Charles University

Faculty of Science

Study programme: Modeling of Chemical Properties of Nano- and Biostructures

Branch of study: 4XMOCHEV

Mgr. Dávid Jakubec

Interactions of Proteins with Nucleic Acids: from Structure to Specificity

Interakce Proteinů s Nukleovými Kyselinami: od Struktury k Specificitě

Doctoral thesis

Supervisor: doc. RNDr. Jiří Vondrášek, CSc.

Prague, 2020

Prohlašuji, že jsem tuto disertační práci vypracoval samostatně a že všechny použité zdroje a literatura byly řádně citovány. Tato práce nebyla využita jako závěrečná práce k získání jiného nebo obdobného druhu vysokoškolské kvalifikace.

V .... dne ....

Podpis autora

Na tomto mieste by som rád poďakoval predovšetkým svojmu školiteľovi doc. RNDr. Jiřímu Vondráškovi, CSc. za pretrvávajúcu podporu a vzdelávanie počas môjho štúdia, ďalej ostatným členom skupiny Bioinformatika na ÚOCHB AV ČR za nespočetné plodné diskusie a tvorbu príjemných pracovných podmienok, a v neposlednej rade i svojim priateľom a rodine za motiváciu a trpezlivosť.

Abstract

Sequence-specific interactions between proteins and nucleic acids play an essential role in the cell biology. While several molecular mechanisms contributing to the binding specificity have been identified empirically, no general protein–DNA recognition code has been described to date. In this thesis, I explore selected characteristics of protein–DNA interactions using computational methods. First, the pairwise interactions between the basic biomolecular building blocks—amino acids and nucleotides—are investigated. It is shown that several statistically enriched, biologically relevant interaction motifs correspond to the most energetically favorable configurations of the respective binding partners. In addition, a relationship between the physicochemical properties of the amino acid residues found at the protein–DNA interface and the local geometric features of the DNA helix is presented. Next, the applicability of molecular dynamics-based setups to the description of binding equilibria in protein–DNA systems is investigated. Discrepancies are observed between the description offered by the computer simulations and experimental results, as well as between the results obtained using two molecular mechanical force fields. Finally, the more general evolutionary aspects of protein organization are explored, and a tool for the study of evolutionary conservation is introduced.

Abstrakt

Sekvenčně-specifické interakce mezi proteiny a nukleovými kyselinami mají zásadní roli v biologii buňky. I když několik molekulárních mechanismů podílejících se na vazebné specificitě bylo empiricky vypozorováno, žádný obecný rozpoznávací kód pro protein–DNA interakce nebyl zatím popsán. V této disertační práci prozkoumávám vybrané charakteristiky protein–DNA interakcí pomocí výpočetních metod. Nejdřív jsou studovány párové interakce mezi základními stavebními prvky biomolekul—aminokyselinami a nukleotidy. Je ukázáno, že některé statisticky nabohacené, biologicky relevantní interakční motivy odpovídají energeticky nejvýhodnějším geometrickým uspořádáním daných vazebných partnerů. Dále je demonstrován vztah mezi fyzikálně-chemickými vlastnostmi aminokyselin nacházejících se na rozhraní proteinu s DNA a lokálními topologickými charakteristikami DNA dvoušroubovice. V další části je prozkoumána využitelnost postupů založených na molekulové dynamice při popisu vazebných rovnováh v systémech protein–DNA. Jsou pozorovány rozdíly nejen mezi popisem získaným na základě počítačových simulací a experimentálními výsledky, ale i mezi výsledky získanými s využitím dvou různých molekulárně mechanických silových polí. V závěru jsou prozkoumány obecnější evoluční charakteristiky struktury proteinů a je představen nástroj pro studium evoluční konzervace.

Contents

Introduction

Function and structure of biomacromolecules

Biological roles of proteins and nucleic acids

Proteins and nucleic acids constitute the fundamental molecules common to all life forms on Earth. Proteins, or polypeptides, are polymers of amino acid residues joined by peptide bonds. Individual amino acids have unique physico-chemical properties (Figure A.1) which give proteins with different amino acid compositions distinct characteristics. In organisms, proteins primarily fulfill enzymatic, structural, and other complex (*e.g.*, signaling) roles. The nucleic acids—deoxyribonucleic acid (DNA) and ribonucleic acid (RNA)—are polymers of (deoxy)ribonucleoside monophosphates [(d)NMPs; Figure A.2] joined by phosphodiester bonds. The primary role of DNA is the storage of "genetic information" in the form of its sequence of nucleotides. In designated regions of the organisms' genomes (protein-coding genes), these sequences encode the primary structures of polypeptide chains (see below). In comparison with DNA, the biological roles of RNA are much more diverse, and include the transport of the information stored in the genome for translation into proteins, enzymatic activities, regulation of gene expression, and other functions (*e.g.*, serving as the primary carrier of genetic information in RNA viruses).[1]

Structural organization of proteins and nucleic acids

In a cellular environment, these biomolecules can adopt complex three-dimensional (3D) shapes which enable them to carry out their biological functions. Examples of protein and nucleic acid 3D structures are shown in Figures 1 and 2, respectively. A hierarchy of structural features exists for both proteins and nucleic acids. The linear ordering of the monomers (amino acid residues or nucleotides) constituting the macromolecule is referred to as its primary structure, or simply sequence. In proteins, the various intra- and intermolecular interactions (see below) can lead to the formation of a few distinct local structural motifs (secondary structure). The major secondary structural elements formed by polypeptide chains are the α-helix and the β-strand. These and other local structural arrangements (*e.g.*, turns, loops, ...) combine to form the tertiary structure of a protein. In many cases, the relative orientations and connectivity of the secondary structural elements in a tertiary structure form a recognizable 3D pattern known as the protein fold. As of April 2020, around $1,400$ distinct protein folds have been recognized according to databases such as CATH[4] and SCOP.[5,6] A domain is a unit of protein structure with a well-defined fold and sequential features which is often able to carry out its biological function independently of its sequential surroundings. Most eukaryotic proteins contain multiple domains.[7,8] Finally, the quaternary structure describes the arrangement of the polypeptide chain subunits in multiprotein assemblies.[9]

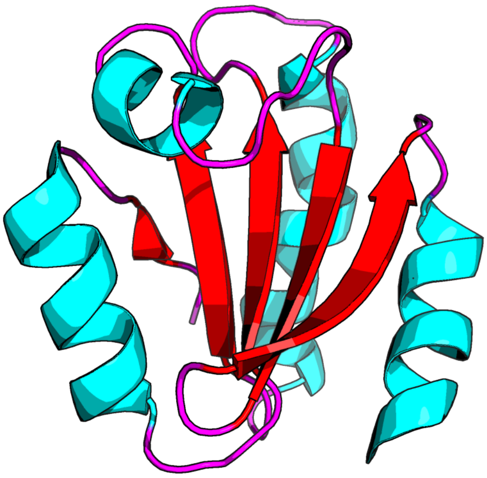


Figure 1: The solution structure of human thioredoxin [Protein Data Bank (PDB) ID 3TRX] depicted in a cartoon representation. α-helices are shown in cyan, β-strands are shown in red, and loops are shown in magenta. Visualized using PyMOL-2.3.0.[2]

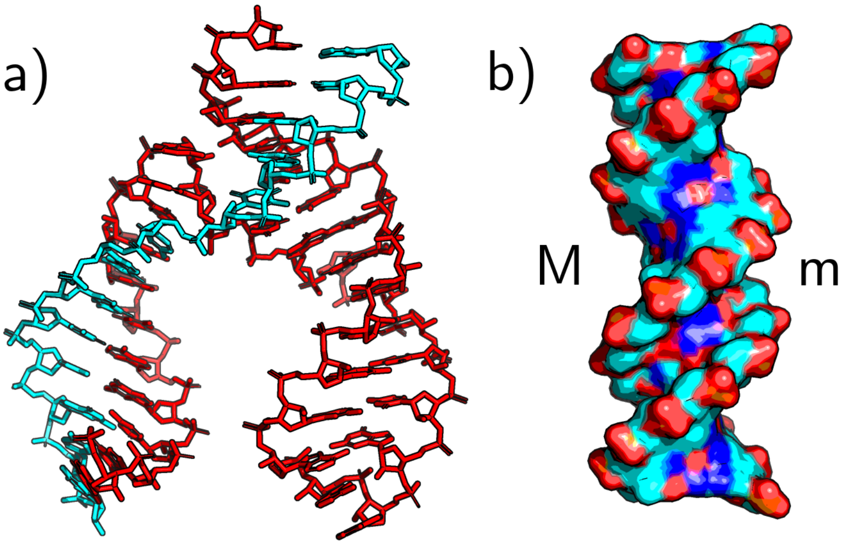


Figure 2: Examples of nucleic acid 3D structures. a) The X-ray crystallographic structure of a hammerhead RNA ribozyme–DNA inhibitor complex at 2.6 Å resolution (PDB ID 1HMH). The RNA and DNA strands are shown in red and cyan, respectively. b) A model of an ideal B-form DNA prepared using 3DNA v2.3-2016apr02.[3] "M" and "m" denote the major and minor grooves, respectively. Visualized using PyMOL-2.3.0.[2]

The major 3D structural motifs adopted by nucleic acids are helices and loops. These structural elements are primarily characterized by the patterns of hydrogen bonding and stacking interactions between the various chemical moieties of the nucleotides involved. Such interactions can occur within individual nucleic acid strands or between the functional groups belonging to multiple strands. In living cells, DNA can be found almost exclusively in the form of a right-handed double helix. The structure of the double-stranded DNA (dsDNA) features two prominent grooves (Figure 2b) which expose different functional groups of the nucleotides to the environment (see below). In contrast with DNA, the 3D structures adopted by RNA in biological contexts can be much more complex, reflecting the much greater functional diversity of the latter.[1,10]

Experimental methods of biomolecular structure determination

X-ray crystallography

The 3D structures of biomolecules can be determined experimentally using a range of methods. The most widely used approaches to obtaining atomic-level views of proteins and nucleic acids are X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and electron microscopy. Each of these methods has its own set of pros and cons which make it advantageous for the study of certain types of biomolecular systems and limit its applicability for others. X-ray crystallography is based on the diffraction of X-ray radiation by the electrons of molecules forming crystals. The 3D structures of the molecules can then be determined in atomic-level resolution by the mathematical manipulation of the recorded diffraction patterns. The main problems of this method are the difficulty of obtaining biomolecular crystals of sufficient size and quality and possible changes in the biomolecular structures resulting from the constraints of the crystal environment. In addition, the structures of the flexible regions of the biomolecules can often not be resolved using this method. Nevertheless, X-ray crystallography has been considered the most accurate method for the determination of 3D structures of proteins and nucleic acids and remains the most widely used (see below).[9]

Nuclear magnetic resonance spectroscopy

When atoms with nonzero nuclear spins are placed in strong external magnetic fields, their nuclear magnetic moments align either parallel or antiparallel to the field lines. The energy difference between these two states is influenced by the local chemical environment of each atom and generally falls into the radio wave range of electromagnetic radiation. In NMR spectroscopy, molecules containing nuclei with nonzero nuclear spins interact with such electromagnetic radiation and their 3D structures can be deduced from the differences between the measured and reference NMR spectra of the nuclei involved (chemical shifts). The main advantages of this method in comparison with X-ray crystallography are the possibility to study the biomolecular structures in solution (as opposed to the artificial crystal environment), as well as the ability to describe structural dynamics and flexible regions of the biomolecules. The limitations of NMR spectroscopy include the necessity to prepare high concentration solutions of the biomolecules and the complexity of the NMR spectra, which makes the determination of the 3D structures of large biomolecules impractical. In addition, while X-ray crystallography yields actual electron density maps, only sets of structural constraints are obtained from an NMR experiment. These constraints are used to construct the 3D models of the studied molecules with the aid of computational tools, which may be a source of other problems (see below).[9]

Electron microscopy

Transmission electron microscopy (TEM) is an emerging method of biomolecular structure determination. In a TEM experiment, individual macromolecules can be directly visualized using a focused electron beam passing through the sample. The 3D structures of the molecules can be constructed by computationally processing a large number of images of the molecules in various orientations. With the introduction of cryogenic cooling of the sample, it has become possible to use electron beams powerful enough to enable the determination of biomolecular structures in atomic-level resolution. The rapid freezing encloses the biomolecules in a layer of amorphous ice which resembles their native biological environment much more closely than the crystal structures required in X-ray crystallography. In addition, it enables the study of structural dynamics and conformational transitions in the investigated biomolecules, as multiple conformational states may be captured. The main disadvantages of this method have traditionally been the cost of the instrumentation and lower resolution of the 3D structures in comparison with X-ray crystallography; however, recent advances in the field have diminished the latter.[9]

Protein Data Bank

The Protein Data Bank (PDB) archive, managed by the Worldwide PDB (wwPDB) organization, is a single repository of experimentally determined 3D structures of biomolecules.[11] The PDB archive is typically accessed through the websites of the wwPDB member organizations, such as the Research Collaboratory for Structural Bioinformatics PDB (RCSB PDB; `https://www.rcsb.org/`)[12] or the PDB in Europe (`https://www.ebi.ac.uk/pdbe/`).[13] These websites provide access to the 3D structural data in computer-readable formats, as well as additional services, such as analysis and annotation of the structures.

As of April 2020, over 163,000 biomolecular 3D structures have been deposited in the PDB archive (Table 1). The majority of these structures (over 145,000) have been resolved using X-ray crystallography; the structures determined using solution NMR spectroscopy and electron microscopy count almost 13,000 and 4,800, respectively. Fewer than 500 structures have been determined using other methods. Almost 160,000 structures contain at least one polypeptide chain; almost 12,000 structures contain at least one nucleic acid strand; over 8,400 structures contain both proteins and nucleic acids.

Table 1: The numbers of biomolecular 3D structures containing various polymer types resolved using the experimental methods discussed in the text. Data obtained from the April 29, 2020, release of the RCSB PDB; the release contained a total of 163,414 structures. Table headers are adopted from the RCSB PDB. “Entry Polymer Types” “Protein (only)” and “Nucleic acid (only)” are associated with the structures containing exclusively the respective biopolymer; “Protein/NA” refers to the structures containing both polypeptide and nucleic acid chains. The 3D structures determined using other experimental methods or those with other “Entry Polymer Types” are not shown.

Experimental Method	Entry Polymer Types	Count
X-RAY DIFFRACTION	Protein (only)	136,217
	Nucleic acid (only)	2,047
	Protein/NA	6,592
SOLUTION NMR	Protein (only)	11,293
	Nucleic acid (only)	1,289
	Protein/NA	266
ELECTRON MICROSCOPY	Protein (only)	3,585
	Nucleic acid (only)	36
	Protein/NA	1,053

The disparity among these numbers stems from different levels of scientific interest in the respective biopolymers, as well as from different states of development of the individual experimental methods.[9]

Protein–nucleic acid interactions

Biological significance of protein–nucleic acid interactions

Interactions between selected proteins and nucleic acids are essential for several vital cellular processes. In some cases, proteins and nucleic acids form stable nucleoprotein complexes in which the biopolymers cooperate to perform the intended function. For example, ribosomes, which facilitate the synthesis of new polypeptide chains according to messenger RNA (mRNA) templates, are complex biomacromolecular assemblies consisting of multiple protein and ribosomal RNA (rRNA) molecules. Other examples of protein–RNA complexes (ribonucleoproteins) are the spliceosomes, which excise introns from precursor mRNA molecules in the nuclei of eukaryotic cells, and the signal recognition particles, which bind to designated signal sequences in newly synthesized proteins and direct them toward their target locations. The most prominent examples of stable protein–DNA assemblies are the nucleosomes, which package genomic DNA in the nuclei of eukaryotic cells.[1]

In other instances, proteins and nucleic acids interact only transiently as a response to some stimulus or to achieve a specific interim goal. For example, DNA repair enzymes recognize and repair DNA loci damaged by, *e.g.*, oxidative stress or ultraviolet radiation; genomic DNA interacts with the replication machinery during DNA replication; and transcription factor

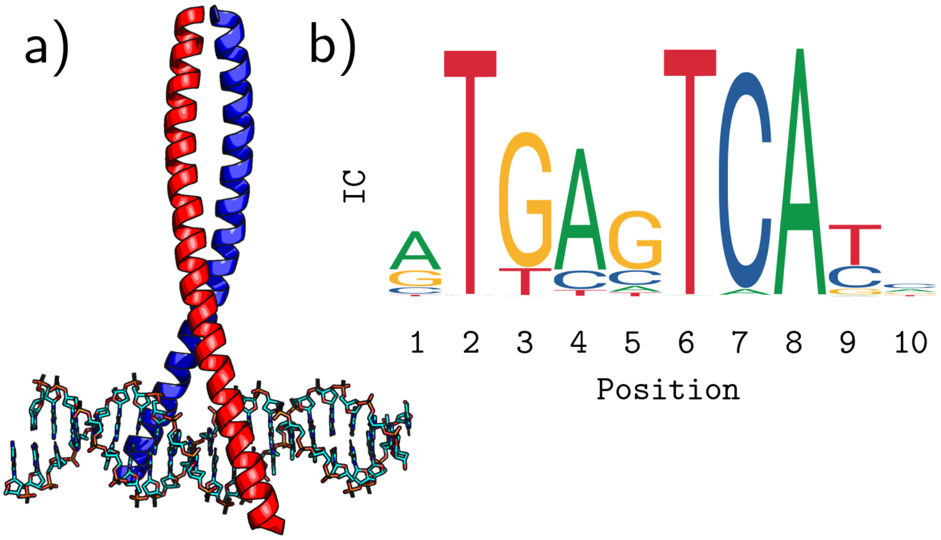


Figure 3: Example of a TF and a model of its DNA binding site. a) The X-ray crystallographic structure of the bZIP domains of human TFs c-Fos (red) and c-Jun (blue) bound to a dsDNA containing an AP-1 site (PDB ID `1FOS`). Visualized using PyMOL-2.3.0.[2] b) Sequence logo representing the sequence composition of the DNA binding sites bound by the human c-Fos–c-Jun TF (JASPAR database matrix profile `MA0099.3`). Letter heights at individual positions in the logo are proportional to the frequencies of the respective nucleotides at that position. Total heights of the letter stacks represent the information content (IC) of the positions.

(TF; Figure 3a) proteins interact with regulatory regions in the organisms' genomes and thus modulate gene expression.[1,14]

Specificity of protein–nucleic acid interactions

Biological significance of binding specificity

It is necessary for the survival of the cells that some of the described processes—such as DNA repair and replication—proceed regardless of the sequence of the nucleic acid involved. On the other hand, interactions of TF and other proteins with precise genomic regions are required for, *e.g.*, the regulation of expression of the respective genes.[1,14] The preferential binding of a protein to a well-defined set of nucleotide sequences is referred to as its binding specificity. In the rest of this text, I will primarily discuss means *via* which DNA-binding specificity of proteins can be studied and factors which are assumed to contribute to it. Many of these methods and results are applicable to protein–RNA interactions as well; however, as these are not the main focus of this work, their specifics are not elaborated on.

Experimental determination of DNA-binding specificity

A dsDNA molecule containing N base pairs (bps) offers 4^N sequence options. If the binding reaction between a DNA-binding protein (DBP) and a particular DNA binding site can be written as protein + DNA $\rightleftharpoons$ protein–DNA, *i.e.*, one protein molecule interacts with one DNA molecule to form a protein–DNA complex, then the affinity of the protein for the binding site can be characterized using the thermodynamic equilibrium constant K as

$$K = \frac{a_{\text{protein}-\text{DNA}}}{a_{\text{protein}}a_{\text{DNA}}} \tag{1}$$

where a_{X} is the activity of species X at equilibrium; equivalently, the affinity can be expressed using the standard Gibbs free energy of reaction $\Delta_{\text{r}}G^{\ominus}$ as

$$\Delta_{\text{r}}G^{\ominus} = \Delta_{\text{r}}H^{\ominus} - T\Delta_{\text{r}}S^{\ominus} = -RT\log K \tag{2}$$

where $\Delta_{\text{r}}H^{\ominus}$ and $\Delta_{\text{r}}S^{\ominus}$ are the standard enthalpy and entropy of reaction, respectively, T is the thermodynamic temperature, R is the universal gas constant, and K is the thermodynamic equilibrium constant. These quantities can be accessed experimentally using, for example, the techniques of isothermal titration calorimetry,[15,16] fluorescence anisotropy titration,[17,18] or microscale thermophoresis.[19,20] The binding specificity of a DBP can be expressed quantitatively in terms of equilibrium constants characterizing the interactions between the protein and all of its possible DNA binding sites;[21] however, determination of the required thermodynamic descriptors on such a scale is usually infeasible using these experimental methods (see below).

The most frequent DNA binding site length among human TFs is $N =$ 10 bps[1];[22] this results in $4^{10} \approx 10^{6}$ sequence options. Unfortunately, the biophysical methods mentioned above offer only very limited throughput, making them impractical for the assessment of affinities of DBPs toward all possible DNA binding sites on such scales. For this purpose, high-throughput (HT) methods, such as protein-binding microarrays (PBMs),[23,24] HT systematic evolution of ligands by exponential enrichment,[22,24] and chromatin immunoprecipitation followed by sequencing (ChIP–seq),[22,25] can be used. For example, PBMs allow the *in vitro* detection of protein binding to dsDNA probes covering all nucleotide sequences of length 10 bps (10-mers), while ChIP–seq enables the identification of genomic regions bound by the DBP of interest *in vivo*.

