Akademie věd České republiky

Disertace
k získání vědeckého titulu "doktor věd"
ve skupině věd molekulárně biologické a lékařské vědy

Structure and Dynamics of Nucleic Acids

Komise pro obhajoby doktorských disertací v oboru
biochemie, biofyzika a molekulární biologie

Jméno uchazeče: Bohdan Schneider

Pracoviště uchazeče: Biotechnologický ústav AV ČR, v.v.i.

Místo a datum: Praha 28. února 2010

Resumé

Disertace shrnuje autorovu práci ve strukturní molekulární biologii. Zahrnuje dvacet publikací, devatenáct z nich jsou originální práce, jedna je kapitola v knize; dalších třináct relevantních publikací je v disertaci citováno. Po krátkém úvodu do problematiky strukturní biologie nukleových kyselin jsou diskutovány dvě krystalové struktury oligonukleotidů, na jejichž řešení se autor podílel. V obou případech je ukázán význam solvatace pro lokální konformaci DNA i uspořádání dvoušroubovic v krystalu.

Těžiště disertace je v popisu strukturního chování nukleových kyselin a struktury jejich solvatačního obalu bioinformatickými přístupy, zejména systematickým porovnáním velkého množství struktur. Analýzou dostupných krystalografických dat se jak pro DNA, tak pro RNA podařilo identifikovat opakující se konformace nukleotidové páteře mezi dvěma bázemi, kroku („step“, analyzovaný motiv ribosa-ribosa. Tato jednotka je použita, protože chemická jednotka nukleových kyselin, nukleotid, je pro konformační analýzu nevhodná). Porovnání konformačních rodin DNA a RNA ukázalo, že přes podobné chování na úrovni jednotlivých torzních úhlů jsou konformační prostory nukleotidů v DNA a RNA velmi různé. Nalezené lokální konformační motivy DNA i RNA jsou diskutovány v kontextu známých architektur těchto molekul.

Vodné prostředí, v němž se nukleové kyseliny vyskytují, aktivita vody i chemická identita a koncentrace rozpuštěných species, přímo ovlivňuje chování nukleových kyselin, jejich konformace, stability a tím i jejich biologickou aktivitu. Rozhodující je přitom první solvatační vrstva v bezprost-ředním okolí molekuly nukleové kysliny, která má fyzikální vlastnosti odlišné od solventu v objemu. Autorovi se podrobnou analýzou poloh molekul vody v krystalových strukturách DNA podařilo ukázat, že tato první hydratační vrstva kolem bází a fosfátové skupiny je v DNA do značné míry uspořádaná a že stereochemie preferovaných poloh molekul vody závisí na typu báze (nukleotidu) i formě DNA. Tento fakt je důležitý, protože potvrzuje měření o jejích odlišnosti od solventu v objemu a má důsledky pro chování izolovaných DNA i při jejich interakcích s jinými molekulami.

Autor pro zpracování dat o polohách molekul vody navrhl a opakovaně použil metodu „Fourierovské průměrování“, která se dá charakterizovat jako postup inverzní protokolu používaném v krystalografii. Metoda je bezparametrická, založená výhradně na experimentálních datech z analyzovaných krystalových struktur. Nepřehledné a komplikované prostorové distribuce experimentálních bodů (poloh molekul vody) interpretuje jejich transformací do hustot, které je možné dále zjednodušit do nemnoha poloh o nejvyšší hustotě. Fourierovské průměrování bylo nejdříve použito pro zjištění preferovaných solvatačních míst, později byla metoda generalizována na interpretaci konformačního chování RNA a DNA v mnohorozměrném torzním prostoru.

Vývoj metod pro řešení struktur nukleových kyselin a databázových systémů, infrastuktura, je důležitou podmínkou rozvoje strukturní biologie, v které se zpracovávají velké objemy dat. Disertace krátce zmiňuje projekty budování obou typů infrastruktury. Sestavení slovníků struk-turních parametrů pro nukleotidy, automatizace fitování molekulárních modelů do elektronových hustot a strukturní interpretace NMR parametrů patří do první kategorie. Základním kamenem infrastruktury strukturní molekulární biologie jsou ovšem databáze, jejichž úkolem je shromažďovat, archivovat, validovat, hledat, a distribuovat deponované struktury, eventuelně vyvíjet metody pro jejich validaci a porovnání. Autor se podílel na budování dvou významných a široce používaných databází prostorových struktur

molekul: Nucleic Acid Database, NDB, a Protein Data Bank, PDB, jejichž historie a současný stav jsou velmi stručně popsány.

Závěrem je zdůrazněno, že přes znalost desetitisíců experimentálně stanovených struktur biomolekul a bouřlivý rozvoj bioinformatických a výpočetně biologických metod, stojí před strukturní biologií rozsáhlé úkoly, které budou mít významné dopady pro základní vědu i pro aplikace v medicíně i biotechnologiích. Splnění těchto cílů bude záviset na rozvoji nových fyzikálních teorií, molekulárně biologických znalostí a technik, ale hlavně těsné spolupráci biologů a fyziků.

Table of Contents

Motivation

to get involved in scientific work is not always a straightforward process and I am certainly not an exception. When I was deciding where to turn for my postdoctoral during happy days of the year 1989, one thing was clear, I would study structures of macromolecules, of biological macromolecules that is. Proteins would have been an obvious choice for multiple roles they play in the living organisms and, let's admit it, for existence of multitude of good protein crystallography laboratories. But then there was this other molecule, DNA. Even a complete novice in structural molecular biology recognized stark contrast between deceiving simplicity of the double helix and the strict control it imposes on all living creatures; a closer look at the esthetically charming molecule evokes immediate interest how simple principles of its composition ensure verbatim self-recognition and translation into its kin molecule, RNA. All this fascination led me to the decision to devote my limited capacity for scientific work to study of the great molecule of DNA.

1. Introduction

Nucleic acids carry out fundamental roles in all living organisms and research of their functions in the second half of the twentieth century has been one of the key impulses for development of biology. DNA double helix became an icon of molecular biology that symbolizes the paradigm of modern biology –*DNA makes RNA, RNA makes protein*. Determination of structures of nucleic acids has significantly contributed to our understanding of biological processes and especially structures of complexes between proteins and nucleic acids at submolecular and atomic resolution unveiled life as intricate web of molecular processes. Many molecular details of key cellular processes of transcription, splicing, and translation and various levels of their regulation would not have been known without solution of structures of double helical DNA complexed with regulatory transcription factors, DNA and RNA processing enzymes, ribozymes, and ribosomes.

Rapid development of structure determination of nucleic acids –and for that matter of all biomolecules– was enabled by advances in methods of structure determination, namely x-ray crystallography, spectroscopic techniques of nuclear magnetic resonance, and electron microscopy. An important part of this development has been growing availability of sophisticated and expensive hardware for structure determination as synchrotrons as well as progress in methods of structure determination incorporated in ever improving computer software and these methodological advances further greatly benefited from revolution in tools of molecular biology and their wide accessibility. Synergy of advances in all these fields then have led to explosion in number and complexity of experimentally determined structures; when PDB catalogued less than 400 structures at the end of year 1989, it contained almost eleven thousand by the end of 1999, and over 62 thousand by the end of year 2009. At the same time, complexity of the released structures also increases as can be evidenced by numbers of residues, amino acids or nucleotides, in these structures. It is important to note that quantitative growth of available structures has been accompanied by their higher accuracy and precision.

Easier, faster and yet more accurate structure determination has been achieved by development of various software tools that integrate supporting data as force filed parameters, dictionaries of geometric parameters, libraries of rotamers, validation protocols, and the like. Evolution of these tools has been an integral part of the overall progress of structural biology especially in the last twenty years.

The growing number of solved structures called for their organization into electronic depositories, databases. Structural databases were very early on inspired by order-loving crystallographers as archives of solved structures but they became ubiquitous tools of structural bioinformatics and computational biology that spur further research in the related fields. Their construction has represented serious challenges for software development, formal data representations, data standardization, and required formulation of data formats for archiving and data exchange.

1.1. Content, scope, and organization of the thesis

The thesis deals with a few aspects of structural biology of nucleic acids: Their local conformational behavior, building software tools for structure determination, validation, and exploration including databases of molecular structures with the goal to summarize my limited contribution to the vast research of nucleic acid structure; the thesis by no means aspires to substitute for a balanced introduction to the issues of nucleic acid structure or even offer an overview of the state of our knowledge about the nucleic acid structure. The Nucleic Acid Database [Berman et al., 1992] archives

over four thousand three-dimensional models of structures containing nucleic acids determined by the x-ray crystallography and NMR spectroscopy and this vast amount of information is recapitulated in several good reviews including works by Neidle [1], Calladine [2], and an older but still useful comprehensive survey by Saenger [3]; also an excellent book by Branden and Tooze [4] has an informative chapter about protein-DNA complexes. Fundamentals of nucleic acid structure can be found in three reviews I co-authored [Schneider & Berman, 2006], [Neidle et al., 2009], [Neidle et al., 2003].

To help to evaluate my contribution to the field the papers I co-authored are cited differently from – obviously far more numerous and important– "other" works: My papers are cited by names of the authors and year of publication, as in [Neidle at al., 2009] above, and listed under the heading "*Papers co-authored by the applicant*" on the top of references. The remaining papers are cited by numbers in order of their appearance in the text and listed under heading "*References cited in the text*". Papers I co-authored, are discussed in greater detail in the text and compiled in the thesis are color-highlighted in the text as [Schneider & Berman, 2006] and listed under the heading "*Publications compiled in the thesis"* on the bottom of the reference section.

2. *Structure of nucleic acids*

Nucleic acids were known as components of all cells and especially of Eukaryotic nuclei but their function was unclear till the year 1944 when Avery, MacLeod, and McCarthy suggested that these macromolecules were responsible for heredity [5]. Soon after this fundamental discovery structural studies showed helical nature of fibrous samples of biological DNA [6,7], observation that led to the Watson-Crick model of right-handed, antiparallel double helical form [8], B-DNA. Atomic model of the DNA molecule as vehicle of heredity changed the paradigm of biology and led to the origin of a new science, molecular biology.

Right-handed double helix is not the only form DNA molecule can adopt, other forms as triple helices, quadruplexes, or left handed duplex are known and some have important biological functions [9]. Even when function of a close kin of DNA, RNA molecule, seemed much more limited, its structural features were explored early in 1970s, first in form of short dinucleotides [10,11] but these small structures were very soon followed by structures of molecule of utmost biological significance, transfer RNA [12-14]. After these breakthrough structures came a period of lower interest in RNA structure and, indeed lower inflow of new RNA structures. Interest in RNA structure renewed in mid-1990s after the RNA molecule, for a long time considered a passive information messenger and a mechanical scaffold of protein synthesis, has been assigned multitude of new functions in gene expression and its control, shown to have catalytic functions, and even function as steering the protein transport across membranes [15]. Breakthrough discovery of enzymatic ability of RNA [16] was soon followed by the first structure of hammerhead ribozyme [17] and more crystal and NMR structures of RNA began to emerge often with exciting and unexpected new features. Surprising was discovery of the role RNA plays in regulation of gene expression by so called RNA interference (RNAi) [18,19]. Forming of short double helical RNA segments emphasized the importance of double stranded RNA [20], so called A-RNA form, once considered of marginal importance and by-product of crystallization. Hand in hand with deeper understanding of various RNA functions has come the need to determine structures of the corresponding RNA molecules. Biochemical and other obstacles related to their handling have been, at least for some types of RNAs, apparently

overcome as evidenced by influx of new NMR, cryo-electron microscopy, and especially crystal structures in the late 1990's, and especially in the new century. Crystallographic structural study led to a surprising discovery that protein synthesis in ribosome particles is performed by ribosomal RNA, that ribosome is a ribozyme [21].

2.1. Nucleic acid building blocks

Concise information about chemical composition and nomenclature of nucleic acid components, (deoxy)ribose chirality, differences between DNA and RNA molecules, phosphate backbone and nitrogenous bases, their basic structural features including sugar ring pucker modes, and descriptors used to define base pairing can be found in reviews [Neidle et al., 2009] and [Schneider & Berman, 2006], the very basics of nomenclature describing the chemical unit of nucleic acids, nucleotide, are summarized in **Figure 1**.

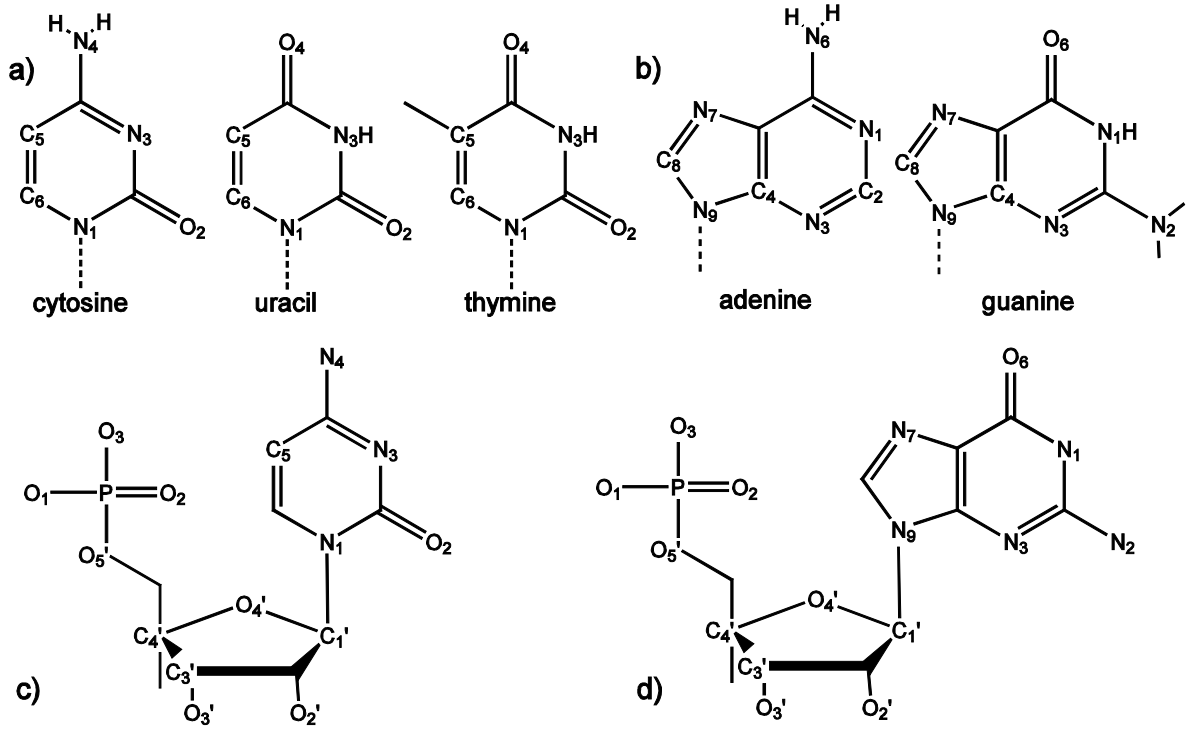

Figure 1. *Nomenclature of the nucleic acid bases and nucleotides. a) Pyrimidine bases cytosine, uracil (only in RNA), and thymine (only in DNA). b) Purine bases adenine and guanine. c) and d) Natural nucleotides cytidine-5'-phosphate (c) and guanosine-5'-phosphate (d).*

2.2. Structural alphabet of nucleic acid moieties

Nucleic acids are chemically complex molecules consisting of three building blocks of quite different chemical nature. Different physico-chemical and structural properties of charged phosphate group (that include phosphorous with available *d*-orbitals), chiral β-ribose, and aromatic nitrogenous bases are reflected by differences in their behavior during self-association of the nucleic acid molecule and during interaction with other molecules. Importance of DNA self-recognition *via* forming Watson-Crick pairs and the resulting coding ability of DNA has naturally led to a great attention to description of base pair topologies and geometries. Systematic description of base pairing topologies has been developed by Leontis and Westhof, [22]; the classification schema is summarized in **Table I** and further illustrated in **Figure 2**. Non-Watson-Crick base pairs represent an important construction element of folded RNA molecules, examples of two important classes are shown in

Figure 2. Frequently used description of these base pairs as "mismatches" is perhaps justified in DNA where they compromise the coded genetic message. Base pairs are formed by virtually planar bases, which can lie in one plane or in deformed folded, or propeller-like arrangement.

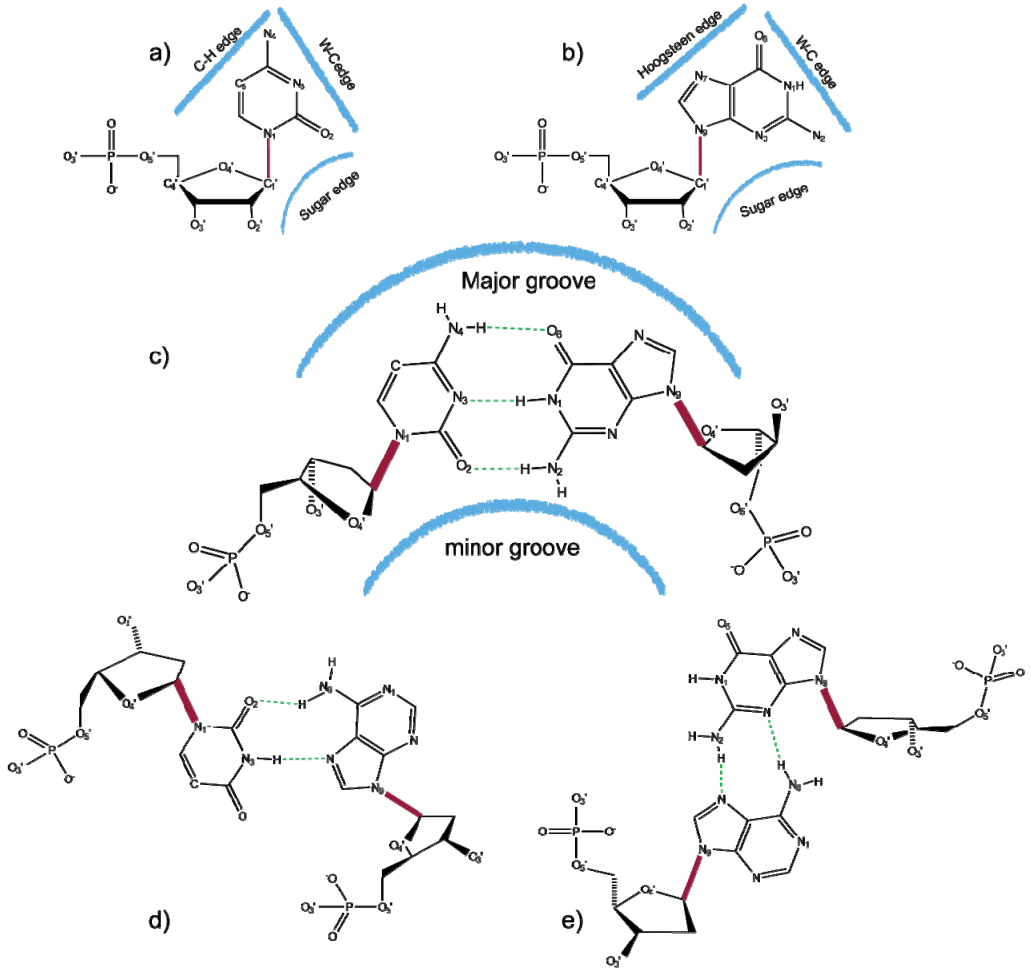

Figure 2. *Base pairing topology. a) Definition of base "edges" according to Leontis-Westhof [22] nomenclature for pyrimidines, and b) for purines. c) Watson-Crick base pairing (L-W pair of type 1, see* **Table I***), shown as the G-C pair. d) Non-canonical base-pair interacting by Watson-Crick and Hoogsteen edges with trans orientation of glycosidic bonds, L-W class 4. e) L-W class 10 formed by the Hoogsteen and Sugar edges with trans orientation of glycosidic bonds. L-W classes 4 (d) and 10 (e) are important for RNA architecture. Glycosidic bonds are dark red, hydrogen bonds between bases green.*

Local geometry of the Watson-Crick pairs and their steps is quite important for formal geometric description of DNA molecule and its deformability as well as for understanding of possible sequence dependencies of DNA recognition by other molecules. Base pair and base step geometries have therefore been described by various geometric parameters as helical twist, propeller twist, buckle, opening, etc.; their definitions can be found e.g. in [Neidle et al., 2009]. Historically, these parameters were calculated by different algorithms that used various reference frames (coordinate systems) with unfortunate consequence that parameter values calculated by various programs could not be easily compared. To rectify this situation, the common reference has been defined at a meeting in Tsukuba, Japan [23]. This "Tsukuba standard reference frame" is currently used by most pro-

grams to calculate nucleic acid geometries (as 3DNA [24]) and the corresponding values of base and base-pair geometries are archived in the Nucleic Acid Database [Berman et al., 1992].

