Charles University

Faculty of Science

Study program: Bioinformatics

Branch of study: Bioinformatics

Barbora Schramlová

Nucleic Acid Geometry Standards

Geometrické standardy nukleových kyselin

Bachelor's thesis

Supervisor: Ing. Jiří Černý, Ph. D.

Prague, 2021

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 dne 6.5.2021 Podpis autora

Chtěla bych poděkovat svému školiteli Ing. Jiřímu Černému, Ph.D., za možnost vypracování bakalářské práce, odborné vedení, trpělivost a ochotu, kterou mi při jejím zpracování věnoval. Dále také děkuji Mgr. Paulíně Božíkové za poskytnutí výsledků a její pomoc a vstřícnost v průběhu projektu.

Abstract

Nucleic acids (NA) are biochemical macromolecular substances essential for all living organisms - their function is to store genetic information and control the process of protein biosynthesis. Nucleic acids are composed of polynucleotide chains. According to the composition of these, we distinguish two basic types of nucleic acids: DNA (Deoxyribonucleic Acid) and RNA (Ribonucleic Acid). Based on the internal nucleotide sequence and external interactions, these chains are formed into different spatial conformations. At the dinucleotide level, these conformations are described and classified by so-called classes of dinucleotide conformers - NtC.

The X-ray crystallography method has been used for over 50 years to reveal the three-dimensional structures of biological macromolecules. The principle of this is the interaction of X-rays with the electron cloud of atoms in the crystal. Atomic positions are then determined based on the calculated electron density. However, with the resolution available for most macromolecular crystals, these crystallographic data are not sufficient to derive a chemically acceptable structure, so stereochemical restraints apply.

Stereochemical restraints are dictionaries describing specific bond lengths, bond angles, torsion angles, planes, and chirality. Using these values, structural macromolecular models are modified and improved. Based on data obtained from structural databases, we studied the latest version of these stereochemical restraints.

Key Words: NA structure, NA conformation, stereochemical restraints, refinement, structure database, CSD, PDB

Abstrakt

Nukleové kyseliny (NA) jsou biochemické makromolekulární látky, nezbytné pro všechny živé organismy – jejich funkcí je uchovávání genetické informace a řízení procesů syntézy bílkovin. Nukleové kyseliny jsou složené z polynukleotidových řetězců a podle složení těchto nukleotidů rozlišujeme dva základní druhy nukleových kyselin: DNA (Deoxyribonukleová kyselina) a RNA (Ribonukleová kyselina). Na základě vnitřní nukleotidové sekvence, a i vlivem vnějších interakcí se tyto řetězce formují do různých prostorových konformací. Na úrovni dinukleotidů jsou tyto konformace popisovány a klasifikovány tzv. třídami dinukleotidových konformerů – NtC.

Pro odhalení trojdimenzionálních struktur biologických makromolekul se již přes 50 let používá především metoda rentgenové krystalografie. Principem této metody je interakce rentgenových paprsků s elektronovým oblakem atomů v krystalu. Na základě elektronové hustoty se poté určují atomové pozice. Při rozlišení dostupném u většiny makromolekulárních krystalů však nejsou tyto krystalografická data dostatečná k tomu, aby bylo možné odvodit chemicky přijatelnou strukturu, proto se používají stereochemická omezení.

Stereochemické standardy jsou slovníky popisující konkrétní délky vazeb, úhly vazeb, torzní úhly, roviny a chiralitu. Pomocí těchto hodnot se zdokonalují a upravují strukturní modely makromolekul. Na základě dat získaných ze strukturních databází jsme studovali nejnovější verzi těchto stereochemických standardů.

Klíčová slova: struktura NA, konformace NA, stereochemické standardy, strukturní databáze, CSD, PDB

Abbreviations

NA	Nucleic Acid
RNA	Ribonucleic Acid
DNA	Deoxyribonucleic Acid
A	Adenosine-5'-phosphate nucleotide
G	Guanosine-5'-phosphate nucleotide
C	Cytidine-5'-phosphate nucleotide
T	Thymidine-5'-phosphate nucleotide
U	Uridine-5'-phosphate nucleotide
W-C pairs/pairing	Watson-Crick (canonical) pairing of bases
NtC	diNucleotide Conformers
CANA	Conformational Alphabet of Nucleic Acids
CSD	The Cambridge Structural Database
CIF	Crystallographic Information Framework/File
mmCIF	Macromolecular Crystallographic Information File
PDB	The Protein Data Bank
PDB format	Protein Data Bank file format
NDB	Nucleic Acid Database

Table of contents

1 Introduction

Nucleic acids play an essential role in all living organisms - they store genetic information. Over the past century, scientists have sought to understand how genetic information is stored and passed on to future generations. The first of a series of successful experiments was Griffith's experiment in 1928 (Griffith 1928). In this experiment, he tried to show that bacteria can transfer their genetic information to future generations through transformation. At that time, however, it was not yet clear what was being transmitted. It was discovered later in 1944 by Avery, MacLeod and McCarty (Avery, Macleod, and McCarty 1944). These three dealt with the chemical nature of the transformation principle, and they found that the cell genetic information is transmitted by DNA. Further experiments in 1952 by Hershey and Chase on bacteriophages later confirmed this (Hershey and Chase 1952). In 1953, the groundbreaking discovery of Watson and Crick followed. They were the first in the world to identify the structure of the deoxyribonucleic acid double-helical form (Watson and Crick 1953). These discoveries led to very rapid research of various types of DNA and RNA and became the basis for the emergence of a new field of science - molecular biology.

Over time, scientists have discovered new conformations of nucleic acids. For DNA, there are three basic ones - A-, B- and Z-forms. RNA in the double-stranded form occurs most commonly in the A form. These three basic conformational forms describe the architecture of nucleic acids at the global level. However, DNA and RNA molecules are very flexible and often change their structure depending on other molecules. The three basic forms are insufficient to describe this flexibility. For this reason, a more detailed conformational alphabet was created, which describes the local conformations of nucleic acids at the dinucleotide level and assigns them into the so-called NtC classes (Jiří Černý et al. 2020).

To construct three-dimensional models of nucleic acids, it is necessary to know the conformations of the monomer components - i.e., nucleosides and nucleotides. These conformations are studied using many different methods, but the most common are X-ray structural analysis and NMR spectroscopy. In this study, we will work with structures that were determined with the help of X-ray crystallography. The principle of X-ray crystallography is the scattering of X-rays by a crystal and their subsequent detection. The intensities of beams on the detector are digitalized and transferred into a computer which, according to a mathematical formula (Fourier transform), determines the position of each

atom in the crystal molecule. Atomic distances are determined in Ångströms (Å). One of the problems with this method in nucleic acid analysis is that it is often challenging to crystallize nucleic acids in sufficiently high quality. Typically, it is necessary to adjust the acquired model with the help of so-called refinement.

Refinement is the last stage in the process of solving nucleic acid structures. During this stage, the discrepancies between the measured diffraction intensities and the calculated stereochemistry data are adjusted. Some stereochemical properties were revealed even before the first structures. As early as the 1950s, Pauling *et al.* studied the planar nature of the peptide bond (Pauling, Corey, and Branson 1951). Stereochemical knowledge also contributed to the groundbreaking discovery of the B-DNA model (Watson and Crick 1953). Later, stereochemical restraints began to limit usually bond lengths, bond angles, planes and chirality. These restraints are described in dictionaries, together with information on where the values come from, how precise they are, what they relate to, atomic types and bonds. Standard dictionaries for nucleic acids have been compiled in 1982 by Taylor and Kennard (Taylor and Kennard 1982) and later in 1982 by Parkinson *et al.* (Parkinson et al. 1996). Since then, further significant improvements have been made in 2020 (Kowiel, Brzezinski, and Jaskolski 2016), (Gilski et al. 2019), (Kowiel et al. 2020). This update was initiated, among other things, by an expansion of the Cambridge Structural Database (CSD), whose accurate crystal structures of the small molecules are the primary source for stereochemical values.

2 Nucleic acids

Nucleic acids are large biomolecules found in all known living organisms and viruses. They are the carriers of genetic information, and therefore they are indispensable for all living organisms. A nucleic acid is made up of several repeating nucleotide subunits connected into linear chains. According to these nucleotides' chemical composition, we distinguish two types of nucleic acids: ribonucleic acid (RNA) and deoxyribonucleic acid (DNA).

2.1 Nucleotides

Nucleotides are the basic units that make up nucleic acids. They are composed of a five-carbon monosaccharide ring, nitrogenous base and a phosphate group. In nucleic acids, these subunits are linked together by a phosphodiester bond between the phosphate and sugar rings (-C3-O3-P- covalent bond, the junction can be seen in Figure 2-5) into linear polymers.

Sugars in nucleotides are derivatives of furan - it is a five-membered ring. They occur in two forms: ribose in RNA or 2'-deoxyribose in DNA (Figure 2-1).

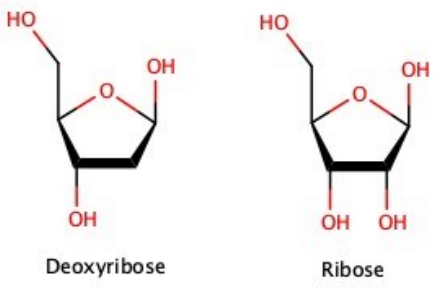

Figure 2-1. Two sugar components of a nucleic acids

Nitrogen bases are planar aromatic heterocycles that can be divided into purines and pyrimidines. Pyrimidines are formed by a six-membered ring where nitrogen atoms are located at positions 1 and 3. Pyrimidines include thymine, cytosine and uracil. Purines consist of a fused pyrimidine and imidazole ring and include adenine and guanine (Neidle, Schneider, and Berman 2005). All these five nitrogen bases are shown in Figure 2-2. Adenine, guanine and cytosine occur in DNA and RNA, while thymine is specific only for DNA and uracil is specific for RNA.

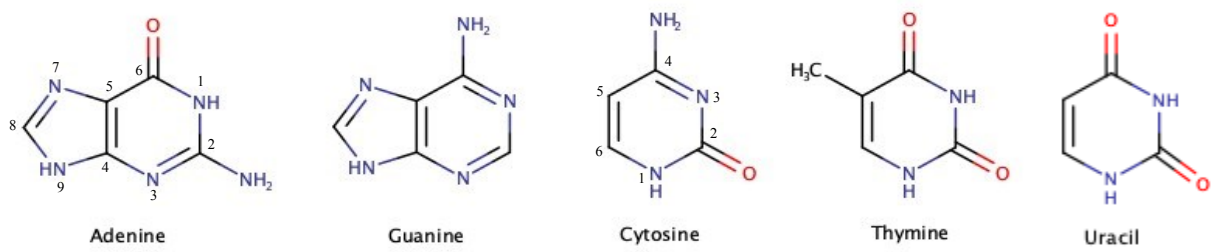

Figure 2-2. The five bases of nucleic acids with examples of numbered atoms.

The nitrogen base and sugar are linked by a glycosidic bond and form a nucleoside. When a nucleoside is phosphorylated on the free hydroxyl group of sugar, a nucleotide is formed (Figure 2-3). Multiple nucleotides joined into linear strands form a single-stranded DNA or RNA.

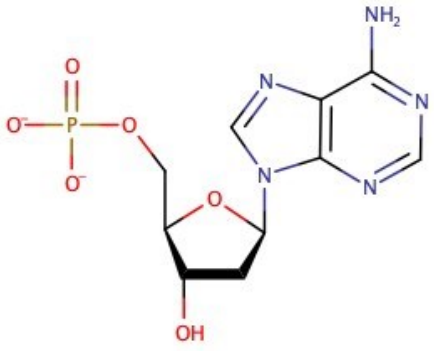

Figure 2-3. Nucleotide.