The HT methods do not typically yield the thermodynamic characteristics of the investigated protein–DNA interactions; instead, they reveal the sequence composition of the DNA binding sites bound by the particular DBP. This information can be used to construct statistical models of the binding sites of individual DBPs. These models, which can be represented graphically using, *e.g.*, sequence logos (Figure 3b),[26] are available for thousands of TFs online in databases such as JASPAR[27] and HOCOMOCO.[28] Multiple sequence alignments (MSAs) of the binding site sequences allow an alternative definition of binding specificity in terms of information content (IC) of individual positions in the alignment.[21,29–31]

Molecular determinants of DNA-binding specificity

While the biophysical and sequencing methods described above yield information about the thermodynamics of protein–DNA interactions or show

[1] Most human TFs, however, specifically recognize binding sites longer than 10 bps.[22]

the large-scale binding preferences of DBPs, they do not directly reveal the mechanistic factors determining the binding specificity. However, knowledge of such factors is highly important due to the central role TFs and other sequence-specific DBPs play in cell physiology, and thus in health and disease. For example, uncovering the roles of individual amino acid residues and nucleotides forming the protein–DNA interface in the molecular recognition is essential for understanding the molecular mechanisms underlying the pathological effects of selected harmful mutations.[32–39] Similarly, understanding the molecular basis of specific protein–DNA recognition is critical for the rational design and engineering of novel DBPs or peptides.[40–47] Over the past few decades, search for a general "protein–DNA recognition code" has been undertaken,[48–50] motivated perhaps by the simplicity of the "rules" describing base pairing in a dsDNA molecule.[51] Structural, biophysical, bioinformatical, and computational analyses have been utilized to gain a detailed insight into the recognition mechanisms of several DNA-binding domains (DBDs), such as the ETS domain,[52,53] forkhead domain,[54,55] or homeodomain (HD),[56–58] while the "programmable" DNA-binding specificities of zinc finger nucleases and transcription activator-like effector nucleases made these chimeric enzymes key tools of genome editing.[45,59]

In spite of these achievements, no single code explaining the *in vivo* binding preferences of all DBPs in the cell has been described to date. While the preferences of individual DBPs for their cognate DNA binding sites can often be rationalized—with the help of biomolecular 3D structures—in terms of interactions between selected amino acid residues and nucleotides, many other factors affect protein–DNA binding in a cellular environment. These include the local and global chromatin states,[60] nucleosome positioning,[61] DNA binding site accessibility, DNA modifications,[62] interactions with other proteins or cofactors, and others. The necessity to capture the complex interplay between these effects raises the question whether a single universal protein–DNA recognition code describing the *in vivo* binding preferences of all DBPs can ever be found or what form would it take.[14,50]

Although the details of how DBPs select their DNA targets in a biological environment are not fully understood, statistical analyses of the 3D structures of protein–DNA complexes have been able to identify the basic mechanisms of sequence-specific recognition at the molecular level.[63–65] These mechanisms are usually divided into the direct and indirect readout of the sequence-specific features of a DNA molecule.[14,49,50] The direct, or base, readout is based on noncovalent interactions between the protein molecule and characteristic chemical groups of individual bases in the dsDNA. While it was originally proposed to only include the interactions between selected amino acids and DNA bases—asparagine/glutamine–adenine and arginine–guanine in the major groove of the DNA; asparagine/glutamine–guanine in the minor groove of the DNA—which feature bidentate hydrogen bonds (HBs) involving the side chain of the respective amino acid,[66] the term can be defined more broadly to encompass any hydrogen bonding, nonpolar, or water-mediated interaction relying on a matching character of chemical groups at the interface between the protein and DNA molecules.[49,67–75]

In contrast, indirect, or shape, readout is based on the specific recognition

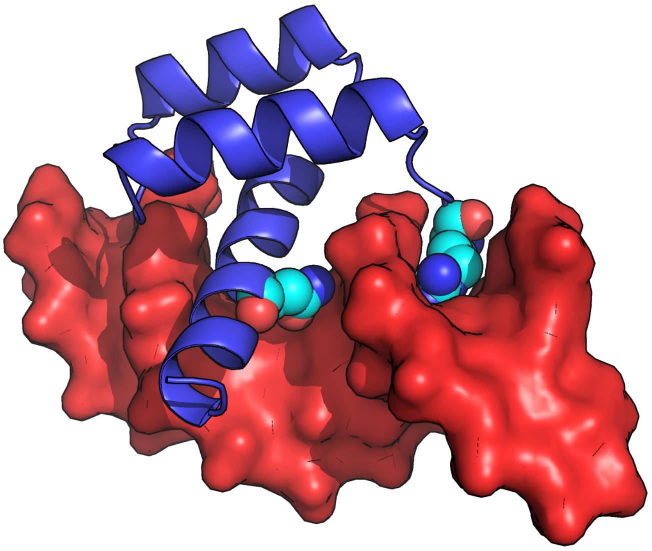


Figure 4: The X-ray crystallographic structure of the Antennapedia HD–DNA complex at 2.4 Å resolution (PDB ID `9ANT`). The HD is depicted in a cartoon representation in blue; DNA is depicted in a surface representation in red. Atoms of selected amino acid residues—Arg5 and Asn51—involved in direct and indirect recognition of the DNA binding site are shown as spheres. Visualized using PyMOL-2.3.0.[2]

of sequence-dependent topological features of the dsDNA molecule. These may manifest at multiple levels. Local effects include widening or narrowing of the major or minor grooves or local deformations of the DNA double helix, such as kinks. Global geometric properties include long-range bending of the DNA fiber or changes of its overall conformation, such as switching to the A- or Z-form. Finally, indirect readout involves the recognition of sequence-dependent dynamic characteristics of the DNA structure, such as its deformability.[49,76–81]

Direct and indirect modes of recognition are not mutually exclusive. In fact, TFs and other sequence-specific DBPs often utilize both mechanism to achieve high fidelity in the selection of their respective DNA binding sites. The level to which each mode is utilized and its role in the protein–DNA recognition depend on the character of the TF family.[49,81,82] A well-studied example is the HD (Figure 4), a small DBD featuring the helix-turn-helix (HTH) DNA-binding motif found in many eukaryotic proteins involved in development. HD proteins recognize their family-specific DNA binding sites through the interactions between selected amino acid residues and nucleotides in the major groove of the DNA; however, the direct read-out alone does not allow individual HD proteins to identify their respective protein-specific binding sites. The latter is made possible through the recognition of subtle sequence-dependent topological features of the adjacent minor groove.[56–58,61,79,83–85] This example illustrates how the direct and indirect readout mechanisms can complement each other in achieving different levels of specificity in protein–DNA recognition.

Computational approaches to the study of biomolecules

Quantum chemical methods

Modern computers and supercomputers make it possible to assess many properties of molecular systems which are difficult or impossible to observe directly using experimental methods. The computational techniques in use today vary vastly with regard to the physical accuracy of the models and approximations used, their computational demands, and the range of properties of the system which they enable one to explore. On one end of the spectrum, quantum chemical methods attempt to characterize the molecular system in terms of properties derived from its quantum mechanics (QM)-based description. These "first principles", or *ab initio*, approaches are used to study the electronic structure of the systems and can be divided into the wave function theory (WFT) and density functional theory (DFT) methods. The WFT techniques are, in principle, able to provide the exact solutions to the time-independent nonrelativistic Schrödinger equation for the electrons in the system under a set of physically meaningful approximations. These solutions take the form of the electronic wave functions and their corresponding energies.[86] In contrast, the DFT methods do not yield the wave functions and operate only with the electron density.[87–90]

The main drawback of the *ab initio* methods are their immense computational requirements. These limit the applicability of the most accurate reference methods, such as coupled clusters with iterative single and double excitations and perturbative triple excitations [CCSD(T)] in the complete basis set (CBS) limit, to systems consisting of tens of atoms. Less computationally demanding methods can suffer from an imbalanced description of intermolecular forces; for example, the second-order Møller–Plesset perturbation theory (MP2) overestimates the π–π stacking energies.[91,92] On the other hand, the basic WFT method (Hartree–Fock) completely ignores the correlation of motion of electrons with opposite spins, leading to severe errors in the description of, *e.g.*, the London dispersion interaction, which plays a critical role in the stabilization of biomolecular complexes.[87–90,93,94] The DFT methods are generally less computationally demanding than WFT methods of similar accuracy; however, they are also plagued by an inadequate description of noncovalent interactions, in particular the London dispersion. This has led to the development and adoption of empirical dispersion corrections and specialized density functionals.[95,96] In fact, the uncertainty about the exact form of the "correct" density functional is one of the problems of DFT.[89,90]

In spite of these limitations, QM methods have been extremely helpful in unveiling the physical properties of the basic biomolecular building blocks. The applications of *ab initio* calculations have included the exploration of the geometries[97] and molecular surface electrostatic potentials[98,99] of DNA and RNA bases, the energetics of hydrogen bonding[91,100–102] and stacking[101,103,104] interactions in nucleic acids, the energies of DNA sugar–phosphate backbone rotamers,[105] the energies of representative interactions between amino acid side chains in proteins,[91,106,107] or the energetics of hydrogen bonding interactions between selected amino acid side chains and DNA or RNA bases.[108] These studies have been instrumental, *e.g.*, for understanding the significance of the various types of noncovalent interactions for the stability of biomolecular complexes.[88]

Empirical methods

Empirical methods stand in contrast to the QM approaches with regard to the truthfulness of the physical model used. Where the *ab initio* methods derive the energy of the system from fundamental physical constants based on the principles of QM, empirical approaches, such as molecular mechanics (MM), make drastic simplifications regarding the nature of the particles in the system and their interactions. In MM, the potential energy $E_{\text{potential}}$ of a particular configuration of the system is usually described as

$$E_{\text{potential}} = E_{\text{bonded}} + E_{\text{nonbonded}} \tag{3}$$

where

$$E_{\text{bonded}} = E_{\text{bonds}} + E_{\text{angles}} + E_{\text{dihedrals}} \tag{4}$$

and

$$E_{\text{nonbonded}} = E_{\text{Coulomb}} + E_{\text{van der Waals}} \tag{5}$$

In all-atom MM, atoms are represented by single point particles with parameters describing their intra- and intermolecular interactions. The terms E_{bonds}, E_{angles}, and $E_{\text{dihedrals}}$ in Equation 4 penalize the deviations from equilibrium bond lengths l_0, angles θ_0, and dihedral angles, respectively, for pairs, triples, and quadruples of neighboring, covalently-bonded atoms. The term E_{Coulomb} in Equation 5 describes the Coulomb interactions between partial atomic charges q_i, q_j; the term $E_{\text{van der Waals}}$ models the exchange repulsion and London dispersion interactions. The last two terms concern interactions between all pairs of atoms i, j in the system, although the "nonbonded" interactions between neighboring, covalently-bonded atoms are usually ignored or downscaled, and a special treatment of long-range interactions may be used.[109,110] The terms E_{bonds} and E_{angles} are usually modeled as quadratic functions of the displacements from the respective equilibrium values:

$$E_{\text{bonds}} = \sum_{\text{bonds}} k_l (l - l_0)^2 \tag{6}$$

$$E_{\text{angles}} = \sum_{\text{angles}} k_\theta (\theta - \theta_0)^2 \tag{7}$$

where k_l and k_θ are the respective force constants; $E_{\text{dihedrals}}$ is a periodic function of the respective dihedral angles ω:

$$E_{\text{dihedrals}} = \sum_{\text{dihedrals}} \sum_{n=0}^{N} k_n [1 + \cos(n\omega - \gamma)] \tag{8}$$

where k_n is a force constant, n is the multiplicity, and γ is the phase factor; E_{Coulomb} is calculated from the Coulomb's law:

$$E_{\text{Coulomb}} = \sum_i \sum_{j>i} \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \tag{9}$$

where ε_0 is the vacuum permittivity and r_{ij} is the distance between the atoms; and $E_{\text{van der Waals}}$ is often represented by the Lennard-Jones potential:

$$E_{\text{van der Waals}} = \sum_i \sum_{j>i} 4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \tag{10}$$

where ε_{ij} and σ_{ij} are parameters of the potential. The precise form of the overall potential energy function, together with the set of the atomic interaction parameters, is called a force field (FF).[111]

Many MM FFs exist. These differ in the form of the potential energy function, the atomic interaction parameters, the parametrization procedure, or the set of molecules which they can describe. The parameters used in the individual FFs are usually self-consistent but not transferable to other FFs. As a result, FFs are usually specialized for the description of a particular set of molecules. The most widely used FFs for the investigation of proteins and nucleic acids belong to the AMBER and CHARMM families. These derive their parameters from fitting to different sets of experimental (*e.g.*, thermodynamic) or theoretical (QM) data. The most recent AMBER FF for the simulations of proteins is ff14SB;[112] this FF can be combined with the parmbsc1[113] or OL15[114] FFs for the simulations of DNA. ff94,[115] ff99,[116] and their various modifications are older AMBER FFs which remain commonly used today, owing perhaps to their inclusion (and absence of the more recent alternatives) in the popular simulation package GROMACS.[117] The most recent CHARMM FF for the simulations of proteins and nucleic acids is CHARMM36;[118,119] since its release, the protein part has been improved and released as CHARMM36m.[120]

MM FFs can be used to perform "single-point" calculations of the potential energies of individual configurations of the system, facilitating the exploration of its potential energy surface. These energies can be directly compared with ones obtained using QM methods for the same configurations, yielding information about the accuracy of the empirical models and possibly also guidelines for their improvement.[105,106,121,122] However, the most important application of MM FFs lies in molecular dynamics (MD). In MD, atoms in a particular (starting) configuration of the system are assigned initial momenta and the system is evolved through time by numerically integrating the classical equations of motion. The primary output of an MD simulation is a many-body trajectory representing the dynamics of the studied system. MD simulations are often performed under virtualized constant temperature[123] and pressure[124,125] conditions; assuming the validity of the ergodic hypothesis, which equates the time averages over a simulation with the averages over the corresponding statistical ensemble, MD simulations enable the estimation of thermodynamic properties of the system based on the principles of statistical mechanics. Nonequilibrium processes can be studied by means of MD simulations as well; for example, in steered MD simulations, external forces are applied to selected parts of the system to drive it away from equilibrium.[111,126]

MD simulations offer insight into molecular behavior with spatial and temporal resolution unparalleled by any experimental technique. Simulations of systems as large as viral capsids (containing millions of atoms) spanning hundreds of nanoseconds are now practically achievable,[127] while simulations of smaller systems, such as single biomolecules is a water environment, can reach millisecond time scales. While QM calculations are usually performed *in vacuo* or using implicit (continuum) solvation models,[105,107,128] in modern MD simulations, molecules of the solvent are usually treated explicitly, enabling the descriptions of, *e.g.*, water bridges or structural water molecules which can play an essential role in molecular recognition and structural dynamics of biopolymers.[18,64,68,69,72,129] Nevertheless, certain problems with MD simulations exist. For complex biomolecular systems, the time scales examined in state-of-the-art all-atom MD simulations allow the exploration of only a very small region of the phase space, making the trajectories highly dependent on the initial conditions.[111] Many biological processes, such as protein folding, can naturally occur over time scales several orders of magnitude longer than those currently achievable in MD simulations, making them difficult to study without the use of dedicated "enhanced sampling" techniques. A second major source of problems are the inaccuracies of present MM FFs. These include both errors which can be remedied by the reevaluation of certain atomic interaction parameters, as well as errors stemming from the highly simplistic form of the potential energy function. Notably, the use of fixed partial atomic charges and omission of polarization and charge transfer effects can make it difficult to balance the various types of noncovalent interactions determining the structure and dynamics of biomolecules, as well as prohibit the quantitative description of certain types of interactions, such as those with multivalent inorganic ions, altogether.[90,130,131]

Scope and structure of this work

In this introductory chapter, I attempted to summarize the essential information regarding the function of fundamental biomacromolecules—proteins and nucleic acids. Their structural organization and methods of its experimental determination were discussed. Special attention was given to the titular topic of this work—protein–nucleic acid (DNA) interactions. The biological significance of specificity in protein–DNA interactions and methods for its experimental characterization were presented. The known molecular determinants of the binding specificity and challenges in deriving a general protein–DNA recognition code were reviewed. Finally, an overview of the methods of computational chemistry applicable to the study of biomolecules was provided, and the respective strengths and weaknesses of the QM and empirical methods were discussed.

Given the breadth of the subject area, many important topics—ones I did not particularly study during the recent years—were omitted from this short introduction. These include, for example, the computational prediction of biomolecular 3D structures or the prediction of the genomic binding sites of DBPs. Other topics, such as the HT sequencing methods, were briefly explored despite the lack of direct relevance to my work owing to their immense role in modern biological research. The applications of the presented experimental and theoretical methods to the study of proteins and nucleic acids have been so many that an exhaustive review is likely beyond any-one's capacity; therefore, for the most part, only key works and up-to-date reviews were referenced. Nevertheless, I wish and believe that this chapter enables an interested reader to grasp the current state of our understanding of specificity in protein–nucleic acid (DNA) interactions and the respective capabilities and limitations of the tools used for its study.

In the following chapters, I present the results of my work exploring selected properties of biomolecules. Chapters 1 and 2 are dedicated to protein–DNA interactions. In Chapter 1, I explore the pairwise interactions between the basic biomolecular building blocks—amino acids and nucleotides. The primary goal here is bridging the gap between the energetics of the interaction motifs determined using the methods of computational chemistry and geometric preferences observed in large-scale bioinformatical analyses of the 3D structures of protein–DNA complexes. The secondary objective is assessing the level of agreement among the computational techniques used to calculate the energies of the interactions. This chapter concludes with a study of DNA shape and its correlation with the presence of amino acid residues with certain physico-chemical properties at the protein–DNA interface.

In Chapter 2, I move beyond the pairwise interactions and investigate whether all-atom, explicit solvent MD simulations can be used to quantitatively describe the binding equilibria in model protein–DNA complex systems. Equilibrium and nonequilibrium approaches are examined and compared, and the effects of the MM FF and reference temperature are explored. Analyses of the populations of the various intra- and intermolecular interactions observed in the respective simulations are then performed in order to understand the microscopic origins of the differences among the results obtained under the various conditions.

As Chapters 1 and 2 already contain critical reflections on the respective results presented, I dedicate the Discussion to the overview of additional works which are not restricted to the topic of protein–nucleic acid interactions but attempt to deepen our general understanding of proteins nevertheless. While the works presented in Chapters 1 and 2 offer, for the most part, physical analyses of protein–DNA complexes, the studies introduced in the Discussion focus on the evolutionary aspects of protein sequences and 3D structures. In this chapter, an analysis of evolutionary coupling among protein domains in multidomain proteins is discussed, and a tool for the visualization of evolutionary conservation in protein 3D structures is presented.

1. Preferences in pairwise amino acid–nucleotide interactions

This chapter summarizes the methods and results of the works Jakubec *et al.* (2015)[1], Hostaš *et al.* (2015), Jakubec *et al.* (2016), Stasyuk *et al.* (2017), Galgonek *et al.* (2017), and Faltejsková *et al.* (2020) presented in the List of publications (Page 69). As discussed in the Introduction, these works concern the pairwise interactions between the basic biomolecular building blocks—amino acids and nucleotides. The works Jakubec *et al.* (2015), Jakubec *et al.* (2016), and Faltejsková *et al.* (2020) are discussed in greater detail given my principal involvement in these; in contrast, only a synopsis is provided for the works Stasyuk *et al.* (2017) and Galgonek *et al.* (2017) given my limited input to these. Only a surface-level overview of the works mentioned is provided here; a more in-depth discussion of the analyses performed is available in the attached publications.

1.1 Jakubec *et al.* (2015)

1.1.1 Synopsis

In this work, MM methods were used to calculate the interaction energies (IEs) of all interacting amino acid residue–DNA base pairs found in the 3D structures of protein–DNA complexes. Geometric preferences of individual amino acids for binding the DNA bases were examined by performing a clustering analysis on the respective sets of 3D configurations of the dimers. A synthesis of the energetic and geometric views of the interactions allowed us to identify spatially well-defined binding motifs which simultaneously corresponded to the IE minima of the respective amino acid residue–DNA base pairs. For selected representative configurations of the binding partners, the IEs were calculated using QM methods as well, and a satisfactory agreement with the results of empirical calculations was observed.

1.1.2 Methods

The data set used in this work was based on the "Atlas of Protein Side-Chain Interactions"[71] and consisted of a total of 50,205 amino acid residue–dNMP pairs extracted from a total of 1,569 3D structures of protein–DNA complexes stored in the PDB archive.[11] Only X-ray crystallographic structures solved to a resolution of 2.5 Å or better and containing a dsDNA region of length at least 4 bps were considered. An interaction was recognized as featuring the DNA base when any atom of the amino acid side chain was within 4.5 Å of any atom of the DNA base; a total of 30,357 dimers featuring interactions

[1] Throughout this thesis, I deliberately use a different (author–date) format to distinguish the works to which I have contributed from standard references to the bibliography.

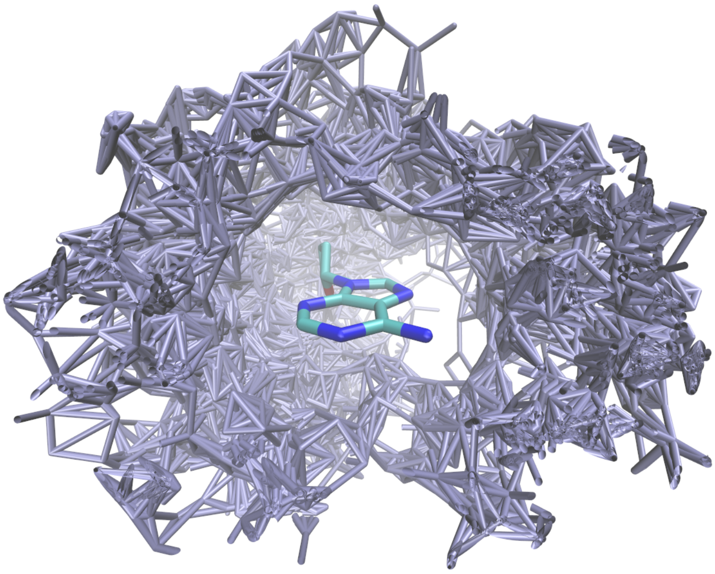


Figure 1.1: A distribution of asparagine side chains around an adenosine nucleoside. From Jakubec *et al.* (2016).

with DNA bases were identified in the data set. For the purpose of clustering, a common frame of reference centered at the DNA base was established for all amino acid residue–dNMP dimers containing a particular combination of the monomers; an ensemble of such dimers was referred to as the 3D *distribution* of the amino acid residue around the DNA base. An example of a distribution is shown in Figure 1.1.

A combination of energetic favorability and geometric constraints can lead to the enrichment of certain amino acid residue–dNMP configurations in the distributions. These *clusters* of amino acid residues in the distributions were identified as follows. The root-mean-square deviation (RMSD) of the positions of three reference atoms was calculated for each pair of amino acid residues in each distribution. The amino acid residue with the greatest number of neighbors within an RMSD of 1.5 Å was considered the first *cluster representative* (CR); the CR together with these neighbors were considered the first cluster. The amino acid residues comprising the cluster were then removed from the search space and the procedure was repeated until a total of 6 clusters were identified in each distribution or until an insignificant cluster appeared. A total of 468 clusters comprising a total of 13,155 amino acid residue–dNMP dimers were identified in the 80 distributions; a total of 272 clusters comprising a total of 6,991 dimers corresponded to amino acid residues interacting directly with the DNA bases. An example of the clusters identified in a 3D distribution is shown in Figure 1.2.

For each amino acid residue–dNMP dimer featuring a direct interaction with the DNA base, a C_α representation of the amino acid was prepared by substituting the backbone carbonyl and amide groups with hydrogen atoms. Only the DNA base was retained from each nucleotide. The IE was calculated for each amino acid residue–DNA base pair in each distribution as the difference between the potential energy of the dimer and the sum of the potential energies of the isolated monomers. Hydrogen atom positions were

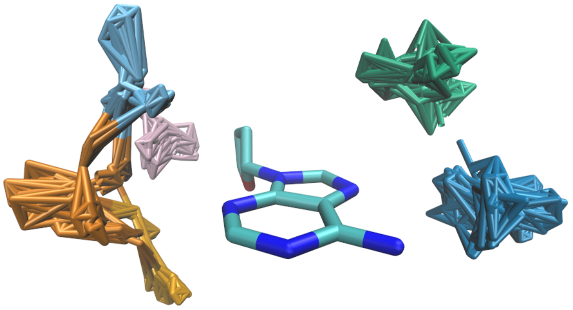


Figure 1.2: The clusters identified in the 3D distribution shown in Figure 1.1. From Jakubec *et al.* (2016).

optimized prior to each potential energy calculation; the positions of nonhydrogen atoms were kept fixed. The MM energy calculations and geometry optimizations were performed *in vacuo* using GROMACS 4.5.5.[117] Three MM FF combinations were used to perform the calculations and optimizations: ff03[132] (protein) + ff94[115] (DNA), ff99SB-ILDN[133] (protein) + ff94 (DNA), and CHARMM22[134] (protein) + CHARMM27[135] (DNA).