Table I. *Leontis-Westof (L-W) classification of base pairing topologies [22]. See also* **Figure 2**.

L-W type	Orientation at the glycosidic bond	Interacting edges	Local strand directions
1	cis	Watson–Crick/Watson–Crick	Antiparallel
2	trans	Watson–Crick/Watson–Crick	Parallel
3	cis	Watson–Crick/Hoogsteen	Parallel
4	trans	Watson–Crick/Hoogsteen	Antiparallel
5	cis	Watson–Crick/Sugar Edge	Antiparallel
6	trans	Watson–Crick/Sugar Edge	Parallel
7	cis	Hoogsteen/Hoogsteen	Antiparallel
8	trans	Hoogsteen/Hoogsteen	Parallel
9	cis	Hoogsteen/Sugar Edge	Parallel
10	trans	Hoogsteen/Sugar Edge	Antiparallel
11	cis	Sugar Edge/Sugar Edge	Antiparallel
12	trans	Sugar Edge/Sugar Edge	Parallel

The conformational behavior of the sugar-phosphate backbone of nucleic acids is complicated. It can be described at the atomic level by six backbone torsion angles labeled α, β, γ, δ, ε, and ζ, five endocyclic torsion angles of the (deoxy)ribose ring, and one torsion around the glycosidic bond describing the rotation of the base plane relative to the sugar, torsion χ. Specific behavior of the backbone and its geometric descriptors in DNA and RNA molecules will be in a greater detail discussed in further paragraphs, a few following sentences and **Figure 3** summarize their definitions and very basics of their behavior.

Conformationally the most complicated is the (deoxy)ribose sugar ring which is intrinsically non-planar, "puckered" *via* pseudorotation around its five single bonds. Bioinformatic studies [25,26], [Gelbin et al., 1996] based on nucleotide geometries from CSD [27] and NDB [Berman et al., 1992] have confirmed theoretically predicted sinusoidal relationships between the five ribose endocyclic torsion angles $\nu_{0\text{-}4}$ and the pseudorotation angle P:

$$tgP = [(\nu_4 + \nu_1) - (\nu_3 + \nu_0)]/2\nu_2[\sin 36° + \sin 72°]$$

The pseudorotation angle P correlates quite tightly with the backbone torsion δ so that for practical purposes the set of six backbone torsions and χ is sufficient for complete conformational description of the sugar-phosphate backbone including sugar pucker. Torsion χ around the glycosidic bond has two populated ranges, the common one dubbed *anti* orientation, the rare one *syn*.

2.3. DNA structures

The prevailing form of DNA molecules is antiparallel double helix. The Watson and Crick [8] model of B-form DNA observed at physiological values of humidity and ionic strength was supported by interpretation of fiber diffraction experiments during the sixties and especially the seventies that are reviewed in [28] but the right-handed double helix had not been experimentally confirmed until a single crystal structure of so called Dickerson-Drew dodecamer was solved in 1981 [29]. This very first single crystal structure of the B-form DNA showed its most typical features: W-C pairs on aver-

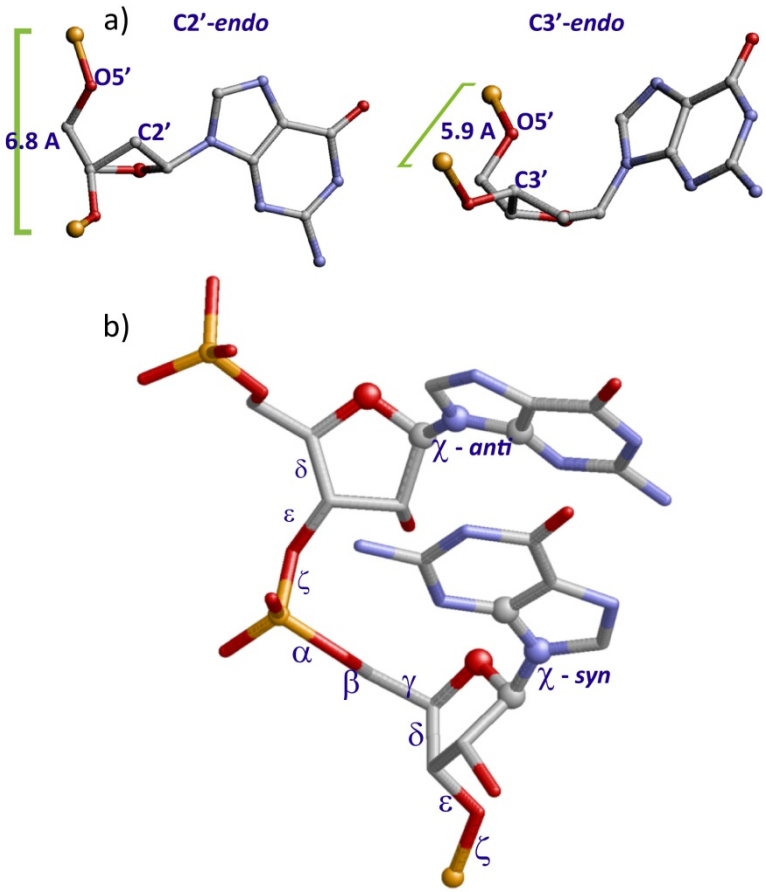

Figure 3. *Nucleotide stereochemistry shown for deoxyguanosine-5'-phosphate. Atoms are drawn in their chemical colors, phosphorus in yellow, oxygen in red, nitrogen in blue.*
a) C2'- and C3'-endo pucker of the sugar ring. With O5' on the top and O4' in the back of the page, the C2'-endo pucker has the ribose atom C2' "above" the plane formed by the remaining four ribose ring atoms, the C3'-endo pucker has the atom C3' above. C2'-endo is typical for the B-form, C3'-endo for the A-form.
b) Orientation of the nucleobase (here guanine) relative to the ribose ring is described as "anti" or "syn". These orientations are defined by torsion angle χ around the glycosidic bond. A large majority of bases is in the anti orientation which has the major groove atoms (for guanine C8, N7, O6) pointing to the same direction as the nucleotide phosphate; for the syn orientation, the major groove points away from the phosphate. Atoms defining torsion χ are shown as small spheres. Conformation of the sugar-phosphate backbone is defined by six torsion angles: α, β, γ, δ, ε, and ζ.

age perpendicular to the helical axis, full helix turn completed after ten base pairs, i.e. the average helical twist per nucleotide step 36°, and the pitch between the stacked base pairs of 3.4 Å. Charged phosphate groups are exposed to the solvent and divide the double helix into two non-equivalent grooves, minor and major. The bottoms of the grooves are formed by hydrophilic and hydrophobic base atoms, their walls are lined mostly by hydrophobic deoxyribose atoms. Both grooves have a similar depth in B-DNA but the major groove is wider and exposes chemically more diverse set of base atoms than narrower and chemically more uniform minor groove. The fact that hydrogen-bonding atoms are more diverse in the major groove than in the minor one and that they form a much more sequence dependent patterns has consequences for ligand binding to DNA, including ways proteins recognize DNA.

The other right handed DNA form, A-form, can be considered deformed B-form (see discussion in paragraph 2.3.2). Change of sugar pucker from C2'-*endo*, typical for the B-form, to C3'-*endo* brings the subsequent phosphates closer together than in the B-form (**Figure 3**), induces a larger tilt of base pairs relative to the helical axis, shifts the center of base pairs away the helical axis by almost 5 Å, lowers value of glycosidic torsion angle χ from 270° to 200°, and lowers rise between base steps to only ~2.6 Å. The result is a wider helix with different groove geometries than in the B-form, deep and narrow major groove and shallow, wide-open minor groove; **Figure 4** illustrates the main features of both B- and A-forms. One important difference between both forms is their deformability. While the B-form is known to be highly deformable by forming abrupt kinks or smooth bends sometimes exceeding 90° –many quite striking examples can be found especially in complexes with pro-

teins– double helical A-DNA and especially A-RNA with all bases in Watson-Crick pairs are known to be straight within a small degree of bending up to 15°.

Despite some previous doubts about biological role of A-DNA, it is now well established that this DNA form is locally induced by interactions with proteins [30,31], [Svozil et al., 2008] where it serves as a local recognition motif. The A-form is however most important in RNA; it is the prevailing conformation of all RNA molecules, forms the scaffold of large folded RNA molecules, and compose the main part of double helices in small interfering RNAs.

The topic of this thesis is not discussion of base pairing patterns from structural or energetic point of view. Let us however stress one important and often underappreciated aspect linking the prevailing, and arguably the most important, pattern of base pairs, the Watson-Crick pairing, and geometry of the sugar-phosphate backbone. Right-handed, antiparallel double helices with bases paired in the Watson-Crick arrangement are the most prevalent forms in both DNA and RNA because the helical arrangement allows the nucleotides to adopt optimal or near-optimal conformations of their sugar-phosphate backbone and at the same time bases to form Watson-Crick pairs. The importance of the Watson-Crick pairing for preserving the genetic information is further enhanced by the fact that both G-C and A-T (or A-U in RNA) pairs in the Watson-Crick arrangement have very similar distances between the points of attachment to the backbone, i.e. the distance between (deoxy)ribose atoms C1' of the paired bases is in both pairs similar: This allows the backbone to run smoothly regardless of sequence.

Figure 4. *The two most important structures of nucleic acids: A-form (green, left) and B-form (red, right). Shown are idealized conformations derived from the fiber diffraction data; sugars and bases are shown as rods, the backbone as ribbon.*

Right-handed double helices are not the only DNA forms, in fact the first single crystal of DNA showed quite unexpectedly a left-handed duplex with architecture distinct from both right-handed forms [32]. This DNA duplex form, called Z-DNA, is formed from two antiparallel DNA strands, but the backbone runs in an irregular zigzag pattern and the repeating unit is a dinucleotide, not a nucleotide as in both right-handed forms. Z-DNA is thermodynamically less stable than A- and especially B-form and strongly prefers alternating pyrimidine-purine sequences; in fact, most known crystal structures are formed mostly by CG units, the TA step is destabilizing, as is any purine-purine or pyrimidine-pyrimidine step. Z-DNA is known in two related forms, ZI and ZII that differ in torsion angle values, the most descriptive is value of torsion ζ at the purine step: ~300° in ZI, ~50° in ZII.

Biological role of Z-DNA, if any, is not clear but it may play a role in transcription and translation regulation [33]; a tempting hypothesis is that regions of "short tandem repeats" of sequence CG known in some genomes can be converted to the structurally different Z-form and serve then as structural markers in the genome.

A distinct DNA form, quadruplexes, is thought to be an important structural element of telomeres, sequences ending eukaryotic chromosomes and essential for their replication. Guanine-rich telomeric sequences have been structurally most investigated; both NMR [34] and crystallography [35] showed that oligonucleotides containing repeat TTTTGGGG form quadruplexes with guanines hydrogen bonded into planar quartets with alternating *anti* and *syn* orientation at the glycosidic bond [36,37] (e.g. structure of NDB code UD0013 [37]). Interestingly, sequences containing tetrads of cytosines that sequentially correspond to the G-tetrads also form tetraplex arrangement, albeit of topology and 3D structure very different from the G-tetraplexes. The cytosine tetraplex, so-called i-motif [38], has two pairs of parallel opposite strands that are linked by paired cytosines (e.g. structure of NDB code UDD024). Each duplex is stabilized by hemi-protonated C-C^+ base pairing between the parallel strands, and a string of water molecules bridging the cytosine N4 to phosphate oxygen atoms.

Two DNA double helices may combine into four-way branched topology, so called four-way junctions, or Holiday junctions, during recombination of chromosomes, e.g. during site-specific recombination assisted by recombinases. Four-way junctions have been crystallized in several different topologies as pure DNA [39,40] as well as from solutions of DNA complexed with proteins, e.g. with its specific topoisomerase Cre recombinase [41]. All these structures share almost planar arrangement of the four DNA double strands with all bases paired in the Watson-Crick pattern (or close to it); links between double helices are formed by single nucleotides with only a few of their backbone bonds rotated from the typical values.

2.3.1. Crystal structures of DNA oligonucleotides

A large majority of structural information at submolecular or atomic resolution originates from analysis of diffraction of x-rays on single crystals ("x-ray crystallography"). X-ray crystallography or discussion of particular crystal structures is not in the focus of the thesis so that here we only briefly mention two structures co-determined by the applicant, complex between dinucleotide dCG and drug proflavine [Schneider et al., 1992a] and hexanucleotide in the Z-form [Harper et al., 1998].

The first structure is a complex between dCG and intercalating drug proflavine, diaminoacridine antibacterial drug with mutagenic effects [Schneider et al., 1992a]. The structure is important because it was measured and independently refined using high-resolution crystallographic data below 1 Å taken at three temperatures, -130 °C, -4 °C, and room temperature [42]. Dinucleotide conformation and its interaction with proflavine are in all three structures similar but both low temperature structures show disorder of the nucleotide backbone, feature typical also for other nucleic acids: Higher resolution and lower temperature enables to determine crystallographic disorder (alternate positions of molecular fragments) that is smeared off at higher temperature or when fewer experimental data are available for the refinement of the molecular model. Quite typical is oscillation of the phosphate group between two positions that belong to the BI and BII backbone conformations. Here we observed disorder of one of the deoxyribose rings between C2'- and C3'-*endo* puckers and oscillations of atoms defining the nearby torsions γ and β. Both these low temperature conformational changes lead to tightening of the cavity between two base pairs available to the intercalating proflavine.

The dCG/proflavine is however an archetypal structure for other reason than disorder of its backbone, it is its extensive hydration network. Already the room temperature (RT) structure showed

well ordered network of hydrogen-bonded water molecules but the network further extended in both low temperature structures so that the cryo structure determined at -130 °C had all the gaps seen in the RT structure filled with water molecules. Water networks found in all three structures contained several pentagons of hydrogen bonded water molecules. Pentagonal arrangement of waters was at that time considered important recurrent motif of the solvation shell of biomolecules. It should however be noted that the water pentagons observed here and in other structures [43] have not proven to be the feature able to explain hydration patterns around biomolecules that would unify our view of the DNA hydration.

The other briefly discussed structure is a DNA hexamer $d(TGCGCA)_2$ that crystallized in the Z-form [Harper et al., 1998]. Z-DNA structures may be considered of lower biological relevance but they have been historically very important because they were actually first monocrystal structures of oligodeoxynucleotides [32] and form crystals diffracting to a higher resolution than most crystallized B- and A-DNA sequences. Well resolved Z-DNA structures allowed analysis of Watson-Crick base pairing, backbone geometries, and solvation effects. Most known Z-DNA structures are hexamers of alternating pyrimidine-purine sequences with a majority of CG steps, a few structures exhibit purine-purine (and pyrimidine-pyrimidine) steps, not many structures contain thymine and adenine bases, and only one dodecamer, one decamer, and one octamer have been crystallized in the Z-form. $d(TGCGCA)_2$ was determined at reasonable crystallographic resolution of 1.3 Å, typical for Z-DNA and refined to validation standards observed usually in small-molecule crystallography. The structure shows most features typical for the Z-form: Watson-Crick base pairing with *syn* orientation of purines and *anti* of pyrimidines, alternating ZI and ZII backbone conformation (detailed description of these conformers is in [Svozil et al., 2008]), locally stabilized by interactions with metal cations (here complex $[Co(NH_3)_6]^{+++}$), and the first hydration shell formed by well resolved water molecules with the minor groove "spine of hydration" [44,45].

The paper [Harper et al., 1998] compared $d(TGCGCA)_2$ to all then known Z-DNA hexamers to analyze their hydration patterns and elicit their possible packing mechanisms. Z-DNA crystallized in the $P2_12_12_1$ space group are packed in the crystal lattice in two ways sometimes called magnesium form [45], "M", and spermine form [46], "S"; packing M is stabilized by polyvalent metal cations, typically Mg^{++}, S is stabilized by polyamines as spermine. The lack or presence of metal ions is not however exclusive for either the S or M arrangement and the "pure spermine form" can be formed also in presence of Mg^{++} or Co^{+++}. We tested also another hypothesis about possible packing mechanism of Z-DNA duplexes put forward in the paper [Schneider et al., 1992]. It had suggested that Z-DNA hexamers were stabilized by interactions between atoms forming the specific ZI/ZII backbone alteration and hydrogen bonded network of waters but even this hypothesis could not be confirmed in all cases despite its statistical validity. Therefore, to our dismay, we had to concede that seemingly simple mechanism of packing of Z-DNA hexamers was still elusive and that mechanistic and causal explanation of the particular packing mode is beyond the scope of the available structural data and our understanding of molecular interactions.

2.3.2. Conformational dynamics of DNA nucleotides

The sugar-phosphate backbone of nucleic acids is conformationally very complex, it consists of seven rotatable torsion angles and sugar rings have five bonds that can undergo pseudorotation. There have been attempts to develop rules that would limit a large number of possible combinations of conformers resulting from such a high number of degrees of freedom [47,48], and also early analyses of crystal structures yielded important information about restrictions of the nucleic acid confor-

mation space –the concept of so called "rigid nucleotide" [49]. Significant advances to our understanding of the nucleic acid conformations occurred after the first crystal structures of RNA [10,11] and DNA [32] began to be determined in the 1970s, and then especially after the first B- [29] and A-DNA [50] monocrystal structures were refined.

