Crick. Molecular structure of nucleic acids, 1953.

[3] W. Saenger. *Principles of Nucleic Acid Structure.* Springer advanced texts in chemistry. Springer-Verlag, 1984.

[4] S. Neidle. *Principles of Nucleic Acid Structure.* Elsevier, 2008.

[5] B. Schneider, Z. Morávek, and Helen M. Berman. RNA conformational classes. *Nucl. Acids Res.*, 32:1666–1677, 2004.

[6] D. Svozil, J. Kalina, M. Omelka, and B. Schneider. DNA conformations and their sequence preferences. *Nucl. Acids Res.*, 36:3690–3706, 2008.

[7] G. Parkinson. Structure of the CAP-DNA Complex at 2.5 ÅResolution: A Complete Picture of the Protein-DNA Interface. *J. Mol. Biol.*, 260:395–408, 1996.

[8] K. Luger, W. Mäder, R. K. Richmond, D. F. Sargent, and T. J. Richmond. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature*, 389:251–260, 1997.

[9] B. Schneider, J. Černý, D. Svozil, Petr Čech, J. C. Gelly, and A. G. De Brevern. Bioinformatic analysis of the protein/DNA interface. *Nucl. Acids Res.*, 42:3381–3394, 2014.

[10] J. S. Richardson, B. Schneider, L. W. Murray, G. J. Kapral, R. M. Immormino, J. J. Headd, D. C. Richardson, D. Ham, E. Hershkovits, L. D. Williams, K. S.

Keating, A. M. Pyle, D. Micallef, J. Westbrook, and H. M. Berman. RNA backbone: consensus all-angle conformers and modular string nomenclature (an RNA Ontology Consortium contribution). *RNA*, 14:465–481, 2008.

[11] A. Gelbin, B. Schneider, L. Clowney, S. H. Hsieh, W. K. Olson, and H. M. Berman. Geometric parameters in nucleic acids: Sugar and phosphate constituents. *J. Am. Chem. Soc.*, 118:519–529, 1996.

[12] V. Sychrovsky, S. Foldynova-Trantirkova, N. Spackova, K. Robeyns, L. van Meervelt, W. Blankenfeldt, Z. Vokacova, J. Sponer, and L. Trantirek. Revisiting the planarity of nucleic acid bases: Pyramidilization at glycosidic nitrogen in purine bases is modulated by orientation of glycosidic torsion. *Nucl. Acids Res.*, 37:7321–7331, 2009.

[13] P. Čech, J. Kukal, J. Černý, B. Schneider, and D. Svozil. Automatic workflow for the classification of local DNA conformations. *BMC Bioinf.*, 14:205, 2013.

[14] R. Taylor and O. Kennard. Molecular structures of nucleosides and nucleotides. 2. Orthogonal coordinates for standard nucleic acid base residues. *J. Am. Chem. Soc.*, 104:3209–3212, 1982.

[15] X. J. Lu, Z. Shakked, and W. K. Olson. A-form conformational motifs in ligand-bound DNA structures. *J. Mol. Biol.*, 300:819–840, 2000.

[16] H. R. Drew, R. M. Wing, T. Takano, C. Broka, S. Tanaka, K. Itakura, and R. E. Dickerson. Structure of a B-DNA dodecamer: conformation and dynamics. *Proc. Natl. Acad. Sci. USA*, 78:2179–2183, 1981.

[17] D. W. Ussery. *DNA Structure: A-, B- and Z-DNA Helix Families*, pages 1–6. John Wiley & Sons, Ltd, 2001.

[18] A. H. et al. Wang. Molecular structure of a left-handed double helical DNA fragment at atomic resolution. *Nature*, 282:680–686, 1979.

[19] R. E. Franklin and R. G. Gosling. The structure of sodium thymonucleate fibres. I. The influence of water content. *Acta Crystallogr.*, 6:673–677, 1953.

[20] B. W. Matthews. Protein-DNA interaction. No code for recognition., 1988.

[21] C.O. Pabo and L. Nekludova. Geometric analysis and comparison of protein-DNA interfaces: why is there no simple code for recognition? *J. Mol. Biol.*, 301:597–624, 2000.

[22] O. G. Berg, R. B. Winter, and P. H. von Hippel. Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory. *Biochemistry*, 20:6929–6948, 1981.

[23] S. E. Halford and J. F. Marko. How do site-specific DNA-binding proteins find their targets? *Nucl. Acids Res.*, 32:3040–3052, 2004.

[24] P. H. Von Hippel and O. G. Berg. Facilitated target location in biological systems. *J. Mol. Biol.*, 264:675–678, 1989.

[25] R. Rohs, X. Jin, S. M. West, R. Joshi, B. Honig, and R. S. Mann. Origins of specificity in protein-DNA recognition. *Annu. Rev. Biochem.*, 79:233–269, 2010.