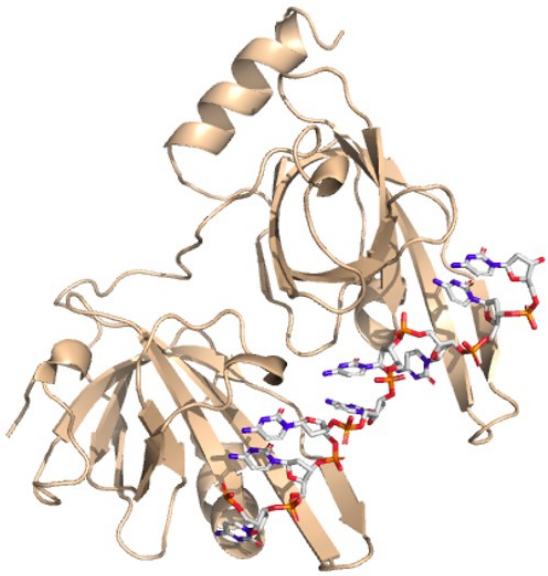

Figure 2-4. An example of nucleotide strand in the structure of the single-stranded-DNA-binding domain of replication protein A bound to DNA (Bochkarev et al. 1997).

2.2 Sugar phosphate backbone

The sugar-phosphate backbone consists of sugars connected by phosphodiester bonds. It is the critical component of nucleic acids – it determines their flexibility. The backbone has six torsion degrees of freedom in every nucleotide, additionally the sugar rings are accessing various non-planar conformations. Due to these two properties, nucleic acids are flexible molecules.

In the sequence of atoms of the phosphate skeleton C5' - C4'- C3' - O3'- P- O5' - C5' - C4'- C3' - O3' (from C5' of one nucleotide to O3'of the other), the torsion angles are denoted in the order α, β, γ, δ, ε, ζ (shown in Figure 2-5). The ranges of these torsion angles are limited because the spatial distribution of atoms in the phosphodiester backbone is restricted by steric hindrance. The atomic sequence mentioned above with the described torsions is part of the smallest descriptive fragment called step (Schneider et al. 2018). The step contains two sugar rings, two bases and phosphate - the whole step is in Figure 2-5.

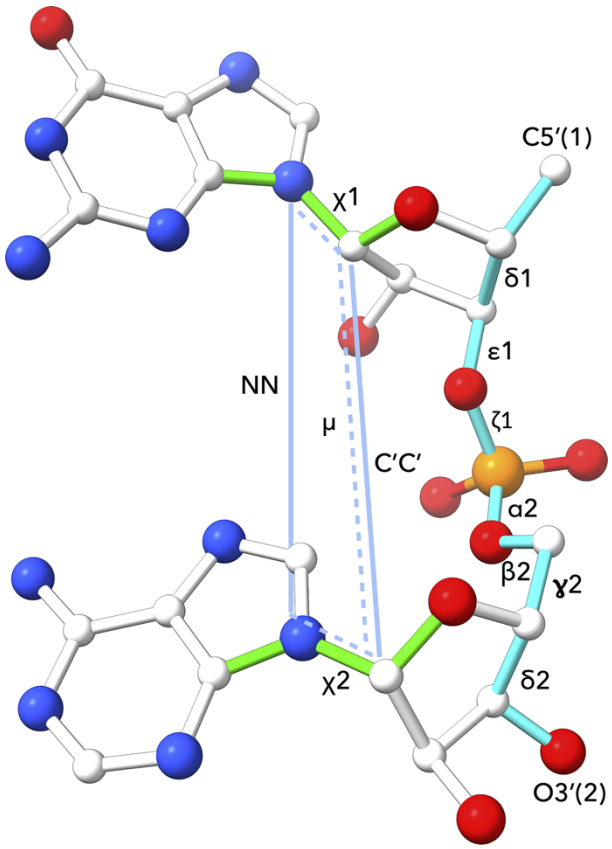

Figure 2-5. The dinucleotide fragment described by seven backbone torsions, two torsions around the glycosidic bonds, one pseudo-torsion angle and two distances (Jiří Černý et al. 2020).

The sugar ring is non-planar and can occupy more spatial conformation. These conformations are called puckers. There are one rotational bond and four pseudorotation bonds between the individual five-carbon sugar atoms (Altona and Sundaralingam 1972). The torsion angles of pseudorotation bonds are interdependent and interact with each other. Because of that, the ribose conformation can be described by two parameters - pseudorotation P and magnitude τ (Altona and Sundaralingam 1972). The most common conformation is an arrangement where four atoms lie in a plane and the fifth C2' or C3' deviates. Whether it deviates to the same side as C5' or to the opposite, it distinguishes the *endo* and *exo* conformations. The most common are the C2'-*endo* and C3'-*endo* conformations.

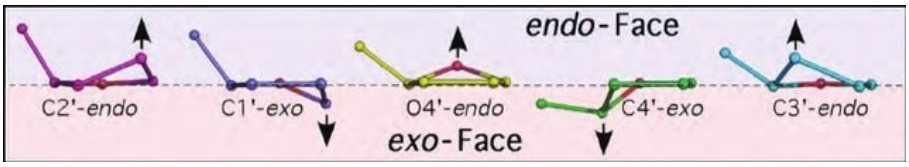

Figure 2-6. Examples of sugar puckers. Arrows indicate the atom that is puckered, and the direction of puckering (Shing and Carter, 2011).

With respect to the sugar, the base usually occupies one of two orientations, *syn* or *anti*. These orientations are defined by the seventh torsion angle χ describing rotation around the glycosidic bond between the C1′ atom on the sugar ring and the base nitrogen. In the *anti* orientation, the N1 and C2 atoms of the purines and the C2 and N3 atoms of the pyrimidines point away from the saccharide. In contrast, in the *syn* orientation, all of these atoms are oriented towards the sugar ring.

2.3 Base pairing

Nitrogen bases (described above) can be paired together by hydrogen bonds. Depending on which two bases are paired together, we divide several types of pairing.

The most common type of pairing is the so-called Watson-Crick (or canonical) pairing, which was first experimentally verified in 1952 by Chargaff (Zamenhof, Brawerman, and Chargaff 1952). As is shown in Figure 2-8, in this type of pairing, guanine and cytosine are connected by three hydrogen bonds. The other pair is adenine and thymine in DNA or uracil in RNA, connected by two hydrogen bonds. These pairs have

unique properties that make them preferred to other types of pairing. The first is that these pairs provide stabilizing energies thanks to their hydrogen bonds. About half of the stabilizing forces are the van der Waals forces, which arise due to stacking interactions. For the second, they guarantee a very similar geometry of G-C and A-T or A-U pairs.

Due to the pairing properties of nitrogen bases, nucleic acids tend to pair, which is why DNA usually occurs in double-stranded form.

The two most widely used nomenclatures to describe base pairing have been developed by Saenger (Saenger 1984) and later by Leontis and Westhof (Leontis and Westhof 2001). The nomenclature proposed by Leontis and Westhof has 16 classes, while the one from Saenger (shown in Figure 2-7) has 28 classes.

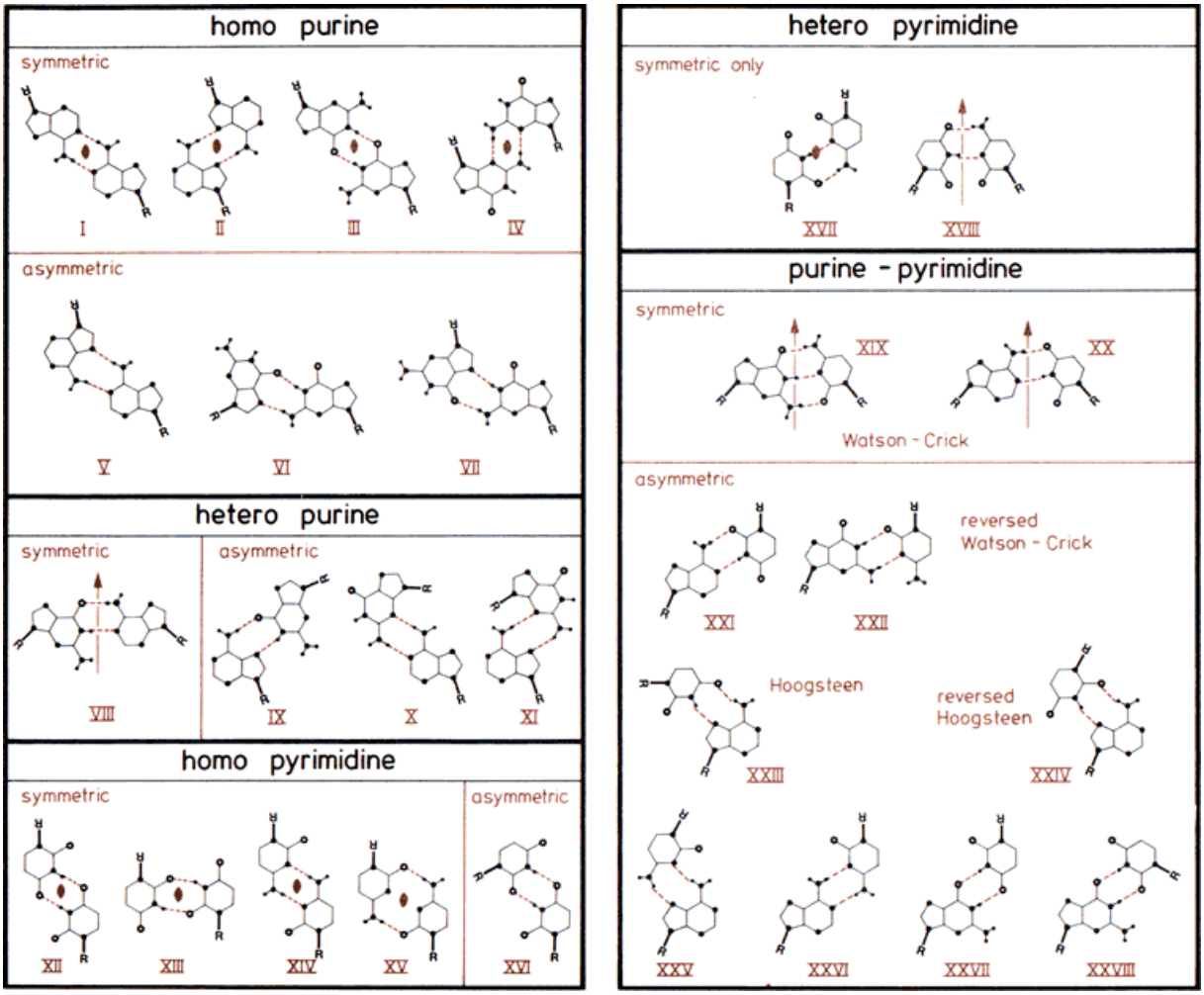

Figure 2-7. The 28 possible base-pairs for A, G, U/T, and C (Saenger, 1984).

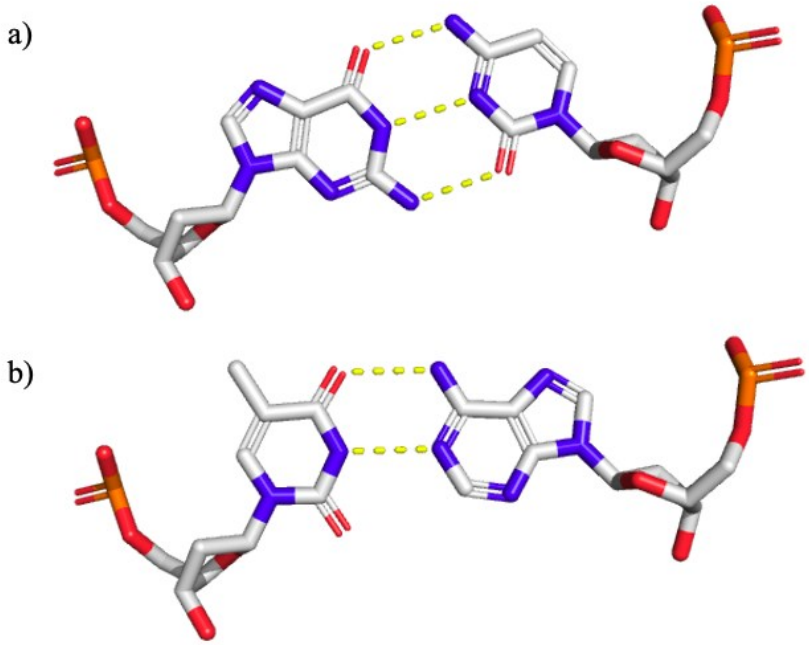

Figure 2-8. Watson-Crick base pairs a) cytosine and guanine b) thymine and adenine.