Number of available crystal structures of DNA in mid-nineties allowed a simple statistical analysis of local conformations of double helical DNA and we characterized nucleotide and dinucleotide conformations in crystal structures of "naked" DNA, i.e. DNA crystallized only with solvent species [Schneider et al., 1997]. This study of the DNA backbone geometry offered reliable averages of the torsion angles for the main double helical forms, BI, BII, A, and the left-handed ZI and ZII [32] that are being widely used. The study analyzed geometric parameters (torsion and also valence geometries) at different values of crystallographic resolution, better than 2.7, 2.4, 2.0 and 1.9 Å, and while not all the results could be shown in the paper they led to –I believe important– general conclusion that the quality of crystallographic data improves significantly for resolutions better than 1.9 Å and that this limiting value should be used for selection of structures for statistical and/or bioinformatic analysis of crystal structures. Albeit this limit is fully empirical, our later studies confirmed its validity. It should be stressed that the "rule of 1.9 Å" is valid only for statistical analysis of many structures and does not give evidence about quality of any individual structure that may be highly reliable at 2.4 Å (as deservedly famous Dickerson-Drew dodecamer [29]) or unreliable at 1.5 Å.

The study [Schneider et al., 1997] corrected erroneous torsion values reported for the "canonical" B-form derived from the fiber data [28]. Scarce diffraction data and the resulting unavoidable averaging of geometry parameters in fiber studies led to smearing off of the differences between the BI and BII forms in the studied DNA polymers and the crystallographic refinement led to incorrect values of torsion angles; especially values of ε, ζ, and β in the fiber model (220°, 200°, and 136°, respectively) do not represent any populated torsion region in any B-DNA form (Table II in [Schneider et al., 1997] or Tables 3 and 4 in [Svozil et al., 2008]).

This conclusion was confirmed by our later extensive compilation of a large "across-the-archive" sample of DNA crystal structures [Svozil et al., 2008]. This study was conducted by finer analytical tools and extended by rigorous tests of statistical significance of the results. The fundamental improvement from our original analysis was the way how we identified DNA conformers: While the earlier study assumed *a priori* existence of the basic double helical forms as BI, BII, or A and determined the average values of their backbone torsions, the study [Svozil et al., 2008] analyzed the available data and independently determined identified conformers. The analysis was based on "Fourier averaging" of 3D data (Fourier averaging is described in the following paragraph, 2.3.2.1) of almost eight thousand nucleotide units from 187 crystal structures of naked (non-complexed) DNA and 260 protein/DNA complexes; all the main DNA conformers, namely BI, BII, AI, AII, and the Z forms, were identified validating thus the analytical procedure and providing highly accurate values of torsions for these conformers (Table 3 in [Svozil et al., 2008]). The analyzed unit was slightly smaller than a dinucleotide, between O5' of the first nucleotide to the O3' of the following one; it contains ten torsion angles, γ, δ, ε, ζ, α+1, β+1, γ+1, χ, χ+1. Some conclusions from the study are summarized below.

Far the most populated conformer in all DNA structures is the BI-form (α 300°, β and ε near 180°, γ 50°, δ 130°, ζ 260°), the related BII is indeed a distinct form with ζ in the *trans* region, high ε (250°) and low β (140°) values. The gap between the BI and BII forms almost disappears in protein/DNA

complexes and a series of intermediate conformers indicates almost continuous transformation that is best described by correlation between torsions ε and ζ: ε = -0.73 ζ + 367°. The A-form has also two sub-classes, the "canonical" AI, and minor AII, characterized by α and γ values close to 180° (as opposed to values of 300° and 60°, resp., in AI). Both these A forms exist also in RNA with virtually identical torsion values. Some conformers can be characterized neither as B nor A. Observed were conformers with mixed B and A features, mainly sugar pucker and value of torsion χ. These conformers have one part of the analyzed unit ("suite", see [Richardson et al., 2008]) in the B-, the other in the A-form; observed were combinations BI/AI and BII/AI. Besides these mixed conformers, we also identified conformers with unusual sugar ring puckers O4'-*endo* and C1'-*exo* that are pseudorotation intermediates between C2'- and C3'-*endo* major puckers (see e.g. treatise by Saenger, [3]). These conformers can be considered true A-to-B transitional geometries and because they exist in both complexed and high-resolution naked DNA, they are most likely thermodynamically stable even when high-energy DNA conformers.

All major conformers of both B- and A-forms and also of some mixed conformers have usually several satellite clusters in close conformational proximity that differ by values of some torsion angles; these torsional differences however tend to compensate each other so that the parent and satellite conformations are similar. The existence of structurally close conformers is a typical feature of DNA that sets it apart from RNA and, in my opinion, the existence of these near-lying energy minima in the conformational space of DNA is responsible for the ability of DNA to respond to the external force by minor deformation rather than large conformational switch.

"Naked" DNA, DNA crystallized only with solvent, water and small ions, is conformationally most compact, and its torsion distributions significantly broaden upon complexation with ligands, drugs and especially proteins. This principle is best exemplified by the existence of the sharply divided BI and BII forms in the naked DNA that become connected by a continuous cloud of conformers in complexes. Besides BI/BII intermediates induce also A-like conformational features in otherwise B-like double helices, as evidenced by the fact that several B/A mixed conformers indeed exist only in protein/DNA complexes.

The naked DNA structures were studied in a greater detail from the point of view of possible relationships between sequence preferences of the assigned conformers (Table 5 and 9 in [Svozil et al., 2008]). The BI-from is numerically dominant in all dinucleotide sequences (steps) but it is statistically over-represented in homogeneous purine-purine (and pyrimidine-pyrimidine) steps. The BII form is often found at dinucleotides TG and CA; note that these steps can form Watson-Crick pairs so that BII is then likely to occur at the facing nucleotides of both strands. Overrepresented in BII is also the GG step. The alleged malleability of the CG step to adopt the BII-form cannot be confirmed. No pyrimidine-purine was classified as B/A conformer (the first sugar C2'-, the second C3'-*endo*), and very few purine-purine and purine-pyrimidine steps adopted the A/B (the first sugar C3'- and the second C2'-*endo*) confirming reluctance of purines to adopt the C3'-*endo* pucker in a B-like double helix. The GC step shows conformationally the most complicated and variable behavior of all steps: It prefers B/A (and to some extent BII) conformers, disfavors BI and A/B, but many GC steps cannot be classified into any cluster of conformers, which means that this step is likely to adopt unusual conformation. Regardless of sequence, two consecutive BII steps are very rare and need to be stabilized by external forces as crystal packing or interaction with a complexed species.

2.3.2.1. *The method of "Fourier averaging"*

The method of Fourier averaging was developed to simplify noisy three-dimensional (3D) distributions of solvent particles around DNA by localizing regions of their high density. It was for the first time used to find preferred hydration sites around DNA bases in double helical DNA [Schneider et al., 1993] and later successfully applied to analysis of base hydration [Schneider & Berman, 1995], [Woda et al., 1998], [Morávek et al., 2002], for interpretations of noisier distributions of solvent species around phosphates in DNA [Schneider et al. 1998] and in organic phosphates [Schneider et al., 1996], [Schneider & Kabeláč, 1998]. The technique was later generalized to interpret complicated and noisy multidimensional torsional distributions in RNA [Schneider et al., 2004] and in DNA [Svozil et al., 2008].

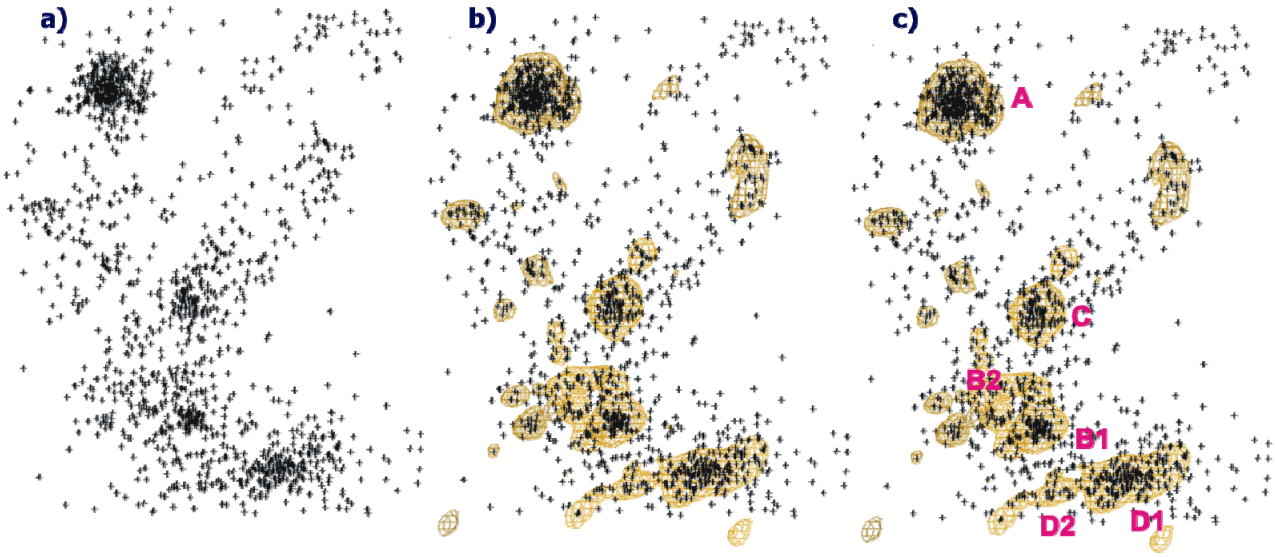

Figure 5. *Schematic description of the method of Fourier averaging. a) The original distribution of points (black crosses) in three-dimensional space. b) The distribution of points is transformed to its density. Density distribution is calculated by Fourier transform and the isodensity cages are drawn (yellow). c) Points of the highest density are identified (manually or computationally) and labeled. The peak centers may have immediate interpretation as preferred hydration sites or are further analyzed. In this case, peak centers are used to label all data points in their neighborhood and the assigned symbols are clustered as described in the text and summarized in* **Table II**.

The gist of the method is to turn distribution of points to their density. Densities are easier to interpret including localization of sites of the highest density of the distribution; the procedure is illustrated in **Figure 5**. The method handles three-dimensional distributions, either positions of atoms in Cartesian space or of three torsion angles in the abstract torsion space. In either case, the distribution of points is transformed into their pseudo-electron densities by Fourier transforms used in crystallography albeit here applied in the reverse order: The first Fourier transform is used to turn distribution of points to the reciprocal space and calculate Fourier coefficients called in crystallography structure factors, F(hkl). The second Fourier transform then converts F(hkl) into pseudo-electron densities that are further analyzed visually or computationally. Employing the procedure established in crystallography has a major advantage of having available a host of tools for mathematical and visual treatment of the data, importantly, the standard crystallographic programs allow visual inspection of distributions, and manual or automated fitting of positions of the highest density, i.e. points where the distribution has local minima. Application of crystallographic protocol brings also a necessity to adapt certain parameters that are technically necessary but have no obvious physical

meaning in the Fourier averaging: Space group, cell dimensions, crystallographic resolution, atom types, their occupancies, and temperature displacement factors. Of all these parameters, crystallographic resolution is the only one, which needs to be optimized in order to extract maximum of information from noisy data. Rationale for selecting all the parameters can be found in published papers, especially [Schneider et al., 1993], [Schneider & Berman, 1995], and [Schneider et al., 2004].

Fourier averaging of solvent positions is straightforward, each solvent atom is directly represented by an atom in the calculated FT, in full analogy with crystallography but its application on analysis of backbone conformations deserves a short discussion. Multidimensional torsional space is divided into 3D subspaces called "torsion maps". Each map consists of points made of the three torsions [τ_1, τ_2, τ_3], these points are equivalent to atoms in real space. The value of crystallographic resolution needs to be determined by trial and error and depends on the quality of the data. The higher the resolution, the higher the number of peaks; since each peak in a map corresponds to a conformer, the optimal resolution should produce maps with number of peaks, conformers, that is reasonable for interpretation. A range between eight to twelve peaks lead to the optimal resolution near 2.5 Å for DNA, for more noisy RNA distributions close to 3.0 Å.

Identification of peak positions in a 3D torsion map solves the problem of noise of the data and indicates frequent combinations of the three analyzed torsions but by itself does not classify the backbone segment with eight or ten torsions, in other words, the problem of multidimensionality of the torsion space is yet to be overcome. To cover all the possible combinations of the ten torsions analyzed in the DNA analysis [Svozil et al., 2008] would require analysis of 10!/7!3! = 120 3D torsion maps. Such an analysis would not be practical but it is not fortunately necessary either because preliminary knowledge about the conformational space and of its one- and two-dimensional torsion distributions allows radical reduction of the number of analyzed maps. In fact, just six combinations of torsions were analyzed in the RNA study [Schneider et al., 2004], nine were considered in the DNA study [Svozil et al., 2008] but only four analyzed in detail.

In any particular map, each of the ten peaks is assigned a two-letter name as indicated by red letters A, B1, B2, C, etc. in **Figure 5**. Peaks are then approximated by spheres of a typical radius between 15° and 40°, the size is estimated from the volume of a particular density contour. All the original data points [τ_1, τ_2, τ_3] lying inside the peak's sphere are labeled by peak's name, data points located near two or more peaks are assigned to the most intense one. The data points located outside the radii of all the peaks are not assigned to any peak. As nine maps were analyzed for the DNA analysis, each DNA dinucleotide in the analyzed dataset is characterized by a nine-letter string referred to as an *imprint*. The imprints are then used to find dinucleotides with similar conformations by simple alphabetical sorting. The clusters of dinucleotides with similar conformations –*conformers*– are

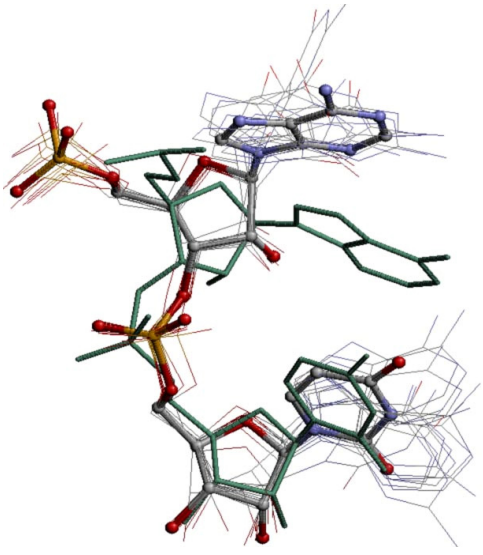

Figure 6. *Superposition of dinucleotides grouped into a possible conformer by Fourier averaging. One dinucleotide is a likely outlier; its first base has a different orientation from the remaining fragments. The averaged conformer's structure is shown in bold lines, individual nucleotides are in thin lines. The canonical A-RNA form is shown in dark green for reference.*

identified as a set of dinucleotides with (nearly) identical imprints. The process is exemplified by data in **Table II**.

The ultimate test as to whether the Fourier averaging indeed produced a family of similar conformations is evaluation of overlaps of contributing dinucleotides in the conventional Cartesian 3D space. The atomic coordinates of the dinucleotides are compared using least-square overlaps; overlap of RNA nucleotides forming a small cluster is shown in **Figure 6**. The resulting root mean square deviations are numerically compared and the overlapped structures visually inspected in order to check for possible outliers that should be removed from the cluster.

Table II. *Clustering in the Fourier averaging. Each line of the table represents one dinucleotide (step). ID is its unique identifier, NDB ID indicates the structure of its origin. Imprints show labels assigned to individual steps for the four torsion maps indicated by torsion names. Torsion angles are values of the particular torsions, and Cluster number is a label of the resulting conformer. Conformer number 25 belongs to the A-DNA family, conformers 99 and 101 are BII forms. Note A-form-like conformers in B-DNA double helices BD0014 [51] and BDL001 [29]. Note that torsion angles within a cluster may vary quite substantially, by more than 30°, so that simple classification by individual torsions is problematic.*

ID	NDB ID	Imprints for torsion maps				Torsion angles [°]									Cluster number
		ζα1γ1	ζα1δ	γζγ1	ζγ1δ1	δ	ε	ζ	α1	β1	γ1	δ1	χ	χ1	
7083	ADJ0113	A	A2	A	ZZ	82	216	271	282	176	65	82	196	197	25
7586	BD0014	A	A2	A	ZZ	105	167	292	293	187	43	105	241	235	25
333	BDL001	A	A2	A	ZZ	99	174	274	301	173	64	109	233	234	25
5014	UD0045	A	A2	A	ZZ	98	213	274	303	179	42	95	188	206	25
7395	BD0005	F	F	H	F	144	249	149	293	138	51	141	275	239	101
7437	BD0007	F	F	H	F	141	266	149	280	157	44	141	271	260	101
755	BD0029	F	F	H	F	143	252	149	294	137	50	141	277	240	101
947	BD0037	F	F	H	F	131	260	151	286	143	49	142	272	260	101
3622	UD0023	F	F	H	F	145	248	169	315	132	37	150	290	285	101
3637	UD0024	F	F	H	F	139	260	161	303	123	44	147	290	284	101
3641	UD0024	F	F	H	F	153	268	150	299	134	37	136	284	251	101
4986	UD0040	F	F	H	F	121	239	166	290	157	52	138	247	213	101
503	BD0017	F	F	ZZ	F	156	258	153	297	151	34	152	288	269	99
6654	BDJB27	F	F	ZZ	F	157	242	166	308	152	34	141	288	273	99
9	BDJB49	F	F	ZZ	F	155	236	177	301	156	40	149	277	269	99
6988	BDL084	F	F	ZZ	F	143	260	146	287	144	51	144	276	247	99

2.3.3. *Solvation of DNA*

DNA solvation, relative humidity, ionic strength, and chemical identity of cations are phenomena directly influencing the conformation and other structural properties of DNA molecules that directly bear on their biological functions [52,53]. The topic has been well reviewed from the thermodynamic point of view, for instance in [54-56]. The hydration shell retains its integrity at both high and low values of relative humidity [57,58] and effectively ties up water molecules to the nucleic acid surface. Various estimates place the number of these bound water molecules between five and twelve per nucleotide. At lower relative humidity, water does not diffuse freely and is located mostly around phosphate groups [59,60]. Stronger binding of water to phosphates than to bases is further confirmed by the fact that water is first removed from the grooves.

Structural features of DNA hydration determined by crystallography and NMR or predicted by theoretical approaches have been described in several reviews, [61-64], [Berman and Schneider, 1998], [65]. Of these approaches, crystallography offers the richest level of detail, especially about the first hydration shell around biomolecules [61], [Berman and Schneider, 1998], [65]; only few waters in the grooves of B-DNA have their relaxation times just below the limit for unequivocal localization [66] and hydration especially of phosphate groups is very dynamic and poorly detectable by NMR methods.

Two following sections show that systematic studies of hydration around bases and phosphates in crystal structures of double helical DNA led to conclusion that the first hydration shell of these DNA constituents is significantly ordered and determined primarily by the stereochemistry of hydrogen bond but modulated by the conformation of the nucleotide residue.

2.3.3.1. *Hydration of DNA bases*

Already our first overview of crystallographic water molecules around DNA double helices [Schneider et al., 1992] suggested that waters concentrate into well ordered sites around bases, the fact clearly demonstrated by our second study [Schneider et al., 1993], that for the first time applied Fourier averaging described in section 2.3.2.1. The analysis observed clear differences in distributions of hydration sites around bases in B-, A-, and Z-DNA duplexes. Not surprisingly, hydration patterns around the left-handed Z-form and both right-handed form are quite different, but there are distinctive differences between spatial distributions and arrangements of the hydration sites between the B- and A-DNA. This is best exemplified by the fact that the B-DNA hydration sites inserted into the A-DNA duplex produce clashes with DNA atoms and vice versa.