For the set of 272 representatives of clusters featuring interactions with the DNA bases, the IEs were calculated using the QM DFT-D3[96] method as well. The B3LYP[136,137] functional and the def2-TZVPP[138,139] basis set were used in these calculations; selected dimers were also investigated using the def2-QZVP basis set. In order to assess the accuracy of the DFT calculations, the IE was calculated for a single glutamate–guanine dimer using the CCSD(T)/CBS method. The IEs were calculated *in vacuo* using the supermolecular approach described above. A correction for the basis set superposition error (BSSE)[88] was not applied. The program package TURBOMOLE 6.5[140] was used to perform the QM calculations.

1.1.3 Results and discussion

For the purposes of the analysis, the 272 dimers featuring the CRs were split into 4 groups according to the physico-chemical character of the amino acid residue involved: nonpolar (G, A, V, I, L, P; 76 pairs), polar (T, S, N, Q, C, M; 69 pairs), charged (K, R, D, E; 64 pairs), and aromatic (F, Y, W, H; 63 pairs). The respective R^2 coefficients of determination between the IEs calculated using the three MM FFs and B3LYP-D3/def2-TZVPP are shown in Table 1.1. It can be seen that the Amber FFs yield mutually similar results, while the performance of the CHARMM FFs differs when applied to dimers featuring polar or aromatic amino acid residues. A global shift toward more negative (*i.e.*, more stabilizing) IEs was observed in the B3LYP-D3/def2-TZVPP results for the set of dimers featuring charged amino acid residues in comparison with the MM FFs; the opposite behavior was observed for the pairs featuring the nonpolar amino acids. No global IE shifts were observed for the two other sets of dimers. In spite of these discrepancies, it was concluded—based on the presented coefficients of determination—that at least a qualitative agreement between the IEs calculated using the MM and QM methods had been achieved; this enabled us to postulate hypotheses

Table 1.1: The R^2 coefficients of determination between the IEs calculated using the MM and QM (B3LYP-D3/def2-TZVPP) methods for the sets of representative dimers featuring amino acid residues with the respective physicochemical properties. The correlation coefficients (R) were positive in each case.

FF (protein + DNA)	nonpolar	polar	charged	aromatic
ff03 + ff94	0.69	0.84	0.94	0.84
ff99SB-ILDN + ff94	0.76	0.84	0.95	0.88
CHARMM22 + CHARMM27	0.78	0.73	0.93	0.61

concerning pairwise amino acid–nucleotide interactions based on the results of empirical calculations.

We thus proceeded to study the MM IE profiles (*i.e.*, histograms) of the entire 3D distributions and, in particular, focus on the respective positions of the IE profiles corresponding to the clusters and CRs therein. Notably, we observed the following characteristics for specific clusters in selected distributions:

1. the cluster comprised of dimers with the most negative IEs (*i.e.*, greatest stabilization energies) found in the respective distribution,

2. a significant portion of the dimers found within that IE range were members of the respective cluster, and

3. the peak corresponding to the cluster was visibly set off from the bulk of the IE profile.

The occurrence of such a cluster in a distribution suggests that only a conformationally narrow set of dimer geometries can lead to the greatest stabilization energies possible. These conditions were met by a single cluster in the asparagine–adenine, glutamine–adenine, lysine–adenine, asparagine–cytosine, and tyrosine–cytosine IE profiles and by two clusters in the glutamine–guanine IE profile. An example of an IE profile with the stated characteristics is shown in Figure 1.3. The 3D structures of the CRs corresponding to these clusters are depicted in Jakubec *et al.* (2015). It can be seen that all these binding motifs feature a hydrogen bonding interaction. In particular, it must be noted that the asparagine–adenine, glutamine–adenine, and glutamine–guanine dimers feature bidentate HBs and thus correspond to the binding modes which enable a unique one-to-one pairing between amino acid residues and DNA bases. Interestingly, additional binding motifs displaying the described characteristics were identified as well. The roles of most of these are unknown; however, we found that one of the identified glutamine–guanine motifs is uniquely associated with a class of DNA methyltransferases, in which it plays a functional role. This example illustrates that the statistical–computational approach presented can be used to identify biologically relevant patterns in the structures of biomolecules.

This work is burdened by several simplifications, some of which are addressed in the following sections. Inclusion of the interactions between amino

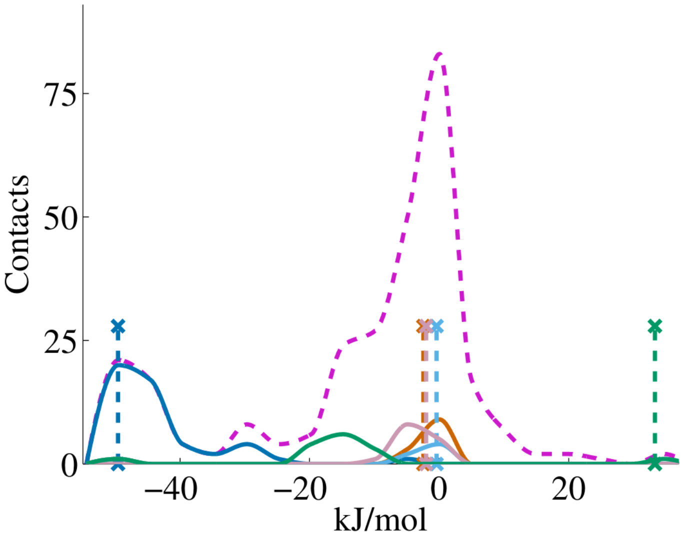


Figure 1.3: The ff03 IE profile corresponding to the glutamine–adenine distribution (dashed purple curve) together with the cluster IE profiles (solid blue, orange, light blue, pink, and green lines) and CRs (dashed vertical lines). From Jakubec *et al.* (2015).

acid residues and sugar-phosphate backbone of DNA is, without a doubt, essential for a full description of protein–DNA binding.[49] The relevance of the gas phase IEs for the description of binding in biomolecular systems can be questionable,[87] regardless of the accuracy of the computational method used. Both of these issues are addressed in Jakubec *et al.* (2016); in that work, the interactions with the sugar-phosphate backbone of DNA as well as the effects of implicit solvation on the IEs are studied. Topics which are not further addressed include calculations of IEs for complexes of amino acid residues with larger blocks of DNA, such as dinucleotide steps, or the energetic roles of structural water molecules.[68,69] Similarly, the omission of the interactions with the protein backbone appears unfortunate in retrospect, as the HBs featuring the peptide bond carbonyl and amide groups can easily contribute to the recognition of DNA sequences.[49]

1.2 Hostaš *et al.* (2015)

1.2.1 Synopsis

In this work, the performance and accuracy of various QM methods was evaluated by calculating the IEs for the set of amino acid residue–DNA base dimers featuring the representatives of the clusters identified in the previous section. The CCSD(T)/CBS IEs were calculated for all dimers and were used as a benchmark to which the results obtained using the MP2.5/CBS, MP2.5/MP2-F12, DFT-D3, PM6, and empirical methods were compared.

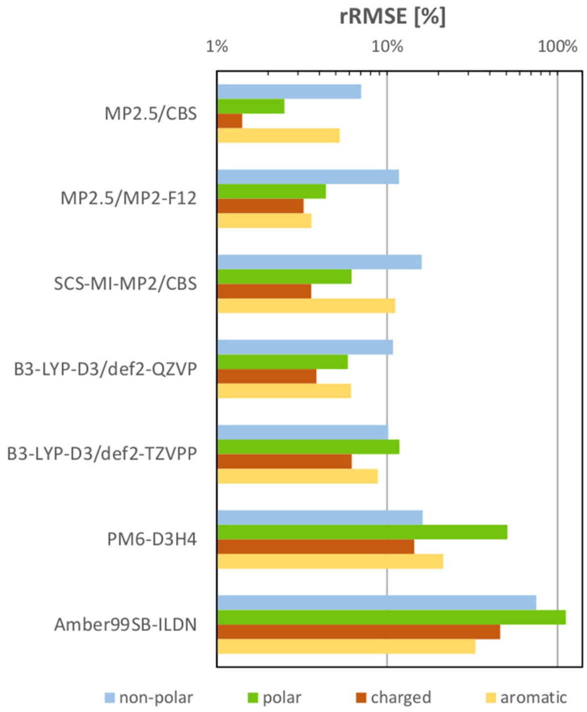


Figure 1.4: The root-mean-square errors of the IEs calculated using the respective computational methods in comparison with the reference CCSD(T)/CBS calculations. The errors are expressed as a percentage of the average IE observed for the groups of dimers featuring the respective sets of amino acid residues. From Hostaš *et al.* (2015).

1.2.2 Methods

The 3D structures of the 272 representatives of clusters featuring interactions with the DNA bases were adopted from Jakubec *et al.* (2015). As in the previous work, only the C_α representations of amino acids were considered. For selected analysis purposes, the set of dimers featuring charged amino acid residues was split into two groups corresponding to the positively (K, R; 33 pairs) and negatively (D, E; 31 pairs) charged amino acids. All IE calculations were performed *in vacuo* and were corrected for the BSSE. The reference IEs were calculated for all dimers using the most accurate and computationally demanding CCSD(T)/CBS method; these were compared primarily to the IEs calculated using the QM MP2.5/CBS, MP2.5/MP2-F12, B3LYP-D3/def2-TZVPP, and B3LYP-D3/def2-QZVP methods, the semiempirical PM6-D3H4 method, and the empirical ff99SB-ILDN (protein) + ff94 (DNA) FF method.

1.2.3 Results and discussion

The magnitude of the disagreement between the IEs calculated using the reference CCSD(T)/CBS method and the computationally less demanding methods is shown in Figure 1.4. The results obtained using the MP2.5/CBS and

MP2.5/MP2-F12 methods were found to be the most accurate and balanced across all binding motifs, making these tools suitable for the calculation of IEs in extended protein–DNA systems. The B3LYP-D3 methods globally overestimated the stabilization energies, in particular for the set of dimers featuring positively charged amino acid residues. Significantly more accurate IEs were observed for the sets of dimers featuring polar or negatively charged amino acid residues when the larger def2-QZVP basis set was used, while both B3LYP-D3/def2-TZVPP and B3LYP-D3/def2-QZVP methods yielded satisfactory results for the dimers featuring nonpolar or aromatic amino acids. The semiempirical PM6-D3H4 method exhibited less than half the error of the MM FF calculations, making it a promising tool for the study of extended protein–DNA systems with thousands of atoms. Finally, the empirical FF performed reasonably well for the dimers featuring nonpolar or aromatic amino acid residues but significantly underestimated the stabilization energies for the complexes of charged amino acids. The geometries of the dimers affected the performance of the MM method greatly, with configurations far from the equilibrium geometries leading to IE outliers.

1.3 Jakubec *et al.* (2016)

1.3.1 Synopsis

This work builds upon the methodology and results of Jakubec *et al.* (2015) and Hostaš *et al.* (2015) in which we have shown that MM FFs can yield qualitatively accurate gas phase IEs for amino acid residue–DNA base dimers. In this work, the studied biomolecular complexes are extended to include the moieties of the sugar-phosphate backbone of DNA, and the energetics of the extended systems are examined using the empirical methods. The effects of the dielectric properties of the environment on the IEs are studied using implicit solvation models. Rigorous criteria of specificity combining the geometric and energetic views of the interactions are defined and used to identify the statistically enriched binding motifs which simultaneously correspond to the most stabilizing arrangements of the respective binding partners.

1.3.2 Methods

The data set of amino acid residue–dNMP pairs was constructed as in Jakubec *et al.* (2015); the 3D distributions and clusters were defined identically. A slightly newer version of the PDB archive was used in the construction of the data set. The statistical bias which could have been introduced to the data set by the abundance of certain families of DBPs in the PDB archive was addressed to various levels by generating a total of 4 nonredundant data sets of amino acid residue–dNMP pairs; the amino acid residues comprising the respective data sets originated from corresponding ensembles of polypeptide chains having less than X % mutual sequence identity (X = 30, 90, 95, or 100, respectively). The removal of identical polypeptide chains from the data set (*i.e.*, X = 100 above) discarded over a half of all dimers (21,709 dimers remained from the original 47,480), while a relatively small difference was seen between the numbers of dimers included in the sets at the $X = 30$ (8,926 dimers) and $X = 90$ (12,505 dimers) levels. For each of these data sets, a subset of dimers featuring a direct interaction between the amino acid residue and the DNA base was identified; these sets comprised of 4,752, 6,546, and 10,572 amino acid residue–dNMP dimers at the $X = 30$, 90, and 100 levels, respectively.

The IE calculations were performed using the ff99SB-ILDN (protein) and ff94 (DNA) FFs as in Jakubec *et al.* (2015). Only the C_α representations of amino acids were considered. The generalized Born/surface area model[141] was used to account for the effects of continuum solvation; the IE calculations were performed in environments with the values of the relative permittivity $\varepsilon_\mathrm{r} = 1$, 4, 16, and 80, respectively.

1.3.3 Results and discussion

To investigate the geometric and energetic specificity in the pairwise interactions, the IE profile was generated for each distribution as described in Jakubec *et al.* (2015). The cluster formed by the dimers which exhibited the most negative (*i.e.*, most stabilizing) average IE was then identified and the histogram of the IEs of the cluster members was plotted against the IE profile of the distribution. We rationalized that a binding motif must meet the following criteria to be viewed as significant for the process of DNA sequence recognition:

1. the configuration must be found within one of the geometric clusters; this condition implies that the respective binding motif is present in multiple protein–DNA complexes and as such is not limited to be functional only in the unique local environment of a single DBP family,

2. the cluster to which the binding motif belongs must represent the most energetically favorable arrangement of the respective binding partners,

3. few to no configurations which are not members of the cluster are to be present within its respective IE range. This leads to a unique association between the IE range and a specific binding geometry; this criterion also implies that all dimers within that particular IE range are geometrically well-defined, as the respective amino acid residues could be recognized as forming a cluster,

4. the IEs corresponding to the cluster must be more negative (*i.e.*, more stabilizing) than the most negative IEs found in the dimers of said amino acid residue with the other nucleotides. Such an analysis must be performed separately for each edge of the nucleotide (Hoogsteen, Watson–Crick, sugar-phosphate), as it may be possible for an amino acid residue to distinguish between individual nucleotides in each of these regions.

The clusters meeting all these criteria correspond to statistically enriched, geometrically well-defined binding motifs which can distinguish between individual nucleotides on the basis of energetic favorability. The work Jakubec

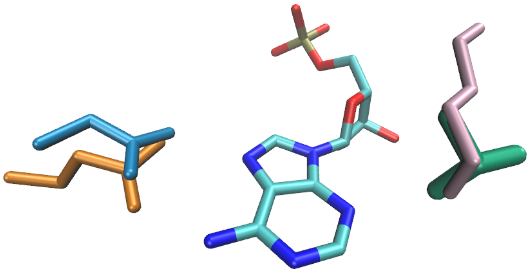


Figure 1.5: The specific recognition of adenine (dAMP) by asparagine (blue), glutamine (orange), lysine (pink), or threonine (green) side chains. From Jakubec *et al.* (2016).

et al. (2016) and its supporting materials list, discuss, and graphically depict the binding motifs meeting these criteria to various levels; a summary of the most significant observations is provided here.

The pairs asparagine–adenine, glutamine–adenine, and arginine–guanine, traditionally associated with the direct readout of DNA sequences, were correctly identified in the "control" group of amino acid residue–DNA base dimers as specific binding motifs according to the criteria described above, supporting our hypothesis that the combined statistical–geometric–energetic view of the interactions can be used to identify the specific binding motifs. These results remained valid regardless of the relative permittivity ε_r of the implicit solvent used. A few other interactions that were described in Jakubec *et al.* (2015) were identified in this set of dimers as well. The asparagine–adenine, glutamine–adenine, and arginine–guanine interactions feature a bidentate HB between the amino acid side chain and the Hoogsteen edge of the DNA base. These interaction motifs appear significant with regard to the criteria described above even when dNMPs are considered; however, the prominent positions of the asparagine–dAMP and glutamine–dAMP pairs diminish when a large relative permittivity ($\varepsilon_\mathrm{r} = 80$) of the environment is assumed. Interestingly, a hydrogen bonding interaction between the side chain of lysine and the N3 atom of adenine emerges as significant in the water-like environment when the sugar-phosphate moiety is included. The specific recognition motifs featuring adenine or dAMP are depicted in Figure 1.5.

In contrast, the interaction of arginine with dGMP maintains its distinctive characteristics even in the water-like environment. An interesting interaction featuring a bidentate HB between the side chain of aspartate and the Watson–Crick edge of guanine emerges as significant in the water-like environment when the sugar-phosphate moiety is included. A further analysis revealed that this motif is utilized in the binding of aptameric, telomeric, or otherwise strained DNA structures. As this interaction obviously interferes with the Watson–Crick pairing between the DNA bases, it is most likely not involved in routine sequence recognition; nevertheless, it can contribute to the recognition of noncanonical forms of DNA. This example illustrates the robustness of the introduced criteria of specificity: without any information regarding the function or evolutionary relationships among the proteins in

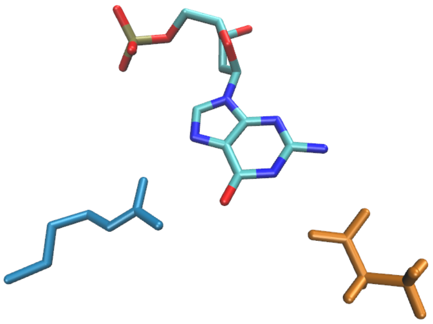


Figure 1.6: The specific recognition of guanine (dGMP) by arginine (blue) and aspartate (orange) side chains. From Jakubec *et al.* (2016).

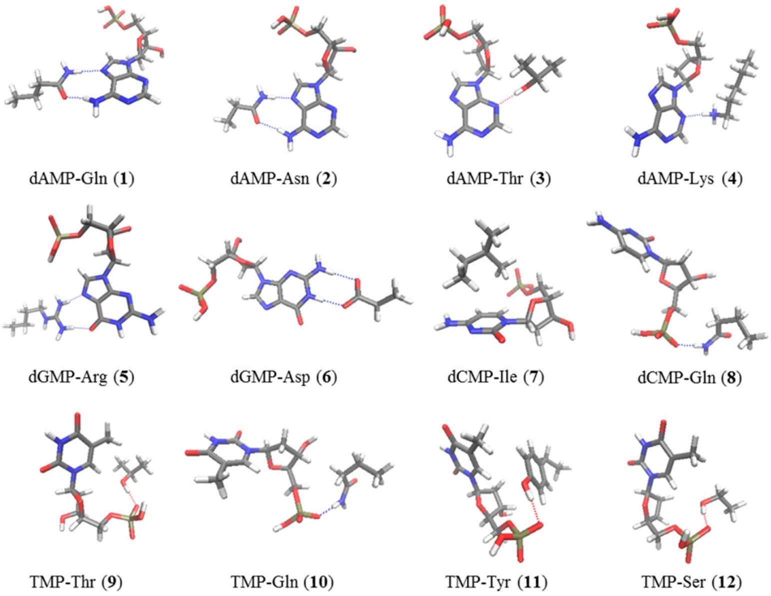
our data set, it was possible to identify a specific binding motif involved in the recognition of a characteristic set of DNA structures. The specific recognition motifs featuring guanine or dGMP are depicted in Figure 1.6.

The binding motifs featuring cytosine or thymine appeared less significant in comparison with the ones described above, given both the lower populations of the respective clusters and the lack of biophysical features (*e.g.*, presence of HBs between the amino acid residue and the DNA base) which could lead to these interactions being described as specific.

This work is burdened by similar issues as Jakubec *et al.* (2015). The size of the data set is limited—in particular, when the redundancy is addressed—prohibiting the discovery of binding motifs which are not abundant in the current state of the PDB archive. The role of structural water molecules and interactions with larger blocks of DNA remain neglected. The pairwise interaction analysis of the protein–DNA interface may ignore important features of DNA sequence and structure recognition which are specific to individual families of DBPs.[49,81,82] The IEs calculated for individual configurations of the binding partners may not be indicative of the respective biophysically relevant free energies of binding.[142] Finally, a more robust statistical analysis could be devised for the identification of significant clusters.

1.4 Stasyuk *et al.* (2017)

In this work, the IEs were calculated using QM methods for a set of 12 amino acid residue–dNMP dimers (Figure 1.7) which displayed the characteristics of specific binding motifs described in the previous work. The CCSD(T)/CBS IEs were calculated as a reference to which the performance of the MP2.5/CBS, MP2-F12/cc-pVDZ-F12, DFT-D3 (BLYP-D3/def2-QZVP), and PM6-D3H4 methods was compared. In addition to the gas phase calculations, the COSMO[128,143] model of implicit solvation was used together with the DFT-D3 and PM6-D3H4 methods to study the effects of the environment. The geometries of the dimers were further optimized in environments with various values of the relative permittivity ε_{r} to examine how close the configurations present in the 3D structures of protein–DNA complexes are to the respective energy minima. Finally, the decomposition of



dAMP-Gln (**1**) dAMP-Asn (**2**) dAMP-Thr (**3**) dAMP-Lys (**4**)

dGMP-Arg (**5**) dGMP-Asp (**6**) dCMP-Ile (**7**) dCMP-Gln (**8**)

TMP-Thr (**9**) TMP-Gln (**10**) TMP-Tyr (**11**) TMP-Ser (**12**)

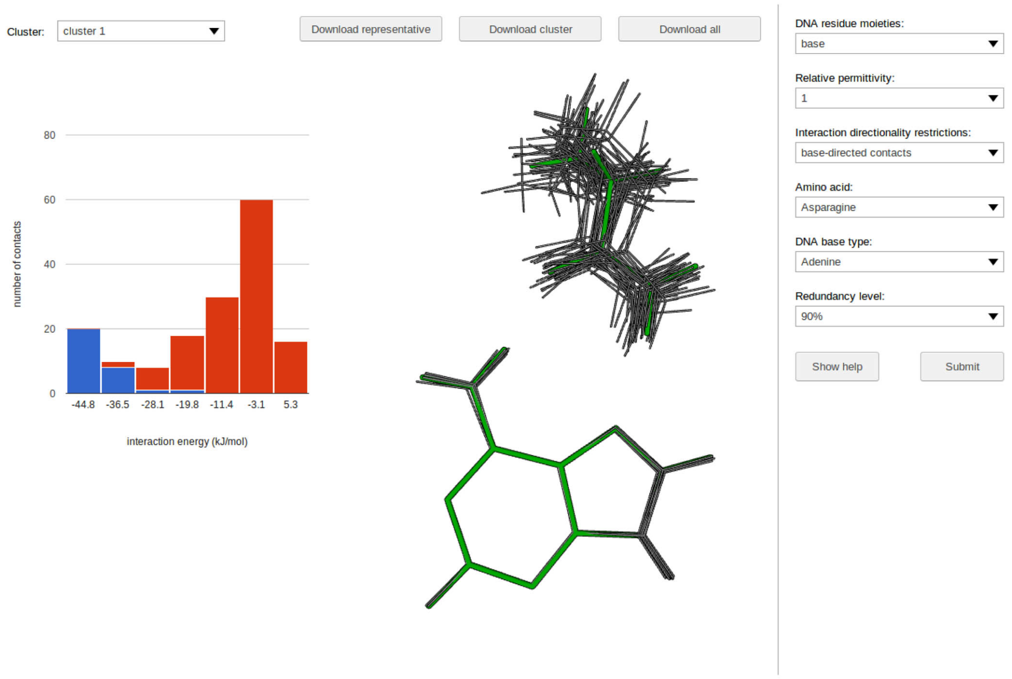
Figure 1.7: The binding motifs investigated in Stasyuk *et al.* (2017). From Stasyuk *et al.* (2017).

the IEs into their various components was performed using the DFT-SAPT[94] method.