3 Conformations of nucleic acids

The conformation models describe how the spatial arrangement of strands joined by base-pairing looks. Nucleic acids may occupy several conformations given by the nucleotide sequence, binding partners and solvation conditions. Watson and Crick created the first model of the secondary structure based on what is now known as the canonical pairing of DNA in 1953 (Watson and Crick 1953). This model shows a DNA duplex formed by two antiparallel polynucleotide strands whose sugar-phosphate backbone envelops the molecule, and the nucleotide bases form pairs inside the helix. This double helix structure is called a B form. In addition to the B form, we distinguish two other basic forms - A and Z (all three forms are shown in Figure 3-1). These forms depend not only on the primary nucleotide sequence but also on the binding partners and various solvation conditions.

In DNA the most important of these main forms are the B form mentioned above. It is a right-handed helix in which the (deoxy)riboses are in the C2'-*endo* conformation. Glycosidic bonds have the *anti* orientation, and the bases form W-C pairs and point perpendicular to the helical axis.

The second form is the A form, which, like the B form, is right-handed, but unlike the B form, the (deoxy)ribose is found in the C3'-*endo* pucker.

The last form is the Z form, which, unlike the previous ones, is left-handed. The difference is also that, in this case, the conformation of the (deoxy)ribose in the nucleotides changes depending on how the base is bound. In the case of purine, the sugar is in the form of C3′-*endo* and in the case of pyrimidine in C2′-*endo*. Another feature of this form is that the glycosidic bond changes in nature depending on its location. It takes the *anti* orientation in cytidine and the *syn* orientation in guanidine.

These three forms, shown in Figure 3-1, are found mainly in DNA, usually in double-stranded form. Thanks to the hydroxyl group on the C2' atom, supporting the C3'-*endo* conformation of the sugar, RNA most often takes forms conformationally indistinguishable from A-DNA (Schneider, Morávek, and Berman 2004).

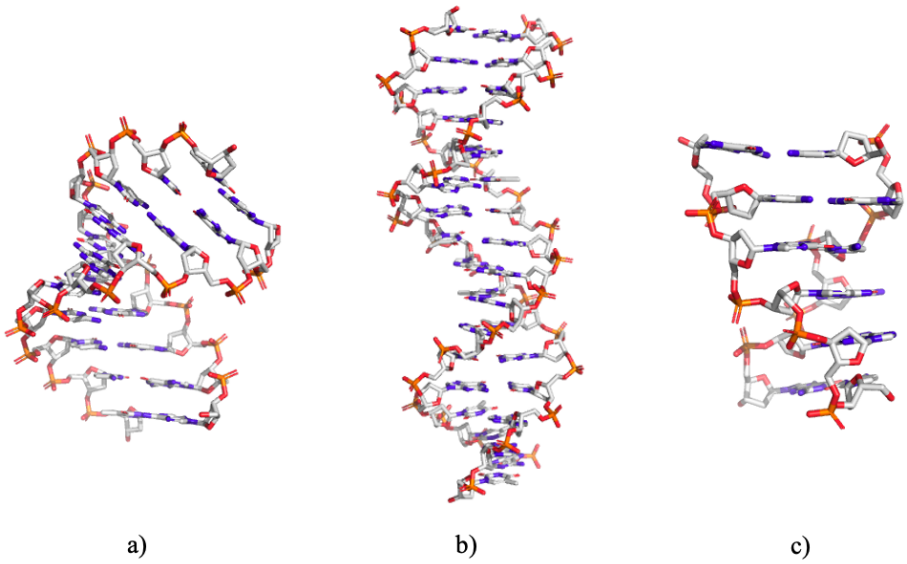

Figure 3-1. The tree basic conformation forms. a) A DNA (Gao, Robinson and Wang, 1999) b) B DNA (Narayana and Weiss, 2009) c) Z DNA (Brzezinski et al., 2011).

3.1 Conformational classes and alphabet

The spatial architecture of DNA is described using three conformational models A, B, and Z mentioned above. However, the DNA strand has a great adaptive capacity and can perform various conformational changes. These cannot be described by classification into one of the three created models. Therefore, it is necessary to be able to describe the conformation of DNA at the local level.

The smallest unit of a nucleic acids molecule that can be conformationally examined is a dinucleotide (Richardson et al. 2008) - two nucleotides linked by a phosphodiester bond - a region of sizeable conformational variability. In 2018, a new system was defined to describe these small DNA sections' conformation more accurately (Schneider et al. 2018). In this system, the conformations are divided into classes called NtC (diNucleotide Conformer classes). Each NtC class contains steps with the same conformation. A step is a DNA unit from two nucleotides of one strand, thus containing two (deoxy)riboses, two bases and one phosphate group (one step is shown in Figure 2-5). During the formation of NtC classes, seven torsion angles of their sugar-phosphate backbone and two torsion angles around the glycosidic bond were analyzed in these fragments (Schneider et al. 2018). All these parameters are also shown in Figure 2-5. Based on this, 44 NtC classes were formed. These classes strictly describe the geometric properties of individual steps. However, about 21% of the observed steps could not be assigned to any of these groups and are labeled NANT. These dinucleotides could not be assigned for two reasons. The main reason is that significantly lower resolution structures are often refined incorrectly. The other reason is the existence of conformationally unique dinucleotides, thanks to the DNA backbone's high flexibility. All these NtC groups can be viewed on the website dnatco.datmos.org.

To describe the agreement between the analyzed dinucleotide structure and the given geometric properties of NtC classes, the validation score called confal was introduced. (Schneider et al. 2018)

In addition to the NtC classes, a higher level of classification of dinucleotide geometries was created to understand DNA structure better- the DNA structural alphabet for nucleic acids CANA - Conformational Alphabet of Nucleic Acids (Schneider et al. 2018). This classification considers strict geometric structures described by NtC classes and distinguishes its categories based on more general properties such as stacking.

The assignment of DNA and RNA structures to these classes is available on the website DNATCO (Jiří; Černý, Božíková, and Schneider 2016) https://dnatco.datmos.org. The website analyzes the structures available in PDB or mmCIF format and further assigns them to NtC classes, CANA codes and provides confal values for all steps in the structure.

4 Structural databases and their file formats

Structural databases are publicly available computer-readable collections of biological structural data stored in relational databases and subsequently distributed in text files (Jones 1997). The most critical databases include the Protein Data Bank and the Cambridge Structural Database.

4.1 The Protein Data Bank

The Protein Data Bank (PDB) is a worldwide best-known database associating biological macromolecular crystal 3D structures. This database was created in 1977 at Brookhaven National Laboratories (BNL), where all data have also been stored and were available for distribution on magnetic tapes (Bernstein et al. 1977). At the beginning, the PDB stored seven structures, but since the 1980s, the number of deposited structures and users have begun to rise sharply (Berman et al. 2000). With the expanding size of the database, it was also moved from magnetic tapes to the web archive. At the end of April 2021, this database contained over 177,000 records (more precise numbers, including the distribution according to the resolution, are in Table 4-1 and Table 4-2).

Table 4-1. Numbers of structures in the PDB database https://www.rcsb.org (Berman et al. 2000) *by polymer entity types as of 25.4.2021.*

Polymer entity type	Count of structures
Protein	173325
DNA	8152
RNA	5356
NA-hybrid	212
Other	5

Table 4-2. Numbers of structures in PDB database https://www.rcsb.org (Berman et al. 2000) divided by nucleic acid and resolution as of 25.4.2020.

Refinement resolution (Å)	Count of DNA structures	Count of RNA structures
x <= 1.2	123	67
1.2 < x <= 1.8	1123	384
1.8 < x <= 2.5	2794	914
2.5 < x <= 5	3001	2968
5 < x	174	147

4.2 The Cambridge Structural Database

The Cambridge Structural Database (CSD) is a database storing over one million complete records of all published organic and metal-organic small molecule crystal structures, whose structures have been determined using crystallography.

The database was established in the Department of Organic Chemistry, the University of Cambridge, in 1965 by The Cambridge Crystallographic Data Centre (CCDC) - X-ray crystallography group lead by Olga Kennard (Groom and Allen 2014). Since the 1990s, all data in this database are stored in CIF format (the Crystallographic Information Framework/File) (Groom et al. 2016). The development of this format took place in cooperation with IUCr and its Commissions and was a turning point in data availability (Groom and Allen 2014).

Because the CSD is updated within a few moments of a new publication, the database contains all published structures, which are checked computationally and expert structural chemistry editors to guarantee the reliability and reusability of stored data (Groom et al. 2016). The CSD has been operational for over fifty years, and, as Groom et al. has shown, the number of newly added structures increases every year. Nowadays, the database is updated every year with approximately 50,000 new and newly modified structures (Bruno and Groom 2014).

4.3 Protein Data Bank file format

All data in Protein Data Bank were stored in a PDB (Protein Data Bank) file format, which provided an annotated description of proteins and nucleic acids, whose structures were derived from X-ray diffraction, NMR studies or Cryo-EM. This format was established in the early 1970s. Nowadays, this format is gradually being replaced by the more diverse dictionary of the mmCIF format.

Unlike the CIF format described below, the PDB format has a fixed structure. It is a flat text file that contains the Cartesian coordinates of each atom in a structure. The whole file consists of partitions, which are initially defined by a keyword with a maximum length of six characters. This keyword then determines the format of the paragraph with the data in the lines below it (Burley et al. 2017). Due to this feature, the format is severely limited. However, this was considered an advantage because the format was easy to read. With the development of structural biology, these properties have become limiting (Burley et al. 2017), so a newer mmCIF format replaced this format.

4.4 The Crystallographic Information File

The Crystallographic Information File (CIF) is a standard file format for exchanging crystallographic and related structural data of small organic molecules. This format was first described in a 1991 publication by Hall, Allen and Brown (S. R. Hall, Allen, and Brown 1991). It is used to capture the results of a diffraction experiment and enables the streamlined publication of results. CIF is promoted and owned by the International Union of Crystallography, which also takes care of defining data items. These data dictionaries are stored and publicly available under the authority of an IUCr Committee (COMCIFS).

The principle of this format is the Self-Defining Text Archival and Retrieval (STAR) File format which means that it uses only ASCII characters, which describe both the data structure and the information itself (Sydney R. Hall 1991). The format is very flexible and self-defining – it consists of data names, data items and loop facilities for repeated items. Each datum has two components - its name, which is defined with the appropriate attributes (for example, the data type) and the data itself, which follow (Brown and McMahon 2006). This structuring means that the format can be easily read by the human eye and by computer and can be edited in a simple text editor.

In 1990 begun the endeavor of IUCr to extend this format to include data items relevant to macromolecular crystallography experiment (Bourne et al. 1997). Thus, an extension of the CIF format - mmCIF - was created.

4.5 The Macromolecular Crystallographic Information File

The Macromolecular Crystallographic Information File - abbreviated mmCIF, is a format derived from the original CIF format and is adapted to contain macromolecular data similar to PDB format but providing more variability and internal consistency due to ontology. This format is based on the same principle as the classic CIF - it contains the *name-value* pairs defined by STARS (Bourne et al. 1997). The variable name is always marked with an underscore at the beginning of the line to distinguish it from the following values (Bourne et al. 1997). The mmCIF has recently begun to replace the older macromolecular PDB format. In 2014 it became the standard format for the PDB archive, and later in 2019, it was announced as a mandatory acceptable file format.