Hydration patterns of biologically most relevant B-form were subject of a more detailed analysis at a time when more available crystal structures allowed analysis of higher quality structures of DNA decamers [Schneider & Berman, 1995]. Hydration sites derived exclusively from decanucleotide structures were virtually identical to the previously determined sites [Schneider et al., 1993], that were predicted mostly from dodecamer structures. The "hydration building blocks" extracted from decamer structures were used to predict hydration around the emblematic Dickerson-Drew dodecamer [29] and the predicted hydration sites were in very good agreement with the actual crystal water positions. This seemingly technical point is important because it proves that even quite non-random sampling of crystal structures as offered by the dodecamer structures with similar crystal packing and poor sequence sampling provides robust results.

Using an analogous protocol as in the previous hydration studies we analyzed intermolecular contacts in protein/DNA and DNA/drug complexes [Morávek et al., 2002]. An extended compilation

of crystal structures allowed us to determine sites of preferred binding of polar, hydrophobic, and water-mediated contacts of proteins and small molecule drugs in the DNA minor groove and compare their stereochemistry to positions of the hydration sites. Analysis showed that most minor-groove interactions are directed to purine N3, guanine N2, and pyrimidine O2, binding to deoxyribose O4' is less frequent and close interactions with hydrophobic base and sugar atoms are rare. A large number of protein contacts, roughly one half, are mediated by water molecules but very few in drug complexes. Explanation is not obvious but we proposed that water mediation is more effective at large protein/DNA interfaces where water molecules can fill up cavities between the two macromolecules improving their packing and interaction energy. However, a clear and unequivocal conclusion that follows from the study [Morávek et al., 2002] is that all binding motifs –hydrophilic and hydrophobic protein residues, hydrophilic drugs, and water-mediated protein contacts– bind in a very similar manner. It is illustrated in **Figure 7** that shows preferential binding sites in both purines and pyrimidines. Pyrimidine distributions are always more diffuse than purine ones, noteworthy are comparable sizes of distributions for hydrophobic and hydrophilic contacts.

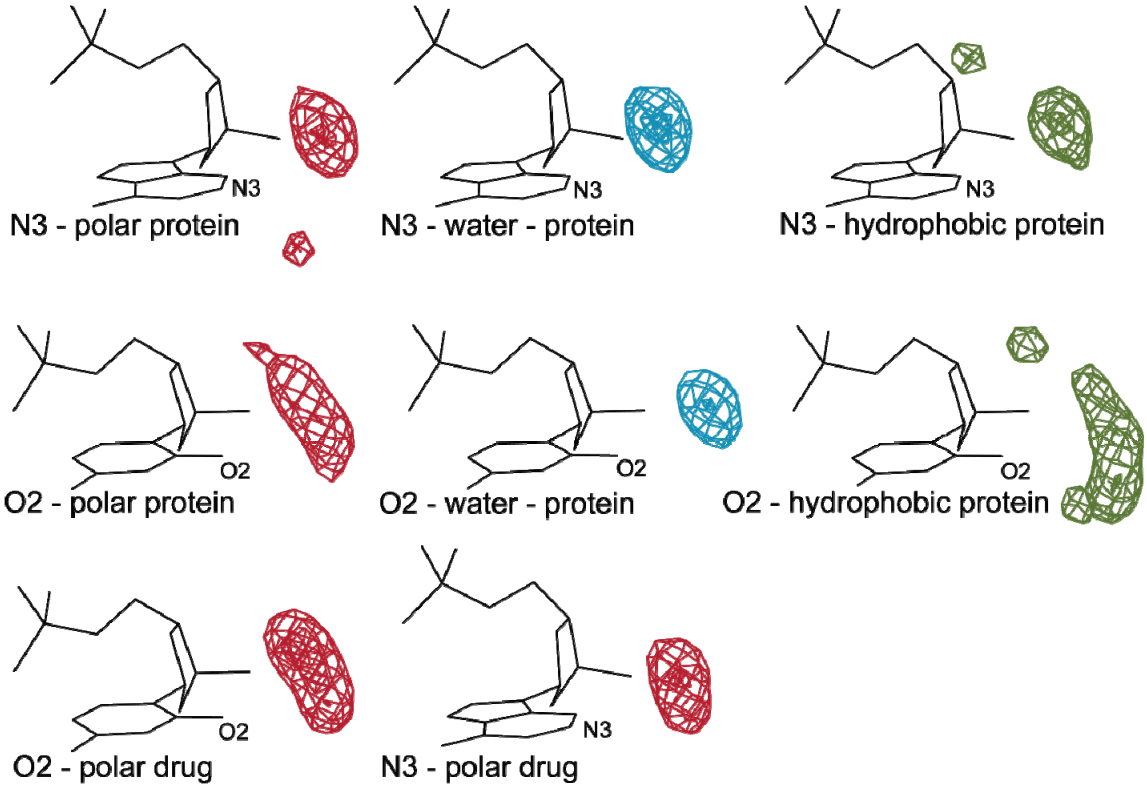

Figure 7. *Preferential binding sites in the B-DNA minor groove [Morávek et al., 2002]. Shown are interactions of polar and hydrophobic amino acid atoms and water-mediated protein contacts and contacts of hydrophilic and hydrophobic atoms from small molecule drugs. DNA atoms are drawn in black, distributions of interacting atoms are color coded as follows: Polar (hydrophilic) atoms in red, hydrophobic in green, water cyan. Purine nitrogen N3 and pyrimidine oxygen O2 are labeled.*

The most compact are distributions of water, all other types of ligands bind in less concentrated regions. However unexpected it is quite logical because the water molecule has the highest degree of freedom of all the ligands and is least restricted in optimization of its binding position. It is not therefore surprising that waters, or in this context hydration sites, may act as effective probes of binding sites; a similar hypothesis has been formulated for protein/DNA recognition by Seeman et al. already in 1976 [67]. We took advantage of the possibility to predict hydration sites around DNA

bases and compared them to amino acid positions in eleven protein/DNA crystal structures [Woda et al., 1998]. We predicted hydration sites around DNA duplexes in their experimental crystal conformations by incorporation of "hydrated building blocks" [Schneider & Berman, 1995]. Comparison of positions of protein atoms in direct hydrogen-bonding contact with DNA and the calculated hydration sites showed a close agreement, and significantly, percentage of good predictions is larger in regions of conserved DNA sequences than in the non-conserved regions; **Figure 8** shows on an example from one analyzed structure that a good correlation between amino acid binding and hydration site is obtained in both minor and major grooves. It can therefore be concluded that positions of protein atoms "recognizing" DNA are predicted well by hydration sites derived from hydration of free (naked) DNA.

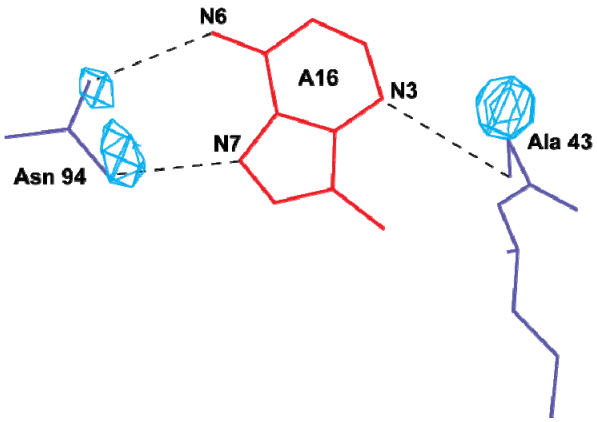

Figure 8. *Interaction of asparagine 94 and alanine 43 of MAT a2 homeodomain protein (grey) with its operator domain adenine base (red) [68]. Hydration site prediction is depicted in cyan density contours [Woda et al., 1998]. Predictions were based on the decamer water distributions [Schneider & Berman, 1995]. Note that alanine 43 binds to adenine N3 by its main chain oxygen.*

The observed distributions of water and amino acids atoms from high resolution crystal structures were transformed into inter-action potentials of their binding [Ge et al., 2005]. The point of the highest density in the distribution ("the peak") represents the site for optimal binding of a given ligand, water, drug or protein atom, and any deviation from the peak is energetically penalized by elliptical function. Distributions were determined by Fourier averaging (called "pseudoelectron density function" in the paper) and by hierarchical clustering methods. The empirically obtained distribution of interacting particles were then approximated by ellipsoids that were subsequently fitted by the energy functions. Potentials derived from the two used clustering techniques were inferior to the potential derived from densities calculated by Fourier averaging.

2.3.3.2. Solvation of DNA phosphates

Hydration around the DNA phosphates has not been studied as thoroughly as base hydration phenomena despite that water has higher affinity to the phosphate charged oxygens than to any other DNA constituent. The paper [Schneider et al., 1998] provides a systematic overview of structural aspects of phosphate hydration in the three main double helical forms compiled using an analogous protocol as in our base-hydration studies. Water is around phosphates more scattered than around the bases, each phosphate charged oxygen is hydrated by three hydration sites, some are split into two overlapping sites; these close sites represent alternative binding positions that are not occupied at the same time. Phosphate hydration by three independent water molecules has been predicted by quantum mechanics [69,70]. Each DNA conformational type shows a particular pattern of hydration of the isolated phosphate group and the stereochemistry of hydration differs also between purines and pyrimidines, and between BI and BII conformers (Figure 2 in [Schneider et al., 1998]). In B-DNA, phosphates are most frequently bridged via their O2P atoms to major groove base atoms of the same nucleotide, to C6 of pyrimidines and slightly less frequently to C8 of purines.

Differences become more pronounced when one compares hydration of two consecutive phosphates; **Figure 9** shows water distributions of purine-pyrimidine steps (hydration of G/A and C/T

nucleotides, respectively, averaged) in the BI and A-DNA. Two B-DNA phosphate groups are too far apart to be bridged by a single water molecule, can only be linked by second shell water molecules. In contrast, charged oxygens O2P from two consecutive phosphates in A-DNA are at near-optimal distance to be bridged by a single water molecule. In fact, the hydration site representing this bridging water is of the highest density of all in A-DNA. Formation of the bridge $O2P_i$ ... W ... $O2P_{i+1}$ is a generalization of the concept of "economy of hydration" [71], that predicted that stability of A-DNA in low-humidity environment is enhanced by sharing hydration spheres of the neighboring phosphates; the concept has also been independently confirmed by high angle neutron scattering studies [72] that have revealed networks of water molecules linking the phosphate groups in the major groove of A-DNA.

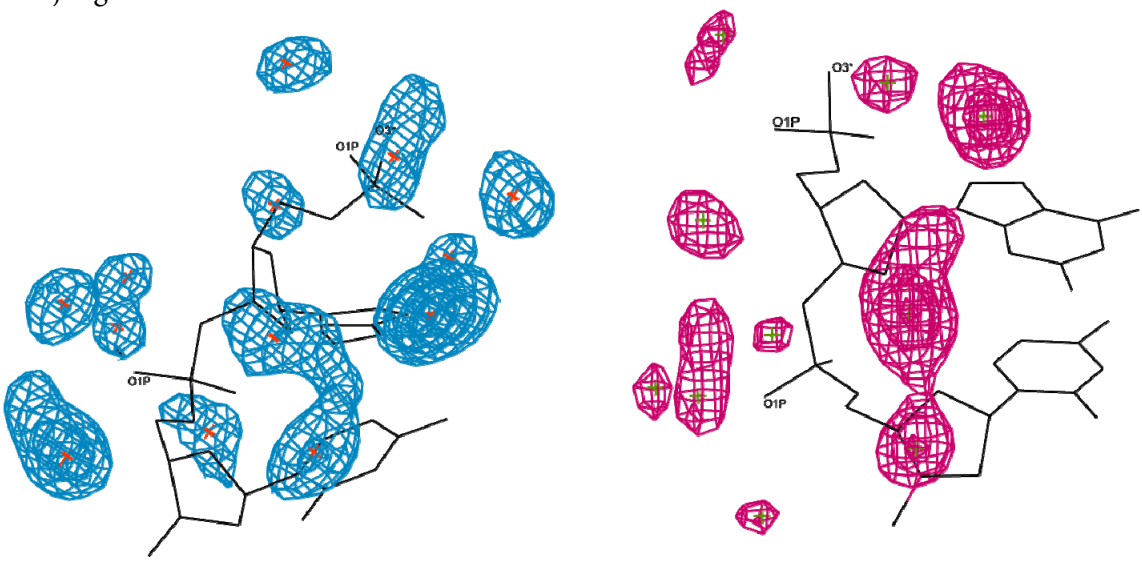

Figure 9. *Hydration pattern of phosphate groups in B (shown is BI conformer) and A-DNA [Schneider et al., 1996]. Two consecutive phosphates in the BI-form (left) have their hydration densities (cyan) isolated around each phosphate while two phosphates in A-DNA (right) share one hydration site. It is the highest density (red) between the two phosphates pointing to the bases. Shown is hydration of a purine-pyrimidine step.*

Crystallographic resolution of oligonucleotide crystal structures is not high enough to always discriminate between water and metal and other solvent species producing in crystals similar electron densities; problem is e.g. with discrimination of Na^+, Mg^{++}, but also NH_3/NH_4^+, all biologically important species. To visualize stereochemistry of important metal cations we therefore turned to high resolution crystal structures of small organic phosphates and analyzed their spatial distributions [Schneider & Kabeláč, 1998] and we were able to determine distributions of several metals, Na^+, K^+, Mg^{++}, Ca^{++}, Zn^{++}, and water. All metals are relatively sharply localized, the best localized are distributions of dications Mg^{++} and Zn^{++}; quite significantly, the stereochemistry of binding is very different for the most biologically relevant metals, Na^+ and Mg^{++}, their distributions are illustrated in **Figure 10**. For all cations, no metal density is observed in the symmetric position along the OP=P lines or in the symmetry axis of the OP=P=OP plane and very low or no density is located in the perpendicular O5'-P-O3' plane.

2.3.3.3. The first hydration shell of DNA is ordered – summary

The main conclusion from our hydration studies, mainly, [Schneider et al., 1993], [Schneider & Berman, 1995], and [Schneider et al., 1998] is that both bases and phosphates have well ordered

first hydration shells; the extent of hydration is larger around phosphates, but water is more organized around bases. The sites of high water densities may represent preferential binding sites for hydrophilic residues and can be used for studying DNA interactions with drugs and proteins ([Woda et al., 1998] and [Morávek et al., 2002]). Various biophysical studies conclude that water has the highest affinity for the phosphate charged oxygens, followed by exocyclic hydrophilic keto groups at the bases, and their exo- and endocyclic nitrogens. The deoxyribose ether oxygen O4' in B-, and A-DNA usually shares water with the minor groove hydrophilic base atom from a previous residue but the distance O_{water}-O4' indicates a weak hydrogen bond. Perhaps surprisingly, the ester oxygens O5' and O3' are hydrated little in double helical DNA; O5' is sterically inaccessible in the right handed forms; the reasons for poor hydration of O3' are not clear.

Hydration sites determined by analysis of oligonucleotide crystal structures allow a rough estimate of the number of water molecules bound in the first hydration shell: Full occupation of all phosphate hydration sites accounts for six water molecules, pyrimidine bases have at least two and purine bases three localized hydration sites. The fully occupied first hydration shell of a nucleotide, therefore, represents between eight and nine water molecules. It accounts for a significant part of the number of five to twelve tightly bound water molecules estimated independently by thermodynamic and spectroscopic methods [55]. Partially ordered water molecules tightly bound to the DNA surface forming the first hydration shell have unique physical properties, the lower mobility [59,60,73], and three to five orders of magnitude slower anisotropic reorientation at low humidity [74] than bulk water. These observations are explained [75] by a model in which blocks of water with limited mobility are bound to the DNA surface. An important feature of this model is that water binding is not due to a larger strength of the DNA – water than water – water interactions but due to the larger anisotropy of the former.

It has been proposed [76] that the cooperative influence of especially polyvalent cations and water is the source of the hydration force [77], which arises from the work of removing water organized at macromolecular surfaces. Cations bound to DNA reconfigure the water at discrete sites complementary to unabsorbed sites between macromolecular surfaces and create these attractive long range forces. The hydration force is detected at intermolecular distances between 5 and 15 Å [78-80]. At a distance of 10 Å, the first hydration shells of both biomolecules are only about 3 - 4 Å apart and only one more layer of water can be placed between them. Bringing the two biomolecules closer would indeed result in release of some interfacial water molecules and accompanied by a favorable entropic effect.

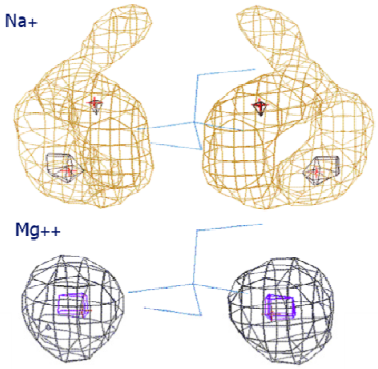

Figure 10. *Distributions of Mg++ and Na+ around the charged phosphate group [Schneider & Kabeláč, 1998]. Distributions derived from well resolved crystals of organic phosphates .*

This concept of self organization of the solvation shell at close intermolecular distances is supported by our analyses of water distributions around DNA, which show more focused bound waters at places where their movement is restricted by another nearby group even when it is hydrophobic. By a simple analogy, we assume that water molecules become more focused as two macromolecules approach each other. The release of water from the significantly ordered first hydration shells into bulk would favorably contribute to the free energy change of the intermolecular interaction by increasing the entropy of the system. Other components of the interaction free energy, such as charge complementarity or steric repulsion, obviously also influence the equilibrium so that the final coor-

dination of the interacting molecules may be a nonspecific interaction at a distance range of 10 Å or a close specific contact with all or a part of the interfacial water removed.

2.4. RNA conformations

In contrast to single –albeit fundamental– biological role of DNA, RNA molecules are involved in various biological processes such as transfer of the genetic information from DNA to proteins by messenger RNA, critical catalytic roles in splicing of pre-mRNA and especially performing protein synthesis in ribosomes, transfer of activated amino acids to the ribosome synthetic center by tRNA, etc. Controlled manipulation of these RNA functions will undoubtedly be of great value for medicine, either by application of traditional "small molecule" drugs, by designing specific ribozymes, anti-sense nucleic acid mimicking drugs, or by designing proteins blocking RNA function by their specific complexation. Variability of RNA functions is reflected by diversity of shapes acquired by RNA molecules and their complex spatial architecture can be compared to that in proteins with domains and complex folding process necessary for proper function of RNA.

The A-form is the prevailing RNA conformation that builds up about two thirds of large folded RNA molecules. A-RNA double helices imply the Watson-Crick base pairing that is however quite often interrupted by regions with nucleotides in "unusual" conformations forming non-W-C base pairs, unpaired bases in bulges, loops of variable lengths, and three- or four-way junctions. Frequently occurring single-stranded regions serve as hinges between double helices or as folding anchors by forming base pairs (often non-W-C) with distant parts of the molecule. A striking difference between DNA and RNA molecules is the presence of a large number of non-W-C, sometimes called non-canonical, pairs in RNA.