A satisfactory agreement between the *in vacuo* IEs calculated using the reference and the tested methods was observed, with MP2.5/CBS and MP2-F12/cc-pVDZ-F12 exhibiting the smallest errors and PM6-D3H4 the greatest. MP2.5/CBS was found to globally underestimate the stabilization energies in comparison with the reference values, while overestimation of the energies was observed for the BLYP-D3/def2-QZVP method. All interactions were found to be stabilizing in environments with the values of relative permittivity ε_r = 1, 4, 16, and 78.5; interestingly, the most negative interaction energy in the water-like environment was observed for the aspartate–dGMP dimer featuring a bidentate HB between the amino acid side chain and the Watson–Crick edge of guanine. The DFT-SAPT analysis revealed that in the complexes of polar or positively charged amino acid residues, the electrostatic component contributes the most to the IE; however, in the complex featuring the negatively charged aspartate, this term is the least significant and the induction component dominates. The dispersion component dominates the IE in the complex featuring the nonpolar isoleucine side chain. Finally, the optimization of the dimer geometries at the BLYP-D3/def2-TZVPP level in environments with large values of relative permittivity ε_r yielded configurations very similar to those found in the X-ray crystallographic structures.

1.5 Galgonek *et al.* (2017)

This work presents the "Amino Acids Interactions Web Server" (`https://bioinfo.uochb.cas.cz/INTAA/`), an online service dedicated to the exploration of pairwise IEs in the 3D structures of biomolecules. The main purpose of this web server is the calculation of pairwise IEs among



Cluster: cluster 1

DNA residue moieties: base

Relative permittivity: 1

Interaction directionality restrictions: base-directed contacts

Amino acid: Asparagine

DNA base type: Adenine

Redundancy level: 90%

Figure 1.8: The user interface for the component of the “Amino Acids Interactions Web Server” dedicated to the exploration of IEs corresponding to specific configurations of amino acid residue–DNA base (dNMP) dimers. From Galgonek *et al.* (2017).

the respective biomolecular building blocks in the publicly available or user-uploaded representations of biomolecular structures. Various MM FFs and implicit solvent models can be used to perform the energy calculations. In addition, this web server enables an interactive exploration of the geometries and IEs of the amino acid residue–DNA base (dNMP) dimers studied in Jakubec *et al.* (2016). An illustration of the user interface for this component of the web server is shown in Figure 1.8. The user can select the amino acid residue and nucleotide to visualize, choose whether only the dimers featuring a direct interaction with the DNA base are to be considered, or select whether the sugar-phosphate backbone of DNA is to be taken into account in the IE calculations. Additional parameters include the relative permittivity ε_{r} of the environment to consider and the level of treatment of redundancy in the data set. The IE profile corresponding to the selected distribution is then visualized. The user can choose from individual clusters identified in the 3D distribution and the respective cluster IE profile will be overlaid on top of that of the distribution. The geometries of the cluster members will be shown inside a web-based molecular viewer; the user can click individual dimers to learn from which 3D structure of a protein–DNA complex they originate. Finally, the representations of the 3D structures of the selected CR, cluster members, or whole distribution can be downloaded.

1.6 Faltejsková *et al.* (2020)

1.6.1 Synopsis

In this work, we employed a statistical analysis to study the relationship between the physico-chemical character of the amino acid residues found at the protein–DNA interface and the local geometry of the DNA. We devised an algorithm which enabled us to associate each amino acid residue involved in binding the minor groove of the DNA with the local dimensions of the DNA double helix. A statistical analysis of thus determined characteristics of the protein–DNA interface revealed that hydrophobic amino acid residues are enriched in widened or otherwise distorted minor groove regions. We showed that this phenomenon is not restricted to a few families of DBPs by repeating the analysis with selected families known to distort the DNA helix discarded. A positive correlation was observed between the GC content of the DNA binding sites and the mean width of the corresponding minor groove regions. Finally, preferences for distinct secondary structural elements were detected for selected hydrophobic amino acid residues binding the distorted minor grooves.

1.6.2 Methods

A total of 4,783 3D structures of protein–DNA complexes were downloaded from the RCSB PDB.[12] Regions corresponding to DNA helices were identified in the 3D structures using the 3DNA[3] software package. A nonredundant set of protein chains corresponding to X-ray crystallographic structures solved to a resolution of 3.5 Å or better was extracted from the set of 3D structures containing a DNA helix using the PISCES[144] web server. The maximal sequence identity between any pair of protein chains was 90 %. The resulting nonredundant data set comprised a total of 976 polypeptide chains found in a total of 857 3D structures of protein–DNA complexes.

For each of the DNA bases, a reference atom associated with the major or minor grooves of the DNA was defined. The amino acid residues for which the minor groove-associated atom was the nearer of the two and within a distance of 6.0 Å of any atom of the amino acid were defined as minor groove-binding. The "contacted" bp and the nearer of the two adjacent bps were then defined as the *dinucleotide step* associated with that particular minor groove-binding amino acid residue. Using this procedure, the data set DS1 comprising a total of 1,293 amino acid residue–dinucleotide step pairs was constructed.

The minor groove width associated with each dinucleotide step was retrieved from the 3DNA analysis. Three width categories were defined according to the local dimensions of the minor groove at the dinucleotide step: narrow (width $\leq$ 11.0 Å), standard (width $>$ 11.0 Å and $<$ 17.0 Å), and wide (width $\geq$ 17.0 Å). The protein families as defined by the Pfam[145] database were identified within the polypeptide chains. A second data set DS2 was then constructed by eliminating from DS1 the pairs featuring amino acid residues which were a part of the *HMG_box* (PF00505) or *TBP* (PF00352) proteins families known from the literature to distort the DNA minor

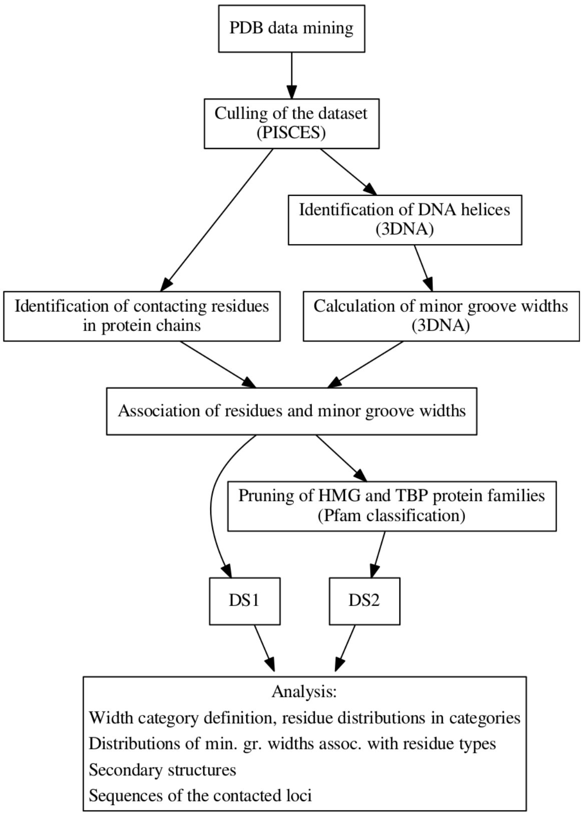

PDB data mining
Culling of the dataset (PISCES)
Identification of DNA helices (3DNA)
Identification of contacting residues in protein chains
Calculation of minor groove widths (3DNA)
Association of residues and minor groove widths
Pruning of HMG and TBP protein families (Pfam classification)
DS1
DS2
Analysis:
Width category definition, residue distributions in categories
Distributions of min. gr. widths assoc. with residue types
Secondary structures
Sequences of the contacted loci

Figure 1.9: A flowchart summarizing the methods and analyses featured in the work Faltejsková *et al.* (2020). From Faltejsková *et al.* (2020).

grooves.[77,146,147] A total of 1,103 amino acid residue–dinucleotide step pairs were included in the DS2 data set.

The Mann–Whitney U test[148] was employed to examine whether the mean of the distribution of minor groove widths associated with a particular amino acid residue was significantly greater than that of the respective outgroup. For this test, the amino acids were separated into the hydrophobic group (G, A, V, I, L, F, Y, W, M) and the rest. For each amino acid in the hydrophobic group, the corresponding distribution of minor groove widths was compared to the aggregate distribution of the nonhydrophobic ones. For each of the nonhydrophobic amino acids, the widths associated with that particular amino acid residue were removed from the aggregate distribution and the comparison was done as before.

The GC content (% of G–C bps) was calculated for each 6-bp region (hexamer) in the parts of the DNA helices which formed continuous protein–DNA interfaces. The minor groove width corresponding to a hexamer was taken to be the width at the central dinucleotide step. Finally, the secondary structural elements featuring the respective amino acid residues were identified using the DSSP[149] program. A flowchart summarizing the methodology of this work is provided in Figure 1.9.

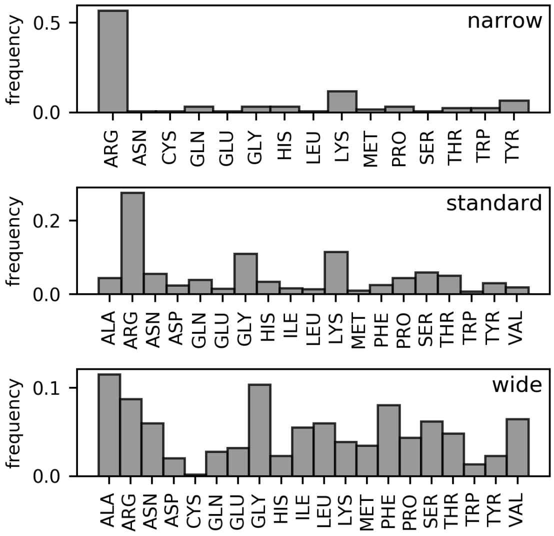


Figure 1.10: The frequencies with which the respective amino acid residues are found at the protein–minor groove interfaces for the respective minor groove width categories (DS1 data set). From Faltejsková *et al.* (2020).

1.6.3 Results and discussion

Figure 1.10 shows the amino acid composition of the protein–minor groove interfaces falling to the corresponding categories of minor groove width for the DS1 data set; the results for the DS2 data set were very similar. It can be seen that arginine is by far the most abundant amino acid residue interacting with the narrow and standard minor grooves, in which it constitutes 56.8 % and 27.6 % of all DNA-binding residues, respectively. The enrichment of arginine residues in the narrow minor grooves was previously observed by Rohs and colleagues,[61,79] who rationalized it on the basis of the enhanced negative electrostatic potential present. However, this dominance is lost in the wide minor grooves, in which arginine constitutes only 8.8 % of the DNA-binding residues; its frequency is thus similar to that of alanine (11.5 %) and glycine (10.4 %). The frequencies of other hydrophobic amino acids, such as phenylalanine and valine, are also considerably higher in this set in comparison with the interactions with narrow or standard minor grooves. In total, the fraction of hydrophobic amino acids in the wide minor grooves (57.3 %) is comparable to the fraction of arginine residues in the narrow minor grooves. In the DS2 data set, the hydrophobic amino acids constitute 55.7 % of all residues interacting with the wide minor grooves, showing that the removal of the known minor groove-distorting protein families affects the overall distribution of the amino acid frequencies only marginally.

Figure 1.11 shows the distributions of minor groove widths associated with the dinucleotide steps bound by the respective amino acids for the DS1 data set. It can be seen that the largest mean minor groove widths are associated with selected hydrophobic amino acids (M, L, V, F, I, A); on the other hand, tyrosine is associated with the lowest mean width. These results remain valid even after discarding the *HMG_box* and *TBP* protein families, although notable differences appear. While I, A, V, F, and L remain

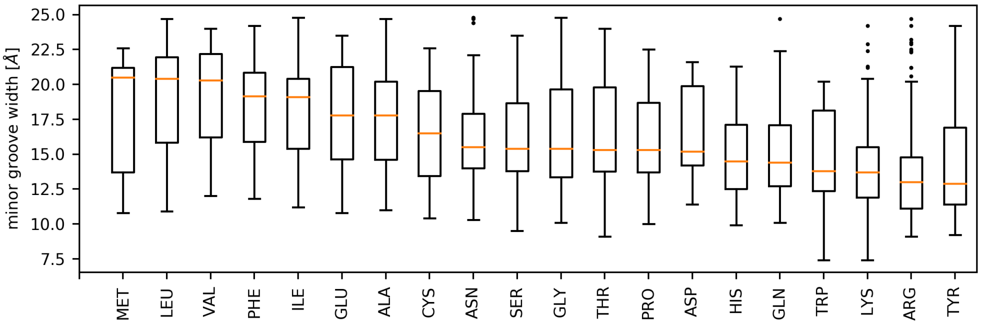


Figure 1.11: Distributions of minor groove widths associated with the dinucleotide steps bound by the respective amino acids (DS1 data set). The amino acids are sorted in descending order according to the median minor groove width. The whiskers of the boxplots denote the 25th and 75th percentile of the distributions. From Faltejsková *et al.* (2020).

the amino acids binding the widest minor grooves, the median groove width decreases by up to 3 Å for these residues in the DS2 data set; in the case of methionine, a decrease of the median minor groove width by over 5 Å is observed. Nevertheless, the results of the Mann–Whitney *U* test show that, on the confidence level of 95 %, the median minor groove width is significantly higher in the dinucleotide steps interacting with A, I, L, G, F, V, and E compared to the outgroup. These results are valid for both the DS1 and DS2 data sets. The fact that no DBPs are significantly enriched in the DS2 data set suggests that the distortion of the DNA minor grooves by hydrophobic amino acid residues is a universal phenomenon not specific to any single DBP family.

Associations between the bound minor groove width categories and secondary structural elements were identified for selected amino acid residues. G, I, and V residues were most commonly found in the β-strand structure when binding wide minor grooves, while this conformation was rare when binding the standard minor grooves. In other instances, such as F, A, and L, the preferences were more obscure or dominated by single families of DBPs.

Figure 1.12 shows the characteristics of the distributions of minor groove widths corresponding to the hexamers of DNA bps in contact with proteins as a function of the GC content of the DNA sequences. It can be seen that the means of the distributions increase with the increasing GC content of the sequences; on the contrary, the variance of the distributions decreases with the increasing GC content. The AT-only sequences were observed to form both extremely narrow and extremely wide minor grooves. The binding of the extremely wide, AT-only sequences often featured an intercalating hydrophobic amino acid residue, although other modes of binding were observed as well. More specific preferences for certain sequences of the dinucleotide step could be detected for selected amino acids. For example, 60 % of leucine and 40 % of isoleucine residues were in contact with dinucleotide steps consisting only of AT and TA bps, of which about a half corresponded to the ApA/TpT dinucleotide steps. On the other hand, a preference for GC-rich sequences was observed for aspartate and glutamate.

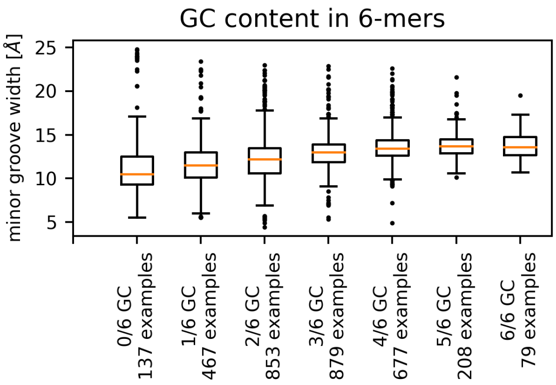


Figure 1.12: Distributions of DNA minor groove widths corresponding to the hexamers with different values of the GC content (DS1 data set). From Faltejsková *et al.* (2020).

While the enrichment of hydrophobic amino acid residues in the wide minor grooves was statistically proven in this work, the physical causes behind it remain unclear. Similarly, we were not able to conclude whether the hydrophobic amino acid residues specifically seek out the deformable regions of DNA, or whether they recognize regions which are already distorted.[49,76,79,80] An MD-based study exploring the energetics of DNA deformation in response to the binding of hydrophobic residues could be useful for the exploration of these topics.[150] Further limitations of this work include the limited size of the data set, which complicates the detection of preferences for binding specific DNA sequences of greater length, an incomplete description of the effects of multiple amino acid residues binding the DNA loci simultaneously, or the omission of global alterations of the DNA shape, such as long-range bends.[49,79]

2. Calculations of absolute binding free energies for protein–DNA systems

This chapter summarizes the methods and results of the works Jakubec and Vondrášek (2019) and Jakubec and Vondrášek (2020) presented in the List of publications (Page 69). As discussed in the Introduction, these works move beyond the pairwise interactions explored in Chapter 1 and concern the applicability of MD simulation setups to the description of binding equilibria in protein–DNA systems. As in the other chapters, only a surface-level overview of the works is provided here; a more in-depth discussion of the analyses performed is available in the attached publications.

2.1 Jakubec and Vondrášek (2019)

2.1.1 Synopsis

The Gibbs free energy difference (FED) $\Delta G \equiv G_{\text{bound}} - G_{\text{unbound}}$ between the bound and unbound states of a molecular complex, alternatively called the absolute binding free energy (BFE), determines the equilibrium ratio of activities of the species in the binding reaction. Numerous methods have been developed to enable the calculation of FEDs by means of MD simulations.[111,151–162] These can be broadly divided into approaches based on equilibrium and nonequilibrium simulations.[162] Although these methods do, in principle, yield identical results, practical limitations might favor single methods for specific applications.[153,154,162]

Unfortunately, calculations of absolute BFEs rank among the most computationally demanding applications of MD simulations.[111,162] This is caused by extensive sampling, and thus long simulation times, required to explore all relevant regions of the phase space. For this reason, they are often limited to small molecular models, and knowledge of their general accuracy for larger, biologically important systems is lacking.

In this work, we systematically examined the applicability of an MD setup to the calculations of standard BFEs of biologically relevant protein–DNA complexes. The FEDs between the bound and unbound states of the respective systems were extracted from one-dimensional potentials of mean force (PMFs)[1]. In order to efficiently sample all intermediate phase space regions, the umbrella sampling (US) technique was employed.[111,159,164,165] The biasing potential was then removed from the PMF calculations using the weighted histogram analysis method (WHAM).[165,166]

[1] The PMF acting between two particles is the free energy profile $G(\xi)$ along some geometric coordinate ξ;[111,151,157] when the coordinate ξ is chosen so that the PMF spans regions of the phase space corresponding to bound and unbound states, the absolute BFE can be deduced from it.[163]

Two protein–DNA systems derived from the Nkx2.5 HD–dsDNA complex were studied in order to investigate the interactions of both disordered and globular proteins. The systems were carefully chosen to possess several desirable features which make them suitable for testing the validity of real protein–DNA complex simulation setups. The FEDs and trajectories obtained using two modern MM FFs were compared to each other and to experimental data. The temperature-dependence of the calculated standard BFEs was further investigated by performing all simulations over a range of temperatures.

Most importantly, we showed that the values of standard BFEs obtained from these MD simulations are overestimated compared to the experimental results. In addition, significant differences were observed between the two protein–DNA systems as well as between the two FFs, which were rationalized by the different propensities to form various inter- and intramolecular interactions. Conclusions about the temperature-dependence of the standard BFEs could not be made with confidence, as the differences among the respective values were on the order of statistical error.

2.1.2 Methods

A model of the Nkx2.5 HD–specific dsDNA complex for which experimentally determined thermodynamic binding data are available[167] was prepared based on the X-ray crystallographic structure `3RKQ`[168] downloaded from the RCSB PDB.[12] The modeled system was equilibrated by performing an unrestrained MD simulation using GROMACS 2016.3.[117] A total of 5 system configurations were selected from the equilibration trajectory to serve as initial geometries for the pulling simulations used to generate the starting configurations for the US simulations. Prior to the pulling simulations, the protein molecules in the individual snapshots of the system were split into two fragments: the globular, 64-residue HTH HD core (HDC) and the disordered, cationic, 15-residue N-terminal peptide (NTP) element; only the central 14-bp region of the DNA duplex was retained. The complexes of the DNA duplex with the respective protein fragments were further treated separately (*i.e.*, each system contained only the HDC or NTP fragment of the original HD). The structures of the NTP and HDC systems in bound and unbound states are shown in Figure 2.1. A total of 10 (2 systems × 5 starting configurations) pulling simulations were then performed; a total of 29 snapshots were selected from each pulling simulation trajectory to serve as initial configurations in the US simulations.

Two MM FF combinations were used to perform the US simulations: CHARMM36m[118,120] (protein) + CHARMM36[119] (DNA) and ff14SB[112] (protein) + parmbsc1[113] (DNA); these will be referred to as CHARMM and Amber, respectively. The US simulations were performed at 4 reference temperatures T = 283.15, 293.15, 303.15, and 313.15 K for each FF and system; for each FF–temperature–system combination, five separate PMFs were calculated, each using the respective configurations derived from one of the five pulling simulations. The individual PMFs were calculated using 29 US simulations corresponding to the 29 configurations extracted from each

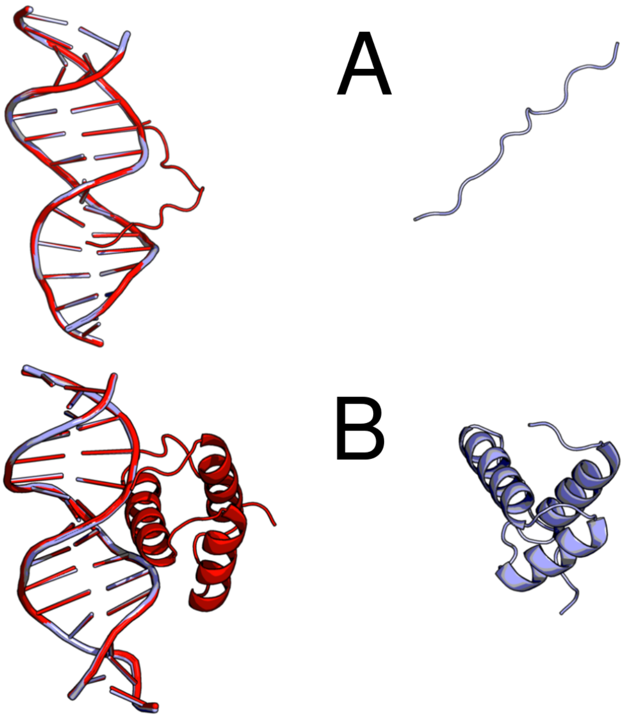


Figure 2.1: Example structures of the NTP (**A**) and HDC (**B**) variants of the protein–DNA complex at the beginning (red) and end (blue) of a pulling simulation. From Jakubec and Vondrášek (2019).

pulling simulation trajectory; therefore, a total of 2 (FFs) × 4 (temperatures) × 2 (complex variants) × 5 (pulling simulations/PMFs) × 29 (US simulation windows) $= 2,320$ US simulations were performed. The US simulations were performed under NpT conditions; the length of each US simulation was 35.0 ns. All simulations were performed using GROMACS 2016.4 in explicit solvent.