[26] S. Chen, A. Gunasekera, X. Zhang, T. A. Kunkel, R. H. Ebright, and H. M. Berman. Indirect readout of DNA sequence at the primary-kink site in the CAP-DNA complex: Alteration of DNA binding specificity through alteration of DNA kinking. *J. Mol. Biol.*, 314:75–82, 2001.

[27] N. C. Seeman, J. M. Rosenberg, and A. Rich. Sequence-specific recognition of double helical nucleic acids by proteins. *Proc. Natl. Acad. Sci. USA*, 73:804–808, 1976.

[28] T. Sunami and H. Kono. Local Conformational Changes in the DNA Interfaces of Proteins. *PLoS ONE*, 8, 2013.

[29] K. Mizuguchi and S. Ahmad. Conformational changes in DNA-binding proteins: Relationships with precomplex features and contributions to specificity and stability. *Proteins: Struct., Funct., Bioinf.*, 82:841–857, 2014.

[30] M. Suzuki and M. Gerstein. Binding geometry of α-helices that recognize DNA. *Proteins*, 23:525–535, 1995.

[31] W. A. McLaughlin and H. M. Berman. Statistical models for discerning protein structures containing the DNA-binding helix-turn-helix motif. *J. Mol. Biol.*, 330:43–55, 2003.

[32] M. Suzuki, M. Gerstein, and N. Yagi. Stereochemical basis of DNA recognition by Zn fingers. *Nucl. Acids Res.*, 22:3397–3405, 1994.

[33] Y. Mandel-Gutfreund and H. Margalit. Quantitative parameters for amino acid base interaction : implications for prediction of protein DNA binding sites. *Genetics*, 26:2306–2312, 1998.

[34] Y. Choo and A. Klug. Physical basis of a protein-DNA recognition code. *Curr. Opin. Struct. Biol.*, 7:117–125, 1997.

[35] D. Garcia de Viedma, R. Giraldo, G. Rivas, E. Fernández-Tresguerres, and R. Diaz-Orejas. A leucine zipper motif determines different functions in a DNA replication protein. *EMBO J.*, 15:925–934, 1996.

[36] N. M. Luscombe, R. A. Laskowski, and J. M. Thornton. Amino acid base interactions: a three-dimensional analysis of protein DNA interactions at an atomic level. *Nucl. Acids Res.*, 29:2860–2874, 2001.

[37] N. M. Luscombe and J. M. Thornton. Protein DNA Interactions: Amino Acid Conservation and the Effects of Mutations on Binding Specificity. *J. Mol. Biol.*, 320:991–1009, 2002.

[38] Z. Otwinowski, R. W. Schevitz, R. G. Zhang, C. L. Lawson, A. Joachimiak, R. Q. Marmorstein, B. F. Luisi, and P. B. Sigler. Crystal structure of trp repressor/operator complex at atomic resolution. *Nature*, 335:321–329, 1988.

[39] J. Woda, B. Schneider, K. Patel, K. Mistry, and H. M. Berman. An analysis of the relationship between hydration and protein-DNA interactions. *Biophys. J.*, 75:2170–2177, 1998.

[40] B. Schneider and H. M. Berman. Hydration of the DNA bases is local. *Biophys. J.*, 69:2661–2669, 1995.

[41] B. Schneider, D. M. Cohen, L. Schleifer, A. R. Srinivasan, W. K. Olson, and H. M. Berman. A systematic method for studying the spatial distribution of water molecules around nucleic acid bases. *Biophys. J.*, 65:2291–2303, 1993.

[42] C. A. Bewley, A. M. Gronenborn, and G. M. Clore. Minor groove-binding architectural proteins : structure , function , and DNA. *Annu. Rev. Biophys. Biomol. Struct.*, 1:105–131, 1998.

[43] M. H. Werner, J. R. Huth, M. Gronenborn, and G. M. Clore. Molecular basis of human 46X,Y sex reversal revealed from the three-dimensional solution structure of the human SRY-DNA complex. *Cell*, 81:705–714, 1995.

[44] D. B. Nikolov, H. Chen, E. D. Halay, A. Hoffman, R. G. Roeder, and S. K. Burley. Crystal structure of a human TATA box-binding protein/TATA element complex. *Proc. Natl. Acad. Sci.*, 93:4862–4867, 1996.

[45] S. C. Schultz, G. C. Shields, and T. A. Steitz. Crystal structure of a CAP-DNA complex: the DNA is bent by 90 degrees. *Science*, 253:1001–1007, 1991.

[46] Z. Du, J. K. Lee, S. Fenn, R. Tjhen, R. M. Stroud, and T. L. James. X-ray crystallographic and NMR studies of protein protein and protein nucleic acid interactions involving the KH domains from human poly(C)-binding protein-2. *RNA*, 13:1043 – 1051, 2007.