The A-RNA double helix can accommodate relatively long stretches of non-W-C base pairs without breaking the double helix and its global parameters (diameter, pitch). This is achieved by local deformations of the backbone including sugar puckers and occasional presence of unusual *syn* orientation of bases. A systematic classification of all possible mutual orientations of two bases has been done by Leontis and Westhof [22] (summarized in **Table I** and illustrated in **Figure 2**); their survey of the observed pairing patterns in the crystal structures [81] allowed for identification of isosteric pairs that can replace each other without disrupting the local geometry. The ability of RNA molecule to accommodate non-W-C pair, bulges, and short internal loops is a prerequisite for forming complex three-dimensional folds. An open issue in RNA structural biology remains how to find and define structural motifs besides the double helix that would simplify large-scale description of RNA folds as helices and β-sheets in proteins. Among the well-known motifs is the kink-turn motif [82], an elbow-like helix-loop-helix of less than twenty nucleotides, UNRA tetraloop, or the S-motif [83].

More detailed description of structural features of various functional types of RNA molecules such as tRNA, varied types of ribozymes, or ribosomal RNAs is beyond the scope of this thesis and can be found in the original papers publishing these structures or in short overview in reviews [Neidle et al., 2009], [Schneider & Berman, 2006].

2.4.1. Conformational dynamics of RNA nucleotides

Advances of RNA crystallography in the seventies promised early description of RNA conformational space and, in fact, first monocrystal structures of nucleic acids were structures of RNA. However, complicated biochemistry of RNA and difficulties with its crystallization slowed the progress

down and shifted most of the interest in the eighties and nineties towards the DNA molecule. The situation started to change in the late nineties with first structures of ribozymes [17,84], but especially with the crystal structures of ribosome subunits that were first solved at rather low resolutions around 5 Å but later in 2000 at reasonable resolution of 3 Å and better [21], [85,86]. By solving and refining these huge structures, structural information about RNA grew by several orders of magnitude and analysis of the multidimensional conformational space of RNA nucleotides became an obvious task. Several laboratories independently started to partition and classify RNA conformations and published their results at different levels of detail in late 2003 [87-89] and early 2004 [Schneider et al., 2004].

For our analysis [Schneider et al., 2004], I decided to analyze the single largest crystal structure available at that time, the large 50S ribosomal subunit refined reliably at the crystallographic resolution of 2.4 Å [21]. The structure provides 2841 dinucleotides that were organized into a data matrix with 2841 rows and fourteen torsion angles as columns. A preliminary analysis confined dinucleotides in the A-form and these were excluded from further analysis. The remaining 830 non-A dinucleotides were analyzed by the Fourier averaging technique (described in section 2.3.2.1). Based on the analysis of one- and two-dimensional distributions of torsion angles, the fourteen dinucleotide torsions were divided into six 3D maps and their analysis resulted in eighteen non-A conformers and several A-form variants.

The conformers can be characterized from different viewpoints; we choose relative base orientation that offers a simple useful classification of conformers: i) non-A, parallel, (nearly) stacked bases, ii) non-A, bases lying in parallel but very close planes, iii) "open" conformers – non-parallel and non-stacked bases, most have a large distance between the successive C1' ribose atoms. Despite our effort we discovered only a few sequence preferences. For instance, in group i) conformer 2 (Table 3 in [Schneider et al., 2004]) has a preference for short, mostly tetra-loops of sequence RNRN; this sequence preference is broader than for the GNRA tetraloops. In group ii), conformers labeled 5 and 6 in the paper occur mostly in purine-rich double helical regions, its bases have very low pitch and form non-Watson-Crick pairs with the opposite strand that has often the same conformation:

```
5'-G A A-3'
   | | |
3'-G G A-5'
```

The color highlighted nucleotides adopt structure of the classified conformations, non-W-C base pairing is indicated by bars. The most "exotic" are open conformers of the group iii) above. The backbone in these conformers can form a sharp U- or S-turns and these conformers form often hinges between single-stranded and double-helical regions, rarely form base pairs and never the Watson-Crick ones. Non-A conformers rarely link to one another and in most cases occur surrounded by A-RNA conformers. The work confirmed the long anticipated [48] central structural role of the phosphodiester link O3'-P-O5' formed by torsion angles ζ and α+1; their scattergram can be to a limited extent compared to the Ramachandran plot of the protein backbone [90], but the full description of conformations of the RNA backbone requires that the other torsion angles, mainly δ and χ are not excluded from explicit considerations and dimensionality of RNA conformational space remains a problem.

The groups competing in search of RNA conformational space [87-89], [Schneider et al., 2004] analyzed different units of the RNA polymer chain (nucleotide, ribose-to-ribose "suite", dinucleotide) and used very different approaches (description of individual techniques is beyond the scope of the thesis, see the references). Encouraging was that despite these differences the results were

comparable, especially between Murray et al. [87] and [Schneider et al., 2004] that published the most detailed data. It was therefore natural to initiate a project to integrate these different approaches and produce "consensus" RNA fragments. The project became a part of the RNA Ontolo-

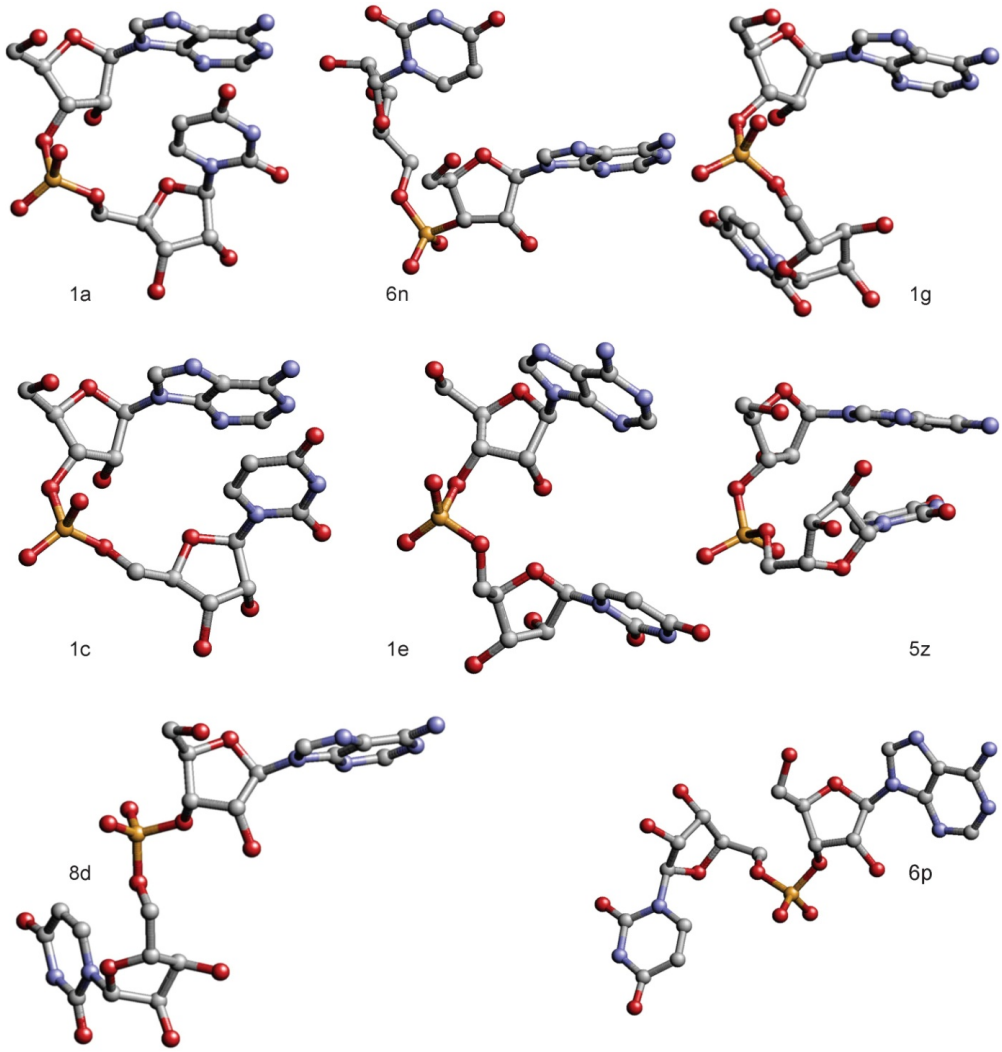

Figure 11. *A few selected important RNA conformers labeled by nomenclature according to* [Richardson et al. 2008]. *1a is canonical A-RNA.*

gy Consortium [91] and resulted in paper [Richardson et al., 2008]. The survey was done on a larger sample of better refined RNA structures and analyzed ribose-to-ribose fragments called "suite". The suite, comprising ribose, the central phosphate group, and the subsequent ribose (**Figure 1**), is formed by seven torsions, δ, ε, ζ, α+1, β+1, γ+1, and δ+1. The work resulted in 46 distinct conformers; a few important examples are shown in **Figure 11**.

The suite is constituted by two overlapping units of "heminucleotides" that were used to develop a new modular nomenclature to describe the 3D backbone conformations as a linear string of two-character conformer names. The first heminucleotide, defined by torsions δ, ε, ζ, is represented by a number, the second heminucleotide, defined by torsions α+1, β+1, γ+1, and δ+1 is represented by a letter. Numbers and letters are semi-mnemonic: Odd numbers of the first half are conformers with

C3'-*endo* ribose, even numbers signify C2'-*endo* pucker. The specific numbers then label particular combination of ε and ζ, for δ in C3'-*endo* –odd numbers ***1***, ***3***, ***5***, ***7***, ***9***– or C2'-*endo* –even numbers ***0***, ***2***, ***4***, ***6***, ***8***. The second heminucleotide is characterized by a letter: Letters ***a***, ***c-n*** represent ribose in the second heminucleotide (described by δ+1) in C3'-*endo*, letters ***b***, ***o-z*** in C2'-*endo*; letter "***b***" is treated as a special case to capture the essence of the DNA most important form, B-DNA, characteristic by the C2'-*endo* pucker. Since two successive suites share torsion δ they are logically required to belong to the common ribose pucker; in terms of the heminucleotide nomenclature, C3'-*endo* conformers ***a***, ***c-n*** can be followed by an odd number while C2'-*endo* ones labeled ***b***, ***o-z*** by an even number. For example, ***1a*** is the A-form conformer with both riboses in C3'-*endo*, ***4s*** is a conformer with C2'-*endo* puckers. The full table of identified conformers can be found in the Table 1 of [Richardson et al., 2008].

The first reading of the above conformer nomenclature may seem confusing and perhaps a bit artificial. However, it brings one essential benefit, namely the possibility to describe three-dimensional RNA as a one-dimensional sequence of symbols. In addition, the suite nomenclature can be combined with the traditional chemical sequence of nucleotides: Symbol

1aG***1a***A***1a***C***1a***U***1a***

represents tetranucleotide of sequence GACU in the A-form. A less trivial example may be a typical conformation of the tetraloop motif of general sequence GNRA:

N***1a***G***1g***N***1a***R***1a***A***1c***N***1a***.

Advantages of automated informatics analysis of such one-dimensional strings over analysis of RNA conformations in Cartesian 3D space or even multidimensional torsion space are obvious.

Software to assign suite code names to RNA structures was published and is available online at the internet address `http://kinemage.biochem.duke.edu/software/suitename.php`. Routine analyses of RNA sequence-conformation space are however held back by several factors that all point to two facts: Lack of RNA structural data and their high noise. High noise in RNA conformational space is partially caused by real complexity of the RNA conformations but its substantial part can likely be attributed to errors in the refinement process. These factors cause that still about 20% of the available RNA fragments cannot be assigned to any conformer. Improvement of dictionaries of RNA conformations used in structure refinement based on the current version of conformers will likely lead to a gradual improvement in quality of the RNA structural data and the possibility to discriminate between the real fluctuations of the RNA conformations and noise caused by errors in the refinement.

2.5. Comparison of RNA and DNA conformations

Molecules of RNA and DNA differ only by the oxygen O2' at ribose sugar ring of RNA that is replaced by a hydrogen in DNA deoxyribose; the difference between uracil and thymine, missing methyl in RNA uracil, is in this context of much lesser importance. Seemingly small difference of a hydroxyl group between DNA and RNA nucleotides has momentous consequences after nucleotides polymerize because the extra hydroxyl group adds hydrogen bonding donor to the molecule with deficit of hydrogen bond donor, actually no donor is present in the DNA backbone.

Both molecules share one fundamental similarity, their most abundant and stable forms are antiparallel right-handed double helices, BI-DNA and AI-RNA, that form more than 2/3 of all conformers. These structural elements are however used in radically different ways in each molecule:

While the dominant characteristic of the DNA molecule is self recognition of two strands, RNA can be characterized by forming 3D folds. Structural differences between the two molecules are striking at the global level of molecular architecture as well as when one compares their local conformations. While DNA does exhibit unusual architectures such as quadruplexes [34,92] or junctions [41,93] these structures are rare, fulfill particular functions, and cannot match complexity and variability of RNA folds.

The situation at the local nucleotide level is comparable: Variability and diversity of RNA conformers contrast with existence of many but mutually similar DNA conformers that can transform from one to another in an almost continuous fashion. More specifically, DNA conformers that form the right-handed duplexes of B- and A-forms all belong to one large ensemble of conformations. Transformation between BII, BI, AI, and AII can indeed occur in almost continuous fashion. Only specific DNA sequences at extreme salt concentrations can form radically different left-handed Z-form but even this form is antiparallel double helix. In contrast, RNA nucleotides form a host of radically different conformers. When RNA molecule is locally diverted from the most stable A-form, it flips to a conformationally distant structure that may be stabilized by hydrogen bonding of the ribose hydroxyl O2'. DNA, with its numerous closely related conformers, is conformationally soft whereas RNA with fewer but conformationally distant conformers is "brittle", not "rigid".

A particular, rather technical but important issue is related to the alleged difference in populated sugar pucker modes in both molecules. Although all RNA "consensus conformers" show strict bimodality between C3'- and C2'-*endo* ribose puckers, I believe that it remains to be seen if other pucker modes undoubtedly observed in DNA are also present in RNA. Scarce population of minor pucker modes in RNA conformations may be a consequence of the refinement protocols using incomplete geometric dictionaries (see also section 3.1.1) rather than their actual non-existence. **Figure 12** compares populations of torsion angle δ in DNA and RNA in the δ/χ scattergram, which illustrates striking absence of puckers between sharply delimited C3'- and C2'-*endo* regions in RNA. An indication that the lack of minor ribose puckers in known RNA structures may be more a consequence of incomplete dictionaries than reflection of reality was already obtained during the analysis of the "consensus conformers" [Richardson et al., 2008] where one cluster with the O4'-*endo* pucker was identified by the Fourier averaging. Because other approaches had not localized this cluster it was decided not to include it to the list of consensual conformers.

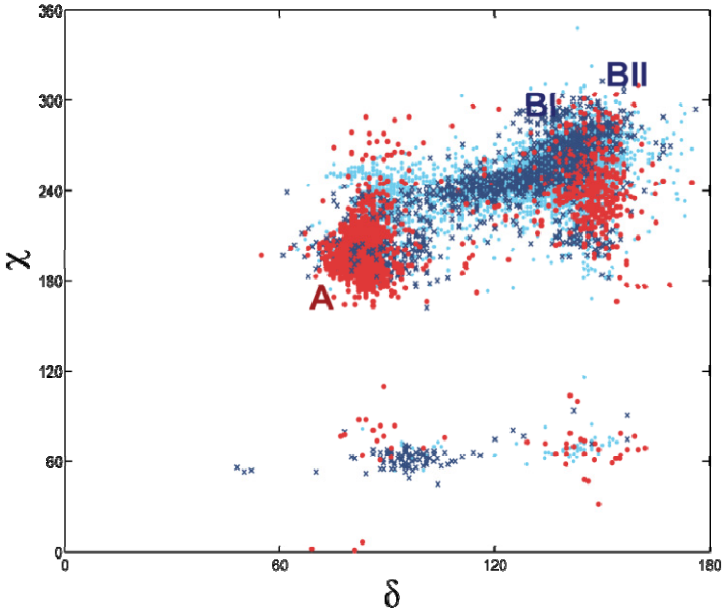

Figure 12. *Scattergram of the backbone torsions δ and χ in RNA (red dots), naked DNA (dark blue), and protein/DNA complexes (light blue).*

3. Infrastructure for structural biology of nucleic acids

3.1. Tools for solution and interpretation of nucleic acid structures

3.1.1. Dictionaries of standard geometries

Reliable geometries of building blocks of nucleic acids and proteins are needed for refinement and interpretation of experimentally determined biomolecular structures and for their modeling. Geometric parameters of biomolecules or rather their building blocks as amino acids or nucleotides are collected in so called dictionaries. Crystallography, NMR spectroscopy, and molecular modeling techniques require dependable values especially for valence geometries of protein and nucleic acid constituents, amino acids and nucleotides but also populated regions of torsion angles in a form of torsion periodic potentials but geometric parameters have also been collected for important classes of other molecules, drugs, ligands, solvent species, and other chemicals.

The values of all parameters in the geometric dictionaries should be as close to the physically accurate values as possible because any deviation would systematically bias refinement of experimental structures or run of computer modeling. The best available source of biomolecular geometric standards are high resolution small-molecule crystal structures that are contained in the Cambridge Structure Database, CSD [27]. Crystal structures used for such analyses need to be expertly selected for as high crystallographic resolution as possible, checked for possible experimental errors, as well as for incorrectly conducted refinement; these structures should also be refined only against experimental data without using any geometric constraints or restraints not to bias the final molecular geometry.

Geometry of the nucleic acid backbone was analyzed in high-resolution crystal structures of RNA and DNA nucleotides from CSD [27] and NDB [Berman et al., 1992] by [Gelbin et al., 1996]. This and the accompanying statistical studies of nucleobase geometries [94] provide a reliable set of valence geometries for crystallographic refinement of nucleic acids, they used three times larger sample size than the previous studies of the valence geometry of nucleotides [25,26]. It was possible to distinguish geometric differences between ribose and deoxyribose rings and each sugar was treated separately for their two main conformational regions, C2'- and C3'-*endo* and provide reliable values of P-O and O-C bond distances and the related bond angles.

Results of the DNA backbone analysis already discussed in section 2.3.2 [Schneider et al., 1997] were instrumental for formulation of an updated geometric parameterization [95] implemented to by then the most popular program for refinement of x-ray and NMR structures, x-plor [96]; the dictionary was later transferred to the x-plor successor, CNS [97].

3.1.2. Fitting of electron density maps

A typical protocol of x-ray crystallography has several steps before the final structural model is determined [98,99]. Each of these steps requires extensive expertise in various fields from molecular biology, physics, and to software engineering. Few steps are however as time consuming and subjective as building of molecular models into electron density maps after the phasing of the diffraction data is completed. When done manually, this work requires a great deal of patience but it is also to a large extent subjective and the resulting molecular model may depend on the experience of the researcher and automation of the fitting procedure would therefore accelerate the process of structure refinement and standardize features of the final model. The issue has partially been resolved for pro-

teins [100-102] but it has barely been touched for nucleic acid crystal structures; especially modeling of large RNA molecules producing typically crystals of lower resolution is a complicated task.