5 Stereochemical restraints

Stereochemical restraints are an essential part of devising chemically acceptable structures when there is not enough data or the data are of insufficient quality. The restraints help to estimate bond lengths, bond angles, planes, and chirality for the structure. This process is called refinement and requires the most accurate knowledge of the geometry of the monomer components of the polymer chains (bond distances, angles, torsion angles, planarity) (Parkinson et al. 1996).

The X-ray data of most macromolecular crystals are not accurate enough to form an exact structure, so these geometric restraints are used to improve the structure with lower resolution, where the quality of experimental data is insufficient (Evans 2006). Besides, these stereochemical restraints are also used in high-resolution structures, for example, to correct disordered fragments that are not defined by diffraction or to create probability functions that require accurate geometric targets and error estimates. However, these restraints are used not only in crystallography but also in NMR models, cryo-EM models, computational modeling of macromolecular structures, and the validation of structural models in macromolecular databases, such as in Protein Data Bank (Kowiel et al. 2020).

Dictionaries containing these standards have been updated several times. Taylor and Kennard first compiled them in 1981. Their work concluded that the protonation states of basic nucleic acid residues could be reliably derived from their molecular dimensions (Taylor and Kennard 1982).

Another significant update was made by Parkinson et al. They selected structures containing bases, sugars or the phosphodiester linkage with atomic resolution up to 1.0 Å not only from Cambridge Structural Database (as the authors mentioned above) but also from the Nucleic Acid Database. Their analysis was focused on the sugar-phosphate backbone and the nucleobase moiety. Based on obtained data, they created a dictionary of nucleic acids from refined X-ray structures containing average values of binding distances, angles and torsions (Parkinson et al. 1996).

At the same time, other dictionaries of these stereochemical restrictions on nucleic acids have emerged. One was focused on the nitrogenous bases (Clowney et al. 1996) and the second on the sugar and phosphate constituents (Gelbin et al. 1996). As a source for defining the nucleic acid standards, the high-resolution small-molecule crystal structures contained in the Cambridge Structure Database (Clowney et al. 1996) and the case of the second research by Gelbin *et al.,* also an atomic resolution oligonucleotides in the Nucleic Acid Database were used (Gelbin et al. 1996). Examples of determined values are in Table 5-1 and Table 5-2 below, where Mean stands for arithmetic mean, esd for estimated standard deviation and N for number of cases considered.

Table 5-1. Parameters of bond lengths and angles estimated for uracil (Clowney et al. 1996).

Bond or Angle	Mean	(esd, N)
N1-C2	1.381	(0.009, 46)
C2-N3	1.373	(0.007, 46)
N3-C4	1.380	(0.009, 46)
C4-C5	1.431	(0.009, 46)
C5-C6	1.337	(0.009, 46)
C6-N1	1.375	(0.009, 46)
C2-O2	1.219	(0.009, 46)
C4-O4	1.232	(0.008, 46)
N1-C1'	1.469	(0.014, 46)
C6-N1-C2	121.0	(0.6, 46)
N1-C2-N3	114.9	(0.6, 46)
C2-N3-C4	127.0	(0.6, 46)
N3-C4-C5	114.6	(0.6, 46)
C4-C5-C6	119.7	(0.6, 46)
C5-C6-N1	122.7	(0.5, 46)
N1-C2-O2	122.8	(0.7, 46)
N3-C2-O2	122.2	(0.7, 46)
N3-C4-O4	119.4	(0.7, 46)
C5-C4-O4	125.9	(0.6, 46)
C6-N1-C1'	121.2	(1.4, 46)
C2-N1-C1'	117.7	(1.2, 46)

Table 5-2. Parameters of bond lengths and angles estimated for furanose rings (Gelbin et al. 1996).

	Ribose		Deoxyribose	
Bond or Angle	**Mean**	**(esd, N)**	**Mean**	**(esd, N)**
C1'-C2'	1.528	(0.010, 80)	1.521	(0.014, 47)
C2'-C3'	1.525	(0.011, 80)	1.518	(0.010, 47)
C3'-C4'	1.524	(0.011, 80)	1.528	(0.010, 47)
C4'-O4'	1.453	(0.012, 80)	1.446	(0.011, 47)
O4'-C1'	1.414	(0.012, 80)	1.420	(0.013, 47)
C3'-O3'	1.423	(0.014, 80)	1.431	(0.013, 47)
C5'-C4'	1.510	(0.013, 80)	1.511	(0.008, 47)
C2'-O2'	1.413	(0.013, 80)	na	
C1'-N1/N9	1.471	(0.017, 80)	1.474	(0.020, 47)
(H)O5'-C5'	1.423	(0.014, 61)	1.420	(0.021, 37)
C1'-C2'-C3'	101.5	(0.9, 80)	102.7	(1.4, 47)
C2'-C3'-C4'	102.7	(1.0, 80)	103.2	(1.0, 47)
C3'-C4'-O4'	105.5	(1.4, 80)	105.6	(1.0, 47)
C4'-O4'-C1'	109.6	(0.9, 80)	109.7	(1.4, 47)
O4'-C1'-C2'	106.4	(1.4, 80)	106.1	(1.0, 47)
C1'-C2'-O2'	110.6	(3.0, 80)	na	
C3'-C2'-O2'	113.3	(2.9, 80)	na	
C2'-C3'-O3'	111.0	(2.8, 80)	110.6	(2.7, 47)
C4'-C3'-O3'	110.6	(2.6, 80)	110.3	(2.2, 47)
C5'-C4'-C3'	115.5	(1.5, 80)	114.7	(1.5, 47)
C5'-C4'-O4'	109.2	(1.4, 80)	109.4	(1.6, 47)
O4'-C1'-N1/N9	108.2	(1.0, 80)	107.8	(0.8, 47)
C2'-C1'-N1/N9	113.4	(1.6, 80)	114.2	(1.6, 47)
(H)O5'-C5'-C4'	111.6	(1.7, 61)	110.9	(2.0, 37)
C1'-N9-C4	127.1	(1.8, 46)	125.9	(1.4, 13)
C1'-N1-C2	117.9	(1.3, 34)	117.8	(1.4, 34)

Recently, after nearly two decades without revision Mariusz Jaskolski and co-workers have been reinvestigating the nucleic acids restraints (Kowiel, Brzezinski, and Jaskolski 2016) (Gilski et al. 2019) (Kowiel et al. 2020). They divided their research (as Parkinson before) into three parts - the phosphodiester group, the nucleobase fragment and the sugar moiety. The result of their work was the creation of the newest dictionaries for each of these parts.

In August 2020, the Nucleic Acid Valence Geometry Working Group was established. Its task is defining and implementing a uniform dictionary for nucleic acid valence geometry parameters for modeling, improvement and validation of nucleic acids. The reason for creating this group was the fact that the software tools used for refinement and validation rely on different values for the valence geometry.

5.1 Phosphodiester group

This part of nucleic acids is a critical node in the sugar-phosphate backbone and a large electron-rich region of nucleic acids. These properties are the reason why the correct parameterization for refinement is essential here as well.

In the dictionary compiled by Gelbin *et al.,* two types of the angle of this bond were defined - *small* and *large* (Gelbin et al. 1996). However, this division did not propagate to software tools because the study was based on a small number of samples (13 structures) and also because it was not possible to determine where which angle would occur. For these reasons, the distinction has been abandoned.

In a later analysis of the Jaskolski group, these geometries were investigated on phosphodiester fragments (C-O-PO_2-O-C) found in CSD (Kowiel, Brzezinski, and Jaskolski 2016). Fragments found in CSD were divided into six groups: our cyclic (R5, R6, R7, R8) and two linear ones (AS, AA), of which category AA was described for the first time (Kowiel, Brzezinski, and Jaskolski 2016). The obtained restraints were compared with data from Nucleic Acid Database and checked on PDB Z-DNA structure 3p4j in ultra-high resolution (Brzezinski *et al.*, 2011). Based on these comparisons, it was confirmed that the restrains are more accurate than the previous ones from the Gelbin dictionary. This research group also created a web server called RestraintLib (http://achesym.ibch.poznan.pl/restraintlib/) that automatically provides the PO4 restraints for a given .pdb file (Kowiel, Brzezinski, and Jaskolski 2016).

5.2 Nucleobase fragment - Watson-Crick base pairing

In 2019, new optimal values of lengths and bond angles in nucleobase fragment of nucleic acids, specifically in the case of Watson-Crick pairing, were published (Gilski et al. 2019). The motivation for this research was the update of previous analyzes performed by the Parkinson's group in 1996 (Parkinson et al. 1996).

These new values were based on structures from CSD, quantum-mechanical (QM) calculations and two ultra-high-resolution nucleic acid crystal structures from the Protein Data Bank (Gilski et al. 2019). Combining these three sources and many more accessible structures has been a significant improvement since the previous study on this topic by Parkinson. The result of this work, in addition to measuring and updating Parkinson's

results for the bond lengths and angles for Watson-Crick pairs and individual nucleotides, also supplemented data for the missing nucleotide pair isocytosine and isoguanine. They also compared the angle and distance between isolated bases and paired bases and found that specific visible differences exist, but they are minimal - bases were found to have fixed geometries. These analyzes were compared on two oligonucleotide structures available in the PDB (3p4j and 1d8g). These structures were both described in high resolution (< 0.8 Å), and no geometry restraints were used during their refinement.

5.3 The sugar moiety

The last part is the glycosidic moiety of nucleic acid chains, which is the most flexible region of nucleic acids. At the same time, it is strongly influenced by the pucker given as the combination of the torsion angles. For these reasons, the analysis of this part is the most difficult one.

Structures from CSD that contained ribose or deoxyribose linked to purine or pyrimidine were selected for Kowiel's study – the dataset consisted of 30 ribose- purine sugar-base cases, 51 deoxyribose- purine cases, 84 ribose - pyrimidine cases and 167 deoxyribose - pyrimidine cases (Kowiel et al. 2020), which was a big difference compared to the previous datasets from which the Parkinson's scientific group drew (Parkinson et al. 1996). The relevant parameters were then calculated from these structures, which were further conformationally grouped and statistically analyzed.

After evaluating the relevant parameters, the structures were divided into subgroups, and it was searched whether there were differences between the individual groups or not. Potential subgroups were defined by (i) ring pucker (C2′-*endo*, C3′- *endo*, Other), (ii) χ torsion angle rotamer (syn, anti), (iii) γ torsion angle rotamer (trans, gauche +, gauche–), (iv) sugar type (ribose, 2′-deoxyribose), (v) base type (purine, pyrimidine) (Kowiel et al. 2020). Ring puckers were partitioned on the same principle like before in the previous study by Gelbin *et al.* (Gelbin et al. 1996), which means that the C3′- *endo* orientation was defined for the pseudorotation angle P in the range $0° \leq P \leq 36°$ and the C2′-*endo* orientation for $144° \leq P \leq 180°$; other options fall into the Other category. When it was found that the individual values differ from each other within the category, then the restraints were calculated separately for each subgroup. When statistically significant multiple sets of subgroups were found, it was verified whether the majority of subgroups

defined using pairs of these variables were also significant. If so, the restraints were based on the variable combination. Based on this research, it was clear that bond lengths and angles are dependent on several variables in these parts of nucleic acids, including conformational parameters.

As a result of Kowiel's work, new conformationally dependent stereochemical restrictions within groups, including their functional relationships have been proposed. The new restraints are available at the RestraintLib site, together with restraints for other parts of nucleic acids structures mentioned above. These restraints were validated against NDB and ultra-high-resolution PDB structures. Then the restraints were also used for re-refinement of PDB structures across a wide range of resolutions.