For each FF–temperature–system combination, the respective histograms of the distance between the centers of mass of the protein and DNA duplex along the z-direction d_z were aggregated from the corresponding 5 (pulling simulations) × 29 (US simulation windows) = 145 US simulations. The PMFs were calculated from the umbrella histograms using the WHAM algorithm implemented in GROMACS (`gmx wham`[169]). Data from the first 5.0 ns of each US simulation were discarded as system equilibration. The PMFs were set to zero at the distance d_z of 5.0 nm. The statistical errors were estimated using Bayesian bootstrapping of complete umbrella histograms. A total of 100 bootstraps were used to obtain the respective average bootstrapped PMFs (PMF$_{\text{bs}}$s) and error estimates. The maximal depth of the PMF minimum $\Delta G_{\text{PMF,max}}$ and the standard deviation of the ΔG_{PMF} estimate at the PMF minimum $\sigma_{\Delta G}$ were obtained from each PMF$_{\text{bs}}$ for use in the calculation of standard BFEs ΔG° using an expression derived by Doudou *et al.*[163]

Table 2.1: Values of thermodynamic parameters calculated using the CHARMM FFs for the binding of the NTP fragment toward its cognate dsDNA at the studied temperatures. ΔG_V is a correction for the standard state concentration; the other terms are explained in the text. Values of $\Delta G_{\mathrm{PMF,max}}$, $\sigma_{\Delta G}$, ΔG_V, and ΔG° are in kJ mol^{-1}.

T (K)	$\Delta G_{\mathrm{PMF,max}}$	$\sigma_{\Delta G}$	ΔG_V	ΔG°
283.15	−103.4	6.2	−3.1	−106.4
293.15	−109.4	6.5	−3.2	−112.6
303.15	−105.2	6.2	−3.3	−108.4
313.15	−107.3	6.8	−3.4	−110.7

Table 2.2: Values of thermodynamic parameters calculated using the CHARMM FFs for the binding of the HDC fragment toward its cognate dsDNA at the studied temperatures; otherwise as Table 2.1.

T (K)	$\Delta G_{\mathrm{PMF,max}}$	$\sigma_{\Delta G}$	ΔG_V	ΔG°
283.15	−67.6	5.9	−3.3	−70.9
293.15	−73.1	4.9	−3.4	−76.5
303.15	−69.6	4.7	−3.5	−73.2
313.15	−65.0	4.0	−3.7	−68.7

2.1.3 Results and discussion

The PMF$_{\mathrm{bs}}$s calculated using the CHARMM FFs for the NTP and HDC variants of the protein–DNA complex are shown in Figures A.3A and A.3B, respectively. The corresponding BFEs are shown in Tables 2.1 and 2.2, respectively. The PMF$_{\mathrm{bs}}$s calculated using the Amber FFs are shown in Figures A.3C and A.3D for the binding of the NTP and HDC variants of the protein–DNA complex, respectively; the corresponding BFEs are presented in Tables 2.3 and 2.4, respectively. All of the presented PMF$_{\mathrm{bs}}$s show good convergence in the unbound (plateau) region. The PMF$_{\mathrm{bs}}$s calculated at various temperatures for the individual FF–complex variant combinations show similar characteristics, with differences between the respective $\Delta G_{\mathrm{PMF,max}}$ values within or on the order of statistical error $\sigma_{\Delta G}$.

Fodor *et al.*[167] have measured the apparent Gibbs free energy $\Delta G'$ of the Nkx2.5 HD binding its cognate dsDNA using isothermal titration calorimetry and found it to decrease from −45.5 to −52.1 kJ mol^{-1} between 283 and 310 K. Little temperature-dependence is observed among the values of ΔG° estimated using the CHARMM FFs, while the values of ΔG° (and

Table 2.3: Values of thermodynamic parameters calculated using the Amber FFs for the binding of the NTP fragment toward its cognate dsDNA at the studied temperatures; otherwise as Table 2.1.

T (K)	$\Delta G_{\mathrm{PMF,max}}$	$\sigma_{\Delta G}$	ΔG_V	ΔG°
283.15	−89.0	5.7	−3.1	−92.1
293.15	−89.0	5.5	−3.2	−92.2
303.15	−95.2	5.7	−3.3	−98.5
313.15	−96.5	5.3	−3.4	−99.9

Table 2.4: Values of thermodynamic parameters calculated using the Amber FFs for the binding of the HDC fragment toward its cognate dsDNA at the studied temperatures; otherwise as Table 2.1.

T (K)	$\Delta G_{\mathrm{PMF,max}}$	$\sigma_{\Delta G}$	ΔG_V	ΔG°
283.15	−82.2	5.3	−3.6	−85.9
293.15	−79.5	4.9	−3.8	−83.3
303.15	−82.1	6.0	−3.9	−86.0
313.15	−88.2	5.8	−4.0	−92.2

Table 2.5: The experimentally measured Gibbs free energies of association ΔG^{a} for selected HD proteins binding their target DNA sequences. Value for Nkx2.5 (C56S) obtained from Fodor *et al.*;[167] value for vnd/NK-2 obtained from Gonzalez *et al.*;[170] remaining data from Dragan *et al.*[171] and Privalov *et al.*[172]

HD	ΔG^{a} (kJ mol^{-1})
Nkx2.5 (C56S)	−49.3
vnd/NK-2	−46.0
engrailed	−45.1
Antennapedia	−54.3
MATα2	−49.8

$\Delta G_{\mathrm{PMF,max}}$) estimated using the Amber FFs seem to decrease with increasing temperature in agreement with the experimental result. However, as the differences between the individual values of ΔG° are on the order of $\sigma_{\Delta G}$, conclusions about the temperature-dependence of ΔG° cannot be made with confidence.

Table 2.5 shows the experimentally determined values of the Gibbs free energy of association ΔG^{a} for a selection of HD proteins binding their cognate dsDNA targets at physiological conditions. The HD proteins used in these studies included the disordered N-terminal arms. Dragan *et al.*[171] have also studied the binding of the Antennapedia and NK-2 HDs lacking the N-terminal arms and found their DNA-binding affinities decreased. It can be seen that the values of ΔG° predicted from our simulations for binding of both the NTP and HDC variants of the Nkx2.5 HD are significantly overestimated (in magnitude) compared to the experimental results. The closest calculated values of ΔG°, obtained for the binding of the HDC fragment using the CHARMM FFs, exceed the experimental results for the Nkx2.5 HD by $\approx$ 20.0 kJ mol^{-1} in magnitude. As the estimated DNA-binding affinity of the disordered NTP was even greater than that of the globular HDC, it can be assumed that simulations of the full HD would yield values of ΔG° even further apart from the experimental results. Unfortunately, it was not possible to perform the pulling and US simulations with the full HD without disrupting the HD tertiary structure.

The differences between the ΔG° values calculated for the NTP and HDC systems using the CHARMM FFs ($\approx$ 35.0 kJ mol^{-1}) are much greater than the corresponding differences between the Amber FFs values ($\approx$ 10.0 kJ mol^{-1}). In addition, the binding of the NTP fragment is predicted to be more favorable using the CHARMM FFs in comparison with the Amber FFs, while the opposite is observed for the binding of the HDC. We rationalized these differences by analyzing the inter- and intramolecular interactions between and within the protein and DNA molecules over the course of the US simulation trajectories. The populations of the various interaction modes are listed in Jakubec and Vondrášek (2019); using this information, we explained the differences between the standard BFEs ΔG° estimated for the binding of the protein fragments using the two FFs as follows. The Amber FFs predict more favorable binding of the HDC compared to the CHARMM FFs on the account of higher populations of intermolecular (*i.e.*, protein–DNA) contacts. The binding of the NTP is predicted to be more favorable compared to the binding of the HDC because of significantly higher populations of intermolecular contacts observed in the respective trajectories. However, despite the higher populations of protein–DNA contacts observed in the simulations performed using the Amber FFs, the predicted affinity of the NTP toward the DNA is weaker compared to the CHARMM FFs due to an increased stabilization of the unbound state by the elevated formation of intramolecular contacts.

Limited sampling of the phase space could pose a possible limitation of this study. While we performed a block analysis to confirm that the BFEs are not significantly affected by the length of the US simulations, we could not exclude the possibility that structural changes occurring on time scales far beyond the reach of current computational capacities could lessen the discrepancies between the computed BFEs and the experimental results. In addition, the positional restraints applied to the DNA duplex in the simulations could have an effect on the BFEs, as the structural flexibility of the DNA molecule can play a major role in protein–DNA recognition.[49,173] Unfortunately, applying such restraints was necessary to preserve the DNA as an immobile reference. Further testing of the calculations of FEDs in complex systems will be necessary before it can be concluded whether systematic errors in the MM description of the systems exist, or whether sampling deficiencies pose the greatest issue.

2.2 Jakubec and Vondrášek (2020)

2.2.1 Synopsis

The Jarzynski equality (JE; Equation 2.1)[153,154] and the Crooks fluctuation theorem (CFT)[155,156] relate the FED $\Delta G_{AB} \equiv G_B - G_A$ between two equilibrium states A and B to the distributions of nonequilibrium work W_{AB} performed to drive the system between the two states:

$$\Delta G_{AB} = -\frac{1}{\beta} \log\langle \exp(-\beta W_{AB}) \rangle \qquad (2.1)$$

where $\beta = \frac{1}{RT}$, with R being the universal gas constant and T the thermodynamic temperature, W_{AB} is the work performed on the system to drive it from state A to state B along a single path, and $\langle \cdots \rangle$ denotes an average.

In this work, we explored the accuracy and efficiency of setups based on nonequilibrium pulling simulations applied to the estimation of binding affinities of DNA-binding proteins. The absolute BFEs were calculated over a range of temperatures and compared to the results obtained previously using an equilibrium method. Most importantly, we showed that realistic binding affinities can be obtained with the presented nonequilibrium approach, which also entails lower computational requirements. Errors of the BFE estimates were investigated and shown to be comparable to those observed previously.

2.2.2 Methods

The same HDC and NTP systems for which the absolute BFEs ΔG_{PMF} and their associated standard deviations $\sigma_{\Delta G_{\text{PMF}}}$ were calculated in Jakubec and Vondrášek (2019) were studied in this work using the same combinations of CHARMM and Amber FFs. First, the end (pure bound and unbound) states were simulated for both the HDC and NTP systems using a series of US simulations for a total of 1.0 μs of the respective sampling times. These equilibration simulations were initiated from system configurations selected from the previous work. A total of 105 configurations were then extracted from the equilibration trajectories for both the HDC and NTP systems to initiate the respective forward- and reverse-switching (unbinding and binding, respectively) pulling simulations.

The pulling simulations were performed with both the CHARMM and Amber FFs at the reference temperatures T = 283.15, 293.15, 303.15, and 313.15 K. The total pull distance was l = 4.2 nm in each pulling simulation. The pull rate in the reference pulling simulations was 0.0005 nm ps^{-1} and the length of the simulations was 8.4 ns. Three other pull rates were tested in combination with the CHARMM FFs and a single reference temperature of T = 293.15 K; the pull rates were 0.001, 0.00025, or 0.000125 nm ps^{-1} and the lengths of the simulations were 4.2, 16.8, or 33.6 ns, respectively.

The mechanical work performed during each pulling simulation was calculated by integrating the pull force applied along the pulling path. Gaussian probability distributions were fit to the ensembles of 105 values of work calculated from the forward- and reverse-switching simulations for each system. The means W_f, W_r and standard deviations σ_f, σ_r of the distributions were utilized to calculate the absolute BFEs ΔG_{CGI} using the Crooks Gaussian intersection (CGI) method.[174] The errors $\sigma_{\Delta G_{\text{CGI}}}$ associated with the individual ΔG_{CGI} estimates were calculated using a bootstrapping analysis. The FED estimates ΔG_{JE} were obtained by directly evaluating Equation 2.1.

2.2.3 Results and discussion

Values of absolute BFEs ΔG_{CGI} and the corresponding standard deviations $\sigma_{\Delta G_{\text{CGI}}}$ estimated using the CGI method based on the nonequilibrium pulling simulations performed using the CHARMM FFs are shown in Table 2.6. These values correspond to the pulling simulations performed at the standard pull rate and are compared to the results obtained in the previous work. The analogous absolute BFEs calculated using the Amber FFs are shown in Table 2.7. The means W_f, W_r and the standard deviations σ_f, σ_r of the work distributions obtained from the respective ensembles of forward- and

Table 2.6: Values of absolute BFEs calculated using the CHARMM FFs. Values of ΔG_{CGI}, $\sigma_{\Delta G_{\mathrm{CGI}}}$, ΔG_{PMF}, and $\sigma_{\Delta G_{\mathrm{PMF}}}$ are in kJ mol^{-1}. Values of ΔG_{PMF} and $\sigma_{\Delta G_{\mathrm{PMF}}}$ are from Jakubec and Vondrášek (2019).

	HDC				NTP			
T (K)	ΔG_{CGI}	$\sigma_{\Delta G_{\mathrm{CGI}}}$	ΔG_{PMF}	$\sigma_{\Delta G_{\mathrm{PMF}}}$	ΔG_{CGI}	$\sigma_{\Delta G_{\mathrm{CGI}}}$	ΔG_{PMF}	$\sigma_{\Delta G_{\mathrm{PMF}}}$
283.15	−56.5	9.3	−67.6	5.9	−83.8	11.4	−103.4	6.2
293.15	−37.0	7.5	−73.1	4.9	−72.7	10.0	−109.4	6.5
303.15	−50.5	7.6	−69.6	4.7	−105.1	10.3	−105.2	6.2
313.15	−42.4	6.4	−65.0	4.0	−88.8	9.3	−107.3	6.8

Table 2.7: Values of absolute BFEs calculated using the Amber FFs. Values of ΔG_{CGI}, $\sigma_{\Delta G_{\mathrm{CGI}}}$, ΔG_{PMF}, and $\sigma_{\Delta G_{\mathrm{PMF}}}$ are in kJ mol^{-1}. Values of ΔG_{PMF} and $\sigma_{\Delta G_{\mathrm{PMF}}}$ are from Jakubec and Vondrášek (2019).

	HDC				NTP			
T (K)	ΔG_{CGI}	$\sigma_{\Delta G_{\mathrm{CGI}}}$	ΔG_{PMF}	$\sigma_{\Delta G_{\mathrm{PMF}}}$	ΔG_{CGI}	$\sigma_{\Delta G_{\mathrm{CGI}}}$	ΔG_{PMF}	$\sigma_{\Delta G_{\mathrm{PMF}}}$
283.15	−42.0	8.0	−82.2	5.3	−73.3	10.1	−89.0	5.7
293.15	−54.6	9.0	−79.5	4.9	−51.4	8.5	−89.0	5.5
303.15	−64.1	9.6	−82.1	6.0	−71.0	9.4	−95.2	5.7
313.15	−54.9	7.7	−88.2	5.8	−88.5	8.6	−96.5	5.3

reverse-switching simulations are listed in Jakubec and Vondrášek (2020). Finally, the absolute BFE estimates $\Delta G_{\mathrm{JE},f}$, $\Delta G_{\mathrm{JE},r}$ obtained by evaluating Equation 2.1 using the values of work calculated from the unbinding and binding process simulations, respectively, are shown in Table 2.8.

All ΔG_{CGI} values show that the bound state dominates under equilibrium conditions. The interaction of the NTP fragment with the dsDNA binding site appears to be more favorable than that of the HDC in all but one case.

Differences are observed between the values of ΔG_{CGI} estimated for the NTP system using the two studied FFs: the values obtained using the CHARMM FFs are, on average, lower by 16.6 kJ mol^{-1} in comparison with the Amber FFs results. This is similar to the earlier PMF calculations, in which the CHARMM FFs consistently yielded values of ΔG_{PMF} lower by 15–20 kJ mol^{-1} compared to the Amber FFs. The opposite trend is observed among the values of ΔG_{CGI} estimated for the HDC system: in this case, the Amber FFs yield values that are, on average, lower by 7.3 kJ mol^{-1} in comparison with the CHARMM FFs. Again, this is consistent with the differences between the FFs observed among the corresponding values of ΔG_{PMF},

Table 2.8: Values of absolute BFEs obtained from the JE. Values of $\Delta G_{\mathrm{JE},f}$ and $\Delta G_{\mathrm{JE},r}$ are in kJ mol^{-1}.

	CHARMM				Amber			
	HDC		NTP		HDC		NTP	
T (K)	$\Delta G_{\mathrm{JE},f}$	$\Delta G_{\mathrm{JE},r}$	$\Delta G_{\mathrm{JE},f}$	$\Delta G_{\mathrm{JE},r}$	$\Delta G_{\mathrm{JE},f}$	$\Delta G_{\mathrm{JE},r}$	$\Delta G_{\mathrm{JE},f}$	$\Delta G_{\mathrm{JE},r}$
283.15	−351.6	10.0	−401.4	2.8	−351.9	28.0	−403.8	−4.9
293.15	−302.8	28.8	−370.5	8.3	−348.2	20.2	−344.0	11.8
303.15	−268.3	8.1	−428.4	0.1	−270.4	6.5	−347.7	5.3
313.15	−240.9	11.9	−376.6	−8.1	−311.3	13.2	−239.2	−5.0

although the magnitude of the differences was previously observed to be greater (14.2 kJ mol^{-1}).

In order to rationalize the differences observed between the FFs, we performed an analysis of inter- and intramolecular interactions similar to the one performed in the previous work. The qualitative differences between the two FFs were very similar to those described previously. The populations of HBs and other contacts between the protein and DNA molecules were lower in the simulations performed using the CHARMM FFs, while a greater number of protein–water HBs was observed in these. A significantly greater number of protein side chain–side chain HBs was observed in the simulations performed using the Amber FFs. These characteristics were shared by both the HDC and NTP systems. In addition, a significantly greater number of HBs between the protein main chain donor and acceptor groups was observed in the simulations of the NTP system performed using the Amber FFs.

The values of ΔG_{CGI} are in all cases greater (*i.e.*, less negative) than the corresponding values of ΔG_{PMF}. As discussed in the previous work, the values of ΔG_{PMF} are overestimated (in magnitude) compared to the experimentally determined Gibbs free energies of association ΔG^{a}. The values of ΔG estimated using the CGI method for the HDC system therefore appear to reproduce the experimental results for the HD proteins more accurately than the values extracted from the PMFs.

Similarly to the previous work, no clear dependence of the ΔG_{CGI} values on temperature is seen in either the HDC or NTP systems, with the differences between the individual observations comparable in magnitude to the statistical errors $\sigma_{\Delta G_{\mathrm{CGI}}}$ of the estimates. It must be noted, however, that the values of ΔG_{CGI} calculated at the different reference temperatures usually span a much broader range than the corresponding values of ΔG_{PMF}.

The statistical errors $\sigma_{\Delta G_{\mathrm{CGI}}}$ of the ΔG_{CGI} estimates are universally greater, but comparable in scale to the corresponding errors $\sigma_{\Delta G_{\mathrm{PMF}}}$. The magnitude of the errors does not change monotonically with the reference temperature or with the magnitude of the BFE estimates: in both Tables 2.6 and 2.7, the respective rankings of ΔG_{CGI} and $\sigma_{\Delta G_{\mathrm{CGI}}}$ differ.

An analysis of the ensembles of pulling simulations utilizing different pull rates revealed that changing the pull rate affects the ΔG_{CGI} estimates to a similar extent as changes to the reference temperature, and that the estimates do not appear to converge toward a single value as the pull rate is reduced. Importantly, however, the errors $\sigma_{\Delta G_{\mathrm{CGI}}}$ decrease as the pull rate is reduced, suggesting that a higher accuracy can be systematically achieved at the expense of further computational resources.

While seemingly more accurate BFE estimates were obtained using the nonequilibrium approach compared to the previous work, it must be highlighted that the values of ΔG_{CGI} presented in Tables 2.6 and 2.7 are several standard deviations σ_f, σ_r apart from the means W_f, W_r of the respective work distributions. For this reason, the convergence of the absolute BFE calculations is difficult to estimate, as the Gaussian distribution fitting used in the CGI method, as well as the exponential averaging present in the JE, may be sensitive to extremal values of work. This might be the reason why no clear trends are observed among the absolute BFE estimates ΔG_{CGI} obtained at different reference temperatures or pull rates.

Similarly to the previous work, the BFE estimates for the NTP system likely remain overestimated (in magnitude). This may be a sign of an insufficient sampling of the unbound state or point to deeper problems with the description of disordered proteins in the FFs used.[175] In addition, the pull rates employed in this work might be too fast to allow a physically-relevant description of the binding process.[126,176] This may be evidenced by the large energy dissipation. As a result, only the overlap of the far-tail regions of the work distributions is observed, contributing to the problematic assessment of convergence. As in other studies of FEDs, the selection of the collective variable along which the free energy profile is calculated might be problematic[131,161] and, indeed, this topic is not explored in sufficient depth in this work. Nevertheless, the qualitative agreement among the populations of the various noncovalent interactions between this and the previous study leads us to believe that a certain convergence sufficient to explore the intrinsic properties of the FFs has been reached. This is further evidenced by the relative stability of the BFE estimates across the simulations utilizing different pull rates.

Discussion

This chapter summarizes the methods and results of the works Jakubec *et al.* (2018) and Jakubec *et al.* (2019) presented in the List of publications (Page 69). As discussed in the Introduction, these works are dedicated to the exploration of the evolutionary aspects of protein organization. As these works do not necessarily concern DBPs or protein–nucleic acid interactions, only a synopsis is provided for each. Nevertheless, I include them in this overview, as the evolutionary approaches to the study of biomolecules complement the physical methods presented so far.

Jakubec *et al.* (2018)

The structural and functional independence of protein domains is reflected by their apparent modularity in the context of multidomain proteins.[7–9] In this work, we examined the coupling of evolution of domain sequences cooccurring within multidomain proteins to see if it proceeds independently or in a coordinated manner. An information-theoretic analysis of the coevolutionary signals among protein domains in multidomain arrangements was conducted. Based on the implications these signals carry, we questioned the notion of evolutionary independence of domains and examined their adaptability to their primary sequence context. For the first time, we examined coevolution as a global property of a domain pair, and introduced an appropriate continuous measure to quantify its extent. Using this measure, we showed that coevolution among protein domains is a much more widespread phenomenon than previously anticipated. This finding challenges the notion of the complete modularity of protein domains and provides new perspective on the evolution of protein sequence and function.