The work [Pavelčík & Schneider, 2008] is the first that successfully developed a procedure for building the double helical regions of the B and A conformations into electron maps of RNA and DNA. The procedure utilizes the technique of the phased rotation, conformation, and translation function originally developed for protein fitting [103] and positions tetranucleotide double helical fragments (two dinucleotide strands with two Watson-Crick paired bases, Figure 1 in [Pavelčík & Schneider, 2008]) in A- and B-conformations. Short double helical segments are connected into longer chains wherever the segments overlap in the right geometry. The sequence of polynucleotide chain is approximated by discriminating between purines (guanine, adenine) and pyrimidines (cytosine, uracil). The method was tested for nine RNA structures of various sizes determined at resolutions from 1.5 to 3.1 Å. Almost all residues in A-RNA double helices in the fully refined models were correctly positioned into the electron density; the average root mean square deviation was 0.8 Å, comparable to the values obtained for fitting of small proteins. The double helical regions were actually over-fitted in some structures. This is a consequence of the fact that Watson-Crick pairs are almost isosteric with some "mismatched" pairs and they have similar electron-density envelopes. The issue while not critical is relatively serious because pairs similar to Watson-Crick ones are quite frequent in RNA structures. The issue will likely be resolved by the development of a procedure fitting of dinucleotide fragments of universal shape; work towards this goal is under progress.

3.1.3. Interpretations of parameters of NMR spectroscopy

NMR spectroscopy is an important method for determination of molecular structures. It should be noted that various NMR techniques have been noticeably successful in structure determination of notoriously hard to crystallize small RNA molecules. NMR spectra strongly depend on local molecular structure. Their successful interpretation in structural terms therefore requires intimate knowledge of the physical fundaments of the measured spectral characteristic but also of structural behavior of the measured molecules. A variety of NMR techniques have been used for determination of nucleic acid structures; they use diverse physical effects including the nuclear Overhauser effect (NOE), the NMR shift, direct and indirect spin-spin coupling, and the cross-correlation relaxation rates [104-109].

Scarcity of hydrogens (protons) in nucleic acids causes that the signal from NOE measurements does not yield enough data to determine all-atom molecular models. Measurement of NMR shifts alone also does not allow building of molecular models without inclusion of relaxation rates between the chemical shift anisotropy and the bond vectors. Experimental as well as computational complexities linked to these parameters are still considerable and their application for structure determination is far from routine. A promising approach is measurement of indirect spin-spin coupling constants, "J-couplings", that provide independent and specific information about conformation around the coupled nuclei [110,111]. J-couplings are experimentally accessible values but their use is sometimes limited by the lack of rules and tools to interpret them in terms of structure.

Possibility to correlate J-couplings between ^{31}P, ^{13}C, and ^{1}H nuclei in the sugar-phosphate backbone with nucleic acid conformation seemed a promising project. High dimensionality of the nucleic acid conformational space and complex relationships between torsion angles of the back-bone make any "brute force" or systematic scanning of the torsions unfeasible for determination of any quantity.

We therefore decided to take advantage of our *a priori* knowledge of the nucleic acid conformers and calculated J-couplings for a subset of the RNA conformers [Schneider et al., 2004], [Richardson et al., 2008] and BI-form of DNA with the goal to characterize backbones of these conformers by unique sets of J-couplings [Sychrovský et al., 2006b]. The predictions of J-couplings were calculated using a molecular model of dinucleoside-3'-5'-phosphate with torsion angles fixed at their known values and bases replaced by methyl groups (Figure 1 in [Sychrovský et al., 2006b]) by the DFT method with inclusion of all four coupling terms, diamagnetic spin-spin, paramagnetic spin-orbit, Fermi contact, and spin-dipolar [112,113]. Solvent effects were considered at the level of "polarized continuum solvent" (PCM) and extensively tested by a model of explicit hydration of the phosphate group by six water molecules [Schneider et al., 1998] immersed in the PCM solvent.

A correlation between J-couplings and conformation was originally proposed for rotameric states in hydrocarbons by Karplus [114]. Well-known "Karplus equations" have since then been parameterized for various spin-spin interactions in different types of molecules including nucleic acids. Karplus equations relate one J-coupling to one torsion angle and are intrinsically one-dimensional. However, nucleic acid conformational space is inherently multidimensional that may –and we know it actually does– limit their applicability. We demonstrated that while a J-coupling primarily depends on one torsion angle, the dependence is often modulated by other torsions. As described in [Sychrovský et al., 2006b], correlations between "soft" torsion angles, e.g. β, and J can still be conceptualized in terms of the classical Karplus equation but other "hard" torsions, typically γ and δ, interfere with one another so strongly that the Karplus equation of one torsion must be calculated separately for separate values of the other torsion.

Unequivocal determination of values of all seven nucleotide torsion angles by J-couplings is unfortunately still not possible especially for torsions at the phosphodiester linkage, α and ζ, clarification of the problem will require further study. However, [Sychrovský et al., 2006b] does offer an algorithm for improved assignment of the backbone torsions based on J-couplings: Assign torsions γ and δ first. Their knowledge then sharpens preliminary determination of torsions ε and β, all four should then be used to limit possible ranges for α and ζ.

Also other NMR parameters have been studied with the goal of their structural interpretation using a similar approach combining empirically determined conformation of a nucleic acid fragment and quantum mechanics computations in approximation of density functional theory (DFT) [Sychrovský et al., 2005], [Sychrovský et al., 2006a], [Vokáčová et al., 2009].

3.2. Design, development, and maintenance of structural databases

Databases are an important part of scientific infrastructure, in biology probably more than in any other scientific field. First databases of crystal structures evolved very soon after the diffraction techniques had emerged as a powerful tool of structural analysis in sixties. Especially the Cambridge Structure Database paved the way to other databases by its query and report capabilities [115] and became electronic with emergence of the first useable computers. Also the founders of the first and the most important database of biological structures, the Protein Data Bank (PDB), had exceptional foresight: At the time when the database was founded in October 1971 there had been seven protein structures solved – today there are almost seventy thousand crystal, NMR, and EM structures in the PDB archives [116].

PDB and CSD were established in response to the needs of crystallographers and users of crystal structures. They have however differed from the very beginning in one important aspect: While

PDB has always been funded by the grant money from the US government and all its resources have been available freely, the CSD was initiated as a self-supporting project financed by licensing the use of the database by a fee. As an indirect consequence, the CSD staff has collected the relevant crystal structures of organic molecules itself from journals and other resources, PDB relied on voluntary deposition of structures by their authors. The vast majority of the crystallographic community understood the need to submit their structures to PDB but there were occasional "islands of resistance" against deposition of structural models (coordinates) and especially the structure factors. The current policy of virtually all journals and grant agencies is deposition of coordinates as well as experimental data of all publically funded structures before they can be published.

3.2.1. Nucleic Acid Database, NDB

As the name of PDB suggests, nucleic acids were not in the focus of its curators. PDB was at the beginning operated by the Brookhaven National Laboratories [117] and despite that most structures of nucleic acids and their complexes were archived in the PDB, the quality of annotation was not always flawless. Because also shorter, di- and tetranucleotide structures were not included to the PDB and were available only from the CSD [115], it became obvious that a specialized database would be required to facilitate archival, querying, and reporting of structural information about nucleic acids. Such a database, the Nucleic Acid Database, NDB, was established in 1991 at Rutgers University [Berman et al., 1992].

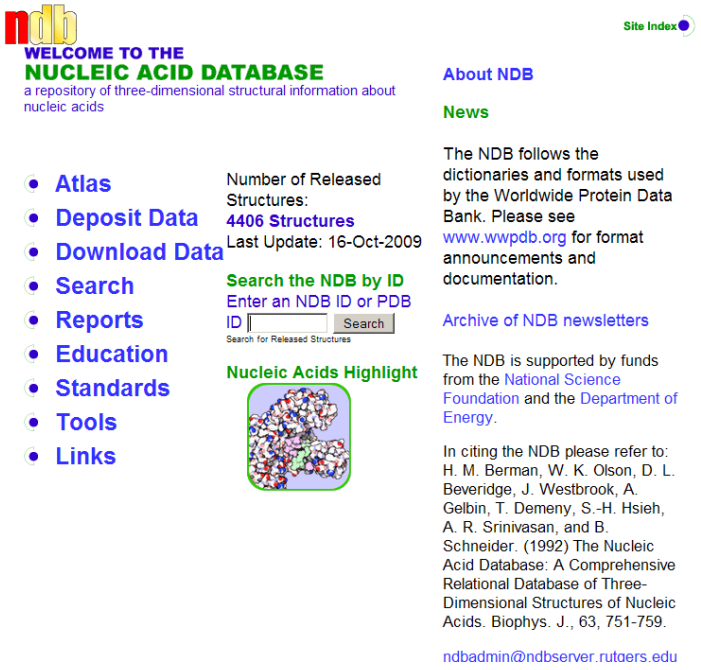

Figure 13. *NDB home page http://ndbserver.rutgers.edu*

In contrast to the BNL-operated PDB [117], NDB has been built as a relational database from the beginning and a substantial part of the NDB project has always been development of computational tools for deposition, validation, querying, and dissemination of information about structures of biological molecules. When NDB was established, the known nucleic acid structures consisted of DNA and RNA oligonucleotides, a few protein/DNA complexes, and some transfer RNAs. Structures were annotated manually and classified into a few known molecular architectures by visual inspection. Since then, entirely new types of structures have emerged, many complexes with proteins, structures of ribozymes, and especially large structures of ribosomes. Mere number of these structures but mainly their complexity and novel, often unexpected features presented considerable challenge for their correct, reliable but quick annotation. Accommodation of this challenge required development of robust data processing system that would produce "flat files" that were easy to load

into the database. Such a system, the Macromolecular Crystallographic Information File (mmCIF, [118]), was developed as a part of the NDB project and is included in the current version of the database [Berman et al., 2002b]. The mmCIF format is now widely accepted as archival and together with the related PDBML format serves as exchange format for bioinformatics [119]. It provides the comprehensive dictionary of terminology for crystallography, NMR, and molecular structure. This self-defining format is built on syntax rules that define relationships between individual data items and allow easy checking of the data and their native loading to relational database(s). Today, the tools developed for the NDB form the basis of the software used by the RCSB PDB (see below) for processing of biomolecular structures of nucleic acid as well as proteins.

NDB contains *primary* and *derived* data. The primary data include coordinates of the molecular models derived from the crystallographic or NMR experimental data, the experimental data themselves –structure factors for crystal, and distance constraints for NMR structures– various collection and refinement statistics, information about molecules as organism(s) of origin, sequence(s), and also "demographics" of the structures as authors and relevant references. Derived data are generated by standard procedures by the annotation staff or calculated by generally accepted algorithms. All this information is available *via* the web interface (**Figure 13**). The web interface also provides simple forms for querying the underlying database, the results can be summarized in custom-made or prepared tabular reports; an effective way of surveying the structures is browsing the Atlas pages that provide a concise overview of a particular structure. The interface provides access to the tools developed by the NDB project, various geometric standards and programs, and allows download of the structures in various archived formats.

3.2.2. Protein Data Bank, PDB

PDB is the single worldwide depository of structures of biological macromolecules. It was established and initially maintained at Brookhaven National Laboratory [117], has now the main operation site at the Research Collaboratory for Structural Bioinformatics, RCSB [116], [Berman et al., 2002a] but independent deposition sites are maintained also by the European Bioinformatics Institute in the UK and the Institute for Bioinformatic Research and Development in Japan. To ensure the existence of a single archive, these three sites joined and created consortium of World Wide PDB, wwPDB [120].

PDB is an essential tool of structural bioinformatics, molecular modeling of macromolecules, many other databases derive their data from the PDB archives. PDB serves a wide user community: It is a primary resource for specialists as are authors of three-dimensional structures, crystallographers and NMR spectroscopists, as well as for their users, bioinformatitians, computer biologists, and natural scientists in general. PDB must however cater also for the needs of non-specialists, mainly high school teachers and students and for the general public. All these needs, sometimes contradictory, require fulfillment of the general demands on any large database facility: Long term archiving and availability of the deposited data, scalability, ability to query and report the database, download the primary as well as derived data in community-accepted format(s) but PDB also provides the community with software tools as programs, dictionaries of geometry parameters, format definitions, tools for structural interpretation, distribution of data, and teaching of biomolecular structures.

Ever growing pace of depositions of structures to PDB presented a great challenge to the PDB staff, especially at the time of the management change from the BNL to RCSB in 1998. The RCSB finished development of a new data processing system based on the mmCIF format, which had been

initiated by the NDB, and integrated deposition and annotation processes [Berman et al., 2002a]. The process of annotation of deposited structures has been described [Burkhardt et al., 2007].

4. Summary

The thesis summarizes author's contribution to structural biology of nucleic acids.

The main part of the thesis, Section 2, offers an overview of structural features of DNA and RNA. Composition and nomenclature of building blocks of nucleic acids, nucleotides, are briefly introduced (2.1) and the basic facts about their spatial motifs and self-assembling ability are presented (2.2). Section 2.3 outlines molecular architectures of DNA [Schneider & Berman, 2006] and then talks about author's contribution to oligonucleotide crystallography in section 2.3.1, [Schneider et al., 1992a], [Schneider et al., 1992b], [Harper et al., 1998]. The rest of the studies presented in section 2 can be characterized as structural bioinformatics and their strength is, to some perhaps paradoxically, in their purely phenomenological nature; these studies are sometimes also called "knowledge-based". DNA local conformations and their variability are described in section 2.3.2, [Schneider et al., 1997], [Svozil et al., 2008]. An original method of "Fourier averaging" is described in section 2.3.2.1, [Schneider et al., 1993], [Schneider & Berman, 1995], [Schneider et al., 2004]. An important aspect of DNA structural integrity, solvation, is discussed in section 2.3.3, [Schneider et al., 1992], [Schneider et al., 1993], [Schneider & Berman, 1995], [Schneider et al., 1998], [Woda et al., 1998], [Schneider & Kabeláč, 1998], [Morávek et al., 2002], [Ge, et al, 2005]. RNA architectures are globally mentioned at the introduction of section 2.4 [Schneider & Berman, 2006] that is followed by discussion of analyses of conformers of RNA dinucleotides [Schneider et al., 2004], [Richardson et al., 2008]. Conformational behavior of DNA and RNA at the local level is compared in section 2.5.

Section 3 recapitulates author's involvement in projects and activities leading to building infrastructure for structural biology and bioinformatics, mainly of nucleic acids. Characterization of geometry parameters and dictionaries of nucleic acid geometry is summarized in section 3.1.1 [Schneider et al., 1997], [Gelbin et al., 1996]. Our contribution to automation of labor-intensive refinement of molecular models derived from crystallographic data, a program for fitting of DNA and RNA double helical fragments into maps of electron densities from diffraction experiments [Pavelčík & Schneider, 2008], is mentioned in section 3.1.2. Effort towards improved structural interpretation of spectral parameters of the important method of determination of 3D structures of biomolecules, nuclear magnetic resonance is summarized in section 3.1.3, [Sychrovský et al., 2005], [Sychrovský et al., 2006a], [Sychrovský et al., 2006b], [Vokáčová et al., 2009]. Structural databases are the cornerstone of the infrastructure for structural biology, bioinformatics, and computational biology; two highly important and widely used databases are briefly mentioned in section 3.2. The Nucleic Acid Database, NDB, section 3.2.1, is the primary deposition site for experimental structures containing nucleic acids [Berman et al., 1992], [Berman et al., 2001], [Berman et al., 2002b], [Schneider et al., 2009]. The Protein Data Bank, PDB, 3.3.2, is the most important database of biological structures [Berman et al., 2002a], [Burkhardt et al., 2007].

Papers cited in the above paragraphs are listed on page 34, the color highlighted ones are compiled in the thesis.

5. Perspective

During the last two decades, structural biology has undergone a rapid development. We now have within our reach structural data about tens of thousands of biomolecules at atomic detail. Most of them are proteins, but our knowledge of nucleic acid structures has exploded especially with the solved crystal structures of ribosomal particles. Our knowledge of intermolecular complexes is however much more limited and the principles underlying their recognition are understood even less, not to mention understanding the dynamics of the interactions: Description of biological processes as interactions of molecules has indeed not been achieved as yet. The following paragraphs are a "laundry list" that comes to mind when thinking about the future of structural biology.

Experimental determination of molecular structures will remain at the heart of structural biology and the number of experimentally determined structures is going to grow. Experimental techniques for protein and nucleic acid production, especially in eukaryotic systems, their purification and crystallization, will keep evolving, obviously not only to fulfill the demands of structural biology. Technologies for data acquisition in crystallography (synchrotrons, new detectors) and NMR spectroscopy will be more powerful and also widely available to the scientific community. Insight into physical nature of diffraction on crystals is likely to bring more effective methods of the phase problem solution and better models of interactions between nuclear magnetic moments will provide for new algorithms of signal assignment in NMR techniques. When we concentrate on bioinformatics, the focus of the thesis, one can foresee near-complete automation of the process of fitting of molecular models into electron maps. Further success of NMR techniques will not depend only on higher magnetic field of newly built spectrometers but also on firmly established interpretation of spectral parameters in structural terms; one can also envisage a firmer procedure for validation of NMR-determined structures.

A growing proportion of newly determined structures will be large multi-molecular machines executing complex chemical or mechanical tasks as light capture, ATP generation, RNA synthesis. Crystallization of such large complexes or their controlled preparation for electron microscopy brings challenges specific for each studied system that will require tight collaboration between biologists and physicists. Two separate areas where inventions and development of fundamentally new techniques would be highly desirable are membrane proteins and their complexes, and research of unstructured or quasi periodic protein formations as amyloids.

Understanding of the way how large molecular complexes and biological nanomachines function in space will require development of techniques able to fill the spatial gap between the submolecular, 10^0 – 10^1 Å, and nano-to-micro scales, 10^3 – 10^4 Å. This will enable integration of the extensive knowledge we acquired about behavior of subatomic systems by spectroscopy, quantum mechanics and the like, about large molecules described crystallographically and thermodynamically, and about micrometer scale subcellular and cellular systems reachable by light microscopy. However, biological systems function in space and *time* and their full understanding without the time dimension is always questionable. Therefore, data about spatial organization of the biological structures need to be acquired with time resolution of microseconds, perhaps nanoseconds. An obvious candidate technique to fill in the spatial gap is (cryo)electron microscopy but the fast measurements will require development of revolutionary techniques of diffraction of single particles, perhaps employing as yet emerging x-ray laser technology. Chromosome with extremely packed DNA millions of pairs long and hundreds of associated proteins is an example of biological nanoparticle which function is strictly orchestrated in space and time. Description and understanding of its structure

and dynamics, the ways how DNA is maintained, copied, and its transcription regulated to synthesize the right genes at the right time symbolize the complexities facing structural molecular biology in the coming years.

Growing numbers of determined structures will put pressure on efficient, reliable, and freely available bioinformatic tools. Databases will remain the main tool of bioinformatics and molecular modeling; they will have to be able to archive, store, query, and distribute more and larger structures. All this will be possible only if newly deposited structures will be carefully and *consistently* annotated. Besides their relational model, new types of databases will become available for more efficient data mining and knowledge.