6 Project

This project aimed to compare dinucleotide data from the CSD with the latest version of stereochemical restraints and the values from PDB structures.

6.1 Procedure

For this study, we collected a set of structures containing all the atoms defining parameters shown in Figure 2-5. Unlike the work of the Jaskolsi group, who chose nucleotides fragments for their research (Kowiel, Brzezinski, and Jaskolski 2016), (Gilski et al. 2019), (Kowiel et al. 2020), we chose dinucleotides directly. We also did not consider the R factor values. This selection was made from CSD using MOGUL (Bruno et al. 2004) (22. 6. 2020). Graphical input was used to select suitable structures. The result was 83 samples (their list is given in Table 6-1). Upon closer examination of the result, it was found that some files do not contain any data of atomic coordinates. Therefore, the respective files could not be included in the study. That is why we continued to work with only fifty structures.

Table 6-1. Alphabetical list of codes of all found files.

ACCYGA10	ACCYGB10	ACRACG40	ACRCYT	ADPAPF	ADPAPF10	ADURAC	ADYPUR10	ALOZIK	ALOZOQ	AMADUR
APAPAD01	APAPAD10	BAJKAY	BAJKAY10	BAWJAK	BAWJAK10	BEJXET	BOLHEP	BOTLAX	BUYGEH	CAADNS
CAADNS10	CACNUP	CAGUCP10	CEXBOW	CEXBUC	CIGWOE	CIGYEW	COSTOT	CPAPRF	CPGTPH	DABRED
DETLIX	DETLIX01	DETLIX10	DETLIX11	DETLIX20	DEZMAW	DINYII	DINYII10	DUVDON	DUVDON10	EICGUA
ETCYGU	ETCYGU10	ETHIUA10	ETHUAD	ETHUAD10	FEDFID	FEDFID01	GIFBAY	GUPCYT20	ICYGET10	IFODUF
KELDAG	LELZIL	PFCYGU10	PFDOCT	PFLCPA	PFLCPB	PFLCPC	PFLCPC01	PFLCPC10	PFLCPC11	PMICGU10
QEXHAC	QOCVIP	QOKJUV	QOTFOV	SAHDIP	SATCEW	SATCEW01	SIKJAX	SIKJEB	SIWWIE	SIWWIE10
SIWWOK	SIWWOK10	SOYWIM	SUKHUB	THYTHY10	URPOAD10					

All of the files needed to be supplemented with additional information (for example atom name, residue name, chain name were missing) to make it possible to be processed by DNATCO web server. This stage was done manually using PyMOL (The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.). Other problems such as too long bonds or missing atoms were discovered during the completion of information in some structures. For these reasons, other steps, for which it was not possible to correctly calculate the required parameters due to the data, were excluded from the analysis. On the other hand, some structures were cyclic, and, thanks to this, two possible step conformations were available for one cyclic step.

After the structures were inspected and supplemented with missing information, their angles, torsions and distances were calculated by DNATCO (Jiří Černý et al. 2020) version 4.0. Based on these measured data, the individual steps were assigned to the given NtC class. A total of 97 different steps were compared, of which 44 steps were assigned to specific NtC classes (exact numbers of steps in NtC classes are shown in Table 6-2), and the remaining 53 steps were assigned to NANT class for geometrically unclassified steps. In total, we obtained samples from 6 NtC groups plus the NANT group. In contrast, a total of 96 classes plus NANT were defined based on the PDB database (Jiří Černý et al. 2020).

Table 6-2. Count of steps associated with NtC classes.

NtC class	AA00	AB04	BB16	IC01	NANT	OP15	ZZ1S	Total amount
Count of steps	9	6	1	25	53	1	2	97

The next stage was to compare bonds and angles with the results from the PDB and the resulting values of the Jaskolsi group (Kowiel et al. 2020). To do this, it was necessary to divide the structures into groups according to the Jaskolsi group. We divided the structures according to whether they contain ribose or deoxyribose and according to the sugar puckers. Then within the individual groups, the averages, esd (estimated standard deviation) and mode in atomic distances and angles were monitored and compared. The comparison was made against the dictionary of restraints and structures from PDB across the different resolution.

6.2 Results and Discussion

After searching through all dinucleotides in the Cambridge Structural Database, we obtained only 97 sufficiently described steps, which is a tiny number for accurate statistical analysis. This is the first difference compared to previous analyzes. These analyses were based on all found fragments composed of the required section of nucleic acids sorted by R-factor. There is another thing connected with that – from this dataset, there were found only 6 NtC classes, which is a big contrast to the PDB database, where the total number of NtC classes is 96. These results show that the conformational space in the PDB database is much richer than in CSD, but stereochemical restraints were created based on the data from CSD. All these steps, including their assigned NtC classes (or the nearest classes in the case of NANT), are in Table 6-3.

Table 6-3. CSD-derived steps and their assignment to the NtC classes, differences for 12 parameters, and root-mean-square deviation of atomic positions (rmsd).