The multidomain proteins obtained from the UniProt reference proteomes[177] data set spanning the tree of life were examined in this work; a domain architecture was defined for each as the vector of the Pfam[145] sequence families identified within the protein sequence ordered according to their proximity to the N-terminus. For each pair of domains in the multidomain architectures, the average value of the interdomain mutual information (MI)[178] between all pairs of positions corresponding to the domains was calculated based on the MSAs of the respective domain sequences. This quantity was normalized to the range $\langle 0.0, 1.0 \rangle$ using the average entropies of the positions corresponding to the domains forming the respective pair, resulting in a global measure of evolutionary coupling between a pair of domains. The multidomain MSAs were then split at the domain boundaries, individual domain sequences were shuffled, and randomly rejoined. This way, any native evolutionary coupling among the protein domains was disrupted. The normalized average value of the interdomain MI was then recalculated for the perturbed MSAs in which fragments corresponding to individual domains almost certainly originated from different proteins.

Most importantly, we showed that the differences between the values of the interdomain MI before and after the domain sequence shuffling are in each case positive, *i.e.*, the values are always greater before the shuffling. Information is thus lost when domain sequences are paired randomly. We further showed that the normalized MI scores between randomly generated "MSAs" possessing large values of entropy at individual positions and MSAs of real protein domains differ significantly from those observed between pairs of real domains, suggesting that the evolutionary couplings observed are unlikely to be the result of random fluctuations. These results imply that, in the context of multidomain proteins, a portion of the domain sequence variation can always be attributed to the coordinated evolution among the different domains. We hypothesize that the domains in multidomain proteins can to some extent act as buffers or reservoirs of evolutionary capacity that can be utilized to either mitigate the impact of the mutations required to preserve the protein function or, alternatively, to optimize the respective functions of the individual domains. The precise mechanism through which this functional modulation is realized and its full impact on protein evolution remain to be established.

Jakubec *et al.* (2019)

Amino acid residues manifesting high levels of evolutionary conservation are often indicative of functionally significant regions within a protein structure.[9] Residues critical for protein folding, hydrophobic core stabilization, intermolecular recognition, or enzymatic activity often manifest lower substitution rates compared to the rest of the protein. Understanding how the sequence conservation profile relates in the 3D space requires a projection onto a protein structure, a potentially time-consuming process. In this work, we present 3DPatch (`https://www.skylign.org/3DPatch/`), a web application that streamlines this task by automatically generating MSAs and finding structural homologs, presenting the user with a choice of structures matching their query, annotated with the residue conservation scores. 3DPatch operates at interactive speeds and can be used without any prior knowledge of the available homologous 3D structures and without having to construct an MSA in advance.

The 3DPatch workflow is summarized in Figure 2.2. 3DPatch usually accepts a single protein sequence as an input. A search for similar sequence regions in a large sequence database is then performed using the HMMER web server,[179] yielding a MSA. A profile hidden Markov model (HMM)[180] is built from this MSA. This profile HMM is used to search for homologous sequences in the PDB archive,[11] yielding a set of 3D structures of proteins similar to the original sequence query. At the same time, the IC at individual positions of the profile HMM is calculated using the tool Skylign.[181] 3DPatch then automatically maps the IC to individual amino acid residues in the 3D structures of proteins. The user can select any of these structures to be visualized with the residue-level IC-based mark-up using an in-built molecular viewer.[182] An illustration of such a visualization is shown in Figure 2.3.

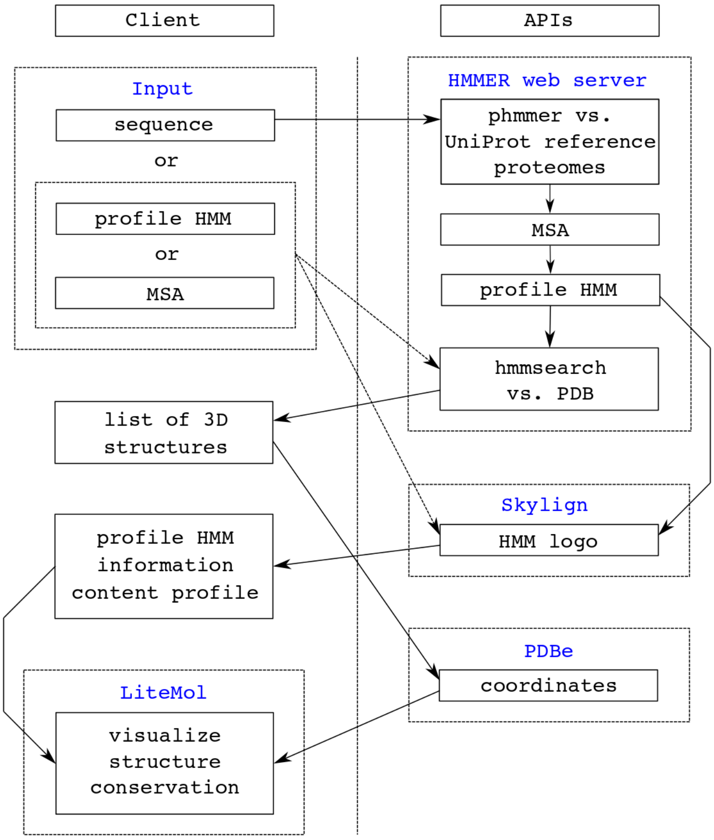


Figure 2.2: The distribution of operations between the client side and application interfaces in 3DPatch. From Jakubec *et al.* (2019).

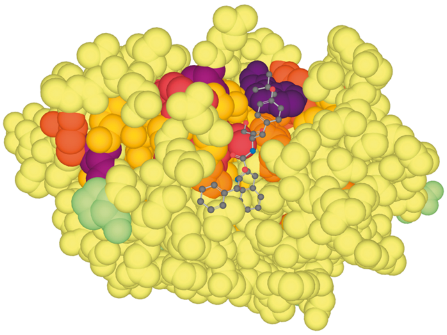


Figure 2.3: The X-ray crystallographic structure of human cathepsin L1 (PDB ID 3OF8) marked-up with the residue IC using 3DPatch. Darker colors correspond to higher conservation levels. Catalytic and binding pocket residues (center top) are clearly distinguished based on the IC. Two conserved cysteine residues forming a disulfide bridge can be seen on the left. From Jakubec *et al.* (2019).

Conclusion

In this thesis, I attempted to capture the current state of our understanding of the mechanisms of sequence-specific recognition of nucleic acids by proteins and present the results of my research in related areas. The function and structure of the essential biomacromolecules were discussed, with special attention given to the biological significance and molecular mechanisms of binding specificity in protein–DNA interactions. Methods of DNA-binding specificity determination were presented and challenges in constructing a "protein–DNA recognition code" were discussed. An overview of the methods of computational chemistry applicable to the study of biomolecules was then given, with a focus on the respective strengths and weaknesses of the individual methods.

In Chapter 1, I presented the results of my work exploring the pairwise interactions between the basic biomolecular building blocks—amino acids and nucleotides. Most importantly, it was demonstrated that selected interaction motifs statistically enriched in the 3D structures of protein–DNA complexes correspond to the most energetically favorable geometric arrangements of the respective binding partners. A qualitative agreement was shown among the IEs determined using the QM and empirical methods, making it possible to draw this conclusion. Finally, a relationship was shown between the presence of hydrophobic amino acid residues at the interface between the protein and the minor groove of the DNA and the local geometric features of the DNA helix, suggesting a universal mechanism of the DNA structure deformation.

In Chapter 2, I presented the results of my work exploring the applicability of MD simulation setups to the description of binding equilibria in protein–DNA systems. It was shown that the bound forms of the complexes are overstabilized in the MD simulations compared to the experimental results. Importantly, significant differences were discovered between the propensities of the modern MM FFs to form the various intra- and intermolecular interactions, leading to consistent, systematic differences between the BFE estimates.

Finally, in the Discussion, I presented the results of my work exploring the evolutionary aspects of protein organization. First, it was demonstrated that the evolution of protein domains in multidomain proteins proceeds to some extent in a coupled manner, challenging the traditional view of the domain modularity. Second, the tool 3DPatch facilitating the interactive exploration of evolutionary conservation in the 3D structures of proteins was presented.

I wish that the methods and results presented in this thesis may serve as a demonstration of how the synthesis of statistical, computational, and evolutionary approaches can be utilized to elucidate the mechanisms of biomolecular function.

Bibliography

[1] Alberts, B.; Johnson, A. D.; Lewis, J.; Morgan, D.; Raff, M.; Roberts, K.; Walter, P. *Molecular Biology of the Cell,* 6th ed.; W. W. Norton & Company: New York, 2014.

[2] The PyMOL Molecular Graphics System, Version 2.3 Schrödinger, LLC.

[3] Lu, X.-J.; Olson, W. K. 3DNA: A Software Package for the Analysis, Rebuilding and Visualization of Three-Dimensional Nucleic Acid Structures. *Nucleic Acids Res.* **2003**, *31*, 5108–5121.

[4] Dawson, N. L.; Lewis, T. E.; Das, S.; Lees, J. G.; Lee, D.; Ashford, P.; Orengo, C. A.; Sillitoe, I. CATH: An Expanded Resource to Predict Protein Function through Structure and Sequence. *Nucleic Acids Res.* **2017**, *45*, D289–D295.

[5] Andreeva, A.; Howorth, D.; Chothia, C.; Kulesha, E.; Murzin, A. G. SCOP2 Prototype: a New Approach to Protein Structure Mining. *Nucleic Acids Res.* **2014**, *42*, D310–D314.

[6] Andreeva, A.; Kulesha, E.; Gough, J.; Murzin, A. G. The SCOP Database in 2020: Expanded Classification of Representative Family and Superfamily Domains of Known Protein Structures. *Nucleic Acids Res.* **2020**, *48*, D376–D382.

[7] Han, J.-H.; Batey, S.; Nickson, A. A.; Teichmann, S. A.; Clarke, J. The Folding and Evolution of Multidomain Proteins. *Nat. Rev. Mol. Cell Biol.* **2007**, *8*, 319–330.

[8] Lees, J. G.; Dawson, N. L.; Sillitoe, I.; Orengo, C. A. Functional Innovation from Changes in Protein Domains and Their Combinations. *Curr. Opin. Struct. Biol.* **2016**, *38*, 44–52.

[9] Kessel, A.; Ben-Tal, N. *Introduction to Proteins: Structure, Function, and Motion,* 2nd ed.; Chapman and Hall/CRC: New York, 2018.

[10] Neidle, S. *Principles of Nucleic Acid Structure;* Academic Press: London, 2008.

[11] Berman, H. M.; Henrick, K.; Nakamura, H. Announcing the Worldwide Protein Data Bank. *Nat. Struct. Mol. Biol.* **2003**, *10*, 980.

[12] Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.

[13] Mir, S.; Alhroub, Y.; Anyango, S.; Armstrong, D. R.; Berrisford, J. M.; Clark, A. R.; Conroy, M. J.; Dana, J. M.; Deshpande, M.; Gupta, D.; Gutmanas, A.; Haslam, P.; Mak, L.; Mukhopadhyay, A.; Nadzirin,

N.; Paysan-Lafosse, T.; Sehnal, D.; Sen, S.; Smart, O. S.; Varadi, M.; Kleywegt, G. J.; Velankar, S. PDBe: Towards Reusable Data Delivery Infrastructure at Protein Data Bank in Europe. *Nucleic Acids Res.* **2018**, *46*, D486–D492.

[14] Carlberg, C.; Molnár, F. *Mechanisms of Gene Regulation,* 2nd ed.; Springer Netherlands: Dordrecht, 2016.

[15] Oda, M.; Furukawa, K.; Ogata, K.; Sarai, A.; Nakamura, H. Thermodynamics of Specific and Non-Specific DNA Binding by the c-Myb DNA-Binding Domain. *J. Mol. Biol.* **1998**, *276*, 571–590.

[16] Jelesarov, I.; Bosshard, H. R. Isothermal Titration Calorimetry and Differential Scanning Calorimetry as Complementary Tools to Investigate the Energetics of Biomolecular Recognition. *J. Mol. Recognit.* **1999**, *12*, 3–18.

[17] Dragan, A. I.; Klass, J.; Read, C.; Churchill, M. E. A.; Crane-Robinson, C.; Privalov, P. L. DNA Binding of a Non-Sequence-Specific HMG-D Protein Is Entropy Driven with a Substantial Non-Electrostatic Contribution. *J. Mol. Biol.* **2003**, *331*, 795–813.

[18] Dragan, A. I.; Frank, L.; Liu, Y.; Makeyeva, E. N.; Crane-Robinson, C.; Privalov, P. L. Thermodynamic Signature of GCN4-bZIP Binding to DNA Indicates the Role of Water in Discriminating Between the AP-1 and ATF/CREB Sites. *J. Mol. Biol.* **2004**, *343*, 865–878.

[19] Jerabek-Willemsen, M.; André, T.; Wanner, R.; Roth, H. M.; Duhr, S.; Baaske, P.; Breitsprecher, D. MicroScale Thermophoresis: Interaction Analysis and Beyond. *J. Mol. Struct.* **2014**, *1077*, 101–113.

[20] Wang, W.; Ding, J.; Zhang, Y.; Hu, Y.; Wang, D.-C. Structural Insights into the Unique Single-Stranded DNA-Binding Mode of *Helicobacter pylori* DprA. *Nucleic Acids Res.* **2014**, *42*, 3478–3491.

[21] Stormo, G. D.; Fields, D. S. Specificity, Free Energy and Information Content in Protein–DNA Interactions. *Trends Biochem. Sci.* **1998**, *23*, 109–113.

[22] Jolma, A.; Yan, J.; Whitington, T.; Toivonen, J.; Nitta, K. R.; Rastas, P.; Morgunova, E.; Enge, M.; Taipale, M.; Wei, G.; Palin, K.; Vaquerizas, J. M.; Vincentelli, R.; Luscombe, N. M.; Hughes, T. R.; Lemaire, P.; Ukkonen, E.; Kivioja, T.; Taipale, J. DNA-Binding Specificities of Human Transcription Factors. *Cell* **2013**, *152*, 327–339.

[23] Newburger, D. E.; Bulyk, M. L. UniPROBE: An Online Database of Protein Binding Microarray Data on Protein–DNA Interactions. *Nucleic Acids Res.* **2009**, *37*, D77–D82.

[24] Orenstein, Y.; Shamir, R. Modeling Protein–DNA Binding via High-Throughput *in Vitro* Technologies. *Brief. Funct. Genomics* **2016**, *16*, 171–180.

[25] Furey, T. S. ChIP–seq and Beyond: New and Improved Methodologies to Detect and Characterize Protein–DNA Interactions. *Nat. Rev. Genet.* **2012**, *13*, 840–852.

[26] Wheeler, T. J.; Clements, J.; Finn, R. D. Skylign: A Tool for Creating Informative, Interactive Logos Representing Sequence Alignments and Profile Hidden Markov Models. *BMC Bioinformatics* **2014**, *15*, 7.

[27] Khan, A.; Fornes, O.; Stigliani, A.; Gheorghe, M.; Castro-Mondragon, J. A.; van der Lee, R.; Bessy, A.; Chèneby, J.; Kulkarni, S. R.; Tan, G.; Baranasic, D.; Arenillas, D. J.; Sandelin, A.; Vandepoele, K.; Lenhard, B.; Ballester, B.; Wasserman, W. W.; Parcy, F.; Mathelier, A. JASPAR 2018: Update of the Open-Access Database of Transcription Factor Binding Profiles and Its Web Framework. *Nucleic Acids Res.* **2018**, *46*, D260–D266.

[28] Kulakovskiy, I. V.; Vorontsov, I. E.; Yevshin, I. S.; Sharipov, R. N.; Fedorova, A. D.; Rumynskiy, E. I.; Medvedeva, Y. A.; Magana-Mora, A.; Bajic, V. B.; Papatsenko, D. A.; Kolpakov, F. A.; Makeev, V. J. HOCOMOCO: Towards a Complete Collection of Transcription Factor Binding Models for Human and Mouse via Large-Scale ChIP–seq Analysis. *Nucleic Acids Res.* **2018**, *46*, D252–D259.

[29] Schneider, T. D.; Stormo, G. D.; Gold, L.; Ehrenfeucht, A. Information Content of Binding Sites on Nucleotide Sequences. *J. Mol. Biol.* **1986**, *188*, 415–431.

[30] Berg, O. G.; von Hippel, P. H. Selection of DNA Binding Sites by Regulatory Proteins: Statistical-Mechanical Theory and Application to Operators and Promoters. *J. Mol. Biol.* **1987**, *193*, 723–743.

[31] Berg, O. G.; von Hippel, P. H. Selection of DNA Binding Sites by Regulatory Proteins II: The Binding Specificity of Cyclic AMP Receptor Protein to Recognition Sites. *J. Mol. Biol.* **1988**, *200*, 709–723.

[32] Harvey, R. P. *NK-2* Homeobox Genes and Heart Development. *Dev. Biol.* **1996**, *178*, 203–216.

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

[34] Vogt, P. K. Fortuitous Convergences: The Beginnings of JUN. *Nat. Rev. Cancer* **2002**, *2*, 465–469.

[35] Verger, A.; Duterque-Coquillaud, M. When Ets Transcription Factors Meet Their Partners. *BioEssays* **2002**, *24*, 362–370.

[36] Yeh, E.; Cunningham, M.; Arnold, H.; Chasse, D.; Monteith, T.; Ivaldi, G.; Hahn, W. C.; Stukenberg, P. T.; Shenolikar, S.; Uchida, T.; Counter, C. M.; Nevins, J. R.; Means, A. R.; Sears, R. A Signalling Pathway Controlling c-Myc Degradation That Impacts Oncogenic Transformation of Human Cells. *Nat. Cell Biol.* **2004**, *6*, 308–318.

[37] Barrera, L. A.; Vedenko, A.; Kurland, J. V.; Rogers, J. M.; Gisselbrecht, S. S.; Rossin, E. J.; Woodard, J.; Mariani, L.; Kock, K. H.; Inukai, S.; Siggers, T.; Shokri, L.; Gordân, R.; Sahni, N.; Cotsapas, C.; Hao, T.; Yi, S.; Kellis, M.; Daly, M. J.; Vidal, M.; Hill, D. E.; Bulyk, M. L. Survey of Variation in Human Transcription Factors Reveals Prevalent DNA Binding Changes. *Science* **2016**, *351*, 1450–1454.

[38] Deplancke, B.; Alpern, D.; Gardeux, V. The Genetics of Transcription Factor DNA Binding Variation. *Cell* **2016**, *166*, 538–554.

[39] Lambert, S. A.; Jolma, A.; Campitelli, L. F.; Das, P. K.; Yin, Y.; Albu, M.; Chen, X.; Taipale, J.; Hughes, T. R.; Weirauch, M. T. The Human Transcription Factors. *Cell* **2018**, *172*, 650–665.

[40] Talanian, R. V.; McKnight, C. J.; Kim, P. S. Sequence-Specific DNA Binding by a Short Peptide Dimer. *Science* **1990**, *249*, 769–771.

[41] Talanian, R. V.; McKnight, C. J.; Rutkowski, R.; Kim, P. S. Minimum Length of a Sequence-Specific DNA Binding Peptide. *Biochemistry* **1992**, *31*, 6871–6875.

[42] Shang, Z.; Isaac, V. E.; Li, H.; Patel, L.; Catron, K. M.; Curran, T.; Montelione, G. T.; Abate, C. Design of a "minimAl" Homeodomain: The N-terminal Arm Modulates DNA Binding Affinity and Stabilizes Homeodomain Structure. *Proc. Natl. Acad. Sci. U. S. A.* **1994**, *91*, 8373–8377.

[43] Pellegrini, M.; Ebright, R. H. Artificial Sequence-Specific DNA Binding Peptides: Branched-Chain Basic Regions. *J. Am. Chem. Soc.* **1996**, *118*, 5831–5835.

[44] Caamaño, A. M.; Vázquez, M. E.; Martínez-Costas, J.; Castedo, L.; Mascareñas, J. L. A Light-Modulated Sequence-Specific DNA-Binding Peptide. *Angew. Chemie Int. Ed.* **2000**, *39*, 3104–3107.

[45] Gaj, T.; Gersbach, C. A.; Barbas III, C. F. ZFN, TALEN, and CRISPR/Cas-Based Methods for Genome Engineering. *Trends Biotechnol.* **2013**, *31*, 397–405.

[46] Lange, O. D.; Wolf, C.; Dietze, J.; Elsaesser, J.; Morbitzer, R.; Lahaye, T. Programmable DNA-Binding Proteins from *Burkholderia* Provide a Fresh Perspective on the TALE-like Repeat Domain. *Nucleic Acids Res.* **2014**, *42*, 7436–7449.

[47] Joyce, A. P.; Zhang, C.; Bradley, P.; Havranek, J. J. Structure-Based Modeling of Protein: DNA Specificity. *Brief. Funct. Genomics* **2015**, *14*, 39–49.

[48] Benos, P. V.; Lapedes, A. S.; Stormo, G. D. Is There a Code for Protein–DNA Recognition? Probab(ilistical)ly... *BioEssays* **2002**, *24*, 466–475.

[49] Rohs, R.; Jin, X.; West, S. M.; Joshi, R.; Honig, B.; Mann, R. S. Origins of Specificity in Protein–DNA Recognition. *Annu. Rev. Biochem.* **2010**, *79*, 233–269.

[50] Slattery, M.; Zhou, T.; Yang, L.; Dantas Machado, A. C.; Gordân, R.; Rohs, R. Absence of a Simple Code: How Transcription Factors Read the Genome. *Trends Biochem. Sci.* **2014**, *39*, 381–399.

[51] Watson, J. D.; Crick, F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **1953**, *171*, 737–738.

[52] Shore, P.; Whitmarsh, A. J.; Bhaskaran, R.; Davis, R. J.; Waltho, J. P.; Sharrocks, A. D. Determinants of DNA-Binding Specificity of ETS-Domain Transcription Factors. *Mol. Cell. Biol.* **1996**, *16*, 3338–3349.

[53] Grishin, A. V.; Alexeevski, A. V.; Spirin, S. A.; Kariagin, A. S. Conserved Structural Features of ETS Domain–DNA Complexes. *Mol. Biol. (Mosk.)* **2009**, *43*, 666–674.

[54] Nakagawa, S.; Gisselbrecht, S. S.; Rogers, J. M.; Hartl, D. L.; Bulyk, M. L. DNA-Binding Specificity Changes in the Evolution of Forkhead Transcription Factors. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 12349–12354.

[55] Vacha, P.; Zuskova, I.; Bumba, L.; Herman, P.; Vecer, J.; Obsilova, V.; Obsil, T. Detailed Kinetic Analysis of the Interaction between the FOXO4–DNA-Binding Domain and DNA. *Biophys. Chem.* **2013**, *184*, 68–78.