Bioinformatic studies will greatly benefit from the increased number of structures, the conformational and sequence spaces will be better sampled so that construction of fully empirical (and therefore robust) force fields will be possible because available experimental structures will represent the potential energy surface of a given studied molecule well by the strength of the ergodic principle. The alphabet of protein and RNA folds will be close to complete and a representative gallery of protein, RNA, and DNA 3D motifs of varying sizes will be put together. New physical models incorporated into publically available modeling software should then be able to predict their possible assembling into new stable (and useful) molecules. New robust methods of 3D comparison will hopefully be developed.

Seemingly more down to earth but practically important task is development of techniques to explain and predict affinity and specificity of molecular interactions at the atomic and molecular levels. The acquired expertise will allow rational design of high-affinity binders to nucleic acids and proteins as drugs, diagnostics, and biotechnological materials. This often mentioned but rarely achieved understanding of "structure-activity relationships" will require much more detailed description of thermodynamics and kinetics of biological processes and their deeper understanding; physicists will perhaps develop new models fundamentally new moving beyond the Boltzmann equilibrium thermodynamics to better reflect dynamic, fluent, and inherently non-equilibrium biological systems.

References

Papers co-authored by the applicant and cited in the text

Alphabetically by the first author

Berman HM, Olson WK, Beveridge DL, Westbrook J, Gelbin A, Demeny T, Hsieh S-H, Srinivasan AR & Schneider B (1992). The Nucleic Acid Database - A Comprehensive Relational Database of Three-Dimensional Structures of Nucleic Acids. *Biophys.J.* **63,** 751-759.

Berman HM, Feng Z, Schneider B, Westbrook, J, & Zardecki C (2001). The Nucleic Acid Database (NDB). In *International Tables for Crystallography*. Volume F, Crystallography of Biological Macromolecules. (eds. MG Rossmann & E Arnold), pp. 657-682, Kluwer Academic, Dordrecht, The Nederlands.

Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, Feng Z, Gilliland G L, Iype L, Jain S, Fagan P, Marvin J, Ravichanran V, Schneider B, Thanki N, Padilla D, Weissig H, Westbrook JD & Zardecki C (2002a). The Protein Data Bank. *Acta Cryst. D* **58,** 899-907.

Berman HM, Westbrook, J, Feng Z, Iype L, Schneider B & Zardecki C (2002b). The Nucleic Acid Database. *Acta Cryst. D* **58,** 889-898.

Burkhardt K, Schneider B & Ory J (2006). A Biocurator Perspective: Annotation at the Research Collaboratory for Structural Bioinformatics Protein Data Bank. *PLOS Comput Biol.* **2**, 1186-1189.

Gelbin A, Schneider B, Clowney L, Hsieh S-H., Olson WK & Berman HM (1996). Geometric parameters in nucleic acids: sugar and phosphate constituents. *J. Am. Chem. Soc.* **118,** 519-528.

Ge W, Schneider B & Olson WKO (2005). Knowledge-Based Elastic Potentials for Docking Drugs or Proteins with Nucleic Acids. *Biophys.J.* **88**, 1166-1190.

Harper A, Brannigan JA, Buck, M, Hewitt L, Lewis RJ, Moore MH & Schneider B (1998). The structure of d(TGCGCA)$_2$ and a comparison to other DNA hexamers. *Acta Cryst. D* **54**, 1273 - 1284.

Kratochvilová I, Král K, Bunček M, Víšková A, Nešpůrek S, Kochalská A, Todorciuc T, Weiter M & Schneider B (2008). Conductivity of natural and modified DNA measured by scanning tunneling microscopy. The effect of sequence, charge and stacking. *Biophys. Chem.* **138,** 3-10.

Morávek Z, Neidle S & Schneider B (2002). Protein and drug interactions in the minor groove of DNA. *Nucleic Acids Res.* **30,** 1182-1191.

Neidle S, Schneider B & Berman HM (2003). Fundamentals of DNA and RNA structure. In *Structural Bioinformatics*, (eds. PE Bourne & H Weissig), pp. 41-73, John Wiley & Sons, Inc., Hoboken, NJ.

Neidle S, Schneider B & Berman HM (2009). Fundamentals of DNA and RNA structure. In *Structural Bioinformatics*, 2nd edition (eds. J Gu & PE Bourne), pp. 41-76, Wiley-Blackwell, Hoboken, NJ.

Pavelčík F & Schneider B (2008). Building of RNA and DNA double helices into electron density. *Acta Cryst. D* **64,** 620-626.

Richardson JS, Schneider B, Murray LW, Kapral GJ, Immormino RM, Headd JJ, Richardson DC, Ham D, Hershkovits E, Williams LD, Keating KS, Pyle AM, Micallef D, Westbrook J & Berman HM (2008). RNA backbone: Consensus all-angle conformers and modular string nomenclature (an RNA Ontology Consortium contribution). *RNA* **14,** 465-481.

Schneider B, Ginell SL & Berman HM (1992a). Low Temperature Structures of dCpG-proflavine: Conformational and Hydration Effects. *Biophys. J.* **63**, 1572-1578.

Schneider B, Ginell SL, Jones R, Gaffney B & Berman HM (1992b). Crystal and Molecular Structure of a DNA Fragment Containing a 2-Amino Adenine Modification: The Relationship between Conformation, Packing, and Hydration in Z-DNA Hexamers. *Biochemistry* **31**, 9622-9628.

Schneider B, Cohen DM, & Berman HM (1992c). Hydration of DNA bases: Analysis of crystallographic data. *Biopolymers* **32,** 725-750.

Schneider B, Cohen DM, Schleifer L, Srinivasan AR, Olson WK & Berman HM (1993). A systematic method for studying the spatial distribution of water molecules around nucleic acid bases. *Biophys. J.* **65,** 2291-2303.

Schneider B & Berman HM (1995). Hydration of the DNA bases is local. *Biophys. J.* **69,** 2661-2669.

Schneider B, Kabeláč M & Hobza P (1996). Geometry of the phosphate group and its interactions with metal cations in crystals and ab initio calculations. *J. Am. Chem. Soc.* **118,** 12207-12217.

Schneider B, Neidle S & Berman HM (1997). Conformations of the sugar-phosphate backbone in helical DNA crystal structures. *Biopolymers* **42,** 113-124.

Schneider B, Patel K & Berman HM (1998). Hydration of the phosphate group in double-helical DNA. *Biophys. J.* **75,** 2422-2434.

Schneider B & Kabeláč M (1998). Stereochemistry of binding of metal cations and water to a phosphate group. *J. Am. Chem. Soc.* **120**, 161-165.

Schneider B, Morávek Z & Berman HM (2004). RNA conformational classes. *Nucleic Acids Res.* **32,** 1666-1677.

Schneider B & Berman HM (2006). Basics of Nucleic Acids Structure. In *Computational Studies of RNA and DNA,* (Šponer J & Lankaš F, eds.), pp. 1-44, Springer, Dordrecht.

Schneider B, de la Cruz J, Feng Z, Chen L, Dutta S, Persikova I, Westbrook, J, Yang H, Young J, Zardecki C & Berman HM (2009). The Nucleic Acid Database. In *Structural Bioinformatics*, 2nd edition (eds. J Gu & PE Bourne), pp. 305-319, Wiley-Blackwell, Hoboken, NJ.

Svozil D, Kalina J, Omelka M & Schneider B (2008). DNA conformations and their sequence preferences. *Nucleic Acids Res.* **36,** 3690-3706.

Sychrovský V, Muller N, Schneider B, Smrecki V, Špirko V, Šponer J & Trantírek L (2005). Sugar pucker modulates the cross-correlated relaxation rates across the glycosidic bond in DNA. *J. Am. Chem. Soc.* **127,** 14663-14667.

Sychrovský V, Šponer J, Trantírek L & Schneider B (2006a). Indirect NMR spin-spin coupling constants (3)J(P, C) and (2)J(P, H) across the P-O … H-C link can be used for structure determination of nucleic acids. *J. Am. Chem. Soc.* **128,** 6823-6828.

Sychrovský V, Vokáčová Z, Šponer J, Špačková N & Schneider B (2006b). Calculation of structural behavior of indirect NMR spin-spin couplings in the backbone of nucleic acids. *J. Phys. Chem. B* **110,** 22894-22902.

Vokáčová Z, Buděšínský M, Rosenberg I, Schneider B, Šponer J & Sychrovský V (2009). Structure and Dynamics of the ApA, ApC, CpA, and CpC RNA Dinucleoside Monophosphates Resolved with NMR Scalar Spin-Spin Couplings. *J. Phys. Chem. B* **113,** 1182-1191.

Woda J, Schneider B, Patel K, Mistry K & Berman HM (1998). An Analysis of the Relationship between Hydration and Protein-DNA Interactions. *Biophys. J.* **75**, 2170-2177.

Other works cited in the text

Numbered in order of appearance in the text

1. Neidle, S (2008): *Principles of Nucleic Acid Structure*, first ed., Academic Press, London.

2. Calladine, CR & Drew, HR (1997): *Understanding DNA. The Molecule & How It Works*, second edition ed., Academic Press, London.

3. Saenger, W (1984): *Principles of nucleic acid structure*, Springer-Varlag.

4. Branden, C & Tooze, J (1999): *Introduction to Protein Structure*, second ed., Garland, New York.

5. Avery, OT, MacLeod, CM & McCarthy, M: 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. *J. Exp. Med.* **79**, 137-158 (1944).

6. Wilkins, MHF, Stokes, AR & Wilson, HR: Molecular structure of deoxypentose nucleic acids. *Nature* **171**, 738-740 (1953).

7. Franklin, RE & Gosling, RG: Molecular configuration in sodium thymonucleate. *Nature* **171**, 740-741 (1953).

8. Watson, JD & Crick, FHC: A structure for deoxyribose nucleic acid. *Nature* **171**, 737-738 (1953).

9. Rich, A: DNA comes in many forms. *Gene* **135**, 99-109 (1993).

10. Sussman, JL, Seeman, NC, Kim, S-H & Berman, HM: The crystal structure of a naturally occurring dinucleotide phosphate uridylyl 3',5'-adenosine phosphate. models for RNA chain folding. *J.Mol.Biol.* **66**, 403-421 (1972).

11. Rubin, J, Brennan, T & Sundaralingam, M: Crystal and molecular structure of a naturally occurring dinucleoside monophosphate uridylyl 3',5' - adenosine hemihydrate. Conformational "rigidity" of the nucleotide unit and models for polynucleotide chain folding. *Biochemistry* **11**, 3112-3128 (1972).

12. Kim, S-H, Suddath, FL, Quigley, GJ, McPherson, A, Sussman, JL, Wang, AH-J, Seeman, NC & Rich, A: Three-dimensional tertiary structure of yeast phenylalanine transfer RNA. *Science* **185**, 435-440 (1974).

13. Robertus, JD, Ladner, JE, Finch, JT, Rhodes, D, Brown, RS, Clark, BFC & Klug, A: Structure of yeast phenylalanine tRNA at 3 A resolution. *Nature* **250**, 546-551 (1974).

14. Sussman, JL, Holbrook, SR, Warrant, RW & Kim, S-H: Crystal structure of yeast phenylalanine transfer RNA. I. Crystallographic refinement. *J.Mol.Biol.* **123**, 607-630 (1978).

15. Gesteland, RF, Cech, TR & Atkins, JF (1999): *The RNA World*, second ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor.

16. Cech, T, Zaug, A & Grabowski, P: In vitro splicing of the ribosomal RNA precursor of Tetrahymena: involvement of a guanosine nucleotide in the excision of the intervening sequence. *Cell* **27**, 487-496 (1981).

17. Pley, HW, Flaherty, KM & McKay, DB: Three-dimensional structure of a hammerhead ribozyme. *Nature* **372**, 68-74 (1994).

18. Napoli, C, Lemieux, C & Jorgensen, R: Introduction of a Chimeric Chalcone Synthase Gene into Petunia Results in Reversible Co-Suppression of Homologous Genes in trans. *Plant Cell* **2**, 279-289 (1990).

19. van der Krol, AR, Mur, LA, Beld, M, Mol, JN & Stuitje, AR: Flavonoid genes in petunia: addition of a limited number of gene copies may lead to a suppression of gene expression. *Plant Cell* **2**, 291-299 (1990).

20. Fire, A, Xu, S, Montgomery, MK, Kostas, SA, Driver, SE & Mello, CC: Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature* **391**, 806-811 (1998).

21. Ban, N, Nissen, P, Hansen, J, Moore, PB & Steitz, TA: The complete atomic structure of the large ribosomal subunit at a 2.4 Å resolution. *Science* **289**, 905-920 (2000).

22. Leontis, NB & Westhof, E: Geometric nomenclature and classification of RNA base pairs. *RNA* **7**, 499-512 (2001).

23. Olson, WK, Bansal, M, Burley, SK, Dickerson, RE, Gerstein, M, Harvey, SC, Heinemann, U, Lu, X-J, Neidle, S, Shakked, Z, Sklenar, H, Suzuki, M, Tung, C-S, Westhof, E, Wolberger, C & Berman, HM: A standard reference frame for the description of nucleic acid base-pair geometry. *J.Mol.Biol.* **313**, 229-237 (2001).

24. Lu, XJ & Olson, WK: 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. *Nat.Protoc.* **3**, 1213-1227 (2008).

25. Taylor, R & Kennard, O: The molecular structures of nucleosides and nucleotides part I. *J. Mol. Struct.* **78**, 1-28 (1982).

26. Taylor, R & Kennard, O: Molecular Structures of Nucleosides and Nucleotides. 2. Orthogonal Coordinates for Standard Nucleic Acid Base Residues. *J. Am. Chem. Soc.* **104**, 3209-3212 (1982).

27. Allen, FH, Davies, JE, Galloy, JJ, Johnson, O, Kennard, O, Macrae, CF, Mitchell, EM, Mitchell, GF, Smith, JM & Watson, DG: The development of versions 3 and 4 of the Cambridge Structural Database System. *J. Chem. Inf. Comp. Sci.* **31**, 187-204 (1991).

28. Chandrasekaran, R & Arnott, S (1989): *Landolt-Börnstein Numerical Data and Functional Relationships in Science and Technology, Group VII/1b, Nucleic Acids*, Springer-Verlag, Berlin.

29. Drew, HR, Wing, RM, Takano, T, Broka, C, Tanaka, S, Itakura, K & Dickerson, RE: Structure of a B-DNA dodecamer: conformation and dynamics. *Proc.Natl.Acad.Sci.USA* **78**, 2179-2183 (1981).

30. Guzikevich-Guerstein, G & Shakked, Z: A novel form of the DNA double helix imposed on the TATA-box by the TATA-binding protein. *Nat.Struct.Biol.* **3**, 32-37 (1996).

31. Lu, X-J, Shakked, Z & Olson, WK: A-form conformational motifs in ligand-bound DNA structures. *J.Mol.Biol.* **300**, 819-840 (2000).

32. Wang, AH-J, Quigley, GJ, Kolpak, FJ, Crawford, JL, van Boom, JH, van der Marel, GA & Rich, A: Molecular structure of a left-handed double helical DNA fragment at atomic resolution. *Nature* **282**, 680-686 (1979).

33. Schwartz, T, Behlke, J, Lowenhaupt, K, Heinemann, U & Rich, A: Structure of the DLM-1-Z-DNA complex reveals a conserved family of Z-DNA-binding proteins. *Nat Struct Biol* **8**, 761 (2001).

34. Schultze, P, Smith, FW & Feigon, J: Refined solution structure of the dimeric quadruplex formed from the *Oxytricha* telomeric oligonucleotide d(GGGGTTTTGGGG). *Structure* **2**, 221-233 (1994).

35. Haider, SM, Parkinson, GN & Neidle, S: Structure of a G-quadruplex-ligand complex. *J.Mol.Biol.* **326**, 117-125 (2003).

36. Neidle, S & Parkinson, GN: The structure of telomeric DNA. *Curr Opin Struct Biol* **13**, 275-283 (2003).

37. Haider, S, Parkinson, GN & Neidle, S: Crystal structure of the potassium form of an Oxytricha nova G-quadruplex. *J.Mol.Biol.* **320**, 189-200 (2002).

38. Chen, L, Cai, L, Zhang, X & Rich, A: Crystal Structure of a Four-Stranded Intercalated DNA: d(C4). *Biochemistry* **33**, 13540-13546 (1994).

39. Vargason, JM & Ho, PS: The effect of cytosine methylation on the structure and geometry of the Holliday junction - The structure of d(CCGGTACm(5)CGG) at 1.5 A resolution. *J.Biol.Chem.* **277**, 21041-21049 (2002).

40. Thorpe, JH, Gale, BC, Teixeira, SC & Cardin, CJ: Conformational and hydration effects of site-selective sodium, calcium and strontium ion binding to the DNA Holliday junction structure d(TCGGTACCGA)(4). *J.Mol.Biol.* **327**, 97-109 (2003).

41. Gopaul, DN, Guo, F & Van Duyne, GD: Structure of the Holliday junction intermediate in Cre-loxP site-specific recombination. *EMBO J.* **17**, 4175-4187 (1998).

42. Neidle, S, Berman, H & Shieh, HS: Highly structured water network in crystals of a deoxydinucleoside-drug complex. *Nature* **288**, 129-133 (1980).

43. Kennard, O, Cruse, WBT, Nachman, J, Prange, T, Shakked, Z & Rabinovich, D: Ordered water structure in an A-DNA octamer at 1.7 A resolution. *J.Biomol.Struct.Dyn.* **3**, 623-647 (1986).

44. Gessner, RV, Quigley, GJ & Egli, M: Comparative studies of high resolution Z-DNA crystal structures.1. Common hydration patterns of alternating DC-DG. *J. Mol. Biol.* **236**, 1154-1168 (1994).

45. Gessner, RV, Frederick, CA, Quigley, GJ, Rich, A & Wang, AH-J: The molecular structure of the left-handed Z-DNA double helix at 1.0-Å atomic resolution. *J.Biol.Chem.* **264**, 7921-7935 (1989).

46. Egli, M, Williams, LD, Gao, Q & Rich, A: Structure of the Pure-Spermine Form of Z-DNA (Magnesium Free) at 1 Angstrom Resolution. *Biochemistry* **30**, 11388-11402 (1991).

47. Olson, WK (1982): Theoretical Studies of Nucleic Acid Conformation: Potential Energies, Chain Statistics, and Model Building. in *Topics in Nucleic Acid Structures, Part 2* (Neidle, S ed.), Macmillan Press, London. pp 1-79.

48. Kim, S-H, Berman, HM, Newton, MD & Seeman, NC: Seven basic conformations of nucleic acid structural units. *Acta Cryst. B* **29**, 703-710 (1973).

49. Yathindra, N & Sundaralingam, SM: Correlation Between the Backbone and Side Chain Conformations in 5'-Nucleotides. The Concept of a "Rigid" Nucleotide Conformation. *Biopolymers* **12**, 297-314 (1973).

50. McCall, M, Brown, T & Kennard, O: The crystal structure of d(G-G-G-G-C-C-C-C). A model for poly(dG).poly(dC). *J.Mol.Biol.* **183**, 385-396 (1985).

51. Liu, J & Subirana, JA: Structure of d(CGCGAATTCGCG) in the presence of Ca(2+) ions. *J.Biol.Chem.* **274**, 24749-24752 (1999).