step_ID	NtC	CANA	nearest_NtC	dd1	de1	dz1	da2	db2	dg2	dd2	dch1	dch2	dNN	dCC	dmu	rmsd
accyga10_A_C_1_G_2	IC01	ICL	IC01	0.60	-2.50	5.00	-3.30	15.90	-1.10	0.30	-2.30	-17.60	-0.04	0.00	1.10	0.18
accyga10_B_C_1_G_2	IC01	ICL	IC01	-1.90	6.20	-5.60	2.20	5.60	5.90	-6.60	1.20	-5.90	-0.04	-0.15	2.50	0.26
accygb10-f_A_C_1_G_2	NANT	NAN	IC01	19.20	4.70	4.50	8.00	50.60	-51.60	-23.10	-4.60	-7.60	0.04	0.30	24.30	0.70
accygb10-f_B_C_1_G_2	NANT	NAN	IC01	-3.70	0.60	-16.70	25.50	-16.90	-1.80	-26.60	0.70	-12.90	-0.28	-0.30	31.40	0.41
acracg40-f_A_C_1_G_2	NANT	NAN	IC01	-10.60	16.70	21.60	-16.90	-2.60	12.10	-35.00	4.60	3.70	-0.01	0.17	1.10	0.46
acracg40-f_B_C_1_G_2	IC01	ICL	IC01	2.60	-4.20	6.10	3.40	5.70	-16.20	-15.00	23.30	5.50	-0.21	-0.27	-7.60	0.23
acracg40-f_C_C_1_G_2	NANT	NAN	IC01	15.70	-3.20	20.70	-2.10	-14.80	3.40	-38.90	-6.90	-29.30	0.21	-0.09	2.40	0.34
acracg40-f_D_C_1_G_2	NANT	NAN	AB04	18.40	-5.60	-10.60	14.60	13.10	-10.20	17.20	10.80	32.90	1.49	0.83	-9.50	0.62
adpapf10_A_A_1_A_2	NANT	NAN	OP17	-2.60	-4.80	-2.90	0.00	39.40	-6.90	1.00	179.30	63.40	0.81	0.95	15.70	0.69
adypur10-f_A_A_1_U_2	AA00	AAA	AA00	-4.00	15.10	-3.90	1.70	-4.30	3.70	-4.60	-10.90	0.50	0.18	0.19	-3.70	0.14
adypur10-f_B_A_1_U_2	AA00	AAA	AA00	1.60	7.20	5.10	-5.40	4.40	2.70	-8.40	-6.20	1.60	0.21	0.26	-5.70	0.18
apapad10_A_A_1_A_2	AA00	AAA	AA00	0.00	16.40	-4.50	3.80	-11.40	1.00	-0.90	-2.70	3.60	-0.14	-0.03	1.90	0.16
apapad10_A_A_2_A_3	NANT	NAN	ZZ01	0.40	-1.00	28.40	-73.30	38.80	12.30	-68.50	-3.60	-22.90	2.09	1.85	43.40	0.97
bajkay10_A_DI_1_DA_2	NANT	NAN	ZZ02	-10.90	-57.40	13.10	-46.80	11.10	7.60	-6.80	72.10	23.80	2.10	2.07	14.60	1.02
bawjak10_A_DC_1_DG_2	BB16	BBw	BB16	7.60	-16.70	11.30	0.20	2.10	5.50	4.80	21.10	11.50	0.03	-0.21	34.00	0.51
bejxet_A_DG_1_DC_2	NANT	NAN	OP21	23.10	-19.60	-25.00	10.90	4.00	-34.00	5.50	41.50	-5.70	0.18	0.16	7.00	0.72
bolhep_A_G_1_C_2	AA00	AAA	AA00	4.30	5.30	10.20	-5.50	7.10	1.40	-5.10	-15.60	-1.30	0.35	0.30	-7.20	0.25
botlax-f_A_DA_1_DA_2	NANT	NAN	BB03	-5.60	33.50	27.40	9.00	-0.20	2.70	4.20	66.40	52.10	1.89	1.33	23.90	1.02
botlax-f_B_DA_1_DA_2	NANT	NAN	BB03	-4.50	35.00	29.40	10.90	-2.00	6.30	0.90	63.00	52.30	1.92	1.40	21.80	1.01
buygeh_A_DA_1_DT_2	NANT	NAN	OP22	-1.10	-42.80	-7.80	-17.70	-3.50	2.90	-14.50	26.10	-3.30	1.28	1.10	14.90	1.03
caadns10_A_A_1_A_2	NANT	NAN	OP05	68.90	15.10	10.40	-8.30	60.20	0.90	-7.20	-160.10	-0.50	0.94	-0.22	19.60	1.16
cagucp10_A_G_1_C_2	AA00	AAA	AA00	-9.00	10.40	3.00	1.00	-0.80	2.50	-2.20	-10.40	-5.90	0.08	0.11	3.00	0.15
cagucp10_B_G_1_C_2	AA00	AAA	AA00	-5.70	16.00	5.70	-2.30	8.60	-7.80	-3.00	-8.30	0.40	0.35	0.35	-9.80	0.22
cagucp10_C_G_1_C_2	AA00	AAA	AA00	5.80	10.10	0.40	-11.00	8.60	-2.60	5.40	-9.60	3.40	0.30	0.37	-3.10	0.22
cagucp10_D_G_1_C_2	AA00	AAA	AA00	-2.20	17.20	1.80	-7.60	-6.00	7.80	-8.20	-9.70	0.70	0.08	0.14	-5.20	0.17
cigwoe-f_B_DC_1_DG_2	IC01	ICL	IC01	-5.80	-12.00	2.20	-7.30	-1.20	20.00	-16.90	-4.60	4.00	-0.11	-0.09	-4.70	0.26
cigwoe-f_A_DC_1_DG_2	IC01	ICL	IC01	-4.50	-14.80	4.30	-5.80	2.40	2.70	13.20	-0.90	-5.00	-0.09	-0.23	-0.70	0.26
cigyew_A_C_1_G_2	IC01	ICL	IC01	-0.60	0.10	-2.80	1.00	5.30	-4.70	-9.80	-2.30	3.60	0.08	0.14	-6.60	0.15
cigyew_B_C_1_G_2	IC01	ICL	IC01	1.70	2.50	-0.90	-12.40	0.80	4.20	-7.40	0.10	10.60	0.05	0.12	-5.50	0.20
costot-f_A_DT_1_DA_2	NANT	NAN	OP17	9.30	4.90	-7.80	7.90	-28.70	-4.20	22.60	15.00	-145.00	-0.35	-0.29	-36.40	0.50
costot-f_B_DT_1_DA_2	NANT	NAN	OP17	-8.00	5.20	-2.60	0.50	19.90	9.10	-55.20	18.40	-132.30	0.76	0.73	7.40	0.81
dabred-f_A_U_1_G_2	NANT	NAN	AB04	-3.10	-10.90	-20.40	17.50	9.70	-3.90	-40.00	1.40	32.70	1.71	0.79	7.30	0.72
dabred-f_B_C_1_A_2	IC01	ICL	IC01	-7.20	-15.60	15.10	-16.00	4.50	7.80	-18.70	-15.00	-14.40	0.12	0.09	5.80	0.29
detlix01-new_A_DG_1_DG_2	AB04	A-B	AB04	15.10	17.60	-6.50	31.10	-9.70	-12.40	-2.60	25.10	0.90	-0.16	0.20	-5.60	0.29
detlix01-new_B_DG_1_DG_2	AB04	A-B	AB04	4.20	19.90	-8.80	23.50	-17.40	-8.70	9.10	28.50	9.90	-0.13	0.22	-2.50	0.34
detlix01-new_C_DG_1_DG_2	NANT	NAN	AB01	7.60	32.20	14.00	-26.20	38.60	-23.70	5.50	43.20	11.60	0.46	0.77	-15.00	0.65
detlix11-f_A_DG_1_DG_2	AB04	A-B	AB04	13.60	12.70	-1.40	24.60	-8.10	-14.20	3.80	23.90	0.20	-0.07	0.29	-5.60	0.29
detlix11-f_B_DG_1_DG_2	AB04	A-B	AB04	4.00	21.50	-9.30	21.60	-17.30	-11.40	8.20	28.10	3.90	-0.14	0.25	0.60	0.34
detlix20_A_G_1_G_2	AB04	A-B	AB04	-17.30	12.00	11.70	3.60	-15.30	1.80	-3.20	15.60	-19.00	-0.37	-0.16	0.30	0.27
detlix20_B_G_1_G_2	NANT	NAN	AA02	0.20	14.70	20.40	47.60	10.60	-40.70	48.80	40.10	28.80	0.38	0.91	-21.00	0.62
detlix20_C_G_1_G_2	AB04	A-B	AB04	-4.40	25.70	-4.20	15.70	-4.50	-9.10	12.20	20.60	-1.20	-0.21	0.01	-3.40	0.28
detlix20_D_G_1_G_2	NANT	NAN	AB01	5.10	29.20	17.00	-26.90	15.60	12.10	-12.70	59.50	-2.80	0.47	0.97	-15.30	0.61
dinyii10-f_A_DC_1_DG_2	NANT	NAN	BB16	0.00	-14.70	5.60	6.10	7.30	7.80	7.80	19.30	11.20	0.24	-0.01	37.60	0.55
dinyii10-f_B_DC_1_DG_2	NANT	NAN	BB16	4.00	-13.80	11.40	0.60	2.60	8.60	6.90	26.50	12.30	0.05	-0.20	38.10	0.54
duvdon10_A_DC_1_DG_2	NANT	NAN	BB16	-9.70	43.00	0.70	12.70	-41.60	-2.00	-23.80	4.40	17.60	-0.07	-0.17	-9.70	0.62
duvdon10_B_DC_1_DG_2	NANT	NAN	BB16	-20.20	50.10	7.40	9.90	-40.50	-1.40	-18.40	-7.60	21.10	0.19	0.03	-11.10	0.71
eicgua-f_A_C_1_G_2	IC01	ICL	IC01	14.70	28.60	-29.50	13.50	-20.00	-7.70	-7.10	-0.90	2.10	0.05	0.13	-2.30	0.25
eicgua-f_B_C_1_G_2	IC01	ICL	IC01	0.80	-12.70	-1.60	-23.00	1.60	13.70	-6.00	2.30	1.40	0.05	-0.12	-2.90	0.24
etcygu10-f_B_C_1_G_2	IC01	ICL	IC01	-0.80	9.60	-8.60	-11.70	2.10	5.20	-7.40	6.20	3.00	0.00	0.13	-6.10	0.17
etcygu10-f_A_C_1_G_2	IC01	ICL	IC01	-11.60	4.40	1.40	-9.80	7.90	0.90	-7.40	9.70	14.00	-0.11	-0.11	-4.60	0.29
ethiua10_A_U_1_A_2	IC01	ICL	IC01	15.40	-12.70	-3.60	-5.70	13.60	-3.10	-12.00	3.10	-4.20	-0.15	-0.17	-5.70	0.19
ethiua10_B_U_1_A_2	IC01	ICL	IC01	11.80	-1.50	12.60	-20.90	7.40	15.20	-27.10	-8.90	-3.00	-0.07	-0.08	-5.70	0.41
ethiuad10-f_A_U_1_A_2	IC01	ICL	IC01	4.60	-1.60	-1.00	-1.10	-0.60	7.30	-20.60	-9.20	-4.10	0.13	0.27	-8.90	0.17
ethiuad10-f_B_U_1_A_2	IC01	ICL	IC01	-2.60	7.80	6.30	-2.20	5.20	-4.10	-9.10	-10.30	8.80	0.08	0.09	-7.10	0.20
fedfid-f_C_DC_1_DG_2	NANT	NAN	OP12	8.10	0.20	5.50	20.70	-23.50	4.50	7.60	35.90	99.70	-0.46	-0.75	28.10	0.67
fedfid-f_C_DG_2_DG_3	NANT	NAN	AA02	2.30	11.40	33.50	-45.00	38.00	46.30	17.70	46.20	-1.10	0.42	0.88	-25.10	0.67
fedfid-f_B_DC_1_DG_2	NANT	NAN	IC06	13.40	-22.30	13.90	-6.80	11.80	5.70	-44.70	-16.00	14.20	-0.17	-0.35	54.60	0.69
fedfid-f_B_DG_2_DG_3	NANT	NAN	AA02	9.10	25.30	22.70	5.60	25.90	-7.70	37.40	38.80	23.90	0.63	0.98	-24.70	0.64
fedfid-f_A_DC_1_DG_2	NANT	NAN	BB16	3.70	14.20	8.90	3.90	0.10	-7.70	-53.40	10.80	16.00	0.75	0.20	52.50	0.62
fedfid-f_A_DG_2_DG_3	NANT	NAN	BB03	-58.60	32.00	20.90	-18.30	-2.00	15.20	3.80	44.90	7.00	0.58	1.08	-39.00	0.74
gupcyt20_A_G_1_C_2	AA00	AAA	AA00	1.40	4.80	3.90	-8.50	11.80	-4.50	-4.70	0.00	8.00	0.28	0.45	-6.10	0.22
icyget10-f_A_C_1_G_2	IC01	ICL	IC01	1.40	5.60	1.80	-6.30	1.30	1.00	-11.00	-2.20	11.40	0.13	-0.03	-9.00	0.18
icyget10-f_B_C_1_G_2	IC01	ICL	IC01	4.30	6.30	-8.20	-10.90	-13.30	18.10	-13.70	6.90	3.80	0.12	0.08	-8.10	0.19
keldag-aa_A_DA_1_DA_2	NANT	NAN	OP05	18.00	-2.20	10.90	-3.00	69.20	9.50	9.30	-25.10	-0.60	-2.01	-1.22	21.00	1.18
keldag-aa_B_DA_1_DA_2	NANT	NAN	OP05	6.50	1.80	27.60	-5.70	66.70	0.80	4.40	55.60	65.30	0.47	-0.05	46.90	0.98
keldag-bb_A_DA_1_DA_2	NANT	NAN	OP05	15.30	-2.90	13.80	-1.50	70.90	7.70	12.00	-14.90	-10.80	-2.01	-1.22	21.00	1.20
keldag-bb_B_DA_1_DA_2	NANT	NAN	OP05	10.40	7.10	22.50	-2.50	64.90	2.60	0.50	51.00	69.90	0.47	-0.05	46.90	1.04
pfcygu10_A_C_1_G_2	NANT	NAN	IC01	-8.10	-15.80	2.20	-10.40	11.20	-1.30	-66.50	-1.00	-8.60	0.29	0.18	16.40	0.48
pfdoct-f_A_DC_1_DG_2	NANT	NAN	AB04	-2.30	-4.70	-6.30	9.90	9.90	-8.70	-44.60	5.20	17.80	1.69	1.04	6.40	0.69
pfdoct-f_B_DC_1_DG_2	IC01	ICL	IC01	-3.30	-16.50	10.20	-9.80	-4.90	18.00	4.30	-8.00	-1.00	-0.08	0.08	-2.00	0.27
pflcpc10-f_A_C_1_G_2	IC01	ICL	IC01	1.80	1.40	-0.30	7.30	6.10	-2.10	-2.60	-6.50	-2.70	0.09	0.20	4.50	0.23
pflcpc10-f_B_C_1_G_2	IC01	ICL	IC01	-3.20	-15.80	21.00	-13.30	11.00	-5.30	-4.80	-1.80	-8.20	0.08	-0.03	2.70	0.17
pflcp11_A_C_1_G_2	IC01	ICL	IC01	4.00	-6.80	13.50	-7.70	5.10	-1.50	-5.70	0.50	-2.20	-0.04	-0.13	3.30	0.24
pflcp11_B_C_1_G_2	IC01	ICL	IC01	-6.60	-8.00	-5.60	9.50	0.10	8.30	-2.80	-4.80	-5.10	0.08	0.05	1.10	0.26
pmicgu10-f_A_C_1_G_2	IC01	ICL	IC01	2.40	-7.80	-5.50	-19.70	10.70	18.70	-10.00	-1.80	0.30	0.12	0.09	-3.50	0.26
pmicgu10-f_B_C_1_G_2	IC01	ICL	IC01	4.70	5.10	-1.00	0.10	-3.10	2.30	-5.70	0.50	-3.80	-0.12	-0.17	5.90	0.22
qocvip-aa_A_G_1_G_2	NANT	NAN	OP05	15.80	1.20	15.50	1.30	61.70	13.60	0.10	-23.70	15.60	-2.15	-1.09	43.10	1.22
qocvip-aa_B_G_1_G_2	NANT	NAN	OP05	10.30	-4.90	13.20	4.10	66.10	8.90	0.20	-21.10	19.60	-2.13	-1.09	37.10	1.18
qocvip-bb_A_G_1_G_2	NANT	NAN	OP05	5.90	-5.40	15.10	2.00	72.70	0.20	9.80	1.30	-9.40	-2.15	-1.09	43.10	1.25
qocvip-bb_B_G_1_G_2	NANT	NAN	OP05	6.20	0.60	17.00	6.50	63.60	4.40	4.30	5.30	-6.80	-2.13	-1.09	37.10	1.22
sikjeb-chaina_A_DC_1_DG_2	ZZ1S	ZZZ	ZZ1S	2.20	7.20	2.80	-2.30	0.50	0.00	-11.00	-11.70	4.80	-0.32	-0.38	-9.20	0.24
sikjeb-chainb_B_DC_1_DG_2	ZZ1S	ZZZ	ZZ1S	4.70	-5.60	9.10	-3.20	4.20	7.60	4.90	-3.00	7.30	-0.13	-0.17	-3.90	0.22
siwwie10-aa_A_G_1_G_2	NANT	NAN	OP05	8.40	0.10	16.80	7.40	64.40	1.20	5.50	6.20	-8.60	-2.17	-1.10	32.70	1.22
siwwie10-aa_B_G_1_G_2	NANT	NAN	OP05	7.60	-3.90	14.70	4.40	66.80	9.30	1.90	2.20	-8.20	-2.12	-1.12	41.20	1.22
siwwie10-bb_A_G_1_G_2	NANT	NAN	OP05	11.50	-2.60	16.10	5.90	65.20	5.40	2.40	-22.90	20.50	-2.17	-1.10	32.70	1.18
siwwie10-bb_B_G_1_G_2	NANT	NAN	OP05	7.90	-2.80	13.40	4.60	63.90	10.60	1.60	-22.50	16.50	-2.12	-1.12	41.20	1.18
soywim-f3_A_DA_1_DG_2	NANT	NAN	OP21	0.10	34.20	26.20	3.50	-38.80	-12.10	10.00	47.30	32.20	-0.17	0.08	-1.20	0.84
soywim-f3_A_DG_2_DA_3	NANT	NAN	OP22	6.70	-28.40	31.00	10.50	6.70	2.80	0.80	43.00	-0.50	0.08	0.17	-20.20	0.58
soywim-f3_B_DA_1_DG_2	NANT	NAN	OP21	-17.70	10.70	-12.80	14.30	-8.10	-3.90	-2.70	71.70	36.40	0.10	0.21	21.20	0.39
soywim-f3_B_DG_2_DA_3	OP15	OPN	OP15	-7.40	-14.20	4.80	16.60	-2.10	0.20	3.30	15.70	6.30	-0.22	-0.28	7.70	0.33
sukhub-aa_A_G_1_G_2	NANT	NAN	OP05	12.60	-3.90	16.40	3.40	66.60	6.60	6.50	-22.40	21.40	-2.16	-1.23	39.10	1.21
sukhub-aa_B_G_1_G_2	NANT	NAN	OP05	9.30	-5.40	18.40	-0.80	64.80	14.50	-6.30	-22.50	16.10	-2.15	-1.20	42.20	1.20
sukhub-bb_A_G_1_G_2	NANT	NAN	OP05	12.50	-3.80	13.50	3.60	66.70	4.40	6.60	7.10	-8.10	-2.16	-1.23	39.10	1.23
sukhub-bb_B_G_1_G_2	NANT	NAN	OP05	0.30	-1.50	24.20	0.70	66.40	7.00	3.30	1.80	-8.20	-2.15	-1.20	42.20	1.24
thythy10_A_DT_1_DT_2	NANT	NAN	OP22	1.30	10.40	38.60	7.80	22.30	-7.20	10.10	-20.20	-10.20	0.76	0.60	4.20	0.92
urpoad10-f2_A_U_1_A_2	NANT	NAN	OP06	-4.70	11.90	21.00	-30.20	19.60	5.80	12.70	2.80	28.10	-0.34	0.16	9.30	0.46
urpoad10-f2_B_U_1_A_2	NANT	NAN	ZZ01	4.90	-4.40	32.30	-84.00	53.60	5.90	-62.30	-8.00	-18.50	2.26	1.76	29.40	1.09