[47] A. G. Gilman. A protein binding assay for adenosine 3':5'-cyclic monophosphate. *Proc. Natl. Acad. Sci. USA*, 67:305–312, 1970.

[48] D. J. Galas and A. Schmitz. DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucl. Acids Res.*, 5:3157–3170, 1978.

[49] D. S. Dimitrova, M. Giacca, F. Demarchi, G. Biamonti, S. Riva, and A. Falaschi. In vivo protein-DNA interactions at a human DNA replication origin. *Proc. Natl. Acad. Sci. USA*, 93:1498–1503, 1996.

[50] R. Blecher-Gonen, Z. Barnett-Itzhaki, D. Jaitin, D. Amann-Zalcenstein, D. Lara-Astiaso, and I. Amit. High-throughput chromatin immunoprecipitation

for genome-wide mapping of in vivo protein-DNA interactions and epigenomic states. *Nat. Protoc.*, 8:539–554, 2013.

[51] H. Im, J. A. Grass, K. D. Johnson, M. E. oyer, J. Wu, and E. H. Bresnick. Measurement of Protein-DNA Interactions In Vivo by Chromatin Immunoprecipitation. *Methods Mol. Biol.*, 284:129–146, 2004.

[52] A. Pascual-Ahuir and M. Proft. Quantification of protein-DNA interactions by in vivo chromatin immunoprecipitation in yeast. *Methods Mol. Biol.*, 809:149–156, 2012.

[53] B. ten Heggeler-Bordier, R. Hipskind, A. Seiler-Tuyns, E. Martinez, B. Corthésy, and W. Wahli. Electron microscopic visualization of protein-DNA interactions at the estrogen responsive element and in the first intron of the Xenopus laevis vitellogenin gene. *EMBO J.*, 6:1715–1720, 1987.

[54] M. Schnos and R. B. Inman. New insights into protein-DNA interactions obtained by electron microscopy. *Mol. Biotechnol.*, 16:77–86, 2000.

[55] S. Furini, P. Barbini, and C. Domene. DNA-recognition process described by MD simulations of the lactose repressor protein on a specific and a non-specific DNA sequence. *Nucl. Acids Res.*, 41:3963–3972, 2013.

[56] C. G. Kalodimos, N. Biris, A. M. J. J. Bonvin, M. M. Levandoski, M. Guennuegues, R. Boelens, and R. Kaptein. Structure and Flexibility Adaptation in Nonspecific and Specific Protein-DNA Complexes. *Science*, 305:386–389, 2004.

[57] H. M. Berman, T. Battistuz, T. N. Bhat, W. F. Bluhm, P. E. Bourne, K. Burkhardt, Z. Feng, G. L. Gilliland, L. Iype, S. Jain, P. Fagan, J. Marvin, D. Padilla, V. Ravichandran, B. Schneider, N. Thanki, H. Weissig, J. D. Westbrook, and C. Zardecki. The Protein Data Bank. *Acta Crystallogr. D*, 58:899–907, 2002.

[58] H. Berman, K. Henrick, H. Nakamura, and J. L. Markley. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucl. Acids Res.*, 35:301–303, 2007.

[59] I. W. Davis, L. W. Murray, J. S. Richardson, and D. C. Richardson. MolProbity: Structure validation and all-atom contact analysis for nucleic acids and their complexes. *Nucl. Acids Res.*, 32:615–619, 2004.

[60] V. B. Chen, W. B. Arendall, J. J. Headd, D. A. Keedy, R. M. Immormino, G. J. Kapral, L. W. Murray, J. S. Richardson, and D. C. Richardson. MolProbity: All-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D*, 66:12–21, 2010.

[61] M. A. Larkin, G. Blackshields, N. P. Brown, R. Chenna, P. A. Mcgettigan, H. McWilliam, F. Valentin, I. M. Wallace, a. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson, and D. . G. Higgins. Clustal W and Clustal X version 2.0. *Bioinformatics*, 23:2947–2948, 2007.

[62] J. D. Westbrook and P. M. Fitzgerald. The pdb format, mmcif formats and other data formats. In H. Bourne, P/ E. & Weissig, editor, *Structural Bioinformatics.* John Wiley & Sons, Hoboken, NJ, 2003.

[63] J. D. Westbrook, H. M. Berman, and S. R. Hall. Specification of a relational dictionary definition language (ddl2). In B. Hall, S. R. & McMahon, editor, *International Tables for Crystallography. Definition and Exchange of Crystallographic Data.* Dordrecht, Heidelberg, 2006.

[64] Černý et al. To be published. 2015.

Svoluji k zapůjčení této práce pro studijní účely a prosím, aby byla řádně vedena evidence vypůjčovatelů.

Jméno a příjmení s adresou	Číslo OP	Datum vypujčení	Poznámka