52. Wang, JH: The hydration of desoxyribonucleic acid. *J. Am. Chem. Soc.* **77**, 258-260 (1955).

53. Falk, M, Poole, AG & Goymour, CG: Infrared study of the state of water in the hydration shell of DNA. *Can. J. Biochem.* **48**, 1536-1542 (1970).

54. Cheng, YK & Pettitt, BM: Stabilities of double- and triple-strand helical nucleic acids. *Prog.Biophys.Mol.Biol.* **58**, 225-257 (1992).

55. Chalikian, TV, Sarvazyan, AP & Breslauer, KJ: Hydration and Partial Compressibility of Biological Compounds. *Biophys.Chem.* **51**, 89-109 (1994).

56. Plum, GE & Breslauer, KJ: Calorimetry of proteins and nucleic acids. *Curr.Opin.Struct.Biol.* **5**, 682-690 (1995).

57. Tao, NJ, Rupprecht, SML & Rupprecht, A: The dynamics of the DNA hydration shell at gigahertz frequencies. *Biopolymers* **26**, 171-188 (1987).

58. Lavalle, N, Lee, SA & Rupprecht, A: Counterion effects on the physical properties and the A-transition to B-transition of calf-thymus DNA films. *Biopolymers* **30**, 877-887 (1990).

59. Edwards, GS, Davis, CC, Saffer, JD & Swicord, ML: Resonant microwave absorption of selected DNA molecules. *Phys. Rev. Lett.* **53**, 1284-1287 (1984).

60. Milton, JG & Galley, WC: Evidence for heterogeneity in DNA-associated solvent mobility from acridine phosphorescence spectra. *Biopolymers* **25**, 1673-1684 (1986).

61. Westhof, E (1993): *Water and biological macromolecules*, CRC Press, Boca Raton.

62. Wüthrich, K: Hydration of biological macromolecules in solution: surface structure and molecular recognition. *Cold Spring Harbor Symposia on Quantitative Biology* **58**, 149-157 (1993).

63. Hummer, G, Garcia, AE & Soumpasis, DM: Hydration of nucleic acid fragments: comparison of theory and experiment for high-resolution crystal structures of RNA, DNA, and DNA-drug complexes. *Biophys. J.* **68**, 1639-1652 (1995).

64. Jayaram, B & Beveridge, DL: Modeling DNA in aqueous solutions: Theoretical and computer simulation studies on the ion atmosphere of DNA. *Annu. Rev. Biophys. Biomol. Struct.* **25**, 367-394 (1996).

65. Auffinger, P & Hashem, Y: Nucleic acid solvation: from outside to insight. *Curr.Opin.Struct.Biol.* **17**, 325-333 (2007).

66. Denisov, VP, Carlström, G, Venu, K & Halle, B: Kinetics of DNA hydration. *J. Mol. Biol.* **268**, 118-136 (1997).

67. Seeman, NC, Rosenberg, JM & Rich, A: Sequence specific recognition of double helical nucleic acids by proteins. *Proc.Natl.Acad.Sci. USA.* **73**, 804-808 (1976).

68. Wolberger, C, Vershon, AK, Liu, B, Johnson, AD & Pabo, CO: Crystal Structure of a MATα2 Homeodomain-Operator Complex Suggests a General Model for Homeodomain-DNA Interactions. *Cell* **67**, 517-528 (1991).

69. Pullman, B, Pullman, A, Berthod, H & Gresh, N: Quantum-mechanical studies of environmental effects on biomolecules. IV. *Ab initio* studies on the hydration scheme of the phosphate group. *Theor. Chim. Acta.* **40**, 90-111 (1975).

70. Langlet, J, Claverie, P, Pullman, B & Piazzola, D: Studies of solvent effects. IV. study of hyrdration of the dimethyl phosphate anion (DMP-) and the solvent effect upon its conformation. *Int. J. Quant. Chem.* **6**, 409-437 (1979).

71. Saenger, W, Hunter, WN & Kennard, O: DNA conformation is determined by economics in the hydration of phosphate groups. *Nature* **324**, 385-388 (1986).

72. Langan, P, Forsyth, VT, Mahendrasingam, A, Pigram, WJ, Mason, SA & Fuller, W: A high angle neutron fibre diffraction study of the hydration of the A conformation of the DNA double helix. *J. Biomol. Struct. Dyn.* **10**, 489-503 (1992).

73. Schreiner, LJ, Pintar, MM, Dianoux, AJ, Volino, F & Rupprecht, A: Hydration of NaDNA by neutron quasi-elastic scattering. *Biophys. J.* **53**, 119-122 (1988).

74. Schreiner, LJ, Mactavish, JC, Pintar, MM & Rupprecht, A: NMR spin grouping and correlation exchange analysis - application to low hydration NaDNA paracrystals. *Biophys. J.* **59**, 221-234 (1991).

75. Zandt, LLV: Why Structured Water Causes Sharp Absorption by DNA at Microwave Frequencies. *J. Biomol. Struct. Dyn.* **4**, 569-582 (1987).

76. Leikin, S, Parsegian, VA & Rau, DC: Hydration forces. *Annu. Rev. Phys. Chem.* **44**, 369-395 (1993).

77. Marcelja, S & Radic, N: Repulsion of interfaces due to boundary water. *Chem. Phys. Lett.* **42**, 129-130 (1976).

78. Rau, DC, Lee, B & Parsegian, VA: Measurement of the repulsive force between polyelectrolyte molecules in ionic solution: hydration forces between parallel DNA double helices. *Proc. Natl. Acad. Sci. U.S.A.* **81**, 2621-2625 (1984).

79. Rau, DC & Parsegian, VA: Direct measurement of the intermolecular forces between counterion-condensed DNA double helices. *Biophys.J.* **61**, 246-259 (1992).

80. Rau, DC & Parsegian, VA: Direct measurement of temperature-dependent solvation forces between DNA double helices. *Biophys. J.* **61**, 260-271 (1992).

81. Leontis, NB, Stombaugh, J & Westhof, E: The non-Watson-Crick base pairs and their associated isostericity matrices. *Nucleic Acids Res.* **30**, 3497-3531 (2002).

82. Klein, DJ, Schmeing, TM, Moore, PB & Steitz, TA: The Kink-Turn: A new RNA secondary structure motif. *EMBO J.* **20**, 4214-4221 (2001).

83. Leontis, NB & Westhof, E: A common motif organizes the structure of multi-helix loops in 16S and 23S ribosomal RNAs. *J.Mol.Biol.* **283**, 571-583 (1998).

84. Cate, JH, Gooding, AR, Podell, E, Zhou, KH, Golden, BL, Kundrot, CE, Cech, TR & J.A., D: Crystal structure of a group I ribozyme domain: Principles of RNA packing. *Science* **273**, 1678-1685 (1996).

85. Wimberly, BT, Brodersen, DE, Clemons Jr., WM, Morgan-Warren, R, Carter, AP, Vonrhein, C, Hartsch, T & Ramakrishnan, V: Structure of the 30S ribosomal subunit. *Nature* **407**, 327-339 (2000).

86. Tocilj, A, Schlunzen, F, Janell, D, Gluhmann, M, Hansen, H, Harms, J, Bashan, A, Bartels, H, Agmon, I, Franceschi, F & Yonath, A: The small ribosomal subunit from *Termus termophilus* at 4.5 Å resolution: Pattern fittings and the identification of a functional site. *Proc. Nat. Acad. Sci. USA* **96**, 14252-14257 (1999).

87. Murray, LJ, Arendall 3rd, WB, Richardson, DC & Richardson, JS: RNA backbone is rotameric. *Proc.Natl.Acad.Sci.USA* **100**, 13904-13909 (2003).

88. Hershkovitz, E, Tannenbaum, E, Howerton, SB, Sheth, A, Tannenbaum, A & Williams, LD: Automated identification of RNA conformational motifs: theory and application to the HM LSU 23S rRNA. *Nucleic Acids Res.* **31**, 6249-6257 (2003).

89. Sims, GE & Kim, S-H: Global mapping of nucleic acid conformational space: dinucleoside monophosphate conformations and transition pathways among conformational classes. *Nucleic Acids Res.* **31**, 5607-5616 (2003).

90. Ramachandran, GN & Sasisekharan, V: Conformation of polypeptides and proteins. *Adv. Protein Chem.* **23**, 283-437 (1968).

91. Leontis, NB, Altman, RB, Berman, HM, Brenner, SE, Brown, JW, Engelke, DR, Harvey, SC, Holbrook, SR, Jossinet, F, Lewis, SE, Major, F, Mathews, DH, Richardson, JS, Williamson, JR & Westhof, E: The RNA Ontology Consortium: An open invitation to the RNA community. *RNA* **12**, 533-541 (2006).

92. Wang, Y & Patel, DJ: Solution structure of a parallel-stranded G-quadruplex DNA. *J.Mol.Biol.* **234**, 1171-1183 (1993).

93. Hargreaves, D, Rice, DW, Sedelnikova, SE, Artymiuk, PJ, Lloyd, RG & Rafferty, JB: Crystal structure of E.coli RuvA with bound DNA Holliday junction at 6 A resolution. *Nat.Struct.Biol.* **5**, 441-446 (1998).

94. Clowney, L, Jain, SC, Srinivasan, AR, Westbrook, J, Olson, WK & Berman, HM: Geometric Parameters In Nucleic Acids: Nitrogenous Bases. *J.Am.Chem.Soc.* **118**, 509-518 (1996).

95. Parkinson, G, Vojtechovsky, J, Clowney, L, Brunger, AT & Berman, HM: New parameters for the refinement of nucleic acid containing structures. *Acta Cryst. D* **52**, 57-64 (1996).

96. Brünger, AT: X-PLOR, version 3.1, a system for X-ray crystallography and NMR. 3.1 Ed., Yale University Press, New Haven, CT (1992).

97. Brünger, AT, Adams, PD, Clore, GM, DeLano, WL, Gros, P, Grosse-Kunstleve, RW, Jiang, J-S, Kuszewski, J, Nilges, M, Pannu, NS, Read, RJ, Rice, LM, Simonson, T & Warren, GL: Crystallographic and NMR system: A new software suite for macromolecular structure determination. *Acta Cryst. D* **54**, 905-921 (1998).

98. Drenth, J (1994): *Principles of Protein X-ray Crystallography*, Springer-Verlag, New York.

99. Rhodes, G (2006): *Crystallography Made Crystal Clear: A Guide for Users of Macromolecular Models*, Third ed., Academic Press.

100. Oldfield, TJ: Applications for macromolecular map interpretation: X-AUTOFIT, X-POWERFIT, X-BUILD, X-LIGAND, and X-SOLVATE. *Methods Enzymol.* **374**, 271-300 (2003).

101. Lamzin, VS & Perrakis, A: Current state of automated crystallographic data analysis. *Nat.Struct.Biol.* **7 Suppl**, 978-981 (2000).

102. Langer, G, Cohen, SX, Lamzin, VS & Perrakis, A: Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. *Nat.Protoc.* **3**, 1171-1179 (2008).

103. Pavelcik, F & Vanco, J: Simple procedure conformation-family search in multidimensional torsion-angle space. *J.Appl.Cryst.* **39**, 483-486 (2006).

104. Schwalbe, H, Marino, JP, King, GC, Wechselberger, R, Bermel, W & Griesinger, C: Determination of a Complete Set of Coupling-Constants in C-13-Labeled Oligonucleotides. *J.Biomol. NMR* **4**, 631-644 (1994).

105. Reif, B, Hennig, M & Griesinger, C: Direct measurement of angles between bond vectors in high-resolution NMR. *Science* **276**, 1230-1233 (1997).

106. Wijmenga, SS & van Buuren, BNM: The use of NMR methods for conformational studies of nucleic acids. *Prog.NMR Spectrosc.* **32**, 287-387 (1998).

107. Marino, JP, Schwalbe, H & Griesinger, C: J-coupling restraints in RNA structure determination. *Acc.Chem.Res.* **32**, 614-623 (1999).

108. Zidek, L, Stefl, R & Sklenar, V: NMR methodology for the study of nucleic acids. *Curr.Opin.Struct.Biol.* **11**, 275-281 (2001).

109. Qin, PZ & Dieckmann, T: Application of NMR and EPR methods to the study of RNA. *Curr.Opin.Struct.Biol.* **14**, 350-359 (2004).

110. Munzarova, ML & Sklenar, V: DFT analysis of NMR scalar interactions across the glycosidic bond in DNA. *J.Am.Chem.Soc.* **125**, 3649-3658 (2003).

111. Munzarova, ML & Sklenar, V: Three-bond sugar-base couplings in purine versus pyrimidine nucleosides: A DFT study of Karplus relationships for (3)J(C2/4-H1 '), and (3)J(C6/8-H1 ') in DNA. *J.Am.Chem.Soc.* **124**, 10666-10667 (2002).

112. Sychrovsky, V, Grafenstein, J & Cremer, D: Nuclear magnetic resonance spin-spin coupling constants from coupled perturbed density functional theory. *J.Chem.Phys.* **113**, 3530-3547 (2000).

113. Helgaker, T, Watson, M & Handy, NC: Analytical calculation of nuclear magnetic resonance indirect spin-spin coupling constants at the generalized gradient approximation and hybrid levels of density-functional theory. *J.Chem.Phys.* **113**, 9402-9409 (2000).

114. Karplus, M: Contact electron-spin coupling of nuclear magnetic moments. *J.Chem.Phys.* **30**, 11-15 (1959).

115. Allen, FH, Bellard, S, Brice, MD, Cartright, BA, Doubleday, A, Higgs, H, Hummelink, T, Hummelink-Peters, BG, Kennard, O, Motherwell, WDS, Rodgers, JR & Watson, DG: The Cambridge Crystallographic Data Centre: Computer-based search, retrieval, analysis and display of information. *Acta Cryst.* **B35**, 2331-2339 (1979).

116. Berman, HM, Westbrook, J, Feng, Z, Gilliland, G, Bhat, TN, Weissig, H, Shindyalov, IN & Bourne, PE: The Protein Data Bank. *Nucleic Acids Res.* **28**, 235-242 (2000).

117. Bernstein, FC, Koetzle, TF, Williams, GJB, Meyer Jr., EF, Brice, MD, Rodgers, JR, Kennard, O, Shimanouchi, T & Tasumi, M: Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535-542 (1977).

118. Fitzgerald, PMD, Westbrook, JD, Bourne, PE, McMahon, B, Watenpaugh, KD & Berman, HM (2006): Macromolecular dictionary (mmCIF). in *International Tables for Crystallography. Definition and exchange of crystallographic data* (Hall, SR & McMahon, B eds.), Springer, Dordrecht. pp 295-443.

119. Westbrook, JD & Fitzgerald, PMD (2009): The PDB Format, mmCIF Formats, and Other Data Formats. in *Structural Bioinformatics* (Gu, J & Bourne, PE eds.), Second Ed., Wiley-Blackwell, Hoboken. pp 271-291.

120. Berman, HM, Henrick, K & Nakamura, H: Announcing the worldwide Protein Data Bank. *Nat.Struct.Biol.* **10**, 980 (2003).

List of publications compiled in the thesis

Listed in the order as discussed in the text

I. Schneider B & Berman HM (2006) Basics of nucleic acids structure. In *Computational Studies of RNA and DNA,* (eds. Šponer J & Lankaš F), pp. 1-44, Springer, Dordrecht.

II. Schneider B, Neidle S & Berman HM (1997) Conformations of the sugar-phosphate backbone in helical DNA crystal structures. *Biopolymers* **42,** 113-124.

III. Svozil D, Kalina J, Omelka M & Schneider B (2008) DNA conformations and their sequence preferences. *Nucleic Acids Res.* **36,** 3690-3706.

IV. Schneider B, Ginell SL & Berman HM (1992a) Low temperature structures of dCpG-proflavine: Conformational and hydration effects. *Biophys. J.* **63**, 1572-1578.

V. Harper A, Brannigan JA, Buck, M, Hewitt L, Lewis RJ, Moore MH & Schneider B (1998) The structure of $d(TGCGCA)_2$ and a comparison to other DNA hexamers. *Acta Cryst. D* **54**, 1273-1284.

VI. Schneider B, Morávek Z & Berman HM (2004) RNA conformational classes. *Nucleic Acids Res.* **32,** 1666-1677.

VII. Richardson JS, Schneider B, Murray LW, Kapral GJ, Immormino RM, Headd JJ, Richardson DC, Ham D, Hershkovits E, Williams LD, Keating KS, Pyle AM, Micallef D, Westbrook J & Berman HM (2008) RNA backbone: Consensus all-angle conformers and modular string nomenclature (an RNA Ontology Consortium contribution). *RNA* **14,** 465-481.

VIII. Schneider B, Cohen DM, Schleifer L, Srinivasan AR, Olson WK & Berman HM (1993) A systematic method for studying the spatial distribution of water molecules around nucleic acid bases. *Biophys.J.* **65,** 2291-2303.

IX. Schneider B & Berman HM (1995) Hydration of the DNA bases is local. *Biophys. J.* **69,** 2661-2669.

X. Morávek Z, Neidle S & Schneider B (2002) Protein and drug interactions in the minor groove of DNA. *Nucleic Acids Res.* **30,** 1182-1191.

XI. Woda J, Schneider B, Patel K, Mistry K & Berman HM (1998) An analysis of the relationship between hydration and protein-DNA interactions. *Biophys. J.* **75**, 2170-2177.

XII. Ge W, Schneider B & Olson WKO (2005) Knowledge-based elastic potentials for docking drugs or proteins with nucleic acids**.** *Biophys.J.* **88**, 1166-1190.

XIII. Schneider B, Patel K & Berman HM (1998) Hydration of the phosphate group in double-helical DNA. *Biophys. J.* **75,** 2422-2434.

XIV. Schneider B & Kabeláč M (1998) Stereochemistry of binding of metal cations and water to a phosphate group. *J. Am. Chem. Soc.* **120**, 161-165.

XV. Pavelcik F & Schneider B (2008) Building of RNA and DNA double helices into electron density. *Acta Cryst. D* **64,** 620-626.

XVI. Sychrovský V, Vokáčová Z, Šponer J, Špačková N & Schneider B (2006b) Calculation of structural behavior of indirect NMR spin-spin couplings in the backbone of nucleic acids. *J. Phys. Chem. B* **110,** 22894-22902.

XVII. Berman HM, Olson WK, Beveridge DL, Westbrook J, Gelbin A, Demeny T, Hsieh S-H, Srinivasan AR & Schneider B (1992) The Nucleic Acid Database - A comprehensive relational database of three-dimensional structures of nucleic acids. *Biophys.J.* **63,** 751-759.

XVIII. Berman HM, Westbrook, J, Feng Z, Iype L, Schneider B & Zardecki C (2002b) The Nucleic Acid Database. *Acta Cryst. D* **58,** 889-898.

XIX. Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, Feng Z, Gilliland G L, Iype L, Jain S, Fagan P, Marvin J, Ravichanran V, Schneider B, Thanki N, Padilla D, Weissig H, Westbrook JD & Zardecki C (2002a). The Protein Data Bank. *Acta Cryst. D* **58,** 899-907.

XX. Burkhardt K, Schneider B & Ory J (2006) A biocurator perspective: Annotation at the Research Collaboratory for Structural Bioinformatics Protein Data Bank. *PLOS Comput Biol*. **2**, 1186-1189.