Another finding was that most of the assigned structures fell into the NtC class IC01, which means intercalated structures. Another compound thus forced the conformations of these structures.

Irrespective of NtC classification we divided all the steps into two groups according to their being a part of RNA or DNA molecules. The mean values, esd and mode are shown in Table 6-4. Figure 6-1 shows the values from the article (Kowiel et al. 2020). As shown in the figure and the table, the data do not differ much from each other.

Table 6-4. Comparison of values measured from structures from PDB and from CSD.

PDB	DNA = 3864 structures		RNA = 2630 structures	
	mean	mode	mean	mode
Bonds				
C5'_C4'	1.511	1.510	1.508	1.504
C4'_C3'	1.525	1.525	1.521	1.519
C3'_O3'	1.429	1.427	1.419	1.415
C4'_O4'	1.444	1.446	1.451	1.450
O4'_C1'	1.431	1.421	1.417	1.414
C1'_C2'	1.518	1.520	1.524	1.525
C2'_C3'	1.517	1.519	1.520	1.521
C1'_N1	1.463	1.461	1.470	1.473
N1_C2	1.383	1.376	1.380	1.376
Angles				
C5'_C4'_C3'	115.02	115.27	115.86	115.91
C4'_C3'_O3'	108.31	109.29	110.57	110.48
C3'_C4'_O4'	105.41	105.91	103.42	103.43
C4'_O4'_C1'	109.02	109.73	109.23	109.50
O4'_C1'_C2'	105.68	105.24	106.80	107.21
C1'_C2'_C3'	102.25	102.48	100.36	100.57
C2'_C3'_C4'	103.67	103.48	101.20	101.84
O4'_C1'_N1	107.04	107.75	108.72	108.79
C2'_C1'_N1	115.49	115.37	112.72	112.96
C1'_N1_C2	122.58	118.38	122.25	125.58

CSD	DNA = 85 structures		RNA = 105 structures	
	mean	mode	mean	mode
Bonds				
C5'_C4'	1.516	1.512	1.510	1.506
C4'_C3'	1.530	1.528	1.529	1.519
C3'_O3'	1.452	1.432	1.430	1.433
C4'_O4'	1.448	1.448	1.456	1.447
O4'_C1'	1.417	1.419	1.413	1.411
C1'_C2'	1.522	1.516	1.524	1.535
C2'_C3'	1.529	1.522	1.523	1.527
C1'_N1	1.485	1.471	1.475	1.478
N1_C2	1.372	1.379	1.385	1.382
Angles				
C5'_C4'_C3'	114.77	114.88	114.55	115.88
C4'_C3'_O3'	108.67	108.11	109.55	110.28
C3'_C4'_O4'	111.92	105.65	110.95	104.61
C4'_O4'_C1'	108.06	108.91	107.40	106.01
O4'_C1'_C2'	107.14	107.87	108.47	108.98
C1'_C2'_C3'	105.13	103.95	104.12	100.80
C2'_C3'_C4'	101.98	103.17	102.00	101.07
O4'_C1'_N1	105.07	105.29	105.74	104.47
C2'_C1'_N1	109.32	109.32	110.52	110.33
C1'_N1_C2	119.99	114.86	117.28	117.03

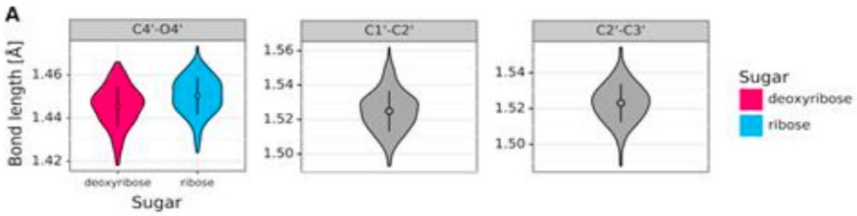

Figure 6-1. Bond length distributions in sugar fragments retrieved from the CSD (Kowiel et al., 2020).

We repeated a similar process for selected puckers. In this case, we primarily dealt with the *C2'-endo* and *C3'-endo* conformations. In both groups, the data matched. A slight deviation was observed in the rest of the data in the so-called "Other" group, so we decided to divide this group into other possible puckers. Unfortunately, we did not have enough data for a deeper statistical analysis.

We also focused on the calculated values within the individual NtC classes. In the Cambridge database, most structures were assigned to the NANT NtC class of conformational outliers and then to IC01, which indicates intercalated structures. The next

most numerous class that could be examined more closely was the AA00 class (representing the canonical A-form nucleic acid). In the Table 6-5 are the average bond lengths and angle sizes, as well as the mode and standard deviations for structures from both CSD and PDB. For a demonstration to this work, the structures from the PDB database were selected only with resolution better than or equal to 1.2 Å. At worse resolutions, the values approached stereochemical standards.

Table 6-5. Comparison of values calculated for structures within the NtC class AA00.

AA00

PDB				CSD		
bond/angle	**mean**	**esd**	**mode**	**mean**	**esd**	**mode**
O3'-P	1.600	0.014	1.602	1.604	0.020	1.614
P-O5'	1.592	0.012	1.591	1.586	0.018	1.586
C5'-O5'	1.429	0.016	1.418	1.446	0.032	1.440
P-OP1	1.482	0.014	1.481	1.469	0.026	1.478
P-OP2	1.478	0.015	1.480	1.480	0.029	1.490
C5'-C4'	1.508	0.012	1.504	1.508	0.047	1.506
C4'-C3'	1.520	0.012	1.519	1.564	0.053	1.529
C3'-O3'	1.420	0.015	1.415	1.409	0.048	1.421
C4'-O4'	1.450	0.009	1.449	1.462	0.053	1.450
O4'-C1'	1.420	0.014	1.415	1.435	0.036	1.418
C1'-C2'	1.522	0.019	1.525	1.548	0.031	1.546
C2'-C3'	1.519	0.013	1.520	1.532	0.048	1.536
C1'-N1	1.471	0.014	1.472	1.472	0.042	1.486
N1-C2	1.382	0.014	1.375	1.392	0.031	1.385
C3'-O3'-P	119.981	1.659	119.916	121.577	1.374	121.059
O3'-P-O5'	104.257	1.452	104.077	103.541	1.560	104.040
P-O5'-C5'	120.215	1.759	120.593	118.736	1.598	119.166
O5'-C5'-C4'	109.528	1.719	109.939	107.488	2.312	108.598
O3'-P-OP1	105.783	1.860	105.880	104.080	0.964	103.842
O3-P-OP2	109.643	1.763	109.767	111.039	1.140	110.835
OP1-P-OP2	119.880	1.579	119.771	119.309	1.703	119.776
C5'-C4'-C3'	116.233	1.641	115.950	116.278	1.935	117.025
C4'-C3'-O3'	110.714	1.952	110.500	112.913	3.860	113.178
C3'-C4'-O4'	103.166	1.087	103.375	112.035	8.505	110.686
C4'-O4'-C1'	109.154	1.128	109.412	109.185	2.038	109.251
O4'-C1'-C2'	107.050	1.019	107.248	107.313	2.925	106.503
C1'-C2'-C3'	100.615	1.256	100.625	106.340	5.565	107.875
C2'-C3'-C4'	101.122	1.365	101.749	101.258	1.911	100.462
O4'-C1'-N9	108.569	1.609	108.734	106.025	2.602	105.483
C2'-C1'-N1	112.685	2.611	112.877	112.038	2.458	112.728
C1'-N1-C2	121.209	4.306	117.297	117.908	2.751	117.910

Figure 6-2 is a histogram of the length of C5' - C4' bond from PDB data in 0 - 1.2 Å resolution belonging to NtC class AA00. Red line indicates the value of the standard mentioned above in the Table 5-2. Green lines indicate -5×esd and +5×esd from the standard value. Blue symbols indicate the values obtained in this work when analyzing CSD data. Figure 6-3, which describes the values at the bond angle C1'-C2'-C3', has the same properties.

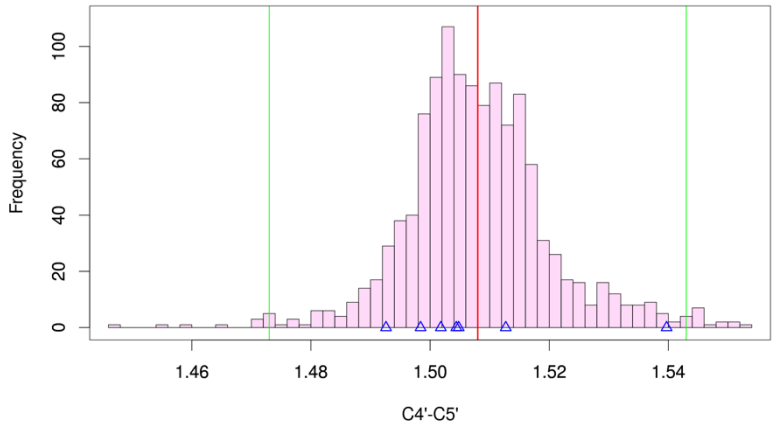

Figure 6-2. Histogram of bond lengths from PDB data of NtC class AA00 in resolution 0 - 1.2 Å.