[56] Berger, M. F.; Badis, G.; Gehrke, A. R.; Talukder, S.; Philippakis, A. A.; Peña-Castillo, L.; Alleyne, T. M.; Mnaimneh, S.; Botvinnik, O. B.; Chan, E. T.; Khalid, F.; Zhang, W.; Newburger, D. E.; Jaeger, S. A.; Morris, Q. D.; Bulyk, M. L.; Hughes, T. R. Variation in Homeodomain DNA Binding Revealed by High-Resolution Analysis of Sequence Preferences. *Cell* **2008**, *133*, 1266–1276.

[57] Noyes, M. B.; Christensen, R. G.; Wakabayashi, A.; Stormo, G. D.; Brodsky, M. H.; Wolfe, S. A. Analysis of Homeodomain Specificities Allows the Family-Wide Prediction of Preferred Recognition Sites. *Cell* **2008**, *133*, 1277–1289.

[58] Dror, I.; Zhou, T.; Mandel-Gutfreund, Y.; Rohs, R. Covariation between Homeodomain Transcription Factors and the Shape of Their DNA Binding Sites. *Nucleic Acids Res.* **2014**, *42*, 430–441.

[59] Seeliger, D.; Buelens, F. P.; Goette, M.; de Groot, B. L.; Grubmüller, H. Towards Computional Specificity Screening of DNA-Binding Proteins. *Nucleic Acids Res.* **2011**, *39*, 8281–8290.

[60] Stricker, S. H.; Köferle, A.; Beck, S. From Profiles to Function in Epigenomics. *Nat. Rev. Genet.* **2017**, *18*, 51–66.

[61] Rohs, R.; West, S. M.; Liu, P.; Honig, B. Nuance in the Double-Helix and Its Role in Protein–DNA Recognition. *Curr. Opin. Struct. Biol.* **2009**, *19*, 171–177.

[62] Pérez, A.; Castellazzi, C. L.; Battistini, F.; Collinet, K.; Flores, O.; Deniz, O.; Ruiz, M. L.; Torrents, D.; Eritja, R.; Soler-López, M.; Orozco, M. Impact of Methylation on the Physical Properties of DNA. *Biophys. J.* **2012**, *102*, 2140–2148.

[63] Jones, S.; van Heyningen, P.; Berman, H. M.; Thornton, J. M. Protein–DNA Interactions: A Structural Analysis. *J. Mol. Biol.* **1999**, *287*, 877–896.

[64] Schneider, B.; Černý, J.; Svozil, D.; Čech, P.; Gelly, J.-C.; de Brevern, A. G. Bioinformatic Analysis of the Protein/DNA Interface. *Nucleic Acids Res.* **2014**, *42*, 3381–3394.

[65] Corona, R. I.; Guo, J. Statistical Analysis of Structural Determinants for Protein–DNA-Binding Specificity. *Proteins Struct. Funct. Bioinforma.* **2016**, *84*, 1147–1161.

[66] Seeman, N. C.; Rosenberg, J. M.; Rich, A. Sequence-Specific Recognition of Double Helical Nucleic Acids by Proteins. *Proc. Natl. Acad. Sci. U. S. A.* **1976**, *73*, 804–808.

[67] Lundbäck, T.; Härd, T. Sequence-Specific DNA-Binding Dominated by Dehydration. *Proc. Natl. Acad. Sci. U. S. A.* **1996**, *93*, 4754–4759.

[68] Morton, C. J.; Ladbury, J. E. Water Mediated Protein–DNA Interactions: The Relationship of Thermodynamics to Structural Detail. *Protein Sci.* **1996**, *5*, 2115–2118.

[69] Schwabe, J. W. R. The Role of Water in Protein–DNA Interactions. *Curr. Opin. Struct. Biol.* **1997**, *7*, 126–134.

[70] Mandel-Gutfreund, Y.; Margalit, H. Quantitative Parameters for Amino Acid–Base Interaction: Implications for Prediction of Protein–DNA Binding Sites. *Nucleic Acids Res.* **1998**, *26*, 2306–2312.

[71] Luscombe, N. M.; Laskowski, R. A.; Thornton, J. M. Amino Acid–Base Interactions: A Three-Dimensional Analysis of Protein–DNA Interactions at an Atomic Level. *Nucleic Acids Res.* **2001**, *29*, 2860–2874.

[72] Davey, C. A.; Sargent, D. F.; Luger, K.; Maeder, A. W.; Richmond, T. J. Solvent Mediated Interactions in the Structure of the Nucleosome Core Particle at 1.9 Å Resolution. *J. Mol. Biol.* **2002**, *319*, 1097–1113.

[73] Cheng, A. C.; Chen, W. W.; Fuhrmann, C. N.; Frankel, A. D. Recognition of Nucleic Acid Bases and Base-Pairs by Hydrogen Bonding to Amino Acid Side-Chains. *J. Mol. Biol.* **2003**, *327*, 781–796.

[74] Coulocheri, S. A.; Pigis, D. G.; Papavassiliou, K. A.; Papavassiliou, A. G. Hydrogen Bonds in Protein–DNA Complexes: Where Geometry Meets Plasticity. *Biochimie* **2007**, *89*, 1291–1303.

[75] Kondo, J.; Westhof, E. Classification of Pseudo Pairs between Nucleotide Bases and Amino Acids by Analysis of Nucleotide–Protein Complexes. *Nucleic Acids Res.* **2011**, *39*, 8628–8637.

[76] Olson, W. K.; Gorin, A. A.; Lu, X.-J.; Hock, L. M.; Zhurkin, V. B. DNA Sequence-Dependent Deformability Deduced from Protein–DNA Crystal Complexes. *Proc. Natl. Acad. Sci. U. S. A.* **1998**, *95*, 11163–11168.

[77] Travers, A. Recognition of Distorted DNA Structures by HMG Domains. *Curr. Opin. Struct. Biol.* **2000**, *10*, 102–109.

[78] Zhang, Y.; Xi, Z.; Hegde, R. S.; Shakked, Z.; Crothers, D. M. Predicting Indirect Readout Effects in Protein–DNA Interactions. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 8337–8341.

[79] Rohs, R.; West, S. M.; Sosinsky, A.; Liu, P.; Mann, R. S.; Honig, B. The Role of DNA Shape in Protein–DNA Recognition. *Nature* **2009**, *461*, 1248–1253.

[80] Yang, L.; Zhou, T.; Dror, I.; Mathelier, A.; Wasserman, W. W.; Gordân, R.; Rohs, R. TFBSshape: A Motif Database for DNA Shape Features of Transcription Factor Binding Sites. *Nucleic Acids Res.* **2014**, *42*, D148–D155.

[81] Yang, L.; Orenstein, Y.; Jolma, A.; Yin, Y.; Taipale, J.; Shamir, R.; Rohs, R. Transcription Factor Family-Specific DNA Shape Readout Revealed by Quantitative Specificity Models. *Mol. Syst. Biol.* **2017**, *13*, 910.

[82] Contreras-Moreira, B.; Sancho, J.; Angarica, V. E. Comparison of DNA Binding across Protein Superfamilies. *Proteins Struct. Funct. Bioinforma.* **2010**, *78*, 52–62.

[83] Affolter, M.; Slattery, M.; Mann, R. S. A Lexicon for Homeodomain-DNA Recognition. *Cell* **2008**, *133*, 1133–1135.

[84] Tóth-Petróczy, A.; Simon, I.; Fuxreiter, M.; Levy, Y. Disordered Tails of Homeodomains Facilitate DNA Recognition by Providing a Trade-Off between Folding and Specific Binding. *J. Am. Chem. Soc.* **2009**, *131*, 15084–15085.

[85] Bürglin, T. R.; Affolter, M. Homeodomain Proteins: An Update. *Chromosoma* **2016**, *125*, 497–521.

[86] Mulliken, R. S. Electronic Population Analysis on LCAO–MO Molecular Wave Functions. I. *J. Chem. Phys.* **1955**, *23*, 1833–1840.

[87] Müller-Dethlefs, K.; Hobza, P. Noncovalent Interactions: A Challenge for Experiment and Theory. *Chem. Rev.* **2000**, *100*, 143–168.

[88] Černý, J.; Hobza, P. Non-Covalent Interactions in Biomacromolecules. *Phys. Chem. Chem. Phys.* **2007**, *9*, 5291–5303.

[89] Riley, K. E.; Pitoňák, M.; Jurečka, P.; Hobza, P. Stabilization and Structure Calculations for Noncovalent Interactions in Extended Molecular Systems Based on Wave Function and Density Functional Theories. *Chem. Rev.* **2010**, *110*, 5023–5063.

[90] Šponer, J.; Šponer, J. E.; Mládek, A.; Banáš, P.; Jurečka, P.; Otyepka, M. How to Understand Quantum Chemical Computations on DNA and RNA Systems? A Practical Guide for Non-Specialists. *Methods* **2013**, *64*, 3–11.

[91] Jurečka, P.; Šponer, J.; Černý, J.; Hobza, P. Benchmark Database of Accurate (MP2 and CCSD(T) Complete Basis Set Limit) Interaction Energies of Small Model Complexes, DNA Base Pairs, and Amino Acid Pairs. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985–1993.

[92] Černý, J.; Pitoňák, M.; Riley, K. E.; Hobza, P. Complete Basis Set Extrapolation and Hybrid Schemes for Geometry Gradients of Noncovalent Complexes. *J. Chem. Theory Comput.* **2011**, *7*, 3924–3934.

[93] Jeziorski, B.; Moszynski, R.; Szalewicz, K. Perturbation Theory Approach to Intermolecular Potential Energy Surfaces of van der Waals Complexes. *Chem. Rev.* **1994**, *94*, 1887–1930.

[94] Szalewicz, K. Symmetry-Adapted Perturbation Theory of Intermolecular Forces. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2012**, *2*, 254–272.

[95] Grimme, S. Accurate Description of van der Waals Complexes by Density Functional Theory Including Empirical Corrections. *J. Comput. Chem.* **2004**, *25*, 1463–1473.

[96] Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A Consistent and Accurate *ab Initio* Parametrization of Density Functional Dispersion Correction (DFT-D) for the 94 Elements H-Pu. *J. Chem. Phys.* **2010**, *132*, 154104.

[97] Šponer, J.; Hobza, P. Nonplanar Geometries of DNA Bases. *Ab Initio* Second-Order Møller–Plesset Study. *J. Phys. Chem.* **1994**, *98*, 3161–3164.

[98] Bonaccorsi, R.; Pullman, A.; Scrocco, E.; Tomasi, J. The Molecular Electrostatic Potentials for the Nucleic Acid Bases: Adenine, Thymine, and Cytosine. *Theor. Chim. Acta* **1972**, *24*, 51–60.

[99] Murray, J. S.; Peralta-Inga, Z.; Politzer, P.; Ekanayake, K.; Lebreton, P. Computational Characterization of Nucleotide Bases: Molecular Surface Electrostatic Potentials and Local Ionization Energies, and

Local Polarization Energies. *Int. J. Quantum Chem.* **2001**, *83*, 245–254.

[100] Šponer, J.; Hobza, P. MP2 and CCSD(T) Study on Hydrogen Bonding, Aromatic Stacking and Nonaromatic Stacking. *Chem. Phys. Lett.* **1997**, *267*, 263–270.

[101] Šponer, J.; Leszczynski, J.; Hobza, P. Electronic Properties, Hydrogen Bonding, Stacking, and Cation Binding of DNA and RNA Bases. *Biopolymers* **2002**, *61*, 3–31.

[102] Šponer, J.; Jurečka, P.; Hobza, P. Accurate Interaction Energies of Hydrogen-Bonded Nucleic Acid Base Pairs. *J. Am. Chem. Soc.* **2004**, *126*, 10142–10151.

[103] Šponer, J.; Gabb, H. A.; Leszczynski, J.; Hobza, P. Base-Base and Deoxyribose-Base Stacking Interactions in B-DNA and Z-DNA: A Quantum-Chemical Study. *Biophys. J.* **1997**, *73*, 76–87.

[104] Hobza, P.; Šponer, J. Toward True DNA Base-Stacking Energies: MP2, CCSD(T), and Complete Basis Set Calculations. *J. Am. Chem. Soc.* **2002**, *124*, 11802–11808.

[105] Mládek, A.; Krepl, M.; Svozil, D.; Čech, P.; Otyepka, M.; Banáš, P.; Zgarbová, M.; Jurečka, P.; Šponer, J. Benchmark Quantum-Chemical Calculations on a Complete Set of Rotameric Families of the DNA Sugar–Phosphate Backbone and Their Comparison with Modern Density Functional Theory. *Phys. Chem. Chem. Phys.* **2013**, *15*, 7295–7310.

[106] Berka, K.; Laskowski, R. A.; Riley, K. E.; Hobza, P.; Vondrášek, J. Representative Amino Acid Side Chain Interactions in Proteins. A Comparison of Highly Accurate Correlated *ab Initio* Quantum Chemical and Empirical Potential Procedures. *J. Chem. Theory Comput.* **2009**, *5*, 982–992.

[107] Berka, K.; Laskowski, R. A.; Hobza, P.; Vondrášek, J. Energy Matrix of Structurally Important Side-Chain/Side-Chain Interactions in Proteins. *J. Chem. Theory Comput.* **2010**, *6*, 2191–2203.

[108] Cheng, A. C.; Frankel, A. D. *Ab Initio* Interaction Energies of Hydrogen-Bonded Amino Acid Side Chain–Nucleic Acid Base Interactions. *J. Am. Chem. Soc.* **2004**, *126*, 434–435.

[109] Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An N·log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98*, 10089–10092.

[110] Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.* **1995**, *103*, 8577–8593.

[111] Leach, A. R. *Molecular Modelling: Principles and Applications,* 2nd ed.; Pearson Education: Harlow, U.K., 2001.

[112] Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **2015**, *11*, 3696–3713.

[113] Ivani, I.; Dans, P. D.; Noy, A.; Pérez, A.; Faustino, I.; Hospital, A.; Walther, J.; Andrio, P.; Goñi, R.; Balaceanu, A.; Portella, G.; Battistini, F.; Gelpí, J. L.; González, C.; Vendruscolo, M.; Laughton, C. A.; Harris, S. A.; Case, D. A.; Orozco, M. Parmbsc1: A Refined Force Field for DNA Simulations. *Nat. Methods* **2015**, *13*, 55–58.

[114] Zgarbová, M.; Šponer, J.; Otyepka, M.; Cheatham, T. E.; Galindo-Murillo, R.; Jurečka, P. Refinement of the Sugar–Phosphate Backbone Torsion Beta for AMBER Force Fields Improves the Description of Z- and B-DNA. *J. Chem. Theory Comput.* **2015**, *11*, 5723–5736.

[115] Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.

[116] Wang, J.; Cieplak, P.; Kollman, P. A. How Well Does a Restrained Electrostatic Potential (RESP) Model Perform in Calculating Conformational Energies of Organic and Biological Molecules? *J. Comput. Chem.* **2000**, *21*, 1049–1074.

[117] Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015**, *1–2*, 19–25.

[118] Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E. M.; Mittal, J.; Feig, M.; MacKerell, A. D., Jr. Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone ϕ, ψ and Side-Chain χ_1 and χ_2 Dihedral Angles. *J. Chem. Theory Comput.* **2012**, *8*, 3257–3273.

[119] Hart, K.; Foloppe, N.; Baker, C. M.; Denning, E. J.; Nilsson, L.; MacKerell, A. D., Jr. Optimization of the CHARMM Additive Force Field for DNA: Improved Treatment of the BI/BII Conformational Equilibrium. *J. Chem. Theory Comput.* **2012**, *8*, 348–362.

[120] Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B. L.; Grubmüller, H.; MacKerell, A. D., Jr. CHARMM36m: An Improved Force Field for Folded and Intrinsically Disordered Proteins. *Nat. Methods* **2017**, *14*, 71–73.

[121] Hobza, P.; Hubálek, F.; Kabeláč, M.; Mejzlík, P.; Šponer, J.; Vondrášek, J. Ability of Empirical Potentials (AMBER, CHARMM, CVFF, OPLS, Poltev) and Semi-Empirical Quantum Chemical Methods (AM1, MNDO/M, PM3) to Describe H-Bonding in DNA Base Pairs; Comparison with *ab Initio* Results. *Chem. Phys. Lett.* **1996**, *257*, 31–35.

[122] Zgarbová, M.; Otyepka, M.; Šponer, J.; Hobza, P.; Jurečka, P. Large-Scale Compensation of Errors in Pairwise-Additive Empirical Force Fields: Comparison of AMBER Intermolecular Terms with Rigorous DFT-SAPT Calculations. *Phys. Chem. Chem. Phys.* **2010**, *12*, 10476–10493.

[123] Bussi, G.; Donadio, D.; Parrinello, M. Canonical Sampling through Velocity Rescaling. *J. Chem. Phys.* **2007**, *126*, 1–7.

[124] Parrinello, M.; Rahman, A. Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method. *J. Appl. Phys.* **1981**, *52*, 7182–7190.

[125] Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular Dynamics with Coupling to an External Bath. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

[126] Isralewitz, B.; Gao, M.; Schulten, K. Steered Molecular Dynamics and Mechanical Functions of Proteins. *Curr. Opin. Struct. Biol.* **2001**, *11*, 224–230.

[127] Perilla, J. R.; Goh, B. C.; Cassidy, C. K.; Liu, B.; Bernardi, R. C.; Rudack, T.; Yu, H.; Wu, Z.; Schulten, K. Molecular Dynamics Simulations of Large Macromolecular Complexes. *Curr. Opin. Struct. Biol.* **2015**, *31*, 64–74.

[128] Tomasi, J.; Mennucci, B.; Cammi, R. Quantum Mechanical Continuum Solvation Models. *Chem. Rev.* **2005**, *105*, 2999–3094.

[129] Kührová, P.; Otyepka, M.; Šponer, J.; Banáš, P. Are Waters around RNA More than Just a Solvent? – An Insight from Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2014**, *10*, 401–411.

[130] Šponer, J.; Banáš, P.; Jurečka, P.; Zgarbová, M.; Kührová, P.; Havrila, M.; Krepl, M.; Stadlbauer, P.; Otyepka, M. Molecular Dynamics Simulations of Nucleic Acids. From Tetranucleotides to the Ribosome. *J. Phys. Chem. Lett.* **2014**, *5*, 1771–1782.

[131] Šponer, J.; Krepl, M.; Banáš, P.; Kührová, P.; Zgarbová, M.; Jurečka, P.; Havrila, M.; Otyepka, M. How to Understand Atomistic Molecular Dynamics Simulations of RNA and Protein–RNA Complexes? *Wiley Interdiscip. Rev. RNA* **2017**, *8*, e1405.

[132] Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J. W.; Wang, J.; Kollman, P. A. A Point-Charge Force Field for Molecular Mechanics Simulations of Proteins Based on Condensed-Phase Quantum Mechanical Calculations. *J. Comput. Chem.* **2003**, *24*, 1999–2012.

[133] Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. Improved Side-Chain Torsion Potentials for the Amber ff99SB Protein Force Field. *Proteins* **2010**, *78*, 1950–1958.

[134] MacKerell, A. D., Jr.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiórkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.

[135] MacKerell, A. D., Jr.; Banavali, N.; Foloppe, N. Development and Current Status of the CHARMM Force Field for Nucleic Acids. *Biopolymers* **2000**, *56*, 257–265.

[136] Lee, C.; Yang, W.; Parr, R. G. Development of the Colle–Salvetti Correlation-Energy Formula into a Functional of the Electron Density. *Phys. Rev. B* **1988**, *37*, 785–789.

[137] Becke, A. D. Density-Functional Thermochemistry. I. The Effect of the Exchange-only Gradient Correction. *J. Chem. Phys.* **1992**, *96*, 2155–2160.

[138] Schäfer, A.; Huber, C.; Ahlrichs, R. Fully Optimized Contracted Gaussian Basis Sets of Triple Zeta Valence Quality for Atoms Li to Kr. *J. Chem. Phys.* **1994**, *100*, 5829–5835.

[139] Weigend, F.; Ahlrichs, R. Balanced Basis Sets of Split Valence, Triple Zeta Valence and Quadruple Zeta Valence Quality for H to Rn: Design and Assessment of Accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297–3305.

[140] Ahlrichs, R.; Bär, M.; Häser, M.; Horn, H.; Kölmel, C. Electronic Structure Calculations on Workstation Computers: The Program System TURBOMOLE. *Chem. Phys. Lett.* **1989**, *162*, 165–169.

[141] Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C. The GB/SA Continuum Model for Solvation. A Fast Analytical Method for the Calculation of Approximate Born Radii. *J. Phys. Chem. A* **1997**, *101*, 3005–3014.

[142] de Ruiter, A.; Zagrovic, B. Absolute Binding-Free Energies between Standard RNA/DNA Nucleobases and Amino-Acid Sidechain Analogs in Different Environments. *Nucleic Acids Res.* **2015**, *43*, 708–718.

[143] Schäfer, A.; Klamt, A.; Sattel, D.; Lohrenz, J. C. W.; Eckert, F. COSMO Implementation in TURBOMOLE: Extension of an Efficient Quantum Chemical Code towards Liquid Systems. *Phys. Chem. Chem. Phys.* **2000**, *2*, 2187–2193.

[144] Wang, G.; Dunbrack, R. L., Jr. PISCES: A Protein Sequence Culling Server. *Bioinformatics* **2003**, *19*, 1589–1591.

[145] Finn, R. D.; Coggill, P.; Eberhardt, R. Y.; Eddy, S. R.; Mistry, J.; Mitchell, A. L.; Potter, S. C.; Punta, M.; Qureshi, M.; Sangrador-Vegas, A.; Salazar, G. A.; Tate, J.; Bateman, A. The Pfam Protein Families Database: Towards a More Sustainable Future. *Nucleic Acids Res.* **2016**, *44*, D279–D285.

[146] Kim, Y.; Geiger, J. H.; Hahn, S.; Sigler, P. B. Crystal Structure of a Yeast TBP/TATA-Box Complex. *Nature* **1993**, *365*, 512–520.

[147] Bewley, C. A.; Gronenborn, A. M.; Clore, G. M. Minor Groove-Binding Architectural Proteins: Structure, Function, and DNA Recognition. *Annu. Rev. Biophys. Biomol. Struct.* **1998**, *27*, 105–131.

[148] Mann, H. B.; Whitney, D. R. On a Test of Whether One of Two Random Variables Is Stochastically Larger than the Other. *Ann. Math. Stat.* **1947**, *18*, 50–60.

[149] Kabsch, W.; Sander, C. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers* **1983**, *22*, 2577–2637.

[150] van der Vaart, A. Coupled Binding–Bending–Folding: The Complex Conformational Dynamics of Protein-DNA Binding Studied by Atomistic Molecular Dynamics Simulations. *Biochim. Biophys. Acta* **2015**, *1850*, 1091–1098.

[151] Kirkwood, J. G. Statistical Mechanics of Fluid Mixtures. *J. Chem. Phys.* **1935**, *3*, 300–313.

[152] Zwanzig, R. High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *J. Chem. Phys.* **1954**, *22*, 1420–1426.

[153] Jarzynski, C. A Nonequilibrium Equality for Free Energy Differences. *Phys. Rev. Lett.* **1997**, *78*, 2690–2693.

[154] Jarzynski, C. Equilibrium Free Energy Differences from Nonequilibrium Measurements: A Master Equation Approach. *Phys. Rev. E* **1997**, *56*, 5018–5035.

[155] Crooks, G. E. Nonequilibrium Measurements of Free Energy Differences for Microscopically Reversible Markovian Systems. *J. Stat. Phys.* **1998**, *90*, 1481–1487.

[156] Crooks, G. E. Entropy Production Fluctuation Theorem and the Nonequilibrium Work Relation for Free Energy Differences. *Phys. Rev. E* **1999**, *60*, 2721–2726.

[157] Lee, M. S.; Olson, M. A. Calculation of Absolute Protein-Ligand Binding Affinity Using Path and Endpoint Approaches. *Biophys. J.* **2006**, *90*, 864–877.

[158] Deng, Y.; Roux, B. Computations of Standard Binding Free Energies with Molecular Dynamics Simulations. *J. Phys. Chem. B* **2009**, *113*, 2234–2246.

[159] Christ, C. D.; Mark, A. E.; van Gunsteren, W. F. Basic Ingredients of Free Energy Calculations: A Review. *J. Comput. Chem.* **2009**, *31*, 1569–1582.

[160] Pohorille, A.; Jarzynski, C.; Chipot, C. Good Practices in Free-Energy Calculations. *J. Phys. Chem. B* **2010**, *114*, 10235–10253.

[161] Gumbart, J. C.; Roux, B.; Chipot, C. Standard Binding Free Energies from Computer Simulations: What Is the Best Strategy? *J. Chem. Theory Comput.* **2013**, *9*, 794–802.

[162] Gapsys, V.; Michielssens, S.; Peters, J. H.; de Groot, B. L.; Leonov, H. In *Molecular Modeling of Proteins;* Kukol, A., Ed.; Methods in Molecular Biology; Springer Science+Business Media: New York, 2015; Vol. 1215, pp 173–209.

[163] Doudou, S.; Burton, N. A.; Henchman, R. H. Standard Free Energy of Binding from a One-Dimensional Potential of Mean Force. *J. Chem. Theory Comput.* **2009**, *5*, 909–918.

[164] Torrie, G. M.; Valleau, J. P. Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling. *J. Comput. Phys.* **1977**, *23*, 187–199.

[165] Roux, B. The Calculation of the Potential of Mean Force Using Computer Simulations. *Comput. Phys. Commun.* **1995**, *91*, 275–282.

[166] Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. The Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules. I. The Method. *J. Comput. Chem.* **1992**, *13*, 1011–1021.

[167] Fodor, E.; Mack, J. W.; Maeng, J.-S.; Ju, J.-H.; Lee, H. S.; Gruschus, J. M.; Ferretti, J. A.; Ginsburg, A. Cardiac-Specific Nkx2.5 Homeodomain: Conformational Stability and Specific DNA Binding of Nkx2.5(C56S). *Biochemistry* **2005**, *44*, 12480–12490.

[168] Pradhan, L.; Genis, C.; Scone, P.; Weinberg, E. O.; Kasahara, H.; Nam, H.-J. Crystal Structure of the Human NKX2.5 Homeodomain in Complex with DNA Target. *Biochemistry* **2012**, *51*, 6312–6319.

[169] Hub, J. S.; de Groot, B. L.; van der Spoel, D. g_wham—A Free Weighted Histogram Analysis Implementation Including Robust Error and Autocorrelation Estimates. *J. Chem. Theory Comput.* **2010**, *6*, 3713–3720.

[170] Gonzalez, M.; Weiler, S.; Ferretti, J. A.; Ginsburg, A. The vnd/NK-2 Homeodomain: Thermodynamics of Reversible Unfolding and DNA Binding for Wild-Type and with Residue Replacements H52R and H52R/T56W in Helix III. *Biochemistry* **2001**, *40*, 4923–4931.

[171] Dragan, A. I.; Li, Z.; Makeyeva, E. N.; Milgotina, E. I.; Liu, Y.; Crane-Robinson, C.; Privalov, P. L. Forces Driving the Binding of Homeodomains to DNA. *Biochemistry* **2006**, *45*, 141–151.

[172] Privalov, P. L.; Dragan, A. I.; Crane-Robinson, C.; Breslauer, K. J.; Remeta, D. P.; Minetti, C. A. S. A. What Drives Proteins into the Major or Minor Grooves of DNA? *J. Mol. Biol.* **2007**, *365*, 1–9.

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

[174] Goette, M.; Grubmüller, H. Accuracy and Convergence of Free Energy Differences Calculated from Nonequilibrium Switching Processes. *J. Comput. Chem.* **2009**, *30*, 447–456.

[175] Rauscher, S.; Gapsys, V.; Gajda, M. J.; Zweckstetter, M.; de Groot, B. L.; Grubmüller, H. Structural Ensembles of Intrinsically Disordered Proteins Depend Strongly on Force Field: A Comparison to Experiment. *J. Chem. Theory Comput.* **2015**, *11*, 5513–5524.

[176] Do, P.-C.; Lee, E. H.; Le, L. Steered Molecular Dynamics Simulation in Rational Drug Design. *J. Chem. Inf. Model.* **2018**, *58*, 1473–1482.

[177] The UniProt Consortium. UniProt: The Universal Protein Knowledgebase. *Nucleic Acids Res.* **2017**, *45*, D158–D169.

[178] Martin, L. C.; Gloor, G. B.; Dunn, S. D.; Wahl, L. M. Using Information Theory to Search for Co-Evolving Residues in Proteins. *Bioinformatics* **2005**, *21*, 4116–4124.

[179] Finn, R. D.; Clements, J.; Arndt, W.; Miller, B. L.; Wheeler, T. J.; Schreiber, F.; Bateman, A.; Eddy, S. R. HMMER Web Server: 2015 Update. *Nucleic Acids Res.* **2015**, *43*, W30–W38.

[180] Eddy, S. R. Profile Hidden Markov Models. *Bioinformatics* **1998**, *14*, 755–763.

[181] Wheeler, T. J.; Clements, J.; Finn, R. D. Skylign: A Tool for Creating Informative, Interactive Logos Representing Sequence Alignments and Profile Hidden Markov Models. *BMC Bioinformatics* **2014**, *15*, 7.

[182] Sehnal, D.; Deshpande, M.; Vařeková Svobodová, R.; Mir, S.; Berka, K.; Midlik, A.; Pravda, L.; Velankar, S.; Koča, J. LiteMol Suite: Interactive Web-Based Visualization of Large-Scale Macromolecular Structure Data. *Nat. Methods* **2017**, *14*, 1121–1122.

List of abbreviations

3D	three-dimensional
BFE	binding free energy
bp	base pair
BSSE	basis set superposition error
CBS	complete basis set
CCSD(T)	coupled clusters with iterative single and double excitations and perturbative triple excitations
CFT	Crooks fluctuation theorem
CGI	Crooks Gaussian intersection
ChIP–seq	chromatin immunoprecipitation followed by sequencing
CR	cluster representative
DBD	deoxyribonucleic acid-binding domain
DBP	deoxyribonucleic acid-binding protein
DFT	density functional theory
DNA	deoxyribonucleic acid
dNMP	deoxyribonucleoside monophosphate
dsDNA	double-stranded deoxyribonucleic acid
FED	free energy difference
FF	force field
HB	hydrogen bond
HD	homeodomain
HDC	homeodomain core
HMM	hidden Markov model
HT	high-throughput
HTH	helix-turn-helix
IC	information content
IE	interaction energy
JE	Jarzynski equality
MD	molecular dynamics
MI	mutual information
MM	molecular mechanics
MP2	second-order Møller–Plesset perturbation theory
mRNA	messenger ribonucleic acid
MSA	multiple sequence alignment
NMR	nuclear magnetic resonance
NTP	N-terminal peptide
PBM	protein-binding microarray
PDB	Protein Data Bank
PMF	potential of mean force
QM	quantum mechanics
RCSB PDB	Research Collaboratory for Structural Bioinformatics Protein Data Bank
RMSD	root-mean-square deviation
RNA	ribonucleic acid
rRNA	ribosomal ribonucleic acid

TEM	transmission electron microscopy
TF	transcription factor
US	umbrella sampling
WFT	wave function theory
WHAM	weighted histogram analysis method
wwPDB	Worldwide Protein Data Bank

List of publications

Jakubec, D.; Hostaš, J.; Laskowski, R. A.; Hobza, P.; Vondrášek, J. Large-Scale Quantitative Assessment of Binding Preferences in Protein–Nucleic Acid Complexes. *J. Chem. Theory Comput.* **2015**, *11*, 1939–1948.

Hostaš, J.; Jakubec, D.; Laskowski, R. A.; Gnanasekaran, R.; Řezáč, J.; Vondrášek, J.; Hobza, P. Representative Amino Acid Side-Chain Interactions in Protein–DNA Complexes: A Comparison of Highly Accurate Correlated *ab Initio* Quantum Mechanical Calculations and Efficient Approaches for Applications to Large Systems. *J. Chem. Theory Comput.* **2015**, *11*, 4086–4092.

Jakubec, D.; Laskowski, R. A.; Vondrášek, J. Sequence-Specific Recognition of DNA by Proteins: Binding Motifs Discovered Using a Novel Statistical/ Computational Analysis. *PLoS One* **2016**, *11*, e0158704.

Stasyuk, O. A.; Jakubec, D.; Vondrášek, J.; Hobza, P. Noncovalent Interactions in Specific Recognition Motifs of Protein–DNA Complexes. *J. Chem. Theory Comput.* **2017**, *13*, 877–885.

Galgonek, J.; Vymětal, J.; Jakubec, D.; Vondrášek, J. Amino Acid Interaction (INTAA) Web Server. *Nucleic Acids Res.* **2017**, *45*, W388–W392.

Jakubec, D.; Kratochvíl, M.; Vymětal, J.; Vondrášek, J. Widespread Evolutionary Crosstalk among Protein Domains in the Context of Multi-Domain Proteins. *PLoS One* **2018**, *13*, e0203085.

Jakubec, D.; Vondrášek, J.; Finn, R. D. 3DPatch: Fast 3D Structure Visualization with Residue Conservation. *Bioinformatics* **2019**, *35*, 332–334.

Jakubec, D.; Vondrášek, J. Can All-Atom Molecular Dynamics Simulations Quantitatively Describe Homeodomain–DNA Binding Equilibria? *J. Chem. Theory Comput.* **2019**, *15*, 2635–2648.

Jakubec, D.; Vondrášek, J. Efficient Estimation of Absolute Binding Free Energy for a Homeodomain–DNA Complex from Nonequilibrium Pulling Simulations. *J. Chem. Theory Comput.* **2020**, *16*, 2034–2041.

Faltejsková, K.; Jakubec, D.; Vondrášek, J. Hydrophobic Amino Acids as Universal Elements of Protein-Induced DNA Structure Deformation. *Int. J. Mol. Sci.* **2020**, *21*, 3986.

A. Figures

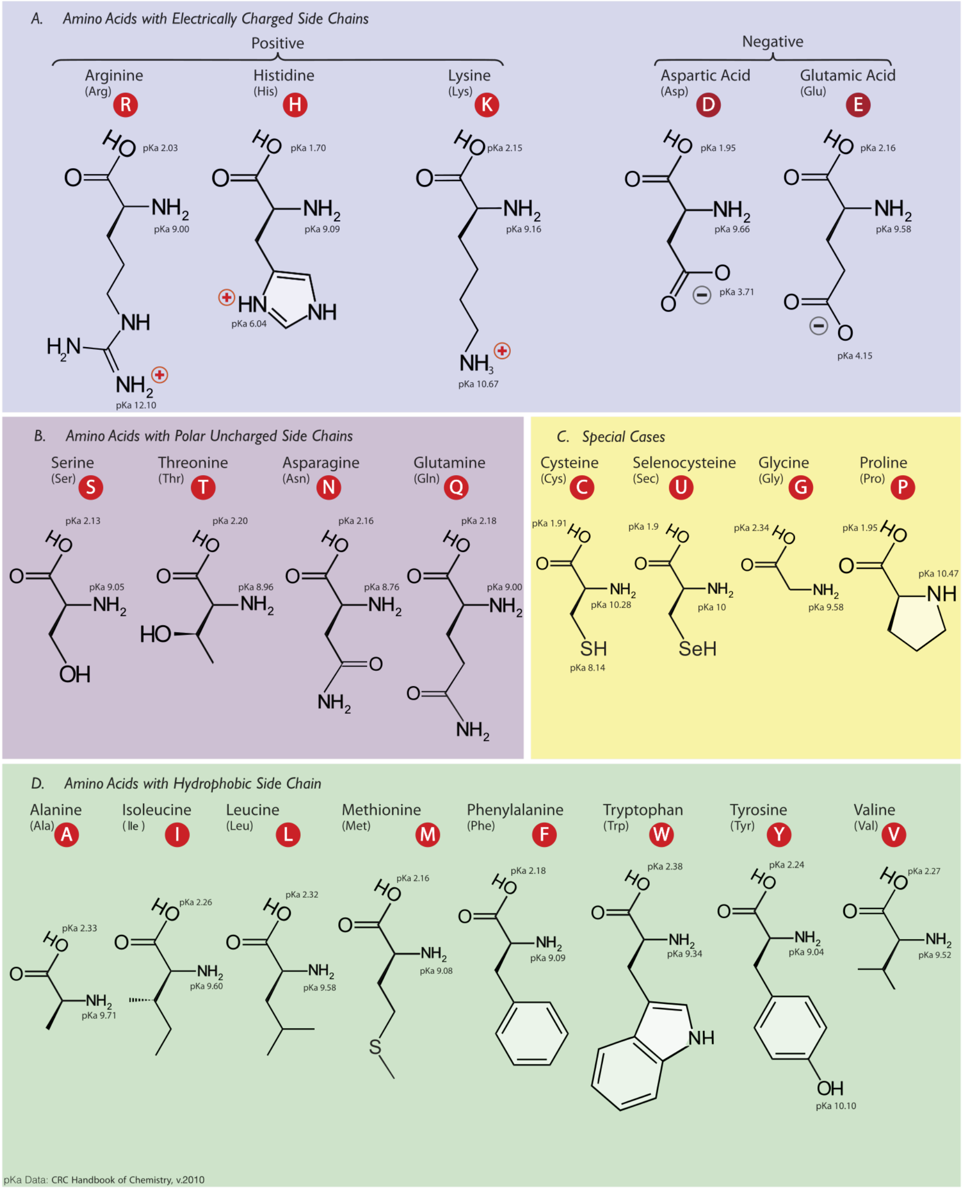


Figure A.1: The structures and properties of the proteinogenic amino acids. Distributed under the terms of the Creative Commons Attribution-ShareAlike 3.0 Unported license (`https://creativecommons.org/licenses/by-sa/3.0/legalcode`). Original author: Dan Cojocari, Department of Medical Biophysics, Faculty of Medicine, University of Toronto, 2010. Source: `https://en.wikipedia.org/wiki/File:Molecular_structures_of_the_21_proteinogenic_amino_acids.svg`. Modified: removed fine print.

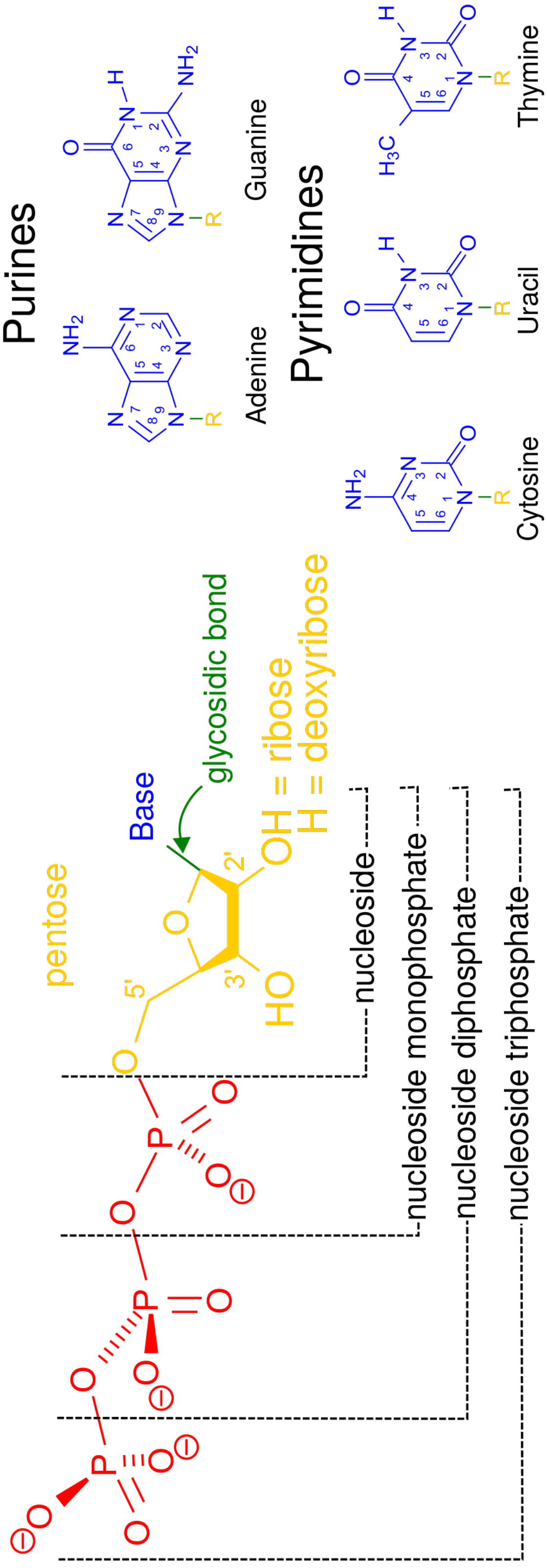


Figure A.2: The structures of the nucleotides forming DNA and RNA. Source: `https://en.wikipedia.org/wiki/File:Nucleotides_1.svg`.

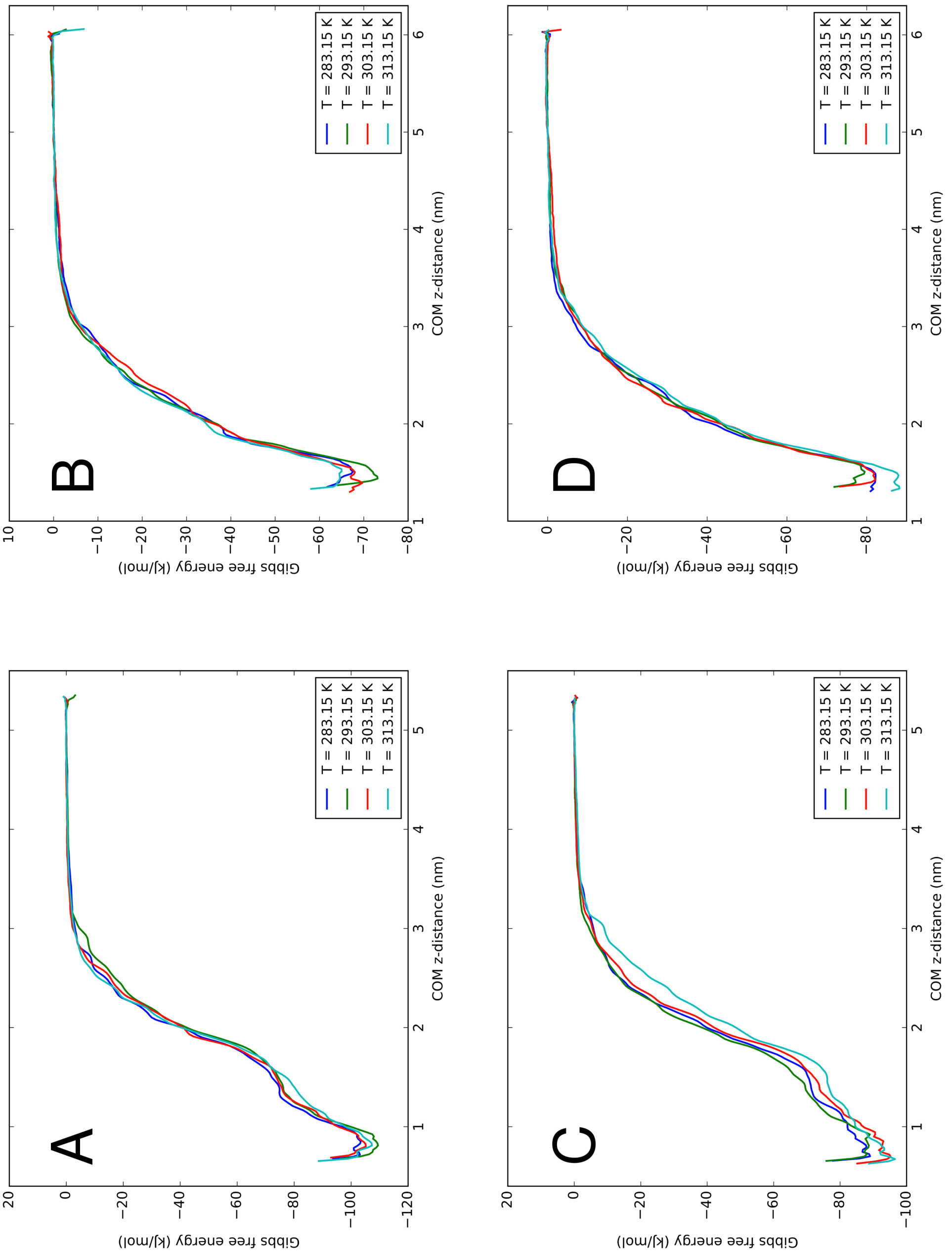


Figure A.3: The PMFs (PMF$_{\mathrm{bs}}$s) for the binding of the NTP (**A**, **C**) and HDC (**B**, **D**) fragments of the protein to the DNA. **A** and **B** calculated using the CHARMM FFs; **C** and **D** calculated using the Amber FFs. From Jakubec and Vondrášek (2019).

B. Attachments

The following attachments are available online as a single ZIP file:

- `att_1_jakubec_JCTC_2015.pdf`
- `att_2_hostaš_JCTC_2015.pdf`
- `att_3_jakubec_PLOS_ONE_2016.pdf`
- `att_4_stasyuk_JCTC_2017.pdf`
- `att_5_galgonek_NAR_2017.pdf`
- `att_6_jakubec_PLOS_ONE_2018.pdf`
- `att_7_jakubec_Bioinformatics_2019.pdf`
- `att_8_jakubec_JCTC_2019.pdf`
- `att_9_jakubec_JCTC_2020.pdf`
- `att_10_faltejsková_IJMS_2020.pdf`

These correspond (in the stated order) to the PDF versions of the publications listed in the List of publications (Page 69). The attachments are nonpublic and intended for the use by reviewers and for archiving purposes only. Copyright issues or technical restrictions prohibit their inclusion in the electronic form of this thesis; nevertheless, in the printed form of the thesis, they are included on the following pages.