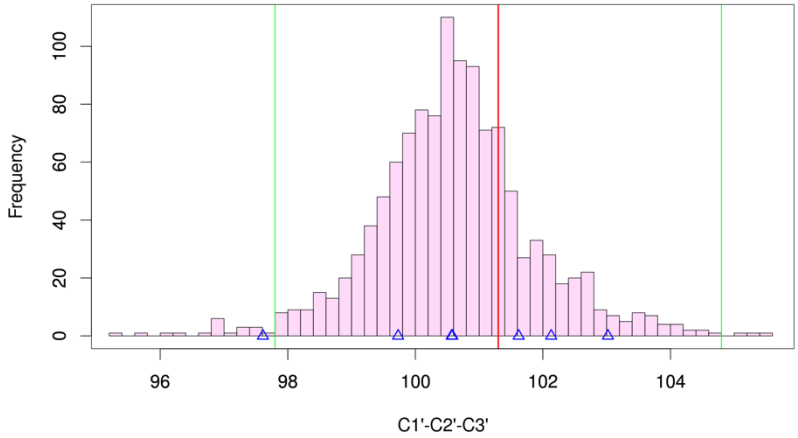

Figure 6-3. Histogram of bond angles from PDB data of NtC class AA00 in resolution 0 - 1.2 Å.

Data analysis revealed that almost all CSD data values belonging to class AA00 fall into the ±5esd range from the standard value as well as from the average of PDB structures with resolution 0 - 1.2 Å.

7 Conclusion

This thesis summarizes the development of stereochemical restraints used for nucleic acid structure refinement. Dictionaries of restraints have not been updated for a long period of time. Only in recent years have their updates begun to appear, which is, among other things, caused by a significant increase in the number of new structures available in the CSD database. This database has been the main source of data for all dictionaries created so far. In this study, the latest published updates of stereochemical dictionaries (Kowiel, Brzezinski and Jaskolski, 2016), (Gilski et al., 2019), (Kowiel et al., 2020) are discussed.

To work with structures from CSD, the NtC class conformational system was used. It describes the conformations of nucleic acids at the local level. Using this system, it was also revealed that the CSD database covers much smaller conformational space than the PDB database. Nevertheless, the stereochemical restraints are derived only from conformationally sparse data from the CSD and based on them, the refinement standards for PDB structures are made.

In this work, the last updated stereochemical restrains with the values obtained from other initial datasets were compared using selected dinucleotide structures from CSD and PDB databases. Comparison of the data within several different groups divided according to their different conformations (NtC, pucker, etc.) revealed that the measured data agree in most aspects, but stereochemical restraints are inaccurate within some groups.

Finally, one can conclude that dictionaries of stereochemical restraints still need new and more precise adjustments so that they are applied properly and do not reduce the quality of experimentally obtained data.

8 References

Altona, C. and Sundaralingam, M. (1972) 'Conformational Analysis of the Sugar Ring in Nucleosides and Nucleotides. a New Description Using the Concept of Pseudorotation', *Journal of the American Chemical Society*, 94(23), pp. 8205–8212. doi: 10.1021/ja00778a043.

Avery, O. T., Macleod, C. M. and McCarty, M. (1944) 'Studies on the chemical nature of the substance inducing transformation of pneumococcal types: Induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type iii', *Journal of Experimental Medicine*, 79(2), pp. 137–158. doi: 10.1084/jem.79.2.137.

Berman, H. M. *et al.* (2000) 'The Protein Data Bank', *Nucleic Acids Research*, pp. 235–242. doi: 10.1093/nar/28.1.235.

Bernstein, F. C. *et al.* (1977) 'The protein data bank: A computer-based archival file for macromolecular structures', *Journal of Molecular Biology*, 112(3), pp. 535–542. doi: 10.1016/S0022-2836(77)80200-3.

Bochkarev, A. *et al.* (1997) 'Structure of the single-stranded-DNA-binding domain of replication protein A bound to DNA', *Nature*, 385(6612), pp. 176–181. doi: 10.1038/385176a0.

Bourne, P. E. *et al.* (1997) 'Macromolecular crystallographic information file', *Methods in Enzymology*, 277, pp. 571–590. doi: 10.1016/S0076-6879(97)77032-0.

Brown, I. D. and McMahon, B. (2006) 'The Crystallographic Information File (CIF)', *Data Science Journal*, pp. 174–177. doi: 10.2481/dsj.5.174.

Bruno, I. J. *et al.* (2004) 'Retrieval of crystallographically-derived molecular geometry information', *Journal of Chemical Information and Computer Sciences*. J Chem Inf Comput Sci, 44(6), pp. 2133–2144. doi: 10.1021/ci049780b.

Bruno, I. J. and Groom, C. R. (2014) 'A crystallographic perspective on sharing data and knowledge', *Journal of Computer-Aided Molecular Design*, 28(10), pp. 1015–1022. doi: 10.1007/s10822-014-9780-9.

Brzezinski, K. *et al.* (2011) 'High regularity of Z-DNA revealed by ultra high-resolution crystal structure at 0.55', *Nucleic Acids Research*, 39(14), pp. 6238–6248. doi: 10.1093/nar/gkr202.

Burley, S. K. *et al.* (2017) 'Protein Data Bank (PDB): The single global macromolecular structure archive', in *Methods in Molecular Biology*, pp. 627–641. doi: 10.1007/978-1-4939-7000-1_26.

Černý, J. *et al.* (2020) 'A unified dinucleotide alphabet describing both RNA and DNA structures', *Nucleic acids research*, 48(11), pp. 6367–6381. doi: 10.1093/nar/gkaa383.

Černý, J., Božíková, P. and Schneider, B. (2016) 'DNATCO: assignment of DNA conformers at dnatco.org', *Nucleic Acids Research*, 44. doi: 10.1093/nar/gkw381.

Clowney, L. *et al.* (1996) 'Geometric parameters in nucleic acids: Nitrogenous bases', *Journal of the American Chemical Society*, 118(3), pp. 509–518. doi: 10.1021/ja952883d.

Evans, P. R. (2006) 'An introduction to stereochemical restraints', in *Acta Crystallographica Section D: Biological Crystallography*, pp. 58–61. doi: 10.1107/S090744490604604X.

Gao, Y.-G., Robinson, H. and Wang, A. H.-J. (1999) 'High-resolution A-DNA crystal

structures of d(AGGGGCCCCT). An A-DNA model of poly(dG).poly(dC)', *European Journal of Biochemistry*, 261(2), pp. 413–420. doi: 10.1046/j.1432-1327.1999.00270.x.

Gelbin, A. *et al.* (1996) 'Geometric parameters in nucleic acids: Sugar and phosphate constituents', *Journal of the American Chemical Society*, 118(3), pp. 519–529. doi: 10.1021/ja9528846.

Gilski, M. *et al.* (2019) 'Accurate geometrical restraints for Watson–Crick base pairs', *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials*, 75(2), pp. 235–245. doi: 10.1107/S2052520619002002.

Griffith, F. (1928) 'The Significance of Pneumococcal Types', *Journal of Hygiene*, 27(2), pp. 113–159. doi: 10.1017/S0022172400031879.

Groom, C. R. *et al.* (2016) 'The Cambridge Structural Database', *Acta Cryst*, 72, pp. 171–179. doi: 10.1107/S2052520616003954.

Groom, C. R. and Allen, F. H. (2014) 'The Cambridge Structural Database in Retrospect and Prospect', *Angewandte Chemie International Edition*, 53(3), pp. 662–671. doi: 10.1002/anie.201306438.

Guan, Y. *et al.* (1993) 'Molecular structure of cyclic diguanylic acid at 1 Å resolution of iwo crystal forms: Self-association, interactions with metal ion/planar dyes and modeling studies', *Journal of Biomolecular Structure and Dynamics*, 11(2), pp. 253–276. doi: 10.1080/07391102.1993.10508725.

Hall, S. R. (1991) 'The STAR File: A New Format for Electronic Data Transfer and Archiving', *Journal of Chemical Information and Computer Sciences*, 31(2), pp. 326–333. doi: 10.1021/ci00002a020.

Hall, S. R., Allen, F. H. and Brown, I. D. (1991) 'The crystallographic information file (CIF): a new standard archive file for crystallography', *Acta Crystallographica Section A Foundations of Crystallography*, 47(6), pp. 655–685. doi: 10.1107/S010876739101067X.

Hershey, A. D. and Chase, M. (1952) 'Independent functions of viral protein and nucleic acid in growth of bacteriophage.', *The Journal of general physiology*, 36(1), pp. 39–56. doi: 10.1085/jgp.36.1.39.

Jones, D. (1997) 'Structural Databases', in *Genetic Databases*, pp. 215–239. doi: 10.1016/b978-012101625-8/50013-x.

Kowiel, M. *et al.* (2020) 'Conformation-dependent restraints for polynucleotides: The sugar moiety', *Nucleic Acids Research*, 48(2), pp. 962–973. doi: 10.1093/nar/gkz1122.

Kowiel, M., Brzezinski, D. and Jaskolski, M. (2016) 'Conformation-dependent restraints for polynucleotides: I. Clustering of the geometry of the phosphodiester group', *Nucleic Acids Research*, 44(17), pp. 8479–8489. doi: 10.1093/nar/gkw717.

Leontis, N. B. and Westhof, E. (2001) 'Geometric nomenclature and classification of RNA base pairs', *RNA*, 7(4), pp. 499–512. doi: 10.1017/S1355838201002515.

Narayana, N. and Weiss, M. A. (2009) 'Crystallographic Analysis of a Sex-Specific Enhancer Element: Sequence-Dependent DNA Structure, Hydration, and Dynamics', *Journal of Molecular Biology*, 385(2), pp. 469–490. doi: 10.1016/j.jmb.2008.10.041.

Neidle, S., Schneider, B. and Berman, H. M. (2005) *Fundamentals of DNA and RNA Structure*, *Structural Bioinformatics*. doi: 10.1002/0471721204.ch3.

Newton, M. G. *et al.* (2005) 'A non-natural dinucleotide containing an isomeric L-related

deoxynucleoside: Dinucleotide inhibitors of anti-HIV integrase activity', *Acta Crystallographica Section C: Crystal Structure Communications*, 61(8), pp. o518–o520. doi: 10.1107/S0108270105019037.

Parkinson, G. *et al.* (1996) 'New parameters for the refinement of nucleic acid-containing structures', *Acta Crystallographica Section D: Biological Crystallography*, 52(1), pp. 57–64. doi: 10.1107/S0907444995011115.

Pauling, L., Corey, R. B. and Branson, H. R. (1951) 'The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain.', *Proceedings of the National Academy of Sciences of the United States of America*, 37(4), pp. 205–211. doi: 10.1073/pnas.37.4.205.

Richardson, J. S. *et al.* (2008) 'RNA backbone: Consensus all-angle conformers and modular string nomenclature (an RNA Ontology Consortium contribution)', *RNA*, 14(3), pp. 465–481. doi: 10.1261/rna.657708.

Saenger, W. (1984) *Principles of Nucleic Acid Structure*. Springer New York. doi: 10.1007/978-1-4612-5190-3.

Schneider, B. *et al.* (2018) 'A DNA structural alphabet provides new insight into DNA flexibility', *Acta Crystallographica Section D: Structural Biology*, 74(Pt 1), pp. 52–64. doi: 10.1107/S2059798318000050.

Schneider, B., Morávek, Z. and Berman, H. M. (2004) 'RNA conformational classes', *Nucleic Acids Research*, 32(5), pp. 1666–1677. doi: 10.1093/nar/gkh333.

Shing, P. and Carter, M. (2011) 'DNA Structure: Alphabet Soup for the Cellular Soul', in *DNA Replication-Current Advances*, doi: 10.5772/18536.

Taylor, R. and Kennard, O. (1982) 'The molecular structures of nucleosides and nucleotides. Part 1. The influence of protonation on the geometries of nucleic acid constituents', *Journal of Molecular Structure*, 78(1–2), pp. 1–28. doi: 10.1016/0022-2860(82)85306-4.

Watson, J. D. and Crick, F. H. C. (1953) 'Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid', *Nature*. doi: 10.1038/171737a0.

Zamenhof, S., Brawerman, G. and Chargaff, E. (1952) 'On the desoxypentose nucleic acids from several microorganisms', *BBA - Biochimica et Biophysica Acta*. doi: 10.1016/0006-3002(52)90184-4.